Strong and Weak Male Mutation Bias at Different Sites in the Primate Genomes: Insights from the Human-Chimpanzee Comparison

James Taylor,† Svitlana Tyekucheva,†‡ Michael Zody,§ Francesca Chiaromonte,†‡|| and Kateryna D. Makova†¶*

*Department of Computer Science and Engineering, Penn State University; †The Center for Comparative Genomics and Bioinformatics, Penn State University; ‡Department of Statistics, Penn State University; §Computational Biology and Bioinformatics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts; ||Department of Health Evaluation Sciences, Penn State University; and ¶Department of Biology, Penn State University

Male mutation bias is a higher mutation rate in males than in females thought to result from the greater number of germ line cell divisions in males. If errors in DNA replication cause most mutations, then the magnitude of male mutation bias, measured as the male-to-female mutation rate ratio (α), should reflect the relative excess of male versus female germ line cell divisions. Evolutionary rates averaged among all sites in a sequence and compared between mammalian sex chromosomes were shown to be indeed higher in males than in females. However, it is presently unknown whether individual classes of substitutions exhibit such bias. To address this issue, we investigated male mutation bias separately at non-CpG and CpG sites using human-chimpanzee whole-genome alignments. We observed strong male mutation bias at non-CpG sites: α in the X-autosome comparison was ~6–7, which was similar to the male-to-female ratio in the number of germ line cell divisions. In contrast, mutations at CpG sites exhibited weak male mutation bias: α in the X-autosome comparison was only ~2–3. This is consistent with the methylation-induced and replication-independent mechanism of CpG transitions, which constitute the majority of mutations at CpG sites. Interestingly, our study also indicated weak male mutation bias for transversions at CpG sites, implying a spontaneous mechanism largely not associated with replication. Male mutation bias was equally strong at CpG and non-CpG sites located within unmethylated "CpG islands," suggesting the replication-dependent origin of these mutations. Thus, we found that the strength of male mutation bias is nonuniform in the primate genomes. Importantly, we discovered that male mutation bias depends on the proportion of CpG sites in the loci compared. This might explain the differences in the magnitude of primate male mutation bias observed among studies.

Introduction

Mutations are the ultimate source of genetic variation in natural populations and the cause of many human genetic diseases; however, spontaneous mutations are very difficult to study directly because of their rarity. To address the question of whether mutations result primarily from errors in DNA replication, one can compare mutation rates and the numbers of germ line cell divisions for males and females. In mammals, the male germ line undergoes more cell divisions (and more DNA replications) than the female germ line (Vogel and Motulsky 1997), giving replication errors more opportunity to accumulate in males. If mutations originate mostly during DNA replication, then the male-to-female ratio of mutation rates (α) should be similar to the male-to-female ratio of germ line cell divisions (c). If α is different from c, then the contribution of replication-independent factors is not negligible. The Y chromosome is transmitted only through the male germ line, the X chromosome is transmitted more often through the female germ line, and autosomes are transmitted equally in the male and female germ lines. Thus, comparisons of mutation rates among X, Y, and autosomes can be used to estimate α (Miyata et al. 1987) and to speculate about the mechanism (replication dependent vs. replication independent) of a particular type of mutation.

Male bias for nucleotide substitutions has been observed in a variety of mammals (Li, Yi, and Makova 2002). The magnitude of such bias in primates has been the subject of a recent controversy. In primates, males undergo approximately five to six times more germ line cell divisions than females (Hurst and Ellegren 1998). Comparisons of nucleotide substitution rates between primate genes homologous between chromosomes Y and X resulted in $\alpha \approx 5$ (Shimmin, Chang, and Li 1993; Huang et al. 1997), consistent with an important role for replication errors in the generation of substitution mutations. However, two other studies (Bohossian, Skaletsky, and Page 2000; Lander et al. 2001) claimed that α was only ~2 in humans and thus suggested a significant input of replication-independent factors. Recently, two additional studies observed a stronger male mutation bias in primates with α equal to ~3 (Ebersberger et al. 2002) and ~5 (Makova and Li 2002).

While the existence of male bias for substitution (point) mutations in mammals is well accepted, not "all" substitution mutations are expected to exhibit this bias. In particular, transitions at CpG dinucleotides occur predominantly due to spontaneous deamination of methylated cytosines (Ehrlich and Wang 1981), a replication-independent process. Thus, it has been suggested that CpG transition rate should scale with time and not with the number of cell divisions (Vogel and Motulsky 1997). Investigations of male mutation bias at CpG dinucleotides have led to contradictory results. Two studies—one in primates (Nachman and Crowell 2000) and the other in rodents (Smith and Hurst 1999)—observed similar rates of transitions at CpG dinucleotides between autosomes and chromosome X. However, a third study (Anagnostopoulos et al. 1999) observed a higher rate of CpG transitions on chromosome Y than on chromosome X in primates. All three studies analyzed a relatively small number of sites.

The draft sequence of the chimpanzee genome (Mikkelsen et al. 2005) and human-chimpanzee genomic

Key words: male mutation bias, male-driven evolution, genome evolution.

E-mail: kdm16@psu.edu.

Mol. Biol. Evol. 23(3):565–573. 2006
doi:10.1093/molbev/msj060
Advance Access publication November 9, 2005

alignments provide a novel opportunity to investigate male mutation bias in primates. Chimpanzee is the closest human relative, and the two species have similar generation times (Ruvolo 1997) and physiologies. Thus, humans and chimpanzees are expected to undergo about the same number of germ line cell divisions. The small evolutionary distance also reduces the chance of multiple substitutions occurring at the same site, requiring minimal, if any, correction (Ebersberger et al. 2002). A genome-wide study allows substitution rates to be estimated for a large number of loci, counterbalancing the effects of regional (intrachromosomal) variation (Waterston et al. 2002; Hardison et al. 2003). Additionally, the determinants of such variation (e.g., GC content and recombination rate) can be assumed to be equivalent in closely related species (although see Ptak et al. 2004).

Male mutation bias estimated for closely related species should be corrected for diversity in the ancestral population (Li, Yi, and Makova 2002; Makova and Li 2002). Indeed, for closely related species, the observed divergence is equal to the sum of divergence since speciation and the diversity present in the ancestral population at speciation (Li 1977). Diversity differs among autosomes, X, and Y due to the different effective population sizes of these chromosomes. Selective sweeps and background selection are additional factors that might contribute to the extremely low levels of diversity on chromosome Y (Begun and Aquadro 1992; B. Charlesworth and D. Charlesworth 2000; Berlin and Ellegren 2004). Because interchromosomal differences in diversity caused by these factors might confound estimates of substitution rate, it is important to subtract diversity at the point of speciation (henceforth referred to as "ancient diversity") from the observed divergence before calculating the magnitude of male mutation bias from closely related species. This approach has been utilized in a number of recent studies (Makova and Li 2002; Bartosch-Harlid et al. 2003; Axelsson et al. 2004).

Here, we use the human-chimpanzee whole-genome alignments to address the following questions. (1) Is male mutation bias observed at CpG dinucleotides and does its magnitude differ from that at non-CpG sites? (2) Can male mutation bias be explained by intrachromosomal variation in substitution rates? (3) Does the magnitude of male mutation bias differ between transitions and transversions? (4) What is the magnitude of male mutation bias at CpG islands, which are usually unmethylated? (5) How do our estimates of male mutation bias compare with those obtained from previous studies?

Methods

Data Sets

BlastZ (Schwartz et al. 2003) alignments of the July 2003 human genome assembly (hg16) with the November 2003 ARACHNE chimpanzee draft assembly (panTro1) were obtained from the University of California Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu). Sequences that aligned between two different sex chromosomes or between a sex chromosome and an autosome were discarded. Due to the draft state of the chimpanzee assembly, we examined only high-quality sites by employing a modified version of the neighborhood quality standard (NQS; Altshuler et al. 2000) used by the Chimpanzee Genome Analysis group (Mikkelsen et al. 2005). Specifically, we excluded any alignment column satisfying any of these conditions: (1) quality score <30 in the chimpanzee sequence; (2) within five columns of another column with quality score <20 in the chimpanzee sequence; or (3) within five columns of an alignment gap. The pseudoautosomal and recently transposed regions of chromosome X and chromosome Y, the palindromic and ampliconic regions on chromosome Y (Skaletsky et al. 2003), and recent segmental duplications (Cheung et al. 2003) were excluded. Additionally, at this stage we removed CpG islands because they are usually unmethylated (Cross and Bird 1995) and thus might confound our analysis of substitutions at CpG sites. As a result, we obtained the "filtered" data set (Supplementary Table 1, Supplementary Material online).

To minimize the effects of selection, we also analyzed the "noncoding nonrepetitive (NCNR) set" (Supplementary Table 1, Supplementary Material online), a subset of the filtered set, from which we excluded interspersed repetitive elements, known genes, and 5-kb flanking regions upstream and downstream of known genes. Interspersed repetitive elements were excluded because they (particularly the young ones) undergo frequent gene conversion (Batzer and Deininger 2002) and have a heterogeneous methylation pattern (Meunier et al. 2005). The locations of known CpG islands, genes, and interspersed repetitive elements were obtained from the UCSC Genome Browser. An additional data set contained CpG islands on autosomes and chromosome X. CpG islands were not analyzed on chromosome Y due to lack of data.

Defining CpG and Non-CpG Sites

Two definitions were used to distinguish between CpG and non-CpG sites. According to the "inclusive" definition, all sites were divided into CpG sites (sites that were CG in at least one of the two species) and non-CpG sites (sites that were not CG in both species). According to the "restricted" definition, sites were considered to be CpG if they were CG in at least one of the two species, however, sites that were part of overlapping CpGs were excluded; non-CpG sites were defined as sites that were not CG in both species and were not immediately preceded or followed by either C or G in either species. Simulations have indicated that the restricted definition leads to minimal errors (potential homoplasies due to high mutation rate at CpG sites) in estimation of substitution rates (Meunier and Duret 2004). However, the use of this definition substantially decreased sample size and might have introduced bias (e.g., non-CpG sites were located in regions with lower GC content when the restricted definition was used). Thus, the results for both site definitions are reported.

Calculating Divergence and Male-to-Female Mutation Rate Ratio

Following Ebersberger et al. (2002), divergence was determined by dividing the number of sites different between humans and chimpanzees by the total number of

aligned sites. For each chromosome, this was done separately for CpG and non-CpG sites (according to the two definitions used) as well as for all sites. The divergence for autosomes (A) was calculated as the average divergence among individual autosomes weighted by their lengths (a similar approach was later used to calculate the diversity for autosomes). Each of the three ratios of these rates (X/A, Y/X, Y/A) was then used to derive α according to the formulas in Miyata et al. (1987). The 95% confidence intervals (CIs) for α were estimated using the bootstrap method. Namely, we divided the human genome into 10-kb windows (a total of 307,002), randomly selected the alignments 1,000 times with replacement, and estimated the divergence from these pseudosamples.

Correcting for Ancient Diversity

To correct for ancient diversity, two times and four times the human diversity (π) was subtracted from the value of observed divergence for X, Y, and autosomes (Ebersberger et al. 2002). Human diversity was estimated by sequencing eight unrelated African-American individuals with an average coverage of 1× (sequencing was done at the Baylor College of Medicine and the Broad Institute (The International HapMap Consortium 2003). Assuming a Poisson distribution, most sites were covered by one sequence read (the reads were clonal and had no heterozygotes), and all the reads were likely to come from independent samples. We counted a base as covered at depth N, if N different reads aligned to that base and the alignment met the NQS at that position. Total coverage for a chromosome was calculated as the sum of the coverage of all the bases. Diversity was calculated as the total number of single-base differences between any reference base and any read (because all reads are expected to be independent and randomly distributed) divided by the total number of bases analyzed. For each individual chromosome, diversity was calculated separately for CpG and non-CpG sites and for CpG sites in CpG islands. Because of differences in methylation, divergence at CpG sites located in CpG islands was corrected using diversity calculated specifically for these sites.

The correction for ancient diversity led to a decrease in $\alpha_{X/A}$ but to an increase in $\alpha_{Y/X}$ and $\alpha_{Y/A}$. The divergence on autosomes was higher than that on chromosome X (Supplementary Table 2, Supplementary Material online). As the diversity on autosomes was also higher than that on chromosome X (Supplementary Table 3, Supplementary Material online), correcting for ancient diversity decreased the autosomal divergence to a greater extent than the chromosome X divergence, thus bringing the two divergence values closer together and leading to lower $\alpha_{X/A}$. Similarly, in the Y/X and Y/A comparisons, correcting for ancient diversity led to lower divergence on autosomes and on chromosome X but to virtually no change on chromosome Y. As a result, $\alpha_{Y/X}$ and $\alpha_{Y/A}$ increased.

Correcting for Regional Variation in GC Content and Recombination Rates

For this part of the analysis, the human-chimpanzee divergence and human diversity were calculated in 3-Mb windows, for which sex-averaged recombination rates were available (Kong et al. 2002). Recombination rates (in cM/Mb) were calculated as the slope of the regression of genetic and physical distances of markers within each 3-Mb window separately. To obtain divergence values corrected for the effect of ancient diversity, two times the human diversity was subtracted from the human-chimpanzee divergence.

To assess the significance of differences in human-chimpanzee divergence between autosomes and chromosome X when correcting simultaneously for GC content and recombination rate, we employed a regression approach. Restricting attention to windows for which both GC content and recombination rate were available (a total of 36 windows on X and 840 windows on autosomes), we regressed divergence on GC content (linear and quadratic terms) and recombination rate (linear) and took residuals. To compare residual divergences between autosome and chromosome X windows, we calculated a two-sample *t*-statistic (Ott 1993). The null distribution of the *t*-statistic was simulated by random permutations. Namely, we randomly permuted the labels that identified windows as belonging to autosomes or X 1,000 times and recomputed the *t*-statistic for each permutation. Thus, we obtained a null distribution and a right-tail *P* value (the probability of a *t*-statistic equal to or larger than the observed one if indeed the human-chimpanzee divergence differed between autosomes and X only by chance). All permutation tests were implemented using the R statistical package (http://www.r-project.org/).

Dependence of α on the Proportion of CpG Sites

We modeled $\alpha_{X/A}$ for all sites as a mathematical function of the proportion of CpG sites on chromosome X and autosomes (p_X and p_A, respectively). Namely, according to Miyata et al. (1987)

$$\alpha_{X/A} = \frac{3\frac{X}{A} - 4}{2 - 3\frac{X}{A}},$$

where

$$\frac{X}{A} = \frac{0.141p_X + 0.007(1 - p_X)}{0.168p_A + 0.009(1 - p_A)}.$$

In the latter formula, the divergence at chromosome X (X) is equal to the average of divergences at CpG sites and non-CpG sites (0.141 and 0.007, respectively, Supplementary Table 2, Supplementary Material online), weighted by the proportion of CpG sites and proportion of non-CpG sites (p_X and $1 - p_X$, respectively). The divergence at autosomes (A) was calculated similarly. The divergence at CpGs and non-CpGs (according to the inclusive definition) was taken as fixed values (the values estimated from our data in the filtered data set, Supplementary Table 2, Supplementary Material online), while p_X and p_A were allowed to vary from 0 to 1. The function represented a complex surface, and the relationship between $\alpha_{X/A}$ for all sites and p_X and p_A was nonlinear. In figure 1, we show part of this surface and its two slices, one obtained fixing $p_X = 0.0168$ (observed proportion of CpG sites at X) and the other obtained fixing $p_A = 0.0199$ (observed proportion of CpG sites at autosomes).

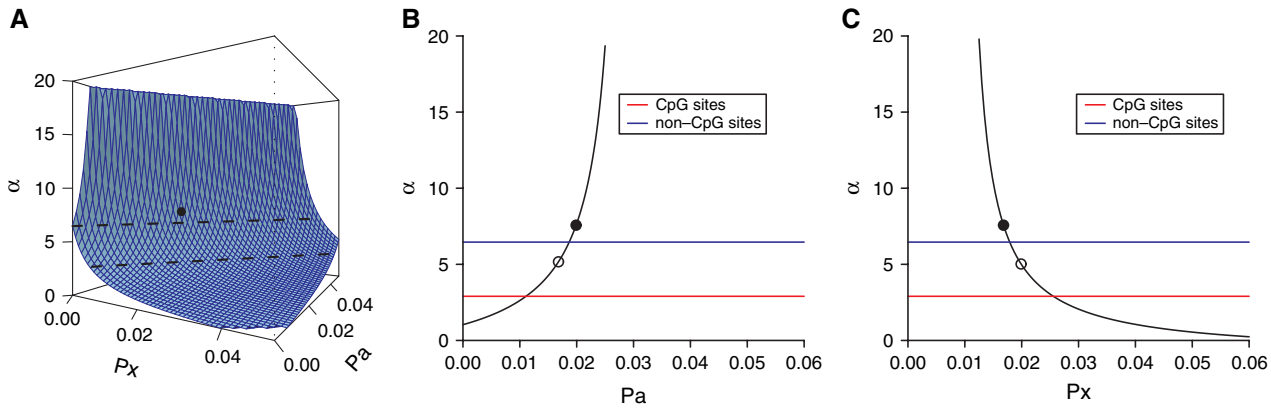

FIG. 1.—The dependence of $\alpha_{X/A}$ for all sites on the proportion of CpG sites in chromosome X (p_X) and autosomes (p_A) in the filtered data set. The $\alpha_{X/A}$ values computed for non-CpG sites and CpG sites are shown as horizontal lines. (*A*) Part of the surface $\alpha = f(p_X, p_A)$. The upper dashed line is $\alpha_{X/A}$ for non-CpG sites and the lower dashed line is $\alpha_{X/A}$ for CpG sites. The observed $\alpha_{X/A}$ for all sites is indicated by the filled circle. (*B*) The slice of the surface $\alpha = f(p_X, p_A)$ with p_X fixed at 0.0168 (the observed proportion of CpG sites in X). The filled circle corresponds to $p_A = 0.0199$, which is the observed proportion of CpG sites at autosomes. The open circle corresponds to $p_X = p_A = 0.0168$. (*C*) The slice of the surface $\alpha = f(p_X, p_A)$ with p_A fixed at 0.0199 (the observed proportion of CpG sites in autosomes). The filled circle corresponds to $p_X = 0.0168$, which is the observed proportion of CpG sites in X. The open circle corresponds to $p_A = p_X = 0.0199$.

Results and Discussion

Inference of Male Mutation Bias from Human-Chimpanzee Genomic Alignments

Three data sets were used in this study (see *Methods*; Supplementary Table 1, Supplementary Material online). Divergence in the filtered data set is expected to reflect the genome-wide average, while that in the NCNR data set should represent the neutral rate at nonrepetitive sites. The third data set consisted of CpG islands. We calculated divergence at CpG and non-CpG sites under inclusive and restricted definitions (see *Methods*) as well as for all sites (CpG and non-CpG sites taken together). This was done separately for the two sex chromosomes and autosomes (Supplementary Table 2, Supplementary Material online) and used to calculate the magnitude of male mutation bias.

Male mutation bias was estimated from the observed divergence and from that corrected by ancient diversity. To apply this correction, we assumed that diversity in the population of the human and chimpanzee common ancestor was two to four times higher than it is in contemporary humans (Supplementary Table 3, Supplementary Material online). Larger effective population size in the common ancestor of humans and chimpanzees than in contemporary humans is supported by population genetic models (Chen and Li 2001; Wall 2003) and by recent molecular data indicating high nucleotide diversity in contemporary chimpanzee populations (Yu et al. 2003; Fischer et al. 2004).

Our analysis focused on comparing human-chimpanzee divergence between chromosome X and autosomes. When chromosome Y was compared with chromosome X or autosomes, the results were less reliable because the divergence and diversity estimates for chromosome Y (particularly for CpG sites on chromosome Y) were obtained from a substantially smaller number of sites than for the other chromosomes (Supplementary Tables 1 and 3, Supplementary Material online).

Male Mutation Bias at Non-CpG Versus CpG Sites

From the X-autosome comparison, we observed strong male bias at non-CpG sites, which implies that mutations at such sites are caused primarily by errors in DNA replication (table 1). The resulting values of $\alpha_{non\text{-}CpG}$ (~6–7) were close to the sex ratio in the number of germ

Table 1
Comparison of Male Mutation Bias (from the X-autosome comparison) Between CpG and Non-CpG Sites Located Outside CpG Islands

		Inclusive site definition			Restricted site definition		
Data Set	Correction for Ancient Diversity	α_{CpG} (95% CI[a])	$\alpha_{non\text{-}CpG}$ (95% CI)	$\alpha_{CpG}/\alpha_{non\text{-}CpG}$ (95% CI)	α_{CpG} (95% CI)	$\alpha_{non\text{-}CpG}$ (95% CI)	$\alpha_{CpG}/\alpha_{non\text{-}CpG}$ (95% CI)
Filtered	None	2.90 (2.76–3.05)	6.46 (5.98–6.99)	2.22 (2.09–2.38)	3.55 (3.35–3.76)	6.13 (5.68–6.58)	1.72 (1.60–1.87)
	2π[b]	2.60 (2.45–2.75)	6.17 (5.65–6.75)	2.37 (2.21–2.56)	3.35 (3.13–3.58)	5.86 (5.39–6.35)	1.75 (1.61–1.91)
	4π[b]	2.20 (2.05–2.36)	5.79 (5.23–6.44)	2.63 (2.42–2.89)	3.09 (2.85–3.34)	5.53 (5.03–6.06)	1.79 (1.62–1.98)
NCNR[c]	None	2.81 (2.60–3.03)	7.90 (7.02–8.84)	2.82 (2.52–3.15)	3.54 (3.21–3.92)	7.22 (6.41–8.07)	2.04 (1.79–2.32)
	2π	2.52 (2.31–2.75)	7.78 (6.78–8.89)	3.09 (2.71–3.52)	3.36 (3.00–3.77)	7.04 (6.16–8.00)	2.10 (1.81–2.43)
	4π	2.15 (1.94–2.40)	7.62 (6.47–8.95)	3.54 (3.03–4.14)	3.13 (2.75–3.58)	6.82 (5.85–7.89)	2.18 (1.84–2.59)

[a] 95% CI.
[b] Assuming the population size of the chimpanzee-human common ancestor to be twice and four times as high as it is in contemporary humans.
[c] NCNR data set.

Table 2
Male Mutation Bias (from the X-autosome comparison) for Transitions and Transversions in the NCNR Data Set

	Inclusive Site Definition		Restricted Site Definition	
Mutation Type	α_{CpG} (95% CI)	$\alpha_{non\text{-}CpG}$ (95% CI)	α_{CpG} (95% CI)	$\alpha_{non\text{-}CpG}$ (95% CI)
Transitions	2.84 (2.60–3.08)	7.93 (7.06–8.85)	3.54 (3.21–3.94)	7.12 (6.33–8.05)
Transversions	2.67 (2.30–3.09)	7.84 (6.70–9.15)	3.57 (3.03–4.36)	7.39 (6.35–8.78)

NOTE.—Correction for ancient diversity was not applied.

line cell divisions in primates ($c \approx 6$), suggesting that most mutations at non-CpG sites occur due to errors in DNA replication. When diversity in the human-chimpanzee common ancestor was assumed to be three to four times greater than that in contemporary humans, we obtained overlapping or nearly overlapping 95% CIs for $\alpha_{non\text{-}CpG}$ calculated from comparisons between X and autosomes, X and Y, and Y and autosomes (table 1, Supplementary Table 4 [Supplementary Material online], and data not shown). This suggests that the distinct numbers of cell divisions for male and female germ lines are the major factors determining the differences in the non-CpG substitution rates among X, Y, and autosomes.

Interestingly, in the X-autosome comparison male mutation bias was approximately two- to threefold lower at CpG sites than at non-CpG sites (table 1). For both data sets, both site definitions, and independent of correction for ancient diversity, the difference between α_{CpG} and $\alpha_{non\text{-}CpG}$ was significant as assessed by the bootstrap method. The α_{CpG} values from the X-Y and Y-autosome comparisons showed substantial variation and wide CIs (Supplementary Table 5, Supplementary Material online) and need to be reevaluated when a high-quality sequence of chimpanzee chromosome Y and more accurate diversity estimates for this chromosome become available.

The low male bias observed at CpG sites from the X-autosomal comparison is consistent with the molecular mechanism of CpG transitions. Because the majority of mutations at CpG sites are transitions resulting from spontaneous deamination of cytosine (Ehrlich and Wang 1981), substitutions at these sites are expected to be less dependent on the number of DNA replications in the two germ lines as compared with substitutions at non-CpG sites. Remarkably, the weak male mutation bias at CpG sites observed here complements another recent finding that also points to the relative independence of CG→TG mutations from DNA replication. Hwang and Green (2004) showed that, compared to other substitutions, CpG transitions accumulate in a relatively clocklike fashion in mammalian evolution: they do not appear to be affected by generation time and are probably not caused by replication errors.

Because substitutions leading to loss and to formation of CpG sites (the latter is expected to be replication dependent) were considered together, we still observed a higher substitution rate at CpG sites in males than in females (i.e., $\alpha_{CpG} > 1$). An outgroup sequence (e.g., macaque) will allow us to differentiate between CG→TG and TG→CG substitutions and to investigate male mutation bias for these substitutions separately. However, even after we separate these, mutations that lead to the loss of CpG sites might still exhibit weak male mutation bias. The level of methylation is known to be lower in the mammalian germ line (Monk 2002) than in somatic cells and thus, some mutations at CpG sites in the germ line might result not from spontaneous deamination of methylated cytosine but from replication errors. Additionally, a potentially higher methylation level of sperm than of oocyte DNA (Monk 1995) could also contribute to $\alpha_{CpG} > 1$. However, the magnitude, timing, and genomic location of methylation differences between the two germ lines have been debated (Yoder, Walsh, and Bestor 1997; Bestor 1998; El-Maarri et al. 1998) and require further studies.

Male Mutation Bias and Variation in GC Content and Recombination Rate

In addition to interchromosomal variation, substitution rate has been shown to fluctuate intrachromosomally (Lercher, Williams, and Hurst 2001; Malcom, Wyckoff, and Lahn 2003) and to correlate with GC content and recombination rate (Hardison et al. 2003). Thus, one might argue that the substitution rate is lower on chromosome X than on autosomes because of differences in GC content and/or recombination rate. To explore this possibility, we evaluated the difference in divergence between autosomal and chromosome X windows accounting for GC content and recombination rate (windows in chromosome Y were not investigated because of the lack of data). Namely, we divided chromosome X and the autosomes into 3-Mb windows, regressed divergence on GC content and recombination in these windows, compared the resulting residuals between autosome and chromosome X windows, and used a random permutation test to determine the significance of the observed differences. Even after we accounted for GC content and recombination, the divergence at both non-CpG and CpG sites (for both site definitions) was still significantly lower for the X chromosome than for autosomes. This was observed for both the NCNR and filtered data sets and when divergence was corrected for ancient diversity ($P < 0.001$ in all comparisons). Thus, our data support male mutation bias despite the variation in GC content and recombination rate that has been linked to variation in substitution rate within chromosomes.

Male Mutation Bias at Transitions and Transversions

Surprisingly, when considered separately, both transitions and transversions at CpG sites displayed weak male mutation bias (table 2). For instance, for CpG sites (under inclusive definition) in the NCNR data set, α calculated for

Table 3
Comparison of Male Mutation Bias (from the X-autosome comparison) Between CpG and Non-CpG Sites Located Within CpG Islands

Correction for Ancient Diversity	Inclusive Site Definition		Restricted Site Definition	
	α_{CpG} (95% CI)	$\alpha_{non\text{-}CpG}$ (95% CI)	α_{CpG} (95% CI)	$\alpha_{non\text{-}CpG}$ (95% CI)
No	6.91 (2.66–65.3)	5.96 (2.45–83.9)	8.93 (2.91–∞)	7.72 (2.11–∞)
2π	4.56 (1.82–20.0)	5.57 (2.02–∞)	5.57 (1.96–∞)	7.58 (1.44–∞)
4π	2.87 (1.10–9.56)	5.02 (1.53–∞)	3.36 (1.16–61.7)	7.38 (0.89–∞)

transitions and for transversions did not differ significantly (2.84 [95% CI: 2.60–3.08] vs. 2.67 [95% CI: 2.30–3.09]). Additionally, similarly to other studies (Ebersberger et al. 2002; Siepel and Haussler 2004), we found transversion rates to be ~10 times higher at CpG sites than at non-CpG sites (0.0347 vs. 0.0035 and 0.0446 vs. 0.0041 for the inclusive and restricted site definitions, respectively, for autosomes in the NCNR data set).

Weaker male bias and higher rates for transversions at CpG sites than at non-CpG sites imply differences in the molecular mechanisms leading to transversions in these two site categories. Interestingly, this is consistent with the suggestion by Blake, Hess, and Nicholson-Tuell (1992) that transversions at CpG sites might result from spontaneous alkylation of guanine at the O-6 position (Fix, Koehler, and Glickman 1990). Such mutations, similar to the spontaneous deamination of methylated cytosine, are expected to be largely independent of DNA replication errors. Both transitions and transversions at non-CpG sites displayed strong male mutation bias (table 2), suggesting that these mutations result from errors in DNA replication.

Male Mutation Bias in CpG Islands

Male mutation bias was equally high for both CpG and non-CpG sites located within CpG islands (table 3), in contrast to what was observed outside CpG islands. In fact, α was approximately two times higher for CpG sites located within versus outside CpG islands (in the NCNR data set) when no or 2π correction for ancient diversity was applied (tables 1 and 3). This difference was not statistically significant due to the wide CIs of the α values for CpG islands (a consequence of the relatively small number of sites analyzed). When 4π correction for ancient diversity was applied, α_{CpG} at CpG islands was similar to that outside CpG islands. This can be explained by potentially inaccurate diversity estimates for CpG sites located within CpG islands as such estimates were based on a small number of sites (Supplementary Table 3, Supplementary Material online).

High male mutation bias at CpG sites located within CpG islands is consistent with the low degree of methylation and, as a result, the more replication-dependent origin of mutations at such sites. Unlike in the rest of the genome, in CpG islands CpG sites tend to be unmethylated (Cross and Bird 1995). Thus, the main source of mutations at such sites is replication errors and not spontaneous deamination of methylated cytosines. We cannot, however, exclude the possibility that CpG islands investigated in the present study were affected by selection.

Male Mutation Bias for All Sites Depends on the Proportion of CpG Sites

To directly compare our results with those of other studies, we investigated male mutation bias considering all (both CpG and non-CpG) sites (table 4). The resulting α from the X-autosome comparison (which is the most reliable) in the filtered data set was 7.58, 6.92, and 6.11 when no, 2π, and 4π correction for ancient diversity were applied, respectively. Thus, the magnitude of male mutation bias calculated here was not significantly different from most previous estimates (Shimmin, Chang, and Li 1993; Chang, Hewett-Emmett, and Li 1996; Huang et al. 1997; Makova and Li 2002), but it was significantly higher than the estimates calculated by Bohossian, Skaletsky, and Page (2000) and Lander et al. (2001). The potential causes of a low male bias in the latter two studies have been discussed elsewhere (Li, Yi, and Makova 2002). The α value calculated here from the X-autosome comparison, uncorrected for ancient diversity, was not significantly different from that reported by Ebersberger and colleagues (2002), who investigated a smaller sample of human-chimpanzee alignments. However, when correction for ancient diversity was applied (4π), $\alpha_{X/A}$ reported here was significantly higher than that reported in the study of Ebersberger et al.(2002). This can be explained by different diversity values used for correction. The α values calculated in our study from the Y-X and Y-autosome comparisons were significantly higher than those obtained by Ebersberger et al. (2002).

We found that the male mutation bias calculated for all sites together critically depends on the proportions of CpG sites in the chromosomes compared. Our original observation was that α for all sites was sometimes higher than both $\alpha_{non\text{-}CpG}$ and α_{CpG} (tables 1 and 4). For instance, without correction for ancient diversity, α calculated for all sites in the filtered data set using X-autosome comparison was 7.58, while $\alpha_{non\text{-}CpG}$ and α_{CpG} were only 6.46 and 2.90, respectively (under inclusive definition). This seems counterintuitive as one might expect α for all sites to have a value intermediate between $\alpha_{non\text{-}CpG}$ and α_{CpG}. However, the observation is easily explained by modeling how α for all sites depends on the proportion of CpG sites on different chromosomes (see *Methods*; fig. 1). For instance, if we assume the proportion of CpG sites to be equal (or nearly equal) between X and autosomes, α for all sites does indeed have an intermediate value between $\alpha_{non\text{-}CpG}$ and α_{CpG}. On the other hand, if we assume that the proportion of CpG sites differs between the two types of chromosomes, α for all sites can be either higher or lower than both $\alpha_{non\text{-}CpG}$ and α_{CpG}. In our data, the proportion of CpG sites in chromosome X was substantially lower than in autosomes

Table 4
Primate Male Mutation Bias for All Sites

Taxa	Chromosome	Length Analyzed	Correction for Ancient Diversity	α (95% CI)	Reference
Humans and chimpanzees	X/A	A total of 2.0 Gb on three types of chromosomes (Supplementary Table 1, Supplementary Material online)	None	7.58 (7.04–8.20)	Filtered data set from this study
			2×	6.92 (6.38–7.58)	
			4×	6.11 (5.58–6.78)	
	Y/X		None	2.68 (2.50–2.89)	
			2×	4.03 (3.67–4.43)	
			4×	8.24 (7.01–9.84)	
	Y/A		None	1.77 (1.64–1.94)	
			2×	3.04 (2.70–3.46)	
			4×	11.2 (8.04–17.6)	
Humans and chimpanzees	X/A	A total of 1.9 Mb on three types of chromosomes	None	5.4 (3.7–8.6)	Ebersberger et al. (2002)
			1×	4.8 (3.3–7.6)	
			4×	3.2 (2.2–4.9)	
	Y/X		None	1.9 (1.7–2.2)	
			1×	2.1 (1.8–2.4)	
			4×	2.8 (2.3–3.4)	
	Y/A		None	1.3 (1.2–1.5)	
			1×	1.5 (1.3–1.7)	
			4×	2.6 (2.2–3.2)	
Humans	Y/X	1.6 Mb on X and 0.4 Mb on Y	None	2.1	Lander et al. (2001)
Humans and chimpanzees	Y/X	39 kb	None	1.70 (1.15–2.87)	Bohossian, Skaletsky, and Page (2002)
Higher primates	Y/A	10 kb	None	5.25 (2.44–∞)	Makova and Li (2002)
Higher primates	Y/X	1.1 kb	None	5.14 (2.42–16.6)	Huang et al. (1997)
Higher primates	Y/X	1.4 kb	None	6.26 (2.63–32.4)	Chang, Hewett-Emmett, and Li (1996)
Higher primates	Y/X	0.9 kb	None	4.20 (2.20–10.0)	Shimmin, Chang, and Li (1993)

(1.68% vs. 1.99% for the filtered data set), resulting in higher α for all sites than for non-CpG or CpG sites. This suggests that for all sites, only α values estimated from loci with similar proportions of CpG sites can be directly compared. Alternatively, the male mutation bias can be compared separately for CpG sites and for non-CpG sites even between loci differing in CpG content.

Conclusions

For nucleotide substitutions at non-CpG sites and CpG sites within unmethylated CpG islands, as well as for small insertions and deletions (Makova, Yang, and Chiaromonte 2004), the magnitude of male mutation bias appears to be similar to the male-to-female ratio in the number of germ line cell divisions. This suggests that these mutations result primarily from errors in DNA replication. However, substitutions at CpG dinucleotides have a lower male mutation bias. This is consistent with spontaneous, replication-independent deamination of methylated cytosines being the major molecular mechanism for most CpG mutations. Our study indicates that male mutation bias is nonuniform in the primate genomes and depends on the type of mutation under investigation. Although here the mechanism of CpG mutations was known, this study illustrates how male mutation bias phenomenon can be used to infer the mechanism of mutations for which such knowledge does not exist. Different proportions of CpG sites at different loci might explain substantial variation in the estimates of primate male mutation bias among studies.

Here, male mutation bias was investigated by comparing humans with our closest relatives, chimpanzees. Studying such closely related species required correcting the observed divergence for ancient diversity and making assumptions about the population size in the common ancestor. With the sequencing of the macaque genome approaching completion, we look forward to comparing our estimates of male mutation bias based on human-chimpanzee alignments to the corresponding estimates from human-macaque alignments. Additionally, using human-chimpanzee-macaque three-way alignments will allow to polarize CpG→TpG versus TpG→CpG mutations and employ neighbor-dependent substitution models (Hwang and Green 2004; Siepel and Haussler 2004).

One of the limitations of the present study is the low number of high-quality sites currently available for the chimpanzee Y chromosome. A targeted sequencing of the chimpanzee Y chromosome will aid in obtaining more reliable estimates of male mutation bias for the Y-X and Y-autosome comparisons. Even though similar α values were obtained for non-CpG sites from the three comparisons (X/A, X/Y, and Y/A), we cannot completely eliminate the possibility that replication-independent factors might act differently on different chromosomes and contribute

substantially to the evolutionary rates for X, Y, and autosomes. One of these factors could be selection for a lower mutation rate on the X chromosome due to a hemizygote state of deleterious recessive mutations in males, as suggested by McVean and Hurst (McVean and Hurst 1997), although this was later disputed (Ellegren and Fridolfsson 1997; Makova and Li 2002). Additionally, if recombination is indeed mutagenic (Lercher and Hurst 2002), this would lead to a lower mutation rate on chromosome X than on autosomes because chromosome X has a lower recombination rate than autosomes do (Kong et al. 2002). However, according to our results this is unlikely: we still observed higher divergence in autosomes than in chromosome X after accounting for variation in recombination rate and GC content, which supports the male-driven evolution hypothesis.

Supplementary Material

Supplementary Tables 1–5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are thankful to Ines Hellmann for providing data on 3-Mb windows, to Webb Miller for his advice about genomic alignments, to Svante Paabo for his suggestion to study CpG islands, to Tarjei Mikkelsen for his help at the earlier stages of this project, to Steve Schaeffer, Ines Hellmann, Cathy Riemer, Paula Goetting-Minesky, and Erika Kvikstad for critical reading of the manuscript. We thank the Chimpanzee Sequencing and Analysis Consortium for including us in the analysis team and all the groups in the consortium who produced data and made it publicly available. Further, we want to thank the members of the Baylor College of Medicine Human Genome Sequencing Center and the Broad Institute of MIT and Harvard for generation and early release of the DNA sequence data used to derive the estimates of human diversity in this study. This study was supported by National Institutes of Health grant R01-GM072264 and start-up funds from the Penn State Eberly College of Science to K.D.M.

Literature Cited

Altshuler, D., V. J. Pollara, C. R. Cowles, W. J. Van Etten, J. Baldwin, L. Linton, and E. S. Lander. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature **407**:513–516.

Anagnostopoulos, T., P. M. Green, G. Rowley, C. M. Lewis, and F. Giannelli. 1999. DNA variation in a 5-Mb region of the X chromosome and estimates of sex-specific/type-specific mutation rates. Am. J. Hum. Genet. **64**:508–517.

Axelsson, E., N. G. Smith, H. Sundstrom, S. Berlin, and H. Ellegren. 2004. Male-biased mutation rate and divergence in autosomal, z-linked and w-linked introns of chicken and Turkey. Mol. Biol. Evol. **21**:1538–1547.

Bartosch-Harlid, A., S. Berlin, N. G. Smith, A. P. Moller, and H. Ellegren. 2003. Life history and the male mutation bias. Evolution Int. J. Org. Evolution **57**:2398–2406.

Batzer, M. A., and P. L. Deininger. 2002. Alu repeats and human genomic diversity. Nat. Rev. Genet. **3**:370–379.

Begun, D. J., and C. F. Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature **356**:519–520.

Berlin, S., and H. Ellegren. 2004. Chicken W: a genetically uniform chromosome in a highly variable genome. Proc. Natl. Acad. Sci. USA **101**:15967–15969.

Bestor, T. H. 1998. Cytosine methylation and the unequal developmental potentials of the oocyte and sperm genomes. Am. J. Hum. Genet. **62**:1269–1273.

Blake, R. D., S. T. Hess, and J. Nicholson-Tuell. 1992. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. J. Mol. Evol. **34**:189–200.

Bohossian, H. B., H. Skaletsky, and D. C. Page. 2000. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. Nature **406**:622–625.

Chang, B. H., D. Hewett-Emmett, and W.-H. Li 1996. Male-to-female ratios of mutation rate in higher primates estimated from intron sequences. Zool. Stud. **35**:36–48.

Charlesworth, B., and D. Charlesworth. 2000. The degeneration of Y chromosomes. Phil. Trans. R. Soc. Lond. B **355**: 1563–1572.

Chen, F. C., and W. H. Li. 2001. Genomic divergences 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.

Cheung, J., X. Estivill, R. Khaja, J. R. MacDonald, K. Lau, L. C. Tsui, and S. W. Scherer. 2003. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol. **4**:R25.

Cross, S. H., and A. P. Bird. 1995. CpG islands and genes. Curr. Opin. Genet. Dev. **5**:309–314.

Ebersberger, I., D. Metzler, C. Schwarz, and S. Paabo. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. Am. J. Hum. Genet. **70**:1490–1497.

Ehrlich, M., and R. Y. Wang. 1981. 5-Methylcytosine in eukaryotic DNA. Science **212**:1350–1357.

El-Maarri, O., A. Olek, B. Balaban, M. Montag, H. van der Ven, B. Urman, K. Olek, S. H. Caglayan, J. Walter, and J. Oldenburg. 1998. Methylation levels at selected CpG sites in the factor VIII and FGFR3 genes, in mature female and male germ cells: implications for male-driven evolution. Am. J. Hum. Genet. **63**:1001–1008.

Ellegren, H., and A. K. Fridolfsson. 1997. Male-driven evolution of DNA sequences in birds. Nat. Genet. **17**:182–184.

Fischer, A., V. Wiebe, S. Paabo, and M. Przeworski. 2004. Evidence for a complex demographic history of chimpanzees. Mol. Biol. Evol. **21**:799–808.

Fix, D. F., D. R. Koehler, and B. W. Glickman. 1990. Uracil-DNA glycosylase activity affects the mutagenicity of ethyl methanesulfonate: evidence for an alternative pathway of alkylation mutagenesis. Mutat. Res. **244**:115–121.

Hardison, R. C., K. M. Roskin, S. Yang et al. (16 co-authors). 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res. **13**:13–26.

Huang, W., B. H. Chang, X. Gu, D. Hewett-Emmett, and W. Li. 1997. Sex differences in mutation rate in higher primates estimated from AMG intron sequences. J. Mol. Evol. **44**:463–465.

Hurst, L. D., and H. Ellegren. 1998. Sex biases in the mutation rate. Trends Genet. **14**:446–452.

Hwang, D. G., and P. Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. USA **101**:13994–14001.

Kong, A., D. F. Gudbjartsson, J. Sainz et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. Nat. Genet. **31**:241–247.

Lander, E. S., L. M. Linton, B. Birren et al. (249 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.
Lercher, M. J., and L. D. Hurst. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. **18**:337–340.
Lercher, M. J., E. J. Williams, and L. D. Hurst. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. Mol. Biol. Evol. **18**:2032–2039.
Li, W. H. 1977. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. Genetics **85**:331–337.
Li, W. H., S. Yi, and K. Makova. 2002. Male-driven evolution. Curr. Opin. Genet. Dev. **12**:650–656.
Makova, K. D., and W. H. Li. 2002. Strong male-driven evolution of DNA sequences in humans and apes. Nature **416**:624–626.
Makova, K. D., S. Yang, and F. Chiaromonte. 2004. Insertions and deletions are male biased too: a whole-genome analysis in rodents. Genome Res. **14**:567–573.
Malcom, C. M., G. J. Wyckoff, and B. T. Lahn. 2003. Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. Mol. Biol. Evol. **20**:1633–1641.
McVean, G. T., and L. D. Hurst. 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. Nature **386**:388–392.
Meunier, J., and L. Duret. 2004. Recombination drives the evolution of GC-content in the human genome. Mol. Biol. Evol. **21**:984–990.
Meunier, J., A. Khelifi, V. Navratil, and L. Duret. 2005. Homology-dependent methylation in primate repetitive DNA. Proc. Natl. Acad. Sci. USA **102**:5471–5476.
Mikkelsen, T. S., L. W. Hillier, E. E. Eichler et al. (67 co-authors). 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**:69–87.
Miyata, T., H. Hayashida, K. Kuma, K. Mitsuyasu, and T. Yasunaga. 1987. Male-driven molecular evolution: a model and nucleotide sequence analysis. Cold Spring Harbor Symp. Quant. Biol. **52**:863–867.
Monk, M. 1995. Epigenetic programming of differential gene expression in development and evolution. Dev. Genet. **17**:188–197.
———. 2002. Mammalian embryonic development—insights from studies on the X chromosome. Cytogenet. Genome Res. **99**:200–209.
Nachman, M. W., and S. L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics **156**:297–304.
Ott, R. L. 1993. An introduction to statistical methods and data analysis. Duxbury Press, Belmont.
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**:849–855.
Ruvolo, M. 1997. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Mol. Biol. Evol. **14**:248–265.
Schwartz, S., W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. 2003. Human-mouse alignments with BLASTZ. Genome Res. **13**:103–107.
Shimmin, L. C., B. H. Chang, and W. H. Li. 1993. Male-driven evolution of DNA sequences. Nature **362**:745–747.
Siepel, A., and D. Haussler. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. **21**:468–488.
Skaletsky, H., T. Kuroda-Kawaguchi, P. J. Minx et al. (40 co-authors). 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature **423**:825–837.
Smith, N. G., and L. D. Hurst. 1999. The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate? Genetics **152**:661–673.
The International HapMap Consortium. 2003. The International HapMap Project. Nature **426**:789–796.
Vogel, F., and A. G. Motulsky. 1997. Human genetics: problems and approaches. Springer-Verlag, Berlin.
Wall, J. D. 2003. Estimating ancestral population sizes and divergence times. Genetics **163**:395–404.
Waterston, R. H., K. Lindblad-Toh, E. Birney et al. (156 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–562.
Yoder, J. A., C. P. Walsh, and T. H. Bestor. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. **13**:335–340.
Yu, N., M. I. Jensen-Seaman, L. Chemnick, J. R. Kidd, A. S. Deinard, O. Ryder, K. K. Kidd, and W. H. Li. 2003. Low nucleotide diversity in chimpanzees and bonobos. Genetics **164**:1511–1518.

Manolo Gouy, Associate Editor

Accepted November 2, 2005