Mapping the Inter- and Intra-genic Codon Usage Landscape in *Homo sapiens*

Maahil Arshad[1], Matthew Uchmanowicz[2], Vanshika Rana[2] and Muhammad Arshad Rafiq[2]

[1]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada; [2]Department of Biology, Algoma University, Sault Ste Marie, ON, Canada.

Abstract

The redundancy of the genetic code permits multiple codons to specify a single amino acid, but codon selection is far from random and reflects underlying functional significance. Most lifeforms exhibit the phenomenon of codon usage bias, where the usage of certain codons is favored over others and hence used more often throughout the genome. Codon bias is thought to provide a layer of post-transcriptional regulation that shapes mRNA stability, translation kinetics and co-translational protein folding. Despite extensive characterization of codon bias in bacteria, viruses and various model organisms, its prevalence and functional consequences remain understudied in humans. Accordingly, this study seeks to (1) characterize the genome-wide distribution of key codon-bias metrics in humans, (2) investigate how codon usage correlates with protein structural features, and (3) assess whether genes with exceptionally restricted codon repertoires are selectively constrained against synonymous mutations. Our study finds that codon usage bias is pronounced in structured protein domains relative to intrinsically disordered regions (IDRs) (Cliff's $\Delta = 0.26$; $p < 2.2 \times 10^{-16}$); that the most frequently used codons are supported by higher tRNA gene copy numbers ($r = 0.34$; $p < 0.05$); and that genes exhibiting extreme codon bias (high ΔENC) harbor significantly fewer synonymous variants than expected ($r = -0.28$, $p < 2.2 \times 10^{-16}$). Collectively, our findings position codon usage bias as a central, evolutionarily conserved regulator of translational efficiency, protein folding, and genomic integrity in humans.

This offers a new framework for interpreting synonymous variation and guiding future efforts in functional genomics.

Keywords: Genetic code; ENC value; Genetic codon bias; CUB; tAI

Declaration of Competing interest: All authors declare that there is no competing interest

The study was supported by AURF (Algoma University Research Funds) provided to Muhammad Arshad Rafiq.

Correspondence address: Muhammad Arshad Rafiq, Department of Biology, Algoma University, 1520 Queen Street E, Sault Ste. Marie, Ontario, P6A 2G4, Canada; arshad.rafiq@algomau.ca

Introduction

Earth hosts a remarkable diversity of life that have adapted to every imaginable environment, yet all share the same molecular foundation: the nucleotides of Deoxyribose Nucleic Acids (DNA). Four bases, adenine, thymine, cytosine, and guanine, combine in myriad sequences to encode the information required for development, growth and survival. Genetic information about proteins is conveyed and disseminated through sequences of three nucleotides or codons, commonly referred to as the triplet genetic code[1,2]. The triplet genetic code is characterized as, universal, non-overlapping and degenerate. Several species of Ribose Nucleic Acids (RNAs) are transcribed from DNA that carry genetic information required for diverse cellular functions. Among them, only messenger RNA (mRNA) serves as a template for protein synthesis requiring codons to encode amino acids, whereas other RNA species are non-coding, yet, perform various functions ranging from translational regulation to other structural or regulatory roles in the cell[3].

The universality of the genetic code, with a few exceptions in prokaryotic and mitochondrial genomes, across all forms of life is widely considered strong evidence for the existence of a single common ancestor[4]. The non-overlapping nature of the genetic code is thought to have evolved primarily to minimize the deleterious effects of transcriptional and translational errors, especially in the absence of RNA repair mechanisms[5]. In an overlapping code, a single mutation could affect multiple amino acids, resulting in a catastrophic disruption of protein structure and function. By contrast, the non-overlapping triplet code confines such errors to a single codon, preserving overall protein integrity and offering greater evolutionary flexibility[6].

Degeneracy refers to the phenomenon in which multiple codons encode the same amino acid. Apart from methionine and tryptophan, which are each specified by a single codon, all other amino acids are encoded by two to six codons[1,2]. Codon families are traditionally classified as F2, F3 (unique

to isoleucine), F4, and F6 based on the number of codons per amino acid. Leucine, serine, and arginine are each encoded by six codons that can be further subdivided into F2 and F4 groups due to differences in evolutionary conservation and tRNA recognition mechanisms[7]. In these cases, four codons share the same first two nucleotides (forming an F4 group), while the remaining two belong to a separate F2 group. This split reflects distinct evolutionary origins and differences in how transfer RNAs (tRNAs) decode them during translation. This layered redundancy allows the genetic code to buffer against mutations and protect proteins' native function. Codon usage bias within these families reflects both translational efficiency and evolutionary pressures[8].

Degeneracy likely evolved as a buffering mechanism against errors caused by spontaneous and endogenous mutations during DNA replication, transcription, and repair. It also provides a regulatory layer that enables differential codon usage, leading to codon bias, the preferential use of certain codons for a given amino acid over others[9].

Beyond general codon usage patterns, several molecular mechanisms contribute to codon bias. These include wobble base pairing at the third codon position, tRNA modifications, codon–tRNA pool correlations, and codon co-occurrence in relation to neighboring codons[2,6]. Codon families are also classified based on nucleotide composition at the third codon position: GC3 (guanine or cytosine) and AT3 (adenine or thymine). GC3-rich codons are associated with mRNA stability and efficient translation, whereas AT3-rich codons are linked to reduced transcript stability[8,10].

Codon usage bias significantly influences mRNA stability and translation efficiency. Studies have shown that optimal codons correlate with longer mRNA half-lives, while non-optimal codons lead to faster degradation. This is partly mediated by RNA-binding proteins and decay factors such as *Dhh1*, which detect elongation kinetics to regulate stability[11].

Codon bias exists both between genes (intergenic codon bias) and within individual genes (intragenic codon bias). Structured protein regions tend to use optimal codons, while intrinsically disordered regions (IDRs) favor non-optimal ones, potentially to modulate translation speed and ensure proper folding[12,13]. In addition, species-specific codon preferences further illustrate the evolutionary tuning of translational regulation[14,15].

Translation speed is influenced by multiple mechanisms that regulate ribosome movement along the mRNA. One well-characterized feature is the presence of 'codon ramps', an initial region enriched in rare codons at the 5′ end of transcripts, which is thought to reduce ribosomal traffic jams by spacing ribosomes early in translation[16,17]. Beyond ramping, translation speed is also modulated by mRNA secondary structures that stall ribosomes, local tRNA availability and aminoacylation levels, interactions with RNA-binding proteins, and the presence of regulatory upstream open reading frames (uORFs)[18,19]. Codon pair bias and non-optimal codon stretches also influence elongation rates and co-translational protein folding[20]. These mechanisms can be experimentally assessed using ribosome profiling, which maps ribosome positions genome-wide at codon resolution, revealing regions of translational pausing[21]. Additionally, reporter assays with engineered codon sequences, tRNA manipulation experiments, and SHAPE-seq or DMS-seq can be used to evaluate the effects of secondary structure and codon usage on translation kinetics[22].

Codon usage also influences transcription termination. In *Neurospora crassa*, rare codons can induce premature transcription termination by promoting cryptic polyadenylation signal formation[23]. Similar mechanisms have been observed in mammalian cells, suggesting evolutionary conservation of codon bias in transcriptional regulation.

Codon bias, once considered a neutral feature of redundancy, is now understood as a dynamic mechanism for regulating genes. It affects not only translation but also mRNA stability, folding,

and decay. Codon usage patterns are modulated by translational demand, tRNA availability, and developmental or stress-related cues[24].

Despite growing interest, comprehensive genome-wide analyses of codon usage bias in human protein-coding genes remain limited[28]. Most previous studies have focused on specific gene families or subsets. To address this gap, we performed a systematic computational analysis of codon bias across ~20,000 human protein-coding genes using various codon bias metrics. These include the Effective Number of Codons (ENC)**,** which quantifies overall codon usage bias (lower values indicate higher bias); the Codon Adaptation Index (CAI)**,** which reflects how well a gene's codons match those in highly expressed genes or used the most across transcripts; the tRNA Adaptation Index (tAI)**,** which assesses the compatibility between codon usage and the cellular tRNA pool or genome-wide copy numbers; Relative Synonymous Codon Usage (RSCU), a measure of how frequently a particular codon is used, relative to how often it would be used if all synonymous codons for the same amino acid were used equally; GC3 content, which measures the guanine-cytosine content at the third codon position and serves as a proxy for mutational pressure and codon preference[7,25–27].

Our analysis revealed several key findings: (1) a positive correlation between codon usage and tRNA copy number across the human genome, (2) evidence of intragenic codon bias distinguishing structured versus IDRs and (3) biased genes tend to harbor fewer synonymous mutations than expected alone from gene length and rate of synonymous mutations reported in humans. These findings suggest that codon bias may reflect evolutionary pressures that optimize translational kinetics and folding efficiency. This study lays a foundation for exploring codon usage as a regulatory layer, with potential implications for cancer biology, neurodevelopmental disorders, and translational control across diverse physiological contexts. While monogenic disorders are

becoming well understood, the complex architecture of polygenic disorders remain largely unknown. Our work does not seek to provide a complete solution to this long-standing challenge, but rather to act as an icebreaker, revealing the tip of a much larger regulatory iceberg, and to chart a path for future investigations into how codon usage may shape the molecular underpinnings of complex human phenotypes.

Methods

To investigate codon usage bias across the human transcriptome, we developed a comprehensive computational pipeline integrating high-quality genomic and transcriptomic resources. Canonical coding sequences were curated from Ensembl to calculate codon usage metrics, including RSCU, CAI, ENC and tAI using the cubar R package and custom Python scripts[28]. To assess codon preference, we incorporated tRNA gene copy numbers for calculating tAI were retrieved from GtRNAdb, and regional structural disorder annotations were integrated using IUPred3 predictions to differentiate between IDRs and structured domains[29,30].

This integrative approach enabled us to examine codon bias at both global and region-specific levels, incorporating translational efficiency, protein structure, and evolutionary constraint. Constraint scores from gnomAD were also included to explore the relationship between synonymous codon usage and mutational intolerance[31]. Together, this framework provides a scalable strategy to dissect the biological implications of codon usage across diverse molecular and evolutionary contexts.

Datasets

Sequencing data was obtained from Ensembl using release 112, GRCh38.p14 assembly. We retrieved the complete CDS FASTA files from each repository, restricted our analysis to transcript sequences for which a CDS was available. Canonical transcript definitions were

retrieved from Ensembl BioMart (release 112), and only transcripts explicitly marked as "canonical" by Ensembl were retained for downstream analyses. This filtering yielded 19 601 high-confidence protein-coding transcripts used throughout the study.

In addition, tRNA genes were obtained from the high-confidence set provided by gtRNAdb v2.0 (downloaded via https://gtrnadb.ucsc.edu/genomes/eukaryota/Hsapi38/hg38-tRNAs.tar.gz), yielding 428 distinct tRNA sequences with their corresponding copy numbers for codon–tRNA pairing analyses.

Gene-level constraint metrics (e.g., synonymous observed vs expected ratio, LOEUF) were obtained from gnomAD v4.1 by downloading gnomad.v4.1.constraint_metrics.tsv from the gnomAD public data release(https://storage.googleapis.com/gcp-public-data--gnomad/release/4.1/constraint/gnomad.v4.1.constraint_metrics.tsv). These metrics were joined to our Ensembl transcript list via stable IDs to compare patterns of synonymous mutation constraint against codon-bias measures.

Codon-Bias Metric Definitions

Most codon-bias metrics such as ENC, CAI, tAI, and GC3s were computed using the cubar (v1.1.0) package in R. Each metric follows the canonical implementation defined by cubar, based on widely used and accepted formulations for each.

RSCU was calculated for each of the 19, 601 protein-coding transcripts using a custom Python script. RSCU quantifies codon usage bias by comparing the observed frequency of a codon against the expected frequency if all synonymous codons for an amino acid were used equally[25]. The RSCU for codon j encoding amino acid i is defined as:

$$\mathbf{RSCU_{(ij)} = X_{(ij)} / ((1 / n_i) * \Sigma_j X_{(ij)})}$$

where $X_{(ij)}$ is the observed count of codon j for amino acid i and n_i is the number of synonymous codons encoding amino acid i. In implementation, each transcript's CDS was parsed codon-by-codon and codon counts were tallied. For each codon, the RSCU value was computed based on the total usage of all its synonymous codons across a transcript. To reduce noise, RSCU values were averaged per codon across all transcripts after removing outliers (defined as those falling outside 1.5× IQR per codon distribution). The resulting average RSCU values for each codon were used for downstream correlation analyses with human tRNA gene copy numbers.

To accurately evaluate codon usage bias across human transcripts, we calculated ΔENC to distinguish intrinsic codon bias from background nucleotide composition. The ENC is a widely used metric that quantifies the extent of codon usage bias in a gene, ranging from 20 (extremely biased) to 61 (no bias). However, ENC is inherently influenced by the GC3 content which can obscure whether observed codon preferences are driven by selection or mutational bias.

To decouple ENC from GC3 effects, we computed the expected ENC value based on the transcript's GC3 content using the established empirical relationship between ENC and GC3[7]. The deviation between the expected and observed ENC was defined as ΔENC:

$$\mathbf{\Delta ENC = ENC_expected\ (GC3) - ENC_observed}$$

where ENC_observed is the directly calculated ENC from a transcript's codon usage and ENC_expected (GC3) is the ENC predicted based on the transcript's GC3 content. This formulation originally developed by John Novembre, allows identification of transcripts whose

codon usage is more (or less) biased than expected from their GC3 composition alone[32]. A positive ΔENC indicates higher codon bias than expected (suggesting selective constraint or functional optimization), while a negative ΔENC suggests lower-than-expected codon bias (possibly due to relaxed selection or stochastic codon usage). By applying this correction across all 19, 601 transcripts, we could systematically explore codon usage patterns that reflect biological selection rather than mere nucleotide composition, enhancing the resolution of our codon bias analysis across the transcriptome.

The relationship between ENC and GC3 is described by Wright's theoretical curve, which provides the expected ENC value for a gene under the assumption that codon usage bias is solely due to GC3 content and not selective pressure. This serves as a baseline to determine whether a gene exhibits more or less codon bias than expected based on its nucleotide composition alone. Wright's curve is mathematically represented as:

$$\mathbf{ENC_expected = 2 + GC3 + 29 / (GC3^2 + (1 - GC3)^2)}$$

where GC3 is the fraction of guanine and cytosine at third codon positions. The constant 29 reflects the average contribution of codon family degeneracy under random codon choice constrained by nucleotide composition. Deviations of observed ENC from this neutral baseline reveal selective or regulatory influences beyond mutational bias, which we leverage to investigate codon usage bias. Wright's Curve enables the calculation of ENC_expected for any transcript, providing a reference point to compute ΔENC[7].

Programs

All analyses were performed on a MacBook Pro (arm64) running macOS 14.5 (23F79; Darwin 23.5.0) with Python 3.12.7 and R 4.4.1.

Ensembl protein-coding FASTA sequences were filtered via a custom R-based quality control pipeline to ensure that every sequence represented a complete open reading frame (ORF). Specifically, we verified that sequences initiated with a canonical AUG start codon, terminated with one of the three standard stop codons, and contained no premature termination codons. Sequences failing these criteria were discarded prior to downstream codon usage analyses. The remaining sequences were then loaded into R using Biostrings, and headers were parsed to extract the transcript ID, Ensembl gene ID, gene symbol, and GRCh38 genomic coordinates (chromosome, start, end, strand). These metadata were compiled into a unified table linking each sequence to its genomic context and used in all subsequent cubar analyses. Every header was parsed to retrieve the transcript ID, Ensembl gene ID, gene symbol, and GRCh38 genomic coordinates (chromosome, start, end, strand). The resulting fields were compiled into a common metadata table that links every sequence to its genomic context and accompanies all downstream **cubar** analyses. FASTA entry name was reset to the transcript ID to ensure a one-to-one link between sequence and metadata. Next, triplet frequencies were tabulated for every transcript against the standard nuclear codon table. The same counts were subsequently used to derive all downstream metrics. Finally, all metrics were merged with the transcript-level metadata to yield a final table containing transcript ID, gene ID, gene symbol, genomic coordinates, CAI, tAI, ENC and GC3s. This file represents a comprehensive codon-bias atlas for the human proteome and serves as the basis for all subsequent comparative analyses.

Coding DNA sequences (CDS) for all human protein-coding transcripts from GRCh38 (Ensembl release 112) served as the starting dataset. Each FASTA header was parsed to extract transcript (ENST) and gene (ENSG) identifiers; only the first instance of every unique ENST-ENSG pair was kept eliminating redundancy. CDS whose lengths were not divisible by three were removed, and the remaining sequences were translated in silico with the standard genetic code, truncating at the first in-frame stop codon. Proteins shorter than ten amino acids were discarded, and every surviving sequence was saved to an individual FASTA file named "ENST_ENSG.fa".

These protein FASTA files were analysed locally with IUPred3 (v3.2) in long mode, which assigns a disorder probability (0–1) to each residue. Residues scoring ≥ 0.50 were labelled intrinsically disordered (IDR), whereas those < 0.50 were considered structured.

For every transcript, nucleotide coordinates of IDR residues and of structured residues were gathered and concatenated into two separate sequences: one representing all disordered segments and the other all structured segments. Together they formed a composite FASTA containing exactly two entries per transcript. All downstream codon-usage analyses were performed in R 4.4.1 with the cubar package, exactly as described previously, to calculate codon bias metrics for IDRs and structured regions for each transcript.

Analyses were performed using the following packages (versions): cubar (1.1.0), Biostrings (2.72.1), data.table (1.16.2), dplyr (1.1.4), tidyverse (2.0.0), ggplot2 (3.5.1), ggpubr (0.6.0), GGally (2.2.1), patchwork (1.3.0), gghalves (0.1.4), ggtext (0.1.2), glue (1.8.0), rstatix (0.7.2), ppcor (1.1), and knitr (1.48). Supplementary processing in Python used biopython (1.85) and pandas (2.2.3).

Results

As established by the 1000 Genomes Project, a typical human genome varies from the reference genome at about 4.1 to 5 million sites, most of which are accounted by Single Nucleotide Polymorphisms (SNP)[33]. Most of these SNPs, occur in non-coding genomic regions as only 1% of the genome accounts for protein-coding regions. On average, this indicates that SNPs occur every 600-1000 nucleotides, in a total protein-coding region of 30 Mb, 30-50k SNPs can exist, and in ~20k protein-coding genes, each gene will harbor 2-3 SNPs. Moreover, an average of ~9600 synonymous variants per individual, suggests that each gene will have 0.5 synonymous variants[34]. Therefore, codon-usage profiles between individuals should not vary significantly and calculation of codon usage metrics from the reference genome should suffice for studying codon usage bias in humans. We chose protein-coding sequences from Ensembl to study codon usage metrics across the human genome, and all subsequent correlations analyses are Pearson-based.

Distribution of Codon Usage Metrics

Across ~20 000 Ensembl protein-coding genes (Fig. 1A), ENC values (blue) range from 33.21 to 57.72 (mean = 49.44), with a median of 50.65 and an interquartile range (IQR) of 46.48–52.96, reflecting moderate gene-to-gene variability in overall codon bias. GC3 content (gold) spans the widest interval of 0.16 to 1.00 (mean = 0.59; median = 0.59; IQR = 0.44–0.73), indicating substantial heterogeneity in third-position GC composition. In contrast, CAI (green) is tightly clustered between 0.56 and 0.96 (mean ≈ median = 0.80; IQR = 0.77–0.83), while tAI (orange) ranges from 0.19 to 0.46 (mean = 0.29; median = 0.29; IQR = 0.27–0.31), both showing comparatively constrained variation in codon-adaptation indices. The lower absolute values and narrow spread of tAI likely reflect incomplete annotation of human tRNAs: because tAI is

calculated from the set of codon-cognate tRNAs in the reference, any uncharacterized or low-abundance isoacceptors are omitted, leading to an underestimation of true translational adaptation. Plot of GC3 versus ENC for all genes and overlays Wright's neutral-expectation curve (red dashed line) (Fig. 1B). Most points fall below this curve, indicating stronger codon bias than GC3 alone predicts. A linear regression (blue line; R = – 0.75) confirms a robust inverse relationship: as GC3 increases, ENC decreases. Consequently, low ENC values can arise in part from elevated GC3 rather than from genuine translational selection, underscoring the need to correct for nucleotide composition when identifying genes with codon-usage bias independent of GC3.

Decoupling ENC from GC3

High GC3 composition inherently depresses ENC, thus apparent codon bias can be driven by nucleotide content rather than selection. Therefore, to detect genes biased independent of GC3 content, we used a metric called ΔENC, which can be calculated by considering expected ENC based on GC3 content (as per Wright's neutral-expectation curve) minus the observed ENC calculated directly from the sequence[32]. The distribution of ΔENC ranges from –16.21 to 20.69, with a median of 5.23, a mean of 4.94 and an interquartile range of 3.27–6.89; because positive ΔENC values indicate stronger codon bias than predicted by GC3 alone, these summary statistics show that most genes exhibit a modest excess of codon bias, whereas only a small subset (the negative tail) is less biased than expected (Fig. 2A).

We next flagged the 3% of genes with the lowest observed ENC (≤ 39.48; ~588 genes) as candidates for highly biased genes (Fig. 2B). Of these, 452 also have ΔENC > 0, indicating their bias cannot be explained by GC3 alone. Applying both thresholds yields a conservative set of highly biased genes likely under genuine translational selection. Panel C repeats the observed ENC distribution and highlights these ΔENC-positive genes in orange, illustrating how many of the low-ENC tail remain biased after GC3 correction (Fig. 2C).

Finally, plotting GC3 against ΔENC, Wright's neutral curve collapses to the ΔENC = 0 baseline because the GC3-driven component of codon bias has been subtracted out (Fig. 2D). Most genes nevertheless lie above this line, indicating additional bias beyond nucleotide composition, and the residual correlation is weak (r = – 0.28 compared with – 0.75 in Fig. 1B). This attenuation shows that GC3 now explains little of the remaining variation, making ΔENC a cleaner, composition-independent metric that more precisely captures true codon-usage bias.

Relationships between Codon-Bias Metrics

As codon–usage bias intensifies (lower ENC), genes gravitate toward optimal codons that are decoded by abundant tRNAs, so both CAI (r = – 0.62) and tAI (r = – 0.60) increase (Fig 3A). This inversely proportional relationship is expected because CAI is explicitly designed to reward codons that appear more often in a given set of sequences, while tAI weights codons by the genomic copy-number of their cognate tRNAs, a quantity still imperfectly measured in human due to extensive modifications and tissue-specific expression of tRNAs, lowering absolute tAI values[35].

When we correct ENC for its GC3 dependence (ΔENC), these links almost vanish (CAI r = 0.02; tAI r = - 0.17; Fig 3B). In other words, once base-composition bias is removed, there is little residual signal of extra translational selection captured by either CAI or tAI. This interpretation is reinforced by Fig 3C: GC3 alone explains the bulk of variation in both CAI (r = 0.76) and tAI (r = 0.80), mirroring earlier genome-wide reports that CAI and tAI track GC-rich codons[36].

ENC's theoretically expected relationship with GC3 comes from Wright's neutral curve, which allows us to compute ΔENC and disentangle true codon-usage bias from mere nucleotide composition. No comparable neutrality model exists for CAI or tAI: CAI is intrinsically tied to high-expression or set of sequences chosen, and tAI inherits any compositional skew in the tRNA gene repertoire, both factors make it difficult to generate a GC3-independent expectation for these

metrics. Consequently, downstream analyses will treat ΔENC as the primary GC3-corrected measure of codon bias, while CAI and tAI will be interpreted cautiously, acknowledging that their apparent adaptation largely reflects underlying GC3 enrichment rather than additional translational-selection pressure.

Intrinsically Disordered Regions and Codon-Usage Bias

Multiple studies have reported that structured protein regions exhibit stronger codon-usage bias than IDRs in various species, leading to the hypothesis that optimal codons may facilitate co-translational folding and stable tertiary structure[12,37]. To test this genome-wide in humans, we retrieved all human protein-coding transcripts from Ensembl and ran IUPred3 to score per-residue disorder (cutoff ≥0.5 to define IDRs). Within each transcript, we concatenated all IDR segments and all structured segments separately, then computed key codon-usage metrics (ENC, ΔENC, CAI, tAI and GC3) for each region class. This region-level design allowed a direct comparison of codon bias between disordered and ordered segments across the proteome (Fig. 4). After doing so, we found that IDRs illustrate a higher median ENC, thus weaker codon-usage bias than structured segments. In our dataset, IDRs have a median ENC of 52.79 (mean 52.69; interquartile range 50.29–55.40; range 34.41–61.00), whereas structured regions show a median ENC of 50.65 (mean 49.44; IQR 46.50–52.90; range 32.84–60.68). This shift is significant and small-to-moderate in magnitude (Cliff's $\Delta = 0.33$, 95 % CI 0.32–0.34; $p < 2.2 \times 10^{-16}$; Fig. 4A). After correcting for GC3 content, IDRs show a lower median ΔENC of 2.85 (mean 0.53; IQR –1.92 – 5.96; range –29.68 – 24.78) compared with structured regions (median 5.02; mean 4.45; IQR 2.79–6.83; range –29.68 – 20.05), with a small-to-moderate effect size (Cliff's $\Delta = 0.26$, 95 % CI 0.25–0.27; $p < 2.2 \times 10^{-16}$; Fig. 4B), confirming that codon-usage is more biased in structured regions and is not simply driven by nucleotide-composition. By contrast, while CAI, tAI and GC3 each reached

nominal significance by Wilcoxon rank-sum ($p < 2.2 \times 10^{-16}$;), the actual differences are relatively small once effect size is considered: Cliff's $\Delta = 0.07$ for CAI, 0.06 for tAI, and 0.05 for GC3 (all considered negligible by Cliff's criteria; Fig. 4C, S1A, S1B, S1C). This trend where even trivial shifts become significant in very large samples, confirms that translational-adaptation indices and base composition are effectively indistinguishable between IDRs and structured regions.

Linking Codon Usage Bias to Mutation Spectra, tRNA Availability & Stop-Codon Frequencies

Across ~20 000 human protein-coding genes we explored the extent of codon-usage bias, background mutation pressure, tRNA supply and termination choice intertwine (Fig. 5A–D). ΔENC was strongly and negatively associated with the ratio of observed-to-expected synonymous variants (O/E) ($r = -0.28$, $p < 2.2 \times 10^{-16}$; Fig. 5A). Thus, the more a gene departs from neutral codon usage, the more its preferred codons appear to be protected from silent changes, consistent with pervasive purifying selection at synonymous sites reported in recent genome-wide analyses[38]. By contrast, the per-site synonymous mutation rate (μ) rose only marginally with ΔENC ($r = 0.09$, $p < 2.2 \times 10^{-16}$; Fig. 5B), indicating that mutational rate alone cannot account for the depletion of silent variants in highly biased genes. Similarly, genes with stronger codon bias tend to have lower synonymous constraint scores (LOEUF), suggesting they are less tolerant to synonymous variants than less biased genes (Fig. 5D).

We next asked whether codon usage tracks the genomic supply of cognate tRNAs. Averaging RSCU for each codon across genes and plotting it against the copy number of its matching tRNA gene revealed a moderate positive correlation ($r = 0.34$, $p = 0.02$; Fig. 5C). This pattern mirrors classical codon–tRNA co-adaptation models, where selection favours codons translated by abundant tRNAs to maximise translational efficiency[13]. Finally, genome-wide stop-codon frequencies were markedly skewed: TGA accounted for almost half of all termination codons

(49.2%), whereas TAA (28.5%) and TAG (22.3%) were less common (Fig. 5D), consistent with previously reported literature[39,40]. Collectively, these observations highlight that codon bias, mutation constraints, tRNA availability and stop-codon choice are tightly coupled elements of translational optimisation in the human genome.

Discussion

The non-random usage of codons is a phenomenon that can be seen throughout the kingdom of life and was one of the first to be known at the advent of molecular biology. However, many of its plausible functions such as protein-folding, mRNA structure stability and translational efficiency, were among the hardest to study due to technological limitations of the time. Over the years, as technology improved, it became easier to correlate changes in protein structure to diseases. Any changes that spared protein structure, would inevitably spare protein function, hence coining such mutations as "silent", earning them lesser importance than mutations that alter protein structure, hence function[41]. Nevertheless, the 1000 Genomes Project revealed that most variation in humans tends to be found in non-coding regions and most plausibly affects gene regulation rather than protein structure[33]. Thus, the quantity and appropriate cellular distribution of proteins are equally important for protein function. This builds a compelling case that any genetic change that alters gene regulation can either drive disease or contribute to phenotypic diversity within a species. Likewise, we argue that codon usage is a means to regulate genes, and genetic changes that alter codon usage have the potential to dysregulate biological function.

Codon usage can be seen as a continuous measure because the extent to which a gene uses codons may reflect the specificity in translational dynamics, that do not need to be the same for all proteins to perform their respective functions. For example, stress-responsive genes such as heat shock proteins often exhibit codon usage patterns optimized for rapid translation under stress conditions, ensuring efficient protein production when cellular demand peaks. Similarly, ribosomal

proteins and other housekeeping genes, which are constitutively expressed at high levels, frequently display codon preferences aligned with abundant tRNAs, supporting stable and efficient protein synthesis. These examples illustrate how codon usage bias may encode regulatory information tuned to the functional and temporal requirements of different protein classes, reinforcing its role as a layer of translational control. The most reliable method to study codon usage is ΔENC, which is derived from decoupling ENC from GC3 content[32]. We can see that majority of the genes depict a positive ΔENC value, indicating that they are more biased or limited in their codon usage than that would be expected from GC3 content alone. The ΔENC calculation is inherently absolute, as it relies directly on the exact codon frequencies observed in each transcript minus the effect from GC3 content. Similarly, this trend is also evident from the ENC values, as most genes fall below Wright's curve (Fig. 1B).

In contrast, tAI can be computed in multiple ways but depends on comprehensive data, such as tissue-specific tRNA expression levels or the precise tRNA gene copy number across the genome. When such information is incomplete, as it is in our study, tAI scores may be underestimated and thus offer limited insight into true codon bias. Likewise, CAI can be derived either from a set of highly expressed genes or from overall codon frequencies in the transcriptome, with the most frequent codons deemed optimal. Although we adopted the transcriptome-wide approach, CAI remains tightly coupled to GC3 content; without an explicit neutrality model like Wright's curve for ENC, it's hard to disentangle mutational bias from selection. As such, CAI adds little beyond what GC3 alone reveals, and the same limitation would apply even if we had complete tRNA copy-number data. Yet, we observed a clear trend between tRNA gene copy number and the average RSCU value: as the cognate tRNA copy number increases for a given codon, its RSCU value also increases. Biologically, this is expected, since more frequently used codons tend to be supported by a larger pool of matching tRNAs.

Additionally, reports that codon usage is more biased in structured protein domains than IDRs hints at an interplay between codon usage and protein folding[42]. The more selective usage of codons in structured domains may be necessary for accurate protein folding, while IDRs may use various codons as certain protein regions may not need to assume a permanent structure. IDRs being less structurally constrained than folded domains, frequently exhibit reduced codon bias, consistent with relaxed translational selection. Such regions may therefore serve as evolutionary 'flexibility zones,' enabling sequence variation that contributes to intra-species diversity without compromising essential function, as seen in proteins like p53, which combines a structured DNA-binding domain with highly disordered regulatory termini[43]. Given that p53 and other cancer-associated proteins often contain extensive IDRs, our findings support the prediction that codon bias may play an underappreciated role in cancer biology by influencing the regulation of such structurally heterogeneous proteins. Consistent with our results, we see that structured protein regions have a greater median ΔENC suggesting that the codons used in these regions may aid with protein folding, as compared to IDRs that use a greater variety of codons. Although the effect size is small, it may nonetheless reflect underlying biology. Our study grouped all structured and IDR regions into two broad meta-segments. Yet both structured domains and IDRs comprise multiple subclasses, each of which may exhibit distinct codon-usage patterns[13]. Thus, future work should investigate codon usage across the various classes of structured and disordered regions. More rigidly structured segments will preferentially employ optimal codons as natural selection promotes efficient and accurate translation especially in regions critical for folding and function[44,45].

Next, we predicted that genes with the most constrained codon usage would exhibit fewer observed synonymous variants than expected, especially if codon choice is functionally important. Indeed, we observed that as ΔENC increases (indicating more restricted codon usage), the number

of reported synonymous variants falls below expectation. These genes also show greater constraint against synonymous mutations, reflected by lower synonymous LOEUF scores for these genes. In contrast, repeating this analysis with raw ENC values yields the opposite pattern, largely because ENC is influenced by GC3 content, and GC-rich regions are known to have higher mutation rates[46]. Moreover, overall synonymous mutation rates account for very little of the variance in ΔENC. Despite similar background synonymous mutation rates (syn mu), genes with more constrained codon usage accumulate fewer synonymous changes, strongly suggesting their functional importance. Future work should therefore investigate synonymous variants in these genes across different disorders. That being said, not every synonymous change in these genes will have an effect: the local codon context and surrounding sequence likely modulate any functional impact. Nonetheless, genes with high ΔENC are prime genome-wide candidates for prioritizing studies that link synonymous mutations to disease.

Finally, the predominance of TGA is well documented in mammalian genomes and is thought to reflect a combination of GC-rich context, differential read-through rates and historical substitution biases[39]. The predominance of TGA as a termination codon in our analysis aligns with previous findings that show TGA usage is enriched in newly evolved genes, possibly due to reduced translational termination signals on the complementary strand and selection-driven context effects that favor TGA over TAA or TAG[40].

Conclusion

In conclusion, codon usage is an important feature of the human genome due to its contribution in a number of biological processes. The more biased usage of codons in structured protein domains as compared to IDRs hints at its involvement in co-translational folding and may serve as evolutionary 'flexibility zones,' enabling sequence variation that contributes to intra-species diversity without compromising essential function The enrichment of tRNA copy

numbers for the most frequently used codons points to a co-evolution between codon usage and the tRNA gene pool, likely reflecting translational selection driven by yet-to-be-defined functional benefits associated with these codons. Most importantly, genes with strongly biased codon usage harbor fewer synonymous mutations than expected, indicating that codon choice is under selective constraint. This highlights the functional importance of synonymous sites and suggests that even silent changes can have evolutionary and biological consequences. Ultimately, our study demonstrates that codon usage bias is a multilayered determinant of translational efficiency, protein folding, plausibly contributing to intra-species diversity without compromising essential function, and evolutionary constraint in the human genome, highlighting its central role in gene regulation and its capacity to reveal functional impacts of 'silent' variation.

Our findings add to a growing appreciation that synonymous changes are not merely silent passengers in the genome but elements of a nuanced genetic language. The choice of codon, like selecting the right word in a sentence, can encode regulatory signals that influence translation dynamics, co-translational folding and protein expression, even when the amino acid sequence remains unchanged. In this view, codon usage acts as a layer of biological communication, embedding regulatory intent into a protein's genetic blueprint. These insights emphasize the need to revisit the so-called 'silent' portion of the genetic code, not as neutral noise, but as a subtle dialect shaping phenotypic diversity, disease susceptibility and evolutionary innovation, with particular relevance to cancer biology and complex polygenic disorders where small regulatory shifts can have outsized phenotypic consequences.

Figures

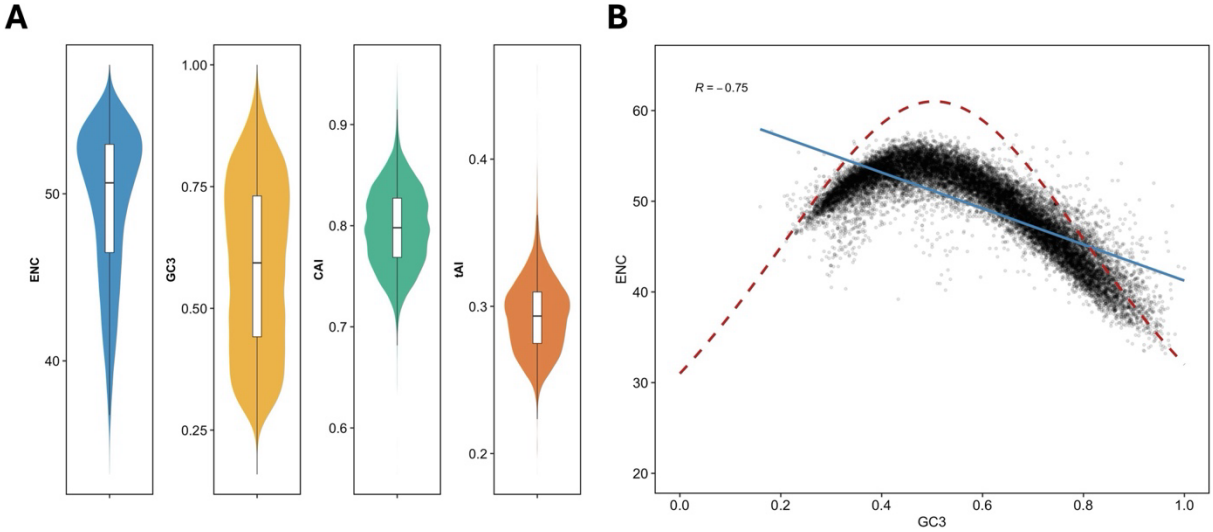

Figure 1. Codon-bias landscapes across the Ensembl protein-coding gene set. **(A)** Violin plots with overlaid boxplots showing the genome-wide distributions of ENC, GC3, CAI, tAI for ~20,000 human protein-coding genes. **(B)** Scatterplot of GC3 versus ENC for every gene, overlaid with Wright's neutral expectation curve (red dashed line) and a linear regression fit (blue line; R = –0.75). Most genes lie below the theoretical curve, indicating stronger codon bias than expected from GC3 alone and pointing to selective forces shaping codon usage.

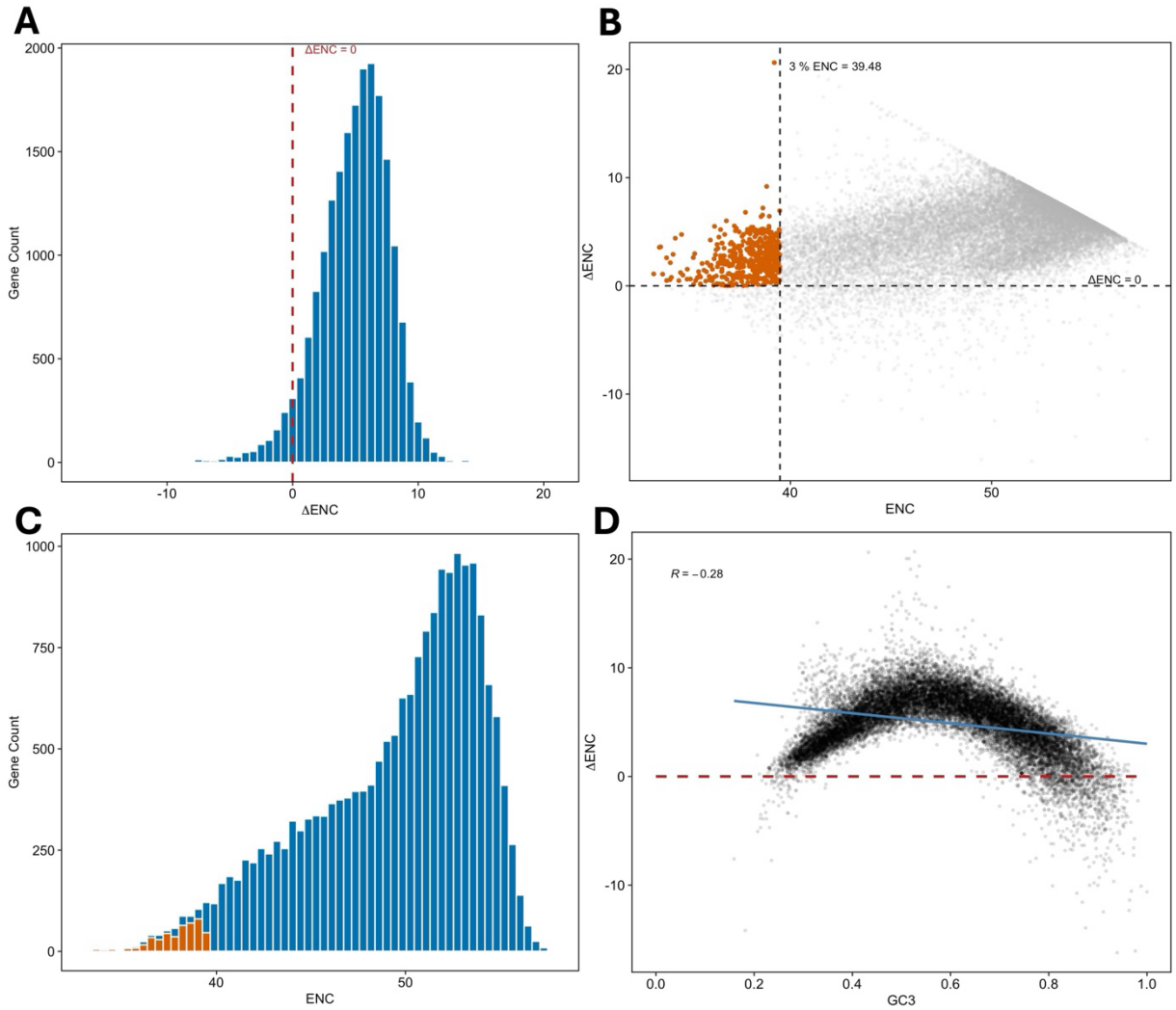

Figure 2. Genome-wide patterns of codon-bias deviation and selection of the most strongly biased genes. **(A)** Histogram of ΔENC, the difference between each gene's observed ENC and its expected ENC based on GC3 content, showing a right-skewed distribution across ~20,000 protein-coding genes; the red dashed line marks ΔENC = 0. **(B)** Scatterplot of ENC versus ΔENC for all genes (gray), with genes meeting both ΔENC > 0 and ENC ≤ 39.48 criteria highlighted in orange and demarcated by dashed horizontal (ΔENC=0) and vertical (ENC=39.48) lines, these represent the subset of genes under the strongest codon-usage selection. **(C)** Histogram shows the distribution of ENC values; genes highlighted in orange have both low ENC and high ΔENC, indicating codon bias independent of GC3 content. **(D)** Scatter of GC3 content versus ΔENC subtracting Wright's neutral expectation drives the entire theoretical curve to the red dashed ΔENC = 0 baseline, while the solid blue line shows the residual linear trend (R = –0.28).

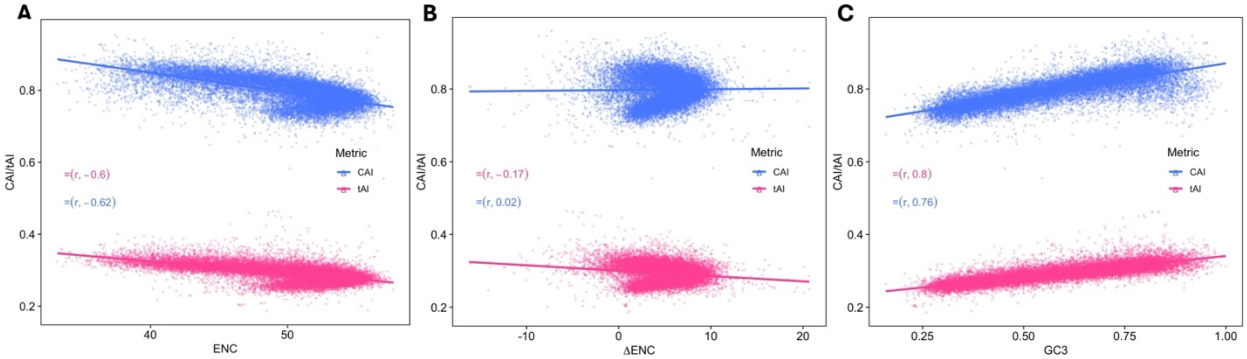

Figure 3. Assessing the effect of GC3, ENC, and ΔENC on translational efficiency metrics CAI and tAI. Gene-level scatterplots of codon adaptation index (CAI, blue) and tRNA adaptation index (tAI, pink) against three codon-bias measures: **(A)** ENC versus CAI and tAI, showing strong negative correlations (CAI: r = –0.62; tAI: r = –0.60), indicating that genes with fewer used codons exhibit higher translational adaptation. **(B)** ΔENC versus adaptation indices, revealing minimal association (CAI: r = 0.02; tAI: r = -0.17), which implies that selection on codon usage beyond GC3 content has little impact on these metrics. **(C)** GC3 versus CAI and tAI, with strong positive correlations (CAI: r = 0.76; tAI: r = 0.80), demonstrating that nucleotide composition is the primary driver of codon-adaptation variation.

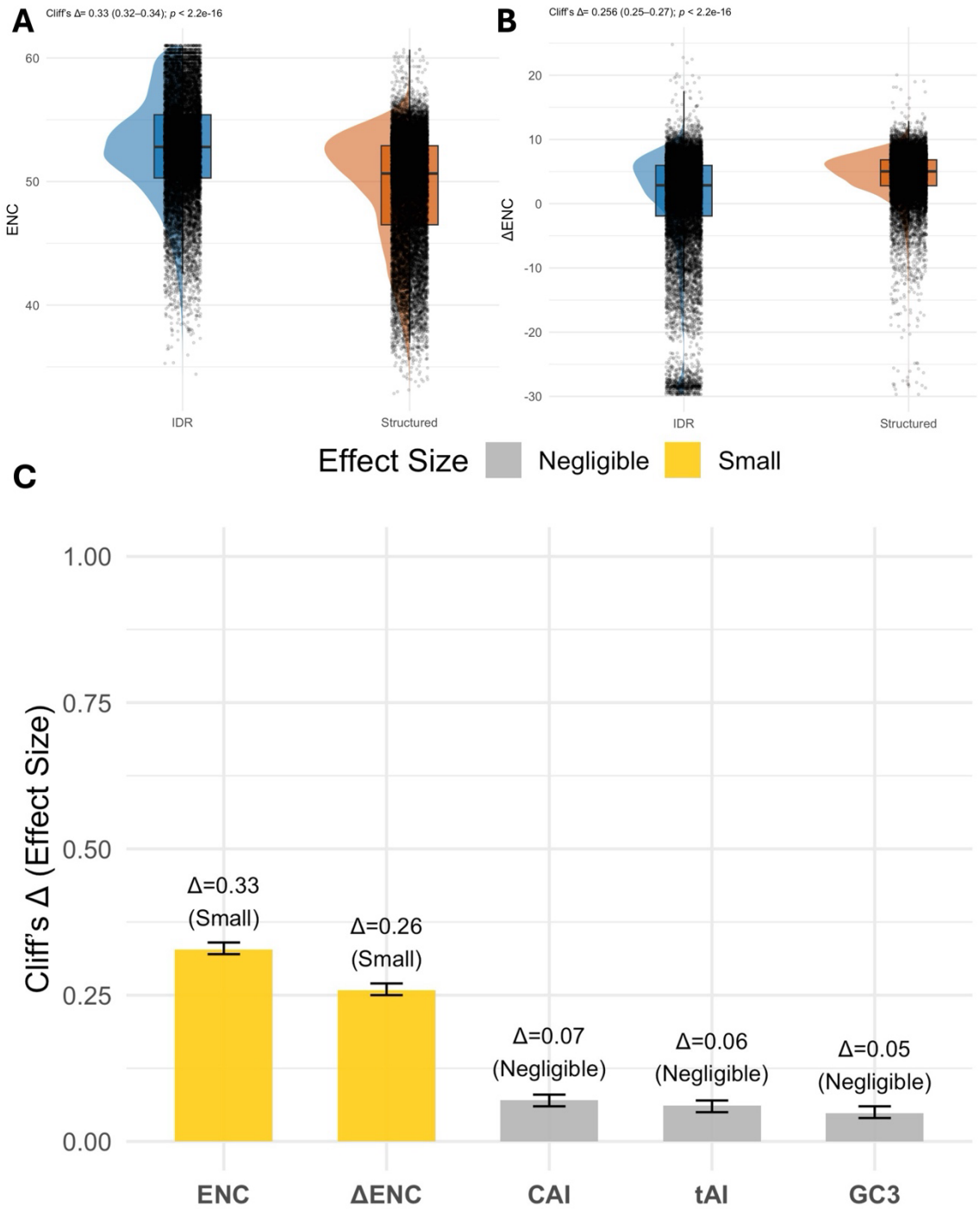

Figure 4. Codon-usage differences between IDRs and structured regions. **(A)** Violin/box plots of ENC show that IDRs (blue) have significantly higher ENC, thus, weaker codon usage bias than structured regions (orange). Statistical comparison by Cliff's Δ yields Δ = 0.33 (95% CI 0.32 – 0.34), $p < 2.2e-16$. **(B)** Violin plots of ΔENC reveal that IDRs exhibit significantly lower ΔENC than structured regions (Cliff's Δ = 0.26, 95% CI 0.25 – 0.27; $p < 2.2e-16$), demonstrating that even after accounting for GC3-driven bias, IDRs maintain weaker codon-usage bias than structured regions. **(C)** Bar chart of Cliff's Δ effect sizes (with 95% confidence intervals) for five metrics comparing IDRs to structured regions: ENC (Δ=0.33, small), ΔENC (Δ=0.26, small), CAI (Δ=0.07, negligible), tAI (Δ=0.06, negligible) and GC3 (Δ=0.05, negligible). Small effects are shown in yellow; negligible effects in gray.

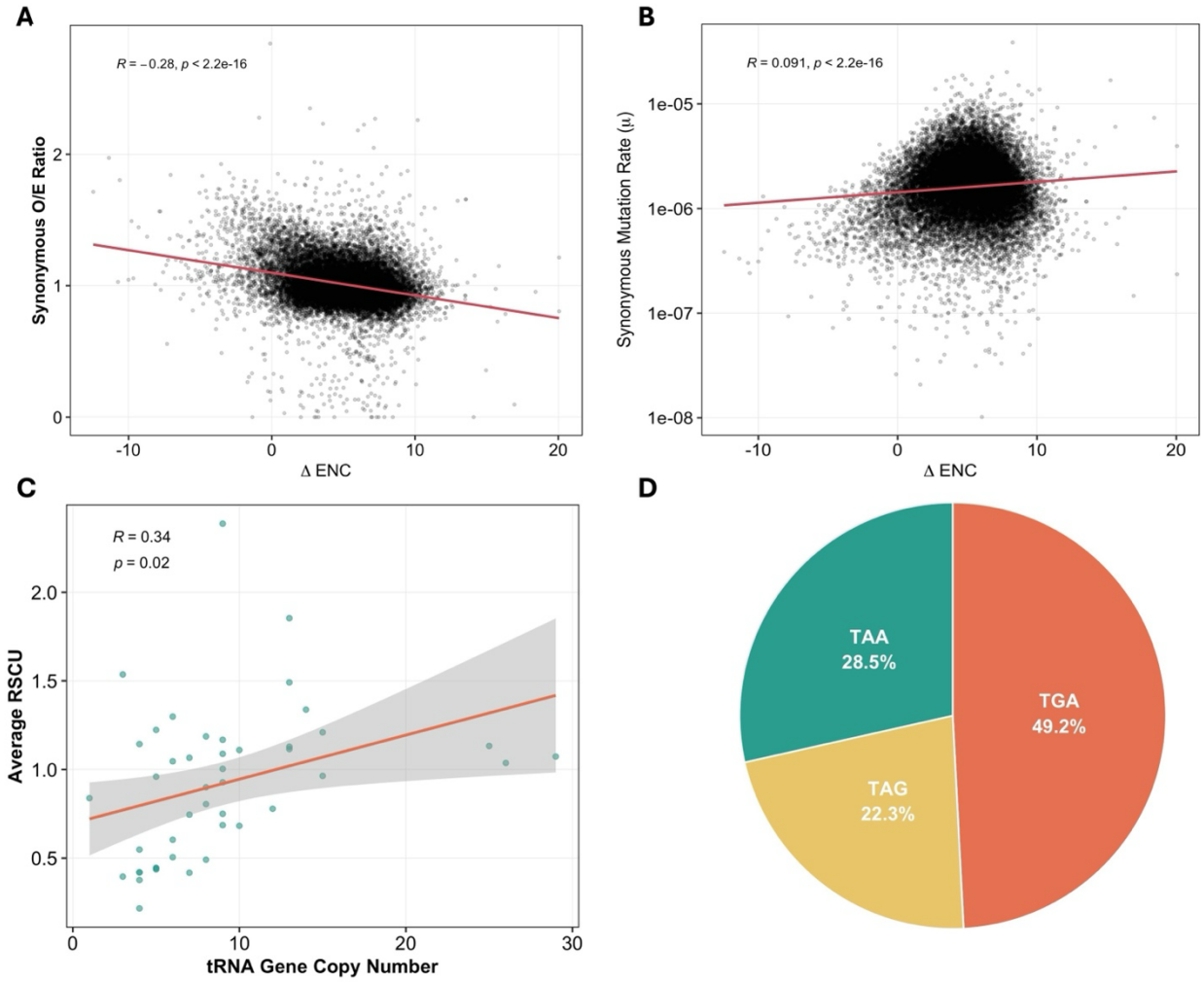

Figure 5. Codon-bias deviation, mutation patterns, tRNA abundance and stop-codon usage. **(A)** ΔENC versus the synonymous observed/expected (O/E) ratio, showing a significant negative correlation ($R = -0.28$, $p < 2.2 \times 10^{-16}$): genes under stronger codon bias (higher ΔENC) have fewer observed synonymous variants than expected. **(B)** ΔENC versus per-site synonymous mutation rate (μ), with a weak but significant positive correlation ($R = 0.091$, $p < 2.2 \times 10^{-16}$), indicating that codon-biased genes exhibit slightly higher underlying mutation rates. **(C)** Average RSCU per codon plotted against cognate tRNA gene copy number, revealing a moderate positive correlation ($R = 0.34$, $p = 0.02$): codons decoded by more abundant tRNAs are used more frequently. **(D)** Pie chart of genome-wide (~ 20 000 protein-coding genes) stop-codon frequencies: TGA is most common (49.2%), followed by TAA (28.5%) and TAG (22.3%).

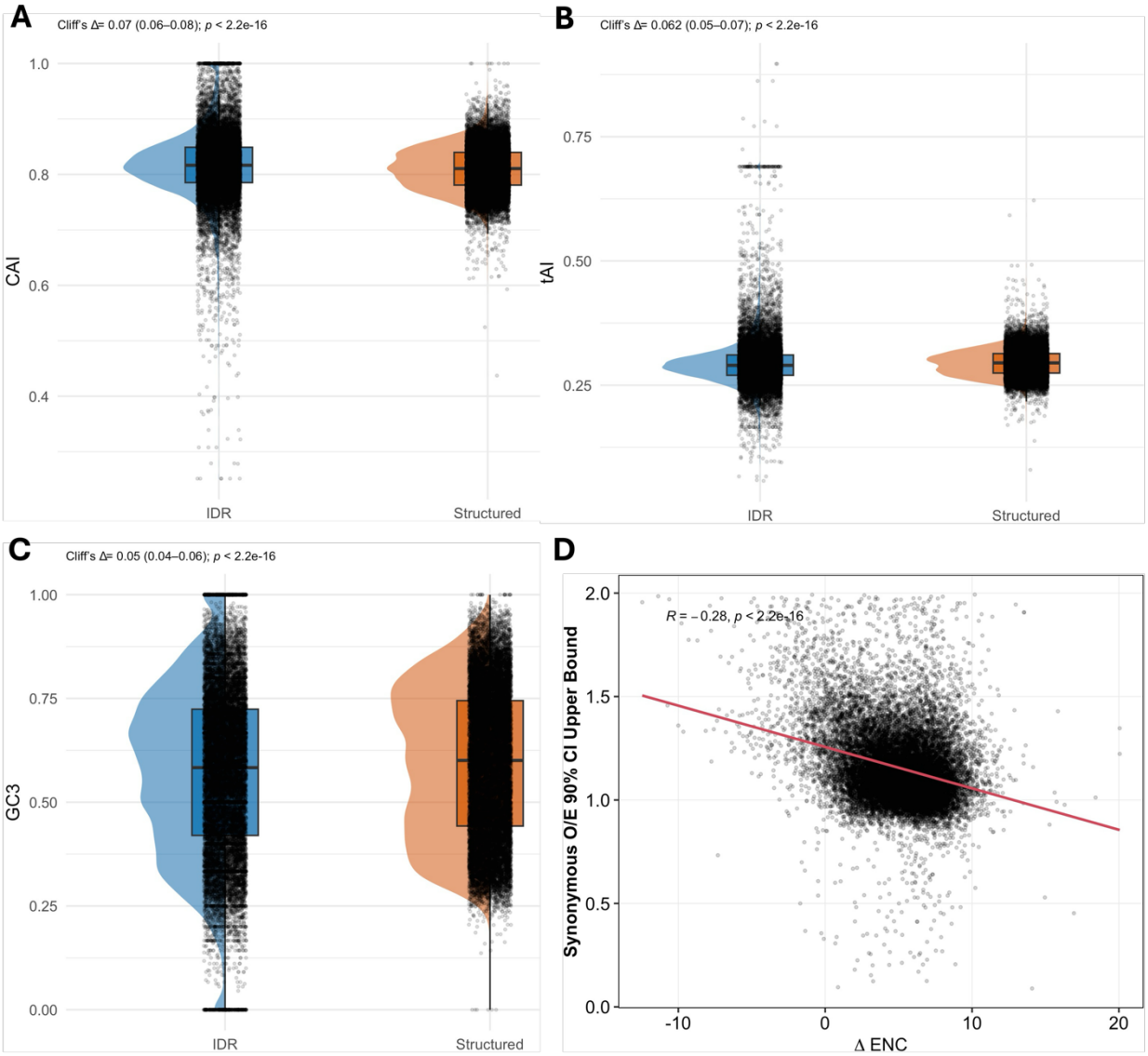

Figure S1. (A) Violin and box plots of CAI in IDRs (blue) versus structured regions (orange). Median CAI values are nearly identical, and although the difference is statistically significant, the effect size is negligible ($p < 2.2 \times 10^{-16}$; Cliff's $\Delta = 0.06$). **(B)** Violin and box plots of tAI show similar distributions in IDRs and structured regions ($p < 2.2 \times 10^{-16}$; Cliff's $\Delta = 0.05$). **(C)** Violin and box plots of GC3 reveal comparable GC3 content between IDRs and structured regions ($p < 2.2 \times 10^{-16}$; Cliff's $\Delta = 0.04$). **(D)** Scatterplot of ΔENC versus the upper bound of the 90% CI for synonymous O/E shows a negative correlation ($r = -0.28$; $p < 2.2 \times 10^{-16}$), indicating that genes with stronger codon bias tend to have lower synonymous constraint scores.

References

1. Crick, F. H. The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).
2. Nirenberg, M. W. & Matthaei, J. H. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **47**, 1588–1602 (1961).
3. Santosh, B., Varshney, A. & Yadava, P. K. Non-coding RNAs: biological functions and applications. *Cell Biochem. Funct.* **33**, 14–22 (2015).
4. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99–111 (2009).
5. Novozhilov, A. S. & Koonin, E. V. Exceptional error minimization in putative primordial genetic codes. *Biol. Direct* **4**, 44 (2009).
6. Crick, F. H., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961).
7. Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).
8. Parvathy, S. T., Udayasuriyan, V. & Bhadana, V. Codon usage bias. *Mol. Biol. Rep.* **49**, 539–565 (2022).
9. Athey, J. *et al.* A new and updated resource for codon usage tables. *BMC Bioinformatics* **18**, 391 (2017).
10. Wu, Q. & Bazzini, A. A. Translation and mRNA Stability Control. *Annu. Rev. Biochem.* **92**, 227–245 (2023).
11. Radhakrishnan, A. *et al.* The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell* **167**, 122-132.e9 (2016).

12. Homma, K., Noguchi, T. & Fukuchi, S. Codon usage is less optimized in eukaryotic gene segments encoding intrinsically disordered regions than in those encoding structural domains. *Nucleic Acids Res.* **44**, 10051–10061 (2016).

13. Liu, Y. A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun. Signal. CCS* **18**, 145 (2020).

14. Quax, T. E. F., Claassens, N. J., Söll, D. & van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol. Cell* **59**, 149–161 (2015).

15. Wu, X., Xu, M., Yang, J.-R. & Lu, J. Genome-wide impact of codon usage bias on translation optimization in Drosophila melanogaster. *Nat. Commun.* **15**, 8329 (2024).

16. Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354 (2010).

17. Fredrick, K. & Ibba, M. How the sequence of a gene can tune its translation. *Cell* **141**, 227–229 (2010).

18. Smith, T. J., Tardu, M., Khatri, H. R. & Koutmou, K. S. mRNA and tRNA modification states influence ribosome speed and frame maintenance during poly(lysine) peptide synthesis. *J. Biol. Chem.* **298**, 102039 (2022).

19. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7507–7512 (2009).

20. Liu, Y., Yang, Q. & Zhao, F. Synonymous but Not Silent: The Codon Usage Code for Gene Expression and Protein Folding. *Annu. Rev. Biochem.* **90**, 375–401 (2021).

21. Limbu, M. S., Xiong, T. & Wang, S. A review of Ribosome profiling and tools used in Ribo-seq data analysis. *Comput. Struct. Biotechnol. J.* **23**, 1912–1918 (2024).

22. Zubradt, M. *et al.* DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods* **14**, 75–82 (2017).

23. 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* **7**, e33569 (2018).

24. Wilusz, J. E. Controlling translation via modulation of tRNA levels. *Wiley Interdiscip. Rev. RNA* **6**, 453–470 (2015).

25. Sharp, P. M. & Li, W. H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).

26. dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).

27. Parvathy, S. T., Udayasuriyan, V. & Bhadana, V. Codon usage bias. *Mol. Biol. Rep.* **49**, 539–565 (2022).

28. Hong Zhang, Mengyue Liu, & Bu Zi. cubar: Codon Usage Bias Analysis. (2025).

29. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184-189 (2016).

30. Erdős, G., Pajkos, M. & Dosztányi, Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.* **49**, W297–W303 (2021).

31. Gudmundsson, S. *et al.* Variant interpretation using population databases: Lessons from gnomAD. *Hum. Mutat.* **43**, 1012–1030 (2022).

32. Novembre, J. A. Accounting for background nucleotide composition when measuring codon usage bias. *Mol. Biol. Evol.* **19**, 1390–1394 (2002).

33. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

34. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).

35. Yu, P. *et al.* Dynamic Landscapes of tRNA Transcriptomes and Translatomes in Diverse Mouse Tissues. *Genomics Proteomics Bioinformatics* **21**, 834–849 (2023).

36. Hia, F. *et al.* Codon bias confers stability to human mRNAs. *EMBO Rep.* **20**, e48220 (2019).

37. Zhou, M., Wang, T., Fu, J., Xiao, G. & Liu, Y. Nonoptimal codon usage influences protein structure in intrinsically disordered regions. *Mol. Microbiol.* **97**, 974–987 (2015).

38. Im, E.-H. & Choi, S. S. Synonymous Codon Usage Controls Various Molecular Aspects. *Genomics Inform.* **15**, 123–127 (2017).

39. Trexler, M., Bányai, L., Kerekes, K. & Patthy, L. Evolution of termination codons of proteins and the TAG-TGA paradox. *Sci. Rep.* **13**, 14294 (2023).

40. Okamura, K., Feuk, L., Marquès-Bonet, T., Navarro, A. & Scherer, S. W. Frequent appearance of novel protein-coding sequences by frameshift translation. *Genomics* **88**, 690–697 (2006).

41. Oelschlaeger, P. Molecular Mechanisms and the Significance of Synonymous Mutations. *Biomolecules* **14**, 132 (2024).

42. Moss, M. J., Chamness, L. M. & Clark, P. L. The Effects of Codon Usage on Protein Structure and Folding. *Annu. Rev. Biophys.* **53**, 87–108 (2024).

43. Bell, S., Klein, C., Müller, L., Hansen, S. & Buchner, J. p53 contains large unstructured regions in its native state. *J. Mol. Biol.* **322**, 917–927 (2002).

44. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* **20**, 237–243 (2013).

45. Hershberg, R. & Petrov, D. A. Selection on codon bias. *Annu. Rev. Genet.* **42**, 287–299 (2008).

46. Fryxell, K. J. & Moon, W.-J. CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.* **22**, 650–658 (2005).