The roles of stop codons and 3' flanking base in bacterial translation termination efficiency

Yulong Wei

Supervisor: Dr. Xuhua Xia

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
In partial fulfilment of the requirement for a
Master of Science degree in Biology, specialization: Bioinformatics
From the Ottawa-Carleton Institute of Biology

Thèse soumise à la
Faculté des Etudes Supérieures et Postdoctorales
Université d'Ottawa
En vue de l'obtention de la maîtrise
L'Institut de Biologie d'Ottawa-Carleton

Abstract

Understanding translation efficiency is crucial to pharmaceutical companies that have invested substantial time and effort in engineering bacteria to produce recombinant proteins. While translation initiation and elongation have been studied intensively, much remains obscure in the subprocess of translation termination. We aim to understand how stop codons and the first 3' flanking (+4) base affect translation termination efficiency.

In chapter two, we hypothesized that stop codon usage of UAG and UGA is dependent on the abundance of their respective decoders, RF1 and RF2. We predicted and observed that bacterial species with high relative proportions of RF1 uses UAG more, and vice versa for UGA. In addition, the usage of UGA, not UAG, is always avoided in highly expressed genes. Thus, we argued against the claim made by a recent study that UAG is a minor stop codon in bacteria. The claim is incorrect because UAG does not meet the two criteria of a minor codon: i) it is most avoided in highly expressed genes, and ii) it corresponds to the least abundant decoder. Interestingly, we found that the proportion of RF2 decreases rapidly towards zero in species with high AT contents; this explains why UGA is reassigned to a sense codon in bacterial lineages with high AT content.

In chapter three, we examined the role of the first downstream (+4) base Uracil in bacterial translation termination. The +4U is associated with a decrease in stop codon read-through in bacteria and yeast. We hypothesized that i) +4U enhances the termination efficiency of stop signals, and ii) +4U may serve to prevent stop codon misreading by near cognate tRNAs (nc_tRNAs). We predicted that i) +4U is preferred in highly expressed genes (HEGs) than lowly expressed genes (LEGs), and ii) +4U usage increases with the frequency of stop codon nc_tRNAs. We found +4U consistently over-represented in HEGs in contrast to LEGs; however, +4U usage in HEGs decreases in GC-rich species where most stop codons are UGA and UAG. In addition, +4U usage increases significantly with UAA usage in the known highly expressed ribosomal protein genes. These results suggest that +4U is a strong stop signal enhancer for UAA, not UAG or UGA. Furthermore, in HEGs, +4U usage also increases significantly with the abundance of UAA nc_tRNAs, suggesting that +4U increases UAA termination efficiency presumably by reducing misreading of UAA by nc_tRNAs.

Résumé

Comprendre l'efficacité traductionnelle est une étape cruciale pour les compagnies pharmaceutiques qui ont fait d'énormes investissements de temps et d'effort dans le génie génétique bactérien pour produire des protéines recombinantes. Même si beaucoup d'études ont été faites sur l'initiation et l'élongation de la traduction génétique, le sous-processus de la terminaison de la traduction demeure peu connu. Cette étude vise à comprendre comment le codon de terminaison et le 3' site adjacent (+4) affectent l'efficacité traductionnelle.

Dans le chapitre 2, j'argumente contre l'étude récente qui déclare que l'UAG est un codon de terminaison mineur dans la bactérie. Cette déclaration est invalide parce qu'elle ne concorde pas avec les deux caractéristiques du codon mineur : i) il est le codon le plus évité parmi les gènes fortement exprimés et ii) il correspond au décodeur le moins abondant. Nous avons comme hypothèse que l'utilisation du codon de terminaison UAG et UGA est dépendante de l'abondance de leurs décodeurs respectifs, RF1 et RF2. Nous observons que les espèces avec une proportion relativement élevée de RF1 vont plus utiliser l'UAG, et vice versa pour UGA. La proportion de RF2 diminue rapidement vers zéro en présence du contenu élevé d'AT ; ceci explique les espèces avec du contenu d'AT élevé ont souvent l'UGA réaffecté à un codon de détection. En plus, l'utilisation d'UGA et non d'UAG est toujours évitée dans les gènes fortement exprimés.

Dans le chapitre 3, nous expliquons pourquoi la traduction des codons de terminaison est réduite en présence de 3' site adjacent (+4), comme observée par les études antérieures. Nous avons comme hypothèse que la présence de +4U réduit la reconnaissance du codon de terminaison par des ARNt correspondants proches, ainsi réduisant la traduction des codons de terminaison. Dans 19 espèces bactériennes, nous trouvons que l'utilisation du +4U augmente avec l'abondance de UAA nc-ARNt, et cette corrélation est particulièrement significative avec les nc-ARNt qui présentent un décalage au premier site du codon. Cette distinction soutient la conjecture faite par des études antérieures qui propose que l'efficacité de reconnaissance du site du premier codon de terminaison par le RF2 est dépendante du +4U. En plus, nous démontrons que le +4U augmente l'efficacité de la terminaison. L'utilisation de +4U dépend de l'utilisation du codon de terminaison UAA et augmente avec l'$I_{TE,}$ notre mandataire de l'expression génétique, mais diminue avec du contenu élevé en GC.

Acknowledgement

I express my sincere respect and gratitude to my supervisor Dr. Xuhua Xia for his support, cooperation and inspiration. This thesis could not have been completed without his assistance and guidance. In addition, I am grateful to Dr. Xuhua Xia for providing me the opportunity to work in his lab as his master's student.

I thank my advisory committee members, Dr. Stéphane Aris-Brosou and Dr. David Sankoff for their inputs and guidance in the development of the thesis. I am grateful to Dr. Stéphane Aris-Brosou for his teachings. In addition, I would like to thank Dr. Stéphane Aris-Brosou, Dr. Douglas Johnson and Dr. Alex Wong for accepting to be my thesis examiners.

I thank my past and present lab mates, Juan Wang, Joran Silke, Caitlyn Vlasschaert and Ramanandan Prabhakaran for their friendship, insights and encouragement. I thank my parents for their love, understanding and support, and I acknowledge and thank the University of Ottawa and NSERC for their generous funding.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A	Adenosine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
AT/CG	Adenosine and Tymine/ Cytosine and Guanine
N	A, C, G or U
CDS	Coding DNA sequence
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger RNA
ssu rRNA	Small subunit ribosomal RNA
tRNA	Transfer RNA
Nc_tRNA	Near cognate tRNA
RF1/RF2	Class I release factor one/ two
RF3	Class I release factor three
GTP	Guanosine triphosphate
GTPase	GTP hydrolase
30S/50S	Small/ large ribosomal subunit
CAI	Codon adaptation index
ITE	Index of translation elongation
RSCU	Relative synonymous codon usage
APE	Analysis of Phylogenetics and Evolution
PHYLIP	Phylogeny Inference Package
PaxDB	Protein abundance Database
DAMBE	Data Analysis in Molecular Biology and Evolution
MAFFT	Multiple Alignment using Fast Fourier Transform
PhyML	Phylogeny reconstruction based on maximum-likelihood principle
UniProtKB	Uniprot Knowledgebase
FastME	Fast phylogeny reconstruction based on minimum-evolution principle
GTR	Generalised time reversible model
TN93	Tamura and Nei 1993 model

Amino acids:

Arg	Arginine
Cys	Cysteine
Gln	Glutamine
Glu	Glutamic acid
Leu	Leucine
Lys	Lysine
Ser	Methionine
Trp	Tryptophan
Tyr	Tyrosine

1. Chapter one: Introduction

1.1 Translation termination

Translation termination involves the recognition of stop codons by release factors. The process begins when one of the stop codons UAA, UAG or UGA enters the 70S ribosomal A site during translation. The class I release factors RF1 and RF2 are encoded by the *prfA* and *prfB* genes, respectively (Scolnick *et al.* 1968; Milman *et al.* 1969; Scolnick and Caskey 1969). RF1 recognizes stop codons UAA and UAG, and RF2 recognizes stop codons UAA and UGA. The precise mechanism of stop codon recognition by class I release factors is unclear, but involves a recognition motif containing amino acid triplets SPF in RF2 and PNT in RF1 (Ito *et al.* 2000). In addition, RF1 and RF2 share a common GGQ motif to facilitate the hydrolysis of the ester bond between the peptide chain and tRNA at the P site (Scarlett *et al.* 2003). Class II release factor RF3 is not directly involved in stop codon recognition, but it associates with the 50S ribosomal subunit and with class I release factors. RF3 is a GTPase and its GTP hydrolysis activity facilitates the dissociation of RF1 and RF2 from the ribosome (Freistroffer *et al.* 1997); while RF3 itself is rapidly dissociated in its GDP conformation from the 70S ribosome (Zavialov *et al.* 2002).

1.1.1 Termination efficiency

Termination read-through occurs when stop codons are misread as sense codons by near cognate tRNAs (nc_tRNAs). In bacteria and eukaryotes, the level of termination read-through varies among stop codons, but UGA is consistently the least efficient stop codon (Parker 1989; Jorgensen *et al.* 1993; Tate *et al.* 1999; Dabrowski *et al.* 2015). In eukaryotes, termination read-through is affected by the penultimate (-2) and ultimate (-1) nucleotide positions preceding the

stop codon (Mottagui-Tabar *et al.* 1998), and -1A significantly decreases termination efficiency (Cassan and Rousset 2001; Loughran *et al.* 2014). Despite this, the first 3' flanking (+4) base is considered more influential with respect to the efficacy of termination in both bacteria and eukaryotes. (Bossi and Ruth 1980; Miller and Albertini 1983; Tate *et al.* 1995; Poole *et al.* 1998).

In addition to read-through, translation frameshift can also occur at the termination site. A well-known example is the auto-regulatory translation of the *prfB* gene by programmed frameshift at a premature UGA (Craigen *et al.* 1985; Craigen and Caskey 1986; Baranov *et al.* 2002), discussed in more detail in Chapter 3.

1.2 Codon usage and translation efficiency

1.2.1 Mutation bias

GC mutation bias (Muto and Osawa 1987) affects the nucleotide composition of a genome. A species with high GC content prefers synonymous codons containing C and G over those containing A and T. For example, comparatively, one expects to observe more AAA than AAG lysine codons in species with low GC contents. The genetic code is degenerate at the third codon site; thus, GC mutation bias can often be measured at this site (Palidwor *et al.* 2010) since it is evolutionarily less constrained than the first and second codon sites.

We can also observe strand asymmetry in nucleotide compositions in a wide spectrum of bacterial species, where the leading strand is AG-rich and the lagging strand is TC-rich (Rudner *et al.* 1968; Lobry and Sueoka 2002). AT skew is often more subtle than GC skew in bacteria, but it is still significant in Firmicutes species such as *Bacillus subtilis*, wherein the leading strand is A-rich and the lagging strand is T-rich (Charneski *et al.* 2011; Xia 2012).

1.2.2 Selection bias

The selection bias in codon usage reflects the need for major and minor codons which dictate the translation elongation efficiency of the species' genes. A codon is often under selection bias when its decoder tRNA is abundant, this observation is termed the tRNA-mediated selection bias (Akashi 1994; Xia 1998), and is first observed in *E. coli* by Ikemura (1981), who proposed the codon-anticodon adaptation hypothesis. In conjunction with this criterion, codons with abundant decoders are termed major codons (Mcpherson 1988) if they are also more favourably selected in highly expressed genes (HEGs) than lowly expressed genes (LEGs). In contrast, minor codons are avoided in HEGs and decoded by less abundant tRNAs. HEGs are efficiently translated, and examples of such genes often include those which encode 30S and 50S ribosomal proteins (Sharp and Li 1987; Guerdoux-Jamet *et al.* 1997).

1.2.3 Translation elongation efficiency

Recent studies have recognized that the influence of translation elongation efficiency in gene expression is dependent on translation initiation efficiency (Xia 2007a; Supek and Smuc 2010; Tuller *et al.* 2010; Prabhakaran *et al.* 2015; Xia 2015a). For example, protein production of an mRNA with low initiation efficiency does not benefit extensively from an efficient elongation, as translation will be rate-limited at the initiation phase. On the other hand, the rate of protein production of an mRNA with high initiation efficiency will be dependent on elongation efficiency. This concept can be further extended to explain the role of termination efficiency in gene expression, discussed in more detail in Chapter 3.

To measure translation elongation efficiency, codon adaptation indices such as CAI and I_{TE} (Sharp and Li 1987; Xia 2015a) can be used to assess the codon usage of an mRNA in reference

to that observed in known highly expressed genes (HEGs). In this manner, elongation efficiency can be inferred without the requirements of extensive wet lab data encompassing abundance and degradation values at the level of both RNA and protein. The index of translation elongation, I_{TE} (Xia 2015a), is a recently formulated codon adaptation index that uses two reference sets: one for highly expressed genes (HEGs), and one for non-HEGs. Having both reference sets allows one to appropriately identify codon usage bias due to selection and mutation. For example, in *Escherichia coli* HEGs compiled by Rice *et al.* (2000), the frequency of GCA and GCG are 1973 and 2654, respectively. This leads one to think that the major codon in this family is GCG; however, in non-HEGs, GCA and GCG frequencies are 25511 and 43261, respectively (Xia 2015a). Thus, GCA is relatively more frequent in HEGs than non-HEGs, suggesting that GC-mutation bias favours GCG, but selection bias favours GCA.

Other measurements of translation efficiency include ribosome profiling and protein abundance. Ribosome profiling (Ingolia *et al.* 2009; Ingolia 2014) produces a global snapshot of ribosome activity in the cell at any given moment. This allows one to observe mRNA translational activities through location-specific ribosome density. Protein abundance data (Wang *et al.* 2015) provides us with the steady-state levels of protein products at the end of translation. However, protein abundance data may not accurately reflect translation efficiency since they often do not account for mRNA abundance and protein degradation rates.

1.3 Significance of the study

Optimizing translation efficiency is crucial to pharmaceutical companies that have invested substantial time and effort in engineering bacteria to produce recombinant proteins. To increase translation initiation efficiency, an expression vector contains a set of prokaryotic sequence

elements in the 5' untranslated region (5' UTR) such as a strong promoter and Shine-Dalgarno sequence (Hannig and Makrides 1998). Additionally, changes in the recombinant coding sequence are necessary in order for the trans-gene to use major codons of the host bacterium for efficient translation elongation.

What affects the stop codon termination efficiency in bacteria, however, is still not fully understood. Studies suggest that UGA has the highest termination read-through in bacteria (Parker 1989; Jorgensen *et al.* 1993; Tate *et al.* 1999; Dabrowski *et al.* 2015), but Povolotskaya *et al.* (2012) and Korkmaz *et al.* (2014) studied stop codon usages in a large number of bacterial species and concluded that UAG is a universal minor stop codon. Furthermore, the presence of +4U decreases stop codon read-through in bacteria; however, the molecular mechanism underlying this site is poorly understood. Studies examining for possible interactions between RF2 and termination signals using crosslinking experiments (Brown and Tate 1994; Tate *et al.* 1996; Poole *et al.* 1997; Poole *et al.* 1998) speculated that +4U increases RF2's decoding efficiency. However, the influence of +4U in translation termination may be alternatively explained by the its avoidance of near-cognate tRNA (nc_tRNA) recognition, based on the results of a recent study in yeast termination read-through (Beznoskova *et al.* 2016b). These possibilities motivate us to study the role of +4U in translation termination.

In chapter two, we hypothesized that stop codon usage of UAG and UGA is dependent on the abundance of their respective decoders, RF1 and RF2. We predicted that species with high abundance of RF1 will use UAG more frequently, while species with high abundance of RF2 will use UGA more frequently. The availability of protein abundance data for a number of bacterial species in PaxDB 4.0 (Wang *et al.* 2015) allowed us to measure the relative abundance of RF1 and RF2 and permits the characterization of HEGs and LEGs. In 14 bacterial species,

usage of UGA relative to UAG increased significantly with relative proportions of RF2. The proportion of RF2 was higher than RF1 over a wide range of AT content, but decreased rapidly towards zero at high AT contents. In addition, usage of UGA, not UAG, was always higher in LEGs over HEGs. Our findings explain that bacterial lineages with high AT-content often reassign UGA as a sense codon because RF2 abundance is significantly reduced at high AT-contents. Neither UAG nor UGA is a universal minor stop codon, because stop codon usage is affected by codon-decoder adaptation and mutation bias.

In chapter three, we examined i) whether +4U enhances the stop signal relative to other nucleotides and ii) if +4U may serve to prevent misreading of stop codons by nc_tRNAs. To test for these hypotheses, we predicted that i) +4U is preferred in HEGs than LEGs, and ii) +4U usage should increase with the frequency of stop codon nc_tRNAs. In 25 bacterial species whose protein abundance data are present in PaxDB 4.0 (Wang *et al.* 2015), +4U was consistently over-represented in HEGs in contrast to LEGs; however, +4U usage in HEGs decreased in GC-rich species where most stop codons are UGA and UAG. In addition, +4U usage increased significantly with UAA usage in the known highly expressed ribosomal protein genes. These results suggest that UGA and UAG do not need +4U as a stop signal enhancer as much as UAA. Furthermore, in HEGs, +4U usage also increased significantly with the abundance of UAA nc_tRNAs, suggesting that +4U increases UAA termination efficiency presumably by reducing misreading of UAA by nc_tRNAs.

2. Chapter two

Coevolution between stop codon usage and release factors in bacterial species

2.1 Abstract

Three stop codons in bacteria represent different translation termination signals, and their usage is expected to depend on their differences in translation termination efficiency, mutation bias, and relative abundance of release factors (RF1 decoding UAA and UAG, and RF2 decoding UAA and UGA). In 14 bacterial species (covering Proteobacteria, Firmicutes, Cyanobacteria, Actinobacteria and Spirochetes) with cellular RF1 and RF2 quantified, UAA is consistently over-represented in highly expressed genes (HEGs) relative to lowly expressed genes (LEGs), whereas UGA usage is the opposite even in species where RF2 is far more abundant than RF1. UGA usage relative to UAG increases significantly with PRF2 [=RF2/(RF1+RF2)] as expected from adaptation between stop codons and their decoders. PRF2 is greater than 0.5 over a wide range of AT content (measured by PAT3 as the proportion of AT at third codon sites), but decreases rapidly towards zero at the high range of PAT3. This can at least partially explains why bacterial lineages with high PAT3 often have UGA reassigned because of low RF2. There is no indication that UAG is a minor stop codon in bacteria as claimed in a recent publication. The claim is invalid because of the failure to apply the two key criteria in identifying a minor codon: i) it is least preferred by HEGs and ii) it corresponds to the least abundant decoder. Our results suggest a more plausible explanation for why UAA usage increases, and UGA usage decreases, with PAT3, but UAG usage remains low over the entire PAT3 range.

2.2 Contribution

The data, results and interpretations in this chapter were published in *Molecular Biology and Evolution* (Wei Y, Wang J, Xia X. 2016. Coevolution between stop codon usage and release factors in bacterial species. *Mol Biol Evol.* 33: 2357-2367). The development of the hypotheses, data analyses and interpretations are contributed among Yulong Wei, Juan Wang and Dr. Xuhua Xia.

2.3 Introduction

Most bacterial lineages share genetic code 11 with three stop codons, UAA, UAG, and UGA, which are decoded by two release factors (RF1 and RF2), with RF1 decoding UAA and UAG and RF2 decoding UAA and UGA (Milman *et al.* 1969). In *Escherichia coli,* RF2 is consistently more abundant than RF1, which is associated with UGA used much more frequently than UAG. This association between the frequency of stop codon and its decoder concentration is consistent with codon-anticodon adaptation documented in bacteria (Ikemura 1981; Gouy and Gautier 1982; Xia 1998; Stoletzki and Eyre-Walker 2007; Palidwor *et al.* 2010), eukaryotes (Chavancy *et al.* 1979) such as yeast (Sharp and Li 1986; Xia 1998; Akashi 2003) and fruit flies (Moriyama and Hartl 1993; Akashi 1994; Moriyama and Powell 1997), viruses (Van Weringh *et al.* 2011; Chithambaram *et al.* 2014c; Chithambaram *et al.* 2014a; Prabhakaran *et al.* 2014), and mitochondria (Xia 2005; Xia 2007a; Carullo and Xia 2008; Jia and Higgs 2008; Xia 2008). Because different stop codons may manifest as different signals to the cellular translation termination machinery, both experimental and bioinformatics approaches have been taken to characterize translation termination efficiency in association with their decoders. The experimental studies on translation termination have focused mainly on *E. coli* (and occasionally

on the yeast, *Saccharomyces cerevisiae*) and addressed two questions: i) which tRNA species tend to misread a stop codon as a near-cognate sense codon, and ii) which release factor tends to misread near-cognate sense codons as stop codons.

All three stop codons can be misread by tRNAs, and UGA appears to be the leakiest of the three, with a read-through frequency of at least 10^{-2} to 10^{-3} in *Salmonella typhimurium (Roth 1970)* and *E. coli* (Sambrook *et al.* 1967; Strigini and Brickman 1973). UAA and UAG can also be leaky in bacteria (Davies *et al.* 1966; Ryden and Isaksson 1984), although their misreading has not been reported as frequently as UGA. Natural UAG read-through frequency is mostly within the range of 1.1×10^{-4} to 7×10^{-3}, depending on the nature of the downstream nucleotides (Bossi and Ruth 1980; Bossi 1983; Miller and Albertini 1983; Ryden and Isaksson 1984). The read-through of UAA seems to occur at frequencies from 9×10^{-4} to $< 1 \times 10^{-5}$ (Ryden and Isaksson 1984). Overall, the available experimental data suggest that in bacteria species, particularly in *E. coli*, read-through is most frequent for UGA, less for UAG, and least for UAA (Strigini and Brickman 1973; Geller and Rich 1980; Parker 1989; Jorgensen *et al.* 1993; Meng *et al.* 1995; Cesar Sanchez *et al.* 1998; Tate *et al.* 1999).

Translation termination error rate depends not only on read-through by tRNA, but also on the efficiency and relative concentration of RF1 and RF2 (Korkmaz *et al.* 2014). Increasing RF2 concentration decreased both UGA read-through and frameshift (reviewed in (Tate *et al.* 1999). The observation that UAA is the most frequently used stop codon in *E. coli, B. subtilis,* and *S. cerevisiae (Sharp and Bulmer 1988)* was interpreted in light of the fact that UAA has the largest number of decoders (being decoded by both RF1 and RF2) and that it is the most reliable stop signal of the three as reviewed above. Early studies suggest that RF1 and RF2, given the same concentration, decode their respective stop codons with roughly equal efficiency (Scolnick *et al.*

1968; Jorgensen *et al.* 1993; Freistroffer *et al.* 1997; Ito *et al.* 2000), and that both are extremely efficient and accurate against near-cognate codons, except for UGG in the case of RF2 and UAU in the case of RF1 (Freistroffer *et al.* 1997). However, given the same codon context, RF2 decoding UGA is less efficient than RF1 decoding UAG in *E. coli* (Björnsson and Isaksson 1996).

The effect of both mutation and selection (mediated by relative concentration of RF1 and RF2) on stop codon usage has been studied. The selection effect is derived as an extension of the well-known codon-anticodon adaptation (Ikemura 1981; Akashi and Eyre-Walker 1998; Xia 1998; Van Weringh *et al.* 2011; Chithambaram *et al.* 2014c; Prabhakaran *et al.* 2014; Prabhakaran *et al.* 2015). As UGA is decoded only by RF2 and UAG only by RF1, one expects UGA to be used more than UAG when RF2 concentration is higher than RF1 (assuming the two have equal decoding efficiency on their respective codons). This is consistent in *E. coli*, where RF2 is about five times more frequent than RF1 (Adamski *et al.* 1994; Mora *et al.* 2007) and UGA is used much more frequently than UAG (Korkmaz *et al.* 2014).

The mutation effect on stop codon usage is mainly studied through genomic GC content which has a strong effect on stop codon usage based on data from 736 species (Povolotskaya *et al.* 2012). An even more comprehensive compilation involving 4684 genomes (Korkmaz *et al.* 2014) has revealed strong effect of GC content on the frequencies of UAA and UGA, but little on the frequency of UAG. However, the effect of GC content on stop codon usage depends on gene expression (Korkmaz *et al.* 2014).

Furthermore, bioinformatics studies (Sharp and Bulmer 1988; Brown and Tate 1994; Cridge *et al.* 2006; Povolotskaya *et al.* 2012; Korkmaz *et al.* 2014) have generally found UAA to be the most frequent stop codon and UAG the least frequent. In particular, Korkmaz *et al.* (2014)

claimed that "TAG is truly a minor stop codon in all aspects". Designating codons as major and minor codons is important not only in understanding the function of the translation machinery, but also in biopharmaceutical industry as many experimental studies have shown that replacing minor codons by major codons increases protein production (Robinson *et al.* 1984; Sorensen *et al.* 1989; Haas *et al.* 1996; Ngumbela *et al.* 2008). However, the term "major (or minor) codon" is often misunderstood. "Major codon" (or optimal codon) originally refers to sense codons preferred by highly expressed genes and decoded by the most abundant tRNA. It is first used by (Mcpherson 1988) in reference to a study (Kurland 1987) showing that highly expressed genes use codons to optimize decoding efficiency of the tRNA pool. A minor codon is the opposite. Major and minor codons are not necessarily the most frequent or least frequent codons when compilation is done for all genes.

Two criteria, one essential and one corroborative, have been used, sometimes implicitly, to identify a minor sense codon. The essential criterion is that a minor codon is the most strongly avoided in highly expressed genes (HEGs, in contrast to lowly expressed genes or LEGs). The corroborative criterion is that a minor codon corresponds to the least abundant tRNA among synonymous codons. Without these two criteria, a minor codon could be identified incorrectly (Xia 2015a). For example, if we compile the codon frequencies of Asp codon family for all genes in *E. coli* (NC_000913), we will get 41806 GAU and 25015 GAC, which would mislead us to conclude that GAU is the major codon, and GAC the minor. However, if we rank *E. coli* genes by the protein abundance data compiled in the integrated dataset in PaxDB (Wang *et al.* 2015) or by the index of translation elongation, I_{TE} (Xia 2015a), then LEGs (100 genes at the low end of abundant proteins) uses more GAU than GAC, but HEGs (100 genes at the high end of gene expression) uses more GAC than GAU. Furthermore, these Asp codons are translated by

three $tRNA^{Asp}$ genes all with the same GUC anticodon forming perfect base-pair with GAC. Thus, both criteria support GAC as the major (optimal) codon, and GAU as the minor.

Korkmaz *et al.* (2014) made an effort to apply these two criteria in identifying major and minor stop codons in bacteria. They compiled 4684 bacterial genomes and concluded that "in all these phyla, TAG is the minor stop codon", and that "TAG is truly a minor stop codon in all aspects". The conclusion, however, is not correct because of misapplication of the two criteria, which may be best illustrated by taking *Microcystis aeruginosa* for example. LEGs use more UGA than UAG as stop codons in this species (PUGA.LEG=0.2970, PUAG.LEG = 0.2393, Table 2.1), but HEGs use more UAG than UGA (PUAG.HEG=0.2536, PUGA.HEG = 0.1556, Table 2.1). This stop codon usage pattern is consistent with the relative RF1 and RF2 concentrations compiled in the integrated dataset available in PaxDB (Wang *et al.* 2015). Protein abundance is 33.3 ppm (parts per million) for RF1 and 18.2 ppm for RF2 in that integrated dataset. The average concentration of RF1 is also higher than RF2 based on multiple separate measurements (Table 2.1). Thus, UAG has more decoders than UGA and is expected to be more preferred than UGA by HEGs, especially given the experimental evidence (reviewed above) that UAG is a more accurate stop signal than UGA. So UAG clearly is not a minor stop codon in *M. aeruginosa*, contrary to what Korkmaz *et al.* (2014). Korkmaz *et al.* (2014) used ribosomal protein and translation factor genes (which are generally highly expressed) as HEGs in a subset of genomes studied, but they did not contrast between HEGs and LEGs, so one does not know the difference in relative stop codon preference between HEGs and LEGs.

Table 2.1. Bacterial species with both RF1 and RF2 concentrations in PaxDB (Wang *et al.* 2012), together with stop codon usage in highly expressed and lowly expressed genes (HEGs and LEGs).

Species	N_{gene}[1]	RF1	RF2	P_{AT3}[2]	P_{RF2}[3]	$P_{UAA.LEG}$[4]	$P_{UAA.HEG}$	$P_{UAG.LEG}$	$P_{UAG.HEG}$	$P_{UGA.LEG}$	$P_{UGA.HEG}$
E coli	1,000	53.1	453	0.4383	0.8951	0.5730	0.7770	0.1070	0.0320	0.3200	0.1910
Y. pestis	300	11.6	672	0.4979	0.9830	0.6100	0.7433	0.1300	0.0700	0.2600	0.1867
M. tuberculosis	800	200.5	548.5	0.2018	0.7323	0.1525	0.1688	0.2713	0.3538	0.5763	0.4775
S. enterica	600	59.2	142.89	0.4008	0.7070	0.5717	0.7650	0.1083	0.0433	0.3200	0.1917
L. lactis	300	45.5	98.05	0.7247	0.6833	0.7167	0.9100	0.1067	0.0467	0.1767	0.0433
P. aeruginosa	500	56.4	167	0.1262	0.7475	0.0560	0.2640	0.1280	0.0480	0.8160	0.6880
H. pylori	300	157.0	214	0.5777	0.5768	0.6267	0.6600	0.1567	0.1567	0.2167	0.1833
L. interrogans	600	139.3	183	0.6969	0.5677	0.5683	0.6467	0.1317	0.0983	0.3000	0.2550
M. aeruginosa	1,000	35.1	27	0.6059	0.4348	0.4639	0.5908	0.2392	0.2536	0.2970	0.1556
S. pyogenes	301	246.5	74.65	0.6766	0.2324	0.5748	0.7902	0.2292	0.1475	0.1960	0.0623
B. subtilis	1,000	216.0	205	0.5518	0.4869	0.5600	0.7300	0.1530	0.1240	0.2870	0.1460
B. anthracis	300	94.3	4.59	0.7349	0.0464	0.7367	0.8567	0.1367	0.0867	0.1267	0.0567
S. aureus	392	496.0	47.7	0.7702	0.0877	0.7398	0.8475	0.1633	0.1025	0.0969	0.0500
A. ferrooxidans	301	425.5	377	0.3096	0.4698	0.2425	0.3033	0.1362	0.1433	0.6213	0.5533

[1] Number of genes in top and bottom 25% on the gene expression scale (ranked by either protein abundance values in PaxDB). If 25% includes more than 1000 genes, then use 1000.

[2] Proportion of AT at third codon site.

[3] Proportion of RF2, i.e., RF2/(RF1+RF2)

[4] Proportion of UAA stop codons in LEGs. The same format applies to the last five columns.

For relative abundance of RF1 and RF2, Korkmaz *et al.* (2014) only confirmed previous findings that RF2 is several folds more abundant than RF1 in *E. coli*, but did not have RF1 and RF2 abundance data for the rest of the 4684 species they studied. For the two other species that they studied in detail, *B. subtilis* and *Mycobacterium smegmatis*, they have only mRNA data for *prfA* (coding RF1) and *prfB* (coding RF2). However, more *prfB* mRNA than *prfA* mRNA does not imply more RF2 than RF1 because RF2 is translationally regulated (Craigen *et al.* 1985;

Donly *et al.* 1990). Thus, their key conclusion that "UAG is truly a minor stop codon in all aspects" may be an unwarranted generalization.

Korkmaz *et al.* (2014) did notice that UAG in some bacterial species is more frequent than UGA. However, they interpreted these observations as likely arising from the process of UGA reassignment to a sense codon. They in particular drew attention to Mollicutes where many lineages use genetic code 4 with only two stop codons (UAA and UAG, with UGA reassigned to tryptophan). However, their Table 2 included bacterial species where UAG is used frequently, with no evidence that UGA is either reassigned or in the process of being reassigned. Korkmaz *et al.* (2014) also speculated that the combination of UAG and RF1 is translationally less efficient and accurate than that of UGA and RF2 which, however, is contrary to available experimental evidence reviewed above.

It may be insufficient to argue that UAG is a nearly universal minor stop codon in bacteria. Those bacterial species that use more UAG than UGA as stop codons may not at all be in the process of having UGA reassigned to sense codons, but instead may simply have more actively decoding RF1 than RF2 in their cells. This hypothesis, which may be termed codon-decoder adaptation hypothesis, is consistent with many previous experimental and bioinformatics studies, including Korkmaz et al. (2014). In fact, one of the key contributions in Korkmaz *et al.* (2014) is the confirmation that stop codon usage in *E. coli* is related to relative abundances of RF1 and RF2.

Proteomic studies have been carried out in many bacterial species, with 14 of them (covering Proteobacteria, Firmicutes, Cyanobacteria, Actinobacteria and Spirochetes) having both RF1 and RF2 quantified and deposited in PaxDB (Wang *et al.* 2015). Of particular value in these data is that relative abundance of RF1 and RF2 varies widely, which paves the way for evaluating the

effect of relative abundance of RF1 and RF2 on stop codon usage. The availability of protein abundance data for thousands of proteins also permits a more objective and comprehensive characterization of HEGs and LEGs and their respective stop codon usage.

We found UAA consistently over-represented in HEGs relative to LEGs, consistent with experimental studies showing UAA to be the most efficient stop codon. In contrast, UGA is always avoided in HEGs relative to LEGs. This is true even in species where UGA accounts for an overwhelming majority of stop codons and RF2 is far more abundant than RF1. In such species, UAA is mostly found in HEGs. UGA usage relative to UAG increases significantly with relative abundance of RF2, following the expectation that synonymous codons increase in usage with the abundance of their decoders (which are tRNAs in the case of sense codons and release factors in the case of stop codons). RF2 is more abundant than RF1 over a wide range of AT content, but decreases rapidly towards zero at extreme AT-richness. This explains why bacterial lineages with high genomic AT content often have UGA reassigned because the low RF2 would select strongly against UGA. There is no indication that UAG is a minor stop codon in bacteria as claimed by Korkmaz *et al.* (2014). Our results suggest a more plausible explanation for why UAA usage increases, and UGA usage decreases, with PAT3, but UAG usage remains low over the entire PAT3 range.

2.4 Materials and Methods

2.4.1 Classifying genes according to gene expression

We have used protein abundance and Index of Elongation Efficiency, I_{TE} (Xia 2015a) as proxies of gene expression. Protein abundance data were downloaded from PaxDB (Wang *et al.* 2015). For species with multiple proteomic studies, only the integrated datasets are downloaded

and used to rank the coding sequences. The protein ID in PaxDB is often the Uniprot ID and needs to be mapped to gene names (or GI or GeneID) in a GenBank file for individual species (e.g., *Bacillus subtilis*). We downloaded the paxdb-uniprot-links file relevant to the species (e.g., 224308-paxdb_uniprot.txt for *B. subtilis*), saved the Uniprot ID (the last column) to a file (e.g., BsUniprotID.txt), browsed to http://www.uniprot.org/uploadlists/, under "Provide your identifiers" uploaded the BsUniprotID.txt file, under "Selection options" selected the mapping from 'UniProtKB AC/ID' to 'Gene name' (or GI or GeneID), and clicked 'Go'. The resulting mapping file was generated with two columns (original input Uniprot IDs and the mapped gene name (or Gis GeneID) corresponding to gene name or other IDs in a GenBank file.

An alternative proxy for gene expression is I_{TE} which requires codon usage data from both HEGs and LEGs. For each species, we ranked the genes by protein abundance, took the top 40 ribosomal protein genes as HEGs and bottom 40 genes with non-zero values as LEGs, and compiled codon usage table for HEGs and LEGs separately. These codon usage tables were then used to compute I_{TE} with DAMBE. The resulting I_{TE} is then used as a proxy of gene expression. The advantage of using I_{TE} is that it can be used for all genes and that it is less affected by differential mRNA abundance and protein degradation.

After genes were ranked by either protein abundance or I_{TE}, we have used top and bottom 25% of genes as HEGs and LEGs, respectively, to compile stop codon usage, so the actual number of genes taken as HEGs and LEGS differ among species. If 25% of genes are greater than 1000, then only 1000 genes were used. The two ways of ranking genes by their expression (i.e., by protein abundance or by I_{TE}) lead to similar results. The results presented are based on the ranking by protein abundance. The results from ranking by I_{TE} have slightly stronger patterns with slightly smaller p values.

2.4.2 *RF1 and RF2 concentrations*

We compiled RF1 and RF2 concentrations from proteomic data at PaxDB. Only 14 species have both RF1 and RF2 measured and were included. An average is used when multiple values are available. Our values are therefore not always the same as those RF1 and RF2 values in the integrated datasets in PaxDB because the latter includes studies in which either RF1 or RF2 is measured.

2.4.3 *Phylogenetic reconstruction*

For computing phylogeny-based independent contrasts, we extracted small subunit ribosomal RNA (ssu rRNA) sequences from genomic sequences in GenBank (with Accession included in Fig. 2.4). For species with multiple ssu rRNA genes, only the first one is used for phylogenetic reconstruction. The sequences were aligned by MAFFT (Katoh *et al.* 2009) with the slow but accurate '–localpair' and '–maxiterate = 1000' options.

Two phylogenetic reconstruction methods were used. The first was PhyML (Guindon and Gascuel 2003) with GTR (or HKY85). The tree improvement option '-s' was set to 'BEST' (best of NNI and SPR search). The '-o' option (optimize starting tree) was set to 'tlr' which optimizes the topology, the branch lengths and rate parameters. The other was a distance-based FastME method (Desper and Gascuel 2002; Desper and Gascuel 2004) implemented in DAMBE (Xia 2013c), with the simultaneously estimated maximum composite likelihood distance (Tamura *et al.* 2004; Xia 2009) based on the TN93 model (MLCompositeTN93). The two trees from the two methods have identical topology and almost perfectly correlated branch lengths. The independent contrasts were generated by using the CONTRAST program in the PHYLIP package (Felsenstein 1989).

2.5 Results and Discussion

We ranked protein-coding genes by i) protein abundance and ii) by index of translation efficiency (I_{TE}), and the top 25% and bottom 25% of genes are taken as HEGs and LEGs (see Methods section for details). We defined PUAA, PUGA and PUAG as the proportion of the three stop codons, and P2UGA = NUGA/(NUGA+NUAG), where NUGA and NUAG are the number of UGA and UAG codons. Note that P2UGA is different from PUGA which is NUGA/(NUGA+NUAA+NUAG). PUAA, PUGA, PUAG and P2UGA based on HEGs or LEGs will be subscripted by 'HEG' or 'LEG', respectively. We also defined PRF2 as [RF2]/([RF1]+[RF2]) where [X] is the concentration of X. We used AT content at the third codon site (PAT3) as a proxy of AT-biased mutation.

To facilitate presentation, we rebranded the conventional codon-anticodon adaptation hypothesis for sense codons as codon-decoder adaptation hypothesis. This generalized hypothesis predicts that a codon, be it sense or stop codon, increases its usage with its decoders, and that such increase is typically more pronounced in HEGs than in LEGs.

2.5.1 UAA is a major codon in all 14 species

PUAA does not increase or decrease with the relative availability of RF2 (PRF2, Fig. 2.1a and Table 2.1) which is expected because RF1 and RF2 can both decode UAA with roughly equal efficiency, at least in *E. coli* (Scolnick *et al.* 1968; Jorgensen *et al.* 1993; Freistroffer *et al.* 1997; Ito *et al.* 2000). What is remarkable is that PUAA is always higher in HEGs than in LEGs in all 14 species (Fig. 2.1), even in extremely GC-biased genomes (Fig. 2.1b, PAT3 is only 0.1262 for *P. aeruginosa* and 0.2018 for *Mycobacterium tuberculosis*). In contrast, UGA is always avoided in HEGs relative to LEGs (Fig. 2.1), even in species where UGA represents an overwhelming

majority of stop codons in all genes. Among the 5925 annotated protein-coding genes in *P. aeruginosa* (NC_011770), 4651 terminate with UGA, 684 with UAG and only 590 with UAA (which are mostly in HEGs). This preponderance of UGA stop codons is associated with greater abundance of RF2 than RF1 (PRF2 = 0.7475 in *P. aeruginosa*). Given so many UGAs and so few UAAs in *P. aeruginosa*, one would have expected RF2 to evolve a higher efficiency to decode UGA, perhaps at the cost of reduced efficiency of decoding UAA, so that HEGs would have an increased preference for UGA relative to UAA. However, this expectation is not supported as UGA is used less frequently in HEGs than in LEGs in these two species (Fig. 2.2a, PUGA.HEG = 0.6880 and PUGA.LEG = 0.8160 for *P. aeruginosa*). Thus, although UAA is rare in *P. aeruginosa*, it is strongly preferred by HEGs. In contrast, UGA in *P. aeruginosa* is frequent (and RF2 more abundant than RF1), yet it is avoided by HEGs. The difference in stop codon usage between 500 HEGs and 500 LEGs is highly significant based on Chi-square test with Yates correction for continuity ($\chi^2 = 91.23$, DF = 2, $p < 0.0001$). One possible explanation for this lack of expected RF2 evolution is that genomic AT content could change very quickly (Marin and Xia 2008; Nikbakht *et al.* 2014), whereas functional modification of a key cellular protein is typically a very slow process. In short, GC-biased mutation can increase UGA at the cost of UAA, but does not change the preference of UAA by HEGs in all 14 species studied.

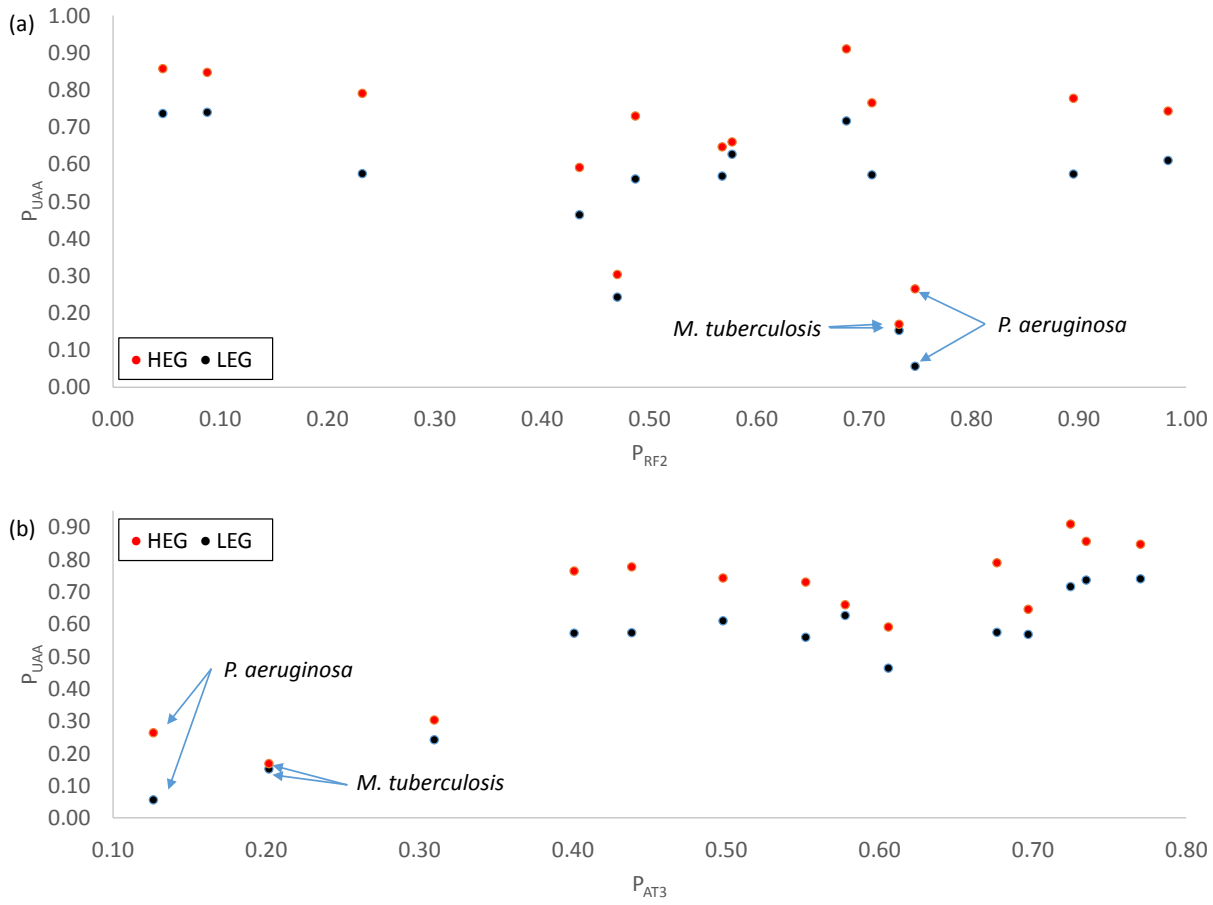

Fig. 2.1 Stop codon UAA is preferred in highly expressed genes (HEGs) relative to lowly expression genes (LEGs) in all 14 species, regardless of (a) relative abundance in RF1 and RF2, measured by PRF2 as RF2/(RF1+RF2), or (b) proportion of AT at third codon site (P_{AT3}).

In model organisms such as *E. coli*, UAA has been shown to be the most efficiently decoded and UGA the least (Strigini and Brickman 1973; Geller and Rich 1980; Parker 1989; Jorgensen *et al.* 1993; Tate *et al.* 1999). Highly expressed genes in *E. coli* were previously observed to prefer UAA as stop codons (Jin *et al.* 2002). Our result, with 14 species covering a wide taxonomic spectrum, suggests that UAA is a more efficient stop signal than other stop codons in bacteria in general. This implies that a transgenic gene expressed in a bacterial species should be terminated with UAA to enhance termination efficiency.

The other AT-poor species, *M. tuberculosis*, also exhibit strong difference between HEGs and LEGs ($\chi^2 = 16.23$, DF = 2, p = 0.0003), but here both UAG (Fig. 2.2) and UAA (Fig. 2.1) are preferred in HEGs relative to LEGs. The strong preference of UAG in HEGs is clearly at odds with the conclusion in Korkmaz et al. (2014) that "UAG is a minor stop codon in all aspects". PUAG.HEG is also higher than PUAG.LEG in *Microcystis aeruginosa* and *Acidithiobacillus ferrooxidans*, and the two are equal in *Helicobacter pylori* (Table 2.1). Thus, UAA is universally preferred in HEGs, UAG is preferred in HEGs in 3 species, and UGA is avoided in HEGs in all 14 species.

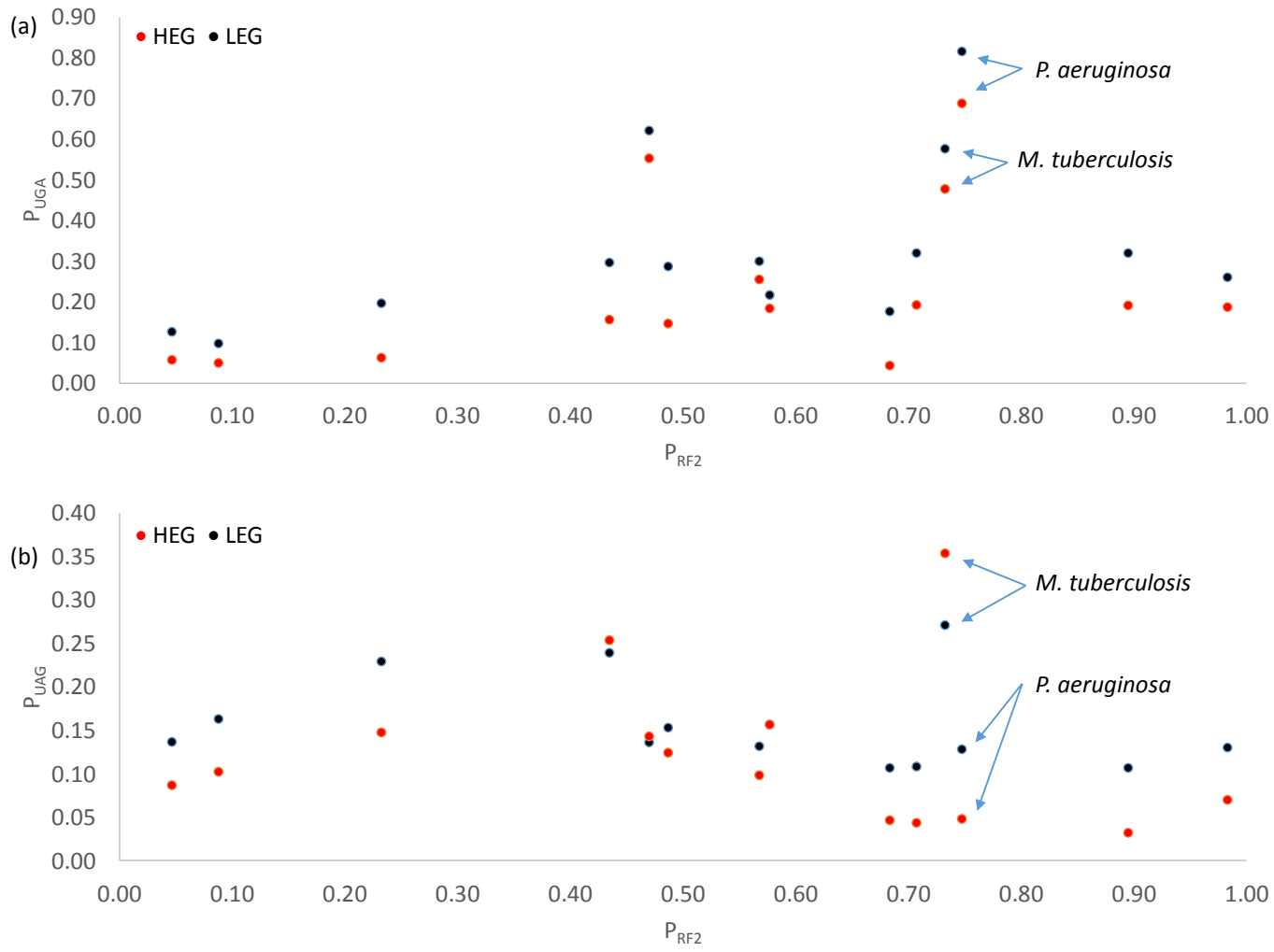

Fig. 2.2. Stop codon UGA is never preferred in HEGs relative to LEGs even RF2 is far more abundant than RF1 (a), and stop codon UAG is preferred in HEGs in 4 of the 14 species (b).

If we do not contrast between HEGs and LEGs, and focus on HEGs only or all genes, then we may arrive at a wrong conclusion that UGA is the major codon and UAA the minor codon in *M. tuberculosis* and *P. aeruginosa* because UGA is more frequent than UAA or UAG. Take HEGs in *M. tuberculosis* for example. PUGA.HEG, PUAG.HEG and PUAA.HEG are 0.4775, 0.35375 and 0.16875, respectively (Table 2.1). However, UGA is not the major codon because UGA is even more frequent than UAA or UAG in LEGs, with PUGA.LEG, PUAG.LEG and PUAA.LEG being 0.57625, 0.27125 and 0.1525, respectively (Table 2.1). It is crucially important to contrast

codon usage between HEGs and LEGs in identifying codons favoured by decoder-mediated selection (Eyre-Walker and Bulmer 1995; Xia 2015a).

2.5.2 *Relative usage of UAG and UGA depends on relative abundance of RF1 and RF2*

Because UAG is decoded by RF1 and UGA by RF2, we expect P2UGA, which is the proportion of UGA within (UGA+UAG), to increase with PRF2. The concentration of RF1 and RF2 vary widely among the 14 bacterial species, with PRF2 varying from 0.046 in *Bacillus anthracis* to 0.9830 in *Yersinia pestis* CO92. The codon-decoder adaptation hypothesis predicts that species like *B. anthracis* should use UAG more frequently than UGA in HEGs and species like *Y. pestis* CO92 should use UGA more frequently than UAG. We tested this prediction by using regression on the original PRF2 and P2UGA and on phylogeny-based independent contrasts (Felsenstein 1985a). The latter method alleviates the problem of data dependence due to sharing of ancestry.

The stop codon usage among the 14 bacterial species is as predicted by the codon-decoder adaptation hypothesis (Fig. 2.3). First, both LEGs and HEGs follow the same trend with P2UGA increasing with PRF2 ($p < 0.01$ in both LEGs and HEGs, Fig. 2.3). Second, the pattern is stronger in HEGs than in LEGs. For example, in the three species with the highest PRF2 values, P2UGA.HEG is greater than P2UGA.LEG (Fig. 2.3). In the three species with the lowest PRF2 values, P2UGA.HEG is lower than P2UGA.LEG (Fig. 2.3). Such a pattern is consistent with that observed in sense codons. There is no indication that "UAG is truly a minor stop codon in all aspects" (Korkmaz *et al.* 2014), and the speculations by Korkmaz *et al.* (2014) that the combination of UAG and RF1 is worse than that of UGA and RF2 in translation termination efficiency and accuracy is insufficiently substantiated. A codon becomes rare when its decoder is

rare and vice versa. One may say that UAG is a minor codon in *E. coli*, but UAG is not a universal minor codon.

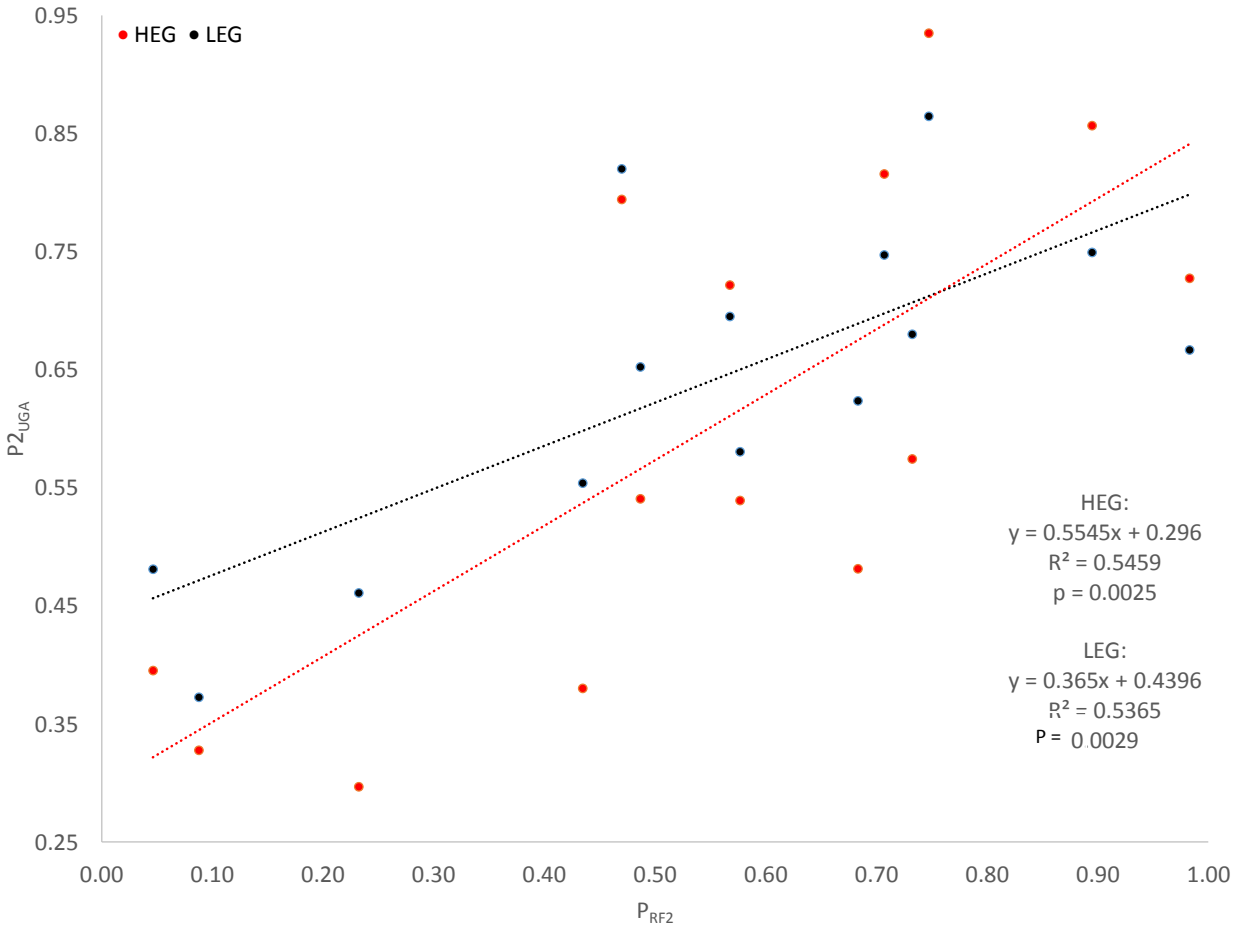

Fig. 2.3. Relative usage of UGA and UAG, measured as $P2_{UGA}$ = UGA/(UGA+UAG), increases significantly with relative abundance of RF2, measured as P_{RF2} = RF2/(RF1+RF2).

Based on the regression line for P2UGA.HEG on PRF2, P2UGA.HEG equals 0.5 when PRF2 = 0.3679 (i.e., when RF2:RF1 is about 0.6:1). Thus, if we may make a liberal interpretation of this result from a limited data of 14 species, then UGA will tend to be less frequent than UAG (i.e., P2UGA.HEG < 0.5) when PRF2 is smaller than 0.3679, but UAG will tend to be less frequent than UGA when PRF2 is greater than 0.3679 (assuming equal efficiency between RF1

decoding UAG and RF2 decoding UGA). In our study, 3 of the 14 species (*Streptococcus pyogenes*, *Bacillus anthracis*, and *Staphylococcus aureus*) have PRF2 smaller than 0.3679 (Fig. 2.3) and their UGA, instead of UAG, is the less frequent of the two, with their P2UGA.HEG values being 0.2969, 0.3953, and 0.3279, respectively.

Strictly speaking, the regression and significance tests of the regression slope in Fig. 2.3 may not be accurate because the P2UGA and PRF2 values are not independent due to the sharing of ancestry among the bacterial species. For example, *E. coli*, *S. enterica*, and *Y. pestis* are closely related, so are *B. subtilis* and *B. anthracis.* In the extreme case when two species are identical, then the two associated data points should really be treated as just one data point. To alleviate this problem, we have built a tree from the small subunit ribosomal RNA from the 14 species (Fig. 2.4) and computed the independent contrasts (Felsenstein 1985a) for PRF2 and P2UGA based on the tree and the data in Table 2.1. The results for regressing P2UGA.HEG on PRF2 are slope = 0.3062, R^2 = 0.5693, P =0.0336, and those for regressing P2UGA.LEG on PRF2 are slope = 0.2663, R^2 = 0.5800, P = 0.0297. This result does not depend heavily on the tree in Fig. 2.4. We have generated 100 bootstrap trees and repeated independent contrast analysis for each tree. The P value is always smaller than 0.05. Thus, P2UGA depends significantly on PRF2, following the prediction of codon-decoder adaptation hypothesis.

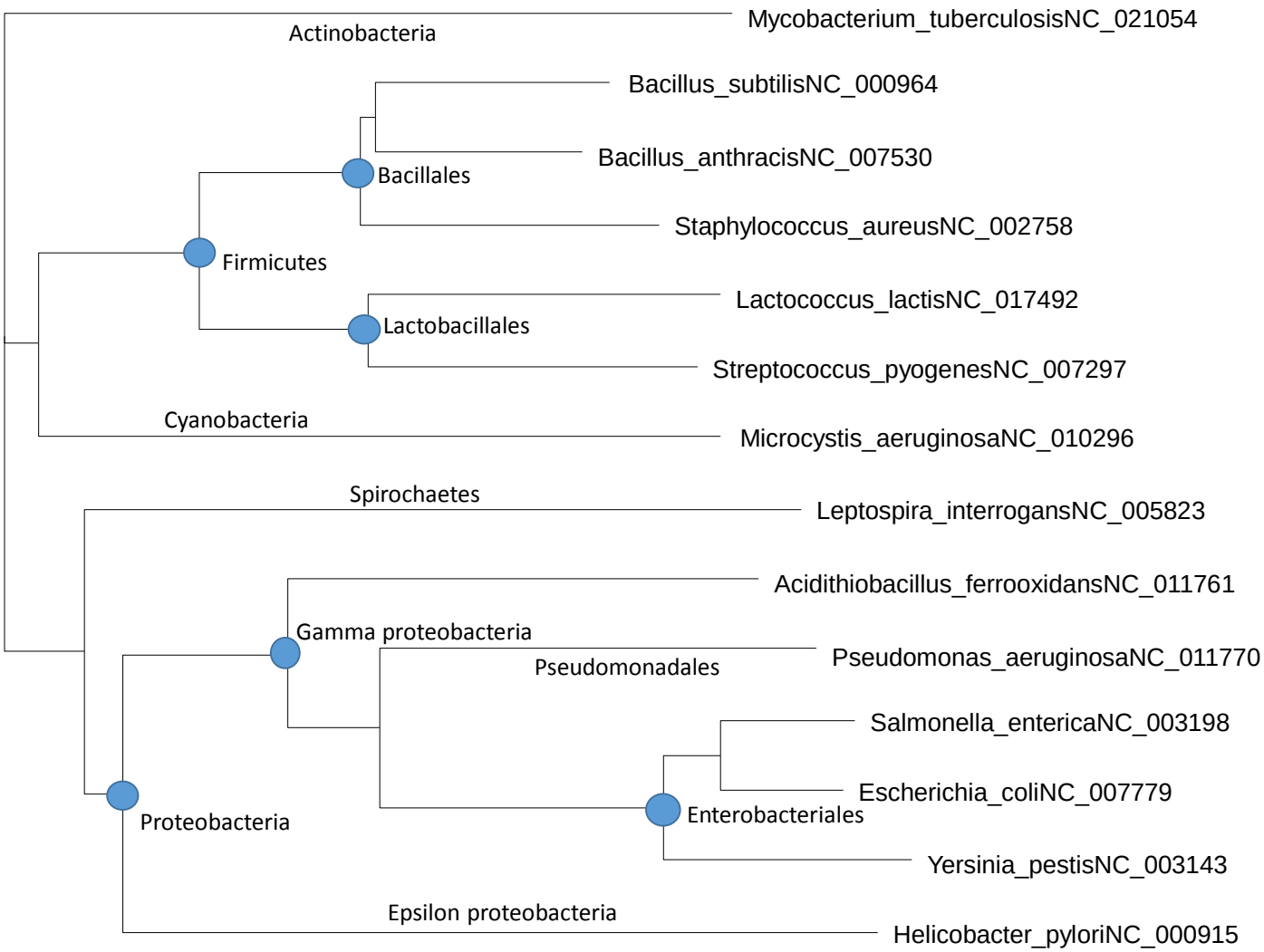

Fig. 2.4. Phylogenetic tree built with small subunit ribosomal RNA sequences (ssu rRNA), used for independent contrasts, with leaves denoted by species name and GenBank accession for genomes from which the ssu rRNA sequences are extracted. Only the first annotated ssu rRNA sequence is used.

2.5.3 P_{RF2} *decreases with genomic AT bias*

The wide variation in relative concentration of RF1 and RF2 (with PRF2 varies from 0.046 to 0.9830) raises the question of what affects PRF2. As previously noted (Korkmaz *et al.* 2014), bacterial species that lack the *prfB* gene and have UGA reassigned as a sense codon are typically associated with high genomic AT content. It is therefore reasonable to hypothesize that RF2

abundance decreases with AT content and disappears in species with extreme AT-bias so that UGA as a stop codon would be strongly selected against and eventually reassigned.

AT bias, measured by either the third codon position or by inter-gene sequences, indeed is negatively and highly significantly related to PRF2 (Fig. 2.5, the Spearman rank correlation is 0.6659, p = 0.0093, where PAT3 is the proportion of AT at third codon sites, and is similar to the proportion of AT in intergenic sequences). The relationship can be fitted well by the following equation:

$$P_{RF2} = \frac{0.81566 - P_{AT3}}{1 - P_{AT3}} \quad (1)$$

The fitted curve (Fig. 2.5), which accounts for 46.94% of the variation in PRF2, implies that PRF2 will rapidly approach 0 when PAT3 approaches 0.81566 or higher. This trend that PRF2 would approach 0 with increasing PAT3 explains why extremely AT-rich bacterial genomes frequently lose *prfB* and have stop codon UGA reassigned. The equation also explains why RF2 is more likely lost than RF1 because the concentration of RF1 does not approach 0 with changes in PAT3 (Fig. 2.5). These results offer empirical substantiation to previous models on stop codon reassignment (Osawa and Jukes 1989; Andersson and Kurland 1991; Sengupta and Higgs 2005; Ogawa *et al.* 2006; Sengupta *et al.* 2007; Higgs and Ran 2008).

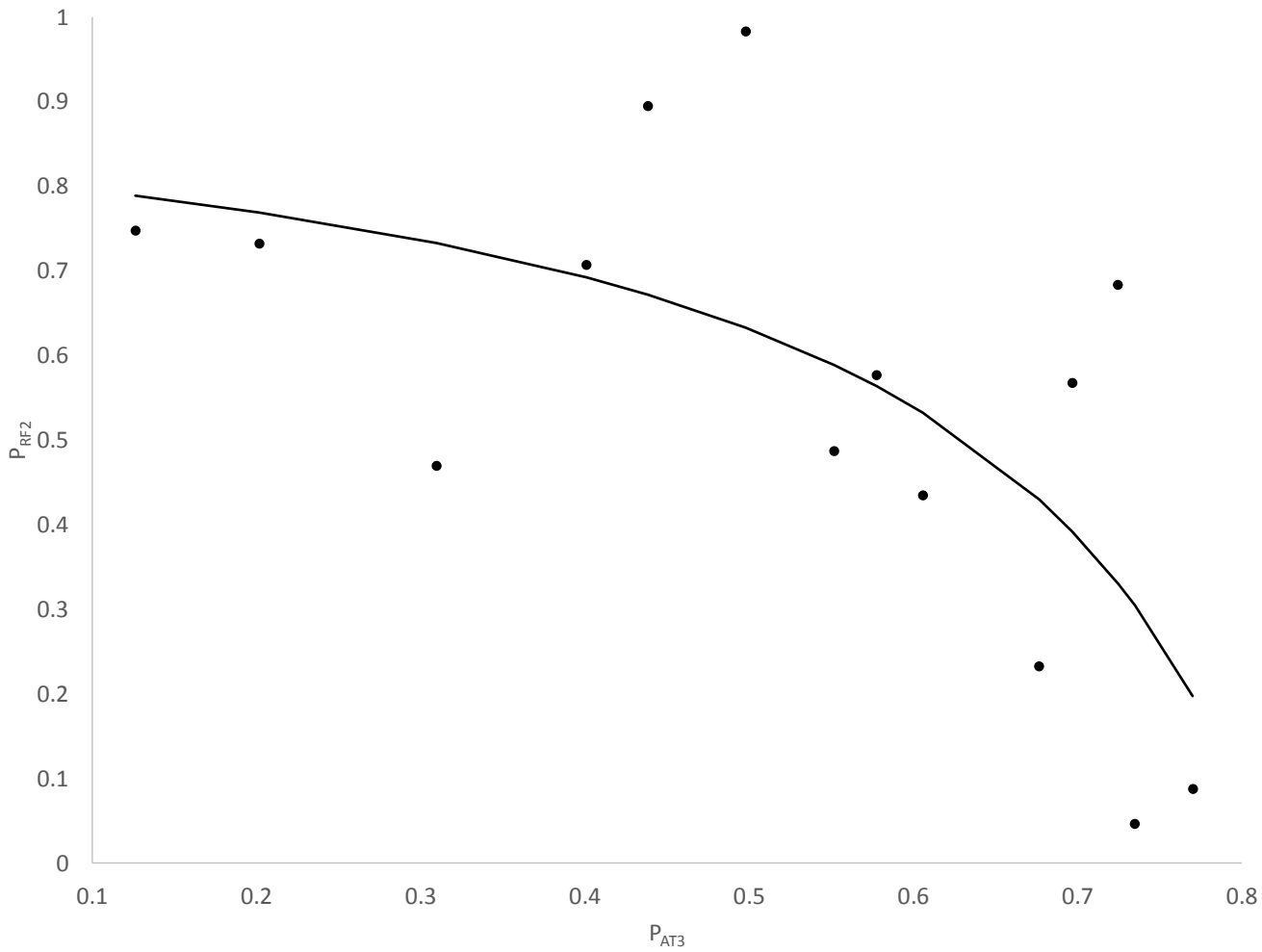

Fig. 2.5. Relative abundance of RF2 decreases rapidly at high range of AT content, measured by proportion of AT at third codon site (P_{AT3}).

We have previously mentioned that P2UGA.HEG tends to be smaller than 0.5 (i.e., more UAG than UGA) when PRF2 is smaller than 0.3679. According to Eq. (1), PRF2 will be smaller than 0.3679 when PAT3 > 0.70835. This result, if interpreted liberally, suggests that UAG will tend to be more frequent than UGA only when PAT3 is greater than 0.70835, and explains why UAG tends to be the least frequent in most bacterial species because relatively few bacterial genomes have PAT3 > 0.70835.

2.5.4 *Dynamic changes of stop codons with AT content*

One conspicuous pattern observed previously (Korkmaz *et al.* 2014) is that UAA usage increases, and UGA usage decreases, with AT content, but UAG usage remains low and hardly changes with AT content. This pattern is also visible in the 14 species here (Fig. 2.6). Korkmaz *et al.* (2014) interpreted this pattern as consistent with UAG being a minor codon that has translation termination efficiency and accuracy problems and is therefore nearly universally avoided. This interpretation by Korkmaz *et al.* (2014) is somewhat far-fetched for two reasons. First, as we have mentioned earlier, experimental evidence suggests that UAG is typically more efficient and accurate than UGA as a termination signal. Second, UAG is favoured by HEGs in 3 of the 14 species whereas UGA is avoided by HEGs in all 14 species. Furthermore, the interpretation does not explain why UGA becomes less frequent than UAG at high AT content which is particularly visible in Figure 2B in Korkmaz *et al.* (2014) for highly expressed genes.

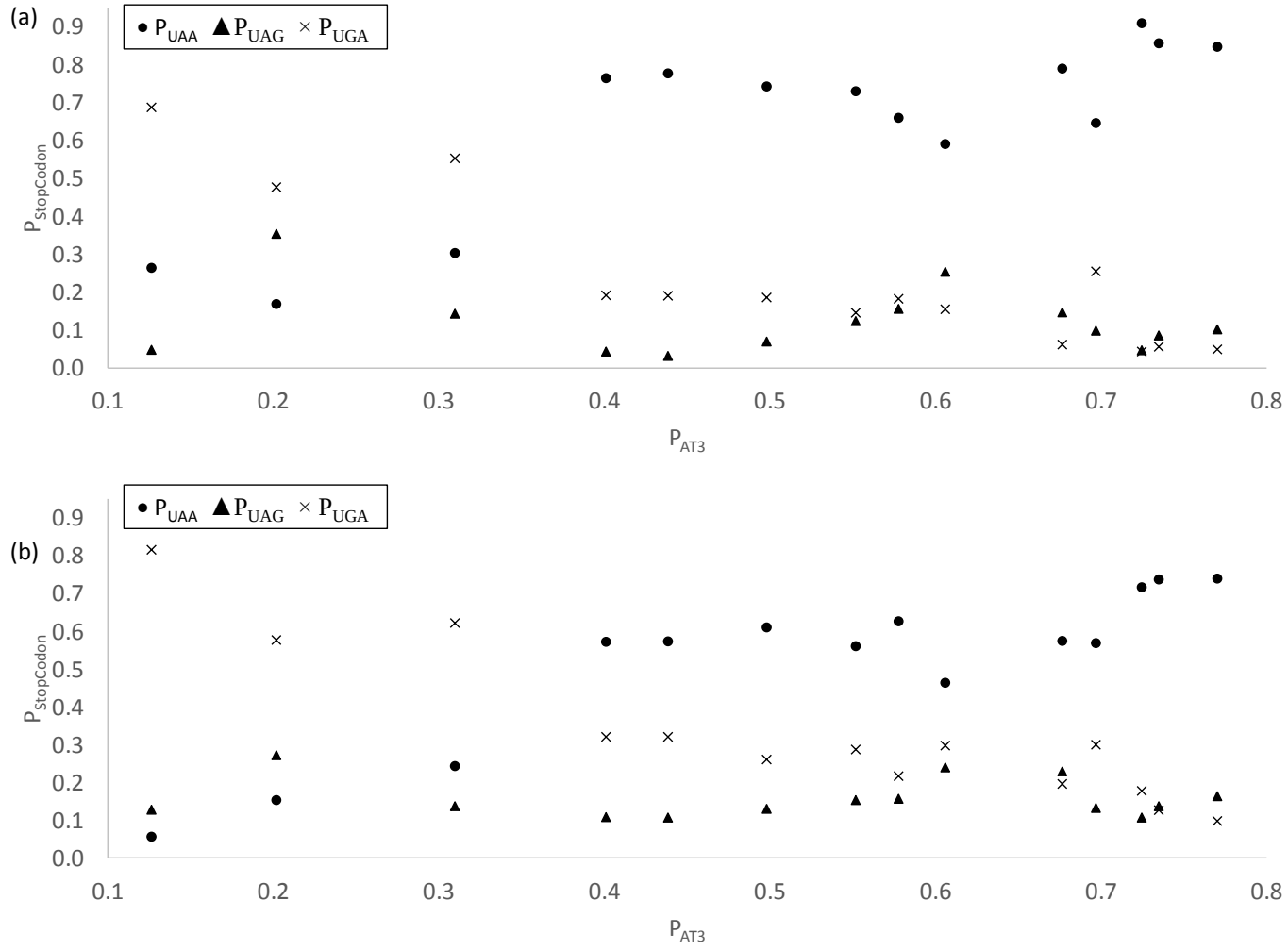

Fig. 2.6. UAA usage increases, and UGA usage decreases, with P_{AT3}, but UAG usage is low and changes little with P_{AT3}.

Our results on the change of PRF2 and PAT3 offers an alternative explanation for the observation of i) low UAG usage and ii) little change in UAG usage over the entire range of AT content in bacterial genomes. At the low PAT3 range, mutation would have favoured both UAG and UGA at the cost of UAA. However, PRF2 is high with low PAT3 (Fig. 2.5) which would favour UGA and select against UAG, keeping the latter at low frequency. At high PAT3, mutation would favour UAA against UGA and UAG and we expect the latter two to decrease. However, PRF2 approaches 0 at high PAT3 (Fig. 2.5), which selects strongly against UGA

codons, but little against UAG codons (as RF1 becomes the dominant release factor at high PAT3). This explains why, at high PAT3, UAG does not decrease as much as UGA in Fig. 2.6A and tend to have its usage higher than that of UGA. This pattern is also visible in Figure 2B in Korkmaz *et al.* (2014). In the mid-range of PAT3, UAA is overused because i) it is favoured by selection and ii) there is no mutation bias against it. Also in this range, PRF2 is still much greater than 0.5 (Fig. 2.5), favouring UGA against UAG and keep the latter at low frequency. So UAG usage is kept low and change little over the entire range of PAT3.

2.6 Acknowledgements

This study was funded by the Discovery Grant from Natural Science and Engineering Research Council of Canada to XX. We thank J. Silke and C. Vlasschaert for discussion and comments, and for suggestions from two reviewers that have led to substantial improvement of the paper.

3. Chapter three

The role of +4U as an extended translation termination signal in bacteria

3.1 Abstract

Termination efficiency of stop codons depends on the first 3' flanking (+4) base in bacteria and eukaryotes. In both *Escherichia coli* and *Saccharomyces cerevisiae*, termination read-through is reduced in the presence of +4U; however, the molecular mechanism underlying +4U function is poorly understood. Here, we perform comparative genomics analysis on 25 bacterial species (covering Actinobacteria, Bacteriodetes, Cyanobacteria, Deinococcus-Thermus, Firmicutes, Proteobacteria and Spirochaetae) with bioinformatics approaches to examine the influence of +4U in bacterial translation termination by contrasting between highly and lowly expressed genes (HEGs and LEGs). We estimated gene expression using the recently formulated Index of Translation Elongation, I_{TE}, and identified stop codon near-cognate tRNAs from well annotated genomes. We show that +4U was consistently over-represented in UAA-ending HEGs relative to LEGs. The result is consistent with the interpretation that +4U enhances termination mainly for UAA. Usage of +4U decreases in GC-rich species where most stop codons are UGA and UAG, with few UAA-ending genes, which is expected if UAA usage in HEGs drives up +4U usage. In highly expressed genes, +4U usage increases significantly with abundance of UAA nc_tRNAs (near-cognate tRNAs which decode codons differing from UAA by a single nucleotide), particularly those with a mismatch at the first stop codon site. UAA is always the preferred stop codon in highly expressed genes, and our results suggest that UAAU is the most efficient translation termination signal in bacteria.

3.2 Contribution

The data, results and interpretations in this chapter were accepted in *GENETICS* (November 2016). The development of the hypotheses, data analyses and interpretations in this chapter is contributed by Yulong Wei and Dr. Xuhua Xia.

3.3 Introduction

Different stop codons have different termination efficiency, and replacing UGA with UAA reduces termination read-through of human genes expressed in *E. coli* (Meng *et al.* 1995; Cesar Sanchez *et al.* 1998). The discrepancies in termination efficiency among stop codons in bacteria are largely attributed to: i) the competition between nc_tRNAs and class I release factors (RF1 and RF2) in decoding stop codon (Nakamura *et al.* 1996; Tate *et al.* 1999; Blanchet *et al.* 2014), mediated by the relative abundance of RF1 and RF2 (Korkmaz *et al.* 2014; Wei *et al.* 2016), and ii) nucleotide sites downstream of stop codons interacting with 18S rRNA and modulating the structural stability of binding sites for RF1 and RF2 (Namy *et al.* 2001) or interacting directly with release factors based on inferences from crosslinking experiments in both bacterial(Poole *et al.* 1998) and eukaryotic species (Bulygin *et al.* 2002).

Termination efficiency of stop codons depends on the first 3' flanking (+4) base in bacterial species such as *E. coli* and *Salmonella typhimurium* (Bossi and Ruth 1980; Miller and Albertini 1983; Tate *et al.* 1995; Poole *et al.* 1998) and in eukaryotes (Manuvakhova *et al.* 2000; Jungreis *et al.* 2011). The inefficiency of translation termination associated with +4C, especially in UGA-ending genes, is well documented in both bacteria (Brown and Tate 1994; Poole *et al.* 1995; Tate *et al.* 1999) and eukaryotes (Manuvakhova *et al.* 2000; Namy *et al.* 2001; Jungreis *et al.* 2011; Dabrowski *et al.* 2015; Beznoskova *et al.* 2016a). UGA-C contributes to the auto-regulation of

prfB (coding RF2) translation (Craigen *et al.* 1985; Craigen and Caskey 1986), with a truncated RF2 produced when functional RF2 is abundant and a full-length functional RF2 produced when functional RF2 is rare. Baranov *et al.* (2002) identified the *prfB* gene in 87 bacterial species using BLAST (Altschul *et al.* 1990), and revealed programmed frameshift in 70% of these bacteria. The segment involved in the frameshift (CUU UGA CNN) and the translated segment (CUU GAC NNN) are always conserved, showing ribosome slippage at UGA-C.

UGA is particularly prone to be misread by $tRNA^{Trp}$ when followed by +4A in *E. coli* (Engelberg-Kulka 1981) and yeast (Geller and Rich 1980). A recent study in yeast by Beznoskova *et al.* (2016a) measured the read-through of termination tetra-nucleotides (e.g. UGA-C) in dual luciferase constructs. Indeed, +4C increases read-through in all three stop codons, but particularly so in UGA in yeast. Furthermore, UGA-A and UGA-G enhance misreading by $tRNA^{Trp}$ and $tRNA^{Cys}$, respectively (Beznoskova *et al.* 2016b). In contrast, UGA-U, UAA-U and UAG-U are all associated with low read-through (Beznoskova *et al.* 2016b). The finding that +4U reduces termination read-through is consistent with the observation that this base is over-represented in *E. coli*, especially in UAA-ending genes (Brown *et al.* 1990; Poole *et al.* 1995; Tate *et al.* 1996).

Early studies in *E. coli* suggest that the decoding efficiency of RF2 depends on the +4 base (Brown and Tate 1994; Tate *et al.* 1996; Poole *et al.* 1997; Poole *et al.* 1998). In particular, (Brown and Tate 1994) and Poole *et al.* (Poole *et al.* 1998) revealed that RF2 crosslinks with UAA and with UGA at the first (+1) base and with the downstream +4 base; and the crosslink efficiency between RF2 and stop codons is promoted in the presence of +4U. Thus, +4U may participate in recruiting RF2 to the stop codon. Similarly, studies in eukaryotes found crosslink between the +4 base and eRF1 in human UAA-ending genes (Bulygin *et al.* 2002).

If +4 site really serves as part of an extended stop signal, and if +4U enhances the stop signal relative to other nucleotides, then one can immediately predict that highly expressed genes (HEGs), which are under selection to evolve toward high translation initiation, elongation and termination efficiency, should prefer +4U more strongly than lowly expressed genes (LEGs). Furthermore, it is possible that different stop codons may require different +4 nucleotides to enhance the stop signal. In particular, GC-rich species may have difficulty maintaining a +4U site and may have different combinations of stop codon and +4 nucleotide from those AT-rich species. Testing these predictions constitute the first part of this paper.

Stop codons can be misread by near cognate tRNAs (nc_tRNAs) in *Escherichia coli* (Sambrook *et al.* 1967; Strigini and Brickman 1973), coliphage (Weiner and Weber 1973), eukaryotic viruses (Beier and Grimm 2001), the yeast *Saccharomyces cerevisiae* (Blanchet *et al.* 2014) and mammals (Geller and Rich 1980). Available data suggest termination read-through is most frequent at UGA, less at UAG and least at UAA, in both bacteria and eukaryotes (Parker 1989; Jorgensen *et al.* 1993; Tate *et al.* 1999; Dabrowski *et al.* 2015).

Stop codon read-through can occur in yeast by the incorporation of nc_tRNAs with wobble-pairing at the third stop codon site (Beznoskova *et al.* 2015; Beznoskova *et al.* 2016a), or at the first stop codon site involving $tRNA^{UUG/Gln}$ and $tRNA^{CUG/Gln}$ misreading UAA and UAG, respectively (Blanchet *et al.* 2014; Roy *et al.* 2015; Roy *et al.* 2016). In the yeast, $tRNA^{Gln}$, $tRNA^{Tyr}$ and $tRNA^{Lys}$ can misread stop codons UAA and UAG, whereas $tRNA^{Trp}$, $tRNA^{Cys}$ and $tRNA^{Arg}$ can misread stop codon UGA (Blanchet *et al.* 2014). Misreading of UAA and UAG by $tRNA^{Gln}$ also occur in *E. coli* (Nilsson and Ryden-Aulin 2003). UGA can be misread by $tRNA^{Trp}$ decoding UGG in both *E. coli* and *B. subtilis* (Engelberg-Kulka 1981; Matsugi and Murao 1999; Matsugi and Murao 2000; Nilsson and Ryden-Aulin 2003).

How +4U may enhance the stop codon signal remains unknown. Namy *et al*. (2001) speculated that, in yeast UAG-ending genes, several bases at the 3' UTR leading with +4C may pair with yeast 18S rRNA and destabilize secondary structures in the ribosome, preventing release factors from binding to stop codons. However, it is possible that +4U may serve to prevent misreading of stop codons by nc_tRNA. If this is the case, then +U usage should increase with the frequency of nc_tRNA which is an easily testable prediction. Testing this prediction constitutes the second part of this study.

We analyzed the genomic and proteomic data in 25 bacterial species (Supplementary Table S1) whose protein abundance data are present in PaxDB 4.0 (Wang *et al.* 2015) to examine the effect of +4 site and nc_tRNA on termination efficiency of the three stop codons. We found +4U consistently over-represented in HEGs in contrast to LEGs in bacteria. However, +4U usage in HEGs decreased in GC-rich bacterial species where most stop codons are UGA and UAG, suggesting that UGA and UAG do not need +4U as a stop signal enhancer as much as UAA. In HEGs, +4U usage also increases significantly with the abundance of UAA nc_tRNAs, suggesting that +4U increases UAA termination efficiency presumably by reducing misreading of UAA by nc_tRNAs.

3.4 Materials and Methods

3.4.1 *Protein expression data*

Proteomic data are available in PaxDB 4.0 (Wang *et al.* 2015) for 26 bacterial species of which one (*Mycoplasma pneumoniae*) is excluded from this study. The reason for the exclusion is that *M. pneumoniae* uses genetic code 4 that is different from the other species which use genetic code 11. *M. pneumoniae* uses only two stop codons (UAA and UAG, decoded by RF1) and does not have *prfB* genes coding for RF2 (which would decode UAA and UGA). The

integrated dataset was chosen when there are multiple datasets for a single species. *Bacillus subtilis* protein IDs in PaxDB are uniprot IDs; the "Retrieve/ID mapping" function in UniProt (Pundir *et al.* 2016) was used to map Uniprot IDs to Gene IDs.

Proteomic data are used to classify genes into HEGs and LEGs for compiling codon usage tables of HEGs and LEGs that are needed for computing the index of translation elongation or (ITE, Xia 2015b). I_{TE} incorporates the tRNA-mediated selection and the effect of background mutation and is therefore advantageous over codon adaptation index (Sharp and Li 1987; Xia 2007b) or tAI (Dos Reis *et al.* 2004) when genomes of diverse GC% are used in analysis. We used I_{TE} as a proxy of translation efficiency. That is, genes with a high I_{TE} are expected to be under stronger selection for translation efficiency than genes with a low I_{TE}.

For each of the 25 species, 40 ribosomal protein genes with the highest protein abundances (ppm) and 40 genes with the lowest non-zero protein abundances were taken from each species to compile codon usage for HEGs and LEGs, respectively. I_{TE} was computed with the option of "Break 8-fold and 6-fold families into 2". Only non-pseudo and non-hypothetical genes were selected in this study.

Among the 25 species, five species (*Bartonella henselae*, *Helicobacter pylori*, *Leptospira interrogans*, *Pseudomonas aeruginosa*, and *Synechocystis sp*.) do not exhibit clear differences in codon usage between HEGs and LEGs. This means that I_{TE} will not be a good proxy for translation efficiency in these five species. *Shigella flexneri* is phylogenetically nested within *E. coli* strains and therefore does not supply an independent data point. For this reason, only those 19 remaining species were used for I_{TE}-related analysis.

3.4.2 *Processing bacterial genomes*

The bacterial genomes were retrieved from GenBank, and coding sequences (CDSs) were extracted by using DAMBE (Xia 2013b) for computing I_{TE}. An alternative set of HEGs consists of all ribosomal protein genes extracted from DAMBE (Xia 2013b) based on genomic annotation. We also extracted small subunit ribosomal RNA (ssu rRNA) genes from each species for building a phylogenetic tree for computing independent contrasts. For each stop codon, their nc_tRNAs (Table 1) were compiled. No tRNA has anticodons AUA or ACA in all these species and these two which are not included in Table 1.

Table 3.1. Anticodons of nc_tRNAs for each of the three stop codons. No tRNA has AUA or ACA anticodon in all bacterial species we studied.

UAA	UAG	UGA
Glu-TTC	Glu-CTC	Gly-TCC
Gln-TTG	Gln-CTG	Arg-TCG
Lys-TTT	Lys-CTT	Arg-TCT
Leu-TAA	Leu-CAA	Leu-TAA
Ser-TGA	Trp-CCA	Ser-TGA
Tyr-GTA	Ser-CGA	Trp-CCA
	Tyr-GTA	Cys-GCA

3.4.3 *Phylogenetic reconstruction and independent contrasts*

Variables measured from a set of species are typically not independent because of shared ancestry. Phylogeny-based independent contrasts (Felsenstein 1985a) were computed to alleviate this problem. We aligned ssu rRNA sequences aligned by MAFFT (Katoh *et al.* 2009) with the LINSI option that generates the most accurate alignment ('–localpair' and '–maxiterate = 1000').

PhyML (Guindon and Gascuel 2003) was used for phylogenetic reconstruction, with GTR substitution model and six categories of gamma-distributed rates. The resulting tree (Fig. 3.1) was used for computing independent contrasts (Felsenstein 1985b) as numerically illustrated in Xia (2013a). The same approach is used to reconstruct the tree for the 19 species (indicated in Fig. 3.1) in I_{TE}-related analysis.

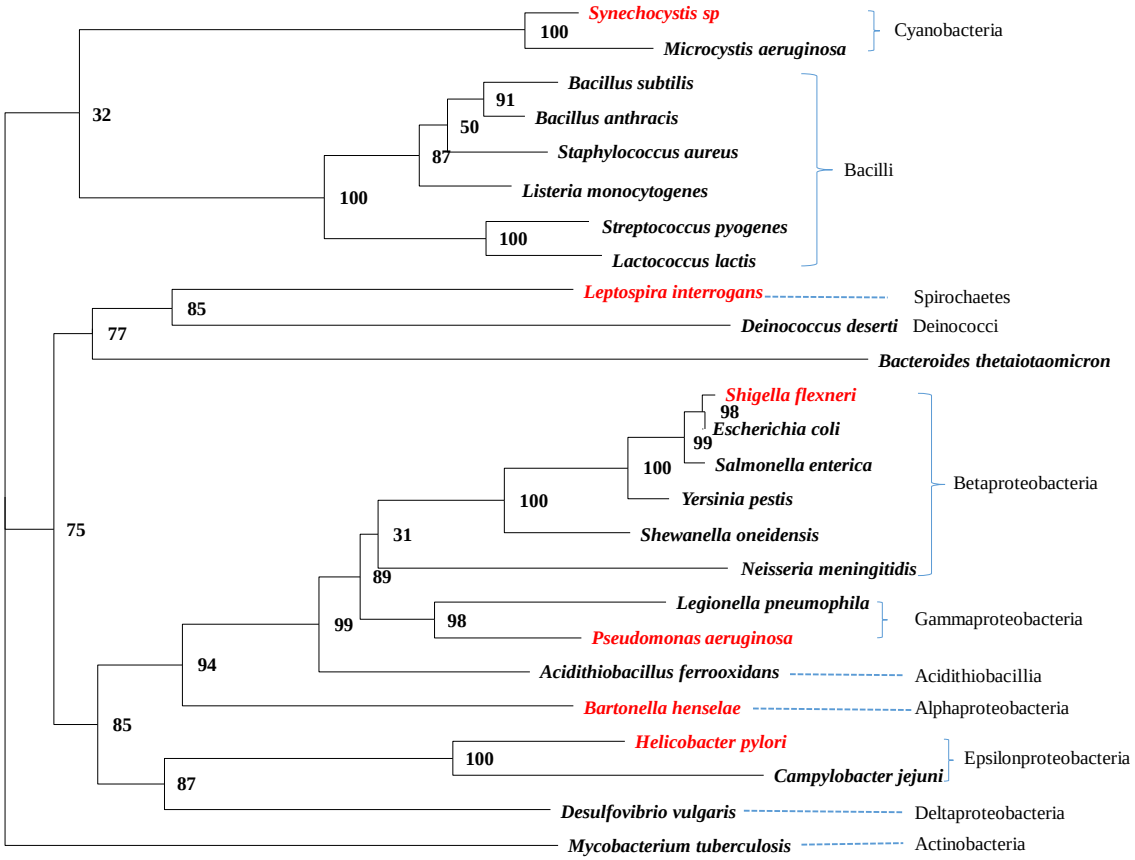

Fig. 3.1. Phylogenetic relationship among the 25 bacterial species. The six species in red were not used in I_{TE}-related analysis (see METHODS for reason of exclusion). The branch length for *Bacteroides thetaiotaomicron* was shortened by nearly 1/3 for a more compact display.

Because the bacterial species involve deep phylogeny with limited resolution close to the root node, we assessed the effect of different trees on the results of independent contrasts by using 100 bootstrapped trees. DAMBE takes a file with the 100 trees and automatically perform independent contrasts for each tree. We have also used a tree built with PhyPA suitable for deep phylogenetic relationships (Xia 2016). The PhyPA is based on pairwise sequence alignment using default options simultaneously estimated distances based on TN93 model (Tamura and Nei 1993).

3.5 Results

3.5.1 *HEGs and LEGs differ in the relationship between +4U and stop codons*

+4U is strongly overrepresented in all stop codons in *E. coli*, especially for UAA-ending and UGA-ending HEGs (Fig. 3.2). In contrast, +4U is overrepresented UAA-ending HEGs relative to UAA-ending LEGs in *B. subtilis* (Fig. 3.2). In each species, the nucleotide distribution at the +4 site depends highly significantly on stop codons ($p < 0.0001$) when tested by log-linear models. The difference between the two species are also highly significant ($p < 0.0001$), with the main contribution to the difference from +4 sites following UAG and UGA. While both species exhibit overuse of +4U in UAA-ending HEGs, only *E. coli* overused +4U in UGA-ending HEGs. A previous experimental study demonstrated a strong effect of +4U in increasing the termination efficiency of UGA (Kopelowitz *et al.* 1992).

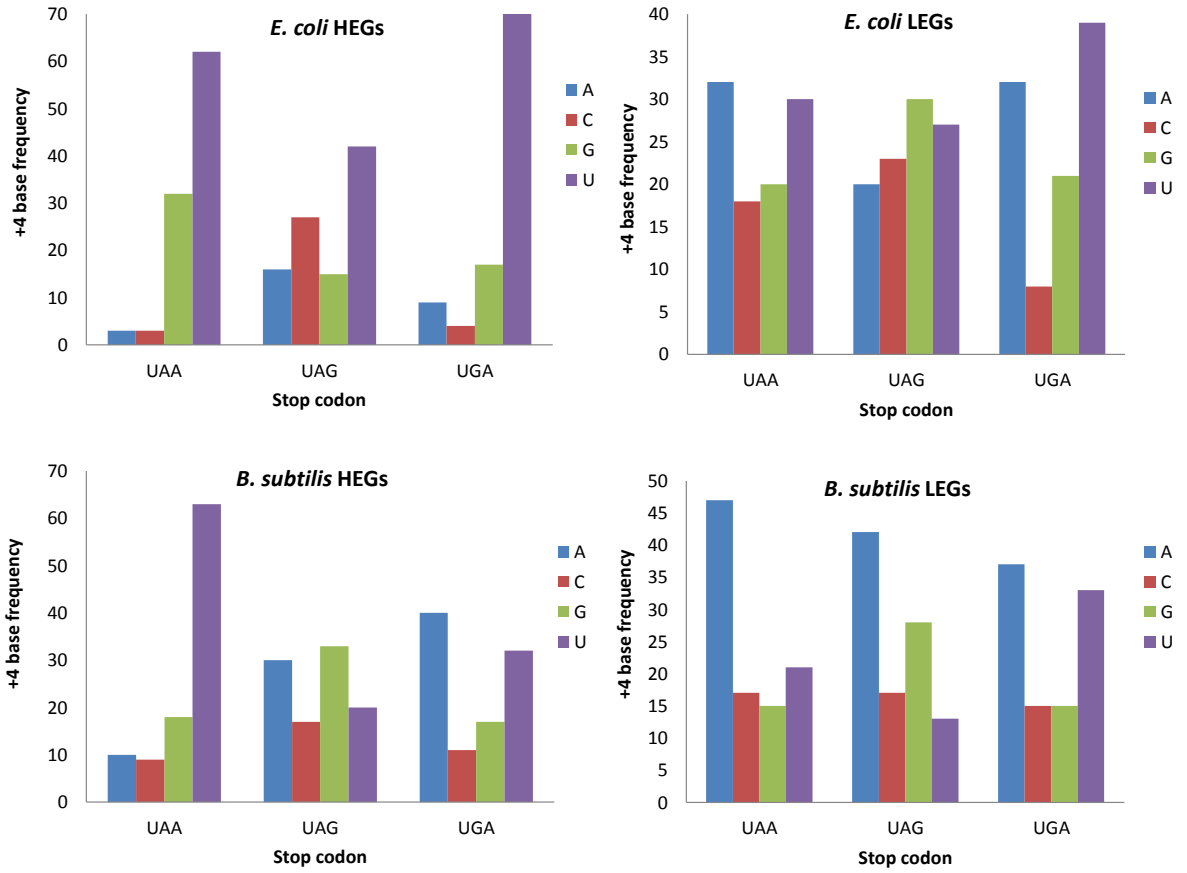

Fig. 3.2. Relationship between +4 nucleotide usage and stop codons in *E. coli* and *B. subtilis*, contrasting between 100 highly and 100 lowly expressed genes (HEGs and LEGs, respectively) for each stop codon, respectively. Only non-pseudo and non-hypothetical genes are used.

All five species belonging to Betaproteobacteria (Fig. 3.1) share the *E. coli* pattern, i.e., +4U overrepresented in all stop codons in HEGs relative to LEGs (Table 3.2) and all seven species belonging to Cyanobacteria and Bacilli share the *B. subtilis* pattern, with strong overrepresentation of +4U in UAA-ending HEGs relative UAA-ending LEGs, but no clear pattern involving UAG and UGA codons (Table 3.2). Species with the *E. coli* pattern generally have far more RF2 than RF1, whereas those with the *B. subtilis* pattern have more RF1 than RF2 (Wei *et al.* 2016). It is likely that +4U increases termination efficiency for RF2 decoding UAA

and UGA, whereas RF1 may benefit from +4U only in decoding UAA. This would suggest that overuse of TAA by HEGs would result in overuse of +4U. This is indeed the case. The species with overrepresented +4U in HEGs, i.e., the seven species belonging Cyanobacteria and Bacilli and the five species belonging to Betaproteobacteria indeed all have UAA overrepresented in HEGs than LEGs.

The usage of +4U changes with genomic GC% (Fig. 3.3), with the overuse of +4U most pronounced in UAA-ending genes with the proportion of genomic GC from low to slightly higher than 50% (Fig. 3.3). Based on the Wilcoxon rank sum test with continuity correction, the difference in +4U usage between HEGs and LEGs is significant in UAA-ending genes (P = 0.000327, two-tailed test), but not significant in UAG-ending genes (P = 0.2538, two tailed test) and UGA-ending genes (P = 0.0795, two tailed test). However, four species with high genomic GC contents (>58.7%): *Mycobacterium tuberculosis*, *Deinococcus deserti*, *Desulfovibrio vulgaris* and *Acidithiobacillus ferrooxidans*, do not have higher P_U in HEGs than LEG do not have higher PU in HEGs than LEG (Wilcoxon rank sum test: P = 0.706, two-tailed test; Table 3.2). These four species, being GC-rich, have few UAA-ending genes, which is consistent with our previous interpretation from Figs. 2-3 that UAA-ending genes are the main driver for increased +4U. Few UAA-ending genes implies little selection driving up +4U usage.

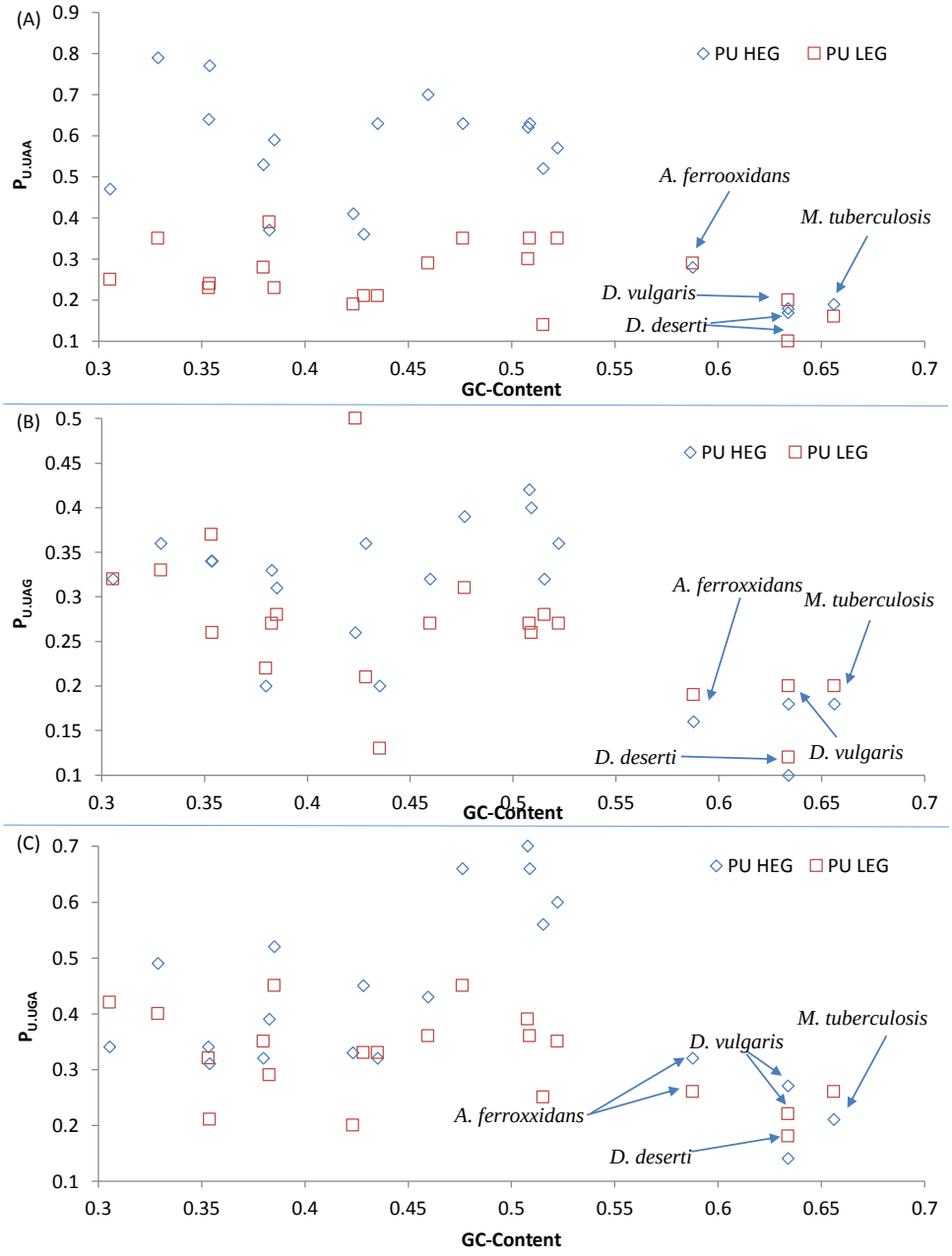

Fig. 3.3. Relationship between genomic GC-content (proportion of G and C in the genome) and +4U usage measured as the proportion of +4U at the +4 site and designated by $P_{U.UAA}$ (A), $P_{U.UAG}$ (B) and $P_{U.UGA}$ (C), respectively, for the three stop codons in 19 bacterial species. 100 HEGs and 100 LEGs are used for each stop codon. Only non-pseudo, non-hypothetical genes are used. The four species with high GC-contents (>58%) are indicated.

Table 3.2. The usage of +4U (P_U) in 100 non-pseudo and non-hypothetical UAA, UAG and UGA-ending HEGs and LEGs, ranked by I_{TE}, in 19 bacterial species, together with the species' accession number and genomic GC content. A value of 0.26 under UAA/$P_{U.HEG}$ means 26 genes out of 100 UAA-ending HEGs have +4U. Horizontal lines delineate major taxonomic groups corresponding to Fig. 3.1.

			UAA		UAG		UGA	
SPECIES NAME	ACCESSION	GC%	$P_{U.HEG}$	$P_{U.LEG}$	$P_{U.HEG}$	$P_{U.LEG}$	$P_{U.HEG}$	$P_{U.LEG}$
Microcystis aeruginosa	NC_010296	42.331	0.41	0.19	0.26	0.5	0.33	0.2
Bacillus anthracis	NC_005945	35.379	0.77	0.24	0.34	0.26	0.31	0.21
Bacillus subtilis	NC_000964	43.514	0.63	0.21	0.2	0.13	0.32	0.33
Staphylococcus aureus	NC_002758	32.878	0.79	0.35	0.36	0.33	0.49	0.40
Listeria monocytogenes	NC_003210	37.981	0.53	0.28	0.2	0.22	0.32	0.35
Streptococcus pyogenes	NC_002737	38.512	0.59	0.23	0.31	0.28	0.52	0.45
Lactococcus lactis	NC_002662	35.329	0.64	0.23	0.34	0.37	0.34	0.32
Deinococcus deserti	NC_002937	63.388	0.17	0.10	0.10	0.12	0.14	0.18
Bacteroides thetaiotaomicron	NC_004663	42.837	0.66	0.33	0.36	0.21	0.45	0.33
Escherichia coli	NC_000913	50.791	0.62	0.3	0.42	0.27	0.7	0.39
Salmonella enterica	NC_003197	52.222	0.57	0.35	0.36	0.27	0.6	0.35
Yersinia pestis	NC_003143	47.636	0.63	0.35	0.39	0.31	0.66	0.45
Shewanella oneidensis	NC_004347	45.961	0.7	0.29	0.32	0.27	0.43	0.36

Neisseria meningitidis	NC_003112	51.528	0.52	0.14	0.32	0.28	0.56	0.25
Legionella pneumophila	NC_002942	38.27	0.37	0.39	0.33	0.27	0.39	0.29
Acidithiobacillus ferrooxidans	NC_011761	58.773	0.28	0.29	0.16	0.19	0.32	0.26
Campylobacter jejuni	NC_002163	30.549	0.47	0.25	0.32	0.32	0.34	0.42
Desulfovibrio vulgaris	NC_002937	63.388	0.18	0.2	0.18	0.2	0.27	0.22
Mycobacterium tuberculosis	NC_000962	65.615	0.19	0.16	0.18	0.2	0.21	0.26

We investigated how stop codon and +4 nucleotide usage change with I_{TE} (a proxy of translation efficiency and gene expression) for three species (*E. coli*, *B. subtilis* and *D. vulgaris*) that appear to represent the three different patterns: i) +4U is over-represented in HEGs, ii) +4U is over-represented in only UAA-ending HEGs, and iii) +4U is not over-represented, respectively. We binned all non-pseudo, non-hypothetical CDSs into 10 gene groups ranked by I_{TE}. I_{TE} is significantly and positively correlated with P_{UAA} in all three species (*E. coli*: R^2 = 0.935, P < 0.0001; *B. subtilis*: P_{UAA}: R^2 = 0.884, P < 0.0001; *D. vulgaris*: R^2 = 0.644, P = 0.00518; Fig. 3.4), even when UAA accounts for a small fraction of the stop codons. This is consistent with a previous study (Wei *et al.* 2016) showing UAA to be always preferred by HEGs. Furthermore, I_{TE} was significantly positively correlated with P_U in *E. coli* (R^2 = 0.9149, P < 0.0001) and in *B. subtilis* (R^2 = 0.773, P < 0.001), but not in *D. vulgaris* (R^2 = 0.0098, P = 0.786). No significant relationship between other nucleotide at the +4 site and I_{TE} is observed (Fig. 3.4). To show that the U bias exists only at the +4 site, we randomly shuffled 20 nucleotides in the 5' untranslated region (5' UTR) for all 4140 non-pseudo, non-hypothetical E. coli genes, and the significant correlation between ITE and PU disappeared (R2 = 0.0301, P = 0.632; Supplementary Fig. S1). To validate that other metrics of codon usage bias return compatible results, we measured HEGs and LEGs by CAI (Supplementary Table S1); the two metrics (CAI, ITE) returns similar +4U usage (Wilcoxon rank sum test with continuity correction: P = 0.845, two tailed test).

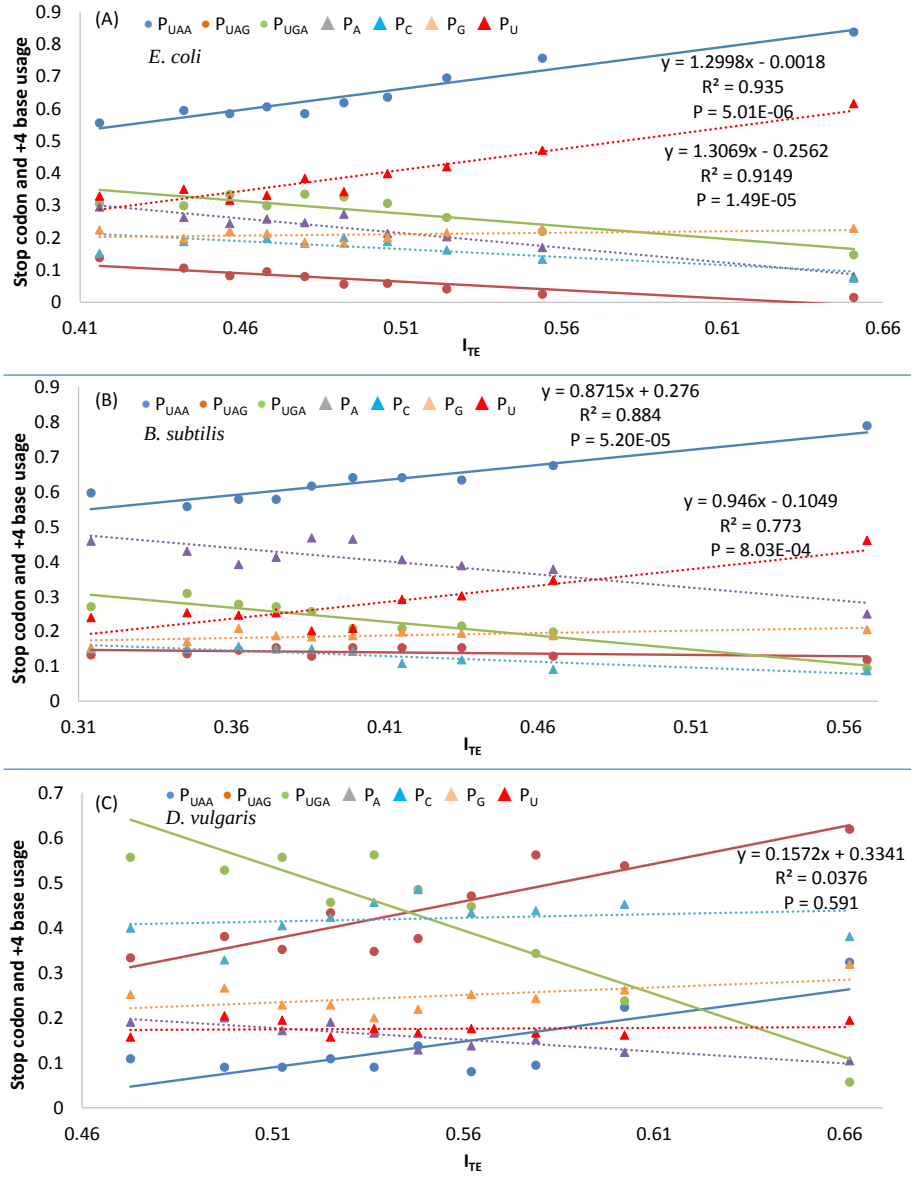

Fig. 3.4. Relationship between I_{TE} and usage of termination signals (stop codons and +4 bases), in *E. coli* (A), *B. subtilis* (B), and *D. vulgaris* (C). All non-pseudo, non-hypothetical CDSs were ranked by I_{TE} and binned into 10 sets, the stop codon usage and +4 base usage was obtained in each set. Stop codon usage (P_{UAA}, P_{UAG}, P_{UGA}) is represented by solid lines; +4 base usage (P_A, P_C, P_G, P_U) is represented by dotted lines.

The overuse of +4U in UAA-ending genes is also visible in the highly expressed 30S and 50S ribosomal protein genes (Fig. 3.5A, Spearman rank correlation = 0.8385, d.f. = 25, P<0.0001), and the fitted non-linear curve (Fig. 3.5A) accounts for 82.93% of the variation in $P_{U.UAA}$. There is no significant correlation between $P_{U.UAG}$ and P_{UAG} (Fig. 3.5B, $R^2 = 0.0032$, P = 0.8237), and a negative linear correlation between $P_{U.UGA}$ and P_{UGA} (Fig. 3.5C, $R^2 = 0.414$, P = 0.003968). Here, all 25 species (Fig. 3.1) were analyzed since ribosomal protein genes were considered. To alleviate the issue of data dependence due to shared ancestry between species (Fig. 3.1), we performed linear regression on Felsenstein's phylogeny-based independent contrasts (Felsenstein 1985), and the correlation between $P_{U.UAA}$ and P_{UAA} was still significant (R^2= 0.5819, P < 0.0001), and the result is consistent with bootstrapped trees or the tree reconstructed by using PhyPA (Xia 2016).

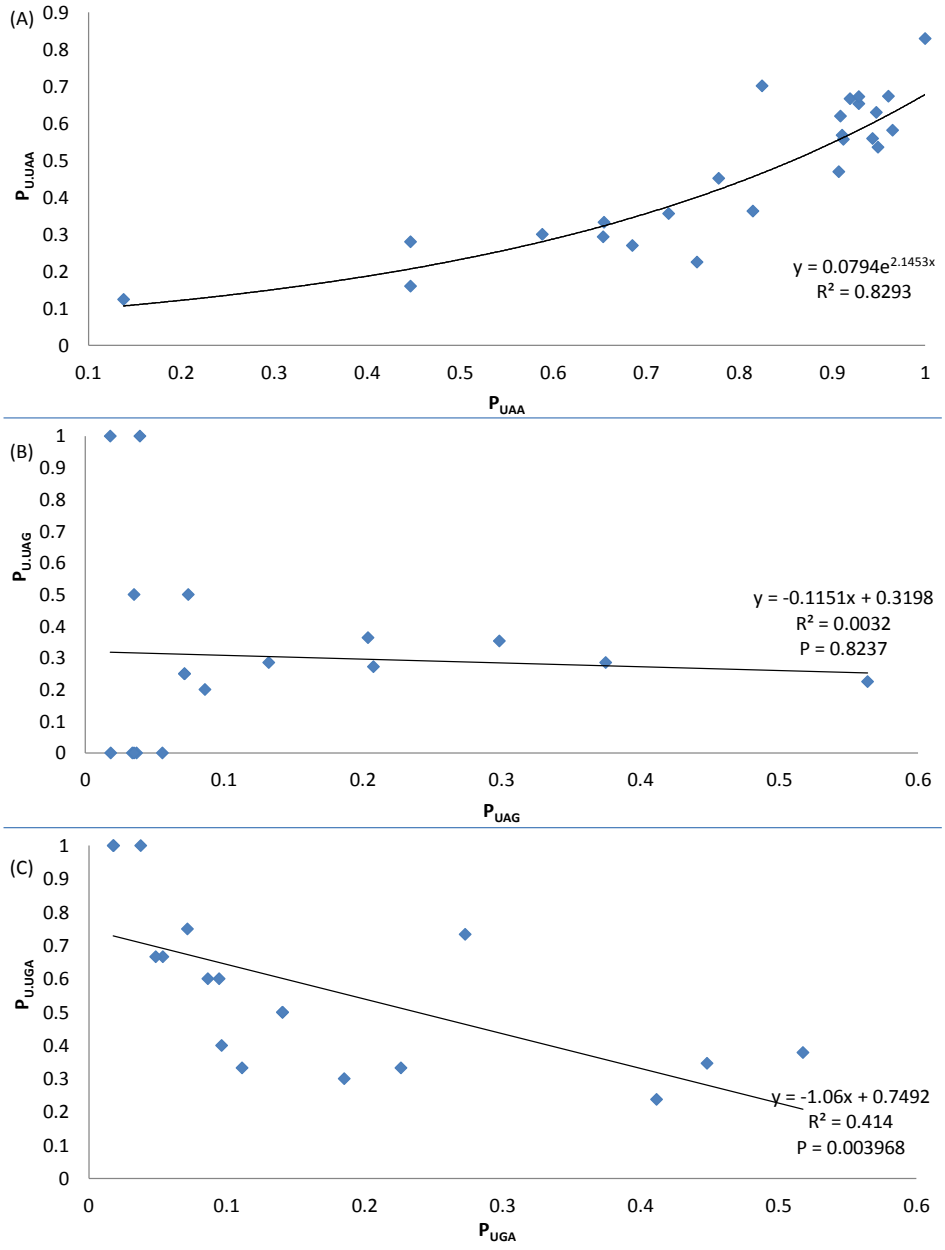

Fig. 3.5. Relationship between stop codon and +4 base usage, represented with regression between the proportions of stop codons (P_{UAA}, P_{UAG} and P_{UGA}) and proportion of their +4U ($P_{U.UAA}$, $P_{U.UAG}$ and $P_{U.UGA}$), and shown in (A), (B) and (C), respectively. Data from all 30S and 50S ribosomal protein genes in 25 bacterial species, excluding the data point if the stop codon usage is zero.

3.5.2 *Relationship between +4U usage and nc_tRNA abundance*

We have hypothesized that +4U reduce misreading of stop codons, especially UAA, by nc_tRNAs (Table 1). We used tRNA gene copy numbers as a proxy of tRNA abundance. This approach has been fruitful in a number of studies (Percudani *et al.* 1997; Duret and Mouchiroud 1999; Kanaya *et al.* 1999; Chithambaram *et al.* 2014b; Chithambaram *et al.* 2014d; Prabhakaran *et al.* 2014). We denoted nc_tRNA1, nc_tRNA2 and nc_tRNA3 as number of nc_tRNAs with a single mismatch at the first, second and third stop codon site, respectively. In each species, P_{nc_tRNA1} was calculated as the number of nc_tRNA1 copies divided by the total number of tRNA copies. In the 19 bacterial species, P_U in UAA-ending HEGs was significantly and positively correlated with P_{nc_tRNA}, the relationship being particularly strong nc_tRNAs with a single mismatch at the first stop codon site (Fig. 3.6). This positive correlation remains highly significant even after excluding nc_tRNA^{Gln} which is a key contribution to UAA read-through (Blanchet *et al.* 2014; Roy *et al.* 2015; Roy *et al.* 2016) ($R^2 = 0.517$, $P = 0.0005$, Fig. 3.6D). The correlation between P_U and P_{nc_tRNA} was, however, not significant in UAG and UGA-ending HEGs (Supplementary Fig. S2).

To alleviate data dependence due to shared ancestry, we performed regression on independent contrasts (Felsenstein 1985b) which showed significant correlation between P_U and P_{nc_tRNA1} ($R^2 = 0.349$, $P = 0.00985$) and between P_U and $P_{nc\_tRNA1-TCC}$ ($R^2 = 0.501$, $P = 0.00101$), but weak linear correlation between P_U and P_{nc_tRNA2} ($R^2 = 0.150$, $P = 0.112$) and P_{nc_tRNA3} ($R^2 = 0.233$, $P = 0.0424$).

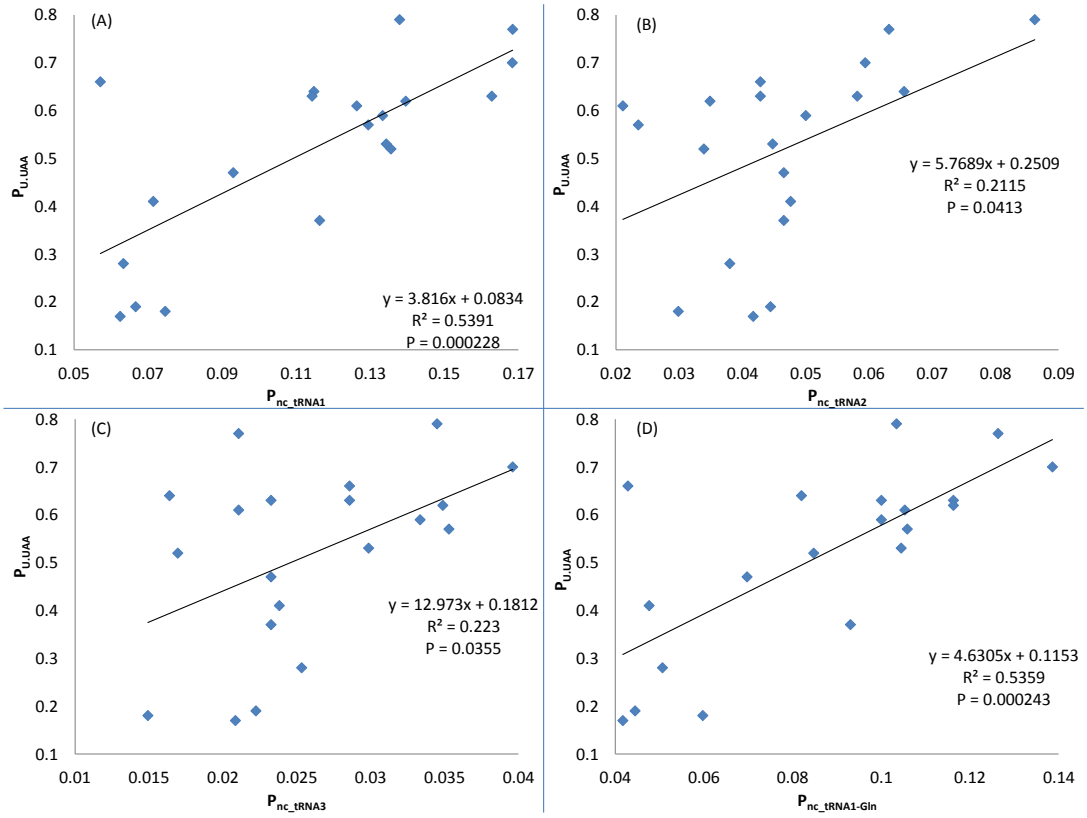

Fig. 3.6. Relationship between nc_tRNA abundance and +4U usage, represented by linear regression between 100 UAA-ending HEGs (highest I_{TE} scores) and abundance of UAA nc_tRNAs with a single mismatch at A) the first stop codon site, B) the second stop codon site, C) the third stop codon site, and D) the first stop codon site, omitting $tRNA^{Gln}$, 5'-TTG-3', in 19 bacterial species.

3.6 Discussion

UAA is consistently preferred stop codon in highly expressed genes in a diverse array of bacterial species (Wei *et al.* 2016), presumably because i) UAA can be decoded by both RF1 and RF2 (Scolnick *et al.* 1968; Milman *et al.* 1969; Nakamura *et al.* 1996), and ii) UAA has the least termination read-through (Parker 1989; Jorgensen *et al.* 1993; Meng *et al.* 1995; Cesar Sanchez *et al.* 1998; Tate *et al.* 1999; Dabrowski *et al.* 2015). Our study advanced these studies by showing i) +4U is strongly associated with UAA in HEGs relative to LEGs, ii) +4U usage increases with increasing number of nc_tRNAs, and iii) both UAA and +4U usage increases with

gene expressed measured by I_{TE}. Taken together, these findings suggest that +4U may enhance the UAA stop signal by reducing misreading by nc_tRNAs. This interpretation is consistent with read-through studies discussed previously and with the finding that termination suppression of stop codons was least efficient in the presence of +4U in *E. coli* (Kopelowitz *et al.* 1992). Consequently, the tetranucleotide UAAU is expected to represent the strongest termination signal a variety of bacterial lineages.

The interpretation above also explains why +4U is not over used in GC-rich species (Table 3.2 and Fig. 3.3) because these species have few genes ending with UAA. If +4U mainly enhances the termination signal of UAA against misreading by nc_tRNAs, the rarity of UAA-ending genes is naturally expected not to associate with overuse of +4U.

The importance of considering gene expression (or translation efficiency) in studying codon adaptation is highlighted by the fact that little +4U bias would be observed in the 19 species when all CDSs were considered (Supplementary Fig. S3) without contrasting between HEGs and LEGs. It is also important to study +4U bias separately for different stop codons because nucleotide distribution at +4 site is heterogeneous among genes ending with different stop codons (Fig. 3.2). Previous studies on termination read-through in yeast (Roy *et al.* 2015; Roy *et al.* 2016) and bacteria (Kramer and Farabaugh 2007) often did not take into consideration of all possible combination of stop codons and +4U nucleotide.

Our study also suggests phylogenetic inertia in the evolution of the stop codon decoding mechanism. For example, all five species in Betaproteobacteria exhibit very similar difference between HEGs and LEGs in +4U usage, so do the seven species belonging to the supercluster

including Cyanobacteria and Bacilli (Fig. 3.1). For this reason, phylogeny-based comparative methods are crucially important for proper assessment of statistical significance among variables.

It is interesting to note that UGA- and UAG-ending genes do not show the same strong preference of +4U observed in UAA-ending genes. Given that RF1 decodes UAA and UAG, and RF2 decodes UAA and UGA, it seems that RF1 must have different binding dynamics between UAA-ending and UAG ending genes, and RF2 between UAA-ending and UGA-ending genes. Structural studies (Matheisl *et al.* 2015; Svidritskiy *et al.* 2016; Tang *et al.* 2016) or cross-linking studies (Brown and Tate 1994; Tate *et al.* 1996; Poole *et al.* 1997; Poole *et al.* 1998) may shed light on the effect of +4U on UAA termination signal.

3.7 Acknowledgement

This study is funded by the Discovery Grant from Natural Science and Engineering Research Council of Canada to X.X. We thank J. Wang, J. Silke and C. Vlasschaert for discussion and comments, and the two reviewers for suggestions that have led to significant improvement of the manuscript.

4. Discussion and Conclusion

Understanding the translation process is crucial for efficient protein production. While translation initiation and elongation have been studied intensively, the inner workings of the translation termination mechanism remain obscure to us. More specifically, although stop codon and the +4 base have been recognized to affect termination read-through, there is inconsistencies among literatures in the attempt to identify the minor stop codon and the function of the +4 site. These issues prompted us to perform comparative genomics analyses to examine the role of stop

codons and the +4 site in bacterial translation termination efficiency using bioinformatics approaches.

In chapter two, we explain that stop codon usage is dictated by mutation bias and decoder abundance. We found that, in bacteria, UAA is a universal major stop codon because it has the most abundant decoders (decoded by RF1 and RF2), and it is always preferred in HEGs over LEGs. The usage of UAG and UGA, on the other hand, is dependent on the relative abundance of their respective decoders, RF1 and RF2. Moreover, usage of UGA, not UAG, is always avoided in HEGs when compared to LEGs. Thus, against the claims made by Korkmaz *et al.* (2014), we explain that UAG does not meet the two criteria of a minor stop codon. Furthermore, RF2 is significantly reduced in species with high GC contents; thus, we suggest that UGA is reassigned to a sense codon due to a significant decrease in RF2 abundance in GC-rich species.

In chapter three, we suggest that +4U increases UAA termination efficiency by reducing misreading of UAA by nc_tRNAs in bacteria. Our study is the first to suggest that +4U usage is dependent on the abundance of stop codon nc_tRNAs. Indeed, we found +4U consistently over-represented in HEGs in contrast to LEGs in species where UAA is the most abundant stop codon. However, in GC-rich species where UAG and UGA are overused, +4U usage in HEGs decreased. In addition, +4U usage in HEGs increases significantly with the usage of stop codon UAA, not UAG or UGA. Thus, our results suggest that +4U is a strong enhancer for the stop codon UAA, not UAG or UGA. Lastly, in HEGs, +4U usage also increased significantly with UAA nc_tRNA abundance, but not with UAG or UGA nc_tRNA abundance.

We demonstrated that UAA is the most efficient stop signal, but termination recognition goes beyond the stop codons, and our results suggest that the tetra-nucleotide UAA-U is the most

efficient termination signal. The studies in this thesis attempt to extend our current understanding in bacterial translation termination efficiency, and will hopefully pique the interest of researchers to examine, more rigorously, the possibility that recognition of termination signals by class I release factors is extended to the first 3' flanking base.

5. Supplementary content

Table S1. The usage of +4U in 100 non-pseudo, non-hypothetical HEGs and LEGs for each stop codon, ranked by CAI, in 25 bacterial species. Together with the number of 30S and 50S ribosomal protein genes in each species. The reference HEGs to compute CAI are the same 40 ribosomal protein genes used for .ITE reference files.

			UAA[b]		UAG		UGA	
SPECIES NAME	ACCESSION	N.R[a]	$P_{U.HEG}$	$P_{U.LEG}$	$P_{U.HEG}$	$P_{U.LEG}$	$P_{U.HEG}$	$P_{U.LEG}$
Bacillus anthracis	NC_005945	56	0.77	0.23	0.37	0.29	0.32	0.24
Bacillus subtilis	NC_000964	56	0.65	0.24	0.19	0.10	0.44	0.34
Bartonella henselae	NC_005956	55	0.25	0.35	0.21	0.19	0.39	0.35
Staphylococcus aureus	NC_002758	53	0.78	0.35	0.34	0.32	0.49	0.343
Listeria monocytogenes	NC_003210	57	0.52	0.29	0.17	0.23	0.36	0.36
Streptococcus pyrogenes	NC_002737	51	0.61	0.23	0.34	0.24	0.51	0.49
Lactococcus lactis	NC_002662	57	0.64	0.17	0.36	0.34	0.32	0.32
Leptospira interrogans	NC_005823	54	0.37	0.26	0.30	0.34	0.41	0.33
Deinococcus deserti	NC_002937	51	0.17	0.14	0.12	0.11	0.12	0.17
Bacteriodes thetaiotaomicron	NC_004663	55	0.62	0.29	0.34	0.21	0.46	0.28
Escherichia coli	NC_000913	57	0.62	0.3	0.36	0.24	0.72	0.30
Salmonella enterica	NC_003197	59	0.60	0.35	0.36	0.25	0.69	0.30

Yesinia pestis	NC_003143	56	0.63	0.39	0.33	0.31	0.65	0.45
Shewanella oneidensis	NC_004347	62	0.71	0.27	0.31	0.27	0.49	0.328
Neisseria meningitidis	NC_003112	57	0.52	0.21	0.33	0.32	0.59	0.27
Legionella pneumophila	NC_002942	54	0.37	0.39	0.29	0.29	0.42	0.26
Acidithiobacillus ferrooxidans	NC_011761	56	0.15	0.30	0.15	0.20	0.24	0.31
Campylobacter jejuni	NC_002163	54	0.45	0.30	0.29	0.28	0.34	0.43
Desulfovibrio vulgaris	NC_002937	56	0.19	0.18	0.14	0.11	0.37	0.21
Mycobacterium tuberculosis	NC_000962	58	0.17	0.14	0.16	0.17	0.20	0.24
Microcystis aeruginosa	NC_010296.1	53	0.39	0.19	0.34	0.37	0.35	0.31
Helicobacter pylori	NC_000915	54	0.21	0.20	0.22	0.34	0.41	0.45
Pseudomonas aeruginosa	NC_002516	58	0.21	0.22	0.11	0.13	0.37	0.23
Shigella flexneri	NC_004337	53	0.62	0.17	0.47	0.20	0.63	0.37
Synechocystis sp.	NC_017277	52	0.32	0.26	0.35	0.18	0.38	0.33

a Number of 30S and 50S ribosomal protein genes.
b The +4U usage in 100 UAA-ending HEGs and LEGs (highest and lowest CAI scores, respectively).

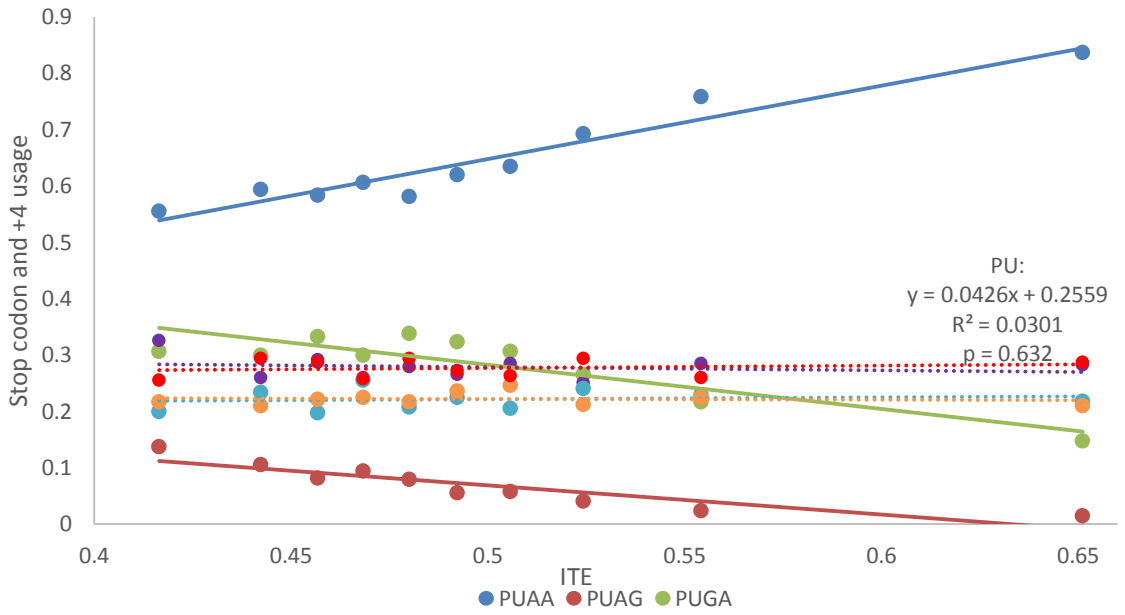

Fig. S1. Relationship between ITE and usage of termination signals (stop codons and +4 bases), in *E. coli*. All non-pseudo, non-hypothetical CDSs were ranked by ITE and binned into 10 sets. Twenty sequences in the 5' UTR were randomly shuffled using Sequence Manipulation Suite: Shuffle DNA (Stothard 2000). The stop codon usage and +4 base usage was obtained in each set. Stop codon usage (PUAA, PUAG, PUGA) is represented by solid lines; +4 base usage (PA, PC, PG, PU) is represented by dotted lines.

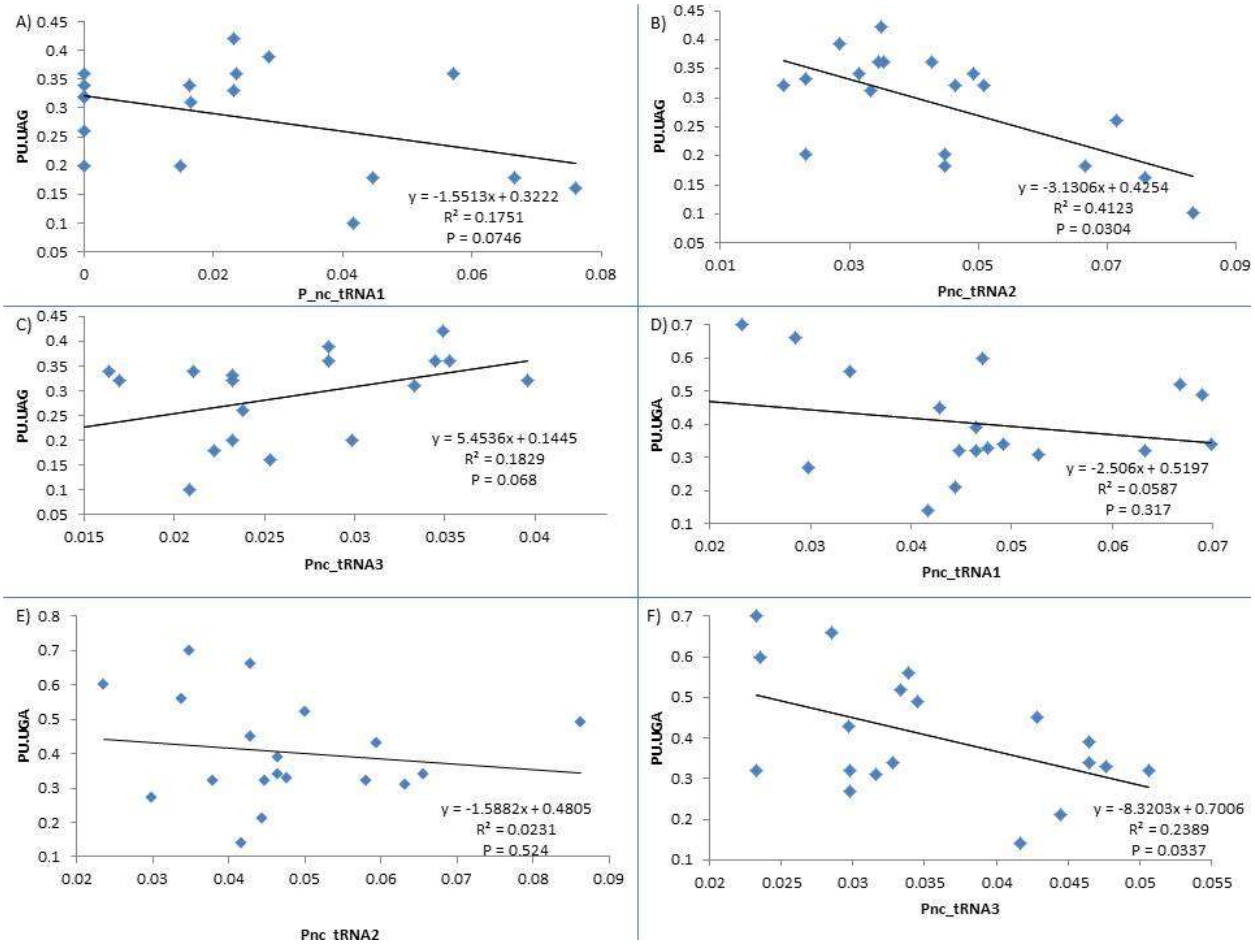

Fig. S2. Relationship between nc_tRNA abundance and +4U usage, represented by linear regression between 100 UAA-ending HEGs (highest ITE scores) and abundance of UAG nc_tRNAs with a single mismatch at A) the first stop codon site, B) the second stop codon site, C) the third stop codon site, and abundance of UGA nc_tRNAs with a single mismatch at D) the first stop codon site, E) the second stop codon site, F) the third stop codon site.

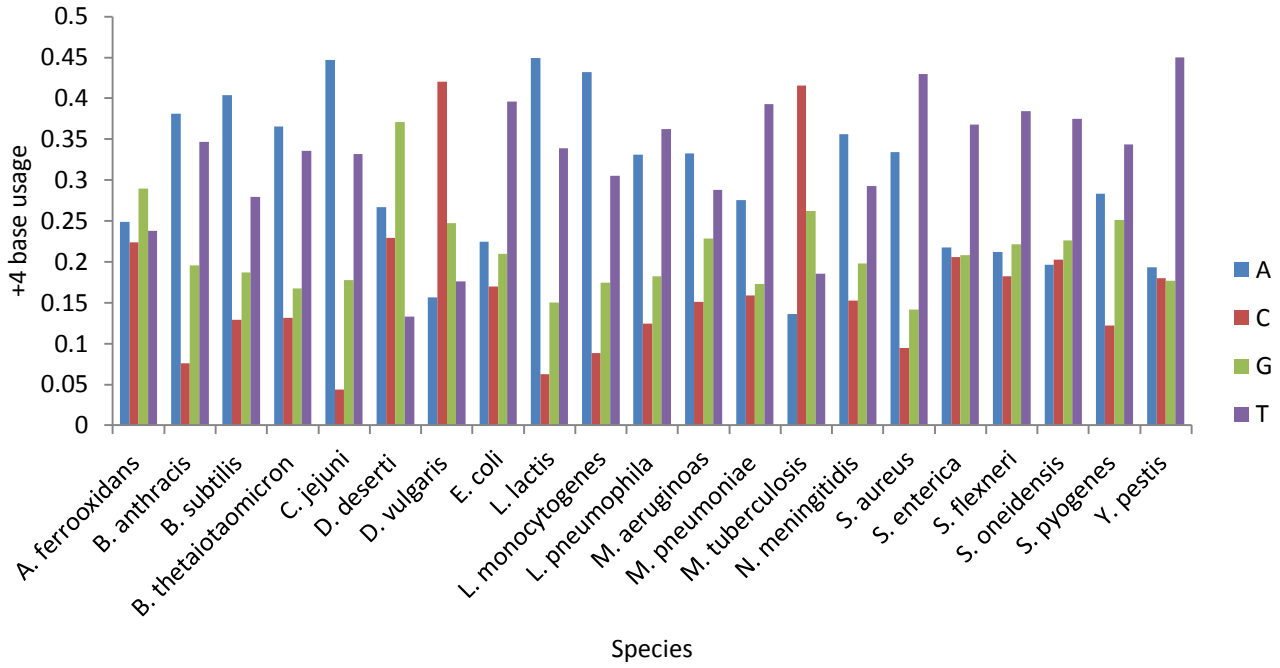

Fig. 3. The +4 base usage in all non-pseudo, non-hypothetical CDSs in 19 bacterial species (Table 3.2).

6. References

Adamski, F. M., K. K. Mccaughan, F. Jorgensen, C. G. Kurland and W. P. Tate, 1994 The concentration of polypeptide chain release factors 1 and 2 at different growth rates of *Escherichia coli*. J Mol Biol 238**:** 302-308.

Akashi, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136**:** 927-935.

Akashi, H., 2003 Translational selection and yeast proteome evolution. Genetics 164**:** 1291-1303.

Akashi, H., and A. Eyre-Walker, 1998 Translational selection and molecular evolution. Curr Opin Genet Dev 8**:** 688-693.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tool. J Mol Biol 215**:** 403-410.

Andersson, G. E., and C. G. Kurland, 1991 An extreme codon preference strategy: codon reassignment. Mol Biol Evol 8**:** 530-544.

Baranov, P. V., R. F. Gesteland and J. F. Atkins, 2002 Release factor 2 frameshifting sites in different bacteria. EMBO Rep 3**:** 373-377.

Beier, H., and M. Grimm, 2001 Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. Nucleic Acids Res 29**:** 4767-4782.

Beznoskova, P., S. Gunisova and L. S. Valasek, 2016a Rules of UGA-N decoding by near-cognate tRNAs and analysis of readthrough on short uORFs in yeast. RNA 22**:** 456-466.

Beznoskova, P., S. Gunisova and L. S. Valasek, 2016b Rules of UGA-N decoding by near-cognate tRNAs and analysis of readthrough on short uORFs in yeast. RNA.

Beznoskova, P., S. Wagner, M. E. Jansen, T. Von Der Haar and L. S. Valasek, 2015 Translation initiation factor *eIF3* promotes programmed stop codon readthrough. Nucleic Acids Res 43**:** 5099-5111.

Björnsson, A., and L. A. Isaksson, 1996 Accumulation of a mRNA decay intermediate by ribosomal pausing at a stop codon. Nucleic Acids Research 24**:** 1753-1757.

Blanchet, S., D. Cornu, M. Argentini and O. Namy, 2014 New insights into the incorporation of natural suppressor tRNAs at stop codons in *Saccharomyces cerevisiae*. Nucleic Acids Res 42**:** 10061-10072.

Bossi, L., 1983 Context effects: translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. J Mol Biol 164**:** 73-87.

Bossi, L., and J. R. Ruth, 1980 The influence of codon context on genetic code translation. Nature 286**:** 123-127.

Brown, C. M., P. A. Stockwell, C. N. Trotman and W. P. Tate, 1990 The signal for the termination of protein synthesis in procaryotes. Nucleic Acids Res 18**:** 2079-2086.

Brown, C. M., and W. P. Tate, 1994 Direct recognition of mRNA stop signals by *Escherichia coli* polypeptide chain release factor two. J Biol Chem 269**:** 33164-33170.

Bulygin, K. N., M. N. Repkova, A. G. Ven'yaminova, D. M. Graifer, G. G. Karpova *et al.*, 2002 Positioning of the mRNA stop signal with respect to polypeptide chain release factors and ribosomal proteins in 80S ribosomes. FEBS Lett 514**:** 96-101.

Carullo, M., and X. Xia, 2008 An extensive study of mutation and selection on the wobble nucleotide in tRNA anticodons in fungal mitochondrial genomes. J Mol Evol 66**:** 484-493.

Cassan, M., and J. P. Rousset, 2001 UAG readthrough in mammalian cells: effect of upstream and downstream stop codon contexts reveal different signals. BMC Mol Biol 2**:** 3.

Cesar Sanchez, J., G. Padron, H. Santana and L. Herrera, 1998 Elimination of an *HuIFN alpha 2b* readthrough species, produced in *Escherichia coli*, by replacing its natural translational stop signal. J Biotechnol 63**:** 179-186.

Charneski, C. A., F. Honti, J. M. Bryant, L. D. Hurst and E. J. Feil, 2011 Atypical at skew in Firmicute genomes results from selection and not from mutation. PLoS Genet 7**:** e1002283.

Chavancy, G., A. Chevallier, A. Fournier and J. P. Garel, 1979 Adaptation of iso-tRNA concentration to mRNA codon frequency in the eukaryote cell. Biochimie 61**:** 71-78.
Chithambaram, S., R. Prabhakaran and X. Xia, 2014a Differential codon adaptation between dsDNA and ssDNA phages in *Escherichia coli*. Mol Biol Evol 31**:** 1606-1617.
Chithambaram, S., R. Prabhakaran and X. Xia, 2014b Differential Codon Adaptation between dsDNA and ssDNA Phages in *Escherichia coli*. Molecular Biology and Evolution 31**:** 1606-1617.
Chithambaram, S., R. Prabhakaran and X. Xia, 2014c The Effect of Mutation and Selection on Codon Adaptation in *Escherichia coli* Bacteriophage. Genetics 197**:** 301-315.
Chithambaram, S., R. Prabhakaran and X. Xia, 2014d The Effect of Mutation and Selection on Codon Adaptation in *Escherichia coli* Bacteriophage. Genetics 197**:** 301-315.
Craigen, W. J., and C. T. Caskey, 1986 Expression of peptide chain release factor 2 requires high-efficiency frameshift. Nature 322**:** 273-275.
Craigen, W. J., R. G. Cook, W. P. Tate and C. T. Caskey, 1985 Bacterial peptide chain release factors: conserved primary structure and possible frameshift regulation of release factor 2. Proc Natl Acad Sci U S A 82**:** 3616-3620.
Cridge, A. G., L. L. Major, A. A. Mahagaonkar, E. S. Poole, L. A. Isaksson *et al.*, 2006 Comparison of characteristics and function of translation termination signals between and within prokaryotic and eukaryotic organisms. Nucleic Acids Res 34**:** 1959-1973.
Dabrowski, M., Z. Bukowy-Bieryllo and E. Zietkiewicz, 2015 Translational readthrough potential of natural termination codons in eucaryotes--The impact of RNA sequence. RNA Biol 12**:** 950-958.
Davies, J., D. S. Jones and H. G. Khorana, 1966 A further study of misreading of codons induced by streptomycin and neomycin using ribopolynucleotides containing two nucleotides in alternating sequence as templates. Journal of Molecular Biology 18**:** 48-57.
Desper, R., and O. Gascuel, 2002 Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol 9**:** 687-705.
Desper, R., and O. Gascuel, 2004 Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol Biol Evol 21**:** 587-598.
Donly, B. C., C. D. Edgar, F. M. Adamski and W. P. Tate, 1990 Frameshift autoregulation in the gene for *Escherichia coli* release factor 2: partly functional mutants result in frameshift enhancement. Nucleic Acids Res 18**:** 6517-6522.
Dos Reis, M., R. Savva and L. Wernisch, 2004 Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32**:** 5036-5044.
Duret, L., and D. Mouchiroud, 1999 Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc Natl Acad Sci U S A 96**:** 4482-4487.
Engelberg-Kulka, H., 1981 UGA suppression by normal tRNA Trp in *Escherichia coli*: codon context effects. Nucleic Acids Res 9**:** 983-991.
Eyre-Walker, A., and M. Bulmer, 1995 Synonymous substitution rates in enterobacteria. Genetics 140**:** 1407-1412.
Felsenstein, J., 1985a Phylogenies and the Comparative Method. The American Naturalist 125**:** 1-15.
Felsenstein, J., 1985b Phylogenies and the comparative method. Amer. Nat. 125**:** 1-15.
Felsenstein, J., 1989 PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5**:** 164-166.
Freistroffer, D. V., M. Y. Pavlov, J. Macdougall, R. H. Buckingham and M. Ehrenberg, 1997 Release factor RF3 in E.coli accelerates the dissociation of release factors RF1 and RF2 from the ribosome in a GTP-dependent manner. EMBO J 16**:** 4126-4133.
Geller, A. I., and A. Rich, 1980 A UGA termination suppression tRNATrp active in rabbit reticulocytes. Nature 283**:** 41-46.
Gouy, M., and C. Gautier, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Research 10**:** 7055-7074.

Guerdoux-Jamet, P., A. Henaut, P. Nitschke, J. L. Risler and A. Danchin, 1997 Using codon usage to predict genes origin: is the *Escherichia coli* outer membrane a patchwork of products from different genomes? DNA Res 4**:** 257-265.

Guindon, S., and O. Gascuel, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52**:** 696-704.

Haas, J., E. C. Park and B. Seed, 1996 Codon usage limitation in the expression of HIV-1 envelope glycoprotein. Curr Biol 6**:** 315-324.

Hannig, G., and S. C. Makrides, 1998 Strategies for optimizing heterologous protein expression in Escherichia coli. Trends Biotechnol 16**:** 54-60.

Higgs, P. G., and W. Ran, 2008 Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol 25**:** 2279-2291.

Ikemura, T., 1981 Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151**:** 389-409.

Ingolia, N. T., 2014 Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet 15**:** 205-213.

Ingolia, N. T., S. Ghaemmaghami, J. R. Newman and J. S. Weissman, 2009 Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324**:** 218-223.

Ito, K., M. Uno and Y. Nakamura, 2000 A tripeptide 'anticodon' deciphers stop codons in messenger RNA. Nature 403**:** 680-684.

Jia, W., and P. G. Higgs, 2008 Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. Mol Biol Evol 25**:** 339-351.

Jin, H., A. Bjornsson and L. A. Isaksson, 2002 Cis control of gene expression in *E.coli* by ribosome queuing at an inefficient translational stop signal. Embo j 21**:** 4357-4367.

Jorgensen, F., F. M. Adamski, W. P. Tate and C. G. Kurland, 1993 Release factor-dependent false stops are infrequent in *Escherichia coli*. J Mol Biol 230**:** 41-50.

Jungreis, I., M. F. Lin, R. Spokony, C. S. Chan, N. Negre *et al.*, 2011 Evidence of abundant stop codon readthrough in *Drosophila* and other metazoa. Genome Res 21**:** 2096-2113.

Kanaya, S., Y. Yamada, Y. Kudo and T. Ikemura, 1999 Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtili*s tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238**:** 143-155.

Katoh, K., G. Asimenos and H. Toh, 2009 Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol 537**:** 39-64.

Kopelowitz, J., C. Hampe, R. Goldman, M. Reches and H. Engelberg-Kulka, 1992 Influence of codon context on UGA suppression and readthrough. J Mol Biol 225**:** 261-269.

Korkmaz, G., M. Holm, T. Wiens and S. Sanyal, 2014 Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. J Biol Chem 289**:** 30334-30342.

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

Kurland, C. G., 1987 Strategies for efficiency and accuracy in gene expression. Trends in Biochemical Sciences 12**:** 126-128.

Lobry, J. R., and N. Sueoka, 2002 Asymmetric directional mutation pressures in bacteria. Genome Biol 3**:** Research0058.

Loughran, G., M. Y. Chou, I. P. Ivanov, I. Jungreis, M. Kellis *et al.*, 2014 Evidence of efficient stop codon readthrough in four mammalian genes. Nucleic Acids Res 42**:** 8928-8938.

Manuvakhova, M., K. Keeling and D. M. Bedwell, 2000 Aminoglycoside antibiotics mediate context-dependent suppression of termination codons in a mammalian translation system. Rna 6**:** 1044-1055.

Marin, A., and X. Xia, 2008 GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. J Theor Biol 253**:** 508-513.

Matheisl, S., O. Berninghausen, T. Becker and R. Beckmann, 2015 Structure of a human translation termination complex. Nucleic Acids Res 43**:** 8615-8626.
Matsugi, J., and K. Murao, 1999 Search for a selenocysteine tRNA in *Bacillus subtilis*. Nucleic Acids Symp Ser**:** 209-210.
Matsugi, J., and K. Murao, 2000 A study of the method to pick up a selenocysteine tRNA in *Bacillus subtilis*. Nucleic Acids Symp Ser**:** 149-150.
Mcpherson, D. T., 1988 Codon preference reflects mistranslational constraints: a proposal. Nucleic Acids Res 16**:** 4111-4120.
Meng, S. Y., J. O. Hui, M. Haniu and L. B. Tsai, 1995 Analysis of translational termination of recombinant human methionyl-neurotrophin 3 in *Escherichia coli*. Biochem Biophys Res Commun 211**:** 40-48.
Miller, J. H., and A. M. Albertini, 1983 Effects of surrounding sequence on the suppression of nonsense codons. J Mol Biol 164**:** 59-71.
Milman, G., J. Goldstein, E. Scolnick and T. Caskey, 1969 Peptide chain termination. 3. Stimulation of in vitro termination. Proc Natl Acad Sci U S A 63**:** 183-190.
Mora, L., V. Heurgue-Hamard, M. De Zamaroczy, S. Kervestin and R. H. Buckingham, 2007 Methylation of bacterial release factors RF1 and RF2 is required for normal translation termination in vivo. J Biol Chem 282**:** 35638-35645.
Moriyama, E. N., and D. L. Hartl, 1993 Codon usage bias and base composition of nuclear genes in *Drosophila*. Genetics 134**:** 847-858.
Moriyama, E. N., and J. R. Powell, 1997 Codon usage bias and tRNA abundance in *Drosophila*. J Mol Evol 45**:** 514-523.
Mottagui-Tabar, S., M. F. Tuite and L. A. Isaksson, 1998 The influence of 5' codon context on translation termination in *Saccharomyces cerevisiae*. Eur J Biochem 257**:** 249-254.
Muto, A., and S. Osawa, 1987 The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A 84**:** 166-169.
Nakamura, Y., K. Ito and L. A. Isaksson, 1996 Emerging understanding of translation termination. Cell 87**:** 147-150.
Namy, O., I. Hatin and J. P. Rousset, 2001 Impact of the six nucleotides downstream of the stop codon on translation termination. EMBO Rep 2**:** 787-793.
Ngumbela, K. C., K. P. Ryan, R. Sivamurthy, M. A. Brockman, R. T. Gandhi *et al.*, 2008 Quantitative Effect of Suboptimal Codon Usage on Translational Efficiency of mRNA Encoding HIV-1 *gag* in Intact T Cells. PLoS ONE 3**:** e2356.
Nikbakht, H., X. Xia and D. A. Hickey, 2014 The evolution of genomic GC content undergoes a rapid reversal within the genus Plasmodium. Genome 57**:** 507-511.
Nilsson, M., and M. Ryden-Aulin, 2003 Glutamine is incorporated at the nonsense codons UAG and UAA in a suppressor-free *Escherichia coli* strain. Biochim Biophys Acta 1627**:** 1-6.
Ogawa, A., S. Sando and Y. Aoyama, 2006 Termination-free prokaryotic protein translation by using anticodon-adjusted E. coli tRNASer as unified suppressors of the UAA/UGA/UAG stop codons. Read-through ribosome display of full-length DHFR with translated UTR as a buried spacer arm. Chembiochem 7**:** 249-252.
Osawa, S., and T. H. Jukes, 1989 Codon reassignment (codon capture) in evolution. J Mol Evol 28**:** 271-278.
Palidwor, G. A., T. J. Perkins and X. Xia, 2010 A general model of codon bias due to GC mutational bias. PLoS One 5**:** 0013431.
Parker, J., 1989 Errors and alternatives in reading the universal genetic code. Microbiol Rev 53**:** 273-298.
Percudani, R., A. Pavesi and S. Ottonello, 1997 Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J Mol Biol 268**:** 322-330.
Poole, E. S., R. Brimacombe and W. P. Tate, 1997 Decoding the translational termination signal: the polypeptide chain release factor in *Escherichia coli* crosslinks to the base following the stop codon. Rna 3**:** 974-982.

Poole, E. S., C. M. Brown and W. P. Tate, 1995 The identity of the base following the stop codon determines the efficiency of in vivo translational termination in *Escherichia coli*. Embo j 14**:** 151-158.
Poole, E. S., L. L. Major, S. A. Mannering and W. P. Tate, 1998 Translational termination in *Escherichia coli*: three bases following the stop codon crosslink to release factor 2 and affect the decoding efficiency of UGA-containing signals. Nucleic Acids Res 26**:** 954-960.
Povolotskaya, I. S., F. A. Kondrashov, A. Ledda and P. K. Vlasov, 2012 Stop codons in bacteria are not selectively equivalent. Biol Direct 7**:** 1745-6150.
Prabhakaran, R., S. Chithambaram and X. Xia, 2014 Aeromonas phages encode tRNAs for their overused codons. Int J Comput Biol Drug Des 7**:** 168-182.
Prabhakaran, R., S. Chithambaram and X. Xia, 2015 *Escherichia coli* and Staphylococcus phages: effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. J Gen Virol 96**:** 1169-1179.
Pundir, S., M. J. Martin and C. O'donovan, 2016 UniProt Tools. Curr Protoc Bioinformatics 53**:** 1.29.21-21.29.15.
Rice, P., I. Longden and A. Bleasby, 2000 EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16**:** 276-277.
Robinson, M., R. Lilley, S. Little, J. S. Emtage, G. Yarranton *et al.*, 1984 Codon usage can affect efficiency of translation of genes in *Escherichia coli*. Nucleic Acids Research 12**:** 6663-6671.
Roth, J. R., 1970 UGA nonsense mutations in Salmonella typhimurium. J Bacteriol 102**:** 467-475.
Roy, B., W. J. Friesen, Y. Tomizawa, J. D. Leszyk, J. Zhuo *et al.*, 2016 Ataluren stimulates ribosomal selection of near-cognate tRNAs to promote nonsense suppression. Proceedings of the National Academy of Sciences of the United States of America 113**:** 12508-12513.
Roy, B., J. D. Leszyk, D. A. Mangus and A. Jacobson, 2015 Nonsense suppression by near-cognate tRNAs employs alternative base pairing at codon positions 1 and 3. Proceedings of the National Academy of Sciences of the United States of America 112**:** 3038-3043.
Rudner, R., J. D. Karkas and E. Chargaff, 1968 Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. Proc Natl Acad Sci U S A 60**:** 921-922.
Ryden, S. M., and L. A. Isaksson, 1984 A temperature-sensitive mutant of *Escherichia coli* that shows enhanced misreading of UAG/A and increased efficiency for some tRNA nonsense suppressors. Mol Gen Genet 193**:** 38-45.
Sambrook, J. F., D. P. Fan and S. Brenner, 1967 A strong suppressor specific for UGA. Nature 214**:** 452-453.
Scarlett, D. J., K. K. Mccaughan, D. N. Wilson and W. P. Tate, 2003 Mapping functionally important motifs SPF and GGQ of the decoding release factor RF2 to the *Escherichia coli* ribosome by hydroxyl radical footprinting. Implications for macromolecular mimicry and structural changes in RF2. J Biol Chem 278**:** 15095-15104.
Scolnick, E., R. Tompkins, T. Caskey and M. Nirenberg, 1968 Release factors differing in specificity for terminator codons. Proc Natl Acad Sci U S A 61**:** 768-774.
Scolnick, E. M., and C. T. Caskey, 1969 Peptide chain termination, V. The role of release factors in mRNA terminator codon recognition. Proceedings of the National Academy of Sciences of the United States of America 64**:** 1235-1241.
Sengupta, S., and P. G. Higgs, 2005 A Unified Model of Codon Reassignment in Alternative Genetic Codes. Genetics 170**:** 831-840.
Sengupta, S., X. Yang and P. G. Higgs, 2007 The Mechanisms of Codon Reassignments in Mitochondrial Genetic Codes. Journal of Molecular Evolution 64**:** 662-688.
Sharp, P. M., and M. Bulmer, 1988 Selective differences among translation termination codons. Gene 63**:** 141-145.
Sharp, P. M., and W. H. Li, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24**:** 28-38.

Sharp, P. M., and W. H. Li, 1987 The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15**:** 1281-1295.
Sorensen, M. A., C. G. Kurland and S. Pedersen, 1989 Codon usage determines translation rate in *Escherichia coli*. J Mol Biol 207**:** 365-377.
Stoletzki, N., and A. Eyre-Walker, 2007 Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol Biol Evol 24**:** 374-381.
Strigini, P., and E. Brickman, 1973 Analysis of specific misreading in *Escherichia coli*. J Mol Biol 75**:** 659-672.
Supek, F., and T. Smuc, 2010 On relevance of codon usage to expression of synthetic and natural genes in *Escherichia coli*. Genetics 185**:** 1129-1134.
Svidritskiy, E., R. Madireddy and A. A. Korostelev, 2016 Structural Basis for Translation Termination on a Pseudouridylated Stop Codon. J Mol Biol 428**:** 2228-2236.
Tamura, K., and M. Nei, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution 10**:** 512-526.
Tamura, K., M. Nei and S. Kumar, 2004 Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences of the United States of America 101**:** 11030-11035.
Tang, X., Y. Zhu, S. L. Baker, M. W. Bowler, B. J. Chen *et al.*, 2016 Structural basis of suppression of host translation termination by Moloney Murine Leukemia Virus. Nat Commun 7**:** 12070.
Tate, W. P., J. B. Mansell, S. A. Mannering, J. H. Irvine, L. L. Major *et al.*, 1999 UGA: a dual signal for 'stop' and for recoding in protein synthesis. Biochemistry (Mosc) 64**:** 1342-1353.
Tate, W. P., E. S. Poole, M. E. Dalphin, L. L. Major, D. J. Crawford *et al.*, 1996 The translational stop signal: codon with a context, or extended factor recognition element? Biochimie 78**:** 945-952.
Tate, W. P., E. S. Poole, J. A. Horsfield, S. A. Mannering, C. M. Brown *et al.*, 1995 Translational termination efficiency in both bacteria and mammals is regulated by the base following the stop codon. Biochem Cell Biol 73**:** 1095-1103.
Tuller, T., Y. Y. Waldman, M. Kupiec and E. Ruppin, 2010 Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A 107**:** 3645-3650.
Van Weringh, A., M. Ragonnet-Cronin, E. Pranckeviciene, M. Pavon-Eternod, L. Kleiman *et al.*, 2011 HIV-1 modulates the tRNA pool to improve translation efficiency. Mol Biol Evol 28**:** 1827-1834.
Wang, M., C. J. Herrmann, M. Simonovic, D. Szklarczyk and C. Von Mering, 2015 Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 15**:** 3163-3168.
Wang, M., M. Weiss, M. Simonovic, G. Haertinger, S. P. Schrimpf *et al.*, 2012 PaxDb, a database of protein abundance averages across all three domains of life. Molecular and Cellular Proteomics 11**:** 492-500.
Wei, Y., J. Wang and X. Xia, 2016 Coevolution between Stop Codon Usage and Release Factors in Bacterial Species. Molecular Biology and Evolution 33**:** 2357-2367.
Weiner, A. M., and K. Weber, 1973 A single UGA codon functions as a natural termination signal in the coliphage q beta coat protein cistron. J Mol Biol 80**:** 837-855.
Xia, X., 1998 How optimized is the translational machinery in *Escherichia coli, Salmonella typhimurium* and *Saccharomyces cerevisiae*? Genetics 149**:** 37-44.
Xia, X., 2005 Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. Gene 345**:** 13-20.
Xia, X., 2007a An improved implementation of codon adaptation index. Evol Bioinform Online 3**:** 53-58.
Xia, X., 2007b An Improved Implementation of Codon Adaptation Index. Evolutionary Bioinformatics 3**:** 53–58.
Xia, X., 2008 The cost of wobble translation in fungal mitochondrial genomes: integration of two traditional hypotheses. BMC Evol Biol 8**:** 1471-2148.

Xia, X., 2009 Information-theoretic indices and an approximate significance test for testing the molecular clock hypothesis with genetic distances. Mol Phylogenet Evol 52**:** 665-676.
Xia, X., 2012 Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. Scientifica (Cairo) 2012**:** 917540.
Xia, X., 2013a *Comparative genomics.* Springer.
Xia, X., 2013b DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Molecular Biology and Evolution 30**:** 1720-1728.
Xia, X., 2013c DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol 30**:** 1720-1728.
Xia, X., 2015a A major controversy in codon-anticodon adaptation resolved by a new codon usage index. Genetics 199**:** 573-579.
Xia, X., 2015b A Major Controversy in Codon-Anticodon Adaptation Resolved by a New Codon Usage Index. Genetics 199**:** 573-579.
Xia, X., 2016 PhyPA: Phylogenetic method with pairwise sequence alignment outperforms likelihood methods in phylogenetics involving highly diverged sequences. Molecular Phylogenetics & Evolution 102**:** 331-343.
Zavialov, A. V., L. Mora, R. H. Buckingham and M. Ehrenberg, 2002 Release of peptide promoted by the GGQ motif of class 1 release factors regulates the GTPase activity of RF3. Mol Cell 10**:** 789-798.