

# Male-Driven Biased Gene Conversion Governs the Evolution of Base Composition in Human Alu Repeats

Matthew T. Webster,\* Nick G. C. Smith,† Lina Hultin-Rosenberg,\*  
Peter F. Arndt,‡ and Hans Ellegren\*

\*Department of Evolution, Genomics and Systematics, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden;  
†Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom; and ‡Max Planck Institute for Molecular Genetics, Berlin, Germany

Regional biases in substitution pattern are likely to be responsible for the large-scale variation in base composition observed in vertebrate genomes. However, the evolutionary forces responsible for these biases are still not clearly defined. In order to study the processes of mutation and fixation across the entire human genome, we analyzed patterns of substitution in Alu repeats since their insertion. We also studied patterns of human polymorphism within the repeats. There is a highly significant effect of recombination rate on the pattern of substitution, whereas no such effect is seen on the pattern of polymorphism. These results suggest that regional biases in substitution are caused by biased gene conversion, a process that increases the probability of fixation of mutations that increase GC content. Furthermore, the strongest correlate of substitution patterns is found to be male recombination rates rather than female or sex-averaged recombination rates. This indicates that in addition to sexual dimorphism in recombination rates, the sexes also differ in the relative rates of crossover and gene conversion.

## Introduction

The causes and significance of the large-scale variation in base composition (GC content) observed in vertebrate genomes are controversial issues (Eyre-Walker and Hurst 2001). This variation is characterized by high heterogeneity between genomic regions and by long stretches of contiguous sequence with relatively homogeneous GC content commonly referred to as isochores (Filipowski, Thierry, and Bernardi 1973; Bernardi et al. 1985; Nekrutenko and Li 2000; IHGSC 2001). In addition to regionality in base composition, vertebrate genomes also exhibit significantly nonrandom organization and complex interrelationships with regard to a number of other genomic features such as gene density, gene expression patterns, chromosome banding patterns, repeat density, and recombination (Mouchiroud et al. 1991; Saccone et al. 1993; Caron et al. 2001; IHGSC 2001; Lercher, Urrutia, and Hurst 2002; Niimura and Gojobori 2002; Lercher et al. 2003; Pal and Hurst 2003; Vinogradov 2003). All these features show significant covariation with GC content, suggesting that isochores reflect a fundamental feature of genome organization, although the causes of these relationships are not clear. An understanding of the evolution of base composition is therefore likely to shed light on the evolution of a number of other features of genome organization.

A highly significant correlation between recombination and GC content has been reported in the human genome (Fullerton, Bernardo Carvalho, and Clark 2001; Kong et al. 2002) and in several other eukaryotes (Birdsell 2002), indicating that it is a widespread phenomenon. One suggestion is that this relationship is generated by high GC content being recombinogenic (Gerton et al. 2000). However, recent work indicates that this correlation has arisen because recombination drives the evolution of GC content (Galtier et al. 2001; Galtier 2003). Hence, it is possible that

the observed variation in GC content is the result of variation in recombination rate. This is supported by the observation that part of the mouse  $F_{XY}$  gene underwent a rapid increase in GC content following translocation into a highly recombining pseudoautosomal region (Montoya-Burgos, Boursot, and Galtier 2003). Additionally, a recent study of human-chimpanzee noncoding alignments showed that the relationship between recombination and the GC content toward which a sequence is evolving ( $GC^*$ ) is stronger than that between recombination and current GC content (Meunier and Duret 2004).

Recombination could bias the substitution pattern toward increased GC content in two main ways: either by a direct mutagenic effect or by a side effect caused by biased gene conversion (BGC). In vitro evidence suggests that recombination-mediated mismatch repair is GC biased (Brown and Jiricny 1988), and it is believed that repair of heteroduplexes formed by gene conversion is also GC biased (Birdsell 2002; Marais 2003). This process results in a bias toward fixation of GC alleles, similar to the effects of weak directional genic selection in a random mating population (Nagylaki 1983; Marais, Charlesworth, and Wright 2004). However, it has also been suggested that recombination events generate mutations (Lercher and Hurst 2002; Hellmann et al. 2003), raising the possibility that such mutations are also biased toward increasing GC content. An alternative possibility is that another factor correlated with recombination could lead to regional biases in the mutation process (Wolfe, Sharp, and Li 1989).

The existence of fixation biases can be inferred by comparing patterns of polymorphism and divergence. The pattern of nucleotide changes observed as substitutions between species reflects the combined action of mutation and fixation processes. Patterns of nucleotide changes in polymorphism data should, however, more closely reflect biases in mutation. Webster, Smith, and Ellegren (2003) showed that  $GC^*$  predicted on the basis of human polymorphism data (referred to here as  $GC^*_{poly}$ ) is significantly lower than  $GC^*$  predicted on the basis of human-chimpanzee divergence data (referred to here as  $GC^*_{div}$ ), indicating a

Key words: isochore, base composition, Alu, mutation, recombination, SNP.

E-mail: matthew.webster@ebc.uu.se.

*Mol. Biol. Evol.* 22(6):1468–1474. 2005

doi:10.1093/molbev/msi136

Advance Access publication March 16, 2005

bias toward fixation of GC alleles. Other studies have shown that mutations from A or T to G or C (AT → GC) segregate at significantly higher frequencies than the opposite (GC → AT) type (Duret et al. 2002; Webster and Smith 2004), supporting this conclusion. However, so far the effects of potential mutational biases caused by recombination have not been studied. Furthermore, the presence of fixation biases has not been demonstrated on a genome-wide scale.

Analysis of substitution patterns in interspersed repeats is a valuable method for understanding regional variation in patterns of mutation and fixation. It allows analysis of large numbers of sequences throughout the genome. In addition, as the inserted sequences are essentially the same, differences in their patterns of nucleotide substitution must be due to the regional effects of where they are inserted. A number of studies have demonstrated biases in sequences inserted in new genomic locations (Filipski, Salinas, and Rodier 1989; Casane et al. 1997; IHGSC 2001). By analyzing substitutions in interspersed repeats of a variety of ages, Arndt, Petrov, and Hwa (2003) demonstrated a weakening of the dependence of patterns of substitution on surrounding GC at the time of the mammalian radiation, suggesting a change from an isochore-preserving to an isochore-decaying pattern in mammalian lineages.

A particularly troublesome methodological problem with inferring patterns of substitution in human repeats is the neighbor dependence of the substitution process, which violates the assumption of independence of sites required by single-nucleotide models with  $4 \times 4$  DNA transition matrices. In particular, the mutation rate at CpG sites is elevated due to methylated cytosine mutagenesis, so the state of one site clearly affects the evolution of its neighbor. Here, we use a maximum-likelihood (ML) implementation of a dinucleotide model developed to account for neighbor-dependent substitution rates (Arndt, Burge, and Hwa 2003; Arndt, Petrov, and Hwa 2003; Arndt and Hwa 2005). Using this ML approach, we have analyzed patterns of substitution in a genome-wide sample of Alu repeats and compared them to patterns of polymorphism, inferred using a combination of the ML approach and parsimony, also accounting for CpG effects. This comparison allows insight into the evolutionary forces that determine base composition across the entire genome.

## Methods

### Identification of Alu Repeats

We downloaded “build 30” of the human genome sequence, which comprises 2,812 Mb of DNA divided among 1,388 contigs. Files containing sequence data and the associated GenBank annotation files were downloaded from National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>). The annotation files were used to determine the position of genes within each contig. Alu repeats were identified using RepeatMasker (<http://www.repeatmasker.org>), which was also used to produce alignments of each Alu repeat with its ancestral sequence as identified in RepBase (Jurka 2000). Only repeats in nongenic regions were considered to eliminate potential biases caused by selection in transcribed regions. Furthermore, we used only autosomal contigs, as repeats on sex chromo-

somes are likely to be subject to different evolutionary pressures relative to the rest of the genome. Contigs were divided into segments with boundaries situated at the midpoints between the locations of markers for which recombination rate estimates were available from the deCODE map (Kong et al. 2002). If only one marker was available in a particular contig, then the entire contig was considered as a single segment. Contigs containing no markers with recombination rate estimates or segments with missing information in the deCODE map were discarded. For each segment, the average GC content at nongenic DNA not masked by RepeatMasker was calculated (those with less than 1,000 bp were discarded).

### Analysis of Substitution Patterns

We produced concatenated alignments of Alu repeats and their ancestral sequences within nongenic regions of each remaining contig segment, dividing the data by the three main classes of Alu repeat (AluY, AluS, and AluJ). Hence, three long alignments were derived from each segment, consisting of all the ancestral and derived repeat sequences of each Alu class. Only alignments containing >20 kb of repeat sequence were retained for further analysis. In order to minimize misinference caused by substitutions occurring in Alu repeats prior to insertion, we masked all “diagnostic” sites (those that differ between Alu classes) from our analysis. This was done by aligning all master sequences in ClustalW (Thompson, Higgins, and Gibson 1994) and identifying diagnostic sites in this alignment.

Following Arndt, Petrov, and Hwa (2003), we assumed a “star” phylogeny for each subfamily of Alu repeats. This assumption implies that the master sequence identified by RepeatMasker is the true ancestral sequence that was inserted at any particular position. All the differences between the present-day and ancestral sequence are assumed to represent independent substitutions. The “master gene” model is the most widely accepted model of Alu proliferation and is supported by comprehensive phylogenetic analyses of Alu subfamilies (Britten et al. 1988; Deininger et al. 1992; Batzer and Deininger 2002). We cannot rule out the possibility that a small proportion of Alu repeats act as secondary source elements after insertions (Cordaux et al. 2004). However, masking the diagnostic sites eliminates any misinference caused by secondary amplification of Alu sequences that are intermediate between known master copies.

We estimated the pattern of substitutions since the insertion of each repeat by analyzing the alignments of repeat and ancestral sequences using an ML approach with a dinucleotide substitution model (see *Introduction*). This method estimates the frequency of seven different types of substitution events (four single-nucleotide transversion rates, two single-nucleotide transition rates, and a CpG transition rate) in any given alignment. For each alignment, the equilibrium GC content predicted by the divergence pattern,  $GC_{div}^*$ , was estimated by applying the seven substitution rates to the dinucleotide model. The equilibrium GC content of a genomic region is the stationary GC content toward which the sequence is evolving given the present-day GC content and substitution pattern.

Arndt, Petrov, and Hwa (2003) used the ML dinucleotide substitution model to reconstruct the ancestral sequences of all human retrotransposons. The reconstructed sequences were found to match the RepBase sequences with >99% accuracy. This indicates that our assumption of a star phylogeny and the dinucleotide substitution model are reliable representations of Alu evolution.

### Analysis of Human Single-Nucleotide Polymorphisms Within Alu Repeats

Single-nucleotide polymorphisms (SNPs) within AluY repeats were identified from the GenBank contig annotation files. For every repeat containing known SNPs, each biallelic SNP was compared with the base at the corresponding position in the ancestral sequence using the alignments created by RepeatMasker. We then inferred the direction of the mutation resulting in each SNP using simple parsimony, assuming that the ancestral state of each SNP is represented by the base in the ancestral repeat sequence. We only used AluY repeats because they were inserted the most recently and hence are the most reliable for predicting the ancestral base of human SNPs. As substitution events can cause parsimony to misinfer the direction of mutations, we checked the simple parsimony analysis by developing a weighted parsimony method based on ML inference of ancestral states. The weighted parsimony method gave results almost identical to those of the simple parsimony method (data not shown), and so only the simple parsimony results are presented here.

We estimated the relative rates of the seven different types of polymorphism events (four transversion rates, two transition rates, and a CpG transition rate) by first inferring the mutation leading to each SNP using parsimony at each individual site. We then used the dinucleotide substitution model to estimate  $GC_{poly}^*$ , the equilibrium GC content predicted by the mutation pattern derived from SNPs within each segment. Only segments containing 100 or more SNPs within AluY repeats for which the ancestral states could be inferred were included in the analysis.

### Statistics

Confidence intervals (CIs) for correlation coefficients relating patterns of substitution and other sequence characteristics were estimated using nonparametric bootstrapping. In this procedure, the data set was resampled with replacement 1,000 times, treating each chromosome as an individual data point, and the correlation coefficient was recalculated for each resampling. The data were grouped by chromosome rather than segment because neighboring segments cannot be considered fully independent. In order to determine which factors have the strongest effect on  $GC_{div}^*$ , we used partial correlation coefficients. Partial correlations consider the correlations between pairs of variables while holding the value of each of the other variables constant.

### Results

We divided human autosomal contigs into segments for which recombination rate estimates were available and analyzed patterns of polymorphism and divergence

**Table 1**  
Summary of Data Sets

	AluY	AluS	AluJ
Number of segments	3,799	3,843	3,819
Mean genomic segment length/kilobases (SD) <sup>a</sup>	598 (467)	592 (468)	595 (467)
Mean alignment length per genomic segment/kilobases (SD)	8.5 (9.0)	36.3 (42.9)	14.5 (17.4)
Total aligned bases/kilobases	32,469	139,339	55,214

<sup>a</sup> SD, standard deviation.

in Alu repeats within these segments by aligning the repeats with their ancestral sequences. A summary of the data set is given in table 1. All Alu subtypes are GC rich (average GC content of Alu sequences in RepBase is 56.8%) and are relatively rich in CpG sites: on average there are 23.3 CpGs in an average length of 308.8 bp. This represents an observed/expected ratio of 0.94, much higher than the base level of an observed/expected ratio of 0.2 in most of the human genome (IHGSC 2001).

The average transversion frequencies of these three types of repeat are 0.0080 for AluY, 0.0124 for AluS, and 0.0208 for AluJ. Given that a transversion frequency of 0.01 corresponds to roughly 35 MYA (Arndt, Petrov, and Hwa 2003), the substitution data indicate average insertion times of 28 MYA for AluY, 43 MYA for AluS, and 73 MYA for Alu J. The pattern of substitution indicates a general trend for reduction in GC content in all Alu classes. The  $GC_{div}^*$  values obtained from each Alu type are significantly different: mean  $GC_{div}^*$  is 39.9% (bootstrap 95% CIs, 39.4%–40.4%) for AluY, 34.7% (34.4%–35.1%) for AluS, and 33.6% (33.2%–34.0%) for AluJ. Given the large differences between the mean insertion times of the three Alu types, such differences in  $GC_{div}^*$  can be attributed to changes in patterns of mutation and/or fixation over evolutionary time-scales. One clear difference between substitutions in AluY and the two older Alu types is in the ratio of the rate of CpG transitions to the rate of transversions: 32.7 for AluY, 44.3 for AluS, and 43.2 for AluJ.

There are strong positive correlations between GC content and  $GC_{div}^*$  for AluY (Pearson's  $r = 0.503$ , bootstrap 95% CIs, 0.34%–0.61%), AluS ( $r = 0.632$ , 0.58%–0.68%), and AluJ ( $r = 0.680$ , 0.61%–0.74%). These figures are higher than, but roughly comparable to, the correlation between GC and  $GC_{div}^*$  of  $r = 0.40$  reported by Meunier and Duret (2004) on the basis of just 33 data points. For all Alu types, repeats in regions of high GC are decaying in GC content (i.e.,  $GC_{div}^* < GC$ ), indicating that the human genome isochore structure is becoming homogenized, in concordance with a number of recent studies (Duret et al. 2002; Smith, Webster, and Ellegren 2002; Arndt, Petrov, and Hwa 2003; Webster, Smith, and Ellegren 2003). As AluY repeats are the youngest class, substitutions in these repeats should most accurately correspond to the present substitution pattern in the human lineage. The correlation between GC and  $GC_{div}^*$  using substitutions in AluY repeats is shown in figure 1. The gradient of the regression line is much less than expected under compositional equilibrium. If we consider substitutions in AluY repeats (fig. 1), it appears that repeats in low GC regions have a higher

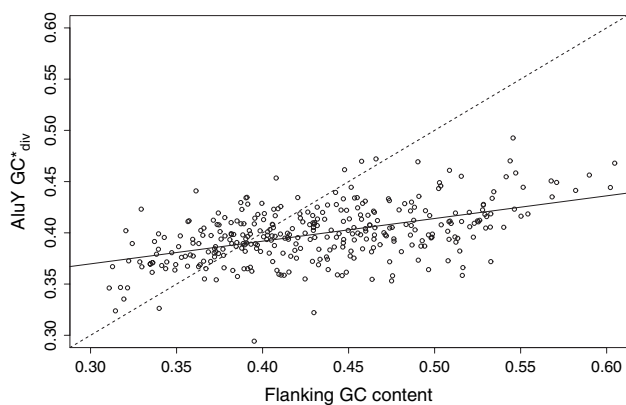


FIG. 1.—Positive correlation between  $GC_{div}^*$  in AluY repeats and surrounding GC content ( $r = 0.503$ ,  $P < 0.0001$ ). The dotted line represents the expected slope under compositional equilibrium ( $GC = GC_{div}^*$ ).

$GC_{div}^*$  than surrounding GC content. However, for AluS and AluJ, which have lower average  $GC_{div}^*$ , it appears that GC content is decreasing in all regions.

We investigated the positive correlation between recombination rate and nonrepetitive nongenetic GC content in all the contig segments retained for analysis (table 2). As reported previously (Kong et al. 2002), GC content is correlated with sex-averaged crossover rates in all data sets. However, the GC-recombination correlations in our data sets (table 2) are weaker than that presented by Kong et al. (2002), where  $r^2$  is reported to be 0.15 for sex-averaged recombination rates ( $r = 0.39$ ). This discrepancy may be due to the effects of pooling data (pooled data of Kong et al. within nonoverlapping 3-Mb windows, whereas we considered segments for which single-recombination estimates were available) or simply because the coverage of the human genome is incomplete for our analyses due to restrictions on the minimum amounts of Alu sequence per contig segment. In all three data sets, we also find that male recombination is a stronger correlate of GC content than female recombination (table 2). In order to eliminate chromosomal effects as the cause of the correlations between recombination and GC or  $GC_{div}^*$ , we recalculated all the correlation coefficients shown in table 2 for individual chromosomes. In all cases, similar trends were observed, indicating that such effects are not solely responsible for the observed variation in GC and  $GC_{div}^*$  (supplementary table 1, Supplementary Material online).

When the correlations between average recombination rate and GC content are compared to the correlations between recombination and  $GC_{div}^*$ , the correlation with  $GC_{div}^*$  is stronger in eight out of nine cases (table 2). A clearer picture is obtained by using partial correlation coefficients. We compared the partial correlation coefficient of  $GC_{div}^*$  versus average recombination rate controlling for GC with the partial correlation coefficient of GC versus average recombination rate controlling for  $GC_{div}^*$  (table 3). In all three cases, the  $GC_{div}^*$ -recombination partial correlation coefficient is significantly greater than the GC-recombination partial correlation coefficient. These results confirm the interpretation of Meunier and Duret (2004) that recombination drives the evolution of GC content and not the other way round.

**Table 2**  
**Correlation Coefficients (Pearson's  $r$ ) Between Recombination Rate and Measures of GC and  $GC_{div}^*$  (95% CIs Are Bootstrap Derived)**

	Sex-Averaged Crossover Rate	Female Crossover Rate	Male Crossover Rate
AluY			
GC	0.344 (0.18–0.47)	0.186 (0.02–0.32)	0.388 (0.21–0.49)
$GC_{div}^*$	0.422 (0.31–0.51)	0.243 (0.12–0.34)	0.434 (0.34–0.48)
AluS			
GC	0.334 (0.25–0.41)	0.250 (0.15–0.35)	0.297 (0.23–0.36)
$GC_{div}^*$	0.392 (0.32–0.46)	0.235 (0.16–0.30)	0.376 (0.30–0.46)
AluJ			
GC	0.269 (0.15–0.37)	0.189 (0.07–0.29)	0.277 (0.17–0.36)
$GC_{div}^*$	0.393 (0.25–0.50)	0.218 (0.09–0.33)	0.409 (0.32–0.49)

The partial correlation between  $GC_{div}^*$  and recombination rate accounting for GC effects is also stronger for male than for female recombination rates (table 4). This difference is significant for AluS and AluJ and suggestive for AluY. Note, however, that significance is calculated by resampling by chromosome (see *Methods*), which is highly conservative, and all correlations are significant when resampling by chromosome segment (AluY,  $P = 0.038$ ; AluS,  $P < 0.001$ ; AluJ,  $P < 0.001$ ). This contrasts with the result of Meunier and Duret (2004) who found that female recombination had a stronger effect on  $GC_{div}^*$  than male recombination, although the sample size was small ( $n = 33$ ), and the result was sensitive to removal of a single datum. Our results indicate that the effect of recombination on GC content is male driven.

In order to determine whether the correlation between recombination and  $GC_{div}^*$  is due to a mutagenic effect or a consequence of fixation bias, we compared patterns of divergence, which reflect the combined effects of mutation and selection, with patterns of polymorphism, where the influence of selection is lower. We first compared  $GC_{poly}^*$  and  $GC_{div}^*$  in AluY repeats contained within the same contig segment. AluY sequences are the youngest and therefore are most likely to represent present-day substitution patterns. They are also the most reliable for estimating the ancestral state of human SNPs. The number of segments containing sufficient polymorphism data in AluY repeats was low (55 segments with  $>100$  SNPs). Due to the small size of the data set, we simply report the Pearson correlation coefficient for analyses involving  $GC_{poly}^*$ , which assumes that all data points are independent. There is a significant correlation between  $GC_{div}^*$  minus  $GC_{poly}^*$  and the

**Table 3**  
**Partial Correlation Coefficients (Pearson's  $r$ ) Between Sex-Averaged Crossover Rates Controlling for GC and  $GC_{div}^*$  (95% CIs Are Bootstrap Derived)**

	AluY	AluS	AluJ
$GC_{div}^*$ ct. GC (95% CI)	0.306 (0.22–0.39)	0.247 (0.17–0.31)	0.298 (0.17–0.40)
GC ct. $GC_{div}^*$ (95% CI)	0.169 (0.02–0.28)	0.122 (0.03–0.16)	0.022 (–0.07–0.08)
$P(GC_{div}^*$ ct. GC) > (GC ct. $GC_{div}^*$ )	0.029	0.014	<0.001

NOTE.—ct., controlling for.

**Table 4**  
**Partial Correlation Coefficients (Pearson's  $r$ ) Between Sex-Specific Crossover Rates and  $GC_{div}^*$  Controlling for GC (95% CIs Are Bootstrap Derived)**

	AluY	AluS	AluJ
Female crossover rate	0.176 (0.06–0.30)	0.103 (0.05–0.17)	0.125 (0.02–0.23)
Male crossover rate	0.300 (0.21–0.37)	0.255 (0.17–0.35)	0.312 (0.21–0.40)
$P$ (male > female)	0.06	0.003	0.007

NOTE.— $P$  is the probability that the correlation between  $GC_{div}^*$  and male crossover rates is stronger than the correlation between  $GC_{div}^*$  and female crossover rates.

sex-averaged recombination rate (fig. 2;  $r = 0.301$ ,  $P = 0.026$ ). If the correlation between  $GC_{div}^*$  and recombination was caused by a biased mutation pattern, then we would expect to see the same effect in  $GC_{poly}^*$ , and hence, this correlation would not be significant. The correlation between  $GC_{div}^*$  and recombination is therefore not solely due to a mutational bias. The difference between  $GC_{div}^*$  and  $GC_{poly}^*$  is most strongly correlated with male-specific recombination ( $r = 0.448$ ,  $P = 0.001$ ), but there is no correlation with female-specific recombination ( $r = 0.048$ ,  $P = 0.727$ ).

We next examined the relationship between recombination rate and  $GC_{poly}^*$  estimated from SNPs in AluY repeats. There is no significant correlation between recombination and  $GC_{poly}^*$  for both sex-averaged rates ( $r = -0.102$ ,  $P = 0.458$ ) and female-specific rates ( $r = 0.080$ ,  $P = 0.557$ ). Male-specific recombination rates are negatively correlated with  $GC_{poly}^*$  ( $r = -0.296$ ,  $P = 0.028$ ). This indicates that the positive correlation between  $GC_{div}^*$  and recombination is unlikely to be due to variation in the underlying mutation pattern.

## Discussion

The results presented here confirm earlier reports that GC-rich isochores are vanishing in the human lineage. This has previously been shown on the basis of human-chimpanzee-baboon noncoding alignments (Smith, Webster, and Ellegren 2002; Webster, Smith, and Ellegren 2003; Meunier and Duret 2004), substitutions in interspersed repeats (Arndt, Petrov, and Hwa 2003), and alignments of coding regions from multiple species (Duret et al. 2002). It has been suggested that previous results suffer from a bias due to inaccuracies with using simple parsimony on sequences with extreme GC contents (Eyre-Walker 1998; Alvarez-Valin et al. 2004). However, such biases are only likely to be important when using coding sequences, where third-codon positions can reach very high GC contents (often >80%). GC content in noncoding DNA and interspersed repeats in mammalian genomes rarely reaches such extreme values and therefore simple parsimony should reliably estimate the substitution pattern. Furthermore, misaligned bases are not expected to result in large misinferences in substitution pattern unless GC content is highly skewed. In this study, we have also implemented an ML dinucleotide substitution model that incorporates variation in rates between different classes of

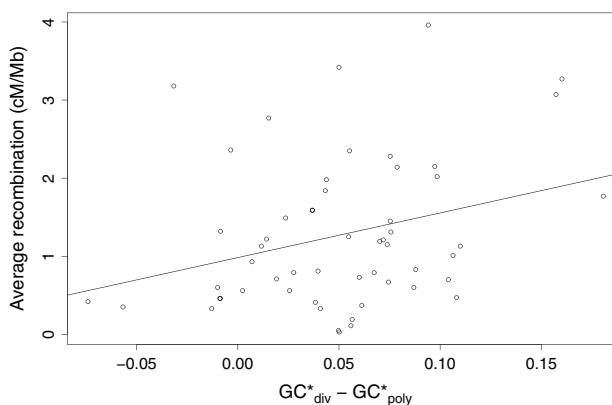


FIG. 2.—Positive correlation between average recombination rate and  $GC_{div}^* - GC_{poly}^*$  within Alu repeats ( $r = 0.301$ ,  $P = 0.026$ ).

substitution and accounts for neighbor effects caused by CpG hypermutability (Arndt, Burge, and Hwa 2003). This adds to the now strong evidence to support the homogenization of isochores along the human lineage since the mammalian radiation.

We find that the correlation between crossover rates and  $GC_{div}^*$  in Alu repeats is stronger than the correlation between crossover rates and present-day GC content in nongenic nonrepetitive DNA, accounting for the correlation between GC and  $GC_{div}^*$ . This is in concordance with the findings of Meunier and Duret (2004) and indicates that recombination drives GC content. The potential mechanisms by which recombination could affect GC content can be divided into two categories: direct mutagenic effects or fixation biases. In order to determine whether recombination results in biases in the process of mutation or fixation, we compared patterns of substitution in Alu repeats with patterns of polymorphism. We found no correlation between crossover rates and  $GC_{poly}^*$ , estimated from patterns of polymorphism within Alu repeats. This suggests that recombination does not influence the pattern of mutation but increases the probability of fixation of GC alleles. This is consistent with the action of BGC.

The correlations between  $GC_{div}^*$  in all classes of Alu repeat and recombination rates are stronger with male than with female recombination. Hence, it is mainly male recombination that governs the evolution of GC content in Alu repeats. Assuming that BGC is the process by which recombination influences GC content, this could suggest that a greater proportion of male recombination events are resolved as gene conversions. Alternatively, male crossover rates could be a better predictor of gene conversion rates than female rates, i.e., the mechanisms of gene conversion and crossing-over may be indirectly, rather than directly, related. Sex dimorphism in recombination is a widespread phenomenon (Lenormand 2003). These results suggest that sexes may differ in both rates and modes of recombination events. Our results may also be in part a consequence of a greater variation in male than female recombination rates. The full genomic data set of Kong et al. (2002) indicates that although the mean female recombination rate is much greater than the mean male recombination rate (1.54 vs. 0.97 cM/Mb), the standard deviation of female recombination rates is lower (0.79 vs. 0.99). The greater

the variation in recombination rates, the greater the potential to affect other features, such as base composition, which vary across the genome.

An interesting finding of this study is that estimates of  $GC_{div}^*$  from different Alu repeats are not the same, with the oldest Alu repeats exhibiting the most AT-biased substitution pattern. This suggests that substitution biases are changing over time, which could potentially be caused by changes in average recombination rate. The human linkage map is twice as large as mouse and rat (Jensen-Seaman et al. 2004) and 28% longer than baboon (Rogers et al. 2000). If recombination drives GC content, then an increase in the average recombination rate in the human lineage could explain why younger repeats have a higher  $GC_{div}^*$ . However, the major difference between evolution in younger AluY elements and older AluS and AluJ elements seems to be a decrease in CpG mutability. This suggests that levels of methylation on the human lineage may have decreased.

In this study, we have used the equilibrium GC content ( $GC^*$ ) to summarize patterns of mutation and fixation in different genomic regions. However, molecular processes may have variable effects on individual mutational classes, and in future studies, it may be valuable to consider the entire spectrum of changes (Arndt, Petrov, and Hwa 2003; Webster and Smith 2004). The evolution of base composition is likely to be a dynamic process, and regional patterns of substitution may change quickly over evolutionary time in response to changes in the pattern and rate of recombination (Meunier and Duret 2004). Recent reports indicate that changes in recombination rate are surprisingly rapid: known hotspots in humans are absent in chimpanzees (Wall et al. 2003; Ptak et al. 2004), and there is a low correspondence between recombination rates in human and mouse (Jensen-Seaman et al. 2004). This implies that the equilibrium GC content of a genomic region is constantly changing.

Factors that alter the mutation pattern, such as changes in levels of methylation, are also likely to play an important role in the evolutionary dynamics of base composition. Such a change could explain an inferred increase in the CpG mutation rate at the start of the mammalian radiation (Arndt, Petrov, and Hwa 2003). The impact of BGC depends on effective population size and levels of inbreeding (Nagylaki 1983; Marais, Charlesworth, and Wright 2004), which may change rapidly over time, thus further increasing heterogeneity in GC content between species. As isochores appear to be the result of variation in a neutral process, their functional significance is unclear. However, greater understanding of the processes governing the evolution of isochores is likely to provide insights into the evolution and significance of many other related features of genome organization.

## Supplementary Material

Supplementary table 1 is available at *Molecular Biology and Evolution* online ([www.mbe.oupjournals.org](http://www.mbe.oupjournals.org)).

## Acknowledgments

This work was supported by the Swedish Research Council. H.E. is a Royal Swedish Academy of Sciences

Research Fellow supported by a grant from the Knut and Alice Wallenberg Foundation.

## Literature Cited

- Alvarez-Valin, F., O. Clay, S. Cruveiller, and G. Bernardi. 2004. Inaccurate reconstruction of ancestral GC levels creates a "vanishing isochores" effect. *Mol. Phylogenet. Evol.* **31**: 788–793.
- 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., and T. Hwa. 2005. Identification and measurement of neighbor dependent nucleotide substitution processes. *Bioinformatics* (in press).
- 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.
- Batzer, M. A., and P. L. Deininger. 2002. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**:370–379.
- 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.
- 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.
- Britten, R. J., W. F. Baron, D. B. Stout, and E. H. Davidson. 1988. Sources and evolution of human Alu repeated sequences. *Proc. Natl. Acad. Sci. USA* **85**:4770–4774.
- Brown, T. C., and J. Jiricny. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54**:705–711.
- Caron, H., B. van Schaik, M. van der Mee et al. (13 co-authors). 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**:1289–1292.
- Casane, D., S. Boissinot, B. H. Chang, L. C. Shimmin, and W. Li. 1997. Mutation pattern variation among regions of the primate genome. *J. Mol. Evol.* **45**:216–226.
- Cordaux, R., D. J. Hedges, M. A. Batzer, P. L. Deininger, J. V. Moran, and H. H. Kazazian Jr. 2004. Retrotransposition of Alu elements: how many sources? *Trends Genet.* **20**: 464–467.
- Deininger, P. L., M. A. Batzer, C. A. Hutchison III, M. H. Edgell, R. Cordaux, D. J. Hedges, J. V. Moran, and H. H. Kazazian Jr. 1992. Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8**:307–311.
- 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. 1998. Problems with parsimony in sequences of biased base composition. *J. Mol. Evol.* **47**:686–690.
- Eyre-Walker, A., and L. D. Hurst. 2001. The evolution of isochores. *Nat. Rev. Genet.* **2**:549–555.
- Filipski, J., J. Salinas, and F. Rodier. 1989. Chromosome localization-dependent compositional bias of point mutations in Alu repetitive sequences. *J. Mol. Biol.* **206**:563–566.
- Filipski, J., J. P. Thiery, and G. Bernardi. 1973. An analysis of the bovine genome by  $Cs_2SO_4$ -Ag density gradient centrifugation. *J. Mol. Biol.* **80**:177–197.
- 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**:1139–1142.
- 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. Lichten, 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. E. Ptak, S. Paabo, and M. Przeworski. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**:1527–1535.
- IHGSC. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.
- Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C. F. Chen, M. A. Thomas, D. Haussler, and H. J. Jacob. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* **14**:528–538.
- Jurka, J. 2000. RepBase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**:418–420.
- Kong, A., D. F. Gudbjartsson, J. Sainz et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31**:241–247.
- Lenormand, T. 2003. The evolution of sex dimorphism in recombination. *Genetics* **163**:811–822.
- Lercher, M. J., and L. D. Hurst. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18**:337–340.
- Lercher, M. J., A. O. Urrutia, and L. D. Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31**:180–183.
- Lercher, M. J., A. O. Urrutia, A. Pavlicek, and L. D. Hurst. 2003. A unification of mosaic structures in the human genome. *Hum. Mol. Genet.* **12**:2411–2415.
- Marais, G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* **19**:330–338.
- Marais, G., B. Charlesworth, and S. I. Wright. 2004. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. *Genome Biol.* **5**:R45.
- Meunier, J., and L. Duret. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21**:984–990.
- 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.
- Nagyilaki, T. 1983. Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. USA* **80**:6278–6281.
- Nekrutenko, A., and W. H. Li. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res.* **10**:1986–1995.
- Niimura, Y., and T. Gojobori. 2002. In silico chromosome staining: reconstruction of Giemsa bands from the whole human genome sequence. *Proc. Natl. Acad. Sci. USA* **99**:797–802.
- Pal, C., and L. D. Hurst. 2003. Evidence for co-evolution of gene order and recombination rate. *Nat. Genet.* **33**:392–395.
- Ptak, S. E., A. D. Roeder, M. Stephens, Y. Gilad, S. Paabo, and M. Przeworski. 2004. Absence of the TAP2 human recombination hotspot in chimpanzees. *PLoS Biol.* **2**:e155. (Epub June 15, 2004).
- Rogers, J., M. C. Mahaney, S. M. Witte et al. (17 co-authors). 2000. A genetic linkage map of the baboon (*Papio hamadryas*) genome based on human microsatellite polymorphisms. *Genomics* **67**:237–247.
- Saccone, S., A. De Sario, J. Wiegant, A. K. Raap, G. Della Valle, and G. Bernardi. 1993. Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci. USA* **90**:11929–11933.
- Smith, N. G. C., M. T. Webster, and H. Ellegren. 2002. Deterministic mutation rate variation in the human genome. *Genome Res.* **12**:1350–1356.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 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.
- Vinogradov, A. E. 2003. Isochores and tissue-specificity. *Nucleic Acids Res.* **31**:5212–5220.
- Wall, J. D., L. A. Frisse, R. R. Hudson, and A. Di Rienzo. 2003. Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. *Am. J. Hum. Genet.* **73**:1330–1340.
- Webster, M. T., and N. G. C. Smith. 2004. Fixation biases affecting human SNPs. *Trends Genet.* **20**:122–126.
- Webster, M. T., N. G. C. Smith, and H. Ellegren. 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. *Mol. Biol. Evol.* **20**:278–286.
- Wolfe, K. H., P. M. Sharp, and W. H. Li. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337**:283–285.

Pekka Pamilo, Associate Editor

Accepted March 10, 2005