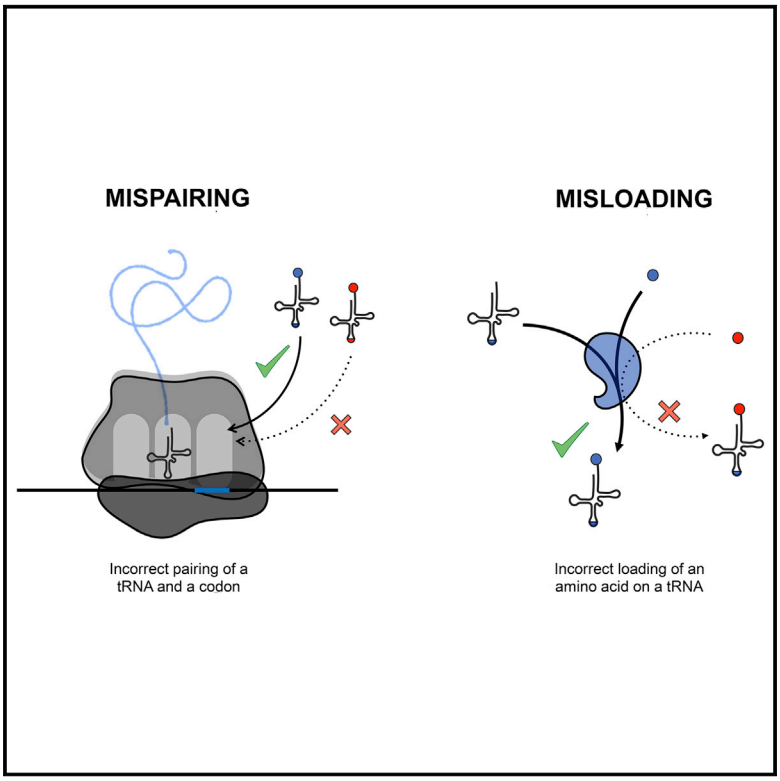
Article

Molecular Cell

Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity

Graphical Abstract



Authors

Ernest Mordret, Orna Dahan, Omer Asraf, ..., Tamar Geiger, Ariel B. Lindner, Yitzhak Pilpel

Correspondence

geiger@tauex.tau.ac.il (T.G.), ariel.lindner@inserm.fr (A.B.L.), pilpel@weizmann.ac.il (Y.P.)

In Brief

Errors in translation are common, occurring almost once per protein. Mordret et al. have developed a methodology to detect and quantify translation errors and applied it to bacteria and yeast. Errors appear to be programmed, reduced in highly expressed proteins and conserved protein sites. Codon identity and ribosome speed participate in governing a protein translation fidelity code.

Highlights

- A new methodology to detect and quantify most translation errors in proteomes
- Most amino acid substitutions result from mis-pairing between codons and anti-codons
- Proteins' error rates are reduced at conserved and highly expressed proteins
- Translation speed is negatively correlated with error rates

Mordret et al., 2019, Molecular Cell *75*, 427–441
August 8, 2019
https://doi.org/10.1016/j.molcel.2019.06.041

CellPress

Molecular Cell
Article

Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity

Ernest Mordret,[1,2,3,6] Orna Dahan,[1,6] Omer Asraf,[1] Roni Rak,[1] Avia Yehonadav,[1] Georgina D. Barnabas,[4] Jürgen Cox,[5] Tamar Geiger,[4,*] Ariel B. Lindner,[2,3,*] and Yitzhak Pilpel[1,7,*]
[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel
[2]Institut National de la Santé et de la Recherche Médicale, U1001
[3]CRI, Université Paris Descartes, Sorbonne Paris Cité, Paris, France
[4]Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel
[5]Computational Systems Biochemistry, Max Planck Institute for Biochemistry, Martinsried, Germany
[6]These authors contributed equally
[7]Lead Contact
*Correspondence: geiger@tauex.tau.ac.il (T.G.), ariel.lindner@inserm.fr (A.B.L.), pilpel@weizmann.ac.il (Y.P.)
https://doi.org/10.1016/j.molcel.2019.06.041

SUMMARY

The translation machinery and the genes it decodes co-evolved to achieve production throughput and accuracy. Nonetheless, translation errors are frequent, and they affect physiology and protein evolution. Mapping translation errors in proteomes and understanding their causes is hindered by lack of a proteome-wide experimental methodology. We present the first methodology for systematic detection and quantification of errors in entire proteomes. Following proteome mass spectrometry, we identify, in *E. coli* and yeast, peptides whose mass indicates specific amino acid substitutions. Most substitutions result from codon-anticodon mispairing. Errors occur at sites that evolve rapidly and that minimally affect energetic stability, indicating selection for high translation fidelity. Ribosome density data show that errors occur at sites where ribosome velocity is higher, demonstrating a trade-off between speed and accuracy. Treating bacteria with an aminoglycoside antibiotic or deprivation of specific amino acids resulted in particular patterns of errors. These results reveal a mechanistic and evolutionary basis for translation fidelity.

INTRODUCTION

Genetic information propagation is subject to errors in DNA replication, transcription, and translation. DNA replication typically manifests the highest fidelity, featuring a mutation rate on the order of 10^{-9}–10^{-10} per nucleotide per genome doubling (Lee et al., 2012; Zhu et al., 2014). "Phenotypic errors"—i.e., errors in transcription and translation—occur at a considerably higher rate. The bacterial RNA polymerase misincorporation rate is 10^{-4}–10^{-5} per nucleotide (Traverse and Ochman, 2016). Amino acid substitutions rates are higher, estimated to be 10^{-4}–10^{-3} per incorporated residue (Kramer and Farabaugh, 2007).

Translation errors can result either from the charging of a tRNA with the wrong, "non-cognate" amino acid (synthetase error or "mischarging") or from the ribosome failing to discriminate against imperfect codon-anticodon complexes in its A-site (ribosome error or "mispairing"). The accuracy of both processes is amplified by kinetic proofreading (Hopfield, 1974; Ninio, 1975), a general mechanism that, through an irreversible energy-consuming step, allows discrimination levels that are inaccessible at thermodynamic equilibrium. Theoretical models have demonstrated that proofreading is subject to an inherent trade-off between speed, accuracy, and energetic cost (Wohlgemuth et al., 2011; Chen et al., 2016; Banerjee et al., 2017).

The resources cells invest to ensure that proteins function properly indicate that errors during protein synthesis strongly affect fitness. Proteins that are translated with errors can misfold, aggregate, be engaged with wrong interactions, and saturate the protein quality control machinery, resulting in proteotoxic stress (Drummond and Wilke, 2009) and diseases; e.g., in aging (Lee et al., 2006; Lindner and Demarez, 2009; Kapur and Ackerman, 2018). Conversely, some errors might be advantageous. Moderate levels of methionine misacylation on non-methionine tRNAs can provide a fitness advantage in oxidative stress in bacteria and humans (Netzer et al., 2009; Jones et al., 2011; Wiltrout et al., 2012). High error rates allow a parasitic yeast to increase its adherence and evasion of immunity (Miranda et al., 2013). Mistranslation is beneficial in response to environmental stresses because it disseminates phenotypic viability in surface proteins (Miranda et al., 2013). On an evolutionary timescale, phenotypic errors might open evolutionary paths otherwise precluded by epistatic interactions (Whitehead et al., 2008), and they may facilitate purging of deleterious mutations (Bratulic et al., 2017). Natural selection constrains the identity of codons at evolutionarily conserved positions, suggesting that evolution favors more accurate codons at these sites (Drummond and Wilke, 2008).

These profound biological implications of translation fidelity reveal the need to measure translation accuracy per site per

protein throughout proteomes under diverse life conditions. However, although the rates of DNA mutation and RNA polymerase errors are now quantifiable thanks to sequencing, errors in protein translation have remained elusive at the proteome-wide level. An early effort by Edelmann and Gallant (1977), who directed at cysteine misincorporation in *E. coli*'s flagellin, suggested an error occurring every 10,000 amino acids. Luminescent reporters allowed quantitative tracking of mistranslation at defined positions within these reporter genes (Kramer and Farabaugh, 2007; Kramer et al., 2010). These methods have highlighted the importance of codon-anticodon recognition and tRNA competition as determinants of these error rates.

However, current methods only estimate an averaged error rate for entire proteomes, and, hence, major questions remain open. Accurate and broad measurements of error rate per protein per position would reveal whether error rates change within and between proteins. This would shed light on the relative contribution of tRNA synthetases and ribosomes to translation errors. Detecting and quantifying amino acid substitutions across the proteome may reveal evolutionary constraints imposed by translation infidelity and whether and how overall translation fidelity responds to environmental and genetic perturbations and allow us to find out whether organisms modulate error levels locally.

Mass spectrometry (MS)-based proteomics enable routine, high-throughput characterization of proteomes and common post-translational modifications (PTMs) and has been described as an upcoming tool for the study of protein mistranslation (Drummond and Wilke, 2009). Proteomics have been harnessed to detect various substitutions from several purified recombinant proteins (Zhang et al., 2013) and to track the incorporation of norvaline at leucine positions in *E. coli* mutants (Cvetesic et al., 2016). However, MS has yet to be harnessed for the systematic study of amino acid substitutions on a proteome-wide scale. Such a study has so far been hindered by the low abundance of substitutions compared with other natural and post-translational protein modifications and a much larger search space.

Here we performed a deep proteomics analysis to detect and quantify translation errors in *E. coli* under normal and perturbed conditions, repurposing the MaxQuant algorithm (Cox and Mann, 2008) from its typical PTM analysis to identify mass shifts in peptides that result from amino acid substitutions. We then validated these identifications using a set of independent analyses that included a shift in high-pressure liquid chromatography (HPLC) retention time because of a change in hydrophobicity of the encoded amino acid. Our dataset of translation error events allowed us to start unraveling the mechanistic basis and evolutionarily selective forces that shape translation errors and fidelity. We found that most errors result from mispairing between codons and near-cognate tRNAs, mostly within the A site of the ribosome. We derived the amino acid error spectrum of each codon in the genetic code to deduce patterns of codon mispairing at each of its three codon positions. We found that the aminoglycoside drug affects the error pattern and revealed that it mainly causes errors because of mispairing at the third codon position. By depriving cells of particular amino acids, we observed misincorporation of specific amino acids instead of the depleted ones. In addition, comparing the substitutions spectra between yeast and bacteria revealed a similar error pattern, probably reflecting shared chemical constraints. Finally, we found that errors are allowed to occur more frequently at evolutionarily rapidly evolving sites, at protein structure sites in which mutations are more tolerated energetically, in lowly expressed proteins, and in positions along mRNAs in which the ribosome progresses more rapidly.

RESULTS

A Pipeline to Confidently Identify Amino Acid Substitutions in a Proteome

MS enables identification of peptides within complex samples. From a computational point of view, amino acid substitutions can be regarded as a particular case of PTMs, many of which are now routinely studied at the proteome level. However, the standard database search algorithm is not suitable for large-scale detection of substitutions. Assuming peptides of an average length of 10 amino acids, there would be on the order of 200 times more singly modified than canonical peptides to search for, leading to impractical search times and a considerable loss of statistical power.

Blind modification searches (Tsur et al., 2005; Savitski et al., 2006; Na et al., 2012) offer a way to identify modified peptides without requiring the user to enter a list of predefined modifications. They rely on the observation that modified peptides are usually less abundant than their unmodified counterparts and are therefore only likely to be present in the sample if their corresponding unmodified peptide has already been detected. We thus adapted MaxQuant (Cox and Mann, 2008), developed for the analysis of PTMs, to identify mistranslated peptides using its "dependent peptide search" algorithm (Figure 1A). Dependent peptides are defined as peptides that show mass shifts in comparison with the unmodified, genome-encoded "base peptides" (Figure 1B). We applied a series of filters to the list of identified dependent peptides to stringently remove known PTMs and known artifacts and conservatively retain only amino acid substitutions (STAR Methods). A full dataset of all identified substitution errors can be found in Table S1 and Table S2.

Most of the High-Quality Hits Are *Bona Fide* Amino Acid Substitutions

Aiming to identify translation errors in *E. coli*, we generated deep proteomics profiles that would allow detection of the rare mistranslation events. In total, we generated error maps of 10 samples, each in two replicates (STAR Methods). First, we grew wild-type *E. coli* cells (MG1655) in defined medium (3-(N-morpholino)propanesulfonic acid [MOPS] complete, 37°C) in biological duplicates, harvested cells at three time points—two time points during the exponential phase (t1 optical density [OD], ~0.5; t2 OD, ~1.5) and one time point during the stationary phase (t3 OD, ~2.3)—and used our pipeline to detect amino acid substitutions. Mass spectrometers sample, in priority, the most abundant peptides for MS2 fragmentation. Because we expect translation errors to be present at a much lower level than their corresponding error-free peptides (Figure S1), we fractionated our samples to reduce their complexity and increase the chances of sampling low-abundance peptides. We separated

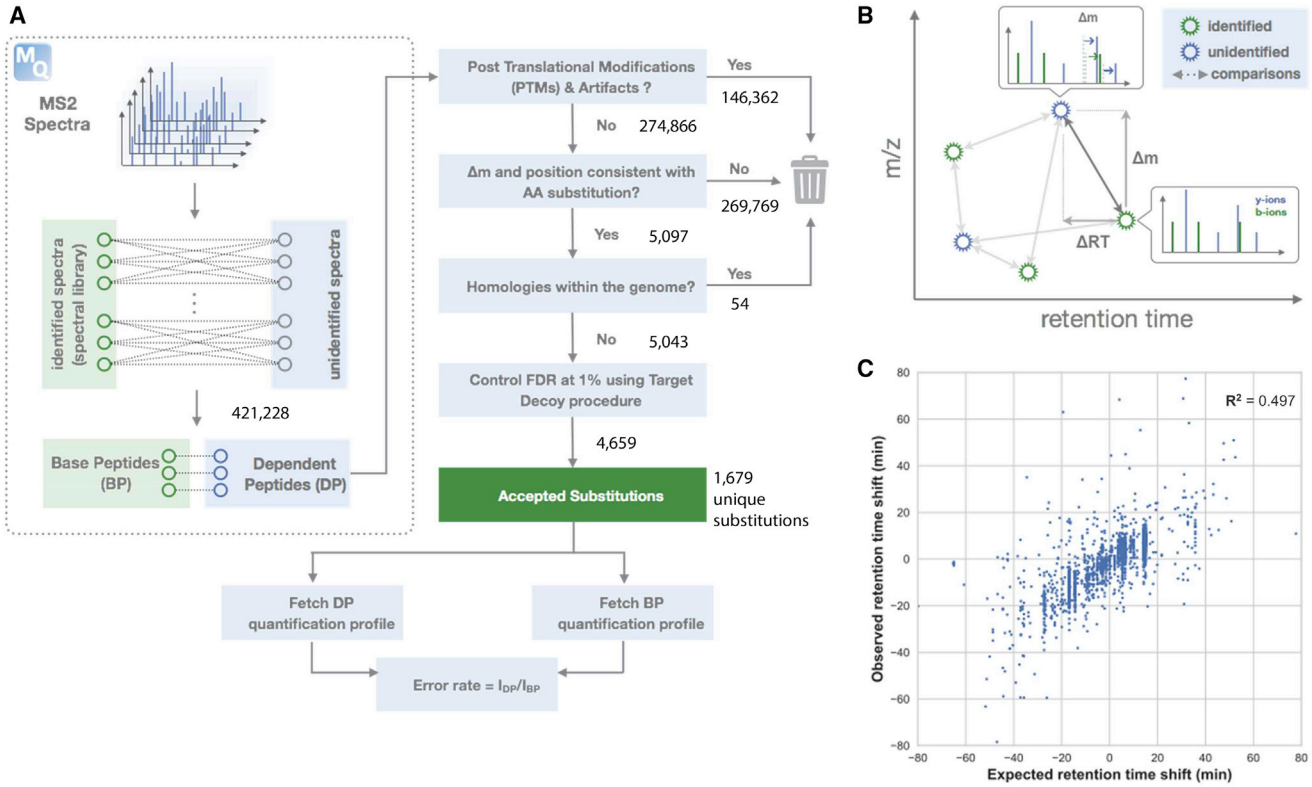


Figure 1. A Computational Pipeline to Confidently Identify Amino Acid Substitutions from Mass Spectrometry Data

(A) Overview of the pipeline. For a detailed description of the different steps, see STAR Methods. Numbers indicate the number of peptides identified in each step of the pipeline.

(B) The MaxQuant-dependent peptide search performs exhaustive pairing of unidentified spectra to a spectral library derived from the identified spectra. For each pair of (identified, unidentified) spectra of the same charge z and found in the same fraction, the algorithm first computes the mass difference $\Delta m = m_{unidentified} - m_{identified}$. It simulates, *in silico* and sequentially, the addition of a single moiety of mass Δm at any position in the identified peptide and generates the corresponding theoretical spectrum for the modified peptide. These spectra are then compared with the experimental spectrum using MaxQuant Andromeda's score formula. The pair with the highest score is retained, and the significance of the match is assessed using a target-decoy false discovery rate (FDR) procedure.

(C) The observed retention time shift induced by our set of substitutions is accurately predicted by a simple sequence-based retention time model.

proteins into a high-solubility and a low-solubility fraction (Khan et al., 2011) that could be enriched with error products and further fractionated each cellular fraction into five chemical fractions by strong cation exchange (SCX) chromatography.

Given mass differences detected between a pair of base and dependent peptides, we must first establish that they represent *bona fide* amino acid substitutions and not methodological artifacts and examine the possibility that they may represent PTMs with exactly the same mass difference. We took advantage of the fact that many amino acid substitutions result in a change of peptide hydrophobicity and that they should hence result in retention time shifts during liquid chromatography (LC). The retention time of a peptide can be predicted with high accuracy ($R^2 > 0.9$) as the sum of the hydrophobicity coefficients of its amino acids (Moruz and Käll, 2017). Therefore, the predicted HPLC retention time of the substituted amino acid can be computed and compared with the observed retention time recorded for the substituted peptide. We trained a retention time prediction algorithm (Goloborodko et al., 2013) on a list of identified unmodified peptides and used it to generate an expectation of the retention time shift induced by the substitutions (Figure S2). We compared this expectation with the observed retention time shift for each of the detected substitutions (Figure 1C), revealing a good agreement, indicating that we typically correctly identify amino acid replacements. Because MS2 spectra are systematically recorded for highly abundant parent ions, our sampling strategy is biased toward detection of substitutions originating from highly abundant peptides (Figure S1). We devised a procedure (Figure S5) that estimated that approximately 8% of the substituted peptides detected by the pipeline are likely un-annotated PTMs or artifacts.

We define a substitution as a combination of a position in a protein, characterized by an "origin" amino acid, its associated codon, and a "destination" amino acid. We then divide substitutions into two sets: a near cognate error (NeCE) is defined to be the case when the error-bearing codon of the origin amino acid matches (any) two of three bases of at least one of the codons of the destination amino acid, and a non-cognate error (NoCE) is defined to be the case when none of the codons of the wrongly incorporated amino acid match, with one mismatch, the codon of the original amino acid (cases in which the original codon matched one of the destination codons by two nucleotides but

the destination codon does not have a perfectly matched cognate tRNA in the organisms' genome were also considered as NoCE). For example, a substitution from serine encoded by AGC into an asparagine is considered a NeCE because one of the codons encoding asparagine, AAC, represents a single mismatch (A instead of G at the second position); in contrast, substitution of a cysteine encoded by a UGU into an alanine must be a NoCE. The structure of the genetic code dictates that, of all detectable codon-to-amino acid substitutions, 30% are expected to be of the NeCE type. In stark contrast, in our data of detected substitutions, 64% of the unique substitutions with the core *E. coli* dataset (wild-type, MOPS complete) are classified as NeCE. Such enrichment for NeCE serves as an indication that we inspect genuine amino acid substitutions.

The ribosome uses small differences in free energy between correct and incorrect codon-anticodon matches to select tRNAs and has been shown to generate NeCEs at much higher rates than NoCEs (Kramer and Farabaugh, 2007). During loading of an amino acid by an aminoacyl-tRNA synthetase, an error can stem from the choice of an incorrect tRNA or loading of an incorrect amino acid. Because the majority of synthetases assess the identity of the tRNA by probing the bases of the anticodon, this first mechanism of error is also likely to generate mostly NeCEs. However, if the error results from loading of the wrong amino acid, then there should be *a priori* no preference for NeCEs over NoCEs. We will consider NoCEs as more likely to stem from a synthetase error, and we will examine below the notion that the majority of NeCEs indeed represent mRNA-tRNA mispairing events.

The *E. coli* Amino Acid Substitution Landscape

We generated 61 × 20 codon-to-amino acid matrices that depict the prevalence of each type of amino acid misincorporation (Figures 2A and S3). The i,j^{th} entry in the matrix depicts the numbers of unique peptides in the proteome in which amino acid j was found to be misincorporated instead of an amino acid encoded by codon i. Because leucine and isoleucine share the same mass, we cannot distinguish them as error destinations. Furthermore, substitution types that transform a codon into its cognate amino acid or that involve a stop codon or substitutions that cannot be detected using our method because they represent a mass shift that corresponds precisely to the mass shift and specificity of a known PTM were discarded from subsequent analyses, and they were grayed out in this graphical depiction (STAR Methods). Even upon normalization of substitution counts to codon usage, the matrices only have minor changes (data not shown).

The substitution matrix is highly structured, and some substitutions appear to be much more prevalent than others. In particular, we observe that the codon used to encode an amino acid position determines error patterns at the corresponding site. This is nicely illustrated with substitutions from Gly to Asp and Glu. When Gly is encoded by the GGC codon, the frequent substitution destination is the near-cognate Asp (which can be encoded by GAC), whereas encoding Gly with GGA often results in substitution of Gly with Glu (presumably because of its near-cognate codon GAA). Similar additional cases were found for other amino acids (Figure 2A).

We calculated the observed error rate, estimated as the ratio of intensity between the dependent and base peptide, for abundantly detected substitutions at each site where it occurred. Because we measured the proteome at different time points during culture growth (Figure S4), we could follow the error rate for those substitutions as the culture grew. As an example, the SerAGC→Asn substitution was detected among a total of 81 different peptides across the proteome in untreated, MOPS-complete samples. Figure 2B summarizes the error rate estimations; each dot in the plot corresponds to one specific substitution on a particular proteomic position, and the error rate is on the x axis. Shown are the 10 most frequent substitutions types in the proteome. The majority of the substitutions that are observable in our dataset span the error rate range around 10^{-3}, with the most highly abundant substitutions showing slightly higher error rates. Because of the MS acquisition strategy, positions that feature a low error rate are less likely to be detected, which could lead to over-estimation of the actual error rates. We note that, for most substitution types, the error rates seem to consistently decline as cells progress from exponential to stationary phase; of the 10 most prevalent substitution types, only two follow the opposite trend, and both involve the common glycine GGC codon, perhaps reflecting an intracellular shortage of glycine. However, it is possible that this measured decline in error level may in part reflect a general potential decrease in total protein synthesis in cells as they enter the stationary phase, a situation that, when accompanied by degradation of erroneous proteins, would result in an apparent decline in error levels.

What are the possible explanations for these error patterns and preferences? One possibility is that imbalances in expression between a cognate and a near-cognate tRNA may lead to translation errors. To examine this possibility, we performed deep RNA sequencing of the bacterial tRNA pool at the same time points where we measured the proteome during culture growth. We could thus detect and quantify tRNA expression level changes and examine the ratio of expression (and, hence, potential imbalances) between pairs of cognate and near-cognate tRNAs. We examined, for all the substitutions shown in Figure 2B, whether dynamic changes in error rates can be explained by dynamic changes in tRNA level imbalances. We found that the patterns in the majority of these cases are consistent with the relative change in tRNA abundance. In particular, we see that, if the ratio of abundance between the cognate and near-cognate tRNA increases between two phases of the growth cycle (e.g., exponential phase to stationary phase), then the error rate between the corresponding amino acid tends to decrease, and, vice versa, when the ratio decreases, the error rate increases (Figure 2C). This indicates that some of the translation errors indeed occur as the cognate and near-cognate tRNAs compete on the ribosome's A site, in part based on their relative concentrations. However, because of a paucity of examples, this result cannot be declared statistically significant.

A Global Nucleotide Mispairing Pattern for Translation Errors

We further classified NeCEs based on the type of codon-anticodon mismatch within the codon and the nucleotide types they involved. We created 4 × 4 "confusion matrices" that depict the mispairing propensity of each nucleotide within the codon to mispair against each nucleotide in the anti-codon.

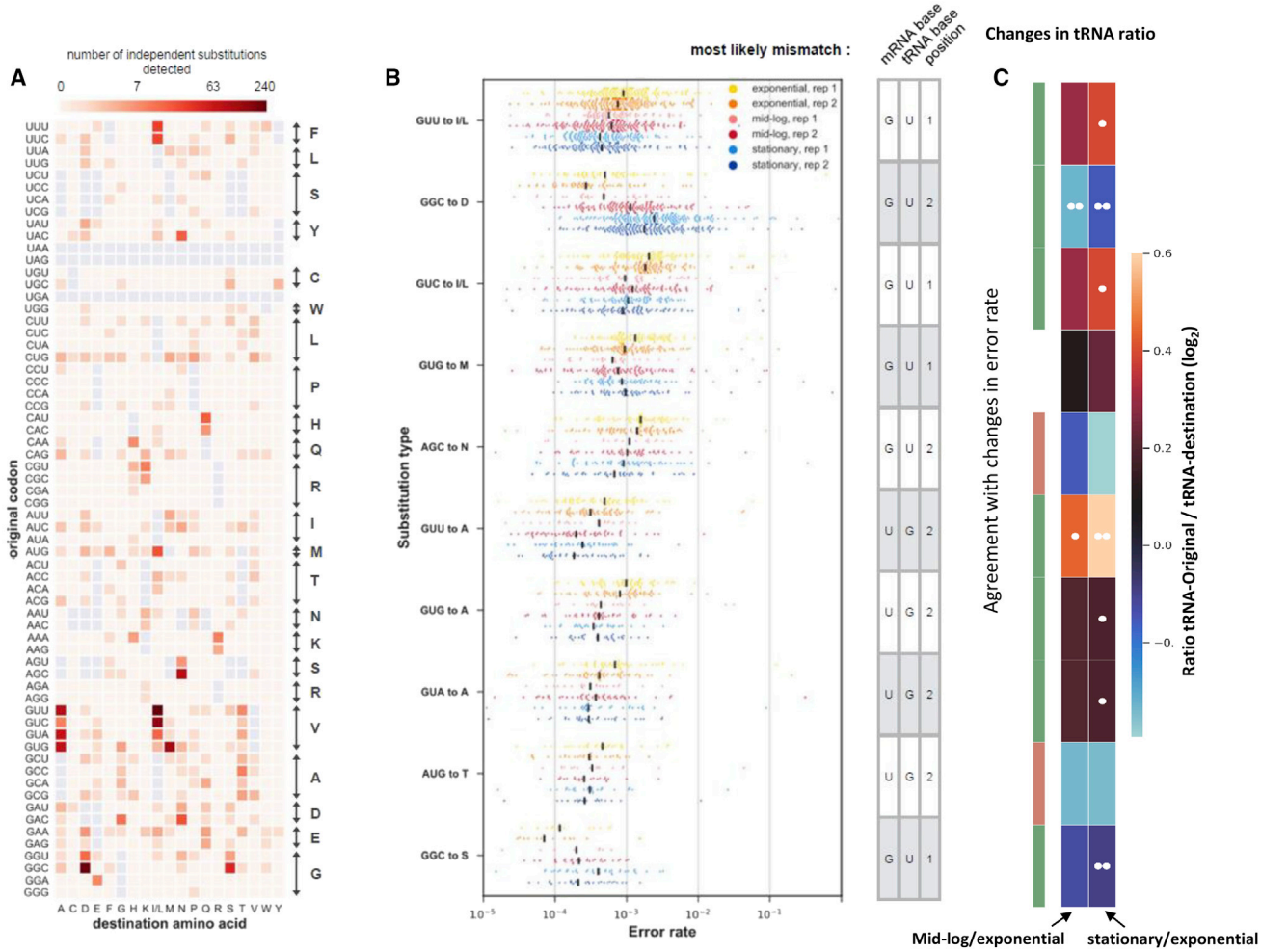


Figure 2. Overview of the Substitution Profile of *E. coli* in MOPS Complete Medium

(A) Matrix of substitution identifications. Each entry in the matrix represents the number of independent substitutions detected for the corresponding (original codon, destination amino acid) pair in the MOPS-complete dataset. The logarithmic color bar highlights the dynamic range of detection. Grey squares indicate substitutions from a codon to its cognate amino acid, substitutions from the stop codon, or substitutions undetectable via our method because they are indistinguishable from one of the PTMs or artifacts in the UniMod (http://www.unimod.org/) database. Substitutions to Leu and Ile are *a priori* undistinguishable and thus grouped together.

(B) Left panel: for each of the top 10 most frequently detected substitution types, we fetched the quantification profile of the dependent peptide and the base peptide. Each dot represents the ratio of intensities I_{DP}/I_{BP} for each of the samples when both peaks were detected and quantified. The black lines indicate the medians of the distributions. Colors indicate the growth phase of the culture in which samples were taken: exponential (t1 OD, ~0.5), mid-log (mid/late log t2 OD, ~1.5), and stationary (t3 OD, ~2.3). Right panel: we inferred the most likely mismatch for each of the substitution types, using a procedure described in the STAR Methods. This allows us to guess that the V-to-I/L substitutions are likely substitutions from Val to Ile, enabled by a G:U mismatch at the first position (in cases in which the corresponding tRNA was missing, we considered a wobble providing tRNA).

(C) tRNA imbalances might explain error patterns. Following tRNA deep sequencing during the culture growth cycle, we calculated, for each of these 10 most common error types, the ratio of expression between the cognate and near-cognate tRNA. The ratio of expression between mid-log to exponential time points is color-coded, and likewise between stationary to exponential time points (• $p < 0.1$, •• $p < 0.05$, $n = 3$). A green mark to the left of the color map indicates that the change in the ratio between the cognate to near-cognate tRNA expression correctly predicts a decline or an increase in the error rate of the corresponding codon-to-amino acid event, a red mark indicates lack of agreement, and the single white mark is a case in which the error pattern in (B) did not show a clear temporal trend.

Because each of the three codon positions may feature different mispairing patterns, we created three such matrices, one for each codon position (Figure 3A). Most of the substitutions were found to originate from a single mismatch type, and substitutions that could not be unambiguously traced back were assigned to the most likely mismatch using a greedy algorithm (STAR Methods). The most frequently observed substitution types involve mismatches between uracil and guanine in the first or the second position of the codon. Interestingly, this observation holds mainly for G:U mismatches in the first and second position of the codon, where the codon base is G and the anticodon base is U. Other common mismatches are G:G in the first position and U:G in the second position of the codon. The differences between the three matrices is

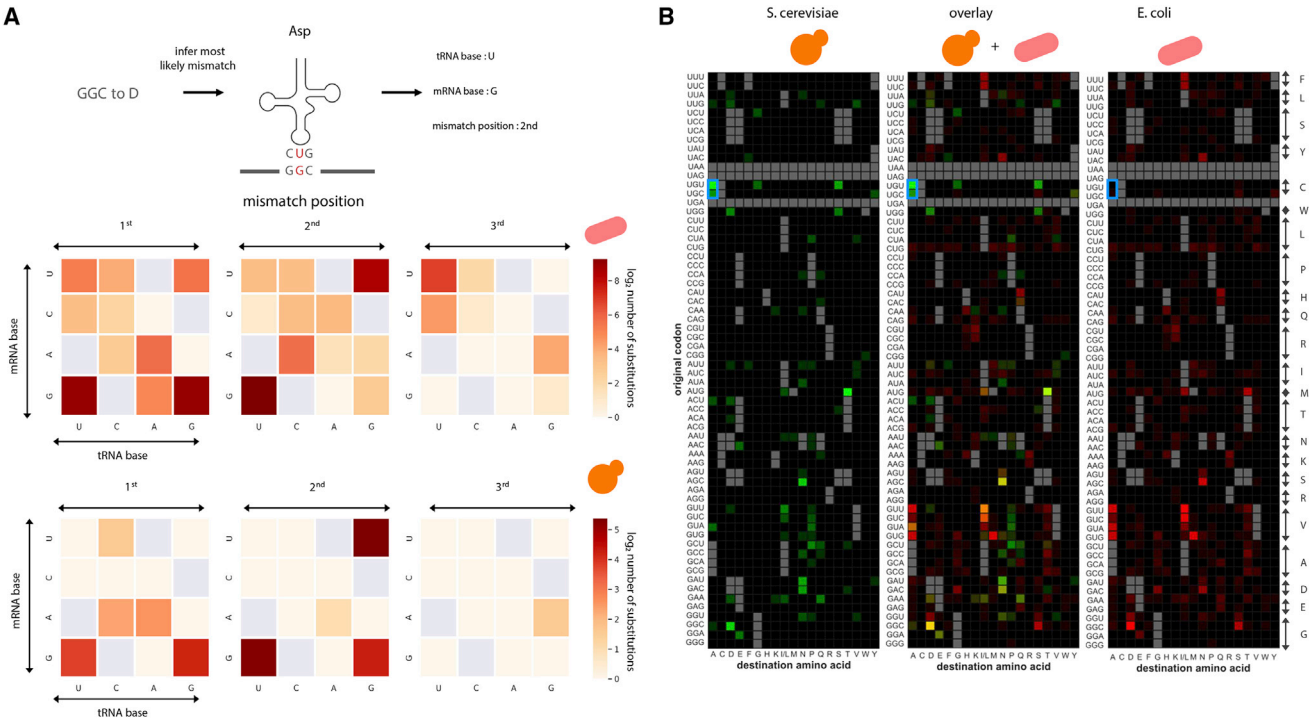


Figure 3. Error Spectra in *E. coli* and *S. cerevisiae* Reveal a Shared Signature

(A) Top panel: an example of a NeCE-originated substitution. Bottom panels: NeCEs are classified by the mismatch most likely to generate them. The shade intensity reflects the logarithm (base 2) of the number of peptides in which the corresponding mismatch was observed. Grey boxes are either correct base pairings or mismatches to which no substitutions could be unambiguously mapped. To measure similarity or dissimilarity between pairs of 4 × 4 matrices, we computed Pearson's correlations between (linearized forms) of pairs of matrices. We compared all three codon position-specific matrices within each species and also each pair of matrices at each codon position between the two species. Of all comparisons between matrix pairs, two were significantly correlated (hence demonstrating similar substitution patterns). The *E. coli* matrix at position 1 was similar to the *S. cerevisiae* matrix at position 1 ($r = 0.94$, $p = 6 \times 10^{-6}$), and the *E. coli* matrix at position 2 was similar to the *S. cerevisiae* matrix at position 2 ($r = 0\ 0.86$, $p = 3 \times 10^{-4}$). All other pairs of matrices were non-significantly similar. (B) The substitution identifications matrices of *S. cerevisiae* (green channel, left) and wild-type *E. coli* grown on MOPS complete (red channel, right) are compared and overlaid (center). The intensity of the color is proportional to the logarithm of the number of independent identifications, with one pseudo-count. Values are normalized by the highest entry in the matrix for each of the two organisms. The blue box highlights the recently described property of eukaryotic AlaRS to mischarge $tRNA^{Cys}$. The same statistical procedure as in (A) was conducted ($r = 0.4$, $p = 1 \times 10^{-40}$).

among the several indications gathered here that we observe translation errors and not transcription errors.

E. coli and *S. cerevisiae* Share Similar Error Profiles

We then wished to compare error profiles between bacteria and an eukaryote to examine whether errors are constrained by chemical or evolutionary necessities. We reanalyzed a previously published MS dataset of strong anion exchange (SAX) and SCX fractionated proteomes of *S. cerevisiae* grown under a single condition, a rich medium (30°C, yeast extract peptone dextrose (YPD)) (Kulak et al., 2014) using our pipeline (Figure 3B, left matrix). Similarly to the core *E. coli* dataset, the majority of the errors were classified as NeCEs (63%). Comparing the error spectrum between the two species (Figure 3B, center matrix), we observed a high overlap between the set of substitution types. For example, the most highly frequent substitution types (e.g., $Met_{AUG} \rightarrow Thr$, $Ser_{AGC} \rightarrow Asn$, $Val_{GUU} \rightarrow Ile/Leu$) are shared between the two species. Among the 34 substitution types observed more than once in the yeast dataset, 25 had been seen in the *E. coli* samples. Among the 25 substitution types also detected in *E. coli*, 16 were NeCEs. Conversely, among the remaining 9 substitution types that were not seen in the bacterial samples, only 4 were NeCEs. This observation reveals a universal error pattern for mistakes that are likely to occur within the ribosome, whereas most NoCEs likely originate from separate factors unique to each of the species. The most notable difference between the two species (marked with a light blue rectangle in the three matrices in Figure 3B) is in the most frequently observed substitution of Cys to Ala in yeast, which is not seen in the bacterium. A recent report (Sun et al., 2016) revealed the basis for this exact observation: the eukaryotic, but not prokaryotic, alanyl-tRNA synthetase (AlaRS) tends to mischarge $tRNA^{Cys}$ with alanine.

For the yeast data, we also computed three 4 × 4 confusion matrices and observed that, similar to the *E. coli* matrices, they also predominantly feature G:U mismatches at the first or second positions (Figure 3A), but the eukaryote also shows an elevated level of G:G mispairing in the second position. Observing such levels of error similarity between such loosely related organisms suggests that these errors depend on universal chemical or genetic constraints.

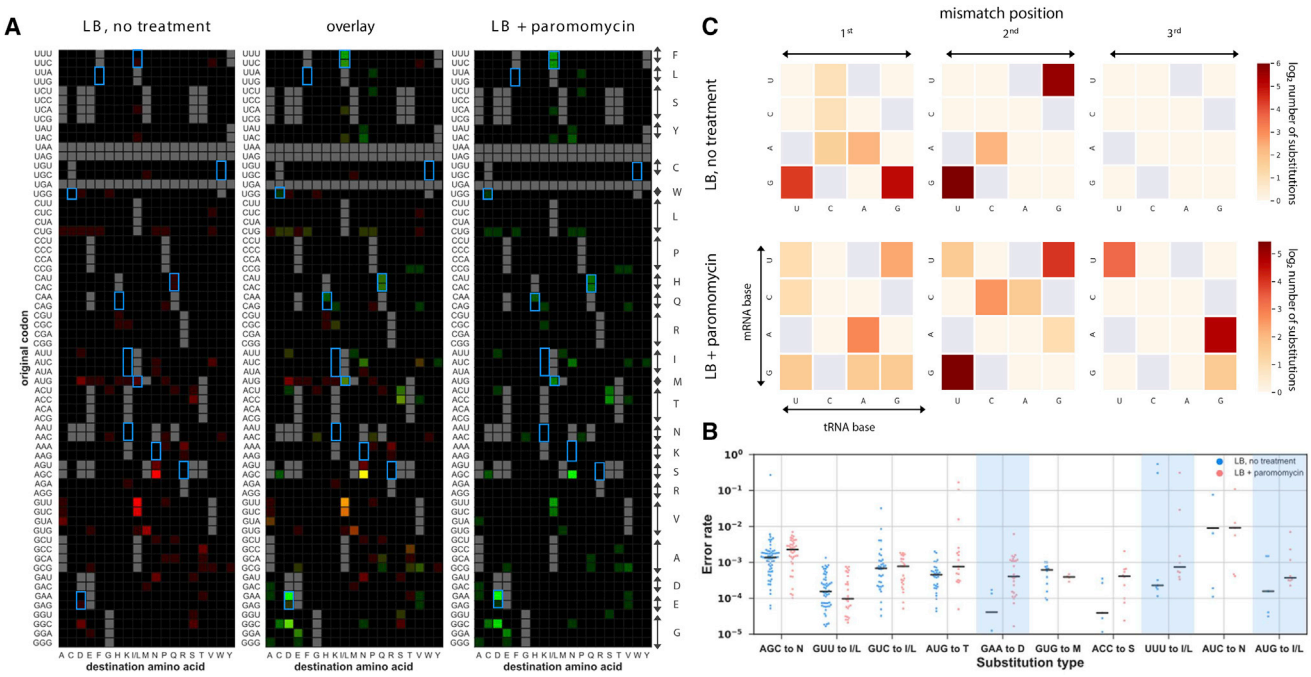


Figure 4. The Error Spectrum Is Affected by Sub-lethal Concentrations of Paromomycin

(A) The substitution identifications matrices of *E. coli* in LB (red channel, left) or LB supplemented with paromomycin (green channel, right) are compared and overlaid (center). The intensity of the color is proportional to the logarithm of the number of independent identifications, with one pseudo-count. Values are normalized by the highest entry in the matrix for each of the two conditions. The blue boxes highlight errors involving third-position mismatches.

(B) Quantification of the top 10 most frequent substitutions in the drugs dataset. Errors involving third-position mismatches are shaded in light blue.

(C) NeCEs are classified using the same procedure as in Figure 3A for the LB samples, with or without paromomycin. The shade intensity reflects the logarithm (base 2) of the number of peptides in which the corresponding mismatch was observed. Pearson correlation was calculated as in Figure 3A. Significantly similar positions found were as follows: LB position 2 against position 2 of LB + paromomycin ($r = 0.92$, $p = 1 \times 10^{-5}$) and LB position 1 against position 3 ($r = 0.83$, $p = 7 \times 10^{-4}$).

Aminoglycoside Treatment and Amino Acid Depletion Affect the Translation Error Spectrum

We characterized error patterns upon perturbations. First, we grew *E. coli* in Luria-Bertani medium (LB) supplemented with sub-lethal concentrations of paromomycin, an aminoglycoside antibiotic known to interfere with the ribosome's proofreading activity (Gromadski and Rodnina, 2004; Kramer and Farabaugh, 2007). We inspected the codon-to-amino acid error matrices under treated and non-treated conditions (Figure 4A), the error rate profiles (Figure 4B), and the nucleotide mispairing matrices (Figure 4C). Comparing the codon-to-amino acid substitution matrices between treated and untreated samples revealed a clear pattern: the drug increased error rates, especially at third codon wobble positions.

Next we examined the effect of amino acid depletion on the cell's error spectrum. In three parallel experiments, we partially depleted either isoleucine, proline, or serine using the corresponding auxotrophic *E. coli* strains or the wild-type strain (STAR Methods). We predicted that, upon depletion of each of the amino acids, we would observe elevated levels of errors specific to the deleted amino acid. Further, we examined which amino acids replaced the depleted one and which codons of the original depleted amino acid were more sensitive to depletion. Overall, we observed increased translation errors, typically leading from certain codons of the depleted amino acid to other amino acid destinations, which were typically NeCEs of the depleted one (Figure 5); there were no major changes in the rate of substitution from other amino acids (data not shown). Depletion of Ile resulted in an elevation of errors in the auxotroph (ΔIlvA) strain, in which Ile is replaced predominantly with Met, Thr, and Val, all destinations that represent NeCE substitutions (Figure 5A). Upon depletion of Pro, we also detect new errors leading mainly from proline, and the main new destinations were Ala, Ile/Leu, Thr, and Ser (all NeCEs of the corresponding codons apart from Ile, which cannot be distinguished from Leu with MS). Curiously, these new substitutions were more prevalent when the auxotroph was grown on complete medium rather than on proline-depleted medium (Figure 5B). Upon depletion of serine, proline (a NeCE substitution) becomes the most abundantly observed new destination of serine. The AGU and AGC codons are often substituted with Asn, but this dynamics is also observed in the non-depleted wild type (Figure 5C). However, when we followed a time course of the translation error dynamics across the growth of the culture toward stationary phase, these codons showed more intricate dynamics. We observed, as expected, an increase in the rate of $Ser_{AGC} \rightarrow Asn$ mistranslation, and the rate of this error intensified as the cells approached stationary phase, when the depletion effect intensified (Figure 5D). This result indicates a clear mechanism that accounts for mistakes in translation in which a shortage of an amino

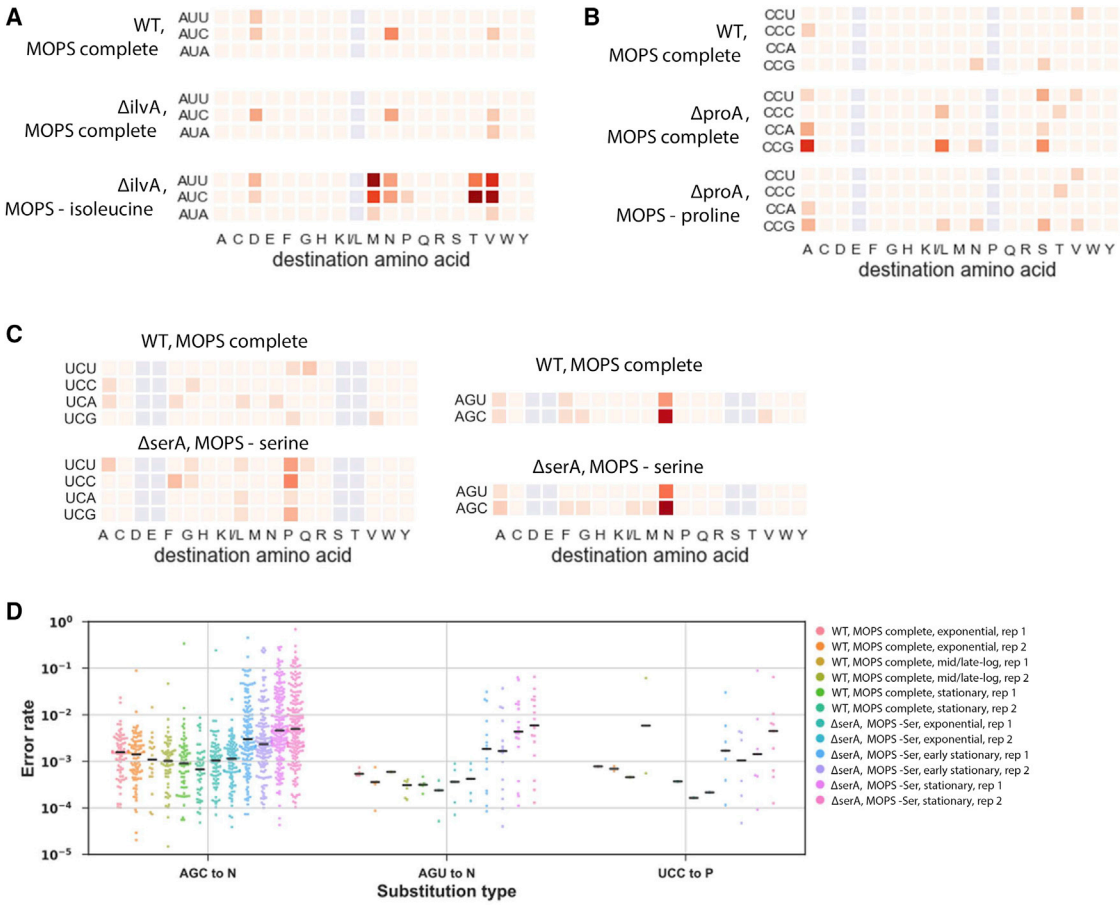


Figure 5. Depleting Cells of Particular Amino Acids Promotes Mistranslation at Certain Codons of the Corresponding Amino Acid

The shade of each cell is proportional to the logarithm of the number of different peptides observed bearing a substitution from the codon on the y axis to the amino acid on the x axis. Only codons of the depleted amino acid are represented for clarity in each panel; other amino acids rarely show a response relative to the wild type, not depleted.

(A) Effects of isoleucine depletion on isoleucine codons. Top: wild type grown on complete medium. Center: *ΔilvA*, an isoleucine auxotroph, grown on complete medium. Bottom: *ΔilvA* grown on medium depleted of isoleucine.

(B) Effects of proline depletion on proline codons: Top: wild type grown on complete medium. Center: *ΔproA*, a proline auxotroph, grown on complete medium. Bottom: *ΔproA* grown on medium depleted of proline.

(C) Effects of serine depletion on serine codons: Top: wild type grown on complete medium (6 samples pooled). Bottom: *ΔserA*, a serine auxotroph grown on medium depleted of serine (6 samples pooled).

(D) Error rate quantification of the three most common error types for serine depletion. Error rates were computed as the ratio of intensities of the substituted and the unmodified parent peptide. Colors indicate the growth phase of each of the three time points when samples were taken for each of the two cultures. Exact OD values can be found in Figure S4.

acid determines its probability to be replaced by others. A theoretical model (Elf et al., 2003) predicts that the UCN serine codons should be affected more strongly by depletion than the AGC and AGU codons due to differential loading of the corresponding tRNAs. Indeed, our method detected increased number of substitutions stemming from the UCN codons upon starvation. On the other hand, the rate of substitution shows an increase, upon deletion, in the AGC and AGU codons. This observation does not necessarily contradict the theory because the AGC and AGU errors are also those most seen in MOPS complete medium, and it is possible that the UCX codons suffer more from starvation, but the substitutions in these codons remained at low levels in absolute terms.

Surprisingly, we also observed a signal for threonine-to-serine substitutions, which is independent of the threonine codon (data not shown). This is in line with the observation made by Ling and Soll (2010), who reported that *E. coli*'s threonine aaRS tended to misincorporate serine for threonine during oxidative stress conditions.

A Tradeoff between Fidelity and Speed of Translation by the Ribosome

The theory of kinetic proofreading (Hopfield, 1974; Ninio, 1975) predicts that the ribosome trades-off speed and accuracy during the aa-tRNA selection step, with a higher error rate expected at positions where the ribosome decodes more rapidly (Banerjee et al., 2017). This trade-off was experimentally

observed *in vitro* as a function of magnesium concentrations (Johansson, Zhang and Ehrenberg, 2012) and i*n vivo* while comparing hyper and hypo-accurate ribosomal mutants (Zhu, Dai and Wang, 2016). Yang et al. (2014) proposed that yeast cells take advantage of mRNA structures downstream of the ribosome to tip this trade-off toward faster or more accurate decoding.

The speed at which ribosomes translate mRNAs can be estimated at nucleotide resolution from ribosome footprint density (Ingolia, 2016). Such data allowed us to test whether substitutions arise preferentially at quickly translated positions. We estimated the A-site ribosome density across the proteome of *E. coli* (MG1655) using a published dataset acquired under similar conditions (Woolstenhulme et al., 2015) and standardized this density score per protein to control for intergenic differences in mRNA abundance and translation initiation levels (STAR Methods). We compared the mean ribosome density at error sites with that of sets of random controls (Figure 6A). In each of the controls, the mean ribosome density of the *bona fide* substitutions was found to be lower than that of most of the random controls (Figure 6C); error sites are less dense and, hence, translated faster than expected by chance. The mean density at sites associated with a substitution was significantly ($p = 0.025$) lower than expected under our null model controlling for biases associated with codon identity.

Misincorporations Occur at Error-Tolerant and Rapidly Evolving Protein Positions

A prediction (Drummond and Wilke, 2008) posited that, to avoid fitness loss because of protein misfolding and aggregation, cells manage their errors by selecting error-proof codons at positions where inserting the correct amino acid is critical for folding or function. Support for that theory was obtained by computational means. To examine this notion, we computed normalized conservation scores for each position in the *E. coli* proteome (STAR Methods) so that a high score indicates that an amino acid position evolves rapidly compared with other positions in the same protein. To account for the fact that some amino acids types tend to be more conserved than others and that some codons are over-represented at conserved positions, we devised three strategies to generate adequate negative controls (Figure 6A). In the first and least stringent control, for each observed substitution, we sampled a normalized conservation score from any position in the same protein. In the second control, the random re-sampling was carried within the same protein but also with the additional constraint that the amino acid type in the randomly sampled position is identical to the amino acid type observed originally at the position where the substitution occurred. Finally, in the most stringent control, we performed random re-sampling within the same protein at sites sharing the same codon as the observed positions. We generated 1,000 such re-samplings in each of the three types of negative controls and compared the mean of the observed distribution of evolutionary rate scores at the observed substitution positions with those of the random control distributions to obtain empirical p values. The mean rate of evolution at substitution sites was similar to that of random sets of positions generated through the first model but significantly higher than that of the random sets generated with the other two (Figure 6B). Consistent with a previous prediction (Drummond and Wilke, 2008), controlling for codon identity reduced the magnitude of the difference between the real error sites and randomly selected sites (second and third control, Figure 6B, "same codon" versus "same amino acid"), supporting the notion that evolution allows or precludes error-prone codons from sites that are correspondingly tolerant or intolerant to errors. However, codon identity did not fully account for the poor conservation at substituted sites, suggesting that other factors allow cells to locally modulate their error levels.

Similarly to conservation, we examined the related possibility that observed substitutions minimally affect the energetic folding stability of the protein in which they occur. We predicted the effect of each of the observed substitutions on its protein's folding stability ($\Delta\Delta G$). We compared the distribution of $\Delta\Delta G$ to mock distributions obtained via three control strategies (Figure 6D). The first is the "identity control," and it allows us to test whether the observed NeCEs are less destabilizing, on average, than other substitution types modeled to occur at the same protein positions. We generated 1,000 random sets of $\Delta\Delta G$ values so that for each observed substitution, we randomly sampled a destination amino acid accessible via a single-nucleotide substitution from the error-bearing codon and computed the difference in free energy resulting from the resulting amino acid substitution. The mean $\Delta\Delta G$ of the set of *bona fide* substitutions, 1.45 kcal/mol, was lower than that of each of the 1,000 mock sets sampled under the identity control, suggesting that error rates are tuned so that substitutions preferentially replace the original amino acid with one that would minimally affect protein folding at the same site. Our control strategy accounts for the fact that substitution types classified as NeCEs tend to swap chemically similar amino acids because of the organization of the genetic code. For a second negative control, we asked whether the observed errors were preferentially mapped to protein positions in which they minimally destabilize folding. We generated 1,000 sets for each observed NeCE and estimated the $\Delta\Delta G$ that would have resulted from that same substitution occurring at a randomly chosen position of the same protein sharing either the same amino acid ("amino acid control") or the same codon ("codon control"). The mean $\Delta\Delta G$ of our set of observed substitutions was lower than that of 98% of the sets generated with the amino acid control and 97% of those generated with the codon control, suggesting that substitutions preferentially occur at protein sites where they would not disturb folding.

Little Indication for Selective Degradation of Proteins with Harmful Misincorporations

From the above analyses, it could be deduced that the translation machinery selectively avoids harmful errors. But an alternative interpretation could be that all errors, benign and harmful, are equally likely to occur but that the protein degradation machinery selectively degrades proteins bearing harmful errors. To address this alternative, we performed another full experiment of our error detection pipeline on *E. coli* deleted for the Lon protease, a protein that degrades misfolded and unfolded proteins (Powers, Powers and Gierasch, 2012). In the *Δlon* strain,

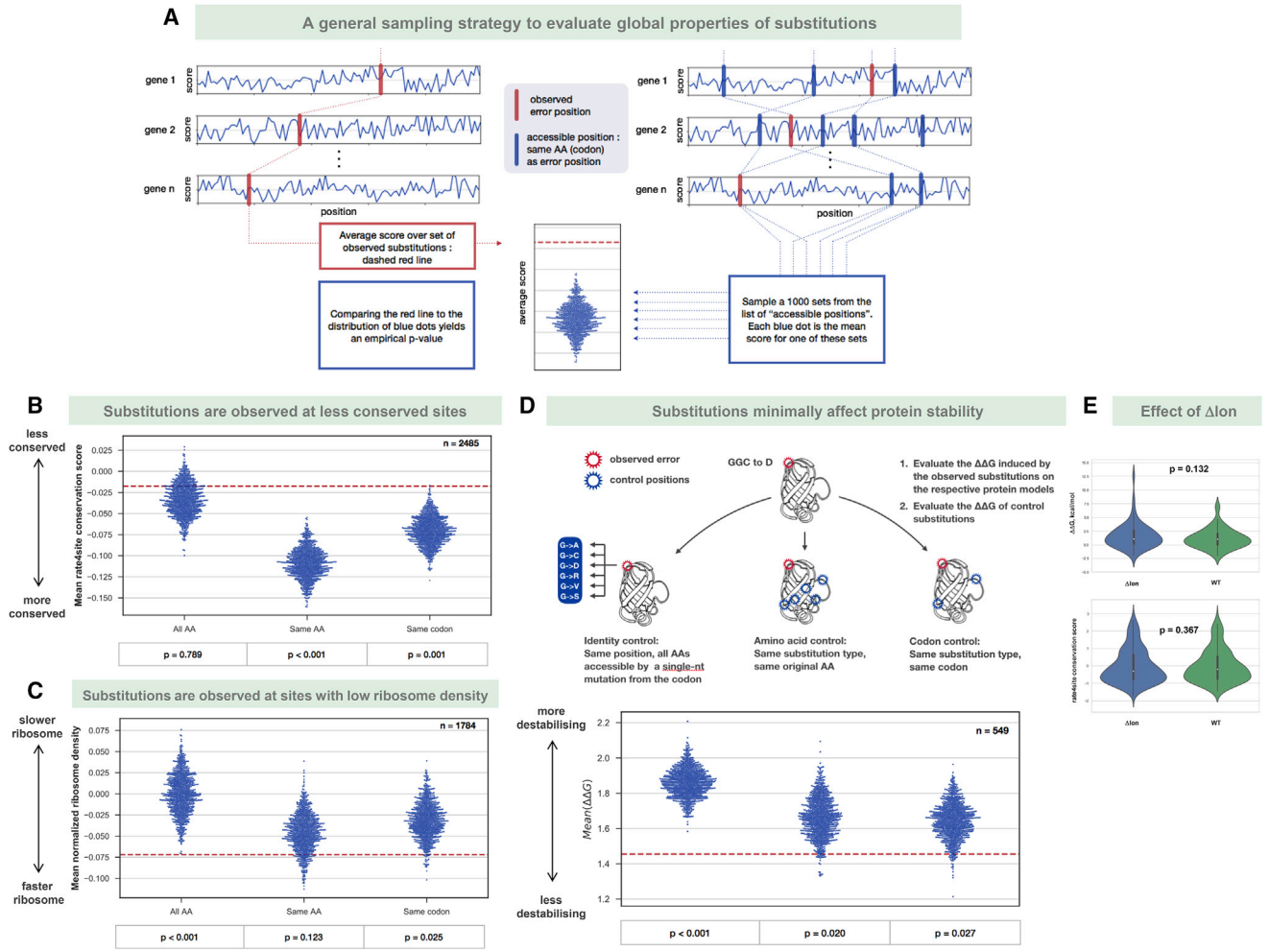


Figure 6. Global Properties of Substitution Errors

(A) General sampling strategy. To test whether the set of detected substitutions differs from expectations in any way, we first need to account for the fact that many local properties of proteins are affected by the protein's expression level and so is our ability to detect substitutions from that protein. First, the local property of interest ("score") is recorded at all positions bearing a substitution. The average of that set is plotted as a red dashed line. To compare this average with an appropriate control, we devised three strategies to eliminate the potential contributions of protein level, amino acid identity, and codon identity on the score. In each of these strategies, we draw 1,000 sets of the same size as the set of observed substitutions and plot the average of each of these sets as a blue dot. In the first strategy, for every *bona fide* distribution, we draw the score from any position within the same protein. In the second strategy, we draw the score from any position within the same protein that shares the same amino acid as the one bearing the *bona fide* substitution. Similarly, in the third control, the codon for the sampled position has to be the same as the substituted codon.

(B) Amino acid conservation. We derived amino acid conservation scores for *E. coli* proteins using the COGs database to fetch 50 homologs, MUSCLE (Edgar, 2004) to align them, and rate4site to estimate the evolutionary rate at each site. The resulting scores are standardized per protein, and a high score indicates low conservation. The empirical p values are computed by dividing the number of blue dots above the red line by 1,000. n indicates the number of positions considered in this analysis.

(C) Ribosome density. Ribosome profiling data from Woolstenhulme et al. (2015) was processed (STAR Methods) to estimate the ribosome density at positions along the *E. coli* transcriptome. Because ribosome speed (density) can only affect errors in *cis*, this analysis was restricted to NeCEs. The empirical p values are computed by dividing the number of blue dots bellow the red line by 1,000. n indicates the number of positions considered in this analysis.

(D) Effect of substitutions on protein stability. For proteins whose 3D structure is known, we evaluated the effect of NeCEs on protein stability using FoldX. In control 1, we test whether the observed substitutions are, on average, less destabilizing than those stemming from other single-nucleotide mismatches between the codon and the anticodon at the same position. In controls 2 and 3, we test whether the observed substitution type observed was less destabilizing, on average, at the observed position than at other positions sharing the same amino acid or the same codon. The empirical p values are computed by dividing the number of blue dots bellow the red line by 1,000. n indicates the number of positions considered in this analysis.

(E) Effect of *Δlon*. Top: distribution of ΔΔG associated with substitutions observed in *Δlon* (blue) and wild type (green) strains. The mean of the two distributions does not differ significantly (two-sided Wilcoxon rank-sum test, p = 0.132). Bottom: distribution of the rate4site conservation score associated with substitutions observed in *Δlon* (blue) and wild type (green) strains. The mean of the two distributions does not differ significantly (two-sided Wilcoxon rank-sum test, p = 0.367).

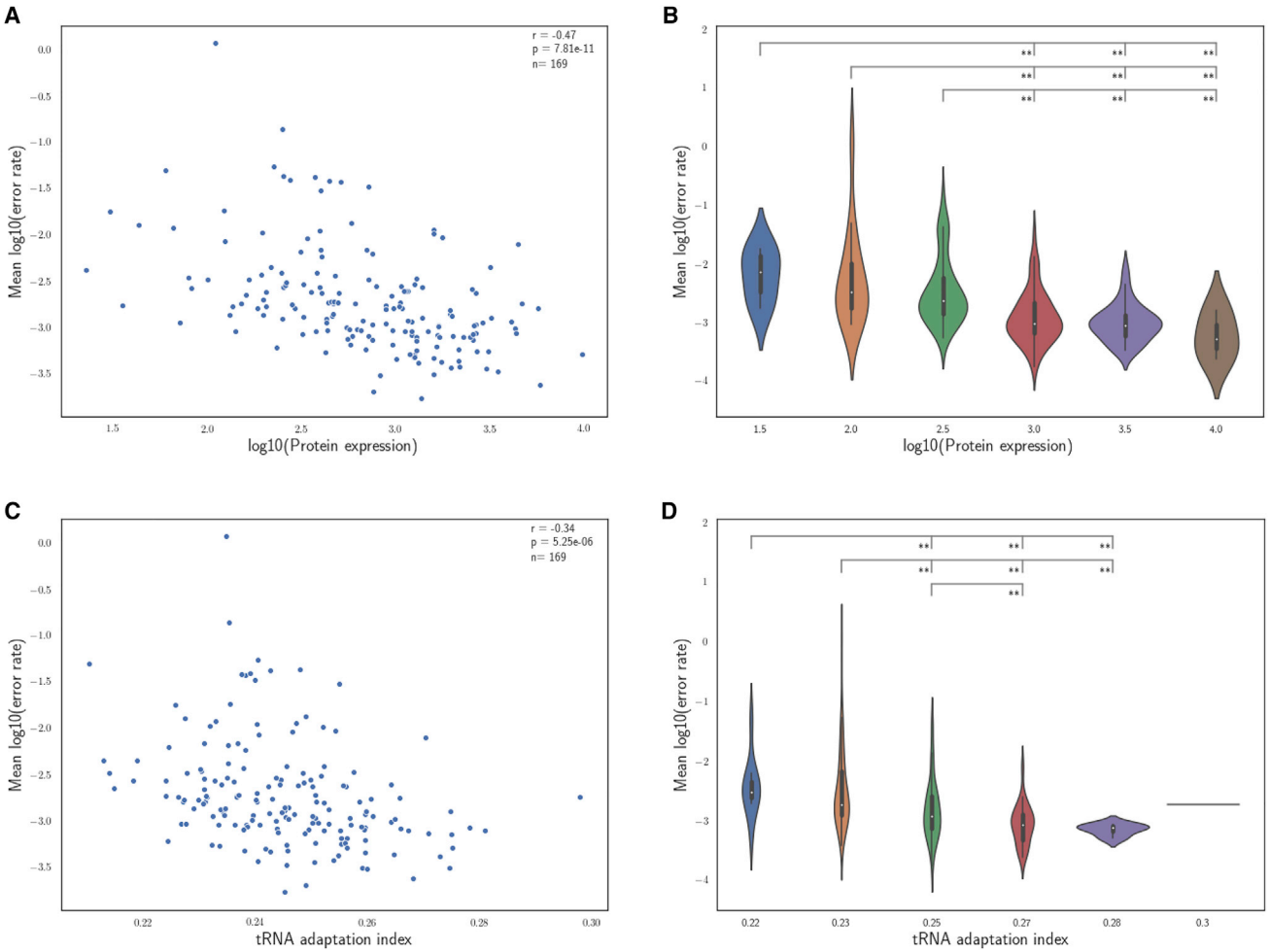


Figure 7. The Error Rate Decreases with Protein Expression

Log 10 of the error rate (dependent peptide [DP]/base peptide [BP] intensity ratio) is plotted against the average protein expression. Log error rates of all detected peptides within a protein are averaged into one score per protein. Only proteins with two or more dependent peptides were considered for this analysis.

(A and B) In the scatterplot (A), each protein is a dot, whereas in the violin plot (B), proteins are binned together based on expression. Protein expression for all *E. coli* proteins are taken from PaxDB.

(C and D) Shows the same analysis but with the tRNA adaptation index (tAI) (dos Reis et al., 2004), a computational proxy for expression, instead of the PaxDB data.

For the scatter plots (A and C), r indicates the Pearson correlation and p is the corresponding p value. n indicates the number of proteins. For the violin plots (B and D), two-sided Wilcoxon rank-sum test was conducted. *$p < 0.05$, **$p < 0.01$.

observed error rates should more closely mirror the actual translation error rate.

We detect more errors upon deletion of *Lon* (Figure S6A), indicating that the protease indeed eliminates proteins with translation errors. However, we found that new errors exposed upon deletion of the protease do not have a different level of de-stabilizing effects on protein structure (Figure 6E, top; Figure S6B), nor do they have any modified tendency to occur in more (or less) slowly evolving sites in proteins (Figure 6E, bottom; Figure S6C). These results indicate that depletion of errors from structurally sensitive and slowly evolving sites results from genuine selection for higher accuracy at these sites. These results do not exclude the parallel possibility that the degradation machinery has selectivity too, to eliminate preferentially proteins with de-stabilizing errors. It is possible that other proteases degrade proteins with de-stabilizing errors.

Misincorporations Occur Less Frequently in Highly Expressed Proteins

Further to observing that sensitive and slowly evolving positions along proteins show fewer errors than other sites on the same proteins, we wanted to examine whether certain proteins display overall lower error rates than others. In particular, we checked whether more highly expressed genes exhibit lower error rates. Indeed, we saw a negative correlation between protein expression level (taken from PaxDB; Wang et al., 2015) and mean error rate among all peptides within a protein (Figures 7A and 7B). This correlation is sustained across the entire dynamic range of protein abundance levels. The correlation was also observed when

we used two additional means to assess protein expression levels, one based on the computational tRNA adaptation index (dos Reis et al., 2004; Figures 7C and 7D) and a second based on the current study's protein expression data (data not shown). This observation supports a model (Yang et al., 2014) that suggested that, upon trading off accuracy and speed of translation, highly expressed genes would be optimized for accurate translation because the cost of mistranslating them would be prohibitive.

DISCUSSION

Here we provide a new methodology to detect and quantify translation errors at protein and codon resolution. We quantify the error rate to be around one misincorporation event per 1,000 translated residues on average, but this rate varies by more than an order of magnitude between sites in proteins, codons, and physiological conditions. Although translation errors have, by definition, a short expiration time. Until the protein is degraded, the reproducible nature of such errors—i.e., the fact that same protein positions tend to repeatedly misincorporate the same amino acid—actually indicate that errors endure throughout the life of the organism as if they were encoded in the DNA.

Although both the ribosome and the amino acid synthetases are responsible for translation errors, the majority of errors detected here appear to be made by the ribosome, which tends to mispair codons against near-cognate anticodons, incorporating wrong amino acids based on codon similarity. Indeed, codon choice can be conserved and can affect translation efficiency and accuracy (Conticello et al., 2000; Frumkin et al., 2018; Rak et al., 2018). Although the error patterns show resemblance between prokaryotes and eukaryotes, potentially because of the chemical nature of base pairing, our analysis did expose particular types of errors, predominantly in eukaryotes, that correspond to known amino acid mischarging tendencies of certain synthetases.

Several lines of evidence indicate that our observed amino acid misincorporations events mainly represent translation errors, not transcription errors. First, transcription error rates are estimated at 10^{-5}-10^{-6} per nucleotide (Traverse and Ochman, 2016); i.e., up to three orders of magnitude lower than the rate of detectable translation errors found here. Further, indications that we observed genuine translation errors and not transcription are that (1) error tendencies correlate with ribosome translation speed; (2) we observed different nucleotide mispairing patterns at each of the three codon positions; and (3), we employed two environmental perturbations that should only affect translation—a drug against ribosome proofreading and amino acid depletion, and they both modified and elevated detected errors.

Measuring error rates per protein and per position within proteins reveals that errors are distributed in a very non-random fashion. Certain positions within proteins appear to be much more error-prone than others, and certain proteins appear to be relatively protected from errors. The fact that the majority of errors observed here are attributed to mispairing between codons and anti-codons suggests that there exists a "fidelity code" for the ribosome that determines the probability of error and type of misincorporation that can occur at each position. The fact that the ribosome was shown here to be translating more rapidly at sites of error, that codon identity dictates errors, the observation that an anti-ribosome drug affected the error pattern, and the observation that error sites are more rapidly evolving and less sensitive thermodynamically to mutations all suggest that the propensity to make translation errors and the type of errors made are programmed into genes' sequences. Together, these analyses reveal that translation errors have a clear mechanistic basis and that their propensity is subject to evolutionary selection.

Examining the error patterns across diverse perturbations revealed a physiological context of translation fidelity. We examined the proteome-wide error spectra at various stages of the growth cycle, upon depletion of particular amino acids, when bacteria encounter anti-ribosomal antibiotics and when a key component of the misfolding response gene is compromised. Each of these perturbations revealed a change in the amount, rate, and type of revealed translation errors, suggesting that translation infidelity is a central constraint that characterizes physiology under many life conditions. An exciting possibility is that genes are programmed to make particular errors and that they use errors to regulate gene expression, similar to stochastic RNA editing (Bar-Yaacov et al., 2017, 2018).

Our experiments with amino acid depletion indicate that the economy of amino acids in the cell can affect quality control during aminoacylation. This conclusion follows from our observation of changes in error patterns involving a given amino acid when that amino acid is depleted from the medium. Together with previous findings showing that oxidation levels in the cell affect the accuracy of tRNA aminoacylation quality control (Ling and Soll, 2010), the picture that now emerges is that diverse factors play a role in determining acylation error rates and patterns.

Our analyses of translation errors could be discussed in the same context as that of noise in gene expression (Bar-Even et al., 2006; Newman et al., 2006; Eldar and Elowitz, 2010; Balázsi et al., 2011). The study of noise in gene expression brought the realization that genetically identical cells can express the same protein at different levels. These studies revealed how cells govern noise levels in different genes the implications of noise for physiology and evolution. Analyses of noise in gene expression revealed several key properties: (1) noise scales with gene expression; (2) the propensity for noise differs between genes, e.g., stress-related genes are more noisy than genes encoding structural proteins; and (3) noise levels are programmed into genes sequences. In parallel to "protein quantitative noise," the current study reveals "protein sequence noise;" i.e., translation errors. Both noise sources appear to be regulated by cells in a gene-specific fashion; most essential and highly expressed genes are protected from both quantitative or sequence noise, whereas rapidly evolving genes or less sensitive positions within genes are more free to display the two types of randomness. For the two manifestations of randomness, an exciting possibility is that randomness generates diversity among genetically identical cells, a state that could be essential under certain life conditions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Proteome extraction
  - Sample preparation, HPLC and Mass Spectrometry
  - Raw file processing
  - The Dependent Peptide search
  - DP identifications filtering
  - Error rate quantification
  - Assignment of NeCE to their most likely nucleotide mismatch
  - Evolutionary rates computation
  - Effect of substitutions on protein stability
  - Deep sequencing of the tRNA pool in *E. coli*
  - Ribosome density computation
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.molcel.2019.06.041.

ACKNOWLEDGMENTS

Y.P. thanks the European Research Council, European Union (Grant ID: 616622); Israel Science Foundation, Israel (Grant ID: 1332/14); Israel Cancer Research Foundation, Israel (Grant ID: 18-205-AG), and Minerva Foundation, Germany (Grant ID: AZ 5746940763) for grant support. Y.P. is an incumbent of the Ben May Professorial Chair. A.B.L. thanks the Axa Foundation Chair on Longevity and the Bettencourt Schueller Foundation. T.G. thanks the Israel Science Foundation, and Israel Center of Research Excellence Program (I-CORE, Gene Regulation in Complex Human Disease Centre No. 41/11). We thank Alon Savidor and Yishai Levin from the Mass Spectrometry unit at INCPM for performing mass spectrometry experiments. We thank Naama Knafo for assistance with MS measurements.

AUTHOR CONTRIBUTIONS

E.M., T.G., A.B.L., and Y.P. conceived the study. T.G. supervised the proteomics and mass spectrometry experimental strategy. O.D. designed and supervised the physiological experiments. J.C. devised the application of MaxQuant to this study. E.M. ran all analyses. E.M. and O.A. performed computational analyses. E.M., O.D., A.Y., and G.D.B. performed the experiments. R.R. performed tRNA sequencing and computational analyses. E.M., O.D., T.G., A.B.L., and Y.P. wrote the paper. T.G., A.B.L., and Y.P. supervised the study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 7, 2018
Revised: March 5, 2019
Accepted: June 26, 2019
Published: July 25, 2019

REFERENCES

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. Mol. Syst. Biol. *2*, 2006.0008.

Balázsi, G., van Oudenaarden, A., and Collins, J.J. (2011). Cellular decision making and biological noise: from microbes to mammals. Cell *144*, 910–925.

Banerjee, K., Kolomeisky, A.B., and Igoshin, O.A. (2017). Elucidating interplay of speed and accuracy in biological error correction. Proc. Natl. Acad. Sci. U.S.A *114*, 5183–5188.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., and Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. Nat. Genet. *38*, 636–643.

Bar-Yaacov, D., Mordret, E., Towers, R., Biniashvili, T., Soyris, C., Schwartz, S., Dahan, O., and Pilpel, Y. (2017). RNA editing in bacteria recodes multiple proteins and regulates an evolutionarily conserved toxin-antitoxin system. Genome Res. *27*, 1696–1703.

Bar-Yaacov, D., Pilpel, Y., and Dahan, O. (2018). RNA editing in bacteria: occurrence, regulation and significance. RNA Biol. *15*, 863–867.

Bratulic, S., Toll-Riera, M., and Wagner, A. (2017). Mistranslation can enhance fitness through purging of deleterious mutations. Nat. Commun. *8*, 15410.

Chen, W.-H., Lu, G., Bork, P., Hu, S., and Lercher, M.J. (2016). Energy efficiency trade-offs drive nucleotide usage in transcribed regions. Nat. Commun. *7*, 11334.

Conticello, S.G., Pilpel, Y., Glusman, G., and Fainzilber, M. (2000). Position-specific codon conservation in hypervariable gene families. Trends Genet. *16*, 57–59.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

Cvetesic, N., Semanjski, M., Soufi, B., Krug, K., Gruic-Sovulj, I., and Macek, B. (2016). Proteome-wide measurement of non-canonical bacterial mistranslation by quantitative mass spectrometry of protein modifications. Sci. Rep. *6*, 28631.

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

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell *134*, 341–352.

Drummond, D.A., and Wilke, C.O. (2009). The evolutionary consequences of erroneous protein synthesis. Nat. Rev. Genet. *10*, 715–724.

Edelmann, P., and Gallant, J. (1977). Mistranslation in E. coli. Cell *10*, 131–137.

Edgar, R.C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. Nature *467*, 167–173.

Elf, J., Nilsson, D., Tenson, T., and Ehrenberg, M. (2003). Selective charging of tRNA isoacceptors explains patterns of codon usage. Science *300*, 1718–1722.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods *4*, 207–214.

Frumkin, I., Lajoie, M.J., Gregg, C.J., Hornung, G., Church, G.M., and Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proc. Natl. Acad. Sci. USA *115*, E4940–E4949.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. *43* (Database issue, D1), D261–D269.

Goloborodko, A.A., Levitsky, L.I., Ivanov, M.V., and Gorshkov, M.V. (2013). Pyteomics–a Python framework for exploratory data analysis and rapid software prototyping in proteomics. J. Am. Soc. Mass Spectrom. *24*, 301–304.

Gromadski, K.B., and Rodnina, M.V. (2004). Streptomycin interferes with conformational coupling between codon recognition and GTPase activation on the ribosome. Nat. Struct. Mol. Biol. *11*, 316–322.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell *38*, 576–589, https://doi.org/10.1016/j.molcel.2010.05.004.

Hopfield, J.J. (1974). Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. Proc. Natl. Acad. Sci. U.S.A *71*, 4135–4139.

Ingolia, N.T. (2016). Ribosome Footprint Profiling of Translation throughout the Genome. Cell *165*, 22–33.

Johansson, M., Zhang, J., and Ehrenberg, M. (2012). Genetic code translation displays a linear trade-off between efficiency and accuracy of tRNA selection. Proc. Natl. Acad. Sci. USA *109*, 131–136.

Jones, T.E., Alexander, R.W., and Pan, T. (2011). Misacylation of specific non-methionyl tRNAs by a bacterial methionyl-tRNA synthetase. Proc. Natl. Acad. Sci. U.S.A *108*, 6933–6938.

Kapur, M., and Ackerman, S.L. (2018). mRNA Translation Gone Awry: Translation Fidelity and Neurological Disease. Trends Genet. *34*, 218–231.

Khan, Z., Amini, S., Bloom, J.S., Ruse, C., Caudy, A.A., Kruglyak, L., Singh, M., Perlman, D.H., and Tavazoie, S. (2011). Accurate proteome-wide protein quantification from high-resolution 15N mass spectra. Genome Biol. *12*, R122.

Kramer, E.B., and Farabaugh, P.J. (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA *13*, 87–96.

Kramer, E.B., Vallabhaneni, H., Mayer, L.M., and Farabaugh, P.J. (2010). A comprehensive analysis of translational missense errors in the yeast Saccharomyces cerevisiae. RNA *16*, 1797–1808.

Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. Nat. Methods *11*, 319–324.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359, https://doi.org/10.1038/nmeth.1923.

Lee, J.W., Beebe, K., Nangle, L.A., Jang, J., Longo-Guess, C.M., Cook, S.A., Davisson, M.T., Sundberg, J.P., Schimmel, P., and Ackerman, S.L. (2006). Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. Nature *443*, 50–55.

Lee, H., Popodi, E., Tang, H., and Foster, P.L. (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. Proc. Natl. Acad. Sci. U.S.A *109*, E2774–E2783.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.

Lindner, A.B., and Demarez, A. (2009). Protein aggregation as a paradigm of aging. Biochim. Biophys. Acta *1790*, 980–996.

Ling, J., and Soll, D. (2010). Severe oxidative stress induces protein mistranslation through impairment of an aminoacyl-tRNA synthetase editing site. Proc. Natl. Acad. Sci. U.S.A *107*, 4028–4033.

Lowe, T.M., and Chan, P.P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. *44* (W1). W54-7. https://doi.org/10.1093/nar/gkw413.

Marx, H., Lemeer, S., Schliep, J.E., Matheron, L., Mohammed, S., Cox, J., Mann, M., Heck, A.J., and Kuster, B. (2013). A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. Nat. Biotechnol. *31*, 557–564.

Miranda, I., Silva-Dias, A., Rocha, R., Teixeira-Santos, R., Coelho, C., Gonçalves, T., Santos, M.A., Pina-Vaz, C., Solis, N.V., Filler, S.G., and Rodrigues, A.G. (2013). Candida albicans CUG mistranslation is a mechanism to create cell surface variation. MBio *4*, e00285–e13.

Moruz, L., and Käll, L. (2017). Peptide retention time prediction. Mass Spectrom. Rev. *36*, 615–623.

Na, S., Bandeira, N., and Paek, E. (2012). Fast Multi-blind Modification Search through Tandem Mass Spectrometry. Mol. Cell. Proteomics *11*, M111.010199.

Netzer, N., Goodenbour, J.M., David, A., Dittmar, K.A., Jones, R.B., Schneider, J.R., Boone, D., Eves, E.M., Rosner, M.R., Gibbs, J.S., et al. (2009). Innate immune and chemically triggered oxidative stress modifies translational fidelity. Nature *462*, 522–526.

Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840–846.

Ninio, J. (1975). Kinetic amplification of enzyme discrimination. Biochimie *57*, 587–595.

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., et al. (2019). The PRIDE database and related tools and resources in 2019. improving support for quantification data. Nucleic Acids Res. *47*, D442–D450.

Powers, E.T., Powers, D.L., and Gierasch, L.M. (2012). FoldEco: a model for proteostasis in E. coli. Cell Rep. *1*, 265–276.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics, S71–S77.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *6*, 841–842, https://doi.org/10.1093/bioinformatics/btq033.

Rak, R., Dahan, O., and Pilpel, Y. (2018). Repertoires of tRNAs: The Couplers of Genomics and Proteomics. Annu. Rev. Cell Dev. Biol. *34*, 239–264.

Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal. Chem. *75*, 663–670.

Savitski, M.M., Nielsen, M.L., and Zubarev, R.A. (2006). ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. Mol. Cell. Proteomics *5*, 935–948.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic Acids Res. *33* (Web Server issue, SUPPL. 2), W382-8.

Sinitcyn, P., Rudolph, J.D., and Cox, J. (2018). Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data *1*, 207–234.

Sun, L., Gomes, A.C., He, W., Zhou, H., Wang, X., Pan, D.W., Schimmel, P., Pan, T., and Yang, X.L. (2016). Evolutionary Gain of Alanine Mischarging to Noncognate tRNAs with a G4:U69 Base Pair. J. Am. Chem. Soc. *138*, 12948–12955.

Traverse, C.C., and Ochman, H. (2016). Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. Proc. Natl. Acad. Sci. USA *113*, 3311–3316.

Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P.A. (2005). Identification of post-translational modifications by blind search of mass spectra. Nat. Biotechnol. *23*, 1562–1567.

Wang, M., Herrmann, C.J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics *15*, 3163–3168.

Whitehead, D.J., Wilke, C.O., Vernazobres, D., and Bornberg-Bauer, E. (2008). The look-ahead effect of phenotypic mutations. Biol. Direct *3*, 18.

Wiltrout, E., Goodenbour, J.M., Fréchin, M., and Pan, T. (2012). Misacylation of tRNA with methionine in Saccharomyces cerevisiae. Nucleic Acids Res. *40*, 10494–10506.

Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods *6*, 359–362.

Wohlgemuth, I., Pohl, C., Mittelstaet, J., Konevega, A.L., and Rodnina, M.V. (2011). Evolutionary optimization of speed and accuracy of decoding on the ribosome. Philos. Trans. R Soc. Lond. B Biol. Sci. *366*, 2979–2986.

Woolstenhulme, C.J., Guydosh, N.R., Green, R., and Buskirk, A.R. (2015). High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. Cell Rep. *11*, 13–21.

Yang, J.R., Chen, X., and Zhang, J. (2014). Codon-by-Codon Modulation of Translational Speed and Accuracy Via mRNA Folding. PLoS Biol. *12*, e1001910.

Zhang, Z., Shah, B., and Bondarenko, P.V. (2013). G/U and certain wobble position mismatches as possible main causes of amino acid misincorporations. Biochemistry *52*, 8165–8176.

Zheng, G., Qin, Y., Clark, W.C., Dai, Q., Yi, C., He, C., Lambowitz, A.M., and Pan, T. (2015). Efficient and quantitative high-throughput tRNA sequencing. Nat. Methods *12*, 835–837.

Zhu, Y.O., Siegal, M.L., Hall, D.W., and Petrov, D.A. (2014). Precise estimates of mutation rate and spectrum in yeast. Proc. Natl. Acad. Sci. U.S.A *111*, E2310–E2318.

Zhu, M., Dai, X., and Wang, Y.-P. (2016). Real time determination of bacterial in vivo ribosome translation elongation speed based on LacZα complementation system. Nucleic Acids Res. *44*, e155.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
E.coli: strain MG1655	ATCC stock center	Cat #: 47076
E.coli: strain BW25113	CGSC stock Center	Cat #: 7636
E.coli: strain JW2880-1	CGSC stock Center	Cat #: 10234
E.coli: strain JW0233-2	CGSC stock Center	Cat#: 8468
E.coli: strain JW3745-2	CGSC stock Center	Cat#: 10733
E.coli: strain JW0429-1	CGSC stock Center	Cat#: 8592
Chemicals, Peptides, and Recombinant Proteins		
MOPS rich defined medium	Teknova	Cat#: M2105
Serine methyl ester	Sigma	Cat #: 412201
Proline methyl ester	Sigma	Cat #: 287067
Isoleucine methyl ester	Sigma	Cat #: 58920
B-PER Bacterial Protein Extraction Reagent	Thermo Fisher Scientific	Cat#: 78248
TRI-reagent	Sigma	Cat#: T9424
Agencourt AMPure XP	Beckman Coulter	Cat#: A63881
AlkB wt and D135S	Zheng et al., 2015	N/A
TGIRT-III	InGex	Cat#: TGIRT50
T4 RNA ligase	New England Biolabs	Cat#: M0204L
Dynabeads myOne SILANE	Thermo Fisher Scientific	Cat#: 37002D
NEBNext PCR mix	New England Biolabs	Cat#: M0541L
Trypsin	Promega	Cat#: V5113
C18 StageTip	Fisher Scientific	Cat#: 14-386-2
Deposited Data		
Raw and analyzed mass spectrometry data	This paper	PRIDE: PXD014341
Raw and analyzed tRNA data	This paper	GEO: GSE128812
E. coli reference genome ASM584v2	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/assembly/GCA_000005845.2#/st
E. coli tRNA sequences	Lowe and Chan, 2016	http://gtrnadb.ucsc.edu/genomes/bacteria/Esch_coli_K_12_MG1655/
E. coli reference CDS	European Nucleotide Archive	http://bacteria.ensembl.org/Escherichia_coli_str_k_12_substr_mg1655/Info/Index
E. coli reference proteome	UNIPROT	https://www.uniprot.org/proteomes/UP000000625
Saccharomyces cerevisiae raw mass spectrometry files	PRIDE	https://www.ebi.ac.uk/pride/archive/projects/PXD000269
Saccharomyces cerevisiae reference CDS		
Saccharomyces cerevisiae reference proteome	UNIPROT	https://www.uniprot.org/proteomes/UP000002311
Oligonucleotides		
Primer for tRNA reverse-transcription	This paper	N/A
DNA		
5′-CACGACGCTCTTCCGATCTT –3′		
RNA		
5′-		
rArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUr		
G-3′		
3′ adaptor for tRNA library ligation 5′- AGATCG GAAGAGCACA-3′	This paper	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Custom python scripts	This paper	https://github.com/ernestmordret/substitutions/
Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Samtools	Li et al., 2009	http://samtools.sourceforge.net/
homertools	Heinz et al., 2010	http://homer.ucsd.edu/homer/download.html
Bedtools	Quinlan and Hall, 2010	https://github.com/arq5x/bedtools2/releases
MaxQuant version 1.5.5.1.	Cox and Mann, 2008	http://www.coxdocs.org/doku.php?id=maxquant:common:download_and_installation

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Yitzhak Pilpel (pilpel@weizmann.ac.il)

EXPERIMENTAL MODEL AND SUBJECT DETAILS

To generate the *E. coli* drug dataset, *E. coli* strain MG1655 (Genotype: *F-, lambda-, rph-1,* "ATCC: 47076") was grown in 3ml LB media (10 g/l tryptone, 5g/l yeast extract, and 10 g/l NaCl) at 30°C until reaching stationary phase. Cultures were diluted 1:100 and grown aerobically in 100ml LB with and without addition paromomycin (final concentration of 5 μg/ml). Cultures were grown in standard Erlenmeyer at 37°C, shaking speed of 200rpm until reaching mid-log phase (OD ≃ 0.5). Each experiment was done in two independent biological repeats.

To generate amino acids depletion dataset, cells were grown in of 3ml of modified MOPS rich defined medium made with the following recipe: 10X MOPS rich buffer, 10X ACGU nucleobase stock and 100X 0.132M K2HPO4 (Teknova, Cat.#: M2105) were used at 1X final concentration. Medium was supplemented with 0.25% glucose as carbon source, 10^{-4}% thiamine, 17.2mM Serine, 640μM Lysine, 661μM Arginine, 592μM Histidine, 464μM Tyrosine and 800μM of all 15 remaining amino acids. pH was adjusted to 7.4 using 1M NaOH. Cultures were grown at 37°C until reaching stationary phase. Cells were then diluted 1:1000 into the appropriate media (see details below) and grown at 37°C. The exact strains, media, growth conditions and growth phase in which samples were taken in the different depletion experiments were:

- Serine depletion: BW25113 (WT) strain was grown in modified MOPS rich defined medium and JW2880-1 (*ΔserA*) strain was grown in modified MOPS defined medium depleted for amino acid serine (8.6mM Serine and 800μM Serine methyl ester (Sigma Cat#: 412201)). Cultures were grown in standard Erlenmeyers, shaking speed of 200rpm and samples were taken at 3 time points during culture growth (see Figure S4 for exact harvesting times and OD values). Each experiment was done in two independent biological repeats.
- Proline depletion: BW25113 (WT) strain was grown in modified MOPS rich defined medium and JW0233-2 (*ΔproA*) strain was grown in either modified MOPS rich defined medium or modified MOPS defined medium depleted for amino acid proline (0μM Proline and 160μM Proline methyl ester (Sigma Cat#: 287067)). Cultures were grown in baffled Erlenmeyers to increase aeration, shaking speed of 200rpm and samples were taken at a single time point at stationary phase ("t1 stationary"). Each experiment was done in three independent biological repeats.
- Isoleucine depletion: BW25113 (WT) strain was grown in modified MOPS rich defined medium and JW3745-2 (*ΔilvA*) strain was grown in either modified MOPS rich defined medium or modified MOPS defined medium depleted for amino acid isoleucine (100μM Isoleucine and 160μM Isoleucine methyl ester (Sigma Cat#: 58920)). Cultures were grown in baffled Erlenmeyers to increase aeration, shaking speed of 200rpm and samples were taken at a single time point at stationary phase ("t1 stationary"). Each experiment was done in three independent biological repeats.

To generate the protease deletion dataset, BW25113 (WT) strain as well JW0429-1 (*Δlon*) strain were grown in 3ml of modified MOPS rich defined medium at 37°C until reaching stationary phase. Cells were then diluted 1:1000 into the same media. Cultures were grown in baffled Erlenmeyers, shaking speed of 200rpm to increase aeration and samples were taken at a single time point at stationary phase ("t1 stationary"). Each experiment was done in three independent biological repeats. All auxotrophs, lon mutant, as well as the corresponding wild-type were obtained from the Keio deletion library (Baba et al., 2006).

METHOD DETAILS

Proteome extraction

We adapted our proteome extraction protocol from Khan et al. (2011). Bacterial cultures were centrifuged at 4000 rpm for 5 min, and washed twice with 10ml PBS. Remaining PBS was vacuum-aspirated and the pellets were frozen in ethanol-dry ice and stored at −80°C until protein extraction. For protein extraction, pellets were re-suspended in 1 mL of B-PER bacterial protein extraction buffer (Thermo Fisher Scientific), and vortexed vigorously for 1 min. The extract was centrifuged at 13,000 rpm for 5 min. The supernatants (high solubility fractions) were collected and frozen in an ethanol-dry ice bath. High solubility fractions were stored at −80°C. The remaining pellets were re-suspended in 2 mL of 1:10 diluted B-PER reagent. The suspensions were centrifuged and washed one more time with 1:10 diluted B- PER reagent. The pellets were re-suspended in 1 mL of Inclusion Body Solubilization Reagent (Thermo Fisher Scientific). The suspensions were vortexed for 1 min, shaken for 30 min, and placed in a sonication bath for 10 min at maximum intensity. Cell debris were removed from the suspension by centrifugation at 13,000 rpm for 15 min. The supernatants were frozen in an ethanol-dry ice bath (low solubility fraction), and stored at −80°C.

Sample preparation, HPLC and Mass Spectrometry

400 μg of protein was taken for Filter aided sample preparation (FASP)(Wiśniewski et al., 2009) trypsin digestion on top of 30kDa Microcon filtration devices (Millipore). Proteins were digested overnight at 37°C and the peptides were separated into five fractions using strong cation exchange (SCX) in a StageTip format. Peptides were purified and concentrated on C18 StageTips (Rappsilber, Ishihama and Mann, 2003) (3M EmporeTM, St. Paul, MN, USA). Liquid-chromatography on the EASY-nLC1000 HPLC was coupled to high-resolution mass spectrometric analysis on the Q-Exactive Plus mass spectrometer (ThermoFisher Scientific, Waltham, MA, USA). Peptides were separated on 50 cm EASY-spray columns (ThermoFisher Scientific) with a 140 min gradient of water and acetonitrile. MS acquisition was performed in a data-dependent mode with selection of the top 10 peptides from each MS spectrum for fragmentation and analysis.

Raw file processing

High and Low solubility fractions were aligned separately using MaxQuant. The amino acid substitutions identification procedure relies on the built-in dependent peptide algorithm of MaxQuant (Cox and Mann, 2008; Sinitcyn et al., 2018).

The Dependent Peptide search

Experimental spectra were first searched using a standard database search algorithm, without any variable modification, and the significance of identifications was controlled to a 1% FDR via a target decoy procedure. Identified spectra are then turned into a spectral library, and a decoy spectral library is created by reversing the sequences of the identified spectra. For each possible pair consisting of an identified spectrum in the concatenated spectral libraries and an unidentified experimental spectrum of the same charge, and recorded in the same raw file, we apply the following steps: first we compute the mass shift Δm by subtracting the mass of the identified (unmodified) spectrum to that of the unidentified (modified) spectrum, then we simulate modified versions of the theoretical spectrum by adding *in silico* this mass shift at every position along the peptide, and finally we evaluate the match between the theoretical spectrum and the experimental spectrum using a formula similar to Andromeda's binomial score.

For each unidentified peptide, the match with the best score is reported, the nature of the match (target or decoy) is recorded, and a target-decoy procedure (Elias and Gygi, 2007) is applied to keep the FDR at 1%. Peptides identified using this procedure are called Dependent Peptides (DP), whereas their unmodified counterparts are named Base Peptides (BP).

Additionally, the confidence of the mass shift's localization is estimated using a method similar to MaxQuant/Andromeda's PTM Score strategy, which returns the probability that the modification is harbored by any of the peptide's amino acid.

DP identifications filtering

The list of all known modifications was downloaded from http://www.unimod.org/, and those marked as AA substitution, Isotopic label or Chemical derivative were excluded. Entries in this list are characterized by a monoisotopic mass shift, and a site specificity (i.e., they can only occur on a specific amino acid or on peptides' and proteins' termini). We removed from our analysis any DP identification that could be explained by any of the remaining modifications, using the following criteria: the recorded Δm and the known modification's mass shift must not differ by more than 0.01 Da, and the modification must be harbored by a site consistent with the uniprot entry with a probability $p \geq 0.05$. Conversely, we computed the list of all possible amino acid substitutions and their associated mass shifts. For every substitution, we only retained DP identifications such that the observed Δm and the AA substitution's mass shift did not differ by more than 0.005 Da, and the mass shift was localized on the substitution's original AA with $p \geq 0.95$.

Among the remaining DP identifications, those such that the peptide sequence after substitution was a substring of the proteome (allowing Ile-Leu ambiguities), were also removed, to prevent pairing of dependent peptides and base peptides between paralogs. Finally, the FDR was controlled once again at 1% using the same procedure as above.

Next we examined the capacity of the Dependent Peptide algorithm of MaxQuant to identify known mass shifts in a well-controlled setup. For that we used a large set of synthetic peptides and their corresponding phospho-peptide counterparts (Marx et al., 2013),

that was subject to LC-MS/MS measurements. This dataset constitutes a ground truth to determine the validity of the DP algorithm. We downloaded data from the PRIDE deposition https://www.ebi.ac.uk/pride/archive/projects/PXD000138/files. We used the Fasta files of the whole human proteome plus the synthetic peptides. We ran MaxQuant with default settings and only oxidation of methionine as variable modification and not including phosphorylation. The dependent peptide peptide-spectrum matches (PSMs) were restricted to a false discovery rate of maximally 1% using a target-decoy approach. The dependent peptide algorithm found 1,315 PSMs whose modification mass corresponds to a singly phosphorylated peptide with 1% FDR on dependent peptides. Among these were neither decoy hits, nor wrongly identified peptides, indicating that the dependent peptides strategy employed in the present work is highly specific and efficient in limiting false positive identifications.

Error rate quantification

In order to assess the error rate, we quantify and compare pairs of base and dependent peptides across many samples. For each independent substitution, we fetched the quantification profile of the base peptide from MaxQuant's peptides.txt table, and similarly fetch the dependent peptide's quantification profile from the matchedFeatures.txt table. Whenever a peak has been detected and quantified for both the dependent and the base peptide, we estimate the translation error rate as the ratio of their MS1 intensities.

Assignment of NeCE to their most likely nucleotide mismatch

The assignment of observed substitutions to appropriate anticodons, done here in order to determine underlying mismatch type, was carried out in two steps. First, substitutions that could unambiguously result from a mispairing of a single tRNA type were counted, and probabilities of base-to-base mispairing were obtained; then, substitutions that could have been explained by utilizing more than one tRNA type were addressed by considering the prior base-to-base substitutions derived from the unambiguous cases. When two or more tRNA molecules could explain a given substitution, and existing priors were missing, we have distributed with equal probability the observations among all possible mismatch types. The procedure was repeated until convergence.

Evolutionary rates computation

For each of the proteins associated to a substitution in the MOPS dataset, we fetched a list of orthologous protein sequences from the COG database (Galperin et al., 2015), excluding partial matches (membership class = 3). Proteins whose list of orthologs contained less than 50 sequences were excluded from this analysis. For the remaining proteins, we randomly selected 50 sequences from the list, and created evolutionary alignments using MUSCLE (Edgar, 2004). The alignments were then used to compute normalized evolutionary rates per site with the rate4site program (Pupko et al., 2002). The mean evolutionary rate of sites associated with detected substitutions was compared to that of a 1,000 randomly sampled positions, using the strategy described in Figure 6A.

Effect of substitutions on protein stability

For each of the proteins associated to a substitution in the MOPS dataset, we attempted to fetch the best 3D structure for its biological assembly in the PDB database to estimate the effect of our substitutions on protein stability using the FoldX software (Schymkowitz et al., 2005). We excluded membrane proteins, whose stability is poorly modeled by FoldX, and excluded ribosomal protein because the ribosome is too big to be modeled entirely. We restricted our analysis to WT proteins from *E. coli*, excluding structures determined from orthologs. Among the remaining structures, we selected those with the lowest R-free score.

These structures were first "repaired" using the repairPDB command. We then evaluated the effect of a set of amino acid substitutions comprising the detected substitutions and the controls described in Figure 6D on protein stability ($\Delta\Delta G$), using the PositionScan command. Finally, the mean($\Delta\Delta G$) of our set of substitutions was compared to the mean($\Delta\Delta G$) of 1000 randomly sampled substitutions, using the strategy described in Figure 6D.

Deep sequencing of the tRNA pool in *E. coli*

Bacterial cultures were grown to OD values roughly corresponding to exponential phase (OD ~0.8), late exponential phase (OD ~1.9), and stationary phase (OD ~2.3), experiment was done in three biological repeats. Culttures were centrifuged at 12,000 rpm for 1 min, frozen in liquid nitrogen and stored at −80 C until RNA extraction. RNA was extracted using TRI-reagent (Sigma-Aldirch), according to standard protocol. tRNA sequencing protocol was adapted from Zheng et al., 2015 (Zheng et al., 2015) with minor modification. Small RNA was isolated using SPRI-beads (Agencourt AMPure XP, Beckman Coulter), using dual side size-selection protocol. First RNA and beads were mixed at 1:1.8 ratio, and supernatant was collected. The small RNA was isolated by mixing the clear supernatant with bead at 1:0.8 ratio, with addition of X1.34 isopropanol. Small RNA was treated for modification removal using AlkB wt, and D135S enzymes, a kind gift from Prof. Tao Pan. Reverse transcription was done using TGIRT-III enzyme, with the indicated DNA-RNA hybrid primer. 3′ adaptor was ligated to the cDNA using T4 ligase (NEB). The cDNA was purified using Dynabeads myOne SILANE (life Technologies) after each step. The library was amplified using NEBNext PCR mix and cleaned using SPRI-beads. Samples were pooled and sequenced using NextSeq illumina.

Primer name	Primer sequence
Reverse-transcription DNA	5′-CACGACGCTCTTCCGATCTT –3′
Reverse-transcription RNA	5′-rArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrG-3′
3′ adaptor	5′- AGATCGGAAGAGCACA-3′

Read were trimmed using homerTool. Alignment was done to the genome and mature tRNA using Bowtie2 with parameters–very-sensitive-local. Read aligned with equal alignment score to the genome and mature tRNA were annotated as mature tRNA. Reads aligned to multiple tRNA genes were randomly assigned when mapping to identical anticodon, and discarded from the analysis if aligned to different anticodon. Read count was done using BedTools-coverage count.

Ribosome density computation

Ribosome profiling data for the MOPS complete experiments were downloaded from Woolstenhulme et al. (2015) ("GEO: GSM1572266, GSM1572267"). Reads were aligned using the 3′ mapping method described in the corresponding article, and shifted by 12 nt to obtain the density at the A-site. Read counts from both replicates were summed to obtain more robust estimates, and 20 codons were excluded from both the 3′ and the 5′ ends to avoid known biases. For the remaining positions, we applied the transformation x: $\log_2(x + 1)$ to stabilize the variance, and standardized the resulting score to obtain the normalized read density (NRD), so that the mean of the NRD per protein is 0 and its standard deviation is 1. The mean(NRD) of the set of observed substitutions was then compared to that of 1000 randomly sampled substitutions, using the strategy described in Figure 6A.

DATA AND CODE AVAILABILITY

Raw files were processed with MaxQuant v. 1.5.5.1. Resulting files were processed using a custom pipeline written in Python. The parameters file for MaxQuant and the scripts can be found at https://github.com/ernestmordret/substitutions/.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al., 2019) partner repository. The accession number for the mass spectrometry protemoics data reported in this paper is PRIDE: PXD014341.

The tRNA sequencing data have been deposited in GEO. The accession number for the tRNA sequences reported in this paper is GEO: GSE128812.