


# SCIENTIFIC REPORTS



OPEN

## An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria

Yulong Wei<sup>1</sup>, Jordan R. Silke<sup>1</sup> & Xuhua Xia<sup>1,2</sup> 

The degree to which codon usage can be explained by tRNA abundance in bacterial species is often inadequate, partly because differential tRNA abundance is often approximated by tRNA copy numbers. To better understand the coevolution between tRNA abundance and codon usage, we provide a better estimate of tRNA abundance by profiling tRNA mapped reads (tRNA tpm) using publicly available RNA Sequencing data. To emphasize the feasibility of our approach, we demonstrate that tRNA tpm is consistent with tRNA abundances derived from RNA fingerprinting experiments in *Escherichia coli*, *Bacillus subtilis*, and *Salmonella enterica*. Furthermore, we do not observe an appreciable reduction in tRNA sequencing efficiency due to post-transcriptional methylations in the seven bacteria studied. To determine optimal codons, we calculate codon usage in highly and lowly expressed genes determined by protein per transcript. We found that tRNA tpm is sensitive to identify more translationally optimal codons than gene copy number and early tRNA fingerprinting abundances. Additionally, tRNA tpm improves the predictive power of tRNA adaptation index over codon preference. Our results suggest that dependence of codon usage on tRNA availability is not always associated with species growth-rate. Conversely, tRNA availability is better optimized to codon usage in fast-growing than slow-growing species.

Codon optimization is critical to researchers seeking to improve protein production. Early experimental studies have shown that replacing rare codons with optimal ones increases protein yields in *Escherichia coli*<sup>1,2</sup>. The optimal codon within a given family is the most frequently used, especially in highly expressed genes (HEGs)<sup>3–5</sup>. In order to explain codon preference, early studies in *E. coli*<sup>6–8</sup> have shown that codon usage coevolves with tRNA abundance. The availability of tRNAs influences the usage of corresponding codons; conversely, high usage of preferred codons drives up the availability of their decoding tRNAs<sup>6,9</sup>.

Additionally, tRNA-mediated codon usage bias has been broadly observed in a variety of organisms, including the gram-negative bacterium *Salmonella enterica* serovar typhimurium<sup>10</sup>, the gram-positive *Bacillus subtilis*<sup>11</sup>, eukaryotes such as yeast<sup>10,12</sup>, a variety of fungal and invertebrate mitochondrial genomes<sup>13,14</sup>, and viruses including HIV<sup>15</sup>, and bacteriophages<sup>16,17</sup>. Nonetheless, the nature of the relationship between tRNA abundance and codon usage across bacterial species has been the subject of debate among researchers, with some suggesting that tRNA availability is the main driving force of codon usage bias<sup>18,19</sup> and others contending that the two are weakly correlated<sup>20,21</sup>.

A number of codon usage indices use tRNA gene copy as proxy of tRNA abundance to identify translationally optimal codons. These include the Codon Bias Index<sup>22</sup>, Frequency of Optimal Codons (F<sub>op</sub>)<sup>10</sup>, and tRNA Adaptation Index (tAI)<sup>18</sup>. All three of these indices define a translationally optimal codon as one that corresponds to the most abundant isoacceptor tRNA, with Codon Bias Index additionally incorporating gene expression information. Nevertheless, the use of tRNA gene copy is often undesirable. This is exemplified in *B. subtilis*<sup>18</sup> in

<sup>1</sup>Department of Biology, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, Ontario, Canada.

<sup>2</sup>Ottawa Institute of Systems Biology, Ottawa, Ontario, K1H 8M5, Canada. Yulong Wei and Jordan R. Silke contributed equally. Correspondence and requests for materials should be addressed to X.X. (email: [xxia@uottawa.ca](mailto:xxia@uottawa.ca))

which tAI (based on tRNA gene copy) fails to accurately predict codon usage, although codon usage correlates strongly with early experimental tRNA abundance<sup>11</sup> and conforms well to the selection-mutation-drift theory<sup>21,23</sup>. Moreover, while tRNA copy number is correlated with codon usage in a number of fast-growing bacterial species with a high variation in tRNA gene copy number, slow-growing ones exhibit little tRNA gene redundancy with many tRNA genes existing as a single copy. Resultantly, codon usage in such slow-growing species is poorly predicted by tRNA gene copy number<sup>24</sup>. For example, *Leptospira interrogans* and *Mycobacterium tuberculosis* have only 25 and 45 annotated tRNA genes according to the Genomic tRNA Database (GtRNAdb)<sup>25</sup>, respectively. Obviously, variation in codon usage cannot be well explained by tRNA gene redundancy if there is little variation in tRNA gene copy number. It stands to reason that the coevolutionary relationship between tRNA abundance and codon usage can be better characterized if tRNA abundance can be measured accurately.

To provide a better estimation for tRNA abundance in bacteria, we employed 14 publicly available RNA Sequencing (RNA-Seq) datasets, two for each of the seven species studied (*E. coli*, *S. enterica*, *B. subtilis*, *Bacteroides thetaiotaomicron*, *L. interrogans*, *M. tuberculosis*, and *Synechocystis* species). We quantified reads mapped to tRNA genes retrieved from GtRNAdb in transcripts per kilobase million (tpm) using kallisto<sup>26</sup>. To improve mapping efficacy, we have recently developed a new tool for processing RNA-Seq data, ARSDA<sup>27</sup>, that stores identical reads as single entries to drastically reduce data storage and computation time for analyzing large RNA-Seq datasets relative to previous methods<sup>28,29</sup>. These species were selected because their protein abundance data are available in PaxDb<sup>30</sup>, their growth rates are described on the basis of generation time (bacteria with >2.5 hour generation times are considered slow growing and all those with lower generation times are fast growing)<sup>24</sup>, and their RNA-Seq data are available (GEO Datasets).

A known issue with tRNA sequencing via standard Illumina protocols in eukaryotes<sup>31,32</sup> is post-transcriptional methylation occurring at a number of specific tRNA sites. Two recent approaches (DM-tRNA-seq<sup>32</sup> and ARM-seq<sup>31</sup>) of tRNA sequencing employ the *E. coli* derived AlkB demethylase enzyme to efficiently remove N<sup>1</sup>-methyladenosine (m<sup>1</sup>A), N<sup>3</sup>-methylcytosine (m<sup>3</sup>C) and N<sup>1</sup>-methylguanosine (m<sup>1</sup>G) structural modifications that hinder the activity of cDNA reverse transcriptase. Specifically, ARM-Seq<sup>31</sup> demonstrates that wild-type AlkB alone is sufficient to remove all three of the aforementioned modifications and generate full length tRNA cDNA. These studies are consistent with a prior investigation that similarly concluded that AlkB could capably demethylate m<sup>1</sup>G<sup>33</sup>. It is important to note that both studies focus on eukaryotes, and AlkB treatments may not be necessary to remove tRNA methylations in bacteria that naturally encode their own AlkB homologs. Besides *E. coli*<sup>31–34</sup>, several lines of evidence suggest AlkB homologous proteins are present in other bacterial species<sup>35–37</sup>. Specifically, AlkB homologs are observed in *B. subtilis*<sup>38</sup>, *S. enterica*, *M. tuberculosis*<sup>36,39</sup>, *Synechocystis* sp.<sup>36,40</sup>, and species in *Leptospira*<sup>36,39</sup> and *Bacteroides*<sup>41</sup> genera. Nonetheless, bacterial tRNA sequencing efficiency is a point of investigation in this study.

To our knowledge, our results are the first to show that tRNA quantification by RNA-Seq data is well correlated with early tRNA abundance derived from RNA fingerprinting (hereafter referred as RNA fingerprinting abundance) reported previously in *E. coli*, *S. enterica* and *B. subtilis*<sup>10–12,42</sup>. Briefly, determining RNA fingerprinting abundances involve separating radiolabelled RNA by 2D gel electrophoresis followed by quantification of radioactivity. Despite the challenges associated with tRNA sequencing in yeast<sup>31,43</sup> and mammals<sup>31,32,44</sup>, our results suggest that tRNA methylation may not appreciably influence tRNA sequencing efficiency in the bacterial species studied herein. We devised an integrated approach to show that tRNA tpm better predicts translationally optimal codons in *E. coli* than  $F_{op}$ , and improves the predictive power of tAI over codon preference. We found that the dependence of codon preference on tRNA availability is not always stronger in fast-growing species, and optimal codons can be well explained by tRNA content in certain slow-growing species. Conversely, tRNA availability is better optimized to codon usage in highly expressed genes of fast-growing than slow-growing species.

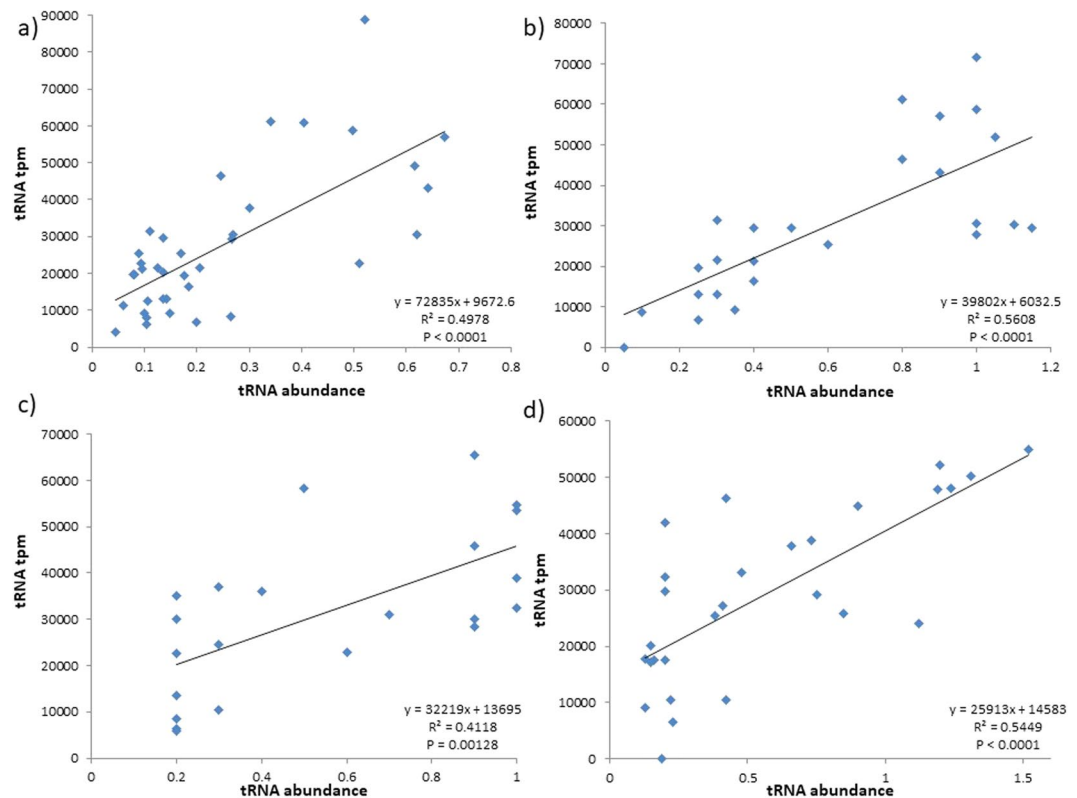
## Results

**RNA-Seq mappings are consistent with tRNA abundance estimates.** To accurately profile tRNA transcripts in tpm, we processed RNA-Seq data to remove adapters and low-quality sequences (See Materials and Methods for more detail) and quantified reads mapped to all unique tRNA sequences (Supplementary File S1) in tRNA tpm. To demonstrate the fidelity of tRNA tpm in bacteria, we compared these values with RNA fingerprinting abundances (Fig. 1, Supplementary File S2) previously reported in *E. coli*<sup>10,42</sup>, *S. enterica*<sup>10</sup>, and *B. subtilis*<sup>11</sup>. In all cases, tRNA tpm correlates with RNA fingerprinting abundance (Fig. 1:  $R^2 > 0.4$ ,  $P < 0.05$ ).

### Documented tRNA methylation does not appreciably affect tRNA sequencing in bacteria studied.

To investigate the potential effects of site-specific tRNA methylation on standard RNA-Seq experiments in bacteria, we visualized the RNA-Seq read depths for all seven species studied (Fig. 2, Supplementary File S1). In *E. coli*, read depths before and after documented tRNA methylation sites (18, 32, 34, 37, 46, and 54) in GenBank annotation (NC\_000913, Supplementary File S1) do not vary substantially, and we observe no partial tRNA mappings (Fig. 2a–d) contrary to the “hard-stops” previously described in both yeast and human tRNAs in the absence of demethylation treatment<sup>31</sup>. Additionally, tpm values associated with the set of tRNAs that can be potentially modified at five or all six documented methylation sites do not differ substantially from the set of tRNAs that can be potentially modified at four or less sites (Fig. 2d,e; two-tailed Student’s t-test with unequal variance:  $P = 0.477$ ). We define tRNAs that can potentially be methylated at >4 sites as heavily methylated with respect to other tRNAs. This cut-off was chosen because it divides the 50 unique tRNA sequences into two subsets of roughly equal size. Similarly, hard-stops were not observed at documented methylated sites in all six other bacteria studied (Supplementary File S1).

**An improved estimation of codon preference using tRNA tpm.** To investigate how well tRNA tpm explains codon preference across bacterial lineages, we first designed an integrative approach to determine

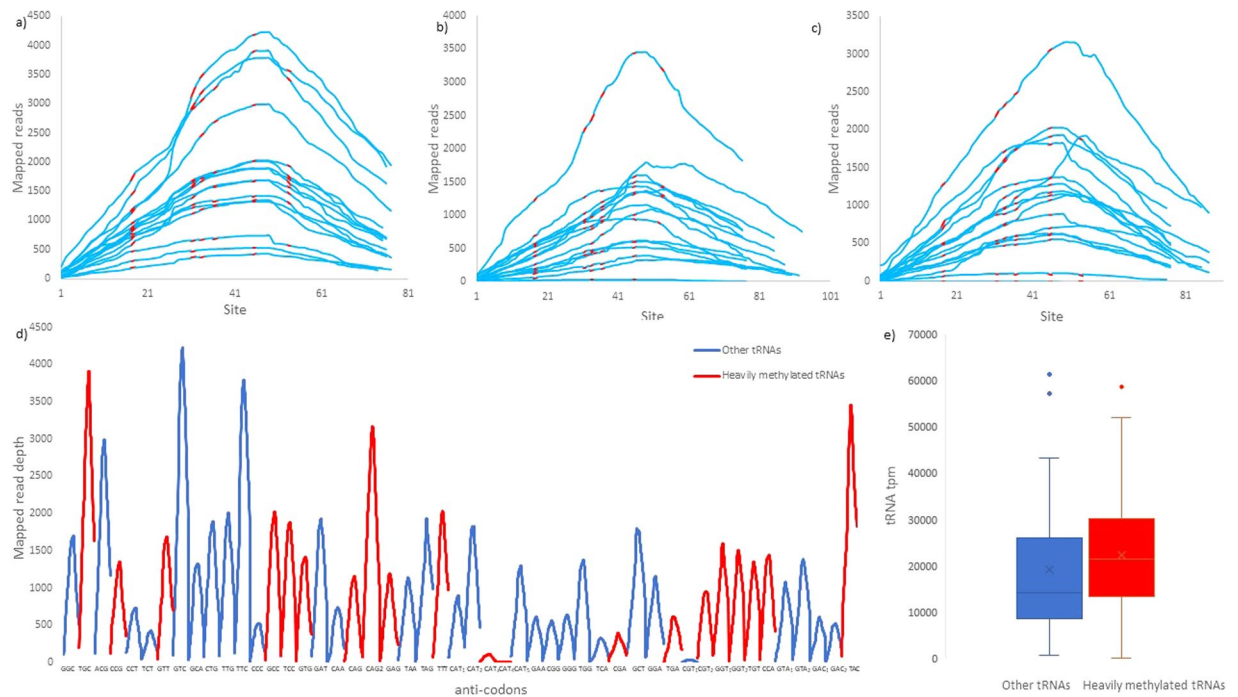


**Figure 1.** Comparison between tRNA tpm and fingerprinting abundance in *E. coli*, *S. enterica*, and *B. subtilis*. In panels (a) the averaged tRNA abundances across five growth phases retrieved from Dong, *et al.*<sup>42</sup>, (b,c) are tRNA abundances retrieved from Ikemura<sup>10</sup>, and in (c) the tRNA abundances retrieved from Kanaya, *et al.*<sup>11</sup>.

translationally optimal codons. We use two criteria to define what constitutes a translationally optimal codon: 1) it has the highest relative synonymous codon usage (RSCU)<sup>45</sup> in HEGs, and 2) its RSCU in HEGs is greater than that in lowly expressed genes (LEGs). These two criteria combine the specifications of optimal codons defined by two previous studies: the first criteria is used by Novoa, *et al.*<sup>19</sup> and the second is adapted from Rocha<sup>24</sup> (HEGs vs. others). These criteria are also consistent with those used in the calculation of the Codon Adaptation Index<sup>4</sup> and Index of Translation Elongation<sup>5</sup>. By our definition, there can be only one codon that is most translationally optimal within each synonymous group, consistent with the original definition of optimal codon determined by  $F_{op}$ <sup>10</sup>. Thus, the characterized translationally optimal codons will increase elongation efficiency relative to others. The availability of protein abundance data in parts per million (ppm) for species studied herein and mRNA transcript abundance (in tpm) determined using kallisto<sup>26</sup> enables us to calculate protein per transcript (ppm/tpm) in the identification of HEGs and LEGs (see Materials and Methods for more detail).

To establish readable isoacceptor tRNA content, we consider all cognate and near-cognate interactions, as well as those enabled by anticodon modifications. Most studies<sup>10,11,18,42</sup> consider anticodon-cognate (e.g., tRNA<sup>Arg</sup><sub>UGC</sub> reading GCA) and near-cognate (e.g., tRNA<sup>Arg</sup><sub>UGC</sub> reading GCG) pairings; while some<sup>19,24</sup> also consider pairings allowed due to anticodon modifications (e.g., tRNA<sup>Arg</sup><sub>UmGC</sub> reading GCC and GCU) which increases the predictive influence of tRNA content on codon usage<sup>19</sup>. In order to explain RSCU, we adapt the idea of Relative tRNA Gene frequency from Novoa, *et al.*<sup>19</sup> to derive Relative tRNA Usage (RTU) (see Materials and Methods for more detail). By considering tRNA abundance for synonymous codons, RTU improves tRNA tpm as an estimator of tRNA abundance (Fig. 3). In particular, the correlation between tRNA tpm and average tRNA abundance<sup>42</sup> in *E. coli* (Fig. 1a;  $R^2 = 0.498$ ,  $P < 0.0001$ ) improves if we consider their RTU values (Fig. 3a:  $R^2 = 0.646$ ,  $P < 0.0001$ ). Furthermore, both RTU values correlate with codon usage (Fig. 3b).

Using RTU, we estimate how well translationally optimal codons match codons with the highest tRNA availability by adapting the four rules of codon-anticodon constraint<sup>10</sup>. Rule one states that codon usage is constrained by tRNA availability<sup>46</sup>. Rules two to four focus on specific base pairing efficiencies and they describe that cognate codon-anticodon pairs are generally more efficient and preferable relative to near-cognate wobble pairs<sup>22,47–50</sup>. Hence, we first rank synonymous codons by highest RTU, and then rank by cognate tRNA abundances (Supplementary File S2). Supplementary Fig. S1 provides a flowchart explaining the approach to predict translationally optimal codons by tRNA availability. For example, applying this two-step identification approach for Threonine codons in *E. coli*, we first select ACC and ACU since they both have the highest RTU and are both readable by the same tRNAs (tRNA<sup>Thr</sup><sub>GGU</sub> and tRNA<sup>Thr</sup><sub>UGU</sub>) due to anticodon modification by ADATs<sup>19</sup>. Next, we rank codon preference by cognate tRNA abundance: tRNA tpm of the cognate tRNA<sup>Thr</sup><sub>GGU</sub> for ACC is 46537.7, and ACU is not decoded by a cognate tRNA. Thus, the predicted optimal codon based on tRNA availability and pairing constraints is ACC in Threonine. Indeed, ACC is the translationally optimal codon for Threonine



**Figure 2.** RNA-Seq read map for all 50 unique tRNA sequences in *E. coli*, split in three sets (a–c). Each line represents the read depth of entire sequence region of one unique tRNA sequence, with sites susceptible to methylation ( $m^2G18$ ,  $m^2C32$  or  $m^2U32$ ,  $m^5U34$  or  $m^2C34$  or  $cmo^5U34$ ,  $m^6A37$ ,  $m^7G46$  and  $m^5U54$ ) highlighted red. In (d) the distribution of mapped reads across the entire length of each unique tRNA sequence. Red indicates tRNAs that are potentially methylated at >4 sites (heavily methylated) and all others are highlighted blue. In (e) the distribution of total tRNA tpm in sets of heavily methylated and other tRNAs.

based on our definition: it has the highest RSCU in HEGs (2.111), and (2) its RSCU HEG is higher than RSCU LEG (1.468). We extended this two-step approach to predict translationally optimal codons in all seven species (Table 1).

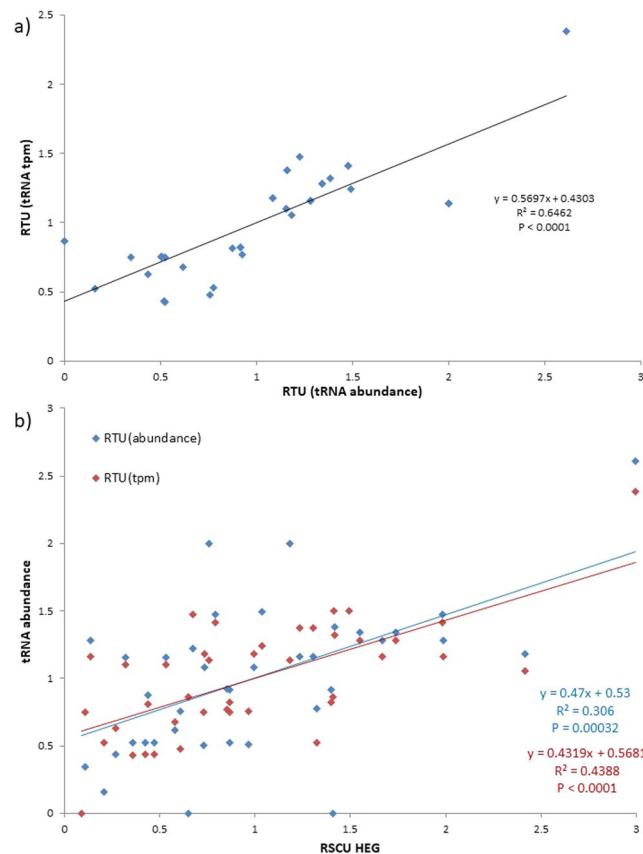
Many studies have demonstrated that bacterial tRNA abundance fluctuates due to growth phase and culture conditions such as temperature and media<sup>42,51–53</sup>. Hence, for each species, we have acquired tRNA tpm values from two independent, but experimentally consistent (i.e., all strains are wildtype, and all cultures are taken during log-phase growth) RNA-Seq datasets (Table 2) that were prepared for sequencing on the same platform (Illumina) to verify consistency in predicting translationally optimal codons using tRNA tpm (Table 1).

**Implementing tRNA tpm in tAI calculation improves the non-parametric S correlation.** We calculated tAI using gene copy number and tRNA tpm (See Materials and Methods for more detail, Supplementary Fig. S2), and Table 3 shows the non-parametric regression  $S$  which reflects the correlation between tAI values and effective number of codons<sup>54</sup> (corrected for silent substitutions<sup>18</sup>). For six out of seven species studied,  $S'$  calculated with tRNA tpm is higher than  $S$  calculated using tRNA gene copy number (Table 3). Hence, tAI' values calculated using tRNA tpm better correlate with codon usage than tAI values obtained from tRNA gene copy number (two-tailed Student's  $t$ -test with unequal variance:  $P < 0.0001$ ; Supplementary Fig. S2) in all species except *L. interrogans*. Note that  $S$  values calculated herein use non-hypothetical and non-pseudo genes, whereas those originally calculated ( $S_0$ )<sup>18</sup> use all coding DNA sequences, although value ranks stay consistent (Table 3).

**Abundance of tRNA depends on codon usage in fast-growing species.** Codon usage preferences can also drive up tRNA content<sup>6,24</sup>. For each tRNA species (distinguished by anticodon), we define its readable codon usage as the sum usage of its cognate and near-cognate codons (e.g., the readable codon usage of  $tRNA^{Ala}_{GCC} = GCC$  usage +  $GCU$  usage). Readable codon usage in HEGs better explains tRNA tpm in fast-growing species than in slow-growing species. More specifically, readable codon usage better correlates with tRNA tpm in *E. coli*, *S. enterica*, and *B. subtilis*, than in *B. thetaiotaomicron*, *Synechocystis* sp., *M. tuberculosis*, and *L. interrogans* (Supplementary Fig. S2).

## Discussion

Two approaches have been taken to characterize the effect of tRNA on codon usage. The first tests whether gain or loss of tRNA genes will lead to predicted changes in codon usage<sup>55–57</sup>. In tunicates and bivalves, an additional  $tRNA^{Met/UAU}$  gene is present in the mitochondrial genome. One would expect that the additional  $tRNA^{Met/UAU}$  would favor increased usage of AUA codons, and this expectation is empirically substantiated<sup>55,56</sup>. One may also reason that, if a bacteriophage encodes many tRNA genes in its own genome, especially when these tRNAs are



**Figure 3.** Relationship between (a) tRNA tpm and averaged tRNA abundance from RNA fingerprinting<sup>42</sup> (tRNA abundance) in *E. coli*, and (b) RTUs and RSCU in *E. coli* HEGs.

rare in the host, then the phage codon usage will be less dependent on the host tRNA pool. This expectation is also consistent with empirical evidence<sup>16,58</sup>. The second approach quantifies within-species association of codon usage with tRNA abundance which is often approximated by tRNA gene copy number, but this proxy has two key shortcomings. First, we do not know if it is generally true that tRNA gene copy number serves as a good proxy for tRNA abundance. Second, some bacterial species, such as *M. tuberculosis*, have only a single tRNA gene for all anticodons. In such cases, we cannot use tRNA gene copy number as a proxy of abundance because of its lack of variability. Thus, accurate quantification of tRNA abundance is crucial for detecting codon adaptation.

Our RNA-Seq-based analysis shows that snapshots of bacterial tRNA pools can be captured using RNA-Seq profiling in spite of claims that standard Illumina RNA Sequencing protocols inefficiently quantify tRNAs in eukaryotes<sup>32,43,44</sup> due to a number of modifications that increase the stability of tRNA secondary structure. While some tRNA modifications are shared among the three kingdoms of life, others are kingdom specific<sup>59</sup>. Nonetheless, tRNA post-transcriptional methylation is extensive in Eubacteria<sup>59</sup>, and we acknowledge that tRNA tpm values may underestimate tRNA abundances since tRNA is highly structured and difficult to denature. Despite this, we do not observe any drastic read count drops or hard-stops in mapped reads at or flanking documented methylated sites in the seven species studied here, nor do we see partially mapped tRNA regions (Fig. 2a–c; Supplementary File S1) that were commonly observed in untreated tRNA sequencing data in Eukaryotes<sup>31</sup>. A caveat is the presence of a hard-stop in most tRNAs at site 50 for *B. subtilis* (SRX2804667), *Synechocystis* sp. (SRX4145044) and *L. interrogans* (SRX2448246) (Supplementary File S1). However, there is no documented methyltransferase activity acting at this site for these species, and the drop in mapped reads is only observed in one of the two SRX datasets retrieved for each species. Nonetheless, we acknowledge that tRNA sequencing may be potentially enhanced with demethylase treatments<sup>31,32</sup> that should be employed in future tRNA-Seq studies in bacteria.

Because eukaryotic tRNA read mapping abundances are considerably higher in demethylated samples than untreated samples<sup>31</sup>, we expected that our read mapping abundances may be similarly impacted by tRNA methylation. In light of this, we considered all annotated methylation sites in bacterial tRNAs, even though they may not be methylated at all times due to structural constraints. Surprisingly, our observations reveal that read mapping abundances of *E. coli* tRNAs that are potentially heavily methylated do not differ from those that are susceptible to methylation at fewer than five methylation sites (Fig. 2e). This suggests that the requirement for demethylase treatment prior to tRNA sequencing in bacterial species with functional AlkB demethylase homologs may be more relaxed, especially since demethylation treatments in current eukaryotic tRNA sequencing approaches are bacterial AlkB-facilitated<sup>31,32</sup>.

Amino acid <sup>1</sup>	Synonymous codons	<i>E. coli</i>	<i>S. enterica</i>	<i>B. subtilis</i>	<i>B. thetaiotaomicron</i>	<i>Synechocystis</i> sp.	<i>M. tuberculosis</i>	<i>L. interrogans</i>
Ala	GCA, GCC, GCG, GCU	— <sup>3</sup>	—	GCA <sup>a</sup>	GCU	<b>GCC</b>	GCC	GCA
Cys	UGC, UGU	<b>UGC<sup>b</sup></b>	UGC	UGC	UGU	—	UGC	—
Asp	GAC, GAU	—	—	—	GAU	—	<b>GAC</b>	—
Glu	GAA, GAG	GAA <sup>ab</sup>	<b>GAA</b>	<b>GAA<sup>a</sup></b>	GAA	<b>GAA</b>	<b>GAG</b>	<b>GAA</b>
Phe	UUC, UUU	<b>UUC<sup>ab</sup></b>	UUC	—	UUC	—	UUC	—
Gly	GGN	<b>GGC<sup>ab</sup></b>	<b>GGC</b>	<b>GGC</b>	GGU	GGU	<b>GGC*</b>	—
His	CAC, CAU	<b>CAC<sup>b</sup></b>	—	—	—	—	<b>CAC</b>	—
Ile	AUA, AUC, AUU	<b>AUC<sup>ab</sup></b>	AUC	—	AUC	—	AUC	—
Lys	AAA, AAG	<b>AAA<sup>ab</sup></b>	AAA	<b>AAA<sup>a</sup></b>	AAA	—	<b>AAG</b>	—
Leu 2-fold <sup>2</sup>	UUA, UUG	<b>UUG<sup>b</sup></b>	UUG	UUA <sup>a</sup>	—	UUG	UUG	—
Leu 4-fold	CUA, CUC, CUG, CUU	<b>CUG<sup>a</sup></b>	CUG	CUU	—	CUG <sup>*</sup>	CUG	—
Asn	AAC, AAU	<b>AAC<sup>ab</sup></b>	AAC	—	—	—	<b>AAC</b>	—
Pro	CCA, CCC, CCG, CCU	<b>CCG<sup>ab</sup></b>	<b>CCG</b>	—	—	CCC <sup>*</sup>	<b>CCG</b>	CCU
Gln	CAA, CAG	<b>CAG<sup>ab</sup></b>	<b>CAG</b>	<b>CAA<sup>a</sup></b>	—	—	<b>CAG</b>	—
Arg 2-fold	AGA, AGG	AGA <sup>a</sup>	AGA	AGA	AGA	<b>AGG</b>	<b>AGG</b>	<b>AGA</b>
Arg 4-fold	CGA, CGC, CGG, CGU	<b>CGU<sup>a</sup></b>	<b>CGU</b>	CGC	<b>CGU</b>	<b>CGG*</b>	CGC	CGU
Serine 2-fold	AGC, AGU	<b>AGC<sup>b</sup></b>	<b>AGC</b>	<b>AGC</b>	—	—	<b>AGC</b>	—
Serine 4-fold	UCA, UCC, UCG, UCU	UCU	<b>UCC</b>	UCA	UCU	UCC	—	UCU
Thr	ACA, ACC, ACG, ACU	<b>ACC<sup>ab</sup></b>	<b>ACC</b>	ACA <sup>a</sup>	ACU	<b>ACC</b>	<b>ACC</b>	ACU
Val	GUA, GUC, GUG, GUU	—	—	GUU	<b>GUA</b>	—	—	—
Tyr	UAC, UAU	—	—	—	—	—	<b>UAC</b>	—

**Table 1.** Translationally optimal codons estimated for synonymous groups in seven bacterial species. Bold are translationally optimal codons that also have the highest tRNA availability estimated by tpm from two independent RNA-Seq datasets. <sup>1</sup>Two amino acids (Met and Trp) are omitted because they are each encoded by a single codon. <sup>2</sup>The 6-fold degenerate codon families (Leu, Arg, and Ser) are broken into 2 and 4-fold families because of differences in the first codon base. <sup>3</sup>An optimal codon cannot be determined by our definition (e.g., RSCU HEG < RSCU LEG violates the second criterion). \*Predicted preferred codons match optimal codons in only one of the two RNA-Seq data analyzed. <sup>a</sup>Translationally optimal codons match optimal codons determined by F<sub>op</sub> in Ikemura<sup>10</sup> and Kanaya, *et al.*<sup>11</sup>. <sup>b</sup>Translationally optimal codons identified using average tRNA abundance from RNA fingerprinting approach, retrieved from Dong, *et al.*<sup>42</sup>.

Species	Strain	Growth Rate <sup>a</sup>	NCBI Accession	Experiment ID <sup>*</sup>
<i>Bacteroides thetaiotaomicron</i>	VPI-5482	Slow	NC_004663	SRX020805, SRX860738
<i>Bacillus subtilis</i>	168	Fast	NC_000964	SRX515181, SRX2804667
<i>Escherichia coli</i>	K-12	Fast	NC_000913	SRX515174, SRX669653
<i>Leptospira interrogans</i>	Fiocruz L1-130	Slow	AE016823	SRX2448246, SRX405952
<i>Mycobacterium tuberculosis</i>	H37Rv	Slow	NC_000962	SRX1372108, SRX4374910
<i>Salmonella enterica</i>	LT2	Fast	NC_003197	SRX1638989, SRX1258668
<i>Synechocystis</i> sp.	PCC 6803	Slow	NC_017277	SRX347145, SRX4145044

**Table 2.** The seven bacterial species studied herein due to their availability of protein abundance, growth rate and RNA-Seq data. <sup>a</sup>Information on species growth-rate are retrieved from Rocha<sup>24</sup>. \* All selected datasets have matching species strains between protein abundance, GtRNAdb, NCBI and RNA-Seq data, with the exception of *L. interrogans* (SRX2448246) due to the lack of a second RNA-Seq data in GEO Datasets for strain Fiocruz L1-130. All cultures in RNA-Seq experiments were isolated during log phase of growth. The first listed SRX dataset was selected for all analyses, and both were used for Table 1 and Fig. S4.

In all bacteria, we combined our RNA quantification approach with available protein abundances to determine the most translationally optimal codon in 21 codon groups (Table 1) based on codon usage differences between HEG and LEG subsets. These subsets of genes are established based on protein per transcript (Supplementary Table S1, File S3), with the aim of providing a more accurate estimate of translation efficiency from protein

Species	$S_0$	$S$	$S'$
<i>E. coli</i>	0.70 <sup>a</sup>	0.61 <sup>b</sup>	0.71 <sup>c</sup>
<i>S. enterica</i>	0.63	0.59	0.69
<i>B. thetaiotaomicron</i>	0.55	0.4	0.62
<i>Synechocystis</i> sp.	0.38	0.27	0.5
<i>B. subtilis</i>	-0.01	0.18	0.33
<i>M. tuberculosis</i>	-0.04	0.1	0.13
<i>L. interrogans</i>	N/A	0.23	0.2

**Table 3.** Non-parametric regression  $S$  correlations between tAI values and effective number of codons. <sup>a</sup> $S$  values retrieved from dos Reis, *et al.*<sup>18</sup>, calculated using all coding DNA sequences. <sup>b</sup> $S$  values calculated using tRNA gene copy number, using genes having non-zero protein abundances. <sup>c</sup> $S$  values calculated using tRNA tpm, using genes having non-zero protein abundances.

abundance by decoupling rates of transcription. However, species-specific translationally optimal codons cannot always be established because the two described criteria are violated in some groups (e.g., the codon with the highest RSCU is more over-represented in LEGs than HEGs). In these cases, there is no evidence that the most abundantly used synonymous codon would contribute to increase translation efficiency. In particular, *L. interrogans* represents a slow-growing species wherein codon optimization is very poor (only seven translationally optimal codons can be characterized in 21 codon groups).

Our tRNA quantification approach better predicts translationally optimal codons over  $F_{op}$ <sup>10</sup>. In *E. coli*, synonymous codons with the highest tRNA tpm (ranked by highest TPU followed by highest cognate tRNA abundance) match 15 out of 17 translationally optimal codons, but 13 out of 17 when we replace tRNA tpm with averaged RNA fingerprinting abundance retrieved from Dong, *et al.*<sup>42</sup> (Table 1). In contrast,  $F_{op}$  determines 12 such translationally optimal codons<sup>10</sup> of which AGA (Arg) is the only codon that our method does not predict to be optimal (Table 1). Additionally, all translationally optimal codons determined herein are consistent with optimal codons determined by  $F_{op}$ , except in the Serine 4-fold family where  $F_{op}$  predicts UCC whereas we predict UCU to be optimal. In the case of *B. subtilis*, both tRNA tpm and  $F_{op}$  determine six translationally optimal codons (Table 1). It is worth mentioning that  $F_{op}$  determines 16 optimal codons in *B. subtilis*<sup>11</sup>, but most may not be translationally optimal. For example, CCA (Pro) was determined to be an optimal codon, but CCG (Pro) is substantially more preferred than CCA (Pro) (RSCU in HEGs are 1.735 and 0.782, respectively).

Nonetheless, bacterial tRNA abundance may not fully explain the variation in usage of all 61 sense codons (Supplementary Fig. S4). First, codon preference cannot always be inferred reliably from tRNA gene redundancy or experimentally measured tRNA abundance. For example, inosine is expected to pair best with C and U, but less with A (presumably because of the bulky I/A pairing involving two purines)<sup>60</sup>. Second, what matters in translation elongation is the availability of charged tRNAs. It is difficult to determine the level of charged tRNAs, and researchers typically would use transcriptionally determined tRNAs or even the number of tRNA genes in the genome as a proxy of charged tRNAs. Unfortunately, the abundance of tRNAs does not always reflect the abundance of charged tRNA<sup>61</sup>. Lastly, other factors such as mutation bias<sup>21,62–65</sup> may exert more pressure on codon usage in certain species.

Conversely, the variation in tRNA tpm is better explained by codon usage (Supplementary Fig. S3) in fast-growing (*E. coli*, *B. subtilis* and *S. enterica*) than in slow-growing species (*B. thetaiotaomicron*, *L. interrogans*, *M. tuberculosis* and *Synechocystis* sp.). This result supports the theory that tRNA translation machinery is better optimized to codon usage in fast-growing than slow-growing species<sup>9,24</sup>. Indeed, duplicating tRNA genes is an effective way to elevate transcript abundance in species that grow and replicate rapidly<sup>10–12,42</sup>, but not in slow-growing species (Supplementary Fig. S5).

One potentially important implementation of tRNA tpm is in the calculation of tAI. Our results (Table 3, Supplementary Fig. S2) show that tAI' (calculated using tRNA tpm) better explains effective number of codons than tAI (calculated using tRNA gene copy number) for all species studied except *L. interrogans*. Considering  $S$  and  $S_0$  calculated using tRNA gene copy numbers, their differences are likely due to our usage of the subset of non-hypothetical and non-pseudo genes that have protein abundance values<sup>30</sup> ( $S$ ) whereas all DNA coding sequences (including hypothetical and pseudo genes) were used in the original calculation<sup>18</sup> ( $S_0$ ). Additionally, both the GtRNAdb<sup>25,66</sup> and DNA coding sequences (GenBank annotations) have been continuously curated since 2004. Lastly, only wobble base pairings were considered in the original introduction of tAI<sup>18,67</sup>; whereas we have also considered possible anticodon modifications<sup>19,24</sup> that further relax codon pairing. These differences improve the calculation of the  $S$  correlation using tRNA copy numbers, notably in *B. subtilis* and *M. tuberculosis*. In contrast, the originally calculated negative  $S_0$  correlation for *B. subtilis* was a major shortcoming of the tAI method<sup>18</sup> and was criticized<sup>21</sup> for suggesting a lack of selective pressure exerted by tRNA abundance on codon preference in this species.

In the case of *M. tuberculosis*, our tRNA quantification approach is much more sensitive to determining tRNA-mediated codon bias than tAI. The  $S$  correlations are consistently the lowest for *M. tuberculosis* (Table 3), yet we identified 17 out of 19 translationally optimal codons using tRNA tpm (Table 1). Our method recaptures the “weak but significant codon usage preference” previously reported<sup>21,68</sup> in this slow-growing species, and show that the degree to which tRNA availability explains optimal codon usage is species-specific and does not always depend on growth-rate.

We studied the coevolution between codon usage and tRNA abundance in three fast-growing species (*E. coli*, *S. enterica*, and *B. subtilis*) and four slow-growing species (*B. thetaiotaomicron*, *L. interrogans*, *M. tuberculosis*, and *Synechocystis* sp.). Our findings indicate that tRNA quantification by tpm offers better predictions of translationally optimal codons over  $F_{op}$  in *E. coli*, and improves the calculation of tAI to better reflect codon preference in all species studied except *L. interrogans*. The usage of translationally optimal codons can be well explained by relative tRNA tpm in *E. coli* and *S. enterica*; however, both tRNA tpm and RNA fingerprinting abundances<sup>11</sup> offer weaker explanations for codon preference in *B. subtilis*. The influence of tRNA availability on codon bias is not always stronger in fast-growing species, and optimal codons can be well explained by tRNA content in certain slow-growing species such as *M. tuberculosis*. Conversely, the tRNA translation machinery is better optimized to codon usage in HEGs of fast-growing than slow-growing species.

## Materials and Methods

**Processing genomic, proteomic and RNA-seq data.** We retrieved the annotated genomes (Table 2) for three fast growing species (*E. coli*, *B. subtilis*, and *S. enterica*) and four slow-growing species (*B. thetaiotaomicron*, *L. interrogans*, *M. tuberculosis*, and *Synechocystis* sp.) in GenBank format from the National Center for Biotechnology Information (NCBI) database (<http://www.ncbi.nlm.nih.gov>). For each species, all documented tRNA methyltransferase genes were retrieved from GenBank annotations (Supplementary File S1).

Protein abundance data corresponding with each of these species were retrieved from PaxDb 4.0<sup>30</sup> <https://pax-db.org/> and abundance values were associated with GeneIDs retrieved from the genomes using DAMBE 7<sup>69</sup>. The integrated protein abundance dataset was taken when available.

RNA-Seq runs of wildtype species were fetched from GEO DataSets (<https://www.ncbi.nlm.nih.gov/gds/>) in FASTQ format. FASTQ files were converted to FASTQ+ format using ARSDA 1.1<sup>27</sup> in order to reduce file sizes by grouping identical reads under a single ID while retaining the copy number for each read (in the format S<read#>\_<copy#>). The FASTQ+ data was then processed using both CutAdapt 1.17<sup>70</sup> and Trimmomatic 0.38<sup>71</sup> to remove flanking adapter sequences and purge low quality reads. For experiments that use the oligo(dT)-adapter primer for cDNA synthesis, RNA fragments are first poly-adenylated at the 3' end. In these cases, we set CutAdapt to recognize "AAAAA". We also used CutAdapt to recognize and remove all possible adapters in experiments that used custom adapters, with 10% mismatch error rate. When the adapters conformed to standard Illumina protocols, we simply used the "ILLUMINACLIP" function built into Trimmomatic to trim the relevant library of adapters from reads (ILLUMINACLIP: <adapters.fa> :2:20:6). After adapters were trimmed, we retained reads that were a minimum of 25 nt long (-m 25 in CutAdapt or MINLEN:25 in Trimmomatic) to mitigate bias in expression levels<sup>72</sup>. We filtered trimmed reads to remove poor quality sequences with average Phred scores lower than 30 (0.1% probability of a base calling error)<sup>73</sup>.

**RNA-Seq read mapping for tRNAs and mRNAs.** We retrieved the sequences of all genomically encoded tRNAs for each organism from the Genomic tRNA Database (GtRNAdb 2.0<sup>66</sup>) in FASTA format and removed predicted pseudo-tRNAs and those with unspecified anticodons. The FASTA files containing tRNA sequences were read into DAMBE to represent identical sequences with one ID indicating the number of identical copies. Since mature tRNAs are modified to have 5'-CCA-3' appended to their 3' end, we manually added CCA to sequences lacking this motif<sup>32</sup>. The modified tRNA FASTA files for all species were indexed and the associated RNA-Seq reads from processed FASTQ+ files were pseudo-aligned to each tRNA index and tRNA tpm was quantified using kallisto v0.44.0<sup>26</sup>. The tRNA pseudo-alignments for each species were subsequently sorted and site-specific depth values were generated for each tRNA using the sorted pseudo-alignments via the 'sort' and 'depth' commands from SAMtools<sup>74</sup>, respectively.

Similarly, all non-pseudo and non-hypothetical DNA coding sequences with non-zero protein abundance values were retrieved using DAMBE in FASTA format and indexed. The associated RNA-Seq reads from processed FASTQ+ files were pseudo-aligned to each mRNA index and mRNA tpm values quantified using kallisto.

### Determination of translationally highly and lowly expressed genes by protein per transcript.

Protein per transcript (ppm/tpm) was estimated by taking gene protein abundance (ppm) divided by its mRNA tpm, for both RNA-Seq datasets in each species except *B. subtilis*. For *B. subtilis*, protein per transcript was obtained with only SRX515181 dataset, because SRX2804667 MiSeq experimental protocol effectively removes large transcripts to study tRNAs<sup>75</sup>. From each dataset, genes with top and bottom 30% ppm/tpm values were selected (Supplementary File S3). A gene is considered to be highly expressed if the gene ID is found in both gene sets for each species; the same was done to identify lowly expressed genes from the bottom 30% ppm/tpm gene sets (Supplementary File S3, Table S1). To verify the validity of this approach, we determined the number of ribosomal protein (30S and 50S subunit) genes that are present in each gene sets. We observed a great number of ribosomal protein genes in the top 30% ppm/tpm gene sets and nearly none in the bottom 30% ppm/tpm gene sets (Supplementary Table S1). This is expected because ribosomal protein genes are commonly accepted and used as highly expressed genes<sup>4,24</sup>.

### Computation of relative synonymous codon usage and tRNA usage metrics.

Relative synonymous codon usage (RSCU)<sup>45</sup> values were computed for each species by loading HEGs and LEGs (Supplementary File S3) into DAMBE and selecting "Seq. Analysis" > "Codon Usage" > "Relative synonymous codon usage". DAMBE's implementation of the RSCU computation automatically splits 6-fold degenerate codon families into a 2-fold and 4-fold degenerate family based on difference at the first codon position.

To acquire relative tRNA usage for each synonymous codon (RTU), we adapt the RSCU formula (1) in the same way that Relative tRNA Gene frequency was employed by Novoa, *et al.*<sup>19</sup> using tRNA gene copy number:



$$RTU_i = \frac{tpm \text{ of } tRNA_i}{\frac{\sum_i tRNA_i}{n}} \quad (1)$$

where  $i$  is any codon within a 2 or 4-fold degenerate codon family, and  $n$  is the total number of codons in the synonymous group. RSCU and RTU are both calculated by breaking 6-fold codon families (Arginine, Leucine, and Serine) into 2 and 4-fold groups (e.g., 4-fold CGN (arg) and 2-fold AGN (arg)). Methionine and Tryptophan have been omitted from RSCU and RTU calculation because they are encoded by a single codon (RSCU and RTU = 1). Similarly, codon groups with RTU = 1 have been omitted from plots when applicable (Supplementary Fig. S4), because RTU will not estimate RSCU for these codons. All correlation coefficients ( $R^2$ ) are calculated by taking the square of Pearson's correlation  $r$  (Figs 1 and 3, Supplementary Figs S3–5).

**Computation of tAI and correlation S.** We first calculated tAI using the original formulation of the model<sup>18,67</sup> which considers the copy number of all isoacceptor tRNAs for each codon via the author's tAI R package version 0.2 (<https://github.com/mariodosreis/tai>) for all species in this study. We additionally computed a modified version of tAI (tAI') which uses the summed tpm values associated with codon-specific isoacceptor tRNAs in lieu of tRNA copy number. Rather than using all annotated DNA coding sequences in these calculations, as was done originally<sup>18</sup>, we only considered genes with non-zero protein abundance values from the integrated datasets stored in PaxDb. The tAI and tAI' values for each species were plotted (Supplementary Fig. S2) against the effective number of codons corrected for silent substitutions at the third codon position ( $f[GC3s] - Nc$ ) to determine the  $S$  and  $S'$  correlation coefficients, respectively.

### Data Availability

Supplementary File S1 contains RNA-Seq read depths and tRNA methylation profile. Supplementary File S2 contains identifications of translationally optimal codons and tRNA abundances (gene copy, tpm and fingerprinting data) and tRNA quantification approaches. Supplementary File S3 contains protein per transcript data, protein abundance information, identified translationally HEGs and LEGs. Supplementary File S4 contains Supplementary Figs S1–S5 and Table S1.

### References

- Robinson, M. *et al.* Codon usage can affect efficiency of translation of genes in *Escherichia coli*. *Nucleic acids research* **12**, 6663–6671 (1984).
- Sorensen, M. A., Kurland, C. G. & Pedersen, S. Codon usage determines translation rate in *Escherichia coli*. *J Mol Biol* **207**, 365–377 (1989).
- McPherson, D. T. Codon preference reflects mistranslational constraints: a proposal. *Nucleic Acids Res* **16**, 4111–4120 (1988).
- Sharp, P. M. & Li, W. H. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research* **15**, 1281–1295 (1987).
- Xia, X. A Major Controversy in Codon-Anticodon Adaptation Resolved by a New Codon Usage Index. *Genetics* **199**, 573–579, <https://doi.org/10.1534/genetics.114.172106> (2015).
- Bulmer, M. Coevolution of codon usage and transfer RNA abundance. *Nature* **325**, 728–730 (1987).
- Gouy, M. & Gautier, C. Codon usage in bacteria: correlation with gene expressivity. *Nucleic acids research* **10**, 7055–7074 (1982).
- Ikemura, T. 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 (1981).
- Higgs, P. G. & Ran, W. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Molecular biology and evolution* **25**, 2279–2291 (2008).
- Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution* **2**, 13–34 (1985).
- Kanaya, S., Yamada, Y., Kudo, Y. & Ikemura, T. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238**, 143–155 (1999).
- Xia, X. How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*? *Genetics* **149**, 37–44 (1998).
- Carullo, M. & Xia, X. An Extensive Study of Mutation and Selection on the Wobble Nucleotide in tRNA Anticodons in Fungal Mitochondrial Genomes. *Journal of Molecular Evolution* **66**, 484, <https://doi.org/10.1007/s00239-008-9102-8> (2008).
- Xia, X. The cost of wobble translation in fungal mitochondrial genomes: integration of two traditional hypotheses. *BMC Evolutionary Biology* **8**, 211, <https://doi.org/10.1186/1471-2148-8-211> (2008).
- van Wieringh, A. *et al.* HIV-1 modulates the tRNA pool to improve translation efficiency. *Molecular biology and evolution* **28**, 1827–1834 (2011).
- Chithambaram, S., Prabhakaran, R. & Xia, X. Differential codon adaptation between dsDNA and ssDNA phages in *Escherichia coli*. *Molecular biology and evolution* **31**, 1606–1617, <https://doi.org/10.1093/molbev/msu087> (2014).
- Prabhakaran, R., Chithambaram, S. & Xia, X. *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 (2015).
- dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**, 5036–5044 (2004).
- Novoa, E. M., Pavon-Eternod, M. & Pan, T. & Ribas de Pouplana, L. A role for tRNA modifications in genome structure and codon usage. *Cell* **149**, 202–213 (2012).
- Rojas, J. *et al.* Codon usage revisited: Lack of correlation between codon usage and the number of tRNA genes in enterobacteria. *Biochemical and Biophysical Research Communications* **502**, 450–455, <https://doi.org/10.1016/j.bbrc.2018.05.168> (2018).
- Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F. & Sockett, R. E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res* **33**, 1141–1153 (2005).
- Bennetzen, J. L. & Hall, B. D. Codon selection in yeast. *J. Biol. Chem.* **257**, 3026–3031 (1982).
- Bulmer, M. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897–907 (1991).
- Rocha, E. P. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res* **14**, 2279–2286 (2004).
- Chan, P. P. & Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* **37**, 4 (2009).

26. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527 (2016).
27. Xia, X. ARSDA: A New Approach for Storing, Transmitting and Analyzing Transcriptomic Data. *G3: Genes[Genomes]Genetics* **7**, 3839–3848, <https://doi.org/10.1534/g3.117.300271> (2017).
28. Kodama, Y., Shumway, M. & Leinonen, R. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res* **40**, 18 (2012).
29. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res* **39**, 9 (2011).
30. Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D. & von Mering, C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**, 3163–3168, <https://doi.org/10.1002/pmic.201400441> (2015).
31. Cozen, A. E. *et al.* ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nat Methods* **12**, 879–884 (2015).
32. Zheng, G. *et al.* Efficient and quantitative high-throughput tRNA sequencing. *Nat Methods* **12**, 835–837 (2015).
33. Falnes, P. Ø., Bjørås, M., Aas, P. A., Sundheim, O. & Seeberg, E. Substrate specificities of bacterial and human AlkB proteins. *Nucleic acids research* **32**, 3456–3461, <https://doi.org/10.1093/nar/gkh655> (2004).
34. Falnes, P. O., Johansen, R. F. & Seeberg, E. AlkB-mediated oxidative demethylation reverses DNA damage in *Escherichia coli*. *Nature* **419**, 178–182 (2002).
35. Schulz, S., Perez-de-Mora, A., Engel, M., Munch, J. C. & Schloter, M. A comparative study of most probable number (MPN)-PCR vs. real-time-PCR for the measurement of abundance and assessment of diversity of alkB homologous genes in soil. *J Microbiol Methods* **80**, 295–298 (2010).
36. van den Born, E. *et al.* Bioinformatics and functional analysis define four distinct groups of AlkB DNA-dioxygenases in bacteria. *Nucleic Acids Res* **37**, 7124–7136 (2009).
37. Wang, L., Wang, W., Lai, Q. & Shao, Z. Gene diversity of CYP153A and AlkB alkane hydroxylases in oil-degrading bacteria isolated from the Atlantic Ocean. *Environmental microbiology* **12**, 1230–1242 (2010).
38. Gao, P. *et al.* An Exogenous Surfactant-Producing *Bacillus subtilis* Facilitates Indigenous Microbial Enhanced Oil Recovery. *Frontiers in microbiology* **7**, 186–186, <https://doi.org/10.3389/fmicb.2016.00186> (2016).
39. Nie, Y. *et al.* Diverse alkane hydroxylase genes in microorganisms and environments. *Scientific reports* **4** (2014).
40. Cassier-Chauvat, C., Veaudor, T. & Chauvat, F. Comparative Genomics of DNA Recombination and Repair in Cyanobacteria: Biotechnological Implications. *Frontiers in microbiology* **7**, 1809–1809, <https://doi.org/10.3389/fmicb.2016.01809> (2016).
41. van den Born, E. *et al.* Viral AlkB proteins repair RNA damage by oxidative demethylation. *Nucleic acids research* **36**, 5451–5461, <https://doi.org/10.1093/nar/gkn519> (2008).
42. Dong, H., Nilsson, L. & Kurland, C. G. Co-variation of tRNA Abundance and Codon Usage in *Escherichia coli* Different Growth Rates. *Journal of Molecular Biology* **260**, 649–663, <https://doi.org/10.1006/jmbi.1996.0428> (1996).
43. Pang, Y. L., Abo, R., Levine, S. S. & Dedon, P. C. Diverse cell stresses induce unique patterns of tRNA up- and down-regulation: tRNA-seq for quantifying changes in tRNA copy number. *Nucleic Acids Res* **42**, 27 (2014).
44. Loher, P., Telonis, A. G. & Rigoutsos, I. Accurate Profiling and Quantification of tRNA Fragments from RNA-Seq Data: A Vade Mecum for MINTmap. *Methods Mol Biol*, 7339–7332\_7316 (2018).
45. Sharp, P. M. & Li, W. H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* **24**, 28–38 (1986).
46. Ikemura, T. & Ozeki, H. Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. *Cold Spring Harb Symp Quant Biol* **2**, 1087–1097 (1983).
47. Grosjean, H. & Fiers, W. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18**, 199–209 (1982).
48. Ikemura, T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol* **158**, 573–597 (1982).
49. Nishimura, S. *Modified nucleosides and isoaccepting tRNA*. (MIT Press, 1978).
50. Weissenbach, J. & Dirheimer, G. Pairing properties of the methylester of 5-carboxymethyl uridine in the wobble position of yeast tRNA<sup>Arg3</sup>. *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis* **518**, 530–534, [https://doi.org/10.1016/0005-2787\(78\)90171-5](https://doi.org/10.1016/0005-2787(78)90171-5) (1978).
51. Avcilar-Kucukgoze, I. *et al.* Discharging tRNAs: a tug of war between translation and detoxification in *Escherichia coli*. *Nucleic Acids Res* **44**, 8324–8334 (2016).
52. Chen, D. & Texada, D. E. Low-usage codons and rare codons of *Escherichia coli*. *Gene Therapy and Molecular Biology* **10**, 1 (2006).
53. Dittmar, K. A., Mobley, E. M., Radek, A. J. & Pan, T. Exploring the regulation of tRNA distribution on the genomic scale. *J Mol Biol* **337**, 31–47 (2004).
54. Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).
55. Xia, X., Huang, H., Carullo, M., Betran, E. & Moriyama, E. N. Conflict between Translation Initiation and Elongation in Vertebrate Mitochondrial Genomes. *PLoS ONE* **2**, e227 (2007).
56. Xia, X. In *Evolution in the fast lane: Rapidly evolving genes and genetic systems* (eds Rama S. Singh, Jianping Xu, & Rob J. Kulathinal) 73–82 (Oxford University Press, 2012).
57. Xia, X. In *Bioinformatics and the Cell* 197–238. (Springer, Cham, 2018).
58. Prabhakaran, R., Chithambaram, S. & Xia, X. *Aeromonas* phages encode tRNAs for their overused codons. *Int J Comput Biol Drug Des* **7**, 168–182 (2014).
59. Hori, H. Methylated nucleosides in tRNA and tRNA methyltransferases. *Frontiers in genetics* **5**, 144–144, <https://doi.org/10.3389/fgene.2014.00144> (2014).
60. Xia, X. "Bioinformatics and Translation Elongation" in *Bioinformatics and the Cell*. 197–238 (Springer, Cham., 2018).
61. Elf, J., Nilsson, D., Tenson, T. & Ehrenberg, M. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* **300**, 1718–1722 (2003).
62. Duret, L. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* **12**, 640–649 (2002).
63. Muto, A. & Osawa, S. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci USA* **84**, 166–169 (1987).
64. Osawa, S. *et al.* Directional mutation pressure and transfer RNA in choice of the third nucleotide of synonymous two-codon sets. *Proc Natl Acad Sci USA* **85**, 1124–1128 (1988).
65. Yang, Z. & Nielsen, R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular biology and evolution* **25**, 568–579 (2008).
66. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic acids research* **44**, D184–D189, <https://doi.org/10.1093/nar/gkv1309> (2016).
67. dos Reis, M., Wernisch, L. & Savva, R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res* **31**, 6976–6985 (2003).
68. Andersson, G. E. & Sharp, P. M. Codon usage in the *Mycobacterium tuberculosis* complex. *Microbiology* **142**, 915–925 (1996).

69. Xia, X. DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. *Molecular biology and evolution* **35**, 1550–1552 (2018).
70. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10, <https://doi.org/10.14806/ej.17.1.200> (2011).
71. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
72. Williams, C. R., Baccarella, A., Parrish, J. Z. & Kim, C. C. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics* **17**, 103, <https://doi.org/10.1186/s12859-016-0956-2> (2016).
73. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175–185 (1998).
74. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352> (2009).
75. Ernst, F. G. M. *et al.* Cold adaptation of tRNA nucleotidyltransferases: A tradeoff in activity, stability and fidelity. *RNA Biol* **15**, 144–155 (2018).

## Acknowledgements

This work was supported by the Discovery Grant of Natural Science and Engineering Research Council of Canada to X.X. (NSERC, RGPIN/2018-03878), and the Ontario Graduate Scholarship 2018-2019 to Y.W.

## Author Contributions

Y.W., J.R.S. and X.X. designed the study and wrote the main manuscript text. Y.W. and J.R.S. collected and analyzed the data. X.X. developed the computer programs and supervised the study. All authors reviewed the manuscript. X.X. supervised the project.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-019-39369-x>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019