

A gradient of silent substitution rate in the human pseudoautosomal region

Dmitry A. Filatov

School of Biosciences, University of Birmingham, Birmingham, B15 2TT, UK

Correspondence:

Dmitry A. Filatov,

School of Biosciences,

University of Birmingham

Edgbaston,

Birmingham B15 2TT

United Kingdom

e-mail: d.filatov@bham.ac.uk

phone: 44-121-4142500

fax: 44-121-4145925

Keywords: human, pseudoautosomal region, recombination, silent substitution rate, mutation rate.

Abstract:

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 ten human, chimpanzee, gorilla and orangutan genes across the p-PAR. Between species silent divergence in the p-PAR forms a gradient, increasing towards the telomere. The correlation of silent divergence with distance from the p-PAR boundary is highly significant ($\rho = 0.911$, $P < 0.001$). After exclusion of the CpG dinucleotides this correlation is still significant ($\rho = 0.89$, $P < 0.01$), thus the substitution rate gradient can not 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 may affect substitution rates due to the increased probability of fixation of the G or C nucleotides at (A or T)/(G or C) segregating sites. BGC, however, does not seem to be the factor creating the substitution rate gradient in the p-PAR, because the gradient is still detectable if only $A \leftrightarrow T$ and $G \leftrightarrow C$ substitutions are taken into account ($\rho = 0.82$, $P < 0.01$). I hypothesize that the substitution rate gradient in the p-PAR is due to 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 $\sim 1\text{cM/Mb}$ (Lien et al. 2000), and may even be as high as 350cM/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 5cM/Mb (Ciccodicola et al. 2000).

Similar to 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), demonstrating 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 fourfold 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 et al. 1995) strengthens the evidence in favour of this hypothesis.

However, several other factors can result in the correlation between recombination and silent substitution rates. Both recombination and substitution rates correlate positively with GC-content (Eyre-Walker 1993; Fullerton et al. 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 \Rightarrow TpG mutations (Robertson and Wolffe 2000) may also impact towards the correlation of recombination and substitution rates, as the frequency of CpG dinucleotides was 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 towards GC (Lamb 1986; Brown and Jiricny 1988), may be another such factor. BGC has been demonstrated to be mathematically equivalent to selection for (A or T) \Rightarrow (G or C) mutations (Nagylaki 1983), thus it may substantially affect fixation rates at (A or T) / (G or C) segregating sites and the overall substitution rates (Eyre-Walker and Bulmer 1995). As BGC might be more frequent in recombinational hotspots, it may also be one of the causes of the correlation between the substitution and recombination rates.

In this paper I report the analysis of substitution rates in ten genes across the human and ape p-PAR, aiming to distinguish between the possible causes of the elevated substitution rate in this peculiar genomic region. I demonstrate that the substitution rate in the p-PAR is not uniform. In fact, it forms a gradient, rising with distance from the p-PAR boundary. The results of the analysis suggest that variation in GC-content, methylation and biased gene conversion are not sufficient to explain the existence of the gradient. I hypothesize that this substitution rate gradient is due to mutagenic effect of recombination, which is very frequent in the distal region (as high as >300 cM/Mb in some regions, May et al. 2002), and might be substantially lower near the PAR boundary, which probably acts as a suppressor of recombination, resembling the suppression of recombination in the proximity of chromosomal inversions (Novitski and Braver 1954; Coyne et al. 1993).

Materials and Methods

To study substitution patterns and rates in the frequently recombining human and ape p-arm pseudoautosomal 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, I will also refer to it as to a separate gene, for convenience.

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 the previous paper (Filatov and Gerrard 2003). The primers for all the other genes are listed in table 2.

All the 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-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 PCR kit under the following conditions: 95C, 2.5 min., 57C, 1 min., 68C 3 min., followed by 34 cycles of 94C 0.5 min., *X* C, 0.5 min. and 68C 2.5min, where *X* is the primer-specific annealing temperature (usually 55C). The PCR products were gel-purified, extracted from the agarose gel using the Qiagen Gel Extraction kit, cloned into pCR4 plasmids using the TA cloning kit (Invitrogen) and sequenced using 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 using 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 using 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 gorilla is the second closest, with orangutan being an outgroup. For the ML analysis of substitution patterns a user-defined general reversible model (REVu, 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 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 χ^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 using the “GC-content” and “show sites” tools in the ProSeq software (Filatov 2002). In all the correlation analyses the Pearson product-moment correlation coefficient, ρ (Sokal and Rohlf 1995) was used.

Results

The ten PAR genes analysed 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 analysed in this study was 13554 bp. As only non-coding sequences were analysed, I will use “silent”, “intron” and “non-coding” substitution rates as synonyms, referring to the substitution rate in introns of the genes analysed.

Substitution rate gradient in the p-PAR: Before proceeding with the estimation of substitution rates in the p-PAR genes, an appropriate substitution model needed 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. This analysis demonstrated that the HKY model (Hasegawa, Kishino, and Yano 1985) gave a drastic improvement compared to the F81 model (Felsenstein 1981), reflecting the fact that C \leftrightarrow T and G \leftrightarrow A transitions occur much more frequently than transversions. Separate substitution parameters for the two transition types (T \leftrightarrow C and A \leftrightarrow G) resulted in a significant improvement of the likelihood value, suggesting that the two transition types occur with different rates. Further sequential addition of separate parameters for G \leftrightarrow C and A \leftrightarrow T transversions also improved the fit of

the model to the data. However, the further addition of separate parameters for G↔T and A↔C transversions does not improve the likelihood value, suggesting that they are not significantly different. Thus, to adequately describe substitution patterns in the PAR, five substitution rate parameters (four free rates) for C↔T and G↔A transitions, C↔G, A↔T, and all the other transversions are required (“HKY+3 rates” model). This model will be used below, unless stated otherwise.

The distribution of intron site divergence in the p-PAR is shown in figure 1. The substitution rate, expressed as a total tree length, forms a gradient with divergence between the species increasing with distance from the PAR boundary. The correlation of the total tree length and the distance from the PAR boundary is highly significant ($\rho = 0.947$, $P < 0.001$).

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”). As different genes have different substitution rates, the analysis was conducted for separate genes and the total log-likelihood 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, nor for the total dataset ($2\Delta\ln L=0.476$, $P > 0.05$), providing no evidence for a significant difference in substitution rates between the species.

GC-content: GC-content is known to correlate with the silent substitution rate (Bielawski, 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 the 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 the distance from the PAR boundary. The proximal part adjacent to

the PAR boundary is relatively GC-poor (intronic GC% < 40%). Intronic GC-content (GC_i%) rises to 42-47% in the region including *L16*, *L15* and *DHRSXY* genes. GC_i% reaches 56% in the region including *ASMT* and *ASMTL* genes further away from the PAR boundary. More distally, in the *SHOX* gene GC_i% = 59%, and increases further to 67% in the *PPP2R3L* gene located 150 kb from the telomere. The positive correlation of GC_i% with the distance from the PAR boundary is highly significant ($\rho = 0.96$, $P < 0.001$). As the substitution rate is also increasing with the distance from the PAR boundary, it is hardly surprising that GC_i% in the p-PAR positively correlates with the silent substitution rate ($\rho = 0.88$, $P = 0.001$).

Methylation: C=>T transitions due to deamination of 5-methylcytosine in methylated CpG pairs are known to occur with at least an order of magnitude higher frequency than the other mutations (Robertson and Wolffe 2000). Transversion rate is also somewhat higher at CpG dinucleotides (Nachman and Crowell 2000; Kondrashov 2002). Thus, mutations at CpG dinucleotides may be an important factor that impacts towards 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 (Eyre-Walker, personal communication). As it is clear from figure 1, this results in reduction of the tree length in the distal genes, while the proximal genes are almost unaffected by this, suggesting that the effect of methylation is weak near the PAR boundary and increases towards the telomere. Thus, methylation may be one of the factors creating the substitution rate gradient. However, the exclusion of all the effect of CpGs does not result in disappearance of the substitutiou rate gradient – the correlation of the total tree length with distance from the PAB is still highly significant ($\rho = 0.89$, $P < 0.01$), suggesting that methylation can not be the only factor creating the substitution rate gradient in the PAR.

Substantial reduction of the substitution rate in the distal, but not the proximal genes after the exclusion of CpGs (figure 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 partly 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. The *PPP2R3L* gene, where the CpG/GpC ratio is especially high ($69 / 68 \sim 1$), might have been a CpG island and became heavily methylated only recently. Indeed, the adjacent human 40 kb intron sequence of the *PPP2R3L* gene (GenBank accession AF215839) is significantly less CpG rich than the region studied in this paper ($\text{CpG} / \text{GpC} = 2025 / 3096 = 0.65$, G-test = 6.38, $P = 0.012$).

Biased gene conversion in the PAR: If BGC operates in the p-PAR, the non-coding regions may not behave as 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 only due to BGC, then the difference in substitution rates should be due to GC-changing substitutions, and not due to $G \leftrightarrow C$ and $A \leftrightarrow T$ substitutions. It is clear from table 4 that in a pairwise human/orangutan comparison the number of $A \leftrightarrow T$ substitutions per 100 A or T sites ($D_{AT\%}$) and the number of $G \leftrightarrow C$

substitutions per 100 G or C sites ($D_{GC\%}$) correlate positively with the distance to the PAR boundary ($\rho = 0.633$, $P < 0.05$ and $\rho = 0.729$, $P < 0.05$, respectively), suggesting that regardless of BGC, the silent substitution rate increases towards the telomere. As neither BGC, nor methylation can affect $G \leftrightarrow C$ and $A \leftrightarrow T$ 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 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 those 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 divergence 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 million years (Goodman et al. 1998) and we assume a generation time of 20 years, then the per-nucleotide per-generation mutation rate ranges from 2.7×10^{-8} in the proximal 200 kb to 9×10^{-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 the estimates 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 *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 ($\theta \sim 0.07\%$), while 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 compatible to the estimates of the non-PAR genes, while 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 dataset 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 due to 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) rate due to more cell divisions in the male germ line (e.g Makova and Li 2002). Apart of that, the *Silene latifolia* Y chromosome was reported to have an elevated per-cell division mutation rate, compared to the X chromosome (Filatov and Charlesworth 2002), which may also be true for human sex chromosomes. As 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 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. However, in this case 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 can not 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, 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 due to 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 can not 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 due to 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 Wolffe 2000). A much stronger reduction in the estimates of the substitution rate in the distal region after removal of CpG dinucleotides (figure 1) suggests that methylation might be more severe in the distal genes compared to 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 as, 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 $A \Leftrightarrow T$ and $C \Leftrightarrow G$ 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 can not reject the hypothesis that BGC impacts to some extent towards the the substitution rate gradient. One would expect BGC to accelerate the substitution rate in the GC-poor proximal region and 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 how 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 $A \leftrightarrow T$ and $G \leftrightarrow C$ 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, compared to more proximal regions (S. Armstrong, unpublished data), suggesting that recombination rate might drop towards 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 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), though it is much steeper 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 the mouse and the human PARs. Indeed, the difference in substitution rates between the PAR and non-PAR genes is substantially higher in mice (Perry and Ashworth 1999) compared to 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 human recombination rate in humans (Filatov and Gerrard 2003). Moreover, the *SHOX* region, which has previously been reported as a hotspot of recombination in humans (as high as 300 cM/Mb, May et al 2002) does not show any signs of significantly elevated substitution rate, compared to adjacent PAR genes (Filatov and Gerrard 2003 and this study). This may be because recombinational hotspots are very short-lived (Boulton, Myers, and Redfield 1997), 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 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. 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*, nor a comparison of the recombination rates between *D. orena* and *D. erecta*, making 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 described a silent substitution rate gradient in the human and ape p-PAR. The existence of this gradient can not 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.

Acknowledgements: 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 grant N^o 068193.

References

- BIELAWSKI, J. P., K. A. DUNN, and Z. YANG. 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. *Genetics* **156**:1299–1308.
- BIRDSELL, J. A. 2002. Integrating genomics, bioinformatics and classical genetics to study the effect 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. Natl. Acad. Sci. USA* **94** :8058-8063.
- BROWN, W. R. A. 1988. A physical map of the human pseudoautosomal region. *EMBO J.* **7**:2377-2385.
- 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.
- BURGOYNE, P. S., S. K. MAHADEVAIAH, M. J. SUTCLIFFE, and S. J. PALMER. 1992. Fertility in mice requires X-Y pairing and a Y-chromosomal spermiogenesis gene mapping to the long arm. *Cell* **71**:391-398.
- 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.
- CICCODICOLA, A., M. D'ESPOSITO, T. ESPOSITO, et al., 2000. Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. *Hum. Mol. Genet.* **9**:395-401.
- COOKE, H. J., W. R., BROWN, and G. A. RAPPOLD. 1985. Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. *Nature* **317**:687-692.

- COYNE, J. A., W. MEYERS, A. CRITTENDEN, and P. SNIEGOWSKI. 1993. The fertility effects of pericentric inversions in *Drosophila melanogaster*. *Genetics* **134**:487-496.
- DURET, L. and N. GALTIER. 2000. The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artefact. *Mol. Biol. Evol.* **17**:1620-1625.
- ELLIS, N., and P. N. GOODFELLOW. 1989. The mammalian pseudoautosomal region. *Trends Genet.* **5**:406-410.
- ELLIS, N., P. YEN, K. NEISWANGER, L. J. SHAPIRO, and P. N. GOODFELLOW. 1990. Evolution of the pseudoautosomal boundary in the old world monkeys and great apes. *Cell* **63**:977-986.
- EYRE-WALKER, A. 1993. Recombination and mammalian genome evolution. *P. Roy. Soc. Lond. B Bio.* **252**: 237-243.
- EYRE-WALKER, A., and M. BULMER. 1995. Synonymous substitution rates in enterobacteria. *Genetics* **140**:1407-1412.
- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**:368-376.
- FILATOV, D. A. 2002. ProSeq: A software for preparation and evolutionary analysis of DNA sequence data sets. *Mol. Ecol. Notes* **2**:621-624.
- FILATOV, D. A., and D. CHARLESWORTH. 2002. Substitution rates in the X- and Y-linked genes of the plants, *Silene latifolia* and *S. dioica*. *Mol. Biol. Evol.* **19** :898-907.
- FILATOV, D. A., and D. T. GERRARD. 2003. High mutation rates in human and ape pseudoautosomal genes. *Gene*, *in press*.
- FREIJE, D., C. HELMS, M. S. WATSON, and H. DONIS-KELLER. 1992. Identification of a second pseudoautosomal region near the Xq and Yq telomeres. *Science* **258**:1784-1787.

- FULLERTON, S. M., A. B. CARVALHO, and A. G. CLARK. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* **18**:1339-1142.
- GLASER, B., D. MYRTEK, Y. RUMPLER, K. SCHIEBEL, M. HAUWY, G. A. RAPPOLD, and W. SCHEMPP. 1999. Transposition of *SRY* into ancestral pseudoautosomal region creates a new pseudoautosomal boundary in a progenitor of simian primates. *Hum. Mol. Genet.* **8**:2071-2078.
- GOODMAN, M., C. A. PORTER, J. CZELUSNIAK, S. L. PAGE, H. SCHNEIDER, J. SHOSHANI, G. GUNNELL, and C. P. GROVES. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Pol. Phylogenet. Evol.* **9**:585-598.
- HARDISON R. C., K. M. ROSKIN, S. YANG, M. DIEKHANS, W. J. KENT, et al. (13 more authors) 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13**:13-26.
- HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:160-174.
- 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.
- HUDSON, R. R., M. KREITMAN, and M. AGUADE. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**:153-159.
- KIM, H.S., and O. TAKENAKA. 1996. A comparison of TSPY genes from Y-chromosomal DNA of the great apes and humans: sequence, evolution, and phylogeny. *Am. J. Phys. Anthropol.* **100**:301-309.

- KONDRASHOV, A. S. 2002. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Human mutation* **21**:12-27.
- KONG, A., D. F. GUDBJARTSSON, J. SAINZ, G. M. JONSDOTTIR, S. A. GUDJONSSON, et al. (12 more authors), 2002. A high-resolution recombination map of the human genome. *Nature Genetics* **31**:241-247.
- LAMB, B. C. 1986. Gene conversion disparity - factors influencing its direction and extent, with tests of assumptions and predictions in its evolutionary effects. *Genetics* **114**:611-632.
- 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.
- LI, W.-H. 1997. *Molecular Evolution*. Sinauer Ass. Sunderland, Mass.
- 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.
- MAKOVA, K. D. and W.- H. LI. 2002. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**:624-626.
- MAY, C. A., A. C. SHONE, L. KALAYDJIEVA, A. SAJANTILA, and A. J. JEFFREYS. 2002. Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene *SHOX*. *Nat. Genet.* **31**:272-275.
- MUSE, S. V., and B. S. WEIR. 1992. Testing for equality of evolutionary rates. *Genetics* **132**:269-276.
- NACHMAN, M. W., and S. CROWELL. 2000. Estimates of the mutation rate per nucleotide in humans. *Genetics* **156**:297-304.

- NAGYLAKI, T. 1983. Evolution of a finite population under gene conversion. Proc. Natl. Acad. Sci. USA **80**:6278-6281.
- NOVITSKI, E., and G. BRAVER. 1954. An analysis of crossing over within a heterozygous inversion in *Drosophila melanogaster*. Genetics **39**:197-209.
- PERRY, J., and A. ASHWORTH. 1999. Evolutionary rate of a gene affected by chromosomal position. Curr. Biol. **9**:987-989.
- PETIT, C., J. LEVILLIERS, and J. WEISSENBACH. 1988. Physical mapping of the human pseudoautosomal region; comparison with the genetic linkage map. EMBO J. **7**:2369-2376.
- ROBERTSON, K. D., and A. P. WOLFFE. 2000. DNA methylation in health and disease. Nat. Rev. Genet. **1**:11-19.
- SCHIEBEL, K., J. MEDER, A. RUMP, A. ROSENTHAL, M. WINKELMANN, C. FISCHER, T. BONK, A. HUMENY, and G. RAPPOLD. 2000. Elevated DNA sequence diversity in the genomic region of the phosphatase *PPP2R3L* gene in the human pseudoautosomal region. Cytogenet. Cell Genet. **91**:224-230.
- SHIMMIN, L.C., B. H. J. CHANG, and W.-H. LI. 1993. Male-driven evolution of DNA sequences. Nature **362**:745-747.
- SOKAL, R. R., and F. J. ROHLF. 1995. Biometry. Third edition. Freeman, San Francisco.
- SORIANO P., E. A. KEITGES, D. F. SCHORDERET, K. HARBERS, S. M. GARTLER, and R. JAENISCH. 1987. High-rate of recombination and double crossovers in the mouse pseudoautosomal region during male meiosis. Proc. Natl. Acad. Sci. USA **84**:7218-7220.
- STRATHERN, J. N., B. K. SHAFER, and C. B. MCGILL. 1995. DNA synthesis errors associated with double-strand-break repair. Genetics **140**:965-972.

- TAKANO-SHIMIZU, T. 2001. Local changes in GC/AT substitution biases and in crossover frequencies on drosophila chromosomes. *Mol. Biol. Evol.* **18**:606-619.
- TAJIMA, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585-595.
- WATTERSON R. H., K. LINDBLAD-TOH, E. BIRNEY, J. ROGERS, J. F. ABRIL et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**:520-562.
- YANG, Z. 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**:105-111.
- YANG, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42**: 587-596.
- YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood *CABIOS* **13**:555-556
- YI, S., D. L. ELLSWORTH, and W.-H. LI. 2002. Slow molecular clocks in old world monkeys, apes, and humans. *Mol. Biol. Evol.* **19**:2191–2198.

Figure legends

Figure 1. The gradient of intron divergence in the human, chimpanzee, gorilla and orangutan p-PAR expressed as the total tree length. The solid line joins the data points for all the intron sites included; the dashed line is the total tree length for the intron sequences after exclusion of the effect of CpGs by removing all the sites preceded by C and followed by G.

Table 1. p-PAR regions analysed.

#	Gene	Kb	Human genomic contig			PCR primers		Positions	GC%	CpG/GpC
	name	to PAB	accession	start	end	forward	reverse	analysed introns ¹⁾	ratio ¹⁾	
0	PAB	0	NT_025302.11	429417	430962	PAB+1	PAB-4x	1507	43	20 / 66
1	XG	7	NT_025302.11	424249	425561	XG+1	XG-2	1298	43	19 / 56
2	MIC2	60	NT_025302.11	369853	372122	MIC2+3	MIC2-4	2189	38	15 / 93
3	L254916	156	NT_025302.11	277306	275299	L254916+4	L254916-3	1857	45	14 / 69
4	L254915	200	NT_025302.11	233017	231298	L254916+4	L254915-5	1620	47	36 / 78
5	DHRXY	350	NT_025302.11	79247	77392	DHR+3	DHR-2	1673	42	20 / 64
6	ASMT	910	NT_033330.5	244101	245253	ASMT+6	ASMT-7	854	56	42 / 62
7	ASMTL	1095	NT_033330.5	57639	56106	ASMTL+4	ASMTL-2	1309	56	55 / 104
8	SHOX	1730	AC137591	29617	30501	SHOX+1	SHOX-3	660	64	47 / 63
9	PPP2R3L	2095	NT_077814.2	28170	27161	PPP+1	PPP-5	587	67	69 / 68

¹⁾ Given for the human sequence; ape sequences may have slightly different values.

Table 2. PCR and sequencing primers. The primers for the *XG*, *SHOX* and *PPP2R3L* genes were described in the previous paper (Filatov and Gerrard 2003).

Gene	Primer name	Sequence (5'-3')
<i>MIC2</i>	MIC2+1	TTCAGATGCTGACCTTGCGG
<i>MIC2</i>	MIC2-2	TTCTTTCCTGTGGCTGCCTC
<i>MIC2</i>	MIC2+3	CGAACCCACCCAAACCGATG
<i>MIC2</i>	MIC2-4	CCATCCGCAAGGTCAGCATC
<i>MIC2</i>	MIC2+5	AAGAGCATAACCACAGCCTCCAAC
<i>MIC2</i>	MIC2-6	AAATGTTCATGGACTATCAGTAAGC
<i>L254916</i>	L254916+1	GGGAAATTAACCTTGCACCTAGCTG
<i>L254916</i>	L254916-2	TTACCTGAAGGGAGATGGTGATG
<i>L254916</i>	L254916-3	AAAATGTAGTCATTTTCAGGAAGGGTC
<i>L254916</i>	L254916+4	AACTCCTTCATCCTCTGATGCAG
<i>L254915</i>	L254915-2	CGCTCACCGTGTTTGTCCATG
<i>L254915</i>	L254915-3	TCCCAACCCTTTGCAGACAC
<i>L254915</i>	L254915+4	GGTGGGTCCAATGCGCTATG
<i>L254915</i>	L254915-5	GCGCCCTACCATTTGAATTCTG
<i>L254915</i>	L254915+6	TTGGGATGGTTAGGGTGTGAGTG
<i>L254915</i>	L254915-13	GGTGGGTTTGTGCTTGGC
<i>DHRSXY</i>	DHR+1	GAGGGATACACCAGAAAACCTAAGC
<i>DHRSXY</i>	DHR-2	AAAGCCTAAGTGTCCACCATCAG
<i>DHRSXY</i>	DHR+3	ATATTTTGACTCTCCCTCTGCCG
<i>DHRSXY</i>	DHR-4	AAATGCCCATCAACGATAGACTG
<i>ASMT</i>	ASMT+6	TGGGCGTGTTTGACCTTCTC
<i>ASMT</i>	ASMT+30	GGTTGCAGTGAGCCGAGATCG
<i>ASMT</i>	ASMT-7	TGACCGTGGTCAGGTAGTCG
<i>ASMTL</i>	ASMTL+1	GCTGATTCTGGAGAAGCCGG
<i>ASMTL</i>	ASMTL-2	CCTGTGAACACGCTGTGTTCTC
<i>ASMTL</i>	ASMTL-3	TCTGTGACTGTAGGGGCTGAGG
<i>ASMTL</i>	ASMTL+4	GTGTCTGTCAGTGGGTAAATGGG
<i>ASMTL</i>	ASMTL+5	GTTGAGGACATCACGGTCAGTG
<i>ASMTL</i>	ASMTL-6	GATCAGTGCTGCAGAATCCTAG

Table 3. The nested ML analysis of substitution models in the introns of ten p-PAR genes in human, chimpanzee, gorilla and orangutan. The notation for substitution rates are according to Yang (1994, 1997). Likelihood ratio tests [$2(\Delta \ln L)$] are for the comparison with a less parameter rich model one row above.

Model	Substitution rates	Intron sites	
		lnL	$2(\Delta \ln L)$
F81	(CT GA GC AT GT AC)	-24159.99	
HKY	(CT GA)(GC AT GT AC)	-23933.76	452.5***
HKY+1	(CT)(GA)(GC AT GT AC)	-23922.2	23.1*
HKY+2	(CT)(GA)(GC)(AT GT AC)	-23891.89	60.6***
HKY+3	(CT)(GA)(GC)(AT)(GT AC)	-23879.7	24.4**
REV	(CT)(GA)(GC)(AT)(GT)(AC)	-23875.14	9.1

* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$

Table 4. The correlation of the A↔T and G↔C substitution rates with distance to the pseudoautosomal boundary (PAB) in a pairwise human/orangutan comparison.

Genes	Kb to PAB	A↔T			G↔C			Sum
		S ¹⁾	# A T ²⁾	D _{AT%} ⁴⁾	S ¹⁾	# G C ³⁾	D _{GC%} ⁵⁾	D _{AT+GC%} ⁶⁾
PAB	0	2	856	0.234	9	638	1.411	1.644
<i>XG</i>	7	2	727	0.275	3	534	0.562	0.837
<i>MIC2</i>	60	4	1345	0.297	8	805	0.994	1.291
<i>L16</i>	156	2	998	0.2	7	814	0.86	1.06
<i>L15</i>	200	4	857	0.467	15	725	2.069	2.536
<i>DHRSXY</i>	350	7	921	0.76	12	665	1.805	2.565
<i>ASMT</i>	910	4	369	1.084	6	435	1.379	2.463
<i>ASMTL</i>	1095	4	575	0.696	16	764	2.094	2.79
<i>SHOX</i>	1730	4	231	1.732	8	410	1.951	3.683
<i>PPP</i>	2095	1	199	0.503	12	468	2.564	3.067
Correlation with distance to PAB				0.633*			0.729*	0.818**

* $P < 0.05$, ** $P < 0.01$

¹⁾ The number of the A↔T and G↔C substitutions between human and orangutan PAR genes.

²⁾ The number of (A or T) positions in the dataset. Includes the number of constant positions with A or T and the positions with A↔T substitutions.

³⁾ The number of (G or C) positions in the dataset. Includes the number of constant positions with G or C and the positions with G↔C substitutions.

⁴⁾ The number of A↔T substitutions per 100 (A or T) sites.

⁵⁾ The number of G↔C substitutions per 100 (G or C) sites.

⁶⁾ The sum of the 4) and 5) values.

