Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Selection for Cheaper Amino Acids Drives Nucleotide Usage at the Start of Translation in Eukaryotic Genes

Na L. Gao[1,2,#], **Zilong He**[3,4,5,#], **Qianhui Zhu**[3,4,6,#], **Puzi Jiang**[1], **Songnian Hu**[3,4,6,*], **Wei-Hua Chen**[1,*]

[1] *Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China*
[2] *Institute for Computer Science and Cluster of Excellence on Plant Sciences, Heinrich Heine University, D-40225 Duesseldorf, Germany*
[3] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[4] *State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China*
[5] *Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Interdisciplinary Innovation Institute of Medicine and Engineering, Beihang University, Beijing 100191, China*
[6] *University of Chinese Academy of Sciences, Beijing 100049, China*

Received 12 July 2018; revised 30 May 2019; accepted 18 August 2019
Available online 17 March 2021

Handled by Jun Yu

Abstract Coding regions have complex interactions among multiple selective forces, which are manifested as biases in nucleotide composition. Previous studies have revealed a decreasing GC gradient from the 5′-end to 3′-end of coding regions in various organisms. We confirmed that this gradient is universal in eukaryotic genes, but the decrease only starts from the ~ 25th codon. This trend is mostly found in nonsynonymous (ns) sites at which the GC gradient is universal across the eukaryotic genome. Increased GC contents at ns sites result in cheaper amino acids, indicating a universal selection for energy efficiency toward the N-termini of encoded proteins. Within a genome, the decreasing GC gradient is intensified from lowly to highly expressed genes (more and more protein products), further supporting this hypothesis. This reveals a conserved selective constraint for cheaper amino acids at the translation start that drives the increased GC contents at ns sites. Elevated GC contents can facilitate transcription but result in a more stable local secondary structure around the start codon and subsequently impede translation initiation. Conversely, the GC gradients at four-fold and two-fold synonymous sites vary across species. They could decrease or increase, suggesting different constraints acting at the GC contents of different codon sites in different species. This study reveals that the overall GC contents at the translation start are consequences of complex interactions among several major biological processes that shape the nucleotide sequences, especially efficient energy usage.

KEYWORDS Macroevolution; Prioritization of selective forces; Energy efficiency; Transcription; Translation initiation

Introduction

GC content is one of the most important aspects of

*Corresponding authors.
E-mail: weihuachen@hust.edu.cn (Chen WH), husn@im.ac.cn (Hu S).
#Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.
https://doi.org/10.1016/j.gpb.2021.03.002

nucleotide sequence composition, and it has been extensively studied [1–6]. Recently, more details of the functional importance of GC content have been revealed. For example, due to an intrinsic trade-off in the codon table, mRNA templates that consist of more energetically costly nucleotides often encode cheaper amino acids. Since GC are energetically more expensive than AT, GC content has become an important indicator for the relative amounts of cellular resources invested in making nucleotides *versus* amino acids [7,8]. Increased GC contents at nonsynonymous (ns) sites in the coding region increase the costs of the mRNA template but decrease the costs of encoded amino acids. Selection for amino acids is prioritized because protein synthesis is at the downstream end of the information amplification process [7]. It is unsurprising that selection for the usage of cheaper amino acids in highly expressed genes always results in higher GC contents, especially at ns sites.

Higher GC contents at synonymous sites could increase the mRNA levels of mammalian genes [9]. This effect is not due to different translation rates of GC-rich and GC-poor mRNAs but due to more efficient transcription or mRNA processing [9]. These results are consistent with the fact that GC contents in coding regions are always higher than in intergenic regions of eukaryotes, while transcribed noncoding RNAs have intermediate GC contents. Furthermore, the results also suggest that different codon sites may be constrained by different selective forces, with some being constrained by multiple forces.

Increased GC contents in coding regions are not always beneficial. In both bacteria and eukaryotes, less stable local secondary structures (LSSs) with higher minimal folding energy (MFE) are preferred at the translation start in order to facilitate translation initiation [10−14], which is the most important speed-limiting step of protein translation [15]. Random sequences with increased GC contents often generate more stable LSSs, because the GC pair is bound by three hydrogen bonds, while the AT/AU pair is bound by only two. Increased GC contents at the translation start may impede efficient translation initiation by forming more stable LSSs. In human cells, mRNAs with less stable LSSs around the translation start have a higher protein/mRNA ratio, suggesting the role of LLSs in translation efficiency. In addition, it is known that eIF4A, an evolutionarily conserved eukaryotic initiation factor-4A family that is required for the binding of mRNA to 40S ribosomal subunits, consists of several helicases that function to unwind double-stranded RNA [16,17].

Selection for less stable LSSs in bacteria is manifested as decreased local GC contents centering at the start codon [10], consequently resulting in the use of rare codons in bacteria that prefer high GC contents at their synonymous sites (*i.e.*, bacteria with overall GC contents higher than ~ 45%) and the use of frequent codons in bacteria that prefer low GC contents [10]. Similar results have been observed in eukaryotes [18]. The use of rare codons may lead to a significant increase in decoding time in yeast [19], likely due to lower concentrations of the corresponding tRNAs [20]. Codon usage significantly contributes to translational efficiency. Selection for translational efficiency has contributed significantly to codon usage and has subsequently been manifested as nucleotide usage biases. Translation appears to typically be limited by initiation [21].

Previous studies have identified several selective forces that constrain the sequence composition in the coding region. These include efficiencies for energy usage, transcription or mRNA processing, and translation initiation. One region or even one class of codon sites could be simultaneously constrained by multiple selective forces. It is therefore interesting to study the complex interactions among these constraints in eukaryotes to determine how they shaped the sequences at both nucleotide and protein levels, and how they are prioritized in different eukaryotes.

Results and discussion

Universal decreasing GC gradients along the coding region

In the majority of the 259 eukaryotic genomes studied, we observed a negative GC gradient at the 5′-ends of the coding sequences (CDSs). The GC content was higher at the translation start and then decreased along the direction of translation (**Figure 1**A, **Figure 2**). In the selected species shown in Figure 1, this negative gradient is apparent at the first 100 codons and is mostly found at the ns sites, which account for ~ 60% of the CDS (see Materials and methods). To quantify the extents of the gradients, for each species we applied linear regression to the first 100 codons and obtained slopes for the CDSs, as well as the ns, two-fold synonymous (2s), and four-fold synonymous (4s) sites. A slope value indicates the GC content change per 100 codons (%). We removed the first 25 codons from the calculation because they were constrained by various selective forces [10,18]. As shown in Figure 2, the slopes of the CDSs are negative in the majority of the species studied.

A previous study has suggested that certain signals such as tRNA-adaptation index (tAI) and selection for weak RNA folding could extend beyond the 25th codon [22]. However, for the GC content, it is evident that the first 25 codons are different from the downstream codons, especially on ns sites (Figure 1B). We found that the decreasing GC gradient could be further extended to 150 codons downstream of the translation start (Figure S3).

Some amino acids, including phenylalanine, valine, threonine, tryptophan, methionine, leucine, isoleucine,

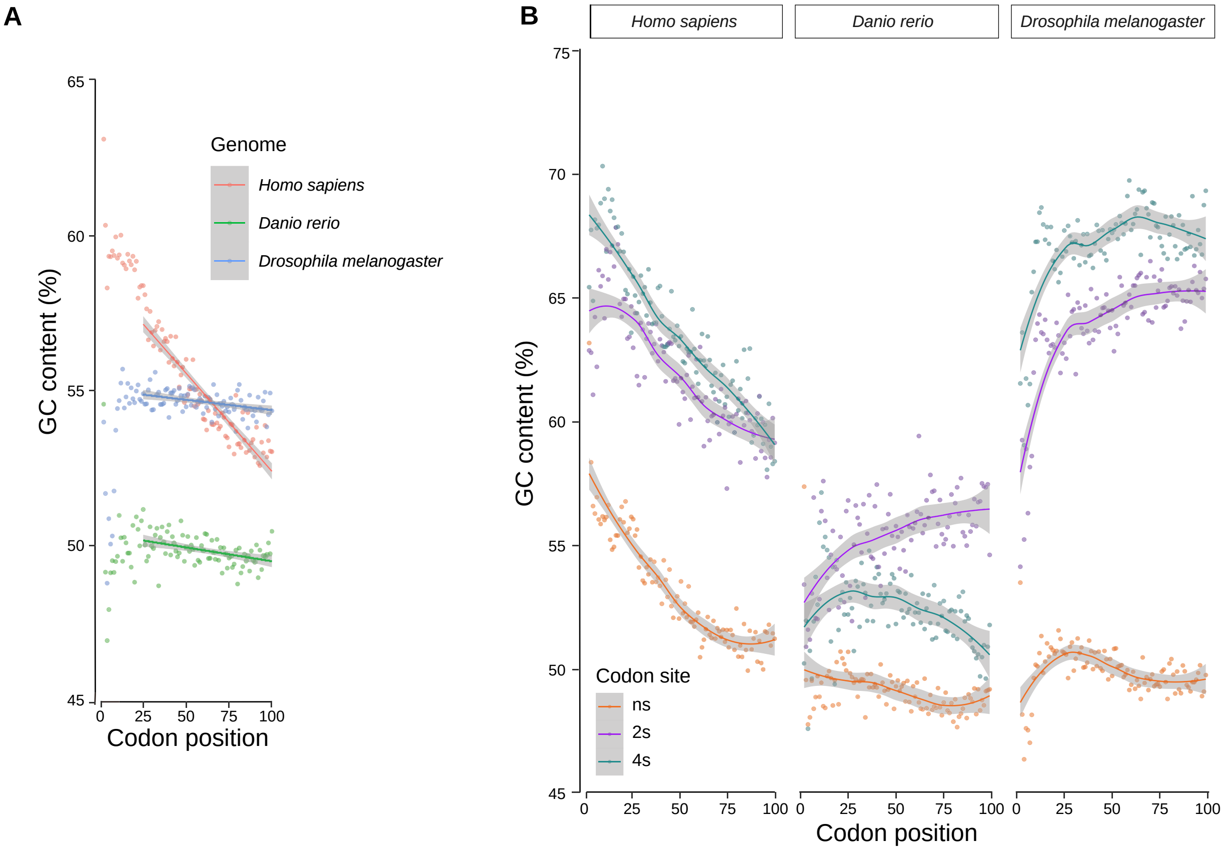

Figure 1 GC gradients for the first 100 codons of the coding regions in selected species
A. Overall GC contents of the first 100 codons for *Homo sapiens* (red), *Danio rerio* (green), and *Drosophila melanogaster* (blue). Each dot represents the average GC content of the codons at the corresponding codon position across all valid genes of a species (see the "Identification of GC gradients in CDSs" section in Materials and methods for details). The linear regression lines were obtained using the 26th–100th codons; the first 25 codons were excluded because they were subjected to constraints for less stable LSSs. **B.** Averaged GC contents of the first 100 codons as a function of codon positions, separated by codon sites. LOESS lines were plotted to better illustrate the general trends. The results for other parts (middle and end) of the coding region are shown in Figure S1. CDS, coding sequence; LSS, local secondary structure; LOESS, locally estimated scatterplot smooth; ns, nonsynonymous; 2s, two-fold synonymous; 4s, four-fold synonymous.

lysine, and histidine, cannot be synthesized from other substrates and thus must be supplied in the diet. They are often referred as to essential amino acids (data obtained from https://en.wikipedia.org/wiki/Essential_amino_acid, accessed on Oct 6, 2018). We found similar GC gradients at the translation start in both essential and non-essential amino acids (Figure S4), suggesting that essential amino acids are not a unique contributor to our observations.

Distinct GC gradients at different codon sites across species indicate multiple selective constraints

By separating codons into codon sites according to their degeneracy, we found distinct patterns of GC contents at these sites in different species. Negative slopes at the ns sites were found in most species. At ns sites, the first parts of the CDSs have the highest GC contents compared with the remainder of the CDSs. G and C are energetically more expensive than A and T/U, respectively, but mRNA templates with high GC contents code for energetically cheaper amino acids [7]. From the perspective of energy efficiency, the use of cheaper amino acids has higher priority than the use of cheaper nucleotides, because protein synthesis is at the end of the information path from DNA to mRNA to protein. This is known as the central dogma of genetics: at each step of the central dogma, the genetic information amplifies and consumes more energy/cellular resources than the previous step. Therefore, higher GC contents at the starting parts of the CDSs may represent an evolutionarily conserved mechanism for energy efficiency. The negative GC gradient from the translation start is thus likely a conserved means to account for premature ribosome drop-offs during translation elongation [23,24], which should impact the protein N-termini more often than downstream. Consistently, we found that the cheaper amino acids are preferably used at the translation start (**Figure 3**).

We estimated that for each protein produced in human, an average of ~ 30 ATPs could be saved in the first 100 codons due to the positive slope (shown in Figure 3A) compared with the 100 codons in the middle. The estimated

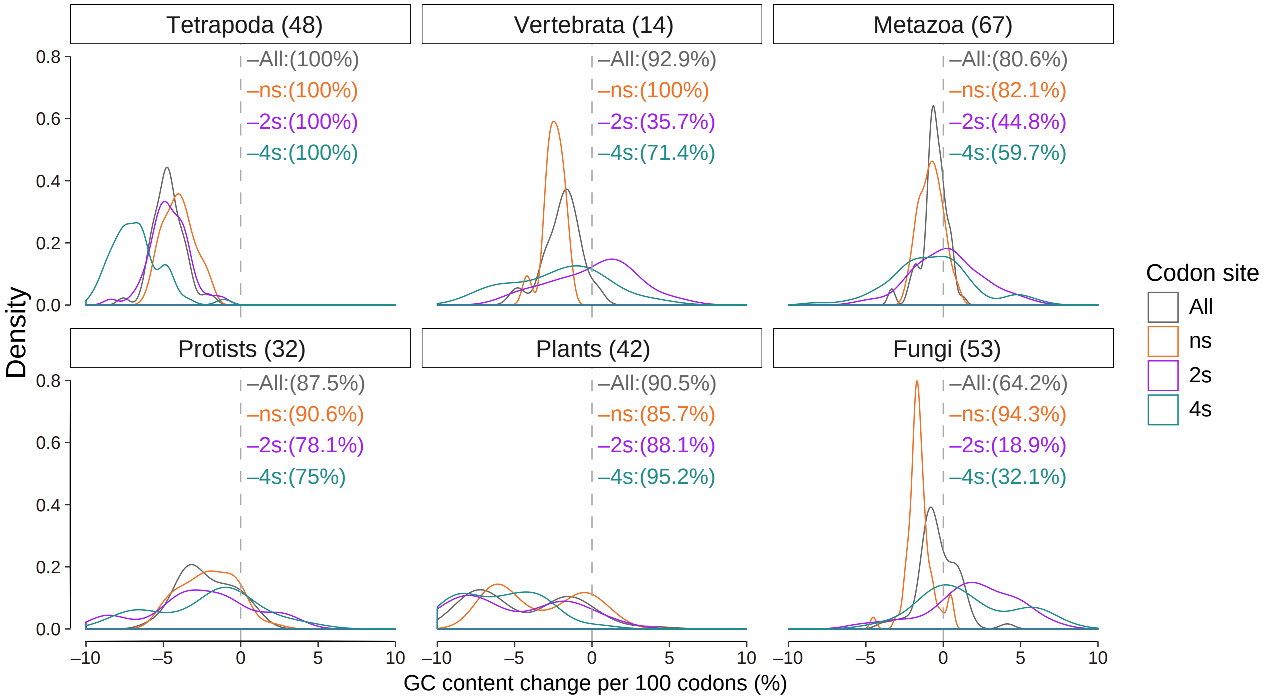

Figure 2 Decreasing GC gradients at the translation start is universal across eukaryotic species

The X-axis shows the GC content change per 100 codons in the starting part of the coding region that spans 100 codons from the start codon. Species are grouped into broad taxonomic groups. The number of species in each group is indicated in the parenthesis next to the group name. The results for GC gradients in the middle and end parts of the coding region are shown in Figure S2.

number of saved ATPs can be calculated using the following equation:

$$\frac{slope \times n}{2} \times 100$$

where *slope* refers to the slope calculated by applying a linear regression model to the first 26–100 codons and *n* refers to the number of codons used in the linear regression model.

Positively charged amino acids are not randomly distributed around the start codon and the following coding region [25,26]. These amino acids could cause ribosome pausing and consequently increase ribosome drop-off rates [27]. In addition, proline could also stop peptide bond formation and stalling translation [28,29]. To test whether the avoidance of those amino acids could explain the negative GC gradient at ns sites, we removed those amino acids from our analysis. Removing the positively charged amino acids and proline did not change our results (Figure S5).

Conversely, GC gradients at 2s and 4s sites are often species- and taxon-specific (Figure 2). Depending on the species, GC gradients at these codon sites could show either similar (Figure 1B, right panel) or distinct patterns (Figure 1B, middle panel). These results also indicate that GC-biased gene conversion (gBGC), which is believed to be the underlying mechanism for the GC gradients in angiosperm genomes [30,31], is insufficient to explain our observations in other species. The gBGC refers to a meiotic repair bias that favors G and C over A and T alleles in high-recombining genomic regions and is believed to be a major contributor to genomic GC evolution [32–35]. Because the A-to-G and C-to-T changes at 2s sites often do not change encoded amino acids, the gBGC model predicts similar overall GC content patterns at 2s and 4s sites, which is not true in most of the vertebrate species (Figure 2). In addition, there is no evidence for gBGC in *Drosophila melanogaster* [36]. This is consistent with the fact that coding genes on the sex chromosomes in *Drosophila* have higher GC contents than autosomal genes, which contradicts the prediction of gBGC. Overall, our results suggest that, independent of gBGC, there are additional selective forces that increase the GC contents in coding regions, including selection for transcription efficiency that increases with higher GC contents at synonymous sites [9]. These forces are likely to be universal, as the GC contents of coding regions of all eukaryotes are higher than those in intergenic regions. However, species-specific constraints must exist to shape the different patterns at 2s and 4s sites in different species (Figures 1 and 2).

GC gradients at ns sites are stronger in genomes with higher coding GC contents

We found a significant negative correlation between the slopes of GC gradients at the ns sites and the overall coding GC contents in five out of six broad taxonomic groups

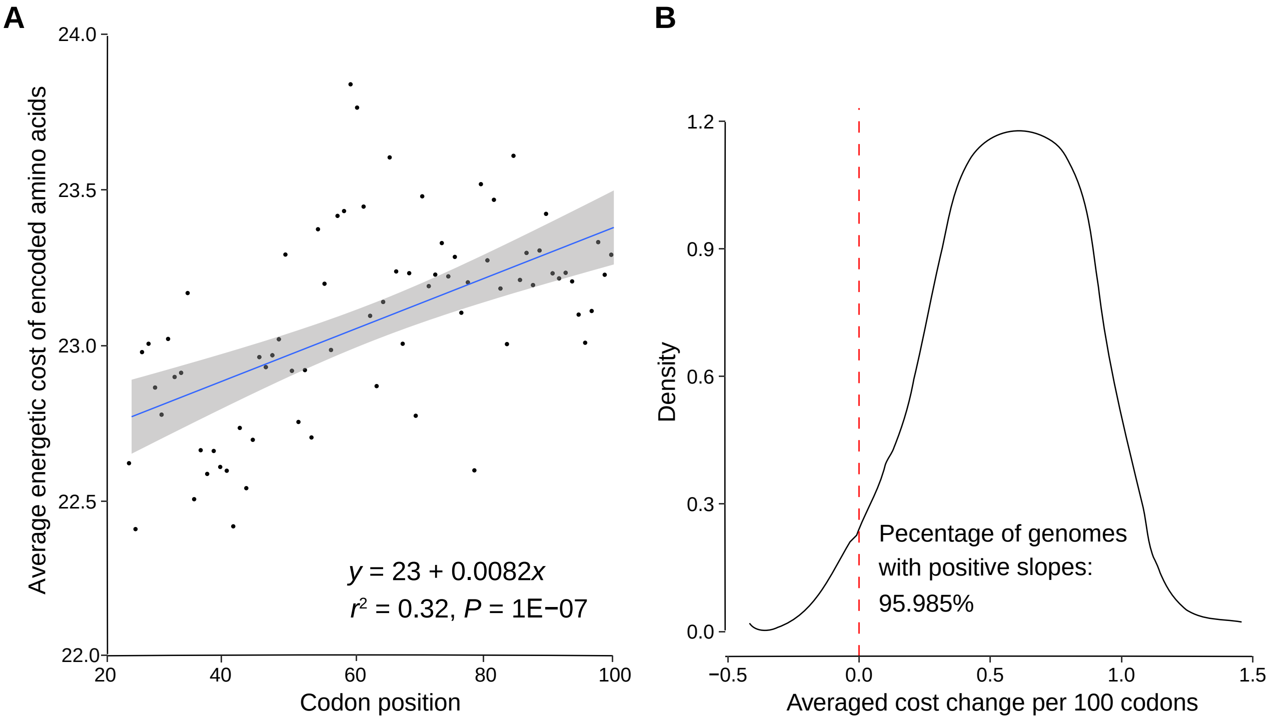

Figure 3 **Energetically cheaper amino acids are used at the N-termini of proteins**
A. An example of the trend in *Homo sapiens*. Each dot represents the average energetic cost of encoded amino acids at the corresponding codon position (from the 26th codon to the 100th codon) across all valid human genes. The blue line represents LOESS over the dots. Linear regression was applied to the 26th codon to the 100th codon of the starting part to describe the relationship between the energetic costs and codon positions. As indicated in the equation at the bottom-right corner, a positive slope value (0.0082 in the equation) indicates that cheaper amino acids are preferably used toward the start codon, and the costs increase along the direction of translation. **B.** Density plot summarizing the distribution of slope values in the 259 eukaryotic genomes.

(**Figure 4**). Applying linear regression to these correlations, we found that about 22% (tetrapoda) to 59% (plants) of variation in the slopes of the GC gradients could be explained by the overall coding GC contents. "Vertebrata" was the only exception, possibly due to the fact that few species (14) were available in Ensembl [37] for this group (see Material and methods). Overall, coding GC contents explained ~ 30% of the variation in the slopes of GC gradients in the 259 species investigated here (Figure S6).

The correlations have two interesting characteristics. First, the slopes of the linear regression (indicated by the solid lines) are different among groups, possibly due to their adaptation to unique environments. Second, the GC content of 40% seems to be the threshold for positive *versus* negative GC gradients. Higher GC contents code for cheaper amino acids; hence, higher overall coding GC content may represent a stronger selective constraint for energy efficiency in the corresponding genomes. Therefore, the higher the overall coding GC content, the higher is the need for energy efficiency and the stronger the negative GC gradient at the translation start.

GC gradients in eukaryotes have been described in plants such as Gramineae species (grasses) [31], angiosperms (flowering plants) [30], and seed plants [38]. Research has focused on inter- and intra-species variations [30,38]; however, global factors affecting the strengths of GC gradients have not been discussed. We found that at ns, 2s, and 4s sites as well as coding regions as a whole, the slopes of the GC gradients at the first 100 codons correlate significantly in eudicot and monocot plants (Figure S7). Overall, coding GC content in a genome could explain 93% of the strength (slopes) of the GC gradients.

Negative GC gradients are increasingly stronger from metazoan to vertebrata to tetrapoda species

We found increasingly stronger (more negative) GC gradients from metazoan to vertebrata to tetrapoda species. As shown in Figure 2, the peaks of the density plots at ns sites shifted from −1 in the metazoa group (GC content decreased 1% per 100 codons) to −2.5 in the vertebrata group and further to −5 in tetrapoda. The 4s sites and the overall CDSs showed similar trends, but the 2s sites did not.

The mRNAs with higher GC contents at the translation start may form more stable secondary structures. This would consequently impede translation initiation, which is the most important speed-limiting step of translation. A less stable secondary structure around the translation start is preferred by bacteria [10,11] as well as human [12]. For example, in human, the translation initiation efficiency, as

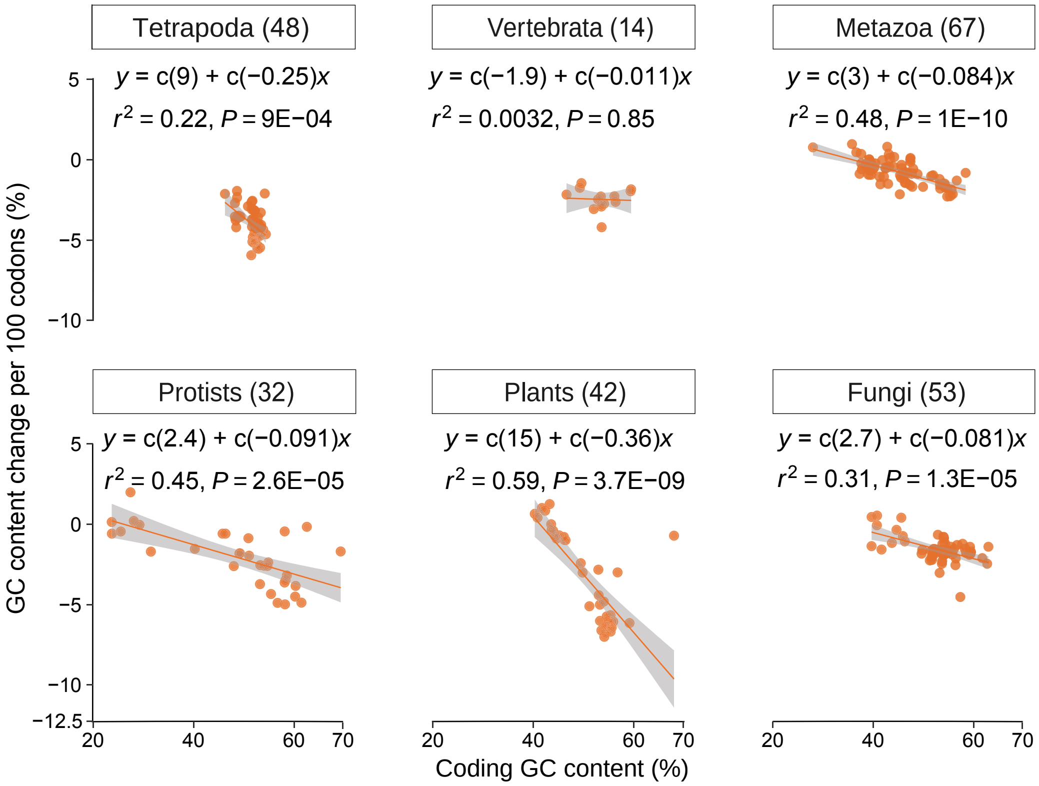

Figure 4 GC gradients at ns sites of the start codon correlate with the overall coding GC content

Each dot represents a genome. Species are grouped into broad taxonomic groups (Table S1). The number of species in each group is indicated in the parenthesis next to the group name.

indicated by the protein/mRNA abundance ratio, correlates significantly with the MFE of the sequence in a 39-bp window near the start codon [12]. More importantly, the less stable the translation start compared to the downstream coding region, the higher the translation initiation efficiency [12]. These results are consistent with our observations that despite the trend of higher GC contents toward the translation start, up to 25 codons after the start codon often showed lower GC contents, although the decrease was weakened in metazoan compared to vertebrata and tetrapoda species (Figure 1).

Stronger GC gradients and higher GC contents at the translation start often mean more stable LSSs for the transcribed mRNAs. We found that the LSSs around the start codon, as measured by MEF, were increasingly stronger from *D. melanogaster* to *Danio rerio* to *Homo sapiens* (**Figure 5**). For example, in *D. melanogaster*, the LSSs around the start codon were less stable as compared with downstream coding regions (Figure 5C). In *D. rerio*, the differences still existed but were less significant (Figure 5B). In *H. sapiens* (Figure 5A) as well as in the mouse (Figure S8), the LSSs around the start codon were more stable than the immediate downstream coding regions (regions from about the 10th codon to the 20th codon). These results indicate that the two selective constraints that shape the coding MEFs, namely the selection efficiencies for translation initiation and GC content, are prioritized in different species with the former being deprioritized in higher animals.

More highly expressed genes often have higher GC contents, likely due to the fact that: 1) higher GC contents at ns sites correspond to cheaper amino acids, and 2) higher GC contents at synonymous sites correspond to higher transcription efficiency in eukaryotes [9]. It is possible that highly expressed genes tend to have more stable LSSs around the start codons. In selected species within different taxonomic groups, this appears to be the case and this trend is more obvious in higher animals such as *D. rerio* and *H. sapiens* (Figure 5).

Conclusion

We describe a universal trend of elevated GC content toward the start codon that manifests as a decreasing GC gradient from the 5′-end to the 3′-end of the coding region. This trend is mostly contributed by the ns sites, which consist of ~ 60% of all codon sites. The GC gradients at the ns sites are universal across most of the 259 eukaryotic genomes studied. Increased GC contents at ns sites result in cheaper amino acids, indicating a universal selection for

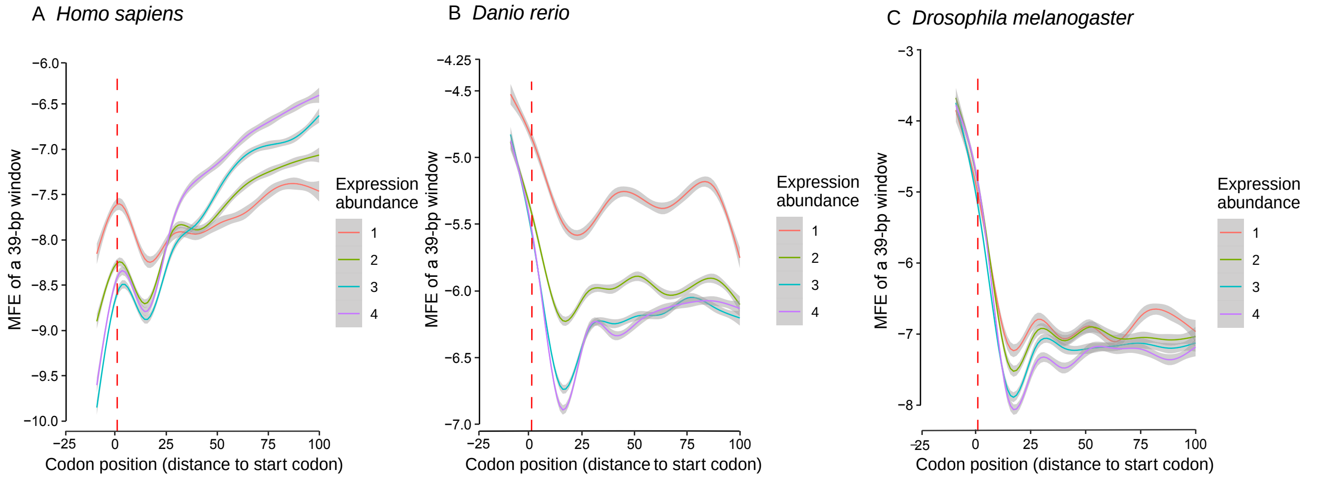

Figure 5 Highly expressed genes and genes in higher species tend to form more stable LSSs at the translation start
Shown here are the results for selected model organisms, including *Homo sapiens* (**A**), *Danio rerio* (**B**), and *Drosophila melanogaster* (**C**). The MFE values were calculated using a 39-bp window, with a codon at the center, plus 18-bp flanking on each side. This window moved along CDSs with a step of 3 bp; it may extend to the 5′-UTR regions. Genes were grouped into equal-sized bins according to their expression abundances in normal tissues: 1 to 4, from lowest to highest. MFE, minimal folding energy.

energy efficiency at the start of translation. Conversely, the GC gradients at 4s and 2s sites could go in either direction, suggesting different constraints acting at the GC gradients of different codon sites. Within the genome of each species, the GC gradient is intensified from lowly to highly expressed genes. Across species, it also is intensified from metazoan to vertebrata to tetrapoda. We have revealed a conserved selective constraint for cheaper amino acids at the translation start that has resulted in increased GC contents at the ns sites. Elevated GC contents facilitate transcription but result in more stable LSSs around the coding start and subsequently impede translation initiation. This study thus reveals universally increased GC contents at the translation starts of ns sites that are driven by selective constraints in eukaryotic genomes.

Materials and methods

Data

cDNA sequences and tabular annotation files in the GFF format for the 259 eukaryotic genomes were downloaded from the Ensembl ftp sites [39]. There are two ftp sites at Ensembl; the main site (ftp.ensembl.org) hosts more frequently used model organisms such as human, mouse, and fruit fly genomes, while the other site (ftp.ensemblgenomes.org) hosts other genomes including fungi, metazoa, plants, and protists. The two sites use different version numbers. We downloaded all available eukaryotic genomes from the release 87 from the main site and the release 33 from the other site. One gene could have multiple cDNAs. In this study, only the one with the longest cDNA was examined.

CDSs and UTR sequences were separated from the downloaded cDNA sequences according to their coordinates in the GFF files using in-house Perl scripts, followed by manual inspection. Unless noted otherwise, a minimal CDS length of 900 bp was required; CDSs shorter than 900 bp were removed from further analyses.

Gene expression data were obtained from the "Baseline experiments" section of Expression Atlas [40], which provides baseline gene expression for different tissues and cell types from selected species.

The amino acid production costs were obtained from Akashi and colleagues [41], and the nucleotide production costs were obtained from Chen and colleagues [7]. These costs refer to *de novo* synthesis costs, *i.e.*, the energetic costs of these information molecules when synthesized from scratch. Detailed information is available in Table S1. Eukaryotic genomes investigated in this study, broad taxonomic groups of their species, and genomic and coding GC contents are listed in Table S2.

Separating CDSs into different fragments and codon sites

Three fragments, including the starting, middle, and end fragments, were taken from each CDS. Each fragment contained 100 codons (300 bp) in length. Sequences between these fragments, if available, were discarded from further analyses.

Each fragment was initially split into codons, and the three nucleotides in each codon were then classified into 4s sites, 2s sites, and ns sites according to the standard codon table.

Identification of GC gradients in CDSs

Codons were numbered from 1 to 100 according to their positions in the fragments. An average GC content value was calculated for all codons at a certain position across all valid CDSs of a genome. The same was performed for each codon site.

GC gradients were identified by linear regression analysis using an "lm()" function in R, with a formula of "y~x", where "x" indicates the codon positions ranging from 1 to 100, and 'y' indicates the averaged GC contents. A "slope" value would be obtained from the analysis, describing the changes in GC contents per 100 codons; for example, a "slope" value of −5 means that, on average, the GC content decreases 5% for every 100 codons along the direction of translation.

A negative "slope" value indicates decreasing GC content, while a positive "slope" value indicates increasing GC content along the direction of translation.

Calculation of MFE

RNAfold of the ViennaRNA package [42] was used with default parameters to calculate the MFE of an input sequence.

Unless otherwise noted, the input sequences were 39 bp in length. When the sliding window technique was used, a step size of 3 bp was chosen.

CRediT author statement

Na L. Gao: Methodology, Software, Formal analysis, Visualization, Writing - original draft. **Zilong He:** Methodology, Software, Formal analysis, Writing - original draft. **Qianhui Zhu:** Software, Formal analysis, Visualization. **Puzi Jiang:** Resources, Data curation. **Songnian Hu:** Conceptualization, Supervision, Writing - review & editing. **Wei-Hua Chen:** Conceptualization, Supervision, Writing - review & editing. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgments

We thank all members of the Wei-Hua Chen and Songnian Hu groups. This work was partially supported by the National Key R&D Program of China (Grant No. 2018YFC0910500).

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.03.002.

ORCID

0000-0002-9095-0987 (Na L. Gao)
0000-0002-7111-7770 (Zilong He)
0000-0002-2506-5535 (Qianhui Zhu)
0000-0001-5872-9943 (Puzi Jiang)
0000-0003-3966-3111 (Songnian Hu)
0000-0001-5160-4398 (Wei-Hua Chen)

References

[1] Foerstner KU, von Mering C, Hooper SD, Bork P. Environments shape the nucleotide composition of genomes. EMBO Rep 2005;6:1208–13.

[2] Yakovchuk P, Protozanova E, Frank-Kamenetskii MD. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res 2006;34:564–74.

[3] Hurst LD, Merchant AR. High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc Biol Sci 2001;268:493–7.

[4] Pozzoli U, Menozzi G, Fumagalli M, Cereda M, Comi GP, Cagliani R, et al. Both selective and neutral processes drive GC content evolution in the human genome. BMC Evol Biol 2008;8:99.

[5] Birdsell JA. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol 2002;19:1181–97.

[6] Romiguier J, Ranwez V, Douzery EJP, Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res 2010;20:1001–9.

[7] Chen WH, Lu G, Bork P, Hu S, Lercher MJ. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. Nat Commun 2016;7:11334.

[8] Bohlin J, Brynildsrud O, Vesth T, Skjerve E, Ussery DW. Amino acid usage is asymmetrically biased in AT- and GC-rich microbial genomes. PLoS One 2013;8:e69878.

[9] Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol 2006;4:e180.

[10] Bentele K, Saffert P, Rauscher R, Ignatova Z, Blüthgen N. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol 2013;9:675.

[11] Goodman DB, Church GM, Kosuri S. Causes and effects of N-terminal codon bias in bacterial genes. Science 2013;342:475–9.

[12] Vogel C, de Sousa Abreu R, Ko D, Le SY, Shapiro BA, Burns SC, et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol 2010;6:400.

[13] Gu W, Zhou T, Wilke CO. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol 2010;6:e1000664.

[14] Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A 2010;107:3645–50.

[15] Jackson RJ, Hellen CUT, Pestova TV. The mechanism of eukaryotic translation initiation and principles of its regulation. Nat Rev Mol Cell Biol 2010;11:113–27.

[16] Rogers GW Jr, Komar AA, Merrick WC. eIF4A: the godfather of the DEAD box helicases. Prog Nucleic Acid Res Mol Biol 2002; 72:307−31.

[17] Schütz P, Bumann M, Oberholzer AE, Bieniossek C, Trachsel H, Altmann M, et al. Crystal structure of the yeast eIF4A-eIF4G

complex: an RNA-helicase controlled by protein–protein interactions. Proc Natl Acad Sci U S A 2008;105:9564–9.

[18] Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, et al. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 2010;141:344–54.

[19] Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B. Measurement of average decoding rates of the 61 sense codons *in vivo*. Elife 2014;3:e03735.

[20] Dana A, Tuller T. The effect of tRNA levels on decoding times of mRNA codons. Nucleic Acids Res 2014;42:9171–81.

[21] Jacques N, Dreyfus M. Translation initiation in *Escherichia coli*: old and new questions. Mol Microbiol 1990;4:1063–7.

[22] Tuller T, Zur H. Multiple roles of the coding sequence 5′ end in gene expression regulation. Nucleic Acids Res 2015;43:13–28.

[23] Sin C, Chiarugi D, Valleriani A. Quantitative assessment of ribosome drop-off in *E. coli*. Nucleic Acids Res 2016;44:2528–37.

[24] Zhang G, Fedyunin I, Miekley O, Valleriani A, Moura A, Ignatova Z. Global and local depletion of ternary complex limits translational elongation. Nucleic Acids Res 2010;38:4778–87.

[25] Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol 2011;12:R110.

[26] Charneski CA, Hurst LD. Positively charged residues are the major determinants of ribosomal velocity. PLoS Biol 2013;11:e1001508.

[27] Sabi R, Tuller T. A comparative genomics study on the effect of individual amino acids on ribosome stalling. BMC Genomics 2015;16:S5.

[28] Pavlov MY, Watts RE, Tan Z, Cornish VW, Ehrenberg M, Forster AC. Slow peptide bond formation by proline and other *N*-alkylamino acids in translation. Proc Natl Acad Sci U S A 2009;106:50–4.

[29] Artieri CG, Fraser HB. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. Genome Res 2014;24:2011–21.

[30] Glémin S, Clément Y, David J, Ressayre A. GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. Trends Genet 2014;30:263–70.

[31] Wong GKS, Wang J, Tao L, Tan J, Zhang JG, Passey DA, et al. Compositional gradients in *Gramineae* genes. Genome Res 2002;12:851–6.

[32] Marais G. Biased gene conversion: implications for genome and sex evolution. Trends Genet 2003;19:330–8.

[33] Figuet E, Ballenghien M, Romiguier J, Galtier N. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. Genome Biol Evol 2014;7:240–50.

[34] Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genom Hum Genet 2009;10:285–311.

[35] Capra JA, Pollard KS. Substitution patterns are GC-biased in divergent sequences across the metazoans. Genome Biol Evol 2011;3:516–27.

[36] Robinson MC, Stone EA, Singh ND. Population genomic analysis reveals no evidence for GC-biased gene conversion in *Drosophila melanogaster*. Mol Biol Evol 2014;31:425–33.

[37] Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. Nucleic Acids Res 2017;45:D635–42.

[38] Serres-Giardi L, Belkhir K, David J, Glémin S. Patterns and evolution of nucleotide landscapes in seed plants. Plant Cell 2012;24:1379–97.

[39] Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al. The Ensembl gene annotation system. Database (Oxford) 2016;2016:baw093.

[40] Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, et al. Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res 2016;44:D746–52.

[41] Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci U S A 2002;99:3695–700.

[42] Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol 2011;6:26.