G3, 2024, jkae207
https://doi.org/10.1093/g3journal/jkae207
Advance Access Publication Date: 30 August 2024
Investigation

Genomic factors shaping codon usage across the Saccharomycotina subphylum

Bryan Zavala,[1,12] Lauren Dineen,[1] Kaitlin J. Fisher,[2,3] Dana A. Opulente,[4,5] Marie-Claire Harrison,[6,7] John F. Wolters,[5] Xing-Xing Shen,[8] Xiaofan Zhou,[9] Marizeth Groenewald,[10] Chris Todd Hittinger,[5] Antonis Rokas,[6,7] Abigail Leavitt LaBella[1,11,*]

[1]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, North Carolina Research Campus, Kannapolis, NC 28081, USA
[2]Department of Biological Sciences, SUNY Oswego, Oswego, NY 13126, USA
[3]Laboratory of Genetics, Wisconsin Energy Institute, Center for Genomic Science Innovation, J. F. Crow Institute for the Study of Evolution, University of Wisconsin–Madison, Madison, WI 53726, USA
[4]Department of Biology, Villianova University, Villanova, PA 19085, USA
[5]Laboratory of Genetics, DOE Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, Center for Genomic Science Innovation, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI 53726, USA
[6]Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA
[7]Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN 37235, USA
[8]Institute of Insect Sciences and Centre for Evolutionary and Organismal Biology, Zhejiang University, Hangzhou 310058, China
[9]Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Center, South China Agricultural University, Guangzhou 510642, China
[10]Westerdijk Fungal Biodiversity Institute, 3584 CT Utrecht, The Netherlands
[11]Center for Computational Intelligence to Predict Health and Environmental Risks (CIPHER), University of North Carolina at Charlotte, 9201 University City Boulevard, Charlotte, NC 28233, USA
[12]Present address: Lab of Mucosal Pathogens and Cellular Immunology, Division of Bacterial Parasitic and Allergenic Products, Office of Vaccines Research and Review, Center for Biologics Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD 20993, USA

*Corresponding author: Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, North Carolina Research Campus, 150 N Research Drive, Kannapolis, NC 28081, United States. Email: alabell3@charlotte.edu

Codon usage bias, or the unequal use of synonymous codons, is observed across genes, genomes, and between species. It has been implicated in many cellular functions, such as translation dynamics and transcript stability, but can also be shaped by neutral forces. We characterized codon usage across 1,154 strains from 1,051 species from the fungal subphylum Saccharomycotina to gain insight into the biases, molecular mechanisms, evolution, and genomic features contributing to codon usage patterns. We found a general preference for A/T-ending codons and correlations between codon usage bias, GC content, and tRNA-ome size. Codon usage bias is distinct between the 12 orders to such a degree that yeasts can be classified with an accuracy >90% using a machine learning algorithm. We also characterized the degree to which codon usage bias is impacted by translational selection. We found it was influenced by a combination of features, including the number of coding sequences, BUSCO count, and genome length. Our analysis also revealed an extreme bias in codon usage in the Saccharomycodales associated with a lack of predicted arginine tRNAs that decode CGN codons, leaving only the AGN codons to encode arginine. Analysis of Saccharomycodales gene expression, tRNA sequences, and codon evolution suggests that avoidance of the CGN codons is associated with a decline in arginine tRNA function. Consistent with previous findings, codon usage bias within the Saccharomycotina is shaped by genomic features and GC bias. However, we find cases of extreme codon usage preference and avoidance along yeast lineages, suggesting additional forces may be shaping the evolution of specific codons.

Keywords: Saccharomycotina; codon usage bias; codon; tRNA; machine learning

Introduction

The genetic code is degenerate, and all but 2 amino acids are encoded by multiple synonymous codons. It is consistently observed that the use of synonymous codons is biased within genes, between genes within a genome and between species (Ikemura 1985; Plotkin and Kudla 2011). The unequal use of synonymous codons is known as codon usage bias (CUB) and has been observed widely across organisms (Ikemura 1985). Some of these biases are associated with the functional properties of codons. Generally, the relative frequency of synonymous codons is proportional to the number of available tRNAs with a corresponding anticodon, which facilitates translation (Grantham 1978). Transcripts containing codons decoded by abundant tRNAs are also frequently expressed at a higher level (Sharp *et al.* 1986). Codon usage bias has been found to have a functional role in many cellular processes, including mRNA stability (Radhakrishnan *et al.* 2016), transcriptional control (Coghlan and Wolfe 2000), protein folding (Zhou *et al.* 2015), chromatin availability (Zhao *et al.* 2021), and ribosome dynamics (Yu *et al.* 2015).

Received on 13 June 2024; accepted on 21 August 2024

The involvement of codon usage bias in diverse cellular processes suggests that codon usage is under natural selection. Natural selection acting on codon usage is typically attributed to translational selection, a form of adaptation associated with increased translational efficiency in highly expressed genes accomplished by tuning CUB to the most abundant tRNAs (dos Reis *et al.* 2004). Nonadaptive forces such as GC-biased gene conversion, mutational bias, and genetic drift also have a signature effect on the nucleotide composition landscape. GC-biased gene conversion is a process that generally influences eukaryotes and is due to recombination that prefers the transmission of GC alleles, thus increasing the GC content (Lesecque *et al.* 2013). However, biased gene conversion may not significantly impact GC content in some species, such as *Saccharomyces cerevisiae* (Liu *et al.* 2018). Mutational biases can also impact global GC composition. Mutational biases are driven by differential transition and transversion rates (Zhu *et al.* 2014), mutational pressure toward AT content driven by the deamination of cytosine to uracil (Lynch *et al.* 2008), and replication strand bias (Pavlov *et al.* 2002). Finally, genetic drift decreases selection efficiency, including translational selection on codon usage (Galtier *et al.* 2018).

The relative degree to which adaptive and nonadaptive forces shape codon usage bias differs across the tree of life. In some species, especially those with small population sizes like large mammals and reptiles, neutral forces are thought to predominately shape codon usage biases (Galtier *et al.* 2018). In this case, codon usage is highly correlated with global GC content and is not predictive of gene expression (dos Reis *et al.* 2004). Other species like *Escherichia coli*, however, exhibit a strong signature of translational selection such that codon composition is highly predictive of gene expression level (Boel *et al.* 2016). Methods have been developed to disentangle these forces to assess the relative contribution of translational selection on genomic codon usage bias (dos Reis *et al.* 2004; Gilchrist *et al.* 2015; Landerer *et al.* 2018). These methods have found a wide range of translational selection levels across organisms (dos Reis *et al.* 2004). This variation has been attributed to factors such as genome size, total number of genomic tRNAs (dos Reis *et al.* 2004), and effective population size (Galtier *et al.* 2018). These features, however, do not fully explain the observed variation, as numerous exceptions exist (LaBella *et al.* 2019). It remains to be seen if other metabolic, phenotypic, or genomic traits impact the degree of translational selection on CUB or CUB itself.

The fungal subphylum Saccharomycotina, 1 of 3 subphyla in the phylum Ascomycota, is an excellent model system for studying codon usage evolution (LaBella *et al.* 2019; 2021; Nalabothu *et al.* 2023). Codon usage in the Saccharomycotina is among the most unusual in eukaryotes as it harbors three orders (Serinales, Alaninales, and Ascoideales) that have undergone nuclear codon reassignments in which the CTG codon encodes for serine or alanine, instead of leucine (Muhlhausen *et al.* 2016; Riley *et al.* 2016; Wada and Ito 2023). The extensive genomic diversity of these yeasts provides insight into the evolution of codon usage and possible factors shaping them. Previous research on a subset of the subphylum of Saccharomycotina (332 or fewer yeasts) has revealed that most, but not all, species within this group are subject to high levels of translational selection (LaBella *et al.* 2019; Cope and Shah 2022; Wint *et al.* 2022). This work found that genomic tRNA copy number was the most robustly associated with levels of translational selection on codon usage, with the highest levels occurring at intermediate tRNA levels (LaBella *et al.* 2019). Surprisingly, genome size was not highly correlated with levels of translational selection. There are likely other forces shaping codon usage bias within this diverse subphylum. These findings warrant additional investigation in the entire yeast subphylum, which the Y1000+ Project has recently made possible through the generation of genomic and phenotypic characterization of 1,154 yeast strains from 1,051 species—nearly all known species of yeasts (Opulente *et al.* 2024).

To understand the evolution of codon usage bias in Saccharomycotina, we analyzed the genomic-wide codon usage metrics and their relationship to phenotypic or genomic features across the subphylum. This analysis builds on previous work (LaBella *et al.* 2019) by more than tripling the species number, including machine learning analysis, and includes metabolic niche breadth data. We measured relative synonymous codon usage (RSCU), which reflects codon preference, and the association between RSCU and various genomic features. Across the subphylum, we found a general preference toward AT-ending codons, but there were significant differences in codon usage between species. A phylogenetic principal component analysis (pPCA) revealed distinct patterns of RSCU values that differentiated the Serinales and Dipodascales from the rest of the subphylum. A random forest classifier was able to classify strains into their taxonomic orders with high accuracy based solely on RSCU values, indicating the presence of distinct patterns in codon usage patterns in the subphylum. Further phylogenetic statistical analysis on codon usage resulted in significant correlations with GC content metrics but also revealed intriguing correlations with specific genomic features. Significant correlations were found between translational selection levels (measured by S-value) and tRNA count, assembly metrics, and the number of protein-coding sequences but not with the niche breadth phenotypes. We also conducted an in-depth analysis of RSCU values of CGN codons, which revealed that no tRNA was computationally predicted to decode CGN codons in the Saccharomycodales. The Saccharomycodales contained 23 species (out of 24) that were predicted to lack CGN-decoding tRNA genes. We conducted additional analyses on gene expression, tRNA alignments, and conserved arginine sites, further suggesting the loss of functional CGN codons. This analysis of codon usage throughout the subphylum highlights the diversity of codon usage strategies and identifies some of the genomic features that may constrain this diversity.

Materials and methods

Codon usage data and metrics

The protein-coding sequence annotations of 1,154 yeast strains from 1,051 species were collected from a previous study from the Y1000+ Project (http://y1000plus.org) of the yeast genomes in the subphylum Saccharomycotina (Supplementary Table 1; Opulente *et al.* 2024). Mitochondrial sequences were previously filtered from these genomes, and the annotations, therefore, do not contain mitochondrial coding sequences. Codon calculations were generated through the sequence analysis tool EMBOSS v6.6.0.0 (Rice *et al.* 2000), which calculated codon frequencies, percentages, and counts for every coding sequence in each yeast strain. Most codon analysis software does not include all possible yeast nuclear translation tables and, therefore, is inaccurate for the Saccharomycotina. To address this, the RSCU for every yeast strain was calculated by in-house scripts (https://github.com/The-Lab-LaBella/RSCU_Calculation_Analysis) that accounted for the nuclear codon reassignments in the Serinales, Alaninales, and Ascoideales. The RSCU is the observed number of occurrences of a synonymous codon divided by the expected number of occurrences if codon usage was random (Sharp *et al.* 1986). We calculate

RSCU by multiplying the frequency of each codon by the number of synonymous codons for that amino acid. For example, if there are 100 valine codons and we observe there are 60 GTT codons, the RSCU would be 2.4 (60/100*4). The RSCU values for every protein-coding sequence were used to calculate each genome's genome-wide average RSCU values.

To estimate the level of translational selection acting on codon usage in our yeast genomes, we applied the S-test to calculate the S-value proposed by dos dos Reis *et al.* (2004) using the tRNA adaptation index (tAI) package in R. We used this package to generate tAI values for every gene. The S-value for each genome was then calculated as the correlation between tAI and a combination of the synonymous third codon position GC content (GC3s) and the effective number of codons. A correlation (S-value) of 1 suggests that codon usage is influenced by translational selection across the genome. This method was chosen because we could conduct high-throughput analysis on the command line instead of a graphical user interface. Other methods have shown similar trends in detecting selection on codon usage in the yeasts (Landerer *et al.* 2018).

tRNA analysis

tRNAscan-SE 2.0.9 (Chan *et al.* 2021) was used to predict tRNAs in the genome for each strain using the standard eukaryotic parameters. We generated a filtered tRNAscan-SE dataset in which we removed pseudogenes, tRNAs lacking isotypes, and tRNAs containing mismatches between anticodon and predicted isotype. Serine or alanine tRNAs with the CAG anticodon were included for the Serinales, Alaninales, and Ascoideales as the orders have undergone a reassignment of the canonical codon (CUG) for leucine to serine [Serinales (Santos and Tuite 1995) and Ascoideales (Krassowski *et al.* 2018)] or alanine [Alaninales (Muhlhausen *et al.* 2016; Riley *et al.* 2016)]. After filtering, the total tRNA-ome size was calculated as the sum of all tRNAs. We conducted additional analyses on the tRNAs of Saccharomycodales yeasts, specifically analyses of the arginine-decoding tRNAs. First, we analyzed the mitochondrial tRNA content of several yeast species in the Saccharomycodales. The mitochondrial genomes were obtained from a recent study (Wolters *et al.* 2023), and they were analyzed for tRNA content using tRNA-scan with the organelle option.

The conservation of Saccharomycodales arginine tRNAs was also assessed using the tRNAviz web browser (Lin *et al.* 2019). In addition to visualization, tRNAviz calculates a penalty score for each position in the supplied tRNAs. Scores range from 0 (identical to the reference) to −15, indicating a highly divergent site. The reference set used in this analysis was the pre-computed Saccharomycotina dataset.

Annotations of tRNA modification enzymes were obtained from KEGG (Kanehisa and Goto 2000) annotations previously conducted (Opulente *et al.* 2024). The annotations examined were KEGG Ortholog groups K15440 (TAD1), K15441 (TAD2), and K15442 (TAD3). They were examined manually to identify any annotation errors.

Genomic and phenotypic traits of yeasts

Next, we aimed to identify genomic and phenotypic traits that may influence levels of codon usage bias and translational selection in yeasts. To do this, we used a variety of yeast traits obtained from and detailed in the publication of the genomes (Opulente *et al.* 2024). Genomic traits analyzed were GC content metrics, tRNA-ome size, S-value, genome size, genome assembly metrics, BUSCO completeness metrics, and number of protein-coding sequences. Phenotypic traits analyzed were metabolic niche breadth (Opulente *et al.* 2024). Metabolic niche breadth is the total number of carbon or nitrogen substrates on which a yeast strain can grow.

Visualization and phylogenetic statistical analyses

The genome-wide average RSCU values across the subphylum were analyzed with hierarchical clustering performed using the ComplexHeatmap package (Gu *et al.* 2016). To decipher patterns and covariance across strains in their codon usage variation, a pPCA was performed in R using the phytools package (Revell 2024). A pPCA was used to take into account the nonindependence of closely related species. The phylogeny of all 1,154 yeasts, which was constructed using maximum likelihood analysis of a concatenated alignment of 1,403 single-copy orthologs, was obtained from the previous study (Opulente *et al.* 2024). Similarly, phylogenetic generalized least squares (PGLS) and phylogenetic independent contrasts (PICs) were used to estimate correlations between biological features, genomic features, and RSCU. The RSCU vs RSCU analysis was conducted under a PGLS with maximum likelihood estimation of Pagel's lambda in the caper v1.7.3 package (Orme *et al.* 2013). In cases where the lambda estimation failed, possibly due to a maximum likelihood outside the bounds of 1×10^{-6} and 1, it was set to 1 to ensure the most conservative analysis. The yeast feature vs yeast feature analysis was also conducted using the same PGLS method. In the feature analysis, 2 genomes were dropped due to naming issues: genome names yHMPu5000026270_Candida_sp._SPAdes, yHMPu5000034970_Candida_sp._SPAdes. Due to differences in scale, the feature vs RSCU analysis was conducted using a PIC in the ape v5.7-1 package (Paradis and Schliep 2019). Strains without metabolic breadth measurements were removed before phylogenetic comparative analysis.

RSCU random forest classifier

To determine if yeast strains can be classified by codon usage, a random forest classifier model was created to determine the order of a particular yeast strain based solely on their genome-wide average RSCU values. The model was built from the R package randomForest v.4.7-1.1 (Liaw and Wiener 2002) with a matrix comprising the genome-wide average RSCU values from 59 codons for each strain. The model was trained with 70% of the data, and 30% was withheld for testing. The trained model included information regarding the significant variables (codons) in classifying yeast strains to orders and error rate stabilization.

Gene expression analysis

We conducted mRNA-sequencing to verify the presence of rare codons in the expressed genes of Saccharomycodales (BioProject PRJNA1144926; Accessions SAMN43045963, SAMN43045964, SAMN43045965.) Triplicate cultures *of Hanseniaspora occidentalis* var. *occidentals* and *Hanseniaspora uvarum* were grown for 18 h in 20 mL rich YPD medium (yeast extract, peptone, 2% glucose) in a room-temperature shaker at 250 rpm. Cell pellets were flash-frozen and stored at −80°C. mRNA was isolated using the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB) with the NEBNext Ultra II Directional RNA Library Prep kit (NEB). One of the 3 *H. occidentals* var. *occidentals* libraries failed, leaving only 2 replicates for sequencing. Sequencing was performed at the University of Wisconsin–Madison Biotechnology Center. Paired-end reads were trimmed and deduplicated using fastp v.0.20.1 (Chen 2023). A guided assembly was conducted using HISAT2 v2.2.1 (Kim *et al.* 2019) to align the sequence reads against their respective reference draft genomes, both of which were retrieved from the Y1000+ Project (Opulente *et al.* 2024). The resulting alignment files were processed through Stringtie v2.2.1 (Shumate *et al.* 2022) for transcript annotation from their respective reference species annotation files, followed by extracting the protein-coding sequences from the reference genomes of each replicate annotation file into FASTA files using TransDecoder v5.5.0. Putative transcript functions were assigned using the NCBI Blast web search (Madden 2013).

Conserved arginine site analysis

Our analyses revealed an extreme bias in arginine codons in the Saccharomycodales. Therefore, we explored the evolutionary context of conserved arginine positions within the genomes. We identified highly conserved arginine amino acid positions in the 1,403 single-copy orthologs previously used to build the Saccharomycotina phylogeny (Opulente *et al.* 2024). The DNA sequences of these orthologs were translated into amino acid sequences using EMBOSS *transeq* (Rice *et al.* 2000) and then aligned using mafft v7.273 (Katoh *et al.* 2002) with the "fftnsi" settings. The DNA-coding sequences were then aligned by codon using the protein alignment as a reference using a custom script available on the Figshare repository. Conserved arginine positions, defined by a simple majority of sequences, were identified in the amino acid alignments using the EMBOSS *cons* function. We then focused on the Saccharomycodales, the Saccharomycodales (the sister order to the Saccharomycodales), and the Phaffomycetales (an outgroup). We identified sites in the alignment where 80% or more of the amino acids were arginine, and the exact codon was conserved within each order at 80% or more. This procedure identified highly conserved arginine positions with highly conserved arginine codons. This allowed us to identify patterns of arginine codon usage within this group.

Results and discussion

Codon usage variation across the subphylum

Our previous work identified significant variation in codon usage in 332 strains of the yeast subphylum (LaBella *et al.* 2019). Here, we have more than tripled the sampling within the subphylum. First, we calculated the strain-level average RSCU values for the 59 degenerate codons. Hierarchical clustering revealed patterns of codon preference throughout the subphylum (Fig. 1; Supplementary Table 2 and Fig. 1). We observed a general preference (RSCU > 1) for A/T-ending codons, while G/C-ending codons were unpreferred (RSCU < 1). However, there were notable deviations from the general patterns in RSCU values. Specifically, the codon TTG, which codes for leucine, was grouped among the preferred A/T-ending codons and was the only G/C-ending codon generally preferred. Another observation was that the A/T-ending codons of CTT (leucine), CGT (arginine), GTA (valine), ATA (isoleucine), CTA (leucine), and CGA (arginine) were among the unpreferred codons in the subphylum. Other RSCU patterns were specific to 1 or a few orders of yeasts. The codon AGA (arginine) exhibited an extreme preference throughout all the orders except for Lipomycetales, Trigonopsidales, and Dipodascales. A similar pattern was observed with the TTG codon. The codon TTA was highly variable across the subphylum and exhibited extreme preferences (preferred and unpreferred).

We also investigated the correlation between RSCU and other codon-associated traits, including synonymous GC3 (GC3 s) composition, translational selection (S-value), and tRNA-ome size (Fig. 1; Supplementary Tables 1 and 2). These traits represent various factors influencing codon usage variation across the subphylum. *S*-value (a measure of translational selection) varies significantly across the subphylum (Fig. 1). A high level of translational selection on codon usage within a genome is characterized by a high correlation between codon adaptation to genomic tRNA numbers (tAI) and compositional bias. Therefore, a value of 1 indicates a perfect correlation and a high inferred level of translational selection on codon usage. An *S*-value of ∼0.5 indicates an intermediate level of translational selection. Across the yeast subphylum, we observed S-values ranging from a low of −0.015 to a high of 0.931 (mean = 0.73, median = 0.76; Supplementary Table 1). GC composition, which can influence codon usage bias in a nonselective way, is also highly variable across yeasts (Supplementary Table 1). We observed synonymous GC3 values from a low of 4.7% to a high of 91.3%. The average across the genomes was 42.36% and a median of 43%. The tRNA-ome is another source of codon usage evolution as the repertoire of tRNA copy number is directly implicated in the presence of translation selection (dos Reis *et al.* 2004). The median tRNA-ome in the subphylum was 208 tRNAs, ranging from a minimum of 49 tRNAs to a maximum of 1,589 tRNAs. *Candida lidongshanica* and *Aciculoconidium aculeatum* from the Serinales have an unusually large predicted tRNA-ome with 1,589 tRNAs and 1,037 tRNAs, respectively. Further analysis of the number of distinct anticodon types in each strain revealed that only 1 species, *Martiniozyma abiesophila* in the Pichiales, violates the theoretical minimum of 30 anticodon types (Marck and Grosjean 2002) by containing only 28 nuclear anticodon types. The tRNA content analyses present here are limited by the references used to model tRNAs within tRNAscan-SE and potentially incomplete genomic sequences. Additional experimental work is required to fully elucidate the expressed tRNA content across the yeast subphylum.

RSCU patterns are a defining feature of yeast orders

To examine interspecies RSCU codon usage variation, we conducted a pPCA. The pPCA revealed that most of the variation (63.56%; Fig. 2) between species is driven by G/C- and A/T-ending codons, which is consistent with previous analyses (LaBella *et al.* 2019). The second principal component (11.96%), which differentiates the Serinales/Ascoideales and Dipodascales/Lipomycetales, is driven by a set of codons that include TTG, CTT, CTA, CTG, AGT, TCA, and TCT. In particular, the usage of TTG, CTT, and CTA (all leucine codons) separates and clusters the Serinales and Ascoideales from the rest. This reflects the avoidance of the reassigned CTG codon in these orders. The Dipodascales also exhibited unique clustering driven primarily by codon preferences for CTG (leucine codon), AGT, TCA, and TCT (serine codons). An interesting deviation was seen for *Dipodascopsis tothii* (order Lipomycetales), which is separate from all the other species in the subphylum. *Dipodascopsis tothii* exhibits a long-branch length from its relatives and an extremely high GC3 content (91%). The results of this species indicate that it has undergone a unique trajectory in its codon usage that is different from all the rest.

The application of machine learning to genomic data is emerging as a powerful tool for studying yeast traits and evolution (Harrison *et al.* 2024; Opulente *et al.* 2024). The pPCA suggested that codon usage can distinguish some, but not all, yeast orders. Machine learning methods, however, can pick up patterns that may be missed in a PCA. To test if codon usage is a distinguishing feature of yeast orders, we constructed a random forest classifier model to classify the order of yeast strains solely on their genome-wide RSCU values (Supplementary Table 2). The model

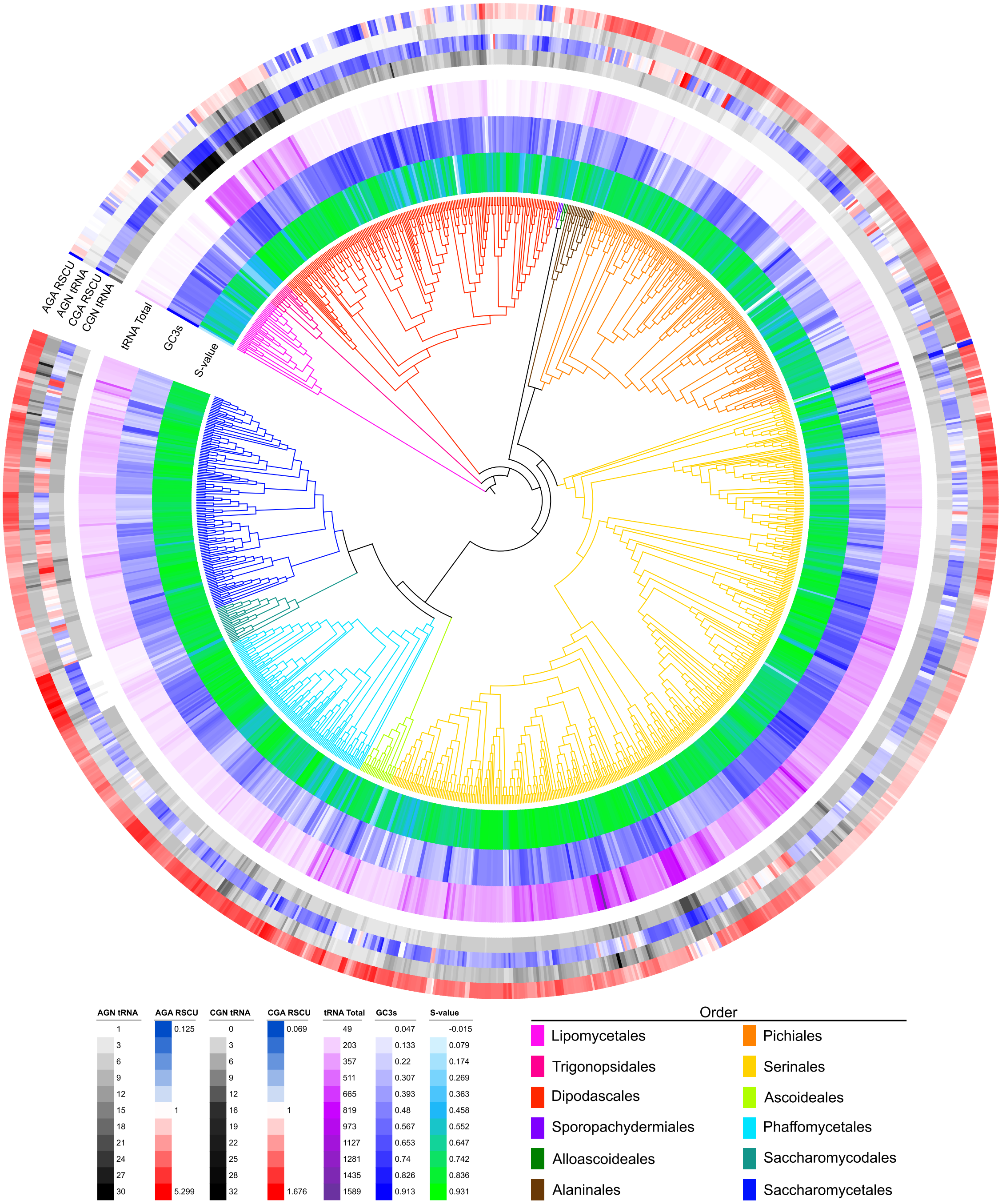

Fig. 1. Variation of codon-associated metrics across the Saccharomycotina subphylum. The S-value is a measure of translational selection on codon usage that varies from no selection (0) to high levels of selection (1). The synonymous GC3s and total genomic tRNA count are also shown. Representative RSCU and individual tRNA counts for arginine are also shown. The RSCU and counts of decoding tRNAs vary widely across the phylogeny.

accurately classified 90.38% of the training data and 93.29% of the test data (30% previously withheld), which indicates that CUB is generally sufficient to differentiate yeast orders (Supplementary Fig. 2). The model was interrogated for the most important variables used for classification using the mean decrease Gini index (Supplementary Table 3). The most important variable in the model was the codon CTA (leucine), followed by the codons TTG (leucine), AGA (arginine), CTG (leucine), GTA (valine), and CGA (arginine). The relatively high importance of the CTA codon is consistent with the finding that it is a rare case of an A/T-ending codon that is generally unpreferred throughout the subphylum. Moreover, this highlights that the reassignment of the CTG codon in 3 orders is a defining feature of these species. The random forest algorithm could distinguish between genomes belonging to different orders using only the 59 RSCU metrics. The RSCU values, therefore, likely contain significant phylogenetic information.

Codon usage biases are correlated with numerous genomic features

Codon usage bias may be shaped by various genomic and ecological factors over the course of evolution. To identify factors shaping codon usage bias of specific codons and overall levels of translational selection, we conducted numerous PGLS

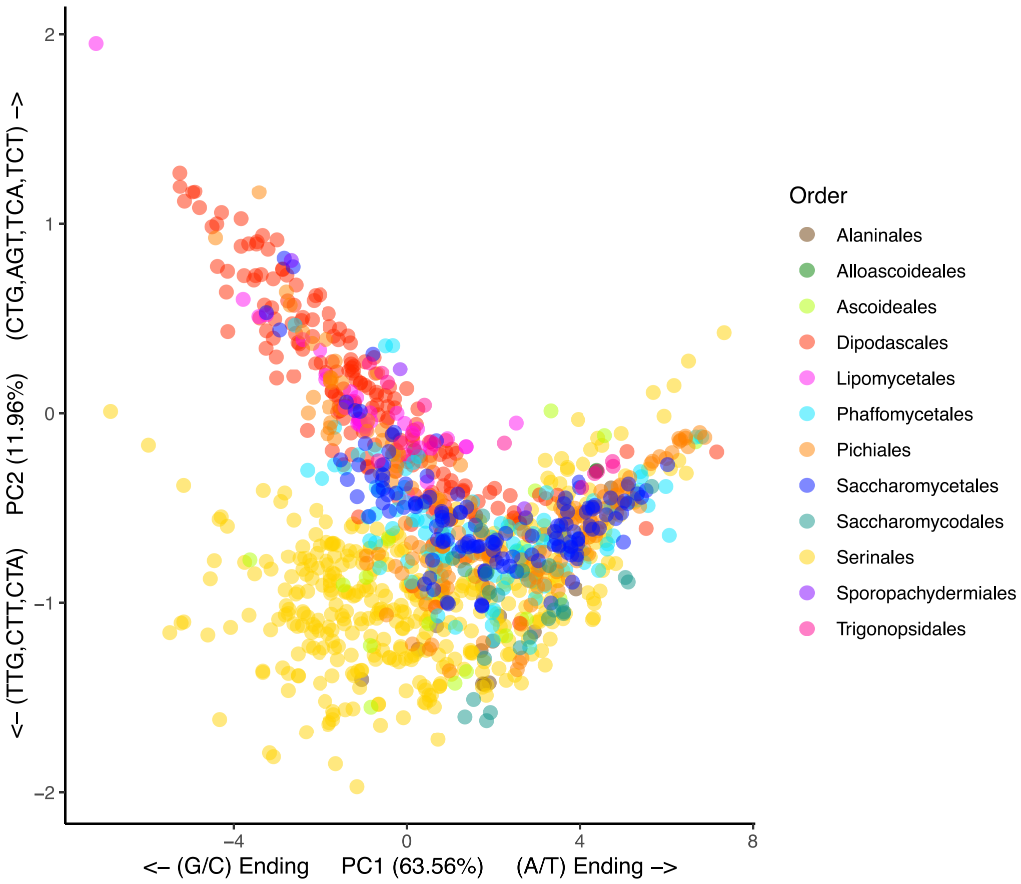

Fig. 2. pPCA of the 59 RSCU values of 1,154 yeast strains from 1,051 species. To derive the patterns and covariance of codon usage throughout the subphylum, a pPCA was conducted to determine their relationships. A pPCA was used to take into account the nonindependence of biological traits due to phylogeny. The results demonstrated that PC1, which explains 63.56% of the variation, was driven primarily by differential usage of G/C- and A/T-ending codons between species. PC2 explained 11.96% of the variation and differentiated the Serinales and Dipodascales orders. The Lipomycetales species at the top left corner, *Dipodascopsis tothii,* is driven by TTG, CTT, CTA, CTG, AGT, TCA, and TCT codons.

regressions. These regressions allowed us to account for the fact that our observations share ancestry and are, therefore, a nonrandom sample.

First, we explored what factors may shape the usage of specific codons by correlating genome-level RSCU values with genomic features, such as other RSCU values, GC content, and tRNA-ome size (Supplementary Table 4). We identified 13 codons strongly associated with GC content and tRNA-ome size. Many of these codons were previously identified in the PCA analysis (CTG, AGT, TCT, TTA, and CTT; Supplementary Fig. 3). We also tested for codon-to-codon correlations by testing 1,830 pairwise combinations of RSCU (Supplementary Table 5 and Fig. 4). After multiple test corrections, 1,785 pairwise combinations were significant at $P < 0.05$. Of the significant comparisons, 1,740 had the expected correlation between G/C- and A/T-ending codons—they were positively correlated within A/T or G/C comparison and negatively correlated for A/T vs G/C comparisons. There were 23 pairwise comparisons that violated our expectation that G/C- and A/T-ending codons should not exhibit positive correlations in RSCU. Comparisons with high slopes include a positive correlation between TTG (leucine) and CTA (leucine; slope = 1.50) and between AGG (arginine) and CTA (leucine; slope = 0.53). Additionally, 22 comparisons between A/T- or G/C-ending codons were not positively correlated. For example, negative correlations within groups include TCA (serine) and CTA (leucine; slope = −1.20) and AGA (arginine) and CGA (arginine; slope = −0.68). Of the 45 correlations producing results that deviated from the hypothesis that G/C- and A/T-ending codons should be anti-correlated, 43 involved arginine ($n = 12$) or leucine ($n = 35$) codons. This result is consistent with the previous study (LaBella *et al.* 2019) and may be associated with the large number of degenerate codons encoding arginine and leucine, leading to more opportunities for poor codon–tRNA pairing (Duret and Mouchiroud 1999; McVean and Vieira 2001).

Second, we explored the factors that shape translational selection on codon usage by identifying features that correlate with the S-value (dos Reis *et al.* 2004), which is a measure of translational selection (Supplementary Table 6 and Fig. 5) We examined the role of metabolic niche breadth for both carbon and nitrogen using a PGLS analysis. Previous work has shown that intrinsic factors, such as gene composition, drive metabolic niche breadth (Opulente *et al.* 2024). We did not find any association between genome-wide levels of selection on codon usage and carbon ($P = 0.966$) or nitrogen ($P = 0.579$) niche breadth (Supplementary Table 6). This suggests that translational selection on codon usage is not specific to generalist or specialist yeasts. Species with high S-values greater than 0.85 can metabolize between 2 (*Ogataea pini*) and 16 carbon sources (*Cyberlindnera nakhonratchasimensis;* Supplementary Table 1). Similarly, species with $S < 0.25$ can metabolize between 3 (*Martiniozyma abietophila*) and 13 carbon sources (*Blastobotrys peoriensis*; Supplementary Table 1).

We tested associations between translational selection on codon usage and other genomic features. This includes factors

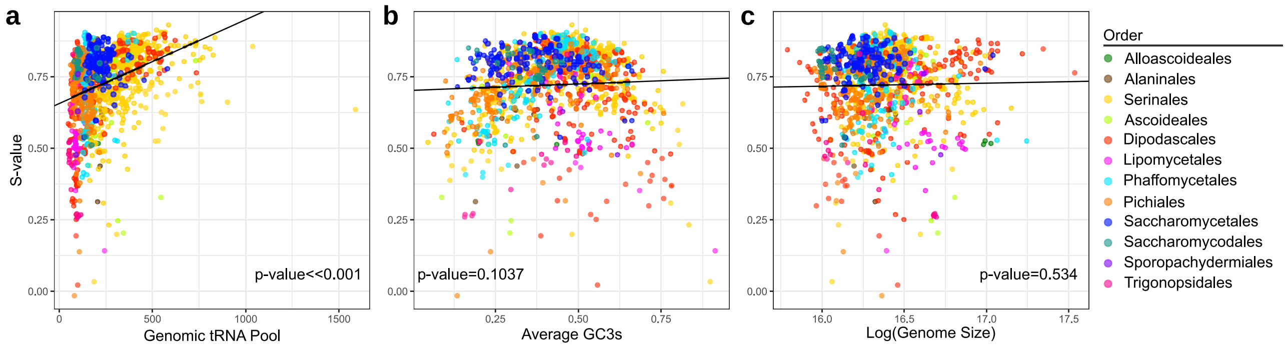

Fig. 3. Analysis of 1,154 yeasts reveals a significant association between selection on codon usage bias and tRNA but not with average GC3 content or genome size. a) PGLS of S-value and tRNA size revealed that there was a significant positive correlation (*P*-value ∼0, slope = 0.00029314). Multiple species in different orders exhibit a wide range of S-values at lower tRNA sizes. Species in orders Dipodascales, Serinales, Saccharomycetales, and Phaffomycetales tended to exhibit higher levels in S-value with increased tRNA size. b) PGLS of S-value and the average GC content in the third position of the codon that is synonymous were not correlated (*P* = 0.1037, slope = 0.04464). However, visualization suggests a nonlinear relationship in which the highest levels of translational selection occur at intermediate GC3 content. c) There was also no association between genome size (log value) and the level of translational selection (*P* = 0.53407, slope = 0.010563).

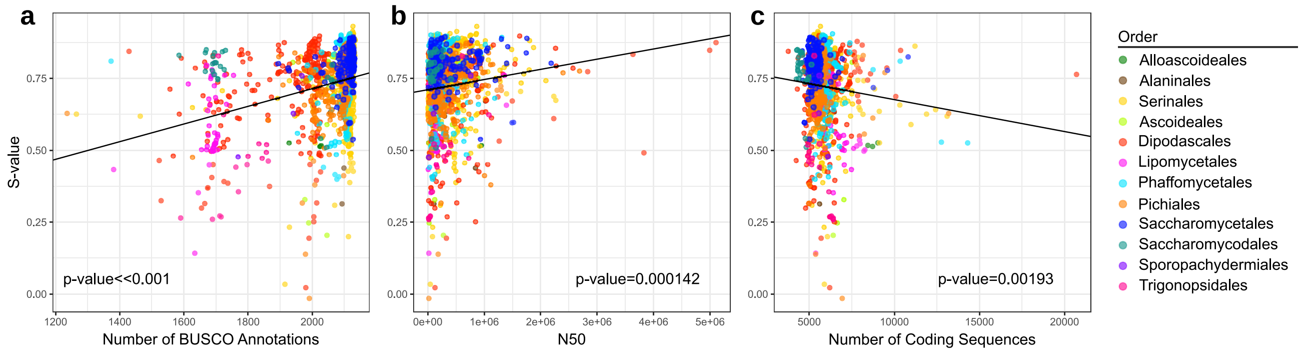

Fig. 4. Selection on codon usage is positively correlated with standard measures of genome completeness but not with the total number of protein-coding sequences predicted in a genome. a) Species with larger numbers of BUSCO-annotated genes also had higher levels of translational selection (PGLS, *P*-value ∼0, slope = 0.0003072). b) Species with higher-quality genome annotations (measured by N50) also exhibited higher levels of translational selection (PGLS, *P* = 0.00014272, slope = 3.54E−08). c) Unlike the association between the S-value and the number of BUSCO annotations, the total number of coding sequence annotations was negatively correlated with translational selection (PGLS, *P* = 0.001934, slope = −1.11E-05).

previously associated with selection on codon usage, such as GC3 content, genome size, and tRNA-ome size (dos Reis *et al.* 2004; LaBella *et al.* 2019). We found positive and significant correlations between the S-value and tRNA-ome (P-value ∼0) size but not with genome size (*P* = 0.534; Fig. 3). Interestingly, both genomic tRNA pool and genome size appeared to serve as lower-bound to high levels of translational selection. Genomes with low levels of translational selection ($S < 0.05$) were limited to genomes with small genomic tRNA pools. Conversely, high levels of translational selection were found across the genome and tRNA pool size spectrum. The results for synonymous GC3 content were similar to findings in previous studies that suggest the highest levels of translation selection occur at intermediate GC3 content (∼50% GC; LaBella *et al.* 2019).

We also found significant correlations between S-value and genome assembly metrics (number of contigs, N50), BUSCO metrics (number of BUSCOs, complete, single, fragmented, and missing), and the number of coding sequences (Supplementary Table 6). We found a positive correlation between S-value and N50 (*P* = 0.000143) and between S-value and the number of BUSCO genes (*P*-value ∼0), but we found a negative correlation between S-value and the number of coding sequences (*P* = 0.00193: Fig. 4). To further explore the role of all the features identified in pairwise comparisons, we built additive regression models for all possible combinations of tRNA pool size, genome size, N50, number of BUSCO genes, and total number of coding sequences (Supplementary Table 7). Based on both AIC and BIC criteria, the best-fitting PGLS model included all five variables. This result suggests that either genome assembly quality biases the estimates of selection on codon usage or underlying genomic features that make assemblies more difficult to assemble also impact selection on codon usage. Interestingly, we find that the number of protein-coding sequences is also negatively correlated with the N50 (*P* = 4.40E−12; Supplementary Table 6). These correlations could indicate that large genomes with a low gene density are harder to assemble and exhibit less translational selection on codon usage. This hypothesis is supported by research that has shown that lower levels of selection associated with low effective population size can lead to larger genomes (Petrov 2002). Our results and this model suggest that low levels of selection in this scenario apply to both synonymous and nonsynonymous changes.

Avoidance of CGN codons associated with arginine tRNA changes

Certain patterns of RSCU values warranted further examination, such as the observation that the CGN codons exhibited extreme avoidances in multiple clades (Fig. 1). Avoidance of CGN has previously been detected in other fungi, including one yeast

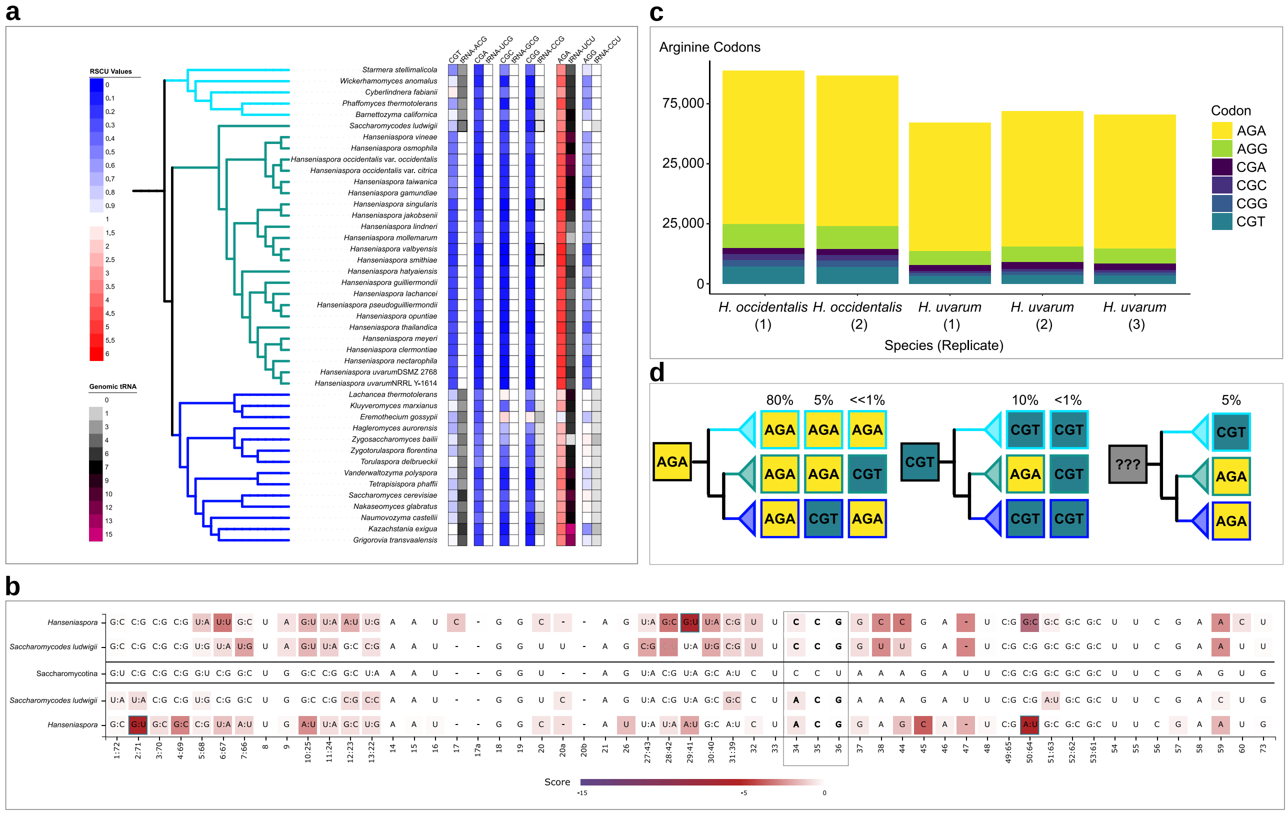

Fig. 5. Loss of CGN-decoding tRNAs in the *Hanseniaspora* and avoidance, but not complete removal, of CGN codons. a) The phylogenetic context tRNA genomic content and corresponding RSCU for arginine codons in *Hanseniaspora* species and representative species from Phaffomycetales (*top*—aqua) and Saccharomycetales (*bottom*—blue). All 4 CGN codons are underrepresented in *Hanseniaspora* (RSCU < 1), while the AGA codon is highly overrepresented (RSCU > 1) with the mean *Hanseniaspora* value being 5 out of 6. Three *Hanseniaspora* species have a hypothesized CGG-decoding tRNA, but their tRNA-scan scores are marginal. b) Comparison of the annotated CCG and ACG anticodon-containing tRNAs and pseudo-tRNAs compared to the rest of the Saccharomycotina. The darker shaded locations indicate positions that are divergent from the reference sequence shown in the middle. c) The total number of arginine codons that were identified in 5 transcriptomes in 2 *Hanseniaspora* species. The majority of identified codons were AGN, but there were still CGN codons detected in the mRNA expression data. d) The evolutionary history of conserved arginine positions (80% across Saccharomycotina) in orthologous genes that are also conserved within the Phaffomycetales (top), Saccharomycodales (middle), and Saccharomycetales (bottom). CGN codons are rare overall at conserved positions. At positions where the ancestral codon was likely CGT, the vast majority of these positions evolved to AGA (134/137). Overall, there were only 4 sites out of 1,404 that retained CGT across the Saccharomycodales.

mitochondrial genome, *Eremothecium gossypii* (Carullo and Xia 2008). This extreme bias against CGN led us to investigate the arginine tRNAs. In the results generated from tRNAscan-SE, the arginine tRNAs demonstrated a notable characteristic primarily in the Saccharomycodales. In the *Hanseniaspora* clade, all but 3 species (*H. singularis*, *H. valbynesis*, and *H. smithiae*) are predicted to be missing the necessary tRNAs to decode CGN codons. However, all the *Hanseniaspora* had predicted tRNAs for AGN codons with an extreme abundance of tRNA-UCU anticodons. The tRNA-UCU anticodons can complementarily base-pair with the codon AGA and AGG through wobble base pairing, which could explain the extreme preference for codon AGA demonstrated in the RSCU heatmap (Fig. 5a) and the low preference of CGN and AGG codons. As noted previously, three *Hanseniaspora* species (*H. singularis*, *H. valbynesis*, and *H. smithiae*) were annotated with a tRNA copy of tRNA-CCG. The presence of this tRNA is particularly interesting as these species are in the fast-evolving lineage (FEL), which has historically experienced significantly higher rates of mutations and gene loss (Steenwyk *et al.* 2019). The outgroup to the *Hanseniaspora* clade, *Saccharomycodes ludwigii*, is the only species in Saccharomycodales with multiple predicted isotypes to decode CGN codons. Previously reported Saccharomycodales mitochondrial genomes (Wolters *et al.* 2023) were also analyzed for tRNA genes using tRNAscan-SE (Supplementary Table 8; Figshare). The only mitochondrial arginine tRNA was tRNA-UCU, which decodes AGA and AGG. This eliminates the possibility that a mitochondrial tRNA is exported to alleviate the nuclear deficiency.

An alternative hypothesis to the degeneration of tRNA genes is that tRNA-scan incorrectly described the CGN-decoding tRNAs as pseudogenes. We obtained the sequences for all tRNA isotypes predicted to decode CGN codons regardless of their reported score or predicted isotype (Supplementary Table 9). All the *Hanseniaspora* have a tRNA with an ACG anticodon (decoding CGT), but the predicted isotype is either lysine, histidine, glycine, or arginine. Similarly, there are an additional 20 *Hanseniaspora* species that have predicted tRNAs with a CCG anticodon (decoding CGG). These are all predicted to have a histidine or glycine isotype. To determine why tRNA-scan predicted a mismatch between codon and tRNA, we used tRNAviz (Lin *et al.* 2019) to compare the *Hanseniaspora* sequences to the Saccharomycotina reference. The alignments of the tRNA sequences revealed positions that were similar and divergent from the consensus sequences (Fig. 5). The *Hanseniaspora* tRNAs with a CCG anticodon diverged from the consensus and *S. ludwigii* at the conserved position 29:41 in the anticodon stem of the tRNA. There is a predicted mismatch here between a G and U base. Previous investigations in S.

cerevisiae found that tRNA with variants in this location is associated with tRNA decay and maturation dynamics (Payea *et al.* 2020), and no tRNA modifications have been described at this location (Suzuki 2021). The *Hanseniaspora* tRNAs with an ACG anticodon diverged from the other sequences at positions 2:71 (mismatch) and 50:64. The mismatch at position 2:71 is located along the acceptor stem, which has been shown to physically interact with the arginyl-tRNA synthetase in *S. cerevisiae* (Delagoutte *et al.* 2000). This position has also been shown to be a positive identity element for the arginyl-tRNA synthetase in bacteria (Giege and Eriani 2023). The change at position 50:64 is along the T-arm. The T-arm is not required for tRNA function, suggesting changes in this region may be neutral (Krahn *et al.* 2020).

Additionally, many *Hanseniaspora* genomes have lost the tRNA adenosine deaminase (*TAD*) genes, which include homologs that have been shown to modify the wobble position arginine tRNAs with an ACG anticodon (decoding CGT codons; Wolf *et al.* 2002; Supplementary Table 10). The enzyme encoded by *TAD1* is found in 9 of the 24 (38%) Saccharomycodales species, a fraction that is significantly lower than across the rest of the yeast subphylum (1065/1130 or 94%, χ^2 with Yates' continuity correction $P < 2.2e-16$) or when compared to the Saccharomycetales (134/135 or 99%, χ^2 with Yates' continuity correction $P < 2.2e-16$). The enzymes encoded by *TAD2* and *TAD3* form a heterodimer (Dance *et al.* 2001). In the Saccharomycodales, 21 out of 24 yeasts (88%) have both components, which is comparable to the rest of the subphylum (963/1130 or 85%). This tRNA modification enzyme complex modifies the wobble position (A) to inosine (I) to allow for better wobble pairing (Wada and Ito 2023). While additional evidence would be required to demonstrate that the *Hanseniaspora TAD* genes modify arginine tRNAs, this loss further narrows down possible explanations for how these species decode CGN codons.

While the CGN codons are rare in the *Hanseniaspora* species without a predicted tRNA (mean RSCU across all CGN of 0.177), they are not completely absent. If the tRNAs were completely nonfunctional, we would expect that genes containing CGN codons would be untranslatable, leading to their extinction or elimination of CGN codons. To determine if transcripts containing CGN codons are expressed, we conducted RNA-sequencing of *H. occidentalis* var. *occidentalis* and *H. uvarum* cultured in a rich glucose medium (Figshare; GenBank). We then counted the total number of arginine codons in the expressed genes. The majority of arginine codons in the transcripts were AGA (mean 76% across samples). The CGN codons only comprised 14% (mean across samples) of the total arginine codons. At the level of transcripts, 37% of genes expressed contained no CGN codons. Of the remaining transcripts that did contain CGN codons, the median number of CGN codons was 4, and the mean was 7.7. Most of the CGN codons are concentrated into a few transcripts, with 1% (258 of 22,498) of the transcripts containing more than 20 CGN codons. We investigated the 20 transcripts with the most CGN codons by BLASTing (tblastx) against the nonredundant protein database (Supplementary Table 11). Twelve transcripts had high similarity to Saccharomycodales genomes but not protein sequences, suggesting that they are either noncoding transcripts or previously unannotated protein sequences. The remaining eight sequences had partial matches to previously reported Saccharomycodales protein-coding genes. The sequence with the highest percent identity to a known gene matched the *FLO1* gene, which encodes a flocculation protein from *H. uvarum*. The relatively high number of CGN codons in this gene (62 codons) may be due to the previously characterized extended tandem repeats in this gene (Bidard *et al.* 1995). The low average percent identity (55%) of these 20 transcripts to known genes suggests they are not likely to be complete translated mRNA sequences. As a comparison, we also analyzed 20 randomly chosen transcripts with no CGN codons. In this case, 18 of the 20 translations had high similarity (average of 91%) to previously characterized proteins in the *Hanseniaspora*. Only 2 sequences showed no significant similarity in the BLAST database. This suggests that the CGN-containing transcripts from our sequencing experiment are not protein-coding mRNA transcripts.

Our analysis indicated that CGN codons are also generally avoided, albeit to a lesser degree, in Phaffomycetales and Saccharomycetales, the 2 orders most closely related to the Saccharomycodales. To examine if the Saccharomycodales are evolving away from CGN codons compared to their relatives, we compared conserved arginine codon positions. These positions were identified in the 1,403 conserved orthologs used to determine the Saccharomycotina phylogeny. Positions were required to be arginine in 80% of all the sequences and the same codon 80% of the time within each order. This allowed us to examine 1,404 conserved arginine positions in 1,400 orthologs (Figshare Repository). We then determined the most parsimonious ancestral codon across the orders. In ∼85% of the positions, the inferred ancestor was AGA. In only 1 conserved position did we infer a change from AGA to CGT in the Saccharomycodales. In positions where CGT was the ancestral codon, the CGT codon was retained in only 3 positions, while 134 transitioned to AGA. We also conducted this analysis with positions that were conserved at a lower threshold (60%) and found similar results (Supplementary Table 12). This analysis suggests that, in conserved arginine positions, the Saccharomycodales have repeatedly switched codons from CGN to AGN. The change from CGN to AGN requires 2 base pair mutations in 6 of the 8 possible ways to go from CGN to AGN.

Collectively, our analysis found that the extreme avoidance of CGN codons in *Hanseniaspora* was likely associated with the accumulation of mutations in the CGN-decoding tRNAs. The transcriptomics data suggests that transcripts containing CGN codons are rare, with over a third of transcripts containing 0 CGN codons. This phenomenon may be associated with the loss of DNA repair and cell cycle genes previously observed in this group (Steenwyk *et al.* 2019). This scenario resembles the hypothesized situation leading to the CTG codon reassignment in 3 orders. In the "tRNA loss driven codon reassignment" model of CTG codon reassignment, loss of function mutations in tRNAs is the driving factor in codon reassignment. The CGN tRNAs appear to have accumulated several mutations, making them unrecognizable to tRNAscan-SE (Kollmar and Muhlhausen 2017). An alternative hypothesis is that codon reassignment is driven by "codon capture," in which a codon is driven to near extinction before changes in tRNAs (Osawa and Jukes 1989). Additional work is needed to test the expression or modification of the CGN tRNAs in *Hanseniaspora*.

Conclusions

The Saccharomycotina exhibits vast diversity in their codon usage and genomic tRNA content. Each order has evolved distinct codon usage patterns—including codon reassignments—that are sufficiently divergent to classify yeasts into their order using RSCU alone. Many forces shape the codon usage of Saccharomycotina, including mutational bias, genomic tRNA pool, and overall genome content. Similar to previous studies (Landerer *et al.* 2018; LaBella *et al.* 2019; Wint *et al.* 2022), we find that the genomic tRNA pool serves as a lower bound for the amount of translational selection acting on codon usage—small pools can exhibit a range of S-values. In contrast, large pools exhibit mostly high S-values.

We also find that the highest levels of translational selection occur at an intermediate GC content of the third codon position. Interestingly, we find that the N50, BUSCO number, and translational selection are all positively correlated with each other and negatively correlated with the total number of predicted protein-coding sequences.

Unlike previous studies, we identified an association between genome assembly and architecture and our measure of translational selection. Our additive model found that 5 variables (tRNA pool size, genome size, N50, number of BUSCO genes, and total number of protein-coding sequences) explained the most variation in S-value. The role of genome assembly features could be technical or biological. Lower-quality genomes may be missing genes and are more likely to be mis-annotated. This could lead to unintentional bias in the codon evaluation due to missing genes or tRNAs (Whibley *et al.* 2021). Conversely, biological genome features, like high GC content (Chen *et al.* 2013), repetitive regions, presence of introns, heterozygosity, and genome size (Jauhal and Newcomb 2021) can result in lower-quality genome assemblies. Many of these features, like GC content and genome size, have previously been found to be associated with codon usage bias (dos Reis *et al.* 2004; LaBella *et al.* 2019; Cope and Shah 2022). Therefore, the improved model fit associated with adding features like N50 and BUSCO may be associated with genome features we did not capture in our model.

Our analysis also uncovered an extreme avoidance of the CGN arginine codons in the Saccharomycodales. This was associated with a widespread predicted loss of function in the *Hanseniaspora* tRNAs, which decode CGN codons. Despite this observation, RNA-sequencing data identified several transcripts rich in CGN codons. However, whether these transcripts result in amino acids or the CGN tRNAs are expressed within the cell remains to be seen. Overall, the tRNAs that decode CGN codons have accumulated multiple mutations that may impact their function. The *Hanseniaspora* have also generally lost the Tad1 enzyme, and 3 have lost the ability to form the Tad2/Tad3 heterodimer—these enzymes are involved in modifying tRNAs to increase wobble base-pairing (Wolf *et al.* 2002; Delannoy *et al.* 2009). The evolution away from CGN codons, the accumulation of mutations in the CGN-decoding tRNAs, and the loss of the *TAD1* genes all support the hypothesis that many *Hanseniaspora* species have a significantly impaired ability to decode CGN codons.

Our analysis of codon usage bias in the Saccharomycotina revealed diverse codon usage biases, widespread selection on codon usage, and an extreme avoidance of CGN codons in an order that has potentially lost the tRNAs to decode CGN codons. Given the diversity in codon usage, the subphylum will likely be critical in answering outstanding questions in the field of codon usage. The *Hanseniaspora* may allow us to observe codon reassignment in action. The various species with incredibly large and very small numbers of tRNA genes may help us answer questions about the role of tRNA copy number and sequence variation in regulation. Finally, as we learn more about the ecology of these yeasts, we may be able to identify life history traits that impact selection on codon usage.

Data availability

The Y1000+ data can be obtained from the project website (http://y1000plus.org) or the associated Figshare repository https://doi.org/10.25452/figshare.plus.c.6714042. The Figshare project (https://figshare.com/projects/Genomic_factors_shaping_codon_usage_across_the_Saccharomycotina_subphylum/187236) contains the raw random forest model data, the assembled transcriptomes from the *Hanseniaspora*, the RSCU for all coding sequences in the subphylum, the conserved arginine analysis, and the mitochondrial tRNA analysis. The *Hanseniaspora* RNA-sequencing data have been deposited in BioProject PRJNA1144926, accessions SAMN43045963, SAMN43045964, and SAMN43045965.

Supplemental material available at G3 online.

Acknowledgments

Computational analyses were run in the UNC Charlotte high-performance computing cluster in Charlotte, North Carolina.

Funding

X.-X.S. was supported by the Science Fund for Distinguished Young Scholars of Zhejiang Province (LR23C140001), the Fundamental Research Funds for the Central Universities (226-2023-00021), and the Major Scientific Project of Zhejiang Lab (2021PE0AC04). This work was supported by the National Science Foundation (grants DEB-2110403 to C.T.H. and DEB-2110404 to A.R.). Research in the Hittinger Lab is also supported by the U.S. Department of Agriculture National Institute of Food and Agriculture (Hatch Projects 1020204 and 7005101), in part by the Department of Energy Great Lakes Bioenergy Research Center [Department of Energy Biological and Environmental Research Office of Science DE–SC0018409, and an H.I. Romnes Faculty Fellowship (Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation)]. Research in the Rokas lab is also supported by the National Institutes of Health/National Institute of Allergy and Infectious Diseases (R01 AI153356), and the Burroughs Wellcome Fund.

Conflicts of interest

A.R. is a scientific consultant for LifeMine Therapeutics, Inc. The other authors declare no other competing interests.

Author contributions

B.Z. conducted computational and statistical analyses, managed data, prepared figures, and co-wrote the manuscript with A.L.L.. L.D. conducted tRNA modification enzyme analysis. K.J.F. generated *Hanseniaspora* mRNA-sequencing data. D.A.O., X.-X.S., X.Z., J.F.W., M.C.H., M.Z., C.T.H., and A.R. provided computational support and reagents. A.L.L. designed and implemented computational analyses, managed data, prepared figures, co-wrote the manuscript, and supervised the project. All authors provided comments and input and approved the manuscript.

Literature cited

Bidard F, Bony M, Blondin B, Dequin S, Barre P. 1995. The *Saccharomyces cerevisiae* FLO1 flocculation gene encodes for a cell surface protein. Yeast. 11(9):809–822. doi:10.1002/yea.320110903.

Boel G, Letso R, Neely H, Price WN, Wong KH, Su M, Luff J, Valecha M, Everett JK, Acton TB, *et al.* 2016. Codon influence on protein expression in *E. coli* correlates with mRNA levels. Nature. 529(7586):358–363. doi:10.1038/nature16509.

Carullo M, Xia X. 2008. An extensive study of mutation and selection on the wobble nucleotide in tRNA anticodons in fungal mitochondrial genomes. J Mol Evol. 66(5):484–493. doi:10.1007/s00239-008-9102-8.

Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. Trnascan-se 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 49(16):9077–9096. doi:10.1093/nar/gkab688.

Chen S. 2023. Ultrafast one-pass fastq data preprocessing, quality control, and deduplication using fastp. iMeta. 2(2):e107. doi:10.1002/imt2.107.

Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC. 2013. Effects of gc bias in next-generation-sequencing data on de novo genome assembly. PLoS One. 8(4):e62856. doi:10.1371/journal.pone.0062856.

Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast. 16(12):1131–1145. doi:10.1002/1097-0061(20000915)16:12<1131::AID-YEA609>3.0.CO;2-F.

Cope AL, Shah P. 2022. Intragenomic variation in non-adaptive nucleotide biases causes underestimation of selection on synonymous codon usage. PLoS Genet. 18(6):e1010256. doi:10.1371/journal.pgen.1010256.

Dance GS, Beemiller P, Yang Y, Mater DV, Mian IS, Smith HC. 2001. Identification of the yeast cytidine deaminase cdd1 as an orphan c–>u rna editase. Nucleic Acids Res. 29(8):1772–1780. doi:10.1093/nar/29.8.1772.

Delagoutte B, Moras D, Cavarelli J. 2000. tRNA aminoacylation by arginyl-tRNA synthetase: induced conformations during substrates binding. EMBO J. 19(21):5599–5610. doi:10.1093/emboj/19.21.5599.

Delannoy E, Le Ret M, Faivre-Nitschke E, Estavillo GM, Bergdoll M, Taylor NL, Pogson BJ, Small I, Imbault P, Gualberto JM, *et al.* 2009. Arabidopsis tRNA adenosine deaminase arginine edits the wobble nucleotide of chloroplast tRNAArg(ACG) and is essential for efficient chloroplast translation. Plant Cell. 21(7):2058–2071. doi:10.1105/tpc.109.066654.

dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 32(17):5036–5044. doi:10.1093/nar/gkh834.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc Natl Acad Sci U S A. 96(8):4482–4487. doi:10.1073/pnas.96.8.4482.

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glemin S, Bierne N, Duret L. 2018. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and gc-biased gene conversion. Mol Biol Evol. 35(5):1092–1103. doi:10.1093/molbev/msy015.

Giege R, Eriani G. 2023. The tRNA identity landscape for aminoacylation and beyond. Nucleic Acids Res. 51(4):1528–1570. doi:10.1093/nar/gkad007.

Gilchrist MA, Chen WC, Shah P, Landerer CL, Zaretzki R. 2015. Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. Genome Biol Evol. 7(6):1559–1579. doi:10.1093/gbe/evv087.

Grantham R. 1978. Viral, prokaryote and eukaryote genes contrasted by mRNA sequence indexes. FEBS Lett. 95(1):1–11. doi:10.1016/0014-5793(78)80041-6.

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 32(18):2847–2849. doi:10.1093/bioinformatics/btw313.

Harrison MC, Ubbelohde EJ, LaBella AL, Opulente DA, Wolters JF, Zhou X, Shen XX, Groenewald M, Hittinger CT, Rokas A. 2024. Machine learning enables identification of an alternative yeast galactose utilization pathway. Proc Natl Acad Sci U S A. 121(18): e2315314121. doi:10.1073/pnas.2315314121.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2(1):13–34. doi:10.1093/oxfordjournals.molbev.a040335.

Jauhal AA, Newcomb RD. 2021. Assessing genome assembly quality prior to downstream analysis: n50 versus BUSCO. Mol Ecol Resour. 21(5):1416–1421. doi:10.1111/1755-0998.13364.

Kanehisa M, Goto S. 2000. Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28(1):27–30. doi:10.1093/nar/28.1.27.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059–3066. doi:10.1093/nar/gkf436.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. Nat Biotechnol. 37(8):907–915. doi:10.1038/s41587-019-0201-4.

Kollmar M, Mühlhausen S. 2017. Nuclear codon reassignments in the genomics era and mechanisms behind their evolution. Bioessays. 39(5):1600221. doi:10.1002/bies.201600221.

Krahn N, Fischer JT, Söll D. 2020. Naturally occurring tRNAs with non-canonical structures. Front Microbiol. 11:596914. doi:10.3389/fmicb.2020.596914.

Krassowski T, Coughlan AY, Shen XX, Zhou X, Kominek J, Opulente DA, Riley R, Grigoriev IV, Maheshwari N, Shields DC, *et al.* 2018. Evolutionary instability of CUG-leu in the genetic code of budding yeasts. Nat Commun. 9(1):1887. doi:10.1038/s41467-018-04374-7.

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. doi:10.1371/journal.pgen.1008304.

LaBella AL, Opulente DA, Steenwyk JL, Hittinger CT, Rokas A. 2021. Signatures of optimal codon usage in metabolic genes inform budding yeast ecology. PLoS Biol. 19(4):e3001185. doi:10.1371/journal.pbio.3001185.

Landerer C, Cope A, Zaretzki R, Gilchrist MA. 2018. Anacoda: analyzing codon data with Bayesian mixture models. Bioinformatics. 34(14):2496–2498. doi:10.1093/bioinformatics/bty138.

Lesecque Y, Mouchiroud D, Duret L. 2013. Gc-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. Mol Biol Evol. 30(6): 1409–1419. doi:10.1093/molbev/mst056.

Liaw A, Wiener M. 2002. Classification and regression by randomforest. R news. 2(3):18–22.

Lin BY, Chan PP, Lowe TM. 2019. tRNAviz: explore and visualize tRNA sequence features. Nucleic Acids Res. 47(W1):W542–W547. doi:10.1093/nar/gkz438.

Liu H, Huang J, Sun X, Li J, Hu Y, Yu L, Liti G, Tian D, Hurst LD, Yang S. 2018. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. Nat Ecol Evol. 2(1):164–173. doi:10.1038/s41559-017-0372-7.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, *et al.* 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A. 105(27):9272–9277. doi:10.1073/pnas.0803466105.

Madden T. 2013. The blast sequence analysis tool. The NCBI Handbook15 2(5):425–436.

Marck C, Grosjean H. 2002. tRNomics: analysis of tRNA genes from 50 genomes of eukarya, archaea, and bacteria reveals anticodonsparing strategies and domain-specific features. RNA. 8(10): 1189–1232. doi:10.1017/S1355838202022021.

McVean GA, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics. 157(1):245–257. doi:10.1093/genetics/157.1.245.

Muhlhausen S, Findeisen P, Plessmann U, Urlaub H, Kollmar M. 2016. A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. Genome Res. 26(7): 945–955. doi:10.1101/gr.200931.115.

Nalabothu RL, Fisher KJ, LaBella AL, Meyer TA, Opulente DA, Wolters JF, Rokas A, Hittinger CT. 2023. Codon optimization improves the prediction of xylose metabolism from gene content in budding yeasts. Mol Biol Evol. 40(6):msad111. doi:10.1093/molbev/msad111.

Opulente DA, LaBella AL, Harrison MC, Wolters JF, Liu C, Li Y, Kominek J, Steenwyk JL, Stoneman HR, VanDenAvond J, *et al.* 2024. Genomic factors shape carbon and nitrogen metabolic niche breadth across Saccharomycotina yeasts. Science. 384(6694):eadj4503. doi:10.1126/science.adj4503.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. 2013. The caper package: comparative analysis of phylogenetics and evolution in R. R package version. 5(2): 1–36.

Osawa S, Jukes TH. 1989. Codon reassignment (codon capture) in evolution. J Mol Evol. 28(4):271–278. doi:10.1007/BF02103422.

Paradis E, Schliep K. 2019. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35(3): 526–528. doi:10.1093/bioinformatics/bty633.

Pavlov YI, Newlon CS, Kunkel TA. 2002. Yeast origins establish a strand bias for replicational mutagenesis. Mol Cell. 10(1): 207–213. doi:10.1016/S1097-2765(02)00567-1.

Payea MJ, Hauke AC, De Zoysa T, Phizicky EM. 2020. Mutations in the anticodon stem of tRNA cause accumulation and met22-dependent decay of pre-tRNA in yeast. RNA. 26(1):29–43. doi:10.1261/rna.073155.119.

Petrov DA. 2002. Mutational equilibrium model of genome size evolution. Theor Popul Biol. 61(4):531–544. doi:10.1006/tpbi.2002.1605.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 12(1):32–42. doi:10.1038/nrg2899.

Radhakrishnan A, Chen YH, Martin S, Alhusaini N, Green R, Coller J. 2016. The dead-box protein dhh1p couples mRNA decay and translation by monitoring codon optimality. Cell. 167(1):122–132.e9. doi:10.1016/j.cell.2016.08.053.

Revell LJ. 2024. Phytools 2.0: an updated r ecosystem for phylogenetic comparative methods (and other things). PeerJ. 12:e16505. doi:10.7717/peerj.16505.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16(6):276–277. doi:10.1016/S0168-9525(00)02024-2.

Riley R, Haridas S, Wolfe KH, Lopes MR, Hittinger CT, Goker M, Salamov AA, Wisecaver JH, Long TM, Calvey CH, *et al.* 2016. Comparative genomics of biotechnologically important yeasts. Proc Natl Acad Sci U S A. 113(35):9882–9887. doi:10.1073/pnas.1603941113.

Santos MA, Tuite MF. 1995. The CUG codon is decoded in vivo as serine and not leucine in *Candida albicans*. Nucleic Acids Res. 23(9): 1481–1486. doi:10.1093/nar/23.9.1481.

Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14(13):5125–5143. doi:10.1093/nar/14.13.5125.

Shumate A, Wong B, Pertea G, Pertea M. 2022. Improved transcriptome assembly using a hybrid of long and short reads with stringtie. PLoS Comput Biol. 18(6):e1009730. doi:10.1371/journal.pcbi.1009730.

Steenwyk JL, Opulente DA, Kominek J, Shen XX, Zhou X, Labella AL, Bradley NP, Eichman BF, Cadez N, Libkind D, *et al.* 2019. Extensive loss of cell-cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts. PLoS Biol. 17(5):e3000255. doi:10.1371/journal.pbio.3000255.

Suzuki T. 2021. The expanding world of tRNA modifications and their disease relevance. Nat Rev Mol Cell Biol. 22(6):375–392. doi:10.1038/s41580-021-00342-0.

Wada M, Ito K. 2023. The CGA codon decoding through tRNA(arg) (ICG) supply governed by tad2/tad3 in *Saccharomyces cerevisiae*. FEBS J. 290(13):3480–3489. doi:10.1111/febs.16760.

Whibley A, Kelley JL, Narum SR. 2021. The changing face of genome assemblies: guidance on achieving high-quality reference genomes. Mol Ecol Resour. 21(3):641–652. doi:10.1111/1755-0998.13312.

Wint R, Salamov A, Grigoriev IV. 2022. Kingdom-wide analysis of fungal protein-coding and tRNA genes reveals conserved patterns of adaptive evolution. Mol Biol Evol. 39(2):2. doi:10.1093/molbev/msab372.

Wolf J, Gerber AP, Keller W. 2002. Tada, an essential tRNA-specific adenosine deaminase from *Escherichia coli*. EMBO J. 21(14): 3841–3851. doi:10.1093/emboj/cdf362.

Wolters JF, LaBella AL, Opulente DA, Rokas A, Hittinger CT. 2023. Mitochondrial genome diversity across the subphylum saccharomycotina. Front Microbiol. 14:1268944. doi:10.3389/fmicb.2023.1268944.

Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, Liu Y. 2015. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Mol Cell. 59(5):744–754. doi:10.1016/j.molcel.2015.07.018.

Zhao F, Zhou Z, Dang Y, Na H, Adam C, Lipzen A, Ng V, Grigoriev IV, Liu Y. 2021. Genome-wide role of codon usage on transcription and identification of potential regulators. Proc Natl Acad Sci U S A. 118(6):e2022590118. doi:10.1073/pnas.2022590118.

Zhou M, Wang T, Fu J, Xiao G, Liu Y. 2015. Nonoptimal codon usage influences protein structure in intrinsically disordered regions. Mol Microbiol. 97(5):974–987. doi:10.1111/mmi.13079.

Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci U S A. 111(22):E2310–E2318. doi:10.1073/pnas.1323011111

Editor: J. Comeron