ELSEVIER

Gene 300 (2002) 161–168

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Isochores, GC_3 and mutation biases in the human genome

Fernando Alvarez-Valin[a], Guillermo Lamolle[a], Giorgio Bernardi[b,*]

[a]Seccion Biomatematica, Facultad de Ciencias, Montevideo, Uruguay
[b]Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Naples, Italy

Received 12 June 2002; received in revised form 4 July 2002; accepted 18 September 2002
Received by T. Gojobori

Abstract

In this work we re-examined the hypothesis that the variation in GC content in the human genome is due to different regional mutational biases. For this purpose we inferred the mutational pattern by using mutation databases that are available for many genes associated with human genetic diseases. The assumption of this approach is that such mutations reflect the actual frequency distribution of mutations as they arise in the population. Four classes of genes, classified according to their GC_3 level, were included in this study: GC_3-poor genes ($GC_3 < 45\%$), genes with intermediate GC_3 content ($45\% < GC_3 < 60\%$), GC_3-rich genes ($60\% < GC_3 < 75\%$) and very GC_3-rich genes ($GC_3 > 75\%$). Our results show that most genes are under AT mutational biases, with very little variation compared to the expectations of neutral GC level. It is noteworthy that the mutational patterns in the GC_3-rich genes do not appear to account for their GC_3-richness. Instead, GC_3-rich and very GC_3-rich genes exhibit patterns of mutations that yield expectations of neutral GC_3 content that are much lower than their actual GC_3. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Compositional evolution; Genomics

1. Introduction

The genomes of warm-blooded vertebrates cover a very broad compositional spectrum (30–60% GC, neglecting satellite DNAs). Contrary to a view that was predominant until the early 1970s, the compositional heterogeneity of these genomes is not continuous but discontinuous (Filipski et al., 1973). Indeed, these genomes are made up, at fragment sizes of 50–100 kb, of a small number of families of DNA molecules which are comparable in heterogeneity to the least heterogeneous natural DNAs, namely bacterial DNAs (Filipski et al., 1973; Cuny et al., 1981). In turn, these families of DNA molecules derive, by degradation during DNA preparation, from much longer chromosomal regions (originally estimated as larger than 300 kb; Macaya et al., 1976) that were called isochores (for ‘similar regions’). In other words, the genomes of warm-blooded vertebrates are mosaics of isochores. In the human genome, isochores belong in two GC-poor families, L1 and L2, and three GC-rich families, H1, H2 and H3. These families represent about 33%, 30%, 24% 7.5% and 4–5%, respectively, of the genome and are characterized by increasing gene densities.

An important point is that protein-coding sequences, which only represent 2–3% of the mammalian and avian genomes, are compositionally correlated with the isochores in which they are located (Bernardi et al., 1985). Corresponding correlations also hold for the GC_3 values (i.e. the average GC levels of the third codon position of genes), as well as for the exons and introns of the same genes. It should be kept in mind that the variability in GC_3 content is much larger than that of isochores, ranging from 20% to almost 100%.

In contrast with the genomes of warm-blooded vertebrates, the genomes of cold-blooded vertebrates are characterized by a much lower GC level heterogeneity, GC-rich isochores being almost absent. Two exceptions to this general rule are some fish genomes (like those of *Tetraodontiformes*) which are GC-rich but compositionally homogeneous and some reptilian genomes (like crocodile and turtle) which show a certain degree of heterogeneity but lack typical features of the genomes of warm-blooded vertebrates, like CpG islands (see Bernardi, 2000 for a recent review).

Abbreviations: Kb, kilobase; GC, molar ratio of guanine + cytosine in DNA GC, GC of third codon positions; CpG, dinucleotide composed by a C followed by a G.

* Corresponding author. Fax: +39-081-2455807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

PII: S0378-1119(02)01043-0

Table 1
Mutation matrices for the phenylalanine hydroxylase locus (phenylketonuria)

Matrix of observed mutations **M**[a]					
From/to	T	C	A	G	
T	224	37	8	13	
C	49	202	16	10	
A	13	10	277	30	
G	19	19	38	199	
Probabilistic (normalized) matrix **P**, obtained from **M**[b]					
From/to	T	C	A	G	
T	0.7411	0.1652	0.0357	0.0580	
C	0.2426	0.6287	0.0792	0.0495	
A	0.0469	0.0361	0.8087	0.1083	
G	0.0955	0.0955	0.1910	0.6181	
Steady state in base frequencies[c]	T	C	A	G	G + C
	0.3151	0.2122	0.3095	0.1632	0.3754

[a] The elements in the diagonal of **M** (m_{ii}) represent the number of times each base appears in the wild-type sequence (i.e. the number of available targets for mutation).

[b] Each off-diagonal entry in this matrix (p_{ij})is obtained by dividing m_{ij}/m_{ii}, while the elements in the diagonal of **P** are: $p_{ii} = 1 - \sum_j p_{ij}, j \neq i$.

[c] This corresponds to the normalized version of vector **x**, which satisfies the following equality expressing equilibrium: $\mathbf{xP} - \mathbf{x} = 0$. Note that **x** is an eigenvector of $\mathbf{P}^t$, associated with the eigenvalue whose numeric value is equal to 1.

The first explanation which was proposed for the origin and the maintenance of the compositional spectra (at the DNA and coding sequence levels) of warm-blooded vertebrates was selection, the selective advantages being associated with the increased thermodynamic stability of both proteins and DNA that were required by the higher body temperature of warm-blooded relative to cold-blooded vertebrates (Bernardi and Bernardi, 1986). Indeed, the increased GC of coding sequences was accompanied by an increase in the frequencies of hydrophobic amino acids which in turn implies a higher thermodynamic stability of proteins.

An alternative explanation was that the compositional changes under discussion had no functional significance and were due to regional variation in the underlying mutation patterns (Sueoka, 1988, 1992). According to this point of view, genes located in GC-rich isochores should exhibit GC mutation bias while AT mutation bias should be in action in GC-poor isochores. Several mechanisms have been postulated for explaining the variation in mutational biases, such as regional differences in DNA repair (Filipski, 1987), and/or differential misincorporation of nucleotides during DNA replication that was postulated to be due, in turn, to a differential depletion of precursor nucleotides during the cell cycle (Wolfe et al., 1989).

We will not reiterate here the numerous and diverse arguments that were raised against these alternative explanations (see Bernardi, 2000 for a review). Instead, we will report novel results on the mutational pattern of the human genome. We have inferred the mutational pattern by using mutation databases that are available for many genes associated with human genetic diseases. The assumption of this approach is that the frequency distribution of these mutations reflects the actual frequency distribution of mutations as they arise in the population. Note that these mutations were not yet subjected to the sieving effect of selection, in contrast to the pattern of nucleotide substitutions, which depends upon two processes, mutation and fixation, the latter being influenced by selection. Inferring the mutational pattern from the observed pattern of mutations responsible for genetic alterations is not new. In effect, several previous investigators have used this kind of information for analyzing the mutational dynamics of human genes (e.g. Cariello and Skopek, 1993; see also Krawczak and Cooper, 1996, and references therein).

2. Materials and methods

Using the observed number of mutations we constructed (for each gene analyzed) a 4 × 4 matrix of counted mutations, where each row in the matrix corresponds to the observed number of mutations from one base to the other ones (Table 1). From this matrix it is possible to obtain a stochastic (probabilistic) matrix by normalizing by rows, from which the expected GC content at equilibrium for a sequence that is evolving only under the effect of mutation and genetic drift (i.e. the neutral GC level) can be readily obtained.

There are some drawbacks of this method to infer the mutational spectrum that should be taken into account. Perhaps the most important one is that mutation databases do not represent random samples of all arising mutations. Rather, the vast majority of them are middle to highly deleterious mutations, whereas advantageous, neutral or slightly deleterious mutations are hardly or not at all represented. This is because for a mutation to be represented in a database, it should come to clinical attention (due to the

Table 2
Mutation databases used in this study

	GC$_3$	Number of mutations
1. GC$_3$-poor genes ($GC_3 < 0.45$)		
1.1. Haemophilia B factor (factor IX)	0.338	1474
http://www.umds.ac.uk/molgen/haemBdatabase.htm		
1.2. Haemophilia A factor (factor VIII)	0.388	428
http://europium.csc.mrc.ac.uk/usr/WWW/WebPages/main.dir/main.htm		
1.3. Ataxia telangiectasia	0.322	126
http://www.vmresearch.org/atm.htm		
1.4. Cystic fibrosis[a]	0.3975	509
http://www.genet.sickkids.on.ca/cftr/		
1.5. HPRT (Lesch–Nyhan syndrome)	0.3791	168
http://www.ibiblio.org/dnam/mainpage.html		
2. Genes with intermediate GC$_3$ content ($0.45 < GC_3 < 0.6$)		
2.1. Phenylalanine hydroxylase locus (Phenylketonuria)[a]	0.519	262
http://ww2.mcgill.ca/pahdb/		
2.2. PHEX X-linked hypophosphatemia	0.457	64
http://data.mch.mcgill.ca/phexdb/		
2.3. PAX6 Developmental eye anomalies	0.5174	80
http://www.hgu.mrc.ac.uk/Softdata/PAX6/		
3. GC$_3$-rich genes ($0.6 \leq GC_3 < 0.75$)		
3.1. Androgen receptor[b]	0.64	314
http://ww2.mcgill.ca/androgendb/		
3.2. P53 gene[c]	0.61	141
http://www.iarc.fr/p53/		
3.3. Wilson disease	0.6057	125
http://www.medgen.med.ualberta.ca/database.html		
4. Very GC$_3$-rich genes ($GC_3 > 0.75$)		
4.1. Glucose-6-phosphate dehydrogenase (Favism)	0.84	118
http://rialto.com/favism/mutat.htm		
4.2. L1CAM, L1 cell adhesion molecule	0.77	65
Van Camp et al. (1996)		
4.3. Haemophilia factor VII	0.7982	148
http://europium.csc.mrc.ac.uk/usr/WWW/WebPages/FVII/database.dir/index.htm		
4.4. LDLR locus (familial hypercholesterolaemia)	0.76	359
http://www.ucl.ac.uk/fh/		
Total		4381

[a] Information on repetitions is not available.

[b] Mutations falling in the first exon were excluded from the analysis because while this exon represents more than 60% of the coding region, its mutations are very scarce (less than 10%) and the amino acid composition of the exon is extremely biased.

[c] Only mutations from exons 5 to 8 were analyzed (they contain more than 90% of all point mutations).

deleterious phenotypes it produces), otherwise it will remain almost undetectable. This problem is particularly serious at the third positions of codons since about two-thirds of mutations that can occur in this position are synonymous and hence have a very low probability of being detected. The remaining third, in turn, generally implies changes between biochemically similar amino acids. Even though some of these mutations between biochemically similar amino acids are indeed detected, again they have lower probability of detection when compared with those mutations affecting the first and second codon positions. This causes two kinds of problems in analyses that include the third position, namely a low representation (a low percentage of mutations are detected) and a biased representation of the position. The latter is due to the fact that most transitions (purine–purine or pyrimidine–pyrimidine mutations) at this position are synonymous, while amino acid altering mutations are almost exclusively transversions (with the exception of those involving Met and Trp codons). As a consequence, the inclusion of third codon positions would lead to an overestimation of transversions. Because of these complications, we decided not to include in our analysis mutations affecting third positions. It is worth mentioning that although the substitution pattern can be very different among the three codon positions, there is no reason to think that these differences could be attributed to differential mutation patterns. Rather, the differences are due to fixation biases

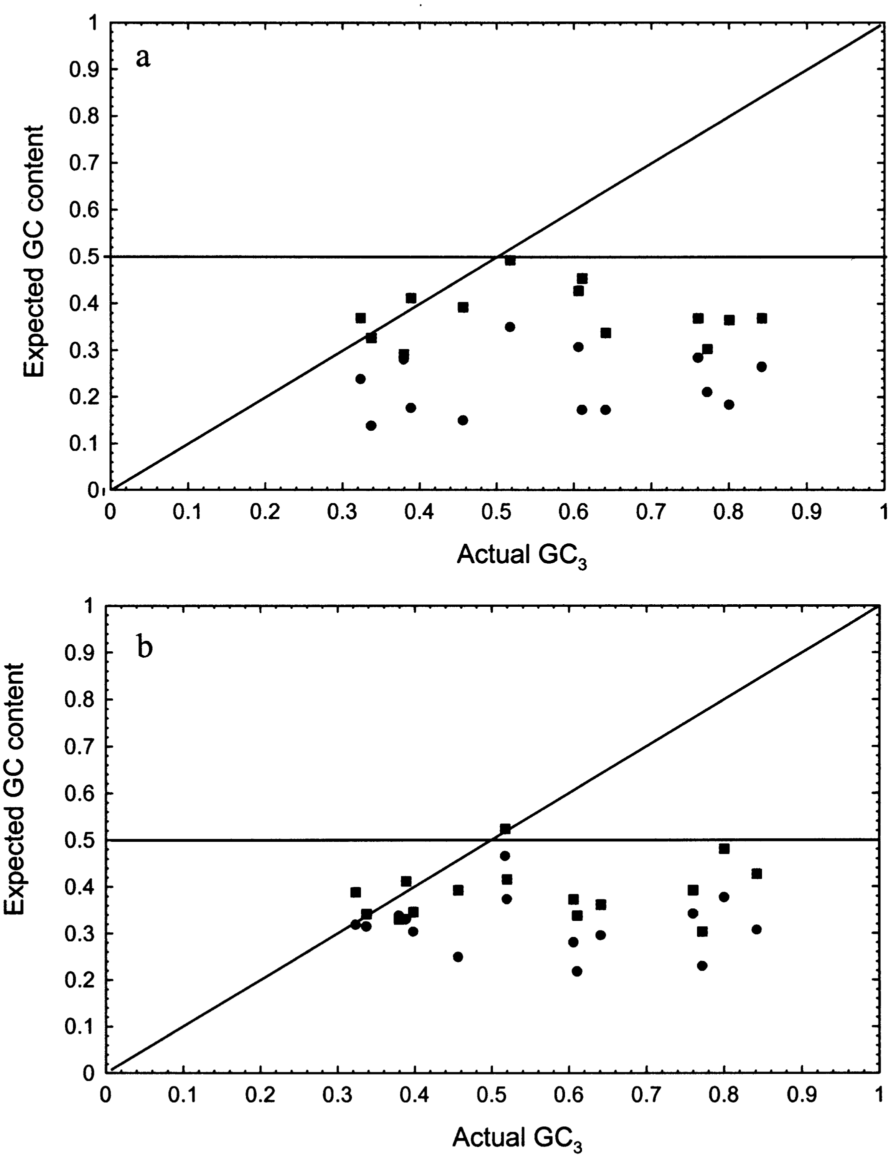

Fig. 1. The actual GC_3 of each gene was plotted against the GC level that would be expected if it depended solely on the mutational pattern. There are four different expectations for each gene. (a) GC expectations that were obtained when repeated mutations were considered. The different symbols indicate whether CpG sites were included (blue circles) or excluded (red squares) in the analysis. (b) GC expectations when repeated mutations were not considered, with symbols as in (a). For cystic fibrosis and phenylketonuria databases information on repeated mutations was not available. The diagonal line shows the hypothetical place where points should be located if the GC_3 levels were determined by the mutational input, so that the vertical distance between each point and this line indicates the distance of the actual GC_3 from the mutational equilibrium. The horizontal line at $G + C = 0.5$ indicates the border line between AT bias and GC bias.

(i.e. selection) owing to the different degrees of functional constraints affecting the different codon positions. As far as the first and second positions of codons are concerned, we would like to mention that with a few exceptions, the mutations affecting them produce amino acids changes, and thus they are potentially detectable.

The second complication that should be dealt with concerns the problem of repeated mutations (i.e. two different persons have the same mutation at the same nucleotide position). There are two types of repeated mutations, recurrent mutations and mutations that are identical by descent. Needless to say, the first type of repetition should be taken into consideration since it represents independent mutational events, while the second type of repetition should be counted as a single mutation. However, it is not always possible to distinguish them. For this reason we have used two different types of mutation matrices, those where repetitions were considered, yet excluding the repetitions that are known to be identical by descent, and matrices where repetitions were not included.

The last complication that should be considered is the hypermutability of the dinucleotide CpG. Indeed, methylated CpG is known to be a hotspot for mutation due to the high deamination rate of methyl-C giving rise to T residues.

Table 3
Expected neutral base composition obtained from three different mutation databases from P53 encoding gene

(a) Including CpG sites	
Database	G + C
Somatic mutations (codon positions 1 and 2)	0.2192
Germline mutations (codon positions 1 and 2)	0.1947
Somatic synonymous mutations	0.3752
(b) Excluding CpG sites	
Database	G + C
Somatic mutations (codon positions 1 and 2)	0.3540
Germline mutations (codon positions 1 and 2)	0.4310
Somatic synonymous mutations	0.3804

In both (a) and (b) the estimates were obtained considering repetitions since for somatic mutations the problem of identity by descent does not exist (all mutations correspond to independent mutational events).

As a consequence, very often we find a disproportionately high number of mutations affecting CpG sites. In spite of their extremely high mutation rate, many CpGs are kept in coding sequences very likely because they participate in encoding crucial amino acids. This could lead to an overestimation not only of the relative mutation rate at CpG sites, but most importantly of their long-term contribution to GC content at equilibrium. To account for this potential distortion, we used both mutation matrices where CpG mutations were taken into consideration and matrices that did not include the effect of CpG dinucleotides.

3. Results

To investigate whether the pattern of mutations differs among genes having different GC_3 levels we analyzed the mutational spectrum of GC_3-poor genes ($GC_3 < 45\%$), genes with intermediate GC_3 content ($45\% < GC_3 < 60\%$), GC_3-rich genes ($60\% < GC_3 < 75\%$) and very GC_3-rich genes ($GC_3 > 75\%$) (see Table 2).

The results presented in Fig. 1 show that although there is some variation in the expected GC level of a neutral sequence evolving only under the effect of mutations, the results contradict what would be expected according to the regional mutation bias hypothesis. In the first place, the results show that independently of the mutation matrix used (whether or not CpGs and/or repetitions are included), the variations of the expected GC content for sequences evolving only under the effect of mutations are always in the region of AT biases (with the exception of one of the results obtained from the PAX6 database). Secondly, there is much less variation than one would expect according to the regional mutation bias hypothesis. Finally, and most important, no correlation can be observed between the expected GC content for a neutral sequence and the actual GC_3 of genes. In fact the highest figures for expected GC level are exhibited by genes that have intermediate GC_3 levels.

It can be argued that the relative homogeneity in the mutation patterns that we show here could be due in part to detection biases, namely that GC → AT mutations are more readily detected, thus causing biased estimations. Although this possibility cannot be completely ruled out, it is very unlikely that all mutation databases analyzed in this work exhibit the same kind of detection bias. Moreover, there are several lines of evidence that suggest that this is not the case. In the first place, in vitro analysis of mutation frequencies (considering both misincorporations and proofreading corrections) in mammal DNA polymerases alpha and beta shows that these enzymes produce a considerably higher number of GC → AT mutations than AT → GC mutations (Kunkel and Alexander, 1986).

Secondly, for the gene encoding the P53 protein, we can obtain three independent estimates of the mutational pattern. Apart from the database containing germline mutations (already listed in Table 1), there is a database of somatic deleterious mutations and a database of synonymous non-deleterious mutations. The latter clearly represents a completely random sample of mutations, since they were isolated not because of their phenotypic effect but simply by chance. Interestingly, the expectations of neutral GC level obtained from the three P53 data sets are very similar (see Table 3). The variability in the expected GC_3 is only 8% among three independent data sets for the same gene when CpG sites are not considered. The variability is indeed larger (but again in the AT bias zone) when CpG sites are taken into account. In this case the expected GC content ranges from 0.195 to 0.375. This almost 20% of GC variation is explained by the fact that CpG sites are almost absent at synonymous positions (there are only two), and hence the expected GC for synonymous positions is almost the same when CpG sites are considered or excluded (the difference is just 0.5%). In summary, the coincidence obtained using the three different P53 mutation databases suggests that the estimations obtained from deleterious germline mutations are not dramatically biased by the detection skew. Although this comparison between the GC expectations obtained from deleterious and synonymous mutations was only possible for the gene encoding the p53 protein, it strongly supports the conclusion that GC_3-rich genes are indeed subjected to an AT bias.

To further investigate if the observed AT mutation bias could be due to detection bias, we have performed the following analysis. We counted the number of GC → AT and AT → GC mutations only in two subsets of codons, those that have either G/C in the first codon position and A/T in the second codon position (Leu4, His, Gln, Val, Asp and Glu) or A/T in the first codon position and G/C in the second position (Ser4, Ser2, Cys, Trp, Asn and Lys). In this analysis the idea is to compare the pattern of mutation inside the same amino acids as well to compare the pattern of mutations between the first and second codon positions.

Table 4
Comparison of mutation frequencies in codons containing either A/T in the first codon position and G/C in the second codon position or codons containing G or C in the first codon position and A or T in the second codon position

	CpGs included							CpGs excluded						
	A/T, G/C codons[a]		G/C, A/T codons[a]											
Codon position	2	1	1	2	1 & 2	1 & 2	**Ratio**	2	1	1	2	1 & 2	1 & 2	**Ratio**
	GC → AT	AT → GC	GC → AT	AT → GC	GC → AT	AT → GC		GC → AT	AT → GC	GC → AT	AT → GC	GC → AT	AT → GC	
1. GC_3-poor genes														
1.1. Haemophilia B factor (factor IX)	248	82	102	64	350	146	**2.40**	140	82	102	64	242	146	**1.66**
1.2. Haemophilia A factor (factor VIII)	43	22	75	45	118	67	**1.76**	38	22	68	44	106	66	**1.61**
1.3. Ataxia telangiectasia	6	1	29	9	35	10	**3.50**	4	1	27	9	31	10	**3.10**
1.4. Cystic fibrosis[b]	47	21	84	52	131	73	**1.79**	43	21	74	52	117	73	**1.60**
1.5. HPRT (Lesch–Nyhan syndrome)	50	3	102	44	152	47	**3.23**	50	3	101	44	151	47	**3.21**
Total	**394**	**129**	**392**	**214**	**786**	**343**	2.29	**275**	**129**	**372**	**213**	**647**	**342**	1.89
2. Genes with intermediate GC_3 content														
2.1. Phenylalanine hydroxylase locus[b]	18	17	31	22	49	39	**1.26**	17	17	25	22	42	39	**1.08**
2.2. PHEX	10	5	6	7	16	12	**1.33**	9	5	6	7	15	12	**1.25**
2.3. PAX6	3	3	14	4	17	7	**2.43**	3	3	12	4	15	7	**2.14**
Total	**31**	**25**	**51**	**33**	**82**	**58**	1.41	**29**	**25**	**43**	**33**	**72**	**58**	1.24
3. GC_3-rich genes														
3.1. Androgen receptor	14	8	48	10	62	18	**3.44**	14	8	37	10	51	18	**2.83**
3.2. P53 gene	9	0	5	2	14	2	**7.00**	7	0	5	2	12	2	**6.00**
(Somatic)[c]	(965)	(248)	(1429)	(573)	(2394)	(821)	**(2.92)**	(953)	(248)	(1411)	(573)	(2364)	(821)	**(2.88)**
3.3. Wilson disease	7	5	18	9	25	14	**1.79**	5	4	17	9	22	13	**1.69**
Total	**30**	**13**	**71**	**21**	**101**	**34**	2.97	**26**	**12**	**59**	**21**	**85**	**33**	2.58
4. Very GC_3-rich genes														
4.1. Glucose-6-phosphate dehydrogenase	8	2	20	6	28	8	**3.50**	8	2	13	6	21	8	**2.63**
4.2. L1CAM. L1 cell adhesion molecule[d]	–	–	–	–	–	–		–	–	–	–	–	–	
4.3. Haemophilia factor VII	10	6	7	5	17	11	**1.55**	8	6	5	5	13	11	**1.18**
4.4. LDLR locus	140	61	165	40	305	101	**3.02**	138	61	126	40	264	101	**2.61**
Total	**158**	**69**	**192**	**51**	**350**	**120**	2.92	**154**	**69**	**144**	**51**	**298**	**120**	2.48

[a] For the subset of A/T, G/C codons AGA and AGG (Arg2) were excluded, because the mutations A → C at the first codon position are synonymous and hence they are likely to be sub-estimated. Likewise CTA and CTG we excluded as well because C → T are synonymous (Leu4 to Leu2).

[b] Although this table contains analyses including repeated mutations, for these two mutation databases unique mutations were considered instead due to the absence of information on repetitions.

[c] The P53 somatic mutation database was included only as a comparison with the germline database given that the germline database is small and the two databases exhibit very similar mutation patterns (see Table 3).

[d] The L1CAM mutation database was not used in this analysis because of the small sample size and the scarcity of theses types of codons in the database.

Take into account that it is reasonable to claim for instance that mutations falling in the second codon position are more likely to produce a deleterious phenotype because in general they imply a change between two biochemical dissimilar amino acids. Consequently GC_2-rich genes could give the erroneous impression that the mutation pattern is AT-biased. However, if GC → AT mutations are more frequent than AT → GC mutations regardless of whether they affect the first or second codon position then it is safe to conclude that this is not due to a overrepresentation of mutations falling in a given codon position nor to a overrepresentation of mutations affecting some amino acid. The results presented in Table 4 show that GC → AT mutations not only are more frequent overall in these codons, but also they are more frequent in each subset when considered separately, namely GC → AT mutations predominate in both G/C starting codons and A/T starting codons. This predominance is observed in the four groups of genes analyzed in this work.

4. Discussion

The problem under consideration here has also been investigated by other authors. For example, the analysis of substitution patterns in pseudogenes can be used to infer the underlying pattern of mutations since pseudogenes are thought to have no function and are thus free of selectional constraints. This approach has given, however, inconclusive results. While some analyses indicate that all pseudogenes (located in both AT-rich and GC-rich isochores) are under AT biases (Gojobori et al., 1982), a widely cited study on the substitution patterns of two globin pseudogenes located in different isochores proposes that the pattern of substitutions is not the same (Francino and Ochman, 1999). According to the latter authors, the expected GC level that is obtained from the β-globin gene (which is located in a GC-poor isochore) would be 0.4, whereas the expected GC level for the α-globin pseudogene (located in a GC-rich isochore) would be 0.57.

Variations in the substitution patterns were also observed among putatively functionless repetitive sequences located in genome regions having different GC levels (The International Human Genome Sequencing Consortium, 2001). In this case too, the bias in the mutational pattern was an AT bias, with expected GC levels ranging from 0.29 to 0.44.

Although the analyses of substitutions in sequences under very weak or no selection (e.g. pseudogenes and repetitive sequences) suggest that the pattern of mutations may change along the genome, it is remarkable that the level of variation reported in these studies is always much less than the variability of the GC level actually observed in the putatively neutral parts of coding sequences, namely the GC level at the third position of codons. Indeed, the average GC_3 of genes located in H3 isochores is 80%, a figure that is much higher than the expectations one obtains from pseudogenes (Gojobori et al., 1982; Francino and Ochman, 1999) or repeated elements located in the GC-rich parts of the genome (The International Human Genome Sequencing Consortium, 2001).

This puzzling situation leads us to raise the following question: is it possible that the mutational spectrum varies in a small scale range in such a way that GC_3-rich genes are under stronger GC mutational biases than the non-coding DNA segments located in the same isochores?

The results presented here indicate that the mutational patterns inside GC_3-rich genes do not appear to explain their GC_3-richness. Indeed, the most remarkable result of this work is that GC_3-rich and very GC_3-rich genes exhibit mutation patterns that yield expectations of a neutral GC level that is much lower than the actual GC_3. In fact, the observed GC_3 levels are between two and three times higher than the expectations from mutational data.

Our results are in agreement with recent analyses on single nucleotide polymorphisms (Eyre-Walker, 1999; Smith and Eyre-Walker, 2001). These authors found that, contrary to what would be expected if the GC level were determined solely by mutational biases (and assuming a mutational equilibrium condition), the number of segregating GC → AT polymorphisms is significantly higher than that of AT → GC ones in genes with high GC_3 levels. It is clear that if these sequences were in mutational equilibrium, the two kinds of polymorphism would be present in equal amounts. The excess of GC → AT polymorphisms, in high GC_3 genes, is an indication that the GC_3-richness of these genes cannot be explained by their mutational input.

It could be argued that the results we present in this work as well as those based on polymorphism data might indicate only the mutational spectrum in present day sequences and that observed GC_3 levels could be reflecting mutational patterns from the past (Piganeau et al., 2002). In other words, it could be argued that present day GC_3-rich genes became GC_3-rich because they were subjected to GC mutation bias in the past. If this were correct one would expect that today they are decreasing their GC content due to the novel mutation pattern, which in turn would imply an excess of GC → AT substitutions over AT → GC substitutions at synonymous positions. However, no evidence has been found that the GC_3 levels are decreasing, as would be expected if this hypothesis were correct. Secondly, analysis of substitution frequencies in primate genes shows that at synonymous positions GC → AT and AT → GC substitutions are observed in approximately similar amounts (Smith and Eyre-Walker, 2001). Moreover, the analysis of repetitive sequences mentioned above, which included sequences that diverged from each other several million years ago (The International Human Genome Sequencing Consortium, 2001), is also in agreement with the claim that AT biases are not a recent phenomenon in the human genome.

The results presented here, along with previous results, strongly suggest that the variability in GC_3 cannot be attributed to variations in the mutational input along the genome but is instead due to fixational biases. In other words, even if GC_3-rich genes are subject to AT biases (or a slight GC bias according to the analysis of the α-globin pseudogene) these genes remain GC_3-rich, implying that the majority of new mutations introduced into the population that are GC → AT are eliminated by negative selection. Recently Eyre-Walker and Hurst (2001) have suggested that the high GC levels could be the result of biased gene conversion, because there is a positive correlation between GC levels and recombination frequencies. However, some genes located in regions where recombination is completely absent (such as the Y chromosome) still display high GC_3 levels.

Rejecting the mutation bias variation hypothesis as a possible explanation for the variability of GC_3 values also suggests that the compositional heterogeneity of isochores cannot be attributed to mutational biases. Indeed, whatever factor is responsible for maintaining high GC_3 levels, it can be expected to affect the whole isochore, because of the strong correlation between the GC level of the isochore and the GC_3 levels of its genes.

Acknowledgements

We wish to thank Oliver Clay for critical reading of the manuscript.

References

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.

Bernardi, G., Bernardi, G., 1986. Compositional constraints and genome evolution. J. Mol. Evol. 24, 1–11.

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.

Cariello, N.F., Skopek, T.R., 1993. In vivo mutations at the human hprt locus. Trends Genet. 9, 322–326.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G., 1981. The major components of the mouse and human genomes: preparation, basic properties and compositional heterogeneity. Eur. J. Biochem. 115, 227–233.

Eyre-Walker, A., 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics 152, 675–683.

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

Filipski, J., 1987. Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett. 217, 184–186.

Filipski, J., Thiery, J.P., Bernardi, G., 1973. An analysis of the bovine genome by Cs_2SO_4-Ag^+ density gradient centrifugation. J. Mol. Biol. 80, 177–197.

Francino, P., Ochman, H., 1999. Isochores result from mutation not selection. Nature 400, 30–31.

Gojobori, T., Li, W.-H., Graur, D., 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18, 360–369.

Krawczak, M., Cooper, D.N., 1996. Mutational processes in pathology and evolution. In: Jackson, M., Strachan, T., Dover, G. (Eds.), Human Genome Evolution, BIOS Scientific, Oxford, pp. 1–33.

Kunkel, T.A., Alexander, P.S., 1986. The base substitution fidelity of eucaryotic DNA polymerases. Mispairing frequencies, site preferences, insertion preferences, and base substitution by dislocation. J. Biol. Chem. 261 (1), 160–166.

Macaya, G., Thiery, J.P., Bernardi, G., 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. J. Mol. Biol. 108, 237–254.

Piganeau, G., Mouchiroud, D., Duret, L., Gautier, C., 2002. Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. J. Mol. Evol. 54, 129–133.

Smith, N.G., Eyre-Walker, A., 2001. Synonymous codon bias is not caused by mutation bias in G + C-rich genes in humans. Mol. Biol. Evol. 18, 982–986.

Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85, 2653–2657.

Sueoka, N., 1992. Directional mutation pressure, selective constraints, and genetic equilibria. J. Mol. Evol. 34, 95–114.

The International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Van Camp, G., Fransen, E., Vits, L., Raes, G., Willems, P.J., 1996. A locus-specific mutation database for the neural cell adhesion molecule L1CAM (Xq28). Hum. Mutat. 8, 391.

Wolfe, K.H., Sharp, P.M., Li, W.H., 1989. Mutation rates differ among regions of the mammalian genome. Nature 337, 283–285.