J Mol Evol (2006) 62:745–752
DOI: 10.1007/s00239-005-0186-0

JOURNAL OF MOLECULAR EVOLUTION

GC Content Evolution of the Human and Mouse Genomes: Insights from the Study of Processed Pseudogenes in Regions of Different Recombination Rates

Adel Khelifi, Julien Meunier, Laurent Duret, Dominique Mouchiroud

Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Claude Bernard—Lyon 1, 16 rue Raphael Dubois, 69622 Villeurbanne Cedex, France

Received: 26 July 2005 / Accepted: 2 February 2006 [*Reviewing Editor:* Dr. Martin Kreitman]

Abstract. Processed pseudogenes are generated by reverse transcription of a functional gene. They are generally nonfunctional after their insertion and, as a consequence, are no longer subjected to the selective constraints associated with functional genes. Because of this property they can be used as neutral markers in molecular evolution. In this work, we investigated the relationship between the evolution of GC content in recently inserted processed pseudogenes and the local recombination pattern in two mammalian genomes (human and mouse). We confirmed, using original markers, that recombination drives GC content in the human genome and we demonstrated that this is also true for the mouse genome despite lower recombination rates. Finally, we discussed the consequences on isochores evolution and the contrast between the human and the mouse pattern.

Key words: Processed pseudogenes — Isochore — GC content — Human genome — Mouse genome — Biased gene conversion — Recombination

Introduction

G + C content varies along mammalian chromosomes at a scale of hundreds of kilobases to megabases (Bernardi 2000; IHGSC 2001; MGSC 2002). These regions of similar composition are referred to as *isochores* (Bernardi et al. 1985). The isochore structure is strongly conserved in all mammalian genomes (Mouchiroud and Bernardi 1993; Clay et al. 2003) except rodents (Mouchiroud et al. 1988). It has been shown previously that the GC content of rodent genomes, especially murid genomes, was less heterogeneous than that of other mammals (i.e., GC-rich genes are less GC rich in murids than in other mammals, and conversely for GC-poor genes [Mouchiroud and Gautier 1988; Mouchiroud et al. 1988]). The comparison of murine and human complete genomes has confirmed the relative homogeneity of base composition in mouse (MGSC 2002). The question of the origin and evolution of isochores is vigorously debated. Several models have been proposed to explain the emergence and evolution of isochores (for review see Eyre-Walker and Hurst 2001). Some authors suggested that isochores are the result variations in the pattern of mutation (variable mutational bias [VMB] [Wolfe et al. 1993; Francino and Ochman 1999; Matassi et al. 1999; Lercher and Hurst 2002; Subramanian and Kumar 2003; Filatov and Gerrard 2003]). Two other models predict that variations in GC content are due to bias in fixation of AT toward GC mutations because of selection for GC (Bernardi and Bernardi 1986; Charlesworth 1994; Eyre-Walker 1999; Bernardi 2000) or a neutral process called GC-biased gene conversion (GC-BGC) (for review see Eyre-Walker 1993; Galtier et al. 2001; Smith et al, 2002; Webster et al, 2003; Galtier 2003; Marais 2003). In meiotic recombination, parental chromosomes form heteroduplexes (DNA with one

Correspondence to: Adel Khelifi; *email:* khelifi@biomserv.univ-lyon1.fr strand from the male and one strand from the female) in which there can be AT/GC heteromismatches (Lamb 1984). If subjected to the effect of GC-BGC, these mismatches are preferentially repaired in GC in mammalian genomes (Brown and Jiricny 1998; Bill et al. 1998; Kudla et al. 2004). This leads to a fixation bias toward GC alleles in regions affected by GC-BGC. Those regions are difficult to identify but here we assumed that GC-BGC and crossing-overs (COs) are correlated because they both are recombination outcomes. Therefore, in genomes with GC-BGC, a link is expected between CO rates (designated recombination rates) and GC content evolution.

Recently, several authors have reported a positive but weak correlation between GC content and CO rate in the human genome ($r^2 = 0.09$ [Fullerton et al. 1999]; $r^2 = 0.15$ [Kong et al. 2002]; $r^2 = 0.13$ [Huang et al. 2005]). But the correlation is very weak (McVean et al. 2004). More recently, using a different approach, Meunier and Duret (2004) have estimated the substitution pattern in human genome by comparing long noncoding regions (introns, intergenic DNA) from humans, chimps, and baboons. They calculated the GC content toward which sequences are evolving according to the current substitution pattern (also called GC content at equilibrium, or GC*). They observed a strong and positive correlation between recombination rate and GC* in the human genome ($r^2 = 0.6$ [Meunier and Duret 2004]). They concluded that (1) recombination drives evolution of GC content in the human genome, and (2) GC*, calculated on recently diverged sequences, is more appropriate than current GC for the study of dynamic processes such as recombination on genome evolution.

The aim of the present article is to determine the impact of recombination on the evolution of GC content in mammalian genomes. For this purpose, we have analyzed the substitution pattern of neutral markers, processed pseudogenes (Vanin 1985; Mighell et al. 2000), located in regions with various recombination rates in the complete human and mouse genomes. Processed pseudogenes arise by reverse transcription of mRNAs and integration of the resulting cDNAs into the genome (Esnault et al. 2000; Pavlicek et al. 2002). They lack promoters and are generally nonfunctional once integrated into the genome. They therefore freely accumulate substitutions or indels. Because of these characteristics, processed pseudogenes are considered good neutral markers for studying genome evolution (Casane et al. 1997; Francino and Ochman 1999; Zhang and Gerstein 2003). Using these neural markers, we reevaluated the relationship between recombination rate and GC* in the human genome, looked for this relationship in the mouse genome, and compared the results for both species.

Materials and Methods

Dataset of Processed Pseudogenes

5206 human and 3428 mouse processed pseudogenes corresponding, respectively, to 2066 human and 1344 murine reverse transcribed genes have been extracted from the HOPPSIGEN database developed in our laboratory (Khelifi et al. 2005) using the WWW-query system (Perriere et al. 2003). The recombination pattern may be subjected to very quick changes in a genome as has been demonstrated in *Drosophila* (True et al. 1996) and more recently in human (Ptak et al. 2004; Yi et al. 2004; Galtier 2004) and mouse PAR regions (Montoya-Burgos et al. 2003). So, in order to observe and compare the impact of recombination on mammalian genome evolution, we need to study only recently inserted processed pseudogenes of the same age. Assuming a neutral substitution rate of 2.2×10^{-9} and 4.5×10^{-9} per nucleotide per year, respectively, for human and mouse (MGSC 2002), we have retained only processed pseudogenes having a divergence with their functional paralogues of less than 5% for human (at least 95% similarity) and 10% for mouse (at least 90% similarity). Under the neutral model, these should correspond to pseudogenes inserted less than about 22.5 Myr ago in the human genome and 22 Myr ago in the mouse genome. This gave 473 human and 1366 mouse processed pseudogenes. After removing sequences generated by duplication of an existing element, we retained for this study 427 human and 1322 mouse processed pseudogenes.

Alignments

In order to estimate GC content at equilibrium (GC*) in the processed pseudogenes, we analyzed the pattern of substitutions. For this purpose, we built alignments among the processed pseudogene, the coding sequence (CDS) of its functional paralogue, and the CDS of the orthologue taken from either mouse or human. To estimate the GC* in the mouse genome, we have used orthologous genes from human, and conversely to estimate the GC* in the human genome. We extracted orthologous genes from the HOVERGEN database (Duret et al. 1994). Half of all human and mouse coding genes have a known orthologous gene in HOVERGEN, but we were not able to retrieve orthologues for all processed pseudogene paralogues. Using CLUSTALW (Thompson et al. 1994) with default parameters, we aligned the two translated CDS from the functional orthologue and paralogue and used the alignment as a nucleic profile in CLUSTALW for alignment of the processed pseudogene sequence with it. We then refined the alignments by hand using SEAVIEW (Galtier et al. 1996). Finally, we generated 239 human and 500 mouse processed pseudogenes triple alignments.

Inferring Substitution Rates

The pattern of substitutions in the processed pseudogene was analyzed using parsimony on informative sites: a substitution from nucleotide X to nucleotide Y was inferred when both the human and the mouse functional orthologous genes shared state X, but the pseudogene showed state Y. These were classified as informative sites. We restricted our analyses to the first and second codon positions (i.e., slowly evolving sites in functional genes). The third codon position was excluded because it evolves rapidly and, hence, may be saturated. Sites where the human and mouse functional genes differed were considered noninformative and were discarded from the analysis. We discriminated CpG and non-CpG sites. CpG sites are known to be hypermutable and are subjected to different mutational mechanism than non-CpG sites (Bird 1980). Each previously selected site was evaluated for being a CpG site using a method described by Meunier and Duret (2004). We used the third codon position to infer CpG and non-CpG sites on first or second codon position. The relatively low overall pseudogene divergence with their paralogous functional CDS and the restrictive definition of CpG and non-CpG classes make such inferences reliable (less than 5% misinference [Meunier, personal communication]). The informative sites were divided into three classes: (i) sites not immediately preceded by a 5′cytosine or followed by a 3′guanine in any of the three sequences (pseudogene or human or mouse functional gene), i.e., sites that are expected never to have been part of a CpG doublet since pseudogene insertion (CpG-free sites); (ii) sites for which the ancestral pseudogene state inferred by parsimony was part of a CpG doublet (CpG-anc sites); and(iii) other sites. Sites directly flanked at one or both sides by an insertion or a deletion, in at least one of the three sequences, were simply discarded.

*Estimation of GC**

The GC content expected at equilibrium was calculated using the model of Arndt and colleagues (2003a, b). The model takes into account substitutions in neighboring sites. It discriminates single nucleotide substitutions and dinucleotide substitutions (CpG sites). Using the first site category, we inferred six rates by parsimony (pooling complementary rates together): four transversion rates (A → T + T → A, G → C + C → G, A → C + T → G, C → A + G → T) and two transition rates (G → A + C → T, A → G + T → C). The transition rates at CpG sites (C → T + G → A) were estimated using the second site category. We used these values to infer GC* using the available web server (http://evogen.molgen.mpg.de/cgi-bin/server/stationary_properties/stationary_properties.cgi).

Estimation of Local Recombination Rates

The recombination rate in each region where a processed pseudogenes is located was estimated from a dataset of human genetic markers (Kong et al. 2002) and a dataset of mouse genetic markers (Dietrich et al. 1996; Blake et al. 2003). We eliminated markers for which the physical and genetic positions were incoherent with the flanking markers. This left 3493 human and 1404 mouse genetic markers. Given the two genome sizes, there is an average number of 1.2 markers/Mb for the human genome and 0.48 markers/Mb for the mouse genome. In order to estimate processed pseudogene recombination rates, we used a local method based on previously published methods (Hey and Kliman 2002) rather than a global method (Chakravati 1991). The major problem we faced was to get accurate local estimations of recombination rates, which was not possible given the low number and quality of recombination markers available. Indeed, a local method resulted in many cases where we could get an estimation of local recombination rates for only one or two markers around processed pseudogenes. So, we decided to use a method that was intermediate between a local and a global method in order to calculate recombination rates. They were calculated using windows centered on processed pseudogenes. The maximal size for a window was fixed at 8 Mb for the human and 10 Mb for the mouse genome in order to get accurate estimations of recombination rates. This is, according to us, a good compromise between local and global methods. Another constraint was put on the number of markers for each window. We built windows with eight markers, four to each side of the processed pseudogene. Overlapping windows were discarded to avoid having the same estimation of recombination rate for several processed pseudogenes. The local recombination rate (Re*c*) was calculated for each marker using the formula:

$$\mathrm{Re}c = \left[\frac{1}{n_{3'}}\sum_{i=1}^{n_{3'}} Pgi - \frac{1}{n_{5'}}\sum_{i=1}^{n_{5'}} Pgi\right] \Bigg/ \left[\frac{1}{n_{3'}}\sum_{i=1}^{n_{3'}} Ppi - \frac{1}{n_{5'}}\sum_{i=1}^{n_{5'}} Ppi\right]$$

with *Pgi* representing the genetic position (cM) for marker *I*, and *Ppi*, its physical position on chromosomes (Mb). The number of markers in 5′ of the sliding window is designated $n_{5'}$, and the number of markers in 3′ of the sliding window $n_{3'}$.

Local recombination rates were estimated for each previous triple alignment. However, the above criteria were not fulfilled for some alignments. The loss was particularly important for the mouse dataset because of the lower marker density.

The data (alignments and tables) are available at http://pbil.univ-lyon1.fr/datasets/khelifi2005/data.html.

Results

Dataset of Processed Pseudogenes

From the 5206 human and 3428 mouse processed pseudogenes in HOPPSIGEN database release 4 (http://pbil.univ-lyon1.fr/databases/hoppsigen.html), 2066 human and 1344 mouse retrotranscribed coding genes were available. From this dataset, 155 human and 126 mouse processed pseudogenes fulfill our criteria as (1) young processed pseudogenes (≤ 5% divergence with the functional paralogous gene for human and ≤ 10% for mouse); (2) comprised in a triple alignment allowing the inference of substitution pattern; and (3) located in a region whose recombination rate could be reliably estimated. These processed pseudogenes are associated with 134 human and 95 mouse functional paralogues. So they were mainly generated by independent reverse transcription events. Substitutions in the first and/or second codon position with a change specific to the processed pseudogene were counted in the alignments of each set of sequences (the processed pseudogene, the corresponding functional gene, and its orthologue). Thus 164,811 human and 102,162 mouse sites were analyzed using parsimony. Among all these sites, 37,945 for human (including 1842 CpG dinucleotides) and 19,574 for mouse (including 1245 CpG dinucleotides) met our criteria and could be used to calculate substitution rates.

The mean recombination rate in the human dataset of processed pseudogenes was approximately twice as high as the mean recombination rate in the mouse dataset (mean human, 1.18 cM/Mb, compared to mean mouse, 0.68 cM/Mb; Wilcoxon test, $p = 0$). Not only was the mean higher, but also the variance of the recombination rate (0.42 compared to 0.13; Fisher test, $p < 0.05$). For each species, processed pseudogenes were classified into three groups: low recombinant regions (LR), medium recombinant regions (MR), and high recombinant regions (HR). For statistical reasons, the limits between these classes were set for each species so as to obtain three samples

Table 1. Distribution of evaluated sites, GC content at insertion and present GC content of *Homo sapiens* processed pseudogenes

	Recombination rate		
	Low	Medium	High
	(A)		
Number of AT sites	6058	5909	5557
Number of GC sites (non CpG)	6190	6254	6135
Number of CpG sites	543	682	617
Ancestral GC content	46.9 (± 2.1)%	48.5 (± 1.9)%	47.8 (± 1.8)%
GC content of processed pseudogenes	45.5 (± 1.9)%	47.6 (± 1.8)%	47.1 (± 1.7)%
	(B)		
Substitution rates			
$f_{\mathrm{A/T} \rightarrow \mathrm{G/C}}$	0.0087	0.0096	0.0137
$f_{\mathrm{G/C} \rightarrow \mathrm{A/T}}$	0.0152	0.0158	0.0183
$f_{\mathrm{A/T} \rightarrow \mathrm{C/G}}$	0.0026	0.0029	0.0045
$f_{\mathrm{G/C} \rightarrow \mathrm{T/A}}$	0.0036	0.0054	0.0051
$f_{\mathrm{A/T} \rightarrow \mathrm{T/A}}$	0.0021	0.0029	0.0038
$f_{\mathrm{G/C} \rightarrow \mathrm{C/G}}$	0.0045	0.0040	0.0060
$f_{\mathrm{CpG} \rightarrow \mathrm{TpG}}$	0.1418	0.1305	0.1394
GC* content of processed pseudogenes	32.7 (± 0.8)%	32.4 (± 0.8)%	37.6 (± 0.8)%

Note: Calculated parameters for the dataset of 155 human processed pseudogenes: The dataset was divided into three parts containing the same number of processed pseudogenes ($r \leq 0.7984$ cM/Mb; $r > 0.7984$ and ≤ 1.2948; $r > 1.2948$). (**A**) Ancestral GC content is an estimate of GC content of processed pseudogenes at insertion. It was calculated using the number of AT and GC informative sites in the functional homologous gene. The GC content of processed pseudogenes was calculated using the complete processed pseudogene sequences. Confidence intervals were calculated using the definition of a confidence interval for a proportion $(100 * 1.96 * \sqrt{\frac{GC*/100 \times (1-GC*/100)}{\text{total number sites}}}$ and are given (brackets) for a p-value of 0.05. (**B**). We calculated 7 substitution rates (see Materials and Methods), including the rate of deamination in CpG sites. $f_{x/\bar{x}} \rightarrow f_{(y/\bar{y})}$-designates the substitution rate from a base X to Y and its complementary. We used the model developed by Arndt and colleagues (Arndt et al., 2003) to calculate GC content at equilibrium (GC*).

containing approximately the same number of evaluated sites.

*Substitution Pattern and GC**

Tables 1A and 2A show the ancestral GC content of processed pseudogenes inferred by measuring the GC content of functional paralogous genes. Ancestral GC content was not correlated with recombination rates in human (Spearman $\rho = 0.08; p = 0.3126$) or in mouse (Spearman $\rho = -0.03; p = 0.709$). The ancestral GC content of processed pseudogenes is not significantly increasing or decreasing with respect to recombination. We calculated the present GC content for each processed pseudogene. Similarly, no significant correlation was found between the present GC content in processed pseudogenes and the recombination rate for both species (Spearman $\rho = 0.11$, $p = 0.156$, for human and Spearman $\rho = -0.02$, $p = 0.850$, for mouse). The present GC content is lower than the ancestral GC content in mouse processed pseudogenes (49.9% versus 51.6%; Student, $p = 0.004$) but not in human (46.8% versus 47.7%; Student, $p = 0.1021$), although we observed the same trend.

For each former group (LR, MR, and HR) of processed pseudogenes, we calculated seven different substitution rates (see Materials and Methods for details) and then estimated GC content at equilibrium (GC*) using the model of Arndt and colleagues. Tables 1B and 2B show the substitution pattern and GC* for human and for mouse processed pseudogenes in LR, MR, and HR recombinant regions. We found that the GC* increases with recombination rates in both species from LR to HR regions. GC* is significantly lower for processed pseudogenes inserted into LR regions than for processed pseudogenes inserted into HR regions for human (32.7% versus 37.6%; Student, $p = 10^{-14}$) (Table 1B) and for mouse (37.7 versus 42.5; Student, $p = 10^{-7}$) (Table 2B). The differences of GC* between HR and LR regions are approximately the same in the human genome compared to the mouse genome (4.9% vs. 4.8%), whereas the recombination variance is three times higher in the human genome. Moreover, the mean GC* calculated for the whole dataset of processed pseudogenes is lower in human compared to mouse (mean ± confidence interval, 34.4 ± 0.5% versus 40.0 ± 0.7%; Student, $p = 0$). Expected values of GC* for all recombination regions are always lower than the corresponding ancestral GC content (mean ± confidence interval, 34.4 ± 0.5% versus 47.7 ± 1.1% in human and 40.0 ± 0.7% versus 51.6 ± 0.8% in mouse; Student, $p = 0$). Even the GC content of processed pseudogenes inserted in HR regions shows a strong decrease.

Table 2. Distribution of evaluated sites, GC content at insertion and present GC content of *Mus musculus* processed pseudogenes

	Recombination rate		
	Low	Medium	High
	(A)		
Number of AT sites	2783	2942	2730
Number of GC sites (non CpG)	3318	3355	3201
Number of CpG sites	482	406	357
Ancestral GC content	51.6 (± 1.5)%	51.9 (± 1.5)%	51.2 (± 1.6)%
GC content of processed pseudogenes	49.8 (± 1.4)%	50.3 (± 1.6)%	49.6 (± 1.7)%
	(B)		
Substitution rates			
$f_{\mathrm{A/T} \rightarrow \mathrm{G/C}}$	0.0194	0.0147	0.0195
$f_{\mathrm{G/C} \rightarrow \mathrm{A/T}}$	0.0234	0.0173	0.0208
$f_{\mathrm{A/T} \rightarrow \mathrm{C/G}}$	0.0044	0.0054	0.0081
$f_{\mathrm{G/C} \rightarrow \mathrm{T/A}}$	0.0074	0.0058	0.0069
$f_{\mathrm{A/T} \rightarrow \mathrm{T/A}}$	0.0062	0.0047	0.0040
$f_{\mathrm{G/C} \rightarrow \mathrm{C/G}}$	0.0049	0.0047	0.0050
$f_{\mathrm{CpG} \rightarrow \mathrm{TpG}}$	0.1718	0.1663	0.1768
GC* content of processed pseudognes	37.7 (± 0.6)%	39.7 (± 0.7)%	42.5 (± 0.6)%

Note: Calculated parameters for the dataset of 126 mouse processed pseudogenes. The dataset was divided into three parts containing the same number of processed pseudogenes ($r \leq 0.4883$ cM/Mb; $r > 0.4883$ and $r \leq 0.7540$; $r >$ 0. 7540). Cf. Table 1 for legends.

Thus, processed pseudogenes are subjected to a strong bias toward AT.

Thus, we found the following. (i) GC content at equilibrium in processed pseudogenes differs with respect to recombination rate. This result supports the hypothesis that, in both species, GC content evolution is driven by recombination. (ii) There is a strong bias toward AT, leading to an erosion of GC content in processed pseudogenes.

Discussion

Distinct Recombination Patterns in Mammalian Genomes

In the dataset of processed pseudogenes, the mean recombination rate in human was higher than the mean recombination rate in mouse (1.18 compared to 0.68 cM/Mb), in agreement with a recent comparison of mean recombination rates in human and mouse genomes (1.26 compared to 0.56 cM/Mb) (Jensen-Seaman et al. 2004). Moreover, the mean recombination rate in rat genome (0.62 cM/Mb) is also lower than the human one (Jensen-Seaman et al. 2004). So, low recombination rates seem to be a general characteristic of the murid lineage compared to the human lineage. The variance of recombination rate in processed pseudogenes was three times higher in human compared to mouse (0.42 compared to 0.13; Fisher test, $p < 0.05$), in agreement with a recent comparison of the variance of human and murid recombination rates (human, 0.396; mouse, 0.113; rat, 0.117 [Jensen-Seaman et al. 2004]).

The Impact of Recombination on GC Content Evolution

Several authors have reported a positive but weak relationship between recombination rate and GC content in the human genome (Fullerton et al. 1999; Kong et al. 2002; Huang et al. 2005). This is probably because recombination rates evolve quite rapidly. Indeed, a strong and positive link was found between current recombination rates and substitution patterns (Meunier and Duret 2004), consistent with the view that recombination drives GC content in the human genome. We report the same result using a different approach (processed pseudogenes) and extend this to the mouse. To our knowledge, this is the first time that a large-scale effect of recombination is shown in the mouse genome. Under the GC-BGC hypothesis, GC content increases with the rate of gene conversion. One limit of our approach is that we did not use the rate of gene conversion. Assuming that gene conversion is positively correlated with the rate of CO (Jeffreys and Neumann 2002), a positive link is expected between the rate of CO (the recombination rate) and the evolution of GC content. We found such a relation in our results. However, recent studies are against our assumption and suggested that the rate of gene conversion is not correlated with the rate of CO in human chromosome 21 (Padhukasahasram et al. 2004) and negatively correlated in *Drosophila*

(Andolfatto and Wall 2003). But the local pattern of gene conversion is poorly known in mammalian genomes, so no strong conclusion can be given for the correlation between the rate of CO and the rate of gene conversion, and our hypothesis still remains valuable. Moreover, gene conversion not only occurs during CO events but also can be observed independently. Kauppi and colleagues (2004) have demonstrated that the rate of gene conversion outside CO is 4 to 15 times higher than the rate of gene conversion inside CO. Recent works have demonstrated that GC-BGC can increase GC content very quickly (Belle et al. 2004; Kudla et al. 2003; Galtier 2004; Webster et al. 2005). The rate of gene conversion is 100 times higher than the rate of neutral substitutions in mammalian genomes (Kudla et al. 2004), which could be enough to induce a tremendous change in GC content and in isochore structure. All this evidence is in favor of the hypothesis that the effect of gene conversion on mammalian genomes is underestimated. As soon as reliable data are available, it should be possible to test directly the link between the rate of gene conversion and GC content evolution.

Evolution of Isochore Structure in Mammalian Genomes

The isochore structure is vanishing in mammalian genomes, according to recent studies (Duret et al. 2002; Arndt et al. 2003b; Belle et al. 2004). Moreover, Meunier and Duret (2004) have demonstrated that the current substitution pattern in human genome should lead to a new isochore structure, with a lower GC mean and a lower variance. Our data show the same erosion for the GC content at equilibrium. The mean GC content at equilibrium is lower than the ancestral GC content in processed pseudogenes (34.4 versus 47.7 in human and 40.0 versus 51.6 in mouse). Processed pseudogenes are more likely subjected to the same AT substitution bias observed in the human genome (IHGSC 2001). The differences of GC* (ΔGC*) between LR and HR are approximately the same in both species (4.9% in human and 4.8% in mouse). Some processed pseudogenes are evolving toward a high GC content at equilibrium and others toward a low GC content at equilibrium, though we demonstrated that the ancestral GC content is not significantly different among LR, MR, and HR regions. The former result supports the hypothesis that isochores, in the human and the mouse genomes, are evolving toward a new structure. More recent processed pseudogenes, from our previous dataset, were analyzed (divergence lower than 3% for the human and 5% for the mouse). They are more likely to have evolved under the current recombination rate. The ΔGC* between LR and HR regions is 12.2% (31.2% versus 43.4%) in the human genome and 16.4% (39.6% versus 56%) in the mouse genome. As expected, the effect of recombination on GC change is stronger for very recent processed pseudogenes. All of these results are in agreement with the view that current isochores are vanishing and that a new isochore structure is emerging.

The Effect of BGC

The efficiency of BGC depends on several factors (for review see Nagylaki 1983; Marais 2003): (1) H, the level of heterozygosity, which is directly related to the effective population size (Ne); (2) γ, the rate of conversion, which is the probability per generation that a given site is affected by a gene conversion tract of length L (γ was estimated as 3×10^{-5} for human and 2×10^{-6} for mouse [for review see Marais 2003]); and (3) c, the bias in favor of the GC allele. We found that the GC* is higher in mouse than in human, whereas the mouse recombination rate is lower than the human one, suggesting a higher efficiency of BGC in the mouse genome. To explain this and given the values of γ, we can only assume that c, Ne, or both are higher in the mouse genome. Ne was estimated to be approximately 10,000 in the human lineage (Takahata et al. 1995; Zhao et al. 2000; Yu et al. 2003; Keightley et al. 2005) and between 450,000 and 810,000 in the mouse lineage (Keightley et al. 2005). These values are in agreement with our results. c is known in the human genome (Birdsell 2002) but not in the mouse genome. Estimation of c in the mouse genome and a precise estimation of γ in both genomes would help us to determine whether BGC can explain differences in substitution patterns among mammals.

Acknowledgments. Thanks go to Tim Greenland and Gabriel Marais for discussions and comments on the manuscript. We thank the Centre de Calcul de l'IN2P3 for providing computer resources. This work was supported by the Centre National de la Recherche Scientique and the Claude Bernard University-Lyon.

References

Andolfatto P, Wall JD (2003) Linkage disequilibrium patterns across a recombination gradient in African Drosophila melanogaster. Genetics 165:1289–1305

Arndt PF, Burge CB, Hwa T (2003a) DNA sequence evolution with neighbor-dependent mutation. J Comput Biol 10:313–322

Arndt PF, Petrov DA, Hwa T (2003b) Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. Mol Biol Evol 20:1887–1896

Belle EM, Duret L, Galtier N, Eyre-Walker A (2004) The decline of isochores in mammals: an assessment of the GC content variation along the mammalian phylogeny. J Mol Evol 58:653–660

Bernardi G (2000) The compositional evolution of vertebrate genomes. Gene 259:31–43

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

Bill CA, Duran WA, Miselis NR, Nickoloff JA (1998) Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese hamster ovary cells. Competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. Genetics 149:1935–1943

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499–1504

Birdsell JA (2002) Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol 19:1181–1197

Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT (2003) MGD: the Mouse Genome Database. Nucleic Acids Res 31:193–195

Brown TC, Jiricny J (1988) Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. Cell 26:705–711

Casane D, Boissinot S, Chang BH, Shimmin LC, Li WH (1997) Mutation pattern variation among regions of the primate genome. J Mol Evol 45:216–226

Chakravarti A (1991) A graphical representation of genetic and physical maps: the Marey map. Genomics 11:219–222

Charlesworth B (1994) Genetic recombination. Patterns in the genome. Curr Biol 4:182–184

Clay O, Douady CJ, Carels N, Hughes S, Bucciarelli G, Bernardi G (2003) Using analytical ultracentrifugation to study compositional variation in vertebrate genomes. Eur Biophys J 32:418–426

Dietrich WF, Miller J, Steen R, Merchant MA, Damron-Boles D, Husain Z, Dredge R, Daly MJ, Ingalls KA, O'Connor TJ (1996) A comprehensive genetic map of the mouse genome. Nature 380:149–152

Duret L, Mouchiroud D, Gouy M (1994) HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res 25:2360–2365

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

Esnault C, Maestre J, Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. Nat Genet 24:363–367

Eyre-Walker A (1993) Recombination and mammalian genome evolution. Proc R Soc Lond B Biol Sci 252:237–243

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 LD (2001) The evolution of isochores. Nat Rev Genet 2:549–555

Filatov DA, Gerrard DT (2003) High mutation rates in human and ape pseudoautosomal genes. Gene 23:67–77

Francino MP, Ochman H (1999) Isochores result from mutation not selection. Nature 400:30–31

Fullerton SM, Bernardo Carvalho A, Clark AG (1999) Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol 18:1139–1142

Galtier N (2003) Gene conversion drives GC content evolution in mammalian histones. Trends Genet 19:65–68

Galtier N (2004) Recombination, GC-content and the human pseudoautosomal boundary paradox. Trends Genet 20:347–349

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12:543–548

Galtier N, Piganeau G, Mouchiroud D, Duret L (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907–911

Hey J, Kliman RM (2002) Interactions between natural selection, recombination and gene density in the genes of Drosophila. Genetics 160:595–608

Huang SW, Friedman R, Yu N, Yu A, Li WH (2005) How strong is the mutagenicity of recombination in mammals? Mol Biol Evol 22:1157

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

Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ (2004) Comparative recombination rates in the rat, mouse, and human genomes. Genome Res 14:528–538

Jeffreys AJ, Neumann R (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat Genet 3:267–271

Kauppi L, Jeffreys AJ, Keeney S (2004) Where the crossovers are: recombination distributions in mammals. Nat Rev Genet 5:413–424

Keightley PD, Lercher MJ, Eyre-Walker A (2005) Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biol 3:e42 (Epub Jan 25)

Khelifi A, Duret L, Mouchiroud D (2005) HOPPSIGEN: a database of human and mouse processed pseudogenes. Nucleic Acids Res 33 (Database Issue):D59–D66

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247

Kudla G, Helwak A, Lipinski L (2004) Gene conversion and GC-content evolution in mammalian Hsp70. Mol Biol Evol 21:1438–1444

Lamb BC (1984) The properties of meiotic gene conversion important in its effects on evolution. Heredity 53:113–138

Lercher MJ, Hurst LD (2002) Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet 18:337–340

Marais G (2003) Biased gene conversion: implications for genome and sex evolution. Trends Genet 19:330–338

Matassi G, Sharp PM, Gautier C (1999) Chromosomal location effects on gene sequence evolution in mammals. Curr Biol 9:786–791

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) Fine-scale structure of recombination rate variation in the human genome. Science 23:581–584

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

Mighell AJ, Smith NR, Robinson PA, Markham AF (2000) Vertebrate pseudogenes. FEBS Lett 468:109–114

Montoya-Burgos JI, Boursot P, Galtier N (2003) Recombination explains isochores in mammalian genomes. Trends Genet 19:128–130

Mouchiroud D, Bernardi G (1993) Compositional properties of coding sequences and mammalian phylogeny. J Mol Evol 37:109–116

Mouchiroud D, Gautier C (1988) High codon-usage changes in mammalian genes. Mol Biol Evol 5:192–194

Mouchiroud D, Gautier C, Bernardi G (1988) The compositional distribution of coding sequences and DNA molecules in humans and murids. J Mol Evol 27:311–320

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Nagylaki T (1983) Evolution of a large population under gene conversion. Proc Natl Acad Sci USA 80:5941–5945

Padhukasahasram B, Marjoram P, Nordborg M (2004) Estimating the rate of gene conversion on human chromosome 21. Am J Hum Genet 75:386–397

Pavlicek A, Paces J, Zika R, Hejnar J (2002) Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. Gene 300:189–194

Perrière G, Combet C, Penel S, Blanchet C, Thioulouse J, Geourjon C, Grassot J, Charavay C, Gouy G, Duret L, Deleage G (2003) Integrated databanks access and sequence/structure analysis services at the PBIL. Nucleic Acids Res 31: 3393–3399

Ptak SE, Roeder AD, Stephens M, Gilad Y, Paabo S, Przeworski M (2004) Absence of the TAP2 human recombination hotspot in chimpanzees. PLoS Biol 2:849–855

Smith NG, Webster MT, Ellegren H (2002) Deterministic mutation rate variation in the human genome. Genome Res 12:1350–1356

Subramanian S, Kumar S (2003) Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res 13:838–844

Takahata N, Satta Y, Klein J (1995) Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol 48:198–221

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

True JR, Mercer JM, Laurie CC (1996) Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics 142:507–523

Vanin EF (1985) Processed pseudogenes: characteristics and evolution. Annu Rev Genet 19:53–272

Webster MT, Smith NG, Ellegren H (2003) Compositional evolution of noncoding DNA in the human and chimpanzee genomes. Mol Biol Evol 20:278–286

Webster MT, Smith NG, Hultin-Rosenberg L, Arndt PF, Ellegren H (2005) Male-driven biased gene conversion governs the evolution of base composition in human Alu repeats. Mol Biol Evol 22:1468–1474

Wolfe KH, Sharp PM, Li WH (1993) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Yi S, Summers TJ, Pearson NM, Li WH (2004) Recombination has little effect on the rate of sequence divergence in pseudoautosomal boundary 1 among humans and great apes. Genome Res 14:37–43

Yu N, Jensen-Seaman MI, Chemnick L, Kidd JR, Deinard AS, Ryder O, Kidd KK, Li WH (2003) Low nucleotide diversity in chimpanzees and bonobos. Genetics 164:1511–1518

Zhang Z, Gerstein M (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res 15:5338–5348

Zhao Z, Jin L, Fu YX et al. (2000) Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc Natl Acad Sci USA 97:11354–113548