Gene 519 (2013) 113–119

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

# Vertebrate codon bias indicates a highly GC-rich ancestral genome

Maryam Nabiyouni [a,1], Ashwin Prakash [a,b,2], Alexei Fedorov [a,b,*]

[a] Program in Bioinformatics and Proteomics/Genomics, University of Toledo, Health Science Campus, Toledo, OH 43614, USA
[b] Department of Medicine, University of Toledo, Health Science Campus, Toledo, OH 43614, USA

ARTICLE INFO

*Article history:*
Accepted 17 January 2013
Available online 31 January 2013

*Keywords:*
Bioinformatics
Genomics
Evolution
Intron
Computational biology

ABSTRACT

Two factors are thought to have contributed to the origin of codon usage bias in eukaryotes: 1) genome-wide mutational forces that shape overall GC-content and create context-dependent nucleotide bias, and 2) positive selection for codons that maximize efficient and accurate translation. Particularly in vertebrates, these two explanations contradict each other and cloud the origin of codon bias in the taxon. On the one hand, mutational forces fail to explain GC-richness (~60%) of third codon positions, given the GC-poor overall genomic composition among vertebrates (~40%). On the other hand, positive selection cannot easily explain strict regularities in codon preferences. Large-scale bioinformatic assessment, of nucleotide composition of coding and non-coding sequences in vertebrates and other taxa, suggests a simple possible resolution for this contradiction. Specifically, we propose that the last common vertebrate ancestor had a GC-rich genome (~65% GC). The data suggest that whole-genome mutational bias is the major driving force for generating codon bias. As the bias becomes prominent, it begins to affect translation and can result in positive selection for optimal codons. The positive selection can, in turn, significantly modulate codon preferences.

© 2013 Elsevier B.V. All rights reserved.

## 1. Introduction

Synonymous codons are nucleotide triplets that are translated into the same amino acid. Codon bias – an unequal frequency usage of the codons from the same synonymous group – is a general phenomenon characteristic to all organisms from bacteria to multicellular eukaryotes. The codon having the highest frequency within its synonymous group is referred to as optimal or preferred. Usually, closely related species have very similar spectrum of codon biases. For instance, all mammals have similar codon usage frequencies. Even more distantly related vertebrates have about the same sets of optimal codons (for example compare Tables 1 and S1). In contrast, codon usage drastically differs among evolutionary distant species, such as human, fruit fly, worm, yeast, and *Arabidopsis* (see Codon Usage Database which represents codon frequencies for thousands of species (Nakamura et al., 2000)). Various explanations for the creation of codon bias have been proposed (Duret, 2002; Hershberg and Petrov, 2008; Plotkin and Kudla, 2011). After a comprehensive examination of codon frequencies in prokaryotes, as described by Chen et al. (2004), a consensus view was established that the whole-genome nucleotide composition bias is the primary cause for codon bias in bacteria. However, this explanation is much more controversial for eukaryotes. Two recent reviews on this issue propounded two main forces that could be responsible for the creation and maintenance of codon bias (Hershberg and Petrov, 2008; Plotkin and Kudla, 2011). The authors suggest that in addition to the genome-wide mutational force, as seen in bacteria, positive selection for optimal codons due to benefits in translational efficiency and fidelity is a major proponent. This notion is supported by bioinformatic analysis of the influence of surrounding nucleotides on codon preferences in various eukaryotic species. Particularly, in about half of the cases, the context-dependent codon bias could not be explained by genome-wide mutational forces (Fedorov et al., 2002). Also, Chamary and co-authors reviewed evidence that variable sites in synonymous codons are important in mRNA stability and proper splicing (Chamary et al., 2006). Thus, these synonymous sites might not only be under the selection for accuracy and effectiveness of protein synthesis, but also under extra selection forces.

Despite significant progress in this field over a number of years, there is no consensus on the predominant force behind eukaryotic codon bias. There are several strong arguments in support of each of the two forces for the leading role. We attempt to resolve this dilemma for vertebrates by re-assessment of extensive genomic datasets.

The uncertainty in explaining codon bias partially exists because this phenomenon is gene-specific. In other words, within the same genome one group of genes could have very strong codon bias, while another group may have a more balanced distribution of all synonymous

*Abbreviations:* A, Adenosine; BGC, Biased Gene Conversion; C, Cytidine; CB, Codon bias; CBI, Codon Bias Index; G, Guanosine; GC3, G or C at the third codon position; SNP, Single Nucleotide Polymorphism; T, Thymidine.

* Corresponding author at: Department of Medicine, University of Toledo, Health Science Campus, Toledo, OH 43614, USA. Tel.: +1 419 383 5270; fax: +1 419 383 3102.
*E-mail addresses:* Maryam.Nabiyouni@rockets.utoledo.edu (M. Nabiyouni), ashwin.prakash@jhu.edu (A. Prakash), Alexei.fedorov@utoledo.edu (A. Fedorov).
[1] Current address: Biomedical Engineering, College of Engineering, University of Toledo, Toledo, OH 43606, USA.
[2] Current address: Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.01.033

codons, and yet another subset of genes may have an inverse bias (where optimal codons become rare and vice versa). As a result several controversies in this field might be attributed to the differences in gene sets being analyzed. In addition, variations in the chromosomal GC composition significantly impact nucleotide composition of a gene and its codon bias. For instance, a recent paper by Romiguier et al. (2010) presents important data on GC-content dynamics across 33 mammalian genomes. They described interesting trends in mammalian codon bias evolution and connected it with biased gene conversion hypothesis (Galtier et al., 2001). In this paper we propose an alternative hypothesis. We suggest that significant enrichment by guanine and cytosine bases in variable codon positions may be attributed to the overall genomic GC-richness of vertebrate ancestors.

## 2. Materials and methods

### 2.1. Gene datasets

Coding and intronic sequences for human, mouse chicken, and zebrafish were obtained from our genomic Exon–Intron Database (Shepelev and Fedorov, 2006). For human and mouse we used complete sets of intron-containing genes obtained from Build 37.1 GenBank release; for *Danio rerio* we used Zv4 genomic release (30-JUN-2005); and for chicken — Build 2.1 release (16-NOV-2006). All genes that contain internal stop codons have been removed. Also, genes with very short coding sequences (<400 nts) have been removed because they are enriched with hypothetical genes. For alternatively spliced genes, we used only one gene isoform that is listed at the top of the GenBank feature table. The final gene samples contain 17,960 human genes; 17,675 mouse; 20,040 zebrafish, and 11,784 chicken. These samples are available at our website: http://bpg.utoledo.edu/~afedorov/lab/eid.html.

### 2.2. Calculation of Codon Bias Index (CBI)

To measure the codon bias in individual genes we used Codon Bias Index as described in Bennetzen and Hall (1982). The details of these calculations and the PERL program are available from Nabiyouni (2011). Alternative to CBI, researchers often use GC3 values to analyze the CB (Chamary et al., 2006). However, interpretation of GC3 depends on the amino acid composition of the coded protein, while the CBI reflects the correlation of codon preferences in the gene to the main codon usage table of the organism.

### 2.3. Gene expression analysis

Publicly available gene array datasets from the BioGPS database (Su et al., 2004) for humans (GC Robust Multi-array Average (GCRMA) normalized Affymetrix microarray probe-level data from Human U133A/GNF1H Gene Atlas) have been used to procure expression values for the genes. Perl scripts were generated to mine the expression levels of individual genes in six different tissues. This expression data was then pooled for genes with close CBI values (each bin of genes has CBI value differences within a range of 0.1). The mean and median gene expression values for the genes within each bin were then plotted in bar graphs.

### 2.4. Statistics

The squared Pearson's correlation coefficients for the GC1, GC2, and GC3 values presented in Fig. 1 and for the CBI values and intronic GC-content in Fig. 2 were calculated using Microsoft Excel (Office 2010) built-in program. The binning of genes by CBI value and the subsequent calculation of mean and median gene expression levels in Fig. 3 were performed using Perl scripts. The data was plotted using Microsoft Excel (Office 2010), and a polynomial 2 trendline was fitted to the plot. For statistical evaluation of the prevalence for optimal codons ending by A/T or G/C over the random model in which optimal codons are chosen arbitrarily, we used Monte-Carlo simulation with our Perl script MonteCarlo.pl available from our web page (http://bpg.utoledo.edu/~afedorov/lab/prog/montecarlo.html).

### 2.5. URLs

The following are the URLs of the databases used in the study:

Codon Usage Database (http://www.kazusa.or.jp/codon/) (Nakamura et al., 2000).

Exon–Intron Database (http://bpg.utoledo.edu/~afedorov/lab/eid.html) (Shepelev and Fedorov, 2006).

BioGPS database (http://biogps.org/downloads/) (Su et al., 2004).

## 3. Results

### 3.1. Regularities in codon bias

Relative frequencies of synonymous codons were computed from 17,960 intron-containing human genes (Table 1). Analogous data for mouse, chicken, and zebrafish codons are in the Supplementary Table S1. The results in Table 1 reveal only minor fluctuations (mainly within 1% interval) from the frequencies presented in the Codon Usage Database (Nakamura et al., 2000) for the entire set of human coding sequences, suggesting that our gene sample is valid and not notably skewed. Essential regularities of codon usage in humans are clearly seen in Table 1. The major pattern in Table 1 is the preference of those synonymous codons in which last nucleotide in the third (wobbling) position is G or C. Exceptions to this pattern are found only in the second column in Table 1 where codons have C in the second codon position. In this column the rarest codons always have G in the third position, yielding CpG dinucleotides at the end of these codons. In vertebrate genomes CpG is the most underrepresented

**Table 1**
Relative frequencies of synonymous codons calculated for 17,960 human intron-containing genes.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| UUU | F | 0.47 | UCU | S | 0.19 | UAU | Y | 0.45 | UGU | C | 0.47 |
| UUC | F | 0.53 | UCC | S | 0.21 | UAC | Y | 0.55 | UGC | C | 0.53 |
| UUA | L | 0.08 | UCA | S | 0.15 | UAA | Stop | | UGA | Stop | |
| UUG | L | 0.13 | UCG | S | 0.05 | UAG | Stop | | UGG | W | 1.0 |
| CUG | L | 0.13 | CCU | P | 0.29 | CAU | H | 0.43 | CGU | R | 0.08 |
| CUC | L | 0.19 | CCC | P | 0.32 | CAC | H | 0.57 | CGC | R | 0.18 |
| CUA | L | 0.07 | CCC | P | 0.28 | CAA | Q | 0.27 | CGA | R | 0.11 |
| CUG | L | 0.39 | CCG | P | 0.12 | CAG | Q | 0.73 | CGG | R | 0.21 |
| AUG | I | 0.37 | ACU | T | 0.25 | AAU | N | 0.48 | AGU | S | 0.15 |
| AUG | I | 0.46 | ACC | T | 0.35 | AAC | N | 0.52 | AGC | S | 0.24 |
| AUA | I | 0.18 | ACA | T | 0.29 | AAA | K | 0.44 | AGA | R | 0.21 |
| AUG | M | 1.0 | ACG | T | 0.11 | AAG | K | 0.56 | AGG | R | 0.21 |
| GUU | V | 0.18 | GCU | A | 0.26 | GAU | D | 0.47 | GGU | G | 0.16 |
| GUC | V | 0.24 | GCC | A | 0.40 | GAC | D | 0.53 | GGC | G | 0.33 |
| GUA | V | 0.12 | GCA | A | 0.23 | GAA | E | 0.43 | GGA | G | 0.25 |
| GUG | V | 0.46 | GCG | A | 0.11 | GAG | E | 0.57 | GGG | G | 0.25 |

Synonymous codon groups composed of two members are shown in light and dark blue; of three codons — in gray; of four codons in yellow; and of six codons in orange and ochre. Amino acids specified by the codons are shown in a single letter form.

dinucleotide, occurring about four times below expectations, because CpG dinucleotides are hot-spots for C→T mutations due to methylation→deamination of the cytosines within this context. Hypermutability of CpG dinucleotides is one of the major causes of codon substitution in mammalian genes (Misawa and Kikuno, 2011). The footprint of this CpG→TpG transition is clearly visible in Table 1. For instance, the alanine codon GCG has the lowest relative frequency (11%) among all synonymous groups comprising four members. Due to recurrent mutation of 5meC into T, this codon should be repeatedly converted into a valine GTG codon, which has the highest relative frequency (46%) among all synonymous groups composed of four triplets. This example of the deficit of NCG codons testifies that genome-wide mutational forces robustly influence codon bias. Due to the observed patterns in synonymous codons preferences, the GC-percentage of the third codon positions (so-called GC3-content) of human genes is 58.6%, despite the fact that the overall GC-composition in the entire human genome is significantly lower — 40.9% (see Table 2). Table 2 demonstrates the same trend for all mammals and vertebrates: species from this taxon have GC3 composition significantly higher than their genomic GC-composition. However, if we consider other branches of eukaryotes with considerably lower genomic GC-composition, their optimal codons end predominantly by A or T. This observation is correct for all but two synonymous codon groups for such organisms as *Plasmodium falciparum* (genome GC-content 19.4%), *Dictyostelium discoideum* (25.7%), *Saccharomyces cerevisiae* (38.3%), *Arabidopsis thaliana* (36%), and *Caenorhabditis elegans* (36%) (see Codon Usage Database (Nakamura et al., 2000)). The only two exceptions to this rule for the aforementioned five species are the following: 1) in the *C. elegans* genome the phenylalanine optimal codon is UUC (relative frequency 50.6% among two synonymous codons) and 2) in *S. cerevisiae* genome the leucine optimal codon is UUG (relative frequency 28.6% among six synonymous codons). Monte-Carlo simulation shows that in case the optimal codons are chosen randomly, the chance to have the observed level of prevalence for optimal codons ending by A or T is less than $10^{-12}$. On the other hand, organisms with GC-rich genomes like *Chlamydomonas reinhardtii* (64%), *Leishmania major* (60%), *Nocardia farcinica* (71%), and *Acidiphilium cryptum* (68%) have opposite codon bias regularities: codons ending with G and C are drastically more abundant than those ending with A or T. In fact, each optimal codon of these four species ends by G or C. Monte-Carlo simulation shows that in case the optimal codons are chosen randomly, the chance to have the observed level of prevalence for optimal codons ending by C or G is less than $10^{-12}$. Taken all together, we observe for the kingdom of eukaryotes a global regularity that organisms with extremely high genomic GC-content also have high GC3-composition while organisms with low genomic GC-content have low GC3-composition. Vertebrates present a notable exception from this rule (as well as *Drosophila* originally considered in Duret and Mouchiroud (1999)). Vertebrate genomic composition is rather GC-poor (in the range 37–46% (Costantini et al., 2009)), in contrast to their much higher GC3 content (ranges between 56 and 64%, see Table 2). Even special genomic regions of mammals and birds with the highest GC-composition (so-called H-isochores) have lower GC-content than GC3. For example, in humans the most GC-rich isochore (H3) has a GC-composition of 53–57% (Costantini et al., 2009), while the overall human GC3 is 58.6%. According to Fig. 3 of Duret and Galtier (2009), the GC3-composition is much higher (about 75–80%) inside the GC-richest (H3) regions of the human genome. Moreover, according to genome-wide computations of Zhao and Jiang, currently the human GC3 value is decreasing (Zhao and Jiang, 2010). This notion implies that human GC3 value was even higher in the past.

Transition of preferable synonymous codons in non-vertebrate eukaryotes from GC3-rich ones in GC-rich genomes to GC3-poor preferable codons in AT-rich genomes perfectly fits into the model of mutational origin of codon bias. This view is concordant with Chen at al. (2004), who concluded that bacterial codon bias is explainable by genome-wide nucleotide composition "with surprising accuracy". On the other hand, mutational theory is totally helpless in the case of vertebrates, where GC3-content is well above whole-genome GC-content. In this case a selectionist view is much plausible (Chamary et al., 2006).

In order to get an insight into this problem we considered other preferences in coding sequences. Fig. 1 illustrates correlations in GC-composition between the first, second, and third codon positions (known as GC1, GC2 and GC3 indices respectively) for a wide spectrum of prokaryotic and eukaryotic species. Strict linear correlations between the GC1, GC2, and GC3 values exist in bacteria (Fig. 1, right column). Prokaryotic GC1, GC2, and GC3 indices are also in direct proportion to the overall genomic GC-content (Nabiyouni, 2011). Similarly eukaryotes also have linear correlations between GC1, GC2, and GC3 albeit with significantly higher fluctuations from the trend line (Fig. 1, left column). Some of the fluctuations from the trend in eukaryotes may be explained by the presence of unusual amino acid abundances in particular species (Nabiyouni, 2011). The described regularities in Fig. 1 suggest that impact of genome-wide GC-content on gene nucleotide composition is very strong. It influences the GC3-content (and thus, codon bias) and also it affects the GC1 and GC2-contents (and thus, amino acid composition of proteins). Such impacts were previously reported for bacteria (Warnecke et al., 2009).

### 3.2. Codon bias in individual genes

We used the Codon Bias Index (CBI) to measure codon preferences in single genes. Fig. 2A presents the distribution of 17,960 human genes by their CBI values. When the value of CBI for a gene is positive, it indicates that the gene has a codon bias similar to that observed in the majority of genes. However, when CBI is negative it means that the gene is mainly populated by non-optimal codons (inverse codon bias). According to Fig. 2A, a majority (83.5%) of human genes have mild codon bias in the 0 to +0.50 range of CBI values. 8.2% of the genes have very strong codon usage preference for the optimal codons (≥0.5 CBI), while 8.3% have inverse codon preferences, which are associated with negative CBI values.

Local GC-content for each of 17,960 genes was calculated by computing the average GC-composition of all introns within the gene. Introns may serve as a reliable control for local chromosomal nucleotide content because their GC-composition is the same as for the whole genome (for animal intronic GC-percentage see Table 1 in Bechtel et al. (2008)). A two dimensional distribution of human genes by their CBI values and local GC-content is represented in Fig. 2B. This figure demonstrates a prominent trend that genes with GC-rich introns have the strongest codon bias, while those with GC-poor introns on average have significantly lower or even inverse codon bias. These findings are in complete agreement with the results of Zeeberg (2002), which employed a smaller gene sample and different algorithms for calculation of codon bias and local GC-content.

Among vertebrates, the most prominent large-scale genome variation in GC-composition, known as isochore structure, exists in warm-blooded amniotes, with much less variation in the genomes of cold-blooded vertebrates (Bernardi, 2007; Costantini et al., 2009; Eyre-Walker and Hurst, 2001). Thus, we also investigated the dependency of CBI on local genomic GC-content for 11,784 genes from chicken (Fig. 2C), which has a similar isochore structure as humans (Costantini et al., 2009). Both species have the same general propensity: the genes with high intronic GC-composition tend to have the strongest codon bias, while genes with low intronic GC-percentage are enriched with coding sequences lacking clear preferences in synonymous codons or with inverse codon bias. On the other hand, studying 20,040 genes of zebrafish, whose genome possesses a more uniform genomic GC-composition, demonstrated much less variation of CBI values for individual genes (Fig. 2D). For mammals and warm-blooded amniotes,

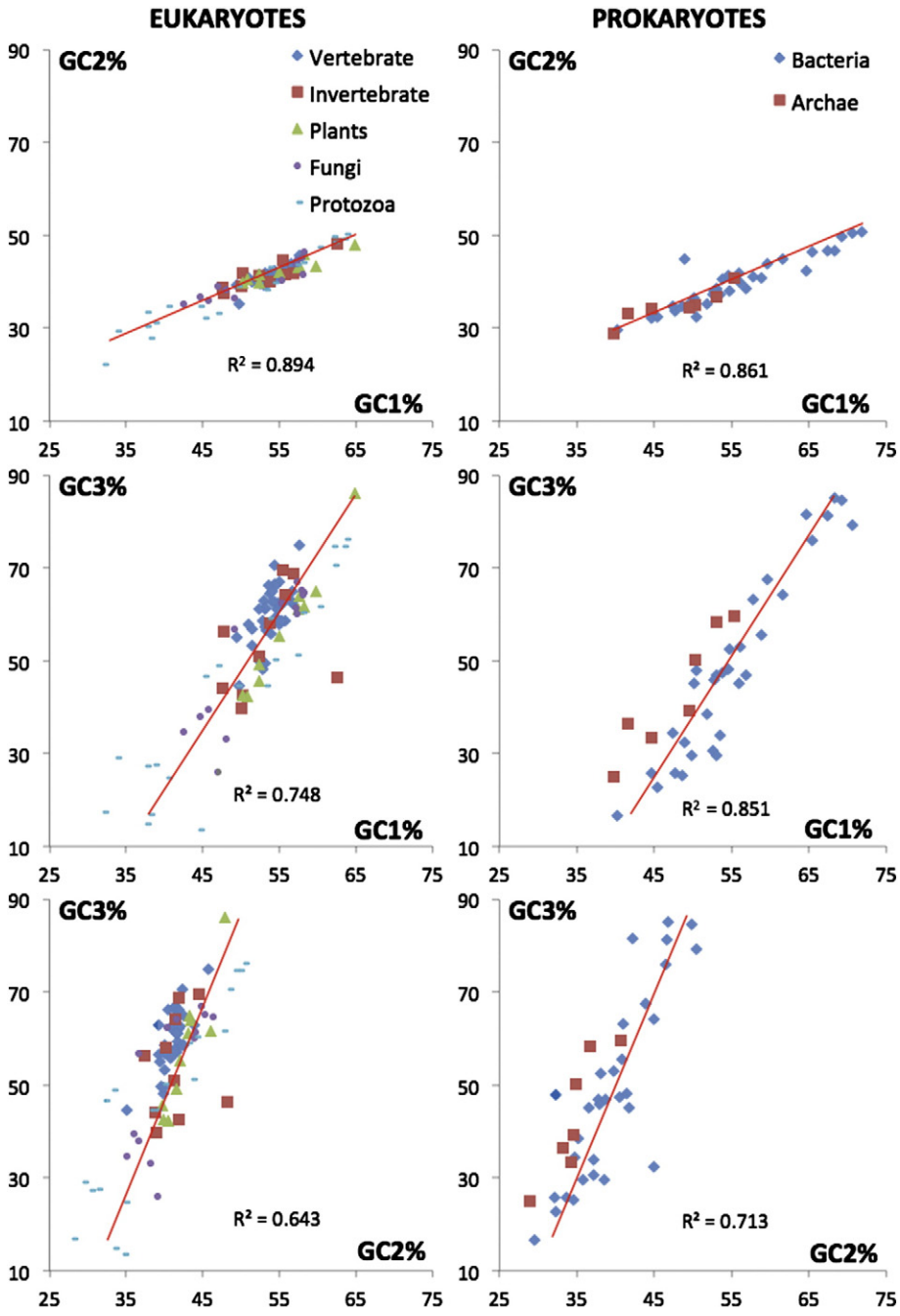


**Fig. 1.** Relationships between GC-content in the first, second, and third codon positions (GC1, GC2, and GC3 respectfully) among a spectrum of organisms. Each dot represents one species. The species are specified in Nabiyouni (2011). Line represents the trend. The squared Pearson's correlation coefficient is shown for each plot. The p-values for all of them are <0.0001, so the correlations are significant.

the data indicate that local genomic nucleotide composition is critical for codon bias in that region.

### 3.3. Association of codon bias with the expression

Recently, Plotkin and Kudla (2011) and Hershberg and Petrov (2008) reviewed numerous reports on the correlations of codon bias with gene expression level in different organisms. These data are complex and sometimes controversial yet they produced an overall consensus that synonymous sites are likely under a weak selection in the efficiency and/or fidelity of protein synthesis. Several biological explanations of this notion are considered in these two papers. The availability of vast public gene-array datasets gave us an opportunity to examine the connection of gene CBI values with mRNA expression levels in our set of 17,960 human genes. The results are shown in the Fig. 3, which presents data for mean and median levels of mRNA expression. Significantly elevated expression was observed only for one group of 84 genes with the highest CBI values (>0.5), while the rest of the genes were found to have median expression levels within a 20% range of variation between gene bins with similar CBI values. These results suggest that highly transcribed genes tend to have high CBI values. We agree with the proposition of weak selection forces that influence on the synonymous codon usage. However, it is unlikely that this weak selection can withstand strong

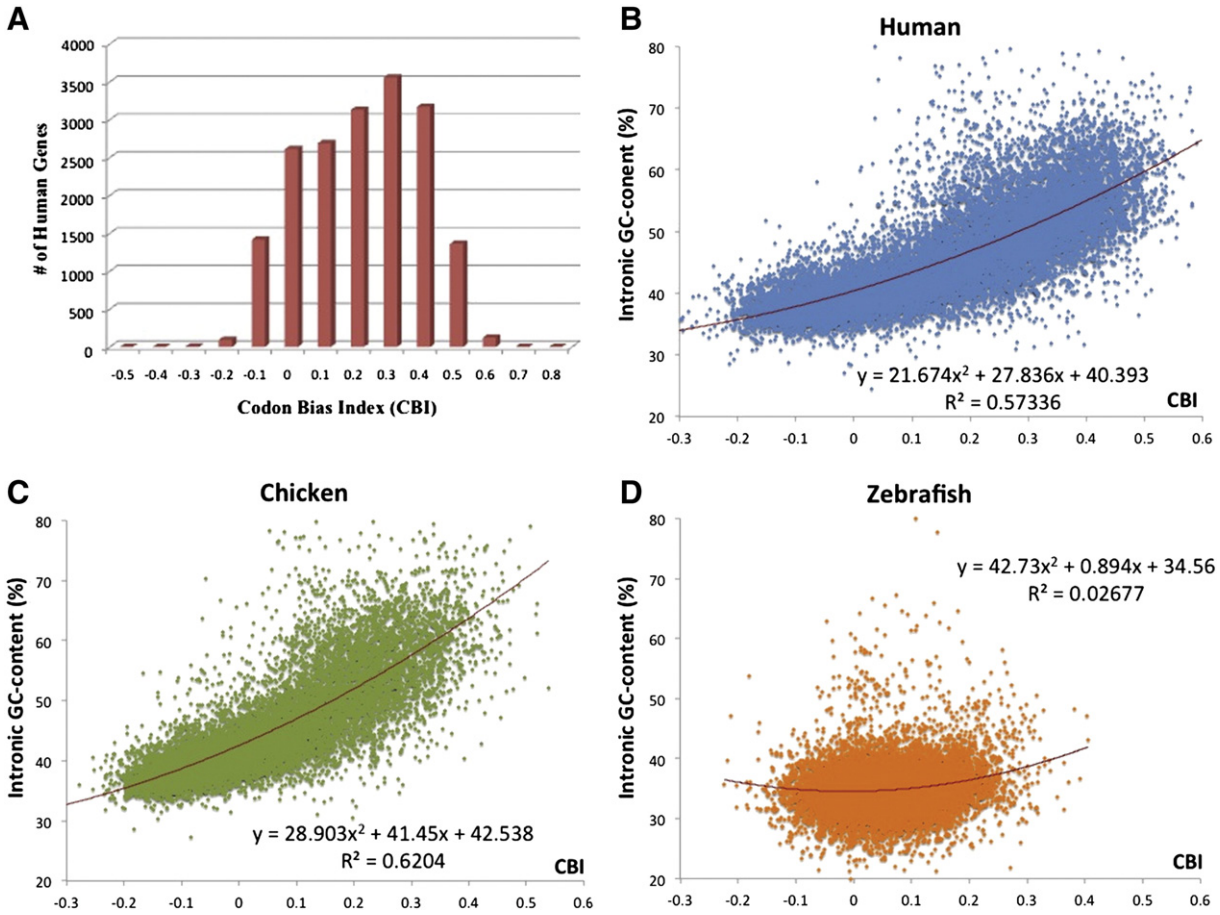


**Fig. 2.** A: Distribution of human genes by CBI values of their coding sequences. A CBI bin $x$ (e.g. 0.1, 0.2 …) represents genes, in which CBI values are in the range $[x, x+0.1)$. B–D: Distribution of human, chicken, and zebrafish genes by CBI values and intronic GC content. Each dot represents a single human gene.

mutational forces working in opposite direction and create considerable GC3-richness (56–65%) in vertebrates despite their overall poor genomic GC-composition (37–46%).

## 4. Discussion

The difference between GC3 and overall genomic GC-composition may exist because GC3 only reflects the nucleotide substitution process, whereas genomic GC-content reflects a balance between substitutions, transposable element insertions, and deletions (Duret and Hurst, 2001). However, in vertebrates, GC-content of unique and repetitive elements is about the same. For example, unique (unmasked) regions and repetitive elements (masked by RepeatMasker and TandemRepeatFinder) for the latest versions of entire human and mouse genomes from UCSC Genome Browser (chrom.Fa.tar.gz files) gave the following results: mouse unique regions GC = 41.2%; mouse repetitive regions GC = 42.2%; human unique regions GC = 40.3%; human repetitive regions GC = 41.5%. The major driving force of vertebrate GC-composition is an intense flow of point mutations ("1000 Genomes" project has confirmed ~50 novel point mutations per individual in humans (2010)). Also, point mutations happen more than ten times often than insertions/deletions. For example, our calculations of the public data of the 1000 Genome project (2010) showed that there are 458 thousand point mutations and only 18 thousand small insertions/deletions on the human chromosome 22. Therefore point mutations are likely the major contributor for GC-composition in vertebrates.

The prevalence of GC3 over genomic GC-composition is used as an argument in consideration of Biased Gene Conversion (BGC) theory (Duret and Galtier, 2009). BGC is an important scheme that explicitly explained non-randomness in nucleotide composition of different genomic regions by non-selective forces occurring via biases in recombination and reparation molecular machineries. However, BGC pathway, which occurs via formation of heteroduplexes spanning over ~1800 nucleotides followed by non-random reparation of mismatches within these heteroduplexes, is unable to explain the significant prevalence of GC3 over genomic GC-composition. Indeed, the average size of exons is only 135 nucleotides in vertebrates, so they should represent only a small portion of heteroduplexes and the BGC effect should spread over neighboring intronic regions. There are other types of fixation biases in which predispositions strongly depend on local nucleotide compositions (Prakash et al., 2009) and which may contribute to the high level of GC3 content in mammals.

This paper suggests another hypothesis that may explain controversies in the codon bias of vertebrates by a simple conjecture that the common ancestor of all vertebrates had a GC-rich genome. Consider organisms having extremes of genomic GC-composition. For instance, *N. farcinica*, *A. cryptum*, and *L. major* have genomic GC percentage from 60% to 71% and considerably higher GC3-content (above 75%, Table 2). Species with abnormally AT-rich genomes demonstrate a similar pattern with respect to AT instead of GC. For example, *D. discoideum* has genomic AT = 74.3% and AT3 = 85.1%; while *P. falciparum* has AT = 80.6% and AT3 = 83.1%. These examples demonstrate that when genomic CG-composition reaches extremes (far from 50%), the CG3-composition has a more pronounced extreme.

Genomic GC-composition is highly variable among species, ranging from 20 to 70% (Table 2). If the common vertebrate progenitor had a GC-rich genomic composition (about 65%), and given that GC3-content often exceeds the genomic GC content when it is well above 50%, the

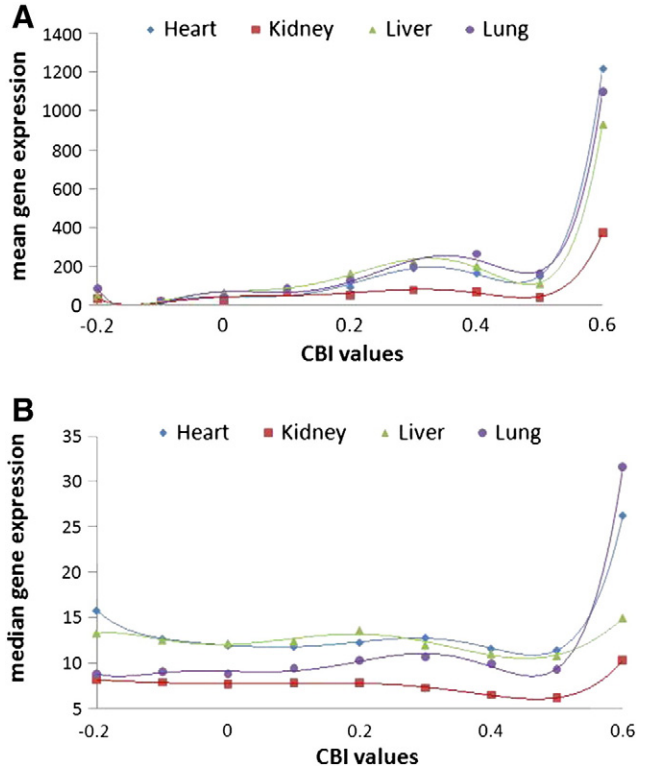


**Fig. 3.** Relationship between expression of human genes and their CBI values. Each dot represents median (A) and mean (B) expression levels of all genes within that range of CBI values (e.g. dot for 0.1 would represent all genes with CBI value ≥ 0.1 and <0.2). Gene expression level was measured in GC Robust Multi-array Average (GCRMA) units and it was obtained from the BioGPS (Affymetrix) database.

common vertebrate ancestor would have had a GC3-content up to 75%. Finally, we assume that this ancient vertebrate progenitor began losing its genomic GC-richness due to mutations in genes controlling DNA replication/reparation systems. All its descendants, which evolved since, have inherited a genome with diminution of GC-content currently in the range of 38–45%, and a GC3-content of >55% (except frogs that have ~51%). It is not currently possible to infer the GC content of the ancestral vertebrate genome from the composition of extant genomes. Our best guess is that this ancestral composition was well above 50% GC-composition, which should have created a prevalence of codons ending by G or C and in turn might have ignited the selection of these GC3-rich codons for efficiency of protein synthesis. Here and in Fig. 4 we made our best approximation of 65%, which is equal to the highest value of GC3 observed among all studied vertebrates (Table 2, fugu: GC3 = 65.0%).

Our hypothesis is illustrated in Fig. 4. The diminution of GC3 should be slower than that of genomic GC-composition since the former is supported by selection forces involved in efficiency/fidelity of protein synthesis while the latter can be faster in the vast non-functional regions such as segments of introns and intergenic regions. In addition, synonymous mutations are subject to additional selection because they affect splicing and/or mRNA stability (Chamary et al., 2006; Sterne-Weiler et al., 2011; Warnecke and Hurst, 2007). In any proposed scenario a selection force favoring GC3-rich codons should have taken place at some period of vertebrate evolution. In our hypothesis this selection force has not experienced a very strong headwind of genome-wide mutational GC-bias. The proposed mutual feedback between neutral and selective forces for shaping codon bias has been previously described by Higgs and Ran for bacteria species (Higgs and Ran, 2008; Ran and Higgs, 2010).

Our hypothesis of GC-rich genome in vertebrate progenitor could be tested in modeling of CG-composition during evolution of different

**Table 2**
Genomic GC- and GC3-compositions among species.

| Taxa | Species | GC% | GC3% |
|---|---|---|---|
| *Mam* | Human | 40.9 | 58.6 |
| *Mam* | Mouse | 41.8 | 58.6 |
| *Mam* | Dog | 41.0 | 62.2 |
| *Mam* | Cow | 41.7 | 62.5 |
| *Mam* | Opossum | 37.7 | 52.4 |
| *Vrt* | Chicken | 41.3 | 57.8 |
| *Vrt* | Frog (*X. tropicalis*) | 40.0 | 51.3 |
| *Vrt* | Zebrafish | 36.9 | 56.0 |
| *Vrt* | Fugu | 45.4 | 65.0 |
| *Chr* | Sea squirt (*Ciona intestinalis*) | 35.8 | 42.5 |
| *Chr* | Lancelet (*Branchiostoma floridae*) | 41.2 | 61.1 |
| *Inv* | Sea urchin (*S. purpuratus*) | 37.0 | 46.5 |
| *Inv* | *D. melanogaster* | 42.2 | 64.3 |
| *Inv* | *C. elegans* | 36 | 39.8 |
| *Fng* | *Neurospora crassa* | 50.0 | 65.1 |
| *Fng* | *S. cerevisiae* | 38.3 | 38.1 |
| *Pln* | *Arabidopsis* | 36 | 42.4 |
| *Ptz* | *Dictyostelium discoideum* | 25.7 | 14.9 |
| *Ptz* | *Leishmania major* | 59.7 | 76.2 |
| *Ptz* | *Plasmodium falciparum* | 19.4 | 17.3 |
| *Ptz* | *Chlamydomonas reinhardtii* | 64 | 86.2 |
| *Bac* | *Acidiphilium cryptum* | 68.0 | 84.6 |
| *Bac* | *Nocardia farcinica* | 71 | 90.7 |

Mam — mammals; Vrt — non-mammalian vertebrates; Chr — chordates; Inv — invertebrates; Fng — fungi; Pln — plants; Ptz — protozoa; Bac — bacteria.

branches of this taxon. Initially, isochore structure (uneven GC-richness of large chromosomal segments) was reported only in amniotes, but recently they have been described in other lineages of vertebrates (Costantini et al., 2009). There are several theories explaining the existence of these isochores (Duret and Galtier, 2009; Pozzoli et al., 2008; Vinogradov, 2005). It is debatable that GC-rich isochores originated due to positive selection within GC-poor mammalian genomes, because of the immense genetic cost required for positive selection to work simultaneously on millions of nucleotides. We

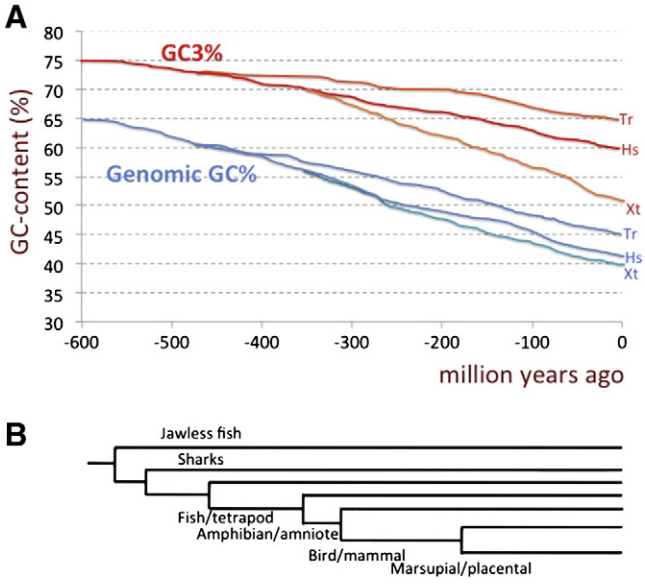


**Fig. 4.** Hypothesis for the evolution of GC-content in vertebrates. A. About 600 million years ago the last common vertebrate ancestor supposedly had around 65% genomic GC-composition (blue lines) and even higher GC3-content (about 75%, red lines). Since that time, the GC- and GC3-percentage started declining (see Discussion). Hs stands for *Homo sapiens*, GC-composition of which is close to other mammals and many vertebrates. Xt — represents a frog *Xenopus tropicalis*; and Tr — a fish *Takifugu rubripes*. They are two species in which, currently, genomic and GC3 nucleotide contents most dramatically differ from humans. B. Phylogeny of vertebrates is shown in the same time-scale as A.

also acknowledge the importance of Biased Gene Conversion (BGC) theory proposing a pure mechanistic force of genomic compositional inhomogeneity (reviewed in Duret and Galtier (2009)). A common view is that BGC could be only a partial reason for the existence of isochores, yet the theory cannot explain all of the dramatic variations in GC composition at different chromosomal regions within mammals (Duret and Galtier, 2009; Pozzoli et al., 2008). However, if the ancestor of vertebrates had highly GC-rich genome, it seems more plausible that the present isochore structure might simply have appeared due to the variation in rates of GC-deterioration in different chromosomal regions. Moreover, our recent analysis of human SNPs demonstrated a strong fixation bias for mutations, which depends on the local nucleotide context (Prakash et al., 2009). Specifically, in GC-rich surroundings A/T→G/C mutations are favorable for fixation, while in AT-rich neighborhood G/C→A/T mutations are preferentially fixed. We plan to verify our hypothesis making a computer model of evolution of GC-composition and taking into account details of mutation rates and their fixation preferences.

## 5. Conclusions

We propose that whole-genome nucleotide composition is the primal force for the creation of codon bias. As the bias becomes prominent, it begins to affect translation and can result in positive selection for optimal codons. The positive selection can, in turn, significantly modulate codon preferences.

## Funding

This work is supported by the National Science Foundation Grant MCB-0643542.

## Conflict of interest statement

None declared.

## Acknowledgments

We are grateful to Dr. Robert Blumenthal, UT, for his insightful discussion of the project.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2013.01.033.

## References

A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073.
Bechtel, J.M., et al., 2008. Genomic mid-range inhomogeneity correlates with an abundance of RNA secondary structures. BMC Genomics 9, 284.
Bennetzen, J.L., Hall, B.D., 1982. Codon selection in yeast. J. Biol. Chem. 257, 3026–3031.
Bernardi, G., 2007. The neoselectionist theory of genome evolution. Proc. Natl. Acad. Sci. U. S. A. 104, 8385–8390.
Chamary, J.V., Parmley, J.L., Hurst, L.D., 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. 7, 98–108.
Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., McAdams, H.H., 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. U. S. A. 101, 3480–3485.
Costantini, M., Cammarano, R., Bernardi, G., 2009. The evolution of isochore patterns in vertebrate genomes. BMC Genomics 10, 146.
Duret, L., 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12, 640–649.
Duret, L., Galtier, N., 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. 10, 285–311.
Duret, L., Hurst, L.D., 2001. The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Mol. Biol. Evol. 18, 757–762.
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, 4482–4487.
Eyre-Walker, A., Hurst, L.D., 2001. The evolution of isochores. Nat. Rev. Genet. 2, 549–555.
Fedorov, A., Saxonov, S., Gilbert, W., 2002. Regularities of context-dependent codon bias in eukaryotic genes. Nucleic Acids Res. 30, 1192–1197.
Galtier, N., Piganeau, G., Mouchiroud, D., Duret, L., 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159, 907–911.
Hershberg, R., Petrov, D.A., 2008. Selection on codon bias. Annu. Rev. Genet. 42, 287–299.
Higgs, P.G., Ran, W., 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol. Biol. Evol. 25, 2279–2291.
Misawa, K., Kikuno, R.F., 2011. Relationship between amino acid composition and gene expression in the mouse genome. BMC Res. Notes 4, 20.
Nabiyouni, M., 2011. Mega-scale bioinformatics investigation of codon bias in vertebrates. University of Toledo Health Science Campus, College of Medicine. University of Toledo, Toledo, p. 65.
Nakamura, Y., Gojobori, T., Ikemura, T., 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 28, 292.
Plotkin, J.B., Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12, 32–42.
Pozzoli, U., et al., 2008. Both selective and neutral processes drive GC content evolution in the human genome. BMC Evol. Biol. 8, 99.
Prakash, A., et al., 2009. Evolution of genomic sequence inhomogeneity at mid-range scales. BMC Genomics 10, 513.
Ran, W., Higgs, P.G., 2010. The influence of anticodon–codon interactions and modified bases on codon usage bias in bacteria. Mol. Biol. Evol. 27, 2129–2140.
Romiguier, J., Ranwez, V., Douzery, E.J., Galtier, N., 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. 20, 1001–1009.
Shepelev, V., Fedorov, A., 2006. Advances in the exon–intron database (EID). Brief. Bioinform. 7, 178–185.
Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D.N., Sanford, J.R., 2011. Loss of exon identity is a common mechanism of human inherited disease. Genome Res. 21, 1563–1571.
Su, A.I., et al., 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. U. S. A. 101, 6062–6067.
Vinogradov, A.E., 2005. Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. Trends Genet. 21, 639–643.
Warnecke, T., Hurst, L.D., 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. Mol. Biol. Evol. 24, 2755–2762.
Warnecke, T., Weber, C.C., Hurst, L.D., 2009. Why there is more to protein evolution than protein function: splicing, nucleosomes and dual-coding sequence. Biochem. Soc. Trans. 37, 756–761.
Zeeberg, B., 2002. Shannon information theoretic computation of synonymous codon usage biases in coding regions of human and mouse genomes. Genome Res. 12, 944–955.
Zhao, Z., Jiang, C., 2010. Features of recent codon evolution: a comparative polymorphism-fixation study. J. Biomed. Biotechnol. 2010, 202918.