Evolution of tRNA pool shapes variation in selection on codon usage across the Saccharomycotina subphylum

Alexander L. Cope[1,2,3,*] and Premal Shah[1,3,*]

[1]Department of Genetics, Rutgers University, Piscataway, NJ, United States

[2]Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ, United States

[3]Human Genetics Institute of New Jersey, Rutgers University, Piscataway, New Jersey, United States

[4]Current: Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee, United States

[*]Corresponding Authors:

alexander.cope@vanderbilt.edu;premal.shah@rutgers.edu

Friday 27th September, 2024

Keywords: Codon usage bias; population genetics; budding yeasts; translational selection

Abstract

Across the major taxonomical domains, synonymous codons of an amino acid are found to be used in unequal frequencies. This codon usage bias – both in terms of the degree of bias and the identity of codons used – is highly variable, even among closely related species. Within a species, genome-wide codon usage bias reflects a balance between adaptive and non-adaptive microevolutionary processes. Variation in these microevolutionary processes results in across-species variation in codon usage bias. As codon usage bias is tightly linked to important molecular and biophysical processes, it is critical to understand how changes to these processes drive changes to the microevolutionary processes. Here we employ a population genetics model of coding sequence evolution to quantify natural selection and mutation biases on a per-codon basis and estimate gene expression levels across the budding yeasts Saccharomycotina subphylum. We interrogate the impact of variation in molecular mechanisms hypothesized to be driving the microevolution of codon usage. We find that natural selection and mutation biases evolved rapidly over macroevolutionary time, with high variability between closely related species. The majority (324/327) of yeasts exhibited clear signals of translational selection, with selection coefficients being correlated with codon-specific estimates of ribosome waiting times within species. Across species, natural selection on codon usage correlated with changes to ribosome waiting times, indicating that tRNA pool evolution is a major factor driving changes to natural selection on codon usage. We find evidence that changes to tRNA modification expression can contribute to changes in natural selection across species independent of changes to tRNA gene copy number, suggesting tRNA modifications also play a role in shaping natural selection on codon usage. Our work firmly establishes how changes to microevolutionary processes can be driven by changes to molecular mechanisms, ultimately shaping the macroevolutionary variation of a trait.

Introduction

The genetic code is "degenerate" – the 61 amino acid-encoding codons are translated into 20 amino acids, meaning multiple codons must be ascribed to the same amino acid. Across all domains of life, synonymous codons are used at unequal frequencies, a phenomenon known as codon usage bias (CUB) [1–6]. The CUB of a genome reflects a balance between the microevolutionary processes of natural selection, mutation, and genetic drift shaping the synonymous codon usage frequencies for a particular amino acid [7]. Natural selection for efficient or accurate translation – often termed "translational selection" – is hypothesized to be the primary driver of adaptive CUB due to the correlations between codon usage and the tRNA pool and the bias towards codons corresponding to more abundant tRNA in highly expressed genes [5, 8–11]. Under translational selection, selection on synonymous mutations is strongest in highly expressed genes due to their effects on potential energetic burden of ribosome pausing or protein misfolding [10, 12–14]. However, beause highly expressed genes constitute only a small portion of protein-coding sequences in a genome, genome-wide CUB (i.e., the most frequently used synonyms genome-wide) is determined by mutation bias and drift. Other microevolutionary processes, such as GC-biased gene conversion [15–17], can also shape patterns of codon usage within a genome and can sometimes obscure signature of selection.

CUB varies across species, both in terms of the degree of bias and the identity of synonymous codons used most frequently [1, 5, 6, 11, 18–22]. The fact that CUB varies across species is clear – the causes are not. Variation in CUB on macroevolutionary timescales ultimately reflects changes to the underlying microevolutionary processes that drive CUB within a genome. Mechanistic evolutionary models providing theoretically justified (i.e., rooted in population genetics theory) estimates of evolutionary parameters (e.g., the strength of natural selection) are necessary to determine how variation in these processes leads to the observed macroevolutionary trends in CUB. Moreover, CUB is tightly linked to the molecular processes of DNA replication (mutation bias) and protein synthesis (translational selection), such that macroevolutionary variations in the microevolutionary processes shaping CUB are hypothesized to reflect changes to the underlying molecular processes. Some studies observed correlations between CUB and the relevant molecular processes [3, 5, 11, 20], but a lack of formal estimates of population genetics parameters at the level of individual codons limits our ability to link changes in microevolution to these processes.

Here, we quantify variation in natural selection and mutation biases that drive CUB across 327 Saccharomycotina budding yeasts [11, 23]. To do so, we employ the Ribosomal Overhead Cost version of the Stochastic Evolutionary Model of Protein Production Rates (ROC-SEMPPR) – a powerful population genetics model for quantifying CUB [24–28]. Unlike many popular heuristic approaches for quantifying CUB, ROC-SEMPPR disentangles the effects of natural selection from mutation biases by quantifying changes in codon usage as a function of gene expression. Specifically, for a given gene g with average gene expression (technically, protein production rate) ϕ_g, the probability $p_{g,i}$ of seeing codon i is

$$p_{g,i} \propto e^{-\Delta M_i - \Delta\eta_i \phi_g}$$

ROC-SEMPPR estimates natural selection $\Delta\eta$ and mutation bias ΔM per codon, allowing us to systematically investigate how and why these microevolutionary processes vary on macroevolutionary timescales. Building from a previous examination of 49 budding yeasts, we find a substantial percentage of the Saccharomycotina subphylum ($\approx 20\%$) exhibit significant across-gene variation in codon usage not attributable to translational selection, suggestive of other non-adaptive evolutionary processes that can shape codon usage bias. We find multiple lines of evidence for translational selection on CUB across most species. Particularly noteworthy, we find codon-specific shifts in natural selection are correlated with changes to the tRNA pool across species. Such a correlation is often presumed, but our work directly shows how a key feature of protein synthesis (the tRNA pool) shapes natural selection on codon usage. Additionally, we explore how mutation biases change across species, finding them to be strongly correlated with changes to GC%, but find inconclusive support for a general role of the evolution of mismatch-repair (MMR) genes in driving changes to mutation biases (see Supplemental Text).

Results

We applied ROC-SEMPPR to quantify the strength and direction of natural selection and mutation bias shaping CUB across the 327 Saccharomycotina budding yeasts. Previous work by us and others indicate that within-genome variation in nucleotide usage can arise due to non-adaptive processes (e.g., GC-biased gene conversion) that can obscure signatures of translational selection [26, 28, 29]. To account for this variation, for each species, we performed correspondence analysis on the absolute codon frequencies of each gene followed by CLARA clustering (a k-medoids clustering) of the first 4 principal axes to separate genes into 2 sets potentially subject to different non-adaptive nucleotide biases. Following [28], we compared two model fits to assess the potential impact of within-genome variation in non-adaptive nucleotide biases – any processes biasing codon usage that does not scale with gene expression ϕ. We refer to these models as “ConstMut” and “VarMut”. The ConstMut model assumes mutation bias is constant across all genes; in contrast, the VarMut model allows mutation bias ΔM to vary between the 2 gene sets determined by the clustering of the correspondence analyses axes (termed the “Lower GC3 Set” and “Higher GC3 Set”, see “Material and Methods: Identifying within-genome variation in codon usage bias”). Using the correlations of ROC-SEMPPR predicted gene expression ϕ with empirical gene expression as our comparisons of the ConstMut and VarMut models, the VarMut model better fit 58 of the 327 species (approximately 18%), consistent with intragenomic variation in non-adaptive nucleotide biases within these species. Across the 58 species better fit by the VarMut model, the predicted-empirical correlations between the VarMut and ConstMut models differ by a median value of 0.31, indicating better predictions of gene expression based on codon usage when using the VarMut model. Across the other 269 species, the predicted-empirical correlations between the VarMut and ConstMut models differ by a median value of -0.05, indicating generally worse or negligible differences in predictions of gene expression based on codon usage when using the VarMut model (Figure S1) – this is consistent with unsubstantial intragenomic variation in non-adaptive nucleotide biases. For subsequent analyses, we use the selection and mutation bias estimates from the best model fit for each species.

Natural selection and mutation biases shaping codon usage are highly-variable among close relatives

To understand the evolution of codon usage on macroevolutionary timescales, it is critical to quantify the variability in the evolutionary forces shaping species-specific CUB. We performed hierarchical clustering of species based on the estimates of (1) (scaled) selection coefficients $\Delta\eta$ and (2) mutation biases to both understand how these parameters varied across species and to what extent these parameters reflect the shared ancestry of the budding yeasts. To further clarify the meanings of these parameters, the selection coefficients $\Delta\eta = s_{i,j}N_e$ in a gene of average expression, where $s_{i,j}$ is the unscaled selection coefficient between synonymous codons i and j and N_e is the effective population size. Mutation bias $\Delta M = \log(\frac{\mu_{i,j}}{\mu_{j,i}})$ where $\mu_{i,j}$ is the mutation rate between synonymous codons i and j. All parameter estimates from our model fit are relative to a reference codon, specifically the alphabetically last codon, with a negative value indicating a codon is "favored" relative to the reference codon.

Based on the clustering of selection coefficients, the direction of natural selection on codon usage is largely consistent across the tree (Figure 1A). For example, NNC codons for most 2/3-codon (Asn, Tyr, His, Phe, Asp, Ser$_2$, Ile) and 4-codon (Val, Thr, Ser$_4$, Ala) amino acids are selectively-favored across the majority of species; this was not the case for either 6-codon amino acid (Arg, Leu). For the 2-codon amino acids Lys, Glu, and Gln (all NNA/NNG), the Lys codon AAG was generally selectively-favored relative to AAA, but the NNA codons for Glu (GAA) and Gln (CAA) are favored across most species (although many species did favor NNG). 34 species exhibit dramatic shifts in natural selection on codon usage relative to the remaining 293 species (Figure 1A, row dendrogram labels 1 and 2). These species have significantly weaker correlations between ROC-SEMPPR predicted gene expression and empirical gene expression (Welch two-sample t-test $p = 3.499E-08$) compared to the other 293 species, suggesting poor model fits (Figure 1A, Obs. vs. Pred. Gene Expr.). Surprisingly, the clustering of selection coefficients only weakly reflects the phylogeny of the Saccharomctoina yeasts, consistent with little similarity between closely-related species. Many species from the same clade fall into the same cluster, but most clades are divided into separate groups (Figure 1A, Clade color bar). To ensure that the poorly fit species did not obscure the similarity of closely related species in our clustering, we removed the 34 species with suspected poor model fits and re-performed the clustering of selection coefficients. We find overall little agreement between the clustering of selection coefficients and phylogeny as measured by the cophenetic correlation to quantify the overall pairwise similarity between species within the dendrograms. A high cophenetic correlation implies variation in selection coefficients recapitulates the evolutionary relationships between the species. We find the cophenetic correlation between the clustering of selection coefficients and the phylogeny was 0.35, consistent with selection coefficients only weakly reflecting the phylogeny of the Saccharomycotina yeasts.

As with the selection coefficients, we examined how mutation biases ΔM varied across the 327 yeasts. Generally, mutation biases favor AT-ending codons over GC-ending codons across the majority of the subphylum (Figure 2 and Figure S2), consistent with the overall AT-bias of the Saccharomycotina subphylum [11]. Similar to the selection coefficients, we performed hierarchical clustering of mutation biases to assess the similarity of mutation biases between closely related species (Figure 2 and Figure S2, Clade color bar). Mutation biases are more variable between closely related species compared to selection coefficients as indicated by the smaller groupings of species by clade (Figure 2 and Figure S2, Clade color bar). Consistent with this qualitative observation, the cophenetic correlations between the hierarchical clustering of mutation biases and the phylogenetic tree after removing the 34 poorly fit species is 0.14 or 0.13 depending on the use of Lower GC3% and Higher GC3% sets for species better fit by the VarMut model. Even more so than selection coefficients, across-species variation in mutation biases poorly reflects the Saccharomycotine phylogeny.

As an orthogonal analysis to quantify the overall similarity of natural selection and mutation biases between closely related yeasts (i.e., phylogenetic signal), we estimated a multivariate version of Blomberg's K (K_{multi}) using the R package **geomorph** [30]. Selection coefficients exhibit a greater phylogenetic signal ($K_{multi} = 0.437$) than mutation biases ($K_{multi} = 0.224$ if using Lower GC3% set, $K_{multi} = 0.156$ if using Higher GC3% set). Taken together, both analyses support greater variation in mutation biases between closely related species, but this should not distract from the fact that both natural selection and mutation biases are highly variable. It is often assumed that a weaker phylogenetic signal is due to higher evolutionary rates. However, this is

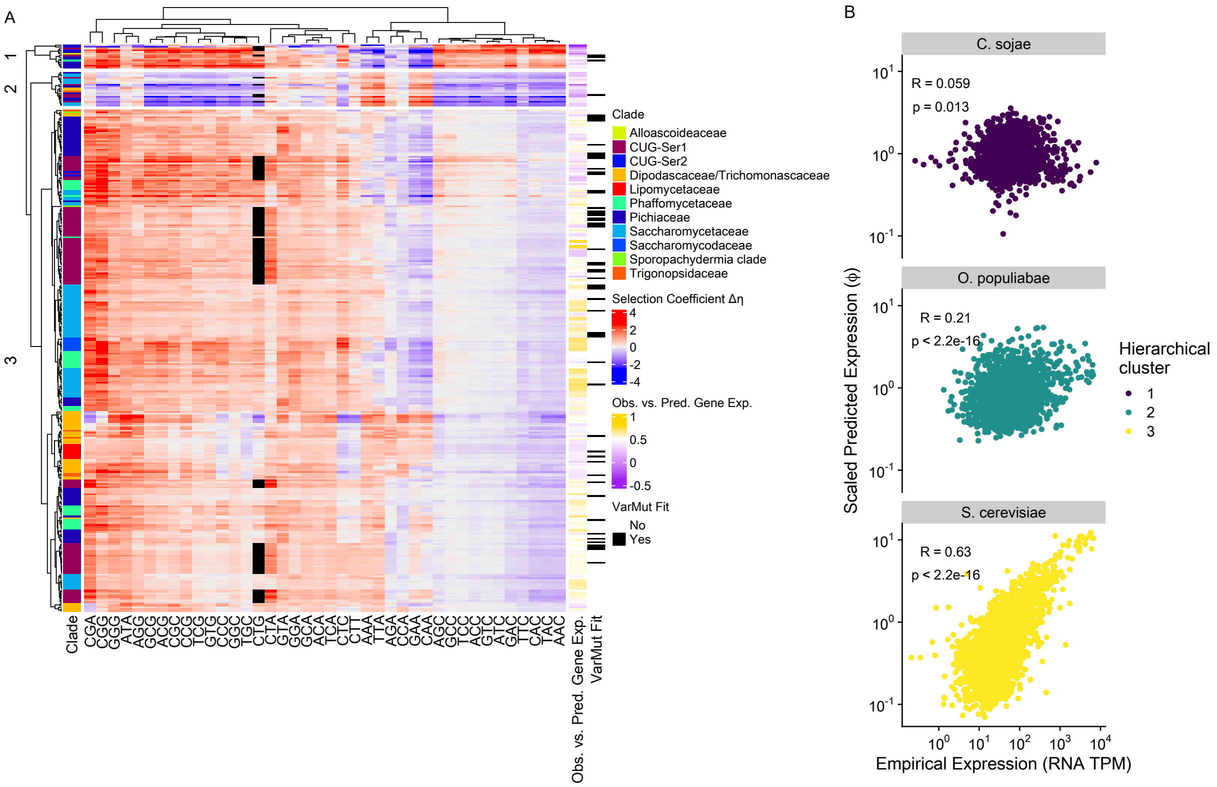

Figure 1: (A) Heatmap representing the selection coefficients $\Delta\eta$ across the 327 Saccharomycotina subphylum. Red indicates a codon is disfavored by selection relative to the reference synonymous codon (the alphabetically last codon for each amino acid), while blue indicates the codon is favored relative to the reference codon. Black indicates cases where the codon CTG codes for serine – in these species, CTG is treated separately from the serine codons (see Materials and Methods). Row-wise dendrogram represents the hierarchical clustering of selection coefficients across species based on dissimilarity as estimated by 1 - R, where R is the Spearman correlation between the selection coefficients of two species. Numbers represent the result of splitting the clustering into 3 groups of species. Column-wise dendrogram represents the hierarchical clustering of selection coefficients across codons based on Euclidean distance. The Clade color bar indicates the major clade of the species, as defined previously[11]. The Obs. vs. Pred. Gene Expr. color bar represents the Spearman correlation between empirical gene expression estimates and ROC-SEMPPR predicted gene expression estimates ϕ per species. The VarMut fit color bar indicates if the model fit used for a given species was VarMut (black) or ConstMut fit (white). (B) Example of the correlation between empirical estimates of gene expression and ROC-SEMPPR predicted estimates of gene expression for species in the three clusters as labeled on the species-wise dendrogram in (A).

not always the case. Phylogenetic signal can degrade due to strong stabilizing selection towards a single optimum [31]. We fit a univariate Ornstein-Uhlenbeck model of trait evolution (via the R package **geiger** [32]) to our estimates of natural selection and mutation biases for each codon to quantify the strength of stabilizing selection for each trait. We find that the strength of stabilizing selection α is generally greater for estimates of mutation biases than estimates of selection (median $\alpha_{\Delta M_{\text{Lower GC3\%}}} = 0.011$, $\alpha_{\Delta\eta} = 0.005$, Wilcox rank sum test $p = 3.051E-9$). We note the highest α values are for estimates of natural selection (Figure S3A). We obtain a similar result if using mutation bias estimates from the Higher GC3% set (Figure S3B). This is consistent with stronger stabilizing selection acting on the factors shaping mutation biases (or other non-adaptive nucleotide biases) relative to natural selection, resulting in a stronger degradation of phylogenetic signal.

Below, we focus on the molecular mechanisms responsible for the observed across-species changes in natural selection on codon usage. We perform similar analyses to examine changes to estimates of mutation biases and find clear correlations between our mutation bias estimates and genome-wide GC%. However, our results relating these changes to a specific molecular mechanism are rather inconclusive (see Supplemental Text).

Selection for translation efficiency is a prevalent force shaping within-genome variation in codon usage

The strength and direction of natural selection on codon usage varies across species, suggesting underlying changes to the molecular processes that shape natural selection. Translational selection (e.g., selection against translation inefficiency) remains the predominant hypothesis regarding genome-wide adaptive CUB, particularly in microbes. The translational selection hypothesis leads to two testable predictions: (1) codon usage will covary with gene expression and (2) highly expressed genes are biased toward codons with faster elongation rates. Such predictions are testable with ROC-SEMPPR's parameters: (1) estimates of evolutionary-average gene expression ϕ are expected to be positively correlated with empirical estimates of gene expression, and (2) selection coefficients $\Delta\eta$ are expected to be positively correlated with ribosome waiting times, the inverse of elongation rates (i.e., slower codons are disfavored by natural selection).

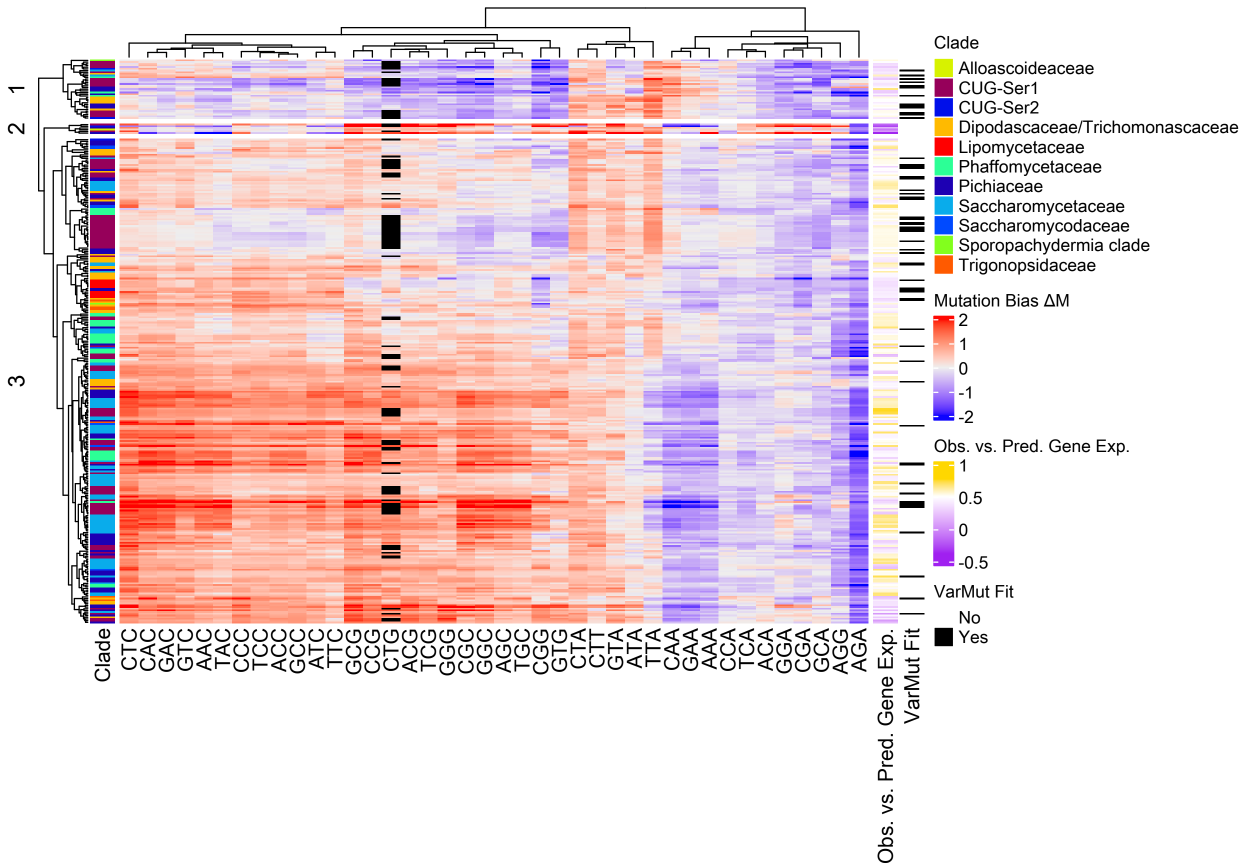

Figure 2: (A) Heatmap representing the mutation biases ΔM across the 327 Saccharomycotina subphylum. Red indicates mutation is biased against a codon relative to the reference synonymous codon (the alphabetically last codon for each amino acid), while blue indicates mutation is biased towards a codon relative to the reference codon. Black indicates cases where the codon CTG codes for serine – in these species, CTG is treated separately from the serine codons (see Materials and Methods). The row-wise dendrogram represents the hierarchical clustering of mutation biases across species based on dissimilarity as estimated by 1 - R, where R is the Spearman correlation between the selection coefficients of two species. Numbers represent the result of splitting the clustering into 3 groups of species. Column-wise dendrogram represents the hierarchical clustering of mutation biases across codons based on Euclidean distance. The Clade color bar indicates the major clade of the species, as defined previously [11]. The Obs. vs. Pred. Gene Expr. color bar represents the Spearman correlation between empirical gene expression estimates and ROC-SEMPPR predicted gene expression estimates ϕ per species. The VarMut fit color bar indicates if the model fit used for a given species was VarMut (black) or ConstMut fit (white).

ROC-SEMPPR predictions of gene expression are well-correlated with empirically measured gene expression

After determining the best model fit between the ConstMut and VarMut models for each species, we find the median Spearman rank correlation between predicted and empirical estimates of gene expression (i.e., ϕ vs. RNA-seq) across all species is 0.51 with 95% of species having a correlation > 0.26 (Figure 3). We emphasize the empirical gene expression data are based on RNA-seq data from a subset of 49 yeasts [28] mapped across species based on a list of one-to-one orthologs [23]. We find the correlation between predicted and empirical estimates decreased relative to the RNA-seq reference species as the divergence time between the two species increased (Figure S4, Spearman rank correlation $-0.14, p = 0.016$). Regardless, the positive correlation between predicted (i.e., based solely on codon usage) and empirical gene expression indicates prevalent translational selection on codon usage across the Saccharomycotina subphylum.

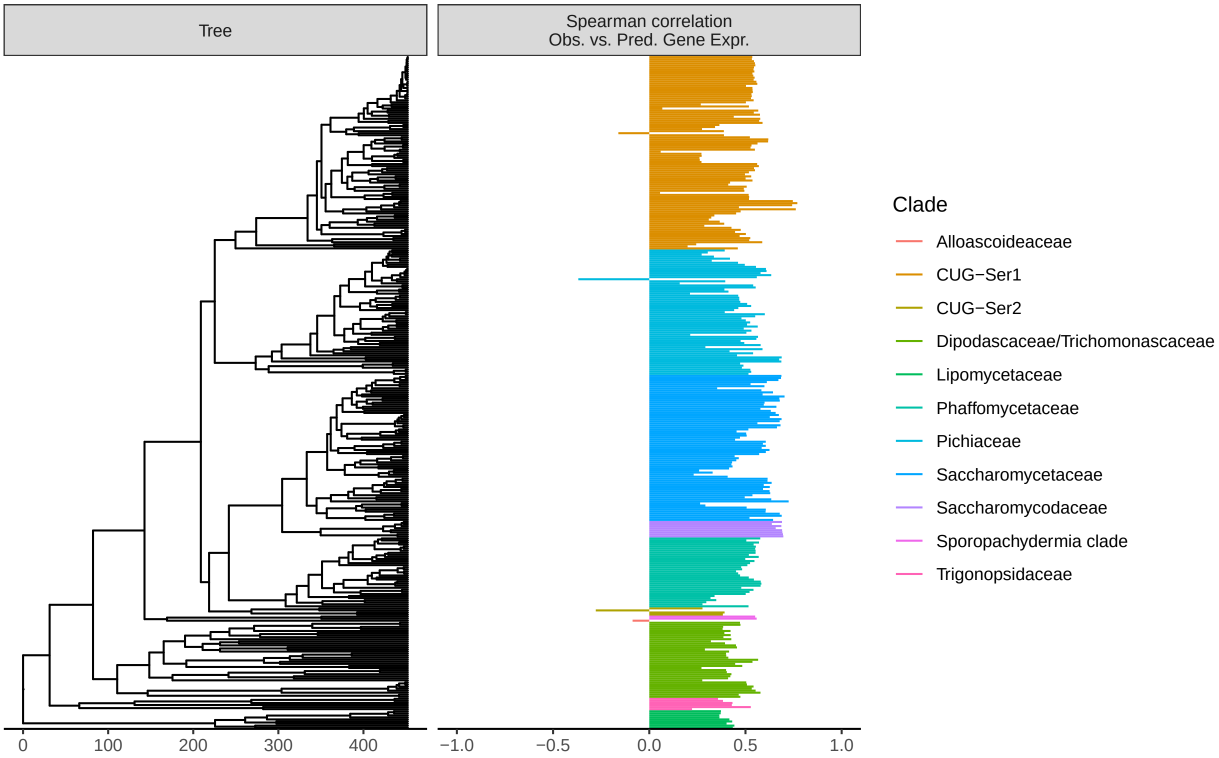

Figure 3: Within-species correlation between observed empirical gene expression measured via RNA-seq and ROC-SEMPPR predicted gene expression ϕ based on codon usage bias. For most species, empirical gene expression were taken from the closest relative for which such estimates were available. Solid lines indicate cases where the Spearman rank correlation was statistically significant; dashed lines indicate non-significance.

Selection coefficients are well-correlated with the predicted speed of elongation within species

As selection coefficients $\Delta\eta$ reflect selection against a codon relative to its synonyms, translational selection will lead to a positive correlation between selection coefficients and ribosome waiting times (from here on out referred to as ”waiting times”). We used the inverse of the relative weights for the tRNA adaptation index (tAI) as a proxy for waiting times [33]. Surprisingly, we find numerous codons do not have a corresponding tRNA based on the identified tRNA genes and standard wobble rules, implying inefficient elongation of these codons. Although many of these are likely spuriously missed tRNA genes by tRNAScan-SE, we note a lack of tRNA genes recognizing proline codons CCC/CCT in the clades CUG-Ser2 (2/4 species), Phaffomycetaceae (34/34 species), Pichiaceae (46/61 species), and Saccharomycodaceae (8/8) species (Figure S5). A gene encoding Pro-tRNA$^{\mathrm{UGG}}$codon is present in each of these species, but not at appreciably different amounts than observed in the other species (Welch two-sample t-test, $p = 0.7742$). Assuming the corresponding tRNA genes are truly missing in these clades, this suggests the occurrence of super-wobbling for proline, whereby an unmodified U_{34} allows a tRNA to bind all 4 codons [34].

We highlight two species: *Candida albicans* and *Starmera amethionina*, both of which were previously determined to have little adaptive codon usage related to translational selection (Figure 4A,B) [11]. We find a strong positive correlation between waiting times and selection coefficients in both species, indicating translational selection is a prevalent force shaping CUB within these yeasts. In the case of *C. albicans*, previous difficulty in detecting translational selection was likely due to the failure to account for within-genome variation in non-adaptive nucleotide biases [28]. However, the VarMut model performed significantly worse for *S. amethionina* (Spearman rank correlation between ROC-SEMPPR predicted and empirical gene expression of 0.348 vs. -0.0819 for ConstMut and VarMut models, respectively). The median Spearman rank correlation between selection coefficients and waiting times across all 327 budding yeasts was 0.78, with a range of 0.17 to 0.93 (324/327 species $p < 0.05$, Figure 4C). The positive correlation between selection coefficients and waiting times indicates translational selection is a prevalent force shaping CUB across the Saccharomycotina subphylum.

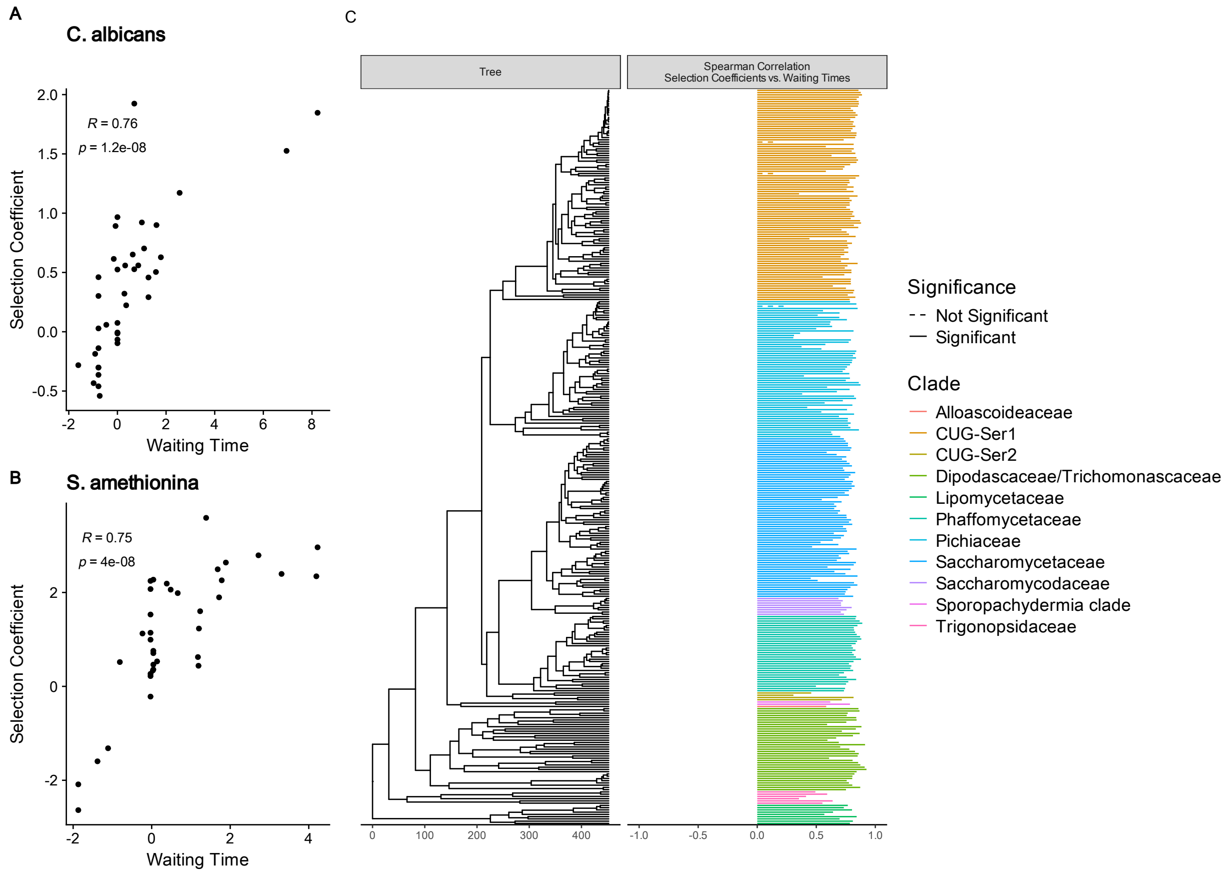

Figure 4: Comparisons of relative elongation waiting times (based on tRNA gene copy numbers) and selection coefficients $\Delta\eta$ for (A) *Candida albicans* and (B) *Starmera amethionina*. R indicates the Spearman rank correlation. (C) Correlation as in (A) and (B) across all species (right panel) ordered by the phylogeny of the budding yeasts (left panel). Species are colored by major clade as defined in [11]. Solid lines indicate cases where the Spearman rank correlation was statistically significant; dashed lines indicate non-significance.

Across-species variation in natural selection on codon usage varies with the tRNA pool

Given the positive correlation between natural selection and waiting times within species, the evolution of the tRNA pool is hypothesized to be a driver of differences in natural selection on codon usage. We exclude the 34 species outlier species indicated by the hierarchical clustering of selection coefficients because these species could obfuscate the general relationship between selection coefficients and the tRNA pool. Previous work found the overall strength of selection on codon usage increased with the number of translation resources, supposedly reflecting increased selection on growth rate [3, 35]. Along these lines, we find the mean absolute selection coefficients per species – which reflects the average strength of selection across all codons – is positively correlated with the total number of tRNA genes per genome, albeit weakly (Figure 5A, Spearman rank correlation $R = 0.22$, $p = 0.00025$). Our finding conflicts with [11], who found no significant relationship between S [33] (based on tAI and the effective number of codons, not to be confused with the population genetics parameter $S = sN_e$) and the total number of tRNA genes after accounting for shared ancestry. We note S is not a theoretically justified estimate of natural selection on codon usage [3, 33]; perhaps unsurprisingly, mean absolute selection coefficients are only weakly correlated with S (Figure S6). Our results suggest selection on codon usage increases slightly with investment in translational resources, but other factors are likely driving differences in adaptive CUB. For example, translational selection is expected to result from differences in the waiting times of synonymous codons, which in turn could be driven by differences in the tRNA pool. We find the selection coefficients for 39 codons (out of 40, not including the ROC-SEMPPR reference codons for each amino acid) are positively correlated with waiting times across species (Figure 5B and C, Benjamini-Hochberg adjusted $p < 0.05$, see also Figure S7), consistent with across-species changes in the strength/direction of natural selection on codon usage being driven by evolution of the tRNA pool.

Certain tRNA modifications alter the waiting times of a codon [36, 37]. Evolutionary changes to the functionality of tRNA modification enzymes may also signal shifts in the strength or direction of natural selection on codon usage. In *S. cerevisiae*, knockouts of the multimeric Elongator Complex protein (responsible for the U_{34} modification mcm^5s^2U) resulted in increased waiting times at codons AAA, CAA, and GAA [36, 37]. Given the impact of tRNA modifications on waiting times, across-species variation in tRNA modification enzyme activity may contribute to variation in natural selection on codon usage. Here, we determine the impact of differences in gene expression ϕ (as a proxy for overall enzyme activity) of the catalytic activity proteins of the Elongator Complex IKI3, ELP2, and ELP3. We observe correlations between predicted gene expression and selection coefficients for individual codons (e.g., Figure 6A for IKI3 expression vs. AAA selection). However, this does not control for other factors likely related to selection (e.g., tGCN) or expression of Elongator Complex proteins (e.g., genome-wide GC%). We performed phylogenetic generalized

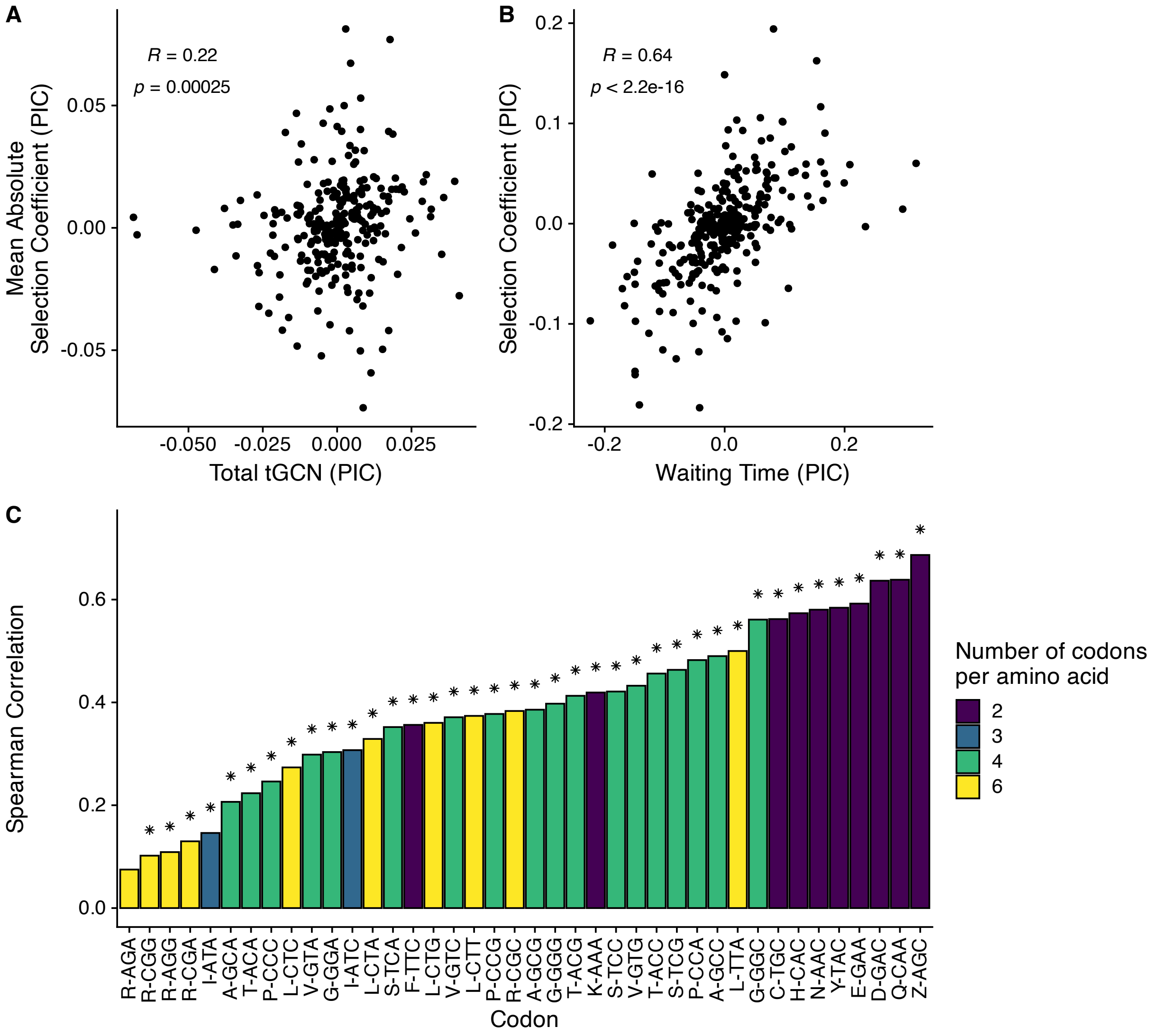

Figure 5: Relationship between codon-specific waiting times (based on tGCN) and selection coefficients $\Delta\eta$. Phylogenetic independent contrasts (PIC) were used in all cases. Spearman rank correlations R and associated p-values are reported. (A) Comparison of mean absolute selection coefficients $|\Delta\eta|$ and the total number of tRNA genes across species. (B) Example scatter-plot showing the relationship between waiting times and selection coefficients $\Delta\eta$ for codon CAA (relative to CAG) across species. (C) Bar plot representing the Spearman rank correlations between waiting times and selection coefficients $\Delta\eta$ across species for all codons. "*" indicate statistical significance $p < 0.05$ after correcting for multiple hypothesis testing via Benjamini-Hochberg.

least squares via the **phylolm** R package to determine the impact of the IKI3, ELP2, and ELP3 on natural selection while controlling for changes to tGCN and genome-wide GC%. Unsurprisingly, tGCN has the overall largest effect on variation in natural selection across species for all 3 codons.

However, we find that shifts in the expression levels of protein IKI3 contribute to variation in natural selection for 2 of the 3 codons. We note there is collinearity between many of our independent variables, which may result in overestimating our standard errors; however, these correlations are weak (for example, Figure S8). Taken together, our results suggest changes to tRNA modification enzyme activity or expression have a modest contribution to changes in selection on codon usage.

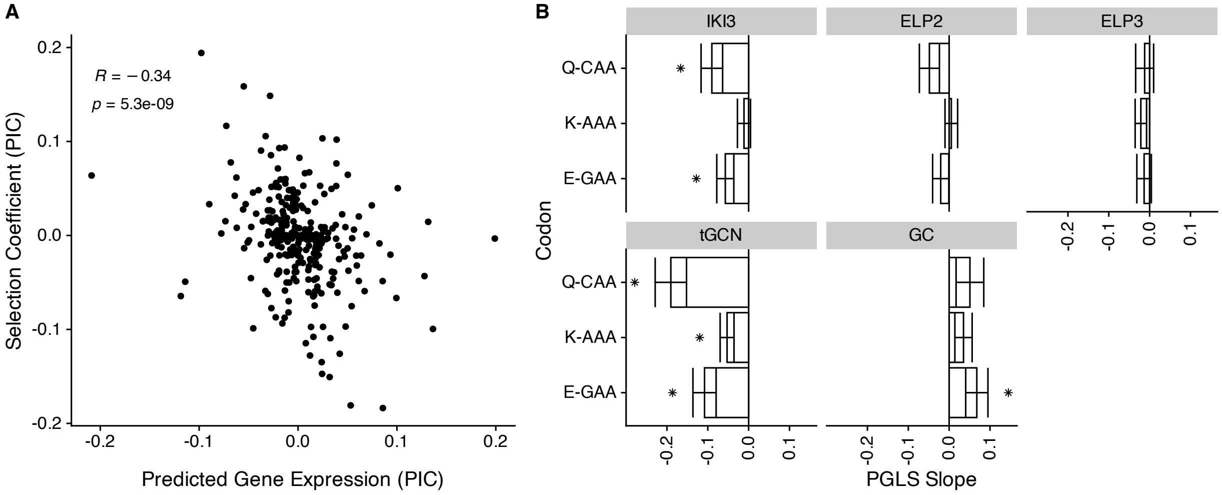

Figure 6: Relationship between across-species variation in natural selection on codon usage and across-species variation in gene expression of proteins forming the tRNA modification enzyme Elongator Complex (IKI3, ELP2, ELP3). (A) Example scatter-plot showing the relationship between the selection coefficients $\Delta\eta$ of codon CAA and gene expression of IKI3. (B) Bar plot representing the effects (i.e., PGLS slopes) of Elongator Complex gene expression, tGCN, and genome-wide GC% on variation in PGLS selection coefficients of codons recognized by tRNA modified by the Elongator Complex.

Discussion

The direction and degree of codon usage bias (CUB) varies across species. Theoretically justified estimates of the underlying microevolutionary processes that shape CUB within a species and how these relate to molecular mechanisms are critical for understanding the causes of the observed macroevolutionary variation in CUB. We applied a population genetics model ROC-SEMPPR to the protein-coding sequences of 327 Saccharomycotina budding yeasts [11, 23] to estimate natural selection and mutation biases at codon-level resolution for all species [10, 24, 38]. The formulation of ROC-SEMPPR and its predecessors assume the only non-adaptive directional force (i.e., favoring one synonym over another) shaping codon usage bias is mutation bias, but many other relevant directional non-adaptive processes exist. This includes but is not limited to GC-biased gene conversion [17] and lateral gene transfer or introgressions [26]. As ROC-SEMPPR quantifies natural selection on codon usage via changes to gene expression, all non-adaptive processes shaping codon usage uncorrelated with gene expression are absorbed into the mutation bias parameter. This is particularly problematic if these processes cause genome-wide variation in the non-adaptive nucleotide background: ROC-SEMPPR is most likely to mistake this variation to be the result of natural selection [28]. We built upon our recent work coupling an unsupervised machine learning approach with ROC-SEMPPR approach to identify protein-coding sequences subject to different non-adaptive nucleotide biases. Across the 327 yeasts, we find 18% of species exhibited significant within-genome variation in non-adaptive nucleotide biases. Previously, we found protein-coding sequences assigned to the different sets were largely differentiated based on GC3% and tended to be colocalized along chromosomes, leading to regions of low and high GC3% content [28]. Within-genome variation in non-adaptive nucleotide biases could be due to several processes; prominent among these is GC-biased gene conversion in which regions of high recombination are expected to have higher GC content [17]. Although we have no direct evidence of GC-biased gene conversion, Saccharomycotina yeasts better fit by the VarMut model often show clear non-adaptive biases towards GC-ending codons. Surprisingly, yeasts exhibiting within-genome variation of non-adaptive biases are often phylogenetically distant. Further work is needed to elucidate the causes of within-genome variation in non-adaptive nucleotide biases.

Translational selection (i.e., natural selection for translation efficiency) is a prominent hypothesis explaining adaptive CUB, particularly in microbes [4]. Most of the Saccharomycotina subphylum exhibits evidence of translational selection, including correlations between predicted and empirical estimates of gene expression, and correlations between selection coefficients with waiting times within species. Under translational selection, we expect across-species changes in selection on codon usage to correlate with changes to the tRNA pool. Across species, natural selection on codon usage is generally correlated with waiting times of codons, strong evidence for a key role of the tRNA pool in shaping natural selection on codon usage. As further evidence for the role of the tRNA pool, changes to the expression levels of the Elongator Complex – a key tRNA modification enzyme

– are correlated with changes to natural selection on codons AAA, CAA, and GAA across species. Previous studies investigated how codon usage changes with the tRNA pool on macroevolutionary timescales [5, 20], but ours is the first to directly test how variation in natural selection on codon usage reflect evolutions variation of tRNA pool at the level of individual codons.

Across species changes to natural selection on codon usage are correlated with the waiting times, but many of these correlations are moderate to weak. Many factors may be obscuring this relationship. First, empirical estimates of waiting times from ribosome profiling data are imperfectly (albeit moderately to strongly) correlated with tRNA-based proxies [39, 40]. Second, our estimates of natural selection are expected to average over different selective pressures that may also scale with gene expression, further obscuring the relationship between selection on codon usage and the tRNA pool. For example, selection against translation errors (missense errors, nonsense errors, etc.) is also expected to scale with gene expression [12, 41, 42], but the most efficient codon may not always be the most accurate codon [43]. Additionally, selective pressures on CUB restricted to specific regions of a protein-coding sequence, such as selection against mRNA secondary structure around the 5'-end [44], also shape adaptive CUB. How different selective pressures interplay to shape the observed CUB remains an open question and will necessitate the development of more nuanced models that can separate different forms of adaptive CUB. Finally, a key question regarding changes in natural selection on codon usage is the relative importance of changes to the effective population size N_e (which modulates the impact of genetic drift) vs. the unscaled selection coefficient s (note that our selection coefficients $\Delta\eta$ reflect the scaled selection coefficients $S = sN_e$ in a gene of average of expression $\phi = 1$). With the current data, we cannot decompose our scaled selection coefficients $\Delta\eta$ into the effective population size and the unscaled selection coefficients (which is a function of both waiting times and the energetic cost of ribosome pausing, see [24] for more details) As such, we cannot say which contributes more to across species variation in adaptive CUB. However, our work indicates changes to the unscaled selection coefficients s via changes to the tRNA pool are prominent in driving changes to natural selection on codon usage, and thus shaping variation in adaptive CUB across species.

Perhaps our most surprising finding was that microevolutionary processes shaping CUB across the Saccharomycotina subphylum are highly variable across closely related species. This was evident by (a) the generally poor agreement between the clustering of parameters and the phylogeny and (b) the overall low estimates of phylogenetic signal based on a multivariate version of Blomberg's K [30]. Interestingly, estimates of natural selection were more similar than mutation biases between closely related species. Consistent with this, estimates of stabilizing selection were generally greater for the mutation bias estimates than natural selection. This suggests the underlying molecular factors shaping mutation biases (e.g., mismatch repair) are generally under stronger stabilizing selection than those relevant to natural selection on codon usage (e.g., the tRNA pool). This should not distract from the larger point that both traits are highly variable across species, ultimately driving variation in CUB.

Hierarchical clustering revealed 34 of the 327 budding yeast were poorly fit by ROC-SEMPPR, at least relative to other species. On average, ROC-SEMPPR parameter estimates for these species were more weakly correlated with empirical estimates of gene expression and codon-specific waiting times. These correlations were often positive, suggesting possible isolated shifts in the direction of natural selection acting on codon usage. Based on the poor model fit, it is possible that these species (1) have reduced natural selection acting on codon usage below the drift barrier [45], possibly due to reduced effective population sizes or (2) additional evolutionary forces acting on codon usage that further obscure signals of natural selection related to protein synthesis. These species serve as excellent starting points for future studies to elucidate the complex interplay of evolutionary forces that shape CUB.

Materials and methods

We obtained genome sequences, associated annotation files, the Saccharomycotina species tree, and a list of one-to-one orthologs from previous work [23]. We excluded mitochondrial genes, protein-coding sequences with non-canonical start codons, internal stop codons, and sequences whose lengths were not a multiple of three from all analyses. We queried all protein sequences against a BLAST database built from sequences in the MiToFun database to identify and remove mitochondrial sequences (`http://mitofun.biol.uoa.gr/`). Empirical gene expression measurements were taken from [28] and the sources cited therein. Briefly, adapters for each sequence were trimmed using **fastp** [46], and genes were quantified using **kallisto** [47]. Transcripts-per-million (TPMs) were re-calculated for each transcript by rounding raw read counts to the nearest whole number [48].

Analyzing codon usage patterns with ROC-SEMPPR

The Ribosomal Overhead Cost version of the Stochastic Evolutionary Model of Protein Production Rates (ROC-SEMPPR) is implemented in a Bayesian framework. This allows for the simultaneous estimation of codon-specific selection coefficients and mutation bias, as well as gene-specific estimates of the evolutionary average gene expression by assuming gene expression follows a log-normal distribution [24]. ROC-SEMPPR does not require empirical gene expression data, meaning it can be applied to any species with annotated protein-coding sequences. For any amino acid with n_{aa} synonymous codons, the probability $p_{i,g}$ of observing codon i in gene g is defined by the equation

$$p_{i,g} = \frac{e^{-\Delta M_i - \Delta \eta_i \phi_g}}{\sum_j^{n_{aa}} e^{-\Delta M_j - \Delta \eta_j \phi_g}} \tag{1}$$

where ΔM_i and $\Delta \eta_i$ represent mutation bias and selection coefficient of codon i relative to a reference synonymous codon (arbitrarily chosen as the alphabetically last codon), and ϕ_g represents gene expression of gene g which follows from the steady-state distribution of fixed genotypes under selection-mutation-drift equilibrium [10, 24, 49]. For each gene, the observed codon counts for an amino acid are expected to follow a multinomial distribution with the probability of observing a codon defined by Equation 1. Given the codon counts and the assumption that gene expression follows a lognormal distribution, ROC-SEMPPR estimates the parameters that best fit the codon counts via a Markov Chain Monte Carlo simulation approach (MCMC). ROC-SEMPPR was fit to 327 species using the R package AnaCoDa [50]. For each species, the MCMC chains were run for 200,000 iterations, keeping every 10^{th} iteration. The first 50,000 iterations were discarded as burn-in. Two separate MCMC chains were run for each species and parameter estimates were compared to assess convergence.

Previous work with ROC-SEMPPR separated serine codons TCN (where N is any of the other four nucleotides) and AGY (where Y is C or T) into separate groups of codons for the analysis

[10, 24, 26]. ROC-SEMPPR assumes each mutation introduced to a population is fixed or lost before the arrival of the next mutation (i.e., "weak mutation" $N_e\mu << 1$). The model also assumes fixed amino acid sequences for all protein-coding sequences. As a result, going between these two groups of serine codons would require the fixation of a non-serine amino acid before returning to serine via the fixation of another mutation, violating the fixed amino acid sequence assumption. A local version of AnaCoDa was created to handle species for which CTG codes for serine. For these species, CTG was treated as a third codon group for serine, similar to ATG (methionine) or TGG (tryptophan), which have no synonyms.

Identifying within-genome variation in codon usage bias

We recently found numerous Saccharomycotina yeasts exhibit variation in the non-adaptive nucleotide biases shaping GC% within a genome that obscures signals of natural selection on codon usage [28]. We followed the same procedure to hypothesize genes evolving under different non-adaptive nucleotide biases across all 327 budding yeasts. For each species, correspondence analysis (CA) was applied to the absolute codon frequencies of each annotated protein-coding sequence using the ca R package [51]. Protein-coding sequences were then clustered into two groups based on the first four principal components from the CA using the CLARA algorithm implemented in the cluster R package, which is designed to perform k-medoids clustering on large datasets [52]. See our previous work for more details on the CLARA clustering algorithm [28]. For each species, the cluster with the lower median GC3% was designated as the "Lower GC3% Set", and the cluster with the higher median GC3% as the "Higher GC3% Set". ROC-SEMPPR was fit to the protein-coding sequences of each species, assuming selection coefficient and mutation bias parameters were the same between the two clusters, which we refer to as the "ConstMut" model. Similarly, the protein-coding sequences of each species were also fit while allowing the mutation bias to vary across sequences based on their assigned cluster, which we will refer to as the "VarMut" model. For the VarMut model, selection coefficients were assumed to be the same across the Higher and Lower GC3% sets.

ROC-SEMPPR predictions of gene expression ϕ for each protein-coding sequence were compared to empirical estimates of mRNA abundance using the Spearman correlation coefficient using processed data from our previous work [28]. For species lacking RNA-seq data, we compared each species' predicted gene expression to the empirical gene expression of its closest relative for which the latter was available. This is reasonable given that mRNA abundances in yeasts evolve under stabilizing selection [53]. The VarMut model was considered an improved fit over the ConstMut model if the correlation between predicted and empirical gene expression estimates was 25% greater relative to the ConstMut model.

Comparing codon-specific parameters across species

Across-species and across-codon variation in selection coefficients $\Delta\eta$ and mutation bias ΔM were compared using hierarchical clustering using the "complete linkage" algorithm with distances determined by the $1 - R$, where R is the pairwise Spearman rank correlation. Results were visualized using heatmaps as implemented in the R package **ComplexHeatmap**. For each codon-specific parameter estimate, we quantified the similarity between the phylogenetic tree and the hierarchical clustering via a cophenetic correlation, which measures how well two dendrograms preserve the pairwise distances between data points. As orthogonal analyses, we quantified the overall phylogenetic signal (i.e., how similar species are to their closest relatives) via a multivariate version of Blomberg's K (K_{multi}) as implemented in the R package **geomorph**. As stabilizing selection toward a single optimum value degrades phylogenetic signal [31], we also fit an Ornstein-Uhlenbeck [54] model of trait evolution via the R package **geiger** [32] to the codon-specific parameters (using the standard deviation of the posterior distribution as measurement error), with the strength of stabilizing selection α compared between selection and mutation bias estimates using a Wilcox rank sum test. For the VarMut species, we performed these analyses using mutation bias estimates from either the Lower GC3% set or the Higher GC3% set.

Determining potential causes for across-species variation in codon-specific parameters

Across-species correlations between traits were assessed after performing phylogenetic independent contrasts (PIC) as implemented in the R package **ape**. Phylogenetic regressions were performed using the R package **phylolm** using Pagel's λ model [55]. Multiple comparisons were accounted for via the Benjamini-Hochberg procedure with a false discovery rate of 0.05.

Estimating codon-specific ribosome waiting times

Estimates of codon-specific elongation rates were obtained based on the weights used to calculate the tRNA Adaptation Index (tAI) [33]. We ran the latest version of tRNA-ScanSE on the genomic FASTA sequences with mitochondrial genomes removed to obtain a tRNA gene copy (tGCN) for each species under consideration [56]. To allow for potential variation in wobble efficiency across species, we estimated wobble parameters by maximizing the Spearman rank correlation coefficient between ROC-SEMPPR predicted gene expression and tAI using the R package **tAI** [57]. ROC-SEMPPR estimates reflect selection against a codon relative to a reference synonymous codon (implemented as the alphabetically last codon for each amino acid in **AnaCoDa**). As we are primarily interested in comparing tRNA-based waiting times to selection coefficients $\Delta\eta$, we calculated the log ratio of the weights between a codon and its respective reference codon i.e.,

$$w_{aa,i} = log(W_{aa_{ref}}/W_i) \tag{2}$$

where aa_{ref} indicates the reference codon for amino acid aa and W_i gives the unnormalized weight as calculated in tAI. This contrasts with the normal formulation of tAI, which typically normalizes all weights relative to the maximum weight across all codons (regardless of amino acid) [33]. In some cases, the reference codon for an amino acid could not be translated based on the given tRNA genes and standard wobble rules. As our goal is to test if across-species changes to selection are generally correlated with changes to the tRNA pool, we opted to drop these cases from our analyses rather than have potentially different reference codons for each species.

Comparing selection coefficients $\Delta\eta$ with tRNA modification gene expression

Removal of the mcmc^{5}s^{2}U modification at U$_{34}$ of tRNA recognizing codon AAA (Lys), GAA (Glu), and CAA (Gln) increases ribosome waiting times at these codons in *S. cerevisiae* [36, 37]. The protein-coding sequences encoding tRNA modifications enzymes that make up the catalytic center of the Elongator Complex – IKI3 (ELP1), ELP2, and ELP3 – were identified in each Saccharomycotina yeasts tRNA modifications known to impact translation efficiency using a previously determined list of one-to-one orthologs [23]. The relevant ROC-SEMPPR gene expression estimates ϕ were obtained for each gene and were compared to the selection coefficients $\Delta\eta$ for AAA, GAA, and CAA using multivariate phylogenetic regressions assuming a Pagel's λ model (3 regressions, one for each codon) as implemented in the R package **phylolm**. In addition to the effects of each of the Elongator Complex proteins, we also included the effects of the corresponding tGCN (specifically, log(tGCN)) for each of the codons and the genome-wide GC% in our regressions, i.e. $\Delta\eta_{Codon} \sim \phi_{IKI3} + \phi_{ELP2} + \phi_{ELP3} + tGCN_{Codon} + GC\% + Intercept$. Each independent variable was transformed into a Z-score to make the effects of each variable more comparable.

Acknowledgments

We thank Mike Gilchrist, Edward Wallace, Matt Pennell, and Abbe LaBella for their helpful comments and discussions throughout this project. A.L.C. was supported by the NIH-funded Rutgers INSPIRE IRACDA Postdoctoral Program (grant #GM093854). Additionally, this work was supported by the National Science Foundation (DBI 1936046 to P.S.) and the National Institutes of Health (R35 GM124976 to P.S).

Competing interests

P.S. is a Director at the stealth mode biotech.

Data availability

All data are publicly available via the citations provided in this article (see doi:10.6084/m9.figshare.5854692.v1). Scripts and R notebooks for re-creating our analysis and visualizations can be found at `https://github.com/acope3/Saccharomycotina_subphylum_analysis`.

References

[1] Knight RD, Freeland SJ, Landweber LF. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome biology. 2001 4;2:1-13. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2001-2-4-research0010.

[2] Drummond DA, Wilke CO. The evolutionary consequences of erroneous protein synthesis. Nature Reviews Genetics. 2009;10:715-24.

[3] Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Research. 2005;33:1141-53.

[4] Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics. 2011;12:32-42.

[5] Novoa EM, Pavon-Eternod M, Pan T, Pouplana LRD. A role for tRNA modifications in genome structure and codon usage. Cell. 2012 3;149:202-13. Available from: http://www.cell.com/article/S0092867412002127/fulltexthttp://www.cell.com/article/S0092867412002127/abstracthttps://www.cell.com/cell/abstract/S0092-8674(12)00212-7.

[6] Novoa EM, Jungreis I, Jaillon O, Kellis M. Elucidation of Codon Usage Signatures across the Domains of Life. Molecular Biology and Evolution. 2019 10;36:2328-39. Available from: https://academic.oup.com/mbe/article/36/10/2328/5492082.

[7] Bulmer M. The selection-mutation-drift theory of synonymous codon usage. Genetics. 1991;129:897-907.

[8] 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. Journal of Molecular Biology. 1981 9;151:389-409. Available from: https://linkinghub.elsevier.com/retrieve/pii/0022283681900036.

[9] Sharp PM, Li W. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research. 1987;15:1281-95.

[10] Shah P, Gilchrist M. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proceedings of the National Academy of Sciences of the United States of America. 2011;108:10231-6.

[11] Labella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A. Variation and selection on codon usage bias across an entire subphylum. PLoS Genetics. 2019 7;15:e1008304. Available from: https://doi.org/10.1371/journal.pgen.1008304.

[12] Drummond DA, Wilke CO. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell. 2008;134:341-52.

[13] Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. Rate-Limiting Steps in Yeast Protein Translation. Cell. 2013;153:1589-601.

[14] Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proceedings of the National Academy of Sciences of the United States of America. 2018 5;115:E4940-9.

[15] Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annual Review of Genomics and Human Genetics. 2009 9;10:285-311. Available from: www.annualreviews.org.

[16] Lassalle F, Périan S, Bataillon T, Nesme X, Duret L, Daubin V. GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. PLOS Genetics. 2015;11:e1004941. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004941.

[17] Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, et al. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. Molecular Biology and Evolution. 2018 5;35:1092-103. Available from: https://academic.oup.com/mbe/article/35/5/1092/4829954.

[18] Botzman M, Margalit H. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. Genome Biology. 2011 10;12:1-11. Available from: https://link.springer.com/articles/10.1186/gb-2011-12-10-r109https://link.springer.com/article/10.1186/gb-2011-12-10-r109.

[19] Clément Y, Sarah G, Holtz Y, Homa F, Pointet S, Contreras S, et al. Evolutionary forces affecting synonymous variations in plant genomes. PLOS Genetics. 2017 5;13:e1006799. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006799.

[20] Wint R, Salamov A, Grigoriev IV. Kingdom-Wide Analysis of Fungal Protein-Coding and tRNA Genes Reveals Conserved Patterns of Adaptive Evolution. Molecular Biology and Evolution. 2022 2;39. Available from: https://academic.oup.com/mbe/article/39/2/msab372/6513383.

[21] Weibel CA, Wheeler AL, James JE, Willis SM, Masel J. A new codon adaptation metric predicts vertebrate body size and tendency to protein disorder. eLife. 2023;12:RP87335. Available from: https://elifesciences.org/reviewed-preprints/87335.

[22] Kotari I, Kosiol C, Borges R. The Patterns of Codon Usage between Chordates and Arthropods are Different but Co-evolving with Mutational Biases. Molecular Biology and Evolution. 2024 5;41. Available from: https://dx.doi.org/10.1093/molbev/msae080.

[23] Shen XX, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, et al. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell. 2018 11;175:1533-45.e20.

[24] Gilchrist MA, Chen WC, Shah P, Landerer CL, Zaretzki R. Estimating Gene Expression and Codon-Specific Translational Efficiencies, Mutation Biases, and Selection Coefficients from Genomic Data Alone. Genome Biology and Evolution. 2015;7:1559-79.

[25] Cope AL, Hettich RL, Gilchrist MA. Quantifying codon usage in signal peptides: Gene expression and amino acid usage explain apparent selection for inefficient codons. Biochimica et Biophysica Acta - Biomembranes. 2018;1860:2479-85.

[26] Landerer C, O'Meara BC, Zaretzki R, Gilchrist MA. Unlocking a signal of introgression from codons in Lachancea kluyveri using a mutation-selection model. BMC Evolutionary Biology. 2020 8;20:109. Available from: `https://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-020-01649-w`.

[27] Cope AL, Gilchrist MA. Quantifying shifts in natural selection on codon usage between protein regions: a population genetics approach. BMC Genomics. 2022 5;23:408. Available from: `https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-022-08635-0http://creativecommons.org/publicdomain/zero/1.0/`.

[28] Cope AL, Shah P. Intragenomic variation in non-adaptive nucleotide biases causes underestimation of selection on synonymous codon usage. PLOS Genetics. 2022 6;18:e1010256.

[29] dos Reis M, Wernisch L. Estimating Translational Selection in Eukaryotic Genomes. Molecular Biology and Evolution. 2009 2;26:461. Available from: `/pmc/articles/PMC2639113//pmc/articles/PMC2639113/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2639113/`.

[30] Adams DC. A Generalized K Statistic for Estimating Phylogenetic Signal from Shape and Other High-Dimensional Multivariate Data. Systematic Biology. 2014 9;63:685-97. Available from: `https://dx.doi.org/10.1093/sysbio/syu030`.

[31] Revell LJ, Harmon LJ, Collar DC. Phylogenetic Signal, Evolutionary Process, and Rate. Systematic Biology. 2008 8;57:591-601. Available from: `https://dx.doi.org/10.1080/10635150802302427`.

[32] Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, Fitzjohn RG, et al. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics. 2014 8;30:2216-8. Available from: `https://dx.doi.org/10.1093/bioinformatics/btu181`.

[33] dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Research. 2004;32:5036-44.

[34] Rogalski M, Karcher D, Bock R. Superwobbling facilitates translation with reduced tRNA sets. Nature Structural and Molecular Biology. 2008 2;15:192-8.

[35] Vieira-Silva S, Rocha EPC. The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. PLoS Genetics. 2010;6:1-15.

[36] Nedialkova DD, Leidel SA. Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. Cell. 2015 6;161:1606-18.

[37] Chou HJ, Donnard E, Gustafsson HT, Garber M, Rando OJ. Transcriptome-wide Analysis of Roles for tRNA Modifications in Translational Regulation. Molecular Cell. 2017 12;68:978-92.e4.

[38] Wallace EWJ, Airoldi EM, Drummond DA. Estimating Selection on Synonymus Codon Usage from Noisy Experimental Data. Molecular Biology and Evolution. 2013;30:1438-53.

[39] Weinberg DE, Shah P, Eichhorn SW, Hussmann JA, Plotkin JB, Bartel DP. Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. Cell Reports. 2016 2;14:1787-99.

[40] Wu CCC, Zinshteyn B, Wehner KA, Green R. High-Resolution Ribosome Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during Cellular Stress. Molecular Cell. 2019 3;73:959-70.e5.

[41] Kurland CG. Translational Accuracy and the Fitness of Bacteria. Annual Review of Genetics. 1992 12;26:29-50. Available from: http://www.annualreviews.org/doi/10.1146/annurev.ge.26.120192.000333.

[42] Gilchrist MA, Shah P, Zaretzki R. Measuring and Detecting Molecular Adaptation in Codon Usage Against Nonsense Errors During Protein Translation. Genetics. 2009;183:1493-505.

[43] Shah P, Gilchrist MA. Effect of Correlated tRNA Abundances on Translation Errors and Evolution of Codon Usage Bias. PLoS Genetics. 2010;6:1-9.

[44] Hockenberry AJ, Sirer MI, Amaral LAN, Jewett MC. Quantifying Position-Dependent Codon Usage Bias. Molecular Biology and Evolution. 2014;31:1880-93.

[45] Lynch M. The frailty of adaptive hypotheses for the origins of organismal complexity. Proceedings of the National Academy of Sciences of the United States of America. 2007 5;104:8597-604. Available from: www.nasonline.org/adaptationandcomplexdesign.

[46] Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 9;34:i884-90. Available from: https://github.com/OpenGene/fastp.

[47] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016 5;34:525-7. Available from: http://www.nature.com/.

[48] Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences. 2012 12;131:281-5. Available from: http://link.springer.com/10.1007/s12064-012-0162-3.

[49] Sella G, Hirsh AE. The application of statistical physics to evolutionary biology. Proceedings of the National Academy of Sciences of the United States of America. 2005 7;102:9541-6. Available from: https://www.pnas.org/content/102/27/9541https://www.pnas.org/content/102/27/9541.abstract.

[50] Landerer C, Cope A, Zaretzki R, Gilchrist MA. AnaCoDa: analyzing codon data with Bayesian mixture models. Bioinformatics. 2018;34:bty138. Available from: http://dx.doi.org/10.1093/bioinformatics/bty138.

[51] Nenadic O, Greenacre M. Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. Journal of Statistical Software. 2007;20:1-13.

[52] Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions; 2019.

[53] Thompson DA, Roy S, Chan M, Styczynsky MP, Pfiffner J, French C, et al. Evolutionary principles of modular gene regulation in yeasts. eLife. 2013 6;2013.

[54] Hansen TF. Stabilizing selection and the comparative analysis of adaptation. Evolution. 1997 10;51:1341-51. Available from: http://doi.wiley.com/10.1111/j.1558-5646.1997.tb01457.x.

[55] Revell LJ. Phylogenetic signal and linear regression on species data. Methods in Ecology and Evolution. 2010 12;1:319-29. Available from: http://doi.wiley.com/10.1111/j.2041-210X.2010.00044.x.

[56] Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. Nucleic Acids Research. 2021;49.

[57] dos Reis M. tAI: The tRNA adaptation index; 2016. R package version 0.2.