World Rabbit Science

World Rabbit Sci. 2003, 11: 117 - 136

GENOME ANALYSIS OF GENBANK KNOWN RABBIT (*Oryctolagus cuniculus*) GENES

Fadiel A. *, Ganji G. †, Farouk A.‡, Marai I.F.M. §

*Center for Computational Biology, Hospital for Sick Children, 180 Dundas Street West, Ontario, Canada.

†Department of Medical Biophysics, University of Toronto, 610 University Ave., Toronto, Ontario, Canada.

‡Department of Biochemistry, Kulliyyah of Science, IIUM, Jalan Gombak, Kuala Lumpur, Malaysia.

§Department of Animal Production, Faculty of Agriculture, Zagazig University, Zagazig, Egypt.

Abstract: After downloading all known rabbit genes, key words specific to complete coding sequences were used to filter out partial coding sequences. As a result, 160 full-length, nuclear, protein-coding and functionally annotated genes were extracted from GenBank database. These genes were subjected to synonymous codon and amino acid usage analysis. The results showed a clear base composition bias in the genes analyzed. The effective number of codons used (Nc values) ranged between 30.07 and 59.98 with a mean of 51.31 with a standard deviation (SD) of 5.71. The frequency of G + C at the synonymous third position of codons (GC3s) varied between 0.3 and 0.96 with a mean of 0.55 and a SD of 0.14, clearly indicating marked variation and heterogeneity in codon usage patterns among the different genes. The distribution of relative synonymous codon usage (RSCU) values calculated for all the genes indicated that codons with a G or C in the third position are widely employed, and although RCSU values for A/T or G/C ending codons are almost equal, G/C appears to play a dominant role. This pattern of dominance was confirmed by the distribution of amino acid occurrence. Leucine (CUN) and serine (UCN) were the most frequent amino acids, followed by arginine (CGN), proline (CCN), glycine (GGN) and alanine (GCN), in relatively equal proportions. The heterogeneity observed in the analyzed genes was then further probed by multivariate statistical analyses. A major trend in codon usage that correlated with gene expression values was revealed. These findings suggest that translational efficiency exerts a stronger influence on codon usage preference than compositional bias in the sampled rabbit genes.

Key words: genome, rabbit, genbank.

Correspondence: I.F.M. Marai
E-mail: marai20022000@yahoo.co.uk

INTRODUCTION

Recently, the rabbit has attracted much attention from the biotechnology community. Several features make it an attractive model for transgenics and cloning, including: the rapid onset of sexual maturity, a short gestation period, a relatively larger number of offspring per litter, year-round reproductive capacity, and an average life-span of 9 years. Also noteworthy is that the rabbit genome is estimated to be 3 billion base pairs long, almost equal to the size of the human genome. In addition, rabbits have similar lipid metabolism to humans, making them good models of atherosclerosis and re-stenosis (Dove, 2000). Together, these features make the rabbit an ideal choice for genomic analysis.

GC (Guanine-Cytosine) content (when GC are found together with G at the second position and C at the third) is an important factor in determining the structure and evolution of genomes (Muto and Osawa, 1987; Bernardi, 1993; Bellgard and Gojobori, 1999). GC content correlates with codon and amino acid usage, and demonstrates that genomes subjected to mutation/selection forces forge a connection between base and amino acid composition (Knight *et al.*, 2001). Such content is used frequently for inferring neutral substitution rates since synonymous substitutions in protein-coding genes are generally free from natural selection. Four-fold degeneration sites are expected to harbour only neutral substitutions because these sites are synonynmous at the amino acid sequences level. However, due to compositional bias, which is, in part, driven by mutational forces at the nucleotide level and/or selection factors, the third nucleotide position in synonymous codons for four-fold degenerated amino acids are not randomly chosen, but follow a pattern that determines the codon preference in the genes under study.

Codon usage has been well characterized in several species, as reviewed by Knight *et al.* (2001). This characterization has broad applications such as in designing PCR (polymerase chain reaction) primers for uncharacterized genomes, maximizing success for *in vivo* genetic manipulation and in molecular phylogenetic reconstruction (Knight *et al.,* 2001).

The aim of the present study was to gain insight into the characteristics of the rabbit genome, to elucidate the characteristic base composition of genes, codon usage bias, amino acid occurrence, their influence on genome function and the forces that shape them in an attempt to comprehend the genomic structure and promote future biotechnological undertakings in this area.

MATERIALS AND METHODS

In the present study, genes were downloaded from the GenBank and compiled into a working dataset after filtering out non-coding and partial gene sequences using specific key word searches, by which only nuclear protein-coding genes were considered for further analysis.

Rabbit gene sequences

Oryctolagus cuniculus sequences were extracted primarily from the GenBank (www.ncbi.nlm.nih.gov/entrez) and other similar databases (EMBL and DDBJ). Despite the fact that the database contains more than one thousand sequences, many were short fragments, redundant or expressed sequence tags (ESTs). These sequences were filtered out to eliminate unreliable open frames or any possible frame shifting. Only complete coding sequences were used to verify the full-length protein-coding genes. The filtered dataset contained 160 genes and was verified to include only full-length nuclear protein-coding genes.

Codon usage bias indices

Codon usage values were converted to relative synonymous codon usage (RSCU) values by dividing the observed frequency by that frequency which would be expected if all codons for the same amino acid were used equally (SHARP *et al.,* 1986). The in-house programs CIAG (Codon Inter-codon and Alien Gene analysis; FADIEL Unpublished) and other publicly available software packages such as CodonW 1.4 (http://enterprise.molbiol.ox.ac.uk/cu/) and GCUA (General codon usage analysis) (MCINERNEY, 1998) were employed for analysis. CIAG calculates codon and inter-

codon frequencies. CodonW first determines whether an amino acid is synonymous or non-synonymous, translates a codon into its cognate amino acid, and then calculates the number of codons in a codon family and the number of synonyms that each codon possesses. GCUA was used for codon usage analysis by calculating RSCU values.

Genomic analysis parameters

RSCU is the ratio of a codon's observed frequency to that frequency which would be expected if all the synonymous codons for those amino acids produced by the gene were distributed equally. To analyze codon usage variation between genes, it is necessary to derive the RSCU values for each gene as follows:

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

where $RSCU_i$ is the Relative Synonymous Codon Usage value for codon I, Obs_i is the observed number of occurrences of codon I, and Exp_i is the expected number of occurrences of codon $_i$. The expected number of occurrences of a codon is calculated as follows (Equation 2):

$$Exp_i = Saa_i/Ssyn_i \qquad (2)$$

Saa_i is the number of times the encoded amino acid is present in the protein sequence, and $Ssyn_i$ is the number of synonyms for the amino acid encoded by codon i. RSCU values are useful in comparing codon usage variation among the genes. Values of RSCU greater than 1.0 indicate that the corresponding codon appears more frequently than expected, while the contrary is true for RSCU values less than 1.0.

GC3s is the frequency of G+C at the synonymous third position of codons. Methionine (Met) and Tryptophan (Trp) were excluded, since each is coded by single codon (Methionine by AUG and Tryptophan by UGG), and therefore no synonymous codons for these amino acids exist.

Nc is the "effective number of codons used by a gene" (WRIGHT, 1990). It is a measure of the codon usage specificity in each organism (WRIGHT, 1990; COMERON and AGUADE, 1998). Nc is generally used to quantify how a gene is utilizing a small subset of codons. Nc values range between 20 and 61 codons. The expected value of Nc under random codon usage is given by:

$$Nc = 2 + GC3 + \{29/[GC3 + (1 - GC3)^2]\}$$

If GC3s deviates from 0.5, the expected value of Nc in the absence of any other source of bias declines.

CA (Correspondence analysis) was used to investigate major trends in codon usage variation among genes (GREENACRE, 1992). Since codon usage by its very nature is not independent but multivariate, it is necessary to analyze this data with a multivariate statistical technique. The main advantage of CA is that data are examined without any *a priori* assumptions. Correspondence analysis constructs a series of orthogonal axes through multidimensional space occupied by the genes with the axes identified in the order of importance as assessed by the fraction of the total variation among RSCU values of genes accounted for.

Fop is the 'frequency of optimal codons' for *O. cuniculus* genes. Codons are identified as translationally optimal based on their significantly higher frequency in highly expressed genes. Using a larger dataset of 575 genes, Sharp and Cowe (1991) identified 18 optimal codons. Eighteen optimal codons for 16 amino acids (excluding Cystine (Cys) and Glycine (Gln)) are used. Fop is calculated as the occurrence of these 18 codons divided by the total occurrence of codons for the 16 amino acids.

CAI is the 'Codon Adaptation Index' applied to yeast (*S. cerevisiae*) genes (SHARP and LI, 1987). Fitness values have been assigned to codons in yeast based on their frequency of use in genes expressed at very high levels. The CAI is calculated as the geometric mean of the fitness values for the codons in a gene and are expressed as values from about 0.07 to 1.0 (SHARP and COWE, 1991).

N and N3 are expressions for the variability at the third synonymous codon position where N represents the individual bases A, T, G, and C. It follows that N3 represents any of the four bases at the third codon position (i.e., A3, T3, G3, or C3). The general base composition at the third position was used for the studies of codon usage variation among the genes of *O. cuniculus*.

Statistical and data analysis

Data reduction and statistical analysis were done using GCUA MCINERNEY (1998). This program performs most of the codon usage/bias parameters. This program also performs multivariate analysis using its imbedded functions. Statistical and graphical analyses were also done using SPSS (SPSS Inc., USA) or Statistica (STATSOFT, Inc, USA).

RESULTS AND DISCUSSION

Codon usage was analyzed in a dataset containing 160 *Oryctolagus cuniculus* genes (Table 1). Two indices for determination of codon usage bias were calculated, namely, effective number of codons used by a gene (Nc) and GC content at the variable third position of synonymous codons (GC3s) in a specific gene. These parameters have been widely used to study codon usage variation among genes. In Table 1, it was observed that Nc values ranged between 30.07 and 59.98 with a mean of 51.31 and standard deviation (SD) of 5.71, while GC3s varied between 0.3 and 0.96 with a mean of 0.55 and SD of 0.14, clearly indicating marked variation and heterogeneity in codon usage patterns among different genes.

The distribution of the relative synonymous codon usage values calculated for all genes in this study indicated that codons with a G or C in the third position were widely employed (Table 2). Although RCSU values for A/T or G/C ending codons were almost equal, G/C appeared to play a dominant role as evidenced by the high GC and GC3 content of genes in the present dataset (Figure 1). The distribution of A, T, G and C (Figure 1) at the third position of codons shows that C occurred most

Table 1: Descriptive list of *Oryctolagus cuniculus* gene sequences: accession number, length (in base pairs; Bp) of the corresponding cDNA, G+C content at the synonymous codon third positions (GC3s) and the effective number of codons (Nc).

Accession No.	Length	GC3s	Nc	Accession No.	Length	GC3s	Nc
Gi* |1772	2122	0.53	53.98	Gi |5819091	193	0.75	45.39
Gi |165094	610	0.59	52.29	Gi |5853353	616	0.46	54.7
Gi |165095	626	0.63	49.53	Gi |5882168	391	0.45	57.25
Gi |217725	1519	0.4	51.05	Gi |5924350	789	0.73	47.58
Gi |217727	1661	0.37	51.17	Gi |5929757	1856	0.42	51.93
Gi |217745	353	0.67	49.37	Gi |5929759	237	0.42	48.31
Gi |402728	518	0.55	52.43	Gi |6007842	1429	0.37	49.45
Gi |463994	1533	0.41	51.91	Gi |6224927	746	0.53	55.49
Gi |463998	1533	0.41	51.71	Gi |6224935	655	0.87	34.58
Gi |483788	646	0.37	54.39	Gi |6526706	430	0.4	55.86
Gi |538244	467	0.47	54.54	Gi |6573131	1322	0.4	55.75
Gi |606794	248	0.52	53.99	Gi |6579182	930	0.59	51.22
Gi |688171	287	0.58	49.34	Gi |6581071	1467	0.69	49.2
Gi |755094	221	0.46	51.62	Gi |6653656	522	0.62	51.58
Gi |755096	314	0.63	50.93	Gi |6653658	869	0.79	41.7
Gi |755098	381	0.5	55.66	Gi |6653660	283	0.65	48.9
Gi |755100	240	0.59	55.6	Gi |6653662	304	0.41	50.31
Gi |862605	291	0.54	54.39	Gi |6653664	292	0.48	48.58
Gi |1052872	431	0.42	51.17	Gi |6714958	605	0.42	55.3
Gi |1098793	866	0.43	53.48	Gi |6715103	1320	0.54	55.03
Gi |1100744	302	0.62	45.74	Gi |6901679	528	0.71	43.16
Gi |1100984	1450	0.44	53.26	Gi |7108516	351	0.83	39.12

Accession No.	Length	GC3s	Nc	Accession No.	Length	GC3s	Nc
Gi |1109677	1732	0.63	52.49	Gi |7140577	219	0.59	48.86
Gi |1144320	440	0.35	48.79	Gi |7243263	441	0.85	36.22
Gi |1144322	346	0.37	47.27	Gi |7259608	530	0.49	56.71
Gi |1144490	674	0.3	49.68	Gi |7321999	520	0.53	57.72
Gi |1236053	1400	0.37	51.39	Gi |7322009	513	0.45	59.98
Gi |1246305	177	0.77	41.67	Gi |7322013	519	0.57	54.52
Gi |1381182	1124	0.58	52.94	Gi |7644353	1550	0.45	56.05
Gi |1399213	1200	0.57	50.93	Gi |7920150	1097	0.57	54.48
Gi |1490412	312	0.82	39.52	Gi |8388948	220	0.59	54.38
Gi |1655929	996	0.46	52.3	Gi |8572238	956	0.83	39.01
Gi |1655931	1013	0.46	52.37	Gi |8886353	744	0.4	51.92
Gi |1816641	4655	0.4	50.87	Gi |9294750	133	0.33	34.8
Gi |1835272	5236	0.38	50.13	Gi |9294752	187	0.48	49.32
Gi |1857434	809	0.59	49.2	Gi |9294754	130	0.56	45
Gi |2058513	341	0.48	55.04	Gi |9294756	61	0.41	47.53
Gi |2058515	371	0.48	56.92	Gi |9294758	67	0.47	47.55
Gi |2098597	107	0.4	50.06	Gi |9294760	144	0.57	48.86
Gi |2197096	371	0.62	50.21	Gi |9294762	136	0.57	52.77
Gi |2197098	343	0.64	49.58	Gi |9294764	136	0.59	56.29
Gi |2326268	403	0.67	51.22	Gi |9294776	139	0.46	53.73
Gi |2460305	427	0.56	53.5	Gi |9367794	2828	0.62	51.11
Gi |2599069	1028	0.87	35.62	Gi |9798667	429	0.5	46.53
Gi |2654530	711	0.43	55.1	Gi |9954411	589	0.61	52.29
Gi |2655056	1488	0.51	58.27	Gi |10121883	177	0.86	35.85
Gi |2745943	282	0.42	50.09	Gi |10567262	469	0.34	51.48

Accession No.	Length	GC3s	Nc	Accession No.	Length	GC3s	Nc
Gi |2997711	193	0.6	55.3	Gi |11041695	991	0.49	52.63
Gi |2997712	608	0.64	50.87	Gi |11041717	910	0.54	49.75
Gi |3089538	1011	0.39	52.2	Gi |11041721	1008	0.8	40.19
Gi |3098310	581	0.96	30.07	Gi |11094229	426	0.54	55.63
Gi |3108210	226	0.44	52.25	Gi |11527284	195	0.8	41.03
Gi |3241846	592	0.56	54.86	Gi |11527286	1403	0.63	52.76
Gi |3241848	1062	0.46	55.9	Gi |11559415	400	0.69	46.34
Gi |3392985	1602	0.48	59.98	Gi |11611536	1050	0.69	45.36
Gi |3411194	696	0.51	53.49	Gi |11611538	1034	0.69	45.38
Gi |3641531	435	0.41	57.28	Gi |11935050	989	0.42	50.87
Gi |3660677	558	0.42	55.57	Gi |12003422	2019	0.75	43.35
Gi |4008113	810	0.64	51.48	Gi |12003424	1926	0.56	47.23
Gi |4039108	619	0.64	53.39	Gi |12003426	1996	0.75	43.07
Gi |4039111	515	0.53	53.59	Gi |12248876	1514	0.54	56.29
Gi |4039112	1322	0.59	53.45	Gi |12248892	746	0.58	51.81
Gi |4063500	687	0.64	54.77	Gi |12743894	376	0.37	54.78
Gi |4102716	254	0.86	33.17	Gi |12743896	376	0.37	54.12
Gi |4103590	810	0.41	55.56	Gi |12743898	335	0.32	48.09
Gi |4105816	207	0.65	41.8	Gi |12958627	1276	0.42	55.49
Gi |4336880	338	0.71	47.44	Gi |13022000	704	0.61	49.06
Gi |4337014	1064	0.55	53.39	Gi |13022003	696	0.68	44.65
Gi |4996370	605	0.35	46.25	Gi |13182930	428	0.47	59.77
Gi |4996372	868	0.38	49.54	Gi |14142009	465	0.4	53.1
Gi |4996374	319	0.37	48.3	Gi |14486160	548	0.75	45.02
Gi |4996376	320	0.37	48.32	Gi |14486162	595	0.76	41.96

Accession No.	Length	GC3s	Nc	Accession No.	Length	GC3s	Nc
Gi |5006362	1014	0.55	53.99	Gi |14583089	578	0.82	39.63
Gi |5360234	484	0.48	53.38	Gi |14701787	263	0.56	54.84
Gi |5442270	606	0.41	54.77	Gi |15216294	552	0.46	55.83
Gi |5708066	1241	0.55	50.33	Gi |15216296	552	0.76	41.97
Gi |5759113	759	0.41	56.5	Gi |17066585	1148	0.58	49.95
Gi |5805340	1072	0.64	50.75	Gi |17646193	1443	0.46	53.9
Gi |5805342	1260	0.68	49.5	Gi |18377357	576	0.62	49.98
Gi |5814018	2084	0.45	55.47	Gi |18409575	476	0.75	42.07

* Gi (GeneBank ID)= I dentification number for gene sequences.

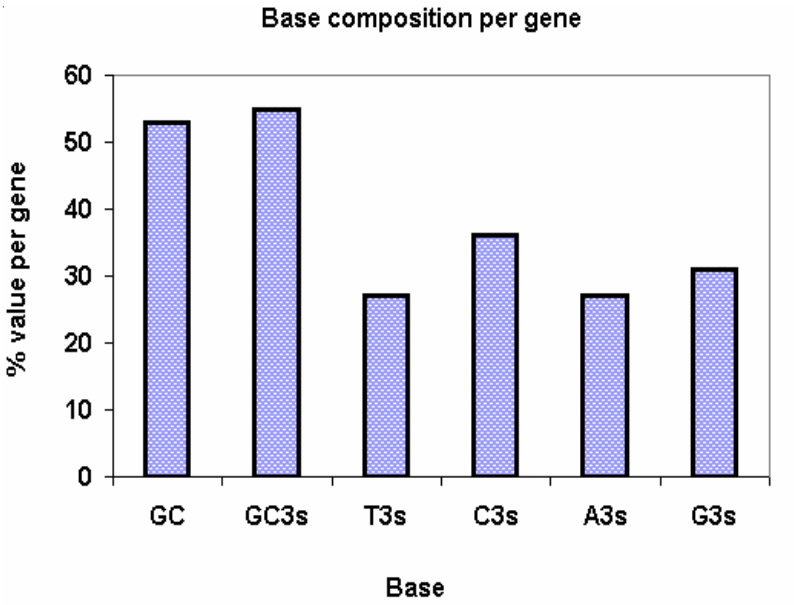

Figure 1: Distribution of base(s) at the synonymous third position of codons in the rabbit genes dataset. The X-axis represents individual nucleotide (T3s, C3s, A3s, G3s) or combinations of nucleotides (GC3s) in the synonymous third positions of codons in the dataset. The Y-axis shows the average percentage of distribution of N3s (A3s, T3s, C3s or G3s) or GC3s (G or C in the third position of synonymous codons) per gene in the same collection. GC is the average G+C content per gene.

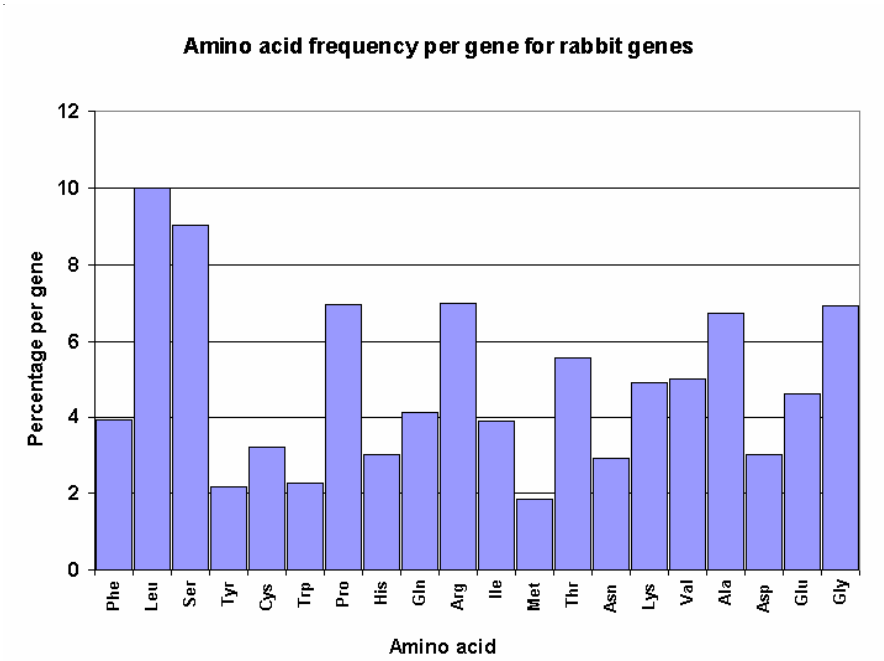

Figure 2: Amino acid usage percentage (per gene) for each amino acid in the entire dataset of rabbit genes. Amino acids are shown on the X-axis and their respective average distributions per gene on the Y-axis.

frequently, followed by G, A and T, respectively. Although the latter two were relatively similar in their frequency of occurrence (i.e., G+C content is higher than A+T). Interestingly, a similar G+C skew has been observed in the rabbit mitochondrial genome (GISSI *et al.,* 1998). Thus, overall RSCU values were consistent in that codon preference among the genes was influenced by compositional or mutational bias. A similar pattern was confirmed by the distribution of amino acid occurrence, which is shown in Figure 2. Leucine (CUN) and Serine (UCN) were the most frequent amino acids, followed by Arginine (CGN), Proline (CCN), Glycine (GGN) and Alanine (GCN), in relatively equal proportions.

In order to further probe the influence of base composition on codon selection in rabbit genes, the effective number of codons (Nc) was plotted against GC3s (Figure 3). This method, first proposed by WRIGHT (1990), examined codon usage bias by comparing the actual distribution of the genes with the distribution expected under no selection (i.e., under random codon usage with the Nc values expected if the

Table 2: Relative synonymous codon usage (RSCU) values for each amino acid (AA) produced by the genes used in our analysis which contain a total of N codons.

AA	Codon*	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	2646	-1.05	Ser	UCU	2232	-1.16
	UUC	2412	-0.95		UCC	2477	-1.28
Leu	UUA	1258	-0.59		UCA	1997	-1.03
	UUG	1865	-0.87		UCG	765	-0.4
Tyr	UAU	1339	-0.95	Cys	UGU	1868	-0.9
	UAC	1488	-1.05		UGC	2295	-1.1
Ter**	UAA	1061	0	ter	UGA	1890	0
Ter	UAG	696	0	Trp	UGG	2882	-1
Leu	CUU	2014	-0.94	Pro	CCU	2439	-1.1
	CUC	2589	-1.21		CCC	2626	-1.18
	CUA	1016	-0.48		CCA	2605	-1.17
	CUG	4088	-1.91		CCG	1215	-0.55
His	CAU	1812	-0.94	Arg	CGU	676	-0.45
	CAC	2033	-1.06		CGC	1210	-0.81
Gln	CAA	1950	-0.74		CGA	746	-0.5
	CAG	3333	-1.26		CGG	1361	-0.91
Ile	AUU	1858	-1.11	Thr	ACU	1770	-1
	AUC	1993	-1.19		ACC	2215	-1.25
	AUA	1160	-0.69		ACA	2119	-1.2
Met	AUG	2374	-1		ACG	986	-0.56
Asn	AAU	1787	-0.95	Ser	AGU	1611	-0.83
	AAC	1982	-1.05		AGC	2505	-1.3

AA	Codon*	N	RSCU	AA	Codon	N	RSCU
Lys	AA<u>A</u>	3114	-1	Arg	AG<u>A</u>	2597	-1.74
	AA<u>G</u>	3145	-1		AG<u>G</u>	2350	-1.58
Val	GU<u>U</u>	1361	-0.85	Ala	GC<u>U</u>	2387	-1.11
	GU<u>C</u>	1608	-1		GC<u>C</u>	3084	-1.43
	GU<u>A</u>	874	-0.54		GC<u>A</u>	2092	-0.97
	GU<u>G</u>	2573	-1.6		GC<u>G</u>	1052	-0.49
Asp	GA<u>U</u>	1688	-0.88	Gly	GG<u>U</u>	1375	-0.62
	GA<u>C</u>	2165	-1.12		GG<u>C</u>	2633	-1.19
Glu	GA<u>A</u>	2680	-0.9		GG<u>A</u>	2461	-1.11
	GA<u>G</u>	3243	-1.1		GG<u>G</u>	2396	-1.08

* Third position codons are underlined. **ter refers to the termination or stop codon.

codon usage was determined only by GC content). In other words, if GC3s were the only determinant of Nc, Nc would fall on the continuous curve when calculated for a random distribution. In Figure 3, the majority of points lie close to this 'expected' curve, signifying a strong codon preference constrained by composition bias. However, several points were seen lying well below the expected curve, suggesting that they were independent of GC3s and that other selection factors might be acting upon them. To investigate other possible trends in codon usage variation among the genes under study, the data were subjected to correspondence analysis, a multivariate statistical method (GREENACRE, 1992). In Figure 4, the distribution of genes on Axis 1 represented 19.5% of total variation, compared to 9.7% on Axis 2. The disparity in values between the first and second axes depicts an important primary trend in codon usage across genes. The genes presented in Table 1 are in their order of appearance on Axis 1, which was derived from correspondence analysis.

Since codon usage is also known to be an indicator of gene expression level, genes found at the extremities of the plot on Axis 1 (Figure 4) would be expected to possess reciprocal expression values. Genes appearing at the extreme right of Axis-

1 (Figure 4 and Table 3), EEF1A-2, PrP, KCND3, PTHrP and NCCT are believed to be highly expressed, while those to the extreme left of Axis 1, IL-2, Fas-Ag, PON3, CT3 and Fas-Ags are predicted to be lowly expressed. Upon consulting the appropriate literature, it was possible to explain some of the present findings. The gene coding for elongation factor 1A-2 (EEF1A-2), seen earlier as an outlier in Figure 3, plays a central role in the protein synthesis machinery, where it mediates the transfer of aminoacylated-tRNA to the acceptor site of the ribosome in a GTP (Guanine triphosphate)-dependent manner. In addition, the same factor has been associated with a large number of other cellular functions, such as cytoskeleton remodeling, activation of phospholipid signaling, and ubiquitin-dependent protein degradation, etc (KAHNS *et al.,* 1998). The literature evidence purports that these genes are highly expressed ubiquitously or in a tissue or organ-specific manner (KAHNS *et al.,* 1998; WANG *et al.*, 1999 and POSTMA *et al*., 2000)

Table 3: List of top five genes that were predicted to be highly or lowly expressed.

Genes that are predicted to be highly expressed			
No	Accession No.	Gene Name	Gene Symbol
1	Gi |3098310,	Elongation factor 1 A2	(EEF1A-2)
2	Gi |4102716,	Prion protein	(PrP)
3	Gi |6224935|	Potassium channel Kv4.3	(KCND3)
4	Gi |10121883|	Parathyroid hormone-related protein	(PTHrP)
5-	Gi |2599069|	Thiazide-sensitive sodium chloride co-transporter	(NCCT)
Genes that are predicted to be lowly expressed			
1	Gi |9294750|	Interleukin 2 variant IL2delta2 (alternatively spliced)	(IL-2)
2	Gi |4996370|	Fas antigen	Fas-Ag
3	Gi |12743898|	Paraoxonase 3	(PON3)
4	Gi |1144490|	Cardiac triadin isoform 3	(CT3)
5	Gi |4996372|	Fas antigen spliced variant	Fas-Ags

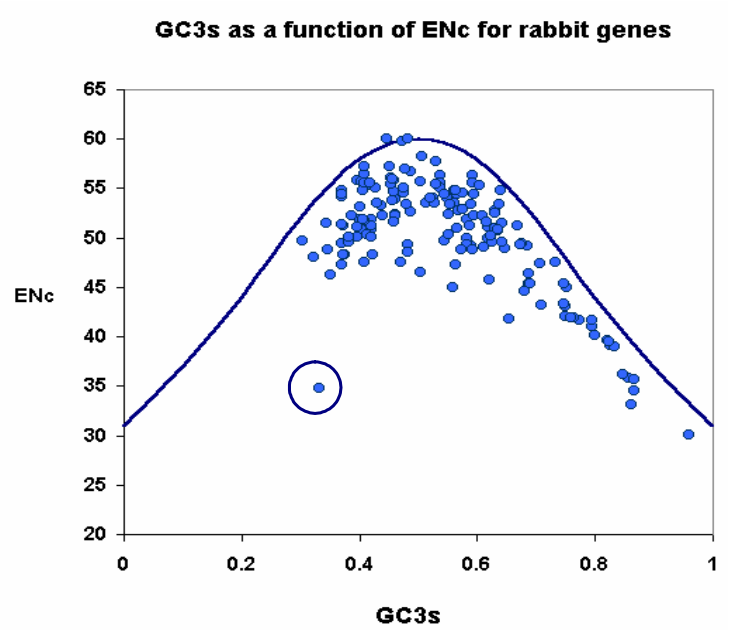

Figure 3: Effective numbers of codons (ENc= Nc) plotted against G+C content at synonymous variable third positions (GC3s) for each of the 160 rabbit genes. The continuous curve represents the relationship between GC3s and Nc under random codon usage where any bias is attributed to base composition (here, G+C content) only. One of the outliers is circled.

On the other hand, as seen in Table 3 and Figure 4, there were five genes with putative low expression values. These proteins included: an IL-2 variant; Fas-Ags, a Fas antigen spliced variant; CT3, a calcium channel protein in skeletal and cardiac muscle; and PON3, a member of the paraxonase gene family. A review of the literature involving the same genes confirmed the hypothesis that predictive high or low gene expression made by the correspondence analysis plot (Figure 4) coincides with the actual levels in the rabbit genome (GUO *et al.,* 1996; KOBAYASHI and JONES, 1999, PERKINS *et al.,* 2000 and WATANABE *et al.,* 2002).

In Figure 5, the highly expressed genes obtained from correspondence analysis were used to calculate the frequency of optimal codons (Fop) then plotted against the CAI of the corresponding orthologs in yeast. The linear correlation further confirms our hypothesis that Axis 1 (Figure 4) represents gene expression as being the primary trend in codon usage. It is known that the codon usage bias in any organism is under the influence of two opposing forces - mutational bias or selection factors such as

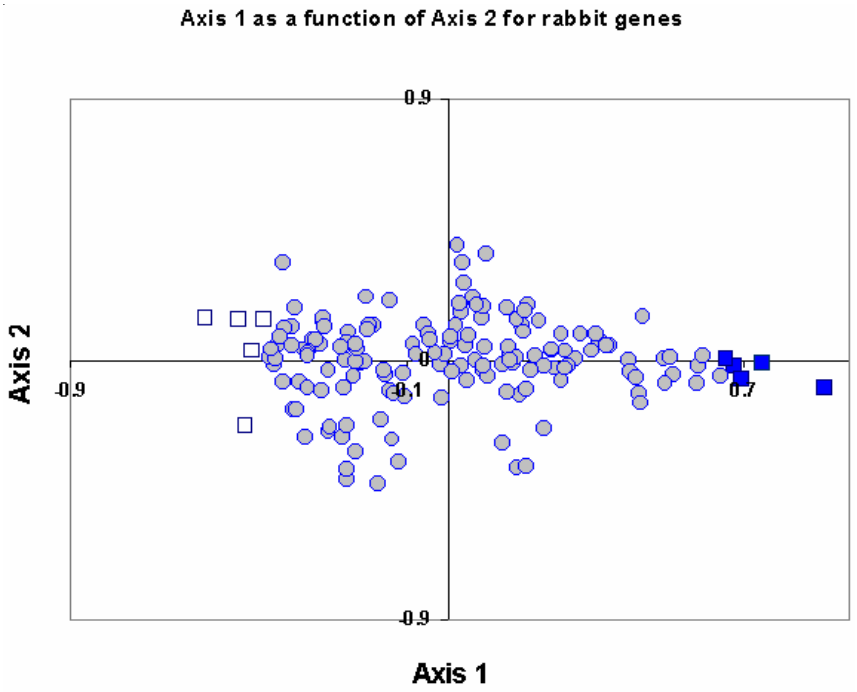

Figure 4: Correspondence analysis of codon usage variation among 160 rabbit genes. Genes are plotted at their co-ordinates on the first two axes. Genes that show extreme outlier behaviour relative to other genes are marked on both sides as open square (left) and closed square (right). The two genes that show clear outliers pattern were the *Oryctolagus cuniculus* elongation factor 1A-2 (EEF1A-2) and *Oryctolagus cuniculus* prion protein (PrP) (see text and Table 3).

translational efficiency. The five genes predicted to be highly or lowly expressed (Table 3) demonstrate correspondingly high and low GC3s values (Table 1), respectively. The high G+C content in the present dataset suggests that codon bias is exclusively governed by mutational bias. However, when Axis 1 (from Figure 4), represented maximum variation in RSCU values of selected genes and is plotted as a function of GC3s, no expected linear relationship exists (Figure 6), revealing a hidden influence. The horizontal nature of this plot may indicate that the force of translational efficiency has a greater influence on the rabbit genes codon usage than mutational bias, hypothetically represented by the dotted line in Figure 6. Investigation of the relative levels of cognate tRNAs for such preferred codons would help in confirming this hypothesis. These findings suggest that selection forces shape the

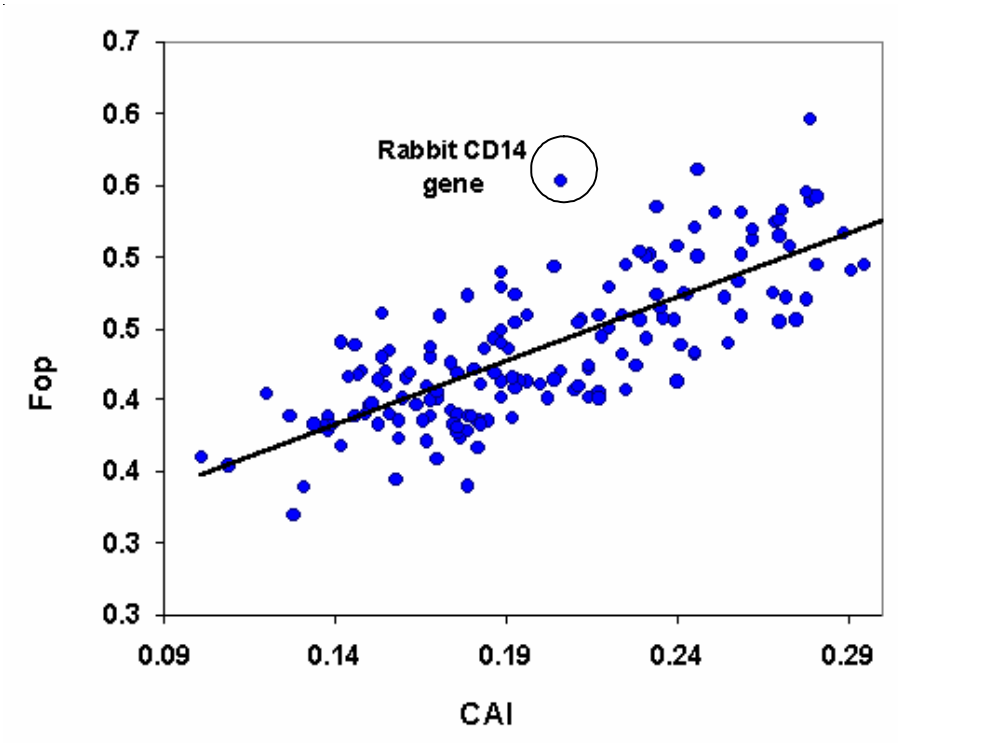

Figure 5: Codon usage bias in homologue genes of the rabbit (measured by Fop) and S. cereviseae (Measured by CAI). Genes that show values out of the pattern range are circled.

codon preference in the rabbit genes under investigation despite the presence of a strong compositional bias.

CONCLUSIONS

It is highly likely that the rabbit's genome possesses similar characteristics to other mammalian genomes which are AT-rich and contains genes with GC-rich regions. The strong codon bias seen in the present study provides useful information on designing more effective PCR primers for Quantitative trait loci (QTL) analysis, and choosing candidates for genetic analysis. Discovering that translational efficiency plays a dominant role in determination of codon bias, despite the strong compositional/

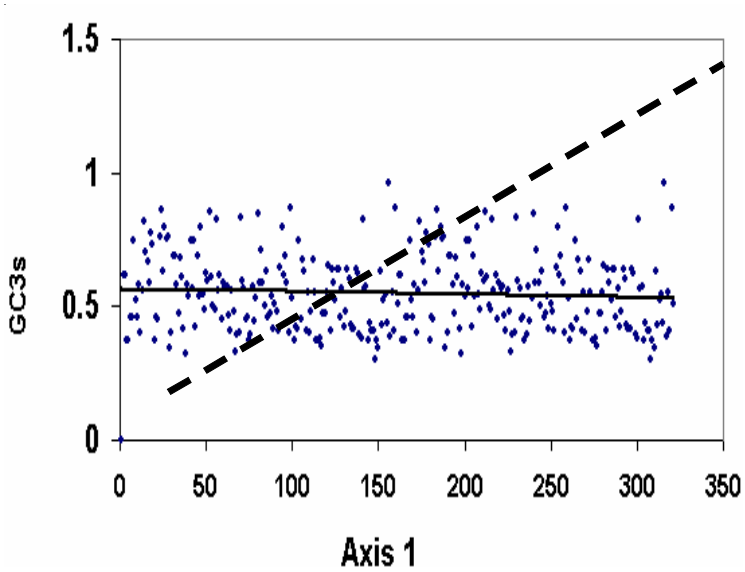

Figure 6: Verification of translational efficiency as a major trend. Axis 1 (Figure 4) is plotted as a function of GC3s. The dotted line represents expected trend for a mutational bias.

mutational bias in *Oryctolagus cuniculus*, reiterates the importance of the differential contributions of the forces of selection, mutation and random drift to codon bias in mammalian evolution. At the same time, as more rabbit genes are sequenced and more knowledge is gathered, it will be of interest to follow the extent that these present findings can be confirmed. With a larger dataset, candidates for horizontal gene transfer showing high expression patterns with unusual codon usage may be identified. CAI values could also be calculated specifically for the rabbit genome. With its many advantages, the importance of the rabbit in modern genetics will surely increase.

Acknowledgements: We thank Dr. Adam Kovac (Humboldt University, Germany) for critical reading of the manuscript and valuable discussions. We also thank Ms. Eshrak Zaky (University of Toronto, Canada) for providing advice throughout the development of this manuscript.

REFERENCES

BELLGARD M I, GOJOBORI T. 1999. Inferring the direction of evolutionary changes of genomic base composition. *Trends Genet., 15(7), 254-256*.

Bernardi G. 1993. The isochore organization of the human genome and its evolutionary history (A review). *Gene, 135(1-2), 57-66.*

Comeron J.M., Aguade M.1998. An evaluation of measures of synonymous codon usage bias. *J. Mol. Evol., 47(3), 268-274.*

Dove A. 2000. Milking the genome for profit. *Nat. Biotechnol., 18(10), 1045-1048.*

Gissi C., Gullberg A., Amason U. 1998. The complete mitochondrial DNA sequence of the rabbit, *Oryctolagus cuniculus*. *Genomics, 50(2), 161-169.*

Greenacre M. 1992. Correspondence analysis in medical research. *Stat. Methods Med. Res., 1(1), 97-117.*

Guo W., Jorgenson A.O., Jones L.R., Campbell K.P. 1996. Biochemical characterization and molecular cloning of cardiac triadin. *J. Biol. Chem. 271(1): 458-465.*

Kahns S., Lund A., Lund A., Kristensen P., Knudsen C.R., Clark B.F., Cavallius J., Merrick W.C. 1998. The elongation factor 1 A-2 isoform from rabbit: cloning of the cDNA and characterization of the protein. *Nucleic Acids Res., 26(8), 1884-1890.*

Knight R.D., Freeland S.J., 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(4), RESEARCH0010.*

Kobayashi Y.M., Jones L.R.. 1999. Identification of triadin 1 as the predominant triadin isoform expressed in mammalian myocardium. *J. Biol. Chem., 274(40), 28660-28668.*

Mcinerney J.O. 1998. GCUA: general codon analysis. *Bioinformatics, 14(4), 372-373.*

Muto A., Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. U.S.A., 84(1), 166-169.*

Perkins H.D., Van Leeuwen B.H., Hardy C.M., Kerr P.J. 2000. The complete cDNA sequences of IL-2, IL-4, IL-6 AND IL-10 from the European rabbit (*Oryctolagus cuniculus)*. *Cytokine, 12(6), 555-565.*

Postma A.V., Bezzina C.R., De Vries J.F., Wilde A.A., Moorman A.F., Mannens M.M. 2000. Genomic organisation and chromosomal localisation of two members of the KCND ion channel family, KCND2 and KCND3. *Hum. Genet., 106(6), 614-619.*

Sharp P.M., Cowe E. 1991. Synonymous codon usage in *Saccharomyces cerevisiae. Yeast, , 657-678.*

Sharp P.M., Li W.H. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res., 15(3), 1281-1295.*

Sharp P.M., Tuohy T.M., Mosurski K.R. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res., 14(13), 5125-5143.*

Wang Z., Feng J., Shi H., Pond A., Nerbonne J.M., Nattel S. 1999. Potential molecular basis of different physiological properties of the transient outward K+ current in rabbit and human atrial myocytes. *Circ. Res., 84(5), 551-561.*

Watanabe K., Hoshiya S., Tokunaga K., Tanaka A., Watanabe H., Nagamatsu S., Ishida H., Takahashi S. 2002. Helicobacter pylori and acetylsalicylic acid synergistically accelerate apoptosis via Fas antigen pathway in rabbit gastric epithelial cells. *Dig. Dis. Sci., 47(4), 809-817.*

Wright F. 1990. The 'effective number of codons' used in a gene. *Gene, 87(1), 23-29*.