PLOS GENETICS

RESEARCH ARTICLE

Variation and selection on codon usage bias across an entire subphylum

Abigail L. LaBella[1], Dana A. Opulente[2], Jacob L. Steenwyk[1], Chris Todd Hittinger[2], Antonis Rokas[1]*

1 Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee, United States of America, **2** Laboratory of Genetics, Genome Center of Wisconsin, DOE Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, J. F. Crow Institute for the Study of Evolution, University of Wisconsin–Madison, Wisconsin, United States of America

* antonis.rokas@vanderbilt.edu

Citation: LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A (2019) Variation and selection on codon usage bias across an entire subphylum. PLoS Genet 15(7): e1008304. https://doi.org/10.1371/journal.pgen.1008304

Editor: Gregory S. Barsh, Stanford University School of Medicine, UNITED STATES

Received: July 5, 2019

Accepted: July 11, 2019

Published: July 31, 2019

Data Availability Statement: All analyses were done on publicly available and published genome assemblies and detailed information, including the publications from where they were retrieved from, is provided in S1 Table. Additional data is shared through figshare (https://figshare.com/collections/Variation_and_selection_on_codon_usage_bias_across_an_entire_subphylum/4498292). This collection includes the Saccharomycotina genome assembly contigs that were removed as putatively mitochondrial contigs, the Saccharomycotina genome annotations filtered for putatively mitochondrial CDSs and, the codon optimization

Abstract

Variation in synonymous codon usage is abundant across multiple levels of organization: between codons of an amino acid, between genes in a genome, and between genomes of different species. It is now well understood that variation in synonymous codon usage is influenced by mutational bias coupled with both natural selection for translational efficiency and genetic drift, but how these processes shape patterns of codon usage bias across entire lineages remains unexplored. To address this question, we used a rich genomic data set of 327 species that covers nearly one third of the known biodiversity of the budding yeast subphylum Saccharomycotina. We found that, while genome-wide relative synonymous codon usage (RSCU) for all codons was highly correlated with the GC content of the third codon position (GC3), the usage of codons for the amino acids proline, arginine, and glycine was inconsistent with the neutral expectation where mutational bias coupled with genetic drift drive codon usage. Examination between genes' effective numbers of codons and their GC3 contents in individual genomes revealed that nearly a quarter of genes (381,174/1,683,203; 23%), as well as most genomes (308/327; 94%), significantly deviate from the neutral expectation. Finally, by evaluating the imprint of translational selection on codon usage, measured as the degree to which genes' adaptiveness to the tRNA pool were correlated with selective pressure, we show that translational selection is widespread in budding yeast genomes (264/327; 81%). These results suggest that the contribution of translational selection and drift to patterns of synonymous codon usage across budding yeasts varies across codons, genes, and genomes; whereas drift is the primary driver of global codon usage across the subphylum, the codon bias of large numbers of genes in the majority of genomes is influenced by translational selection.

Author summary

Synonymous mutations in genes have no effect on the encoded proteins and were once thought to be evolutionarily neutral. By examining codon usage bias across codons, genes,

scores generated by stAI-calc for all coding sequences in all genomes examined. All other results and generated data are within the paper and its Supporting Information files.

Funding: This work was supported in part by the National Science Foundation (https://www.nsf.gov) (DEB-1442113 and DEB-1442148) and the DOE Great Lakes Bioenergy Research Center (https://www.glbrc.org) (DOE Office of Science DE-SC0018409). CTH is a Pew Scholar in the Biomedical Sciences, Vilas Faculty Early Career Investigator, and H. I. Romnes Faculty Fellow, supported by the Pew Charitable Trusts (https://www.pewtrusts.org), the Vilas Trust Estate (https://www.rsp.wisc.edu/Vilas), and the Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation (https://www.warf.org), respectively. AR is supported by a Guggenheim fellowship (https://www.gf.org/about/fellowship). This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University (https://www.vanderbilt.edu/accre). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

and genomes of 327 species in the budding yeast subphylum, we show that synonymous codon usage is shaped by both neutral processes and selection for translational efficiency. Specifically, whereas codon usage bias for most codons appears to be strongly associated with mutational bias and largely driven by genetic drift across the entire subphylum, patterns of codon usage bias in a few codons, as well as in many genes in nearly all genomes of budding yeasts, deviate from neutral expectations. Rather, the synonymous codons used within genes in most budding yeast genomes are adapted to the tRNAs present within each genome, a result most likely due to translational selection that optimizes codons to match the tRNAs. Our results suggest that patterns of codon usage bias in budding yeasts, and perhaps more broadly in fungi and other microbial eukaryotes, are shaped by both neutral and selective processes.

Introduction

One of the first insights drawn from DNA sequence analyses was that synonymous codons are used both non-randomly and in taxon-specific patterns [1–3]. These results were surprising given that synonymous codon changes do not alter primary protein structure (i.e., they are silent) and were therefore previously assumed to be selectively neutral. Two major explanations have been put forth to account for the non-random variation in codon usage seen within and across species, namely natural selection and neutral processes, such as mutational bias coupled with genetic drift.

The discovery that codon usage is correlated with both the abundance of transfer RNA molecules in the genome and with gene expression levels raised the hypothesis that optimization of codons to match the available tRNA pool (or tRNAome) promotes or regulates translation and suggested a key role for codon usage in translational dynamics [4–10]. It is now well established that codon usage influences multiple cellular processes, especially translation. For example, usage of codons corresponding to the tRNA pool, known as codon optimization, has been linked to increased translation speed [11–14], accurate tRNA pairing [15, 16], suppressed premature cleavage and polyadenylation of transcripts [17], and mRNA stability [11, 18]. Conversely, non-optimal codon usage has been associated with translation initiation [19], accurate protein folding [20–22], and signal recognition particle detection [23]. These molecular discoveries are complemented by a plethora of examples where specific synonymous substitutions have substantial fitness [24–27] and phenotypic effects in organisms across the tree of life, including *Escherichia coli* [28], *Saccharomyces cerevisiae* [29, 30], *Drosophila melanogaster* [31], and humans [32–34]. In summary, there is now substantial evidence to suggest that codon usage bias of certain codons in certain species is under strong selection—often through translational mechanisms.

In the absence of selection or in populations where genetic drift is more powerful than selection, patterns of codon usage bias will reflect the effects of genome-wide mutational pressures, such as mutational bias or GC-biased gene conversion [35–39]. This was first suspected for species with extreme GC composition biases, such as the Gram positive bacterium *Mycoplasma capricolum*, which has a genomic GC composition of 25%, and only 2% of its codons end with G or C [40]. For species like *M. capricolum*, it was hypothesized that biased genome-wide mutational processes, such as mutational bias towards A/T bases and GC-biased gene conversion, would drive patterns of codon usage bias. GC-biased gene conversion has been shown to influence the GC content of third codon positions in an evolutionarily neutral

manner in mammals, as well as at recombination hotspots in yeasts [41, 42]. Mutational bias has been proposed as the major driver of codon usage bias in diverse studies in a variety of lineages, including bacteria, archaea, plants, and animals [37, 38, 43, 44]. Even in the presence of selection on synonymous codon sites, it has been proposed that background substitution drives codon preference in organisms with widely different GC compositions [45]. Thus, major differences in codon usage patterns between species are often considered to be primarily driven by neutral mutational changes in GC content [36, 37].

Selective and neutral explanations of codon usage bias are not mutually exclusive, and pioneers in this field were quick to suggest that codon bias is due to a balance between neutral and selective processes [40, 46, 47]. It is unclear, however, what that balance is, how it varies across levels of biological organization (e.g., codons, genes, genomes) and across lineages, and what factors influence the balance [12, 36, 38, 40, 48, 49].

Budding yeasts (subphylum Saccharomycotina, phylum Ascomycota) present a unique opportunity to examine the impact of neutral and selective processes on codon usage bias for several reasons. First, genomes and genome annotations of 332 species across the subphylum recently became available [50], providing a state-of-the-art data set for the study of codon usage bias. Second, the genomic diversity across budding yeasts is comparable to the divergence between different animal phyla or between *Arabidopsis* and green algae, offering us the opportunity to examine variation in patterns of codon usage bias across a highly diverse lineage. Third, budding yeasts exhibit genetic code diversity and are the only known lineage with nuclear codon reassignments. Specifically, three different clades of buddying yeasts have undergone a reassignment of the CUG codon from leucine to serine (two clades) or alanine (one clade) [51–55]. Codon reassignments in the Saccharomycotina provide both a challenge and an opportunity in comparing codon usage bias across the subphylum. Finally, for the majority of budding yeast species in our data set we also have metabolic trait (285 species) and isolation environment (174 species) information, which not only illustrates the ecological diversity of this group but allows us to test for other contributors to codon usage bias [56, 57].

To examine codon usage bias at the codon, gene, and genome levels, we examined the genomes of 327 budding yeast species in the subphylum Saccharomycotina. Analysis of codon usage bias, measured by relative synonymous codon usage (RSCU) revealed diversity in usage at all three levels (codon, gene, genome) examined. This variation in RSCU was highly correlated with GC composition when assessed broadly across the subphylum. Furthermore, the relationship between the relative frequency of each codon and the GC composition of the 3rd codon position showed very small deviations from the neutral expectation, except for codons for three amino acids (proline, arginine, and glycine). However, at the gene level, nearly a quarter of all genes surveyed (381,174/1,683,203; 23%) did not fit the neutral expectation of the relationship between the effective number of codons and synonymous GC composition. In 94% (308/327) of the budding yeast genomes, the overall fit of genes to the neutral expectation was very low. Investigation of possible causes of this deviation revealed that 81% (264/ 327) of budding yeast genomes exhibited moderate-to-high levels translational selection on codon usage bias. While there was no significant correlation between the total number of metabolic traits or isolation environments and selection, the strength of selection was significantly correlated with genomic tRNA gene content (tRNAome). These results suggest that translational selection on codon bias is widespread, but not ubiquitous, in the budding yeast subphylum. Our inference of strong translational selection on codon usage bias suggests that translational regulation has played a major role in the evolution of this group.

Methods

Sequence data

Genomic sequence and annotation data were obtained from a recent comparative genomic study of 332 budding yeast genomes [50] (S1 Table). Genomes of five species from the CUG-Alanine clade were removed from this analysis as their codon reassignment was discovered recently [53, 54] and could not be accounted for by any existing software. To remove mitochondrial genome sequences from the remaining 327 budding yeast genomes, we employed blastn, version 2.6.0+ [58, 59] with 56 partial or complete Saccharomycotina mitochondrial genomes (S2 Table) as our input queries. Hits that had 30 percent or more sequence identity to mitochondrial sequences were removed from our analyses. Similarly, protein-coding gene sequence data from the 327 genomes were filtered for mitochondrial genes by blasting (blastx) against mitochondrial protein-coding sequence data from 37 Saccharomycotina species (S3 Table). The coding sequences were further filtered to conform to the required input for the species-specific tRNA adaptation calculations by stAIcalc, version 1.0 [60]. This filtering step removed all coding sequences that did not begin with the start codon ATG, did not have a whole number of codons, or were shorter than 100 codons (S1 Table). Codons containing ambiguous bases were also removed.

Codon usage bias calculations

To examine the variation in codon usage across the yeast subphylum, we calculated the relative synonymous codon usage (RSCU) for each codon in the 1,683,203 protein-coding genes of the 327 budding yeast genomes that remained after filtering. RSCU is the observed frequency of a synonymous codon divided by the frequency expected if all the synonymous codons were used equally [9]. We computed RSCU values using DAMBE7, version 7.0.28 [61], because it allowed us to accommodate the known nuclear codon reassignment in the CUG-Ser1 and CUG-Ser2 clades [51–55].

To examine broad patterns of codon usage, hierarchical clustering of all RSCU values for each species was calculated and visualized in the R programming environment. To investigate which codons drive between-species differences in codon usage, we performed correspondence analysis of RSCU values [3]. This technique is highly suitable and informative because it reduces the high number of dimensions present in codon usage statistics into a very small number of axes [62, 63].

To examine the influence of phylogeny on the observed variation in codon bias, we computed two measures of phylogenetic signal in R, Pagel's λ [64] and Blomberg's K [65]. The phylogeny used for this analysis was obtained through maximum likelihood-based inference from a data matrix comprised of 2,408 genes obtained from Shen et al. [50].

Mutational bias and codon usage

To assess the role of mutational bias in determining the observed patterns of codon bias in the yeast subphylum, we tested the observed patterns against neutral expectations, both across species and across codons. Between-species patterns in codon usage bias were measured by calculating the Pearson's correlation of the RSCU of each codon against the GC composition of the 3rd codon position (GC3) across all genes in each genome, for each of the 327 species. To account for the observed phylogenetic dependence within both variables, we also assessed the relationship between RSCU and GC3 using the phylogenetic generalized least squares (PGLS). The influence of mutational bias within each set of codons encoding an amino acid was assessed by comparing the equilibrium solutions for relative codon frequencies based on GC3

content generated by Palidwor et al. [38] to the empirical values. Observed relative codon frequencies were calculated as the total number of observations of a codon divided by the total number of observations of the corresponding amino acid. Total codon counts within the genomes were calculated in DAMBE version 7.0.28 [61]. For each codon, predicted values of relative frequency were generated from the corresponding equilibrium solution. R^2 values were then calculated based on the predicted and empirical relative frequency values. Data from the 98 genomes present in the CUG-Ser1 and CUG-Ser2 clades were removed from the analyses of the amino acids leucine and serine.

To assess the influence of mutational bias within every genome, we compared the effective number of codons (ENC) [66] of each gene to the synonymous GC3 proportion of that gene. The ENC for each gene within the 327 genomes was computed in codonW (v1.4.2; http://codonw.sourceforge.net/) which does not allow for CUG codon reassignment. This distribution was compared against the predicted neutral distribution proposed by dos Reis et al. [67] using the suggested parameters. This neutral distribution is a modified version of Wright's proposed function [66] for calculating ENC [67]. We computed an R^2 value between the observed and empirical ENC values based on the GC3 of each gene. To ensure that R^2 values were not driven by phylogenetic signal, we calculated Blomberg's K for the R^2 values. Additionally we investigated the role of gene length in the deviation from the neutral expectation by comparing the distribution of lengths between neutral genes and those that deviate by 10% or 20% from the neutral expectation using a Wilcoxon Rank Sum test [68, 69].

Calculation of selection on codon usage

To determine if selection on translational processes has optimized the codon usage within each species, we tested if there is a significant correlation between the selective pressure on a gene and its level of optimization to the tRNAome for every genome. First, the species-specific value for each codon's relative adaptiveness (wi) was calculated in stAIcalc, version 1.0 [60]. Calculation of wi values requires genomic tRNA counts, which we calculated in tRNAscan-SE 2.0 for all species [70]. The results from tRNAscan-SE 2.0 correctly identified the CUG-Ser1 and CUG-Ser2 tRNAs that have a CAG anticodon but the recognition elements for serine (S4 Table). The species-specific tRNA adaptation index of each gene was then calculated by taking the geometric mean of all wi values for the codons (except the start codon). One drawback of stAIcalc is that it does not account for the nuclear codon reassignment in the CUG-Ser1 and CUG-Ser2 clades. Therefore, we also tested all genomes after removing all CUG codons from all sequences.

To test whether selection has influenced codon usage bias, we calculated the S-value proposed by dos Reis et al. [67]. This metric is the correlation between the tRNA adaptation index (stAI) and the confounded effects of the selection effect of the codon usage of a gene and uncontrollable random factors. Ultimately, the S-value measures the proportion of codon bias variance that cannot be explained by mutational bias or random factors alone. S-values were calculated with the R package tAI.R, version 0.2 (https://github.com/mariodosreis/tai) for each genome using the previously calculated stAI values. We calculated the S-value twice for each genome: once with CUG codons included and once without CUG codons. We also investigated the impact of gene length on the S-value by testing for a correlation between stAI value and gene length within a genome as well as comparing the S-value for a subset of genes whose protein products are over 1000 amino acids with the whole-genome value.

To test whether the S-value for a given genome significantly deviated from what would be expected under neutrality, we ran a permutation test. Specifically, we ran 10,000 permutations where each genome's wi values were randomly assigned to codons, the tAI values were then

recalculated for each gene, and the S-test was run on that permutation. A genome's observed S-value was considered statistically significant if it fell in the top 5% of the distribution formed by the 10,000 values obtained by the permutation analysis.

To investigate which features may influence the level of translational selection occurring within a genome, we tested the contributions of tRNAome size (calculated from tRNA-scan-SE), genome size, number of predicted coding sequences, total number of reported metabolic traits, and total number of reported isolation environments [50] on S-value variation. We performed linear regression analysis on individual and combinations of variables in R. In addition to the linear models, we tested a Gaussian distribution on a subset of features based on visual inspection. We also tested a PGLS analysis on S-value distribution to examine correlations that may be corrected by phylogenetic consideration. Finally, to check that genome completeness did not significantly influence our results, we measured the correlation between genome assembly N50 value and i) total number of tRNA genes, ii) the fit of genes to the neutral expectation of GC and ENC, and iii) the genome wide S-value.

Results

Budding yeast genomes exhibit substantial variation in codon usage

To measure variation in codon usage bias across budding yeast genomes, we measured the RSCU of each codon in each Saccharomycotina species. Hierarchical clustering of the codons revealed three major groups of codons (Fig 1). One group contained codons that were generally overrepresented (RSCU $>$ 1) in budding yeast genomes, which included A/U-ending codons and one G/C-ending codon (UUG). The next group contained mostly G/C-ending codons and two A/U-ending codons (AUA and GUA) that were generally underrepresented (RSCU $<$ 1) across budding yeast genomes. Finally, the smallest group contained A/U-ending codons (CUA, UUA, CGA, GGA, AUA, CCU, and GUA) that were relatively underrepresented across some budding yeast genomes as compared to the first set of A/U-ending codons. Interestingly, the underrepresentation of the CUA codon, which encodes leucine, was driven most strongly by the CUG-Ser1 and CUG-Ser2 clades where the CAG leucine codon has been recoded as serine (Fig 1).

Genome-level variation in codon usage corresponds with mutational bias

To summarize the overall variation in codon usage between species, we conducted a correspondence analysis on RSCU across all 327 species. The majority of the variation in codon usage between species was described by the first dimension of the correspondence analysis (66.891%; Fig 2), which was driven by differential usage of codons that vary at the third codon position, with the codons UUA, CGU, GGC and GUG making the largest contributions (S1A Fig). The second axis, which explained 7.093% of the variation in codon usage, showed some clustering by clade, with the CUG-Ser clade, the CUG-Ser2 clade and the only member of the Alloascoidea clade (*Alloascoidea hylecoeti*) clustering separately from the rest of the clades. This clustering was driven primarily by the codons CUA, CUG, UUG, and UUA (S1B Fig), with species in the CUG-Ser, CUG-Ser2 and *A*. *hylecoeti* being underrepresented in CUA and CUG and overrepresented in UUA and UUG. These four codons are all canonically decoded as leucine, suggesting that the reassignment of the CUG codon in the CUG-Ser1 and CUG-Ser2 clades is largely responsible for the separation of CUG-Ser1 and CUG-Ser2 clades from the rest. This result, however, does not explain the clustering of *A*. *hylecoeti*, which had the second highest overrepresentation of the UUA codon among the sampled Saccharomycotina, including the CUG-Ser1 and CUG-Ser2 clades. *A*. *hylecoeti* is the only representative genome of the major clade Alloascoideaceae in the dataset, and its genome contains tRNAs that decode

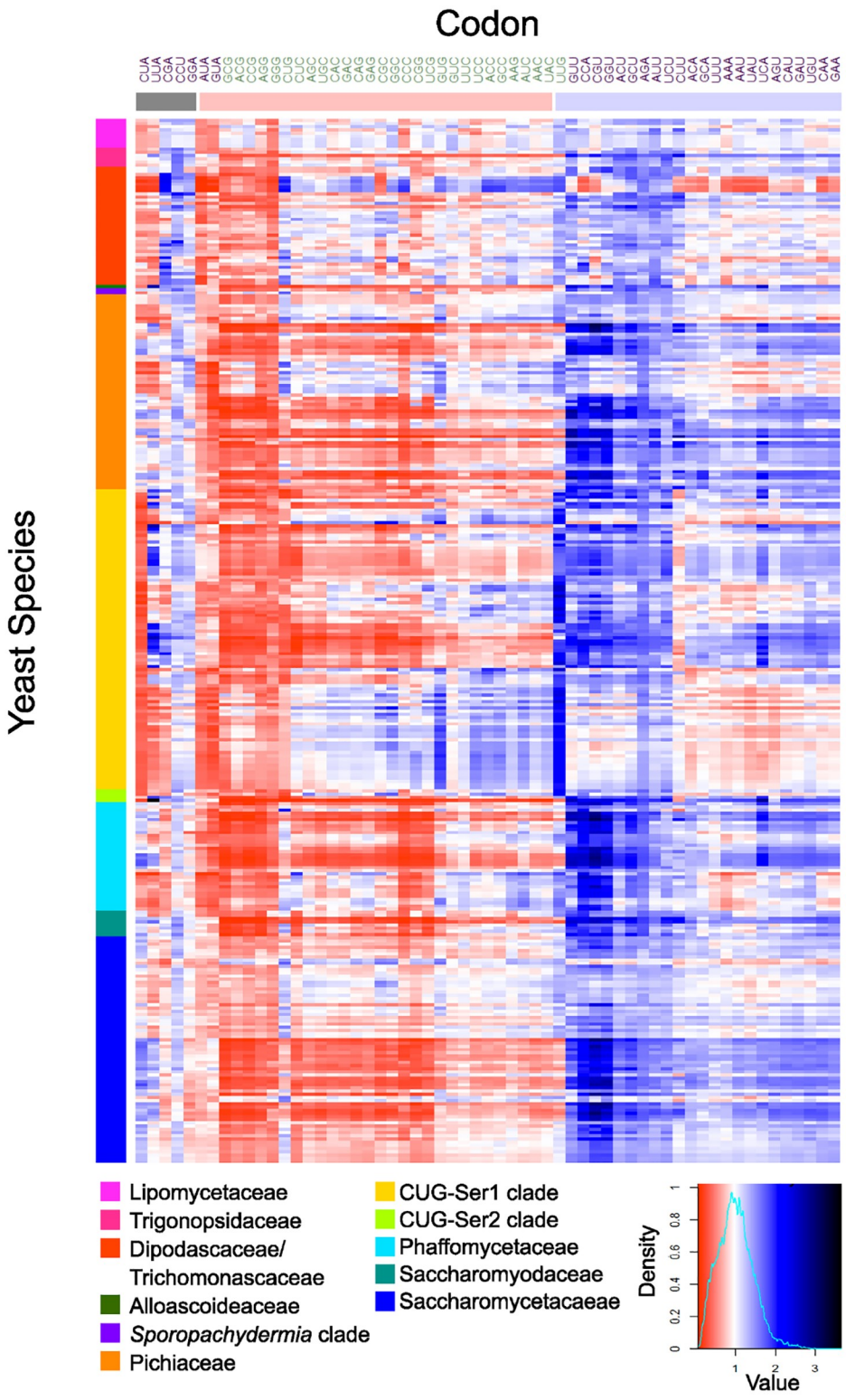

Fig 1. Relative synonymous codon usage (RSCU) analysis revealed an overrepresentation of A/Uending codons across most of the Saccharomycotina subphylum. Columns correspond to the 59 nondegenerate, non-stop codons; A/U-ending codons are shown in in purple font, and GC-ending codons are shown in green font. Rows correspond to the 327 Saccharomycotina species colored by major clade, following the recent genome-scale phylogeny of the subphylum [50]. Blue cells indicate overrepresented codons (RSCU > 1) and red cells indicate underrepresented

codons (RSCU < 1). Codons were clustered (using hierarchical clustering) by RSCU value into three general groups (shown by horizontal bars of different colors): underrepresented A/U-ending codons (grey bar), underrepresented codons mostly ending in G/C (red bar), and overrepresented codons mostly ending in A/U (blue bar).

https://doi.org/10.1371/journal.pgen.1008304.g001

all of the leucine codons, except for CUC. Moreover, there is no evidence of alternative codon usage in this species [71]. Additional species in this major clade will need to be sequenced to further understand why *A. hylecoeti* is an outlier in the relative usage of the UUA codon.

We next tested whether values of the RSCU metric across species had phylogenetic signal by measuring Pagel's λ [64] and Blomberg's K [65, 72, 73] (S5 Table). Pagel's λ tests for the presence of phylogenetic signal in a given trait using tree transformation—making the tree more or less star-like. Values for Pagel's λ vary from 0, which denotes that the trait absence of any phylogenetic signal, to 1, which denotes that the trait varies according to a Brownian model of random genetic drift. Codons' values for Pagel's λ ranged from 0.953 (for CUU) to 1 (for multiple codons) with p-values of <<0.001. These data suggest that codon usage between closely related species is more similar than expected under a Brownian motion model. Blomberg's K measures the ratio of trait variation among species to the contrasts variance. If the trait varies according to a Brownian model of random genetic drift Blomberg's K will equal 1. Blomberg's K however can be greater than 1 which indicates that variance in the trait occurs between clades (versus within). Interestingly, examination of Blomberg's K identified

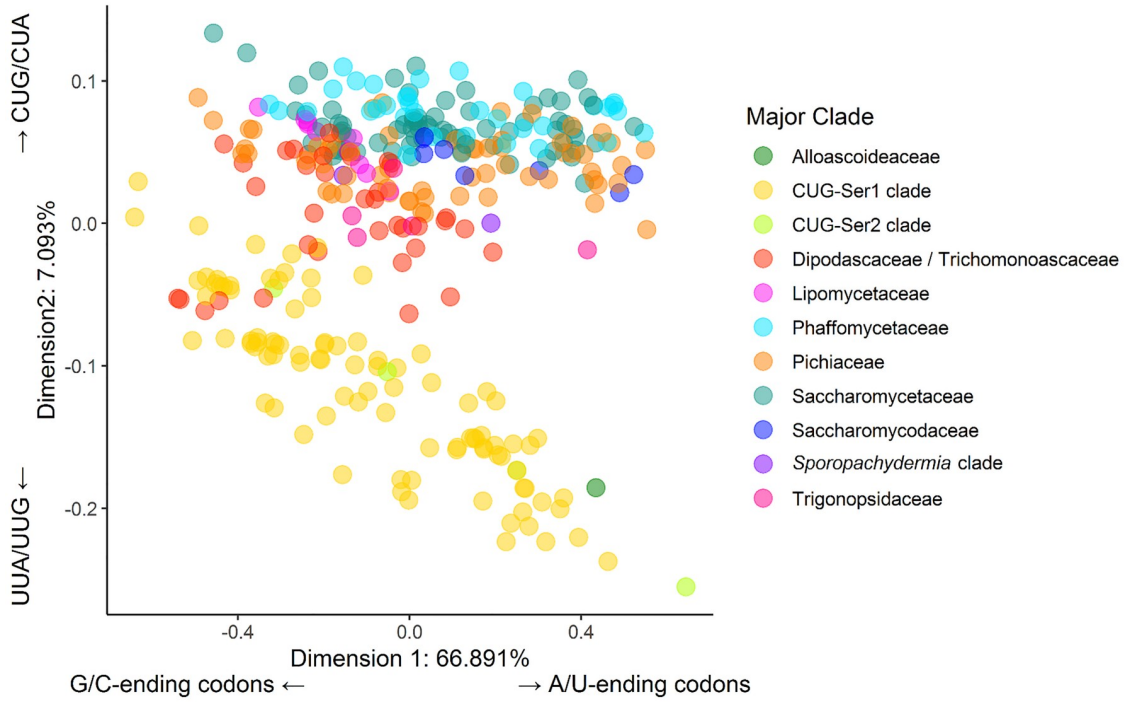

Fig 2. Differences in relative synonymous codon usage values between species are largely driven by variation in the usage of G/C- and A/U-ending codons. The plot shows each of the 327 budding yeast species examined in this study along the first two dimensions (the X and Y axes) of a correspondence analysis. Each axis is labeled with the percent variance explained by the corresponding dimension and the codons that are the major drivers of the observed variance. The first dimension, which explains nearly 67% of the variation between species, is driven by the differential usage of G/C- versus A/U-ending codons. The second dimension, which differentiates the CUG-Ser1 clade, the CUG-Ser2 clade, and one Alloascoideaceae species from the rest of the species in the subphylum, explains a much smaller fraction of the observed variation (about 7%) and is primarily driven by differential usage of the CUA, CUG, UUG, and UUA codons in the two groups.

https://doi.org/10.1371/journal.pgen.1008304.g002

between-clade variance (K>1) for only the codons CGA, CCA, UUG, and CUA, with the majority of the variance of the remaining codons present within major clades (K<1). Taken together, Pagel's λ and Blomberg's K suggest that the phylogenetic signal for most codons resides towards the tips of the phylogeny and explains variation in RSCU between closely related species. Two of the four codons that have phylogenetic signal deeper in the phylogeny (UUG and CUA) canonically encode leucine and were identified as drivers of the second explanatory axis in the correspondence analysis. This result suggests that the phylogenetic correlation between CGA, CCA, UUG and CUA is not restricted to closely related species and represents phylogenetically-driven differences between major clades, whereas the phylogenetic correlation of most other codons is only between closely related species and not between major clades.

Individual codon usage is driven by neutral and non-neutral forces

The correspondence analysis of RSCU revealed that major differences in codon usage are largely explained by differences in the usage of G/C- and A/U-ending codons (Fig 2). To determine the influence of neutral mutational bias on the usage of individual codons, we used Pearson's correlation and phylogenetic generalized least squares (PGLS) to examine the relationship between codon usage and mutational bias. Across all species, the Pearson's correlation of GC3 and RSCU revealed that all G/C-ending codons and two A/U-ending codons were positively correlated with GC3 (p-value < 0.001 in all cases) (S6 Table). The two A/U-ending codons that were positively correlated with GC composition bias were CUU and CGA. Interestingly, CGA was one of the codons identified by Blomberg's K as being phylogenetically differentiated between clades. It is, therefore, not surprising that CGA and CUU are negatively correlated with GC3 in the phylogenetically corrected PGLS analysis (Fig 3, S7 Table). In the PGLS analysis all A/U-ending codons are negatively correlated with GC3 and all G/C-ending codons are positively correlated with GC3. These results reveal that there is a strong correlation between mutational bias and codon usage at the genome level.

While the Pearson's correlation and PGLS analyses suggest that codon bias and GC composition due to mutational bias are correlated, these metrics do not account for the non-linear relationship between GC composition and codon usage. Therefore, we compared observed relative codon frequencies with equilibrium solutions generated by Palidwor et al. [38]. We compared the observed relative codon frequencies for every codon with the equilibrium solutions and measured fit using R^2 (Fig 4; S8 Table). All but one of the 2-fold degenerate codons had an R^2 value > 0.5 when compared to the neutral expectation (Fig 4C). For example, the codon GCC fit the neutral expectation very well ($R^2 = 0.671$; Fig 4a). The only 2-fold degenerate amino acid encoded by a codon that had an $R^2 < 0.5$ was phenylalanine ($R^2 = 0.236$). For the 3-fold and 4-fold degenerate codons, the R^2 values for the individual codons varied but, as previously noted [38], the summed predictions for G/C-ending codons and A/T-ending codons better fit the neutral expectation (Fig 4C: second column). The exceptions to this were proline, arginine, and glycine, which showed deviations from the neutral expectation even with the summed statistics (Fig 4B). To ensure that phylogenetic signal was not driving the deviations from the neutral expectation, we assessed Blomberg's K of the individual species' residuals used to compute the R^2 value. A total of 7 codons had Blomberg's K variances over 1 (Fig 4C: S8 Table), suggesting that deviations from the neutral expectation were driven by differences between major clades. Even after accounting for phylogenetic signal and the improved fit of the summed predictions, codons for proline, glycine, and arginine still showed deviations from the neutral expectation, suggesting that their usages are at least partially driven by

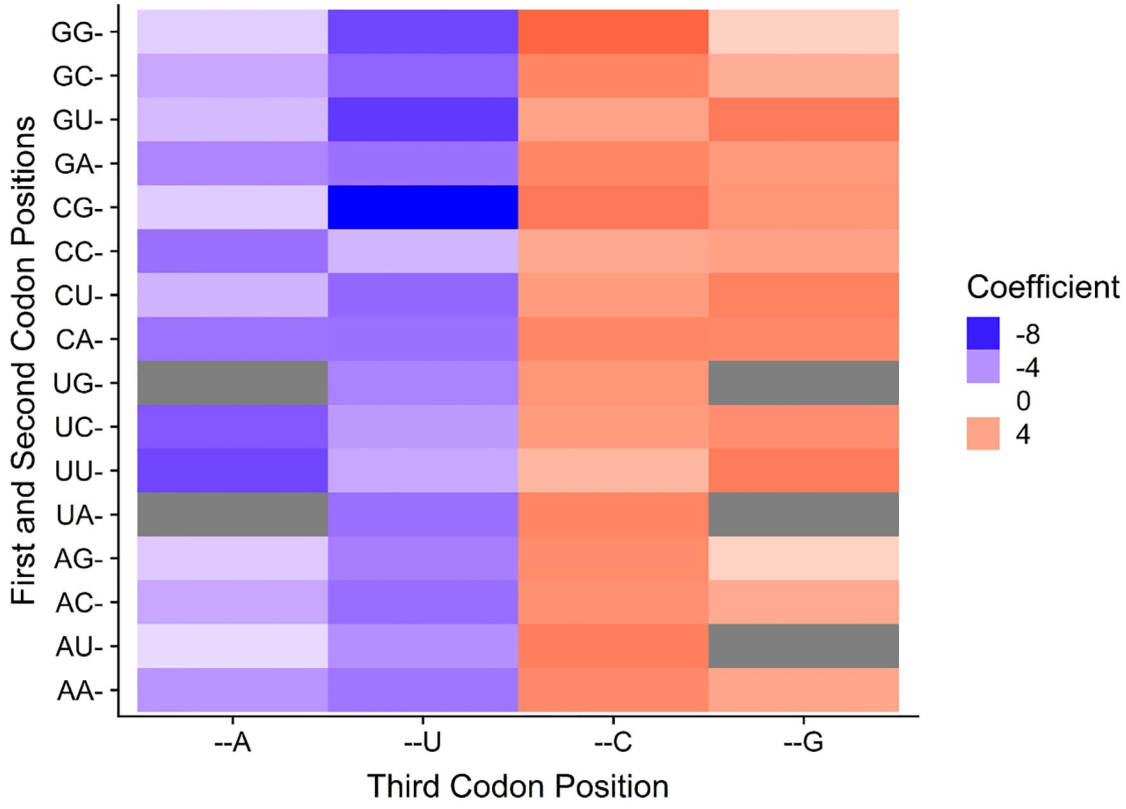

Fig 3. The high correlation between codon usage and GC composition of the third codon position suggests that codon usage bias at the level of individual codons is likely driven by genetic drift. The graph illustrates a phylogenetic generalized least squares comparison between relative synonymous codon usage values and third codon position GC composition (GC3) for each codon across the 327 budding yeast species. Colors toward the red spectrum indicate a positive correlation between CG-ending codons and increasing GC3. Blue colors indicate a negative correlation between A/U-ending codons and increasing GC3. Grey cells denote non-degenerate codons encoding methionine or tryptophan or stop codons.

https://doi.org/10.1371/journal.pgen.1008304.g003

selection. Finally, there was no correlation between genome completeness and S-value (0.14), the fit of genes to the neutral expectation of GC and ENC (0.11), or tRNA count (0.00).

Gene-level codon usage does not fit the neutral expectation

To assess the role of mutational bias across all genes within each genome, we next examined the relationship between the ENC of each gene and its GC3s vis-a-vis the neutral expectation (i.e., the relationship between ENC and GC3s if neutral mutational bias were the only force acting on codon usage). For each genome, we computed the number of genes that fell 10% and 20% of the maximum value outside of the neutral expectation between NC and GC3s [67]. Out of a total of 1,683,203 genes, 381,174 (23%) genes fell outside the 10% threshold and 205,558 (12%) fell outside of the 20% threshold (Fig 5A; S9 Table). We tested the role of gene length in this analysis by comparing the length distribution of genes that deviated either 10% or 20% from the neutral expectation and those that fell within the neutral expectation. In 309 of the 327 species analyzed (~95%), genes that were outside either the 10% or 20% threshold were significantly longer than neutral genes (S9 Table). In 44 species, only those genes that fell outside the 20% threshold were significantly longer than the neutral genes. Interestingly, which species exhibit the pattern of longer non-neutral genes is not associated with major clade, average gene length or the level of translational selection (measured using the S-value; see below).

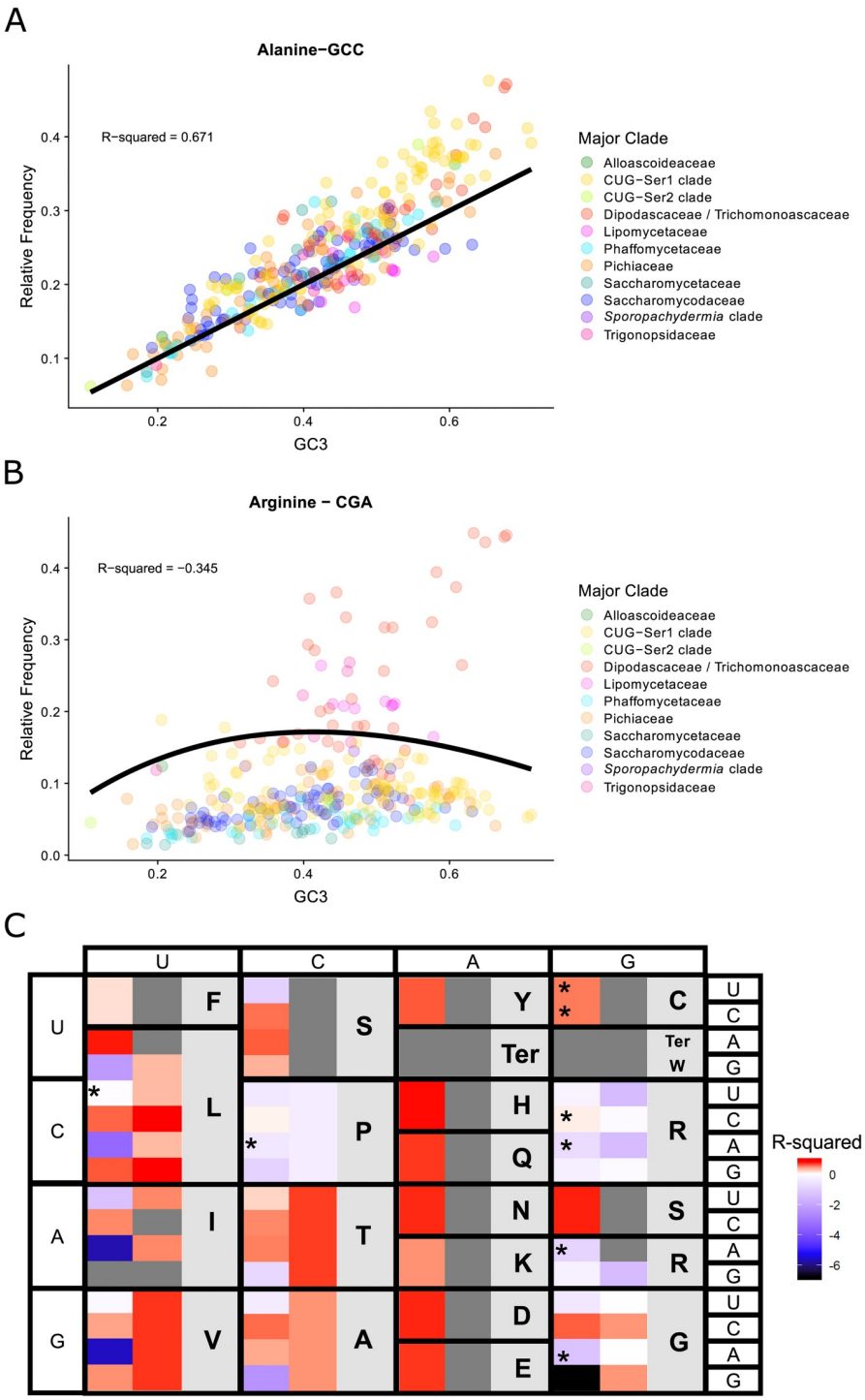

Fig 4. The complex relationship between relative frequency and genome-wide average base composition of the third codon position (GC3) suggests that individual codons vary in their fit to the neutral expectation (i.e., that codon usage is solely driven by GC mutational bias and genetic drift). The neutral expectations for the different codons were obtained from the models developed by Palidwor et al. [38]. A) Observed relative frequency of the alanine codon GCC (shown on the Y axis) plotted against GC3(shown on the X axis) for each of the 327 budding yeast species analyzed in this study. The codon GCC had agood fit to the neutral expectation (black line, R-squared value = 0.671). B) Observed relative frequency of the arginine codon CGU plotted against GC3 composition for each species. The codon CGU had a poor fit to the neutral expectation (black line, R-squared value = -0.165); the same trend was also observed in the other Group-2 arginine codons (CGA and AGG). C) R-squared values for each of the codons (first

column) and the sum of all codons for an amino acid (second column) compared to their neutral expectations. Boxes colored towards the red spectrum indicate a better fit to the neutral model, while boxes colored towards the blue spectrum indicate a poorer fit (i.e., worse than the mean) to the neutral model. Grey-colored boxes in the first column indicate non-degenerate amino acids or stop codons; grey boxes in the second column indicate codons that either have their own models (e.g., ATC) or have values that stem from the same model (e.g., all amino acids encoded by two codons, such as tyrosine (Y), which is encoded by TAT and TAC). Asterisks indicate codons with a Blomberg's K variance over 1 when comparing GC3 and relative frequency, suggesting that the GC3 and relative frequency values for these codons are correlated due to phylogeny (i.e., closely related species tend to have more similar GC3 and relative frequency values due to shared ancestry).

https://doi.org/10.1371/journal.pgen.1008304.g004

We also tested each species' overall fit to the neutral expectation by calculating an R^2 fit to the neutral expectation (Fig 5B & 5C). This analysis revealed that 7 genomes had R^2 values greater than 0.5, suggesting that codon usage in these species can largely be explained by neutral mutational bias. Twelve species had an intermediate R^2 value between 0.25 and 0.5 (or [0.25–0.50]), suggesting that neutral mutational bias is partially responsible for codon usage in most genes in these species. Finally, 72 species had low R^2 values between 0.00 and 0.25, while the remaining 277 species had values below 0. The species with low and negative R^2values deviate from the neutral expectation, suggesting that mutational bias is not the sole driving factor of codon bias within these genomes.

Codon usage in most budding yeast genomes is under translational selection

The previous analysis suggested that most Saccharomycotina species deviate from the strictly neutral expectation between GC3s and NC within their genomes (Fig 5). To test whether translational selection influenced codon usage in budding yeast genomes, we calculated the S-value or the amount of selection on codon usage due to tRNA adaptation. To determine the effect of not accounting for CUG codon reassignment in our analysis, we calculated S-values for genomes with CUG and with all CUG codons removed (S10 Table). The R^2 value when comparing the S-value for the CUG and CUG-removed datasets was 0.99. This suggests that our results are valid despite not accounting for the codon reassignment. S-values could not be produced for the species *Martiniozyma abiesophila*, *Nadsonia fulvescens* var. *fulvescens*, and *Botryozyma nematodophila*, because they did not produce viable wi values from stAI-calc due to software issues (S11 Table). S-values were computed for the remaining 324 species, and significance was assessed using a permutation test (Fig 6A). Thirty-four species from 6 of the 9 clades did not have S-values that were significant at the 0.05 or 0.95 level in the permutation test (S10 Table). These non-significant results ranged in S-value between -0.252 and 0.577, with a median value of 0.273. This result suggests that, in these species, gene-level codon usage could not be distinguished from neutral mutational bias; therefore, it is unlikely that translational selection is broadly acting in these species. In contrast, 27 species exhibit moderate S-values between 0.28 and 0.5 (Fig 6B), on par with levels of translational selection observed in *C. elegans* [S-value of 0.45; 67]. A moderately high S-value between 0.5 and 0.75 was observed in 157 species. Finally, a very high S-value above 0.75 was observed for 107 species, including *S. cerevisiae* (Fig 6C), as previously reported [67]. Overall, 291 / 324 (94%) of genomes examined showed moderate to very high S-values, suggesting that translational selection is widespread across budding yeast genomes.

We also investigated the role of gene length on our measures of translational selection. We found that there was no correlation between our gene level measurement of codon adaptation to the tRNA pool (stAI) and gene length (largest correlation was 0.097 for *Candida tammaniensis*). We did, however, find that when we examined only genes whose protein products are

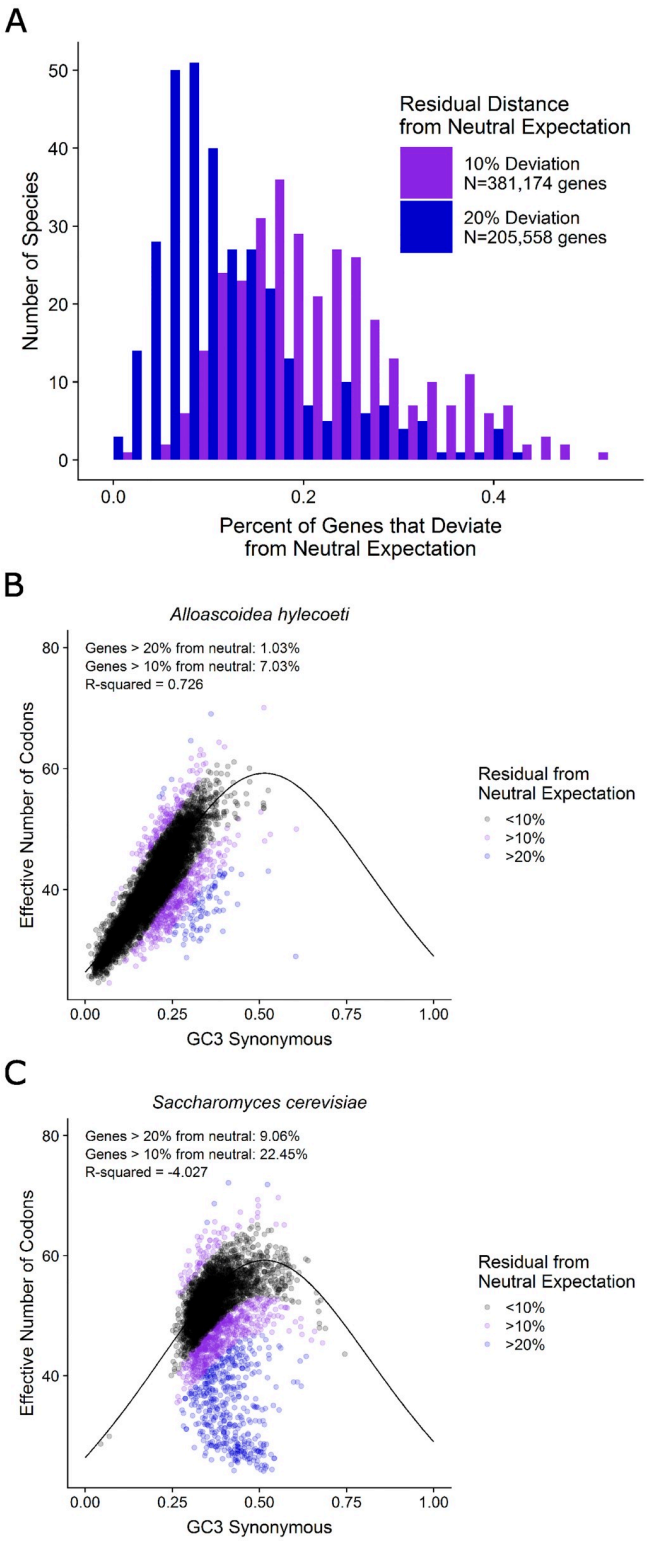

Fig 5. Comparison of the silent third position GC composition of the third codon position (GC3) suggests that individual codons vary in their fit to the neutral expectation (i.e., that codon usage is solely driven by GC mutational bias and genetic drift). The neutral expectations for the different codons were obtained from the models developed by Palidwor et al. (2010). A) Observed relative frequency of the alanine codon GCC (shown on the Y axis) plotted against GC3 (shown on the X axis) for each of the 327 budding yeast species analyzed in this study. The codon

GCC had a good fit to the neutral expectation (black line, R-squared value = 0.671). B) Observed relative frequency of the arginine codon CGU plotted against GC3 composition for each species. The codon CGU had a poor fit to the neutral expectation (black line, R-squared value = -0.165); the same trend was also observed in the other Group-2 arginine codons (CGA and AGG). C) R-squared values for each of the codons (first column) and the sum of all codons for an amino acid (second column) compared to their neutral expectations. Boxes colored towards the red spectrum indicate a better fit to the neutral model, while boxes colored towards the blue spectrum indicate a poorer fit (i.e., worse than the mean) to the neutral model. Grey-colored boxes in the first column indicate non-degenerate amino acids or stop codons; grey boxes in the second column indicate codons that either have their own models (e.g., ATC) or have values that stem from the same model (e.g., all amino acids encoded by two codons, such as tyrosine (Y), which is encoded by TAT and TAC). Asterisks indicate codons with a Blomberg's K variance over 1 when comparing GC3 and relative frequency, suggesting that the GC3 and relative frequency values for these codons are correlated due to phylogeny (i.e., closely related species tend to have more similar GC3 and relative frequency values due to shared ancestry).

https://doi.org/10.1371/journal.pgen.1008304.g005

over 1000 amino acids, the S-value increased by an average of 0.14 in 288 of the 324 genomes analyzed. The largest increase in S-value was observed in the genome of *Eremothecium gossypii*, whose S-value increased from -0.08 to 0.64. Of the 30 species for which the S-value of the longest genes was 0.25 or more greater than the whole genome value, 18 did not have a significant p-value in the permutation test of the genome S-value calculation (i.e., their genome-wide patterns of codon usage bias were consistent with neutrality). This analysis further illustrates that translational selection varies within the genome—even species for which codon usage patterns at the level of the whole genome are consistent with neutrality, translational selection may still act strongly on some of their genes.

Translational selection is weakly associated with tRNAome size

To determine which features are associated with S-values, we examined the relationship between S-values with the combinations of two or more of the following features: genome size, tRNAome size, gene number, number of metabolic traits, and number of isolation environments (S12 Table). The linear model with the highest explanatory power, which accounted for 17.47% of the variation in S-value, includes genome size, tRNAome size, gene number, and total metabolic traits (S13 Table). Among the four features in the model, tRNAome size had the biggest contribution, followed by genome size, gene number, and reported metabolic traits (0.612 versus 0.229, 0.119, and 0.039, respectively). To gain further insight into the contribution of the tRNAome size, we tested a Gaussian model (Fig 7) based on previously reported analyses [67]. The R^2 value of the Gaussian model was higher than that of the linear model (0.11 vs 0.04), although neither model had a very good fit. The Gaussian model suggests that the maximum selection occurs at an intermediate tRNAome size. Interestingly, the estimated maximum for S-value occurs at a tRNAome size of 336 tRNA genes, a value similar to the tRNAome size that corresponds with the maximum modeled S-value from previous models (tRNAome of about 300) [67]. The phylogenetically corrected PGLS analysis revealed no correlation between S-value and either genome size or tRNAome (S2 Fig). Overall, none of the features we tested had strong associations, individually or additively, with S-value, even when phylogenetically corrected.

Discussion

In this study, we surveyed the patterns and forces underlying codon bias across 327 budding yeasts from the subphylum Saccharomycotina. Cluster, correspondence, and correlation analyses of the relative synonymous codon usage across the subphylum is consistent with mutational bias as a significant driver of codon bias—A/U ending codons are generally overrepresented and G/C ending codons are generally underrepresented. This finding is consistent

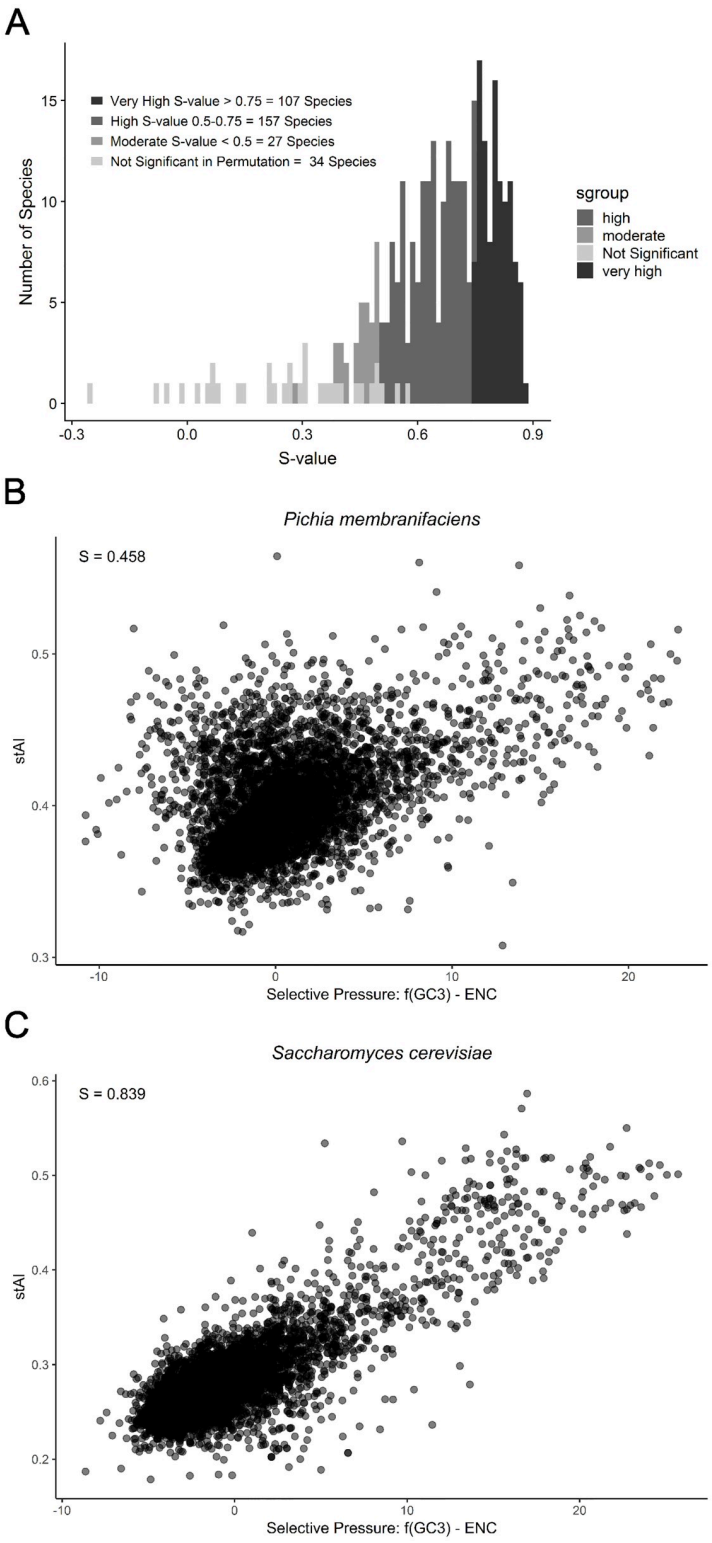

Fig 6. Most genomes in the budding yeast subphylum exhibit moderate to high levels of translational selection on codon bias. Translational selection on codon bias was measured using the S-test, which examines the correlation between the stAI value and the selective pressure (estimated by f(GC3)-ENC where f(GC3) is a modified function of Wright's neutral relationship between the silent GC content of a gene and the effective number of codons) on all coding sequences in a genome. Each point in the comparison between stAI and selective pressure is a single coding

sequence in one genome. Higher S-values indicate higher levels of translational selection on codon bias. A) Distribution of the significant S-values ($p<0.05$ in permutation test; 293 species out of 327) and non-significant S-values ($p>0.05$ in permutation test; 34 / 327 species). B) *Pichia membranifaciens*, an example of a species that exhibits low translational selection on codon bias ($p<0.05$ in permutation test; n = 10,000). C) *Saccharomyces cerevisiae*, an example of a species that exhibits high translational selection on codon bias ($p < 0.01$ in permutation test; n = 10,000).

https://doi.org/10.1371/journal.pgen.1008304.g006

with the low GC content (average silent GC context of 42%) found across the subphylum. Several previous studies have suggested that genome-wide mutational processes are the primary drivers of genome-wide codon usage [36, 37, 44], and we clearly observed the influence of these neutral processes at the genome level. Notably, we also found evidence of selection in both specific codons and genes, which we discuss below.

At the level of individual codon usage, two codons in particular—CGA and CUA—had multiple lines of evidence for violating assumptions of neutral GC-mutational bias and we present biological hypotheses for why these particular codons may be subject to increased selective pressure. For CGA, our results are consistent with previous reports that decoding of the CGA codon in *S. cerevisiae* is inhibitory to translation due to codon-anticodon interactions [74, 75]. This effect, however, may not be universal across the Saccharomycotina: CGA was underrepresented (RSCU < 1) in 222 species but overrepresented (RSCU > 1) in 105 species. RSCU of CGA also varies between major clades of the Saccharomycotina with the Dipodascaceae/Trichomonascaceae clade having the highest average RSCU (1.47) and the Phaffomycetaceae clade having the lowest average RSCU (0.66). Given that Dipodascaceae/Trichomonascaceae clade is distantly related to Saccharomycetaceae, the major clade that *S. cerevisiae* belongs to, it is likely that the two independent defects in translation that result in the inhibitory nature of CGA in *S. cerevisiae* [75] evolved within Saccharomycetaceae, after the divergence of the two clades. The codon CGA is not the only arginine encoding codon to violate the neutral assumptions (Fig 4C). Deviations in the remaining arginine codons may be a

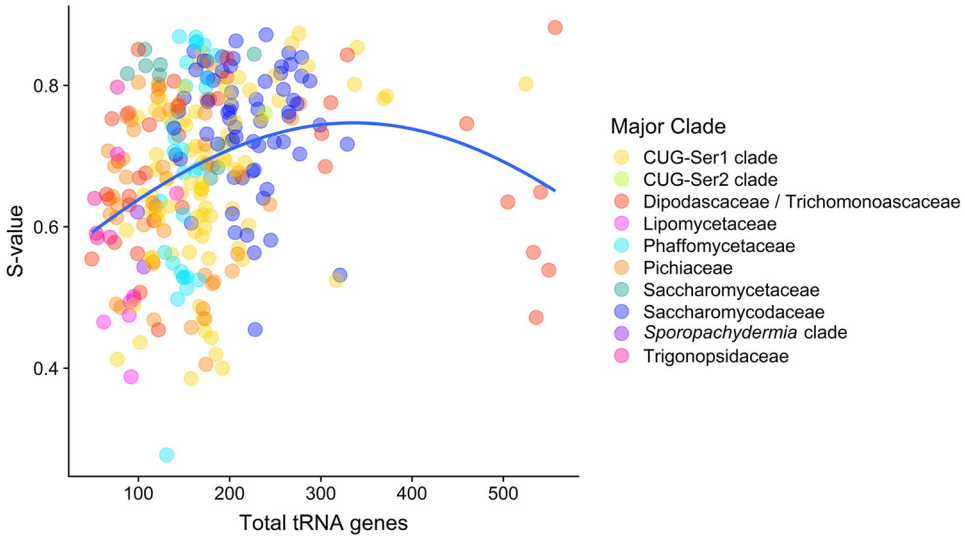

Fig 7. Maximum translational selection occurs at an intermediate number of total tRNA genes in the genome. This plot shows the relationship between the total number of tRNA genes in a genome (tRNAome size) and S-value for each the 327 budding yeast species analyzed in this study. The best fitting model (blue) was a Gaussian distribution with a maximum S-value at 336 tRNA genes. This suggests that species with either low or high numbers of total tRNA genes exhibit lower levels of translational selection.

https://doi.org/10.1371/journal.pgen.1008304.g007

result of strong directional selection due to the large number of degenerate codons encoding arginine, which may result in more opportunities for poor codon-tRNA pairing [76, 77].

For CUA, departure from assumptions of neutral GC-mutational bias are likely driven by the reassignment of CUG in the CUG-Ser1 and CUG-Ser2 clades, which had profound effects on the remaining leucine codons since the majority of CUG codons that remained leucine were reassigned to UUG or UUA [52, 78]. This conclusion is supported by the observation that the CUA codon is underrepresented in the CUG-Ser1 and CUG-Ser2 clades (Fig 1; S14 Table) compared to other major clades in the subphylum (Fig 1: S14 Table). Underrepresentation of CUA is not exclusive to the CUG-Ser2 and CUG-Ser1 clades—the Dipodascaceae/Trichomonascaceae major clade had an average RSCU of 0.60 and includes 12 species (of 37) with a very low RSCU less than 0.5. This may suggest that the Dipodascaceae/Trichomonascaceae major clade experienced similar evolutionary pressures to those that may have contributed to codon reassignment, such as the hypothesized presence of a Virus-Like Element with killer activity in the CUG-Ser1 and CUG-Ser2 clades [55]. The most studied member of the Dipodascaceae/Trichomonascaceae major clade, *Yarrowia lipolytica*, possesses virus-like particles, but these particles do not appear to be associated with a killer phenotype [79, 80]. This finding highlights the strong impact of codon reassignment on codon usage.

We also observed deviations from the neutral expectation in all codons that encode proline that may be associated with the chemical structure of the proline peptide-bond. Biases in proline codon usage may be related to proline-induced stalling in translation [81]. This stalling was observed in *S. cerevisiae* riboprofiling data [81] and may be related to the slow incorporation of proline into the growing amino acid chain due to its imino side-chain [82, 83]. Additionally, in *S. cerevisiae*, codons for proline and glycine (which also deviate from the neutral expectation) are involved in frameshift suppression via suppressor tRNAs that contain four-base anticodon sequences that allow for frameshift read-through [84, 85]. As a whole, the results of the codon-specific analysis suggest that while many codons are highly correlated with mutational bias, specific codons may be under a variety of selective forces—especially translational selection—that alter codon usage.

Almost a quarter of the 1,683,203 genes found in the 327 budding yeast genomes deviate from the neutral expectation by at least 10%. These results are consistent with the observation that codon bias varies between transcripts within a species [37, 86] and is associated with increased expression. In fact, for the species *Saccharomyces mikatae*, the degree to which a transcript differs from the neutral expectation (greater residual) is moderately associated with greater expression at steady state [87]. For the majority of the species examined (320), mutational bias is not the only force influencing codon bias among transcripts.

We also determined that gene length is likely associated with levels of translational selection for many of the species we investigated. This is not surprising given previous work suggesting that gene length and translational selection are not independent [16, 76, 88, 89]. For example, in *S. cerevisiae* and *Escherichia coli*, increased selective pressure on longer genes may be required to reduce missense errors during the translation of energetically expensive large products [16, 88, 89]. In contrast, the opposite pattern has been observed in *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana*, where shorter genes exhibit higher levels of optimal codons [76]. While our results are generally consistent with an increased deviation from neutral expectation for longer genes, this is not the case for all budding yeast genomes—for 62 of the 327 species, the genes that deviate from neutrality by 10% were not longer than neutral genes. Interestingly, we could not associate this pattern with average gene length, total number of genes or whole genome S-value. Furthermore, 36 of the 324 measurements of translational selection did not increase by including only genes over 1,000 amino

acids. Overall, we find support for increased translational selection in longer genes but caution that this is not a universal feature of the subphylum.

Assessing how translational selection may influence codon usage bias within species, we found that the majority of species exhibited moderate or high contribution of selection to the variation in codon bias (Fig 6A). Previous work suggested a model in which the highest amount of selection on synonymous codon usage occurs at intermediate genome size. At the lower end of genome size, low selection is hypothesized to be due to the correlation between small genomes and small tRNAomes with low tRNA gene redundancy. In turn, low tRNA gene redundancy restricts the ability of selection to act on codon bias [67, 90]. At the larger end of genome size, low selection is hypothesized to be due to drift in species with small effective population sizes: this drift would increase the genome size and decrease the ability of selection to shape codon usage [12]. Within Saccharomycotina, the role of tRNAome size is consistent with these predictions, except for genome size. This exception is likely due to a low correlation between genome size and tRNAome size in this group. While tRNAome size and genome size are positively correlated when analyzed using a phylogenetically independent contrast (PIC) [91], this correlation is not very strong (adjusted R^2 of 0.1629). It is likely that other biological and ecological features play a significant role in the amount of translational selection occurring within these genomes. For example, generalist and specialist parasitic fungi have been shown to have significantly different amounts of translational selection occurring on codon usage [92].

In summary, we find that the balance between neutral and selective forces on codon usage varies between genomes, between codons, and between genes within a genome. Some Saccharomycotina species exhibit nearly neutral codon usage in line with those observed in humans or bacteria, such as *Helicobacter pylori*, while other budding yeast species show extremely high adaptation to the tRNA pool through translational selection [67]. This range in the magnitude of forces acting on codon usage in the Saccharomycotina and the low explanatory power of the factors examined suggest that it is difficult to predict *a priori* selection on codon bias based on lineage, cellularity, genome size, tRNAome, or GC composition.

There is moderate to strong evidence for translational selection in most budding yeast genomes examined. This trend may be due to the rapid growth that characterizes most budding yeasts: growth efficiency has been linked to translational selection in codon usage [93, 94]. One interesting implication of this abundance of translational selection is that codon optimization may be a useful proxy for highly expressed genes. It has long been known that ribosomal genes are among both the most highly expressed and highly codon usage-optimized genes across species [49, 95], leading to their use as the basis for the codon adaptation index [35, 96]. In our dataset, there are 11,047 genes (average of 35 per species) that are as highly or more highly optimized than the ribosomal genes, suggesting there is a wealth of information about which genes may be highly expressed or differentially highly expressed across this lineage.

Supporting information

S1 Fig. Percent contribution of individual codons to the correspondence analysis of RSCU. A) The contributions of codons to the first dimension (66.891% of the overall variation) was distributed among multiple codons. Each of the codons made a relatively small contribution to the variation but collectively accounted for most of the differences in RSCU observed between species. B) The contributions of codons to the second dimension (7.093% of the overall variation) was dominated by four codons that contribute more than 10% each to the variation. (TIF)

S2 Fig. Low association between selective pressure on a genome (S-value) and the genomic features of number of tRNA genes and genome size after phylogenetic correction with phylogenetic generalized least squares. A) The association between number of tRNA genes and S-value (slope = 0.00012) after correction for phylogenetic relatedness was nearly flat, suggesting that total tRNA genes do not linearly reflect the selective pressure on codon bias within a genome. B) The association between genome size (in base pairs) was nearly flat (slope ~ 0), suggesting that, after phylogenetic correction, there is no relationship between genome size and the selective pressure on codon bias within a genome.
(TIF)

S3 Fig. Adaptation of codon usage to the tRNA pool in genes of *Saccharomyces mikatae* is correlated with expression at steady state. For each gene in the *Saccharomyces mikatae* genome we measure codon adaptation to the tRNA pool (tAI). This is positively correlated with expression at steady-state.
(TIF)

S1 Table. Genome and annotation data for the 332 Saccharomycotina species considered. This table includes the relevant source information for each genome. Additionally, the number of contigs and basepairs removed in each filtering step is reported—this includes filtering for mitochondrial sequences, short sequences, sequences without a start codon, and sequences with ambiguous codons removed.
(XLSX)

S2 Table. Reference information for the Saccharomycotina mitochondrial genomes that were used as a reference for genomic filtering.
(XLSX)

S3 Table. Reference information for the Saccharomycotina mitochondrial coding sequences that were used as a reference for annotation filtering.
(XLSX)

S4 Table. tRNA annotation for the 332 Saccharomycotina species considered in this analysis. The total number of each tRNA is listed for each species as well as the total number of tRNA genes annotated. The reassigned CUG codons are also listed.
(XLSX)

S5 Table. Testing of phylogenetic concordance of the RSCU for each codon across all 327 species. The Blomberg's K, Pagel's Lambda, and corresponding P-value are reported for each codon.
(XLSX)

S6 Table. The Pearson's correlation and P-value between each codon and the GC content of the third codon position across all 327 Saccharomycotina species.
(XLSX)

S7 Table. The PGLS correlation between each codon and the GC content of the third codon position across all 327 Saccharomycotina species.
(XLSX)

S8 Table. The fit (in r-squared) of each codon to the neutral frequency proposed by Palidwor et al. 2010. Additionally, the Blomberg's K of the individual species' residuals used to compute the R2 value is reported. The individual neutral formulas from Palidwor et al. 2010 are also listed.
(XLSX)

S9 Table. The individual genome results for the comparison between ENC and GC3s including the percent of genes that deviate from neutral by 10 or 20 percent and the R-squared of the fit for each genome. Additionally the results of the gene length analysis are reported including the comparison of length between deviant and neutral genes, the average sequence length and the total number of sequences analyzed.
(XLSX)

S10 Table. The S-values calculated for each genome to assess translational selection on codon usage within the genomes. This also includes the analysis with the CUG codon removed, the permutation test, the correlation between stAI and gene length, and the S-value of sequences longer than 1,000 amino acids.
(XLSX)

S11 Table. The wi values (adaptation) values for each codon in each of the genomes.
(XLSX)

S12 Table. Genomic and biological features against which we compared the S-value for each genome. This includes genomic GC average, total tRNA genes, total genome size, the fit of each genome's sequences to the neutral expectation, total characterized metabolic traits, and total environments from which the strain was isolated.
(XLSX)

S13 Table. The models used to test the correlation between S-value and other genomic and biological features. The fit (R-squared), relative importance, ANNOVA, log likelihood, and correlation are reported where appropriate.
(XLSX)

S14 Table. The relative synonymous codon usage (RSCU) of each codon for each genome considered.
(XLSX)

Acknowledgments

We thank the members of the Rokas and Hittinger labs, in particular Xing-Xing Shen, for their feedback and discussions on this project. We would also like to thank the other members of the Y1000+ project (http://www.y1000plus.org/) including, Jacek Kominek and Xiaofan Zhou, for their feedback. We would also like to thank Renana Sabi, Renana Volvovitch Daniel and Tamir Tuller, the creators of stAIcalc, for their assistance in troubleshooting the codon adaptation analysis.

Author Contributions

Conceptualization: Abigail L. LaBella, Chris Todd Hittinger, Antonis Rokas.

Data curation: Abigail L. LaBella, Dana A. Opulente.

Formal analysis: Abigail L. LaBella, Jacob L. Steenwyk.

Funding acquisition: Chris Todd Hittinger, Antonis Rokas.

Investigation: Abigail L. LaBella, Dana A. Opulente, Jacob L. Steenwyk.

Methodology: Abigail L. LaBella, Dana A. Opulente, Jacob L. Steenwyk.

Project administration: Abigail L. LaBella, Chris Todd Hittinger, Antonis Rokas.

Resources: Abigail L. LaBella, Dana A. Opulente, Jacob L. Steenwyk, Chris Todd Hittinger, Antonis Rokas.

Software: Abigail L. LaBella.

Supervision: Chris Todd Hittinger, Antonis Rokas.

Validation: Abigail L. LaBella.

Visualization: Abigail L. LaBella.

Writing – original draft: Abigail L. LaBella, Antonis Rokas.

Writing – review & editing: Abigail L. LaBella, Dana A. Opulente, Jacob L. Steenwyk, Chris Todd Hittinger, Antonis Rokas.

References

1. Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature. 1976; 260(5551):500–7. Epub 1976/04/08. https://doi.org/10.1038/260500a0 PMID: 1264203.
2. Air GM, Blackburn EH, Coulson AR, Galibert F, Sanger F, Sedat JW, et al. Gene F of bacteriophage phiX174. Correlation of nucleotide sequences from the DNA and amino acid sequences from the gene product. J Mol Biol. 1976; 107(4):445–58. Epub 1976/11/15. https://doi.org/10.1016/s0022-2836(76)80077-0 PMID: 1088826.
3. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 1981; 9(1):r43–74. Epub 1981/01/10. https://doi.org/10.1093/nar/9.1.213-b PMID: 7208352.
4. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol. 1981; 146(1):1–21. Epub 1981/02/15. https://doi.org/10.1016/0022-2836(81)90363-6 PMID: 6167728.
5. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol. 1981; 151(3):389–409. https://doi.org/10.1016/0022-2836(81)90003-6 PMID: 6175758.
6. Post LE, Strycharz GD, Nomura M, Lewis H, Dennis PP. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. Proc Natl Acad Sci U S A. 1979; 76(4):1697–701. Epub 1979/04/01. https://doi.org/10.1073/pnas.76.4.1697 PMID: 377281.
7. Nakamura K, Pirtle RM, Pirtle IL, Takeishi K, Inouye M. Messenger ribonucleic acid of the lipoprotein of the Escherichia coli outer membrane. II. The complete nucleotide sequence. J Biol Chem. 1980; 255(1):210–6. Epub 1980/01/10. PMID: 6765942.
8. Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 1982; 10(22):7055–74. Epub 1982/11/25. https://doi.org/10.1093/nar/10.22.7055 PMID: 6760125.
9. Sharp PM, Li WH. Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons. Nucleic Acids Res. 1986; 14(19):7737–49. Epub 1986/10/10. https://doi.org/10.1093/nar/14.19.7737 PMID: 3534792.
10. Thomas LK, Dix DB, Thompson RC. Codon choice and gene expression: synonymous codons differ in their ability to direct aminoacylated-transfer RNA binding to ribosomes in vitro. Proc Natl Acad Sci U S A. 1988; 85(12):4242–6. Epub 1988/06/01. https://doi.org/10.1073/pnas.85.12.4242 PMID: 3288988.
11. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, et al. Codon optimality is a major determinant of mRNA stability. Cell. 2015; 160(6):1111–24. https://doi.org/10.1016/j.cell.2015.02.029 PMID: 25768907.
12. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. Genetics. 1991; 129(3):897–907. Epub 1991/11/01. PMID: 1752426.
13. Chevance FF, Le Guyon S, Hughes KT. The effects of codon context on in vivo translation speed. PLoS Genet. 2014; 10(6):e1004392. Epub 2014/06/06. https://doi.org/10.1371/journal.pgen.1004392 PMID: 24901308.
14. Xia X. How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? Genetics. 1998; 149(1):37–44. Epub 1998/05/28. PMID: 9584084.

15. Zhou T, Weems M, Wilke CO. Translationally optimal codons associate with structurally sensitive sites in proteins. Mol Biol Evol. 2009; 26(7):1571–80. https://doi.org/10.1093/molbev/msp070 PMID: 19349643.

16. Stoletzki N, Eyre-Walker A. Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol Biol Evol. 2007; 24(2):374–81. Epub 2006/11/15. https://doi.org/10.1093/molbev/msl166 PMID: 17101719.

17. Zhou Z, Dang Y, Zhou M, Yuan H, Liu Y. Codon usage biases co-evolve with transcription termination machinery to suppress premature cleavage and polyadenylation. Elife. 2018; 7. Epub 2018/03/17. https://doi.org/10.7554/eLife.33569 PMID: 29547124.

18. Radhakrishnan A, Chen YH, Martin S, Alhusaini N, Green R, Coller J. The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. Cell. 2016; 167(1):122–32 e9. https://doi.org/10.1016/j.cell.2016.08.053 PMID: 27641505.

19. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A. 2010; 107(8):3645–50. https://doi.org/10.1073/pnas.0909910107 PMID: 20133581.

20. Zhou M, Guo J, Cha J, Chae M, Chen S, Barral JM, et al. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature. 2013; 495(7439):111–5. https://doi.org/10.1038/nature11833 PMID: 23417067.

21. Buhr F, Jha S, Thommen M, Mittelstaet J, Kutz F, Schwalbe H, et al. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. Mol Cell. 2016; 61(3):341–51. Epub 2016/02/06. https://doi.org/10.1016/j.molcel.2016.01.008 PMID: 26849192.

22. Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, et al. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. Mol Cell. 2015; 59(5):744–54. Epub 2015/09/01. https://doi.org/10.1016/j.molcel.2015.07.018 PMID: 26321254.

23. Pechmann S, Chartron JW, Frydman J. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. Nat Struct Mol Biol. 2014; 21(12):1100–5. https://doi.org/10.1038/nsmb.2919 PMID: 25420103.

24. Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Mol Biol Evol. 2013; 30(3):549–60. Epub 2012/12/12. https://doi.org/10.1093/molbev/mss273 PMID: 23223712.

25. Mittal P, Brindle J, Stephen J, Plotkin JB, Kudla G. Codon usage influences fitness through RNA toxicity. Proc Natl Acad Sci U S A. 2018; 115(34):8639–44. Epub 2018/08/08. https://doi.org/10.1073/pnas.1810022115 PMID: 30082392.

26. Fragata I, Matuszewski S, Schmitz MA, Bataillon T, Jensen JD, Bank C. The fitness landscape of the codon space across environments. Heredity (Edinb). 2018; 121(5):422–37. Epub 2018/08/22. https://doi.org/10.1038/s41437-018-0125-7 PMID: 30127529.

27. Ballard A, Bieniek S, Carlini DB. The fitness consequences of synonymous mutations in Escherichia coli: Experimental evidence for a pleiotropic effect of translational selection. Gene. 2019; 694:111–20. Epub 2019/02/11. https://doi.org/10.1016/j.gene.2019.01.031 PMID: 30738968.

28. Krisko A, Copic T, Gabaldon T, Lehner B, Supek F. Inferring gene function from evolutionary change in signatures of translation efficiency. Genome Biol. 2014; 15(3):R44. Epub 2014/03/04. https://doi.org/10.1186/gb-2014-15-3-r44 PMID: 24580753.

29. She R, Jarosz DF. Mapping Causal Variants with Single-Nucleotide Resolution Reveals Biochemical Drivers of Phenotypic Change. Cell. 2018; 172(3):478–90 e15. Epub 2018/01/27. https://doi.org/10.1016/j.cell.2017.12.015 PMID: 29373829.

30. Kliman RM, Irving N, Santiago M. Selection conflicts, gene expression, and codon usage trends in yeast. J Mol Evol. 2003; 57(1):98–109. Epub 2003/09/10. https://doi.org/10.1007/s00239-003-2459-9 PMID: 12962310.

31. Carlini DB, Stephan W. In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. Genetics. 2003; 163(1):239–43. Epub 2003/02/15. PMID: 12586711.

32. Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell. 2014; 156(6):1324–35. Epub 2014/03/19. https://doi.org/10.1016/j.cell.2014.01.051 PMID: 24630730.

33. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet. 2011; 12(10):683–91. Epub 2011/09/01. https://doi.org/10.1038/nrg3051 PMID: 21878961.

34. Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 2006; 7(2):98–108. Epub 2006/01/19. https://doi.org/10.1038/nrg1770 PMID: 16418745.

35. Sharp PM, Li WH. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987; 15(3):1281–95. https://doi.org/10.1093/nar/15.3.1281 PMID: 3547335.

36. 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 Biol. 2001; 2(4):RESEARCH0010. Epub 2001/04/18. https://doi.org/10.1186/gb-2001-2-4-research0010 PMID: 11305938.

37. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci U S A. 2004; 101(10):3480–5. Epub 2004/03/03. https://doi.org/10.1073/pnas.0307827100 PMID: 14990797.

38. Palidwor GA, Perkins TJ, Xia X. A general model of codon bias due to GC mutational bias. PLoS One. 2010; 5(10):e13431. Epub 2010/11/05. https://doi.org/10.1371/journal.pone.0013431 PMID: 21048949.

39. Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glemin S, et al. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. Mol Biol Evol. 2018; 35(5):1092–103. Epub 2018/02/02. https://doi.org/10.1093/molbev/msy015 PMID: 29390090.

40. Sharp PM, Stenico M, Peden JF, Lloyd AT. Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans. 1993; 21(4):835–41. Epub 1993/11/01. https://doi.org/10.1042/bst0210835 PMID: 8132077.

41. Galtier N, Piganeau G, Mouchiroud D, Duret L. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics. 2001; 159(2):907–11. Epub 2001/11/06. PMID: 11693127.

42. Harrison RJ, Charlesworth B. Biased gene conversion affects patterns of codon usage and amino acid usage in the Saccharomyces sensu stricto group of yeasts. Mol Biol Evol. 2011; 28(1):117–29. Epub 2010/07/27. https://doi.org/10.1093/molbev/msq191 PMID: 20656793.

43. Clement Y, Sarah G, Holtz Y, Homa F, Pointet S, Contreras S, et al. Evolutionary forces affecting synonymous variations in plant genomes. PLoS Genet. 2017; 13(5):e1006799. Epub 2017/05/23. https://doi.org/10.1371/journal.pgen.1006799 PMID: 28531201.

44. Wan XF, Xu D, Kleinhofs A, Zhou J. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evol Biol. 2004; 4:19. Epub 2004/06/30. https://doi.org/10.1186/1471-2148-4-19 PMID: 15222899.

45. Sun Y, Tamarit D, Andersson SGE. Switches in Genomic GC Content Drive Shifts of Optimal Codons under Sustained Selection on Synonymous Sites. Genome Biol Evol. 2017; 9(10):2560–79. Epub 2016/08/20. https://doi.org/10.1093/gbe/evw201 PMID: 27540085.

46. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 1985; 2(1):13–34. Epub 1985/01/01. https://doi.org/10.1093/oxfordjournals.molbev.a040335 PMID: 3916708.

47. Shields DC, Sharp PM. Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases. Nucleic Acids Res. 1987; 15(19):8023–40. https://doi.org/10.1093/nar/15.19.8023 PMID: 3118331.

48. Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet. 2008; 42:287–99. Epub 2008/11/06. https://doi.org/10.1146/annurev.genet.42.110807.091442 PMID: 18983258.

49. Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF. DNA sequence evolution: the sounds of silence. Philos Trans R Soc Lond B Biol Sci. 1995; 349(1329):241–7. Epub 1995/09/29. https://doi.org/10.1098/rstb.1995.0108 PMID: 8577834.

50. 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; 175(6):1533–45 e20. Epub 2018/11/13. https://doi.org/10.1016/j.cell.2018.10.023 PMID: 30415838.

51. Kawaguchi Y, Honda H, Taniguchi-Morimura J, Iwasaki S. The codon CUG is read as serine in an asporogenic yeast Candida cylindracea. Nature. 1989; 341(6238):164–6. Epub 1989/09/14. https://doi.org/10.1038/341164a0 PMID: 2506450.

52. Miranda I, Silva R, Santos MA. Evolution of the genetic code in yeasts. Yeast. 2006; 23(3):203–13. Epub 2006/02/25. https://doi.org/10.1002/yea.1350 PMID: 16498697.

53. Muhlhausen S, Findeisen P, Plessmann U, Urlaub H, Kollmar M. A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. Genome Res. 2016; 26(7):945–55. https://doi.org/10.1101/gr.200931.115 PMID: 27197221.

54. Riley R, Haridas S, Wolfe KH, Lopes MR, Hittinger CT, Goker M, et al. Comparative genomics of biotechnologically important yeasts. Proc Natl Acad Sci U S A. 2016; 113(35):9882–7. https://doi.org/10.1073/pnas.1603941113 PMID: 27535936.

55. Krassowski T, Coughlan AY, Shen XX, Zhou X, Kominek J, Opulente DA, et al. Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. Nat Commun. 2018; 9(1):1887. Epub 2018/05/16. https://doi.org/10.1038/s41467-018-04374-7 PMID: 29760453.

56. Opulente DA, Rollinson EJ, Bernick-Roehr C, Hulfachor AB, Rokas A, Kurtzman CP, et al. Factors driving metabolic diversity in the budding yeast subphylum. BMC Biol. 2018; 16(1):26. Epub 2018/03/04. https://doi.org/10.1186/s12915-018-0498-3 PMID: 29499717.

57. Kurtzman C, Fell JW, Boekhout T. The yeasts: a taxonomic study: Elsevier; 2011.

58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10. Epub 1990/10/05. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712.

59. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. Epub 2009/12/17. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500.

60. Sabi R, Tuller T. Modelling the efficiency of codon-tRNA interactions based on codon usage bias. DNA Res. 2014; 21(5):511–26. https://doi.org/10.1093/dnares/dsu017 PMID: 24906480.

61. Xia X. DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. Mol Biol Evol. 2018; 35(6):1550–2. Epub 2018/04/19. https://doi.org/10.1093/molbev/msy073 PMID: 29669107.

62. Grantham R, Gautier C, Gouy M. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res. 1980; 8(9):1893–912. Epub 1980/05/10. https://doi.org/10.1093/nar/8.9.1893 PMID: 6159596.

63. Suzuki H, Brown CJ, Forney LJ, Top EM. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. DNA Res. 2008; 15(6):357–65. Epub 2008/10/23. https://doi.org/10.1093/dnares/dsn028 PMID: 18940873.

64. Pagel M. Inferring the historical patterns of biological evolution. Nature. 1999; 401(6756):877–84. Epub 1999/11/30. https://doi.org/10.1038/44766 PMID: 10553904.

65. Blomberg SP, Garland T Jr., Ives AR. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution. 2003; 57(4):717–45. Epub 2003/06/05. PMID: 12778543.

66. Wright F. The 'effective number of codons' used in a gene. Gene. 1990; 87(1):23–9. https://doi.org/10.1016/0378-1119(90)90491-9 PMID: 2110097.

67. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004; 32(17):5036–44. Epub 2004/09/28. https://doi.org/10.1093/nar/gkh834 PMID: 15448185.

68. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics. 1947:50–60.

69. Wilcoxon F. Individual comparisons of grouped data by ranking methods. J Econ Entomol. 1946; 39:269. Epub 1946/04/01. https://doi.org/10.1093/jee/39.2.269 PMID: 20983181.

70. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 2016; 44(W1):W54–7. Epub 2016/05/14. https://doi.org/10.1093/nar/gkw413 PMID: 27174935.

71. Muhlhausen S, Schmitt HD, Pan KT, Plessmann U, Urlaub H, Hurst LD, et al. Endogenous Stochastic Decoding of the CUG Codon by Competing Ser- and Leu-tRNAs in Ascoidea asiatica. Curr Biol. 2018; 28(13):2046–57 e5. Epub 2018/06/19. https://doi.org/10.1016/j.cub.2018.04.085 PMID: 29910077.

72. Ives AR, Midford PE, Garland T. Within-species variation and measurement error in phylogenetic comparative methods. Systematic Biol. 2007; 56(2):252–70. https://doi.org/10.1080/10635150701313830 PMID: 17464881

73. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012; 3(2):217–23. https://doi.org/10.1111/j.2041-210X.2011.00169.x

74. Letzring DP, Dean KM, Grayhack EJ. Control of translation efficiency in yeast by codon-anticodon interactions. RNA. 2010; 16(12):2516–28. Epub 2010/10/26. https://doi.org/10.1261/rna.2411710 PMID: 20971810.

75. Letzring DP, Wolf AS, Brule CE, Grayhack EJ. Translation of CGA codon repeats in yeast involves quality control components and ribosomal protein L1. RNA. 2013; 19(9):1208–17. Epub 2013/07/05. https://doi.org/10.1261/rna.039446.113 PMID: 23825054.

76. Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci U S A. 1999; 96(8):4482–7. Epub 1999/04/14. https://doi.org/10.1073/pnas.96.8.4482 PMID: 10200288.

77. McVean GA, Vieira J. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics. 2001; 157(1):245–57. Epub 2001/01/05. PMID: 11139506.

78. Massey SE, Moura G, Beltrao P, Almeida R, Garey JR, Tuite MF, et al. Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in Candida spp. Genome Res. 2003; 13(4):544–57. Epub 2003/04/03. https://doi.org/10.1101/gr.811003 PMID: 12670996.

79. Tréton BY, Le Dall M-T, Heslot H. Virus-like particles from the yeast Yarrowia lipolytica. Current genetics. 1985; 9(4):279–84.

80. el-Sherbeini M, Bostian KA, Levitre J, Mitchell DJ. Gene-protein assignments within the yeast Yarrowia lipolytica dsRNA viral genome. Curr Genet. 1987; 11(6–7):483–90. Epub 1987/01/01. PMID: 3502458.

81. Artieri CG, Fraser HB. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. Genome Res. 2014; 24(12):2011–21. Epub 2014/10/09. https://doi.org/10.1101/gr.175893.114 PMID: 25294246.

82. Pavlov MY, Watts RE, Tan Z, Cornish VW, Ehrenberg M, Forster AC. Slow peptide bond formation by proline and other N-alkylamino acids in translation. Proc Natl Acad Sci U S A. 2009; 106(1):50–4. Epub 2008/12/24. https://doi.org/10.1073/pnas.0809211106 PMID: 19104062.

83. Doerfel LK, Wohlgemuth I, Kothe C, Peske F, Urlaub H, Rodnina MV. EF-P is essential for rapid synthesis of proteins containing consecutive proline residues. Science. 2013; 339(6115):85–8. Epub 2012/12/15. https://doi.org/10.1126/science.1229017 PMID: 23239624.

84. Donahue TF, Farabaugh PJ, Fink GR. Suppressible four-base glycine and proline codons in yeast. Science. 1981; 212(4493):455–7. Epub 1981/04/24. https://doi.org/10.1126/science.7010605 PMID: 7010605.

85. Gaber RF, Culbertson MR. The yeast frameshift suppressor gene SUF16-1 encodes an altered glycine tRNA containing the four-base anticodon 3'-CCCG-5'. Gene. 1982; 19(2):163–72. Epub 1982/09/01. https://doi.org/10.1016/0378-1119(82)90002-6 PMID: 6293925.

86. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. Nucleic Acids Res. 1988; 16(17):8207–11. Epub 1988/09/12. https://doi.org/10.1093/nar/16.17.8207 PMID: 3138659.

87. Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. The role of nucleosome positioning in the evolution of gene regulation. PLoS Biol. 2010; 8(7):e1000414. Epub 2010/07/14. https://doi.org/10.1371/journal.pbio.1000414 PMID: 20625544.

88. Coghlan A, Wolfe KH. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. Yeast. 2000; 16(12):1131–45. Epub 2000/08/23. https://doi.org/10.1002/1097-0061(20000915)16:12<1131::AID-YEA609>3.0.CO;2-F PMID: 10953085.

89. Eyre-Walker A. Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? Mol Biol Evol. 1996; 13(6):864–72. Epub 1996/07/01. https://doi.org/10.1093/oxfordjournals.molbev.a025646 PMID: 8754221.

90. Kanaya S, Yamada Y, Kudo Y, Ikemura T. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene. 1999; 238(1):143–55. Epub 1999/11/26. https://doi.org/10.1016/s0378-1119(99)00225-5 PMID: 10570992.

91. Felsenstein J. Phylogenies and the comparative method. The American Naturalist. 1985; 125(1):1–15.

92. Badet T, Peyraud R, Mbengue M, Navaud O, Derbyshire M, Oliver RP, et al. Codon optimization underpins generalist parasitism in fungi. Elife. 2017; 6. Epub 2017/02/06. https://doi.org/10.7554/eLife.22472 PMID: 28157073.

93. Kurland CG. Codon bias and gene expression. FEBS Lett. 1991; 285(2):165–9. Epub 1991/07/22. https://doi.org/10.1016/0014-5793(91)80797-7 PMID: 1855585.

94. Andersson GE, Kurland CG. An extreme codon preference strategy: codon reassignment. Mol Biol Evol. 1991; 8(4):530–44. Epub 1991/07/01. https://doi.org/10.1093/oxfordjournals.molbev.a040666 PMID: 1921708.

95. Shields DC, Sharp PM, Higgins DG, Wright F. "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol. 1988; 5(6):704–16. Epub 1988/11/01. https://doi.org/10.1093/oxfordjournals.molbev.a040525 PMID: 3146682.

96. Nakamura Y, Tabata S. Codon-anticodon assignment and detection of codon usage trends in seven microbial genomes. Microb Comp Genomics. 1997; 2(4):299–312. Epub 1997/01/01. https://doi.org/10.1089/omi.1.1997.2.299 PMID: 9689228.