USPC
Université Sorbonne
Paris Cité

PARIS DIDEROT

Thèse de
doctorat

de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

École doctorale Frontières du Vivant 474

Errors and edits: plasticity in Escherichia coli's gene expression

Par Ernest Mordret

Thèse de
doctorat de
Biologie

Dirigée par Yitzhak Pilpel

Présentée et soutenue publiquement à Paris le 12 décembre 2017

Rapporteurs : Pr. Rachel Green, HHMI/John Hopkins University
Pr. Olivier Namy, CNRS/ Paris Sud
Examinateurs : Pr. Ariel Lindner, INSERM/ Paris Descartes
Pr. Bertrand Cosson, Paris Diderot
Directeur de thèse : Pr. Yitzhak Pilpel, Weizmann Institute

Title : Errors and edits: plasticity in Escherichia coli's gene expression

Abstract : Transmission of genetic information from DNA to proteins is not as rigid as one may think. Despite their marvellous sophistication, the machines at the core of the central dogma rely on chemical processes and complex interactions, and are not exempt from errors, that organisms can either choose to mitigate, or use to their own advantage. I will present a pair of studies which reflect the two sides of this dichotomy. First, we developed a new methodology which allowed us to detect a large number of amino acid misincorporations within the proteome of *Escherichia coli* from high precision mass spectrometry data. We show that these errors are mostly the result of a competition between cognate and non cognate tRNAs in the ribosome, and that they respond dynamically to environmental perturbations, such as amino acid starvation and anti ribosome drugs. Furthermore, we demonstrate that cells tend to encode their proteins in a way that minimizes the deleterious effects of translation errors. In a second study, we discovered that mRNA editing is not restricted to eukaryotes, but is present in bacteria. We found that *Escherichia coli* exploits a tRNA adenosine to inosine editing enzyme to stimulate bacterial drug persistence through the recoding of a toxin's mRNA.

Titre : Erreurs et corrections : plasticité de l'expression génétique chez Escherichia coli

Résumé : La transmission de l'information génétique est généralement décrite comme un processus déterministe. Malgré leur sophistication, les machineries moléculaires en charge de l'expression génétique fonctionnent grâce à des réaction chimiques stochastiques par nature, et ne sont pas à l'abri d'erreurs, dont un organisme vivant peut choisir de minimiser les effets, ou au contraire de les utiliser à son avantage. Je présente dans ce manuscrit deux études illustrant ces deux possibilités. Dans un premier temps, nous avons développé une nouvelle méthodologie nous permettant de détecter un grand nombre de misincorporation d'acides aminés à travers le proteome d'*Escherichia coli*, à l'aide de données de spectrométrie de masse. Ces erreurs réagissent de manière dynamique à des perturbations environnementales telles que la privation d'un acide aminé ou la présence d'antibiotiques visant le ribosome. De plus, nous démontrons que la cellule encode ses protéines d'une manière qui minimise leurs effets délétères. Dans une deuxième étude, nous montrons que l'édition d'ARN messagers n'est pas restreinte aux Eukaryotes. *Escherichia coli* utilise le surplus d'activité d'une enzyme modifiant l'adenosine d'un anticodon d'ARNt en inosine, afin de modifier la sequence codante d'une toxine, accentuant en retour le phénomène de persistance bactérienne.

Acknowledgements

First and foremost, I would like to thank my PhD supervisor Tzachi Pilpel for his unwavering support. I learned a lot from him as a scientist and a person. It has been fantastic to work with someone so kind-hearted, and with such contagious enthusiasm for the beauty of science and nature. I am very grateful that he granted me such freedom to explore my own interests, and always trusted me, especially when I doubted myself. I hope to carry with me the same open-mindedness and optimism that made him such an exceptional mentor.

This collaboration would not have been possible without the original intuition of Ariel Lindner. Not only did he provide invaluable expertise from a scientific viewpoint, I could always count on him during the inexorable hardships of a PhD. I truly appreciate the efforts that he undertook to make sure that my degree stayed on the right tracks, and the time that he devoted to me.

I want to thank Tami Geiger for her support with the experimental mass spectrometry, and Jürgen Cox for helping with some of the computational aspects of the project and hosting me in Münich for a week. Both of them demonstrated a great sense of patience towards our naïve understanding of mass spectrometry.

This PhD could not have taken place within any other french graduate school than Frontières du Vivant. The flexibility they offered me is unparalleled, and they were unconditionally sympathetic and helpful. I would like to express my gratitude to David Tareste, and especially to Élodie Kaslikowski, for getting me out of trouble more than once.

I want to thank my two reviewers, Rachel Green and Olivier Namy, for the time they will devote to the reading of this manuscript, and Bertrand Cosson for accepting to take part in my jury.

I would like to thank all the past and present members of the Pilpel lab, who gave me a warm welcome from the very beginning and made me feel at home in Israel. I dearly miss already the passionate discussions, the honesty, and the interminable lab meetings. In particular, I would like to thank Orna Dahan and Avia Yehonadav for their help and support with experiments that would otherwise have challenged my limited abilities for wet work. I would also like to thank Dan Bar-Yaakov, who was the main driving force for the RNA editing

paper presented in this thesis. It has always been a lot of fun working with him.
I hope to see again very soon my friends, the climbers, the rugby players and the music nerds, who made this PhD such an enjoyable and rewarding experience.

Finally, I would like to thank warmly both of my parents, who triggered and nourished my curiosity and my appetite for learning from an early age. They always supported me morally (and sometimes materially) along this PhD.

Table of contents:

TABLE OF CONTENTS: 7

INTRODUCTION 9

THE PROKARYOTIC TRANSLATION MACHINERY 11
AMINO ACIDS 11
TRANSFER RNAS 13
AMINOACYL-TRNA SYNTHETASES (AARSS) 14
THE PROKARYOTIC RIBOSOME 15
INITIATION PHASE 16
ELONGATION PHASE 17
TERMINATION AND RECYCLING 18
FROM FOLDING TO DEGRADATION: A PROTEIN'S LIFE CYCLE. 19
CHAPERONE INDEPENDENT FOLDING 19
CHAPERONE ASSISTED FOLDING IN E. COLI 20
DEGRADATION OF NATIVE, MISFOLDED, AND UNFOLDED PROTEINS. 22
AGGREGATION: TOXIC SIDE PRODUCT OF MITIGATION STRATEGY? 23
THE PROTEOSTASIS NETWORK FUNCTIONS AT THE EDGE OF AGGREGATION 23
PROTEIN LOCALIZATION 24
MULTI-PROTEIN ASSEMBLIES 25
MECHANISMS AND RATES OF PHENOTYPIC MUTATIONS 25
TRANSCRIPTION ERRORS 26
FRAMESHIFTING ERRORS 27
READTHROUGH ERRORS 28
PREMATURE TERMINATION 29
SINGLE AMINO ACID MISINCORPORATIONS 30
HOW DOES THE CELL MITIGATE THE EFFECTS OF AMINO ACID SUBSTITUTIONS? 33
MOLECULAR MECHANISMS OF TRANSLATIONAL ACCURACY 33
THE GENETIC CODE MINIMIZES THE EFFECTS OF AMINO ACID SUBSTITUTIONS. 36
ORGANISMS BALANCE THEIR POOL OF TRNAS WITH THE CODONS THEY EXPRESS. 36
TRANSLATION ACCURACY AFFECTS THE EVOLUTION OF PROTEIN SEQUENCES. 38

CHAPTER 1: SYSTEMATIC DETECTION OF AMINO ACID SUBSTITUTIONS IN PROTEOME REVEALS THE MECHANISTIC BASIS OF RIBOSOME ERRORS 40

ABSTRACT 41
INTRODUCTION 42
RESULTS 45
A PIPELINE TO CONFIDENTLY IDENTIFY AMINO ACID SUBSTITUTIONS IN A PROTEOME 45
MOST OF THE HIGH QUALITY HITS ARE BONA FIDE AMINO ACID SUBSTITUTIONS. 47
OVERVIEW OF AMINO ACID SUBSTITUTION LANDSCAPE IN E. COLI 48
A GLOBAL NUCLEOTIDE MISPAIRING MECHANISM FOR TRANSLATION ERRORS 50
E. COLI AND *S. CEREVISIAE* SHARE SIMILAR ERROR PROFILES 51
THE EFFECT OF DRUGS AND AMINO ACID STARVATION ON SUBSTITUTION PATTERNS 52
MISINCORPORATIONS OCCUR AT ERROR-TOLERANT AND RAPIDLY TRANSLATED POSITIONS 55
DISCUSSION 59

MATERIAL AND METHODS **63**
STRAINS AND GROWTH CONDITIONS 63
PROTEOME EXTRACTION 63
SCX FRACTIONATION, HPLC AND MASS SPECTROMETRY 64
COMPUTATIONAL METHODS 64
THE DEPENDENT PEPTIDE SEARCH 64
DP IDENTIFICATIONS FILTERING 65
ERROR RATE QUANTIFICATION 65
EVOLUTIONARY RATES COMPUTATION 66
EFFECT OF SUBSTITUTIONS ON PROTEIN STABILITY 66
RIBOSOME DENSITY COMPUTATION 66

CHAPTER 2: RNA EDITING IN BACTERIA RECODES MULTIPLE PROTEINS AND REGULATES AN EVOLUTIONARILY CONSERVED TOXIN-ANTITOXIN SYSTEM **68**

APPENDIX: PREDICTION OF IONIZATION EFFICIENCY FROM AMINO ACID COMPOSITION **77**

REFERENCES: **82**

Introduction

Proteins enable most of chemical reactions in cells. They serve as an interface between the information world, stored in an organism's DNA sequence, and its chemical environment. They typically fold into distinct patterns, dictated by their amino acid sequence, which is itself defined almost deterministically by their DNA coding sequence. These folds in turn create a dynamic, three dimensional environment, locally decreasing the energetic barrier of specific chemical reactions, and potentially allowing thermodynamically unfavorable reactions by coupling them to favored ones. By doing so, they open the realm of an out of equilibrium chemistry necessary for the appearance of the order which characterizes living organisms. The regulation of their expression levels, determined by a constant sensing of the environment, offers a formidable way for organisms to navigate through the space of possible metabolisms, and thus to fine-tune their inner-workings to the available resources or challenges they encounter, towards the goal of generating all the necessary building blocks to the replication of the organism.

Proteins are synthesized by polymerization of amino acids using mRNA as a template, through a process called translation. The order in which different amino acids are assembled is crucial, and will eventually determine the 3D conformation of the protein, and its function. Whereas replication and transcription can take advantage of simple base-pairing rules to ensure that the information stored in the DNA in faithfully transmitted over time, translation pairs any of the 64 possible triplets of RNA bases, or codons, to one of the 20 types of amino acids (sense codons) or a translation termination signal (stop codons).

This matching relies on a complex machinery. First, free amino acids are linked to small RNA molecules called tRNAs (transfer RNAs) by a set of proteins, the aminoacyl-tRNA synthetases (aaRS). tRNAs share a common core 3D structure, and their identity is defined by a triplet of bases on one of their loops, the anticodon. tRNAs loaded with an amino acid, (aminoacyl-tRNA) are then ready to enter the ribosome, a large molecular machine composed of RNA and proteins. After an initiation phase in which the ribosome positions itself at the beginning of the mRNA's coding sequence and starts a polypeptide chain, it proceeds to elongate this chain by ratcheting along the mRNA three bases at a time, and matching the newly examined codon to an aminoacyl-tRNA with the complementary anticodon. It evaluates the validity of the codon anticodon

match by probing the stability of the base pairing between the two RNA segments.

The correspondence table between codons and amino acids, dubbed the genetic code, appears to be near universal, and offers the intriguing property of being error tolerant – single-letter DNA mutations will lead to either no changes in the encoded amino acid (synonymous mutation) or a substitution to a chemically similar amino acid. Most amino acids are encoded by more than one codon, and bioinformatic studies have shown that, despite being interchangeable in theory, the frequencies of synonymous codons deviate significantly from the expectations from mutational biases alone in an organism-dependent manner, a phenomenon called codon usage bias (CUB). In particular, the intensity of CUB correlates with gene expression levels.

Several hypotheses have been put forward to explain CUB. In higher organisms, it is generally accepted that selection plays less of a role due to typically small effective population sizes. Codons frequencies are therefore best explained by mutational biases, with local preferences of one codon over another deriving from fluctuations of the mutational spectrum. However, in organisms with large effective population sizes, the very small fitness effect of choosing one codon over another can be selected for. The cause of these fitness effects is a topic of debate. Some claim that codons are primarily selected for speed: codons supported by a large number of tRNA genes tend to reduce the ribosome's waiting time, and thus allow for a better use of this costly machinery. It has also been suggested that slower codons are selected in the 5' end of highly expressed genes, in order to space translation initiation events and to prevent the occurrence of downstream molecular "traffic jams". Similarly, the translation of linkers between protein domains appears to be slower, to let these domains fold sequentially. Additionally, RNA structure requirements can constrain the identity of bases in the coding sequence. Another school of thoughts holds the view that some codons are more accurate than others, and are therefore enriched in the coding sequence of highly expressed genes in order to mitigate the deleterious effects of erroneous protein synthesis. The driving force behind this phenomenon is selection against misfolding, as misfolded proteins tend to be dysfunctional, generate spurious protein-protein interaction, and saturate the protein quality control machinery.

Whereas DNA polymerases typically make a mistake every 10^9 to 10^{10} bases, proteins are synthesized at a much higher error level, with current estimates ranging between 10^{-3} and 10^{-4} errors per inserted amino acid. This high error rate implies that a sizeable fraction (~15%) of a population of typical 200 aa long proteins contains at least one mistake. As a result, protein sequences have evolved to be robust to most single amino acid changes, and these constraints limit the choice of codons, in turn funneling the proteins' potential evolutionary paths. A recent trend even suggests that controlled levels of mistranslation can be beneficial to the organism's fitness. Mistranslation selectively affecting a codon or group of codons was shown to help parasitic cells evading their host's immune system, or deal with oxidative stress. In extreme cases of adaptive selection, low abundance mistranslated proteins can be selected for their ability to solve a problem better than the native sequence, thereby indirectly favoring sequences whose mutational neighbors have higher fitness.

In this introduction, I will outline the players and mechanisms of the prokaryotic protein translation, with a particular focus on *Escherichia coli*, and the various ways these mechanisms can fail and lead to errors. I will describe the evolutionary pressures that shape the evolution of the translation machinery and protein sequences, and review previous attempts to estimate the rates and spectrum of phenotypic errors.

The prokaryotic translation machinery

Here, I will present the main actors of the translation process in Prokaryotes, and review their structural characteristics, functions and involvement in the different stages of translation. I will then briefly present the mechanisms of the three core stages of mRNA translation by the ribosome (initiation, elongation and termination) in the case of faithful translation.

Amino acids

Amino acids are small organic molecules characterized by a carboxylic acid (-COOH) and a primary or secondary amine group (-NH2 or -NH). The general geometry of these molecules is depicted in Figure 1A. The R group, in magenta, is called the residue, and can theoretically be any group of atoms. The α-carbon bearing the residue is chiral, but Life has universally preferred the L geometry represented in the figure to the D stereoisomers. The

carboxylic acid end of an amino acid (C-terminus) can react with the amine group of another (N-terminus) and form a peptide bond (Fig. 1B), releasing a water molecule in the process. This condensation reaction offers a natural way to polymerize amino acids. The peptide bond is stable under cellular conditions, and structurally rigid: the 6 atoms within the dashed box all lie on the same plane, thus restricting the number of possible conformations of the resulting peptide.

Proteins are usually composed of a combination of the 20 proteinogenic amino acids presented in Fig. 1C. The 20 amino acids are usually grouped by chemical properties: their charge and polarity will affect their ability to form hydrogen bonds with the surrounding water molecules, while their volume and 3D conformation will restrict the flexibility of the peptide chain.

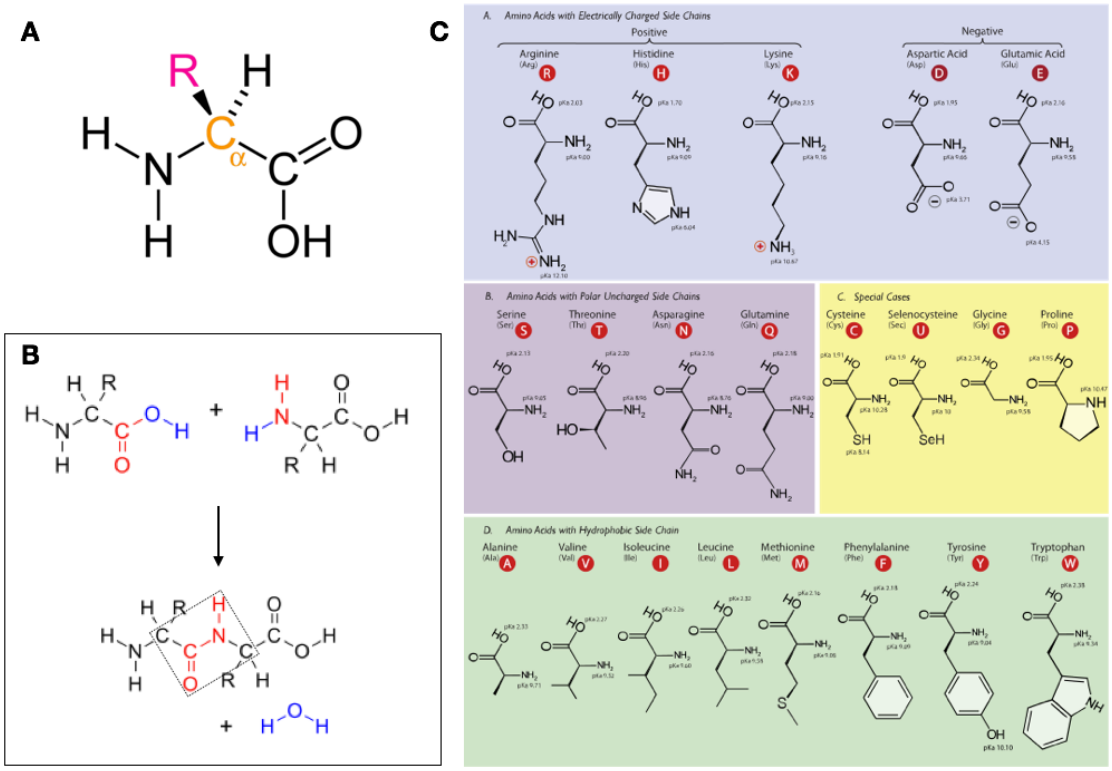
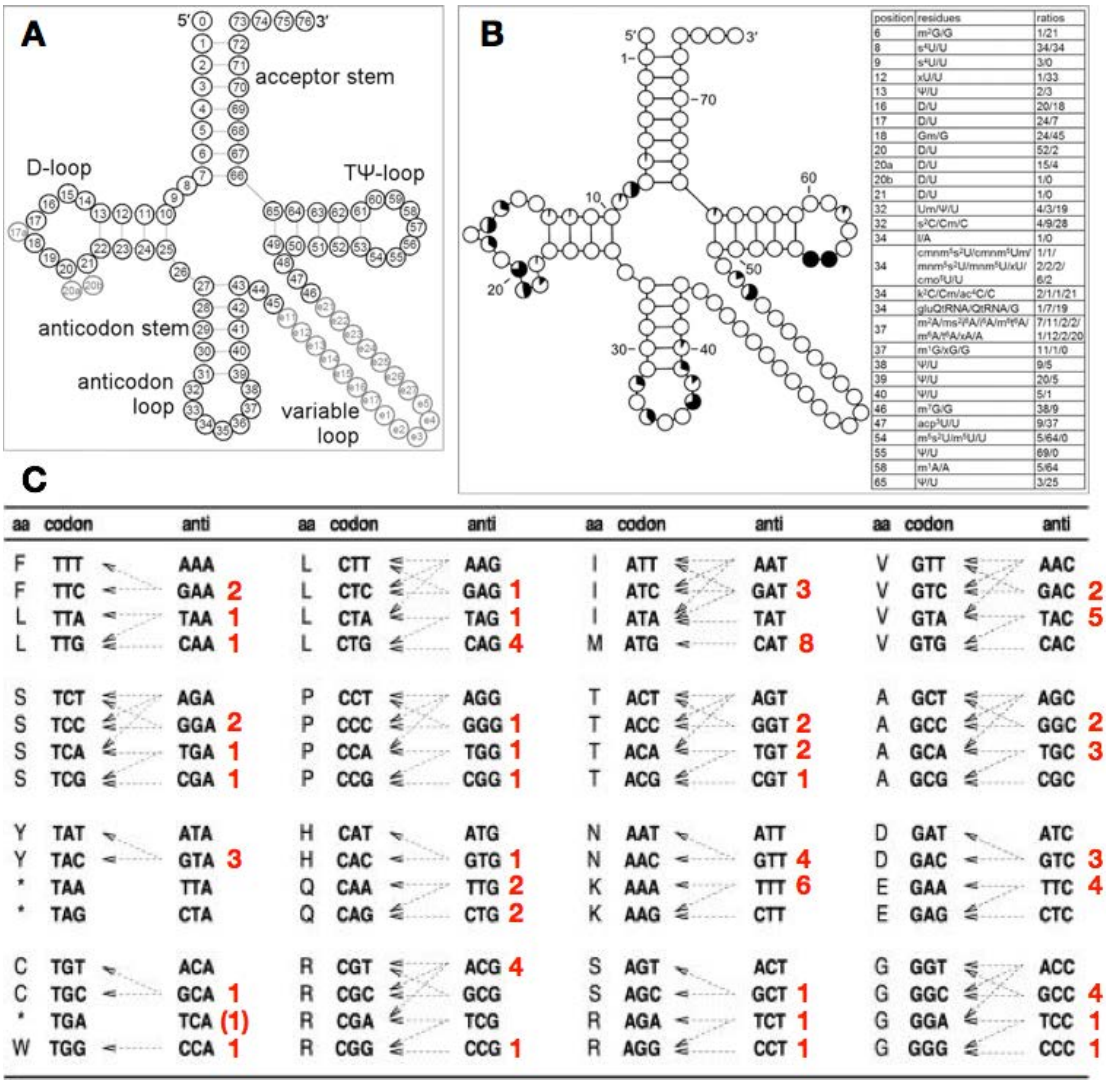


Figure 1: structure and properties of L-amino acids. A: General structure of an amino acid. The asymmetric carbon in green is called the α-carbon . The R group, in magenta, is called the residue. **B:** Formation of the peptide bond. The carboxyl group of amino acid 1 interacts with the amine group of amino acid 2, forming a peptide bond and releasing water. All atoms within the rectangle lie on the same plane. **C:** Structure and properties of the proteogenic amino acids. All figures adapted from Wikipedia.

Transfer RNAs

Transfer RNAs, or tRNAs are short (70-80 nt) RNA fragments which serve as adapter molecules during translation. They are characterized by a shared general "cloverleaf" structure (Fig. 2A), which allows them to be non specifically recognized by different players of the translation machinery. The middle loop harbors a 3-nt sequence called the anticodon, which determines the identity of the tRNA. A tRNA bearing a given anticodon will be loaded with the appropriate amino acid by a set of enzymes called aminoacyl-tRNA synthetases (aaRS). Later, it is this anticodon that will allow the ribosome to test whether it is inserting the appropriate amino acid during translation, by assessing the stability of the base pairing between codon and anticodon. tRNAs represent as much as 15% of the RNA molecules of the cell, and it is generally accepted that there relative intra-cellular abundance closely matches the tDNA gene copy number in the organism's genome. They are typically long lived, and can serve many rounds of translation. In E. coli, the 61 sense codons are served by only 39 different tRNA types (Fig. 2C). This implies that, despite all of the 20 amino acids being associated to at least one tRNA type, some codons cannot be translated by a perfectly matching tRNA. Dotted arrows in figure 2C represent the general matching rules in prokaryotes. In addition to these canonical tRNAs, which all serve a similar role as adaptor molecules during translation elongation, *E. coli* also harbors a distinct class of tRNA for translation initiation, $tRNA^{f\text{-}Met}$, and is able to conditionally insert the non-canonical amino acid Selenocysteine at amber stop codons (UAG) via a suppressor $tRNA^{SelCys}$.

Premature tRNA transcripts undergo several modifications before serving their role in translation. Figure 2B summarizes the post-transcriptional modifications known to affect tRNAs in Gram-negative bacteria. These modifications stabilize the 3D structure of the molecule, affect its interaction with other players of the translation machinery, or even fine-tune its ability to base pair with cognate and near cognate codons when they occur directly on the anticodon loop.



position	residues	ratios
6	m^2G/G	1/21
8	s^4U/U	34/34
9	s^4U/U	3/0
12	xU/U	1/33
13	Ψ/U	2/3
16	D/U	20/18
17	D/U	24/7
18	Gm/G	24/45
20	D/U	52/2
20a	D/U	15/4
20b	D/U	1/0
21	D/U	1/0
32	$Um/Ψ/U$	4/3/19
32	$s^2C/Cm/C$	4/9/28
34	I/A	1/0
34	$cmnm^5s^2U/cmnm^5Um/mnm^5s^2U/mnm^5U/xU/cmo^5U/U$	1/1/2/2/2/6/2
34	$k^2C/Cm/ac^4C/C$	2/1/1/21
34	gluQtRNA/QtRNA/G	1/7/19
37	$m^2A/ms^2i^6A/i^6A/m^6t^6A/m^6A/t^6A/xA/A$	7/11/2/2/1/12/2/20
37	$m^1G/xG/G$	11/1/0
38	Ψ/U	9/5
39	Ψ/U	20/5
40	Ψ/U	5/1
46	m^7G/G	38/9
47	acp^3U/U	9/37
54	$m^5s^2U/m^5U/U$	5/64/0
55	Ψ/U	69/0
58	m^1A/A	5/64
65	Ψ/U	3/25

aa	codon	anti	aa	codon	anti	aa	codon	anti	aa	codon	anti
F	TTT	AAA	L	CTT	AAG	I	ATT	AAT	V	GTT	AAC
F	TTC	GAA 2	L	CTC	GAG 1	I	ATC	GAT 3	V	GTC	GAC 2
L	TTA	TAA 1	L	CTA	TAG 1	I	ATA	TAT	V	GTA	TAC 5
L	TTG	CAA 1	L	CTG	CAG 4	M	ATG	CAT 8	V	GTG	CAC
S	TCT	AGA	P	CCT	AGG	T	ACT	AGT	A	GCT	AGC
S	TCC	GGA 2	P	CCC	GGG 1	T	ACC	GGT 2	A	GCC	GGC 2
S	TCA	TGA 1	P	CCA	TGG 1	T	ACA	TGT 2	A	GCA	TGC 3
S	TCG	CGA 1	P	CCG	CGG 1	T	ACG	CGT 1	A	GCG	CGC
Y	TAT	ATA	H	CAT	ATG	N	AAT	ATT	D	GAT	ATC
Y	TAC	GTA 3	H	CAC	GTG 1	N	AAC	GTT 4	D	GAC	GTC 3
*	TAA	TTA	Q	CAA	TTG 2	K	AAA	TTT 6	E	GAA	TTC 4
*	TAG	CTA	Q	CAG	CTG 2	K	AAG	CTT	E	GAG	CTC
C	TGT	ACA	R	CGT	ACG 4	S	AGT	ACT	G	GGT	ACC
C	TGC	GCA 1	R	CGC	GCG	S	AGC	GCT 1	G	GGC	GCC 4
*	TGA	TCA (1)	R	CGA	TCG	R	AGA	TCT 1	G	GGA	TCC 1
W	TGG	CCA 1	R	CGG	CCG 1	R	AGG	CCT 1	G	GGG	CCC 1

Figure 2. A: Consensus tRNA secondary structure presented in the "cloverleaf" form with the universal numbering system. **B:** Modification profile for tRNA sequences from Gram-negative bacteria (69 sequences from 8 species). The pie charts within each position in the cloverleaf correspond to the percentage of all modified nucleosides (modified being drawn in black). In the tables the series of numbers next to the series of symbols indicate the frequency of occurrence of listed nucleosides at the particular position. A and B were reproduced, and their legends adapted from Machnicka *et al.*, 2014[1]. **C:** Genetic code and general codon–anticodon recognition rules for tRNA genes, and tDNA gene copy number in *E. coli* (in red). Figure adapted from Dos Reis *et al.*, 2004[2].

Aminoacyl-tRNA synthetases (aaRSs)

aaRSs catalyze the specific loading of an amino acid to the 3′ end of its cognate tRNA. They use the energy of a phosphate bond to catalyze the aminoacylation reaction. In a first step, the enzyme binds an ATP molecule and its cognate amino acid to form an aa–AMP complex, releasing a phosphate ion in the process. It then recognizes its cognate tRNA molecule, transfers the amino acid

to the last tRNA nucleotide (position 76), on its 2′ or 3′ end, and releases the AMP molecule. aaRSs are divided into two evolutionarily distinct classes that differ by the structure of their catalytic domain. Class I enzymes, responsible for the tRNA aminoacylation of Cys, Ile, Leu, Met, Val, Arg, Gln, Glu, Trp and Tyr, bind the tRNA acceptor helix on the minor groove side, and can load the amino acid on the 2′ and 3′ –OH groups of the respective tRNA, with a preference for the 2′ -OH. Except for TrpRS and TyrRS, which work as dimers, the rest of class I synthetases are monomeric. Conversely, class II synthetases (Gly, His, Pro, Ser, Thr, Asn, Asp, Lys, Ala, Gly and Phe) usually work as monodimers or tetramers, bind the acceptor helix of the tRNA from the major groove site, and generally load the amino acid on the 3′ -OH. In *E. coli*, there is one aaRS gene for each of the 20 amino acids, with the exception of lysine, which is associated to two genes (lysS, constitutively expressed, and lysU, induced during heat shock). The genes are scattered across the genome, and typically expressed at similar relative concentrations[3]

The prokaryotic ribosome

The ribosome is a large molecular complex of ribosomal RNAs (rRNA) and a number of ribosomal proteins, which serves as a catalytic hub for the process of translation. It is made of two subunits, named 50S and 30S in prokaryotes after their characteristic sedimentation rate in Svedberg units. The large (50S) subunit is composed of 33 proteins and two rRNA fragments, called the 23S (2904 nt) and 5S (120 nt) rRNAs. The small (30s) subunit is made of a single 16S rRNA (1542 nt) and 21 proteins. The full ribosome (70S) is around 20 nm in diameter, and can be found bound to cytoplasmic mRNAs, where it translates cytosolic proteins, or to the inner membrane via the signal-recognition-particule's receptor, for the translation of inner membrane, periplasmic, outer membrane and secreted proteins. mRNAs are commonly translated by more that one ribosome, forming a complex called a polysome. The 50S subunit contains three cavities capable of accommodating tRNAs: the A-site (Aminoacyl-tRNA binding site) performs the tRNA selection step by probing that the tRNA anticodon matches the codon under scrutiny, the P-site (Peptidyl-tRNA binding site) holds the peptidyl-tRNA attached to the nascent polypeptide, and the E-site (Exit site) hosts the uncharged tRNA after the transfer. The part of the ribosome that catalyzes the addition of the new AA to the nascent peptide chain, called the Peptidyl Transferase Center (PTC), is situated between the A-site and the P-site, and leads to the ribosome exit tunnel, from which the peptide chain will eventually emerge and be released.

Despite the ribosome being usually presented as a monolithic complex with fixed stoichiometry of its different components, several lines of evidence have suggested that it actually adapts its composition in response to environmental cues. In particular, it is long known that E. coli ribosomes purified at various growth rates differ slightly in their proteins' ratios[4]. Similarly, the seven rRNA operons of E. coli are not perfectly identical in sequence, and are differentially regulated. Finally, rRNA and ribosomal proteins are subjected to post-transcriptional and post-translational modifications in a condition-dependent manner[5]. Taken together, these observations suggest that the cell might harness ribosome heterogeneity to fine-tune translation.

Initiation phase

The ribosome, an initiator $tRNA^{f\text{-}Met}$, and three proteins (IF1, IF2 and IF3) are the molecular players of the initiation stage. First, the initiator $tRNA^{f\text{-}Met}$, which is structurally distinct from the elongator $tRNA^{Met}$, is charged with a methionine by the MetRS. The Met-$tRNA^{f\text{-}Met}$ complex is then recognized by a methionyl-tRNA formyltransferase (MTF), which formylates the bound methionine.
Initiation Factor 3 (IF3) recognizes an inactive 70S ribosome, and promotes the dissociation of the two subunits. Initiation Factor 1 (IF1) binds to the base of the A-site of the 30S subunit and helps the dissociation. Initiation Factor 2 (IF2), fMet-$tRNA^{f\text{-}Met}$ and the mRNA proceed to associate with the 30S subunit in a random order to form the 30S pre-initiation complex. A base pairing interaction between the Shine-Dalgarno sequence of the mRNA and the anti Shine-Dalgarno sequence of the 16S rRNA mediates the recognition of the mRNA by the 30S subunit, and directs it towards the 5' end of the mRNA, usually 8-nt upstream of the AUG codon indicating the start of the coding sequence. The fMet-$tRNA^{f\text{-}Met}$ complex is positioned in the P-site, and, following a conformational rearrangement that promotes an interaction between the tRNA and the start codon, IF1 and IF3 are ejected. IF2 facilitates the association of the 30S initiation complex to a free 50S subunit, hydrolyzing a GTP molecule in the process, and leaves the newly formed 70S initiation complex[6].

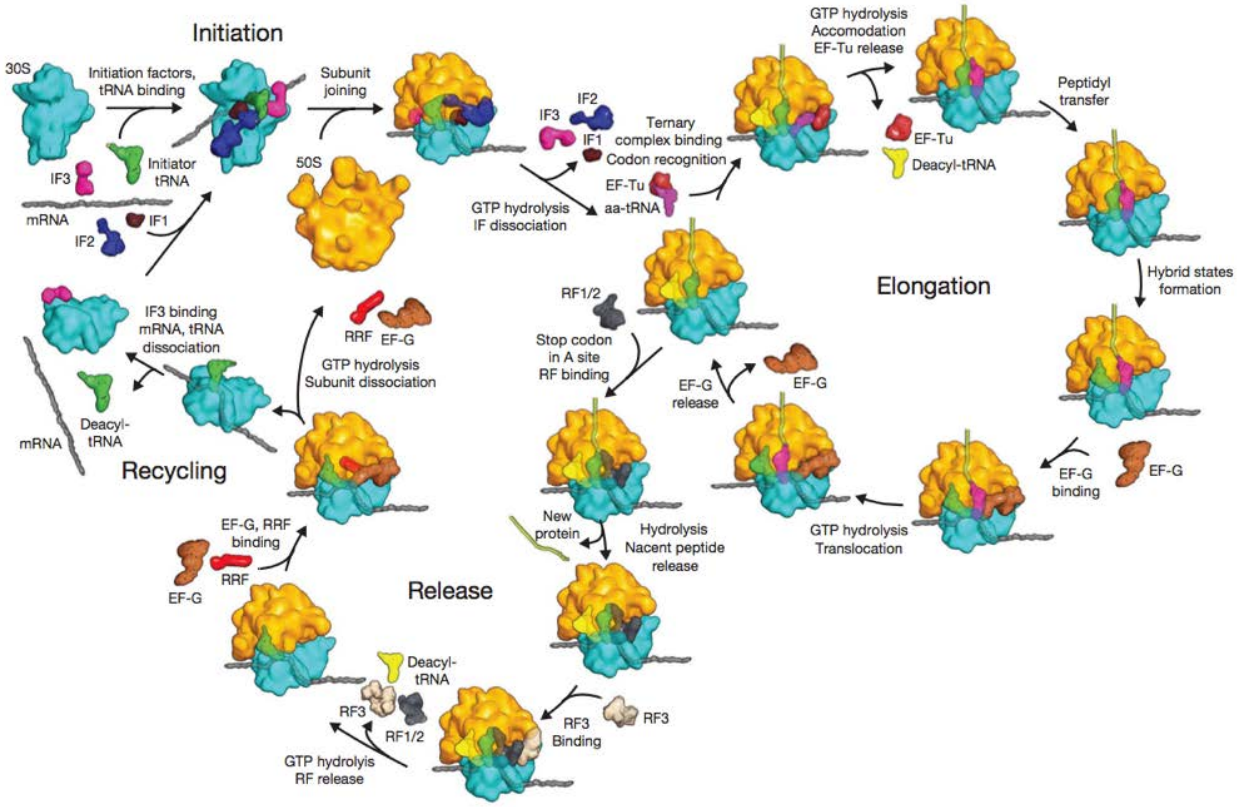


Figure 3 : overview of the different stages of prokaryotic translation. Figure reproduced from Schmeing & Ramakrishan, 2009

Elongation phase

Following initiation, the 70S initiation complex is bound to the mRNA with the start codon facing the $tRNA^{f\text{-}Met}$ in the P-site, while the A-site is empty. The ribosome can now start elongating the peptide chain by repeating the following elongation cycle. Elongation Factor Thermo unstable (EF-Tu) first binds a free aa-tRNA and a GTP molecule. These complexes will repeatedly enter the A-site until an aa–tRNA•EF-Tu•GTP molecule whose anticodon matches the codon is found. The ability of the ribosome to discriminate between cognate and non cognate tRNAs relies on differences in free energy between correct and incorrect codon-anticodon matches, and its accuracy is further improved by the addition of an irreversible step (the hydrolysis of the GTP molecule), through a mechanism called kinetic proofreading (KPR), whose role will be discussed in details in a dedicated section. Once a aa–tRNA•EF-Tu•GTP complex has been accepted, GTP is hydrolyzed by EF-Tu, which is itself released, and the remaining aa-tRNA complex is moved to the (PTC). Following release, EF-Tu•GDP transfers its GDP to another elongation factor,

EF-Ts, binds a new GTP molecule, and dissociates from EF-Ts, allowing it to bind a new aa-tRNA. The ribosome enters the peptide-bond formation step, in which the amine group of the amino acid in the A-site nucleophilically attacks the ester carbon of the peptidyl-tRNA in the P-site, in a step is catalyzed by the 23S rRNA. Eventually, the peptide chain is transferred from the tRNA in the P-site to the aa-tRNA in the A-site. This transfer enables a translocation step, in which the ribosome will reposition itself by exactly one codon towards the 3' end of the mRNA, and the A-site peptidyl-tRNA moves to the P-site, while the P-site tRNA is transferred to the E-site. First, the two tRNAs move with respect to the 50S subunit: the "head" of the P-site tRNA rotates towards the E-site, the ribosome undergoes a conformational change called ratcheting, which is stabilized by the binding of the elongation factor G GTPase (EF-G) to the 30S subunit A-site. EF-G replaces the A-site peptidyl-tRNA and pushes it towards the P-site, while the tRNA in the P-site is transferred to the E-site. Following GTP hydrolysis, the 30S subunit ratchets, and moves together with the mRNA. EF-G dissociates from the ribosome, which can then proceed to another cycle of elongation, or terminate translation if it reaches a stop codon.

Termination and recycling

After many cycles of elongation, the ribosome should eventually reach one of the stop codons (UAA, UGA and UAG). Two release factors, RF1 and RF2, perform the recognition of the stop codon; both RF1 and RF2 can recognize the UAA stop, but UAG is only read by RF1, and UGA only by RF2. They enter the A-site, and interact with PTC via a conserved GGQ motif, exposing the ester bond between the tRNA and the nascent peptide chain to a nucleophile attack by a water molecule. The glutamine from the GGQ motif stabilizes the deacylated P-site tRNA, thus favoring the reaction. The newly synthesized protein is released, and a third release factor, RF3, binds and destabilizes the ribosome•RF1/RF2 complex. RF3 hydrolyses a GTP molecule, and both RF3 and the tRNA in the E-site dissociate from the ribosome-mRNA complex, which is left with an empty A-site and E-site, and a deacylated tRNA in the P-site. In order to recycle the ribosome, EF-G and the ribosome recycling factor (RRF) bind the remaining complex, and promote the dissociation of the two subunits. The 30S subunit, still bound to the decylated tRNA and the mRNA, finally dissociates from these two molecules thanks to the action of IF3, and a new full cycle of translation can start again.

From folding to degradation: a protein's life cycle.

Before accomplishing its function, a protein must first undergo several steps, which start with folding into a defined 3D structure, but might also include targeting to a specific location within the cell, undergoing post-translational modifications, and assembling into complexes. Proteins are assisted in their folding by a suite of proteins called chaperones. They are eventually diluted by growth, and those that fail to fold properly are preferentially degraded to recycle their amino acids or simply aggregated to mitigate their toxic effects. In this section, I will review our current knowledge of these different steps, as it is important them to understand them to fully comprehend the effects of translation errors.

Chaperone independent folding

Due to its wide use in recombinant protein production, protein folding has been extensively studied in *E. coli*. It is known since Anfinsen's experiment in 1973 that, at least for small globular proteins, "the three dimensional structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other components such as metal ions or prosthetic groups) is the one in which the Gibbs free energy of the whole system is lowest; that is, that the native conformation is determined by the totality of inter-atomic interactions and hence by the amino acid sequence, in a given environment[8]." Despite the astronomical number of confirmation that can be adopted even by a relatively small peptide chain, folding can happen at very fast time-scales, on the order of milliseconds, an observation dubbed Levinthal's paradox. Levinthal himself noted that the paradox could be easily resolved if "protein folding [was] sped up and guided by the rapid formation of local interactions which then determine the further folding of the peptide; this suggests local amino acid sequences which form stable interactions and serve as nucleation points in the folding process[9]", *i.e.* if folding happened sequentially. Proteins typically fold in a way that hides hydrophobic, sticky residues in their core, while their surface harbors by more hydrophilic residues. Unfolded or misfolded proteins tend to expose hydrophobic residues, thereby increasing the risk of disturbing cellular processes through spurious protein-protein interactions and aggregation in the crowded intra-cellular environment. Most natural proteins were measured to exhibit a difference in

free energy between folded and unfolded states ΔG on the order of 5-10 kcal/mol[10]. Assuming thermodynamic equilibrium, the ratio of unfolded to folded proteins is given by the formula $\frac{P_{unfolded}}{P_{folded}} = e^{\frac{-\Delta G}{kT}}$, where k is Boltzmann's constant (k = 1.986 cal/mol/K). For a typical ΔG = 5 kcal.mol^{-1}, this ratio is approximately 2.9 × 10^{-4} at 37°C, indicating that the folded vastly outnumbers the unfolded (or misfolded) forms.

Because folding happens fast and co-translationally, the nascent peptide chain can start folding as soon as it exits the ribosome tunnel. Proteins are organized into independently folding subunits called domains. Evolutionary evidence revealed that cells slow down translation of the regions between domains called linkers[11], thus favoring the formation of stable partial structures during the elongation of the nascent chain.

Chaperone assisted folding in E. coli

Despite the fact that protein folding is a thermodynamically favorable process, it is assisted by the action of the trigger factor protein (TF) in *E. coli*. TF, present in dimeric form in the cytoplasm, binds monomerically to the large ribosomal subunit, close to the exit tunnel, and interacts with the nascent peptide chain, as it is still bound to the translation ribosome. *In vivo*, TF preferentially binds ribosomes whose nascent peptide chain reaches at least 100 amino acids in length. It recognizes motifs of 8 amino acids enriched in hydrophobic or basic residues, and aids de novo folding through ATP-independent cycles of binding and release from both the ribosome and the nascent chain, until hydrophobic residue are effectively positioned in the core of the nascent protein and inaccessible to the TF monomer. Approximately 70% of proteins undergo TF-assisted folding, whereas the remaining 30% require the action of additional chaperones (Fig 4A).
These proteins include, in *E. coli*, the Hsp70 chaperone DnaK and its co-chaperones DnaJ and GrpE, and the Hsp60 chaperonin GroEL/GroES system. DnaK assists the folding of ~700 mostly cytosolic proteins. Similarly to TF, it binds 5-7 long stretches of amino acids enriched for hydrophobic residues, and interacts with them through cycles of ATP-dependent binding and release. DnaJ identifies misfolded proteins and transfers them to DnaK, and stimultates DnaK binding through ATP hydrolysis. GrpE releases ADP from DnaK, which upon binding a new molecule of ATP will dissociate from its substrate, thus completing the cycle. Conversely to TF, however, these enzymes can function co-translationally and post-translationally. Approximately 250 different protein

substrates can interact with GroEL/GroES, but it is only necessary for the folding of ~85 of them, thanks to redundancy with the other chaperones. The GroEL chaperonin complex serves as a molecular cage for protein folding, which means it can only function post-translationally. It is composed of two rings of 7 subunits, and interacts with its co-chaperonin GroES, which closes the cage as a "molecular lid". Substrates are bound at multiple sites, and concertedly released by the 14 subunits (*cf.* Fig. 4C). Since only one substrate protein is allowed at once in the cage, the GroEL/GroES system prevents aggregation of the substrate proteins[12].

Finally, ClpB is a stress-induced chaperone whose function is to process aggregates. It works together with the DnaK, DnaJ and GrpE enzymes: clpB binds to aggregated proteins and addresses them to DnaK for resolubilization after heat shock. For a complete review of the role of chaperones in *E. coli*'s protein folding, see Kim et al. 2013[13].

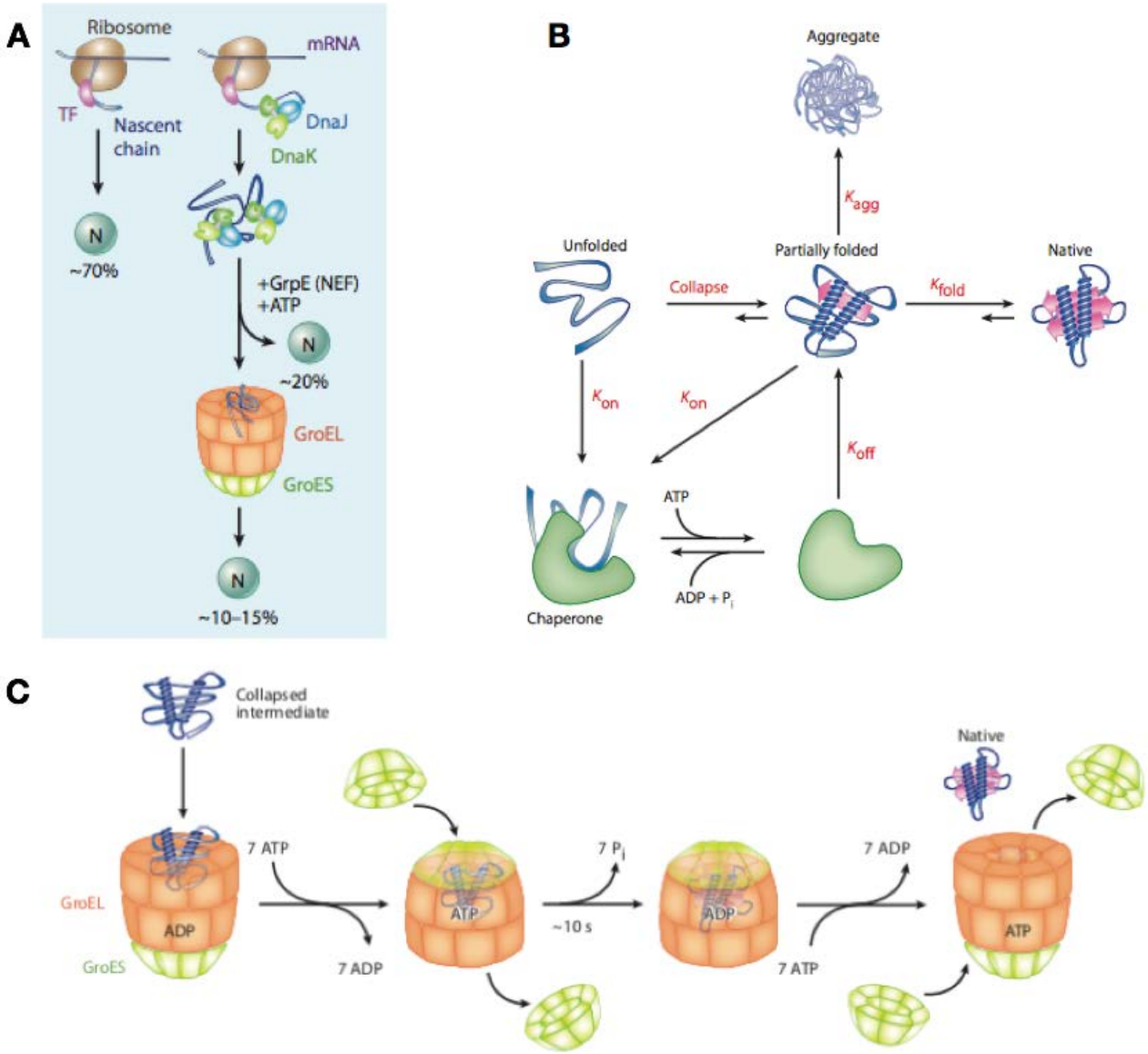


Figure 4. A: E. coli folding pathways. **B:** General chaperone mechanism. **C:** GroEL/GroES refolding mechanism. Figures reproduces from Kim et al. 2013[13]

Degradation of native, misfolded, and unfolded proteins.

In steady state, a protein's expression level is determined by its production rate, α, its dilution rate β_{dil} and its degradation rate β_{deg}[14]:

$$[prot] = \frac{\alpha}{\beta_{dil}+\beta_{deg}}$$

The dilution rate is directly determined by the growth rate, and inversely proportional to the cells doubling time. At high growth rate, the half-life of

most proteins is greater than the doubling time, which means that degradation does not affect their expression level very much. However, the cell uses degradation to purge misfolded and aggregated proteins, and to perform rapid regulation of protein levels. Distinct proteins carry these two aspects of protein degradation. The Lon protease is in charge of degrading unfolded or misfolded proteins, and requires ATP to unfold and ratchet the misfolded protein through its proteolytic chamber.
The ClpAP proteases are in charge of degrading proteins tagged with a degradation signal (degron). They also require ATP to perform degradation.[15]

Aggregation: toxic side product of mitigation strategy?

As mentioned above, misfolded proteins tend to expose hydrophobic residues that "stick" and perturb protein-protein interactions (PPIs). When the chaperone network is overloaded as a result of proteotoxic stress, these misfolded proteins clump together and form aggregates called inclusion bodies. While these aggregates were initially though to be toxic for the cells, it now appears that they serve as damage control strategy. First of all, if the toxicity of unfolded and misfolded proteins stems from its tendency to disturb PPIs, the cost of misfolding is roughly proportional to the surface of contact of these proteins. However, if these proteins aggregate in a roughly spherical shape, for any increase of volume of the aggregate ΔV, the associated increase in surface in only proportional to $\Delta V^{2/3}$, compared to the linear increase expected from an aggregate free situation. Additionally, it serves as a bet-hedging strategy: during cell division, aggregates are asymmetrically transmitted from the mother cell to the daughter cells, and preferentially associate with the cells older poles[16]. This bet hedging strategy allows a higher growth rate at the population level. Finally, a recent study in *S. cerevisiae* revealed that aggregation is not necessarily restricted to misfolded protein, but rather that it can be a reversible, adaptive strategy in the response to heat shock by momentarily protecting functional proteins[17].

The proteostasis network functions at the edge of aggregation

A computational model of the proteostasis network, at the systems level, revealed that it balanced energy and chaperone utilization efficiently[15,18]. The proteostasis network performs sorting of misfolded proteins in a way that resembles the way hospitals sort patients. The network efficiently addresses

the sickest proteins to the most ATP-expensive chaperones, and the chaperone concentrations are just high enough to keep the proteome from aggregating. Their protein levels are adjusted to the growth rate, and are therefore higher in fast growing bacteria, when the synthesis rate is high, but also at very low growth rate to prevent degradation of misfolded proteins when they cannot be balanced by protein synthesis.
A study in *S. cerevisiae*[19] addressed the direct cost of expression of misfolding-prone proteins by comparing the growth rate of cells expressing a wt-YFP to that of cells carrying destabilized mutants of YFP. When the YFP variants were induced, so that the YFP would represent 0.1% of the total protein content of the cell, the most destabilized mutant suffered from a growth rate reduction of 3.2% compared to the wild type control. This emphasizes the notion that the cost of dealing with misfolding is much higher than the mere synthesis cost of a properly folding protein.

Protein localization

Even though the bacterial cell is much simpler in its organization than its eukaryotic counterpart due to its absence of organelles, it has been estimated that nearly one fifth of *E. coli* proteins is actively targeted to a defined subcellular localization, which include the inner and outer membranes, the periplasm, and the extra-cellular space (secretion)[20]. Even cytosolic proteins can be locally restricted to subcellular localization such as the nucleoid (the region of the cytoplasm in which the DNA is stored and condensed), the z-ring (a short-lived structure indicating the middle of the cell before division) or the cell poles. For most proteins, the localization process appears to be driven by diffusion and binding to "anchor proteins", which are actively directed to these subcomponents of the cell. Approximately 96% of the exportome (proteins targeted to one of the membranes, the periplasm, or the extracellular space) requires the action the prokaryotic translocon[20]. The translocon, embedded in the inner plasma membrane, is a channel composed of proteins SecYEG. Proteins are addressed to the translocon co or post-translationally by the recognition of an n-terminal signal peptide by the SRP and SecA proteins. The signal peptide consists of a stretch of positively charged amino acids, followed by 6-18 hydrophobic amino acids and finally 1-11 polar amino acids containing an cleavage site recognized by the membrane anchored SPase I. Substrates of the Sec translocon can be co-translationally integrated to the membrane, or post-translationally translocated.

Multi-protein assemblies

Most proteins in E. coli carry their function as complexes. These can range in complexity from a simple homo-dimeric form to large hetero complexes. A combination of affinity-purification and mass spectrometry on one side, and co-occurrence of proteins in orthology across species, revealed that the *E. coli* proteome contains more than 400 complexes[21]. Since complexes only function when all of their components are present, the cell developed strategies to express the various members of a complex at the right stoichiometry. First of all, members of a complex are often transcribed from the same operon, and therefore can be simultaneously regulated. Furthermore, recent ribosome profiling data revealed that the synthesis rate of the different proteins within an operon closely match their stoichiometry in the resulting complex[22]. Residues facing the interface of a complex are usually more conserved than other surface residues[23]. In Eukaryotes, dominant negative mutations are often associated with complexes, indicating that a single miscoded protein can inactivate an entire complex[24].

A recent study indicated that even monomeric proteins evolve at the edge of multimeric assembly[25]. Garcia-Seisdedos *et al.* expressed point mutated *E. coli* proteins in vitro and heterologously in *S. cerevisiae*, and observed that in some case they could form up to 1µm long fibrils. Together these results indicate that the oligomerization state of proteins can be easily disturbed, even by single nucleotide mutations.

Mechanisms and rates of phenotypic mutations

Phenotypic mutations are defined as "errors that occur when a DNA coded gene is transcribed to mRNA and subsequently translated to protein[26]". They resemble DNA mutations (insertion, deletion and point mutation), but are not transmittable to the cell's progeny, and typically occur at much higher rates[26]. They can be divided into transcription errors (insertion, deletion, point mutation or spurious editing), and translation errors (amino acid substitution, frameshift, readthrough, and premature termination). I will discuss the molecular mechanisms that lead to these errors, and the various ways cells cope with, and even harness them to adapt to the environment.

Transcription errors

Strand specific RNA-seq technologies have shone light on the imprecise nature of transcription initiation and termination[27]. However, the transcription elongation phase of transcription appears to be mostly devoid of errors, and until recently its errors were too rare to be detected by standard RNA sequencing because of machine errors and mutations introduced during reverse-transcription. Traverse & Ochman[28] applied the CircSeq method[29] to solve this problem and directly measure transcription error rates in *E. coli*. During CircSeq, mRNAs are first circularized, and then repeatedly reverse transcribed, so that the final cDNA contains tandem repeats of the mRNA's sequence. Errors found across several tandem repeats cannot originate from reverse transcription or sequencing mistakes, and are therefore present at the RNA level, regardless of whether they occurred during transcription of after. They measured the rate of nucleotide substitutions, and found it to be fairly constant across conditions, with an average error rate of 5.10^{-5} errors per base, and a tendency to replace C with U (implying a G:U mismatch during transcription). Errors did not localize to the leading or the lagging strand, and were very moderately affected by their local context. In a following article[30], the same authors studied the rate and spectrum of insertion and deletion transcription errors, and found it to be about an order of magnitude lower than the rate of substitutions, with deletions prevailing over insertions. Surprisingly, these deletions to be more likely to preserve the reading frame, with observed error rates peaking for 3 and 6 nt deletions. However, we cannot exclude that these peaks are the results of degradation of mRNAs containing frame-disturbing deletions, as they are likely to cause premature translation termination and trigger mRNA degradation. Insertions, on the other end, appear to happen mostly within repeating sequences, and usually consist in adding an additional repeat.

In order to polymerize at such high level of accuracy, it has been suggested that RNA polymerases are able to backtrack to correct their mistakes. Specifically, "the polymerase jumps forward and backward along the template DNA with a net movement that is driven in the forward direction by thermodynamically favorable nucleotide addition in the forward-translocated state[31]". It not only relies on the base pairing ability between the next template DNA base and the incoming RNA base, but also on the base stacking free energy difference between a correctly inserted nucleotide and a mismatch. A mismatch affects the base stacking energy of the preceding and following nucleotide pairs. This standard free energy difference, *i.e.* the melting energy of the pair of base pairs formed by the two last incorporated nucleotides,

serves to discriminate against mismatches in initial selection. A misincorporation also slows down the incorporation of the next nucleotide. The proofreading stems from the ability of the enzyme to cleave the 5' end of the transcript after backtracking. For a detailed review of the mechanisms and determinants of transcriptional accuracy, see Gamba & Jenkin, 2018 [32].
Despite being relatively rare compared to other phenotypic errors, the effect of transcription errors is amplified by the fact that many copies of a protein can be produced from the same mRNA. This property has been harnessed to engineer bi-stable switches in E. coli that allow epigenetic inheritance[33].

In addition to RNA polymerase errors, other processes can affect RNA sequences post-transcriptionally. 8-oxo-guanine, a derivative of guanine generated when the ribonucleotide pool is exposed to reactive oxygen species, can be created within RNAs, where it affects its base pairing preferences and can induce protein recoding[34]. Similarly, adenosine to inosine and cytosine to uracil RNA editing are now believed to be common across higher eukaryotes. Inosine base-pairs with A, C and U, and was shown to induce protein recoding in human, mouse and zebrafish[35]. In the second chapter of my thesis I will challenge the prevalent notion that RNA editing does not recode bacterial mRNAs.

Frameshifting errors

Translation usually occurs in a defined frame, that the ribosome maintains along the length of the transcript. However, some sequences tend to confuse the ribosome, and induce a slippage towards one of the neighboring frames. In *E. coli*, +1 and -1 nucleotides frameshifts are the most frequent, and have been harnessed by the cell to regulate the production of key enzymes (programmed frameshift), including an interesting case of self regulation of the frameshifting propensity: the expression of the release factor protein RF2, whose primary role is to terminate translation via its recognition of the UAA and UGA stop codons, is stimulated by a +1 frameshift, resulting in the bypass of a UGA stop codon and the production of a full length, functional protein[36].
Other well-studied cases of programmed frameshift in *E. coli* include the dnaX frameshifting element, which regulates the relative expression of the τ and γ subunits of DNA polymerase III by redirecting the ribosome to the -1 frame approximately 50% of the time[37], or the joint production of a copper transporter and a copper chaperone by the same gene in different frames[38].
These cases are usually characterized by specific RNA structures, such as slippery homo repeats and pseudo-knots, which have been selected to

generate high levels of frameshifting. In order to evaluate the basal frameshift error rates in vivo, Meyerovich et al.[39] introduced a plasmid containing a frameshifted GFP in the gram positive bacteria *B. subtilis* and compared the residual fluorescence to that of a wild type GFP. They found that 2% of the GFP encoding a frameshift at the DNA level were able to revert and produce a functional protein. Since the DNA frameshift was inserted near the beginning of the sequence (10th codon), it is likely that this figure represent a cumulative rate of frameshifting over the first codons. Assuming that frameshift errors are evenly distributed across the first 10 codons, the resulting figure of 0.2% errors per codon is still astonishingly high compared to other phenotypic errors, especially since early frameshifts are likely to create non functional, truncated proteins.

Until recently, large-scale detection of programmed frameshift relied on the fact that coding sequences show a decreased conservation of the 3rd codon position (wobble) in evolutionary alignments. Ribosome profiling technologies now generate single nucleotide resolution maps of ribosome density across the transcriptome. The characteristic 3-way periodicity of these profiles reveals the dominant frame in which an mRNA is translated. In their analysis of the translational changes induced by the [PSI+] prion in *S. cerevisiae*, Baudin-Baillieu *et al.*[40] took advantage of this feature to identify frameshifts throughout the transcriptome, and showed that they were stimulated by the prion.

Readthrough errors

Stop codon suppression or translation readthrough occurs when the ribosome bypasses a stop codon and interprets it as a sense codon. Like in the frameshift case, one can divide these errors into basal readthrough errors, which happen at any given stop codon, and programmed readthroughs, which are selected for and potentially regulated. In the strict sense, the definition of a readthrough error only applies to cases where a tRNA competes with the release factors, resulting in the ribosome inserting an amino acid in place of the stop codon, and proceeding to translate until it reaches the next stop codon (or the end of the transcript, resulting in the degradation of the mRNA transcript[41]. However, cases of frameshifts that bypass a stop codon and add a peptide extension using a non canonical frame share similarities with bona fide readthrough, and can be included in a broader definition of the term.

In E. coli, the most archetypal case of programmed readthrough is probably the insertion of selenocysteine at UGA stop codons. A suppressor tRNA bearing the TCA anticodon is first charged with serine by the SerRS, and the selA enzyme converts the Ser-tRNASec to Sec-tRNASec. Insertion of

selenocysteine at UGA only happens at sites where the UGA stop codon is associated to a specific RNA structure, the SECIS element[42].
Ribosome profiling experiments in E. coli revealed that readthrough is a pervasive phenomenon, and estimated that as many as 50 genes showed signs of translating ribosomes in sequences' C-termini[43]. This phenomenon was mostly observed at UGA codons, and stimulated by the deletion of RF2. RF2 depletion also disturbed the expression of biosynthetic genes under attenuation control. Fan et al.[44] used a synthetic reporter construct to assess the variability of readthrough efficiency at UGA codons in a population of *E. coli* cells. The readthrough frequency, around 2% on average, varied substantially from cell to cell, and was correlated with a reduced protein synthesis. High levels of readthrough reduced the lag time necessary to exit stationary phase. The notion that global readthrough can be evolutionarily selected for and adaptive was supported by the finding that it was pervasive and regulated in the fruit fly *D. melanogaster*[45]. The C-terminal extensions were added at specific developmental stages, and often contained localization tags, but were not phylogenetically conserved, suggesting that readthrough is an adaptive mechanism to increase the proteome's plasticity, and might perhaps serve as a "stepping stone" for more complex evolutionary processes. Yanagida *et al.*[46] compared the way different species of yeasts encode their IDP genes (responsible for fatty acids oxidation). They showed that pre whole-genome-duplication (WGD) species encoded IDP with a single gene, which is conditionally addressed to the peroxisome thanks to the addition a localization tag via a regulated +1 frameshift to bypass the canonical stop codon. Post WGD species, on the other hand, simplified this system by differentially expressing two different copies of the gene, with or without the peroxisome tag.

Premature termination

Whenever a ribosome terminates translation before reaching the canonical stop codon of an mRNA transcript, an incomplete protein is formed. Since an incomplete protein is unlikely to perform its function, and will probably misfold, premature termination is considered to be very deleterious, and has been proposed to a major driver of codon usage bias[47]. It can result from a nonsense transcriptional error, in which a stop codon would appear in the mRNA coding sequence, or from ribosome drop-off, a process stimulated by extensive stalling at a codon due to the lack of cognate aa-tRNAs. Conversely to readthrough, ribosome drop-off is too subtle to be directly measured at the gene level by ribosome profiling, but Sin et al.[48] developed a sensitive

statistical procedure to quantify the decrease of ribosome density along transcripts globally, in a range of conditions. They estimated the drop off rate to be on the order of 4×10^{-4} per codon, which implies that for a typical, 300 codon long ORF, only about 90% of the ribosomes would reach the canonical stop codon. Whether the decrease in ribosome density comes from frameshifts leading to out of frame termination or genuine termination events at sense codons, and the fate of these prematurely terminated protein, remains unclear. Zaher & Green[49] revealed an intriguing interplay between amino acid misincorporation and premature termination. They showed that strains deleted for RF3, which was believed to primarily serve in the dissociation of RF1 and RF2 at the end of the translation process, suffered from increased sensitivity to errors in protein synthesis, and that RF3 tended to stimulate premature termination when a the tRNA in the P-site formed a mismatch with its codon. Since frameshifting can, at least transiently, form mismatches in the E and P-sites, RF3 was also shown to mitigate its effects. Finally, premature translation termination by RF3 appears to decrease mRNA stability, but not protein stability.

While ribosomes do not appear to dwell longer on any particular codon type in rich conditions, amino acid starvation has been shown to lift the degeneracy of the genetic code, and to induce ribosomal pausing at codons associated to the amino acid depleted from the medium[50]. The severity of the effected on synonymous codons was well predicted by a simple (but counter-intuitive) model of tRNA charging[51], which showed that tRNAs associated to rare codons were more readily charged than more common isoacceptors during starvation for the associated amino acid. In line with these observations, cassettes expressing YFP in which all codons for a given amino acid were systematically recoded to only one of the codons for this amino acid resulted in measurable differences in fluorescence during starvation for this amino acid[52]. The codon robustness index that they proposed to evaluate the amount of premature termination at a given codon during starvation correlated well with the observed dwelling time of the ribosome, suggesting that premature termination is mostly dictated by competition between aa-tRNAs and release factors.

Single amino acid misincorporations

Until now, our experimental knowledge regarding rates of amino acid misincorporations originates almost exclusively from reporter constructs studies. One of the first reliable estimates of these error rates *in vivo* came from the ingenious experiment of Edelmann & Gallant[53]. They took advantage of the

fact E. coli's flagellin does not encode for any cysteine in its sequence, and fed cells with cysteine marked with the radioactive 35S sulfur isotope. After purifying the protein and running it on a SDS polyacrylamide gel, measuring radioactivity levels allowed them to reveal the amount of 35S-Cys inserted per flagellin. They argued that this insertion was likely occurring at CGU/C arginine codons, and deduced that the error rate of misincorporation of cysteine at these codons was on the order of 1.0×10^{-4}. They also confirmed that this error rate was increased in the presence of small concentrations of streptomycin, and during starvation for arginine, therefore highlighting the role of aa-tRNA competition in determining translation accuracy. Similar tricks were used to estimate specific misincorporation levels in reporter constructs, but were usually limited in scope[54–57].
Kramer & Farabaugh[58] designed a series of firefly luciferases to estimate a wider range of codon-specific error rates. They used the fact that luciferase requires a lysine to be present in position 529 to perform its enzymatic activity, and created luciferase constructs in which codon 529 was systematically mutated to all near-cognate codons and some non-cognate codons. For each of these constructs, the residual luminescence served as a proxy for the rate of misreading of the variable codon by Lys-tRNA^{UUU}. They found error levels to vary widely, and to be mostly determined by competition between cognate and near cognate tRNAs: overexpressing the rare arginine tRNA^{UCU} drastically reduced the Arg→Lys error levels at cognate AGA and AGG codons. Error rate from the 14 near cognate codons to lysine varied by a factor of 10. The highest error levels were associated to U:U or G:U mismatches, as would be expected from the thermodynamics of base pairing. They also characterized the effects of two aminoglycosides, streptomycin and paromomycin, and two ribosomal mutations. The two drugs increased error levels at near cognate codons, and the rpsD mutation increased the errors at already error prone codons. They were able to measure decreased error levels in the rpsL hyper accurate mutant, with error prone codons reverting to background levels of mistakes.

The first peek at the full translation error spectrum was provided by a mass spectrometry analysis of six recombinant proteins purified from E. coli[59]. They first identified canonical, error free peptide, and then use a blind modification search strategy to identified "modified versions" of these peptides. They excluded PTMs and MS artifacts using a set of ad hoc rules, and retained among the remaining identifications those consistent with amino acid misincorporations. The vast majority of the errors they detected could be rationalized as originating from an mRNA/tRNA mismatch, rather than a

transcriptional error or a synthetase error. Furthermore, G_{mRNA}:U_{tRNA} mismatches, but also C_{mRNA}:U_{tRNA} and U_{mRNA}:U_{tRNA} mismatches at the wobble position, were the most frequently observed. They confirmed that the identity of the codons determined its errors by synonymously recoding one of the proteins many times. The error rates they measured from these recoded proteins were well predicted by the nature of the mRNA/tRNA mismatch.

As shown by these experiments, misloading errors, *i.e.* errors in which an aaRS pairs a tRNA to a non-cognate amino acid, are much rarer than ribosomal errors. aaRS tend to be precise, and perform their function with an accuracy in the 10^{-4}–10^{-5} range[60]. However, several cases of regulated, adaptive mistranslation have been reported, which usually take advantage of the tRNA charging step to generate high levels of a specific subset of translation errors. C. albicans, a pathogenic yeast, is part of a clade that reassigned the CUG codon from leucine to serine. However, it is able to partially revert and insert high levels of leucine at this codon during invasive growth, recoding predominantly proteins expressed at its surface. This processed is believed to promote cell-adherence (leucine is more hydrophobic than serine), and evasion of the host immune response thanks to the increased sequence variability[61].

Another well-characterized case of adaptive mistranslation is the controlled misacylation of methionine onto various non-cognate tRNAs during oxidative stress. This phenomenon was observed in E. coli[62], S. cerevisiae[63] and H. sapiens[64]. In mammalian cells, the levels of mismethionylation shoot from 1% to 10% during ROS exposure. The adaptiveness of this phenomenon stems from the ability of methionine residues to protect proteins against ROS-mediated damage.

Similarly, in E. coli, oxidative stress appears to trigger another type of mistranslation. The editing domain of the threonine aaRS (thrRS) is inactivated by the oxidation of a cysteine residue. The modified enzyme is not able to discriminate against serine, which is then inserted at high levels at threonine codons[65]. For a complete review on adaptive mistranslation mechanisms, see Pan 2013[66].

How does the cell mitigate the effects of amino acid substitutions?

Since amino acid substitutions tend to be detrimental to fitness, organisms have developed mechanisms to minimize their error rates, and strategies to ensure that the residual errors are well accepted. Here, I will review the various mechanisms that allow the translation machinery to perform at high accuracy, and the ways its components have co-evolved with mRNA sequences to ensure that amino acid substitutions are minimally disruptive to fitness.

Molecular mechanisms of translational accuracy

During translational elongation, the ribosome repeatedly samples aa-tRNAs from the cytosol, with the help of EF-Tu. The process of aa-tRNA selection is blind, and governed by diffusion only. In order to discriminate between cognate and non-cognate tRNAs, the ribosome has to rely exclusively on the difference in free energy (ΔΔG) between correct and incorrect matches in the A-site. Assuming that the selection process relies on thermodynamic equilibrium, an error rate of 10-4 would imply a difference in free energy between cognate and near cognate tRNA to the A-site codon on the order of 10kT, or 0.5 kcal.mol-1 at a temperature of 37°C, but this value is actually higher than the ΔG associated to the perfect binding of a tRNA to its cognate codon. Hopfield[67] and Ninio[68] independently recognized this contradiction, and both proposed that the accuracy of tRNA selection was in fact increased by the addition of an irreversible step, through a mechanism that they respectively termed kinetic proofreading or kinetic amplification. They correctly identified that GTP hydrolysis, which though to be a wasteful reaction, actually provided the necessary boost in accuracy by introducing irreversibility in the selection process (Fig 5A). Recent advances relying on fluorescence resonance energy transfer (FRET) allowed a precise determination of the different rates constants of the elongation cycle, shinning light on the interactions between the ribosome and the tRNAs in the A-site, and its ability to discriminate cognate from non-cognate tRNA at several steps through conformational changes (Fig 5B, reviewed in Wohlgemuth *et al.,* 2010[69].

Despite of the molecular complexity of the proofreading mechanisms, Banerjee *et al.* employed a simplified, general model of proofreading (Fig 5C) and used the method of first passage processes to model the speed and accuracy of the ribosome, and their relationship to the different rate constants.

They were able to show that the ribosome usually operates in a regime that sacrifices accuracy for speed, *i.e*. that it could easily achieve higher accuracy at a lower speed, by simply reducing its rate of GTP hydrolysis (Fig 5D). A linear trade-off between speed and accuracy was observed within an *in vitro* translation system, in response to variation in the Mg^{2+} concentration[70].

Most aaRS also rely on energy consuming proofreading mechanisms to achieve high acylation accuracy[71]. They typically discriminate well between cognate and non-cognate tRNAs, thanks to information encoded both in the tRNA's anticodon and its backbone structure. Discrimination against non-cognate amino acid relies on a double sieve mechanism: the active site first accepts amino acids chemically similar to the cognate AA, but sterically excludes larger ones. In a second, energy consuming step, the editing site probes the chemical properties of the amino acid in the catalytic site and hydrolyses non cognate amino acids[72]. Despite being quite accurate, tRNA acylation is not perfectly error-proof. In particular, it has been suggested that it would be more difficult for the synthetases to exclude small non-cognate amino acids than larger one. This tendency might be at least partially corrected by EF-Tu's binding preferences. LaRiviere et al.[73] probed the affinity of EF-Tu to correctly and incorrectly loaded aa-tRNAs, in vitro. They showed that EF-Tu binds correctly charged tRNAs within a limited range of affinities, but binds incorrectly charged tRNAs over a much wider range. The binding strength seems to be determined by a linear combination of two factors, one determined by the tRNA backbone and the other by the amino acid. Among the correct matches, a amino acid which binds EF-Tu with a low affinity is usually associated to a tRNA whose backbone binds the elongation factor with high affinity, and vice versa. Therefore, incorrect matches do not benefit from this compensation effect, possibly leading to very strong or very weak binding to EF-Tu. The authors suggested that the cell might use this mechanism as a safeguard against mistranslation.

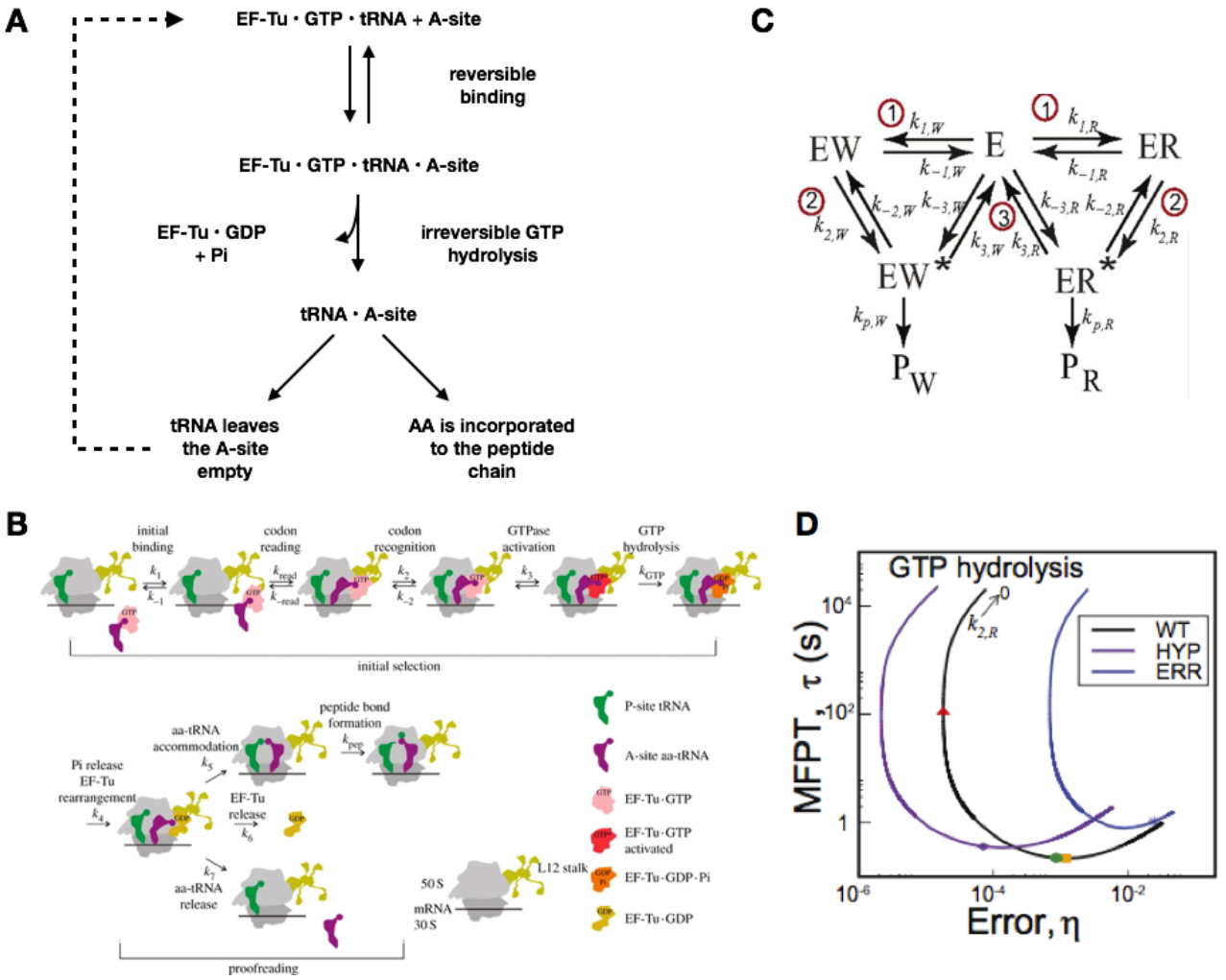


Figure 5: kinetic proofreading in the ribosome. A: General scheme of kinetic proofreading (KPR) in the ribosome. The EF-Tu • GTP • tRNA complex bind the A-site reversibly. Because of the difference in affinity to the A-site between correct and incorrect tRNA, the correct complex will be bound to the A-site after equilibration. The GTP hydrolysis forces the system to either reject the tRNA and repeat the cycle, or incorporate the amino acid in the nascent peptide chain. **B:** State of the art model of tRNA selection in the ribosome. The rate constants were measured for cognate and non-cognate tRNA using FRET. k_{-2}, k_3, k_5 and k_7 depend on the identity of the tRNA, and favor the insertion of cognate over non cognate tRNAs. Reproduced from Wohlgemuth *et al.*, 2011. **C:** Simplified model of KPR used by Banerjee et al. E : free ribosome. ER/EW: ribosome associated with the right (R) or wrong (W) tRNA. ER*/EW*: activated ribosome • tRNA complexes. PR/PW: incorporation of the R or W tRNA in the peptide chain. **D:** The rate of GTP hydrolysis governs a trade-off between speed and accuracy in the ribosome. MFPT : mean first passage time. The black curve represents the simulated speed and accuracy of translation elongation as the rate of GTP hydrolysis tends to 0, using parameters measured on the wild type E. coli ribosome. The blue and purple curves are generated using parameters measured on an error prone and hyper accurate ribosome, respectively. The observed value of $k_{2,R}$ (red circle) indicates that the ribosome usually operates in a regime where speed is optimized rather than accuracy. Reproduced from Banerjee et al., 2017

The genetic code minimizes the effects of amino acid substitutions.

The structure of the genetic code appears is nearly universal, and virtually conserved through all forms of life, with the exception of some minor codon reassignments. Its is characterized by an exceptionally high robustness to point mutations: codons of the same amino acid usually share their first and second positions, and amino acids that share similar chemical properties tend to have similar codons. Although there exist more robust codes, Freeland & Hurst[74] evaluated than only one in a million randomly generated codes (in which the observed sets of codons are randomly reassigned amino acid) would surpass the observed one in terms of robustness to point mutations and translation errors. This result is at first at odds with the notion that the code is near universal, and therefore likely to be poorly evolvable, or as Francis Crick would call it, "frozen". However, it is now believed that the robust properties of the genetic code emerged through the combined effects of co-evolution and selection[75]. First of all, primordial proteins were likely composed of a subset of the current proteogenic amino acids. These primordial amino acids, which were naturally present in the environment, did not require the existence of a complex metabolism. Similarly, primordial translation was likely statistical, and therefore had to be robust to very high error levels. As more complex biosynthetic pathways emerged, the set of codons for an amino acid could be split to encode this amino acid and a newly synthesized one. These two amino acids could, for example, first share an aaRS that would then duplicate and develop an increased specificity towards either one or the other. This process guarantees that amino acids sharing part of their biosynthetic pathways would be encoded by similar codons.

An early genetic code would have likely treated the branched chain amino acids (valine, leucine and isoleucine) interchangeably. Phylogenetic techniques revealed that indeed, there biosynthetic pathways are evolutionarily intertwined and that their aaRS share a common ancestor, and were able to retrace the history of the genetic code's evolution.[76]

Organisms balance their pool of tRNAs with the codons they express.

Systems biology often treats the fluxes in the cell like the fluxes of an economy. Processes like translation seem to obey the law of supply and demand: a pool

of codons needs to be efficiently translated by a pool of tRNAs, and commonly translated codons should be matched with abundant tRNAs. This necessity of matching the codon pool to the tRNA pool is mostly driven by the pressure to translate proteins both fast and accurately. A common codon that is translated by a rare aa-tRNA will induce ribosome stalling, because the ribosome will have to sample many non cognate aa-tRNA complexes before finding the correct one. The ribosome is a large molecular complex in which the cell invested a lot of resources, and assigning it to an inefficiently translated codon represents an opportunity cost: it could be translating other, more efficiently encoded mRNAs during the same time window. Shah & Gilchrist[77] formalized this concept and developed an elegant population genetics based model to explain how codon usage bias results form the conflicting forces of mutation, selection for translational speed, and drift. Their model is based on the assumption that the force of selection counteracts the cost of ribosome stalling, and therefore acts proportionally to a gene's protein synthesis level. Wallace *et al.*[78] modified Shah & Gilchrist's model to account for noisy experimental data, yielding higher estimates of the selection coefficients for "good codons". They rightfully noted that selection could not be assumed to work exclusively against ribosomal stalling, because selection against translation errors would lead to a similar signature. The cost associated to translation errors is indeed likely to be proportional to both the expression level of the genes in which they occurred, and to the time the ribosome spends sampling for the correct aa-tRNA. In order to disentangle the effects of speed and accuracy on the codon usage bias of genes, Drummond & Wilke[79] relied on a statistical test that excluded the confounding influence of gene expression. They hypothesized that cells would likely use good, error-proof codons at positions that are crucial for protein folding, and that these positions would be more conserved in evolutionary alignments. They showed that, within genes, conserved positions were indeed encoded with a different set of codons than non-conserved positions, suggesting that selection for translation accuracy was a major determinant of codon usage bias. Whether cells optimize their translation primarily for speed or accuracy remains an open question. Yang et al.[80] were the first to notice that yeast cells seem to control the trade-off between speed and accuracy by the way they encode their mRNAs. They observed that conserved positions correlated to stronger RNA structures 12nt downstream, consistent with the notion that this structure would slow down the ribosome while it probes the conserved codon in its A-site.

The other side of the supply-demand balance is determined by the expression of tRNA genes, the availability of amino acids, and the activity of aaRS genes. This balance is maintained by processes occurring at physiological and evolutionary timescales. On a rapid, physiological timescale, the levels of free amino acids are tightly regulated: their biosynthesis pathways rely on feedback loops such as transcriptional attenuation[81], and the interconnectivity of the metabolism usually allows cells to efficiently reallocate metabolite fluxes to mitigate the effects of amino acid starvation[82]. Severe amino acid starvation results in the production of ppGpp, a metabolite produced during amino acid induced ribosome stalling. ppGpp in turn activates the stringent response, which redirect resources from high growth rate associated function such as replication, transcription and translation toward amino acid biosynthesis pathways. Impairing the production of ppGpp by the ribosome associated GDPase RelA resulted in a 10-fold reduction of translation accuracy[83]. As seen previously[50–52], synonymous codons are differentially affected when the cell is starved for the associated amino acid. In order to respond efficiently to this challenge, genes involved in the biosynthesis of the depleted amino acid tend to be preferentially encoded with codons that are robustly translated when the tRNA charging levels are low.[51]

On evolutionary timescales, codon usage bias and tDNA gene copy number co-evolve to maintain the balance of supply and demand[2]. Yona *et al.* revealed that the tRNA pool could rapidly adapt in the laboratory and restore the supply-demand balance after they deleted a rare $tRNA_{Arg}^{CCU}$, by spontaneously mutating the anticodon of one copy of the common $tRNA_{Arg}^{UCU}$ to CCU[84]. Bioinformatic analyses revealed that these anticodon reassignments are in fact common, and can even happen between tRNAs decoding different amino acids[84,85]. This suggests that the tRNA pool is extremely plastic, and can rapidly adapt to changes in translation demand. One can suppose that the high evolvability of the tRNA pool implies that it is indeed optimized for a fast and accurate protein synthesis.

Translation accuracy affects the evolution of protein sequences.

Selection pressures can act on the coding sequences of highly expressed proteins to minimize the impact of their translation errors by choosing appropriate codons, but to what extent does translation accuracy affect the evolution of their primary amino acid sequences? In the context of adaptive selection, where mutations are selected because they confer a significant fitness advantage, Whitehead *et al.*[86] investigated the potential role of

phenotypic errors with regard to epistasis. Taking the example of a cysteine bridge in which both cysteines need to be present for the protein to gain activity, they showed that in case of strong selection the intermediate genotypes in which only one cysteine is present could be positively selected because phenotypic errors would lead to a fraction of the protein bearing the two cysteine residues. Bratulic *et al.*[87] tested the effects of translation accuracy on the evolution of a plasmid-borne betalactamase. In order to speed up the evolution of sequences, they performed cycles of *in vitro* PCR mutagenesis, transformation, and plasmid selection based on the fitness advantage they conferred to the host in a medium containing antibiotic. They carried the experiment in parallel a wild type and in an error prone strain. Evolving in an error prone environment conferred sequences a higher folding stability, which was the result of stabilizing non-synonymous mutations on the proteins surface. They did not show sign of synonymous codon selection, and the occurrence of synonymous SNPs at sites where mutations have destabilizing effects was not reduced in error-prone populations. However, despite the large population size, their evolutionary system can hardly be compared to the evolution of natural sequences, as the selection coefficients are very large and the number of generations was limited to about 50. Observing the subtle effects of mistranslation on the evolution of proteins working near optimally would be very difficult in a laboratory setup, and these questions would probably be best addressed through a combination of population genetics and simulation.

Chapter 1: Systematic detection of amino acid substitutions in proteome reveals the mechanistic basis of ribosome errors

Abstract

Translation is limiting the accuracy of information transmission from DNA to proteins. Understanding how cells ensure proper translation of proteins amidst trade-off between accuracy, speed and energy expenditure and whether translation accuracy is modulated across environmental conditions, expression levels or gene locations is largely hindered by lack of a quantitative experimental methods to detect and quantify amino acid misincorporation at the full proteome level. Here we systematically detect and quantify errors in entire proteomes from mass spectrometry data. Following HPLC MS-MS data acquisition, in *E. coli* and in *S. cerevisiae*, we identify peptides whose mass deviate from genome-encoded peptide sequence by one amino acid, verifying that the mass shift cannot be explained by a post-translational modification. Our analysis reveled that most substitutions occur between amino acids that share near-cognate codons. Further analyses suggest that the majority of these near-cognate substitutions occur due to codon-to-anticodon mispairing within the ribosome. Patterns of errors due to mispairing were similar in *E. coli* and yeast, suggesting a universal mechanism that accounts for ribosomal errors. Focusing further on the *E. coli*, we treated the cells with two drugs that decrease ribosomal proofreading and found that they increase error rate due to mispairing at the wobble codon position. Generally, amino acid substitutions tended to occur in positions that are less evolutionarily conserved, and that minimally affect protein energetic stability, indicating a selective pressure to minimize phenotypic errors when potentially detrimental. Genome wide ribosome density data indicate that mistakes tend to occur in sites where ribosome velocity is relatively high, supporting the notion of a trade-off between speed and accuracy as predicted by proofreading theories. Starving the cells for particular amino acids results in specific patterns of amino acid substitutions reflecting the amino acid deficiency. Together our results reveal a mechanistic basis for ribosome errors in translation.

Introduction

Genetic information propagation along the Central Dogma is subject to errors in DNA replication, RNA transcription and protein translation. DNA replication typically manifests the highest fidelity among these processes, featuring genetic mutation rate on the order of 10^{-10} per nucleotide per genome doubling[88,89]. "Phenotypic mutations", i.e. errors in RNA transcription and in protein translation, in which the wrong RNA nucleotide or amino acid are respectively incorporated, occur at considerably higher rate. A recent estimate made in bacteria, is that transcription error rate ranges between 10^{-5} to 10^{-6} per incorporated nucleotide[28]. As for translation, the classical kinetic proof reading theory[67] suggested that the error rate per amino acid would have been extremely high (10^{-2}) at chemical equilibrium, and it is only due to the investment of energy in the form of hydrolysis of GTP that it can be reduced to about 10^{-4} on average.

The two main steps that account for errors in translation are the mischarging – where the wrong amino acid is acylated to a tRNA, or mispairing where a tRNA mispairs with the wrong codon within the ribosome. To reduce errors due to mispairing proofreading is made within the ribosome in a process that consumes energy and that compromises translation speed[90,91]. Mischarging of tRNA with the wrong amino acids is also subject to proof reading working at the aminoacyl tRNA synthetase level[92]. Like in any information channel, translation systems must thus face a "trade-off between energy, speed and accuracy[90].

The heavy investment of cells in proofreading the translation process, in energy and in time is a clear indication that too high error rate would be detrimental. Indeed proteins that contain amino acid substitutions tend to misfold and aggregate, promote spurious protein-protein interactions, and they may saturate protein quality control machinery, resulting in proteotoxic stress[93]. Conversely, some mistakes may be tolerated and a certain level of error might even prove to be advantageous. It has been shown that mistranslation is beneficial in response to environmental stresses as it can help sustain and disseminate cellular phenotypic viability[94]. On an evolutionary time scales too, phenotypic errors might be essential in facilitating adaptation of complex traits when combined with genetic mutations[86] [Whitehead, the look ahead effect], and by the purging of deleterious mutations[95]. A computational analysis of codon usage patterns across genomes revealed that a subset of codons are preferred over others at positions crucial for folding in highly

expressed proteins, suggesting that evolution indeed favors more accurate codons at these sites[79].

Recent development in RNA sequencing technologies quantified the rate of translation errors to reside in the range between 10^{-5} and 10^{-6}, an order of magnitude or two lower than the rate reported in protein translation[58]. In contrast, errors in protein translation have remained elusive and difficult to detect. An early effort by Edelmann and Gallant[53], who quantitatively tracked the insertion of radioactively labeled cysteine in *E. coli*'s flagellin, a cysteine free protein, revealed a first global estimate of mistranslation, with misincorporations happening on average every 10,000 amino acids. Since then, the use of fluorescent or luminescent reporter constructs allowed the quantitative tracking of specific types of mistranslation, at defined sites. These methods have highlighted the importance of codon-anticodon recognition and tRNA competition as determinants of these error rates, and were used to characterize the effects of aminoglycoside antibiotics and ribosome ambiguity mutations (ram)[58,96].
Yet, major questions still remain open. While error rates could be measrured precisely within specific positions of reporter constructs,, the overall error spectrum across the proteome has not yet been characterized. Such measurements would allow the assessment of the relative contribution of mischarging and mispairing.. Further, identification of error positions would allow to study the dependency of error on codon identity, and reveal whether specific positions within proteins are more prone to errors.

Mass Spectrometry (MS), which now permits routine, high throughput characterization of canonical proteomes and common post translational modifications (PTMs), was described as an upcoming tool for the study of protein mistranslation for almost a decade[93] and more recently harnessed to detect various substitutions from several purified recombinant proteins [59], and to detect and quantify the incorporation of norvaline at leucine positions across the proteome of E. coli mutants[97]. Yet, MS has yet to be harnessed for the unbiased study of amino acid substitutions on a proteome wide scale. Such full study was hitherto hindered by the low rate of substitutions compared to other natural and post-processing protein modifications and a much larger search space.
Here, we used Strong Cation-Exchange chromatography (SCX) fractionation and high resolution Liquid Chromatography (LC)-MS-MS to achieve a deep coverage of *E. coli*'s proteome, and assessed the effects of two aminoglycoside

antibiotics, streptomycin and paromomycin, on the bacteria's translation error rates and spectrum. In addition, we also assessed the effect of starvation to a particular amino acid on the mistranslation rate of its cognate codons. We have carried out our analysis with MaxQuant[98], repurposing its dependent peptide algorithm to identify mass shifts consistent with amino acid substitutions, and stringently filtering out potential artifacts. We then validated these identifications using a set of independent analyses that include a shift in HPLC retention time due to change in hydrophobicity of the encoded amino acid. Performing these experiments and analyses on *E. coli* in several growth conditions and analyzing similar data in the yeast *S. cerevisiae*, we could detect over 3500 site-unique substitution events.

This observed set of substitutions could, for the most part, be explained by a single mismatch in the codon-anticodon complex. In particular, G:U mismatches at the 1st and 2nd positions prevail, despite the recent observation that the geometry of the small ribosomal subunit's decoding center prohibits G:U wobble interactions at these positions[99,100]. The increased error rates observed in the presence of aminoglycoside drugs support the conclusion that these mistakes arise in the ribosome due to codon-anticodon mispairing. The set of errors that we detected in published MS data of *S. cerevisiae* shared a strikingly similar pattern of mismatches with *E. coli*, suggesting that errors are deeply constrained by base pairing chemistry. Furthermore, we show that rapidly evolving amino acid positions are more likely to bear amino acid substitutions. Observed substitutions tended to minimally affect protein energetic stability, and analyzing transcriptome-wide ribosome density data revealed low density at sites of mistakes, indicating at speed-accuracy trade-off. Our experimental observations support the view that organisms do mitigate the effects of translation errors by locally fine-tuning the way they encode proteins. Starving the cells for serine increased errors from this amino acid in a codon dependent manner. Our method offers quantitative estimates of error levels at a much larger scale than previously achieved, and offers a way to systematically study the response of the translation machinery to various stresses and conditions.

Results

A pipeline to confidently identify amino acid substitutions in a proteome

Mass spectrometry allows for large-scale identification of peptides at the proteome level. The task of identification of peptides with amino acid substitutions could thus resemble that of detecting known peptides that underwent post-translational modifications (PTMs). Despite the fact that the detection of common PTMs, such as phosphorylation or acetylation, has become commonplace[101], detecting amino acid substitutions by specifying a full list of all possible substitutions would result in a dramatic increase in the size of the peptide database. For example, assuming peptides of average length 10, 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[102–104], *i.e.* approaches that offer a way to identify (singly) modified peptides without requiring the user to input a list of predefined modifications, take advantage of the fact that modified peptides are usually less abundant than their unmodified counterparts. Therefore, a modified peptide is only likely to be detected if the canonical peptide has already been detected. We used MaxQuant to identify modified peptides with its "dependent peptide search" algorithm. "Dependent Peptides" are defined as peptides that show mass shifts in comparison to the unmodified, genome-encoded "Base Peptides" (Fig. 1B). We then applied a series of filters to the list of dependent peptides, in order to stringently remove known PTMs and artifacts and conservatively retain only amino acid substitutions. The outline of our pipeline is described in Figure 1A. For a detailed description of the pipeline, see Methods.

We generated a deep coverage, high resolution map of the *E. coli* proteome in rich medium at 37°C, and in addition evaluated the effect of two aminoglycosides antibiotics at sub-lethal concentration, and the effect of starvation to serine on the accuracy of its translation machinery. In total we generated error maps of 9 samples, each in two replicates (see Methods). All together we detect 3596 independent amino acid substitutions (each defined here by a unique position within a specific protein and a unique amino acid substitution) in the *E. coli* proteome. Similarly we analyze an existing proteome

dataset[105] from the yeast *S. cerevisiae* at a single type, non-treated, condition that yielded 225 substitutions for comparison.

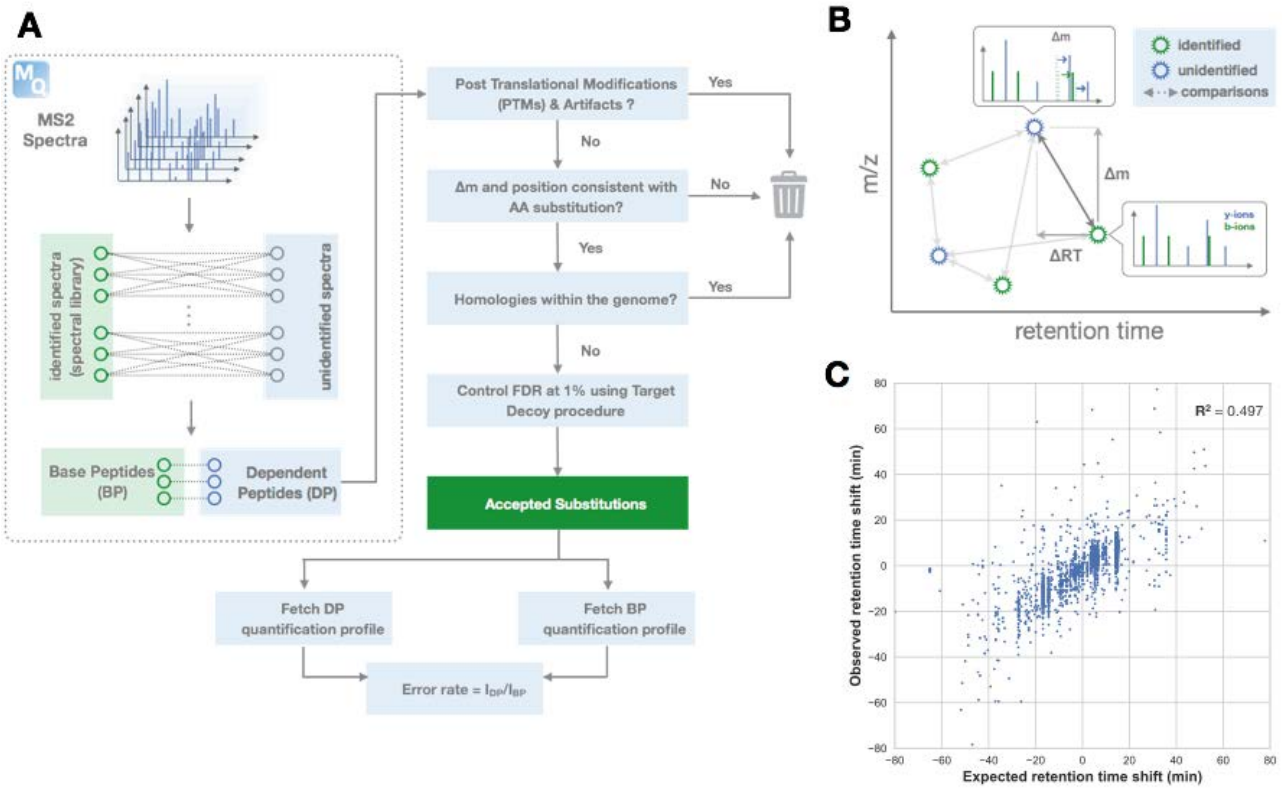


Figure 1: A computational pipeline to confidently identify amino acid substitutions from Mass Spectrometry data. A: Overview of the pipeline. For a detailed description of the different steps, see Material & Methods. **B**: MaxQuant Dependent Peptide search performs exhaustive pairing of unidentified 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 to 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 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.

Most of the high quality hits are bona fide amino acid substitutions.

Given mass differences detected between base and dependent peptides we must first establish that they represent amino acid substitutions. For that, we took advantage of the fact that many amino acid substitutions change peptide hydrophobicity and they hence result in predictable retention time shifts during liquid chromatography. The retention time of a peptide can be predicted with high accuracy ($R^2 > 0.9$) approximately as the sum of the hydrophobicity coefficients of its amino acids[106]. Therefore, the predicted HPLC retention time of the substituted amino acid can be computed and compared to the observed retention time recorded for the substituted peptide. We trained a retention time prediction tool on a list of confidently identified peptides, and generated an expectation of the retention time shift induced by the detected substitutions. We compared this expectation to the observed retention time shift for each of the detected substitutions in the MOPS dataset (Fig. 1C). This analysis supports the notion that most of the substitutions detected are genuine amino acid replacements.
Note that our sampling strategy allows us to detect substitutions originating from the highly expressed proteins only.

We define a substitution as a combination of a position in a protein, an "origin" amino acid (and its associated codon), and a "destination" amino acid. We then divide all substitutions in two sets: a substitution is classified as a Near Cognate Error (NeCE) if the error-bearing codon of the origin amino acid matches with one nucleotide difference at least to the codons of the destination amino acid, and as Non Cognate Error (NoCE) otherwise. The structure of the genetic code dictates that only a minority of the substitutions would be classified as NeCE. In particular, of all detectable codon to amino acid substitution types 30% are expected to be of the NeCE type. In stark contrast, 88% of the unique substitutions detected by our method with the full *E. coli* dataset are classified as NeCE. Thus, the great majority of observed substitutions in our data can be rationalized by a similarity between the origin's codon and a codon of the destination amino acid. Such enrichment for NeCE compared to expectation serves as an indication that we inspect genuine amino acid substitutions (see SOM for a formal statistical test)
An intriguing possibility is that NeCE substitutions might predominantly represent codon-anticodon mis-pairing events that occur within the A-site of

the ribosome, and that NoCE substitutions might occur elsewhere, *i.e.* in the amino acid charging phase by the relevant aaRS. We attempt below to support the notion that indeed the majority of NeCE events represent mRNA-tRNA mis-pairing events.

Overview of amino acid substitution landscape in E. coli

Substitution matrices are common in biological research, for example decades of research in genomics revealed 4*4 nucleotide substitutions matrices for DNA and RNA polymerases, and 20*20 matrices of substitutions between amino acids in evolution. Our amino acid substitution data allow us to generate 64*20 codon to amino acid matrices that depict the prevalence of each type of amino acid error in a dataset. Note that in no way these matrices represent real relative probabilities of mistakes as our ability to detect an error depends on the original protein's expression level, which also influences its codon choices. The numbers of unique peptides supporting any codon to amino acid substitution type is show in Fig. 2A; the intensity of the shade is proportional to the logarithm (base 2) of the number of unique genome positions in which substitutions were observed. Because leucine and isoleucine are isomers and thus share the exact same mass, our method is not able to distinguish the two amino acids as destinations of a substitution; thus, we grouped together substitutions towards Ile and Leu. Furthermore, substitution types that transform a codon into its cognate amino acid, 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 PTM, were grayed out, and discarded from subsequent analyses (see Methods).

An interesting observation we make on this matrix is that the codon that encodes for an amino acid affects its substitution destination. This is nicely illustrated with substitutions from Gly to Asp and Glu. We see that when Gly is encoded by the GGC codon, the frequent substitution destination is the near-cognate Asp (that can be encoded by the near cognate codon GAC), while encoding Gly with GGA often results in substitution of Gly by Glu (presumably due to its near cognate codon GAA). Similar cases in which different codons for the same amino acid tends to show different amino acid substitution pattern can be found in the matric

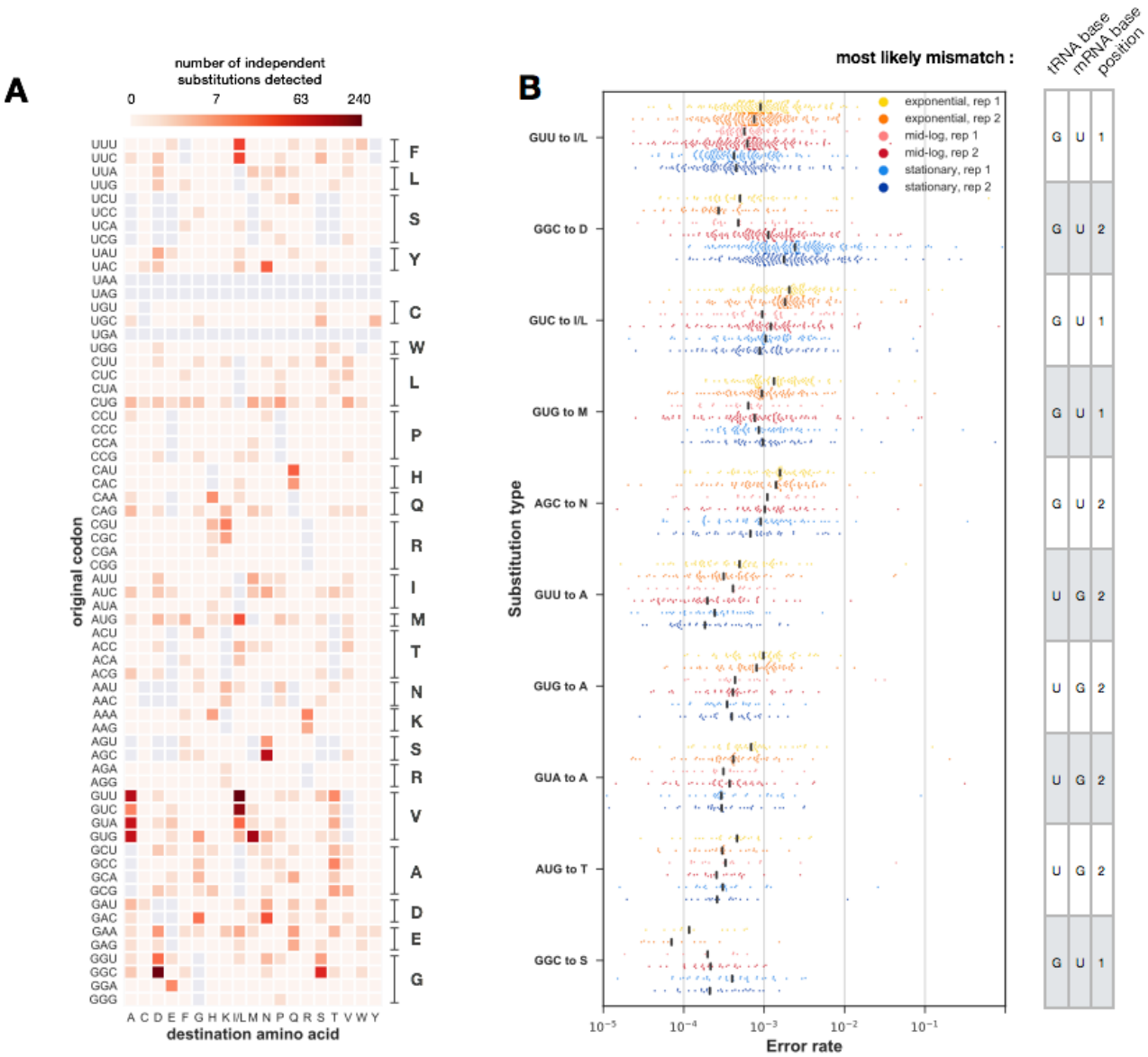


Figure 2 : overview of the substitution profile of *E. coli* in MOPS complete medium. A : Matrix of substitutions identifications. Each entry in the matrix represents the number of independent substitutions detected for the corresponding (original codon, destination amino acid) pair, in the SC-complete 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 stop codon, substitutions undetectable via our method because they are indistinguishable from one of the PTMs or artifacts in the 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 have been detected and quantified. The black line indicates the medians of the distributions. Right panel: we inferred the most likely mismatch for each of the substitution types, using a procedure described in the Material and Methods. This allows us to guess that the V -> I/L substitutions are likely substitutions from Val to Ile, enabled by a G:U mismatch at the 1st position.

Once we validated the amino acid substitutions we calculated the observed error rate (i.e. the ratio of intensity between the dependent and base peptide) for the all detected substitutions. As example, the Ser_{AGC}→Asn substitution, was detected in total in 81 peptides across the *E. coli* proteome of the non-treated samples. Figure 2B shows the error rate estimations in each of these substitutions – each dots in the plot corresponds to one specific Ser_{AGC}→Asn substitution on a particular genomic position, and the error rate is on the y-axis. Likewise the 10 most frequent substitutions types in the proteome are shown. 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 types showing slightly higher error rates. Due to the MS acquisition strategy, positions that feature a low error rate are less likely to be detected, which could lead to an over-estimation of the actual error rates. We measure the proteome, and detect translation errors, in three time points along the growth cycle (beginning in exponential growth phase and ending with the stationary phase (Fig S1). An intriguing trend we observed is that error rates seem to consistently decline as cells enter the stationary phase. The actual decline in error rate might be under estimated here, due to the fact that we measure errors from the whole current proteome without restriction to mistakes made at newly synthesized proteins.

A global nucleotide mispairing mechanism for translation errors

We further classified NeCE substitutions based on the location of the mismatch within the codon and the nucleotide types they involved. We define the count density for a given mismatch type as the number of substitutions that can be explained by that type of mismatch divided by the number of substitution types that can be explained by the same mismatch, and report the count density for the two biological repeats in Fig. 4B. This analysis results in three 4*4 “mismatch matrices” that depict the prevalence of mismatching for each nucleotide in the codon with each of the three non-perfectly matching nucleotides in the anticodon Fig 3B. Substitutions that could be caused to multiple mismatches were assigned to the most likely mismatch using an expectation-maximization scheme (see Material and Methods). The most frequently observed substitution type involves G:U mismatches in the first or the second position of the codon. Interestingly, this rule holds only for mismatches where the codon base is G and the anticodon base in U; the fact that the opposite geometry (i.e. errors in which a U is in the codon and a G in

the anticodon) seems to be less error prone is surprising at first, but might be explained by the numerous modifications affecting uracil at the tRNA level.

E. coli and *S. cerevisiae* share similar error profiles

While both characterized by a mostly planktonic lifestyle and high growth rates, *E. coli* and *S. cerevisiae* have been diverging from one another for at least 2.7 billion years. Comparing the error profiles of these two organisms, thus, allows us to look at how strongly these errors are constrained, both by chemical and evolutionary necessities. We reanalyzed a previously published mass spectrometry dataset of strong anion exchange (SAX) and SCX fractionated proteome of *S. cerevisiae* grown in a single condition, a rich medium (30°C, YPD)[105] using our pipeline.

We were able to detect a total of 225 substitutions in the yeast proteome. Here too the majority of the errors, 143, were classified as NeCE. Comparing the error spectrum between the eukaryote and the prokaryote we observed a high overlap between the set of substitution types seen in the two organisms. This observation reveals a universal error pattern for mistakes that are likely to occur within the ribosome, while most NoCE substitutions likely originate from separate factors unique to each of the species. The most notable difference between the two species is in the most frequently observed substitution of Ala to Cys in yeast, which is not seen in the bacterium. Indeed a recent report[107] reveals the basis for this observation – that eukaryotic, but not prokaryotic Alanyl-tRNA synthetase (AlaRS) have precisely the tendency of mischarging $tRNA_{Cys}$ with Alanine.

For the yeast data too we computed the 3 4*4 substitution matrices and observed that in similarity to the *E. coli* matrices they also feature G:U mismatch at the first or second positions (Fig. 4B). Observing such levels of error similarity between such loosely related organisms, exhibiting distinct codon usage biases and a relying on very different translation machineries, hints at the possibility that these errors depend on universal constraints. Whether these constraints are of a purely chemical nature, or the observed substitutions happen to be more tolerable by these organisms remains to be determined.

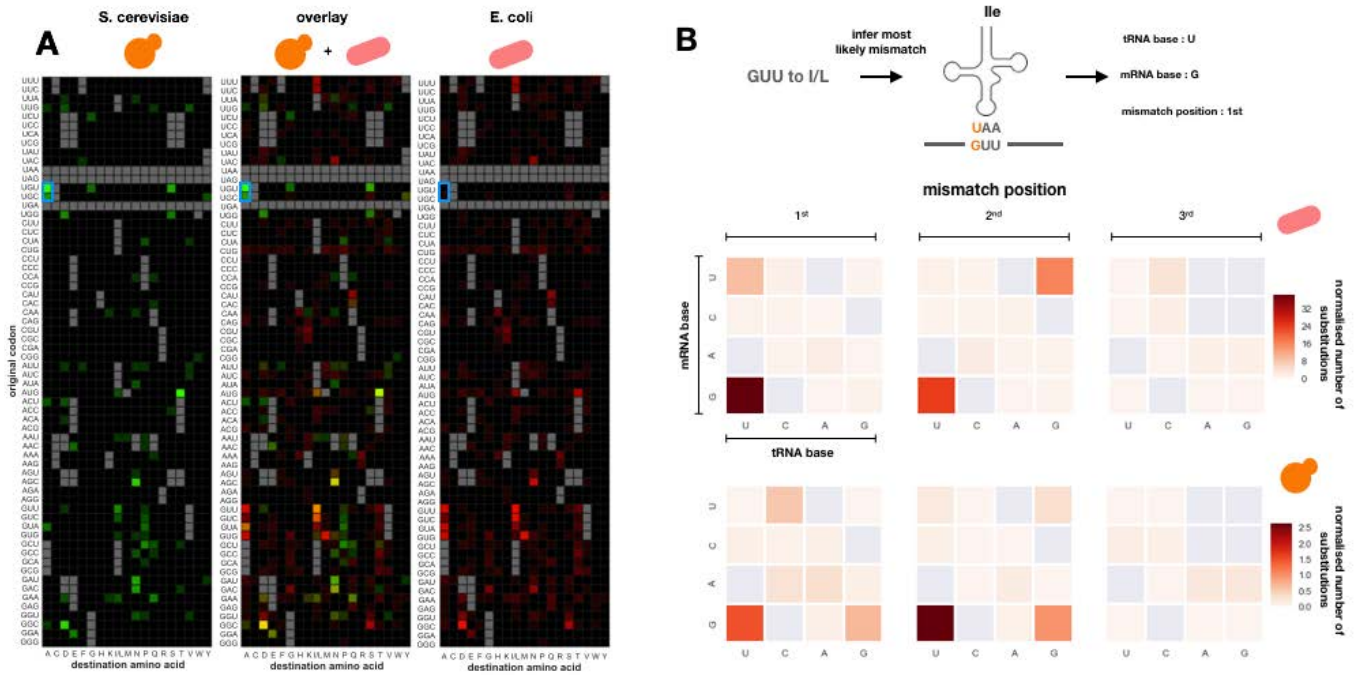


Figure 3 : Comparing the error profiles for *E. coli* and *S. cerevisiae* reveals a shared signature of errors. A: the substitutions identifications matrices of *S. cerevisiae* (green channel, left) and *E. coli* (red channel, right) are compared and overlaid (middle). The intensity of the color is proportional to the logarithm of the number of independent identification, 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. **B:** NeCE are classified by the mismatch most likely to generate them. The shade intensity reflects the ratio of independent substitution to number of substitution types associated with the corresponding mismatch. Grey boxes are either correct base-pairings, or mismatches to which no substitutions could be unambiguously mapped. Upeer panel indicates results obtained from E.coli lower panel was generated based on S.cerevisiae data.

The effect of drugs and amino acid starvation on substitution patterns

To gain further insight on error patterns and how they are affected by various perturbations, we either treated *E. coli* cells with two types of antibiotics that reduce ribosome proof reading capability, or starved them for an amino acid, serine. We applied two aminoglycoside antibiotics, paromomycin and streptomycin, to the bacteria. These two drugs are believed to interfere with the ribosome's proofreading activity[108], and their effect on translation accuracy was previously measured using a luciferase reporter construct[58]. We measured the proteome under each of the drug treatments by the MS-MS procedure and re-ran our error detection pipeline. To compare between the error patterns induced by the drugs, we again inspected the 64*20 codon to amino acid matrices (Fig. 4A), the error rate profiles (Fig. 4CZ) and the three 4*4 nucleotide mispairing matrices (Fig. 4C). Comparing the 64*20 matrices

between the non-treated and drug treated samples reveals a clear pattern – the drugs increased error rates especially at 3rd codon wobble positions, while other mismatch positions remained relatively unaffected. This observation is confirmed by the three 4*4 matrices. The increased error rate at the 3rd position can be quantified using MS1 information, as reported in Fig. 4B.
We have next starved the bacteria to serine, measured again the proteome by the MS-MS procedure and re-ran our error detection pipeline. The prediction was that upon starvation to this amino acid we should observe elevated level of errors leading from this amino acid to others. Indeed, we observe that the rate of Ser_{AGC}→Asn steadily increased upon starvation. We quantified further, as cells enter more deeply into the stationary phase, when the effect of starvation is supposed to intensify the rate of the substitution from Ser to Asn increases. This result indicates to a clear mechanism that accounts for mistakes in translation in which a shortage of an amino acid determines its probability to be replaced by others.

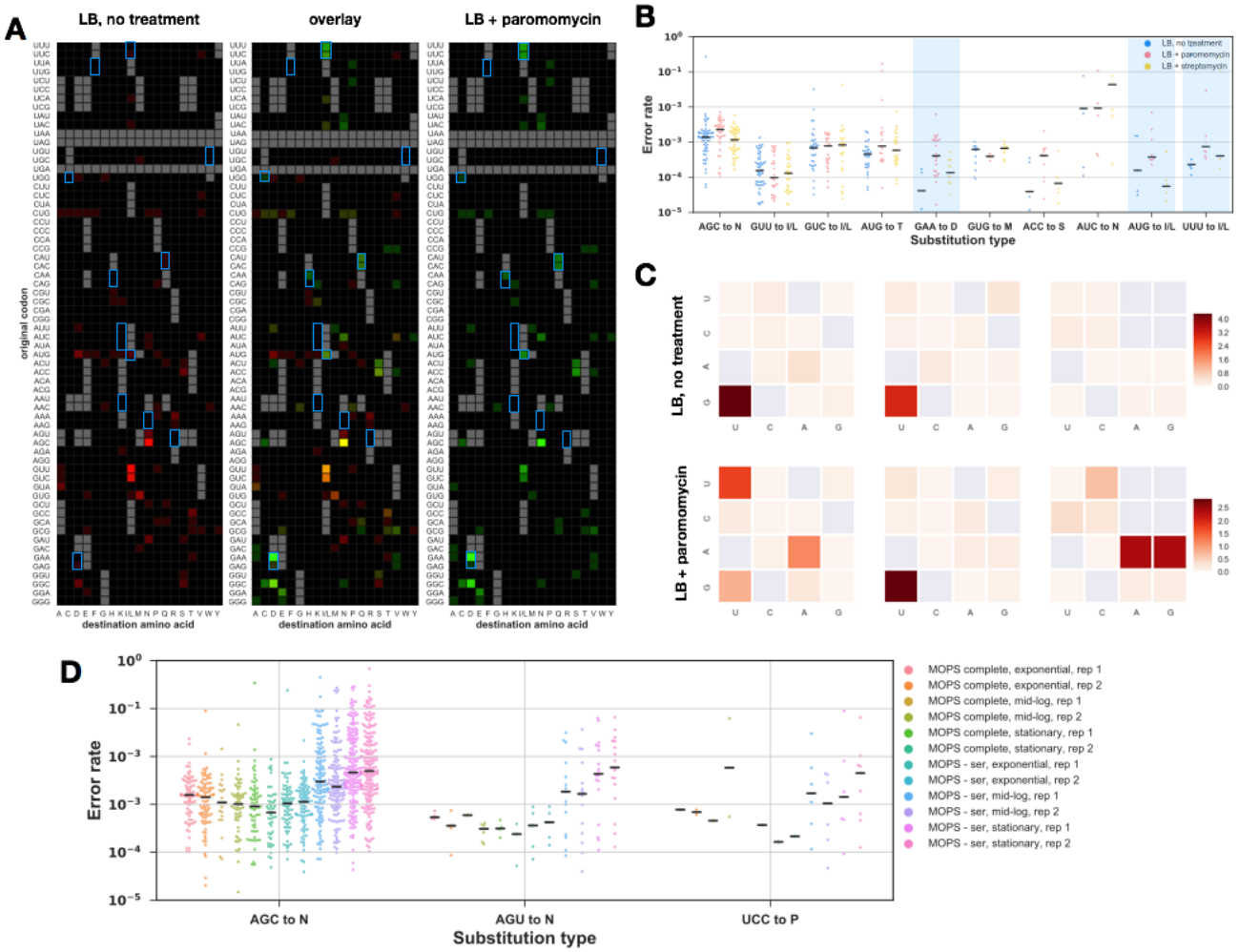


Figure 4 : the error spectrum is affected by external perturbation. A: the substitutions identifications matrices of *E. coli* in LB (green channel, left), or LB supplemented with paromomycin (red channel, right), are compared and overlaid (middle). The intensity of the color is proportional to the logarithm of the number of independent identification, with one pseudo-count. Values are normalized by the highest entry in the matrix for each of the two organisms. The blue boxes highlight errors involving 3rd position mismatches. **B:** Quantification of the top 10 most frequent substitutions in the drugs dataset. Errors involving 3rd position mismatches are shaded in light blue. **C:** NeCE are classified using the same procedure as in Fig. 2B, for the LB samples, with or without paromomycin. **D:** Effect of serine starvation on errors at serine codons, for the three most frequently detected substitutions affecting serine codons.

Misincorporations occur at error-tolerant and rapidly translated positions

Drummond & Wilke[79] posited that cells, in order to avoid the fitness loss due to protein misfolding and aggregation, manage their errors by selecting error-proof codons at positions where inserting the correct amino acid is critical to folding or function. They were able to support that theory using computational means, but had to rely on key assumptions. In particular, they used evolutionary conservation as a proxy for sensitivity to phenotypic errors, and they derived the identity of error prone and error proof codons from conservation data. Correspondingly, fast evolving positions within protein are predicted to be less critical for protein folding and function thus correspond to sites where rate of mistranslation is expected to be higher. This assumption that evolutionary conservation correlates with phenotypic error rate was indeed made in several additional recent publications. Yet, the lack of a systematic set of translation error events within a proteome precluded so far the examination of the notion that they occur preferentially in rapidly evolving sites, or in positions that minimally affect protein structure and function. A careful analysis of the classical model of kinetic proofreading revealed an complex trade-off between speed and accuracy during the aa-tRNA selection step by the ribosome: ribosomes are more likely to misincorporate amino acids at sites where they translate rapidly[109]. This trade-off was examplified by mutants that featured modified translation speed[110], and by *in vitro* conditions that affect ribosome velocity[70]. Yet examination of the theory in natural sites within genes, in which ribosome's speed can now be deduced[111], was so far impossible to obtain due to lack of ability to measure translation errors genome-wide .

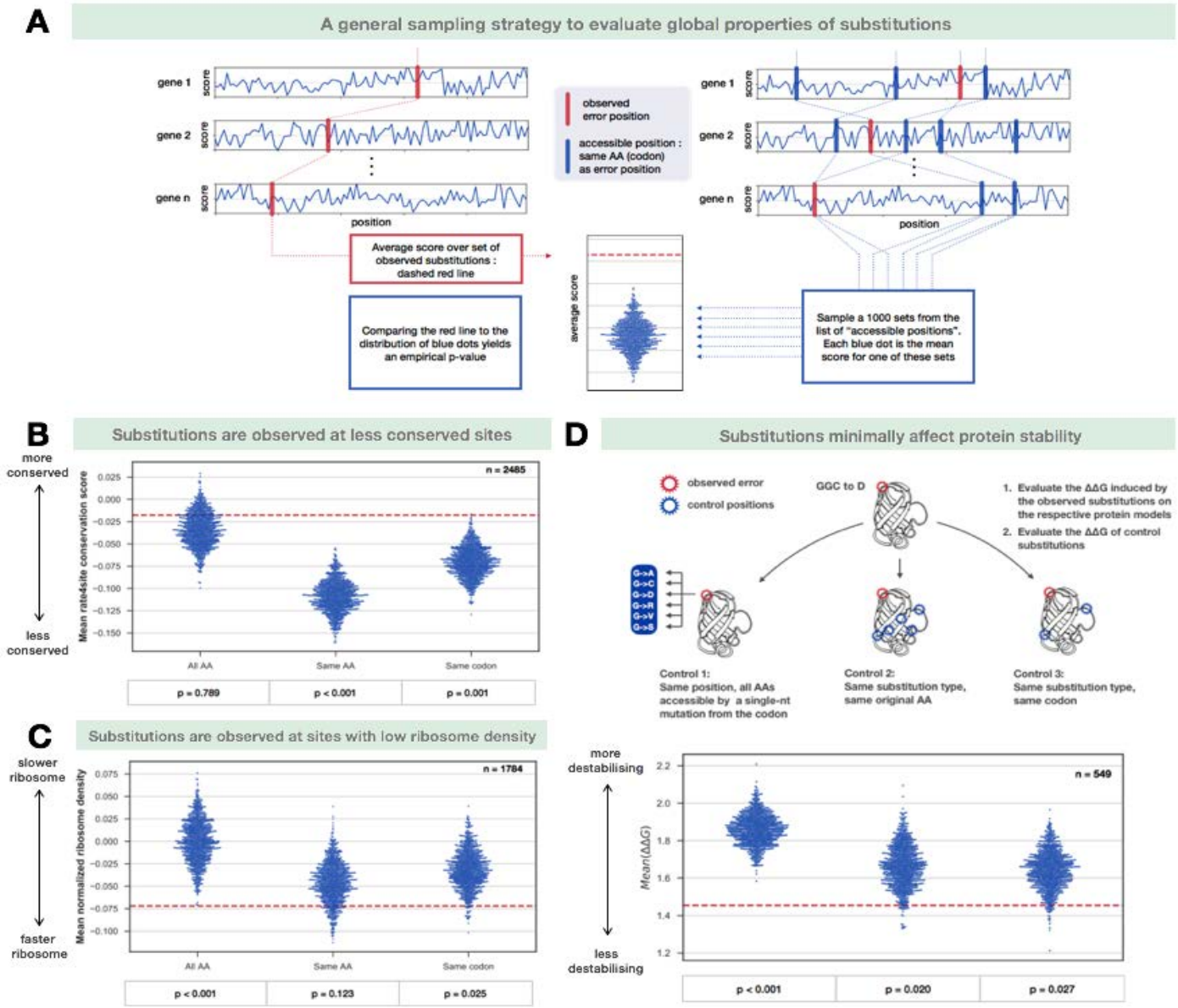


Figure 5: general properties of substitutions. A. Sampling strategy: In order to test if the set of detected substitutions differs from expectations in any way, we first need account for the fact that many local properties of proteins are affected by the protein's expression level, and so is our ability to detect substitutions from that protein. First, the local property of interest ('score') is recorded at all the positions bearing a substitutions. The average of that set is plotted as a red dashed line. To compare this average to 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 1000 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 to align them, and rate4site to estimate the evolutionary rate at each site.

The resulting scores are standardised per proteins, and a high score indicates low conservation. The empirical p-values are computed by dividing the number of blue dots above the red line, divided by 1000. n indicates the number of positions considered in this analysis. **C. Ribosome density:** Ribosome profiling data from [ref] was processed (see methods) to estimate the ribosome density at positions along the E. coli transcriptome. Since ribosome density/speed can only affect errors in cis, this analysis was restricted to NeCE. **D. Effect of substitutions on protein stability:** for proteins whose 3D structure is known, we evaluated the effect of NeCE on protein stability using FoldX. In control 1, we test if the observed substitutions are on average less destabilising than those stemming from other single-nt mismatches between the codon and the anticodon, at the same position. In control 2 and 3, we test if the observed substitution type observed was less destabilising on average at the observed position than at other positions sharing the same AA, or the same codon.

We decided to test if the mis-incorporations we observed indeed occurred at less conserved, rapidly evolving, positions by comparing the distribution of conservation scores for our observed substitutions to that of carefully selected control positions. Normalized conservation scores were computed for each protein by fetching homologs, aligning their sequences, and running rate4site[112] to determine the evolutionary rate at each site in the protein. We recorded the normalized score for each of the positions for which a substitution was detected. We use standardized rate4site evolutionary rate scores per protein; a high score indicates low conservation of the amino acid position across orthologous proteins. In order to account for the fact that some amino acids tend to be more conserved than other, and that some codons are over-represented at conserved positions, we devised three strategies to generate adequate negative controls (Fig. 5A). In the first and least stringent strategy, for each observed substitution, we sampled a normalized conservation score from any position in the same protein. In the second strategy, the random re-sampling was carried not only within the same protein, but also with the additional constraint that the amino acid identity in the randomly sampled position has to match the same amino acid type observed at the position at which the substitution occurred. Finally, in the most stringent of these negative controls, we performed a 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 control, and compared the mean of the observed distribution of scores at the observed substitution positions to those of the random control distributions to obtain empirical p-values. The mean rate of evolution at substitution sites is similar to that of random sets of positions generated though the first model, but significantly higher than that of the random generated with the other two (fig 5B). Consistent with the previous prediction[79], controlling for the codon reduced the magnitude of the difference between the real error sites and random sites (fig 5B "same codon *vs* "same AA") , supporting the notion that

evolution allows or precludes error-prone codons from sites that are correspondingly tolerant or intolerant to errors.

Similarly, to the conservation test, we examined the non-independent possibility that observed amino acids substitutions in the *E. coli* proteome tend to minimally affect the energetic folding stability of protein in which they occur. To this end, we used FoldX[113] to compute ΔΔG, *i.e.* the difference in folding energy between each original, genome-encoded protein and its corresponding substituted version. After obtaining distributions of such scores, we compared these distributions to those obtained upon three random sampling negative control strategies (Fig. 5D). In the first control strategy we analyze the stability difference between the wild-type protein and all the proteins that could be obtained by mutating a single nucleotide of the error-bearing codon ("identity control"). In the other two negative controls, we maintained the identity of the originally observed pair of substituted and substituting residues, but modeled the effect on ΔΔG upon substituting between these two amino acids, albeit at randomly chosen positions within same proteins sharing the same amino acid, or the same codon ("amino acid position and codon position controls"). In the identity control, we test if codons are preferentially mistranslated to amino acids that are accepted at the position of error, controlling for the established property of the genetic code of allowing substitutions between chemically similar amino acids. The two position controls test if substitutions happen at sites at which they better tolerated, and if the codon identity explains this effect.

We find that observed substitutions tend minimally disrupt protein folding, with a mean ΔΔG of 1.454 kcal/mol. This value is very significantly lower than that obtained under the identity control (mean ΔΔG ~ 1.9 kcal/mol). Among the possible single nucleotide mismatches that could lead to a mistake, the cell seems to be more permissive to those less disruptive to protein stability. Consistent with the conservation findings, errors are seen preferentially at sites that minimally affect folding, suggesting positional information within genes that allows mistakes to happen where they would be minimally disruptive. Controlling for the codon identity did not explain this effect. We cannot exclude an equally interesting alternative that some substitutions that destabilize protein structure lead to a more rapid degradation, and are thus precluded from being sampled in our method.

Lastly, we aimed to test the notion that the ribosome is prone to make an error at positions in which it translates more rapidly. An indirect means to deduce the ribosome speed on each position in each gene is to measure its read density in a ribosome footprint experiment. At steady state flow the product of

speed and density should be constant. Hence, region of locally high density in ribosome footprinting indicate a locally low speed of the ribosome. We computed the normalized read density profile of most *E. coli* protein using ribosome profiling data of bacteria grown on MOPS complete medium (see Material & Methods). We could then ask if error sites feature the expected high speed, i.e. low density. Computing the mean ribosome density among all the error sites, and comparing that mean to the mean of 1000 randomly sampled positions indeed showed a small-effect but statistically significant trend (Fig. 5C) – error sites are less dense, and are hence deduced to be translated more rapidly than matched controls.

Discussion

Here we report on a new method to observe single amino acid misincorporations, which we used to detect over 3500 distinct translation errors across the proteome of *E. coli*. Our method takes advantage of the very high accuracy of modern mass spectrometer to generate high confidence identifications. Orbitrap mass spectrometers can be tuned to detect mass differences on the order of thousandth of Daltons, during both the MS1 and MS2 acquisition phases. This accuracy in turn allows us to distinguish peptides and peptide fragments of almost identical masses, but of different atomic and isotopic compositions, and thus greatly improves the performance of database search algorithms. Our method is therefore able to distinguish amino acid substitutions from PTMs of similar masses. Despite the FDR procedure applied at the end our pipeline, we cannot exclude with absolute certainty that some of the substitution types we detect are in fact un-annotated PTMs that cannot be distinguished from amino acid substitutions. However, the retention time shifts in HPLC observed for our set of identifications correlate very well the expected retention time shifts predicted from sequence information alone, an observation that could not be explained by the identification of spurious PTMs.
One cannot guarantee *a priori* that these substitutions stem from errors in the translation machinery, because non-synonymous errors at the DNA or RNA levels could generate the same mistakes at the protein level. However, our samples originate from clonal populations, which implies that DNA mutations are unlikely to reach a detectable level in the absence of strong adaptive selection, and would be very rarely observed to occur across multiple samples. Since we analyze samples in which the number of cells (~10^9) is greatly superior to the inverse of the observed lower bound of transcription error rates

(~10^5), and the average number of mRNA per cell for the genes we detect is greater than one, the relative abundance of errors is expected correspond to the transcription error rate at any examined site, and should not fluctuate from sample to sample thanks to the assumption of ergodicity. This estimate is two orders of magnitude lower than the average observed error rates quantified by our method. Even though transcription error rates were shown to be fairly constant over a range of conditions in *E. coli,* we cannot rule out the possibility that local transcription error rates hotspots could yield peptides detectable by our method.

The set of observed substitutions therefore likely derives from errors within the translation machinery. Two distinct processes have been shown to generate high levels of errors: aaRS can mistakenly load an amino acid to a non-cognate tRNA, and the ribosome can pair a correctly charged aa-tRNA complex to a non-cognate codon. Both processes rely on small energetic differences between correct an incorrect pairings. For the ribosome, the recognition process exploits the difference of free energy between cognate and non-cognate codon-anticodon pairs. Some aaRS also probe the nature of the anticodon of the tRNA before loading, and additionally rely on clues from the tRNA backbone to achieve a high specificity. The amino acid recognition step can be challenging due to similarities between amino acid types, and a subset of these enzymes have to rely on an editing step to achieve higher specificity. Differential binding of EF-Tu to misacylated tRNAs was shown to discriminate against common aaRS mistakes[114], and thus provides an additional layer of specificity. We argue that most of the substitutions detected in our work stem from errors in the ribosome. Indeed, the overwhelming majority (88%) of the substitutions could be explained by a single codon-anticodon mismatch, a fraction much higher than expected by chance due to the organization of the genetic code (30%). Additionally, treating the cells with aminoglycoside antibiotics known to perturb the accuracy of the ribosome affected the rate and spectrum of errors, increasing in particular the error rates for substitutions involving mismatches at the 3rd codon position. However, we were able to identify several instances Cys→Ala subtitutions (NoCE) in the *S. cerevisiae* samples, consistent with a recent report that eukaryotic, but not prokaryotic AlaRS had a tendency to mischarge non-cognate cysteine tRNAs[107].
Comparing the error spectrum of the *E. coli* and *S. cerevisiae* in untreated, rich conditions revealed a large overlap between the set of observed substitution types, and a striking prevalence of G_{codon}:$U_{anticodon}$ mismatches at the first and second positions. Structural analysis of G:U and U:G mismatches within the

ribosome revealed that they typically adopted a Watson-Crick G:C like geometry rather than the expected wobble one due to spatial constraints in the decoding center. These errors are therefore believed to originate from rare enolic or anionic states of nucleobases, as proposed by Rozov *et al.*[99]. The surprising observation that G:U mismatches are typically much more prevalent than the symmetrical U:G conformation could be explained by the abundance of uracil modifications on the anticodons of tRNAs.

The *E. coli* MOPS dataset allowed us to quantify a large number of substitutions. The mean error rate detected was on the order of 10^{-3}, in the higher end of the range of previously reported estimates. Several reasons can be invoked to explain this observation. First, MS detectability is intimately linked to MS1 intensity levels: since the mass spectrometer systematically samples the most intense peptides in each scan, we are bound to preferentially detect and quantify substitutions associated to high error rates. Similarly, a peptide's MS1 intensity depends on its abundance in the sample and on its ability to ionize well. The abundance of the correct peptide is usually much higher than that of the error-bearing one, which means that it will be sampled more often. The quantification depends on the sampling of the lower abundance, error-bearing peptide. Substitutions that increase the peptide's ionization efficiency are therefore bound to increase its detectability, and will result in an inflated error rate. While it is generally accepted that ionization efficiency depends on a peptide's sequence in a very non-linear fashion, we trained a linear regressor to evaluate the mean effects of amino acid composition on ionization efficiency. Our model gave satisfactory results (see appendix : Prediction of Ionization efficiency from amino acid composition), and indicated that, except in a few cases, substitutions should not result in a dramatic change in ionization efficiency. It remains difficult to assess to what extent the standard deviation of the error rate for each substitution type reflects biological variability or technical biases.

Comparing the median error rates of several substitutions across the different physiological states during bacterial growth revealed that they react dynamically to the changing environment: substitutions rate from valine codons tended to decrease with time, while glycine codons became more error-prone in later stages. The extent of this change might be underestimated due to the fact that we are not specifically quantifying the error rate of newly synthesized protein, but rather quantifying the errors in batch. Starving the cells for serine revealed a striking increase in the error rate of two substitutions involving serine codons, Ser_{AGC}→Asn and Ser_{AGT}→Asn. The median error rate

for these codons rose to almost 10^{-2} in the stationary phase time point, with some sites reaching an error rate approaching 10^{-1}. Other serine codons were also affected, but the scarcity of sampling for these rarer errors precluded a reliable quantification of the process. Theory predicts that the 4-box codons of serine (TCN) should suffer more from serine depletion than the 2-box codons (AGY) because of a differential charging of the tRNA isoacceptors[51]. Our failure to detect a large quantity of errors at TCN sites might be partially explained by the preferential usage of AGY codons in genes over-expressed during serine starvation[51].

Translation errors have been hypothesized to be a major constraint in protein evolution, and to drive the long known anti-correlation between gene expression and evolutionary rate at the protein level[79]. According to this theory, the selective pressure to prevent translation errors constrains the synonymous encoding of amino acids critical to protein folding, and organisms must select preferred, error-proof codons at positions where errors are likely to disturb protein stability. These highly constrained sites are characterized by a higher evolutionary conservation, and a slow rate of evolution. Our set of substitutions enabled us to directly test if errors indeed happen preferentially at fast evolving sites. Our analysis carefully controlled for the effects of protein expression level on the detectability of translation errors, the codon usage of proteins, and their evolutionary conservation. It confirmed that indeed, substitutions occur on average at less conserved sites, but also that the choice of codons could not entirely explain this effect, suggesting that other factors might affect translation accuracy *in cis*. Similarly, simulating the effects of the set of observed substitutions on protein stability revealed that they tended to occur at sites where they minimally affected protein folding. Observed NeCE were also less destabilizing than randomly sampled NeCE at the corresponding sites, suggesting that the spectrum of ribosome errors is even more conservative than the effect of naïve single substitutions at the DNA level. Together, these results confirm that the cells encode their proteins and tune their translation machinery in ways that minimize the deleterious effects of amino acid misincorporations.
Since codon identity does not entirely account for the fact that substitutions are preferentially observed at sites where they are tolerated, we tested if the ribosome itself might modulate its accuracy locally. Several lines of evidence indicate that ribosomes optimize both speed and accuracy, and must therefore perform a trade-off between theses two constraints. In particular, decreasing the ribosome's GTP hydrolysis rate should result in an lower processing speed,

but a better discrimination between cognate and non cognate aa-tRNAs.[109] We hypothesized that the ribosome might rely on external clues to locally slow down in order to increase its accuracy at critical sites. Our analysis of a published ribosome profiling dataset indeed revealed a subtle but significant shift in ribosome density: the sites at which we observed substitutions were characterized by a lower ribosome density, i.e. a higher speed.

Material and Methods

Strains and growth conditions

To generate the E. coli drugs dataset, MG1655 cells were plated on LB agar and incubated at 30°C overnight. 6 colonies of MG1655 were picked and grown until stationary phase in 3 ml LB, 30°C. All 8 cell cultures were diluted 1/100 and grown aerobically in 100 ml LB supplemented with the relevant antibiotics (see table X) in 500 ml Erlenmeyer flasks at 37°C until they reach mid-log phase (OD ≃ 0.5). For the serine starvation dataset, BW25113 (WT) and JW2880-1 (ΔserA, obtained from the Keio deletion library) cells were plated on LB agar and incubated at 37°C overnight. 2 colonies of each strain were picked and grown in 3 ml of modified MOPS rich defined medium made according to Cluzel et al recipe (SI Appendix) and incubated at 37°C until stationary phase. BW25113 and JW2880-1 cell cultures were diluted 1/1000 and grown aerobically in 220 ml of modified MOPS rich defined medium and MOPS serine starvation medium accordingly in 500 ml Erlenmeyer flasks at 37°C (mediums were made according to Cluzel et al 2012 SI Appendix).

Proteome extraction

We adapted our proteome extraction protocol from Khan *et al.,* 2011 [115]. Samples were each split into two 50 ml falcon tubes, centrifuged at 4000 rpm for 5 min, and washed twice with PBS (add 10 ml PBS, vortex, centrifuge for 5 min). Remaining PBS was vacuumed and the pellets were frozen in ethanol-dry ice. Pellets were re-suspended in 1 ml of B-PER bacterial protein extraction buffer (Thermo Fisher Scientific), pooled together, and vortexed vigorously for 1 min. The mixture was centrifuged at 13,000 rpm for 5 min. The supernatant (high solubility fraction) was collected and frozen in an ethanol-dry ice bath. The pellet was re-suspended in 2 ml of 1:10 diluted B-PER reagent. The suspension was centrifuged and washed one more time with 1:10 diluted B-PER reagent. The pellet was re-suspended in 1 ml of Inclusion Body

Solubilization Reagent (Thermo Fisher Scientific). The suspension was vortexed for 1 min, shaken for 30 min, and placed in a sonic bath for 10 min at maximum intensity. Cellular debris was removed from the suspension by centrifugation at 13,000 rpm for 15 min. The supernatant was frozen in an ethanol-dry ice bath (low solubility fraction).

SCX fractionation, HPLC and Mass Spectrometry

400µg of protein was taken for in-solution digestion and processed by Filter aided sample preparation (FASP)[116] protocol using 30k Microcon filtration devices (Millipore). Proteins were subjected to on-filter tryptic digestion for overnight at 37°C and the peptides were fractionated using strong cation exchange (SCX) followed by desalting on C_{18} StageTips[117] (3M Empore™, St. Paul, MN, USA). Peptides were analyzed by liquid-chromatography using the EASY-nLC1000 HPLC 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

Computational methods

Raw files were analyzed with MaxQuant v. 1.5.5.1. The list of parameters is available in the supplementary materials. High and Low solubility fractions were aligned separately. The amino acid substitutions identification procedure relies on the built-in dependent peptide algorithm of MaxQuant.

The Dependent Peptide search

Experimental spectra are first searched using a canonical database search, without any variable modification, and a False Discovery Rate (FDR) of 1% is guaranteed by a target decoy procedure. Identified spectra are 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 :

Compute the mass shift Δm by subtracting the mass of the identified (unmodified) spectrum to that of the unidentified (modified) spectrum,

Generate modified versions of the theoretical spectrum by adding *in silico* this mass shift at every position along the peptide, and

Evaluate the match between the theoretical spectrum and the experimental spectrum using a formula similar to Andromeda's binomial score.

Finally, 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[118] 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 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.

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.

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[119], 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[120]. The alignments were then used to compute normalized evolutionary rates per site with the rate4site program[112]. The mean evolutionary rate of sites associated with detected substitutions was compared to that of a 1000 randomly sampled positions, using the strategy described in Fig. 5A

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[113]. 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 Fig. 5D on protein stability (ΔΔG), using the PositionScan command. Finally the mean ΔΔG of our set of substitutions was compared to the mean ΔΔG of 1000 randomly sampled substitutions, using the strategy described in Fig. 5A.

Ribosome density computation

Ribosome profiling data for the MOPS complete experiments were downloaded from Woolstenhulme *et al.*, 2015[121] (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. Genes whose read density (i.e. number of reads mapped divided by gene length) was lower than 10 were also excluded. 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 Fig. 5A.

Chapter 2: RNA editing in bacteria recodes multiple proteins and regulates an evolutionarily conserved toxin-antitoxin system

Research

RNA editing in bacteria recodes multiple proteins and regulates an evolutionarily conserved toxin-antitoxin system

Dan Bar-Yaacov,[1] Ernest Mordret,[1] Ruth Towers,[1] Tammy Biniashvili,[1] Clara Soyris,[1] Schraga Schwartz,[1] Orna Dahan,[1,2] and Yitzhak Pilpel[1,2]

[1] *Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, 76100 Israel*

Adenosine (A) to inosine (I) RNA editing is widespread in eukaryotes. In prokaryotes, however, A-to-I RNA editing was only reported to occur in tRNAs but not in protein-coding genes. By comparing DNA and RNA sequences of *Escherichia coli*, we show for the first time that A-to-I editing occurs also in prokaryotic mRNAs and has the potential to affect the translated proteins and cell physiology. We found 15 novel A-to-I editing events, of which 12 occurred within known protein-coding genes where they always recode a tyrosine (TAC) into a cysteine (TGC) codon. Furthermore, we identified the tRNA-specific adenosine deaminase (tadA) as the editing enzyme of all these editing sites, thus making it the first identified RNA editing enzyme that modifies both tRNAs and mRNAs. Interestingly, several of the editing targets are self-killing toxins that belong to evolutionarily conserved toxin-antitoxin pairs. We focused on hokB, a toxin that confers antibiotic tolerance by growth inhibition, as it demonstrated the highest level of such mRNA editing. We identified a correlated mutation pattern between the edited and a DNA hard-coded Cys residue positions in the toxin and demonstrated that RNA editing occurs in *hokB* in two additional bacterial species. Thus, not only the toxin is evolutionarily conserved but also the editing itself within the toxin is. Finally, we found that RNA editing in *hokB* increases as a function of cell density and enhances its toxicity. Our work thus demonstrates the occurrence, regulation, and functional consequences of RNA editing in bacteria.

[Supplemental material is available for this article.]

RNA editing is a post-transcriptional process in which RNA bases are being altered (Knoop 2011). Adenosine (A) to inosine (I) RNA editing is the most prevalent form of editing in metazoans (Bazak et al. 2014). Inosine in turn can be identified by the translational or genetic machinery (e.g., reverse transcriptase) as a guanosine (G). A-to-I editing can recode proteins in eukaryotes (for example, humans and fungi) (Knoop 2011; Liu et al. 2016; Wang et al. 2016). The majority of editing events found in humans occur in untranslated regions, while only a small fraction of editing events are found in coding regions, of which only a few lead to nonsynonymous recoding (Ramaswami and Li 2014). All A-to-I editing events in mRNA are mediated by enzymes belonging to the ADAR (adenosine deaminase, RNA specific) family that was suggested to constitute a metazoan innovation (Grice and Degnan 2015). In bacteria, however, RNA editing was only reported in a single nucleotide site, within a tRNA for arginine, and it is mediated by the enzyme tRNA-specific adenosine deaminase (tadA) (Wolf et al. 2002).

Recent advances in sequencing technologies have facilitated the discovery of RNA modifications and edited sites in an unprecedented level both in the nucleus (Ramaswami et al. 2013; Bazak et al. 2014; Schwartz et al. 2014; Liu et al. 2016; Wang et al. 2016) and within organelles (Bar-Yaacov et al. 2013; Bentolila et al. 2013; Oldenkott et al. 2014). However, editing events in mRNA were so far not reported in bacteria.

Results

In order to identify novel editing events, we deep sequenced in parallel the RNA and DNA from two *Escherichia coli* strains (Fig. 1A). We used stringent parameters (Supplemental Fig. S1; Methods) to identify editing events that can manifest themselves as base differences between the DNA and RNA sequences. We identified 15 novel A-to-G RNA editing events (12 within known ORFs) in addition to the known editing site in tRNA-Arg (Fig. 1A; Supplemental Table S1). Strikingly, examining all 12 sites in which we detected editing within ORFs revealed that they are all predicted to recode a tyrosine (Tyr) encoded by the TAC codon into a cysteine (Cys) encoded by the TGC codon. While the majority of editing events were A-to-G, we also detected one additional genomic site which constituted a C-to-U substitution (which results in a synonymous substitution at the protein level) (Supplemental Table S1). All A-to-G editing events were embedded within a four-base-long motif TACG, with the edited A on the second position (Fig. 1B). Interestingly, this motif is completely identical to the known tadA recognition motif (Wolf et al. 2002) present on tRNA-Arg. In addition, tadA was previously shown to require for its activity a specific RNA secondary structure loop conformation around the edited site (Wolf et al. 2002). Indeed, RNA secondary structure modeling (Gruber et al. 2008) predicts that the edited base is also embedded within a loop in most of the newly identified sites (Fig. 1C; Supplemental Fig. S2). This raised the suspicion that

[2] **These authors contributed equally to this work.**
Corresponding author: pilpel@weizmann.ac.il
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.222760.117.

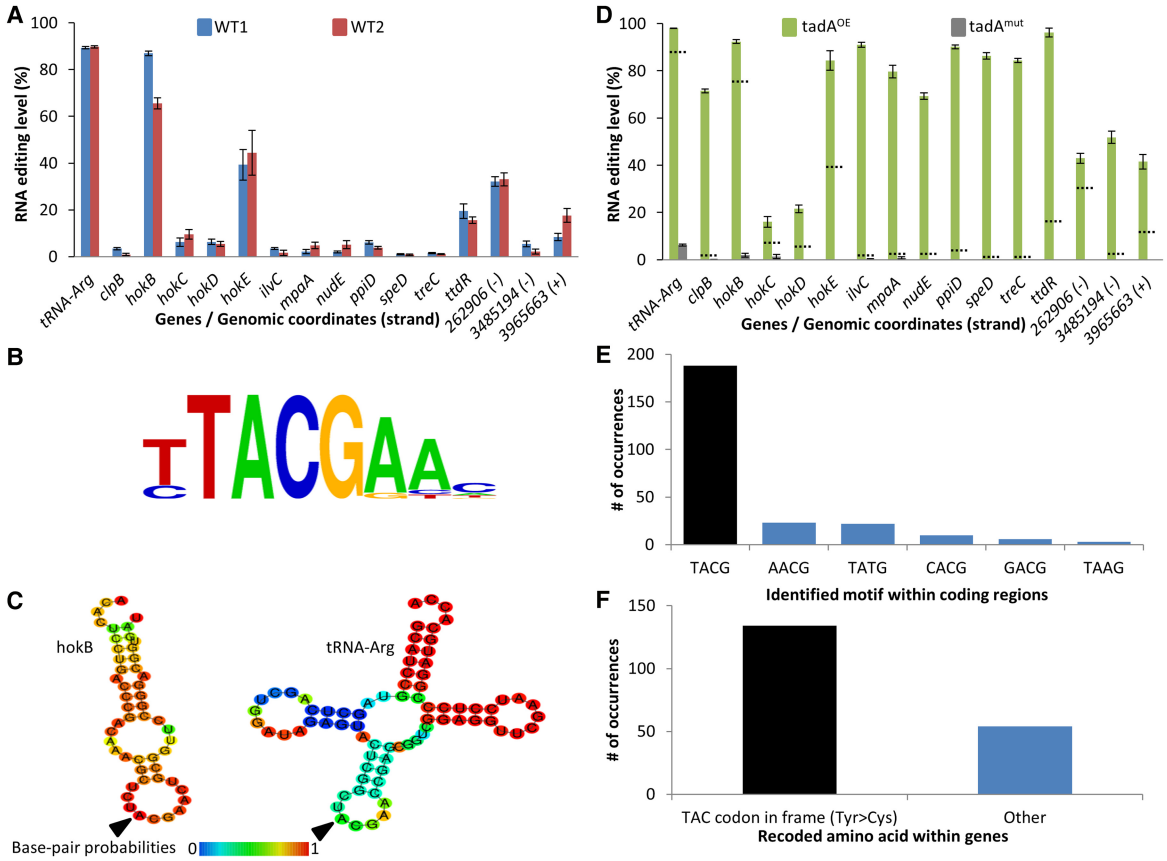


Figure 1. RNA editing occurs in *E. coli* and it is mediated by tadA. (*A*) RNA-seq (and DNA-seq) from two WT *E. coli* strains (Top10 and MG1655-EcM2.1, blue and red, respectively) reveals 15 novel A-to-G(I) RNA editing sites in *E. coli* in addition to the known editing site in tRNA-Arg. Notably, all sites found in known genes (12 out of the 15 sites) recode a tyrosine (TAC) into a cysteine (TGC) codon. The three RNA editing sites that do not occur in known ORFs are denoted by their genomic coordinates and genomic strand (+ or −). RNA editing levels are defined here as the number of reads with a G at the position out of all reads that cover the position. RNA samples were extracted in mid-log phase at $OD_{600} \sim 0.7$. (*B*) All sites share a common four-base DNA motif which is identical to tadA's recognition motif. (*C*) RNA secondary structure modeling predicts that edited sites are embedded within a loop. Here, the secondary structure of *hokB* (as well as tRNA-Arg) is presented (the RNA secondary structure modeling of all other targets found in this work is shown in Supplemental Fig. S2). (*D*) Overexpressing (green) or mutating (gray) *tadA* increases or reduces the editing level, respectively. Dotted lines represent the average editing levels measured for each gene in the two WT strains. RNA samples were extracted in mid-log phase at $OD_{600} \sim 0.5$–0.6. (*E*) Overexpression of *tadA* reveals additional putative editing sites, of which 75% are embedded within the canonical motif (TACG, black bar), while the rest deviate by one base from the canonical motif. (*F*) Out of 188 editing sites which occur within genes, 134 (black bar) recode a Tyr into a Cys codon (71%). Error bars in parts *A* and *D* represent standard errors of measuring editing level in a given coverage. Exact values can be found in Supplemental Tables S1 and S2.

tadA, which was so far believed to exclusively edit the anticodon of the tRNA-Arg, might be responsible for the editing of the aforementioned positions. Therefore, we performed RNA-seq on two additional strains, one overexpressing *tadA* from a plasmid and another harboring a *tadA* mutation (Supplemental Fig. S3) reported to completely abolish its activity in vitro and slightly reduce it in vivo in the NWL37 strain (Poulsen et al. 1992; Wolf et al. 2002). Since this strain was generated through an experimental lab evolution technique (interestingly, by exposing it to constant expression of the toxin hokC), we reasoned that it is possible that the evolved strain might also contain additional mutations. Therefore, rather than using this evolved strain, we have "surgically" introduced only the inactivating mutation into *tadA*'s genome version in the background of our strain (see Methods). Consistent with our hypothesis, overexpressing *tadA* dramatically increased the editing levels in all A-to-G sites, while in the *tadA* mutant, editing levels were abolished or dramatically reduced in all sites, including in the tRNA-Arg (Fig. 1D; Supplemental Fig. S4; Supplemental Table S1). Thus, all A-to-G substitutions in our RNA-seq data are likely to represent an adenosine to inosine editing event (hereafter A-to-I). Notably, editing levels of the C-to-U event (which did not harbor the tadA motif) were unaffected upon overexpression of *tadA* as well as in the *tadA* mutant (Supplemental Table S1). Overexpressing *tadA* revealed 252 additional A-to-I sites in coding regions that demonstrate RNA editing levels of at least 10% (Fig. 1E; Supplemental Table S2). Of these, 188 (75%) are embedded within the TACG recognition motif and show a significant enrichment for recoding a Tyr into a Cys codon ($\chi_1^2 = 1.35 \times 10^{-12}$) (Fig. 1E,F; Supplemental Tables S2, S3), raising the hypothesis that they might represent additional targets for this enzyme. Thus, we showed that A-to-I RNA editing in protein-coding genes occurs in bacteria, recodes Tyr into a Cys codon, and is mediated by tadA, an enzyme previously thought to be a tRNA-specific deaminase. Notably, there is no correlation between RNA editing levels and mRNA expression levels of the 12 genes that are edited by tadA in wild-type strains (Fig. 1A; Supplemental Fig. S5).

Interestingly, we found that four of the A-to-I editing sites observed in wild-type cells occur within the ORF of genes belonging

to the hok family of host-killing toxins (Fig. 1A). Proteins encoded by this family belong to the hok-sok toxin-antitoxin module that confer membrane de-polarization, which results in growth inhibition and potentially cell death (Pedersen and Gerdes 1999; Verstraeten et al. 2015). Multiple sequence alignment of the toxins belonging to the hok family revealed that the editing event which recodes a Tyr (TAC) into a Cys (TGC) codon at position 29 of hokB aligns against a conserved genome-encoded Cys residue in the other hok members. Remarkably, in the other toxins, hokC, hokD, and hokE, editing recodes another position in the peptide, 46, there too converting a Tyr (TAC) into a Cys (TGC) codon (Fig. 2A). hokB in turn harbors a DNA-encoded Cys at position 46. Thus, across the hok family of *E. coli* toxins, there are two positions, of which one is always hard-coded with Cys in the genome and the second contains a hard-code Tyr that can be converted into a second Cys upon editing of the RNA (except in hokA). The conserved Tyr is always encoded in these positions through the TAC codon which is contained in the editing motif TACG, and never with the synonymous codon for this amino acid, TAT, which does not confer to the editing motif consensus. The five genome-encoded hok genes share about a third of their sequence (i.e., conserved positions). At the RNA level, in all of them the edited site resides within a predicted secondary structure loop; however, *hokB*'s sequence around the edited site is the only one with complete identity to the loop of tRNA-Arg, which is the known substrate of tadA (Fig. 1C; Supplemental Fig. S2). All five Hok genes encode a short peptide (~50 amino acids) with an N-terminus that is embedded within the membrane, while the C-terminus is located within the periplasm (Poulsen et al. 1991). We therefore modeled the 3D structure of all five hok peptides which displayed a conserved 3D structure of an alpha helix at the N-terminus and two beta strands at the C-terminus (Supplemental Fig. S6). Notably, the residues at position 29 and 46 of hokB reside each in one of the two beta strands and are predicted here to be in close proximity to each other. Four of the toxins (hokA, hokC, hokD, and hokE) were reported to be inactive in *E. coli* (Pedersen and Gerdes 1999), thus raising the interesting possibility that high levels of RNA editing can be found in functional, rather than non-functional, hok members. Nevertheless, our results suggest that these genes are at least transcribed and that *hokC*, *hokD*, and *hokE* are edited (Fig. 1A; Supplemental Fig. S5; Supplemental Table S1). We therefore focused further on *hokB* that was shown

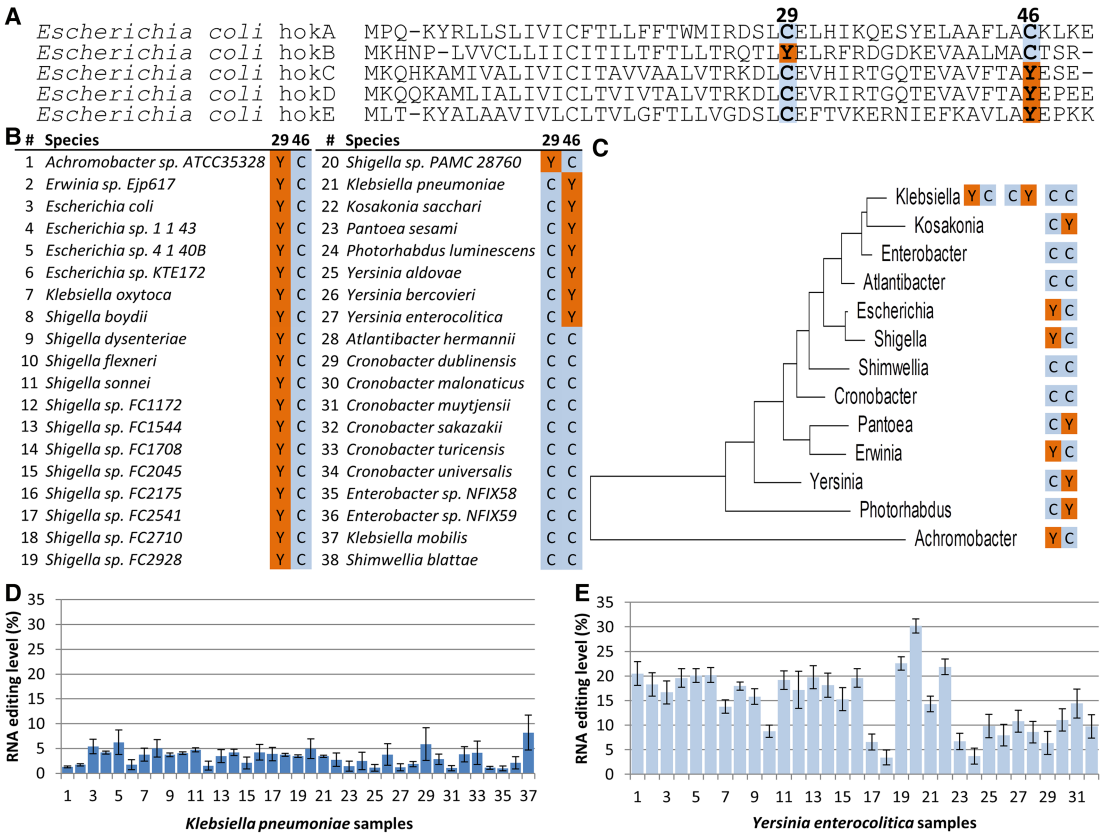


Figure 2. Evolutionary analyses suggest an interplay between the recoded and a hard-coded cysteines in hokB. (*A*) Multiple sequence alignment of five hok proteins encoded by the *E. coli* genome (NC_000913.3). The hokB edited version recapitulates the cysteine at position 29 which is hard-coded in the genome of all other hok protein family members. Symmetrically, hokC, hokD, and hokE editing sites (position 46) recapitulate the cysteine at the same position of hokB. (*B*) Multiple sequence alignment of hokB of a representative nonredundant set of orthologs from bacterial species harboring an annotated *hokB* gene suggests interplay between peptide residues at positions 29 and 46. Notably, all the Tyr codons at position 29 or 46 are encoded by the editable codon (TAC, embedded within the TACG motif). The complete alignment can be found in Supplemental Table S4. (*C*) A maximum likelihood phylogenetic tree based on the 16S rRNA gene, showing the amino acid composition at hokB's positions 29 and 46 in each bacterial genus with species harboring an annotated *hokB*. (*D*,*E*) RNA editing in *hokB* was identified in publicly available *Klebsiella pneumoniae* (37) and *Yersinia enterocolitica* (32) samples with sufficient coverage (≥51×) of RNA reads and at least two reads supporting an editing event. This editing event is predicted to recode position 46 (Tyr>Cys) in hokB. SRA accession numbers can be found in Supplemental Tables S5 and S6. Error bars represent standard errors of measuring editing level in a given coverage.

to be active (Verstraeten et al. 2015) and demonstrated the highest level of RNA editing. When we analyzed multiple sequence alignments of annotated *hokB*'s orthologs from different bacterial species, we found that most orthologs either have a Tyr encoded by the editable TAC codon (embedded within the TACG motif) at position 29 and a Cys at position 46, or, in other orthologs, a Cys at position 29 and an editable Tyr codon at position 46. Note that some species in this sample have Cys at both positions (Fig. 2B, C; Supplemental Table S4). This remarkable correlated pattern raised the question whether *hokB* mRNA editing can occur and can be detected in other species. Indeed, by analyzing multiple publicly available RNA-seq data sets (Leskinen et al. 2015), we observed A-to-I mRNA editing that recodes position 46 in two pathogenic bacteria, *Klebsiella pneumonia* and *Yersinia enterocolitica* (Fig. 2D,E; Supplemental Tables S5, S6). It is possible that *hokB* RNA is edited in additional species which currently are lacking publicly available RNA-seq data sets or only have an insufficient number of RNA reads (>10×) that cover *hokB*. Thus, not only is hokB evolutionarily conserved, but also RNA editing within it can be identified in species other than *E. coli*.

hokB was implicated in arresting cellular growth via membrane depolarization, and by doing so, it was found to mediate antibiotic tolerance through a mechanism of persistence (Verstraeten et al. 2015). Additionally, it was demonstrated that expression of *hokB* is elevated in response to starvation (Verstraeten et al. 2015). Since one of the characteristics of reaching culture stationary phase is a lack of nutrients, we aimed to examine if the editing levels of *hokB* change as a function of cellular density of the bacterial culture. Indeed, editing levels of this toxin's mRNA site were found to increase from ~28% at early logarithmic phase to ~93% when the culture enters stationary phase (Fig. 3A; Supplemental Table S7). Thus, in addition to elevation in toxin expression, *hokB*'s RNA editing levels are elevated during culture growth. We further asked how the predicted change in amino acid, from Tyr to Cys at position 29, can affect hokB's activity. To answer this question, we first mutated the genomic *hokB* gene. The first version was a positive control, as it contained the WT version of the toxin (*hokB*-WT) with the codon TAC coding for Tyr; the second version mimicked constitutive editing, with the codon TGC encoding for a Cys yet hard-coded into the DNA (*hokB*-Cys29); the third version of the mutated toxin had the Tyr at the edited position, yet with the synonymous codon TAT that is noneditable (*hokB*-Tyr29) (Supplemental Table S7). No observable difference in growth was detected between the three strains (Supplemental Fig. S7). This lack of observable phenotype could be expected given that *hokB*'s expression is governed by high levels of the alarmone (p)ppGpp, a condition that is observed in only 1/10,000 cells (Gerdes and Maisonneuve 2015) during logarithmic phase. Therefore, we utilized a previously established strategy (Verstraeten et al. 2015) of overexpressing *hokB* from a plasmid to

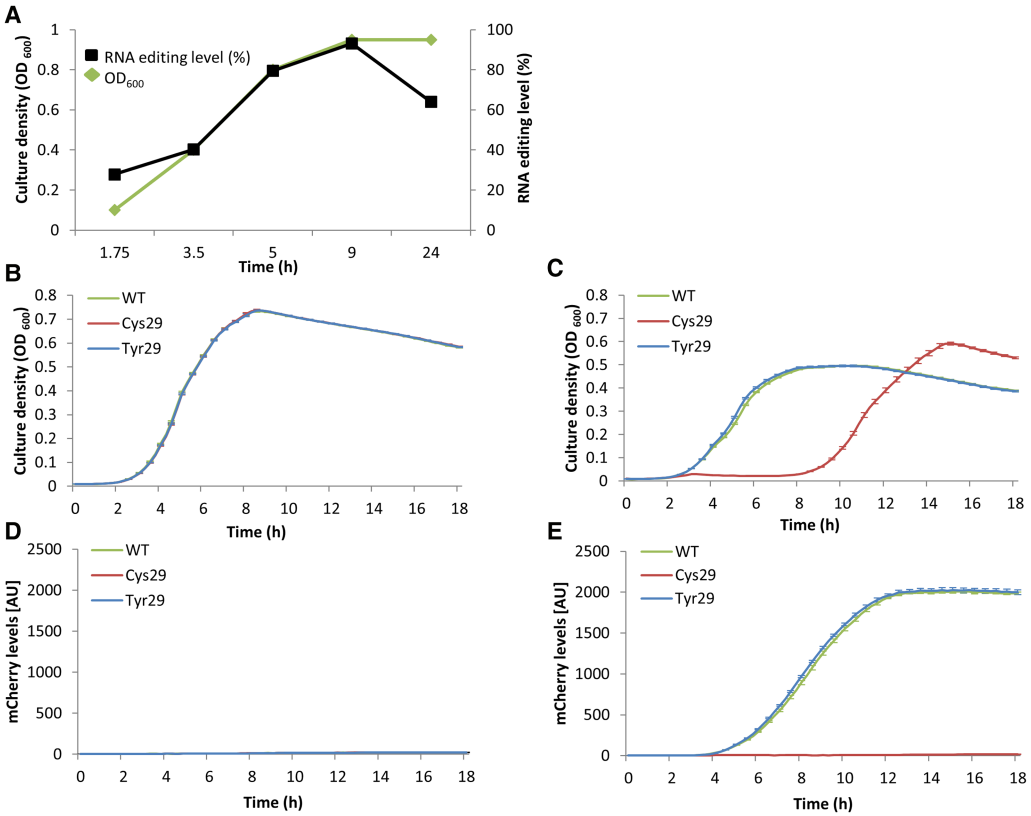


Figure 3. *hokB* mRNA editing increases with cell density and enhances its toxicity. (*A*) *hokB* mRNA editing levels (black) in *E. coli* MG1655-EcM2.1 WT strain as measured in different culture densities (green). Notably, the standard error of measuring editing levels in a given coverage in all samples was smaller than 0.00012%. (*B–E*) The *E. coli* Top10-Δ*hokB* strain was transformed with inducible plasmids harboring the WT (green), constitutively edited (Cys29, red), and noneditable (Tyr29, blue) *hokB* versions fused to mCherry reporter protein (N-terminus). (*B*) Growth analysis without induction of *hokB* (0% arabinose). (*C*) Growth analysis with induction of *hokB* (0.2% arabinose). (*D*) mCherry levels without induction (0% arabinose). (*E*) mCherry levels with induction (0.2% arabinose). Error bars represent standard error for 14 replicates (*B–E*).

acquire a detectable signal in order to facilitate the study of the activity of the three versions of the toxin we created. In order to detect potential functional differences between *hokB*-WT, *hokB*-Cys29, and *hokB*-Tyr29, we expressed the three versions of the toxin from an inducible plasmid in an *E. coli* strain that lacks its genome-encoded *hokB*. When each of the *hokB* versions is induced, we observed reduction in growth rate and yield compared to cells not expressing *hokB* (Fig. 3B,C). Yet, while the toxicity observed in the *hokB*-WT and *hokB*-Tyr29 versions is relatively mild, it is very dramatic in the *hokB*-Cys29 version (Fig. 3B,C). In fact, the Cys29 strain shows signs of growth only 9 h after induction of the toxin. Note that the growth observed in the Cys29 strain after 9 h is probably a result of some (potentially genetic) adaptation that allows the cells not to express the toxin and therefore support growth, while the two other strains still express it (Fig. 3D,E). In other words, the Cys29 version is so toxic that it only allows a small fraction of cells in the population that do not express *hokB* to grow. Since the frequency of such cells is very low (below the technical detection), growth of Cys29 cells is only visible after 9 h. Nonetheless, if we allow all three strains to grow to mid-logarithmic phase without expression of *hokB* and only then inducing *hokB* expression, we observe a clear mCherry signal, including in the Cys29 strain, indicating that *hokB* gene is intact and can be induced in all three variants (Supplemental Figs. S8–S10). This result further supports our conclusion that the lack of growth observed upon induction of the *hokB*-Cys29 version is due to hokB toxicity rather than lack of *hokB* expression from the plasmid. Thus, RNA editing that converts Tyr to Cys at position 29 of hokB enhances its toxicity.

Discussion

Why do bacteria exercise RNA editing? In mammalian cells, editing occurs mostly in noncoding regions (Bazak et al. 2014). In bacteria, however, as we have shown, out of the 15 novel A-to-I sites, 12 occur within known protein-coding genes and recode a Tyr into a Cys codon. RNA editing in bacterial coding regions could provide another layer of post-transcriptional regulation, and it can contribute to proteome diversity, as was recently suggested in cephalopods (e.g., octopus and squids) (Liscovitch-Brauer et al. 2017). In *hokB*'s case, RNA editing appears to provide another layer of regulation of toxicity. This editing-induced increase in toxicity of hokB could either represent a change in toxicity per protein molecule or an increase in its amount (e.g., by a stabilizing effect).

Turning editing on or off can affect the RNA and even the protein sequence within relatively short physiological time scales. This is demonstrated here in our observation that RNA editing levels in *hokB* increase with culture density. In addition, editing may allow cells to obtain both the edited and unedited versions of hokB, and even "play" with the ratio between them, generating phenotypic heterogeneity between genetically identical cells. Such cell-to-cell variability, when exercised in the activation pattern of host-killing toxins, can potentially affect the antibiotic persistence they confer and thus might form an even more complex bet-hedging mechanism than was previously suggested (Verstraeten et al. 2015).

Why do we observe different editing levels among the 15 newly discovered mRNA editing sites? It was shown that RNA secondary structure is important for tRNA-Arg editing (Wolf et al. 2002). Therefore, difference in secondary structure and/or additional sequence features might affect editing levels by affecting tadA-RNA interaction. Indeed, tadA's structure in complex with a tRNA-Arg loop (Losey et al. 2006) demonstrates that the enzyme interacts with seven of the tRNA substrate nucleotides that constitute the entire hepta-loop. Thus, loop size and additional sequence features could affect enzyme-substrate interaction and hence editing level. Indeed, the only newly discovered RNA target that has complete sequence identity to the tRNA seven loop nucleotides is *hokB*, which reassuringly shows the highest editing level, second only to the tRNA. All other targets either differ in their sequence (surrounding the core TACG motif) or loop size. Future studies are needed to examine if all detected mRNA editing events have functional consequences or whether some of them represent an "accidental" activity due to sequence/structure similarity to tadA's substrates. *tadA* is found in most bacterial species (Yokobori et al. 2013). Therefore, our work sets the stage for investigating RNA editing in other bacterial species that harbor this enzyme. Moreover, *tadA*'s orthologs (such as *Tad1p* and *ADAT*) are found in eukaryotes (yeast [Wang et al. 2016] and human [Grice and Degnan 2015], for example). Since we now implicated tadA in mRNA editing, in addition to its established role in tRNA editing, future studies should examine whether its orthologs are involved in mRNA editing in other organisms too. In conclusion, RNA editing occurs in bacteria and can recode protein sequences, potentially affecting their function as well as cell physiology, at least in *hokB*'s case. Thus, sequence variation among bacteria should also be examined at the RNA level.

Methods

RNA and DNA purification

RNA and DNA were purified using the GeneJET RNA Purification kit (Thermo Fisher Scientific) and Wizard Genomic DNA Purification kit (Promega), respectively, according to the manufacturer's protocol. Cultures were grown on LB supplemented with ampicillin (100 µg/mL). RNA was purified from a culture at the middle of logarithmic phase (OD_{600} in a 1-cm cuvette ~0.8) for whole transcriptome sequencing and *hokB* MAGE (multiplex automated genome engineering) strains; different ODs as specified in Figure 3A. DNA was extracted and purified at stationary phase.

cDNA synthesis

One microgram of total RNA was subjected to cDNA synthesis using either the M-MLV cDNA Synthesis kit (Promega) or SuperScript II (Thermo Fisher Scientific), following the manufacturer's protocol.

PCR reaction mix, primers, and conditions

All data regarding PCR reaction mix, conditions, and primers can be found in Supplemental Tables S8 and S9. PCR products were visualized by an EtBr-stained 1% agarose gel. PCR fragments were purified using a Wizard SV Gel and PCR Clean-Up System (Promega), following the manufacturer's protocol.

Massively parallel deep sequencing

RNA was treated with a Ribo-Zero rRNA Removal kit (Illumina). Libraries for sequencing RNA to examine RNA editing levels in different optical densities of microbial cultures (OD_{600}) were constructed by designing PCR primers targeting the *hokB* gene with tails that match Illumina adapters (PCR1). A second PCR (PCR2) was carried out to attach the adapters for the Illumina run. Total DNA and RNA libraries of wild-type and of *hokB* versions that are expressed from plasmids were sequenced using 151-nt or 75-nt

paired-end reads, respectively, on the NextSeq platform (Illumina). Bacterial strains that were sequenced were: RNA-seq – WT1 (Top10), WT2 (MG1655-EcM2.1), *tadA* overexpression (Top10 harboring a pBAD plasmid overexpressing *tadA*), and *tadA* mutant (MG1655-EcM2.1 strain with an introduced *tadA* mutation); and DNA-seq – WT1 (Top10) and WT2 (MG1655-EcM2.1).

Analysis of massively parallel sequencing data

E. coli sequencing reads were aligned against the MG1655 Genome (NC_000913.3). For multiple sequence alignment, we utilized BWA-MEM (Li and Durbin 2009) with default parameters. Only reads that were aligned to the corresponding genome were used for further analyses. SAMtools (Li et al. 2009) was used to convert the SAM to the BAM sequence format. MitoBam Annotator (Zhidkov et al. 2011) transformed the BAM files into tables containing all parameters (e.g., base composition, coverage per base, etc.) for each position in the *E. coli* genome. These tables were used to identify sites that differ between RNA samples and their corresponding DNA base. Initial RNA editing sites were considered high quality only if identified in at least 30 sequence reads that contain a high-quality base call (≥30 Phred quality score); if their minimal read fraction was at least 3%; if each site contained at least five forward and reverse reads; they presented a mixture of only 2 nucleotides; and after manual inspection using the Integrative Genomics Viewer (Robinson et al. 2011) to exclude signal stemming from the edges of reads or low-complexity regions (mononucleotide regions). We also aimed to identify edited sites below 3% (but more than 1%). Therefore, after establishing that tadA was the mediating enzyme, we searched for sites shared between the two WT strains and the *tadA* overexpressing strain that display an A-to-I editing level of at least 1% in the WT and harbor a tadA recognition motif. Three additional sites were detected, demonstrating a dramatic increase in editing levels when *tadA* was overexpressed (Fig. 1A,D; Supplemental Table S1). In addition, BLAST analysis excluded that the RNA editing signal (a mixture of nucleotides at the identified edited sites) stems from paralogous regions in the *E. coli* genome (Supplemental Table S10). BLAST was performed by searching against the *E. coli* genome with a 101-bp fragment encompassing the edited site plus 50 bases upstream of and downstream from it. Finally, we examined and compared the final editing sites between all four RNA-seq and two DNA-seq data sets we obtained (some did not pass the initial thresholds). RNA expression analysis was performed by using python/2.7.6 HTSeq (Anders et al. 2015) for read count per gene. FPKM values were calculated manually.

Multiplex automated genome engineering

In order to create a mutation in *tadA* (T2697699G, D64E) that was previously shown to reduce its activity (Wolf et al. 2002), we used the *E. coli* strain MG1655-EcM2.1 (a specially designed strain for high MAGE efficiency) to carry out one successive MAGE cycle as previously described (Wang et al. 2009; Bar-Yaacov et al. 2016). We used 77-bp single-strand oligonucleotides to target the lagging strand in the *tadA* gene: G*A*TAATTTTGCATCACC AGACCACCCTGCCGCAGGGCCATGATTTCTGCATGTGCGGTG GGCTCATGGCGACCAA*T*C. Similarly, we used MAGE to recode *hokB* gene sequence once into a TGC Cys codon (T1491986C) mimicking constitutive editing and once into a noneditable TAT Tyr codon (C1491985T, synonymous Tyr mutation). We used the following 90-bp single-strand oligonucleotides to target the lagging strand in the *hokB* gene: A*T*CTGCATTACGATTCTGACA TTCACACTCCTGACCCGACAAACGCTCTGCGAACTGCGGTTC CGGGACGGTGATAAGGAGGTTGCTG*C*G (for T1491986C) and A*T*CTGCATTACGATTCTGACATTCACACTCCTGACCCGA CAAACGCTCTATGAACTGCGGTTCCGGGACGGTGATAAGGAG GTTGCTG*C*G (for C1491985T). The mutated base is underlined, and asterisks represent phosphorothioate bonds. Briefly, cells were grown overnight at 30°C. Then, 30 µL of the saturated culture were transferred into fresh 3 mL of LBL (10 g of tryptone, 5 g of NaCl, and 5 g of yeast extract per liter) medium until reaching an OD = 0.4 (measured in a 1-cm cuvette in this section) and then moved to a shaking water bath (350 RPM) at 42°C for 15 min, after which it was moved immediately to ice. Next, 1 mL was transferred to an Eppendorf tube and cells were washed twice with double-distilled water (DDW) at a centrifuge speed of 13,000*g* for 30 sec. Next, the bacterial pellet was dissolved in 50 µL of DDW containing 2 µM of SS-DNA oligo and transferred into a cuvette. Electroporation was performed at 1.78 kV, 200 ohms, 25 µF. After electroporation, the bacteria were transferred into 2 mL of fresh LBL and incubated at 34°C until reaching an OD = 0.8, diluted in $1:10^{-4}$ and $1:10^{-5}$ ratios, and seeded on LB+ampicillin agar plates (100 µg/mL). To identify positive MAGE colonies (referred to as bacterial strains throughout the text), we PCR-amplified fragments encompassing the *E. coli* genomic region with primers corresponding to the mutated and WT form (differing in one base in their 3′ end – PCR3). Successful PCR amplification implies successful MAGE mutagenesis. To verify this interpretation, we amplified a second fragment encompassing the mutated position in *tadA* (PCR4) and Sanger sequenced it. Similarly, we used PCR5 (to identify colonies) and PCR6 (to validate the colonies using Sanger sequencing) to validate MAGE mutagenesis in *hokB*. The sequences were aligned and visualized using SnapGene Viewer 3.1.2 (GSL Biotech LLC).

Plasmid construction and transformation

In order to examine the functional role of RNA editing in *hokB*, we utilized the plasmid previously used to examine hokB's activity (Verstraeten et al. 2015) and constructed two additional plasmids (also harboring an ampicillin resistance cassette): pBAD-mCherry-linker-*hokB*(WT)—a generous gift from Prof. Jan Michiels from KU Leuven–University of Leuven—pBAD-mCherry-linker-*hokB*(Cys29), and pBAD-mCherry-linker-*hokB*(Tyr29). By using PCR7 and PCR8, we mutated the TAC codon corresponding to position 29 in *hokB* into a TGC (Tyr>Cys) and TAT (editable Tyr>noneditable Tyr) codons, respectively. All three plasmids were transformed into a Top10-Δ*hokB* strain (another generous gift from Prof. Jan Michiels) (Verstraeten et al. 2015).

We also constructed a pBAD-mCherry-linker-*tadA*(WT) plasmid. Specifically, we amplified the *tadA* gene and the plasmid backbone with overlapping (~20 nt) tails using PCR9 and PCR10. These fragments were subjected for NEB-assembly (New England Biolabs) according to the manufacturer's protocol. The plasmid was transformed into a Top10 WT strain, and single colonies were isolated, grown, and frozen (−80°C) for future assays.

PCR and Sanger sequence of *hokB*

PCR11 was performed to sequence the *hokB* gene/transcript from corresponding DNA and RNA samples from the WT strain (MG1655-EcM2.1) as well as RNA from the *tadA* mutant.

Liquid growth measurements

Cultures were grown at 30°C and 37°C for the genomic (MG1655-EcM2.1)- and plasmid (Top10-Δ*hokB*)-encoded *hokB* versions for 48 h in LB medium, back diluted in a 1:100 ratio, and dispensed on 96-well plates containing LB medium supplemented with 100 µg/mL ampicillin (150 µL per well) and either with arabinose (final concentration of 0.2% (Fig. 3C,E) or without arabinose (no *hokB*

expression; control) (Fig. 3B,D). Wells were measured for optical density at OD_{600} and mCherry fluorescence levels at 575 nm (excitation) and 620 nm (emission) wavelengths. Measurements were taken at 15-min intervals. Growth comparisons were performed using 96-well plates (Thermo Scientific). For each strain harboring a different version of *hokB*, a growth curve was obtained by averaging over well-dispersed 14 wells. The 96-well plate was divided as following: 12 wells are blank control and the remaining 84 wells were divided between the bacterial stains harboring the plasmid encoding for the three different *hokB* versions, with and without induction; thus, 84/6 = 14. Measurements for Figure 3A were conducted in a 1-cm cuvette, and values were divided by 2 for qualitative presentation purposes and comparison to the measurements shown in the growth curves (Fig. 3B,C).

Detecting mRNA editing in *hokB* of *Yersinia enterocolitica* and *Klebsiella pneumoniae*

RNA-seq data sets were downloaded from the SRA database (https://www.be-md.ncbi.nlm.nih.gov/sra). Accession numbers and parameters (e.g., coverage per base) of samples with identified mRNA editing in *hokB* are found in Supplemental Table S5 (*K. pneumoniae*) and S6 (*Y. enterocolitica*). We analyzed 46 *Y. enterocolitica* and 338 *K. pneuomoniae* samples and detected editing in 32 and 37 samples, respectively. Alignment and file manipulation were performed as described above. Notably, the rest of the species in Figure 2B were not assessed for their editing level since they did not meet the criteria of having a sufficient number of RNA reads (>10×) that cover *hokB* or did not have publicly available RNA-seq data sets.

Identifying tadA's motif

We used weblogo at http://weblogo.berkeley.edu/logo.cgi to identify the conserved, four-base motif which is identical to the tadA recognition motif.

RNA secondary structure prediction

In order to examine the RNA secondary structure, we extracted the RNA sequence 25 bases upstream of and downstream from the edited site (inclusive). We then used this sequence in the Vienna RNA Websuite (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi) to predict the RNA secondary structure with minimum free energy.

hok proteins 3D structure prediction

We used the RaptorX (Källberg et al. 2012) package at http://raptorx.uchicago.edu/StructurePrediction/predict/ to predict the 3D structure of hokA, B, C, D, and E (using default parameters).

Analysis of *hokB* orthologs in other bacterial species

In order to examine the amino acid composition at positions 29 and 46 of hokB, we downloaded *hokB* gene sequence from organisms with annotated *hokB* from the NCBI nucleotide website (https://www.ncbi.nlm.nih.gov/nuccore/). We constructed a nonredundant set of orthologs with one *hokB* sequence (gene and protein) per species (Supplemental Table S4). The sequence identities at positions 29 and 46 are presented in Figure 2B.

Multiple sequence alignment

MSA was performed by using ClustalW (default parameters) embedded in the MEGA5 package (Tamura et al. 2011) and the MAFFT server (http://mafft.cbrc.jp/alignment/server/).

Phylogenetic analyses

We used the 16S ribosomal RNA to build a genus phylogenetic tree to visualize the amino acid composition in hokB's positions 29 and 46 in an evolutionary context. We used the 16S ribosomal RNA from one representative from each genus (Supplemental Table S11). The evolutionary tree was inferred by using the maximum likelihood method based on the Tamura–Nei model (Tamura and Nei 1993). The tree with the highest log likelihood (−4907.0796) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Joining and BIONJ algorithms to a matrix of pairwise distances estimated using the maximum composite likelihood (MCL) approach and then selecting the topology with superior log likelihood value (Saitou and Nei 1987; Gascuel 1997). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 13-nt sequences. All positions containing gaps and missing data were eliminated. There were a total of 1398 positions in the final data set. Evolutionary analyses were conducted in MEGA5 (Tamura et al. 2011).

Statistical analysis

In order to examine whether the enrichment for recoding a Tyr into a Cys codon is significant, we performed a test for goodness of fit. Specifically, we counted how many times TACG occurs in coding regions and in which frame it occurs, and thus, what is the amino acid change predicted to occur upon RNA editing. We then compared it to the distribution we obtained from sites detected to be edited (>10%) after overexpressing *tadA*. See Supplemental Table S3 for numbers and calculations.

Confocal microscopy

Cells were grown for 3.5 h without arabinose until reaching midlogarithmic phase and then induced with 0.2% arabinose (final concentration) for 1 h. Cells were visualized under a confocal microscope (LSM 780, Zeiss) to obtain high-resolution images (10 µm). As a control, we used uninduced cells that were taken from the same culture prior to adding arabinose. Image processing was performed using FIJI (Schindelin et al. 2012).

Data access

The RNA and DNA sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRP103577. The Sanger sequences that correspond to the chromatograms in this study have been submitted to the NCBI GenBank database (https://www.ncbi.nlm.nih.gov/genbank/) under accession numbers MF554632–MF554636.

Acknowledgments

We thank the DIP foundation for grant support. We thank the Minerva Foundation for establishment of the Minerva Center for Live Emulation of Genome Evolution in the Lab. We thank Professor Jan Michiels and Dr. Natalie Verstraeten from KU Leuven–University of Leuven for providing strains and plasmids. We thank Dima Zabezhinsky for his help in the confocal microscopy analysis. We thank Shlomit Gilad and the INCPM center at the Weizmann Institute for their part in sequencing the total DNA and RNA samples of *E. coli*. We thank Nathalie Balaban, Erez Levanon, Avigdor Eldar, Rotem Sorek, Deborah Fass, and all the people in the Pilpel lab for useful discussion and comments.

Author contributions: D.B.Y. raised the original idea and performed all the experiments; E.M., R.T., C.S., and O.D. participated in experiments; D.B.Y., E.M., O.D., and Y.P. designed the experiments; D.B.Y., E.M., T.B., O.D., and Y.P. analyzed the data; Y.P. supervised the project; D.B.Y., E.M., S.S., O.D., and Y.P. interpreted the results; D.B.Y. and Y.P. wrote the manuscript.

References

Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31:** 166–169.

Bar-Yaacov D, Avital G, Levin L, Richards AL, Hachen N, Jaramillo BR, Nekrutenko A, Zarivach R, Mishmar D. 2013. RNA–DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA. *Genome Res* **23:** 1789–1796.

Bar-Yaacov D, Frumkin I, Yashiro Y, Chujo T, Ishigami Y, Chemla Y, Blumberg A, Schlesinger O, Bieri P, Greber B. 2016. Mitochondrial 16S rRNA is methylated by tRNA methyltransferase TRMT61B in all vertebrates. *PLoS Biol* **14:** e1002557.

Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, Isaacs FJ, Rechavi G, Li JB, Eisenberg E, et al. 2014. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res* **24:** 365–376.

Bentolila S, Oh J, Hanson MR, Bukowski R. 2013. Comprehensive high-resolution analysis of the role of an *Arabidopsis* gene family in RNA editing. *PLoS Genet* **9:** e1003584.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* **14:** 685–695.

Gerdes K, Maisonneuve E. 2015. Remarkable functional convergence: alarmone ppGpp mediates persistence by activating type I and II toxin-antitoxins. *Mol Cell* **59:** 1–3.

Grice LF, Degnan BM. 2015. The origin of the ADAR gene family and animal RNA editing. *BMC Evol Biol* **15:** 4.

Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. 2008. The Vienna RNA Websuite. *Nucleic Acids Res* **36:** W70–W74.

Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J. 2012. Template-based protein structure modeling using the RaptorX web server. *Nat Protoc* **7:** 1511–1522.

Knoop V. 2011. When you can't trust the DNA: RNA editing changes transcript sequences. *Cell Mol Life Sci* **68:** 567–586.

Leskinen K, Varjosalo M, Skurnik M. 2015. Absence of YbeY RNase compromises the growth and enhances the virulence plasmid gene expression of *Yersinia enterocolitica* O:3. *Microbiology* **161:** 285–299.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760.

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

Liscovitch-Brauer N, Alon S, Porath HT, Elstein B, Unger R, Ziv T, Admon A, Levanon EY, Rosenthal JJ, Eisenberg E. 2017. Trade-off between transcriptome plasticity and genome evolution in cephalopods. *Cell* **169:** 191–202.e111.

Liu H, Wang Q, He Y, Chen L, Hao C, Jiang C, Li Y, Dai Y, Kang Z, Xu J-R. 2016. Genome-wide A-to-I RNA editing in fungi independent of ADAR enzymes. *Genome Res* **26:** 499–509.

Losey HC, Ruthenburg AJ, Verdine GL. 2006. Crystal structure of *Staphylococcus aureus* tRNA adenosine deaminase TadA in complex with RNA. *Nat Struct Mol Biol* **13:** 153–159.

Oldenkott B, Yamaguchi K, Tsuji-Tsukinoki S, Knie N, Knoop V. 2014. Chloroplast RNA editing going extreme: more than 3400 events of C-to-U editing in the chloroplast transcriptome of the lycophyte *Selaginella uncinata*. *RNA* **20:** 1499–1506.

Pedersen K, Gerdes K. 1999. Multiple *hok* genes on the chromosome of *Escherichia coli*. *Mol Microbiol* **32:** 1090–1102.

Poulsen L, Refn A, Molin S, Andersson P. 1991. Topographic analysis of the toxic Gef protein from *Escherichia coli*. *Mol Microbiol* **5:** 1627–1637.

Poulsen LK, Larsen NW, Molin S, Andersson P. 1992. Analysis of an *Escherichia coli* mutant strain resistant to the cell-killing function encoded by the *gef* gene family. *Mol Microbiol* **6:** 895–905.

Ramaswami G, Li JB. 2014. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res* **42:** D109–D113.

Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB. 2013. Identifying RNA editing sites using RNA sequencing data alone. *Nat Methods* **10:** 128–132.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative Genomics Viewer. *Nat Biotechnol* **29:** 24–26.

Saitou M, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406–425.

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B. 2012. Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9:** 676–682.

Schwartz S, Bernstein DA, Mumbach MR, Jovanovic M, Herbst RH, León-Ricardo BX, Engreitz JM, Guttman M, Satija R, Lander ES. 2014. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell* **159:** 148–162.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10:** 512–526.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28:** 2731–2739.

Verstraeten N, Knapen WJ, Kint CI, Liebens V, Van den Bergh B, Dewachter L, Michiels JE, Fu Q, David CC, Fierro AC. 2015. Obg and membrane depolarization are part of a microbial bet-hedging strategy that leads to antibiotic tolerance. *Mol Cell* **59:** 9–21.

Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM. 2009. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460:** 894–898.

Wang IX, Grunseich C, Chung YG, Kwak H, Ramrattan G, Zhu Z, Cheung VG. 2016. RNA–DNA sequence differences in *Saccharomyces cerevisiae*. *Genome Res* **26:** 1544–1554.

Wolf J, Gerber AP, Keller W. 2002. tadA, an essential tRNA-specific adenosine deaminase from *Escherichia coli*. *EMBO J* **21:** 3841–3851.

Yokobori S-i, Kitamura A, Grosjean H, Bessho Y. 2013. Life without $tRNA^{Arg}$–adenosine deaminase TadA: evolutionary consequences of decoding the four CGN codons as arginine in Mycoplasmas and other Mollicutes. *Nucleic Acids Res* **41:** 6531–6543.

Zhidkov I, Nagar T, Mishmar D, Rubin E. 2011. MitoBamAnnotator: a web-based tool for detecting and annotating heteroplasmy in human mitochondrial DNA sequences. *Mitochondrion* **11:** 924–928.

Received March 11, 2017; accepted in revised form August 7, 2017.

Appendix: Prediction of ionization efficiency from amino acid composition

1 Ionization Efficiency (IE) prediction.

1.1 Notations

Let $peptide_i$ be a product of digestion of $protein_j$.

I_i : Intensity of $peptide_i$

IE_i : Ionization Efficiency of $peptide_i$

c_i : concentration of $peptide_i$ after trypsin digestion

c_j : concentration of $protein_j$ before trypsin digestion

n : total number of detected peptides

1.2 Model

By definition, $I_i = IE_i c_i$. Assuming a perfectly efficient trypsin digestion, if $peptide_i$ is tryptic, then $c_i = c_j$. Thus, for any peptide i belonging to protein j :

$$\log_{10}(I_i) = \log_{10}(IE_i) + \log_{10}(c_j)$$

Let $y_i = \log_{10}(I_i)$, and $b_j = \log_{10}(c_j)$.

$$y_i = \log_{10}(IE_i) + b_j$$

We model $\log_{10}(IE_i)$ as a linear combination of a K-long feature vector $X_{i,k}$ derived from $peptide_i$'s amino acid sequence, weighted with weights $W = [w_1, ..., w_K]$

$$\log_{10}(IE_i) = \sum_k (w_k X_{i,k}) + \epsilon_i$$
$$y_i = \sum_k (w_k X_{i,k}) + b_j + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma^2)$, and ϵ_i is independent of X. The equation resembles that of a Gaussian-noise simple linear regression model, except for the fact that there is a separate intercept b_j per protein. The likelihood of the system is given by:

$$\begin{aligned}\mathcal{L} &= \prod_j \prod_{i \in protein_j} p(y_i | x_i; W, B, \sigma^2) \\ &= \prod_j \prod_{i \in protein_j} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (b_j + \sum_k w_k X_{i,k}))^2}{2\sigma^2}} \\ \log(\mathcal{L}) &= \frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_j \sum_{i \in protein_j} (y_i - b_j - \sum_k w_k X_{i,k})^2\end{aligned}$$

For a refresher concerning maximum likelihood fitting of parameters in the case of simple linear regression, see this article.

1.3 Data preparation

We selected unmodified peptides of charge 2 to train our model. Using the MQ output file 'evidence.txt' as our input. Peptides belonging to proteins detected by only one peptide were excluded from the analysis. The remaining peptides were grouped by sequence, and their intensity was defined as the sum of their intensities across all fractions.
For each peptide, we then computed features capturing their amino acid composition, their length, and potential mis-cleavage issues (see Table 1)

Features were then centered and reduced to unit variance to create the feature matrix X

1.4 Parameter fitting and results

We obtained maximum likelihood estimates for W, B and σ by maximizing the log likelihood described in section 1.2., using the L-BFGS-B implementation of Python's sklearn module. Figure 1 shows a good agreement between the predicted $\log_{10}(IE)$ (defined as $\sum_k(w_k X_{i,k})$) and the observed $\log_{10}(IE)$ (defined as $y_i - b_j$), with a Pearson correlation coefficient R = 0.67. In order to control for over-fitting, we further divided the set of proteins in two equally sized groups, and fitted the weights W and $\log_{10}(protein\,levels)$ B separately for both groups. We found a very good agreement between the W vectors in both groups, confirming that we were not over-fitting the data (Figure 2). We report the value of the fitted W coefficients in Figure 3

Table 1: Features computed for this analysis

Name	Description	Length
$count_{AA}$	# of occurences of AA in peptide	20
$count_{RP}$	# of occurences of the subsequence 'RP' in peptide	1
$count_{KP}$	# of occurences of the subsequence 'KP' in peptide	1
$N_{term}\,Pro$	1 if peptide starts with Pro, 0 otherwise	1
-2 is R	1 if the aa in position -2 relative to the N_{term} cleavage site is 'R', 0 otherwise	1
-2 is K	1 if the aa in position -2 relative to the N_{term} cleavage site is 'K', 0 otherwise	1
-1 is R	1 if the aa in position -1 relative to the N_{term} cleavage site is 'R', 0 otherwise	1
-1 is K	1 if the aa in position -1 relative to the N_{term} cleavage site is 'K', 0 otherwise	1
+1 is R	1 if the aa in position +1 relative to the C_{term} cleavage site is 'R', 0 otherwise	1
+1 is K	1 if the aa in position +1 relative to the C_{term} cleavage site is 'K', 0 otherwise	1
+1 is P	1 if the aa in position +1 relative to the C_{term} cleavage site is 'P', 0 otherwise	1
inverse length	inverse of the peptide's length	1
length	length of the peptide	1
	Total	32

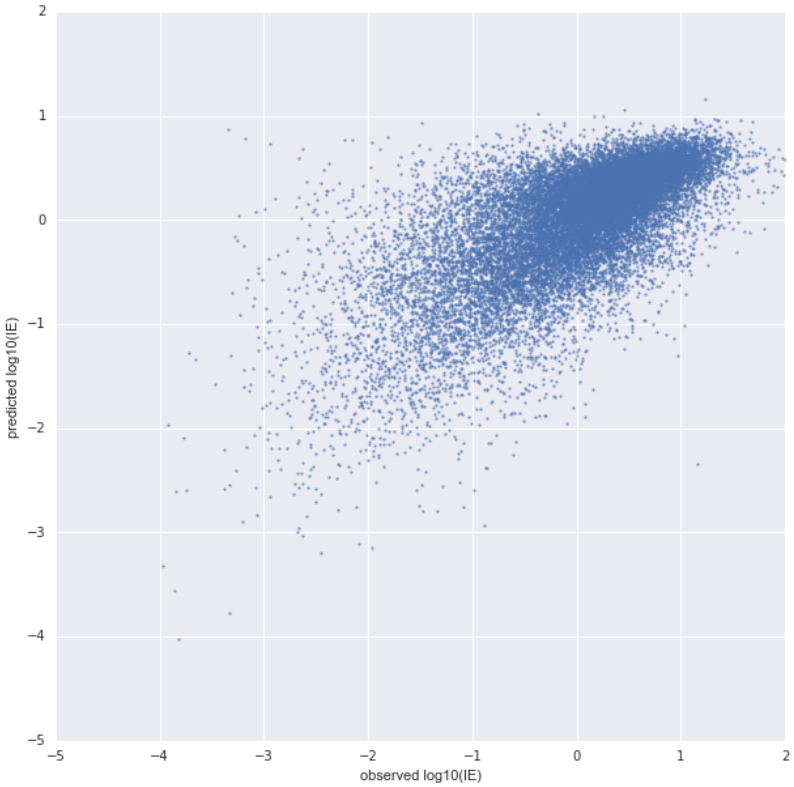


Figure 1: The predicted $\log_{10}(IE)$ was computed as $\sum_k(w_k X_{i,k})$, and the observed $\log_{10}(IE)$ was defined as $y_i - b_j$. Pearson correlation coefficient = 0.69, $\sigma = 0.58$

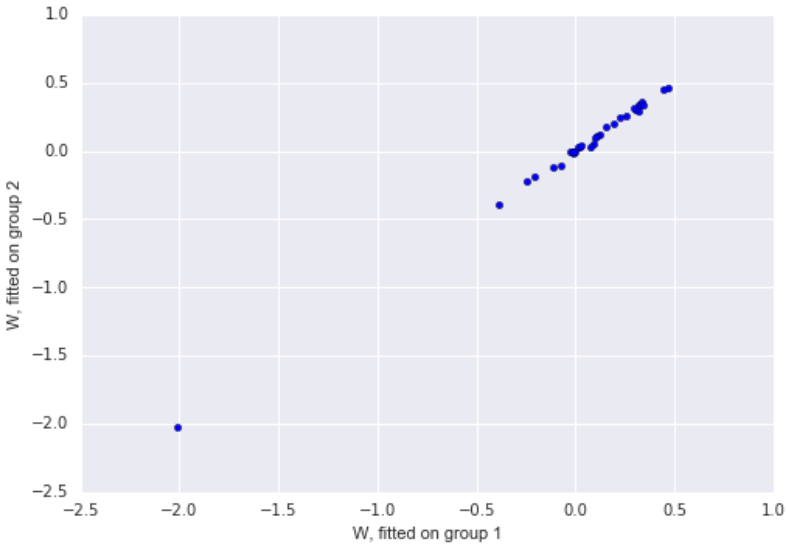


Figure 2: Weights of the regression coefficients W, fitted separately on each half of the dataset, are plotted against one another. Pearson correlation coefficient > 0.999

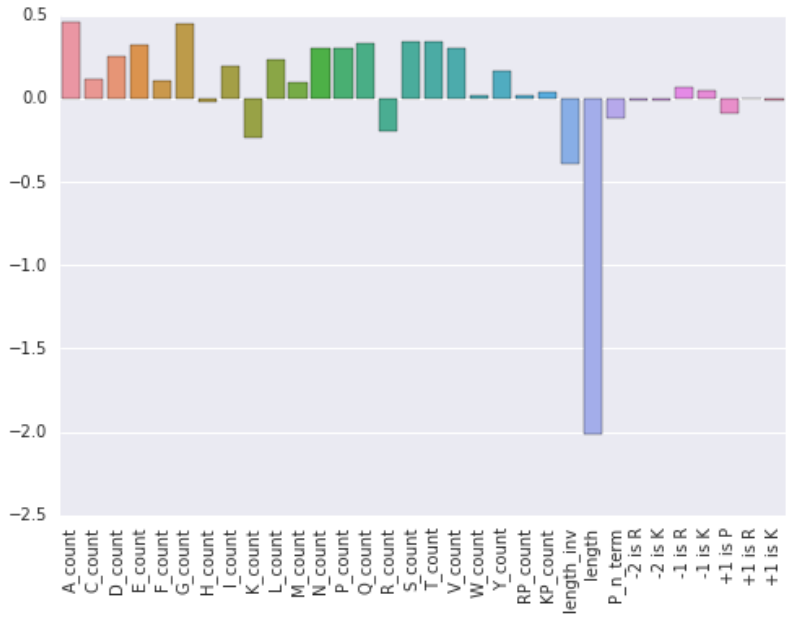


Figure 3: Weights of the regression coefficients W, fitted on the entire dataset

References:

1. Machnicka, M. A., Olchowik, A., Grosjean, H. & Bujnicki, J. M. Distribution and frequencies of post-transcriptional modifications in tRNAs. *RNA Biol.* **11,** 1619–1629 (2014).
2. Dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. doi:10.1093/nar/gkh834
3. Bullwinkle, T. J. & Ibba, M. Emergence and evolution. *Top. Curr. Chem.* **344,** 43–87 (2014).
4. Deusser, E. Heterogeneity of ribosomal populations in Escherichia coli cells grown in different media. *Mol. Gen. Genet.* **119,** 249–58 (1972).
5. Simsek, D. & Barna, M. An emerging role for the ribosome as a nexus for post-translational modifications. *Curr. Opin. Cell Biol.* **45,** 92–101 (2017).
6. Laursen, B. S., Sørensen, H. P., Mortensen, K. K. & Sperling-Petersen, H. U. Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev.* **69,** 101–23 (2005).
7. Schmeing, T. M. & Ramakrishnan, V. What recent ribosome structures have revealed about the mechanism of translation. *Nature* **461,** 1234–1242 (2009).
8. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181,** 223–30 (1973).
9. Rooman, M., Dehouck, Y., Kwasigroch, J. M., Biot, C. & Gilis, D. What is Paradoxical about Levinthal Paradox? *J. Biomol. Struct. Dyn.* **20,** 327–329 (2002).
10. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* **32,** 120D–121 (2004).
11. Jacobs, W. M. & Shakhnovich, E. I. Evidence of evolutionary selection for co-translational folding. (2017). doi:10.1073/pnas.1705772114
12. Kerner, M. J. *et al.* Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli. *Cell* **122,** 209–20 (2005).
13. Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M. & Ulrich Hartl, F. Molecular Chaperone Functions in Protein Folding and Proteostasis. *Annu. Rev. Biochem.* **82,** 323–355 (2013).
14. Alon, U. *An introduction to systems biology : design principles of biological circuits*. (Chapman & Hall/CRC, 2007).
15. Powers, E. T., Powers, D. L. & Gierasch, L. M. FoldEco: A Model for Proteostasis in E. coli. *Cell Rep.* **1,** 265–276 (2012).
16. Lindner, A. B., Madden, R., Demarez, A., Stewart, E. J. & Taddei, F. Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 3076–81 (2008).
17. Wallace, E. W. J. *et al.* Reversible, Specific, Active Aggregates of Endogenous Proteins Assemble upon Heat Stress. *Cell* **162,** 1286–1298 (2015).
18. Santra, M., Farrell, D. W. & Dill, K. A. Bacterial proteostasis balances energy and chaperone utilization efficiently. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E2654–E2661 (2017).
19. Geiler-Samerotte, K. A. *et al.* Misfolded proteins impose a dosage-dependent fitness

cost and trigger a cytosolic unfolded protein response in yeast. *Proc. Natl. Acad. Sci.* **108,** 680–685 (2011).
20. Tsirigotaki, A., De Geyter, J., Šoštaric´, N., Economou, A. & Karamanou, S. Protein export through the bacterial Sec pathway. *Nat. Rev. Microbiol.* **15,** 21–36 (2016).
21. Caufield, J. H., Abreu, M., Wimble, C. & Uetz, P. Protein Complexes in Bacteria. *PLoS Comput. Biol.* **11,** 1–23 (2015).
22. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. (2014). doi:10.1016/j.cell.2014.02.033
23. Guharoy, M. & Chakrabarti, P. Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *BMC Bioinformatics* **11,** 286 (2010).
24. Reiner, W. & Veitia, A. Exploring the Molecular Etiology of Dominant-Negative Mutations. doi:10.1105/tpc.107.055053
25. Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N. & Levy, E. D. Proteins evolve on the edge of supramolecular self-assembly. *Nature* **548,** 244 (2017).
26. Bürger, R., Willensdorfer, M. & Nowak, M. A. Why Are Phenotypic Mutation Rates Much Higher Than Genotypic Mutation Rates? doi:10.1534/genetics.105.046599
27. Creecy, J. P. & Conway, T. Quantitative bacterial transcriptomics with RNA-seq. *Curr. Opin. Microbiol.* **23,** 133–40 (2015).
28. Traverse, C. C. & Ochman, H. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc. Natl. Acad. Sci.* **113,** E4257–E4258 (2016).
29. Acevedo, A. & Andino, R. Library preparation for highly accurate population sequencing of RNA viruses. *Nat. Protoc.* **9,** 1760–1769 (2014).
30. Traverse, C. C. & Ochman, H. Genome-Wide Spectra of Transcription Insertions and Deletions Reveal That Slippage Depends on RNA:DNA Hybrid Complementarity. doi:10.1128/mBio.01230-17
31. Mellenius, H. & Ehrenberg, A. Transcriptional accuracy modeling suggests two-step proofreading by RNA polymerase. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx849
32. Gamba, P. & Zenkin, N. Transcription fidelity and its roles in the cell. *Curr. Opin. Microbiol.* **42,** 13–18 (2018).
33. Gordon, A. J. E., Satory, D., Halliday, J. A. & Herman, C. Heritable Change Caused by Transient Transcription Errors. *PLoS Genet.* **9,** e1003595 (2013).
34. Sekiguchi, M. & Tsuzuki, T. Oxidative nucleotide damage: consequences and prevention. *Oncogene* **21,** 8895–8904 (2002).
35. Sie, C. P. & Maas, S. Conserved recoding RNA editing of vertebrate C1q-related factor C1QL1. *FEBS Lett.* **583,** 1171–1174 (2009).
36. Craigen, W. J. & Caskey, C. T. Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature* **322,** 273–275 (1986).
37. Flower, A. M. & McHenry, C. S. The gamma subunit of DNA polymerase III holoenzyme of Escherichia coli is produced by ribosomal frameshifting. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 3713–7 (1990).
38. Meydan, S. *et al.* Programmed Ribosomal Frameshifting Generates a Copper Transporter and a Copper Chaperone from the Same Gene. *Mol. Cell* **65,** 207–219 (2017).
39. Meyerovich, M., Mamou, G. & Ben-Yehuda, S. Visualizing high error levels during gene expression in living bacterial cells. *Proc. Natl. Acad. Sci.* **107,** 11543–11548 (2010).

40. Baudin-Baillieu, A. *et al.* Genome-wide Translational Changes Induced by the Prion [PSI+]. *CellReports* **8,** 439–448 (2014).
41. Richards, J., Sundermeier, T., Svetlanov, A. & Karzai, A. W. Quality control of bacterial mRNA decoding and decay. *Biochim. Biophys. Acta* **1779,** 574–82 (2008).
42. Walczak, R., Westhof, E., Carbon, P. & Krol, A. A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *RNA* **2,** 367–79 (1996).
43. Baggett, N. E., Zhang, Y. & Gross, C. A. Global analysis of translation termination in E. coli. *PLOS Genet.* **13,** e1006676 (2017).
44. Fan, Y. *et al.* Heterogeneity of Stop Codon Readthrough in Single Bacterial Cells and Implications for Population Fitness. *Mol. Cell* **67,** 826–836.e5 (2017).
45. Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R. & Weissman, J. S. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *Elife* **2,** e01179 (2013).
46. Yanagida, H. *et al.* The Evolutionary Potential of Phenotypic Mutations. (2015). doi:10.1371/journal.pgen.1005445
47. Gilchrist, M. A., Shah, P. & Zaretzki, R. Measuring and Detecting Molecular Adaptation in Codon Usage Against Nonsense Errors During Protein Translation. doi:10.1534/genetics.109.108209
48. Sin, C., Chiarugi, D. & Valleriani, A. Quantitative assessment of ribosome drop-off in E. coli. *Nucleic Acids Res.* **44,** 2528–37 (2016).
49. Zaher, H. S. & Green, R. A Primary Role for Release Factor 3 in Quality Control during Translation Elongation in Escherichia coli. *Cell* **147,** 396–408 (2011).
50. Subramaniam, A. R., Zid, B. M. & O'Shea, E. K. An integrated approach reveals regulatory controls on bacterial translation elongation. *Cell* **159,** 1200–1211 (2014).
51. Elf, J. Selective Charging of tRNA Isoacceptors Explains Patterns of Codon Usage. *Science (80-.).* **300,** 1718–1722 (2003).
52. Subramaniam, A. R., Pan, T. & Cluzel, P. Environmental perturbations lift the degeneracy of the genetic code to regulate protein levels in bacteria. *Proc. Natl. Acad. Sci.* **110,** 2419–2424 (2013).
53. Edelmann, P. & Gallant, J. Mistranslation in E. coli. *Cell* **10,** 131–7 (1977).
54. Toth, M. J., Murgola, E. J. & Schimmel, P. Evidence for a unique first position codon-anticodon mismatch in vivo. *J. Mol. Biol.* **201,** 451–4 (1988).
55. Loftfield, R. B. & Vanderjagt, D. The frequency of errors in protein biosynthesis. *Biochem. J.* **128,** 1353–6 (1972).
56. Parker, J., Johnston, T. C. & Borgia, P. T. Mistranslation in cells infected with the bacteriophage MS2: direct evidence of Lys for Asn substitution. *Mol. Gen. Genet.* **180,** 275–81 (1980).
57. Khazaie, K., Buchanan, J. H. & Rosenberger, R. F. The accuracy of Q beta RNA translation. 1. Errors during the synthesis of Q beta proteins by intact Escherichia coli cells. *Eur. J. Biochem.* **144,** 485–9 (1984).
58. Kramer, E. B. & Farabaugh, P. J. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA* **13,** 87–96 (2006).
59. Zhang, Z., Shah, B. & Bondarenko, P. V. G/U and Certain Wobble Position Mismatches as Possible Main Causes of Amino Acid Misincorporations. *Biochemistry* **52,** 8165–8176 (2013).
60. Zaher, H. S. & Green, R. Fidelity at the molecular level: lessons from protein synthesis.

Cell **136,** 746–62 (2009).
61. Miranda, I. *et al.* Candida albicans CUG mistranslation is a mechanism to create cell surface variation. *MBio* **4,** e00285-13 (2013).
62. Jones, T. E., Alexander, R. W. & Pan, T. Misacylation of specific nonmethionyl tRNAs by a bacterial methionyl-tRNA synthetase. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 6933–8 (2011).
63. Wiltrout, E., Goodenbour, J. M., Fré Chin, M. & Pan, T. Misacylation of tRNA with methionine in Saccharomyces cerevisiae. doi:10.1093/nar/gks805
64. Netzer, N. *et al.* Innate immune and chemically triggered oxidative stress modifies translational fidelity. *Nature* **462,** 522–526 (2009).
65. Ling, J. & Söll, D. Severe oxidative stress induces protein mistranslation through impairment of an aminoacyl-tRNA synthetase editing site. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 4028–33 (2010).
66. Pan, T. Adaptive Translation as a Mechanism of Stress Response and Adaptation. *Annu. Rev. Genet.* **47,** 121–137 (2013).
67. Hopfield, J. J. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. U. S. A.* **71,** 4135–9 (1974).
68. Ninio, J. Kinetic amplification of enzyme discrimination. *Biochimie* **57,** 587–595 (1975).
69. Wohlgemuth, I., Pohl, C. & Rodnina, M. V. Optimization of speed and accuracy of decoding in translation. *EMBO J.* **29,** 3701–3709 (2010).
70. Johansson, M., Zhang, J. & Ehrenberg, M. Genetic code translation displays a linear trade-off between efficiency and accuracy of tRNA selection. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 131–6 (2012).
71. Hussain, T., Kamarthapu, V., Kruparani, S. P., Deshmukh, M. V & Sankaranarayanan, R. Mechanistic insights into cognate substrate discrimination during proofreading in translation. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 22117–21 (2010).
72. Moras, D. Proofreading in translation: Dynamics of the double-sieve model. doi:10.1073/pnas.1016083107
73. LaRiviere, F. J., Wolfson, A. D. & Uhlenbeck, O. C. Uniform binding of aminoacyl-tRNAs to elongation factor Tu by thermodynamic compensation. *Science* **294,** 165–8 (2001).
74. Freeland, S. J. & Hurst, L. D. The Genetic Code Is One in a Million.
75. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the genetic code: The universal enigma. *IUBMB Life* **61,** 99–111 (2009).
76. Caetano-Anollé, G., Wang, M., Caetano-Anollé, D. & Maga, G. Structural Phylogenomics Retrodicts the Origin of the Genetic Code and Uncovers the Evolutionary Impact of Protein Flexibility. *PLoS One* **8,** (2013).
77. Shah, P. & Gilchrist, M. A. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 10231–6 (2011).
78. Wallace, E. W. J., Airoldi, E. M., Drummond, D. A. & Mcdonald, J. H. Estimating Selection on Synonymous Codon Usage from Noisy Experimental Data. doi:10.1093/molbev/mst051
79. Drummond, D. A. & Wilke, C. O. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* **134,** 341–352 (2008).
80. Yang, J.-R., Chen, X. & Zhang, J. Codon-by-Codon Modulation of Translational Speed and Accuracy Via mRNA Folding. *PLoS Biol.* **12,** e1001910 (2014).
81. Naville, M., Gautheret, D., Naville, M. & Gautheret, D. Transcription attenuation in

bacteria: theme and variations. *Brief. Funct. Genomics* **9,** 178–189 (2010).
82. Scott, M., Klumpp, S., Mateescu, E. M. & Hwa, T. Emergence of robust growth laws from optimal regulation of ribosome synthesis. *Mol. Syst. Biol.* **10,** 747 (2014).
83. Sørensen, M. a. Charging levels of four tRNA species in Escherichia coli Rel(+) and Rel(-) strains during amino acid starvation: a simple model for the effect of ppGpp on translational accuracy. *J. Mol. Biol.* **307,** 785–98 (2001).
84. Yona, A. H. *et al.* tRNA genes rapidly change in evolution to meet novel translational demands. *Elife* **2,** (2013).
85. Rogers, H. H. & Griffiths-Jones, S. tRNA anticodon shifts in eukaryotic genomes. *RNA* **20,** 269–81 (2014).
86. Whitehead, D. J., Wilke, C. O., Vernazobres, D. & Bornberg-Bauer, E. The look-ahead effect of phenotypic mutations. doi:10.1186/1745-6150-3-18
87. Bratulic, S., Gerber, F. & Wagner, A. Mistranslation drives the evolution of robustness in TEM-1 β-lactamase. *Proc. Natl. Acad. Sci.* **112,** 12758–12763 (2015).
88. Zhu, Y. O., Siegal, M. L., Hall, D. W. & Petrov, D. A. Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E2310-8 (2014).
89. Lee, H., Popodi, E., Tang, H. & Foster, P. L. 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,** E2774-83 (2012).
90. Wong, F., Amir, A. & Gunawardena, J. An energy-speed-accuracy relation in complex networks for biological discrimination.
91. Banerjee, K., Kolomeisky, A. B. & Igoshin, O. A. Elucidating interplay of speed and accuracy in biological error correction. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 5183–5188 (2017).
92. Schimmel, P. Mistranslation and its control by tRNA synthetases. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **366,** 2965–71 (2011).
93. Allan Drummond, D. & Wilke, C. O. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.* **10,** 715–724 (2009).
94. Mohler, K. & Ibba, M. Translational fidelity and mistranslation in the cellular response to stress. *Nat. Microbiol.* **2,** 17117 (2017).
95. Bratulic, S., Toll-Riera, M., Wagner, A., Marx, C. J. & Tawfik, D. S. Mistranslation can enhance fitness through purging of deleterious mutations. *Nat. Commun.* **8,** 15410 (2017).
96. Kramer, E. B., Vallabhaneni, H., Mayer, L. M. & Farabaugh, P. J. A comprehensive analysis of translational missense errors in the yeast Saccharomyces cerevisiae. *RNA* **16,** 1797–808 (2010).
97. Cvetesic, N. *et al.* Proteome-wide measurement of non-canonical bacterial mistranslation by quantitative mass spectrometry of protein modifications. *Sci. Rep.* **6,** 28631 (2016).
98. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372 (2008).
99. Rozov, A., Westhof, E., Yusupov, M. & Yusupova, G. The ribosome prohibits the G•U wobble geometry at the first position of the codon–anticodon helix. *Nucleic Acids Res.* **44,** gkw431 (2016).
100. Rozov, A., Demeshkina, N., Westhof, E., Yusupov, M. & Yusupova, G. New Structural Insights into Translational Miscoding. *Trends Biochem. Sci.* **41,** 798–814 (2016).

101. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355 (2016).
102. Tsur, D., Tanner, S., Zandi, E., Bafna, V. & Pevzner, P. A. Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23,** 1562–1567 (2005).
103. Na, S., Bandeira, N. & Paek, E. Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell. Proteomics* **11,** M111.010199 (2012).
104. Savitski, M. M., Nielsen, M. L. & Zubarev, R. A. 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 (2006).
105. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11,** 319–324 (2014).
106. Moruz, L. & Käll, L. Peptide retention time prediction. *Mass Spectrom. Rev.* **36,** 615–623 (2017).
107. Sun, L. *et al.* Evolutionary Gain of Alanine Mischarging to Noncognate tRNAs with a G4:U69 Base Pair. *J. Am. Chem. Soc.* **138,** 12948–12955 (2016).
108. Gromadski, K. B. & Rodnina, M. V. Streptomycin interferes with conformational coupling between codon recognition and GTPase activation on the ribosome. *Nat. Struct. Mol. Biol.* **11,** 316–322 (2004).
109. Banerjee, K., Kolomeisky, A. B. & Igoshin, O. A. Elucidating interplay of speed and accuracy in biological error correction. *Proc. Natl. Acad. Sci.* **114,** 5183–5188 (2017).
110. Zhu, M., Dai, X. & Wang, Y.-P. Real time determination of bacterial in vivo ribosome translation elongation speed based on LacZ␣ complementation system. *Nucleic Acids Res.* **44,** (2016).
111. Ingolia, N. T. Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* **165,** 22–33 (2016).
112. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18 Suppl 1,** S71-7 (2002).
113. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33,** W382–W388 (2005).
114. LaRiviere Frederick, Wolfson Alexey D., U. O. C. Uniform Binding of Aminoacyl-tRNAs to Elongation Factor Tu by Thermodynamic Compensation. *Science (80-.).* **294,** 165–168 (2001).
115. Khan, Z. *et al.* Accurate proteome-wide protein quantification from high-resolution 15N mass spectra. *Genome Biol.* **12,** R122 (2011).
116. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6,** 359–362 (2009).
117. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and Go Extraction Tips for Matrix-Assisted Laser Desorption/Ionization, Nanoelectrospray, and LC/MS Sample Pretreatment in Proteomics. *Anal. Chem.* **75,** 663–670 (2003).
118. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4,** 207–214 (2007).
119. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome

coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43,** D261–D269 (2015).

120. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–7 (2004).
121. Woolstenhulme, C. J., Guydosh, N. R., Green, R. & Buskirk, A. R. High-Precision Analysis of Translational Pausing by Ribosome Profiling in Bacteria Lacking EFP. *Cell Rep.* **11,** 13–21 (2015).