A Gradient of Silent Substitution Rate in the Human Pseudoautosomal Region

Dmitry A. Filatov
School of Biosciences, University of Birmingham, Birmingham, United Kingdom

It has been demonstrated that recombination in the human p-arm pseudoautosomal region (p-PAR) is at least twenty times more frequent than the genomic average of ~1 cM/Mb, which may affect substitution patterns and rates in this region. Here I report the analysis of substitution patterns and rates in 10 human, chimpanzee, gorilla, and orangutan genes across the p-PAR. Between species silent divergence in the p-PAR forms a gradient, increasing toward the telomere. The correlation of silent divergence with distance from the p-PAR boundary is highly significant (ρ = 0.911, P < 0.001). After exclusion of the CpG dinucleotides this correlation is still significant (ρ = 0.89, P < 0.01), thus the substitution rate gradient cannot be explained solely by the differences in the extent of methylation across the p-PAR. Frequent recombination in the PAR may result in a relatively strong effect of biased gene conversion (BGC), which, because of the increased probability of fixation of the G or C nucleotides at (A or T)/(G or C) segregating sites, may affect substitution rates. BGC, however, does not seem to be the factor creating the substitution rate gradient in the p-PAR, because the only gradient is still detectable if only A→T and G→C substitutions are taken into account (ρ = 0.82, P < 0.01). I hypothesize that the substitution rate gradient in the p-PAR is due to the mutagenic effect of recombination, which is very frequent in the distal human p-PAR and might be lower near the p-PAR boundary.

Introduction

Human X and Y chromosomes pair and recombine in two small pseudoautosomal regions (PARs) at the ends of the sex chromosomes (Cooke, Brown, and Rappold 1985; Freije et al. 1992). The short arms (Xp/Yp) of the sex chromosomes contain the p-PAR, which is over 2 Mb in size (Brown 1988; Petit, Levilliers, and Weissenbach 1988). Chiasmata between X and Y chromosomes in the p-PAR are essential for the correct segregation of the sex chromosomes in male meiosis (Ellis and Goodfellow 1989; Burgoyne et al. 1992). Obligate crossing over in the relatively small p-PAR results in a very high recombination rate in the region. High-resolution sperm typing has demonstrated that the p-PAR recombination rate in male meiosis is greater than 20 times the genomic average of ~1 cM/Mb (Lien et al. 2000), and may even be as high as 350 cM/Mb (May et al. 2002). The long (Xq/Yq) arms contain the much smaller q-PAR, which is about 0.4 Mb long (Ciccodicola et al. 2000). The recombination rate in the q-PAR is lower than in the p-PAR at about 5 cM/Mb (Ciccodicola et al. 2000).

As in the human p-PAR, recombination in the mouse PAR is also very frequent (Soriano et al. 1987). Perry and Ashworth (1999) demonstrated that the silent substitution rate in the pseudoautosomal part of the mouse Fxy gene substantially accelerated after this region was translocated into the PAR, suggesting that recombination may accelerate the silent substitution rate. This work motivated us to conduct a study of substitution rates in the human PARs, revealing a significantly accelerated silent substitution rate in three human p-PAR genes (Filatov and Gerrard 2003). This finding demonstrates that the elevated silent substitution rate (and perhaps, an elevated mutation rate) is probably a general feature of the pseudoautosomal regions in mammals.

As the elevated recombination rate is the most prominent feature of the PARs, it is tempting to associate elevated substitution and recombination rates. Indeed, there is growing evidence that recombination may affect substitution patterns and rates. Several studies recently reported a weak but significantly positive correlation of recombination rate in humans with human/mouse four-fold degenerate site divergence (Lercher and Hurst 2002; Waterston et al. 2002; Hardison et al. 2003). Recombination rate was demonstrated to correlate positively with human/chimpanzee and human/baboon non-coding divergence (Hellmann et al. 2003). As the silent substitution rate is usually assumed to be neutral, its correlation with recombination suggests a causal relationship between the processes of recombination and mutation. The mutagenic effect of recombination reported for yeast (Strathern, Shafer, and McGill 1995) strengthens the evidence in favor of this hypothesis.

Several other factors can also result in the correlation between recombination and silent substitution rates. Both recombination and substitution rates correlate positively with GC-content (Eyre-Walker 1993; Fullerton, Carvalho, and Clark 2001; Bielawski, Dunn, and Yang 2000; Yi, Ellsworth, and Li 2002), which may result in a covariation between the substitution and recombination rates. Frequent methylation-induced CpG→TpG mutations (Robertson and Wolffe 2000) may also influence the correlation of recombination rates and substitution rates, as the frequency of CpG dinucleotides has been reported to be associated with recombination rate in humans (Kong et al. 2002). Biased gene conversion (BGC), the preferential resolution of (A or T)/(G or C) heteroduplexes toward GC (Lamb 1986; Brown and Jiricny 1988), may be another such factor. In fact, BGC has been demonstrated to be mathematically equivalent to selection for (A or T)→(G or C) mutations (Nagylaki 1983); thus it may have a substantial effect on fixation rates at (A or T)/(G or C) segregating sites and on the overall substitution rates.
Table 1
p-PAR Regions Analyzed

<table>
<thead>
<tr>
<th>Gene # and Name</th>
<th>Kb to PAB</th>
<th>Human Genomic Contig</th>
<th>PCR Primers</th>
<th>Positions Analyzed</th>
<th>GC%</th>
<th>Introns</th>
<th>Cpg/CpC Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 PAB</td>
<td>0</td>
<td>NT_025302.11</td>
<td>PAB+1</td>
<td>1,507</td>
<td>43</td>
<td>20/66</td>
<td></td>
</tr>
<tr>
<td>1 XG</td>
<td>7</td>
<td>NT_025302.11</td>
<td>PAB-4x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 MIC2</td>
<td>60</td>
<td>NT_025302.11</td>
<td>XG+1</td>
<td>1,298</td>
<td>43</td>
<td>19/56</td>
<td></td>
</tr>
<tr>
<td>3 L254916</td>
<td>156</td>
<td>NT_025302.11</td>
<td>XG-2</td>
<td>2,189</td>
<td>38</td>
<td>15/93</td>
<td></td>
</tr>
<tr>
<td>4 L254915</td>
<td>200</td>
<td>NT_025302.11</td>
<td>MIC2-3</td>
<td>1,857</td>
<td>45</td>
<td>14/69</td>
<td></td>
</tr>
<tr>
<td>5 DHR5XY</td>
<td>350</td>
<td>NT_025302.11</td>
<td>L254916-4</td>
<td>1,620</td>
<td>47</td>
<td>36/78</td>
<td></td>
</tr>
<tr>
<td>6 ASMT</td>
<td>910</td>
<td>NT_033330.5</td>
<td>L254916-3</td>
<td>1,620</td>
<td>47</td>
<td>36/78</td>
<td></td>
</tr>
<tr>
<td>7 ASMTL</td>
<td>1095</td>
<td>NT_033330.5</td>
<td>L254915-4</td>
<td>1,620</td>
<td>47</td>
<td>36/78</td>
<td></td>
</tr>
<tr>
<td>8 SHOX</td>
<td>1730</td>
<td>AC137591</td>
<td>DHR+3</td>
<td>1,673</td>
<td>42</td>
<td>20/64</td>
<td></td>
</tr>
<tr>
<td>9 PPP2R3L</td>
<td>2095</td>
<td>NT_077814.2</td>
<td>ASMT+6</td>
<td>854</td>
<td>56</td>
<td>42/62</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ASMTL+4</td>
<td>1,309</td>
<td>56</td>
<td>55/104</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SHOX+1</td>
<td>660</td>
<td>64</td>
<td>47/83</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SHOX-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PPP+1</td>
<td>587</td>
<td>67</td>
<td>60/68</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PPP-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Given for the human sequence; ape sequences may have slightly different values.

Materials and Methods

To study substitution patterns and rates in the frequently recombining human and ape p-arm pseudo-autosomal region, nine p-PAR loci were selected (table 1). One region per gene was sequenced for all the genes, except the XG gene, for which two regions were sequenced, a 1.5 kb pseudoautosomal region adjacent to the pseudoautosomal boundary (referred to as PAB below) and a 1.3-kb region located 7 kb distally from the PAB (referred to as XG below). Although the PAB region is just an intronic sequence of the XG gene, for convenience, I will refer to it as to a separate gene.

The primers for PCR amplification and sequencing of PAR genes were designed based on human genomic sequences (table 1). The primers for the PAB, XG, SHOX, and PPP2R3L genes were described in a previous paper (Filatov and Gerrard 2003). The primers for all the other genes are listed in table 2.

All human sequences used in this study were taken from GenBank, and the orangutan sequences for PAB, XG, SHOX, and PPP2R3L were taken from the previous study (Filatov and Gerrard 2003). All the other ape sequences have not been previously published. The GenBank accession numbers for these sequences are AY296087 through AY296112.

Chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), and orangutan (*Pongo pygmaeus*) genomic DNA samples were purchased from the Coriel cell repository. All the genes were PCR amplified using the Roche High Fidelity polymerase chain reaction (PCR) kit under the following conditions: 95°C, 2.5 min, 57°C, 1 min, 68°C 3 min, followed by 34 cycles of 94°C 0.5 min, 57°C, 0.5 min, and 68°C 2.5 min, where X is the primer-specific annealing temperature (usually 55°C). The PCR products were gel-purified, extracted from the agarose gel using the Qiagen Gel Extraction kit, cloned into pCR4 plasmids with the TA cloning kit (Invitrogen) and were sequenced with the BigDye v3 sequencing system (ABI) on an ABI3700 automated sequencer. Chromatograms were checked and corrected by eye, and contigs were assembled manually using ProSeq software (Filatov 2002). Pairwise alignments were constructed manually or with the mcalign program (Peter Keightley and Toby Johnson, unpublished, available at the Web site http://homepages.ed.ac.uk/eang33/mcinstructions.html). The multiple alignments were constructed manually from pairwise alignments with the ProSeq software.

The maximum-likelihood (ML) analysis of the substitution patterns and rates was conducted using the baseml program (Yang 1997). All the ML analyses assumed that the closest human relative is chimpanzee, and that gorilla is the second closest, with orangutan being an outgroup. For the ML analysis of substitution patterns a user-defined general reversible model (REVs, Yang 1994) with different numbers of allowed rate parameters was used. For every model the analysis was conducted separately for individual genes, and the overall log-likelihood was obtained by summation of the log-likelihoods for individual genes (table 3). The significance in all the likelihood ratio tests was assessed using the approximation that the log-likelihood ratios are \( \chi^2 \)-distributed with the degrees of freedom equal to the difference in the number of parameters of the models compared (e.g., Muse and Weir 1992).

The GC-content and CpG analyses were conducted with the “GC-content” and “show sites” tools in the
ProSeq software (Filatov 2002). In all the correlation analyses the Pearson product-moment correlation coefficient, $r$ (Sokal and Rohlf 1995) was used.

**Results**

The 10 PAR genes analyzed in this study are listed in table 1. After exclusion of exon sequences and regions deleted in one of the four species (indels), the total number of aligned intron positions analyzed in this study was 13,554 bp. Because only noncoding sequences were analyzed, I will use “silent,” “intron,” and “noncoding” substitution rates as synonyms, referring to the substitution rate in introns of the genes analyzed.

**Substitution Rate Gradient in the p-PAR**

Before proceeding with the estimation of substitution rates in the p-PAR genes, an appropriate substitution model had to be selected. The nested ML-ratio analysis of substitution models (table 3) was conducted on all the genes considered separately, with the total log-likelihood value obtained by the summation of the log-likelihoods for individual genes. The analysis was conducted for separate substitution rates, the analysis was conducted for separate genes and the total log-likelihood was obtained by the summation of the log-likelihoods for individual genes. The no-clock model did not fit the data significantly better than the one-rate model in any of the genes or for the total data set (2$\Delta$lnL).

<table>
<thead>
<tr>
<th>Model</th>
<th>Substitution Rates</th>
<th>Intron Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>F81</td>
<td>Ct, Ga GC AT AT AC</td>
<td>-24159.99</td>
</tr>
<tr>
<td>HKY</td>
<td>(CT) (GA) (GC AT AT AC)</td>
<td>-23933.76 452.5***</td>
</tr>
<tr>
<td>HKY+1</td>
<td>(CT) (GA) (GC AT AT AC)</td>
<td>-23922.2 23.1*</td>
</tr>
<tr>
<td>HKY+2</td>
<td>(CT) (GA) (GC AT AT AC)</td>
<td>-23891.89 60.6***</td>
</tr>
<tr>
<td>HKY+3</td>
<td>(CT) (GA) (GC AT AT AC)</td>
<td>-23879.7 24.4**</td>
</tr>
<tr>
<td>REV</td>
<td>(CT) (GA) (GC AT AT AC)</td>
<td>-23875.14 9.1</td>
</tr>
</tbody>
</table>

Note. — The notation for substitution rates are according to Yang (1994, 1997).

**GC-Content**

GC-content is known to correlate with the silent substitution rate (Bilewski, Dunn, and Yang 2000; Yi, Ellsworth, and Li 2002); hence the distribution of GC-content across the p-PAR may be of interest for this study. It is clear from table 1 that GC-content is not equal among the studied genes and that there is a tendency for the GC-content to rise with distance from the PAR boundary. The correlation of the total tree length and the distance from the PAR boundary is highly significant ($r = 0.947, P < 0.001$).

To study whether the differences in the total tree length across the PAR are due to an acceleration of the substitution rate in some of the species, I compared the model with a single rate for all the branches to the model with branches having separate substitution rates (“no clock”). Because different genes have different substitution rates, the analysis was conducted for separate genes and the total log-likelihood was obtained by the summation of the log-likelihoods for individual genes. The no-clock model did not fit the data significantly better than the one-clock model in any of the genes or for the total data set (2$\Delta$lnL = 0.476, P > 0.05), providing no evidence for a significant difference in substitution rates between the species.
of GCi% with the distance from the PAR boundary is located 150 kb from the telomere. The positive correlation PAR boundary. More distally, in the

Thus, methylation may be one of the factors creating the

This is the most efficient way to correct for the effect of CpGs by removing all the sites preceded by C and followed by G.

Mutations at CpG dinucleotides may be an important factor that influences the substitution rate gradient. To test this hypothesis, I repeated the substitution rate analysis, excluding all the sites preceded by C or followed by G. This is the most efficient way to correct for the effect of CpG hypermutability (A. Y. Eyre-Walker, personal communication). As is clear from figure 1, this type of analysis results in reduction of the tree length in the distal genes, whereas the proximal genes remain almost unaffected, suggesting that the effect of methylation is weak near the PAR boundary and increases toward the telomere. Thus, methylation may be one of the factors creating the substitution rate gradient. However, the exclusion of the effect of CpGs does not result in disappearance of the substitution rate gradient; the correlation of the total tree length with distance from the PAB is still highly significant ($p = 0.89, P < 0.01$), however, suggesting that methylation cannot be the only factor creating the substitution rate gradient in the PAR.

Substantial reduction of the substitution rate in the distal genes, but not the proximal genes, after the exclusion of CpGs (fig. 1) suggests that methylation might be fairly strong in the distal region. Surprisingly, the CpG/GpC ratio is fairly high in the distal genes, especially in the the PPP2R3L gene (table 1). This ratio reflects methylation-driven depletion of the number of CpGs and is expected to be close to unity in the absence of methylation. The CpG/GpC ratio was reported to underestimate the effect of methylation in GC-rich regions (Duret and Galtier 2000), which may partially account for the observed high values of this ratio in the distal genes. Another explanation for this intriguing excess of CpGs in the region where methylation is apparently quite strong is that the number of CpGs in this region is not stationary, perhaps because it started to loose CpGs only recently. Indeed, the adjacent human 40-kb intron sequence of the PPP2R3L gene (GenBank accession number AF215839) is significantly less CpG rich than the region studied in this paper (CpG/GpC = 2025/3096 = 0.65, G-test = 6.38, $P = 0.012$).

Biased Gene Conversion in the PAR

If BGC operates in the p-PAR, the noncoding regions may not be neutral (Nagylaki 1983), and the substitution rate may be substantially accelerated or reduced by the BGC, depending on the GC-content and the mutation matrix (Eyre-Walker and Bulmer 1995). Although BGC may affect probabilities of fixation at (A or T)/(C or G) segregating sites, the probability of fixation at G/C and A/T segregating sites should remain unaffected by the process of BGC. If the gradient of silent substitution rate in the p-PAR is caused only by BGC, then the difference in substitution rates should be due to GC-changing substitutions, and not to G→C and A→T substitutions. It is clear from table 4 that in a pairwise human/orangutan comparison the number of A→T substitutions per 100 A or T sites ($D_{AT\%}$) and the number of G→C substitutions per 100 G or C sites ($D_{GC\%}$) correlates positively with the distance to the PAR boundary ($p = 0.633, P < 0.05$ and $p = 0.729, P < 0.05$, respectively), suggesting that, regardless of BGC, the silent substitution rate increases toward the telomere. As neither BGC nor methylation can affect G→C and A→T substitutions rates, there should be another major factor creating the substitution rate gradient in the p-PAR. Below, I argue that a gradient in recombination rate in the p-PAR may be such a factor.

Discussion

Substitution Rate Gradient

Here I have reported a significant difference in the silent substitution rates between the genes in the human and ape pseudoautosomal region. The silent substitution rate forms a gradient, with significantly more substitutions occurring in the more distal p-PAR genes. Average silent
divergence across the genome for human/chimpanzee (1%), human/gorilla (1.2%), and human/orangutan (3%) comparisons (Chen and Li 2001; Filatov and Gerrard 2003) is not significantly different from that found in the proximal p-PAR genes, but divergence beyond the proximal PAR region is significantly higher than that found in both the proximal PAR and in the non-PAR genes (Filatov and Gerrard 2003 and this study).

Assuming that substitution rate differences at non-coding sites reflect the differences in underlying mutation rate, we can use silent diversity in these regions to estimate mutation rates across the PAR. Under neutrality, divergence is equal to twice the product of the divergence time and the mutation rate, which may be used to estimate the mutation rate for relatively distant species like humans and orangutans (e.g., Li 1997). If the time since human/orangutan divergence is ~12 myr (Goodman et al. 1998) and we assume a generation time of 20 years, then the per-nucleotide per-generation mutation rate ranges from $2.7 \times 10^{-8}$ in the proximal 200 kb to $9 \times 10^{-8}$ in the distal PAR genes. The estimate of the mutation rate in the proximal PAR region is similar to the estimates published for other human genes, but the mutation rate estimates for the the distal PAR regions are somewhat higher than those for the human non-PAR genes (Nachman and Crowell 2000; Kondrashov 2002).

If the mutation rate is indeed higher in the p-PAR, we have to expect elevated DNA diversity in the p-PAR genes. Only two estimates of DNA diversity in the PAR are available, for the SHOX (May et al. 2002) and for PPP2R3L (Schiebel et al. 2000) genes. The two estimates contradict each other: according to May et al. (2002), the DNA diversity in the SHOX gene is not higher than elsewhere in the genome ($\theta \sim 0.07\%$), whereas the estimate of DNA diversity from the PPP2R3L sequences reported by Schiebel et al. (2000) is almost an order of magnitude higher ($\theta \sim 0.5\%$). Given the much higher mutation rate (taken into account by the HKA test [Hudson, Kreitman, and Aguade 1987]) in the p-PAR genes, the level of diversity in PPP2R3L is comparable with the estimates of the non-PAR genes, whereas the diversity in SHOX appears to be significantly reduced compared to PPP2R3L and to the non-PAR genes (Filatov and Gerrard 2003). As the DNA diversity in the SHOX gene was obtained mostly by genotyping of known SNPs (May et al. 2002), the rare segregating sites might be substantially under-represented in this data set, and it is not possible to apply the standard frequency spectrum-based techniques (i.e., Tajima 1989) to test whether the reduced diversity is caused by a recent selective sweep in this region. In principle, DNA diversity can be reduced as a consequence of frequent BGC, resulting in much faster fixation of (A or T)/(G or C) segregating sites (Nagylaki 1983). This hypothesis, as well as the level of DNA diversity in the pseudoautosomal genes, requires further investigation.

Possible Causes of the Substitution Rate Gradient in the PAR

The existence of the substitution gradient in the human PAR raises questions as to why the substitution rate is elevated in the PAR and why it is not uniform across the PAR.

Partial Y-Linkage

Y chromosomes are known to have elevated substitution (and mutation) rates because there are more cell divisions in the male germ line (e.g., Makova and Li 2002). Apart from that, the Silene latifolia Y chromosome was reported to have a higher per-cell division mutation rate than the X chromosome (Filatov and Charlesworth 2002), a finding that may also be true for human sex chromosomes. Because pseudoautosomal genes spend an equal amount of time in male and female gametes, the first factor does not apply to the PAR genes. However, if the human Y chromosome has an elevated per-cell division mutation rate, partial Y-linkage of the pseudoautosomal genes could account for some elevation of the mutation rate in the PAR. If this is the case, however, one would expect the substitution rate in the PAR genes to be only a fraction of that in the Y-linked genes. The human/orangutan silent divergence in Y-linked regions ranges from 3.9% in ZFY introns (Shimmin et al. 1993) to 8.2% in TSPY introns (Kim and Takenaka 1996), and is not significantly higher than in the p-PAR (data not shown), suggesting that a higher mutation rate on the Y chromosome cannot explain the elevated substitution rate in the p-PAR. Moreover, the partial Y-linkage of the PAR genes cannot explain the gradient in substitution rate observed, as one would expect partial Y-linkage to affect all the PAR genes equally.

Location of the PAR Boundary

The gradient in substitution rate could be explained by different evolutionary histories of the proximal and distal PAR regions. If the proximal PAR region is X-linked in one of the species studied, then the total tree length for this region would be shorter compared to more distal regions. However, this would result in substantial differences in the branch lengths of the phylogeny for the proximal region, which is not the case. Moreover, the position of the p-PAR boundary is well described and is known to be the same in humans, apes, and Old World monkeys (Ellis et al. 1990). It is thought to have been formed in the progenitor of simian primates as a result of the translocation of the SRY locus into the larger ancestral PAR (Glaser et al. 1999). Thus, the lower substitution rate in the proximal PAR region cannot be explained by X-linkage of this region in one or several of the species studied here.

Methylation

Methylation may substantially affect substitution patterns and rates, as frequent C $\Rightarrow$ T transitions from the deamination of 5-methylcytosine in methylated CpG dinucleotides occur at least an order of magnitude more frequently than other types of mutations (Robertson and Wolfe 2000). A much stronger reduction in the estimates of the substitution rate in the distal region after removal of CpG dinucleotides (fig. 1) suggests that methylation might
Table 4
Correlation of the A ↔ T and G ↔ C Substitution Rates With Distance to the Pseudoautosomal Boundary (PAB) in a Pairwise Human/Orangutan Comparison

<table>
<thead>
<tr>
<th>Genes</th>
<th>Kb to PAB</th>
<th>S^a</th>
<th># A (\rightarrow) T^b</th>
<th>(D_{AT}^d)</th>
<th>S^b</th>
<th># G (\rightarrow) C^c</th>
<th>(D_{GC}^e)</th>
<th>Sum (D_{AT-GC}^f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAB</td>
<td>0</td>
<td>2</td>
<td>856</td>
<td>0.234</td>
<td>9</td>
<td>638</td>
<td>1.411</td>
<td>1.644</td>
</tr>
<tr>
<td>XG</td>
<td>7</td>
<td>2</td>
<td>727</td>
<td>0.275</td>
<td>3</td>
<td>534</td>
<td>0.562</td>
<td>0.837</td>
</tr>
<tr>
<td>MIC2</td>
<td>60</td>
<td>4</td>
<td>1345</td>
<td>0.297</td>
<td>8</td>
<td>805</td>
<td>0.994</td>
<td>1.291</td>
</tr>
<tr>
<td>LI6</td>
<td>156</td>
<td>2</td>
<td>998</td>
<td>0.2</td>
<td>7</td>
<td>814</td>
<td>0.86</td>
<td>1.06</td>
</tr>
<tr>
<td>LI5</td>
<td>200</td>
<td>4</td>
<td>857</td>
<td>0.467</td>
<td>15</td>
<td>725</td>
<td>2.069</td>
<td>2.536</td>
</tr>
<tr>
<td>DHRXSY</td>
<td>350</td>
<td>7</td>
<td>921</td>
<td>0.76</td>
<td>12</td>
<td>665</td>
<td>1.805</td>
<td>2.565</td>
</tr>
<tr>
<td>ASMT</td>
<td>910</td>
<td>4</td>
<td>369</td>
<td>1.084</td>
<td>6</td>
<td>435</td>
<td>1.379</td>
<td>2.463</td>
</tr>
<tr>
<td>ASMTL</td>
<td>1095</td>
<td>4</td>
<td>575</td>
<td>0.696</td>
<td>16</td>
<td>764</td>
<td>2.094</td>
<td>2.79</td>
</tr>
<tr>
<td>SHOX</td>
<td>1730</td>
<td>4</td>
<td>231</td>
<td>1.732</td>
<td>8</td>
<td>410</td>
<td>1.951</td>
<td>3.683</td>
</tr>
<tr>
<td>PPP</td>
<td>2095</td>
<td>1</td>
<td>199</td>
<td>0.503</td>
<td>12</td>
<td>468</td>
<td>2.564</td>
<td>3.067</td>
</tr>
</tbody>
</table>

Correlation with distance to PAB

\[0.633^*\] \[0.729^*\] \[0.818^{**}\]

\(^{a}\) The number of the A \(\rightarrow\) T and G \(\rightarrow\) C substitutions between human and orangutan PAR genes.

\(^{b}\) The number of (A or T) positions in the data set. Includes the number of constant positions with A or T and the positions with A \(\rightarrow\) T substitutions.

\(^{c}\) The number of (G or C) positions in the data set. Includes the number of constant positions with G or C and the positions with G \(\rightarrow\) C substitutions.

\(^{d}\) The number of A \(\rightarrow\) T substitutions per 100 (A or T) sites.

\(^{e}\) The number of G \(\rightarrow\) C substitutions per 100 (G or C) sites.

\(^{f}\) The sum of the values in footnotes d and e.

\(\ast P < 0.05; \^{*\ast} P < 0.01.\)

be more severe in the distal genes than in those of the proximal PAR region, and it may be one of the factors creating the substitution rate gradient. However, the effect of methylation is clearly not sufficient to explain all the variation in substitution rates among the PAR genes because, after removal of all the CpGs, the substitution rate difference between the proximal region and the rest of the PAR is still significant.

Biased Gene Conversion

Taking into account only the A \(\rightarrow\) T and C \(\rightarrow\) G transversions, which are not affected by BGC, the substitution rate gradient is still significant, demonstrating that BGC is insufficient to explain the observed substitution rate gradient. However, I cannot reject the hypothesis that BGC influences, to some extent, the substitution rate gradient. One would expect BGC to accelerate the substitution rate in the GC-poor proximal region and to reduce it in the more distal GC-rich genes, as it accelerates fixation of AT \(\rightarrow\) GC mutations (which should be fairly frequent in the AT-rich proximal region) and reduces the probability of fixation of the GC \(\rightarrow\) AT mutations (which should be fairly frequent in the GC-rich distal region). Thus, if anything, BGC is likely to diminish rather than create the observed gradient in substitution rates in the PAR.

Variation in GC-Content

The substitution rate in the p-PAR demonstrates a highly significant correlation with GC-content. To a large extent this correlation seems to be due to abundant CpG dinucleotides in GC-rich distal genes. However, even after the removal of CpG, the value of the correlation coefficient is quite high \((\rho = 0.57, P = 0.08)\), and with more data points it may well become significant, suggesting that even after exclusion of the effect of methylation there may be an association between GC-content and substitution rate. One way that GC-content can affect substitution rate is to change the frequencies of nucleotides which mutate with different rates. For example, if G and C mutate more frequently than A or T, then regions with higher GC-content will have a higher mutation (and silent substitution) rate. If this is the main cause of the substitution rate gradient, then the correction for the GC-content should result in the disappearance of the substitution rate gradient. However, this is clearly not the case for the A \(\rightarrow\) T and G \(\rightarrow\) C substitution rates after the correction for the GC-content (table 4).

Mutagenic Recombination

The mutagenic effect of recombination seems to be a very attractive explanation for the elevated substitution rate in the PAR. If the substitution gradient in the PAR is driven by variation in the recombination rate in different PAR regions, one would expect the frequency of recombination close to the pseudoautosomal boundary to be lower than in the distal PAR. Unfortunately, the best available estimates of recombination in the p-PAR (Lien et al. 2000) are not detailed enough, and no reliable estimates of the recombination rate close to the PAR boundary are available. According to cytological observations, pairing in the human p-PAR is much more frequent near the telomere than in the more proximal regions (S. Armstrong, unpublished data), suggesting that the recombination rate might drop toward the pseudoautosomal boundary. If the recombination rate is indeed lower near the PAR boundary, this would resemble the suppression of recombination near chromosomal inversions (Novitski and Braver 1954; Coyne et al. 1993). However, we need more precise estimates of the recombination rate in the p-PAR genes to corroborate or reject this hypothesis.

Interestingly, in the mouse PAR there is also a gradient of substitution rate (e.g., Fig. 6 in Birdsell...
2002), although it is much steeper than in the human PAR and extends over only a few kilobases. However, the mouse PAR is much smaller than the human p-PAR, thus, the per-nucleotide recombination rate might be much higher in the mouse PAR, which could result in this difference in gradient scale between mouse and human PARs. Indeed, the difference in substitution rates between the PAR and non-PAR genes is substantially higher in mice (Perry and Ashworth 1999) than the difference observed in humans (Filatov and Gerrard 2003; and this paper), suggesting that recombination (if it is the cause of the higher substitution rate) may be much more frequent in the mouse PAR.

The mutagenic recombination hypothesis is supported by several reports of the positive correlation of human recombination rate with human/mouse (Lercher and Hurst 2002; Waterston et al. 2002; Hardison et al. 2003), human/chimpanzee, and human/baboon divergence (Hellmann et al. 2003). However, we failed to detect a positive correlation of human/orangutan silent divergence with the recombination rate in humans (Filatov and Gerrard 2003). Moreover, the SHOX region, which has previously been reported to be a hot spot of recombination in humans (as high as 300 cM/Mb [May et al 2002]) does not show any signs of a significantly elevated substitution rate, when compared with adjacent PAR genes (Filatov and Gerrard 2003 and this study). This may be because recombinational hot spots are very short-lived (Boulton, Myers, and Redfield 1997), and therefore it is possible that there has not been enough time for the SHOX region to accumulate substitutions.

If recombination is mutagenic, it is surprising that it has not been detected in extensive Drosophila genetics and genomics studies. Takano-Shimizu (2001) reported changes in recombination rate, substitution rate, and GC/AT substitution bias in a sub-telomeric region of the X chromosome in three Drosophila species. On the one hand, a 10-fold reduction in the recombination frequency in the D. melanogaster lineage was coupled with an AT-biased substitution pattern and a significant acceleration of the rate of silent substitutions, which is in contrast to what would be expected if recombination were mutagenic. On the other hand, in D. orena a significant increase in the silent substitution rate coupled with GC-substitution bias was observed, compared to a closely related D. erecta (Takano-Shimizu 2001), resembling the mutagenic effect of recombination observed in human and mouse PARs. Unfortunately, the study does not contain the data on the recombination rate in D. orena or a comparison of the recombination rates between D. orena and D. erecta, and this makes it difficult to assess whether the acceleration of the substitution rate in the D. orena subtelomeric region was due to changes in the recombination rate.

Conclusions

I have described a silent substitution rate gradient in the human and ape p-PAR. The existence of this gradient cannot be fully explained by the differences in GC-content, methylation, or biased gene conversion across the p-PAR. I suggest that the substitution rate gradient might be due to an underlying gradient of recombination rate; however, the available estimates of recombination rate across the p-PAR do not allow a test of this hypothesis.

Acknowledgments

I thank Brian Charlesworth for a helpful discussion, Dave Gerrard for critical reading of the manuscript, Adam Eyre-Walker and the anonymous reviewers for criticism and helpful comments. The work was supported by the Wellcome Trust under grant number 068193.

Literature Cited


Accepted September 29, 2003