Chromosomal location effects on gene sequence evolution in mammals

Giorgio Matassi*†, Paul M. Sharp* and Christian Gautier‡

Background: Nucleotide substitution rates and G+C content vary considerably among mammalian genes. It has been proposed that the mammalian genome comprises a mosaic of regions – termed isochores – with differing G+C content. The regional variation in gene G+C content might therefore be a reflection of the isochore structure of chromosomes, but the factors influencing the variation of nucleotide substitution rate are still open to question.

Results: To examine whether nucleotide substitution rates and gene G+C content are influenced by the chromosomal location of genes, we compared human and murid (mouse or rat) orthologues known to belong to one of the chromosomal (autosomal) segments conserved between these species. Multiple members of gene families were excluded from the dataset. Sets of neighbouring genes were defined as those lying within 1 centiMorgan (cM) of each other on the mouse genetic map. For both synonymous substitution rates and G+C content at silent sites, neighbouring genes were found to be significantly more similar to each other than sets of genes randomly drawn from the dataset. Moreover, we demonstrated that the regional similarities in G+C content (isochores) and synonymous substitution rate were independent of each other.

Conclusions: Our results provide the first substantial statistical evidence for the existence of a regional variation in the synonymous substitution rate within the mammalian genome, indicating that different chromosomal regions evolve at different rates. This regional phenomenon which shapes gene evolution could reflect the existence of 'evolutionary rate units' along the chromosome.

Addresses: *Institute of Genetics, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH, UK. ‡Laboratoire de Biometrie, Génétique et Biologie des Populations, UMR 5558, Université Claude Bernard-Lyon 1, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne Cedex, France.

Present address: †INSERM U76, 6 rue Alexandre Cabanel, 75015 Paris, France.

Correspondence: Giorgio Matassi
E-mail: gmatassi@infobiogen.fr

Received: 17 May 1999
Revised: 18 June 1999
Accepted: 21 June 1999

Published: 14 July 1999

Current Biology 1999, 9:786–791
http://biomednet.com/elecref/0960982200900786

Background

In mammalian genomes, variation at silent sites — that is, synonymously variable positions within protein-coding genes — has been considered to be largely neutral [1] and so may be indicative of the fundamental rates and patterns of mutation [2–5]. It is therefore interesting that the evolution of silent sites varies dramatically among genes in two respects. First, genes vary enormously in their patterns of codon usage, with G+C content at silent sites ranging from about 30% to 95% [6]. Second, genes differ substantially in their rates of evolution at silent sites, with an approximately 20-fold variation among genes compared between the same pair of species [7].

Extensive analyses of G+C content variability in mammalian genomes have been carried out using both experimental and statistical approaches. These have provided strong evidence that the G+C content of a gene is related to its chromosomal location. Indeed, it has been proposed that the mammalian genome comprises a mosaic of regions with differing G+C content, which have been termed 'isochores' (see [8] for a review). As sequence data for mammalian genes accumulated, it became apparent that the G+C content at silent sites, within introns, and in 5′ and 3′ flanking regions are all correlated [9–11], consistent with the idea that this variation reflects a local chromosomal effect. There have also been reports that neighbouring genes have similar G+C content at silent sites [4,12,13], but these analyses have included only very small numbers of genes, and in many cases these were members of gene families which may be similar because of their common functional constraints.

Mammalian genes vary systematically in their rates of synonymous substitution, in so far as the rate of any particular gene is maintained across different mammalian lineages [14–16]. Previous attempts to explain variation in substitution rates at silent sites among mammalian genes have mainly focused on whether this variation is linked to G+C content. Initially, it appeared that there was a strong correlation between G+C content and evolutionary rate [3,4,17], but as more gene sequences have become available for comparison this relationship has become less evident [7,15,18]. It has been proposed that the variation of the rates of synonymous substitution could be related to the variation of patterns of mutation across the genome [2–4].

Here, we propose a new statistical approach to detect and analyse regional organisations in the mammalian genome. We analysed a large dataset of mammalian genes and found that, for both G+C content and substitution rate at silent sites, neighbouring genes were significantly more similar than random sets of genes within the dataset, suggesting that chromosomal location influences both evolutionary parameters. These similarities define two aspects of genome structure, which we found to be independent of each other.

Results

Mammalian genes exhibit extensive variation in both G+C content and substitution rate at silent sites. To examine whether these evolutionary parameters are related to chromosomal location, we compiled a dataset of orthologous human and murid (mouse or rat) genes known to lie within one of the chromosomal (autosomal) segments of conserved synteny between human and mouse. These gene pairs were ordered according to their position on the consensus mouse genetic map [19], and any genes not lying (on the mouse genetic map) within 1 centiMorgan (cM) of another gene in the dataset were discarded. In addition, and importantly, where closely linked genes included members of families resulting from gene duplications, only one member of the family was retained. This yielded a dataset of 520 genes (see Materials and methods).

For each gene, the following evolutionary parameters were calculated: the G+C content at synonymously variable third positions of codons ($GC3_S$) in each species, and the estimated numbers of synonymous and nonsynonymous substitutions per site, K_S and K_A, respectively, between the human and murid genes (see Materials and methods). In the present dataset, $GC3_S$ varied from 26% to 96% and thus covered the range previously reported for other compilations of (not necessarily homologous) human and murid genes [7,20]. The estimated number of nonsynonymous substitutions per site (K_A) varied from 0.001 to 0.453 in the dataset; such variation is expected given the known variation in rates of protein evolution [21]. The estimated number of synonymous substitutions per site (K_S) also varied considerably among genes, with values ranging from 0.16 to 1.00. As synonymous substitutions are expected to be largely neutral in mammals, this observation is more surprising. The mean and standard deviation of synonymous and nonsynonymous distances found within our dataset ($K_S = 0.518 \pm 0.136$; $K_A = 0.094 \pm 0.083$) are comparable to those previously reported for orthologous gene pairs between human and mouse ($K_S = 0.468 \pm 0.169$; $K_A = 0.090 \pm 0.102$) [22]. The variation of K_S and K_A values found in the present study appears therefore to be representative of the variation seen among mammalian genes.

To investigate whether each of these evolutionary parameters is influenced by genomic location, we determined whether the values could be correlated among genes lying close to one another on the chromosome. Neighbouring genes were defined as those located within the 0–1 cM range on the mouse genetic map. Among all possible pairwise comparisons of the 520 genes, 1344 pairs were defined as neighbours by this criterion. Spatial autocorrelation analysis among neighbouring genes was performed by estimating Moran's *I* statistic (Table 1). The significance of these *I* values was estimated first by simulations in which the values for all genes were permuted (model 1). Highly significant similarity between neighbouring genes appeared for both K_S and $GC3_S$: the observed value was

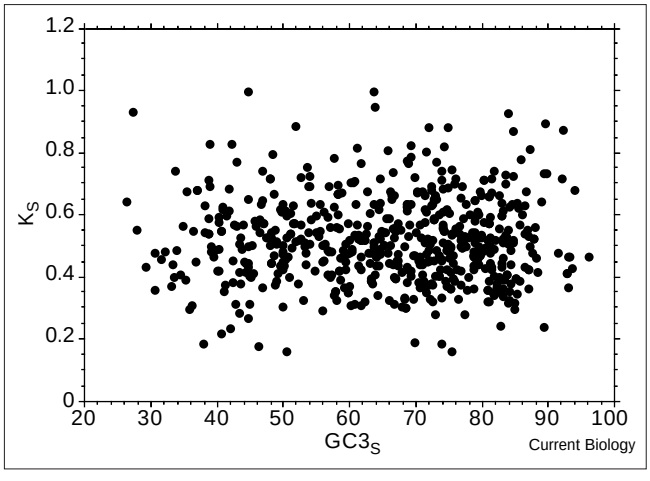
Table 1

Spatial correlation of evolutionary statistics between neighbouring genes in the 0–1 cM genetic distance range.

Evolutionary parameter	*I*	*P*, model 1	*P*, model 2
K_S	0.17	$< 10^{-4}$	$< 10^{-4}$
K_A	0.07	0.080	0.088
$GC3_S$ human	0.24	$< 10^{-4}$	0.212
$GC3_S$ murid	0.23	$< 10^{-4}$	0.076

K_S, estimated number of synonymous nucleotide substitutions per site; K_A, estimated number of nonsynonymous nucleotide substitutions per site; $GC3_S$, G+C content at synonymously variable third positions of codons; *I*, Moran's correlation coefficient; *P*, probability value for the correlation coefficient (estimated as the fraction of 1,000 permutations in which a coefficient of at least this magnitude was obtained); model 1, all genes were permuted; model 2, only those genes within a 10% $GC3_S$ range were permuted (see Materials and methods).

Figure 1



Relationship between silent-site divergence between human and murids (K_S) and the silent-site base composition ($GC3_S$) of the human gene. A similar lack of correlation was seen when the $GC3_S$ values of the murid gene were used.

higher than any of the 1,000 simulation values. K_A showed no significant spatial correlation.

The significant results for $GC3_S$ and K_S could reflect a correlation between these parameters. There was, however, no obvious relationship between $GC3_S$ and K_S values across genes (Figure 1). Nevertheless, to test the possible influence of $GC3_S$ on the correlation of K_S values, a second series of simulations was performed, in which only genes falling within classes of 10% $GC3_S$ were permuted (Table 1, model 2). Under this model, the correlation coefficients for $GC3_S$ became (as expected) non-significant, but the probability value for K_S remained highly significant. Therefore, the similarity of K_S between neighbouring genes is not the indirect consequence of the similarity of their G+C content.

Discussion

Isochores

A spatial organisation of G+C content within the mammalian genome (the isochore hypothesis) was first proposed by Bernardi and colleagues in 1976 on the basis of ultracentrifugation results [23]. Statistical analyses of sequence data have provided results consistent with the isochore hypothesis, but have been limited either to very local effects, such as comparisons of different sites in and around a gene (see [8] for a review), or to a small number of larger genomic regions [12,13,24]. In this study, we have shown that neighbouring genes in both the human and mouse genomes have similar G+C contents at silent sites (Table 1, model 1). Previous studies of neighbouring genes have used much smaller datasets, and those datasets were dominated by multiple members of gene families [4]. Our results provide a statistical confirmation of the isochore hypothesis as we analysed a large number of genes that, firstly, were scattered across the entire genome (the dataset included gene pairs located on all 19 mouse autosomes) and, secondly, were located on chromosomal regions as large as 1 cM (on average, 1 cM corresponds to ~1.7 Mb in the mouse genome [25]).

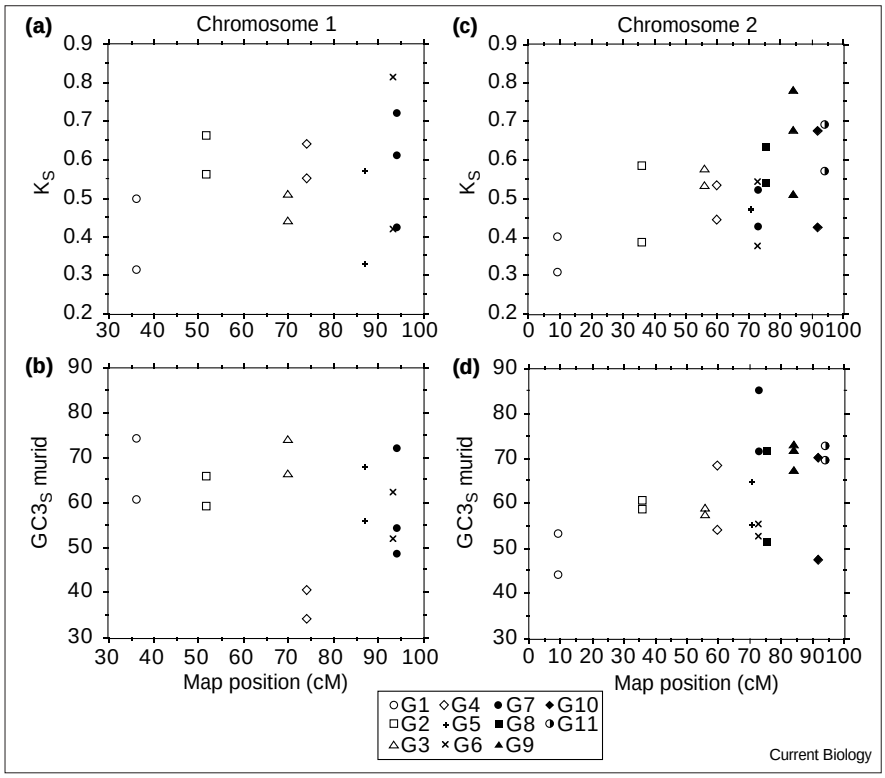
A 'regional' effect on K_S variation across the genome

There have been previous suggestions that variation among mammalian genes in the rate of synonymous substitution may be influenced by gene location [4,26–28], but the data supporting this have been limited to very small numbers of genes. Our results indicate that, for a very large dataset of genes drawn from across the entire genome, pairs of neighbouring genes have significantly more similar rates of synonymous substitution than random pairs of genes (Table 1, model 1). Members of gene families from the same vicinity were excluded and so the only common feature shared by these neighbouring genes is their location. As synonymous substitutions in mammalian genes are likely to be neutral [1,5,26], we infer that there are chromosome-location-dependent effects on the rate of mutation.

Moreover, we have demonstrated that the location-dependent variation in evolutionary rate is independent of the location-dependent variation in base composition (isochores). The permutation test in model 2 confirmed that the correlation of K_S values among neighbouring genes was not attributable to their similarity in $GC3_S$. The similarity of neighbouring genes with respect to both $GC3_S$ and K_S, without a correlation between the two parameters, is illustrated for genes from the two mouse chromosomes best represented in our dataset (Figure 2). For example, on chromosome 1 at 69.9 cM, IL10 and REN (open triangles, Figure 2a) have very similar K_S and $GC3_S$, as do F13B and HF1 (open diamonds) at 74.1 cM. Nevertheless, while the K_S values of the two pairs of genes are quite similar, the $GC3_S$ values are very different (Figure 2b). Many similar examples are found on the other chromosomes (data not shown).

Overall, our results provide the first statistical evidence, on a whole-genome scale, demonstrating that nucleotide substitution rate varies among different chromosomal regions in mammals and suggest the existence of a 'regional' effect driving gene evolution. Results from several studies of small numbers of genes, or of particular chromosomal regions, are consistent with our finding of location-dependent rates of evolution. A regional effect on K_S variability has been shown in the regions surrounding three arginosuccinate-synthetase-processed pseudogenes located at different sites in the primate genome [28]. Moreover, evidence is slowly gathering from large-scale sequencing projects in favour of the existence of genomic regions characterised by high [29–31] and low sequence conservation [32–36]. An intermediate level of sequence conservation between human and mouse was found in the ~24 kb region spanning the immunoglobulin Cμ–Cδ heavy chain loci [37]. Although these studies provide comparisons of noncoding regions, they nevertheless deal with only one or a few loci and analyse relatively short distances (10–100 kb) from a genome perspective. A remarkable piece of evidence on a region showing high sequence conservation has come from a comparison between a 227 kb region on mouse chromosome 6 and its syntenic region on human chromosome 12p13: these regions were found to contain 17 homologous gene pairs [38]. We have carried out a variance analysis comparing the K_S values of 16 gene pairs (one duplicate gene was excluded) in this region with those of the whole set of homologous genes used in the present study and, as expected, found that, within the 227 kb region, the variance of K_S is significantly smaller ($F = 20.2$, $p < 10^{-4}$). Finally, a different kind of evidence consistent with the regional variation in evolutionary rate comes from the finding that in human isogenic lymphoblastoid cell lines, the mutation rate of hamster *hprt* cDNAs was found to vary by 60-fold in five retroviral sequences integrated in different genomic locations [39].

Figure 2



Chromosomal location effect on gene evolution. (a,c) K_S and (b,d) murid $GC3_S$ values are plotted against the position of the corresponding genes on the mouse genetic map for (a,b) chromosome 1 and (c,d) chromosome 2. Only groups of genes (G) at the 'same map position', represented by the same symbol, are shown for clarity (duplicate genes were excluded). On chromosome 1, G1: FN1 and IGFBP5; G2: ALPI and UGT1A1; G3: IL10 and REN; G4: F13B and HF1; G5: CD3Z and POU2F1; G6: CD48 and FCER1G; G7: APCS, ATP1A2 and FCER1A. On chromosome 2, G1: BMI1 and GAD2; G2: GCG and SCN2A1; G3: RAG1 and CAT; G4: FSHB and LM02; G5: FBN1 and HDC; G6: SLC20A1 and IL1A; G7: OXT and PDYN; G8: PCNA and PRNP; G9: CST3, PYGB and THBD; G10: PLCG1 and TOP1; G11: ADA and SDC4.

The regional variation of the synonymous substitution rate shown in this study may reflect the existence of 'evolutionary rate units' along the chromosome (whose size(s) remain to be determined) and therefore may suggest the existence of a novel mosaic structure in the mammalian genome. Even though this proposal must be viewed with caution, as at present we cannot rule out the possibility that the regional similarity of K_S values is independent of any regular underlying genome structure, it nonetheless sets out a working hypothesis that is consistent with all the lines of experimental evidence cited above and integrates them in a general framework for the understanding of the organisation and evolution of the mammalian genome.

Origin and maintenance of the regional variation of K_S

The existence of a regional variation in the silent substitution rate raises the question of the nature of the evolutionary mechanisms that have engendered and maintained such a structure in the mammalian genome. Let us consider first the possible role of natural selection. Regional similarity in K_S might reflect varying levels of functional constraints that are pervasive across entire chromosomal regions, but it is not clear what such selective constraints might be. Moreover, mechanisms such as background selection against deleterious mutations [40], or hitchhiking with advantageous mutations [41], do not seem likely to act either across longer evolutionary timescales (such as the divergence between primates and rodents) or across large genomic regions [42]. Thus, natural selection does not appear to be the most likely explanation of the regional variation of K_S.

In genomic contexts, where selective constraints are absent or weak, substitution rate is highly correlated with mutation rate [1]. Therefore, the most likely explanation we favour to explain the regional variation of silent substitution rate is the variation of mutation rate in different chromosomal regions. A number of hypotheses have been proposed to explain why mutational patterns might vary across the genome [3,4,43–45]. Among them, the proposal that this variation could be related to the different efficiency of repair of DNA lesions caused by endogenous and/or exogenous factors in different regions of the genome is particularly appealing [3,43,44]. Previous attempts to test this hypothesis by modelling focused on the effect of the mismatch repair system [46]. The efficiency of DNA repair mechanisms such as base excision repair and nucleotide excision repair have, however, been shown to vary around the genome [47–50]. Moreover, these processes seem to act over entire chromosomal regions [47,51], for example, the DNA repair domain that comprises the ~50–70 kb region surrounding the p53 gene [52]. Therefore, the different fidelities of the DNA repair machinery may be one of the factors causing the regional

variation of the silent substitution rate observed in the present study. This hypothesis obviously calls for experimental evidence.

Conclusions

Here, we have provided strong statistical evidence demonstrating the existence of a regional variation in the synonymous substitution rate, which can account for the variability of K_S among mammalian genes. Our results show that the regional variation of K_S extends over relatively large regions (≤ 1 cM) of the mammalian genome and therefore tend to suggest the existence of evolutionary rate units along chromosomes. Moreover, we have demonstrated that regional variations in synonymous substitution rate and G+C content (isochores) are superimposed and independent of one another: this leads to a very complex view of the forces shaping mammalian genome evolution. Finally, the characterisation at the molecular level of the chromosomal regions harbouring neighbouring genes evolving at different rates may shed light on whether the regional variation of K_S is somehow related to any of the fundamental processes of DNA metabolism: replication, recombination, transcription and repair.

Materials and methods

The dataset

Homologous human and murid gene pairs were identified by using the similarity search programs BLASTN and BLASTP [53] as well as the Hovergen database [54]. Where both mouse and rat sequences were available, the mouse gene was preferred. If a mouse sequence was not available, the rat sequence was used instead, provided there was evidence that the genomic region is conserved between mouse and rat [55]. Sex-linked genes were excluded owing to their lower nucleotide substitution rates as compared with autosomal genes [7]. Partial sequences were discarded, to avoid possible systematic biases related to intragenic position. Duplicate genes were identified by BLAST analysis: all sequences in the dataset were compared, and any pairs with a value greater than 150 for the S parameter were regarded as duplicates. In cases where, among three genes A1, A2 and A3, two pairs (say A1 plus A2, and A2 plus A3) were designated by this criterion as duplicates, but the third pair (A1 plus A3) was not, we nevertheless chose to consider A1 and A3 as duplicates even though S < 150. When two (or more) genes were recognised as duplicates, only one of them was retained: either the longer sequence, or (in the case of equal sequence length) a random choice. The final dataset consisted of 520 genes (available upon request).

Statistical analysis

Human and murid protein-coding sequences were extracted from GenBank using the ACNUC retrieval system [56] and the deduced protein sequences were aligned using the multiple alignment program CLUSTALV [57]. The number of synonymous (K_S) and nonsynonymous (K_A) substitutions per site were calculated using the method of Li [58] which uses a correction for multiple substitution at single sites based on the two-parameter method of Kimura [59].

Spatial autocorrelation analysis was performed using the *I* statistic by Moran [60]:

$$I = \frac{n}{2A}\frac{\sum_{i\neq j}\delta_{ij}z_i z_j}{\sum z_i^2} \quad \text{where} \quad z_i = x_i - \bar{x}\ ,\ A = \frac{1}{2}\sum \delta_{ij}$$

$\delta_{ij} = \{$ 1 *if i and j are joined (i.e., belong to the* 0–1 *cM distance segment) and* $i \neq j$ *else* 0

where *n* is the number of genes and $\bar{x} = \frac{\sum x_i}{n}$. Here, *n* = 520 and A = 1344.

$[\delta_{ij}]$ defines a joint matrix and therefore allows to explore the spatial correlation of the *X* variable (that is, K_S, K_A, human $GC3_S$ and murid $GC3_S$) as *i* and *j* are joined when they lie in the same 0–1 cM segment.

The *I* statistic is similar to a correlation coefficient, but is able to handle situations where *k* is joined to both *i* and *j*, and yet *i* and *j* are not joined, as can occur here when gene *k* lies between genes *i* and *j*. It should be noted that the mathematical behaviour of *I* is somewhat different from a correlation coefficient, in particular its maximum is not 1 and its expectation, in the absence of correlation, is –1/(n–1), so slightly different from 0 (here –1/519 = –0.002) [61].

Two models were used to test the significance of the *I* values obtained. In model 1, the observed values of *X* were randomly permuted; *I* is known to be approximately normally distributed under such a model, and this was confirmed by Monte Carlo simulations (only simulation results are presented in this study). In model 2, 1,000 simulations were carried out permuting only those genes falling within classes of 10% $GC3_S$ content (random number generator was ran1 [62]). The *I* statistic was computed for each of these simulations, and the fraction of these values greater than the observed value gave an estimate of the significance of the observed value. The 10% $GC3_S$ range was chosen as the largest one for which the isochore structure is no more significant in this dataset.

Acknowledgements

We thank D. Mouchiroud, J.F.Y. Brookfield, G. Bernardi and O. Clay for helpful discussion. This work was supported by grants from the European Commission (CHRX-CT-0196 and BIO4-CT95-0130), the French GREG (92.H.0929) and the BBSRC (G04905). G.M. most warmly thanks A. Falaschi for encouragement and also UNIDO/ICGEB for financial support in the early stages of this project.

References

1. Kimura M: Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 1977, 267:275-276.
2. Sueoka N: Directional mutation pressure, selective constraints and genetic equilibria. *Proc Natl Acad Sci USA* 1991, 85:2653-2657.
3. Filipski J: Why the rate of silent codon substitutions is variable within a vertebrate's genome. *J Theor Biol* 1988, 134:159-164.
4. Wolfe KH, Sharp PM, Li WH: Mutation rates differ among regions of the mammalian genome. *Nature* 1989, 337:283-285.
5. Eyre-Walker A: An analysis of codon usage in mammals: selection or mutation bias? *J Mol Evol* 1991, 33:504-510.
6. Ikemura T, Wada K: Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data. *Nucleic Acids Res* 1991, 19:4333-4339.
7. Wolfe KH, Sharp PM: Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J Mol Evol* 1993, 37:441-456.
8. Bernardi G: The human genome: organization and evolutionary history. *Annu Rev Genet* 1995, 29:445-476.
9. Aota S, Ikemura T: Diversity in G+C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res* 1986, 14:6345-6355.
10. Mouchiroud D, D'Onofrio G, Aissani B, Macaya G, Gautier C, Bernardi G: The distribution of genes in the human genome. *Gene* 1991, 100:181-187.
11. D'Onofrio G, Mouchiroud D, Aissani B, Gautier C, Bernardi G: Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J Mol Evol* 1991, 32:504-510.

12. Ikemura T, Wada K, Aota S: Giant G+C% mosaic structures of the human genome found by arrangement of GenBank human DNA sequences according to genetic positions. *Genomics* 1990, 8:207-216.
13. Fukagawa T, Sugaya K, Matsumoto K, Okumura K, Ando A, Inoko H, *et al.*: A boundary of long-range G+C% mosaic domains in the human MHC locus: pseudoautosomal boundary-like sequence exists near the boundary. *Genomics* 1995, 25:184-191.
14. Bulmer M, Wolfe KH, Sharp PM: Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. *Proc Natl Acad Sci USA* 1991, 88:5974-5978.
15. Mouchiroud D, Gautier C, Bernardi G: Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. *J Mol Evol* 1995, 40:107-113.
16. Mouchiroud D, Robinson M, Gautier C: Impact of changes in GC content on the silent molecular clock in murids. *Gene* 1997, 205:317-322.
17. Ticher A, Graur D: Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. *J Mol Evol* 1989, 28:286-298.
18. Bernardi G, Mouchiroud D, Gautier C: Silent substitutions in mammalian genomes and their evolutionary implications. *J Mol Evol* 1993, 37:583-589.
19. The Jackson Laboratory: *Chromosome Committee Reports*. 1997.
20. Zoubak S, Clay O, Bernardi G: The gene distribution of the human genome. *Gene* 1996, 174:95-102.
21. Nei M: *Molecular Evolutionary Genetics*. New York: Columbia University Press; 1987.
22. Makalowski W, Boguski MS: Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci USA* 1998, 95:9407-9412.
23. Macaya G, Thiery JP, Bernardi G: An approach to the organization of eukaryotic genomes and their evolutionary implications. *J Mol Biol* 1976, 108:237-254.
24. Bettecken T, Aissani B, Muller CR, Bernardi G: Compositional mapping of the human dystrophin-encoding gene. *Gene* 1992, 122:329-335.
25. Silver LM: *Mouse Genetics. Concepts and Applications.* Oxford: Oxford University Press; 1995.
26. Sharp PM, Matassi G: Codon usage and genome evolution. *Curr Opin Genet Dev* 1994, 4:851-860.
27 Koop BF: Human and rodent DNA sequence comparison: a mosaic model of genomic evolution. *Trends Genet* 1995, 11:367-371.
28. Casane D, Boissinot S, Chang BH, Shimmin LC, Li WH: Mutation pattern variation among regions of the primate genome. *J Mol Evol* 1997, 45:216-226.
29. Koop BF, Hood L: Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat Genet* 1994, 7:48-53.
30. Epp TA, Wang R, Sole MJ, Liew CC: Concerted evolution of mammalian cardiac myosin heavy chain genes. *J Mol Evol* 1995, 41:284-292.
31. Oeltjen JC, Malley TM, Muzny DM, Miller W, Gibbs RA, Belmont JW: Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res* 1997, 7:315-329.
32. den Dunnen JT, van Neck JW, Cremers FP, Lubsen NH, Schoenmakers JG: Nucleotide sequence of the rat gamma-crystallin gene region and comparison with an orthologous human region. *Gene* 1989, 78:201-213.
33. Shehee WR, Loeb DD, Adey NB, Burton FH, Casavant NC, Cole P, *et al.*: Nucleotide sequence of the BALB/c mouse beta-globin complex. *J Mol Biol* 1989, 205:41-62.
34. Hardison R, Krane D, Vandenbergh D, Cheng JF, Mansberger J, Taddie J, *et al.*: Sequence and comparative analysis of the rabbit alpha-like globin gene cluster reveals a rapid mode of evolution in a G+C-rich region of mammalian genomes. *J Mol Biol* 1991, 222:233-249.
35. Lamerdin JE, Montgomery MA, Stilwagen SA, Scheidecker LK, Tebbs RS, Brookman KW, *et al.*: Genomic sequence comparison of the human and mouse XRCC1 DNA repair gene regions. *Genomics* 1995, 25:547-554.
36. Lamerdin JE, Stilwagen SA, Ramirez MH, Stubbs L, Carrano AV: Sequence analysis of the ERCC2 gene regions in human, mouse, and hamster reveals three linked genes. *Genomics* 1996, 34:399-409.
37. Koop BF, Richards JE, Durfee TD, Bansberg J, Wells J, Gilliam AC, *et al.*: Analysis and comparison of the mouse and human immunoglobulin heavy chain JH-Cmu-Cdelta locus. *Mol Phylogenet Evol* 1996, 5:33-49.
38. Ansari-Lari MA, Oeltjen JC, Schwartz S, Zhang Z, Muzny DM, Lu J, *et al.*: Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res* 1998, 8:29-40.
39. Lichtenauer-Kaligis EG, Van Der Velde-Van Dijke I, Den Dulk H, Van De Putte P, Giphart-Gassler M, Tasseron-De Jong JG: Genomic position influences spontaneous mutagenesis of an integrated retroviral vector containing the hprt cDNA as target for mutagenesis. *Hum Mol Genet* 1993, 2:173-182.
40. Maynard-Smith J, Haigh J: The hitch-hiking effect of a favourable gene. *Genet Res* 1974, 23:23-25.
41. Charlesworth D, Charlesworth B: Sequence variation: looking for effects of genetic linkage. *Curr Biol* 1998, 8:R658-R661.
42. Birky CW, Walsh JB: Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci USA* 1988, 85:6414-6418.
43 Holmquist GP, Filipksi J: Organization of mutations along the genome: a prime determinant of genome evolution. *Trends Ecol Evol* 1994, 9:65-69.
44. Boulikas T: The evolutionary consequences of nonrandom damage and repair of chromatin domains. *J Mol Evol* 1992, 35:156-180.
45. Holmquist GP: Endogenous lesions, S-phase-independent spontaneous mutations, and evolutionary strategies for base excision repair. *Mutation Res* 1998, 400:59-68.
46. Eyre-Walker A: DNA mismatch repair and synonymous codon evolution in mammals. *Mol Biol Evol* 1994, 11:88-98.
47. Bohr VA, Smith CA, Okumoto DS, Hanawalt PC: DNA repair in an active gene: removal of pyrimidine dimers from the DHFR gene of CHO cells is much more efficient than in the genome overall. *Cell* 1985, 40:359-369.
48. Bohr VA, Phillips DH, Hanawalt PC: Heterogeneous DNA damage and repair. *Cancer Res* 1987, 47:6426-6436.
49. Hanawalt PC: Preferential repair of damage in actively transcribed DNA sequences ***in vivo***. *Genome* 1989, 31:605-611.
50. Svetlova MP, Solovjeva LV, Pleskach NA, Tomilin NV: Focal sites of DNA repair synthesis in human chromosomes. *Biochem Biophys Res Commun* 1999, 257:378-383.
51. Kantor GJ, Barsalou LS, Hanawalt PC: Selective repair of specific chromatin domains in UV-irradiated cells from xeroderma pigmentosum complementation group C. *Mutation Res* 1990, 235:171-180.
52. Tolbert DM, Kantor GJ: Definition of a DNA repair domain in the genomic region containing the human p53 gene. *Cancer Res* 1996, 56:3324-3330.
53. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215:403-410.
54. Duret L, Mouchiroud D, Gouy M: HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res* 1994, 22:2360-2365.
55. The Jackson Laboratory: *Encyclopedia of the Mouse Genome*. URL: http://www.informatics.jax.org/
56. Gouy M, Gautier C, Attimonelli M, Lanave C, Di Paola G: ACNUC — a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. *Comput Appl Biosci* 1985, 1:167-172.
57. Higgins DG, Bleasby AJ, Fuchs R: CLUSTAL V: improved software for multiple sequence alignment. *Comput Appl Biosci* 1992, 8:189-191.
58. Li WH: Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol* 1993, 36:96-99.
59. Kimura M: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 1980, 16:111-120.
60. Moran PAP: Notes on continuous stochastic phenomena. *Biometrika* 1950, 37:17-23.
61. Cliff AD, Ord JK: *Spatial Autocorrelation*. London: Pion; 1981.
62. Press WH, Teukolsky SA, Vetterling WT: *Numerical Recipes in C*. Cambridge: Cambridge University Press; 1992.