Journal of Agricultural Science (2003), **140**, 235–244. © 2003 *Cambridge University Press*
DOI: 10.1017/S0021859602002903 *Printed in the United Kingdom*

Structural and functional relationships shown by genomic analysis of *Capra hircus* genes

A. FADIEL[1], S. LITHWICK[1], G. GANJI[1] AND I. F. M. MARAI[2]*

[1] *The Center for Computational Biology, Hospital for Sick Children, Toronto, Ontario, Canada*
[2] *Department of Animal Production, Faculty of Agriculture, Zagazig University, Egypt*

(*Revised MS received 30 October 2002*)

SUMMARY

The effect of base composition biases on codon usage patterns was investigated in the goat species *Capra hircus*, using custom-designed computational tools available within the public domain. Nucleotide frequencies were nearly equal and a slight increase of adenine–thymine (AT) over guanine–cytosine (GC) was detected throughout the dataset. However, this increase showed no influence on the bases at the third codon position (N3). C3 and G3 were found more often than A/T3, suggesting that there was a small or almost no influence of the general base composition on the N3 base composition. To understand more and analyse in-depth influence and interactions between base compositions and codons, further relative synonymous codon usage (RSCU) was investigated. Amino acid usage and the correlation between its usages were also investigated, using both basic sequence analysis and statistical analysis means (measures of correlation). These analyses were utilized to probe whether there were correlations between genes, genomic characteristics and their function. Genes with high GC and those with low GC were also investigated to see to what extent how a gene functions could influence its sequence structure and impose certain structural modifications. This investigation may shed light on many genomic features of *Capra hircus* genes and would be of significance for future biotechnology/research projects considering *Capra* for transgenic and advanced genomic initiatives.

INTRODUCTION

Goats belong to the Phylum Chordata, of which the most noted are the Angora, the Nubian and the Kashmiri type. Goats are mainly reared for their meat, milk and pelt and play a major role within many agricultural communities. However, interest in the goat has also recently developed significantly in the field of biotechnology.

Little attention has yet been paid to the genomics of goat species, despite being a favourable animal model for the future of biotechnology and transgenics. Specifically, the use of goats for the production of heterologous proteins has met with success. Furthermore, heterogeneous gene expression has been shown to be possible. Gene products have been directed to milk and blood allowing for simple collection. Thus, transgenic goats are very attractive systems for the *in vivo* production of recombinant proteins.

The genetic code is highly degenerated in that most amino acids are encoded by multiple codons. This raises many questions with respect to both relative synonymous codon usage (RSCU) and its effect on amino acid composition. Grantham *et al*. (1980) observed that certain unique species had characteristic codon biases. This led to the development of the 'genome hypothesis', which proposes that all genes within a given genome use the same coding strategy with respect to synonymous codon usage. This hypothesis has stood the test of time, although some exceptions have arisen. Subsequently, Ikemura (1981) suggested that the codon usage biases correlated with relative tRNA abundances by comparing codon use with the abundance of the tRNAs that correspond to these codons. Furthermore, it appeared that codon usage patterns were correlated with gene expression levels. Although this influence of codon bias on gene expression level was found to be present in *Drosophila* genes, no such pattern has been identified in either birds or mammals.

Codon usage has been found to vary between classes and their related species. Codon usage changes and sequence dissimilarity between human and rat

* To whom all correspondence should be addressed.

Table 1. *Mononucleotide and dinucleotide frequencies in* Capra*, based upon a dataset consisting of 90 genes (W represents A and T nucleotides, while S represents the frequencies of G and C)*

No.	Gene	C	T	A	G	W (AT)	S (GC)
1	gi|412637	32·23	19·90	22·05	30·37	41·94	62·59
2	gi|126987	24·22	29·47	26·21	24·69	55·68	48·92
3	gi|235333	33·46	21·17	21·11	26·11	42·28	59·57
4	gi|129584	25·11	22·08	25·32	28·79	47·40	53·90
5	gi|617939	21·25	25·14	33·89	20·56	59·03	41·81
6	gi|120043	36·19	20·14	24·63	20·27	44·76	56·46
7	gi|116927	32·46	24·61	20·68	22·51	45·29	54·97
8	gi|618000	24·30	26·05	25·88	24·04	51·93	48·33
9	gi|618000	24·89	25·59	25·33	24·45	50·92	49·34
10	gi|618000	24·89	25·33	25·59	24·46	50·91	49·35
11	gi|618000	27·29	24·35	24·78	23·83	49·14	51·12
12	gi|617999	25·32	26·59	25·51	22·87	52·10	48·19
13	gi|617999	24·12	26·32	25·18	24·65	51·49	48·77
14	gi|617999	24·56	26·67	25·53	23·51	52·19	48·07
15	gi|617999	23·68	25·88	26·49	24·21	52·37	47·89
16	gi|617999	23·42	26·40	25·35	25·09	51·75	48·51
17	gi|617998	26·60	25·89	24·29	23·49	50·18	50·09
18	gi|617998	23·77	27·28	25·70	23·51	52·98	47·28
19	gi|110669	20·00	26·24	34·62	19·78	60·86	39·78
20	gi|541962	33·73	21·07	15·39	31·70	36·46	65·44
21	gi|997240	26·82	27·18	28·20	24·07	55·38	50·90
22	gi|437751	33·85	16·60	21·62	29·09	38·22	62·93
23	gi|843928	21·39	25·69	34·31	19·86	60·00	41·25
24	gi|764926	29·54	22·75	23·98	24·10	46·73	53·64
25	gi|217664	28·83	21·56	21·88	31·02	43·44	59·85
26	gi|733115	39·44	17·30	12·98	31·04	30·28	70·48
27	gi|217666	24·74	27·67	27·36	22·75	55·03	47·48
28	gi|663406	18·69	35·12	34·79	18·87	69·92	37·56
29	gi|663406	18·69	35·17	34·87	18·89	70·04	37·58
30	gi|410716	26·46	22·34	29·90	27·49	52·23	53·95
31	gi|410716	27·37	21·75	29·82	27·37	51·58	54·74
32	gi|380611	20·77	44·44	35·27	15·46	79·71	36·23
33	gi|380611	16·67	46·67	28·89	14·44	75·56	31·11
34	gi|380611	19·29	35·30	38·32	12·99	73·62	32·28
35	gi|606352	29·47	19·62	25·03	29·93	44·65	59·41
36	gi|606309	32·47	19·79	21·35	26·91	41·15	59·38
37	gi|513936	21·62	25·58	29·11	27·05	54·69	48·66
38	gi|513935	24·73	20·43	23·84	32·08	44·27	56·81
39	gi|513935	26·81	21·01	24·28	28·99	45·29	55·80
40	gi|513935	28·26	19·90	25·73	28·01	45·63	56·27
41	gi|440647	24·35	20·44	25·52	30·08	45·96	54·43
42	gi|440646	24·35	20·44	25·65	29·95	46·09	54·30
43	gi|173027	19·26	33·06	35·52	19·95	68·58	39·21
44	gi|261861	23·99	29·84	28·42	21·83	58·27	45·81
45	gi|261860	27·14	23·45	24·37	25·43	47·83	52·57
46	gi|257582	26·07	20·18	22·56	31·58	42·73	57·64
47	gi|257582	26·32	20·05	21·55	32·46	41·60	58·77
48	gi|385134	16·62	27·63	38·54	23·82	66·17	40·44
49	gi|162103	23·97	28·85	27·67	24·86	56·52	48·82
50	gi|607058	24·17	21·81	31·11	23·33	52·92	47·50
51	gi|148006	24·04	21·63	32·05	22·76	53·69	46·79
52	gi|633053	22·46	26·50	28·39	24·43	54·90	46·88
53	gi|225343	25·17	21·79	27·86	25·77	49·65	50·95
54	gi|219707	27·80	24·02	26·93	21·69	50·95	49·49
55	gi|862328	29·47	20·59	23·95	26·66	44·54	56·13
56	gi|148316	32·68	19·16	20·87	27·69	40·03	60·37
57	gi|183430	24·26	21·94	29·41	24·75	51·35	49·02
58	gi|183430	23·94	25·61	34·49	23·29	60·10	47·23

https://doi.org/10.1017/S0021859602002903 Published online by Cambridge University Press

Table 1. (*cont.*)

No.	Gene	C	T	A	G	W (AT)	S (GC)
59	gi|191166	31·43	13·73	26·05	29·03	39·79	60·46
60	gi|173027	20·22	26·24	34·41	19·78	60·65	40·00
61	gi|173027	22·68	28·21	32·54	21·89	60·75	44·58
62	gi|707033	20·72	28·71	32·49	19·35	61·20	40·06
63	gi|955|em	20·34	28·77	33·13	18·95	61·90	39·29
64	gi|114961	25·23	21·88	26·13	31·02	48·01	56·24
65	gi|556806	27·47	21·52	24·19	30·70	45·71	58·17
66	gi|551229	27·96	25·89	25·76	23·69	51·65	51·65
67	gi|400442	24·35	20·44	25·39	30·21	45·83	54·56
68	gi|979|em	23·95	29·15	25·83	23·23	54·98	47·19
69	gi|977|em	26·54	26·92	33·21	19·10	60·13	45·64
70	gi|551225	27·96	25·89	25·64	23·81	51·53	51·77
71	gi|494966	30·95	20·47	19·94	30·72	40·41	61·67
72	gi|416002	20·65	25·38	33·76	20·86	59·14	41·51
73	gi|311942	23·61	34·39	27·65	17·13	62·04	40·74
74	gi|975|em	34·23	22·18	21·15	27·05	43·33	61·28
75	gi|973|em	33·33	21·88	20·18	26·56	42·06	59·90
76	gi|971|em	21·55	28·82	29·38	26·07	58·19	47·63
77	gi|969|em	21·97	28·67	29·86	25·01	58·54	46·98
78	gi|967|em	33·73	16·47	21·29	29·32	37·75	63·05
79	gi|961|em	27·17	18·84	27·17	31·16	46·01	58·33
80	gi|959|em	22·98	35·60	28·80	16·50	64·40	39·48
81	gi|953|em	24·04	30·70	30·70	20·19	61·41	44·23
82	gi|841157	25·16	25·73	26·67	24·02	52·40	49·19
83	gi|128004	27·76	21·87	23·69	30·43	45·56	58·19
84	gi|164169	22·93	30·25	26·86	26·01	57·11	48·94
85	gi|164145	21·48	28·70	29·26	25·93	57·96	47·41
86	gi|164135	22·22	29·17	27·11	25·21	56·28	47·43
87	gi|164133	22·44	30·13	27·68	25·55	57·81	47·99
88	gi|164125	34·18	18·09	18·82	31·45	36·92	65·63
89	gi|164123	34·21	19·68	20·78	29·11	40·46	63·32
90	gi|164116	28·45	25·65	27·93	24·82	53·58	53·27

suggests that, in some cases, modifications of DNA-based constraints could lead to accelerated inter-genomic divergence (Mouchiroud & Gautier 1990). This evolutionary separation was also clear in many other species, namely *Escherichia coli*, *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Homo sapiens*. Each species has a unique combination of eight least-used codons, but all species make equal use of all remaining codons. Proteins containing a high proportion of low-usage codons are associated with cases in which an excess of the protein is detrimental. Low-usage codons are relatively insensitive to gross base composition changes. However, dinucleotide usage can sometimes have a significant influence on codon usage as in the CG frequencies in several species of primates (Zhang *et al*. 1991).

Inter- and intra-species gene variation have been identified in many species (Bernardi *et al*. 1985). In addition, the rate of synonymous codon substitution in many animal classes has often been correlated with base composition (Filipski 1988). In *D. melanogaster*, some loci are subject to a discrete nucleotide bias favouring C at the third codon position (N3) (Moriyama & Hartl 1993). This might not reflect relative tRNA pools, but instead might be a function of mRNA secondary structure or stability. Conversely, in mammals such as *H. sapiens* significant codon usage biases are observed (Eyre-Walker 1991), and are thought to be a product of both mutational biases (Smith & Eyre-Walker 2001) and of selection. In addition, a correlation has been found between gene location in mammals and GC content, in which the nucleotide content of certain genes seems to reflect their surrounding isochors (Iida & Akashi 2000). Thus, variations in the rate of synonymous nucleotide substitution may likely be the consequence of selection against A and T at synonymous positions.

Investigation of base composition and codon usage patterns present within the coding sequences of *Capra* will probably allow for identification of the important conserved motifs, which would be of value in evaluation of this organism as a potential transgenic model, that would be of interest to both science and biotechnology.

https://doi.org/10.1017/S0021859602002903 Published online by Cambridge University Press

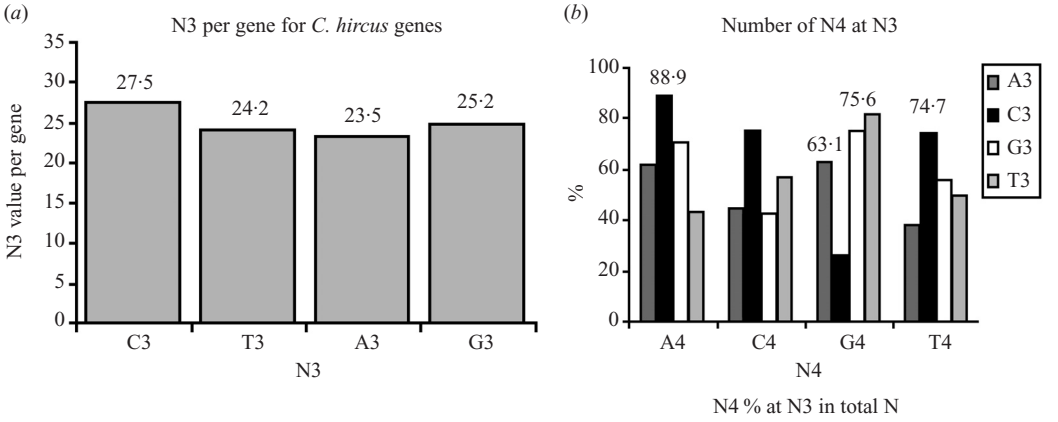

Fig. 1. (*a*) Mean mononucleotide frequency at the third codon position in *Capra*, based upon a 90-gene dataset (the y-axis shows the number of N at the third position N3). (*b*) Frequency of each mononucleotide at the third position (A3, C3, G3 and T3) in correspondence to all other mononucleotides at the intercodon position N4 where A4, C4, G4 and T4 represent each base in the fourth position located after the triplet.

In the present investigation, gene sequences from the goat species *Capra hircus* were characterized with respect to base composition, RSCU and amino acid usage.

MATERIALS AND METHODS

Capra hircus coding sequence patterns were investigated using a dataset consisting of the cDNA sequences for all known protein-encoding genes obtained from GenBank (www.ncbi.nlm.nih.gov/Entrez). The sequence dataset consisted of 131 466 base pairs making up 90 genes, coding for 43 822 amino acids. In order to analyse codon usage variation between genes, it is necessary to derive the RSCU values for each gene:

$$RSCU_i = Obs_i / Exp_i \quad (1)$$

where $RSCU_i$ is the Relative Synonymous Codon Usage value for codon i, Obs_i is the observed number of occurrences of a codon i and Exp_i is the expected number of occurrences of codon i. The expected number of occurrences of a codon is calculated according to

$$Exp_i = \sum aa_i / \sum syn_i \quad (2)$$

where Exp_i is the expected frequency of occurrence of codon i, $\sum aa_i$ is the number of times the encoded amino acid is present in the protein sequence and $\sum syn_i$ is the number of synonyms for the amino acid encoded by codon i. An *RSCU* value greater than 1 means that a codon is used more often than expected, whilst values less than 1 indicate its relative rarity. Data analysis was performed using an IBM-compatible microcomputer. Nucleotide, codon and intercodon analyses were performed using the Perl-based program 'Codon Intercodon Analyzer' (CIAN) program (unpublished), and GCUA (McInerney 1998). Data were then subjected to statistical and graphical analysis using SPSS (SPSS Inc., USA), Statistica (STATSOFT, Inc., USA) and Microsoft Excel 2000 (MS, Inc., USA).

RESULTS

Only a very slight AT-bias was identified throughout the coding sequences examined (A = 26·2, C = 25·2, G = 24·2 and T = 24·4), suggesting that general nucleotide usage was not subject to distinct patterning in *Capra*. Comparatively, when considering only the N3, a slight bias for C was discovered (Fig. 1 *a* and Table 1). However, this frequency difference was only 2·5% greater than what would be expected purely by chance (25%). Thus, a mononucleotide usage pattern does not appear to be present within coding sequences from *Capra*.

A pattern was identified, however, when examining nucleotide frequencies specifically at the intercodon position (N4) (Fig. 1 *b*). A was the least common nucleotide at N3 in the presence of T at N4. Similarly, T was the least common nucleotide at N3 in the presence of A at N4. Equivalent results were also obtained for G and C. This suggests that a distinct pattern may be present at the intercodon position in the absence of a global nucleotide bias. Analysis for the influence of base composition on RSCU is shown in Table 2.

Based upon the GC frequency at the third codon position (GC_3), the variable N_c equivalent to the mean number of unique codons used per gene, was calculated and plotted as a function of GC_3 (Fig. 2). N_c is given approximately by the equation $N_c = 2 + s + [29/(s^2 + (1-s)^2)]$, where $s = GC_3$. The presence of bias within the plotted data would suggest that a bias

https://doi.org/10.1017/S0021859602002903 Published online by Cambridge University Press

Table 2. *Relative Synonymous Codon Usage* (*RSCU*) *and mean frequency per gene in* Capra (*the RSCU value acts as a proportional measure of codon usage, as compared with what would be expected if all synonymous codons were used at equal frequencies*)

Amino acid	Codon	Mean RSCU	Mean N/gene	Amino acid	Codon	Mean RSCU	Mean N/gene
Phe	UUU	4	1·14	Ser	UCU	4	1·2
	UUC	3	0·86		UCC	5	1·5
Leu	UUA	3	0·27		UCA	0	0
	UUG	2	0·18		UCG	2	0·6
	CUU	22	2	Pro	CCU	13	0·84
	CUC	9	0·82		CCC	21	1·35
	CUA	15	1·36		CCA	17	1·1
	CUG	15	1·36		CCG	11	0·71
Ile	AUU	3	1·5	Thr	ACU	6	1·5
	AUC	2	1		ACC	5	1·25
	AUA	1	0·5		ACA	3	0·75
Met	AUG	1	1		ACG	2	0·5
Val	GUU	8	1·28	Ala	GCU	22	1·29
	GUC	9	1·44		GCC	19	1·12
	GUA	5	0·8		GCA	13	0·76
	GUG	3	0·48		GCG	14	0·82
Tyr	UAU	0	0	Cys	UGU	6	0·55
	UAC	1	2		UGC	16	1·45
TER	UAA	4	0·46	TER	UGA	22	2·54
UAG	0	0		Trp	UGG	12	1
His	CAU	26	1·24	Arg	CGU	7	0·68
	CAC	16	0·76		CGC	10	0·97
Gln	CAA	29	1·35		CGA	19	1·84
	CAG	14	0·65		CGG	9	0·87
Asn	AAU	2	2	Ser	AGU	5	1·5
	AAC	0	0		AGC	4	1·2
Lys	AAA	6	1·5	Arg	AGA	6	0·58
	AAG	2	0·5		AGG	11	1·06
Asp	GAU	11	1·1	Gly	GGU	14	0·82
	GAC	9	0·9		GGC	15	0·88
Glu	GAA	14	1·27		GGA	21	1·24
	GAG	8	0·73		GGG	18	1·06

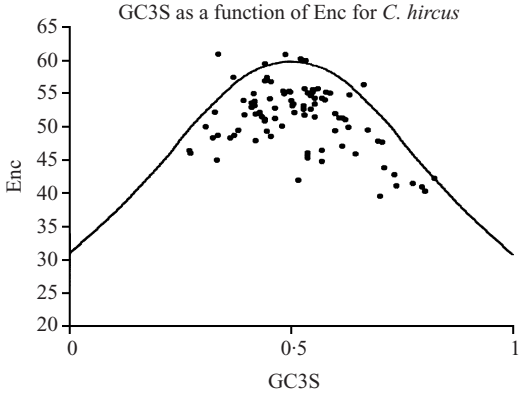

Fig. 2. GC3S (x-axis) plotted as a function for Enc (in this graph it is clear that Enc values are higher than 40).

is present at the codon usage level. However, no such bias was identified; all of the points were distributed evenly around a GC3S value of 50%, as would be expected from random codon selection. This finding argues against the presence of a codon usage bias in *Capra* sp. coding sequences. As shown in Fig. 2, genes have a degree of codon bias that can be explained in terms of GC mutation. This also indicates that these genes are good candidates for genes whose codon usage has been determined by mutational pressure rather than natural selection or translational efficiency.

However, when considering codon frequencies directly, the slight bias identified at the mononucleotide level does appear to be present to a greater extent. An overall bias for the use of A/U ending codons was identified in *Capra* as compared with *Homo sapiens* (Fig. 3). Thus, it appears that although a nucleotide bias is not present at the basal level, a pattern may exist when considering codon usage.

https://doi.org/10.1017/S0021859602002903 Published online by Cambridge University Press

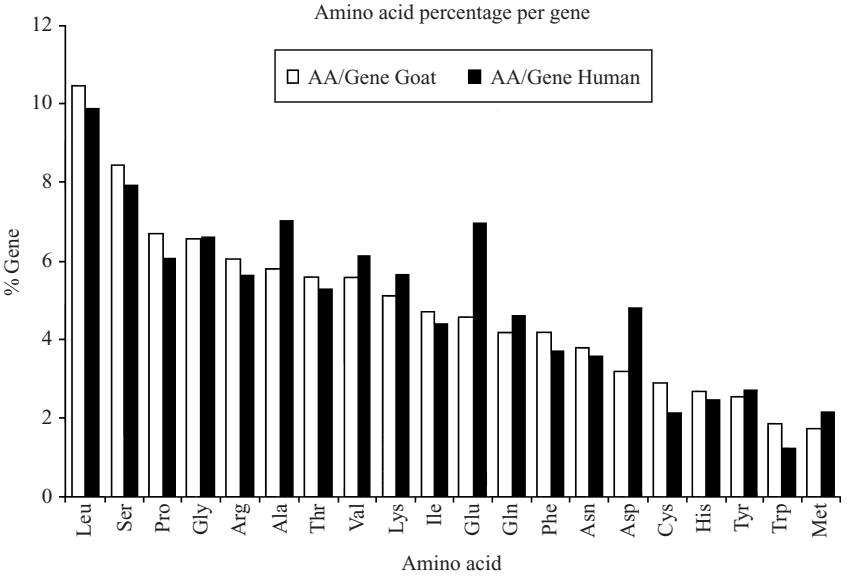

Fig. 3. Graphical representation of the mean proportion of unique amino acids per gene in *Capra*, based upon a 90-gene dataset. Leucine, serine and proline are present more often than any other amino acids. Tryptophan and methionine are present at extremely low levels. These values were compared with those for humans to show level of amino acid usage variations between the two species.

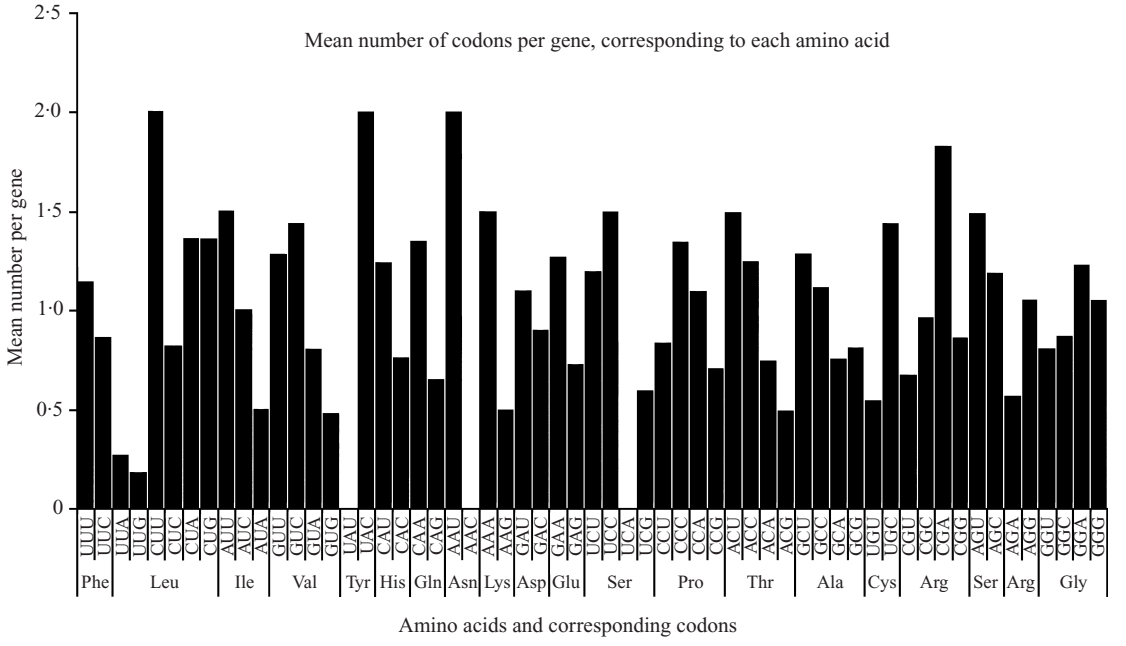

Fig. 4. Codons used for each amino acid in each gene were plotted to facilitate the comparison between GGC versus AT-rich codons and their values.

Differences were also present at the amino acid level in *Capra* as compared with *Homo sapiens*. Levels of glutamine, aspartate and alanine were lower in goat as compared with human (Fig. 3). All other amino acid levels were nearly identical. This suggests that the putative bias identified at the codon usage level may also have an effect on amino acid frequencies. To further investigate whether there is an influence of

https://doi.org/10.1017/S0021859602002903 Published online by Cambridge University Press

Table 3. *Statistical analysis using chi-square, for codon usage in 90* Capra *genes*

AA	Codon	Chi
Phe	UUC	23·5
Leu	CUC	28·8
	CUG	69·8*
Ile	AUC	69·6*
Val	GUG	37·2*
Tyr	UAC	27·7
His	CAC	21·0
Gln	CAG	13·5
Asn	AAC	14·8
Lys	AAG	15·7
Asp	GAC	19·1
Glu	GAG	18·1
Ser	UCC	14·5
	UCG	22·3
Pro	CCC	8·3
	CCG	34·6*
Thr	ACC	35·8*
	ACG	16·7
Ala	GCC	7·1
	GCG	22·2
Cys	UGC	13·6
Arg	CGU	8·9
	CGC	31·7*
	CGG	44·8*
Gly	GGC	7·8
	GGG	7·1

* Codons with very high chi-squared values; AA = Amino acids.

base composition on amino acid codon usage, codons for each amino acid were plotted. As is shown in Fig. 4, the highest usage values were for CUU in Leu, UAC for Tyr, AAU for Asn and CGA for Arg, taking into consideration that the highest amino acid usage values were for Leu, Arg then Asn and considering that CUU, UAC and AAU are AT-rich, while CGA is AT-poor (Fig. 4). To examine further significance levels for codon selection, a chi-squared *t*-test was applied to the whole data set (see Table 3).

Utilizing the power of multivariate analysis, which seeks to identify the most relevant trends governing choice of codon in a given organism, RSCU values were graphically plotted for each codon in each gene using the axis 1 as a function for axis 2 (Fig. 5). This method 'Correspondence analysis' was designed to investigate trends in the dataset and proceeded by plotting all of the codon usage values in an *N*-dimensional hyperspace (the number of dimensions was determined by the number of synonymously degenerate codons in that particular genetic code). The points on this high-dimensional space could resemble a 'cloud'. As shown in Fig. 5, there is a clear absence of any pattern of codon usage variation.

Coding sequences at the extremes of nucleotide composition were then examined and classified. The most GC-biased sequences from *Capra* (Table 4) tended to encode protein members of the globin and globulin protein families, while GC-poor sequences corresponded predominantly to interleukins, interferons and casein proteins.

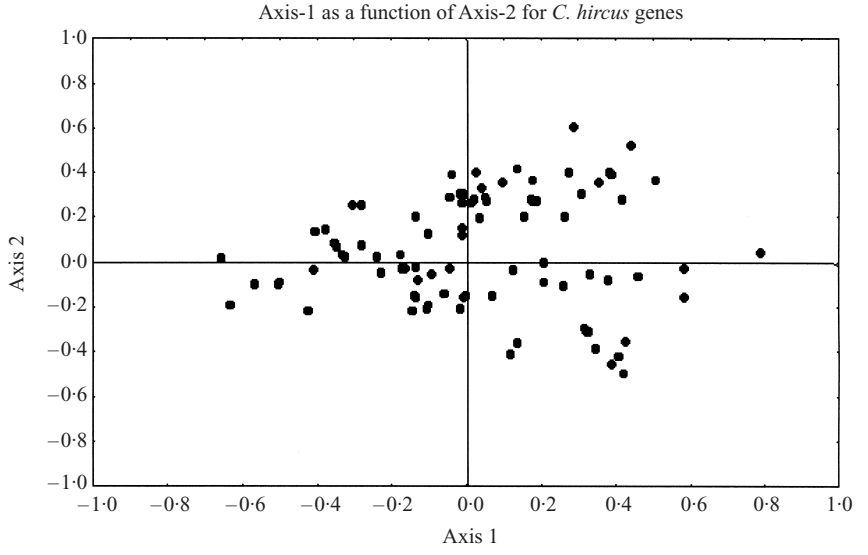

Fig. 5. Metric multidimensional scaling of *Capra* gene data, used to analyse the pattern of codon usage among synonymous codons showing individual gene sequences in the data set. Axes 2 and 1 are plotted against one another, representing the two main sources of variation within the data. Data points are mostly found near to axis 1, indicating that a single main trend is involved in the codon bias in *Capra*.

https://doi.org/10.1017/S0021859602002903 Published online by Cambridge University Press

Table 4. *Genes with high GC content versus those with very low GC content* (*i.e. AT low v. AT-rich genes*)

No.	Accession number	GC %	Gene name
	Genes with highest GC content		
1	Gi|7331159	70·48	Relaxin-like factor; INSL-3
2	Gi|164125	65·63	Alpha-ii-globin gene
3	Gi|5419628	65·44	Beta 3 adrenergic receptor (B3AR) gene
4	Gi|164123	63·32	Alpha-i-globin gene
5	Gi|967	63·05	Beta-lactoglobulin
6	Gi|4126376	62·6	Membrane-type matrix metalloproteinase-1 (MT1MMP)
7	Gi|975	61·3	Growth hormone
8	Gi|1911665	60·5	Tau protein isoform B
	Genes with lowest GC content		
1	Gi|11066965	39·78	Interleukin 2 (IL-2)
2	Gi|955	39·29	As2-casein 1
3	Gi|1730273	39·21	Interferon-gamma
4	Gi|6634063	37·58	Goat beta-casein
5	Gi|6634062	37·56	Goat beta-casein
6	Gi|3806116	36·23	Goat As2-casein 2
7	Gi|3806110	32·28	AlphaS2-casein type C gene
8	Gi|3806115	31·11	Alpha2S-casein type A gene

DISCUSSION

Coding sequences corresponding to *Capra hircus* were investigated at the nucleotide, codon and amino acid levels for the presence of sequence patterns that might be of functional significance. Sequence-based patterns have been discovered in other similar genera and have been shown to affect gene expression levels. The present dataset which was screened for annotated, protein-coding sequences of full-length nuclear DNAs, contains a relatively higher AT frequency. Similar AT biases have been identified in many single-celled organisms such as *Entamoeba histolytica* (Ghosh *et al.* 2000) and *Plasmodium falciparum* (Musto *et al.* 1999), as well as multicellular organisms such as *Brugia malayi* (Fadiel *et al.* 2001). Comparatively complex organisms such as *Caenorhabditis elegans* (Stenico *et al.* 1994), *Drosophila melanogaster* (Marin *et al.* 1998), *Mus musculus* and *Homo sapiens* (Marin *et al.* 1989) have more often been associated with biases for GC base pairs. In certain cases, biases have been shown to affect gene statement levels (Stenico *et al.* 1994). Such base compositional constraints may shape the genomic architecture of *Capra*, thereby influencing gene statement and regulation.

In forming codon analysis, a sequence-specific pattern was identified at the intercodon position as well, with T being uncommon at N3 in the presence of A at N4, and A being uncommon at N3 in the presence of T at N4. Similar results were obtained for G and C, as well. Biases at the intercodon position have previously been identified in species ranging from *Arabidopsis thaliana* (de Amicis & Marchetti 2000), to *Wuchereria bancrofti* (Fadiel *et al.* 2001). Since the nucleotide combinations not commonly found at N3 and N4 represent potential nucleotide binding partners (A with T, G with C), selection might be taking place to inhibit the formation of secondary structure at codon boundaries. It is known that dinucleotide repeat regions made up of AT or GC repeats may often fold upon themselves to form perfect hairpins (Biet *et al.* 1999). However, it is more likely that stem-loop would require longer stretches of base complementarities and/or, in some cases, non-canonical base pairing. Additionally, it has been shown that poly-AT and poly-GC are often bound by proteins involved directly or indirectly in recombination such as single strand binding protein (Ssbp) (Biet *et al.* 1999). Thus, selection against sites that are prone to frequent recombination, may be bound similarly.

In contrast to the slight bias that was identified at the mononucleotide level, codons with A or T at N3 were clearly used more frequently than those with G or C at the wobble position. Therefore, this suggests that a bias favouring the use of codons with A or T at the wobble position may be present within the transcriptome of *Capra*. Functionally, significant AT and GC biases have previously been identified at the codon usage level in many evolutionary diverse organisms, and have often been found to correlate with changes in gene expression level. Therefore, the bias present at the level of codon usage in *Capra* may be of functional importance for proper gene expression. The codon bias thus observed may be reflective of the level of cognate endogenous tRNA species (Kane 1995). Such non-random patterns of synonymous codon usage may cause translational problems for mRNA species containing excess of rare tRNA

codons. It would also be interesting to explore the relationship between gene length and codon usage.

To further elucidate the mechanism of codon usage, a comparison of the goat gene set with the human counterpart was performed. Frequencies of amino acid occurrence in *Capra* were very similar to those observed, in general, throughout *Homo sapiens* coding sequences (see Fig. 4). Cys and Trp, to a lesser extent, exist in relatively greater proportions in the dataset, while in human genome, more Ala, Glu, Asp and Asp exist. The former are coded by codons with N1 = T and the latter are coded by N1 = G. It is possible that the trend in amino acid prevalence may be an artefact of the size of the present dataset. Cys residues could offer a selection advantage by the formation of strong disulphide bonds, while tryptophan may play a more neutral or secondary role. Thus, along with translational efficiency criteria, which may dictate the codon bias in goats, nucleotide composition may qualify the mutational bias in the event of selection pressure. In this respect, it would be of interest to verify the thermodynamic stability of individual mRNA species in the dataset.

To mine for patterns in protein characteristics, the top five proteins coded by genes with high and low GC content were examined (see Table 4). High GC content has been correlated with increased levels of gene expression (Stenico *et al*. 1994). Therefore, genes encoding globin proteins, which are expressed at high levels, are highly GC-rich. With respect to the GC-poor proteins, it is interesting to note that casein proteins are specific to goats and tend to be localized to milk. This class of peptide would be ideal for use in the development and application of goat transgenics. The protein could be used as a vector for therapeutic compounds that could then be provided to patients through milk.

CONCLUSIONS

The goat is an animal model that promises to be of great value to biotechnology and medicine, but not yet characterized from a molecular standpoint. Sequence patterns were identified at the intercodon position in goat coding sequences. Furthermore, codons with A or T at the wobble position tended to be used more often than GC-ending codons. Furthermore, GC content of genes appeared to be functionally correlated with GC-rich genes encoding globins, and GC-poor genes corresponding to caseins. Thus, further characterization of this animal model will be a critical step in the development of novel transgenic approaches for the treatment of a wide range of medical disorders.

REFERENCES

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. & Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958.

Biet, E., Sun, J. S. & Dutreix, M. (1999). Conserved sequence preference in DNA binding among recombination proteins: an effect of ssDNA secondary structure. *Nucleic Acids Research* **27**, 596–600.

de Amicis, F. & Marchetti, S. (2000). Intercodon dinucleotides affect codon choice in plant genes. *Nucleic Acids Research* **28**, 3339–3345.

Eyre-Walker, A. C. (1991). An analysis of codon usage in mammals: selection or mutation bias? *Journal of Molecular Evolution* **33**, 442–449.

Fadiel, A., Lithwick, S., Wanas, M. Q. & Cuticcia, A. J. (2001). Influence of intercodon and base frequencies on codon usage in filarial parasites. *Genomics* **74**, 197–210.

Filipski, J. (1988). Why the rate of silent codon substitutions is variable within a vertebrate genome. *Journal of Theoretical Biology* **134**, 159–164.

Ghosh, T. C., Gupta, S. K. & Majumdar, S. (2000). Studies on codon usage in *Entamoeba histolytica*. *International Journal for Parasitology* **30**, 715–722.

Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pave, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research* **8**, r49–r62.

Iida, K. & Akashi, H. (2000). A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene* **261**, 93–105.

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology* **151**, 389–409.

Kane, J. F. (1995). Effects of rare codon clusters on high-level statement of heterologous proteins in *Eschericia coli*. *Current Opinion in Biotechnology* **6**, 494–500.

Marin, A., Bertranpetit, J., Oliver, J. L. & Medina, J. R. (1989). Variation in G+C-content and codon choice: differences among synonymous codon groups in vertebrate genes. *Nucleic Acids Research* **17**, 6181–6189.

Marin, A., Gonzalez, F., Gutierrez, G. & Oliver, J. L. (1998). Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Research* **26**, 4540.

McInerney, J. O. (1998). GCUA: general codon usage analysis. *Bioinformatics* **14**, 372–373.

Moriyama, E. N. & Hartl, D. L. (1993) Codon usage bias and base composition of nuclear genes in *Drosophila*. *Genetics* **134**, 847–858.

Mouchiroud, D. & Gautier, C. (1990). Codon usage changes and sequence dissimilarity between human and rat. *Journal of Molecular Evolution* **31**, 81–91.

Musto, H., Romero, H., Zavala, A., Jabbari, K. & Bernardi, G. (1999). Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection. *Journal of Molecular Evolution* **49**, 27–35.

SMITH, N. G. & EYRE-WALKER, A. (2001). Why are translationally sub-optimal synonymous codons used in *Escherichia coli*? *Journal of Molecular Evolution* **53**, 225–236.

STENICO, M., LLOYD, A. T. & SHARP, P. M. (1994). Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Research* **22**, 2437–2446.

ZHANG, S. P., ZUBAY, G. & GOLDMAN, E. (1991). Low-usage codons in *Escherichia coli*, yeast, fruit fly and primates. *Gene* **105**, 61–72.