INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Liu SS, Hockenberry AJ, Jewett MC, Amaral LAN. 2018 A novel framework for evaluating the performance of codon usage bias metrics. *J. R. Soc. Interface* **15**: 20170667.
http://dx.doi.org/10.1098/rsif.2017.0667

Received: 11 September 2017
Accepted: 4 January 2018

Subject Category:
Life Sciences–Physics interface

Subject Areas:
bioinformatics, biophysics, computational biology

Keywords:
codon usage bias, theoretical benchmarking, translational regulation

Authors for correspondence:
Michael C. Jewett
e-mail: m-jewett@northwestern.edu
Luís A. N. Amaral
e-mail: amaral@northwestern.edu

†These authors contributed equally to this work.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3986280.v1.

THE ROYAL SOCIETY PUBLISHING

A novel framework for evaluating the performance of codon usage bias metrics

Sophia S. Liu[1,†], Adam J. Hockenberry[1,2,†], Michael C. Jewett[1,2,3,4,5] and Luís A. N. Amaral[1,5,6]

[1]Department of Chemical and Biological Engineering, [2]Interdisciplinary Program in Biological Sciences, [3]Center for Synthetic Biology, [4]Simpson Querrey BioNanotechnology Institute, [5]Northwestern Institute on Complex Systems, and [6]Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA

SSL, 0000-0003-2340-3008; AJH, 0000-0001-9476-0104; MCJ, 0000-0003-2948-6211; LANA, 0000-0002-3762-789X

The unequal utilization of synonymous codons affects numerous cellular processes including translation rates, protein folding and mRNA degradation. In order to understand the biological impact of variable codon usage bias (CUB) between genes and genomes, it is crucial to be able to accurately measure CUB for a given sequence. A large number of metrics have been developed for this purpose, but there is currently no way of systematically testing the accuracy of individual metrics or knowing whether metrics provide consistent results. This lack of standardization can result in false-positive and false-negative findings if underpowered or inaccurate metrics are applied as tools for discovery. Here, we show that the choice of CUB metric impacts both the significance and measured effect sizes in numerous empirical datasets, raising questions about the generality of findings in published research. To bring about standardization, we developed a novel method to create synthetic protein-coding DNA sequences according to different models of codon usage. We use these benchmark sequences to identify the most accurate and robust metrics with regard to sequence length, GC content and amino acid heterogeneity. Finally, we show how our benchmark can aid the development of new metrics by providing feedback on its performance compared to the state of the art.

1. Introduction

As many as six different synonymous codons can be used to code for a single amino acid in protein-coding genes. However, these synonymous codons may be translated with varying degrees of speed and accuracy owing to different tRNA concentrations and interactions with the ribosome [1–8]. Unequal synonymous codon utilization has important consequences for a variety of processes, including mRNA degradation and translation, protein folding, horizontal gene transfer and viral resistance [9–16]. Additionally, codon usage optimization is a widely used strategy to engineer protein-coding sequences for increased expression, and CUB has also been used to decrease the expression of viral genes to aid in vaccine development [17–20].

Given its biological significance, it is not surprising that a variety of metrics have been developed to quantify the level of CUB in a genetic sequence. These metrics fall primarily into two classes: (i) those that require knowledge of highly expressed genes, preferred codons, genomic-context or tRNA copy numbers/modification patterns, and (ii) those that calculate deviations from random expectation [6,21–23]. Here, we focus on metrics from the latter class.

One of the first metrics proposed to quantify CUB was the 'effective number of codons', N_C, which is based on the concept of heterozygosity in population genetics [23]. Over the years, a number of variations of N_C have been proposed to address its theoretical and practical shortcomings [24–30]. Meanwhile, researchers have developed a number of other CUB metrics that take into account various facets of coding sequences, such as uneven amino acid distributions and GC contents [31–33].

Few studies have attempted to systematically evaluate the performance of these different metrics [22]. Most commonly, researchers proposing new metrics focus on the ability to identify highly expressed genes in particular datasets. As we demonstrate below, this measure of performance is problematic. Furthermore, it has not been demonstrated whether a metric is measuring solely codon usage bias (CUB) or accounting for other features of coding sequences. This caveat is due to the fact that there has been little theoretical benchmarking to test performance limitations with regard to sequence length, GC content and amino acid composition—all important features of real genes that may obscure the signature resulting from CUB.

Here, we introduce a rigorous framework for evaluating the performance of CUB metrics comprising of benchmarking tests with synthetic 'ground-truth' datasets. We extend a recently published maximum entropy framework to generate synthetic random coding sequences with known levels of CUB, and then test the ability of metrics to differentiate between sets of sequences under increasingly realistic constraints [34]. We run six different CUB metrics through our benchmarking pipeline: three variations of the effective number of codons (N_C, N_C' and N_C^+) [23–25] and three other metrics (Relative Codon Bias Score (RCBS) [33], Codon Deviation Coefficient (CDC) [31] and Synonymous Codon Usage Score (SCUO) [32]). These metrics were selected because they represent a diversity of approaches and are highly cited and/or recently published. We have developed a python package that allows researchers to rapidly test the performance of novel CUB metrics and to compare results against a selected set of metrics, which is also available at https://github.com/amarallab/cub_benchmarking.

2. Material and Methods

2.1. Generation of random coding sequence with known codon usage bias

The amount of CUB in a sequence can be intuitively expressed through the number of codons used to construct the sequence. Given no prior knowledge, it is safe to assume that a sequence that uses all 61 codons is less biased than a sequence that uses only 20 codons, one for each amino acid.

In addition to controlling for the number of different codons that are used in each sequence, we also incorporated sequence attributes such as length, amino acid content and GC content to investigate how combinations of various sequence attributes affect how metrics calculate CUB. To ensure that no additional bias is incorporated into the sequences when controlling for these attributes, we have devised a mathematical formulation controlling how each sequence is created. Table 1 is a list of variable definitions that will be used.

2.1.1. Creating condensed translation tables

To generate sequences with known CUB, we first create condensed translation tables, where codons have been randomly chosen to be removed from the standard translation table. For example, to generate sequences that uses 40 codons, we would randomly select 21 codons to be removed from the standard translation table. If for a synthetic translation table, an amino acid turns out to have 0 sense codons, then we discard that translation table and draw a new one.

Table 1. Definitions of variables.

variable	definition
A_i	ith amino acid in a given translation table
c_{ij}	jth synonymous codon of A_i
p_{ij}	probability of picking c_{ij}
$p(c_{ij} \mid A_i)$	probability of picking c_{ij} given A_i
$p(A_i)$	probability of observing A_i
k_i	number of synonymous codons that code for A_i
K	total number of sense codons, $K = \sum_{i=1}^{20} k_i$
E_{ij}	number of G/C nucleotides in c_{ij}
f_{GC}	fraction of nucleotides that are G or C
n_{A_i}	number of times A_i is observed in the gene or sequence
L	number of codons in the sequence

2.1.2. No constraints

These sequences are designed to simulate an ideal case of very long coding sequences (2000 codons) that are not constrained by either defined amino acid probabilities or G/C nucleotides. In this simplest case, 2000 codon long sequences (not including stop and start codons) are generated by randomly sampling the codons from a given condensed translation table with replacement according to the following probability:

$$p_{ij} = \frac{1}{K}. \tag{2.1}$$

This process is repeated 1000 times for all 42 CUB levels. We expect the sequences generated under these perfectly random conditions to be easy to discriminate.

2.1.3. Length constraints

To generate sequences in which we control for the length of the sequences, we first randomly choose a sequence length from the distribution of gene lengths in the *E. coli* genome. A condensed translation table is created for a given CUB level and a random sequence is generated by randomly sampling the codons in the condensed translation table according to the probabilities in equation (2.1). This process is repeated 1000 times for all 42 CUB levels. The sequences generated under these conditions will elucidate how CUB metrics deal with the variability in sequence length that is present in real sequences.

2.1.4. Amino acid constraints

To generate sequences in which we control for the amino acid content of the translated sequence, we first randomly choose a target amino acid content from a population of sets of amino acid probabilities. These sets of amino acid probabilities are calculated for each gene in the *E. coli* genome, where $p(A_i) = n_{A_i}/L$. A condensed translation table is created for a given CUB level and a random sequence that is 2000 codons in length is generated by randomly sampling codons from the condensed translation table according to the probability

$$p_{ij} = \frac{1}{k_i} p(A_i). \tag{2.2}$$

This process is repeated 1000 times for all 42 CUB levels. The sequences generated under these conditions will elucidate how CUB metrics deal with the variability in the amino acid content that is present in real sequences.

2.1.5. GC constraints

To generate sequences in which we control for the number of G/C nucleotides in the sequence, we first randomly choose a target GC content from the GC content distribution for all genes in the *E. coli* genome. A condensed translation table is created for a given CUB level and a random sequence that is 2000 codons in length was generated by randomly sampling with replacement codons from the condensed translation table according to the probability

$$p_{ij} = \frac{1}{Z}\exp(-\beta E_{ij}), \tag{2.3}$$

where Z is a normalization factor so that $\sum_i \sum_j p_{ij} = 1$, and β is a constant that fulfills the following equation:

$$f_{GC} = \frac{1}{3}\sum_{i=1}^{20}\sum_{j=1}^{k_i} E_{ij} p_{ij}. \tag{2.4}$$

This process is repeated 1000 times for all 42 CUB levels. Here, we also use artificially long sequences (2000 codons in length) to minimize the uncertainty associated with the estimator (the observed probability of using each codon in the sequence). The maximum entropy approach generates sequences with GC contents with extremely small variance (0.0001); thus, all sequences created are accepted. The sequences generated under these conditions will elucidate how CUB metrics deal with the GC content variability that is present in real sequences.

2.1.6. Amino acid and length constraints

To generate sequences that control for the amino acid content and sequence length, we first randomly choose a target amino acid content from a population of sets of amino acid probabilities for all genes in the *E. coli* genome and a sequence length from the distribution of gene lengths in the *E. coli* genome. A condensed translation table is created for a given CUB level and a random sequence was generated by randomly sampling with replacement codons from the condensed translation table according to the probability given by equation (2.2). This process is repeated 1000 times for all 42 CUB levels. The sequences generated under these conditions will elucidate how CUB metrics deal with simultaneous variability in the amino acid content and sequence length that is present in real sequences.

2.1.7. GC content and length constraints

To generate sequences that control for the GC content and sequence length, we first randomly choose a target GC content from the GC content distribution for all genes in the *E. coli* genome and a target sequence length from the distribution of gene lengths in the *E. coli* genome. A condensed translation table is created for a given CUB level and a random sequence was generated by randomly sampling with replacement codons from the condensed translation table according to the probability given by equation (2.3). This process is repeated 1000 times for all 42 CUB levels. The sequences generated under these conditions will elucidate how CUB metrics deal with simultaneous variability in the amino acid content and sequence length that is present in real sequences.

2.1.8. Amino acid and GC constraints

To generate sequences that control for both the amino acid content and GC content of the sequence, we first randomly choose a target amino acid content from a population of sets of amino acid probabilities for all genes in the *E. coli* genome and randomly choose a target GC content from the GC content distribution for all genes in the *E. coli* genome. A condensed translation table is then created for a given CUB level and a random sequence that is 1000 in length is generated by randomly sampling codons from the condensed translation table according to the probability given by the following equation:

$$p(c_{ij}) = p(c_{ij} \mid A_i)p(A_i), \tag{2.5}$$

where

$$p(c_{ij} \mid A_i) = \frac{1}{Z}\exp(-\beta E_{ij}), \tag{2.6}$$

where Z is a normalization factor so that $\sum_j p(c_{ij} \mid A_i) = 1$ for a given set of synonymous codons, and β is a constant that fulfills the following equation:

$$f_{GC} = \frac{1}{3}\sum_{i=1}^{20} p(A_i)\sum_{j=1}^{k_i} E_{ij} p(c_{ij} \mid A_i). \tag{2.7}$$

This process is repeated 1000 times for all 42 CUB levels.

2.1.9. Organismal genome

To combine all of these features together and create a challenging discriminative task, we simulate sequences where sequence length, GC content and amino acid contents are constrained simultaneously. This essentially codes all genes in the organism's genome with sequences of known levels of CUB. In order to accomplish this, a condensed translation table of known CUB level is created for each gene and one random sequence is created in accordance with the length, GC content primary amino acid sequence of that gene. That is to say that for a given gene and condensed translation table, a synonymous codon is chosen for each amino acid in the gene according to the probability given by equation (2.6). This process is repeated for all protein-coding genes in an organism and for all 42 levels of CUB.

Here, we are interested in asking whether a metric can determine, for instance, whether a 300 amino acid long, 55% GC content protein composed of mainly hydrophobic residues uses more or fewer codons than a 150 amino acid long, 48% GC content, protein composed of mainly hydrophilic residues.

2.1.10. Combined

In order to simulate a situation where sequences may come not from a single defined organism but rather from a community, we combined all sequences created in the organismal genomes into one large dataset with sequences that have properties of each of the three genomes. For each CUB level, the sequences that correspond to that CUB level in the *E. coli*, *B. subtilis* and *S. coelicor* datasets are aggregated into one larger dataset. This task is particularly challenging because the organisms in question come from three very different degrees of GC content making the spread on GC contents far more variable than in the 'Organismal Genome' task.

2.2. Performance of codon usage bias metrics

2.2.1. Calculating theoretical performance

For each constraint, we have generated 1000 (unless otherwise stated) random sequences for each of the 42 CUB levels. A metric's performance is determined by its ability to differentiate sequences with two different CUB levels.

Specifically, for a given constraint, 1000 random sequences were generated for each CUB level. We then chose two CUB levels and for each sequence in the two sampled sets, we measured the CUB level using one of the metrics in question. Based on the measured CUB level, the metric's ability to classify the sequences into the correct 'high codon bias' and 'low codon bias' category was then determined using AUC. The process was repeated to get the *AUC* for all 861 pairs of CUB levels (electronic supplementary material, figures S2 and S3). To summarize this information, the average of the AUCs ($\overline{AUC}$) for all pairs of

CUB level was calculated and linearly transformed to a 0 to 1 scale:

$$\text{Score} = 2\overline{AUC} - 1. \tag{2.8}$$

This score quantifies a metric's ability to correctly identify two sets of sequences of known CUB under a given constraint. The entire process was repeated 10 times and the average and standard deviation of the score was determined. The entire process was then repeated for all constraints and all CUB metrics to yield the results are presented in figure 4.

2.2.2. Calculating Δ_{rg}

Δ_{rg} is a genome-wide measure that quantifies the relative difference between the CUB of the ribosomal genes and the CUB in the genome. This relationship is described using the following equation:

$$\Delta_{rg} = \frac{B_g - B_r}{B_g}, \tag{2.9}$$

where B_g is the calculated CUB of the concatenated sequence of all protein-coding genes in the genome and B_r is the calculated CUB of the concatenated sequence of the ribosomal genes in the genome.

2.3. Data assembly

2.3.1. Organismal minimum growth rate

We first assembled a database of prokaryotic genomes from NCBI using the GBProks software (https://github.com/hyattpd/gbproks), including only 'complete' genomes in our download and subsequent analysis (accessed 10 March 2016). For the data on minimum doubling time, we downloaded the data table from Vieira-Silva *et al.* [35], and paired each bacterial species with a complete genome from our database, resulting in 187 data points. To control for shared ancestry in subsequent analyses, we constructed a phylogenetic tree based on the rRNA sequences for this set of species. We first used RNAmmer to extract the 16S and 23S rRNA sequences, followed by MUSCLE (v. 3.8.31) on each individual rRNA to produce a multiple-sequence alignment [36]. These were concatenated together and we conducted a partitioned analysis using RAxML to construct a final tree. We performed 100 rapid Bootstrap searches, 20 maximum-likelihood searches and selected the best maximum-likelihood tree for subsequent analysis [37].

2.3.2. tRNA gene copy number

For the larger dataset of tRNA gene copy numbers, we relied on a previously computed high-quality dataset published by Hug *et al.* [38]. We used custom scripts to match entries in this tree with genomes from our complete-genome database, and pruned all species without a high-quality match resulting in 618 species in our final dataset for subsequent analyses. We ran tRNAscan-SE on each of these genomes to calculate the number of tRNA genes [39].

2.4. Phylogenetically generalized least-squares analysis

We use phylogenetically generalized least-squares (PGLS) regression in order to mitigate the effects of shared ancestry in statistical analyses relating to growth rates and tRNA abundances. Our PGLS analysis relies on the most common null model, which assumes a Brownian motion model of trait evolution. For all statistical analyses presented in the paper, we use the R package 'caper' and perform a simultaneous maximum-likelihood estimate of Pagel's λ, a branch length transformation, alongside the coefficients for independent variables of interest in order to control for false-positive and false-negative rates.

2.5. Information-based codon usage bias

Information theory provides a solid framework that allows us to quantify the amount of information in a message by using Shannon's information entropy which is described by the following equation:

$$H = -\sum_i p_i \log_2 p_i, \tag{2.10}$$

where p_i is the probability of the occurrence of the ith source symbol.

This means that if we take a sequence of ones and zeros, in which ones appear with a probability of $p(1)$ and zeros appear with a probability of $p(0)$, the information entropy of the sequence will be $H = -p(0)\log_2 p(0) - p(1)\log_2 p(1)$.

To determine the amount of information entropy within the coding of an amino acid, we take the framework given in equation (2.10) to yield the following:

$$H_{A_i} = -\sum_{j=1}^{k_i} p(c_{ij} \mid A_i) \log_2 p(c_{ij} \mid A_i). \tag{2.11}$$

The information entropy for all $\{H_{A_i}\}$ can be aggregated into the information entropy of the entire gene given by

$$H_g = \sum_{i=1}^{20} p(A_i) H_{A_i}. \tag{2.12}$$

However, this only gives us the raw information entropy of a gene. In order to gain useful information from this, H_g should be compared to the maximum possible information entropy of a gene (H_n) given the various constraints including GC content.

The maximum possible information entropy of a gene can be determined by many different ways. Given no additional information and no further constraints, we assume that all codons are used equally for a given amino acid. Thus, H_n can be defined as follows:

$$H_n = \sum_{i=1}^{20} p(A_i) \log_2 k_i. \tag{2.13}$$

The problem with this definition is that genes have specific GC and amino acid content requirements they need to fulfil in order to ensure their functionality. As a result, Hn as given by equation (2.13) is not an appropriate null model to compare against H_g. Instead, we propose that H_g be compared against the information entropy of a random nucleotide sequence that fulfils the GC and amino acid contents of the gene. That is to say $p(c_{ij} \mid A_i)$, should obey equations (2.6) and (2.7).

Using the $\{p(c_{ij} \mid A_i)\}$ from equation (2.6) to determine H_n, we define the raw CUB score as follows:

$$S_g = \frac{H_g}{H_n}. \tag{2.14}$$

We define iCUB on a 20–61 scale, similar to the implementation of the effective number of codons. As such the final value of iCUB is given by

$$\text{iCUB} = 20 + S_g(61 - 20). \tag{2.15}$$

3. Results and discussion

3.1. Broad range of correlations among metrics

All CUB metrics aim to quantify the level of CUB within protein-coding sequences. Therefore, if a coding sequence

can be said to have a given level of CUB, one would expect the measurements obtained with different metrics for the same set of sequences to be highly correlated despite slight differences in their underlying methodologies.

To test this expectation, we measured the CUB of 3740 coding sequences in the *E. coli* genome using the six CUB metrics listed earlier. One of the most well-cited and frequently used metrics is the previously mentioned effective number of codons (N_C) [23]. Subsequently, a variant of N_C was proposed that explicitly accounts for GC bias (N_C') [24], and a more recent implementation of N_C ($N_C^{\dagger}$) purports to alleviate a number of theoretical shortcomings that are apparent in the original method [25]. In addition to these three metrics that are all based on the principle of the effective number of codons, we chose three more metrics to include in our analysis: RCBS [33], CDC [31] and SCUO [32] (see electronic supplementary material text for a more thorough discussion of each method including mathematical derivations).

Overall, we observe that these six metrics show a surprisingly broad range of correlation values with one another—$\rho \in [0.36, 0.92]$ (electronic supplementary material, figure S1). The wide range of correlations between measured CUB points to several non-exclusive scenarios: (i) some or all CUB metrics have large systematic biases due to, for example, not taking into account GC content, or (ii) some or all CUB metrics have large random measurement errors due to, for example short sequence length.

3.2. Correlating codon usage bias to gene expression

One of the most striking and reproducible findings obtained with CUB metrics is the observation that highly expressed genes exhibit higher levels of CUB. For this reason, correlations between CUB and endogenous gene expression have been widely used as a proxy in determining the ability of a metric to measure CUB. Indeed, a number of experimental investigations have attempted to quantify the latter question [40–42]. However, errors associated with gene expression measurements and the fact that one does not know the true magnitude of the impact that CUB has on gene expression, make it very difficult to extract uncontroversial conclusions from this approach.

One of the pitfalls of using correlations with gene expression to determine the performance of CUB metrics is that it is very susceptible to which organisms and datasets are chosen for the evaluation. To illustrate this point, we evaluated the correlation between individual CUB metrics and bacterial protein abundances for the 26 bacterial species collected by Wang *et al.* [43]. We found that, depending on the organism that is chosen for the evaluation, one can reach markedly different conclusions. For example, if one considers *E. coli*, the conclusion that could be drawn is that all metrics exhibit a similar degree of performance, with N_C' only slightly 'outperforming' the other metrics (figure 1*a*). However, if one were to choose *B. subtilis* instead, one would conclude that RCBS, SCUO and CDC are 'superior' CUB metrics (figure 1*b*).

A way to eliminate the bias associated with poorly chosen datasets is to look at the strength of the correlation across diverse species. One could make the case that the metric that exhibits the strongest average correlations across a set of diverse species is the better metric. However, our analysis demonstrated that this approach also does not yield conclusive results concerning the performance of CUB metrics. We found that for a given metric, the strength of the correlation between CUB and protein abundances varies markedly across species (figure 1*c*). Overall, N_C' exhibits the highest median correlation. The 'performance' of this metric, however, is not significantly higher than that of RCBS or SCUO (Wilcoxon signed-rank test, $p > 0.05$). Indeed, of the 15 different pair-wise comparisons between the six different metrics, only five yielded statistically significant results (electronic supplementary material, table S1).

The large degree of heterogeneity in the correlation coefficients could arise from uncertainties associated with measuring both CUB and protein abundances, and of the variability associated with the underlying biological mechanisms that control gene expression in different species and conditions. Regardless of the underlying causes, the large heterogeneity that we observe makes it difficult to determine which metric exhibits the highest performance.

3.3. Evaluation of biological findings that use codon usage bias

Without understanding what a specific CUB metric is truly measuring, biological findings that stem from the use of these measurements are limited. Here, we provide two examples where biological findings could potentially depend on which metric is chosen.

Vieira-Silva *et al.* [35] reported that the relative difference in CUB between ribosomal protein-coding genes and the rest of the genome (Δ_{rg}) is highly predictive of the minimum doubling time for 187 bacterial species. Vieira-Silva *et al.* [35] concatenated sets of genes into two groups so that biases associated with short sequence lengths would be eliminated, and measured CUB using N_C'. We repeat this analysis for the set of six CUB metrics, while also performing phylogenetically generalized least-squares (PGLS) regression to account for shared ancestry. Figure 2*a* shows that there is substantial variation in the performance of these metrics, with N_C' and RCBS clearly showing a stronger correlation between Δ_{rg} and growth rates than the other metrics. In the most extreme case, there is a sixfold difference in the magnitude of correlation observed for the lowest performing metric (CDC) and the highest (N_C').

In another study, Rocha [2] showed that CUB and tRNA gene copy numbers are correlated with one another, suggesting co-evolution between codon preference and the tRNA anti-codon pool. We repeated this analysis using PGLS regression for the set of CUB metrics, and again found substantial variation in the strength of the underlying correlation, with N_C' and RCBS exhibiting the highest correlations (figure 2*b*). In this case, there is a fourfold change in magnitude between the lowest correlating metrics ($N_C^{\dagger}$) and the highest correlating metric (RCBS).

This lack of replicability puts into question the use of CUB metrics for the purpose of drawing biological conclusions. To know whether or not an association exists, and to determine the overall size of these effects, one must have a rigorous understanding of the accuracy of a given metric. In order to rectify the current lack of rigour in determining metric accuracy, we sought to develop a series of benchmark tests that are able to determine the strengths and limitations of CUB metrics.

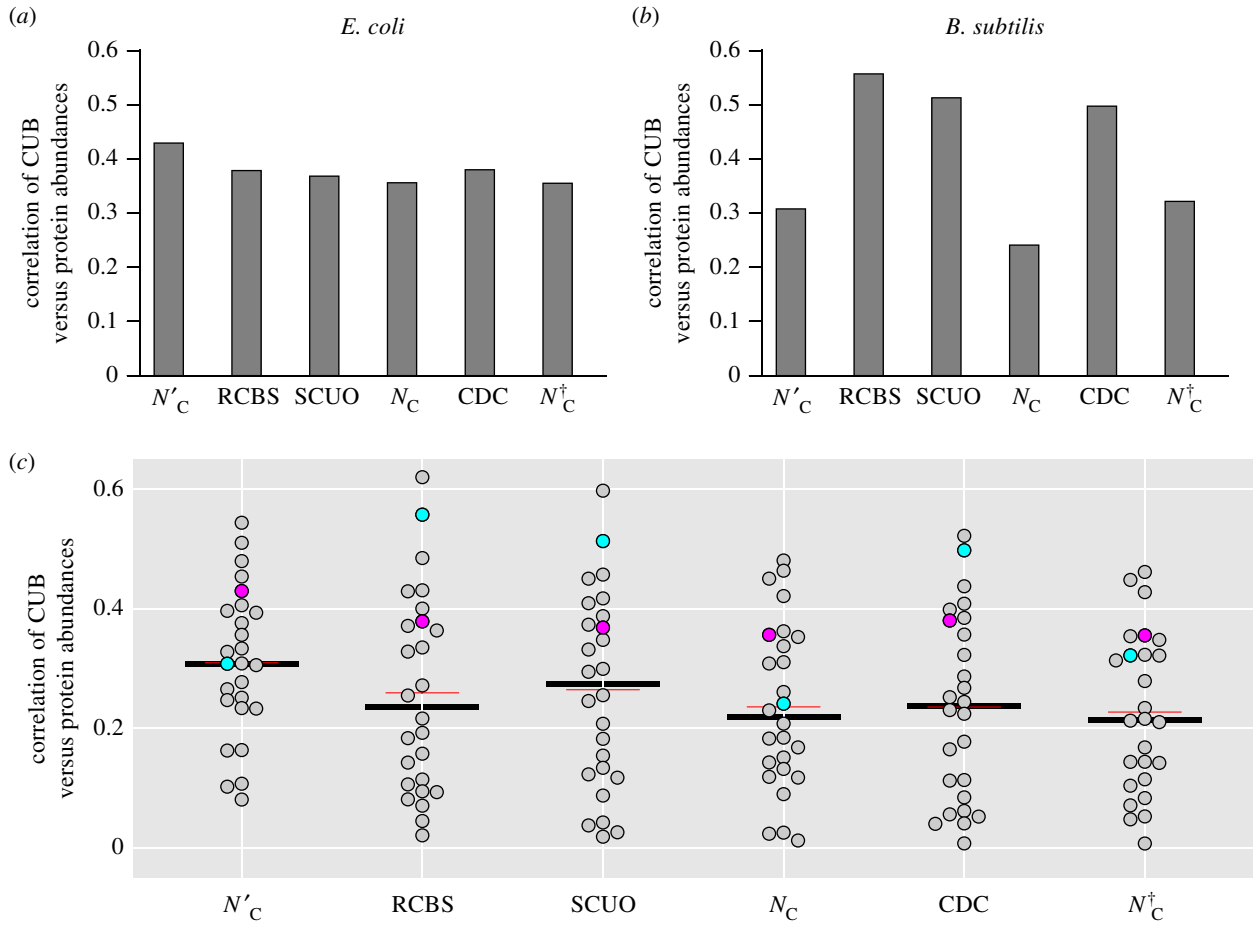

Figure 1. Correlations of gene expression with CUB for a select set of metrics. Comparison of Spearman's ρ correlation coefficients between the CUB and protein abundances for (*a*) *E. coli* and (*b*) *B. subtilis*. All metrics have similar magnitude of correlation coefficients for *E. coli*. RCBS, SCUO and CDC yield higher correlation coefficients than N'_C, N_C and $N^{\dagger}_C$ for *B. subtilis*. (*c*) Comparison of Spearman's ρ correlation coefficients between the CUB and protein abundances for genes for 26 bacterial genomes. Thick black bars indicate the median and thin red bars indicate the mean. Fuchsia and blue circles highlight *E. coli* and *B. subtilis*, respectively.

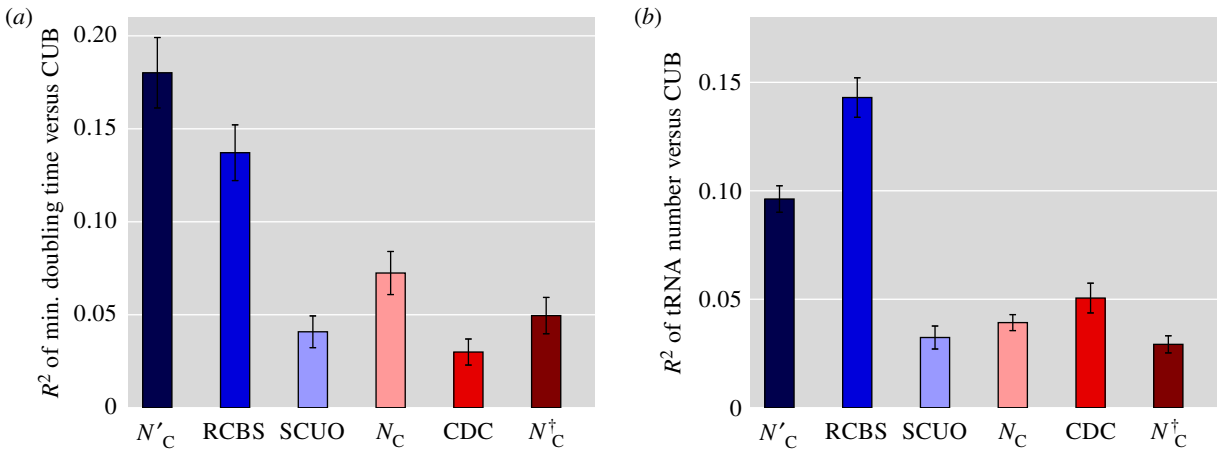

Figure 2. Biological findings can vary depending on the CUB metric that is used. Comparison of R^2 values obtained using PGLS regression of Δ_{rg} against (*a*) the minimum doubling times of 187 bacterial species and (*b*) the tRNA copy number in the genomes of 618 different bacterial species. The height of each bar reflects the average of 10 bootstrap samples and the error bars reflect the standard error of the estimated values.

3.4. Ground truth benchmarking

We use ground truth benchmarking to determine whether a metric is accurately measuring the level of CUB. We developed a pipeline that tests the performance of individual metrics on synthetic sequences with *a priori* known levels of CUB (figure 3). This approach allows us to test metrics in a controlled manner that accounts for changes in confounding features.

In addition to different levels of CUB, different genes within an organism have variable lengths, amino acid utilization and GC contents, all of which will pose measurement challenges that can affect the accuracy of a CUB metric

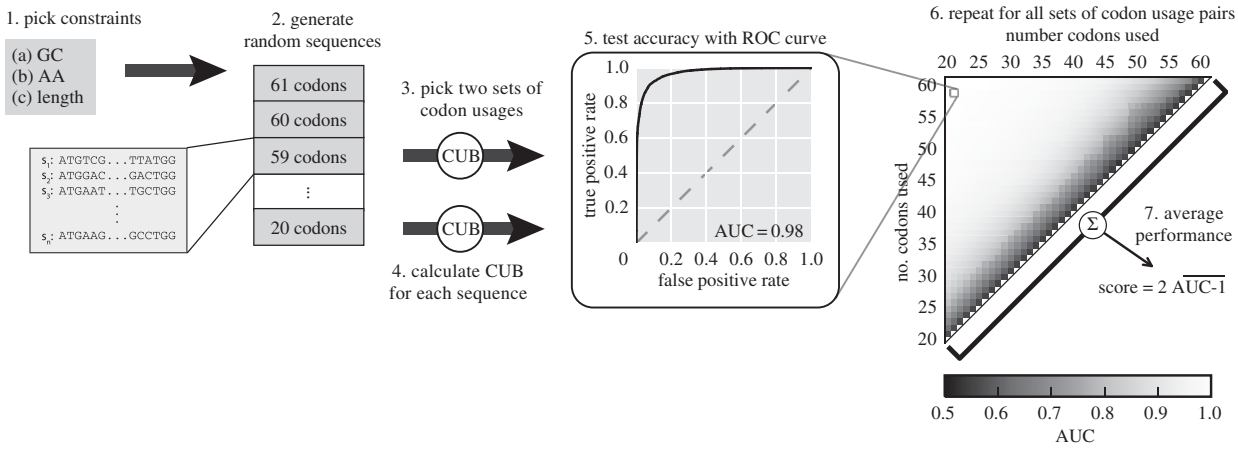

Figure 3. Work flow for the evaluation of performance of CUB metrics. A set of constraints are chosen and 42 sets of 1000 random sequences are generated, each set is characterized by a different level of CUB, which we impose by constraining the number of codons in the translation table. Two sets of random sequences are chosen and the CUB of all sequences in the sets is calculated. The AUC of the ROC curve is determined based on the metric's ability to categorize the two sets of sequences. The AUC of the ROC is calculated for all 861 codon usage pairs. The overall performance of the metric is captured by taking the average of all AUCs and normalizing the result to a 0–1 scale.

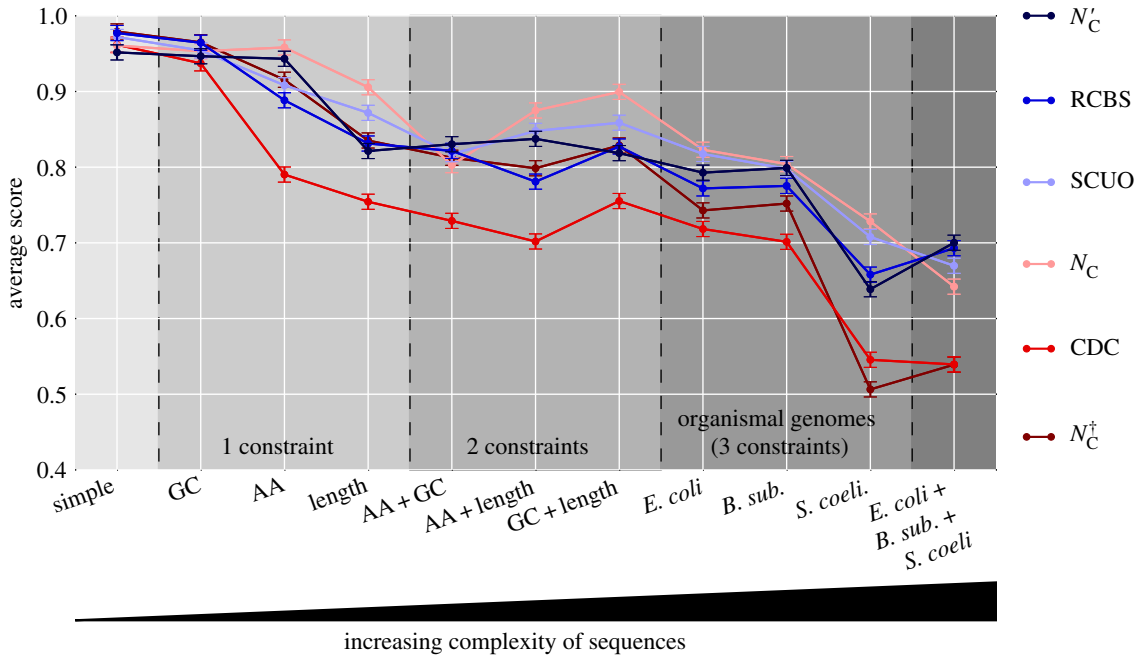

Figure 4. Ground truth benchmarking for a representative set of CUB metrics. All metrics perform similarly well when sequences do not incorporate additional constraints; however, their performance drops as more constraints are incorporated in the testing sequences. The error bars represent the estimated standard deviation associated with each value.

[44,45]. While these features may covary in real coding sequences (i.e. short genes may have more pronounced CUB or skewed GC contents, etc.), we develop an approach for the evaluation of synthetic sequences that allows us to study the impact of each factor in isolation.

We first choose a set of constraints to impose on synthetically generated random sequences in order to test whether methods are capable of isolating signals of CUB from other biases that may confound calculations in real sequences. Here, we evaluate null models to explicitly test how variability in GC content, amino acid usages and gene lengths can affect the accuracy of CUB metrics. These null models we consider are not meant to be exhaustive, and future work that evaluates more complex constraints such as dinucleotide biases, codon pair biases, position-dependent signals, avoidance of particular sequence motifs [46–52], etc., may provide further insight into the performance of different metrics. For each of the null models that we consider, we generate 42 unique sets of sequences with increasing CUB by progressively removing codons from the available genetic code, and apply a maximum entropy approach to generate random sequences according to the different constraints (see Material and methods). We use this extreme version of CUB, where entire codons are absent from the available genetic code, as a simplified test case. We pursued this approach because in order to simulate sequences with more subtle 'degrees' of CUB, we would first have to define those degrees somehow according to a metric, creating a tautology and favouring whatever metric was used to simulate sequences.

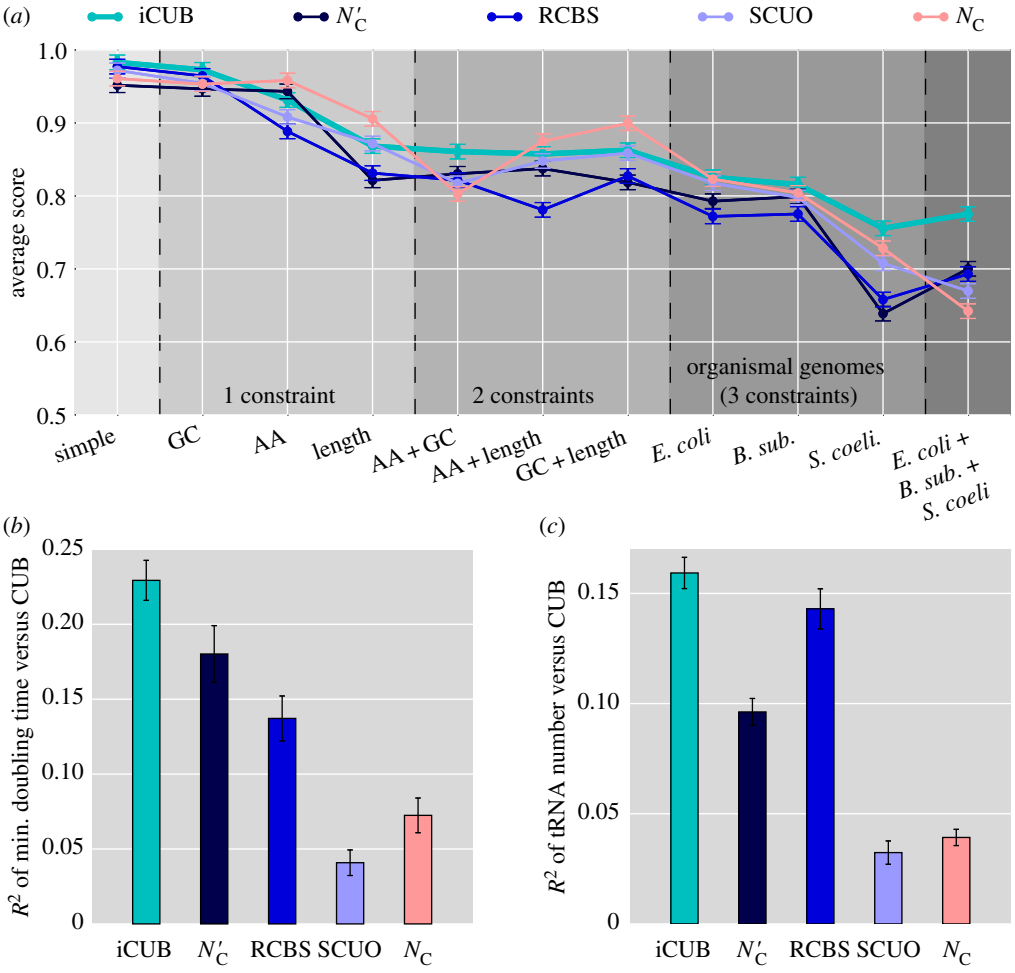

Figure 5. Output from benchmarking package comparing the new method (iCUB) with the best four current metrics. (*a*) Ground truth theoretical benchmarking results. The estimated standard deviation is smaller than 0.01 and is denoted by the error bars. (*b*) Comparison of R^2 values derived from PGLS regression of Δ_{pg} against the minimum doubling time of 187 bacterial species. The height of each bar reflects the average of 10 bootstrap samples and the error bars reflect the standard error of the estimated values. (*c*) Comparison of R^2 values derived from PGLS regression of Δ_{pg} against the tRNA copy number in the genomes of 618 different bacterial species. The height of each bar reflects the average of 10 bootstrap samples and the error bars reflect the standard error of the estimated values.

For a given metric, we then compare its ability to discriminate between two sets of coding sequences generated with differentially constrained genetic codes. We use the area under the curve (AUC) of the receiver operating characteristic as an assessment of a metric's ability to correctly classify items from the two sequence sets [53]. To summarize the theoretical performance of a given metric, we repeat this process for all 861 unique pairs of sets of sequences, take the average of all the AUCs and convert it to a 0 to 1 scale (see Material and methods; figure 4).

We find that all metrics perform equally well on the simplest cases of sequences with no imposed GC or amino acid constraints and with artificially long sequences to reduce the sampling variability (figure 4). Predictably, metrics perform much better at discriminating between genes with highly disparate levels of CUB; all metrics are able to perfectly discriminate (AUC=1) between sets of sequences constructed using only 20 codons from those constructed using all 61 codons (electronic supplementary material, figures S2 and S3). By contrast, the ability to discriminate codon usage differences of 1 (such as 40 versus 41 or 55 versus 56) is a more challenging task (AUC $\approx$ 0.55).

The performance of all CUB metrics decreases as the synthetic sequences more closely approximate real gene sequences by incorporating amino acid, length and GC heterogeneity—illustrating the difficulty that existing metrics have in isolating the impact of CUB (see Material and methods). Specifically, we observe that there is a notable drop in performance for all metrics when length constraints are introduced. This indicates that accounting for the sampling variability that arises from short sequences is one of the most difficult tasks associated with measuring CUB. There is a fundamental limit with regard to sequence length past which no metric could be expected to accurately estimate CUB (i.e. a 20 amino acid long sequence). Rather than drawing a hard threshold for required gene sequence length, researchers should be aware that confidence in CUB estimates for *a particular method* is strongly tied to the observable sequence length. Individual methods, however, are differently able to address the confounding effect of sequence length.

Overall, N'_C, RCBS, SCUO and N_C have approximately similar performances in the most difficult tasks, while the average performance of N^+_C and CDC is substantially lower. The two best performing metrics on the most challenging task, which incorporates a wide degree of GC heterogeneity (N'_C and RCBS), both explicitly control for the underlying nucleotide content of the sequences in their calculations.

3.5. Implementation of benchmarking framework

Developing a pipeline for the benchmarking of CUB metrics is essential not only for the assessment of the current state of

the field but also for the continued advancement of CUB metrics. To demonstrate how this pipeline could potentially be used to evaluate the efficacy of new metrics, we present here a new metric and ran it through our benchmarking pipeline in order to compare its performance against existing metrics. Our method—termed information-based codon usage bias (iCUB)—relies on formulating CUB in the framework of information entropy, while explicitly controlling for the GC and amino acid contents of the sequence (see Material and methods).

Our benchmarking package automatically compares a new CUB metric (in this case iCUB) with the four most accurate metrics in the literature and outputs a figure with the comparison results (figure 5). This comparison will allow researchers to instantaneously gauge the performance of their metric against the current state of the art.

iCUB performs comparably if not better than other metrics under most conditions. Nevertheless, performance in the most stringent benchmarking test still shows considerable room for improvement (average score ≈ 0.78 compared to theoretically perfect performance of 1). Interestingly, analysis with iCUB suggests that the correlation between minimum doubling time and CUB is stronger than previously reported. A Python package allowing researchers to calculate iCUB scores for input sequences is available at https://github.com/amarallab/iCUB.

4. Conclusion

Rigorous evaluation and standardization of metrics is essential to limit the occurrence of false-negative and false-positive results in the literature. Measuring CUB is far from trivial. Short coding sequences, GC contents far from 50% and variable amino acid compositions all conspire to make estimation of the level of CUB from a gene sequence quite a challenging task. We have demonstrated here that there is still ample room for improvement based on the performance of existing metrics.

Our pipeline provides a tool to aid in the continued development of CUB metrics. Researchers can quickly assess the performance of their new metric by running it through the computational pipeline which we have packaged and released at https://github.com/amarallab/cub_benchmarking.

It is possible that different CUB metrics, such as those we evaluated here, are measuring different aspects of coding sequence biases. This is problematic only if conclusions based on a single metric with a particular set of assumptions are generalized to apply to CUB as a whole. Our goal is to highlight that the choice of an individual metric for a particular application should be a rational and explicit choice determined by the hypothesis one is choosing to investigate. Our study can serve as an important guide for how to make this choice.

Data accessibility. Code to run the benchmarking framework is available at https://github.com/amarallab/cub_benchmarking. Code to run iCUB is available at https://github.com/amarallab/iCUB.

Authors' contributions. S.S.L., A.J.H., M.C.J. and L.A.N.A. conceived the study. S.S.L. and A.J.H. developed and analysed the model. M.C.J. and L.A.N.A. provided guidance on study design and analysis. S.S.L. and A.J.H. drafted the manuscript. S.S.L., A.J.H., M.C.J. and L.A.N.A. reviewed and edited the manuscript. All authors gave their final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. This work was funded by the National Institute of General Medical Science (T32 GM008449) to S.S.L., the Northwestern University Presidential Fellowship to A.J.H., and the David and Lucile Packard Foundation (2011-37152); and the Camille Dreyfus Teacher Scholar Award to M.C.J. The John and Leslie McQuown Gift to L.A.N.A. and M.C.J.

Acknowledgements. The authors thank Andrea Lancichinetti for the initial discussions regarding the development of iCUB and critical reading of early drafts of the manuscript.

References

1. Dong H, Nilsson L, Kurland CG. 1996 Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.* **260**, 649–663. (doi:10.1006/jmbi.1996.0428)
2. Rocha EPC. 2004 Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* **14**, 2279–2286. (doi:10.1101/gr.2896904)
3. Dana A, Tuller T. 2014 The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res.* **42**, 9171–9181. (doi:10.1093/nar/gku646)
4. Shah P, Gilchrist M. a. 2010 Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. *PLoS Genet.* **6**, e1001128. (doi:10.1371/journal.pgen.1001128)
5. Higgs PG, Ran W. 2008 Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol. Biol. Evol.* **25**, 2279–2291. (doi:10.1093/molbev/msn173)
6. dos Reis M, Savva R, Wernisch L. 2004 Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids. Res.* **32**, 5036–5044. (doi:10.1093/nar/gkh834)
7. 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. (doi:10.1016/0022-2836(81)90003-6)
8. Zouridis H, Hatzimanikatis V. 2008 Effects of codon distributions and tRNA competition on protein translation. *Biophys. J.* **95**, 1018–1033. (doi:10.1529/biophysj.107.126128)
9. Bahir I, Fromer M, Prat Y, Linial M. 2009 Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol.* **5**, 311. (doi:10.1038/msb.2009.71)
10. Tuller T, Girshovich Y, Sella Y, Kreimer A, Freilich S, Kupiec M, Gophna U, Ruppin E. 2011 Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids. Res.* **39**, 4743–4755. (doi:10.1093/nar/gkr054)
11. Xu Y, Ma P, Shah P, Rokas A, Liu Y, Johnson CH. 2013 Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. *Nature* **495**, 116–120. (doi:10.1038/nature11942)
12. Zhou M, Guo J, Cha J, Chae M, Chen S, Barral JM, Sachs MS, Liu Y. 2013 Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **495**, 111–115. (doi:10.1038/nature11833)
13. Frenkel-Morgenstern M, Danon T, Christian T, Igarashi T, Cohen L, Hou Y-M, Jensen LJ. 2012 Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. *Mol. Syst. Biol.* **8**, 1–10. (doi:10.1038/msb.2012.3)
14. Pechmann S, Frydman J. 2012 Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* **20**, 237–243. (doi:10.1038/nsmb.2466)
15. Saunders R, Deane CM. 2010 Synonymous codon usage influences the local protein structure

observed. *Nucleic Acids. Res.* **38**, 6719–6728. (doi:10.1093/nar/gkq495)

16. Stoletzki N, Eyre-Walker A. 2007 Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol. Biol. Evol.* **24**, 374–381. (doi:10.1093/molbev/msl166)
17. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. 2008 Virus attenuation by genome-scale changes in codon pair bias. *Science* **320**, 1784–1787. (doi:10.1126/science.1155761)
18. Welch M, Govindarajan S, Ness JE, Villalobos A, Gurney A, Minshull J, Gustafsson C. 2009 Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS ONE* **4**, e7002. (doi:10.1371/journal.pone.0007002)
19. Terai G, Kamegai S, Taneda A, Asai K. 2017 Evolutionary design of multiple genes encoding the same protein. *Bioinformatics* **33**, 1613–1620. (doi:10.1093/bioinformatics/btx030)
20. Guimaraes JC, Rocha M, Arkin AP, Cambray G. 2014 D-Tailor: automated analysis and design of DNA sequences. *Bioinformatics* **30**, 1087–1094. (doi:10.1093/bioinformatics/btt742)
21. Sharp PM, Li W-h. 1987 The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids. Res.* **15**, 1281–1295. (doi:10.1093/nar/15.3.1281)
22. Supek F, Vlahovicek K. 2005 Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics*. **6**, 182. (doi:10.1186/1471-2105-6-182)
23. Wright F. 1990 The 'effective' number of codons' used in a gene. *Gene* **87**, 23–29. (doi:10.1016/0378-1119(90)90491-9)
24. Novembre JA. 2002 Accounting for background nucleotide composition when measuring codon usage bias. *Mol. Biol. Evol.* **19**, 1390–1394. (doi:10.1093/oxfordjournals.molbev.a004201)
25. Sun X, Yang Q, Xia X. 2013 An improved implementation of effective number of codons (Nc). *Mol. Biol. Evol.* **30**, 191–196. (doi:10.1093/molbev/mss201)
26. Fuglsang A. 2008 Impact of bias discrepancy and amino acid usage on estimates of the effective number of codons used in a gene, and a test for selection on codon usage. *Gene* **410**, 82–88. (doi:10.1016/j.gene.2007.12.001)
27. Liu X. 2013 A more accurate relationship between 'effective number of codons' and GC3s under assumptions of no selection. *Comput. Biol. Chem.* **42**, 35–39. (doi:10.1016/j.compbiolchem.2012.11)
28. Banerjee T, Gupta SK, Ghosh TC. 2005 Towards a resolution on the inherent methodological weakness of the 'effective number of codons used by a gene'. *Biochem. Biophys. Res. Commun.* **330**, 1015–1018. (doi:10.1016/j.bbrc.2005.02.150)
29. Fuglsang A. 2004 The 'effective number of codons' revisited. *Biochem. Biophys. Res. Commun.* **317**, 957–964. (doi:10.1016/j.bbrc.2004.03.138)
30. Fuglsang A. 2006 Estimating the 'effective number of codons': the Wright way of determining codon homozygosity leads to superior estimates. *Genetics* **172**, 1301–1307. (doi:10.1534/genetics.105.049643)
31. Zhang Z, Li J, Cui P, Ding F, Li A, Townsend JP, Yu J. 2012 Codon deviation coefficient: a novel measure for estimating codon usage bias and its statistical significance. *BMC. Bioinformatics.* **13**, 43. (doi:10.1186/1471-2105-13-43)
32. Wan X-F, Xu D, Kleinhofs A, Zhou J. 2004 Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC. Evol. Biol.* **4**, 19. (doi:10.1186/1471-2148-4-19)
33. Roymondal U, Das S, Sahoo S. 2009 Predicting gene expression level from relative codon usage bias: an application to *Escherichia coli* genome. *DNA Res.* **16**, 13–30. (doi:10.1093/dnares/dsn029)
34. Liu SS, Hockenberry AJ, Lancichinetti A, Jewett MC, Amaral LAN 2016 NullSeq: a tool for generating random coding sequences with desired amino acid and GC contents. *PLoS. Comput. Biol.* **12**, e1005184. (doi:10.1371/journal.pcbi.1005184)
35. Vieira-Silva S, Rocha EPC. 2010 The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* **6**, e1000808. (doi:10.1371/journal.pgen.1000808)
36. Edgar RC. 2004 MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids. Res.* **32**, 1792–1797. (doi:10.1093/nar/gkh340)
37. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)
38. Hug LA *et al.* 2016 A new view of the tree of life. *Nat. Microbiol.* **1**, 16048. (doi:10.1038/nmicrobiol.2016.48)
39. Lowe TM, Chan PP. 2016 tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids. Res.* **44**, W54–W57. (doi:10.1093/nar/gkw413)
40. Chen S, Li K, Cao W, Wang J, Zhao T, Huan Q, Yang Y-F, Wu S, Qian W. 2017 Codon-resolution analysis reveals a direct and context-dependent impact of individual synonymous mutations on mRNA level. *Mol. Biol. Evol.* **34**, 2944–2958. (doi:10.1093/molbev/msx229)
41. Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009 Coding-sequence determinants of gene expression in *Escherichia coli*. *Sci.* (*New York, N.Y.*) **324**, 255–258. (doi:10.1126/science.1170160)
42. Powell JR, Dion K. 2015 Effects of codon usage on gene expression: empirical studies on *Drosophila*. *J. Mol. Evol.* **80**, 219–226. (doi:10.1007/s00239-015-9675-y)
43. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. 2015 Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**, 3163–3168. (doi:10.1002/pmic.201400441)
44. Hildebrand F, Meyer A, Eyre-Walker A. 2010 Evidence of selection upon genomic GC-content in bacteria. *PLoS. Genet.* **6**, e1001107. (doi:10.1371/journal.pgen.1001107)
45. Hershberg R, Petrov DA. 2010 Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* **6**, e1001115. (doi:10.1371/journal.pgen.1001115)
46. Kunec D, Osterrieder N. 2016 Codon pair bias is a direct consequence of dinucleotide bias. *Cell Rep.* **14**, 55–67. (doi:10.1016/j.celrep.2015.12.011)
47. Shen SH, Stauft CB, Gorbatsevych O, Song Y, Ward CB, Yurovsky A, Mueller S, Futcher B, Wimmer E. 2015 Large-scale recoding of an arbovirus genome to rebalance its insect versus mammalian preference. *Proc. Natl Acad. Sci. USA* **112**, 4749–4754. (doi:10.1073/pnas.1502864112)
48. Simmonds P, Tulloch F, Evans DJ, Ryan MD. 2015 Attenuation of dengue (and other RNA viruses) with codon pair recoding can be explained by increased CpG/UpA dinucleotide frequencies: table 1. *Proc. Natl Acad. Sci. USA* **112**, E3633–E3634. (doi:10.1073/pnas.1507339112)
49. Hockenberry AJ, Sirer MI, Amaral LAN, Jewett MC. 2014 Quantifying position-dependent codon usage bias. *Mol. Biol. Evol.* **31**, 1880–1893. (doi:10.1093/molbev/msu126)
50. Diwan GD, Agashe D. 2016 The frequency of internal Shine–Dalgarno-like motifs in prokaryotes. *Genome. Biol. Evol.* **8**, 1722–1733. (doi:10.1093/gbe/evw107)
51. Itzkovitz S, Hodis E, Segal E. 2010 Overlapping codes within protein-coding sequences. *Genome. Res.* **20**, 1582–1589. (doi:10.1101/gr.105072.110)
52. Yang C, Hockenberry AJ, Jewett MC, Amaral LAN. 2016 Depletion of Shine–Dalgarno sequences within bacterial coding regions is expression dependent. *Adv. Genet.* **6**, 3467–3474. (doi:10.1534/g3.116.032227)
53. Fawcett T. 2006 An introduction to ROC analysis. *Pattern. Recognit. Lett.* **27**, 861–874. (doi:10.1016/j.patrec.2005.10.010)