

# Vanishing GC-rich isochores in mammalian genomes.

Laurent Duret<sup>¶</sup>, Marie Semon<sup>¶</sup>, Gwenaël Piganeau<sup>§</sup>, Dominique Mouchiroud<sup>¶</sup>, and Nicolas Galtier<sup>\*</sup>

<sup>¶</sup>Laboratoire de Biométrie, et Biologie Evolutive

UMR CNRS 5558 Université Claude Bernard Lyon 1

16 rue Raphaël Dubois

69622 Villeurbanne Cedex, France

<sup>§</sup> Centre for the Study of Evolution

School of Biological Sciences

Falmer, Brighton BN1 9QG, United Kingdom

<sup>\*</sup> Laboratoire Génome, Populations, Interactions

UMR CNRS 5000 Université Montpellier 2

Place Eugène Bataillon

34095 Montpellier Cedex 5, France

Correspondence should be addressed to L.D.

E-mail: [duret@biomserv.univ-lyon1.fr](mailto:duret@biomserv.univ-lyon1.fr)

Phone: (+33) 4 72 44 62 97 Fax: (+33) 4 78 89 27 19

*Keywords:* isochores; synonymous substitution; biased gene conversion; mutation; selection; CpG; SNP.

*Running head:* Vanishing GC-rich isochores in mammals

## **Abstract.**

To understand the origin and evolution of isochores - the peculiar spatial distribution of GC-content within mammalian genomes - we analyzed the synonymous substitution pattern in coding sequences from closely related species in primates, cetartiodactyls, and rodents. GC-rich genes are undergoing a large excess of GC→AT substitutions over AT→GC substitutions: GC-rich isochores are slowly disappearing from the genome of these three mammalian orders. This observation questions the conclusions of previous works that assumed that base composition was at equilibrium. Analysis of allele frequency in human polymorphism data, however, confirmed that in the GC-rich parts of the genome, GC alleles have a higher probability of fixation than AT alleles (probably because of biased gene conversion, not selection). This fixation bias appears not strong enough to overcome the large excess of GC→AT mutations. Thus, whatever was the evolutionary force (neutral or selective) at the origin of GC-rich isochores, this force is no longer effective in mammals. We propose a model based on the biased gene conversion hypothesis that accounts for the origin of GC-rich isochores in the ancestral amniote genome, and for their decline in present-day mammals.

## Introduction

The mammalian genome is heterogeneous with respect to base composition. In human, the GC-content of large genomic DNA fragments (>100 kb) ranges from 35% to 65% (Nekrutenko and Li 2000; Lander et al. 2001). Bernardi and colleagues first discovered this pattern from cesium chloride centrifugation experiments (Bernardi et al. 1985). Subsequent analyses showed that this variation of base composition affects all parts of the genomes: not only intergenic regions, but also introns and exons (and notably the third position of codons) (D'onofrio et al. 1991). Interestingly, this long range variability of base composition is correlated with various genomic features: GC-rich regions show a lower density of LINE and a higher density of Alu repeated elements, a higher level of methylation, a higher rate of recombination, and a much higher gene density (Jabbari and Bernardi 1998; Mouchiroud et al. 1991; Eyre-Walker 1993; Duret, Mouchiroud, and Gautier 1995; Smit 1999; Fullerton, Bernardo Carvalho, and Clark 2001; Lander et al. 2001). Thus, this long-range variability of GC-content clearly reflects a fundamental feature of genome organization.

Bernardi and colleagues proposed a model of mammalian genomes consisting in a mosaic of compositionally homogeneous regions - the so-called isochores (see Bernardi 2000). The complete sequence of the human genome modified a bit this picture: in general, the GC-content varies continuously (i.e. there is no clear boundaries between GC-poor and GC-rich regions), and isochores (notably, the GC-rich ones) are not as homogeneous as proposed initially (Nekrutenko and Li, 2000; Lander et al. 2001). However, a highly significant spatial autocorrelation of GC-content was found, with most of the structure detectable at a large (300 kb) scale (Lander et al. 2001), indicating indeed a *relative* local homogeneity in base

composition along mammalian chromosomes. Following Eyre-Walker and Hurst (2001), we will use the term isochores to denote these large regions of relatively homogeneous base composition.

Comparison of base composition in different vertebrate species indicates that the ancestral genome of tetrapodes was probably relatively homogeneous and AT-rich, and that the acquisition of GC-rich isochores occurred in the amniote lineage, after the split with amphibians, but before the divergence of mammals, birds and reptiles, i.e. about 310-350 Myrs ago (Bernardi and Bernardi 1990; Bernardi, Hughes, and Mouchiroud 1997; Hughes, Zelus, and Mouchiroud 1999). How did isochores originate? How were they maintained? Two models (one selectionist, and one neutralist) have been first invoked to explain the existence of isochores: (i) selection for a high GC-content in some regions of the genome (*e.g.* Bernardi et al. 1985; Bernardi 2000), (ii) variable mutational bias (VMB) along chromosomes (*e.g.* Wolfe, Sharp and Li 1989; Francino and Ochman 1999). Another neutral model, called biased gene conversion (BGC), originally proposed by Holmquist (1992), recently gained consideration (Galtier et al. 2001, Eyre-Walker and Hurst 2001). BGC is linked to the process of recombination. During meiosis, heteroduplex DNA segments are formed from the hybridization of one maternal and one paternal DNA strands. When a heterozygous site is involved in such a heteroduplex, this results in a DNA mismatch that may be repaired by one of the cellular DNA repair systems. This repair will thus result in the loss (i.e. conversion) of one of the two alleles. This process is said to be biased if the probability of conversion is not symmetric (*e.g.*, if GC-alleles convert AT-alleles more frequently than the reverse). Like selection, the impact of BGC on the fixation process depends on the population size (Nagylaki 1983). It also depends

on the rate of recombination (more precisely on the rate of formation of DNA heteroduplex), the size of DNA heteroduplex, and the bias in the repair of mismatches. Several authors proposed that GC-rich isochores might result from BGC, the fixation of GC alleles being favored in genomic regions of high recombination/conversion rate (Holmquist 1992, Eyre-Walker, 1993; Galtier et al. 2001; Eyre-Walker and Hurst 2001).

The major selectionist hypothesis, namely adaptation to homeothermy, was dismissed after isochores were discovered in several cold-blooded species (Hughes, Zelus, and Mouchiroud 1999). It should be stressed that the particular base composition of GC-rich isochores concerns not only coding regions, but also introns and intergenic regions, and notably pseudogenes (Francino and Ochman 1999). Therefore, if the acquisition of GC-rich isochores resulted from natural selection, this could not be due to a selective advantage at the RNA or protein level. In our opinion, neutralist models (VMB or BGC) appear more realistic.

An important progress toward the solution of this controversial issue has been brought by the analysis of polymorphism data. Indeed, the VMB model can be distinguished from the other two (selection or BGC) by studying the process of allele fixation. Under the VMB model, the probability of fixation is expected to be the same for alleles resulting from a AT→GC mutation (i.e. from A or T to G or C) as for alleles resulting from a GC→AT mutation. Hence, the pattern of GC↔AT substitution is expected to be identical to the pattern of mutation. Conversely, the BGC and the selectionist models predict a fixation bias in favor of GC alleles (i.e. GC alleles should have a higher probability of fixation than AT alleles). Analyses of human and mouse polymorphism data provided evidence for such a fixation bias,

leading to the conclusion that GC-rich isochores result from BGC or selection, but not from mutation bias (Eyre-Walker 1999, Smith and Eyre-Walker 2001). However, this conclusion was based on the assumption that the isochore structure was at equilibrium, *i.e.* that the GC-content of a specific genomic region is not varying in time. The stationarity appeared likely because the GC3 (GC-content at third codon positions) of orthologous genes taken in Primates, Cetartiodactyla, Lagomorpha and Carnivora are very close to each other (Mouchiroud and Bernardi 1993), which suggests that their GC-content has remained unchanged since the divergence of these different mammalian orders. Only the rodents showed a peculiar distribution of GC-content (Mouchiroud and Gautier 1988; Robinson, Gautier, and Mouchiroud 1997). An estimation of the ancestral GC3 of 27 genes sampled in various orders suggested that the ancestral mammalian isochore pattern was similar to the present-day human one (Galtier and Mouchiroud 1998), consistent with the hypothesis of a stationary evolution of isochores in non-rodent mammals.

The GC-content of a genomic fragment is at equilibrium when the number of GC→AT and AT→GC substitutions occurring in this fragment are equal. Interestingly, the analysis of the complete sequence of the human genome has shown that some sequences are not at equilibrium (Lander et al. 2001): in defective transposons (presumably free of any selective pressure), GC→AT substitutions are significantly more frequent than AT→GC substitutions. This excess of GC→AT substitutions in transposons is observed both in GC-rich and GC-poor isochores. This suggests that the GC-content of the human genome might be decreasing (*i.e.* not stationary), questioning the conclusions mentioned above. It should be noted, however, that the pattern of substitution in transposable elements might be different

from that of non-repetitive DNA, because repeated sequences can be subject to specific mutation patterns (Krickler, Drake, and Radman 1992).

The aim of this paper was therefore to directly determine **whether the GC-content of non-repetitive sequences is at equilibrium or not in mammalian genomes.** For this purpose, we analyzed orthologous gene sequences in closely related species from three mammalian orders (primates, rodents, cetartiodactyls). We show that in all these mammals, the GC-content **of genes located in** GC-rich isochores is decreasing. **These results indicate that** GC-rich isochores are slowly disappearing from mammalian genomes. The predictions of the three models for the origin of isochores are re-evaluated in this non-equilibrium situation, and confronted to human sequence polymorphism data. The issue of the origin and evolution of isochores is discussed in the context of a nonstationary process.

## **Material and methods**

We selected orthologous genes from closely related species in three mammalian orders, namely Primates, Rodentia and Cetartiodactyla. In each order, a triplet of species of the form `[[ingroup1, ingroup2], outgroup]` was defined. In Primates the ingroups were human and another Hominidae (chimpanzee, gorilla, or orangutan), the outgroup any Catarrhini (external but as close as possible to the two ingroups, generally *Papio*). In Rodentia, the triplet was generally `[[rat, mouse], hamster]`, but any triplet of species matching the `[[Rattus, Mus], non-Murinae Muridae]` pattern was considered acceptable. In Cetartiodactyla all comparisons involved `[[Caprinae, Bovinae], Suidae]`, excepting one gene for which the outgroup was a Cervidae. The coding sequences of every gene available in a triplet of species as defined above were

collected using the Hovergen database (release 40, May 2000) (Duret, Mouchiroud, and Gouy 1994), and phylogenetic trees were checked to retain only orthologous genes. Average synonymous substitution rates (computed with the method of Li 1993) in each data set are indicated in Table 1. The list of orthologous genes is available at <http://pbil.univ-lyon1.fr/datasets/Isochore2002/data.html>.

The GC-content expected at equilibrium at synonymous codon position ( $GC3eq$ ) was estimated using only the third position of quartet codons.  $GC3eq$  was computed by the ratio  $u / (u + v)$ , where  $u$  is the rate of AT→GC substitutions, and  $v$  the rate of GC→AT substitutions. We measured  $u$  (and  $v$ ) at the third position of quartet codons by dividing the number of AT→GC substitutions (GC→AT substitutions) observed in the two ingroups by the number of AT (GC) in the ancestral sequence inferred by parsimony (ambiguous bases were ignored).

### **Synonymous substitution patterns in mammalian genes**

*Excess of GC→AT synonymous substitutions in GC-rich genes.*

To determine whether the GC content of mammalian genes is at equilibrium, we analyzed orthologous coding sequences from different taxa: rodents (194 genes), cetartiodactyls (79 genes) and primates (55 genes) (Table 1). The synonymous substitution process was approached using the parsimony criterion. A substitution from nucleotide X to nucleotide Y was inferred when both the outgroup and one ingroup shared state X, but the other ingroup showed state Y. We analyzed all informative substitutions: 7058 substitutions in rodents, 817 substitutions in cetartiodactyls, and 197 substitutions in primates. Genes were classified in three groups depending on their GC3 (<57%, 57-75%, >75%) (Mouchiroud et al. 1991).

The results are displayed in Table 2. A large, significant excess of GC→AT over AT→GC changes is found for GC-rich genes in the three orders. Hence, the GC-content of GC-rich genes is decreasing, not only in rodents, but also in primates and cetartiodactyls.

In primates and cetartiodactyls, the present GC-content is very far from the value expected at equilibrium given the observed substitution pattern (Table 2). The bias is less strong in GC-median genes, and virtually no bias is found in GC-poor genes.

Is it possible that the excess of GC→AT substitution be due to recent translocations of GC-rich genes into a GC-poor genomic context ? To test this hypothesis, we examined in primates the synonymous substitution pattern according the the GC-content of the surrounding genomic region. For this purpose, we retrieved for each human gene the largest overlapping genomic sequence available in GenBank (163 kb in average). The 55 genes were split into three groups, according to the GC-content of their genomic context. As shown in table 3, the results of our analysis remain unchanged (compare with Table 2): there is a large excess of GC→AT substitutions, especially for the genes located in regions of high GC-content.

#### *Methodological artifact ?*

Is it possible that this result be due to a methodological artifact? The pattern [[ingroup1=X, ingroup2=Y], outgroup=X] is interpreted by the parsimony method as a single X→Y substitution in the ingroup2 lineage. Parsimony can fail in case of multiple substitutions (e.g. the above pattern can occur through two independent Y→X substitutions in the ingroup1 and outgroup lineages). The method remains unbiased as long as the base composition is balanced (i.e. X%=Y%): failures are

equiprobable for  $X \rightarrow Y$  and  $Y \rightarrow X$  actual changes. When base composition is unbalanced, however, the parsimony method tends to overestimate the proportion of substitutions from common bases to rare bases (Perna and Kocher 1995). In an X-rich sequence data set (i.e. when the rate of  $Y \rightarrow X$  substitution is higher than the rate of  $X \rightarrow Y$  substitution), the method will overestimate the proportion of  $X \rightarrow Y$  changes. The bias can be strong even for moderate levels of divergence between sequences (Eyre-Walker 1998, Galtier and Boursot 2000). To assess the extent of this bias in the current analysis, we made use of a restricted data set of 20 primate genes for which an additional species was available. This fourth species was used to measure the rate of error of the parsimony method. We selected quartets of the form [[ingroup1, ingroup2], ingroup3], outgroup] (generally [[human, chimpanzee], gorilla or orangutan], a non-Hominidae Catarrhini]). We examined all the substitutions previously inferred from the three-species analysis, and we counted as an error all the cases where the nucleotide in the ingroup3 was different from the outgroup (i.e. patterns different from [[X, Y], X], X]). Among the 78 substitutions analyzed, 9 (12%) were erroneously oriented (i.e. counted as  $X \rightarrow Y$  instead of  $Y \rightarrow X$ ). The error rate was similar for  $GC \rightarrow AT$  (13%, 7/55) and for  $AT \rightarrow GC$  (9%, 2/23) substitutions. In GC-rich genes (i.e.  $GC3 > 75\%$ ), there are four times as many  $GC \rightarrow AT$  as  $AT \rightarrow GC$  substitutions (12 vs. 3, a ratio very close to what was measured in the full three-species data set, see Table 2), and this ratio remains unchanged when erroneously orientated substitutions (13%, 2/15) are removed. In other words, the bias in the parsimony method is weak for this data set (presumably because evolutionary distances are very short), and cannot account for the four-fold excess of  $GC \rightarrow AT$  substitutions over  $AT \rightarrow GC$  substitutions observed in GC-rich genes from primates.

### *Hypermutable CpG dinucleotides*

It is well known that in mammals, CpG dinucleotides are hotspots of C→T mutations because of the deamination of methylated cytosines (Coulondre et al. 1978; Bird 1980). To determine whether these mutations could account for the excess of GC→AT substitutions, we removed from our analyses all GC→AT substitutions that occurred within a CpG dinucleotide in the reconstructed ancestral sequence. The substitutions at CpG dinucleotides correspond to 14% (rodents), 28% (primates) and 31% (cetartiodactyls) of GC→AT substitutions in GC-rich genes. But in all taxa, the number of GC→AT substitutions remained higher than the number of AT→GC substitutions after CpG were removed (Table 2). Hence, the decrease in GC-content of GC-rich genes is not due solely to the hypermutability of CpG dinucleotides.

### **Fixation bias in favor of GC alleles**

An asymmetric pattern of GC/AT polymorphisms in human has been reported by Eyre-Walker (1999, MHC sequence data) and Smith and Eyre-Walker (2001, SNP data): a higher number of GC→AT than AT→GC polymorphic sites was found. At that time, it was admitted that the GC content was at equilibrium (i.e. that the number of GC→AT substitution was equal to the number of AT→GC substitutions). Hence, the difference between the mutation pattern (revealed by polymorphism data) and the substitution pattern (inferred from the hypothesis of stationarity) was interpreted as the result of a biased fixation process, namely either

selection or BGC, favoring GC alleles (Eyre-Walker 1999; Smith and Eyre-Walker 2001). However, our analysis of synonymous substitution patterns (see above) shows that the hypothesis of stationarity does not hold: the GC-content is not at equilibrium. The present results, therefore, question the existence of a fixation bias favoring GC alleles - a major argument against the mutation bias hypothesis.

We re-analyzed the SNP data set (Cargill et al. 1999) used by Smith and Eyre-Walker (2001) to check further the mutation bias hypothesis. This data set includes SNP's sampled in 114 human chromosomes and in one chimpanzee. Instead of analyzing the number of polymorphic sites of each category (either GC→AT or AT→GC), we focused on the distribution of the frequencies at which these alleles occur in the sample. We chose this approach because it is not affected by the fact that the GC-content is not at equilibrium: the fixation dynamics of a mutation, given that it has occurred, is independent of the stationary status of base composition. Hence, in the absence of any fixation bias, similar distributions of allele frequency are expected for GC→AT or AT→GC polymorphisms, whatever the mutation process. Conversely, if GC alleles had a higher probability of fixation than AT-alleles (because of selection or BGC), the allele frequency distribution of AT→GC polymorphisms should be shifted to the right (advantageous mutations segregate at higher frequencies, on average).

Polymorphisms were oriented using the chimpanzee as an outgroup. Noncoding and synonymous polymorphisms were pooled. Non-synonymous polymorphisms were excluded from this analysis as they undergo selection at the protein level. The total number of SNP's analyzed was 410. Cargill et al. (1999) did not indicate exact allele frequencies, but classified polymorphic sites in four classes of

frequency: low (<5%), median, (5-15%), high (15-50%) or very high (>50%). The distributions of allele frequencies of GC→AT and AT→GC polymorphisms in GC-rich (GC3>75%, 169 SNP's), GC-median (96 SNP's) and GC-poor (GC3<57%, 145 SNP's) genes are shown (Fig. 1). A visual inspection suggests that AT→GC mutations segregate at higher frequencies than GC→AT ones (i.e. that GC alleles have a higher probability of fixation than AT alleles), especially in GC-rich genes.

A likelihood-ratio test was performed to assess the significance of this apparent discrepancy between the two observed distributions. Two multinomial models were fitted to the allele frequency data. The Joint Multinomial model (null hypothesis) assumes that the allelic frequencies of AT→GC and GC→AT polymorphic sites are drawn from a unique multinomial distribution (three parameters). The Full Multinomial model allows one free parameter for every frequency class of AT→GC and GC→AT polymorphisms (6 parameters): distinct distributions are assumed for the two categories of SNP's. The likelihood was calculated under the two models according to the multinomial formula. Optimal parameter values are obvious : the expected proportions of each class of allele frequency are taken as the observed ones, separating (Full Multinomial model) or pooling (Joint Multinomial model) the distributions of AT→GC and GC→AT polymorphisms. Twice the difference in log-likelihood between the two models follows a  $\chi^2$  distribution (three degrees of freedom) under the null hypothesis of a common distribution. Results are displayed in Table 4. No difference between the two distributions was detected in GC-poor genes, but AT→GC polymorphisms appeared to segregate at significantly higher alleles frequencies, on average, than GC→AT polymorphisms in both GC-median and GC-rich genes.

A population genetics model was also fitted to the allele frequency data set. This model assumes that a given fixation bias (BGC or selection coefficient  $w$ ) favouring GC over AT alleles applies independently at every site in a panmictic population of effective size  $2N$ , at selection(BGC)/mutation/drift equilibrium. The expected probability density of allele frequencies under these assumptions has been derived from diffusion equations (*e.g.* Sawyer & Hartl 1992). It depends on a single parameter, namely the  $4.N.w$  product. The density was numerically integrated over the relevant frequency boundaries (0-5%, 5-15%, ...) to provide the required discrete expected distribution. Using this model, one can test the neutral hypothesis by comparing the maximum likelihood (leaving  $w$  free to vary) to the likelihood obtained by setting  $w=0$  (null, neutral hypothesis). The results were essentially consistent with the multinomial tests: no departure from neutrality was detected for the GC-poor data set, but neutrality was rejected for GC-median genes ( $p$ -value<1%). As far as the GC-rich data set was concerned, the population genetics model fitted badly to the observed distribution of allele frequency. Indeed, a comparison between this model and the Full Multinomial model led to rejection of the hypothesis of a panmictic population at equilibrium (not shown), making the test of neutrality impracticable. This peculiar feature of GC-rich genes is a bit surprising since the population history is the same for every part of the genome. Explaining this result appears challenging, but lies outside the scope of this paper.

These results confirm, on a larger gene data set, a previous analysis of allele frequency distribution in human and mouse MHC genes (Eyre-Walker 1999). Eyre-Walker's finding - a fixation bias favouring GC over AT alleles - appears valid despite the nonstationarity of the substitution process. This bias, however, appears restricted

to GC-median and GC-rich regions. Note that an alternative hypothesis can be proposed to explain the difference in frequency distribution between AT and GC alleles: the excess of AT alleles segregating at low frequency could be due to a recent increase in GC→AT mutation rate. This scenario, however, seems very unlikely because the putative change in mutation pattern must have occurred very recently (less than  $\sim 4.N$  generations ago), and simultaneously in the human and mouse lineages.

## **Discussion**

### *Erosion of GC-rich isochores*

The analysis of substitution pattern in defective DNA transposons had revealed an excess of GC→AT substitutions over AT→GC substitutions, suggesting that the human genome was not at equilibrium (Lander et al. 2001). We show here that this pattern of substitution is observed not only in repetitive DNA but also in unique sequence: the synonymous GC-content of genes located in GC-rich isochores is decreasing in the mammalian genomes, presumably as a consequence of an AT-biased mutation process.

We did not analyze non-coding sequences because there are presently not enough species for which such data are available. However, numerous works have shown that in mammals, there is a strong correlation between the GC content at synonymous codon positions and the GC content of the genomic region in which a gene is located (e.g. Aïssani et al 1991; Clay et al. 1996). As shown in figure 2, this correlation is confirmed here with a much larger data set. Thus, whatever the process that drives the evolution of isochores, this process acts not only in non-coding regions but also at

synonymous codon positions. It should also be noted that the subset of 55 human genes that we have used to infer synonymous substitutions in primates is clearly representative of the whole data set (Fig. 2). This suggests that the pattern of synonymous substitution that we measured directly reflects the evolution of isochores. This conclusion is further supported by the fact that the same pattern is observed in transposable elements, that are essentially located in introns or intergenic regions (Lander et al. 2001). In other words, these different results indicate that GC-rich isochores are disappearing from the genome of these three mammalian orders.

It had been shown previously that the GC-content of rodent genomes was less heterogeneous than in primates (i.e., GC-rich genes are less GC-rich in rodents than in primates, and conversely for GC-poor genes) (Mouchiroud and Gautier 1988), and that this change in GC-content occurred in the rodent lineage (Robinson, Gautier, and Mouchiroud 1997; Galtier and Mouchiroud 1998). However, in all other taxa analyzed (Primates, Lagomorpha, Cetartiodactyla, Carnivora) the GC3 of orthologous genes are very close to each other (Mouchiroud and Bernardi 1993). Hence, it was generally thought that the GC-content had remained unchanged in these mammalian orders since the time of their divergence, 80-100 Myrs ago. Our results show that the decrease of GC-rich isochores, previously reported in rodents, is actually an ongoing process of rodent, primate, and cetartiodactyl genomes. Their presently similar GC-content simply reflects the fact that they have undergone similar substitution patterns, and/or that there was not enough time to accumulate enough substitutions to significantly affect their GC-content. Note that the time necessary to reach equilibrium can be very long. For example, consider a GC-rich sequence (say 85% GC) subject to a pattern of

substitution leading to equilibrium GC-content of 50% (i.e. similar to the substitution pattern observed in primates or cetartiodactyls GC-rich genes). As illustrated figure 3, after 0.33 substitution per site (which corresponds to the average evolutionary distance measured at synonymous sites in human and bovine orthologous genes), the GC-content is still 77%. It is therefore possible that the decline in GC-content presently observed in mammalian GC-rich isochores began a long time ago, before the divergence between the different mammalian orders.

**The second major conclusion of our study**, based on the analysis of allele frequency distribution, is that a directional evolutionary force is biasing the fixation process of GC/AT polymorphism toward GC in GC-rich, but not in GC-poor regions of the genome. However, this force is not strong enough to maintain GC-rich isochores, since GC→AT substitutions are more frequent than AT→GC substitutions. In other words, in GC-rich genes, GC alleles have a higher probability of fixation than AT alleles, but this bias is not sufficient to overcome the large excess of GC→AT mutations.

#### *Fixation bias favoring GC alleles: biased gene conversion*

The fixation bias favoring GC alleles is likely to be a result of BGC, not selection. Indeed, different observations support the BGC model: experiments in mammalian cells have shown that the repair of DNA mismatches is biased in favor of GC; genes that frequently undergo conversions are GC-rich; and there is an overall correlation between GC-content and recombination rate (reviewed in Galtier et al. 2001). Eyre-Walker and Hurst (2001) also concluded that BGC was the most likely scenario for

the origin of isochores. They raised, however, an important point apparently not consistent with this model: a positive correlation between substitution rate and GC-content has been reported (Bielawski, Dunn, and Yang 2000), whereas the BGC model (like selection) predicts that the correlation should be negative. However, this prediction holds only if the GC-content is at equilibrium. On the contrary, when the GC-content is decreasing, one expects a positive correlation between substitution rate and GC-content (Piganeau et al. 2002). In conclusion, the data are consistent with the hypothesis that the origin of GC-rich isochore is due to BGC, but that this process is no longer effective to maintain their GC-content.

#### *Origin and evolution of GC-rich isochores*

The origin and evolution of isochores becomes even more mysterious in this non-equilibrium context. As mentioned in the introduction, the acquisition of GC-rich isochores occurred in the amniote lineage, about 310-350 Myrs ago. However, the substitution process measured from the comparison of closely related species indicates that GC-rich isochores are disappearing from mammalian genomes. The fact that similar patterns are found in three different mammalian orders suggests that this erosion may have started more than 80 Myrs ago (see Fig. 4).

How to explain the rapid acquisition of GC-rich isochores in the amniote ancestral genome, and their slow decay in mammals? A part of the answer probably lies in the asymmetry between the processes of GC-increase and GC-decrease. The data reported in this study are consistent with the existence of an AT-biased mutation process ( $u < v$ , where  $u$  is the AT→GC mutation rate and  $v$  the GC→AT mutation rate), vs. a GC-biased fixation process ( $w$ ). In regions of the genome/periods of time of

negligible fixation bias ( $4.N.w \ll 1$ ), the GC-content therefore approaches the mutational equilibrium through mutation/drift, at neutral rate  $\theta.v(1-\theta).u$ , where  $\theta$  is the current GC-content. When the fixation bias is strong ( $4.N.w \gg 1$ ), however, advantaged GC mutations fix with a probability  $2.w$  (vs.  $1/2N$  in absence of fixation bias), so that the GC-content will increase at a rate close to  $4.(1-\theta).u.N.w$ , possibly much higher than the neutral rate. Therefore, short and/or localized episodes of GC-increase can significantly contribute to the total amount of G and C, even if they concern a small fraction of the genome/time scale. In the light of these notions, we now propose several scenarios about the origin of GC-rich isochores, accounting for the newly reported non-equilibrium situation.

*Model 1: the "GC-factory" hypothesis*

This scenario invokes a spatial, but not a temporal heterogeneity of BGC coefficient. Imagine that a small fraction of the genome is undergoing a strong fixation bias toward GC ("GC-factory"), while the major part of the genome is evolving neutrally (*i.e.* without selection nor BGC). Now assume that these two fractions communicate in some way (translocations or duplications). Sequences incoming a GC-factory would undergo a rapid increase of GC-content. Such GC-rich sequences would then be released in the major component, and form GC-rich isochores. Now when randomly sampling in the genome one would essentially sample genes from the major, neutrally evolving component, and observe a global decrease of GC-content. An example of such a process is provided by the *Fxy* gene in mouse. This gene used to undergo a rapid increase of GC-content after it was translocated in the pseudoautosomal region (Perry and Ashworth 1999).

In this hypothesis, the genome would be at a mutation-selection-migration equilibrium, where migration refers to movements of genes between regions of the genome. Although appealing, this scenario appears incompatible with the observed conservation of isochores among mammalian orders. The GC3 of orthologous genes sampled in Primates, Cetartiodactyla, Carnivora and Lagomorpha are highly correlated (Mouchiroud and Bernardi 1993). Such a correlation is not expected if genes were randomly moving to/from GC-factories independently in distinct lineages.

*Model 2: a unique origin of GC-rich isochores*

The hypothetical scenario that we propose for the origin of GC-rich isochores is the following. In the ancestral amniote, a change occurred in a DNA repair system, resulting in the preferential repair of AT:GC mismatches into GC, and hence to a biased gene conversion favoring the fixation of GC alleles. As suggested by Fryxell and Zuckerkandl (2000), this biased repair process might be an adaptation to the high rate of mutation of cytosine in heavily methylated genomes. Like in many present-day genomes, there was probably a high variability of recombination rate along the ancestral genome of amniotes. Thus, GC-rich isochores would have resulted from BGC in region of high recombination of this ancestral amniote genome. Note that the repair of G:T mismatches in *Xenopus* is unbiased (Varlet, Radman, and Brooks 1990) and, consistent with that model, *Xenopus* does not have GC-rich isochores. The enrichment in GC-content in regions of high recombination rate may have continued independently in different amniote lineages, after the split between mammals, birds and reptiles. Then, BGC ceased to be effective - at least in mammals, and the GC-content of GC-rich isochores began to decrease (Fig. 4).

Why did BGC ceased to be effective ? As mentioned previously, BGC significantly affects the fixation of GC alleles only if the product  $4.N.w$  is greater than one (where  $N$  is the population size and  $w$  the BGC coefficient). We propose two possible explanations for the decline of GC-rich isochores.

First, the evolution of GC-content might be linked to variations in population size. For a given distribution of  $w$  among regions of the genome, the fraction of the genome undergoing significant fixation bias (and eventually becoming GC-rich) increases with  $N$ , as illustrated by Figure 5. GC-rich isochores might therefore have appeared in an ancestral species of larger than today population size. After a subsequent decrease of population size, most of the genome would be under too low a fixation bias ( $4.N.w$ ) for GC-rich isochores to be maintained.

The second possible explanation is that the evolution of GC-content might be linked to variations in recombination (and, therefore, BGC) rate. Such variations could have occurred as a result of chromosome rearrangements. The GC-content of autosomal chromosomes is negatively correlated to chromosome size in human ( $R^2=0.32$ ,  $p=0.006$ , data from Venter et al. 2001). This is probably because the per nucleotide recombination rate is higher in small chromosomes (*e.g.* Kaback 1996; Lander et al. 2001). Imagine that the ancestral amniote genome used to have some very small chromosomes. Their GC-content would have increased rapidly as a consequence of their high recombination rate. These pieces of DNA would later have formed GC-rich isochores when fused with standard chromosomes. This scenario was suggested by the observation of such GC-rich microchromosomes in the genome of birds (Kadi et al. 1993), in which events of fusion between micro- and macrochromosomes have been reported (Shetty, Griffin, and Graves 1999). It would also account for the GC-

richness of telomeric regions: the random fusion of many GC-rich microchromosomes and few GC-poor macrochromosomes should result in a high proportion of GC-rich-ending fused chromosomes.

These mechanisms are possibly simplistic, and not mutually exclusive. We leave them as open hypothesis for future work in this issue. Note, however, that the hypothesis of a decrease of GC-content after a unique origin of GC-rich isochores in an ancestral vertebrate is concordant with the observation of a faster decay in the most rapidly evolving groups, namely rodents *vs.* other mammals (Li, Tanimura and Sharp 1987), and snakes *vs.* other reptiles (S. Hughes, personal communication). To conclude, it should be stressed that whatever the evolutionary force (neutral or selective) that was at the origin of GC-rich isochores, this force is no longer effective to maintain them in mammalian genomes.

### **Acknowledgments**

We thank Nick Smith and Adam Eyre-Walker for their helpful comments. This work was supported by the CNRS (Centre National de la Recherche Scientifique).

### **Literature cited.**

AISSANI, B., G. D'ONOFRIO, D. MOUCHIROUD, K. GARDINER, C. GAUTIER, and G. BERNARDI. 1991. The compositional properties of human genes. *J. Mol. Evol.* **32**:493-503.

BERNARDI, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* **241**: 3-17.

BERNARDI, G., and G. BERNARDI. 1990. Compositional Transitions in the

Nuclear Genomes of Cold-Blooded Vertebrates. *J. Mol. Evol.* **31**:282-293.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS, G. CUNY, M. MEUNIER-ROTIVAL, and F. RODIER. 1985. The mosaic genome of warm-blooded vertebrates. *Science* **228**: 953-958.

BERNARDI, G., S. HUGHES, and D. MOUCHIROUD. 1997. The major compositional transitions in the vertebrate genome. *J. Mol. Evol.* **44**:S44-S51.

BIELAWSKI, J. P., K. A. DUNN, and Z. YANG. 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. *Genetics*. **156**:1299-308.

BIRD, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**:1499-504.

CARGILL, M., D. ALTSHULER, J. IRELAND et al. (18 co-authors) 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**: 231-238.

CLAY, O., S. CACCIO, S. ZOUBAK, D. MOUCHIROUD, and G. BERNARDI. 1996. Human coding and noncoding DNA: compositional correlations. *Mol. Phylogenet. Evol.* **5**:2-12.

COULONDRE, C., J. H. MILLER, P. J. FARABAUGH, and W. GILBERT. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature*. **274**:775-80.

D'ONOFRIO, G., D. MOUCHIROUD, B. AISSANI, C. GAUTIER, and G. BERNARDI. 1991. Correlation between the compositional properties of human genes, codon usage, and amino-acid composition of proteins. *J. Mol. Evol.* **32**:504-510.

- DURET, L., D. MOUCHIROUD, and C. GAUTIER. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* **40**: 308-317.
- DURET, L., D. MOUCHIROUD, and M. GOUY. 1994. HOVERGEN - a database of homologous vertebrate genes. *Nucleic Acid Res.* **22**: 2360-2365.
- EYRE-WALKER, A. 1993. Recombination and Mammalian Genome Evolution. *Proc. R. Soc. Lond. B - Biol.Sci.* **252**:237-243.
- EYRE-WALKER, A. 1998. Problems with parsimony in sequences of biased base composition. *J. Mol. Evol.* **47**:686-90.
- 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., and L. D. HURST. 2001. The evolution of isochores. *Nat. Rev. Genet.* **2**:549-55.
- FRANCINO, M. P., and OCHMAN, H. 1999. Isochores result from mutation, not selection. *Nature* **400**: 30-31.
- FRYXELL, K. J., and E. ZUCKERKANDL. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* **17**:1371-183.
- FULLERTON, S. M., A. BERNARDO CARVALHO, and A. G. CLARK. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* **18**:1139-142.
- GALTIER, N., and D. MOUCHIROUD. 1998. Evolution of isochores in mammals: a human-like ancestral pattern. *Genetics* **150**: 1577-1584.

GALTIER, N., and P. BOURSOT. 2000. A new method to locate changes in a tree reveals unequal polymorphism vs. divergence pattern in mouse mitochondrial DNA. *J. Mol. Evol.* **50**: 224-231.

GALTIER, N., G. PIGANEAU, D. MOUCHIROUD, and L. DURET. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159**: 907-911.

HOLMQUIST, G. P. 1992. Chromosome bands, their chromatine flavours, and their functional features. *Am. J. Hum. Genet.* **51**: 17-37.

HUGHES S., D. ZELUS, and D. MOUCHIROUD. 1999. Warm-blooded isochore structure in Nile crocodile and turtle. *Mol. Biol. Evol.* **16**: 1521-1527.

JABBARI, K., and G. BERNARDI. 1998. CpG doublets, CpG islands and Alu repeat elements in long human DNA sequences from different isochores families. *Gene* **224**: 123-128.

KABACK, D. B. 1996. Chromosome-size dependent control of meiotic recombination in humans. *Nat. Genet.* **13**:20-1.

KADI, F., D. MOUCHIROUD, G. SABEUR, and G. BERNARDI. 1993. The compositional patterns of the avian genomes and their evolutionary implications. *J. Mol. Evol.* **37**:544-551.

KRICKER, M. C., J. W. DRAKE, and M. RADMAN. 1992. Duplication-targeted DNA methylation and mutagenesis in the evolution of eukaryotic chromosomes. *Proc. Natl. Acad. Sci. USA.* **89**:1075-1079.

LANDER, E. S., L. M. LINTON, B. BIRREN et al. (250 co-authors) 2001. Initial sequencing and analysis of the human genome. *Nature.* **409**:860-921.

- LI, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* **36**:96-99.
- LI, W.-H., M. TANIMURA, and P. M. SHARP. 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* **25**: 330-342.
- MOUCHIROUD, D., and C. GAUTIER. 1988. High Codon-Usage Changes in Mammalian Genes. *Mol. Biol. Evol.* **5**:192-194.
- MOUCHIROUD, D., and G. BERNARDI. 1993. Compositional properties of coding sequences and mammalian phylogeny. *J. Mol. Evol.* **37**: 109-116.
- MOUCHIROUD, D., G. D'ONOFRIO, B. AISSANI, G. MACAYA, C. GAUTIER, and G. BERNARDI. 1991. The distribution of genes in the human genome. *Gene*. **100**:181-187.
- NAGYLAKI, T. 1983. Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. U. S. A.* **80**:6278-681.
- NEKRUTENKO, A., and W. H. LI. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome. Res.* **10**:1986-195.
- PERNA, N. T., and T. D. KOCHER. 1995. Unequal base frequencies and the estimation of substitution rate. *Mol. Biol. Evol.* **12**: 359-361.
- PERRY, J., and A. ASHWORTH. 1999. Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* **9**: 987-989.
- PIGANEAU, G., D. MOUCHIROUD, L. DURET, and C. GAUTIER. 2002. Expected Relationship Between the Silent Substitution Rate and the GC Content: Implications for the Evolution of Isochores. *J. Mol. Evol.* **54**:129-33.

- ROBINSON, M., C. GAUTIER, and D. MOUCHIROUD. 1997. Evolution of isochores in rodents. *Mol. Biol. Evol.* **14**: 823-828.
- SAWYER, S. A., and D. L. HARTL. 1992. Population genetic of polymorphism and divergence. *Genetics* **132**: 1161-1176.
- SHETTY, S., D. K. GRIFFIN, and J. A. GRAVES. 1999. Comparative painting reveals strong chromosome homology over 80 million years of bird evolution. *Chromosome. Res.* **7**:289-95.
- SMIT, A. F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**:657-663.
- SMITH, N.G.C., and A. EYRE-WALKER. 2001. Synonymous codon bias is not caused by mutation bias in human. *Mol. Biol. Evol.* **18**: 982-986.
- VARLET, I., M. RADMAN, and P. BROOKS. 1990. DNA mismatch repair in *Xenopus* egg extracts: repair efficiency and DNA repair synthesis for all single base-pair mismatches. *Proc. Natl. Acad. Sci. U. S. A.* **87**:7883-787.
- VENTER, J. C., M. D. ADAMS, E. W. MYERS et al. (274 co-authors) 2001. The sequence of the human genome. *Science*. **291**:1304-151.
- WOLFE, K. H., P. M. SHARP, and W. -H. LI. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337**: 283-285.

Table 1: Average synonymous substitution rates (Ks) in the data sets of orthologous genes used in this study

Order	Rodentia	Cetartiodactyla	Primates
Typical triplet	(hamster, (mouse, rat))	(pig, (cow, sheep))	(papio, (human, chimpanzee))
Number of genes	194	79	55
Ks ingroup1/ingroup2	0.19	0.08	0.02
Ks ingroup/outgroup	0.29	0.11	0.07

Table 2. Pattern of synonymous substitutions (AT↔GC) in mammalian genes of different GC content.

Order	GC3	All codons			Quartet codons		No CpG		
	class	GC→AT <sup>a</sup>	AT→GC <sup>a</sup>	ratio <sup>b</sup>	GC→AT <sup>a</sup>	AT→GC <sup>a</sup>	GC→AT <sup>c</sup>	GC3q <sup>d</sup>	GC3eq <sup>e</sup>
	<57%	934	1378	0.7	374	585	327	46%	58%
Rodentia	57-75%	2143	1923	1.1	910	904	806	64%	61%
	>75%	431	249	1.7**	182	128	157	78%	69%
	<57%	155	161	1	65	65	53	44%	46%
Cetartio-	57-75%	185	104	1.8**	97	45	85	64%	48%
dactyla	>75%	162	50	3.2**	88	20	64	83%	56%
	<57%	40	25	1.6*	18	11	13	41%	34%
Primates	57-75%	53	32	1.7*	31	23	24	63%	54%
	>75%	37	10	3.7**	18	4	13	80%	43%

Data sets were split into three groups of genes of low, medium and high GC3 content.

<sup>a</sup> Total number of synonymous substitutions counted in the branches leading to the two ingroups (see table 1 for details). <sup>b</sup> ratio of GC→AT over AT→GC substitutions.

Significance was assessed by a binomial test (probability of observing that ratio or higher assuming equal expected number of the two kinds of changes). \*  $p < 0.05$ ; \*\*

$p < 0.01$ . <sup>c</sup> Number of GC→AT substitutions at the third position of quartet codons, excluding all positions corresponding to a CpG dinucleotide in the ancestral sequence.

<sup>d</sup> GC3q : average GC-content the third position of quartet codons. <sup>e</sup> GC3eq : GC-content expected at equilibrium at the third position of quartet codons (see text). Note that given the low number of substitutions analyzed in primates GC-rich genes, this estimate might not be very accurate for that subset.

Table 3. Pattern of synonymous substitutions (AT↔GC) in primates genes located in different isochores contexts.

GC class <sup>a</sup>	GC→AT <sup>b</sup>	AT→GC <sup>b</sup>	Ratio <sup>c</sup>
<43%	60	39	1.5
43-49%	40	18	2.2
>49%	30	10	3.0

<sup>a</sup> Genes were split into three groups according to the GC content of the GenBank genomic sequence containing them (average length of the genomic sequence = 163 kb ± 50 kb). The limits of GC-content correspond to the 33% lowest and highest GC-content in the entire data set of 1892 complete human genes (see legend Fig. 2) <sup>b</sup> Total number of synonymous substitutions counted in the branches leading to the two ingroups. <sup>c</sup> ratio of GC→AT over AT→GC substitutions.

Table 4. Maximum-likelihood analysis of the allele frequency distribution of GC→AT and AT→GC polymorphisms.

Multinomial				
Data set	Model	$np^a$	$\ln(L)^b$	LRT <sup>c</sup>
	Full	6	-11.93	
GC-poor	Joint	3	-13.65	NS
	Full	6	-10.67	
GC-median	Joint	3	-16.12	*
	Full	6	-12.44	
GC-rich	Joint	3	-18.59	**

<sup>a</sup> number of free parameters

<sup>b</sup> Maximum log-likelihood

<sup>c</sup> Likelihood ratio test. NS: not significant. \*  $p < 0.05$ ; \*\*  $p < 0.01$ .

### **Legend to figures.**

Figure 1: Distribution of allele frequencies of 410 human SNP's in three classes of gene GC-content. SNP's were classified in four classes depending on the frequency of the mutant allele. The proportions of allele frequency classes are represented separately for GC→AT (white bars) and AT→GC (grey bars) polymorphic sites.

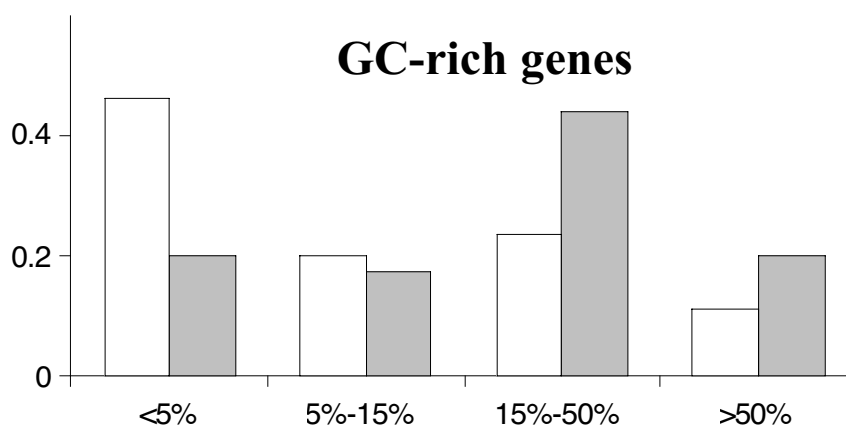
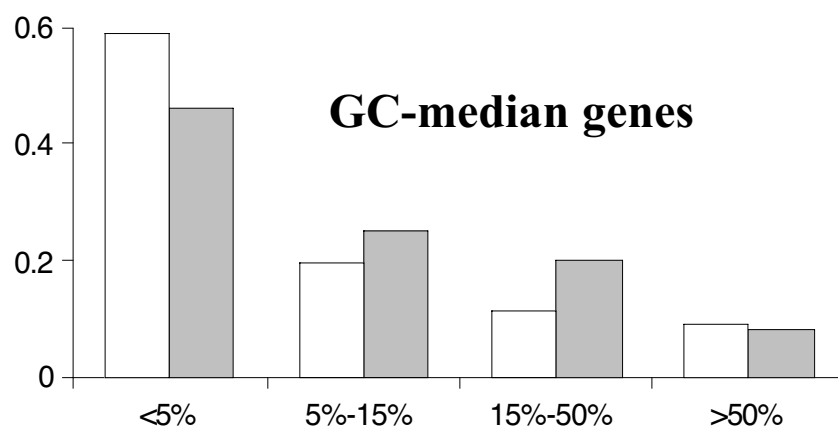
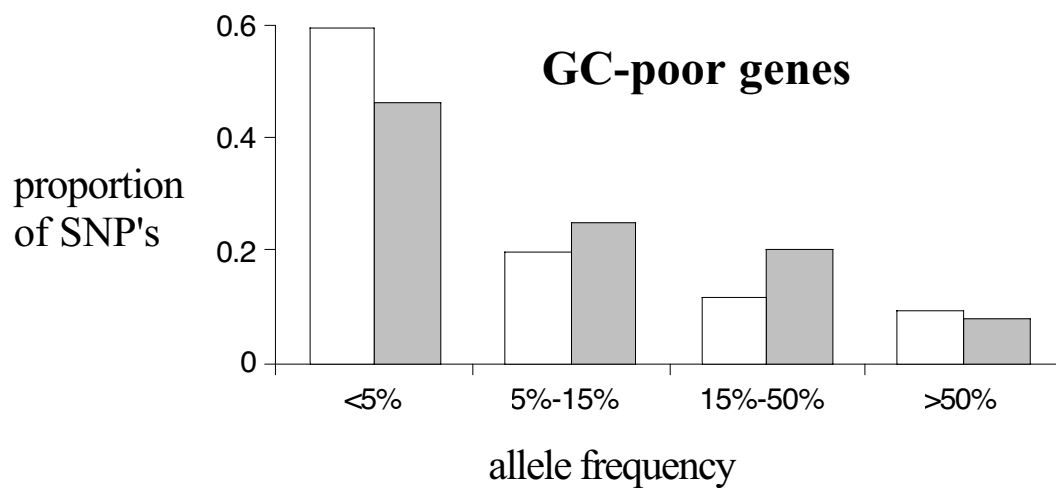
Figure 2: Correlation between the GC content at third codon positions (GC3) of human protein-coding genes and the GC content of the genomic region (> 100 kb) in which they are located. We have selected in GenBank (release 128 February 2002) all complete human genes (excluding those with less than 100 codons) annotated in genomic sequences larger than 100 kb. This data set contains 1892 genes (about  $10^6$  codons), from 874 sequences (total = 137 Mb). Correlation coefficient:  $R^2=0.43$ . The subset of genes used in this work to infer synonymous substitution patterns is indicated by black points.

Figure 3: Evolution of the GC-content of a sequence (initially at 85% GC) and subject to a substitution pattern leading to an equilibrium GC-content of 50% (with a ratio transition/transversion = 2).

Figure 4: Phylogenetic distribution of GC-rich isochores in vertebrates. Genomes of fishes and amphibians are weakly heterogeneous in base composition. Genomes of mammals birds and reptiles are generally highly heterogeneous and contain GC-rich isochores (Hughes, Zelus, and Mouchiroud 1999).

Figure 5: Hypothetic distribution of the selection (or BGC) coefficient  $w$  within mammalian genomes, and its implication with respect to isochore evolution.

A constant-in-time, lognormal distribution of  $w$  is assumed for convenience. With a low effective population size (here,  $10^4$ ), the fraction of the genome under significant effective fixation bias ( $4.N.w > 1$ ) is minute (black). A larger ancestral effective population size ( $10^5$ ) would have resulted in a GC-increase of a non-negligible part of the genome (gray).



□ GC→AT

■ AT→GC

Figure 1

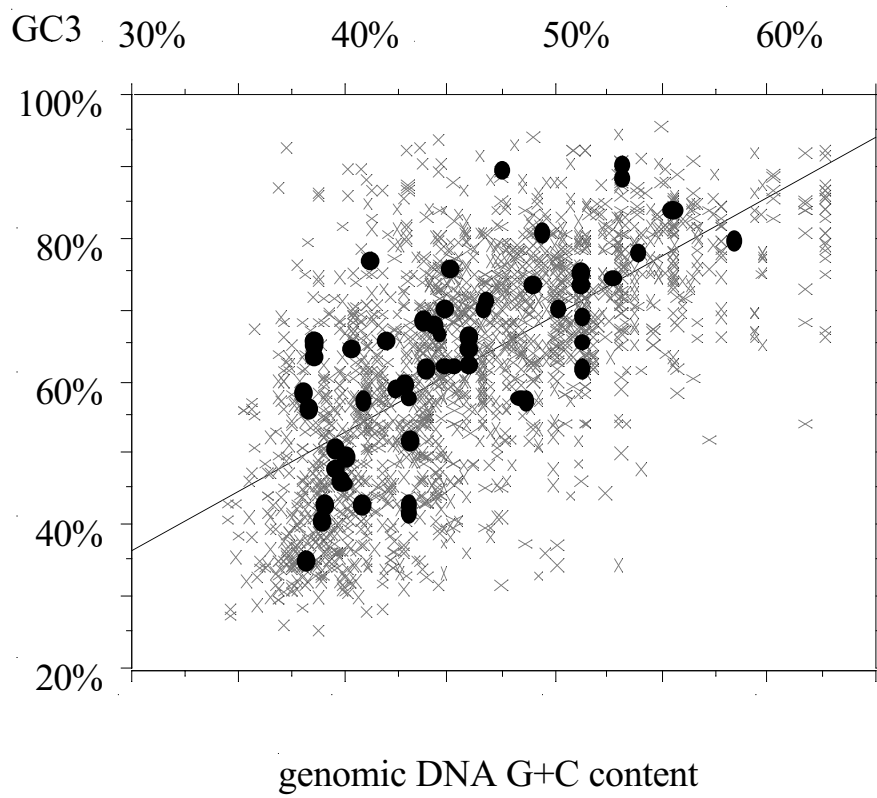


Figure 2

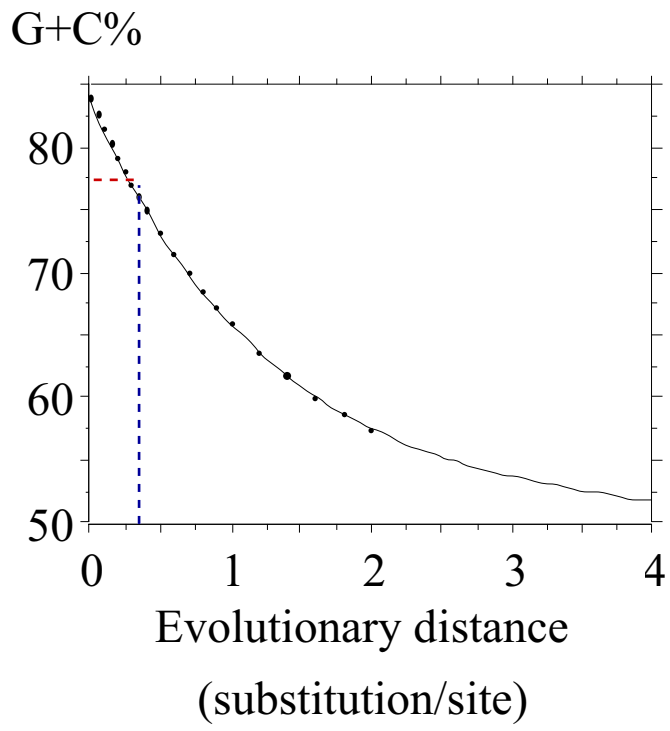


Figure 3

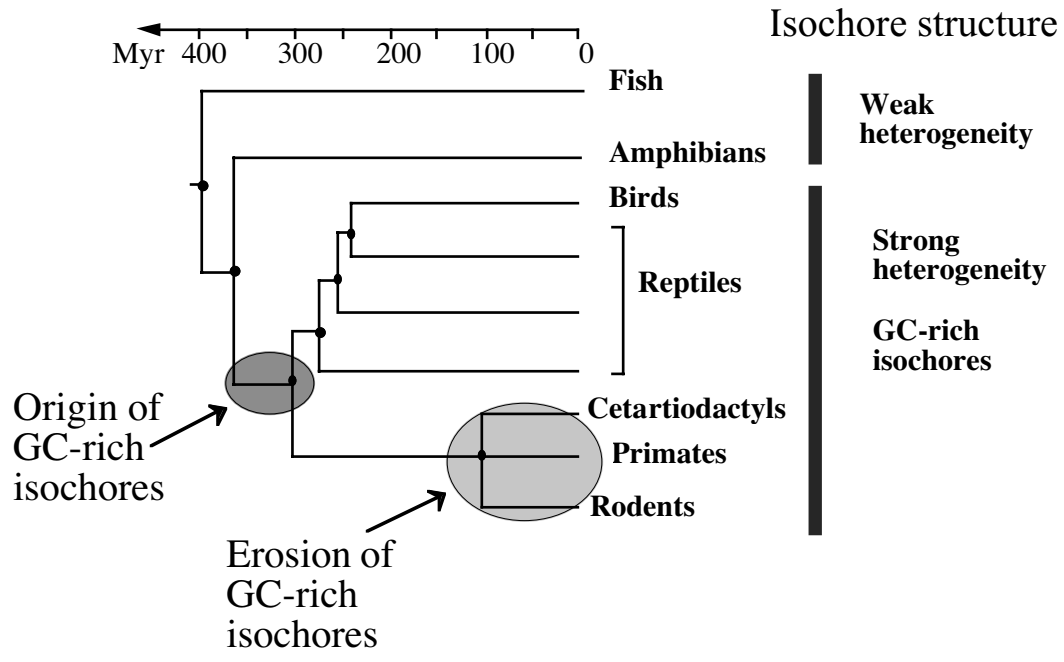


Figure 4

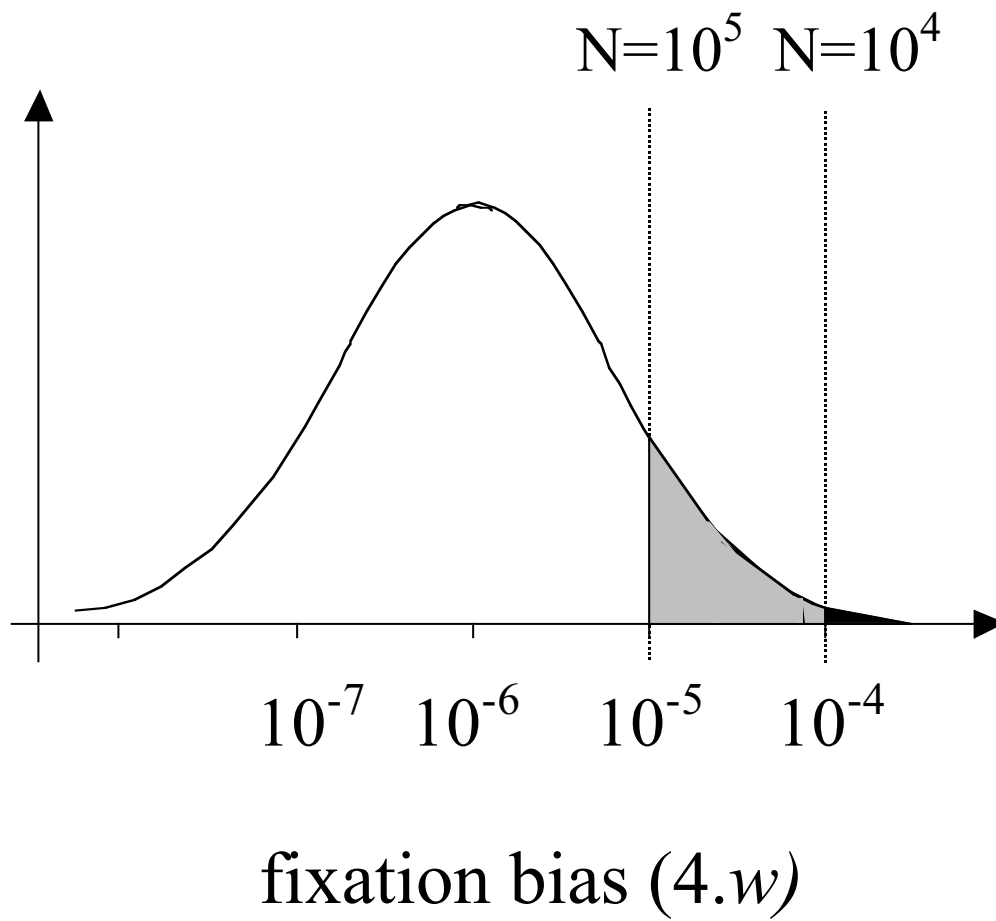


Figure 5