Molecular Evolution of Recombination Hotspots and Highly Recombining Pseudoautosomal Regions in Hominoids

Soojin Yi[1] *and Wen-Hsiung Li*

Department of Ecology and Evolution, University of Chicago

We examined the effects of recombination on the molecular evolution of noncoding regions in pseudoautosomal regions (PARs) and recombination hotspots in hominoids. The PAR-linked regions analyzed had on average longer branch lengths than those of the recombination hotspots. Moreover, contrary to previous observations, we found *no* correlation between recombination rate and silent site divergence in our data set and little change in the GC content during recent hominoid evolution. This suggests that the current rate of recombination is not a good indicator of the past rates of recombination for these highly recombining regions. Furthermore, human recombination hotspots show increased AT to GC substitutions in the human lineage, while no such pattern is detected for PAR-linked regions. Together, these observations suggest that recombination hotspots in hominoids are transient in the evolutionary timescale. Interestingly, the *16p13.3* recombination hotspot locus violates a local molecular clock, though the locus appears to be noncoding and should evolve neutrally. We hypothesize that sudden changes in recombination rate have caused the changes in substitution rate at this locus.

Introduction

Many recent studies have investigated the role of recombination on the rate and pattern of nucleotide substitution (Fullerton, Bernardo Carvalho, and Clark 2001; Meunier and Duret 2004). Recombination may affect nucleotide substitution by one or both of the following two mechanisms: direct mutagenic effect and biased gene conversion (BGC). The mutagenic role of recombination is supported by studies in yeast (Strathern, Shafer, and McGill 1995) and in mammals (Brown and Jiricny 1988; Perry and Ashworth 1999). In particular, Perry and Ashworth (1999) estimated that the elevated recombination in the pseudoautosomal region (PAR) during the last 3 Myr of evolution in the *Mus* lineage has increased the rate of synonymous substitution at the F_{XY} locus by about 170 times. Although a recent study of introns of the F_{XY} locus gave an estimate of only twofold to fivefold increase, the mutagenic effect of recombination was still clear (Huang et al. 2005). Surprisingly, the equivalent region in the human genome does not show such an effect (referred to as the "pseudoautosomal boundary paradox;" Filatov and Gerrard 2003; Galtier 2004; Yi et al. 2004). These recent studies suggest that the mutagenic effect of recombination is considerably weaker than previously thought.

The BGC hypothesis posits that gene conversion leads to preferential fixation of G and C alleles over A and T alleles (e.g., Galtier et al. 2001). This can potentially explain the positive correlation between recombination and genomic GC content (Eyre-Walker 1993; Fullerton, Bernardo Carvalho, and Clark 2001). According to the BGC model, the relative frequency of AT to GC mutation should be positively correlated with the recombination rate. Thus, we can discern the effects of mutagenic recombination and the BGC hypothesis by examining the mutation spectra of genomic regions that have experienced different rates of recombination.

Analyses of recombination in the human genome have shown that recombination events are not evenly distributed along the chromosomes but rather occur nonrandomly, often being clustered within small (1–2 kb) regions, called "recombination hotspots" (Arnheim, Calabrese, and Nordborg 2003; Jeffreys et al. 2004). Recombination rates in such recombination hotspots are often extremely high (sometimes >300 the genome average). Therefore, the effect of recombination on the rate and pattern of nucleotide substitution may be readily detectable at recombination hotspots. In addition, understanding molecular evolution of recombination hotspots can reveal whether recombination hotspots are stable or transient features of the mammalian genome (Wall et al. 2003; Ptak et al. 2004). In this study, we investigated the molecular evolution of four known human recombination hotspots in five hominoid species. For comparison, we also studied eight noncoding regions located inside PARs, which harbor much more frequent recombination than the genome average, and two X-linked regions, which have recombination rates close to the genome average.

Materials and Methods

Regions Sequenced

Recombination hotspot loci and PAR-linked noncoding regions analyzed in this study are shown in table 1. Sequence data of the *16p13.3*, *TAP2*, and *DNA3* hotspots in common chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), and gibbon (*Hylobates lar*) were obtained in this study. Three PAR1 regions, orthologous to the orangutan *ASMT* (*AY181053*), *AY181054*, and *AY181056* loci (as published in Filatov and Gerrard 2003), were newly sequenced in this study in chimpanzee, gorilla, and gibbon. In addition, a PAR2-linked region, *SYBL1*, was sequenced in chimpanzee, gorilla, and gibbon. We were in the process of generating comparative sequence data of PAR regions when Filatov (2004) published his data set that overlapped with some of the regions we were studying. We therefore used his published data for *L254915*, *MIC2*, and *DHRSXY* for chimpanzee, gorilla, and orangutan but obtained new data from gibbon. In addition, we sequenced the recombination hotspot in the human β-globin region, as defined in Wall et al. (2003), in orangutan and gibbon. This corresponds to a ~1.70-kb region just 5′ of

[1] Present address: School of Biology, Georgia Institute of Technology.

Key words: recombination hotspot, pseudoautosomal region, biased gene conversion.

E-mail: soojin.yi@biology.gatech.edu.

Mol. Biol. Evol. 22(5):1223–1230. 2005
doi:10.1093/molbev/msi106
Advance Access publication February 9, 2005

Table 1
Regions Analyzed in This Study

Region	Chromosomal Location	Average Positions Sequenced	Positions Analyzed	GC in Human	Average GC in five Hominoids	Total Tree Length	Human-Orangutan Divergence[b]	Recombination Estimates	Reference	Remark
16p13.3	16p13.3	1790	1673	65.3	64.5	0.14	0.0761**	300	Badge et al. (2000)	Hotspot
DNA3	6p21.32	1063	677	34.6	34.4	0.056	0.0365	140	Jeffreys, Kauppi, and Neumann (2001)	Hotspot
HBD	11p15.4	1526	1383	35.3	35	0.082	0.0458*	140	Wall et al. (2003)	Hotspot
TAP2	6p21.32	1333	1274	40.6	40.6	0.068	0.0398	10	Jeffreys, Ritchie, and Neumann (2000)	Hotspot
ASMT	Xp22.33	1035	706	51.2	51.3	0.146	0.0671**	12.86	Lien et al. (2000)	PAR1
AY181054[a]	Xp22.33	911	664	64.3	64.6	0.123	0.0543**	37.8	Lien et al. (2000)	PAR1
AY181056[a]	Xp22.33	834	583	54.9	54.8	0.134	0.0651**	37.8	Lien et al. (2000)	PAR1
DHRSXY	Xp22.33	1862	1192	42.1	42.1	0.111	0.0506**	12.86	Lien et al. (2000)	PAR1
L254915	Xp22.33	1739	1251	44.9	44.4	0.097	0.0429*	12.86	Lien et al. (2000)	PAR1
MIC2	Xp22.33	2231	1668	34.3	34.5	0.087	0.0282	26	Lien et al. (2000)	PAR1
XG1	Xp22.33	1208	578	40.7	41	0.075	0.0301	26	Lien et al. (2000)	PAR1
SYBL1	Xq28	2464	583	34.5	34.4	0.055	0.0296	6	Lien et al. (2000)	PAR2
XG11	Xp22.33	1350	355	52.6	51.7	0.109	0.0618**	0.87	Lien et al. (2000)	X
XG12	Xp22.33	1073	561	45.3	45.9	0.078	0.02	0.87	Lien et al. (2000)	X
		20419	13148							

NOTE.—Location: based on the UCSC genome browser. Positions analyzed: excluding protein-coding exons, 5′-UTRs, and significant EST matches and alignment gaps (see *Materials and Methods* for details).

[a] Regions homologous to orangutan fragments, AY181054 and AY181056 (Filatov and Gerrard 2003), have been amplified from other hominoid species.

[b] Significantly larger than expected from Poisson distribution assuming 3.08% average divergence: $^{**}P < 0.01$; $^{*}P < 0.05$.

the β-globin gene. Finally, we sequenced portions of the *XG* gene, a pseudoautosomal segment (corresponding to the *XG1* amplicon in Yi et al. 2004), and two X-linked segments (corresponding to amplicons *XG11* and *XG12* in Yi et al. 2004) from gibbon.

Sequencing, Alignment, and Data Curation

Target loci were amplified from genomic DNA by polymerase chain reaction (PCR) using a high-fidelity PCR kit (Roche Diagnostics, Indianapolis, Ill.). The primers used for the amplification and sequencing are available upon request. We used the BigDye terminator cycle sequencing kit version 3 (Applied Biosystems, Foster City, Calif.) to analyze on an ABI 377 sequencer. We used ClustalW (Thompson, Higgins, and Gibson 1994) to align the sequences. In some cases the alignments were manually adjusted to account for complex indels. Human reference sequences were extracted from GenBank. Because we are interested in the mutational patterns of neutral regions, we excluded protein-coding exons, 5′-untranslated regions (UTRs), and other known and annotated regulatory regions from our sequence data set. In addition, we excluded significant expressed sequence tag (EST) matches from our alignments as they may reflect transcribed regions under some functional constraint. We used an arbitrary cutoff E-value of $<e^{-40}$ and alignment length >100 bp. These criteria efficiently removed all sequences matching known ESTs with $>90\%$ sequence identity. The final numbers of nucleotide positions used for analyses, after excluding potentially functional sites, are shown in table 1. All sequences obtained in this study are deposited in the GenBank (accession numbers AY880846–AY880880).

Recombination Rate Data

For all PAR loci, recombination rate estimates are taken from Lien et al. (2000), who provided a recombination map of PARs by typing more than 1,900 single sperm for 25 genetic markers for PAR1 and one marker for PAR2. To date, this is the most extensive study to directly estimate the recombination rates in the human PAR1. Recombination rate estimates for recombination hotspots are taken from the literature (see table 1).

Estimation of Ancestral GC Content

We used two maximum likelihood methods to estimate the ancestral GC content of the hominoid ancestor. First, we used the method by Galtier and Guoy (1998, implemented in NHML). This method estimates the ancestral GC content using a nonhomogeneous, nonstationary sequence evolution model. Second, we reconstructed the ancestral sequence following the Hasegawa, Kishino, and Yano (1985) model (the HKY model) in PAML (Yang 1997) and then estimated the GC content of the ancestral node connecting all five hominoid species.

Ancestral-Derived Substitutions

To elucidate the patterns of nucleotide substitution in the highly recombining regions of the human genome, we inferred the substitution in hominoids by two approaches: a maximum likelihood method and a parsimony method. For the maximum likelihood approach, we used the baseml program in the PAML package (Yang 1997). Specifically, we reconstructed the ancestral nucleotide sequences, using the marginal reconstruction method, following the HKY sequence evolution model. When we used the joint reconstruction instead, the results did not qualitatively differ (results not shown). For parsimony reconstruction, we extracted all sites with a 4:1 segregation pattern in the alignment of the five hominoid species (human, chimpanzee, gorilla, orangutan, and gibbon). Then, at each site the nucleotide with four occurrences was inferred to be ancestral, while

the nucleotide with one occurrence was taken as derived. As this method does not correct for multiple hits, it underestimates the numbers of substitutions and is likely to be more stringent than the maximum likelihood method.

Results

Nucleotide Substitution Rates in Highly Recombining Regions

We analyzed the pattern of nucleotide substitutions of 14 genomic regions, including 4 recombination hotspots, 7 PAR1-linked sequences, 1 PAR2-linked sequence (*SYBL1*) and 2 X-linked sequences of the *XG* gene, which straddles the PAR1 boundary in hominoids, from the following five hominoid species: human, chimpanzee, gorilla, orangutan, and gibbon (table 1). The lengths of sequenced regions, and the lengths analyzed after excluding potentially functional regions, and their recombination rate estimates taken from literature are shown in table 1. Their genomic locations according to the UCSC genome browser (http://genome.ucsc.edu) are also shown.

For all the 14 sequence segments, bootstrap trees support the species tree of (gibbon, (orangutan, (gorilla, (chimpanzee, human)))) under a variety of tree-building schemes (fig. 1*a*). This corresponds to the well-accepted species relationship in hominoids (Yoder and Yang 2000). The total branch lengths for each locus are shown in table 1. They vary between 0.055 (*SYBL1* locus) to 0.146 (a sequence fragment orthologous to orangutan *ASMT*). The PAR1-linked segments showed longer branch lengths than recombination hotspots (average 0.110 vs. 0.087), with the exception of the *16p13.3* recombination hotspot (see below).

To have a better idea of the relative divergences of these regions, we compared the divergence estimates between the human and the orangutan sequences of the genomic regions included in our data set with the average genomic divergence between human and orangutan, which has been rather extensively documented and estimated to be ~3% (3.08 ± 0.11% by Chen and Li 2001, 3% by Filatov and Gerrard 2003; divergence based upon synonymous sites only may be even lower, 2.00 ± 0.11%, Shi et al. 2003). We tested whether the observed numbers of differences between the human and orangutan sequences were greater than the expected 3.08% divergence, assuming the Poisson distribution. Five of the seven PAR1-linked segments showed significantly greater divergence than 3.08% by this test. Two recombination hotspots, the *HBD* and the *16p13.3* locus showed deviation from the genome average. Overall, this result is in accord with the observation that the total branch lengths of the PAR1-linked sequences are greater than those of the four recombination hotspots. A previous study documented that the PAR1-linked regions tend to evolve faster than the rest of genome when compared between the human and orangutan (Filatov and Gerrard 2003). The recombination hotspot *16p13.3* showed the greatest divergence between the human and orangutan sequences in our data set. This may be due to the extended orangutan branch length (see fig. 1*b* and the next section).

Recombination may increase the substitution rate by the direct mutagenic effect and/or via BGC. However, we find *no* correlation between the estimated recombination rates and total branch lengths ($\rho = 0.2$, $P > 0.1$). When only the PAR1-linked segments are considered, there is still no correlation between the estimated recombination frequencies and the total branch lengths.

Maximum Likelihood Analyses of Substitution Rate

Neutral regions of genomes of closely related species are known to evolve at very similar rates among species. In other words, local molecular clocks exist among closely related species (Yoder and Yang 2000). The punctuality of local molecular clocks in noncoding regions is the basis for developing effective phylogenetic markers from such regions. Because we only consider data from noncoding, nonfunctional nucleotides (free from exons, UTRs, other regulatory regions, and EST matches) from closely related hominoid species, we expect that the evolution of these regions should approximately follow a canonical local molecular clock.

The maximum log likelihoods of a substitution model in which all branches are evolving at one rate (referred to as the "one-rate model"), which assumes a hominoid molecular clock, are shown in table 2. This model has four parameters, corresponding to the internal nodes in the tree with five branches. For comparison, the maximum log likelihoods of an alternative model in which all branches are allowed to evolve at independent rates are also shown (referred to as the "free-rate model," table 2). This model assumes no molecular clock and has seven parameters, including four external and three internal branches. For all but one of the 14 regions studied, the free-rate model was not significantly better by the log-likelihood ratio test. This is in accord with the hominoid molecular clock model. Surprisingly, for the *16p13.3* locus, the free-rate model is a significantly better fit than the one-rate model ($2(\ln L_1 - \ln L_2) = 8.38$, $P < 0.05$, $df = 3$).

The neighbor-joining tree of *16p13.3* is shown in figure 1*b*. We note that the branch leading to the orangutan from the ancestor of human, chimpanzee, and gorilla is extraordinarily long. This leads to an erroneous species tree in which the lesser ape gibbon is connected to the three great ape species before the orangutan lineage branches out. Taking this observation into account, we implemented an alternative model in which the orangutan branch was allowed to evolve at a different rate than the rest of the tree. This two-rate model has a maximum log likelihood of −3,476.03, significantly better than the one-rate model ($2(\ln L_1 - \ln L_2) = 8.2$, $df = 1$, $P < 0.005$). This two-rate model is also better than the free-rate model, which estimates two more parameters. Alternative maximum likelihood models in which other branches were allowed to vary did not perform better than this particular two-rate model (results not shown).

GC Content Versus Rate of Nucleotide Substitution

An important factor in determining sequence divergence is sequence context, particularly the proportions of G and C nucleotides. The average GC contents of the hominoid sequences are strongly correlated with the total tree lengths ($\rho = 0.82$, $P < 0.01$). When only the PAR1-linked loci are considered, the correlation is slightly weaker, but still significant ($\rho = 0.66$, $P < 0.05$). In contrast, there is *no*

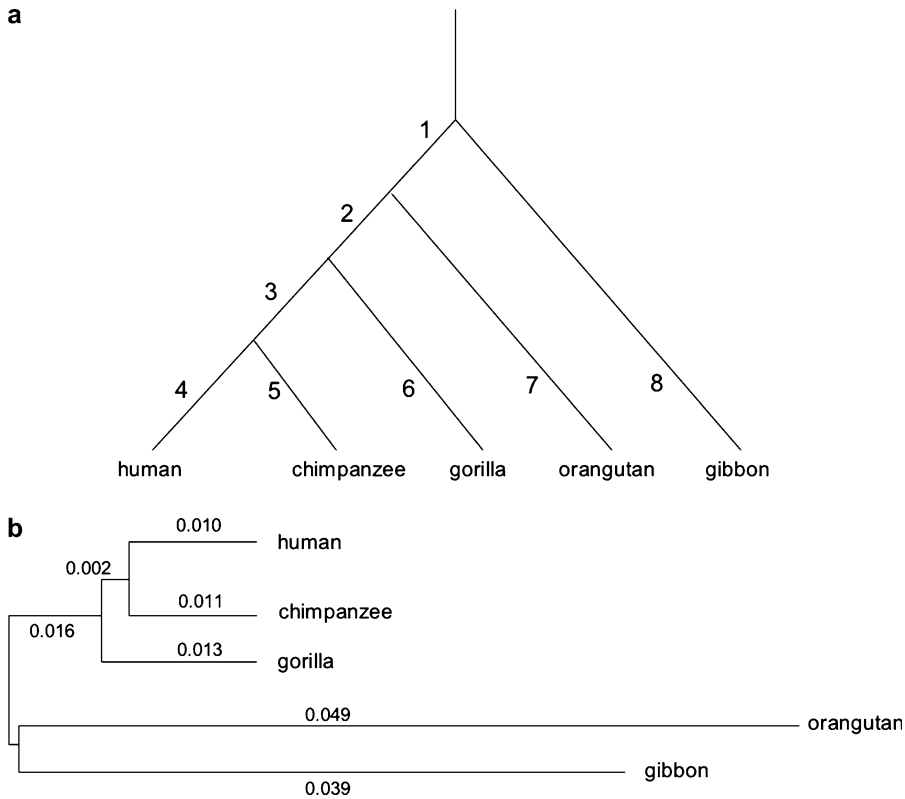

FIG. 1.—(*a*) The phylogenetic relationships among human, chimpanzee, gorilla, orangutan, and gibbon. All eight branches, including three internal and five external branches are labeled with numbers. (*b*) Neighbor-joining tree of the recombination hotspot *16p13.3* locus. The length of each branch is shown. The tree is drawn roughly to scale.

correlation between the GC contents and the estimated recombination rates.

From the ancestral GC contents estimated by maximum likelihood methods (table 3), we note several points. First, the current GC contents of the five hominoid species are very similar to each other (results not shown). Second, the ancestral GC content of each locus differs very little from the current GC contents. Indeed, the two maximum likelihood methods we employed provided similar estimates. There is no strong trend toward either an increase or a decrease in GC content during the recent evolution of the highly recombining regions in the hominoid genomes. Nor is there any correlation between the recombination frequency and the current or ancestral GC content. In other words, increased recombination rates in the human recombination hotspots

Table 2
Log Likelihoods of Different Evolutionary Models

Gene	Location	Hominoid Molecular Clock (One-Rate Model)	No Molecular Clock (Free-Rate Model)	Difference	2 × Difference
16p13.3	Hotspot	−3,480.13	−3,475.94	−4.19	−8.38*
DNA3	Hotspot	−1,141.68	−1,139.97	−1.71	−3.42
HBD	Hotspot	−2,543.5	−2,541.71	−1.79	−3.58
TAP2	Hotspot	−2,230.99	−2,230.08	−0.91	−1.82
ASMT	PAR1	−1,535.16	−1,535.04	−0.12	−0.24
DHRSXY	PAR1	−2,386	−2,384.93	−1.07	−2.14
L254915	PAR1	−2,448.79	−2,445.05	−3.74	−7.48
MIC2	PAR1	−3,044.69	−3,045.22	0.53	1.06
SHOX	PAR1	−2,574.2	−2,573.9	−0.3	−0.6
XG1	PAR1	−1,043.76	−1,041.75	−2.01	−4.02
SYBL1	PAR2	−3,685.51	−3,684.52	−0.99	−1.98
XG	X	−1,739.52	−1,739.14	−0.38	−0.76

NOTE.—To increase the statistical power, we merged sequence data from regions sampled from the same gene for these analyses and renamed the merged segment of *AY181054* and *AY181056* as *SHOX* and the merged segment of the *XG11* and the *XG12* as *XG*.

*$P < 0.05$.

Table 3
Ancestral GC Contents for Each Locus Estimated by Maximum Likelihood Methods

Locus	Remark	Number of Sites	Ancestral GC by Maximum Likelihood Methods	
			PAML	NHML
16p13.3	Hotspot	1,673	65.5	65.6
DNA3	Hotspot	677	34.0	34.3
HBD	Hotspot	1,383	35.9	35.3
TAP2	Hotspot	1,274	40.4	39.4
ASMT	PAR	706	54.5	50.8
DHRSXY	PAR	1,192	42.4	41.3
L254915	PAR	1,251	45.6	43.9
MIC2	PAR	1,668	34.5	35.0
SHOX	PAR	1,247	59.9	58.5
XG1	PAR	578	40.7	39.2
SYBL1	PAR2	2,184	34.2	33.8
XG	X	916	47.4	46.3

Table 4
Inferred Substitutions During Hominoid Evolution

	AT to AT	GC to GC	AT to GC	GC to AT	r[a]	Total
Hotspots						
16p13.3	3 (0)	30 (8)	52 (4)	142 (5)	2.73	227 (17)
DNA3	2 (0)	4 (0)	21 (2)	10 (0)	0.48	37 (2)
HBD	10 (2)	10 (1)	55 (8)	36 (2)	0.65	111 (13)
TAP2	3 (0)	7 (0)	37 (2)	37 (2)	1	84 (4)
Subtotal	18(2)	51(9)	165(16)	225(9)	1.36	459(36)
PAR regions						
ASMT	9 (0)	14 (0)	33 (2)	47 (3)	1.5	103 (5)
DHRSXY	7 (0)	15 (0)	51 (3)	57 (3)	1	130 (6)
L254915	4 (0)	17 (4)	61 (13)	37 (1)	0.61	119 (18)
MIC2	10 (2)	8 (1)	85 (7)	46 (2)	0.54	149 (12)
SHOX	6 (0)	29 (4)	50 (4)	86 (10)	1.7	171 (18)
SYBL1	8 (2)	11 (0)	64 (6)	37 (3)	0.58	120 (11)
XG1	0 (0)	2 (0)	22 (0)	47 (3)	2.14	71 (3)
Subtotal	44(4)	96(9)	366(35)	357(25)	0.98	863(73)
X						
XG	3 (0)	12 (0)	47 (4)	18 (0)	0.38	80 (4)
Total	65(6)	159(18)	578(55)	600(34)	1.04	1402(113)

NOTE.—Substitutions specific to the human branch are shown in parentheses. Substitutions on all branches are shown in supplementary table 1.

[a] r = (GC to AT)/(AT to GC).

and the PAR regions are not correlated with any change in the GC contents of the region in recent evolutionary times.

AT to GC Versus GC to AT Substitutions

All substitutions that occurred during the evolution of the five hominoid species are partitioned into eight branches (see figure 1*a* and supplementary table 1). We divided the substitutions into four categories, AT to AT, GC to GC, AT to GC, and GC to AT, pooling complimentary substitutions (table 4). The total number of AT to GC mutations (578) is somewhat smaller than that of GC to AT mutations (600). Five loci (*XG1*, *DHRSXY*, *ASMT*, *SHOX*, and *16p13.3*) had excess of GC to AT substitutions. A previous study of the 1.8 Mb of orthologous human and chimpanzee genomic sequences from five large regions of human chromosome seven revealed that regions with more than 40% GC content had excess of GC to AT mutations (table 1 in Webster, Smith, and Ellegren 2003). We observe the same pattern in our data set, except for the *L254915* locus, which had 61 AT to GC substitutions and 37 GC to AT substitutions, even though its GC content is over 40%. At this locus, the great ape lineages have a large excess of AT to GC substitutions than GC to AT substitutions (see supplementary table 1) in contrast to the pattern in the rest of the tree. The two types of substitutions at the *L254915* locus are significantly heterogeneous among different branches (χ^2 = 45.6, df = 6 [excluding the branch 1, which has no substitution], P = 0.002). No other locus had a statistically significantly heterogeneous substitution pattern.

Interestingly, in the human branch the sum of the numbers of AT to GC substitutions in the recombination hotspot loci is much greater than (almost twice) the number of GC to AT substitutions (table 5). This increase of AT to GC in the human lineage is in accord with the effect of BGC associated with increased recombination in the recombination hotspots. In contrast, in the PAR1-linked sequences, the pattern in the human branch is similar to that of the rest of the tree. Indeed, the substitution pattern in the recombination hotspots in the human lineage is significantly different from the rest of the tree (P = 0.035, by the two-tailed Fisher's exact test), while it is not the case for the PAR1-linked segments (P = 0.24).

Table 5
Numbers of AT to GC and GC to AT Substitutions Along the Human Branch and the Rest of the Tree

	Hotspots		PAR1-Linked Regions	
	Human	Other Branches	Human	Other Branches
AT to GC	16 (13)	149 (118)	29 (28)	273 (259)
GC to AT	9 (6)	216 (181)	22 (14)	298 (196)
Fisher's exact test (two-tailed)		P = 0.035 (P = 0.016)		P = 0.24 (P = 0.25)

NOTE.—Substitutions inferred by a parsimony approach are shown in parentheses.

To check whether this was due to erroneous assignments of substitutions by the maximum likelihood method, we reanalyzed the data using a stringent parsimony criterion (see *Materials and Methods*). We observed the same excess of AT to GC substitutions in the human branch compared to the rest of the tree for the recombination hotspot regions (13 AT to GC vs. 6 GC to AT in the human branch, while 118 AT to GC vs. 181 GC to AT for the rest of the tree). The difference is significant by the two-tailed Fischer's exact test (P = 0.016). Again, no such heterogeneity is observed in PAR1-linked segments. Therefore, both the maximum likelihood inference and the parsimony inference show that the AT to GC and GC to AT substitution patterns along the human branch may be different from the substitution patterns in other hominoid species for the recombination hotspots, while no such pattern is observed in the PAR1 regions.

Discussion

Recombination Versus Nucleotide Substitution

One of our main questions was whether neutral sequence divergence in highly recombining regions is positively correlated with recombination rate. If recombination is mutagenic, it will accelerate sequence divergence. The effect of BGC is similar to that of weak selection in favor of the G and C nucleotides (Nagylaki 1983) but may not affect the rate of sequence divergence if the base composition is at equilibrium. However, the mutagenic effect of recombination can still increase the rate of sequence divergence even if the GC content is at equilibrium. On the other hand, if the base composition is not at equilibrium, BGC may rapidly increase sequence divergence (Piganeau et al. 2002; Montoya-Burgos, Boursot, and Galtier 2003). Therefore, both mechanisms may increase sequence divergence in highly recombining regions. When analyzed in detail, the molecular evolution of one recombination hotspot locus, *16p13.3*, is accelerated in some lineages, which can be explained by the mutagenic effect of recombination. The evolution of a PAR1-linked region, the *L254915* locus, shows a much more frequent AT to GC substitution in some lineages than is expected given the substitution pattern in

other species, in accord with the prediction of BGC. However, despite the fact that there is some evidence in favor of both the BGC and mutagenic recombination in our data, overall there is *no* significant correlation between recombination rate and neutral divergence. We propose that this is because the recombination estimates in the human genome differ from the past recombination rates during the last 20 Myr or so of hominoid evolution. More specifically, we propose that intense recombination hotspots are mostly short-lived (see below).

Recombination Hotspots Are Likely Transient

Several characteristics of our data support the view that recombination hotspots are transient in evolutionary time (Boulton, Myers, and Redfield 1997; Wall et al. 2003; Ptak et al. 2004). First, the GC contents of the present-day hominoid sequences are very similar to each other and to that of the ancestral sequence. GC content and recombination rate are positively correlated in many taxa (Eyre-Walker 1993; Fullerton, Bernardo Carvalho, and Clark 2001; Marais, Mouchiroud, and Duret 2001; Birdsell 2002), presumably as a result of BGC. Our observation that the GC contents of the highly recombining regions have not significantly changed since the divergence of hominoids cannot be reconciled with the extremely high (sometimes >300 times the genome average) recombination rates. If the intense recombination has been a constant feature during the evolution of these regions, the ancestral GC content is expected to be rather different from the current GC content, which was not what we observed. In addition, neither current nor ancestral GC content is correlated with the estimated recombination rate. It is likely that the increased recombination rate is a recent phenomenon and the GC content has not yet been affected.

Second, we observe that the recombination hotspots and the PAR1-linked regions have evolved differently, as expected given the age differences between the two types of regions. The PAR1 boundary was established in the common ancestor of the hominoids and the Old World monkeys (Ellis et al. 1990; Yi et al. 2004), so the regions in PAR1 have experienced elevated recombination frequencies for over the last 25 Myr or so of primate evolution. Recombination hotspots, on the other hand, have been characterized within the human population. According to a recent population study, the recombination hotspot in the human β-globin region is not a recombination hotspot in Macaque, an Old World monkey species (Wall et al. 2003). Moreover, the human TAP2 recombination hotspot is absent in Western Chimpanzees (Ptak et al. 2004). While there is no information on the antiquity of the recombination hotspots *16p13.3* and *DNA3*, the above studies suggest that recombination hotspots are short-lived. Theoretical work also suggests that recombination hotspots are transient (Boulton, Myers, and Redfield 1997). Therefore, the recombination hotspots in the human genome studied above are likely to be younger than the divergence of the Old World monkeys and the hominoids or, in other words, younger than the PAR1. This can explain why the total branch lengths for PAR1-linked regions are generally longer than those for recombination hotspot loci; because PAR1-linked sequences have experienced increased recombination for a longer time, they have accumulated more sequence divergences.

Another difference between the recombination hotspots and PAR1-linked regions is the difference in substitution pattern between the human and other hominoid branches. In the four recombination hotspots, the human branch has accumulated more AT to GC substitutions than the rest of the tree. This can be taken as evidence for BGC associated with the newly increased recombination rate in the human lineage for the hotspots. As explained above, for PAR1 regions the pattern of molecular evolution of the human branch is not expected to be different from that of the rest of the hominoid tree.

We note that there may be some other possible causes for the lack of correlation between recombination rate and neutral divergence in our data set, in particular for the PAR1-linked regions. Genomic recombination rates vary greatly along a chromosome, often dramatically in a relatively small scale (Wall et al. 2003; McVean et al. 2004). The recombination rate estimates for the PAR1 regions were inferred from markers separated by tens of thousands of nucleotides (Lien et al. 2000) and may not correspond to the actual regional recombination rates. If we had the actual recombination rate estimates for each of the PAR1-linked regions, not the estimates obtained in a relatively gross scale as is available currently, then we may observe the correlation between the recombination rates and the sequence divergence and, the GC contents, because the PAR1 regions have spent similar evolutionary time in their unusually highly recombining environment. However, the four recombination hotspots have been defined by precise, microscale analyses, so that the estimates should be close to the true values. We also note that male recombination rates measured in sperm-typing studies may not reflect the correct recombination frequencies because they may differ from the female recombination rates for some regions (e.g., Meunier and Duret 2004). However, intense male recombination hotspots should still increase average recombination rates.

Violation of the Local Molecular Clock at the *16p13.3* Recombination Hotspot

One interesting finding in this study is the violation of the hominoid local molecular clock at the *16p13.3* locus. This is surprising because we analyzed only noncoding regions, free from UTRs, protein-coding exons, and EST matches. The rates from these regions should only reflect the differences in the neutral mutation rates, assumed to follow a "local" molecular clock. This is often an advantage for using nonfunctional regions for molecular phylogenetic studies.

The cause for the observed rate disparity is not clear. This region is not closely associated with any identifiable coding sequences, suggesting that selection does not influence haplotypic association in this region (Badge et al. 2000). One attractive hypothesis has to do with the observation that the *16p13.3* region corresponds to a well-defined recombination hotspot in the human genome. Even though we do not know whether this locus also corresponds to a recombination hotspot in the orangutan genome, it is possible that a sudden change in recombination frequency in this

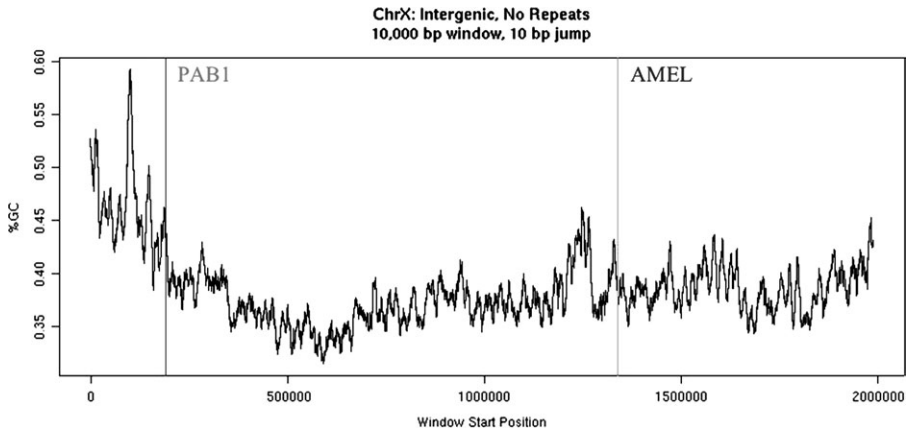

FIG. 2.—A sliding-window analysis of the GC content of a noncoding, nonrepetitive genome sequence along the human Xp arm. The window size is 10,000 bp with 10 bp overlaps. The location of the current pseudoautosomal boundary (PAB1) and the ancient pseudoautosomal boundary (AMEL) are shown for the purpose of comparison.

lineage has led to an elevated mutation rate either by mutagenic effect or BGC. This hypothesis can in principle be tested by analyzing the pattern of SNP variation at the *16p13.3* locus in humans, orangutans, and other hominoids to detect the signature of recombination and gene conversion on the polymorphism data.

The pattern of molecular evolution of *L254915* suggests that this locus may be another target for a newly evolving recombination–gene conversion hotspot in the hominoid genomes, especially in the genomes of the great apes, as it has accumulated excess of AT to GC substitutions, possibly as a result of BGC.

Recombination Rates Along the Pseudoautosomal Region 1

Whether or not recombination rate increases along the chromosomal length toward the telomere is an interesting question. It is generally assumed that the recombination rate may increase along a PAR toward the telomere. Indeed, Filatov (2004) reported an increase of total tree length along the Xp PAR. In our data set, there are nine segments that are collinearly located on the Xp, and the distance from the PAB and the total tree lengths are strongly correlated ($\rho = 0.76$, $P < 0.02$), in agreement with Filatov (2004).

We also observed a strong linear relationship between the GC content of the sequence and the neutral mutation rate for PAR1-linked regions; the GC contents and the neutral divergence are also significantly correlated in the whole data set (see *Results*). To examine whether the GC content of the PAR1 increases toward the telomere, we performed a sliding-window analysis of the GC contents of the intergenic sequences from the human X chromosome (fig. 2). We observe a generally increasing trend of the GC content near the telomere, though it is not strictly linear.

Therefore, the substitution rate, the GC content, and the distance from the telomere all appear to be correlated in the p-arm of the human X chromosome. The cause for such a relationship is of great interest. It is tempting to hypothesize that the increase in recombination rate toward the telomere is the cause of these observations. In this regard, we point out that telomeric regions are generally GC rich (Perani et al. 2000; Arndt and Hwa 2004). In particular, telomeric 1-Mb regions of human chromosomes show a bias in substitution patterns to accumulate G and C nucleotides and become GC rich (Arndt and Hwa 2004). Such a bias in substitution pattern near telomeres may well be due to BGC associated with increased recombination. If so, the observation that the substitution rates increase along the human Xp arm may be part of a general phenomenon in mammalian chromosomes. Investigating other telomeres, in particular in regions that have recently became part of telomeres and vice versa, will be a good approach to study the effect of recombination on substitution patterns and rates in mammals.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online (www.mbe.oupjournals.org).

Acknowledgments

We thank Tyronne Summers for his help in analyzing the GC counts along the human pseudoautosomal region 1. This study was supported by the start-up fund from the Georgia Institute of Technology to S.Y. and NIH grants to W.H.L.

Literature Cited

Arndt, P. F., and T. Hwa. 2004. Regional and time-resolved mutation patterns of the human genome. Bioinformatics **20**:1482–1485.

Arnheim, N., P. Calabrese, and M. Nordborg. 2003. Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. Am. J. Hum. Genet. **73**:5–16.

Badge, R. M., J. Yardley, A. J. Jeffreys, and J. A. Armour. 2000. Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. Hum. Mol. Genet. **9**:1239–1244.

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.

Boulton, A., R. S. Myers, and R. J. Redfield. 1997. The hotspot conversion paradox and the evolution of meiotic recombination. Proc. Nat. Acad. Sci. USA **94**:8058–8063.

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.

Chen, F. C., and W.-H. Li. 2001. Genomic divergence between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. **68**:444–456.

Ellis, N., P. Yen, K. Neiswanger, L. J. Shapiro, and P. N. Goodfellow. 1990. Evolution of the pseudoautosomal boundary in old world monkeys and great apes. Cell **63**:977–986.

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

Filatov, D. A. 2004. A gradient of silent substitution rate in the human pseudoautosomal region. Mol. Biol. Evol. **21**:410–417.

Filatov, D. A., and D. T. Gerrard. 2003. High mutation rates in human and ape pseudoautosomal genes. Gene **317**:67–77.

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. 2004. Recombination, GC-content and the human pseudoautosoomal boundary paradox. Trends Genet. **20**:347–349.

Galtier, N., and M. Gouy. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. **15**:871–879.

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.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

Huang, S.-W., R. Friedman, A. Yu, and W.-H. Li. 2004. How strong is the mutagenecity of recombination in mammals? Mol. Biol. Evol. (in press).

Huang, S.-W., R. Friedman, N. Yu, A. Yu, and W.-H. Li. 2005. How strong is the mutagenecity of recombination in mammals? Mol. Biol. Evol. **22**:426–431.

Jeffreys, A. J., J. K. Holloway, L. Kauppi, C. A. May, R. Neumann, M. T. Slingsby, and A. J. Webb. 2004. Meiotic recombination hot spots and human DNA diversity. Philos. Trans. R. Soc. Lond. B Biol. Sci. **359**:141–152.

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., A. Ritchie, and R. Neumann. 2000. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Hum. Mol. Genet. **9**:725–733.

Lien, S., J. Szyda, B. Schechinger, G. Rappold, and N. Arnheim. 2000. Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. Am. J. Hum. Genet. **66**:557–566.

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

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

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. 2004. The fine-scale structure of recombination rate variation in the human genome. Science **304**:581–584.

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

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

Perani, P., S. Caccio, S. Saccone, L. Andreozzi, and G. Bernardi. 2000. Telomeres in warm-blooded vertebrates are composed of GC-rich isochores. Biochem. Genet. **38**:227–239.

Perry, J., and A. Ashworth. 1999. Evolutionary rate of a gene affected by chromosomal position. Curr. Biol. **9**:987–989.

Piganeau, G., D. Mouchiroud, L. Duret, and C. Galtier. 2002. Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. J. Mol. Evol. **54**:129–133.

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.

Shi, J., H. Xi, Y. Wang et al. (20 co-authors). 2003. Divergence of the genes on the human chromosome 21 between human and other hominoids and variation of substitution rates among transcription units. Proc. Natl. Acad. Sci. USA **100**:8331–8336.

Strathern, J., B. Shafer, and C. McGill. 1995. DNA synthesis errors associated with double-strand-break repair. Genetics **140**:965–972.

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.

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., N. G. C. Smith, and J. Ellegren. 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. Mol. Biol. Evol. **20**:278–286.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.

Yi, S., T. J. Summers, N. M. Pearson, and W.-H. Li. 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.

Yoder, A. D., and Z. Yang. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. **17**:1081–1090.

Naruya Saitou, Associate Editor

Accepted January 25, 2005