Recombination Drives the Evolution of GC-Content in the Human Genome

Julien Meunier and Laurent Duret

UMR 5558 CNRS, Université Lyon I, Villeurbanne, France

Unraveling the evolutionary forces responsible for variations of neutral substitution patterns among taxa or along genomes is a major issue in the identification of functional sequence features. Mammalian genomes show large-scale regional variations of GC-content (the isochores), but the substitution processes at the origin of this structure are poorly understood. We have analyzed the pattern of neutral substitutions in 14.3 Mb of primate noncoding regions. We show that the GC-content toward which sequences are evolving is strongly correlated ($r^2 = 0.61$, $P \leq 2\ 10^{-16}$) with the rate of crossovers (notably in females). This demonstrates that recombination drives the evolution of base composition in human (probably via the process of biased gene conversion). The present substitution patterns are very different from what they had been in the past, resulting in a major modification of the isochore structure of our genome. This non-equilibrium situation suggests that changes of recombination rates occur relatively frequently during evolution, possibly as a consequence of karyotype rearrangements. These results have important implications for understanding the spatial and temporal variations of substitution processes in a broad range of sexual organisms, and for detecting the hallmarks of natural selection in DNA sequences.

Introduction

The base composition of genomes varies substantially between taxa, and even along chromosomes. Notably, the genome of mammals and birds is structured into large genomic regions (several hundreds of kilobases) of relatively homogeneous base composition (the so-called isochores), mostly ranging from 30% to 60% of GC-content (Bernardi 2000). This isochore organization reflects some fundamental aspects of genome organization and evolution. Notably, it has been shown that the GC-content of isochores is correlated with many other genomic features: gene density (Mouchiroud et al. 1991; Lander et al. 2001), intron length (Duret, Mouchiroud, and Gautier 1995; Lander et al. 2001), replication timing (Watanabe et al. 2002), recombination (Kong et al. 2002), methylation pattern (Jabbari and Bernardi 1998), and distribution of transposable elements (Smit 1999; Lander et al. 2001). Thus, determining the underlying mechanism driving the evolution of isochores is a major issue in understanding the organization of genomes.

The question of the evolutionary forces and molecular processes responsible for this peculiar genomic landscape of mammalian chromosomes remains highly debated (Eyre-Walker and Hurst 2001; Galtier et al. 2001). It has been recently conjectured that recombination might influence the evolution of base composition (Galtier et al. 2001). In mammals, several observations support this hypothesis. First, there is a positive correlation between the GC-content of genomic DNA and the local rate of crossover (Kong et al. 2002). Second, multigenic families that are subject to concerted evolution (i.e., ectopic recombination) are GC-rich (Galtier et al. 2001; Galtier 2003). Finally, it has been shown that the translocation of the *Fxy* gene into the pseudoautosomal region (where the recombination rate is extremely high) has resulted in a striking increase of its GC-content in the mouse lineage (Montoya-Burgos, Boursot, and Galtier 2003). The positive correlation between recombination rate and GC-content seems to be a widespread phenomenon, because it has also been observed in yeast, Drosophila, and nematode (Gerton et al. 2000; Marais, Mouchiroud, and Duret 2001; Birdsell 2002; Piganeau and Marais 2002).

Note, however, that such a correlation does not indicate which is the relation of causality between the two variables. According to some authors, the correlation observed in yeast does not occur because recombination affects substitution patterns, but rather because high GC-contents promote recombination (Gerton et al. 2000; Blat et al. 2002; Petes and Merker 2002). Moreover, it should be stressed that the correlation between crossover rate and GC-content is relatively weak in human ($r^2 = 0.15$; Kong et al. 2002). Thus, at first sight, one might conclude that recombination is, at most, a minor determinant of base substitution patterns. Yet, it is important to note that this correlation was measured between two variables that reflect processes operating on different time scales: the rate of crossover is derived from linkage maps measured on extant populations, whereas the GC-content results from the average substitution pattern to which sequences have been subject over a long evolutionary period. Furthermore, many observations indicate that recombination rates may change rapidly during evolution as a consequence of inversions, translocations, and possibly point mutations (True, Mercer, and Laurie 1996; Depaulis et al. 2000; Jeffreys and Neumann 2002; Montoya-Burgos, Boursot, and Galtier 2003). For example, the human linkage map is 30% larger than that of baboon (Rogers et al. 2000). Variations of recombination rate have also been reported among more closely related species, and even within populations (True, Mercer, and Laurie 1996; Depaulis et al. 2000; Kong et al. 2002; Navarro and Barton 2003). To determine whether recombination drives the evolution of base composition, it is therefore necessary to use estimates of recombination rates and substitution patterns measured on similar time scales.

For this purpose, we analyzed the recent pattern of neutral substitutions, which occurred after the human and chimpanzee divergence, in 14.3 Mb of orthologous

Key words: isochore, recombination, biased gene conversion, GC-content, human genome.

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

Mol. Biol. Evol. 21(6):984–990. 2004
DOI:10.1093/molbev/msh070
Advance Access publication February 12, 2004

noncoding DNA sequences. This analysis demonstrates for the first time a genome-wide correlation between the substitution pattern and the rate of crossover, which indicates that recombination is the major determinant of the evolution of isochores in the genome of primates. The mechanism by which recombination affects the GC-content, as well as the consequences of variation of recombination rate on the evolution of base composition in the genome of sexual organisms, is discussed.

Methods

Alignments

To construct genomic human/chimpanzee/baboon alignments, we retrieved large ($\geq$20 kb) chimpanzee and baboon (i.e., Pan and Papio species) DNA sequences (respectively 291 and 233) from GenBank (Rel. 133, February 2003). We conducted a similarity search against human chromosomes (Ensembl, release 8.3) using Megablast to roughly map chimpanzee and baboon sequences on their orthologous loci. We then used human/chimpanzee and human/baboon pairwise alignments computed by MGA (Holn, Kurtz, and Ohlebusch 2002) to generate an accurate mapping, which enabled us to identify potential triple alignments. Finally, the alignments were generated using ClustalW, and they comprised a total of 14.3 Mb of orthologous sites distributed on 12 human autosomes. More details on the methodology and the aligments are available at http://pbil.univ-lyon1.fr/datasets/MeunierDuret2004/data.html and in the Supplementary Infomation section.

Recombination Rate

The rates of crossover in the human genome were taken from Kong et al. (2002). The average recombination rates in mouse chromosomes were computed from the Whitehead Mouse Genetic Map URL: http://carbon.wi.mit.edu:8000/cgi-bin/mouse/index#genetic.

Inferring Substitution Rates

Substitutions were inferred in human and chimpanzee lineages using unweighted parsimony on informative sites, with the baboon as outgroup. In our analyses, we only considered substitutions occurring in noncoding regions (as defined by Ensembl annotations). It is known that because of multiple substitutions, parsimony may be misleading. Given the evolutionary distances considered here, only hyper-mutable 5′-CpG-3′ dinucleotides (hereafter noted CpG) are expected to generate homoplasy. We therefore considered three classes of sites: (1) sites not immediately preceded by a C in 5′ or followed by a G in 3′ in any of the human, chimpanzee, or baboon sequences—i.e., sites that are expected never to have been part of a CpG doublet since the last common ancestor of the three species (CpG-free sites); (2) sites for which the ancestral human/chimpanzee state inferred by parsimony was part of a CpG (CpG-anc sites); (3) other sites. All simulations (see below) revealed that (1) sites that we inferred as CpG-free truly evolved without being part of a CpG; (2) sites that we inferred as CpG-anc were truly part of an ancestral CpG before the human/chimpanzee split; (3) all substitution rates could be accurately estimated on CpG-free and CpG-anc sites, simply by dividing the number of observed changes by the number of inferred ancestral sites (rate estimation errors $\leq$3%). Using the first site category, we inferred by parsimony six rates (pooling together complementary rates): 4 transversion rates (A → T + T → A, G → C + C → G, A → C + T → G, C → A + G → T) and 2 transition rates (G → A + C → T, A → G + T → C. The transition rate at CpG sites (C → T + G → A) was estimated using the second site category. For better rate estimates, we pooled substitutions in human and chimpanzee lineages.

Simulations

We checked the quality of parsimony inferences on the three classes of sites by performing simulations with the following parameters: ratio transition over transversion: 2.75; increase of the transition rate in CpG: ×10; human/chimpanzee divergence: 1%; human/baboon divergence: 5%. Simulations were driven as follows: we first let a sequence evolve until each site experienced an average of 10 substitutions (i.e., to reach equilibrium), using 0.36, 0.42, and 0.52 as GC-bias values; we then simulated the sequence evolution on the human/chimpanzee/baboon phylogenetic tree topology either leaving the GC-bias value unchanged or setting it to 0.35 (i.e., mimicking a non-equilibrium situation).

We also used simulations to estimate the time needed to reach the equilibrium GC-content. As above, we started with sequences at equilibrium (initial GC-content: GC_i), and then changed the GC-bias parameter and let sequences evolve until they reach their new equilibrium (GC_n). We plotted the GC-content as a function of the number of substitutions (d), and we measured on this graph the value $d_{\frac{1}{2}}$ corresponding to half of the distance between GC_i and GC_n. We repeated simulations for different values of GC_i (0.35, 0.40, 0.45, 0.50, and 0.55), and of GC_n (0.35 or 0.40), similar to that observed in our data. The average value of $d_{\frac{1}{2}}$ was 0.75 substitutions per site (varying from 0.7 to 0.8, depending on the values of GC_i and GC_n). Given the per-year substitution rate observed between human and chimpanzee, this value of $d_{\frac{1}{2}}$ corresponds approximatively to 750 Myr.

Analysis of Regional Substitution Patterns

Human autosomal chromosomes were divided into 1 Mb non-overlapping windows (referred to as loci). A locus was retained for the analysis only if (1) it was associated with a marker from the genetic map (if a locus was associated with more than one marker, we used the mean crossover rate value for the analysis); (2) it contained alignments with more than 100 AT ↔ GC (i.e., from A or T to G or C or the opposite) substitutions in human and chimpanzee lineages; (3) AT → GC (i.e., from A or T to G or C) and GC → AT substitution distributions presented no significant difference among both lineages (Fisher exact test, $P > 0.05$). A total of 33 loci were finally selected for the analysis, with a mean of 1,743 inferred AT ↔ GC substitutions in human and chimpanzee lineages. It is possible to compute the GC-content toward which a sequence is evolving, given its present substitution pattern (the equilibrium GC-content, hereafter denoted as *GC**). *GC** is generally computed by using the model proposed by Sueoka (1962): under the assumption that all sites within a sequence evolve independently of each other, then *GC** should converge to $u/(u + v)$, where u and v are, respectively, the AT → GC and GC → AT substitution rates. This assumption is, however, not valid in vertebrates, where it is known that the rate of mutation of a given base depends on the nature of its neighboring bases, essentially because of the hypermutability effects of CpG dinucleotides (Arndt, Petrov, and Hwa 2003). We therefore used the seven substitution rates described above as input to derive *GC** using the sequence evolution model of Arndt, Burge, and Hwa (2003), which takes into account not only the different transversion and transition rates, but also the high transition rate at CpG dinucleotides. The error in *GC** estimates resulting from the use of parsimony appeared negligible in all our simulations (less than 1% error).

Results and Discussion

Substitution Pattern in Human and Chimpanzee

To analyze the pattern of substitution in the lineages of human and chimpanzee, we aligned 14.3 Mb of orthologous noncoding DNA sequences from human, chimpanzee, and baboon (*Methods*). Subsequently, human chromosomes were split into 1 Mb non-overlapping windows (referred to as loci). We retained 36 loci dispersed over 12 autosomal chromosomes, each of which contained sufficiently long triple alignments. The average rate of divergence (excluding indels) between human and chimpanzee is 1.1%, and that between baboon and the couple human-chimpanzee is 5.7%. The whole data set includes 62,746 informative substitutions (on average, 1,743 substitutions per locus). To determine whether the substitution pattern remained unchanged since the divergence between human and chimpanzee, we compared the number of AT → GC (i.e., from A or T to G or C) and GC → AT substitutions in each lineage. In three of the 36 loci (8%), the pattern was significantly different in the two lineages (Fisher exact test, $P < 0.05$). The three deviating loci were excluded from further analysis, as the substitution patterns inferred in these loci may not reflect the present situation in human and chimpanzee. Then, for each of the 33 remaining loci, we pooled substitution data in the human and chimpanzee lineage and computed by parsimony seven substitution rates, taking into account the high transition rate at the CpG dinucleotides (*Methods*). It is possible, given these rates, to estimate the GC-content toward which a locus converges (i.e., the equilibrium GC-content, hereafter noted *GC**; see *Methods*). A summary of the 33 genomic loci used in the following analysis is given in table 1.

Substitution Pattern and Recombination

The main result is presented in figure 1, where *GC** is plotted against the crossover rate in 33 loci dispatched across the human genome. This analysis clearly demonstrates for the first time a strong genome-wide correlation between crossover rate and the pattern of base substitution ($r^2 = 0.61$, $P < 2.10^{-16}$). The strength of the correlation indicates that the crossover rate is a strong predictor of the evolution of base composition, and it therefore suggests a direct causal relation between recombination rate and *GC**. Several authors have proposed that the correlation between crossover rate and the present base composition (*GC*) exists because regions of high GC-content stimulate recombination (Gerton et al. 2000; Blat et al. 2002; Petes and Merker 2002). As previously noted (Kong et al. 2002), we observed a positive correlation between the present base composition (*GC*) and the crossover rate ($r^2 = 0.21$, $P < 2.10^{-16}$, fig. 2). Yet, this correlation is relatively weak compared to that obtained with *GC** ($r^2 = 0.61$), and it is entirely due to the relationship between *GC* and *GC** (figs. 2 and 3). This clearly shows that it is not the base composition that affects recombination rate, but instead that recombination drives the evolution of base composition. These results suggest that this phenomenon lies at the origin of the evolution of GC-rich isochores in the ancestor of amniotes (about 300–350 MYA).

It should be stressed that given the resolution of the present genetic map of the human genome, and given the limited number of substitutions in human and chimpanzee lineages, we had to perform our analyses on relatively large DNA regions (1 Mb). However, patterns of substitution might vary at a finer scale. Indeed, there is growing evidence for recombination hot spots no more than a few kilobases long (Jeffreys, Kauppi, and Neumann 2001; Jeffreys and Neumann 2002; Reich et al. 2002; de Massy 2003). We therefore expect important local variations of substitution patterns (at the kilobase scale). In other words, the overall GC-content of a region (at the megabase scale) would reflect the average density in recombination hot spots. This nonuniform distribution of recombination events would explain why GC-rich isochores show a relatively strong heterogeneity of GC-content as compared to GC-poor isochores (Nekrutenko and Li 2000).

It is known that the two sexes contribute differently to the number of mutations in the human genome: the rate of mutation is 4 to 6 times higher in male than in females (reviewed in Li, Yi, and Makova 2002). There are also some important variations in the rate of crossover according to sex. In humans, the average rate of crossover is 1.65 times higher in the autosomes of females than in those of males. Thus, females probably contribute more than males to the increase in GC-content. Interestingly, *GC** correlates more strongly with sex-averaged rates ($r^2 = 0.61$, fig. 1) or female crossover rates ($r^2 = 0.69$), than with male crossover rates ($r^2 = 0.21$), suggesting that the impact of recombination on base composition might be female-driven. However, more data will be necessary to determine whether this difference is significant or not (the correlation is significantly weaker in males, but the test becomes nonsignificant after the removal of only one highly recombinant locus [data not shown]).

How Does Recombination Influence GC-Content?

Three hypotheses have been proposed to explain how recombination might influence base composition: (1)

Table 1
Description of the 33 Genomic Loci

					CpG-free		CpG-anc	
Chr	Pos	Rec	Sites	*GC*	*AT→GC*	*GC→AT*	*GC→AT*	GC*
1	61	1.36	87,152	0.431	255	351	142	0.356
4	78	0.86	189,291	0.368	845	919	263	0.340
6	57	0.67	35,212	0.344	173	172	44	0.338
7	28	1.26	35,712	0.397	134	133	75	0.374
7	29	1.86	396,709	0.409	1706	1478	677	0.418
7	30	1.92	257,883	0.422	1082	1130	572	0.394
7	32	1.27	315,137	0.401	1291	1391	504	0.372
7	73	0.80	94,533	0.485	212	384	362	0.340
7	74	0.83	199,379	0.517	382	866	878	0.338
7	90	0.85	483,250	0.379	1642	1696	653	0.358
7	91	0.92	205,908	0.372	663	699	229	0.346
7	92	0.94	453,557	0.380	1444	1495	591	0.358
7	93	0.94	310,933	0.358	1167	1068	314	0.364
7	95	0.63	398,631	0.370	1445	1401	493	0.362
7	96	0.54	184,911	0.384	640	622	266	0.374
7	102	0.69	151,884	0.494	323	522	612	0.372
7	118	0.35	225,552	0.366	804	884	295	0.334
7	128	0.48	458,030	0.430	1289	1631	859	0.368
8	133	2.07	44,051	0.460	155	194	134	0.394
10	7	2.09	37,349	0.451	140	182	130	0.378
11	37	0.55	19,899	0.491	38	64	45	0.352
14	57	0.95	31,273	0.451	100	163	97	0.336
14	58	1.30	42,496	0.470	102	164	106	0.356
16	20	1.49	20,361	0.450	57	76	61	0.368
16	65	0.51	15,961	0.427	75	93	82	0.360
17	11	2.80	22,700	0.523	50	76	39	0.418
17	52	1.34	78,130	0.527	177	327	232	0.390
19	51	1.15	21,132	0.544	44	91	98	0.378
19	56	3.23	36,677	0.503	116	176	123	0.392
19	57	3.91	155,372	0.464	630	750	424	0.408
22	2	3.42	43,190	0.504	132	146	113	0.464
22	7	1.83	11,076	0.482	36	66	23	0.388
22	9	2.45	73,135	0.514	167	322	198	0.372

NOTE.—Chr: chromosome number; pos: position (Mb) of the 1-Mb window (i.e., locus) on the chromosome; rec: recombination rate (cM/Mb; sex-averaged) of the locus; sites: number of sites on which substitutions could be safely inferred by parsimony (*Methods*); *GC*: present GC-content of the human sequences contained in a triple alignment; CpG-free: sites inferred as CpG-free (*Methods*); CpG-anc: sites inferred as CpG-anc (*Methods*); *AT→GC*: number of substitutions from A or T to G or C inferred by parsimony; *GC→AT*: number of substitutions from G or C to A or T inferred by parsimony; GC*: equilibrium GC-content, as computed by the method proposed by Arndt, Burge, and Hwa (2003).

recombination promotes the creation of new GC-alleles via mutation; (2) recombination favors the spread of GC-alleles when both AT and GC alleles are present in a population (i.e., there is a fixation bias toward GC alleles), via biased gene conversion (BGC) (Galtier et al. 2001); (3) there is a selective pressure in favor of a high GC-content, and GC* increases with high crossover rates because selection is more efficient (Charlesworth 1994). Among these three models, the selectionist one appears very unlikely. Indeed, the increase in GC-content induced by a single AT → GC mutation, occurring in megabase-long noncoding sequences, is extremely small. Given the limited population sizes of mammals, it is totally unrealistic that this tiny increase in GC-content might be gripped by natural selection. Conversely, various observations support the BGC model. Notably, there is a fixation bias toward GC-alleles (Duret et al. 2002; Webster, Smith, and Ellegren 2003), which is consistent with the BGC hypothesis, but not with the simple mutational model. Moreover, in vivo repair of GC:AT mismatches is indeed biased in favor of GC-alleles (Birdsell 2002), and such bias is expected to induce BGC (Galtier et al. 2001). It should be borne in mind that the BGC model does not exclude the mutational model. Indeed, various observations suggest that recombination has an influence on the mutation rate in humans (Lercher and Hurst 2002; Hellmann et al. 2003); yet, it is not clear whether recombination favors the creation of new GC-alleles. More polymorphism data will be needed to assess the role of mutational bias related to recombination with respect to GC-content evolution.

It is important to note that the crossover rate is not necessarily perfectly correlated with the recombination rate. Each recombination event involves nonreciprocal DNA exchange (i.e., gene conversion), but only a fraction of these recombination events are resolved to give a crossover. There is evidence that the proportion of recombination events that lead to crossover varies according to species (Kleckner 1996) and within genomes (Langley et al. 2000). If the BGC model is correct, we expect that GC* will correlate more strongly with the rate of recombination than with the rate of crossover. Thus, the correlation presented in figure 1 might in fact underestimate the real relationship between GC* and recombination.

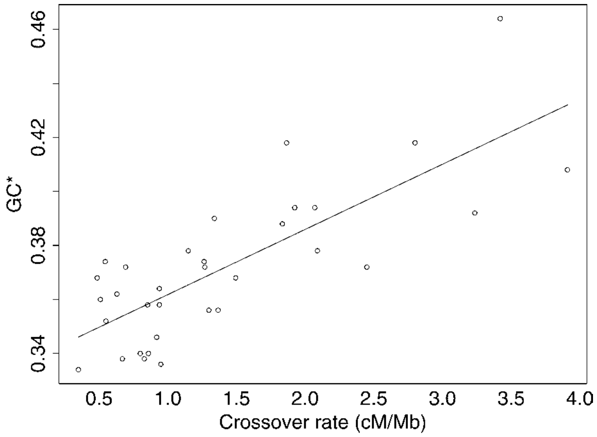

FIG. 1.—Relationship between the base composition toward which a locus is evolving (*GC**) and its crossover rate (cM/Mb; sex-averaged). N = 33 loci from 12 human autosomes. $r^2 = 0.61$, $P < 2.10^{-16}$ (Student *t*-test).

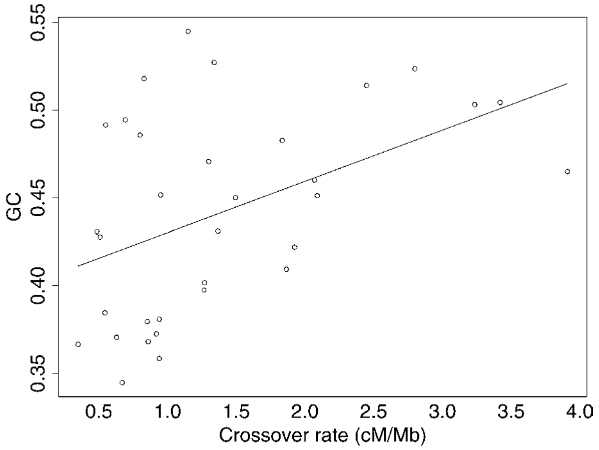

FIG. 2.—Relationship between the present base composition of a loci (*GC*) and its crossover rate (cM/Mb; sex-averaged). N = 33 loci from 12 human autosomes. $r^2 = 0.21$, $P < 2.10^{-16}$ (Student *t*-test). We investigated whether *GC* was correlated with the crossover rate independently of the relationship between *GC* and *GC** by computing the residuals of the regression of *GC* versus *GC**. These residuals are not correlated with the crossover rate ($r^2 = 0.03$, $P = 0.36$, Student *t*-test).

Evolution of the Isochore Structure in Human and Chimpanzee

The second important result is presented in figure 3, which shows the relationship between the present base composition of a locus (*GC*), and the composition toward which it is evolving (*GC**). *GC* and *GC** are only weakly correlated, which indicates that the present substitution pattern is very different from what it had been in the past. Thus, the present pattern is profoundly modifying the isochore organization of our genome, which is evolving toward a more homogeneous and GC-poor base compositon (*GC*: 35%–55%; *GC**: 33%–42%). This observation confirms recent works showing a strong erosion of GC-rich isochores (Duret et al. 2002; Arndt, Petrov, and Hwa 2003; Webster, Smith, and Ellegren 2003). Note that contrary to our findings, Webster, Smith, and Ellegren (2003) reported an increase of GC-content in GC-poor isochores. However, this observation appears to be an artifact, because their estimate of *GC** was based on the simple Sueoka model (Sueoka 1962), and did not take into account the effect of CpGs (data not shown).

The relative independence of *GC* and both of the current crossover rates ($r^2 = 0.21$; fig. 2) and *GC** ($r^2 = 0.16$; fig. 3) is probably due to changes of recombination rates during evolutionary time. As previously mentioned, rates of recombination are highly variable (True, Mercer, and Laurie 1996; Depaulis et al. 2000; Jeffreys and Neumann 2002; Montoya-Burgos, Boursot, and Galtier 2003). For example, the rate of crossover is twice as high in the human genome as in mice (fig. 4). Furthermore, under the BGC model, the overall *GC** also depends on the effective population size (*Ne*) (Nagylaki 1983), another parameter expected to vary rapidly during evolution. Conversely, the evolution of base composition is a relatively slow process: we estimated that the time needed to reduce the distance from equilibrium in base composition by half is about 750 Myr (*Methods*). It seems therefore likely that the *GC** of most loci will significantly change before equilibrium is met. This suggests that the situation of equilibrium in base composition would be rather an exception than the rule in genomes which GC-content is driven by recombination. Interestingly, after only 6 Myr of divergence, about 8% of human and chimpanzees orthologous loci (3 over 36) have experienced drastic changes in *GC**. We propose that these changes are a consequence of variations in recombination rates.

What Are the Factors Influencing Recombination?

Numerous theoretical models investigated the selective forces acting on crossover rates in genomes (Birdsell and Wills 2003), and answers to this important question are still debated. Also, little is known about the

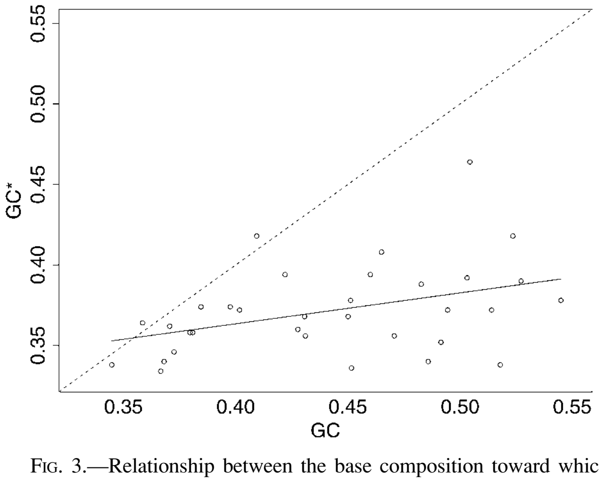

FIG. 3.—Relationship between the base composition toward which a locus is evolving (*GC**) and its present base composition (*GC*). N = 33 loci from 12 human autosomes ($r^2 = 0.16$, $P < 5.10^{-9}$, Student *t*-test). The first diagonal (slope = 1) is indicated by a dashed line. The slope of the correlation between *GC* and *GC** significantly differs from unity (Student *t*-test, $P < 5.10^{-9}$). In GC-poor loci, *GC* and *GC** are quite similar, which indicates equilibrium in base composition. In GC-rich loci, *GC** is lower than *GC*: the current base composition is far from equilibrium.

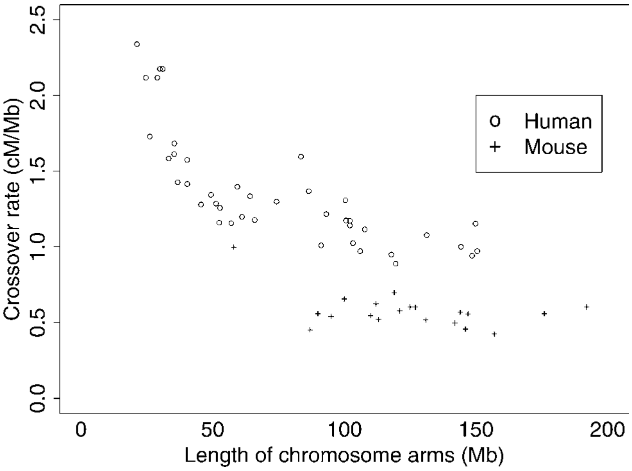

FIG. 4.—Crossover rate (cM/Mb) of human and mouse chromosome arms as a function of their length (Mb). The human karyotype consists of 41 chromosomal arms (18 metacentric and 5 acrocentric chromosomes), ranging from 21 to 150 Mb in length (on average, 74.5 ± 39 Mb). The mouse karyotype (*Mus musculus*) consists of 20 chromosomal arms (all chromosomes are acrocentric), ranging from 58 to 192 Mb in length (on average, 1,24.6 ± 32 Mb). Y chromosomes are not included.

mechanistic determinants of recombination (Kong et al. 2002; de Massy 2003). It has been recently suggested that there is a requirement of at least one crossover per chromosome arm per meiosis (Pardo-Manuel de Villena and Sapienza 2001). In agreement with this model, we observed a strong negative correlation between crossover rate and the length of human chromosomal arms (fig. 4). As their length increases, the crossover rate converges to a basal value. Thus, this correlation holds only for short chromosome arms. The karyotype of mice (*Mus musculus*) consists of 20 chromosomal arms (all chromosomes are acrocentric), ranging from 58 to 192 Mb (on average, 1,24.6 ± 32 Mb). Mouse chromosome arms are therefore all relatively long (only one is shorter than 85 Mb), and as a consequence there is a much smaller inter-chromosomic variability in crossover rate in mouse than in human (fig. 4). Thus, relatively homogeneous recombination rates could well explain the observed homogenization of GC-content (often referred to as the “minor shift”) in the genome of mice as compared to human (Mouchiroud, Gautier, and Bernardi 1988; Mouchiroud et al. 1991; Waterston et al. 2002). Furthermore, the genome of chicken, which presents a very heterogenous karyotype (9 macro-chromosomes and 30 micro-chromosomes) and hence a very strong inter-chromosomal variation of crossover rate (from 3.3 cM/Mb in average in macro-chromosomes, to 8.3 cM/Mb in average in minichromosomes; Burt 2002), is also characterized by an extreme heterogeneity of base composition (Bernardi 2000). Thus, the evolution of karyotypes (number and length of chromosomal arms) is probably an important determinant of the evolution of genomic base composition.

Concluding Remarks

Our analysis is restricted to primates, but the results presented here are likely to apply for a wide range of organisms. Indeed, GC-biased DNA repair has been demonstrated in yeasts and vertebrates and conjectured in most organisms, possibly reflecting an adaptation to frequent GC → AT mutations (Birdsell 2002; Marais 2003). Moreover, positive correlations between recombination rate and GC-content have been observed not only in mammals but also in yeast, Drosophila, and nematode (Gerton et al. 2000; Marais, Mouchiroud, and Duret 2001; Birdsell 2002; Kong et al. 2002; Piganeau and Marais 2002). We therefore expect recombination to drive base composition in genomes of a broad range of sexual organisms with GC-biased DNA repair. Such an evolutionary mode for base composition profoundly modifies our views of neutral evolution. Indeed, it violates fundamental population genetic statements, stipulating that under neutral evolution, the substitution rate equals the mutation rate and that the dynamic of neutral alleles depends only on genetic drift (Otto 2000). These assumptions are the basis of many tests for selection extensively used in the literature (Otto 2000). Thus, this study highlights the importance of neutral side-effects associated with recombination on genome evolution (Marais 2003) and the need to incorporate them into current evolutionary models to accurately detect natural selection within genomes.

Acknowledgments

We thank Adam Eyre-Walker, Dominique Mouchiroud, Nicolas Galtier, Manolo Gouy, Bernard de Massy, and Gabriel Marais for valuable comments and discussions, and Peter Arndt for kindly providing useful software. We are also very grateful to the members of various projects that generated human, chimpanzee, and baboon sequences available in GenBank. This work was supported by the Centre National de la Recherche Scientifique.

Literature Cited

Arndt, P. F., C. B. Burge, and T. Hwa. 2003. DNA sequence evolution with neighbor-dependent mutation. J. Comput. Biol. **10**:313–322.

Arndt, P. F., D. A. Petrov, and T. Hwa. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. Mol. Biol. Evol. **20**:1887–1896.

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

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

Birdsell, J. A., and C. Wills. 2003. The evolutionary origin and maintenance of sexual recombination: a review of contemporary models. Evol. Biol. **33**:27–38.

Blat, Y., R. U. Protacio, N. Hunter, and L. Kleckner. 2002. Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. Cell **111**:791–802.

Burt, D. W. 2002. Origin and evolution of avian microchromosomes. Cytogenet. Genome Res. **96**:97–112.

Charlesworth, B. 1994. Genetic recombination: patterns in the genome. Curr. Biol. **4**:182–184.

de Massy, B. 2003. Distribution of meiotic recombination sites. Trends Genet. **19**:514–522.

Depaulis, F., L. Brazier, S. Mousset, A. Turbe, and M. Veuille. 2000. Selective sweep near the In(2L)t inversion breakpoint in an African population of Drosophila melanogaster. Genet. Res. **76**:149–158.

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

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

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

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.

Gerton, J. L., J. DeRisi, R. Shroff, M. Litchen, P. O. Brown, and T. D. Petes. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc. Natl Acad. Sci. USA **97**:11383–11390.

Hellmann, I., I. Ebersberger, S. Ptak, S. Pbo, and M. Przeworski. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. **72**:1527–1535.

Holn, M., S. Kurtz, and E. Ohlebusch. 2002. Efficient multiple genome alignment. Bioinformatics **S1**:S312–S320.

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.

Jeffreys, A. J., L. Kauppi, and R. Neumann. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. **29**: 217–222.

Jeffreys, A. J., and R. Neumann. 2002. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat. Genet. **31**:267–271.

Kleckner, N. 1996. Meiosis: how could it work ? Proc. Natl. Acad. Sci. USA **93**:8167–8174.

Kong, A., D. F. Gudbjartsson, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, et al. 2002. A high-resolution recombination map of the human genome. Nat. Genet. **31**:241–247.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen, and J. M. Braverman. 2000. Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the Drosophila melanogaster X cromosome. Genetics **156**:1837–1852.

Lercher, M. J., and L. D. Hurst. 2002. Human SNP variability and mutation rate are higher in region of high recombination. Trends Genet. **18**:337–340.

Li, W., S. Yi, and K. Makova. 2002. Male-driven evolution. Curr. Opin. Genet. Dev. **12**:650–656.

Marais, G. 2003. Biaised gene conversion: implications for genome and sex evolution. Trends Genet. **19**:330–338.

Marais, G., D. Mouchiroud, and L. Duret. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl Acad. Sci. USA **98**:5688–5692.

Montoya-Burgos, J. I., P. Boursot, and N. Galtier. 2003. Recombination explains isochores in mammalian genomes. Trends Genet. **19**:128–130.

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.

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

Nagylaki, T. 1983. Evolution of a finite population under gene conversion. Proc. Natl Acad. Sci. USA **80**:6278–6281.

Navarro, A., and N. H. Barton. 2003. Chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes. Science **300**:321–324.

Nekrutenko, A., and W. H. Li. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. **10**:1986–1995.

Otto, S. P. 2000. Detecting the form of selection from DNA sequence data. Trends Genet. **16**:526–529.

Pardo-Manuel de Villena, F., and C. Sapienza. 2001. Female meiosis drives karyotypic evolution in mammals. Genetics **159**:1179–1189.

Petes, T. D., and J. D. Merker. 2002. Context dependence of meiotic recombination hotspots in yeast. The relationship between recombination activity of a reporter construct and base composition. Genetics **162**:2049–2052.

Piganeau, G., and G. Marais. 2002. Hill-Robertson interference is a minor determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. Mol. Biol. Evol. **19**:1399–1406.

Reich, D. E., S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin, J. C. Mullikin, J. M. Higgins, D. J. Richter, E. S. Lander, and D. Altshuler. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. Nat. Genet. **32**:135–142.

Rogers, J., M. C. Mahaney, S. M. Witte, S. Nair, D. Newman, S. Wedel, L. A. Rodriguez, K. S. Rice, S. H. Slifer, A. Perelygin, et al. 2000. A genetic linkage map of the baboon (Papio hamadryas) genome based on human microsatellite polymorphisms. Genomics **67**:237–247.

Smit, A. F. A. 1999. Interspersed repeats and others mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. **9**:657–663.

Sueoka, N. 1962. On the genetic basis of variation and heterogeneity of dna composition. Proc. Natl. Acad. Sci. USA **48**:582–592.

True, J. R., J. M. Mercer, and C. C. Laurie. 1996. Differences in cross-over frequency and distribution among three sibling species of Drosophila. Genetics **142**:507–523.

Watanabe, Y., A. Fujiyama, Y. Ichiba, M. Hattori, T. Yada, Y. Sakaki, and T. Ikemura. 2002. Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. Hum. Mol. Genet. **11**:13–21.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–562.

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

Kenneth H. Wolfe, Associate Editor

Accepted November 26, 2003