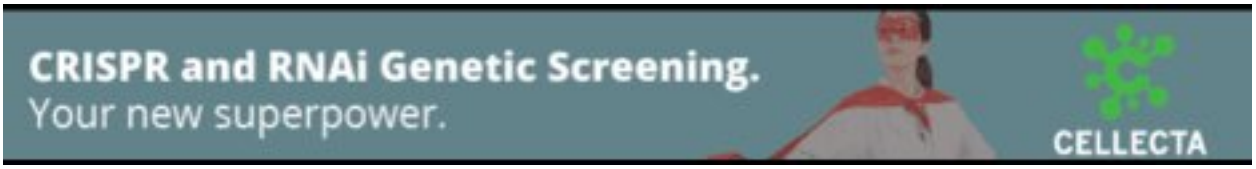
Insertions and Deletions Are Male Biased Too: A Whole-Genome Analysis in Rodents

Kateryna D. Makova, Shan Yang and Francesca Chiaromonte

Genome Res. 2004 14: 567-573
Access the most recent version at doi:10.1101/gr.1971104



Letter

Insertions and Deletions Are Male Biased Too: A Whole-Genome Analysis in Rodents

Kateryna D. Makova,[1,5,6] Shan Yang,[2,5] and Francesca Chiaromonte[3,4,5]

Departments of [1]Biology, [2]Biochemistry and Molecular Biology, [3]Statistics, [4]Health Evaluation Sciences, and [5]the Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

It is presently accepted that, in mammals, due to the greater number of cell divisions in the male germline than in the female germline, nucleotide substitutions occur more frequently in males. The data on mutation bias in insertions and deletions (indels) are contradictory, with some studies indicating no sex bias and others indicating either female or male bias. The sequenced rat and mouse genomes provide a unique opportunity to investigate a potential sex bias for different types of mutations. Indeed, mutation rates can be accurately estimated from a large number of orthologous loci in organisms similar in generation time and in the number of germline cell divisions. Here we compare the mutation rates between chromosome X and autosomes for likely neutral sites in eutherian ancestral interspersed repetitive elements present at orthologous locations in the rat and mouse genomes. We find that small indels are male biased: The male-to-female mutation rate ratio (α) for indels in rodents is ~2. Similarly, our whole-genome analysis in rodents indicates an approximately twofold excess of nucleotide substitutions originating in males over that in females. This is the same as the male-to-female ratio of the number of germline cell divisions in rat and mouse. Thus, this is consistent with nucleotide substitutions and small indels occurring primarily during DNA replication.

In mammals the male germline undergoes more cell divisions (and more DNA replications) than does the female germline. If germline mutations are replication-driven, then (1) mutations should originate more frequently in males than in females, and (2) the male-to-female mutation rate ratio (α) should be equal to the male-to-female ratio of the number of germline cell divisions (c). If α is smaller than c, then the role of replication-independent factors (e.g., environmental damaging agents such as oxygen radicals) is significant in generating mutations. Thus, knowing the precise value of α is critical for assessing whether germline mutations are caused by errors in DNA replication.

Two lines of evidence indicate that nucleotide substitutions are more frequent in males than in females. First, de novo point mutations leading to human genetic diseases are predominantly of paternal origin (for review, see Crow 2000; Li et al. 2002). Second, most molecular evolutionary studies comparing rates of nucleotide substitutions between mammalian sex chromosomes (or between one of the sex chromosomes and the autosomes) have concluded that the nucleotide substitution rate is indeed higher among males than among females (for review, see Li et al. 2002).

Despite the growing support of the male-mutation bias hypothesis in mammals, the exact value of α for nucleotide substitutions remains controversial. In particular, estimates of α within primates and within rodents, the two most studied mammalian orders, have disagreed. In humans, $c \approx 6$ if the father's age is 20 (Hurst and Ellegren 1998). However, estimates of α range from two to five for the higher primates (Shimmin et al. 1993; Chang et al. 1996; Bohossian et al. 2000; Lander et al. 2001; Ebersberger et al. 2002; Makova and Li 2002).

In rodents, c is approximately two (Chang et al. 1994). Studies comparing mutation rates of intronic sequences homologous between X and Y (a total of 2 kb) estimated α to be approximately two (Chang et al. 1994; Chang and Li 1995). However, an X-to-autosome comparison claimed an infinite excess of male over female mutations (Wolfe and Sharp 1993), whereas a Y-to-autosome comparison concluded that $\alpha = 1$ (McVean and Hurst 1997; for possible explanations of this discrepancy, see Smith and Hurst 1999). A recent X-to-autosome comparison of substitution rates at synonymous sites orthologous between mouse and rat led to $\alpha = 3.52$ (Malcom et al. 2003).

A potential sex bias in mutations other than nucleotide substitutions has not been investigated extensively. Do insertions and deletions (indels) occur predominantly in males compared with females? The published data addressing this question are contradictory. First, small indels causing some human genetic diseases originate with the same frequency in males and females (e.g., in tuberous sclerosis; Roberts et al. 2002). Other genetic diseases originate more frequently in males (e.g., in neurofibromatosis type 2 and Rett syndrome; Kluwe et al. 2000; Trappe et al. 2001), whereas still others originate predominantly in females (e.g., in hemophilia B; Sommer et al. 2001). These results indicate either no or a locus-specific sex bias in indel mutations. Notably, the conclusions are based on the analysis of a few (usually <10) observed mutations in each of these studies. Second, a recent sequence comparison between ~6 kb on X and ~5 kb on Y in primates indicated similar indel frequencies (Sundstrom et al. 2003), indicating no sex bias for indels in primates. Interestingly, the same study (Sundstrom et al. 2003) observed male bias for indels in birds.

The availability of the rat and mouse genomic sequences provides a unique opportunity to test the hypothesis of male-mutation bias and to estimate the male-to-female mutation rate ratio in rodents. First, the reproductive physiology of both rat and mouse has been well studied, and the two organisms have a similar ratio of cell divisions between male and female germlines by the time of reproduction (Chang et al. 1994). Second, mutation rates can be estimated from a large number of loci, compensating for the effects of local (within chromosome) variation in mutation rates (Waterston et al. 2002). Here we test whether there are differences in average mutation rates between a sex chromosome and autosomes despite local variation (Lercher et al. 2001; Hardison et al. 2003). Third, the evolutionary distance

[6]Corresponding author.
E-MAIL kdm16@psu.edu; FAX (814) 865-9131.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1971104.

between mouse and rat is small enough to secure sequence alignment with little ambiguity and large enough to minimize the effects of ancient genetic polymorphisms on the estimation of α (Li et al. 2002; Makova and Li 2002).

In this study we estimate the male-to-female mutation rate ratio for indels and nucleotide substitutions in rodents. As female rat and female mouse have been selected for the Rat Genome Project (Rat Genome Sequencing Project Consortium 2004) and the Mouse Genome project (Waterston et al. 2002), respectively, we conduct a large-scale comparison of mutation rates between chromosome X and autosomes. Chromosome X spends one-third of the time in the male germline and two-thirds of the time in the female germline. In contrast, autosomes spend an equal amount of time in each of the two germlines. Thus, if mutations are driven by replication, chromosome X should have a lower mutation rate than do autosomes. The mutation rates are investigated in ancient interspersed repetitive elements (the ancestral repeats [ARs]) present at orthologous locations in the rat and mouse genomes.

We focus on transposable elements fixed in the common ancestor of primates and rodents. It has been suggested that the ARs represent a suitable model of neutral evolution (Waterston et al. 2002; Hardison et al. 2003). First, ARs presumably evolve under no selection, whereas other data sources for modeling neutral evolution might be selectively constrained. For instance, a significant portion of noncoding DNA is likely to be under selection as it possesses regulatory elements (promoters, enhancers, etc.) and unidentified protein- and RNA-coding exons, or might be involved in structural functions (Dermitzakis et al. 2002; Waterston et al. 2002). Some intronic sites are under selection due to their roles in splicing and transcription regulation (see Kolb 2003). Fourfold degenerate sites are sometimes also involved in splicing, can be selected due to codon bias, and often have unequal frequencies of flanking bases (Hardison et al. 2003). Second, the detectable ARs comprise ~5% of the mouse or of the rat genome (Waterston et al. 2002; Rat Genome Sequencing Project Consortium 2004), allowing us to analyze a substantial proportion of rodent genomes. Third, the ARs can be identified and studied at orthologous locations (Waterston et al. 2002; Schwartz et al. 2003). The AR copies present at each orthologous locus inserted into the genome of the common ancestor of primates and rodents, and, most importantly, were already present in the common ancestor of mouse and rat. Thus, contrary to the studies comparing divergences of repetitive elements from the same genome (Erlandsson et al. 2000; Lander et al. 2001), in our study there is no uncertainty about the timing of repeat integration. Fourth, although the density of particular AR classes varies depending upon the GC content of the surrounding region, when all classes of the ARs are considered together, the whole spectrum of the genomic GC content is represented (Lander et al. 2001; Waterston et al. 2002). This is important because the variation in substitution rates was shown to correlate with the GC content (see Hurst and Williams 2000; Hardison et al. 2003).

RESULTS

ARs Used to Model Neutral Evolution

Below, we examine two sets of the ARs at orthologous locations in the rat and mouse genomes. The first set contains all ARs detected in the four-way (human–rat–mouse-repeat consensus) alignments (a total of 5.5 Mb on chromosome X and 100.4 Mb on autosomes). It includes 3 families of SINEs, 64 families of LINEs, 156 families of LTRs, and 100 families of DNA elements. Because of a high mutation rate in rodents, a significant portion of the ARs of more ancient origin is usually undetected in their genomes (Waterston et al. 2002). The ARs of more recent origin, on the other hand, have a higher probability of being identified by the RepeatMasker (A.F.A. Smit and P. Green; http://ftp.genome.washington.edu/RM/RepeatMasker.html) compared with the older ARs. Thus, our second set of ARs contains the ARs that were active just before the primate–rodent speciation ("the young ARs"; Table 1 from Waterston et al. 2002—seven families of LINE1, five families of LTR elements, and three families of DNA elements). The second set encompasses 0.3 Mb on chromosome X and 6.4 Mb on autosomes. Although we study the ARs inserted in the rodent–primate ancestor, only the branches from the rat–mouse common ancestor to rat and to mouse are analyzed.

Focusing only on the portion of the genome occupied by the AR set described above may introduce some biases (perhaps, this part of the rodent genome is particularly conserved as it is alignable with the human genome). Thus, we also perform our calculations on the set of all sites alignable between rat and mouse; there are a total of 1691.0 Mb on autosomes, and 78.3 Mb on X were examined.

Indels

The frequency of small indels (<50 bp) was investigated in the alignments of mouse and rat ARs present at orthologous locations. As shown in Table 1, indels in all of the ARs are less frequent on chromosome X than on autosomes ($P < 0.001$, χ^2 test). On average, between rat and mouse, there are 12.9 and 14.8 indels per 1000 nucleotides on X and on autosomes, respectively.

Table 1. Male Mutation Bias in Insertions and Deletions That Accumulated in the Ancestral Repeats in Rat and Mouse Since Their Divergence From the Common Ancestor

Indel type	N_X	$Rate_X$	N_A	$Rate_A$	X/A	α (95% CI)
Mouse deletions	22,278	0.0041	468,707	0.0047	0.871	2.260 (1.961–2.630)
Rat deletions	24,924	0.0046	542,992	0.0054	0.841	2.818 (2.440–3.291)
Total deletions	47,202	0.0086	1,011,699	0.0101	0.855	2.538 (2.200–2.957)
Mouse insertions	12,375	0.0023	251,339	0.0025	0.902	1.828 (1.561–2.157)
Rat insertions	11,001	0.0020	226,459	0.0023	0.890	1.980 (1.679–2.361)
Total insertions	23,376	0.0043	477,798	0.0048	0.897	1.898 (1.616–2.251)
Indels	70,578	0.0129	1,489,497	0.0148	0.869	2.304 (1.986–2.699)

All ARs ("younger" and "older") are considered together. N_X and N_A indicate the numbers of indels that accumulated on chromosome X and autosomes, respectively; $rate_X$ and $rate_A$, the indel rates (a number of indels per base pair of alignment) on chromosome X and autosomes, respectively; X/A, the ratio of indel rates on chromosome X versus autosomes; and α (95% CI), the male-to female mutation rate ratio and its 95% confidence interval: 5,477,266 bp and 100,388,643 bp in the ancestral repeats were analyzed on chromosome X and autosomes, respectively.

Table 2. The Indel Frequency on Each Rat Chromosome

Rat chromosome	N	L (bp)	Indel rate
1	152,719	10,317,729	0.0148
2	142,873	9,571,504	0.0149
3	120,685	8,124,122	0.0149
4	110,783	7,501,031	0.0148
5	109,469	7,410,038	0.0148
6	91,973	6,267,407	0.0147
7	59,734	3,962,829	0.0151
8	86,007	5,937,094	0.0145
9	68,680	4,614,189	0.0149
10	76,434	5,217,465	0.0146
11	50,951	3,460,479	0.0147
12	16,958	1,031,534	0.0164
13	57,912	3,943,344	0.0147
14	65,296	4,406,887	0.0148
15	58,299	3,922,303	0.0149
16	48,130	3,255,162	0.0148
17	50,054	3,306,985	0.0151
18	57,308	3,865,351	0.0148
19	38,513	2,595,183	0.0148
20	26,719	1,678,007	0.0159
X	70,578	5,477,266	0.0129

N indicates the total number of indel events; L, the total length of the ancestral repeats analyzed.

The resulting ratio of indel rates between chromosome X and autosomes (X/A) is 0.869, which leads to $\alpha = 2.304$ (bootstrap-based 95% confidence interval [CI], 1.986 to 2.699). Next, we analyzed the indel frequency in the ARs active just before the rodent–primate split. Such analysis led to similar results. Namely, X/A = 0.858, and the resulting $\alpha = 2.482$. Hence, the further analysis described in this section was performed on all ARs.

The higher prevalence of indels on autosomes compared with X is observed for both insertions and deletions (Table 1). Interestingly, the male-to-female ratio is higher for deletions ($\alpha = 2.538$) than for insertions ($\alpha = 1.898$), although this is not significant (Table 1). Both rat and mouse lineages exhibit lower indel rates on chromosome X than on autosomes.

The frequency of indels is significantly lower on chromosome X than on any rat autosome (Table 2; $P < 0.001$, χ^2 test; the test remains significant after applying the Bonferroni correction for multiple tests), although there is substantial variation in average indel rates among autosomes. Interestingly, the frequency spectrum of indels that occurred in rat and mouse since their divergence from the common ancestor is similar between X and autosomes (Fig. 1; $P = 0.67$, Kolmogorov-Smirnov test). We do not observe the more pronounced male bias for 1-bp indels ($P = 0.412$, χ^2 test) that was suggested by Sundstrom et al. (2003).

When all sites aligned between rat and mouse are analyzed, we once again observe a lower indel rate on X than on autosomes. The resulting X/A = 0.894, leading to $\alpha = 1.930$, comparable to α obtained from the analysis of the ARs.

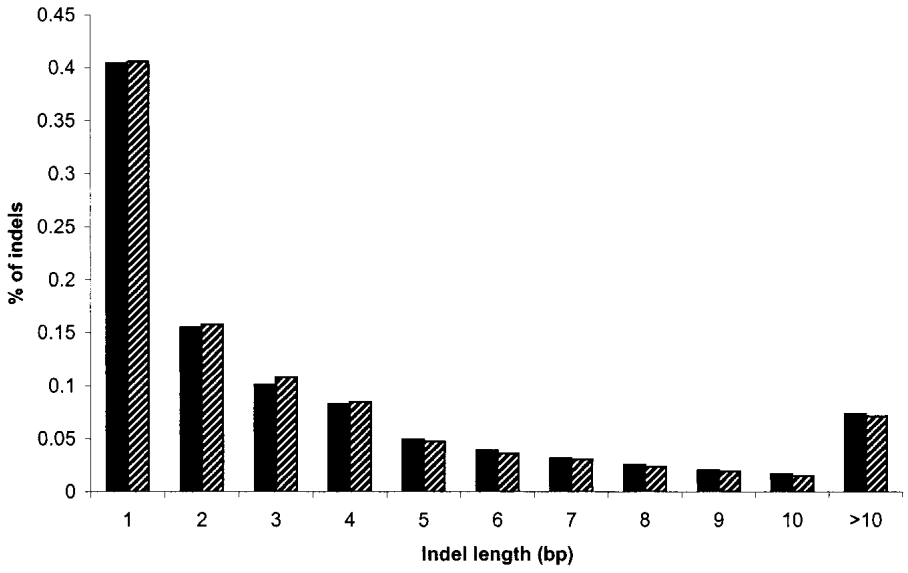


Figure 1 Relative size distribution of indels in rodent autosomes (solid bars) and chromosome X (hatched bars), based on the analysis of the ancestral repeats predating the rodent–primate split.

Nucleotide Substitutions

To obtain the genome-wide estimate of the male-to-female nucleotide substitution rate ratio in rodents, we compared the average substitution rates between chromosome X and autosomes, estimated from ARs located at orthologous locations in rat and mouse. First, we estimated the substitution rates for each alignment of orthologous ARs by applying the general time-reversible Markov model (REV; Whelan et al. 2001). Next, we calculated the average substitution rate weighted by repeat length. When all ARs are analyzed together, the average mouse–rat substitution rate is 0.150 on chromosome X (5.2 Mb) and 0.168 on autosomes (95.0 Mb; Table 3). The difference is statistically significant ($P < 0.001$, unpaired *t*-test). The ratio of substitution rates between X and autosomes for the mouse–rat comparison is 0.897, giving $\alpha = 1.892$.

When the analysis is restricted to ARs active shortly before the rodent–primate speciation, the results are comparable (Table 3). Namely, the average substitution rate is significantly lower on chromosome X than on the autosomes ($P < 0.001$, unpaired *t*-test). The resulting ratio of the substitution rates between chromosome X and autosomes is 0.895, which leads to $\alpha = 1.914$ (the bootstrap-based 95% CI is 1.651 to 2.271). Thus, a higher proportion of older ARs does not introduce a bias in the α estimates for nucleotide substitutions, similar to what we observed for indels.

It is known that substitution rates correlate with fluctuations in GC content (see Castresana 2002). In particular, the relationship between substitution rate and GC content is fit by a quadratic regression (Hardison et al. 2003). We computed a quadratic regression of substitution rate on GC content for all "young ARs" (the second subset) as well as for individual young AR families. The determination coefficient (R^2) of a quadratic regression is low: $R^2 = 1.51\%$ for all "young ARs" and is <3% for individual young AR subfamilies. Thus, variation in GC content explains only a small portion of variation in mouse–rat substitution rates among young ARs. Thus, we did not factor out the effect of GC content on variation in substitution rate.

Similar to the indel rate, the mouse–rat substitution rate is significantly lower on X than on any autosome ($P < 0.001$, unpaired *t*-test; Table 4). The significance remains after applying the Bonferroni correction for multiple comparisons.

A genome-wide comparison of substitution rates for all sites confirmed our findings from the ARs. Indeed, there is a

Table 3. Nucleotide Substitution Rates Between Mouse and Rat Estimated From the AR Sequences

Data set	N_X	L_X (bp)	$Rate_X$	N_A	L_A (bp)	$Rate_A$	X/A	α (95% CI)
All ARs	28,091	5,234,761	0.150 ± 0.072	526,062	94,984,067	0.168 ± 0.076	0.897	1.892
Young ARs	1,652	31,508	0.155 ± 0.060	31,969	6,180,668	0.174 ± 0.067	0.895	1.914 (1.651–2.271)

N_X and N_A indicate the number of individual ancestral repeats examined on chromosome X and autosomes, respectively; L_X and L_A, the lengths of the ARs examined on chromosome X and autosomes, respectively; $Rate_X$ and $Rate_A$, the nucleotide substitution rates (weighted by the length of ARs) and their standard deviations on chromosome X and autosomes, respectively; X/A, the ratio of nucleotide substitution rates on chromosome X versus autosomes; and α (95% CI), the male-to female mutation rate ratio and its 95% confidence interval (given only for young ARs, as bootstrap procedure for all ARs is very computationally intensive).

lower substitution rate on chromosome X than on autosomes when all sites aligned between rat and mouse are compared. The average nucleotide substitution rate (weighted by the length of individual alignments) is 0.152 per site for chromosome X and 0.167 per site for autosomes. Hence, the X/autosome substitution rate ratio is 0.908, so $\alpha = 1.760$.

DISCUSSION

In this study, we conducted a whole-genome comparison of mutation rates between chromosome X and autosomes in rat and mouse and came to the following conclusions. First, there is strong evidence of male mutation bias for indel mutations in rodents. To our knowledge, this is the first report of the sex-specific bias for indels in mammals. Second, we confirm male mutation bias for nucleotide substitutions in rodents. Remarkably, for both types of mutations, the male-to-female mutation rate ratio is approximately two, similar to the male-to-female ratio in the number of germline cell divisions. This strongly supports the replication-driven origin for both indels and nucleotide substitutions. One should be aware that an estimate of $c \sim 2$ in rodents is an approximation (Chang et al. 1994) and, as such, might include some error.

Previous studies indicated that indel mutations might occur during meiosis (Ketterling et al. 1994; Crow 2000) and are related to recombination (Sundstrom et al. 2003). As recombination rates are lower on rodent X chromosome than on rodent autosomes, one might argue that our results are consistent with this hypothesis. Indeed, the adjusted recombination rate for rodent chromosome X is 0.25 cM/Mb (averaged between rat and mouse and multiplied by two-thirds to account for the fact that X spends two-thirds of its time in a recombining sex; Jensen-Seaman et al. 2004), whereas the recombination rate for rodent autosomes is 0.59 cM/Mb (averaged between rat and mouse; Jensen-Seaman et al. 2004). Thus, the X/autosome recombination rate ratio is ~0.42. However, the observed X/autosome indel rate ratio is ~0.87, more than twice as high as the X/autosome recombination rate ratio. Thus, if recombination were to explain the observed difference in indel rates between chromosome X and autosomes, we would expect to see a lower X/autosome indel rate ratio. The most parsimonious explanation for our results is that most indels occur during DNA replication and/or during DNA repair after DNA replication. We cannot exclude the possibility that recombination also contributes to indel formation; however, its influence is probably smaller than that of replication. It will be interesting to further investigate the role of replication and recombination in indel mutations from a genome-wide perspective in primates, particularly in light of the recent study by Sundstrom et al. (2003) indicating input of both factors. The replication-driven origin of indels in humans is supported by the recent study of potential indel mutation mechanisms (Chuzhanova et al. 2002). These mechanisms, for instance, include slipped mispairing (misalignment of short direct repeats) during DNA replication and excision repair–mediated resolution of short inverted repeats (Chuzhanova et al. 2002).

The recombination origin of indels was hypothesized because of maternal mutation bias for large indels (usually deletions) causing some human genetic diseases (e.g., in neurofibromatosis type 1 and hemophilia B; Lazaro et al. 1996; Sommer et al. 2001). As the recombination rate is higher in females than in males (Kong et al. 2002), it is conceivable that large indels are related to recombination. In contrast, our results indicate that small indels are mostly replication driven. Thus, the present study supports a view that small and large indels originate by different molecular mechanisms (Chuzhanova et al. 2002).

Recently, it has been shown that there is a substantial variation in mutation rates within chromosomes, and point mutation rates, insertion rates of repetitive elements, and recombination rates covary locally (Lercher et al. 2001; Hardison et al. 2003). This variation is usually attributed to regional differences in GC content, DNA repair, or nuclear localization and metabolism, although the exact causative agents are presently unknown (Hardison et al. 2003). Despite this variation within chromosomes, our results indicate that there is a significantly lower mutation rate for both point mutations and indels on chromosome X than on autosomes. Thus, we suggest that there is a hierarchy of forces and effects directing variation in mutation rates in a genome, with male mutation bias being the force leading to dif-

Table 4. Average Substitution Rates per Rat Chromosome

Rat chromosome	N	L	Rate	SD
1	54,373	9,733,644	0.164	0.075
2	49,132	9,080,067	0.172	0.067
3	42,909	7,708,489	0.163	0.070
4	38,965	7,100,017	0.168	0.073
5	39,891	6,994,686	0.166	0.067
6	32,561	5,931,910	0.166	0.067
7	20,919	3,740,719	0.172	0.069
8	31,024	5,622,942	0.161	0.139
9	23,799	4,365,545	0.168	0.069
10	28,465	4,929,227	0.160	0.066
11	17,872	3,268,215	0.174	0.071
12	5,865	971,121	0.170	0.070
13	20,545	3,730,183	0.168	0.073
14	22,957	4,169,096	0.172	0.074
15	20,306	3,715,979	0.172	0.068
16	16,714	3,086,705	0.174	0.071
17	17,286	3,132,485	0.170	0.074
18	19,792	3,667,697	0.169	0.068
19	13,817	2,452,234	0.170	0.070
20	8,870	1,583,106	0.176	0.070
X	28,091	5,234,761	0.150	0.073

N indicates the number of ancestral repeats; L, the length of alignments; rate, the average substitution rate weighted by the repeat length; and SD, the standard deviation of the substitution rate.

ferences in mutation rates between sex chromosomes and autosomes (as well as between the two sex chromosomes).

Here, we use orthologous ARs as a data source for modeling neutral evolution. A potential problem associated with using ARs is in gene conversion. Roy et al. (2000) have shown that gene conversion is frequent (10% to 20%) in young subfamilies of *Alu* repeats. Gene conversion has also been shown to occur in older families of repetitive elements as well as between older and younger repeat families (Roy-Engel et al. 2002; Salem et al. 2003). It is known that the frequency of gene conversion is positively correlated with sequence similarity (Modrich and Lahue 1996), and thus, it is expected to be negatively correlated with the age of repetitive elements. To minimize effects of gene conversion, we focused our analysis on the ancient ARs inserted in the rodent–primate common ancestor. The individual copies of these ARs have already diverged greatly from their consensus sequences as well as from each other. For instance, for our younger ARs, the median divergence of individual copies from the family consensus sequence is 26% to 31% (in mouse). This number is even higher when all (younger and older) ARs are considered together. Thus, although a possibility of gene conversion cannot completely be excluded from this data set and should be further investigated, it is unlikely to be favored between such divergent sequences. Gene conversion leads to elevated nucleotide substitution rates at orthologous sites (Drouin et al. 1999). The fact that similar results were obtained from the analysis of all ARs, young ARs, and all sites aligned between mouse and rat further indicates that gene conversion is not a major factor driving the evolution of the ARs used in this study. In addition, there is no reason to hypothesize that gene conversion is biased with respect to chromosome X and autosomes. It would be informative to compare results obtained here using the ARs as the data source for modeling neutral evolution with that from other data sources (e.g., orthologous pseudogenes).

Our results based on a comparison of rates between chromosome X and autosomes await confirmation from the comparisons of mutation rates between chromosome Y and autosomes as well as between the two sex chromosomes. These latter studies will become possible once the complete sequences of the mouse and rat Y chromosome are available. Differing male-to-female mutation rate ratios among the three types of comparisons would indicate that other factors, in addition to the number of DNA replications, determine differences in mutation rates between sex chromosomes and between sex chromosomes and autosomes. One of such factors could be a specifically reduced mutation rate of the X chromosome (McVean and Hurst 1997). Interestingly, our study identified a twofold higher substitution rate in rodent males than in females, similar to the results obtained from the X–Y comparisons (Chang et al. 1994; Chang and Li 1995). However, our results contradict an infinite excess of male over female substitutions obtained in an earlier comparison of substitution rates between chromosome X and autosomes (Wolfe and Sharp 1993).

Conclusions

Nucleotide substitutions and small indels represent the two most common types of mutations causing human genetic disease (Stenson et al. 2003) and are an important source material for molecular evolution. Thus, unraveling the mechanism of their mutagenesis is of great interest. Here we have shown that in rodents small indel mutations, as well as nucleotide substitutions, originate more frequently in males than in females. In rodents, the magnitude of the male mutation bias is similar for the two types of mutations ($\alpha \approx 2$) and is close to the male-to-female ratio in the number of germline cell divisions. This is consistent with the hypothesis that DNA replication errors are the major source of small indels and nucleotide substitutions.

METHODS

Alignments

The mutation rates between orthologous rat and mouse ARs were estimated from four-way multiple alignments. The April 2003 human genome assembly (hg15), the February 2003 mouse genome assembly (mm3), and the June 2003 rat genome assembly (rn3) were aligned along with the consensus sequences for the ARs, as described in Blanchette et al. 2004). First, BLASTZ (Schwartz et al. 2003) was used to generate the pairwise human–mouse and human–rat alignments with human as a reference. The optimal alignment parameters were derived according to an alignment-scoring scheme developed by Chiaromonte et al. (2002). Second, HUMOR (Blanchette et al. 2004), a special variant of MULTIZ, was used to create the three-way human–mouse–rat alignments with human as a reference. These alignments are symmetrical in respect to the mouse and rat sequences (Blanchette et al. 2004). Third, the resulting three-way alignments were aligned to the consensus sequences of the ARs with MULTIZ (Blanchette et al. 2004). The orthology of the ARs was confirmed by extending alignments of adjacent unique genomic DNA (Schwartz et al. 2003).

The mutation rates at all sites were studied from the whole-genome alignments of rat and mouse. Namely, the February 2003 mouse genome assembly (mm3) and the June 2003 rat genome assembly (rn3) were aligned with BLASTZ (Schwartz et al. 2003). The program axtBest (Schwartz et al. 2003) was used to process the alignments to obtain single coverage of rat sequences with mouse sequences.

Estimating the Male-to-Female Mutation Rate Ratio

The mutation rates were compared between the alignments that included only chromosome X sequences and the alignments that included only autosome sequences. The alignments of sequences not assigned to chromosomes ("Unidentified") or to chromosome coordinates ("random") were excluded from the analysis. Miyata's formula (Miyata et al. 1987) was used to calculate α. Namely, if $R = X/A$, then $\alpha = (3R - 4) / (2 - 3R)$, where X and A are the mutation rates on chromosome X and autosomes.

Indels

Only small indels (<50 bp) were analyzed. A gap in one of the aligned sequences (either rat or mouse) was considered to be a result of one indel mutation. The availability of four-way alignments allowed us to distinguish between insertions and deletions as well as to assign them to either the rat or the mouse lineage. For each alignment, indel rate was calculated by dividing the number of rat- or mouse-specific insertion/deletion events by the total number of aligned nucleotides (gaps excluded) between mouse and rat.

The significance in the difference of indel frequencies between chromosome X and autosomes was tested by a χ^2-test on a two-way counts table with "indel, nonindel" versus "X, autosomes". To compare the frequencies of indels between chromosome X and each autosome, the same test was performed for each chromosome separately, and then the Bonferroni correction was applied.

To obtain confidence interval for the estimates of α, indels were resampled by using a nonparametric bootstrap (resampling with replacement), following the method of Sundstrom et al. (2003). The 95% CIs were calculated from 10,000 bootstrap replicates.

The indel frequency spectrum was compared between chromosome X and autosomes by the Kolmogorov-Smirnov test on the two distributions of indels (by length) for chromosome X and pooled for all autosomes. To test a hypothesis of a more pronounced male bias for 1-bp indels, we performed a χ^2-test on a

two-way counts table with "1-bp indel, all other indels" versus "X, autosomes".

Nucleotide Substitutions

Nucleotide substitution rates were estimated by using the general time-reversible Markov model REV (Whelan et al. 2001), as implemented in PAML software package (Yang 1997), with default parameters. Only alignments >100 bp were included in the analysis.

An unpaired *t*-test was used to compare the mean substitution rates between chromosome X and autosomes. Later, the same test was used to compare the mean substitution rates between chromosome X and each autosome, applying the Bonferroni correction for multiple tests. A weighted estimate of the variance for chromosome X and autosomes was calculated as suggested by Erlandsson et al. (2000): $s_z^2 = \{[n(\mathrm{X}) - 1]s(\mathrm{X})^2 + [n(\mathrm{A}) - 1]s(\mathrm{A})^2\} / \{[n(\mathrm{A}) - 1] + [n(\mathrm{A}) - 1]\}$, where $n(\mathrm{X})$ and $n(\mathrm{A})$ are the numbers of the ARs on X and autosomes, respectively, and $s(\mathrm{X})^2$ and $s(\mathrm{A})^2$ are the variances of the divergences on X and autosomes, respectively. The 95% CIs of α were calculated by using a two-step bootstrap procedure. In the first step, we resampled sites for each of the alignments. In the second step, within each replicate data set from step one, we resampled alignments. These two steps were used to generate 100 replicate data sets that were subsequently processed by PAML.

To assess the effects of GC content on nucleotide substitution rate, we proceeded as in Hardison et al. (2003). Namely, we regressed the substitution rate between rat and mouse on a quadratic function of GC content (the quadratic regression fit the data better than the linear or the cubic regressions did). GC content associated with each AR sequence was calculated as follows: For mouse and rat separately, we computed the percentage of bases being either "G" or "C" in a 1-Mb window surrounding the AR sequence and centered at it (with the sequence itself excluded). Then, we averaged the GC percentages computed for mouse and rat.

Statistical tests and the quadratic regression were performed with the MINITAB software package.

ACKNOWLEDGMENTS

We thank Webb Miller and Ross Hardison for comments improving the manuscript and for helpful discussions. Thanks to Hans Ellegren for suggestions on the manuscript. The sequence alignments used in this project were generated on a 1024-node Pentium III cluster at the University of California at Santa Cruz (UCSC). We acknowledge David Haussler, Krishna Roskin, and the UCSC Genome Browser staff for generating these alignments. Scott Schwartz helped with managing the alignments at Penn State. We are thankful to Paula Goetting-Minesky and Laura Elnitski for critical readings of the manuscript. We are grateful to Michael Jensen-Seaman for sharing his results before publication. This study was supported by the startup funds available to K.D.M. from the Eberly College of Sciences at Penn State University.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning muliple genomic sequences with the threaded blockset aligner. *Genome Res.* (this issue).

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

Castresana, J. 2002. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. *Nucleic Acids Res.* **30:** 1751–1756.

Chang, B.H.J. and Li, W.-H. 1995. Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked Ube 1 genes and pseudogenes. *J. Mol. Evol.* **40:** 70–77.

Chang, B.H.J., Shimmin, L.C., Shyue, S.-K., Hewett-Emmett, D., and Li, W.-H. 1994. Weak male-driven evolution in rodents. *Proc. Natl. Acad. Sci.* **91:** 827–831.

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

Chiaromonte, F., Yap, V.B., and Miller, W. 2002. Scoring pairwise genomic sequence alignments. *Pac. Symp. Biocomput.* 115–126.

Chuzhanova, N.A., Annassis, E.J., Ball, E.V., Krawczak, M., and Cooper, D.N. 2002. Meta-analysis of indels causing human genetic disease: Mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mutat.* **21:** 28–44.

Crow, J.F. 2000. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1:** 40–47.

Dermitzakis, E.T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B.J., Flegel, V., Bucher, P., Jongeneel, C.V., et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420:** 578–582.

Drouin, G., Prat, F., Ell, M., and Clarke, G.D. 1999. Detecting and characterizing gene conversions between multigene family members. *Mol. Biol. Evol.* **16:** 1369–1390.

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

Erlandsson, R., Wilson, J.F., and Pääbo, S. 2000. Sex chromosomal transposable element accumulation and male-driven substitutional evolution in humans. *Mol. Biol. Evol.* **17:** 804–812.

Hardison, R.C., Roskin, K., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13:** 13–26.

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

Hurst, L.D. and Williams, E.J.B. 2000. Covariation of GC content and the silent site substitution rate in rodents: Implications for methodology and for the evolution of isochores. *Gene* **261:** 107–114.

Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C-F., Thomas, M.A., Haussler, D., and Jacob, H.J. 2004. Comparative recombination rates in rat, mouse, and human. *Genome Res.* (this issue).

Ketterling, R.P., Vielhaber, E.L., Lind, T.J., Thorland, E.C., and Sommer, S.S. 1994. The rates and patterns of deletions in the human factor IX gene. *Am. J. Hum. Genet.* **54:** 201–213.

Kluwe, L., Mautner, V., Parry, D.M., Jacoby, L.B., Baser, M., Gusella, J., Davis, K., Stavrou, D., and MacCollin, M. 2000. The parental origin of new mutations in neurofibromatosis 2. *Neurogenetics* **3:** 17–24.

Kolb, A. 2003. The first intron of the murine β-casein gene contains a functional promoter. *Biochem. Biophys. Res. Commun.* **306:** 1099–1105.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lazaro, C., Gaona, A., Ainsworth, P., Tenconi, R., Vidaud, D., Kruyer, H., Ars, E., Volpini, V., and Estivill, X. 1996. Sex differences in mutational rate and mutational mechanism in the NF1 gene in neurofibromatosis type 1 patients. *Hum. Genet.* **98:** 696–699.

Lercher, M.J., Williams, E.J.B., and Hurst, L.D. 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., Yi, S., and Makova, K.D. 2002. Male-driven evolution. *Curr. Opin. Genet. Dev.* **12:** 650–656.

Makova, K.D. and Li, W.-H. 2002. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416:** 624–626.

Malcom, C.M., Wyckoff, G.J., and Lahn, B.T. 2004. Genic mutation rates in mammals: Local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol. Biol. Evol.* **20:** 1633–1641.

McVean, G.T. and Hurst, L.D. 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* **386:** 388–392.

Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K., and Yasunaga, T. 1987. Male-driven evolution: A model and nucleotide sequence analysis. *Cold Spring Harb. Symp. Quant. Biol.* **LII:** 863–867.

Modrich, P. and Lahue, R. 1996. Mismatch repair in replication fidelity, genetic recombination and cancer biology. *Annu. Rev. Biochem.* **65:** 101–133.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence

of the Brown Norway Rat yields insights into mammalian evolution. *Nature* (in press).
Roberts, P.S., Chung, J., Jozwiak, S., Dabora, S., Franz, D.N., Thiele, E.A., and Kwiatkowski, D. 2002. SNP identification, haplotype analysis, and parental origin of mutations in TSC2. *Hum. Genet.* **111:** 96–101.
Roy, A.M., Carroll, M.L., Nguyen, S.N., Salem, A.-H., Oldridge, M., Wilkie, A.O.M., Batzer, M., and Deininger, P.L. 2000. Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.* **10:** 1485–1495.
Roy-Engel, A.M., Carroll, M.L., El-Sawy, M., Salem, A.-H., Garber, R.K., Nguyen, S.V., Deininger, P.L., and Batzer, M.A. 2002. Non-traditional Alu evolution and primate genomic diversity. *J. Mol. Biol.* **316:** 1033–1040.
Salem, A.-H., Kilroy, G.E., Watkins, W.S., Jorde, L.B., and Batzer, M.A. 2003. Recently integrated Alu elements and human genomic diversity. *Mol. Biol. Evol.* **20:** 1349–1361.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human–mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.
Shimmin, L.C., Chang, B.H., and Li, W.-H. 1993. Male-driven evolution of DNA sequences. *Nature* **362:** 745–747.
Smith, N.G.C. and Hurst, L.D. 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.
Sommer, S.S., Scaringe, W.A., and Hill, K.A. 2001. Human germline mutation in the factor IX gene. *Mutat. Res.* **487:** 1–17.
Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M., and Cooper, D.N. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21:** 577–581.
Sundstrom, H., Webster, M.T., and Ellegren, H. 2003. Is the rate of insertion and deletion mutation male biased?: Molecular evolutionary analysis of avian and primate sex chromosome sequences. *Genetics* **164:** 259–268.
Trappe, R., Laccone, F., Cobilanschi, J., Meins, M., Huppke, P., Hanefeld, F., and Engel, W. 2001. MECP2 mutations in sporadic cases of Rett syndrome are almost exclusively of paternal origin. *Am. J. Hum. Genet.* **68:** 1093–1101.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Whelan, S., Lio, P., and Goldman, N. 2001. Molecular phylogenetics: State-of-the-art methods for looking into the past. *TIG* **17:** 262–272.
Wolfe, K.H. and Sharp, P.M. 1993. Mammalian gene evolution: Nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.* **37:** 441–456.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.

WEB SITE REFERENCES

http://ftp.genome.washington.edu/RM/RepeatMasker.html; RepeatMasker.

Received September 12, 2003; accepted in revised form December 9, 2003.