DNA Research **18**, 499–512, (2011) doi:10.1093/dnares/dsr035
Advance Access Publication on 27 October 2011

Mutation Bias is the Driving Force of Codon Usage in the *Gallus gallus* genome

Yousheng Rao[1,2,*], Guozuo Wu[1], Zhangfeng Wang[1], Xuewen Chai[1], Qinghua Nie[2,3], and Xiquan Zhang[2,3,*]

Department of Biological Technology, Jiangxi Educational Institute, Nanchang, Jiangxi, China[1]; Guangdong Provincial Key Laboratory of Agro-animal Genomics and Molecular Breeding, South China Agricultural University, Guangzhou, Guangdong, China[2] and College of Animal Science, South China Agricultural University, Guangzhou, Guangdong, China[3]

*To whom correspondence should be addressed. Tel: 86 0791 88300376. Fax: 86 0791 83812190.
E-mail: rys8323571@yahoo.com.cn (Y.R.); xqzhang@scau.edu.cn (X.Z.)

Edited by Hiroyuki Toh
(Received 28 March 2011; accepted 21 September 2011)

Abstract

Synonymous codons are used with different frequencies both among species and among genes within the same genome and are controlled by neutral processes (such as mutation and drift) as well as by selection. Up to now, a systematic examination of the codon usage for the chicken genome has not been performed. Here, we carried out a whole genome analysis of the chicken genome by the use of the relative synonymous codon usage (RSCU) method and identified 11 putative optimal codons, all of them ending with uracil (U), which is significantly departing from the pattern observed in other eukaryotes. Optimal codons in the chicken genome are most likely the ones corresponding to highly expressed transfer RNA (tRNAs) or tRNA gene copy numbers in the cell. Codon bias, measured as the frequency of optimal codons (Fop), is negatively correlated with the G + C content, recombination rate, but positively correlated with gene expression, protein length, gene length and intron length. The positive correlation between codon bias and protein, gene and intron length is quite different from other multi-cellular organism, as this trend has been only found in unicellular organisms. Our data displayed that regional G + C content explains a large proportion of the variance of codon bias in chicken. Stepwise selection model analyses indicate that G + C content of coding sequence is the most important factor for codon bias. It appears that variation in the G + C content of CDSs accounts for over 60% of the variation of codon bias. This study suggests that both mutation bias and selection contribute to codon bias. However, mutation bias is the driving force of the codon usage in the *Gallus gallus* genome. Our data also provide evidence that the negative correlation between codon bias and recombination rates in *G. gallus* is determined mostly by recombination-dependent mutational patterns.

Key words: *Gallus gallus*; codon usage; mutation; selection

1. Introduction

Synonymous codons are used with different frequencies both among species and among genes within the same genome. Highly expressed genes (such as those encoding translation elongation factors and ribosomal proteins) tend to use optimal (preferred) codons and exhibit very high levels of codon bias.[1–5] The optimal codons also tend to correspond to highly expressed tRNAs and tRNA gene copy numbers.[4–13] These patterns have been interpreted as natural selection for more efficient and accurate translation.[6,14–18] In contrast, some studies have demonstrated that the first factor shaping codon usage is nucleotide composition (G + C content) of genes and intergenic regions.[12,19–22] As G + C

content is more likely determined by genome-wide processes rather than by selective forces acting specifically on coding sequence, these findings have been inferred to reflect the genome-wide patterns of codon usage by mutational biases. Based on the fact that both mutational pressures and selective forces are involved in the phenomenon of codon bias in a variety of organisms, an integrated model, known as the mutation–selection–drift balance model, has been proposed.[15,23–25] This model proposes that selection favours optimal codons over minor codons, while mutational pressure and genetic drift allow the minor codons to persist.[17] Population genetics has shown that the selection of codon bias is generally weak[17,23,26] ($|N_e s| \approx 1$), therefore, selection coefficients are expected to be more efficient in species with large effective population sizes (N_e) such as prokaryotes and unicellular eukaryotes.[6,27] In species with low N_e values, genetic drift should be the main force shaping codon usage and overpowering translational selection of codon variants.

Codon bias has been determined to be positively correlated with recombination rates in *Drosophila*, as well as in many other species.[28–35] This observation has been explained by two hypotheses. The first proposed that the reduction of codon bias in the regions with limited recombination is consistent with Hill–Robertson interference.[28,30] However, another hypothesis, called the GC-biased gene conversion model, suggested that the correlation between recombination and codon usage patterns is caused by recombination-related mutational bias rather than by Hill–Robertson interference, as the heteroduplex DNA appears to be biased toward the preferential fixation of AT → GC mutations.[32,36] Except for nucleotide composition, gene expression and recombination rates, other additional parameters such as protein length, gene length and intron length also have been found to play an important role in shaping codon usage in a wide variety of organisms.[2,5,37–39]

The chicken (*Gallus gallus*) is an important model organism that bridges the evolutionary gap between mammals and other vertebrates. The chicken karyotype comprises 39 pairs of chromosomes, which are divided into 8 pairs of cytologically distinct macrochromosomes, Z and W sex chromosomes, and 30 pairs of micro-chromosomes.[40] Compared to other vertebrate genomes, the chicken genome has many distinctive characteristics such as a smaller genome size (less than half of humans and mouse), higher recombination rates and higher G + C content.[41,42] Base composition is found to vary greatly between different genomic regions in many eukaryotes. In vertebrates, such as mammalian and birds, one of the most striking features of their genomes is the variation of G + C content that occurs over scales of hundreds of kilobases to megabases, the so-called 'isochore structure'.[43,44] Although subsequent study indicated that the isochore model might need slight revision,[45] it is clear that the genomes of vertebrates are highly heterogeneous in G + C content and have acquired GC-rich regions.[42] This results in that a large proportion of variance in codon usage bias is explained by G + C content.[20,21] Mank *et al*. investigated the chicken's properties of sex-biased genes (female biased genes 155, male biased genes 286) through a microarray data. They found that the codon usage of sex-biased genes showed some sex-biased effects, primarily for autosomal genes expressed in the gonad. Codon bias is greatest when GC_3 (the G + C content at third coding positions) is skewed away from equal usage of GC or AT.[46] Up to now, a systematic examination of the codon usage for the *G. gallus* genome has not been performed. In the present study, we carried out a whole analysis of the chicken genome and showed that codon bias is negatively correlated with G + C content and recombination rates, but positively correlated with tRNA abundance, gene expression, protein length, gene length and intron length. Our data clearly displayed that regional G + C content explains a large proportion of the variance of codon bias in *G. gallus* genome. This study will benefit our understanding of how natural selection and mutation impacts codon usage in the *G. gallus* genome.

2. Materials and methods

2.1. Sequence data

Only nuclear genes with complete information on protein-coding sequences (CDSs) with no evidence of multiple-splicing forms were included in this study. CDSs corresponding to all annotated genes in the chicken genome were downloaded from http://www.ncbi.nlm.nih.gov/sites/gquery and peptide information coded by the genes was derived from http://www.ncbi.nlm.nih.gov/protein. Some CDSs lengths are obviously not consistent with the total length of corresponding exons and these genes were defined as annotation errors and were not included for analysis. CDSs that did not begin with an ATG start codon, did not have a length of >300 bp, did not contain a multiple of three or that contained an internal stop codon were also ruled out. The final sequence collection contained 8631 CDSs with each corresponding to a unique gene in the *G. gallus* genome. For each gene, total gene length, protein length, first intron length and average intron length were determined.

2.2. Expression data

Expression databases were taken from the NCBI FTP website (http://www.ncbi.nlm.nih.gov/sites/entrez) and a total of 633 321 expressed sequence tag (EST) sequences were available. We used the number of EST sequences in this database that align unequivocally to a given gene, and compared the set of chicken mRNA/cDNA sequences with the ESTs using the program BLASTN. We accepted EST hits of >400 nt and with $>96\%$ identity to a mRNA/cDNA sequence as matches. If they showed $>98\%$ identity, we accepted hits of 100–400 nt and discarded hits of <100 nt.[47,48] An EST matched to multiple genes was discarded. After excluding genes with multiple-splicing forms and genes with obvious annotation errors, the data on the 8631 genes from 18 tissues (including the blood, brain, bursa of fabricius, cecum, connective tissue, embryonic tissue, epiphyseal growth plate, gonad, head, heart, limb, liver, muscle, ovary, pancreas, spleen, testis and thymus) were taken into account. Tags per million were then calculated for each gene in each tissue. Total expression levels are defined as ESTs of a gene in the total number of tissues. Expression breadth is defined as the number of tissues in which the ESTs were found. The tissue specificity index (τ) is measured by both qualitative (i.e. presence/absence) and quantitative variations of expression levels among tissues, and is defined as:

$$\tau = \frac{\sum_i^N (1 - x_i/x_{\max})}{N - 1}$$

where N is the number of tissue samples examined, x_i is the expression level of the gene in sample i and $x_{\max}$ is the highest expression level of the gene across the N samples examined.[49]

2.3. Identification of optimal codons and synonymous codon usage

Optimal codons are defined as those that occur significantly more often in highly expressed genes relative to their frequency in lower expressed genes. We used 5% of the total genes with extremely high and low expression levels inferred from EST counts, as the high and low data set, respectively, and calculated the average RSCU (relative synonymous codon usage) of the two gene samples. RSCU was calculated by dividing the observed codon usage by that expected when all codons for the same amino acid are used equally. RSCU values close to 1.0 indicate a lack of bias. Putative optimal codons were inferred based on departures from equal codon usage by sets of loci with high and low gene expression.[2,50] ΔRSCU for a given codon is the difference between the average RSCU of genes with high and low expression [significance tested using the one-way analysis of variance (ANOVA) by SAS]. If ΔRSCU is >0.1 at $P < 0.05$, this codon will be identified as an optimal codon. Then, we calculated Fop values using the codonW 1.4.2 program with customized optimal codon tables (J Peden, http://codonw.sourceforge.net). Fop is the ratio of optimal codons to synonymous codons, ranging from 0 (where no optimal codons are used) to 1 (where only optimal codons are used). The nucleotide composition indices including GC_3 and G + C content of CDSs were also calculated using codonW 1.4.2.

2.4. tRNA gene copy number data

The tRNA gene copy numbers for each codon in the *G. gallus* genome was taken from http://gtrnadb.ucsc.edu/Ggall/. In these data, pseudogenes have already been removed. We used tRNA gene copy numbers as an assumed estimate of cellular tRNA abundance. The relative gene frequency (RGF) of tRNAs is the observed frequency of an isoacceptor tRNA gene in the genome divided by the frequency expected if all isoacceptor tRNA genes for that amino acid occurred with equal frequencies.[11]

2.5. Recombination rate estimation

The recombination rates for 1 Mbp windows were estimated. The versions of the genome assemblies (NCBI build 2.1, released November, 2006) and the latest chicken consensus linkage map (sex-averaged map) were used.[51] This high-resolution consensus map included 9268 markers, consisting of 34 linkage groups. It enabled us to estimate the local recombination rates using a narrower region. Locations of individual markers were determined based on alignments of the full sequence of the markers using BLAST. The linear function was fitted to the points representing genetic and physical map positions in the 1 Mbp windows. The slope of this line was interpreted as an estimate of recombination rates.[52] Windows were removed that contained $>50\%$ '*N*' in the sequence assembly, as were windows at the beginning, end and centromere of chromosomes with no markers detected in them. Some windows with large discrepancies between the genetic map and the sequence assembly were also removed. A total of 745 windows were included, which covered ~70% of the chicken genome.

2.6. Statistical analysis

Correlation analysis between variables was performed by SAS Proprietary Software Release 8.1. In order to assess the actual strength of association, correlation coefficients reported in this study were obtained

Table 1. The putative optimal codons and tRNA abundance

Acid	Codon	tRNA gene	High	Low	Acid	Codon	tRNA gene	High	Low
Phe	UUU	AAA (1)	0.92	0.90	Ser	**UCU****	AGA (8)	1.32	1.07
	UUC	GAA (9)	1.08	1.10		UCC	**GGA** (NA)	1.21	1.26
Leu	UUA	TAA (2)	0.40	0.47		UCA	TGA (4)	0.90	0.90
	UUG	CAA (2)	0.88	0.83		UCG	CGA (2)	0.33	0.33
	CUU*	GAA (3)	0.95	0.78	Pro	**CCU***	AGG (6)	1.30	1.15
	CUC	**GAG** (NA)	1.01	1.15		CCC	**GGG** (NA)	1.05	1.18
	CUA	TAG (1)	0.39	0.40		CCA	TGG3	1.23	1.20
	CUG	CAG (6)	2.36	2.37		CCG	CGG2	0.42	0.45
Ile	**AUU***	AAT (7)	1.16	0.97	Thr	**ACU***	AGT (4)	1.14	1.00
	AUC	**GAT** (NA)	1.34	1.45		ACC	**GGT** (NA)	1.13	1.22
	AUA	TAT (2)	0.48	0.57		ACA	TGT (4)	1.21	1.26
Met	AUG	13	1.00	1.00		ACG	CGT (2)	0.51	0.52
Val	**GUU***	AAC (7)	0.98	0.81	Ala	**GCU***	AGC (19)	1.34	1.22
	GUC	**GAC** (1)	0.83	0.92		GCC	GGC (1)	1.13	1.22
	GUA	UAC (3)	0.57	0.52		GCA	TGC (7)	1.13	1.18
	GUG	CAC (5)	1.63	1.75		GCG	CGC (2)	0.40	0.38
Tyr	UAU	ATA (NA)	0.83	0.76	Cys	**UGU***	ACA (NA)	0.90	0.80
	UAC	GTA (6)	1.12	1.20		UGC	**GCA** (10)	0.94	1.17
Stop	UAA		1.40	0.80	Stop	UGA		1.69	1.31
	UAG		0.70	1.15	Trp	UGG		1.00	1.00
His	CAU	ATG (1)	0.84	0.80	Arg	**CGU****	ACG (6)	0.87	0.48
	CAC	GTG (7)	1.09	1.18		CGC	**GCG** (NA)	1.00	0.95
Gln	CAA	TTG (3)	0.52	0.57		CGA	TCG (2)	0.58	0.59
	CAG	CTG (3)	1.47	1.42		CGG	CCG (2)	0.78	0.96
Asn	AAU	ATT (NA)	0.87	0.83	Ser	AGU	ACT (NA)	0.90	0.85
	AAC	GTT (9)	1.12	1.16		AGC	GCT (6)	1.34	1.59
Lys	AAA	TTT (3)	0.86	0.87	Arg	AGA	TCT (3)	1.53	1.57
	AAG	CTT (6)	1.14	1.11		AGG	CCT (3)	1.23	1.46
Asp	**GAU***	ATC (NA)	1.08	0.95	Gly	**GGU****	ACC (NA)	0.92	0.64
	GAC	**GTC** (8)	0.91	1.03		GGC	**GCC** (5)	1.11	1.22
Glu	GAA	TTC (7)	0.94	0.86		GGA	TCC (8)	1.20	1.15
	GAG	CTC (7)	1.06	1.13		GGG	CCC (4)	0.77	0.99

Putative optimal codons were inferred based on departures from equal codon usage by sets of loci with high (5% top) and low (5% down) gene expression (△RSCU). △RSCU for a given codon is the difference between the average RSCU of genes with high and low expression (significance tested using the one-way ANOVA). If △RSCU is >0.1 at $P < 0.05$, this codon will be identified as optimal codon. Total optimal codons identified in this study are 11. The transfer RNA gene (tRNA) copy numbers for each codon was taken from http://gtrnadb.ucsc.edu/Ggall/. There is a good correspondence between tRNA abundance and optimal codons within codon classes. However, the above correlation reflects only partially the real co-adaptation of tRNA abundance and codon usage, as the same tRNA can decode several codons. Since we have no experimental data on base modifications in *Gallus gallus* tRNAs, we predicted the codons decoded by the different anticodons according to the 'parsimony of wobbling' criterion. The putative optimal codons with △RSCU is >0.1 at $P < 0.05$ and △RSCU >0.2 at $P < 0.05$ are denoted by '*' and '**', respectively.

using all genes independently and avoided the approach of subdividing genes into groups to later investigate relationships among them. The significance tests were corrected for multiple testing by the Bonferroni step-down correction.[53] To determine the variables contributing to codon bias and how they may interact, we performed multiple linear regressions with the variables, excluding those not contributing significantly through the use of the *t*-statistical logarithm with backward stepwise regression.

3. Results

3.1. Putative optimal codons and tRNA abundance

A total of 8631 genes were included in this study. As shown in Table 1, 11 codons have been identified as putatively optimal. Interestingly, all putative optimal codons in the chicken genome are ended by uracil (U). Previous studies suggested that the nucleotide composition (G + C content) plays an important role in the identities of optimal codons as selection for

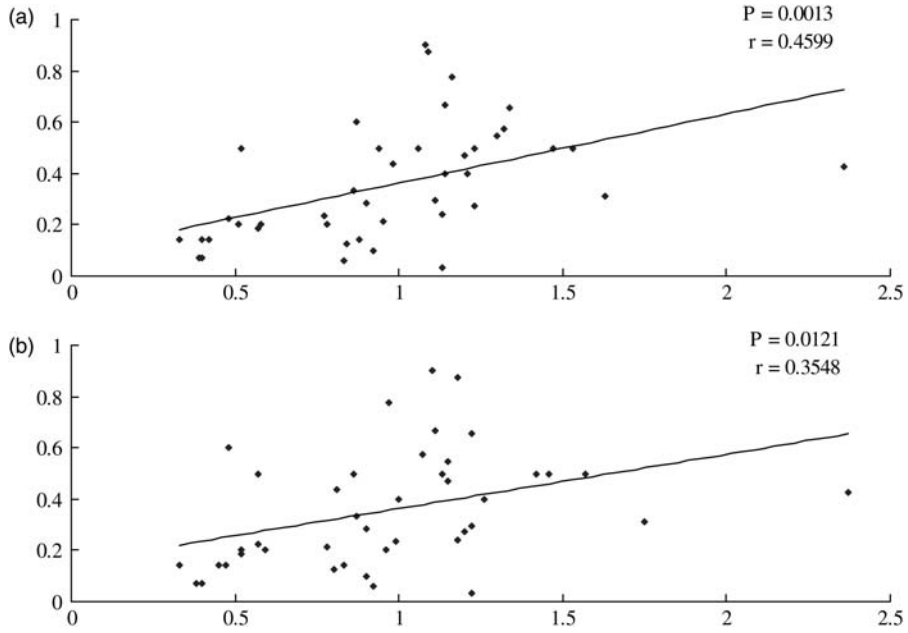

Figure 1. Scatter plots of RGF versus RSCU. The RGF of tRNA genes is the observed frequency of an isoacceptor tRNA gene in *Gallus gallus* genome divided by the frequency expected if all isoacceptor tRNA genes for that amino acid were equally frequent in the genome. The RSCU is the observed frequency of a codon divided by the frequency expected if all synonyms for that amino acid were used equally. (a) RSCU was measured in the highly expressed genes and (b) RSCU was measured in the lowly expressed genes.

optimal codons for transcription and translation is not high enough to overcome compositional skews.[12,20,21] Base composition is found to vary greatly between different genomic regions in many eukaryotes. In vertebrates, such as mammalian and birds, one of the most striking features of their genomes is the variation of G + C content that occurs over scales of hundreds of kilobases to megabases.[43,44] In order to further test whether the G + C content had a significant effect on the identities of optimal codons in the *G. gallus* genome, we produced a high G + C content sample (20% of the highest G + C content of the CDSs) and a low G + C content sample (20% of the lowest G + C content of the CDSs) and inferred the optimal codons by the use of the ΔRSCU method as described above. For the low G + C content sample, 11 codons were identified as optimal and 9 overlapped with the result of the whole data analysis, lacking the alanine (coded by CGU), the threonine (coded by ACU), plus the glutamine (coded by CAG) and the threonine (coded by ACA). For high G + C content sample, 13 codons were identified as optimal and among them, 10 codons overlapped with the result of the whole data analysis, lacking only the cysteine (coded by UGU), plus the phenylalanine (coded by UUU), the serine (coded by UCA) and the proline (coded by CCG, see additional files, Supplementary Tables S1 and S2). The analyses of samples of high and low G + C content revealed that G + C content of CDS has a significant effect on the identities of the optimal codons in the chicken genome. However, it should be noted that most putative optimal codons (9−10 codons) occurred coincidently in three samples. We believe that large samples can give more accurate estimates and, therefore, the codon bias (Fop, the ratio of optimal codons to synonymous codons) estimate next was based on the putatively optimal codons identified by the whole data set of 8631 genes.

For any given set of synonymous codons, the relevant isoacceptor tRNAs might not be equally abundant. Previous studies suggested that the most abundant tRNA for a given amino acid is predominantly recruited by the codons of highly expressed genes.[54] The optimal codons are most likely the ones corresponding to the most abundant and efficient cognate aa-tRNAs present in the cell.[7,13] This trend also existed in the human genome but with lower coefficient.[55−57] We conducted an analysis to test whether this trend also exists in the chicken genome. As tRNA gene copy numbers are generally correlated with cellular levels of tRNAs in both prokaryotes and eukaryotes,[9,58,59] we used the abundance of tRNA genes as a substitute for the levels of tRNAs in the cell. We found that there is a good correspondence between tRNA abundance and optimal codons within codon classes (see Table 1). We also computed the RGF of each isoacceptor tRNA and made a regression analysis between the RGF and RSCU in highly expressed genes and lesser expressed

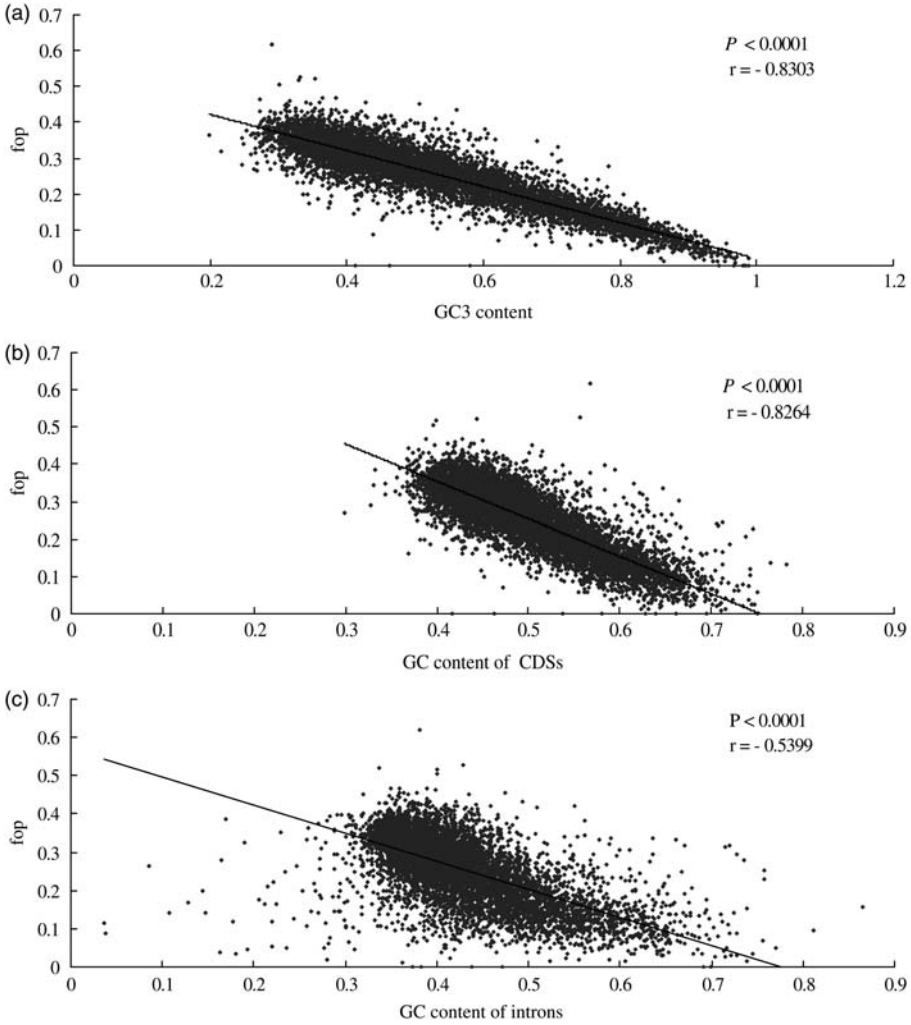

Figure 2. Scatter plots of GC_3, G + C content of coding sequences, G + C content of intronic sequences versus the frequency of optimal codons (Fop). Total gene included is 8631. Fop was estimated by codonW 1.4.2 with customized optimal codon table (see Table 1). (a) Fop versus GC_3; (b) Fop versus G + C content of CDSs and (c) Fop versus G + C content of intronic sequences.

genes. As shown in Fig. 1, there is a significant correlation between RGF and the RSCU of complementary codons in highly expressed genes ($r = 0.4599$, $P = 0.0013$). This positive trend also existed in lesser expressed genes, but with a lower correlation coefficient ($r = 0.3548$, $P = 0.0121$).

3.2. *Relationships between codon bias and GC_3, G + C content of CDSs and G + C content of intronic sequences*

Codon bias, measured as Fop, averaged 0.2560 ± 0.0042 (ranging from 0.0100 to 0.6183) across the 8631 genes in the *G. gallus* genome. The Fop values for genes residing on the macro-chromosomes, micro-chromosomes and Z chromosome are 0.2736 ± 0.0011, 0.2236 ± 0.0015 and 0.2992 ± 0.0032, respectively. There is a significant difference among them ($P < 0.0001$, using the one-way ANOVA). This means that genes residing on the Z chromosome have the highest codon bias. The reasons for this significant difference is most likely owing to the different G + C content, CpG island motifs, gene density and recombination rates of the three types of chromosomes.

Regression analysis demonstrated that Fop is highly correlated with GC_3 and G + C content of CDSs, respectively ($r = -0.8308$, $P < 0.0001$; $r = -0.8264$, $P < 0.0001$, see Fig. 2a and b). It appears that variation in GC mutational bias explains over 60% of the variation of the codon usage bias. This negative correlation is expected, as all putative optimal codons inferred in this study ended with U. We also retrieved all intronic sequences for each gene and provided the combined length of all introns for a particular gene exceeding 200 bp and calculated the G + C content of the intronic sequences. We found that Fop values also negatively correlated with

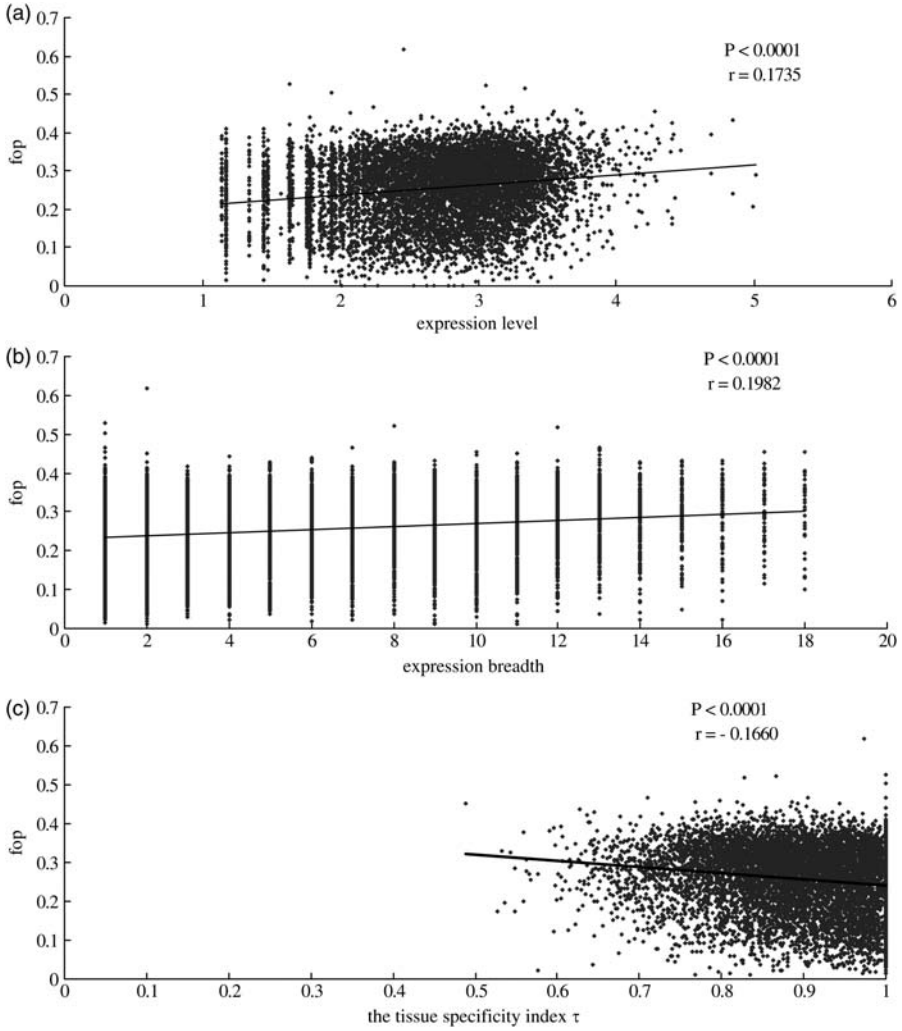

Figure 3. Scatter plots of Fop versus expression level, expression breadth and the tissue specificity index τ. Expression data on 8631 genes for 18 tissues, blood, brain, bursa of fabricius, cecum, connective tissue, embryonic tissue, epiphyseal growth plate, gonad, head, heart, limb, liver, muscle, ovary, pancreas, spleen, testis and thymus, were taken into account. Expression level is defined as the total expression level of a gene, which is the sum of the total 18 tissues' EST (transformed to logarithm with base 10). Expression breadth defined as the numbers of tissues in which EST was found. The calculation of τ can be seen from materials and methods. (a) Fop versus expression level; (b) Fop versus expression breadth and (c) Fop versus the tissue specificity index τ.

the G + C content of the intronic sequences significantly ($r = -0.5399$, $P < 0.0001$, see Fig. 2c).

3.3. *Relationships between codon bias and gene expression*

We assessed the effect of expression levels and expression breadth on codon usage bias in our samples. Our data demonstrated that codon bias is positively correlated with gene expression levels ($r = 0.1735$, $P < 0.0001$). The correlation between codon bias and expression breadth was also positive and significant ($r = 0.1982$, $P < 0.0001$, see Fig. 3a and b). These results suggest that the genes with broader expression breadth and higher expression levels show a higher degree of codon usage bias. Total gene expression is known to be highly influenced by the number of different tissues where a gene is expressed (expression breadth) when expression data are calculated from pooled EST libraries.[60,61] If expression breadth is the predominant force affecting codon usage, a spurious correlation between codon bias and gene expression is more likely to be generated. In order to alleviate this problem, we divided genes into ubiquitously or narrowly expressed groups, if they were expressed in ≥ 15 tissues or ≤ 3 tissues. We regressed expression levels on codon bias for each group, and found that this significant trend also existed for the ubiquitous group ($r = 0.1855$, $P = 0.0016$) and for narrowly expressed groups ($r = 0.0497$, $P = 0.0101$). When the parameters were expanded to ≥ 16 or ≤ 2, we obtained similar results.

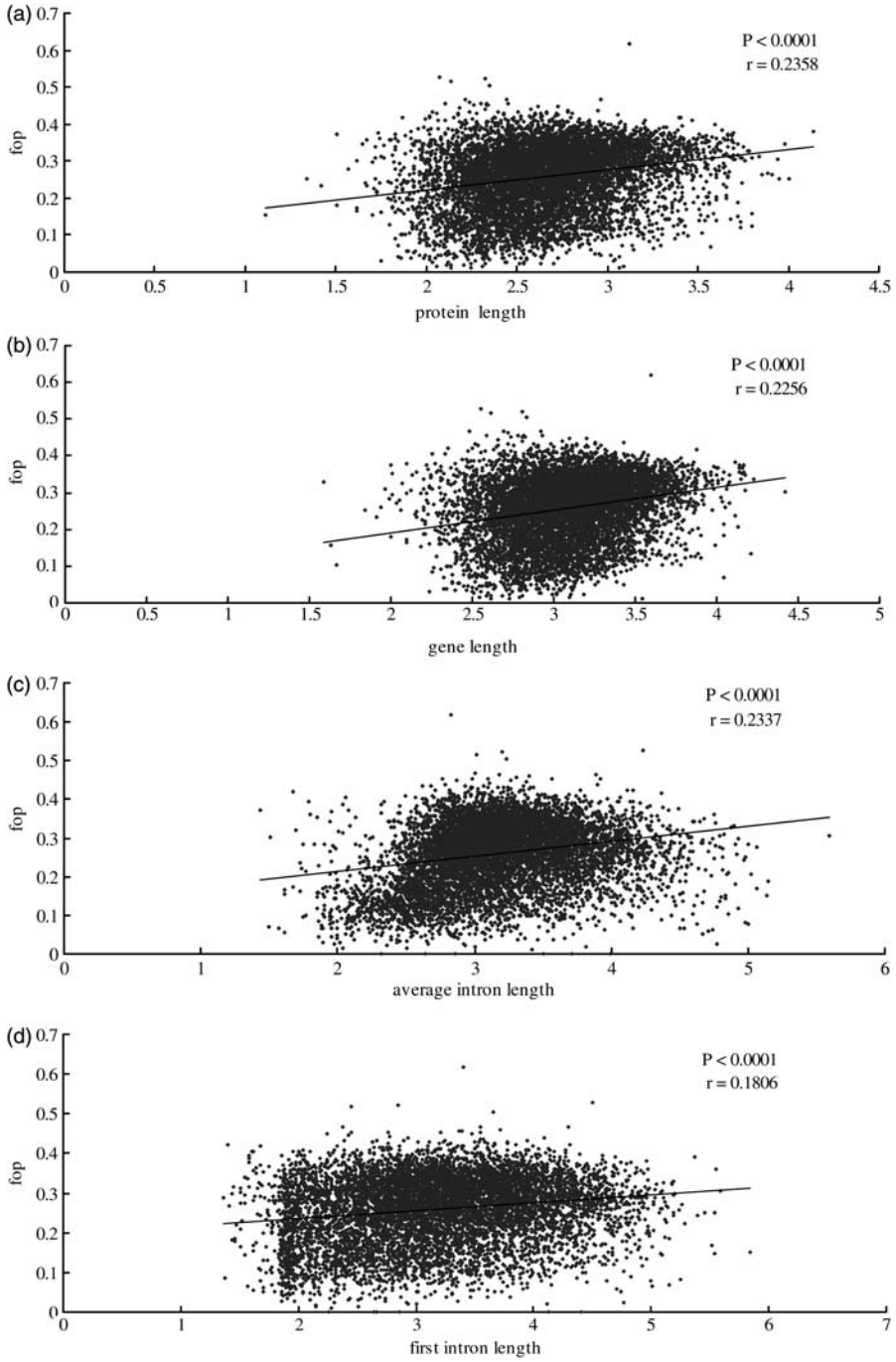

Figure 4. Scatter plots of Fop versus protein length, gene length, average intron length and first intron length (logarithmically transformed). (a) Fop versus protein length; (b) Fop versus gene length; (c) Fop versus average intron length and (d) Fop versus first intron length.

The tissue specificity index τ measures both qualitative and quantitative variations of expression levels amongst tissues.[49] Obviously, τ is more representative than the expression breadth and expression levels alone for the expression pattern of a gene. We calculated the tissue specificity index τ for each gene, and made regression analyses between codon bias and the tissue specificity index τ. As shown in Fig. 3c, codon bias is significantly correlated with the tissue specificity index τ ($r = -0.1660$, $P < 0.0001$).

3.4. *Relationships between codon bias and protein, gene and intron length*

Our data demonstrated that codon bias is significantly and positively correlated with protein length ($r = 0.2358$, $P < 0.0001$) and gene length ($r = 0.2256$, $P < 0.0001$, Fig. 4a and b). Since protein length and expression levels displayed a strong correlation with codon bias,[48] we tested whether the correlation between protein length and codon bias can be explained by gene expression

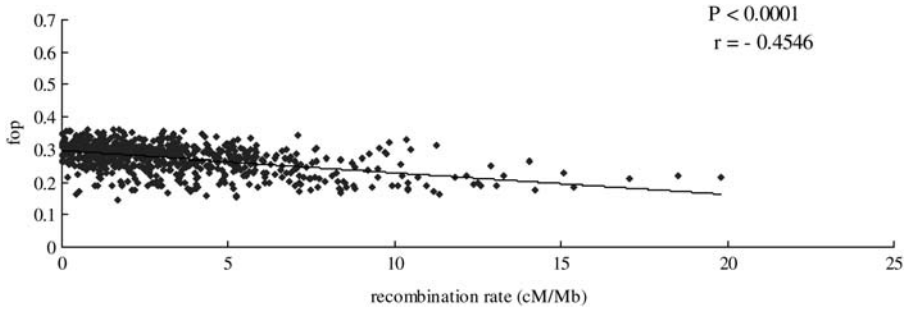

Figure 5. Scatter plots of Fop versus recombination rate. The recombination rates for 1 Mb windows were estimated. The versions of the genome assemblies (NCBI build 2.1, released November, 2006) and the latest chicken consensus linkage map were used. Total windows included is 745, covering ~70% of the chicken genome. For each window, the average Fop for all genes residing in this window was calculated.

levels. We fitted linear models of protein length and codon bias against expression levels. The results from these models correlated at the same levels as uncorrected protein lengths and codon bias. In the narrowly expressed gene samples as described above (breadth is three or less than three tissues), we conducted a regression analysis between codon bias and protein length at similar expression levels (EST counts ranging from 57 to 60) and found that the positive trend also existed ($r = 0.2823$, $P < 0.0001$). The negative correlation between codon usage and protein length and gene length has been observed in many organisms, such as yeast, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Populus tremula* and *Silene latifolia*.[5,29,36,62–64] However, the positive correlation identified in this study has been found only in *Saccharomyces cerevisiae* and *Escherichia coli*, and was been explained through the model of selection on translational accuracy.[37,65]

As shown in Fig. 4c and d, codon bias also shows a positive correlation with average intron length and first intron length ($r = 0.2337$, $P < 0.0001$; $r = 0.1806$, $P < 0.0001$, respectively). The opposite trend has been found in *D. melanogaster*, *C. elegans* and *P. tremula*.[5,66,67] This positive trend between codon bias and intron length has been only found in unicellular organisms;[68] however, the underlying mechanism for this is not clear. Stoletzki and Eyre-Walker[69] and Stoletzki[70] suggested that this trend (negative/positive) is related to whose optimal codons are biased towards codons that end with GC or AU. We found that the G + C content of introns is significantly and negatively correlated with average intron length and first intron length ($r = -0.4024$, $P < 0.0001$; $r = -0.2432$, $P < 0.0001$, respectively). Obviously, the positive trend between codon bias and intron length is related to the G + C content of intronic sequences. If mutation bias is the driving force of the codon usage in the *G. gallus* genome, the mutation bias hypotheses seems reasonable to explain this result.

3.5. *Relationship between codon bias and recombination rate*

Fig. 5 shows that there is a negative correlation between the Fop values and the local recombination rates ($r = -0.4546$, $P < 0.0001$). We also found that the relationship between codon usage bias and recombination rate is independent of the expression levels and protein length (data not shown). This significant trend seems to be expected, since the mutagenic effects of recombination result in a mutational bias toward G and C bases in regions of high recombination rates (GC-biased gene conversion).

4. Discussion

4.1. *The identities of optimal codons in the* G. gallus *genome*

In the present study, we carried out a systematic examination of the codon usage in the chicken genome. By the use of the ΔRSCU method, we identified 11 codons as putatively optimal. All putative optimal codons in the chicken genome end with U. This is significantly departing from the pattern observed in other eukaryotes genomes, such as *Schizosaccharomyces pombe*, *D. melanogaster*, *C. elegans* and *Homo sapiens*. The rules governing the identities of optimal codons in different organisms remain obscure. Recently, Hershberg and Petrov[21] investigated the optimal codons in 675 bacteria, 52 archea and 10 fungi. They found that across all studied organisms, the identities of optimal codons mirrors the G + C content of the genomes. GC-rich organisms tend to have GC-rich optimal codons, while AT-rich organisms tend to have AT-rich optimal codons. However, in *Drosophila*, *C. elegance* and *Populus tremula*, most optimal codons end with G or C (majority are C ending), while their genomes contain 35, 36 and 45% G + C-rich content, respectively.[5,71,72] In humans, optimal codons seem to be driven in two opposite directions, toward AT richness and GC richness. In other words,

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.[73] The human genome comprises a mosaic of long stretches of GC-rich and AT-rich regions, the so-called isochore structure. Not only do they occur in silent sites of coding regions but also introns and flanking regions in the gene have a similar base composition.[43] This isochore structure was also found in the avian genomes.[41] Recently, some studies have suggested that the G + C content is becoming homogenized in humans.[45,55] However, Webster *et al*.[42] found that heterogeneity in the G + C content is being reinforced in the chicken genome. In order to test whether the G + C content influences the identities of optimal codons in *G. gallus*, we produced a high G + C content sample (20% of the highest G + C content of the CDSs) and a low G + C content sample (20% of the lowest G + C content of the CDSs), and inferred the optimal codons by the use of the ΔRSCU method. We found that most putative optimal codons (9−10 codons) occur coincidently in three samples. Using the abundance of tRNA genes as a substitute for the levels of tRNAs in the cell, we found that there is a good correspondence between tRNA abundance and optimal codons within codon classes. This implies that the optimal codons in the chicken genome are most likely the ones corresponding to the highly expressed tRNAs or tRNA gene copy numbers in the cell.

4.2. *Mutation bias is the driving force of the codon usage in the* G. gallus *genome*

Codon bias, as measured by Fop, is significantly correlated with GC_3, G + C content of CDS and G + C content of the intronic sequences. Our data clearly displayed that regional G + C content explains a large proportion of the variance of codon bias in chicken. Our data also provide strong evidence for the mutational bias hypothesis. However, we found that the G + C content of the intronic sequences is significantly lower than that of CDSs, which are not fully consistent with this hypothesis, as it predicts that G + C content is determined by genome-wide processes rather than by selective forces acting specifically on coding regions. It also should be noted that although the G + C content of intronic sequences show a negative trend with codon usage, the correlation is significantly lower than that between codon usage and G + C content of CDSs. These findings indicate that, except for mutation bias, other factors (such as selection) may have contribution to the codon usage.

Our data also show that codon bias is significantly and positively correlated with gene expression. The positive correlations between gene expression and codon bias have been shown in many organisms.[1−2,4−5] In vertebrates such as mammals, data also support a weak relationship between gene expression and codon usage.[56,61] Both the match between tRNA abundance and optimal codons, and the high codon bias of the highly expressed genes, has been interpreted as natural selection for more efficient and accurate translation.[14−18] Our data provide evidence that natural selection also plays an important role in shaping the codon usage in the chicken genome.

Codon bias also shows a significant trend with protein length, intron length and recombination rate. To determine what all variables (G + C content of CDS, expression level, expression breadth, protein length, intron length and recombination rate) were contributing to the differences in codon bias and how they may interact, we performed multiple linear regressions with the above variables, excluding those not contributing significantly through the use of the *t*-statistical analysis and with backward stepwise regression. The best combinations of variables were G + C content of CDSs and expression breadth ($R^2 = 0.7829$, $P < 0.0001$). Stepwise selection model analyses indicated that the G + C content of CDSs are the most important factor responsible for codon bias ($R^2 = 0.6831$, $P < 0.0001$). It appears that variation in the G + C content of CDSs explains over 60% of the variation of codon bias. Recently, a continuous-time Markov chain model to quantify the contribution of GC-biased synonymous substitution on codon usage was developed by Palidwor *et al*.[73] Although many other important factors such as selection, GC skew, did not included in their model, it also provided an informative clue to understand the mechanism of codon usage across a broad variety of organisms. This model indicated that GC bias is the dominant factor in determining codon bias for prokaryotes, plants and human. In the present study, our data suggested that both mutation bias and selection contributed to the codon bias. This seems to be consistent with the few studies in other vertebrates such as *Xenopus laevis* and fishes of Cyprinidae.[74−75] However, it should be noticed that mutation bias is the driving force of the codon usage in the *G. gallus* genome.

4.3. *The negative association between codon bias and recombination in* G. gallus *is determined by recombination-dependent mutational patterns*

In contrast to *D. melanogaster* and *C. elegans*, a negative correlation between codon bias and local recombination rates was found in the chicken genome. The positive pattern in *D. melanogaster* and *C. elegans* has been interpreted by Hill−Robertson effects (hitchhiking and background selection)[28,29]

or by recombination-dependent mutational patterns (gene conversion).[32,36] The *C. elegans* has 21 optimal codons, of which 16 end in G or C bases, and *D. melanogaster* has 22 optimal codons, of which 21 end in G or C bases. Marais *et al*. demonstrated that, in *C. elegans*, the frequency of GC-ending optimal codons (Fop-GC) increases with recombination rate, whereas the frequency of AU-ending optimal codons (Fop-AU) decreases with recombination rate. In *Drosophila*, the frequency of AU-ending non-optimal codons (Fnop-AU) decreases with recombination rate, whereas the frequency of GC ending non-optimal codons (Fnop-GC) increases with recombination rate.[32] In yeasts, about 60% of the optimal codons end by GC, an overall positive correlation is also observed between recombination rate and Fop. However, there is a strong negative correlation between Fop-AU and recombination rate.[35] Marais and Piganeau[32] suggested that the positive correlation between the frequency of optimal codons and recombination rates in *C. elegans* and *D. melanogaster* is not due to improved selection but to a mutational bias toward G and C bases in regions of high recombination rates (GC-biased gene conversion). If mutation bias variation patterns do occur, they should affect all base positions within the gene, including coding and non-coding sequences. We surely found this positive trend between recombination rate and G + C content of introns in the chickens ($r = 0.2467$, $P < 0.0001$). The mutational bias explanation seems to be the case in *G. gallus*. As 11 putative optimal codons identified in this study all ended in U, a negative association between codon bias and recombination rate is expected.

Another question that should be addressed is whether the correlation between the codon usage bias and recombination rate in *G. gallus* is a direct consequence of the recombination process. Based on the chicken consensus linkage map,[51] we selected some chromosome centromere region (chromosome 1–13, chromosome 17, chromosome 23, chromosome 25, chromosome 28 and chromosome Z), and some non-centromere regions where the estimated recombination rate is null (chromosome 1: 96.4173–99.0780 Mbp and 102.5560–105.82 39 Mbp; chromosome 2: 24.8946–27.1233 Mbp; chromosome 3: 62.328–65.3834 and 91.8723–93.5081 Mbp; chromosome 4: 20.8076–22.4734 and 72.3143–75.3621 and, 76.1848–79.2 626 Mbp; chromosome 5: 31.8527–33.3023 Mbp; 33.3360–35.0246 Mbp; chromosome 6: 13.24 26–16.2627 Mbp; chromosome 7: 8.8062–11.1527 Mbp; chromosome 8: 11.7298–12.7 581 Mbp; chromosome 19: 7.8870–8.84 82 Mbp; chromosome 20: 7.1329–8.3848 Mbp), and compared their G + C content of CDSs, G + C content of introns with those of the top recombination regions (same number of intervals on the same chromosome were selected). Although the recombination rates are likely to vary in different populations, the above non-centromere regions have been identified in that no recombination occurred in a outbreed chicken population established from a crossbreeding between a Xinghua line and a White Recessive Rock line by a high-density SNP microarray (556 individuals, unpublished data). We found that the G + C content of CDSs at the high recombination regions (0.5164 ± 0.0056) is significantly higher than that of the regions incurring no recombination (0.4695 ± 0.0056; $P < 0.0001$, *t*-test). The G + C content of introns of genes residing in the high recombination regions (0.4489 ± 0.0061) is also significantly higher than that of regions having no recombination(0.3912 ± 0.0044; $P < 0.0001$, *t*-test). This implies that the correlation between codon usage bias and recombination rates in *G. gallus* is determined predominantly by recombination-dependent mutational patterns. However, this does not mean that selection did not act on the synonymous sites. It is more likely that their impact on codon usage has been masked by variations in mutation pressures associated with the high recombination rates in chickens.

4. Conclusion

In this whole genome analysis of the chicken, we identified 11 putative optimal codons, which all ended with U. There is a good correspondence between tRNA abundance and optimal codons within codon classes. Codon bias is negatively correlated with G + C content and recombination rates, but positively correlated with gene expression, protein length and intron length. The G + C content of coding sequences are the most important factors responsible for codon bias. It appears that variation in the G + C content of CDSs explains over 60% of the variation of codon bias. Our study suggests that both mutation bias and selection contribute to the codon bias. However, mutation bias is the driving force of the codon usage in the *G. gallus* genome. Our data also provide evidence that the negative correlation between codon bias and recombination rates in *G. gallus* is determined predominantly by recombination-dependent mutational patterns.

Acknowledgements: We thank three reviewers for their helpful comments on the manuscript.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by the Science and Technology Program of Jiangxi Education Department, Project No. GJJ8469, and the High Tech Program (863), China, Project No. 2006AA10A120.

References

1. Gouy, M. and Gautier, C. 1982, Codon usage in bacteria: correlation with gene expressivity, *Nucleic Acids Res.*, **10**, 7055–74.
2. Duret, L. and Mouchiroud, D. 1999, Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*, *Proc. Natl. Acad. Sci.*, **96**, 4482–7.
3. Ghaemmaghami, S., Huh, W.K., Bower, K.R., et al. 2003, Global analysis of protein expression in *yeast*, *Nature*, **425**, 737–41.
4. Goetz, R.M. and Fuglsang, A. 2005, Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*, *Biochem. Biophys. Res. Commun.*, **327**, 4–7.
5. Ingvarsson, P.K. 2007, Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*, *Mol. Biol. Evol.*, **24**, 836–44.
6. Ikemura, T. 1985, Codon usage and tRNA content in unicellular and multi-cellular organisms, *Mol. Biol. Evol.*, **2**, 13–34.
7. Andersson, S.G.E. and Kurland, C.G. 1990, Codon preferences in free-living microorganisms, *Microbiol. Rev.*, **54**, 198–210.
8. Percudani, R., Pavesi, A. and Ottonello, S. 1997, Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*, *J. Mol. Biol.*, **268**, 322–30.
9. Kanaya, S., Yamada, Y., Kudo, Y. and Ikemura, T. 1999, Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis, *Gene*, **238**, 143–55.
10. Yamao, F., Andachi, Y., Muto, A., Ikemura, T. and Osawa, S. 1991, Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins, *Nucleic Acids Res.*, **19**, 6119–22.
11. Duret, L. 2000, tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes, *Trends Genet.*, **16**, 287–9.
12. Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. and 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**, 290–8.
13. Rocha, E.P. 2004, Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization, *Genome Res.*, **14**, 2279–86.
14. Sorensen, M.A., Kurland, C.G. and Pedersen, S. 1989, Codon usage determines translation rate in *Escherichia coli*., *J. Mol. Biol.*, **207**, 365–77.
15. Duret, L. 2002, Evolution of synonymous codon usage in metazoans, *Curr. Opin. Genet. Dev.*, **12**, 640–9.
16. Stoletzki, N. and Eyre-Walker, A. 2007, Synonymous codon usage in *Escherichia coli*: selection for translational accuracy, *Mol. Biol. Evol.*, **24**, 374–81.
17. Hershberg, R. and Petrov, D.A. 2008, Selection on codon bias, *Annu. Rev. Genet.*, **42**, 287–99.
18. Sharp, P.M., Fmery, L. and Zeng, K. 2010, Forces that influence the evolution of codon bias, *Phil. Trans. R. Soc. B*, **1544**, 1203–12.
19. Knight, R.D., Freeland, S.J. and Landweber, L.F. 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**, RESEARCH0010.
20. Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L. and McAdams, H.H. 2004, Codon usage between genomes is constrained by genome-wide mutational processes, *Proc. Natl. Acad. Sci.*, **101**, 3480–5.
21. Hershberg, R. and Petrov, D.A. 2009, General rules for optimal codon choice, *PLoS Genet.*, **5**, 1–10.
22. Guo, X., Bao, J. and Fan, L. 2007, Evidence of selectively driven codon usage in rice: implications for G + C content evolution of Gramineae genes, *FEBS Lett.*, **581**, 1015–21.
23. Bulmer, M. 1991, The selection-mutation-drift theory of synonymous codon usage, *Genetics*, **129**, 897–907.
24. Akashi, H. 1995, Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA, *Genetics*, **139**, 1067–76.
25. Akashi, H., Kliman, R.M. and Eyre-Walker, A. 1998, Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*, *Genetica*, **102–103**, 49–60.
26. dos Reis, M. and Wernisch, L. 2009, Estimating translational selection in eukaryotic genomes, *Mol. Biol. Evol.*, **26**, 451–61.
27. Sharp, P.M. and Li, W.H. 1986, An evolutionary perspective on synonymous codon usage in unicellular organisms, *J. Mol. Evol.*, **24**, 28–38.
28. Kliman, R.M. and Hey, J. 1993, Reduced natural selection associated with low recombination in *Drosophila melanogaster*, *Mol. Biol. Evol.*, **10**, 1239–58.
29. Comeron, J.M., Kreitman, M. and Aguade, M. 1999, Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*, *Genetics*, **151**, 239–49.
30. Comeron, J.M. and Kreitman, M. 2002, Population, evolutionary and genomic consequences of interference selection, *Genetics*, **161**, 389–410.
31. Hey, J. and Kliman, R.M. 2002, Interactions between natural selection, recombination and gene density in the genes of *Drosophila*, *Genetics*, **160**, 595–608.
32. Marais, G. and Piganeau, G. 2002, Hill-Robertson interference is a minor determinant of variations in codon

bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genome, *Mol. Biol. Evol.*, **19**, 1399–406.
33. Marais, G., Mouchiroud, D. and Duret, L. 2003, Neutral effect of recombination on base composition in *Drosophila*, *Genet. Res.*, **81**, 79–87.
34. Haddrill, P.R., Halligan, D.L., Tomaras, D. and Charlesworth, B. 2007, Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over, *Genome Biol.*, **8**, R18.
35. Harrison, R. and Charlesworth, B. 2010, Biased gene conversion affects patterns of codon usage and amino acid usage in the Saccharomyces *sens stricto* group of yeasts, *Mol. Biol. Evol.*, **28**, 117–29.
36. Marais, G., Mouchiroud, D. and Duret, L. 2001, Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes, *Proc. Natl. Acad. Sci.*, **98**, 5688–92.
37. Moriyama, E.N. and Powell, J.R. 1998, Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*., *Nucleic Acids Res.*, **26**, 3188–93.
38. Lemos, B., Bettencourt, B.R., Meiklejohn, C.D. and Hartl, D.L. 2005, Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length and number of protein-protein interactions, *Mol. Biol. Evol.*, **22**, 1345–54.
39. Stenoien, H.K. 2005, Adaptive basis of codon usage in the haploid *moss Physcomitrella patens*, *Heredity*, **94**, 87–93.
40. Groenen, M.A., Cheng, H.H., Bumstead, N., et al. 2000, A consensus linkage map of the chicken genome, *Genome Res.*, **10**, 137–47.
41. International Chicken Genome Sequencing Consortium (ICGSC). 2004, Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution, *Nature*, **432**, 695–716.
42. Webster, M.T., Axelsson, E. and Ellegren, H. 2006, Strong regional biases in nucleotide substitution in the chicken genome, *Mol. Biol. Evol.*, **23**, 1203–216.
43. Bernardi, G. 2000, Isochores and the evolutionary genomics of vertebrates, *Gene*, **241**, 3–17.
44. Costantini, M. and Bernardi, G. 2008, Correlations between coding and contiguous non-coding sequences in isochore families from vertebrate genomes, *Gene*, **410**, 241–8.
45. International Human Genome Sequencing Consortium. 2001, Initial sequencing and analysis of the human genome, *Nature*, **409**, 860–921.
46. Mank, J.E., Hultin-Rosenberg, L., Axelsson, E. and Ellegren, H. 2007, Rapid evolution of female-biased, but not male-biased, genes expressed in the avian brain, *Mol. Biol. Evol.*, **24**, 2698–706.
47. Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V. and Kondrashov, F.A. 2002, Selection for short introns in highly expressed genes, *Nat. Genet.*, **31**, 415–8.
48. Rao, Y.S., Wang, Z.F., Chai, X.W., et al. 2010, Selection for the compactness of highly expressed genes in *Gallus gallus*, *Biology Direct*, **5**, 35.
49. Yanai, I., Benjamin, H., Shmoish, M., et al. 2005, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification, *Bioinformatics*, **21**, 650–9.
50. Cutter, A.D. and Charlesworth, B. 2006, Selection intensity on preferred codons correlates with overall codon usage bias in Caenorhabditis remanei, *Curr. Biol.*, **16**, 2053–7.
51. Groenen, M.A., Wahlberg, P., Foglio, M., et al. 2009, A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate, *Genome Res.*, **19**, 510–9.
52. Payseur, B.A. and Nachman, M.W. 2000, Microsatellite variation and recombination rate in the human genome, *Genetics*, **156**, 1285–98.
53. Holm, S. 1979, A simple sequentially rejective Bonferroni test procedure, *Scand. J. Stat.*, **6**, 65–70.
54. Dong, H., Nilsson, L. and Kurland, C.G. 1996, Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates, *J. Mol. Biol.*, **260**, 649–63.
55. Comeron, J.M. 2004, Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence, *Genetics*, **167**, 1293–1304.
56. Lavner, Y. and Kotlar, D. 2005, Codon bias as a factor in regulating expression via translation rate in the human genome, *Gene*, **345**, 127–38.
57. Shah, P. and Gilchrist, M.A. 2010, Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias, *PLoS Genet.*, **6**, e1001128.
58. Cognat, V., Deragon, J.M., Vinogradova, E., Salinas, T., Remacle, C. and Maréchal-Drouard, L. 2008, On the evolution and expression of Chlamydomonas reinhardtii nucleus-encoded transfer RNA genes, *Genetics*, **179**, 113–23.
59. Parmley, J.L. and Huynen, M.A. 2009, Clustering of codons with rare cognate tRNAs in human genes suggests an extra level of expression regulation, *PLoS Genet.*, **5**, e1000548.
60. Akashi, H. 2001, Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.*, **11**, 660–6.
61. Urrutia, A.O. and Hurst, L.D. 2001, Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection, *Genetics*, **159**, 1191–9.
62. Mouchiroud, D. and Duret, L. 1999, Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*, *Proc. Natl. Acad. Sci.*, **96**, 4482–7.
63. Charlesworth, B. and Loewe, L. 2007, Background selection in single genes may explain patterns of codon bias, *Genetics*, **175**, 1381–93.
64. Qiu, S., Bergero, R., Zeng, K. and Charlesworth, D. 2010, Patterns of codon usage bias in *Silene latifolia*, *Mol. Biol. Evol.*, **28**, 771–80.
65. Walker, A.E. 1996, Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Mol. Biol. Evol.*, **13**, 864–72.

66. Marais, G., Nouvellet, P., Keightley, P.D. and Charlesworth, B. 2005, Intron size and exon ewolution in *Drosophila*, *Genetics*, **170**, 481–85.
67. Stenico, M., Lloyd, A.T. and Sharp, P.M. 1994, Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases, *Nucleic Acids Res.*, **13**, 2437–46.
68. Vinogradov, A.E. 2001, Intron length and codon usage, *J. Mol. Evol.*, **52**, 2–5.
69. Stoletzki, N. and Eyre-Walker, A. 2007, Synonymous codon usage in *Escherichia coli*: selection for translational accuracy, *Mol. Biol. Evol.*, **24**, 374–81.
70. Stoletzki, N. 2011, The surprising negative correlation of gene length and optimal codon use—disentangling translational selection from GC-biased gene conversion in yeast, *BMC Evol. Biol.*, **11**, 93.
71. Vicario, S., Moriyama, E.N. and Powell, J.R. 2007, Codon usage in twelve species of *Drosophila*, *BMC Evol. Biol.*, **7**, 226.
72. Nekrutenko, A. and Li, W.H. 2000, Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.*, **10**, 1986–95.
73. Palidwor, G.A., Perkins, T.J. and Xia, X. 2010, A general model of codon bias due to GC mutational bias, *PLoS One*, **5**, e13431.
74. Musto, H., Cruveiller, S., D'Onofrio, G., Romero, H. and Bernardi, G. 2001, Translational selection on codon usage in Xenopus laevis. *Mol. Biol. Evol.*, **18**, 1703–07.
75. Romero, H., Zavala, A., Musto, H. and Bernardi, G. 2003, The influence of translational selection on codon usage in fishes from the family Cyprinidae. *Gene*, **317**, 141–7.