No Evidence for Tissue-Specific Adaptation of Synonymous Codon Usage in Humans

Marie Sémon, Jean R. Lobry, and Laurent Duret

Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), Centre National de la Recherche Scientifique, Université Claude Bernard Lyon 1, Villeurbanne, France

It has been proposed that the synonymous codon usage of human tissue-specific genes was under selective pressure to modulate the expression of proteins by codon-mediated translational control (Plotkin, J. B., H. Robins, and A. J. Levine. 2004. Tissue-specific codon usage and the expression of human genes. Proc. Natl. Acad. Sci. USA **101**:12588–12591.) To test this model, we analyzed by internal correspondence analysis the codon usage of 2,126 human tissue-specific genes expressed in 18 different tissues. We confirm that synonymous codon usage differs significantly between the tissues. However, the effect is very weak: the variability of synonymous codon usage between tissues represents only 2.3% of the total codon usage variability. Moreover, this variability is directly linked to isochore-scale (>100 kb) variability of GC-content that affect both coding and introns or intergenic regions. This demonstrates that variations of synonymous codon usage between tissue-specific genes expressed in different tissues are due to regional variations of substitution patterns and not to translational selection.

Introduction

Although synonymous codons encode the same amino acid, some are used more frequently than others. Such biases exist in most taxa and may result from neutral evolutionary processes (e.g., mutational bias or biased gene conversion) or from a selective pressure on synonymous codon positions. These two models are not mutually exclusive; indeed, codon usage can reflect a balance between selective and neutral evolutionary forces (Bulmer 1991). In some metazoan species (e.g., Drosophila and Nematode), there is evidence that synonymous codon usage is under selective pressure to optimize the efficiency of translation of highly expressed genes (the "translational selection" model) (for a review, see Duret 2002). In mammals, it has been shown that some synonymous sites are under selective pressure, possibly because of constraints acting on regulatory elements (such as splicing enhancers) located within exons (Blencowe 2000; Hurst and Pal 2001; Willie and Majewski 2004). However, up to now, there was no evidence of an impact of translational selection on synonymous codon usage in mammals (Urrutia and Hurst 2001), and codon usage was therefore generally considered to be essentially neutral (for a review, see Duret 2002).

Interestingly, Plotkin, Robins, and Levine (2004) recently reported significant differences in synonymous codon usage between genes specifically expressed in different human tissues. For example, they found that brain-specific genes can be distinguished from testis-specific genes on the basis of their codon usage. Furthermore, they showed that synonymous codon usage in brain has been conserved since human and mouse divergence. These observations are important because they suggest that there might be a selective pressure on synonymous codon usage in humans to optimize translation by adapting codon usage of tissue-specific genes to the pool of tRNAs available in each tissue (Plotkin, Robins, and Levine 2004).

Key words: codon usage, tissue specific, expression, GC-content, multivariate analysis, correspondence analysis.

E-mail: duret@biomserv.univ-lyon1.fr.

Mol. Biol. Evol. 23(3):523–529. 2006
doi:10.1093/molbev/msj053
Advance Access publication November 9, 2005

The aim of the work presented here was first to quantify the effects detected by Plotkin and colleagues, to determine the predictive value of their model. For this purpose, we measured the part of the variance in synonymous codon usage of tissue-specific genes that can be explained by their specific expression in a given tissue. Secondly, we wanted to identify which are the synonymous codons that allow the distinction between tissue-specific genes. Finally, we wanted to address a potential problem in the analyses of Plotkin and colleagues. Indeed, the method they used to measure the distance between the codon usage of two genes consists in counting the number of amino acids that exhibit a significantly different use of synonymous codons (Fisher exact test). The problem with this measure is that it is sensitive to the length and the amino acid composition of proteins simply because the statistical significance of the comparison of synonymous codon usage depends on the number of compared codons. Thus, in average, genes encoding short proteins will appear to be more similar than genes encoding long proteins. The authors mentioned in their article that the distributions of gene sizes were similar in their different data sets of tissue-specific genes. However, even for a same gene length, their measure of distance between codon usage depends on the amino acid composition of proteins (tests are more powerful for abundant than for rare amino acids). It is therefore not totally clear whether the differences in codon usage that they observed between genes specifically expressed in different tissues are due to differences in "synonymous" codon usage or in "amino acid" usage, or both.

To cope with this problem, we propose here to reevaluate the tissue specificity of synonymous codon usage using a multivariate method, internal correspondence analysis (Cazes, Chessel, and Doledec 1988; Bécue, Pagès, and Pardo 2005) which is an extension of correspondence analysis (Benzecri 1973). These analyses confirm that there is a significant difference in synonymous codon usage between tissue-specific genes expressed in different tissues. However, this effect is very weak, and the variability of synonymous codon usage between tissues is much smaller than the variability within tissues. More important, we show that the synonymous codon usage variability is only due to

GC-content differences between the genes expressed in different tissues, and this variability affects not only synonymous codon positions but also introns and intergenic regions. Hence, the tissue-specific variability of synonymous codon usage is not due to translational selection but to isochore scale variation of substitution patterns.

Materials and Methods

Genome Data

Human protein-coding sequences (CDSs) were extracted from Ensembl (release 16.3: August, 2003) (Hubbard et al. 2005). When there were several alternative splicing variants, we randomly selected one CDS per gene. The total data set contains 19,482 human CDSs.

Gene expression patterns were estimated using EST data. We selected from GenBank (Benson et al. 2005, release 133, December, 2002) 4,906,743 expressed sequence tags (ESTs) from human tissues. cDNA libraries from cell culture, tumors, pooled organs, or unidentified tissues were excluded. We only retained cDNA libraries that had been sampled enough by removing those with less than 10,000 ESTs. We retained 44 tissues corresponding to 141 libraries.

CDSs were compared to the EST data set with MEGABLAST (Zhang et al. 2000). MEGABLAST alignments showing at least 95% identity over 100 nt or more were counted as a sequence match. This criterion was chosen to be low enough to allow the detection of most ESTs despite sequencing error, but stringent enough to distinguish in most cases different members of highly conserved gene families.

A gene was considered to be tissue specific if its transcript is detected in only one of the 44 tissues. We only retained tissues for which at least 30 tissue-specific genes were available. The final data set contains 2,126 tissue-specific genes from 18 tissues. This represents 10 times more genes than in the study by Plotkin, Robins, and Levine (2004) and three times more tissues.

The regional genomic GC-content of each gene was computed on 25 kb segments extracted upstream and downstream of the gene using ACNUC software (Gouy et al. 1984). Intronic GC-content was computed for all genes for which cumulative introns length was larger than 1,000 bp ($N = 1{,}451$ genes).

Data Analysis

All computations were done using R (R Development Core Team 2003) with the seqinr (Charif and Lobry, in preparation) and ade4 (Thioulouse et al. 1997) packages. The table of observed codon frequencies has 2,126 rows, each corresponding to a CDS, and 61 columns corresponding to the 61 possible codons. Each CDS is expressed in only one tissue. This row-block structure was used to study inter- and intratissue variability. Each codon codes for only one amino acid. This defines the column-block structure used to analyze the synonymous and nonsynonymous codon usage variability. These data were analyzed by correspondence analysis (Benzecri 1973) which is a standard multivariate tool in codon usage studies (Perriere and Thioulouse 2002). The total variability was then decomposed according to internal correspondence analysis (Cazes, Chessel, and Doledec 1988; Bécue, Pagès, and Pardo 2005). To obtain the between-tissues analysis, all rows corresponding to the same tissue are summed. Then, correspondence analysis on the merged table retains only the between-tissue variability. To obtain the within-tissues analysis, all rows corresponding to the same tissue are centered at their average tissue value. Then, correspondence analysis on the resulting table retains only within-tissue variability. This decomposition applies to both row-block and column-block structure of the table. It is therefore straightforward to combine them to focus on the between-tissue synonymous codon usage variability. These analyses are additive, for instance, the variability explained by the analysis between plus within tissues equals the variability of the global analysis.

Results

Variability of Synonymous Codon Usage Between Tissues

To analyze the variability of synonymous codon usage among tissue-specific genes from different tissues, we determined the expression pattern of 19,482 human genes in 44 normal tissues for which enough EST data were available (see *Materials and Methods*). Genes for which expression was detected in only one out of these 44 tissues were considered as tissue specific. Then we retained for analyses only the tissues for which there was at least 30 tissue-specific genes. The final data set includes 18 tissues and 2,126 human tissue-specific genes. This data set is much larger than the one used by Plotkin, Robins, and Levine (2004) (6 tissues, 198 genes). We also repeated the analyses with serial analysis of gene expression (SAGE) and microarray expression data. These analyses gave the same results as EST data. Because the SAGE data set contains a smaller number of tissues and genes (12 tissues and 1,190 genes) and microarray data set is even smaller (8 tissues and 460 genes), we will only present here the results obtained with ESTs (but all other results are shown in Supplementary Material online).

We have tabulated codon usage for each gene and submitted the data set to correspondence analysis. Correspondence analysis is a multivariate method that aims to summarize data structures in high dimension space by projection onto low-dimension subspaces, while loosing as little information as possible (Benzecri 1973). Here the codon usage table consists of 2,126 rows (the tissue-specific genes) and 61 columns (the 61 codons). The rows can be split into 18 blocks corresponding to the different tissues (each gene is expressed in only one tissue) and the columns into 20 blocks corresponding to the different amino acids. Internal correspondence analysis allows one to split the total variability into between-block and within-block variability. It is therefore possible to quantify which part of the total codon usage variability is due to amino acid usage variability (between–amino acid variability), to synonymous codon usage variability (within–amino acid variability), to variability between different tissues, or to variability within tissues.

This decomposition of the global variability according to amino acids and tissues yields nine elementary analyses. The figure 1 shows the top 10 eigenvalues for each of these analyses.

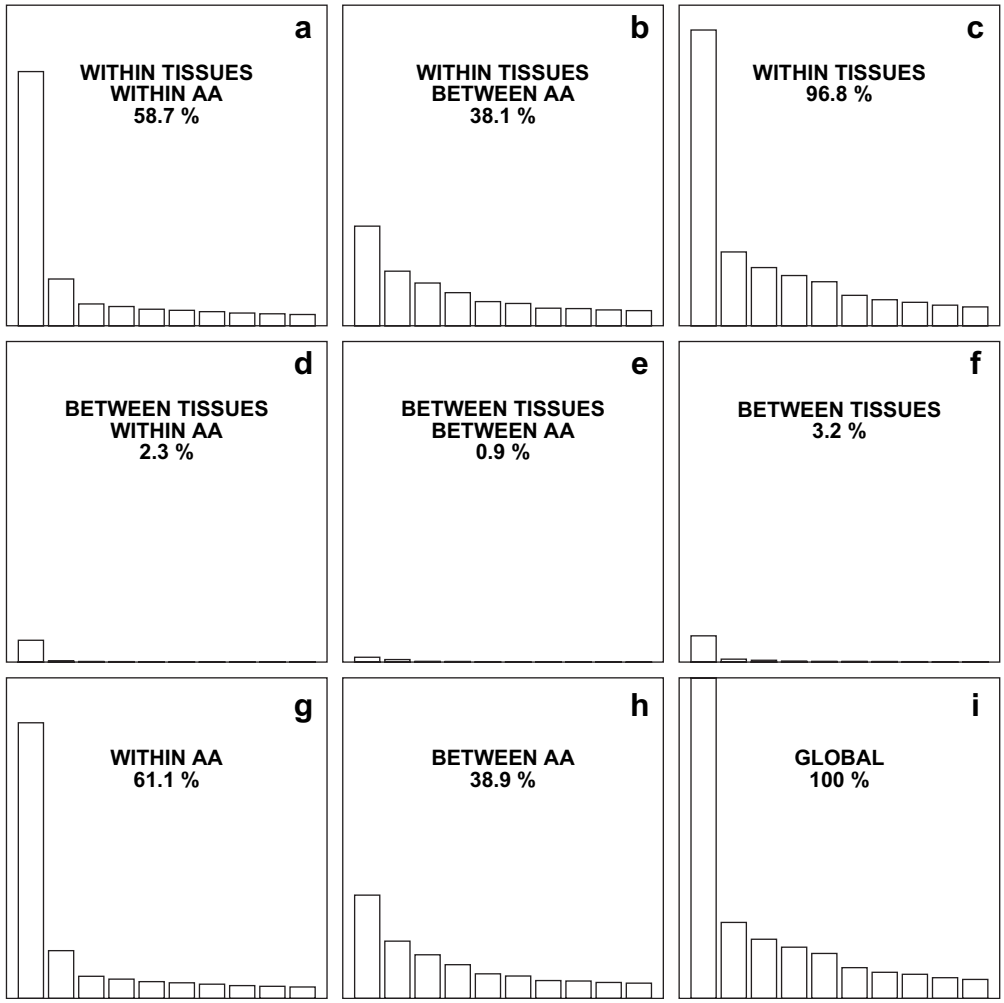

FIG. 1.—Contribution to the global codon usage variability of synonymous, nonsynonymous, between-tissues, and within-tissues effects. Eigenvalue graphs for internal correspondence analysis and associated analyses of codon usage in 2,126 human coding sequences representing tissue-specific genes from 18 tissues. The eigenvalue for a given factor is proportional to the variance in the table under analysis that is accounted for by that factor. The contribution to the total variance of a peculiar analysis is indicated. All the graphs are on the same scale (only the first 10 eigenvalues are represented) to allow a direct visual comparison. These analyses are additives: for example, the variability explained by the analysis between plus within tissues equals the variability of the global analysis. The part of global variability due to synonymous codon usage (*a*, *d*, *g*) is much more important than the part explained by nonsynonymous codon usage (*b*, *e*, *h*). The part explained by the difference of codon usage within tissues (*a*, *b*, *c*) is much more important than the part explained by the difference between tissues (*d*, *e*, *f*).

Overall, 38.9% of the total variability in codon usage is due to variability in amino acid usage (fig. 1*h*), and 61.1% is due to variability in synonymous codon usage (fig. 1*g*). The variability in synonymous codon usage between tissues appears to be a minor phenomenon as compared to other sources of variability. Indeed, the synonymous codon usage variability between tissues (fig. 1*d*) is 25 times smaller than the synonymous codon usage variability within tissues (fig. 1*a*). Overall, the variability of synonymous codon usage between tissues represents only 2.3% of the total variability in codon usage.

To test whether this part of the variance is higher than expected, we permuted randomly the association between genes and tissues and repeated the correspondence analysis. After 1,000 independent permutations, we obtained a distribution of the expected variability. The observed value (2.3%) is greater than expected (P value $< 10^{-16}$). In other words, tissue-specific genes have a significantly different synonymous codon usage depending on the tissue where they are expressed.

To check that the association between synonymous codon usage and tissues was not the result of a few outliers, we have resampled genes with replacement. After 1,000 independent resamplings, we obtained a distribution of the expected variability. The borders of a 95% confidence interval of the observed value are 2.2% and 3.6 %. This suggests that the observed difference of synonymous codon usage between tissues is a general trend in the data set.

Note that the model of translational selection predicts that biases in synonymous codon usage should be stronger for highly expressed genes. To test whether highly expressed tissue-specific genes from different tissues show specific codon usage biases, we repeated the analyses on a subset of the tissue-specific genes obtained after selecting

the 30% most expressed genes in each tissues ($N = 510$ genes). Again, we found that the variability of synonymous codon usage between tissues accounts for only a very small fraction (4.3%) of the total variability in codon usage.

It should also be noticed that the identification of tissue-specific genes is not straightforward because it depends on the sensitivity of the method used to detect expression. For example, genes that are expressed in many tissues but at low levels might be classified as "tissue specific" simply because the probability of detection in any given tissue is low. Conversely, genes that are expressed at especially high levels in a given tissue but also at low levels in a number of other tissues are not considered as tissue specific according to the definition we used. To test whether this problem could affect our results, we examined an alternative definition of tissue specificity: we considered as tissue specific all genes having 10-fold higher EST counts in a given tissue relative to the sum of counts in all other tissues. This defined a data set of 2,426 genes, expressed in one out of 18 tissues (we only retained the tissues containing at least 30 genes). After running the correspondence analysis on this new data set, we found that 0.5% of the global variability is due to the variation of synonymous codon usage between tissues.

In summary, these analyses confirm that there is a significant difference in synonymous codon usage between tissue-specific genes from different tissues (Plotkin, Robins, and Levine 2004), but show that this difference accounts for a very small fraction of total variability.

Tissue-Specific Synonymous Codons and GC-Content

To understand the origin of this effect (weak but significant), we analyzed the features of codons that are responsible for the tissue specificity of synonymous codon usage. For this purpose we looked at the first axis of the analysis of within–amino acids and between-tissues variability in codon usage (see fig. 1*d*). This first axis accounts for 82% of the variability of synonymous codon usage between tissues and can therefore be used as a summary. The first axis is very strongly correlated with GC-content at third codon positions (GC3) (Spearman's correlation coefficient $R^2 = 99\%$, 2,126 genes, 18 tissues; see fig. 2). The observed differences of synonymous codon usage between tissues are therefore almost entirely due to weak differences of GC-content between tissue-specific genes expressed in different tissues.

Note that the total variability of synonymous codon usage is also very strongly correlated with GC3-content (the first axis of this analysis accounts for 38.9% of the total variability in synonymous codon usage, fig. 1*g*, and correlates with GC3 with a Spearman's correlation coefficient $R^2 = 98\%$, 2,126 genes). This phenomenon is not restricted to tissue-specific genes. Indeed, the within–amino acid correspondence analysis of codon usage of all human genes (19,482 genes) shows that the first axis (that accounts for 37.8% of the total variability in synonymous codon usage) is very strongly correlated with GC3-content ($R^2 = 90.3\%$). This demonstrates that variability of synonymous codon usage in human genes is mainly due to variability in GC-content.

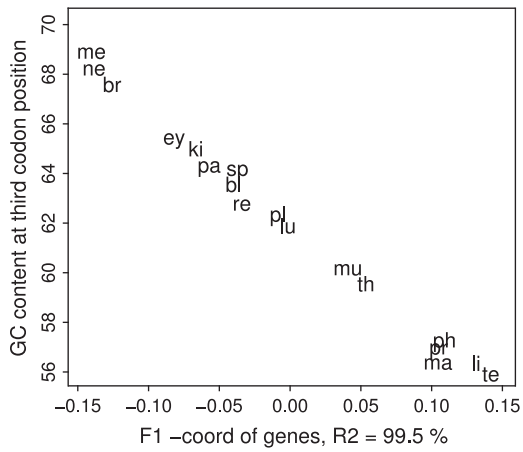

FIG. 2.—Average GC-content at the third position of coding sequences and the positions on the first axis of the factorial map (see fig. 1*d*) for synonymous codon usage variability between tissues (their names are indicated). The correlation is almost perfect ($R^2 = 99\%$). The most important factor (82.5% of total variability for this analysis) for synonymous codon usage between tissues is therefore the difference of average GC3-content of genes between tissues. Initials of the names of tissues are indicated: "bl": blood, "br": brain, "ey": eye, "ki": kidney, "li": liver, "lu": lung, "ma": mammary gland, "me": medulla oblongata, "mu": muscle, "ne": nerve, "pa": pancreas, "ph": pharynx, "pl": placenta, "pr": prostate, "re": retina, "sp": spleen, "te": testis, "th": thalamus.

Discussion

We analyzed the variability in synonymous codon usage of human tissue-specific genes. In agreement with Plotkin, Robins, and Levine (2004), our analyses show that there is a significant difference in synonymous codon usage between tissue-specific genes from different tissues. However, in contradiction with Plotkin, Robins, and Levine (2004), we found that this tissue-specific variability in synonymous codon usage is almost entirely due to variations in GC-content. The fact that the GC3-content of tissue-specific genes varies according to the tissue had already been noticed (Vinogradov 2003*a*). This effect is however very weak: the variability of synonymous codon usage between tissues accounts for only 2.3% of the total variability in codon usage.

As shown in figure 3, within a given tissue, there is a huge variability of GC3-content, much larger than the variability between tissues. The distributions of GC3 are largely overlapping, even for the most different tissues (e.g., testis and brain, see fig. 3). In other words, the relationship between synonymous codon usage and tissue specificity, although statistically significant, has a very low predictive value: it is impossible to predict with reliability the tissue in which a gene is expressed from its synonymous codon usage.

In their article, Plotkin, Robins, and Levine claim that the difference in codon usage between tissues cannot be simply explained by differences in GC3-content. However, they did not indicate the nature of those synonymous codons that differ between tissues, and hence, it is difficult to directly verify their assertion. They showed that their measure of codon usage can separate genes by tissue types, much better than GC3-content. However, as mentioned in

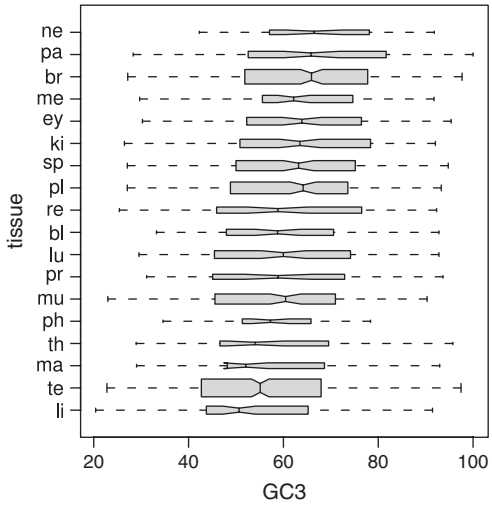

FIG. 3.—Relationship between GC-content at third position for genes specifically expressed in 18 tissues. For each tissue, the distribution of GC-content is represented by a boxplot: the lower, middle, and top horizontal lines of the boxes represent the 25%, 50%, and 75% quantiles, respectively. The notches in the boxes represent a confidence interval ($\alpha = 5\%$) for the median. The whiskers represent extreme values (within 1.5 times the interquartile range from the upper or lower quartile). Tissues are sorted according to the average GC-content of the corresponding genes. The GC3-content is different between the tissues (because of the large number of points in the data set, the analysis of variance is extremely efficient at rejecting the null hypothesis yielding a very small P value of 10^{-14}), but this should be put in the context of the huge variability of GC-content within tissues, depicted here by the vertical size of the boxes. Initials of the names of tissues are indicated: "li": liver, "te": testis, "ma": mammary gland, "th": thalamus, "ph": pharynx, "mu": muscle, "pr": prostate, "lu": lung, "bl": blood, "re": retina, "pl": placenta, "sp": spleen, "ki": kidney, "ey": eye, "me": medulla oblongata, "br": brain, "pa": pancreas, "ne": nerve.

the *Introduction*, the measure of codon usage they used depends not only on synonymous codon usage but also on gene length and amino acid usage. It is therefore possible that the discrimination they obtained was not due to synonymous codon usage but to other gene features.

The reason for this weak tissue-specific variation of GC3-content is not known. What is clear is that these variations of GC-content affect not only the third codon positions but also intergenic regions and introns. Indeed, the first axis of the analysis of between-tissues variability of synonymous codon usage (fig. 1*d*) strongly correlates to intronic GC-content ($R^2 = 61\%, N = 1{,}451$ genes, 18 tissues) and to local intergenic genomic GC-content ($R^2 = 63\%$, $N = 2{,}126$ genes, 18 tissues, intergenic GC-content computed on 25 kb segments extracted upstream and downstream of the gene). Indeed, the intronic GC-content of tissue-specific genes also varies significantly according to the tissues (fig. 7 in Supplementary Material online, analysis of variance, $P < 10^{-16}$). If the tissue-specific variation of synonymous codon usage was due to a selective pressure to optimize translation, then we would have expected this phenomenon to affect only exons and not introns or intergenic regions. Our observations therefore indicate that this tissue-specific variability of synonymous codon usage is linked to regional variations in genomic base composition and not to translational selection.

In mammals, it is well known that the variability in synonymous codon usage is strongly linked to regional variations in genomic GC-content (for review, see Duret 2002). Indeed, mammalian genomes are characterized by wide-scale (>100 kb) variations in GC-content that affect both coding and noncoding regions (the so-called isochores) (Bernardi 1989; Eyre-Walker and Hurst 2001). This phenomenom is demonstrated by the existence of very strong correlations between the GC-content in introns, intergenic regions, and third codon positions (Mouchiroud 1986; Bernardi 1989, 1993; Duret and Hurst 2001). Thus, the evolution of the synonymous codon usage of a given gene essentially depends on the genomic environment in which this gene is located. Indeed, it has been shown that in mammals the pattern of silent substitution vary according to the GC-content of isochores (Duret et al. 2002; Smith, Webster, and Ellegren 2002). Notably, the pattern of substitution is more biased toward AT in GC-poor isochores than in GC-rich isochores (Webster, Smith, and Ellegren 2003; Meunier and Duret 2004). In other words, the evolution of the base composition of isochores is more conservative than if subsitutions were randomly distributed throughout the genome. Plotkin, Robins, and Levine (2004) observed that patterns of synonymous substitutions in human and mouse orthologs tended to preserve tissue-specific codon usage and concluded that this was evidence for selection. However, this observation can be simply explained by the fact that orthologous genes are subject to similar patterns of silent substitutions because they are located in similar genomic environments.

The base composition of mammalian genes is also affected by transcription-associated mutational biases (TAMB) (Duret 2002; Green et al. 2003; Majewski 2003; Touchon et al. 2003, Comeron 2004). Note that such a process can only affect genes that are transcribed in the germ line. Hence, the impact of TAMB on the base composition of a given gene only depends on its level of expression in the germ line. Interestingly, Comeron (2004) has shown that genes expressed in different somatic tissues show diverse degrees of TAMB. Such variations of degrees of TAMB among tissues probably reflect variations in the proportion of genes expressed in a given somatic tissue that are also expressed in the germ line (i.e., the impact of TAMB can be detected in tissues that have a pattern of expression similar to that of the germ line). This phenomenom can partly explain the variation in synonymous codon usage observed among tissue-specific genes expressed in different tissues. However, this process can only affect the base composition of transcription units (introns and exons) but not of untranscribed intergenic regions. Thus, this process of TAMB cannot explain the strong correlation observed between the first axis of the between-tissues variability of synonymous codon usage and the GC-content of intergenic regions flanking genes (see above).

Another possible hypothesis might be proposed to explain tissue-specific variation of GC3-content. Indeed, it has been shown that there is a significant clustering of tissue-specific coexpressed genes along the human genome (Lercher, Urrutia, and Hurst 2002). Thus, tissue-specific genes expressed in a same tissue might have a similar GC-content simply because they are neighbor along

chromosomes and therefore are located in the same isochores. However, in our data set, we did not find any significant clustering of tissue-specific genes along human chromosomes (data not shown). This is in agreement with Lercher and colleagues who found that the clustering of coexpressed genes was essentially due to housekeeping genes (Lercher, Urrutia, and Hurst 2002). Hence, tissue-specific variation of GC3-content cannot be explained by the clustering of tissue-specific genes along chromosomes.

It has also been proposed that regional variations in GC-content might reflect some selective constraints on DNA stability or bendability (Bernardi 1993; Vinogradov 2003*b*). However, there is a priori no reason why genes expressed in different tissues should require different DNA stability or bendability. Moreover, it is difficult to imagine a selective pressure strong enough to be able to detect the change in GC-content induced by a single base mutation in megabase–long genomic sequences.

Our analyses show that variations in synonymous codon usage between tissues are not due to translational selection. However, this does not prove that translational selection does not have any impact on synonymous codon usage in humans. This question has been highly debated for many years. Up to recently, most studies had failed to demonstrate an impact of translational selection in mammals (Sharp et al. 1995; Iida and Akashi 2000; Urrutia and Hurst 2001; Lander et al. 2001, Duret 2002, dos Reis, Savva, and Wernisch 2004). However, Urrutia and Hurst (2003) reported a weak but significant correlation between codon bias (corrected for regional variations in base composition) and expression level of human genes. Moreover, after taking into account the effect of isochores and of TAMB, Comeron (2004) found evidence of translational selection in humans. It should be noted, however, that these effects are very weak (e.g., the correlation coefficient, R^2, reported by Urrutia and Hurst [2003] is less than 1.5%). Given the huge amount of data presently available, it is now possible to demonstrate the statistical significance of such weak effects. However, we would like to stress that the real issue is to discuss the quantitative importance of the effect and not simply its statistical significance. In other words, there is evidence for translational selection in humans (Urrutia and Hurst 2003; Comeron 2004), but this factor explains only an extremely small fraction of the variability of synonymous codon usage. It is especially important to mention this point because researchers who read all these articles are not necessarily experts in statistics and may therefore keep in mind only the positive result (the *P* value is highly significant, therefore "there is translational selection in human"), without any clear idea of the intensity of this effect.

In conclusion, we confirm the existence of significant differences in synonymous codon usage between tissue-specific genes expressed in different tissues. However, this effect is linked to regional variations of GC-content, and not to a selective pressure to optimize translation, as suggested by Plotkin, Robins, and Levine (2004). The reason for these tissue-specific variations of genomic GC-content remains unclear, but it should be stressed that whatever the correct explanation is (neutralist or selectionist), the effect is very small, and hence is of an extremely poor predictive value.

Supplementary Material

To allow the reproducibility of our results, all our analyses can be run again online at the following URL: http://pbil.univ-lyon1.fr/datasets/SemonLobryDuret2005/. The data set is also available at the same address. Figure 7 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Centre National de la Recherche Scientifique. We warmly thank Hiroshi Akashi and two anonymous referees for their helpful comments.

Literature Cited

Bécue, M., J. Pagès, and C. E. Pardo. 2005. Contingency table with a double partition on rows and colums. Visualization and comparison of the partial and global structures. Pp. 355–364 *in* J. Janssen and P. Lenca, eds. Applied stochastic models and data analysis. ENST Bretagne, Brest.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2005. GenBank. Nucleic Acids Res. **33**(Database Issue):D34–D38.

Benzecri, J. P. 1973. L'analyse des correspondances. Bordas, Paris.

Bernardi, G. 1989. The isochore organization of the human genome. Annu. Rev. Genet. **23**:637–661.

———. 1993. The vertebrate genome: isochores and evolution. Mol. Biol. Evol. **10**:186–204.

Blencowe, B. J. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem. Sci. **25**:106–110.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics **129**:897–907.

Cazes, P., D. Chessel, and S. Doledec. 1988. L'analyse des correspondances internes d'un tableau partitionné: son usage en hydrobiologie. Rev. Stat. Appl. **36**:39–54.

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.

dos Reis, M., R. Savva, and L. Wernisch. 2004. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. **32**:5036–5044.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12**:640–669.

Duret, L., and L. D. Hurst. 2001. The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Mol. Biol. Evol. **18**:757–762.

Duret, L., M. Semon, G. Piganeau, D. Mouchiroud, and N. Galtier. 2002. Vanishing GC-rich isochores in mammalian genomes. Genetics **162**:1837–1847.

Eyre-Walker, A., and L. D. Hurst. 2001. The evolution of isochores. Nat. Rev. Genet. **2**:549–555.

Gouy, M., F. Milleret, C. Mugnier, M. Jacobzone, and C. Gautier. 1984. ACNUC: a nucleic acid sequence data base and analysis system. Nucleic Acids Res. **12**:121–127.

Green, P., B. Ewing, W. Miller, P. J. Thomas, and E. D. Green. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. **33**:514–517.

Hubbard, T., D. Andrews, M. Caccamo et al. (52 co-authors). 2005. Ensembl 2005. Nucleic Acids Res. **33**(Database Issue):D447–D453.

Hurst, L. D., and C. Pal. 2001. Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet. **17**:62–65.

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

Lander E. S., L. M. Linton, B. Birren et al. (256 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Lercher, M. J., A. O. Urrutia, and L. D. Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. **31**:180–183.

Majewski, J. 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. Am. J. Hum. Genet. **73**:688–692.

Meunier, J., and L. Duret. 2004. Recombination drives the evolution of GC-content in the human genome. Mol. Biol. Evol. **21**:984–990.

Mouchiroud, D. 1986. Relationship between base composition in non-coding DNA of genes and codon composition. C. R. Acad. Sci. III **303**:743–748.

Perriere, G., and J. Thioulouse. 2002. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res. **30**:4548–4555.

Plotkin, J. B., H. Robins, and A. J. Levine. 2004. Tissue-specific codon usage and the expression of human genes. Proc. Natl. Acad. Sci. USA **101**:12588–12591.

R Development Core Team. 2003. R: a language and environment for statistical computing. Vienna, Austria.

Sharp, P. M., M. Averof, A. T. Lloyd, G. Matassi, and J. F. Peden. 1995. DNA sequence evolution: the sounds of silence. Phil. Trans. R. Soc. Lond. B **349**:241–247.

Smith, N. G., M. T. Webster, and H. Ellegren. 2002. Deterministic mutation rate variation in the human genome. Genome Res. **12**:1350–1356.

Thioulouse, J., D. Chessel, S. Doledec, and J. M. Olivier. 1997. ADE-4: a multivariate analysis and graphical display software. Stat. Comput. **7**:75–83.

Touchon, M., S. Nicolay, A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes. 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. FEBS Lett. **555**: 579–582.

Urrutia, A. O., and L. D. Hurst. 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**:1191–1199.

———. 2003. The signature of selection mediated by expression on human genes. Genome Res. **13**:2260–2264.

Vinogradov, A. E. 2003*a*. Isochores and tissue-specificity. Nucleic Acids Res. **31**:5212–5220.

———. 2003*b*. DNA helix: the importance of being GC-rich. Nucleic Acids Res. **31**:1838–1844.

Webster, M. T., N. G. Smith, and H. Ellegren. 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. Mol. Biol. Evol. **20**:278–286.

Willie, E., and J. Majewski. 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet. **20**: 534–538.

Zhang, Z., S. Schwartz, L. Wagner, and W. Miller. 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. **7**:203–214.

Dan Graur, Associate Editor

Accepted November 2, 2005