academicJournals

Vol. 14(4), pp. 341-349, 28 January, 2015
DOI: 10.5897/AJB2014.14110
Article Number: 886C72649965
ISSN 1684-5315

http://www.academicjournals.org/AJB

African Journal of Biotechnology

Full Length Research Paper

Genomic composition factors affect codon usage in porcine genome

Khobondo, J. O.[1]*, Okeno, T. O.[1,2] and Kahi, A. K.[1]

[1]Animal Breeding and Genomics Group, Department of Animal Sciences, Egerton University, P. O. Box 536, 20115 Egerton, Kenya.
[2]Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, P. O. Box 50, 8830 Tjele, Denmark.

Received 18 August, 2014; Accepted 15 January, 2015

The objective of the study was to determine the codon usage bias in the porcine genome and decipher its determinants. To investigate the underlying mechanisms of codon bias, the coding sequence (CDS) from the swine reference sequence (ssc10.2) was extracted using Biomart. An in house built Perl script was used to derive various genomic traits and codon indices. Analysis was done using R statistical package, and correlations and multivariate regressions were performed. We report the existence of codon usage bias that might suggest existence of weak translational selection. The codon bias is feebly related to nucleotide composition (GC%, GC3, CDS length). This study can be explored for designing degenerate primers, necessitate selecting appropriate hosts expression systems to manipulate the expression of target genes *in vivo* or *in vitro* and improve the accuracy of gene prediction from genomic sequences thus maximizing the effectiveness of genetic manipulations in synthetic biology.

Key words: Coding sequence, synomynous codons, selection, translational mutation, pig genome.

INTRODUCTION

The availability of nearly complete genome sequences from different taxa has enabled tremendous advances in evolutionary biology, providing insight to the actions of natural selection on genomes (Whittle et al., 2012). These biological breakthroughs revealed the importance of studying the degeneracy of genetic code, which enables most amino acids to be coded by more than one o called 'synonymous' codon (Wright, 1990). Synonymous codons usage (SCU) bias has been documented both within and between genomes, with huge interspecific and even intragenomic variation (Jia et al., 2009).

Several biological factors such as tRNA abundance (Kanaya et al., 2001), strand specific mutational bias, replicational, transcriptional and translational selection (Hershberg and Petrov, 2008), secondary structure of proteins, mRNA structure, GC composition (Knight et al., 2001) and environmental factors (Basak and Ghosh,

*Corresponding author. E-mail: jkhobondo@gmail.com.

2005; Behura et al., 2013a) have been reported to influence the synonymous codon usage in various organisms. The afore mentioned factors led to two hypotheses on the evolution of codon bias; mutation bias and natural selection for translation accuracy and efficiency (Sharp et al., 2005).

The mutational bias hypothesis predicted that genes in the GC-rich regions of the genome preferentially use G- and C-ending codons, while those in the AT-rich regions use A- and T-ending codons (Zhang et al., 2009) as observed in mammals.

In *E. coli*, Stoletzki and Eyre-Walker (2007) found strong support for the selection for translational accuracy hypothesis; they reported that highly conserved sites and genes have higher codon bias than less conserved sites and genes. In their report, codon bias was positively correlated to gene length and production costs, both indicating selection against missense and nonsense mutations. This was further corroborated in plants such as *Arabidopsis thaliana*, *Oryza sativa* and *Zea mays* where codon usage bias was correlated to the base composition of genes, gene expression level and CDS length (Morton and Wright, 2007).

In higher animals like humans there are reported codon bias which are thought to maximize the speed of elongation, minimize the costs of proofreading thus maximizing the accuracy of translation (Bulmer, 1991). Several studies have failed to disentangle between translation accuracy and efficiency, as both are believed to be intertwined.

However, there are reports correlating the translational efficiency with expression levels and use of codons that match common tRNAs. For instance, in eukaryotes, codon bias have been associated with translation efficiency (Qian et al., 2012), so that the most abundant tRNA can be recognized easily in highly expressed genes. In the study, Qian et al. (2012) hypothesized that different synonymous codons were translated at different speeds due to disparities in codon selection time, and that faster translation were important because it minimize ribosome sequestering and so help alleviate ribosome shortage.

On the contrary, some reports dismiss translation efficiency and accuracy from synonymous codon bias usage in mammals (Reis et al., 2004). Instead splice related biases are evident (Parmley and Hurst, 2007) and selection for the preservation of exonic splicing enhancers (ESEs) where there are high density of regulatory elements, explains low SNP density, low protein evolutionary rates, and low synonymous substitution near intron- exon boundaries (Parmley and Hurst, 2007).

Despite reports on codon usage in mammals, there is no such literature in porcine genome. The study of codon usage would shade light to the known disparity in gene expression levels and quantitative trait loci that may related to genome architectures. In this study, we tested these mutation and translation hypotheses using genomic traits at our disposal. Our results confirm evidence of codon usage bias that was affected by CDS length, GC content and GC3s in the CDS. The analysis of codon usage pattern in pigs might give insight for understanding the mechanism of biased usage of synonymous codons in silent sites.

This could necessitate selecting appropriate hosts expression systems to improve or decrease the expression of target genes *in vivo* or *in vitro*. The codon usage profiles can be explored for designing degenerate primers and improve the accuracy of gene prediction from genomic sequences and protein functional classification thus maximizing the effectiveness of genetic manipulations in synthetic biology (Qian et al., 2012). The study of gene expression traits in the porcine genome is relevant to many fundamental biological processes including species and breed diversity, gene expression and evolution, and adaptation to micro environment. The objective of the study was to determine determinants of codon usage bias in the pig genome to further decipher plasticity of genes.

MATERIALS AND METHODS

Sequence data

Two complete genome sequences were used for analysis. A total of 23,269 coding sequences was extracted from the female Duroc pig breed as the reference genome, (Sus scrofa build 10.2) using BioMart (Ensembl v 68). Only 21,550 coding sequence (CDS) that were more than 50 amino acids (150 bp) were included for analysis. The short CDS were excluded due to large estimation errors for codon usage which are associated with short sequence length. The majority of the excluded genes were microRNAs which are averagely 22 bp in length. The second data was extracted from gene coordinates and Ssc 10.2 reference genome.

Codon indices

Relative synonymous codon usage (RSCU)

Relative synonymous codon usage (RSCU) is the proportion of the observed codon divided by its expected frequency at equilibrium. An RSCU value close to 1 indicates lack of bias, RSCU >1 indicates a codon used more frequently than expected, and RSCU < 1 indicates a codon used less frequently than expected (Sharp et al., 2005).

RSCU values are largely independent of amino acid composition and are particularly useful in comparing codon usage among genes, or sets of genes that differ in their size and amino acid composition. In this study we developed an in house Perl script to calculate RSCU as;

$$\text{RSCU} = j \times \frac{Ci}{\sum_{i=1}^{j} Ci}$$

Where, C denotes actual codon counts, j denotes number of synonyms and i denote the codon counts within the synonyms.

Table 1. Comparison of GC and length of both CDS and gene. The GC content of CDS is higher than the content of genes.

Parameter	CDS and gene	Minimum	1st Quad	Mean	3rd Quad	Maximum
GC content	gene	0.0818	0.4065	0.4789	0.47890	0.9800
	CDS	0.2852	0.4587	0.5312	0.6013	0.8123
Length bp	gene	150	1422	27338	1722	830146
	CDS	152	639	1415	1722	22503

We finally derived mean RSCU as;

$$RSCUmean = \frac{\sum_{I=1}^{I} RSCU}{N}$$

Where, *RSCU* is the values derived per gene, *N* is the total number of codon counts per gene. The value of RSCU_Mean ranged from one to infinity depending on the biasness of the gene. It was assumed to be a sister index to genomic Codon Adaptation Index (gCAI) that is explained below.
Another parameter genomic RSCU (RSCU') was calculated as;

$$RSCU' = \frac{RSCU}{n}$$

Where, n is the number of synonymous codons of an amino acid. This parameter measures and compares the usage of all 61 sense codons and is the proportion of use of a codon in all genes.

Genomic codon adaptation index (gCAI)

Classical Codon adaptation index (CAI) was first used to measure gene expression. This measure is species dependent and is the empirical measure for gene in studies investigating mutational and selectional components of codon usage (Goetz and Fuglsang, 2005). A CAI value is always between 0 and 1, and a higher value means a stronger codon usage bias and higher expression level and / or translation efficiency. In most research CAI of a coding sequence (CDS) is computed from the two parameters; the codon frequencies of the CDS and the codon frequencies of a set of known highly expressed genes (often referred to as the reference set). This computation leads to CAI which is used as a proxy for gene expression. In this case CAI values are normalized using codon frequencies in highly expressed gene sets. According to Xia (2007), CAI computation involves first derivation of a column of W values;

$$Wif = \frac{Fij.ref}{MAXfi.ref}$$

Where, Fij.ref is the frequency of codon j in synonymous codon family i, and MAXfi.ref is the maximum codon frequency in synonymous codon family i from a set of highly expressed genes.

The codon adaptation index for a given gene is then given by:

$$CAI = \prod_{i=1}^{L} Wif^{1/L}$$

Where, L is the number of codons from synonymous families in the gene.

In our study, genomic codon adaptation index (gCAI) is calculated as the geometric mean RSCU divided by the highest possible geometric mean of RSCU given the same Amino Acid sequence.

$$gCAI = \frac{\sqrt[n]{\prod_{i=1}^{n} rscu}}{\sqrt[m]{\prod_{i=1}^{m} RSCU}}$$

This value (gCAI) is a proxy for codon bias but not gene expression. This is because the gCAI values are normalized using codon frequencies at equilibrium, thus there is no assumption of gene expression bias.

Analysis tools

An in house Perl script was used to derive codon indices, gene length, GC and GC3 (the frequency of G+C at the third position) for all the CDS. Statistical analysis was conducted using R (V 2.15.0). We used a Sperman's rank correlation to relate codon indices (gCAI, RSCU) with different nucleotide composition variables (that is, GC, GC3 and CDS length). Multivariate regression model was used to predict the biasness and determine contribution of genomic factors to the biasness.

RESULTS

Variation in the CDS length, GC content and GC3s in the Coding sequence

The coding sequence GC content ranged from 0.285 to 0.812 with a mean of 0.531. For the CDS length, the shortest and the longest gene were 151 and 45618 bp respectively, with a mean of 1415 bp. The comparison of GC content between genes (intron and exons) and CDS were on average 47 and 53%, respectively. The mean CDS and gene lengths were 1415 and 27338 bp (Table 1). This confirms that the CDS are generally GC richer than the genes.

Codon usage bias analyses

The observed relative synonymous codon usage (RSCU) clearly indicated that there was a nonrandom usage of

synonymous codons for individual genes (Table 2). To investigate if the observed biasness, favoring specific codons, were beyond specific genes, we performed an overall genome wise analysis by concatenating all the genes into one large sequence string. The rationale was to exclude factors specific for individual genes. Preference of certain synonymous codons was observed in Figure 1. Table 2 shows the variation in RSCU values across codons coding for the same amino acid. This table highlights the biasness for a representative gene. For example valine a four degenerate amino acid has more preference for GTG = 1.5688 and GTC = 1.0139 than GTT = 0.8546 and GTA = 0.565. For Aspartate, GAC = 1.0598 was more preferred than GAT = 0.9401. Figure 1 depicts the codon usage of serine (RSCU) and may act as a representative of all synonyms. In this figure, the codon AGC, TCC and TCA were the most preferred in that order. To compare the usage of the 61 codons, RSCU' was used. In this analysis, the codons from two degenerate amino acids had higher values (for example, TAC = 5.50 x 10^{-6} and TAT = 4.96 x 10^{-6}) as compared to three fold degenerate amino acids (for example, ATC = 3.09 x 10^{-6}, ATT = 3.09 x 10^{-6}, ATA = 2.08 x 10^{-6}). Generally it was noticed that the usage of the codons reduced with the increase of the degeneracy with the six fold degenerate having the lowest usage codons or observed bias.

Correlation between codon indices and nucleotide composition

We found a significant correlation between nucleotide content and codon bias indices. The genomic RSCU (RSCU') correlated positively with the GC content (r^2 = 0.796, p = 2.0e-16) and GC3 (r^2 = 0.162, p = 2.0e-16) but negatively with the CDS length (r^2 = 0.84 p = 2.0e-16). Contrary to our expectation the genomic codon adaptation index (gCAI) significantly correlated negatively with the GC content (r^2 = -0.355, p = 2e-16), GC3 (r^2 = -0.321) and the CDS length (r^2 = -0.773, p = 2e-16) (Table 3). It is worth noting the two indices gCAI and genomic RSCU only differs in mathematical calculation with former using geometric means while the later using arithmetic mean.

We further fitted a model to predict the codon usage bias (CUB) using nucleotide composition factors as independent variables. The fitted model;

[$y = \beta_o + \beta_1$ GC% + β_2 CDS Length +β_3GC3 + error] was used.

Where, y is either gCAI or genomic RSCU.

In this model all the factors/variables were highly significant. The model explained 29% of the observed gCAI (Table 4). Almost similar results were realized with genomic RSCU except the change in the sign of the coefficients. As can be seen in Table 5, there was negative association between genomic RSCU with GC content, CDS length and the intercept. However, this parameter was positively associated with GC3 and GC frequencies. It is worth noting that coefficients had minimal effect. This showed that nucleotide composition factors play minor but significant roles in shaping codon bias.

DISCUSSION

Previous analyses of codon usage in different taxa have suggested that there exists a huge interspecific variation and clear intragenomic variability (Gagnaire et al., 2012) and this study is no exception. Several biological factors such as tRNA abundance, strand-specific mutational bias, gene expression level, gene length, amino acid composition, protein structure, mRNA structure, nucleotide composition, intron splicing, recombination, gene conversion, DNA packaging, intron number (Qin et al., 2013) and selection for increased translational efficiency or accuracy have been demonstrated to relate to codon usage bias (CUB) (Ingvarsson, 2007). Despite abundance of these reports very few studies have focused on mammalian genomes. It is commonly accepted that both natural selection and genetic drift shape CUB across taxa (Zhao et al., 2007). Selection of codon bias is generally viewed as being weak; therefore, it is expected that selective forces, such as purifying selection against unfavored codons, should be more prominent in organisms with large effective population size (Ne) such as prokaryotes and unicellular eukaryotes or even fruit flies.

Species with low Ne are expected to be more prone to genetic drift and therefore, should show relaxed selective pressure on codon usage. In order to examine synonymous codon usage in the pig genome we first deciphered the genomic composition of the genes (intron and exons) and the CDS as well, followed by analysis of the CUB. We hereby present evidence suggesting that the pattern of synonymous codon choices in the S*us scrofa* is as a result of a complex equilibrium between different forces, namely the natural selection at the translational level, nucleotide compositional, mutation bias and the length of each gene.

We report conclusive evidence for codon usage bias in the pig genome. The CUB is evident in the nonrandom usage of synonymous codons as shown by the codon indices. This finding is consistent with other studies involving prokaryotes (Karlin et al., 1997) and eukaryotes (Waldman et al., 2011). The observed preference of some codons could be suggestive of a weak selection force acting on codon pool in the pig genome. The observed

Table 2. Codon usage table of a representative gene (ENSSSCG0000000015) showing biased synonyms of 20 amino acids. The numbers in bold are the subtotal synonyms per amino acid.

Amino acid	Codon	Codon count	RSCU
S (Serine)	TCC	16	1.1707
S	AGC	18	1.3170
S	TCA	12	0.8780
S	TCG	7	0.5121
S	TCT	18	1.3170
S	AGT	11	0.8048
Subtotal		82	
F (Phenylalamine)	TTT	20	1.4814
F	TTC	7	0.5185
Subtotal		27	
T (Threonine)	ACT	8	0.6956
T	ACC	19	1.6521
T	ACA	12	1.0434
T	ACG	7	0.6086
Subtotal		46	
N (Asparagine)	AAC	20	1.0810
N	AAT	17	0.9189
Subtotal		37	
Y (Tyrosine)	TAC	10	1.6666
Y	TAT	2	0.3333
Subtotal		12	
E (Glutamate)	GAA	22	0.9361
E	GAG	25	1.0638
Subtotal		47	
V (Valine)	GTT	10	0.9090
V	GTC	13	1.1818
V	GTG	17	1.5454
V	GTA	4	0.3636
Subtotal		44	
Q (Qlutamine)	CAG	16	1.6000
Q	CAA	4	0.4000
Subtotal		20	
M (Methionine)	ATG	8	1.0000
K (Lysine)	AAA	25	1.0000
K	AAG	25	1.0000
Subtotal		50	
C (Cysteine)	TGC	8	1.6000
C	TGT	2	0.4000
Subtotal		10	
L (Leucine)	TTG	11	1.1379
L	CTT	2	0.2068
L	CTA	2	0.2068
L	CTG	22	2.2758
L	TTA	10	1.0344
L	CTC	11	1.1379
Subtotal		58	

Table 2. Contd.

A (Alaline)	GCG	5	0.3508
A	GCT	10	0.7017
A	GCC	29	2.0350
A	GCA	13	0.9122
Subtotal		57	
W (Tryptophan)	TGG	4	1.0000
P (Proline)	CCA	6	0.5714
P	CCC	19	1.8095
P	CCT	7	0.6666
P	CCG	10	0.9523
Subtotal		42	
H (Histidine)	CAT	5	1.1111
H	CAC	4	0.8888
Subtotal		9	
D (Aspartate)	GAT	14	0.8750
D	GAC	8	1.1250
Subtotal		22	
I (Isoleucine)	ATA	3	0.4090
I	ATT	5	0.6818
I	ATC	14	1.9090
Subtotal		22	
R (Arginine)	AGA	5	1.3636
R	CGG	6	1.6363
R	CGA	1	0.2727
R	AGG	7	1.9090
R	CGT	1	0.2727
R	CGC	2	0.5454
Subtotal		22	
G (Glycine)	GGC	25	1.6949
G	GGG	11	0.7457
G	GGA	12	0.8135
G	GGT	11	0.7457
Subtotal		59	

CUB is further proved to be influenced significantly by nucleotide composition.

However, in contrast to other papers (Rao et al., 2011), we report negative correlation between genomic codon adaptation index (gCAI) or CUB and the CDS length. The GC content and GC3s were consistent with their findings. In humans, the GC content and mutational biases were reported as major factors that influence codon usage. In plants several factors like nucleotide composition of genes, the levels of gene expression and length of the coding sequence contributed to the observed codon usage bias.

In *B. pseudomallei* genome, highly expressed genes had the highest GC content and it tended to use G or C at the third position of the codon (Hershberg and Petrov, 2010). The highly expressed genes in *B. pseudomallei* also had high GC content positively correlated with CAI value and GC3s. Their result purport that the highly expressed genes tend to use ‘C’ or ‘G’ at synonymous positions compared with lowly expressed genes. In this study our results points to preferred usage of both C or G and A or T at the synonyms sites as shown in Table 2, with the C or G ending codons being the majority.

However a negative correlation between gCAI and GC content or GC3s is unique. This might be due to the difference in the genome isochore structure, ambiguity (vary with space and time) of the gene expression in mammals, or due to difference in methodology of calculating CAI variants The negative correlation found between gCAI and gene length is consistent with other

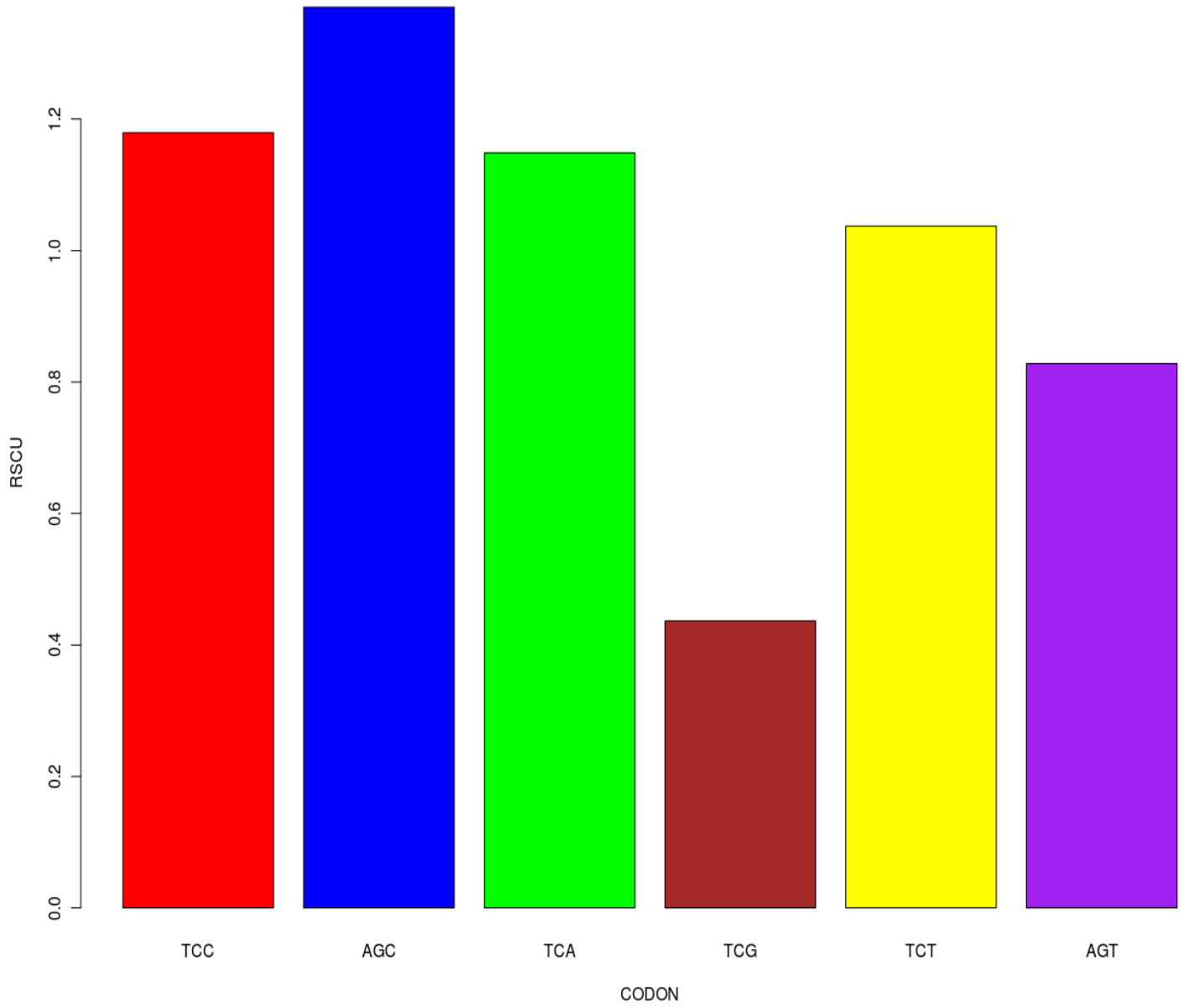

Figure 1. The relative synonymous codon usage of serine showing codon AGC and TCG as the most and the list preferred codons respectively.

Table 3. The correlation between the genome composition factors and the selected codon indices.

RSCU'	gCAI	GC% Content	CDS Length	GC3 Ratio	GC3 Counts	CODON counts
RSCU'	0.725***	0.796***	-0.840***	0.162***	-0.733***	-0.847***
gCAI	1	-0.355***	-0.773***	-0.321***	-0.823***	-0.777***
GC% Content		1	0.066***	0.914***	0.367***	0.066***
CDS Length			1	0.008***	0.931***	0.999***
GC3 Ratio				1	0.338***	0.008***
GC3 Counts					1	0.931***
CODON counts						1

*Denotes level of significance

Table 4. The coefficients of the multivariate regression analysis explaining the genomic composition factors affecting genomic codon adaptation index.

Coefficients	Estimates	Std. error	t- value	p-value
Intercept	4.153e-02	1.308e-03	31.750	< 2e-16
GC_CONTENT	-9.768e-02	3.240e-03	-30.149	< 2e-16
CDS_LENGTH	1.179e-06	2.474e-07	4.764	1.91e-06
GC3_RATIO	-5.018e-02	1.823e-03	-27.531	< 2e-16
GC3_COUNTS	-1.117e-05	1.273e-06	-8.774	< 2e-16

Table 5. The coefficients of the multivariate regression analysis explaining the genomic composition factors affecting the genomic RSCU (RSCU').

Coefficients	Estimates	Std. error	t- value	p-value
Intercept	-1.225	5.763e-04	-21.247	< 2e-16
GC_CONTENT	-6.091e-02	1.461e-03	-41.690	< 2e-16
CDS_LENGTH	-1.002e-06	1.114e-07	-8.997	< 2e-16
GC3_RATIO	2.078e-02	8.206e-04	25.318	< 2e-16
GC3_COUNTS	4.997e-06	5.724e-07	8.730	< 2e-16

reports in organism such as yeast, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Populus tremula and Silene latifolia* (Qiu et al., 2011). Previous studies have shown that metabolic systems prefer to express those genes that are less costly (Hahn and Kern, 2005). Moreover, there have been reports of longer genes having higher expression level, CAI values and higher codon usage bias in some unicellular genomes, specifically *E. coli* (Stoletzki and Eyre-Walker, 2007). These contradicting reports (positive and negative correlation) indicate there are no universal rules about gene length and codon bias. In this study, the longer genes had lower gCAI values or lower codon usage bias. The general consensus is that there should exist a positive correlation between gCAI values and gene length, which could be explained by selection of the preferred codons to avoid errors during translation. Since the cost of producing a protein is proportional to its length, selection in favor of codons which increase accuracy should be greater in longer genes, and long genes should therefore have higher synonymous codon bias.

In such genes, by using optimal codons, translation is faster whereby ribosomes move faster along the mRNA and are released quickly to be available to translate other mRNA (Zhao et al., 2007). The use of optimal codons increases the accuracy of translation by reducing translational errors that can occur. The errors include missense in which an incorrect amino acid is incorporated into the growing peptide chain and nonsense in which the peptide synthesis terminates prematurely by incorporating stop codons. It is believed that both missense and nonsense errors that produce non- and misfunctional proteins respectively, are costly to the cell because they consume amino acids and energy both in their production and during breakdown (Stoletzki and Eyre-Walker, 2007). Besides, missense errors may have other serious consequences, for example, a missense error in a DNA polymerase may temporally increase the mutation rate (Ninio, 1991).

Pig genome just like other mammals is found to vary greatly in base composition between different genomic regions. In vertebrates, such as mammalian and birds, one of the most striking features of their genomes is the difference in G+C contents isolated regions called isochore structures. In pig there exists heterogeneity in G+C content that results in variation in codon usage bias as was revealed elsewhere (Hershberg and Petrov, 2010).

Having a relatively high GC content, we expected the pig preferred codons to mirror the genome composition. GC rich organisms tend to have GC rich optimal codons, while AT rich organisms tend to have AT rich optimal codons. This observation is manifested in RSCU as most preferred codons end with G or C albeit with some ending with A or T. This phenomenon is dependent on the isochore structure of the pig genome that we confirmed by observed variation in GC content. The data analyzed provide evidence for the mutational bias hypothesis. In our view the codon bias is skewed towards the AT ending codons as was revealed by inverse correlation between gCAI and GC3s. Indeed, AT rich genes were shorter in length and could imply efficient protein translation to minimize energy consumption. We also suspect that

other factors besides mutation bias may have contributed to codon usage. Amongst the other factors, we hypothesize that selection for preferred codons is affected by the abundance of tRNA in the cells or the ones that bind those tRNAs with optimal binding strength (Ikemura, 1985; Kanaya et al., 2001). We could not confirm this due to lack of information on tRNA . However, this hypothesis has been proven in other organisms like *E. coli, B. subtilis and C. cerevisae.* In that study cellular tRNAs correlates positively and closely to tRNA gene copy numbers; by extension this suggests that in these species there is correlation between optimal codon use and tRNAs abundance. However such correlation was not found in studies involving *D. melanongaster* and humans (Kanaya et al., 2001).

The positive correlation observed between GC content, GC3 and gene length explains the computed low codon bias. This is because long genes tend to have more G and C, abundant G or C at the third codons which are negatively related to gCAI. The nucleotide composition factors only play significant but minor roles in shaping the codon usage in the pig genome as revealed in low R^2 value and statistical interpretation exhibited in multivariate regression analysis. These statistical inferences are clear indication that the pig genome is so complex and molecular functions are controlled by several factors.

Conclusions

We confirm the existence of codon usage bias in the porcine genome which might suggest there is weak selection of preferred codons for translation accuracy. The codon usage bias is influenced slightly by nucleotide composition factors among others.

Conflict of interests

The authors have not declared any conflict of interests.

REFERENCES

Basak S, Ghosh TC (2005). On the origin of genomic adaptation at high temperature for prokaryotic organisms. Biochem. Biophys. Res. Comun. 330:629-632.

Behura SK, Severson DW (2013a). Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. Biol. Rev. 88:49-61.

Bulmer M (1991). The Selection-Mutation-Drift Theory of Synonymous Codon Usage. Gen.129:897-907.

Gagnaire PA, Normandeau E, Bernatchez L (2012). Comparative genomics reveals adaptive protein evolution and a possible cytonuclear incompatibility between European and American eels. Mol. Biol. Evol. 29(10):2909-2919.

Goetz RM, Fuglsang A (2005). Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. Biochem. Biophysic. Res. Commun. 327:4-7.

Hahn MW, Kern AD (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol. 22:803-806.

Hershberg R, Petrov DA (2010). Evidence That Mutation Is Universally Biased towards AT in Bacteria, PLoS Genet. e1001115.

Ikemura T (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2:13-34.

Ingvarsson PK (2007). Gene Expression and Protein Length Influence Codon Usage and Rates of Sequence Evolution in *Populus tremula*. Mol. Biol. Evol. 24:836-844.

Jia R, Cheng A, Wang M, Xin H, Guo Y, Zhu D, Qi X, Zhao L, Ge H, Chen X (2009). Analysis of synonymous codon usage in the UL24 gene of duck enteritis virus. Virus Genes. 38:96-103.

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T (2001). Codon usage and tRNA genes in eukaryotes: Correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. 53:290-298.

Karlin S, Mrázek J, Campbell AM (1997). Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. 179:3899-3913.

Knight R, Freeland S, Landweber L (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. 2:1001-1013.

Morton BR, Wright SI (2007). Selective constraints on codon usage of nuclear genes from *Arabidopsis thaliana*. Mol. Biol. Evol. 24:122-129.

Ninio J (1991). Transient mutators: a semiquantitative analysis of the influence of translation and transcription errors on mutation rates. Genetics. 129:957-962.

Parmley JL, Hurst LD (2007). Exonic Splicing Regulatory Elements Skew Synonymous Codon Usage near Intron-exon Boundaries in Mammals. Mol. Biol. Evol. 24:1600-1603.

Qian W, Yang JR, Pearson NM, Maclean C, Zhang J (2012). Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet. 8. e1002603.

Qin Z, Zhengqiu Cai Z, Guangmin Xia G, Wang, M (2013). Synonymous codon usage bias is correlative to intron number and shows disequilibrium among exons in plants. BMC Genomics 14:56-67.

Qiu S, Bergero R, Zeng K, Charlesworth D (2011). Patterns of codon usage bias in *Silene latifolia*. Mol. Biol. Evol. 28:771-780.

Rao Y, Wu G, Wang Z, Chai X, Nie Q, Zhang X (2011). Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. DNA Res.18:499-512.

Reis MD, Savva R, Wernisch L (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 32:5036-5044.

Sharp PM, Bailes E, Grocock RJ, Peden F, Sockett RE (2005). Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 33:1141-1153.

Stoletzki N, Eyre-Walker A (2007). Synonymous Codon Usage in Escherichia coli: Selection for Translational Accuracy. Mol. Biol. Evol. 24:374-381.

Waldman YY, Tuller T, Keinan A, Ruppin E (2011). Selection for Translation Efficiency on Synonymous Polymorphisms in Recent Human Evolution. Genome Biol. Evol. 3:749-761.

Whittle CA, Sun Y, Johannesson H (2012). Genome-wide selection on codon usage at the population level in the fungal model organism *Neurospora crassa*. Mol. Biol. Evol. 29(8):1975-1986.

Wright F (1990). The ‘effective number of codons’ used in a gene. Gen. 87:23-29.

Zhang Q, Zhao S, Chen H, Liu X, Zhang L and Li, F (2009). Analysis of the codon use frequency of AMPK family genes from different species. Mol. Biol. Reports 36:513-519.

Zhao S, Zhang Q, Chen Z, Zhao Y, Zhong J (2007). The Factors Shaping Synonymous Codon Usage in the Genome of Burkholderia mallei. J. Gen. Genomics. 34:362-372.