Transitions at CpG dinucleotides: are replication errors to blame?

Kateryna Makova[1,3] & James Taylor[2,3]

[1]Departments of Biology, Penn State University, University Park, Pennsylvania 16802, USA
[2]Department of Computer Science and Engineering, Penn State University, University Park, Pennsylvania 16802, USA and
[3]Center for Comparative Genomics and Bioinformatics, Penn State University, University Park, Pennsylvania 16802, USA

Received line

ABSTRACT

The sequenced human and chimpanzee genomes allow one to investigate mutations with high rates, since the divergence between the two genomes is low. We compared the rates of transitions at CpG dinucleotides among sex chromosomes and autosomes at noncoding sequences orthologous between human and chimpanzee. We discovered that in primates the transition rates at CpG dinucleotides are lowest for chromosome X, intermediate for autosomes, and highest for Y. Thus, these mutations primarily originate in the male germline which undergoes a larger number of DNA replications compared with the female germline. This suggests that transitions at CpG dinucleotides are replication-dependent.

Contact: kdm16@psu.edu (Kateryna Makova)

INTRODUCTION

Mutations provide material for molecular evolution and are the source of genetic variation in natural populations. Thus, unraveling the mechanisms of mutagenesis is critical for our understanding of the evolutionary process. However, direct studies of mutagenesis are limited by the low frequency of mutations. Here we infer the origin of mutations at CpG dinucleotides based on the comparative genomic analysis of DNA sequences.

The higher number of cell divisions in the male than in the female germline, as observed in mammals (Vogel & Motulsky 1997), can be utilized to investigate whether mutations result primarily from replication errors. In males, the germline cell divisions are continuous throughout reproductive life. In females, almost all (all except one) germline cell divisions are completed before birth. As a result, male germline cells undergo more cell divisions and can accumulate more replication errors than female germline cells do. If the male-to-female ratio in mutation rates (α) is similar to the male-to-female ratio in the number of germline cell divisions (c), then mutations result primarily from errors in DNA replication. If α is smaller than c, then the role of replication-independent factors (e.g., free radicals) is significant.

Mutation sex bias can be studied by comparing mutation rates between the two sex chromosomes or between a sex chromosome and autosomes (Miyata et al. 1987). If replication is important in generation of mutations, then the mutation rate for a chromosome depends on how often this chromosome is transmitted through the male and female germlines. Thus, in mammals, mutation rate is expected to be the highest for Y, the intermediate for autosomes, and the lowest for X.

Male bias has been shown in rates of point mutations (reviewed in Li et al. 2002) and indels (Makova et al. 2004). However, it is unclear whether all point mutations exhibit this bias, and, thus, result primarily from errors in DNA replication. In particular, unlike other point mutations, transitions at CpG dinucleotides occur due to spontaneous deamination of methylated cytosines. As a result, mutations at CpG dinucleotides were suggested to be time-dependent and not replication-dependent (Vogel & Motulsky 1997). If this hypothesis is true, then CpG transitions are predicted to have similar rates among X, Y, and autosomes.

The few investigations of potential male bias for transitions at CpG dinucleotides with the evolutionary approach led to controversial results. No differences in transition rates at CpG sites were observed between X and autosomes in primates (Nachman & Crowell 2000) and rodents (Smith & Hurst 1999). In contrast, CpG transition rates were higher on primate Y than on X (Anagnostopoulos et al. 1999), suggesting that the CpG mutation rate may depend on the number of cell divisions.

The rates of transitions at CpG dinucleotides are ~10-fold higher than that for other sites (Nachman & Crowell 2000). As a result, approximately 20% of mutations causing human genetic diseases are nucleotide substitutions at CpG dinucleotides (Krawczak et al. 1998). Therefore, it is of great importance to study the potential role of replication in generation of these mutations.

The availability of the human and chimpanzee genomic sequences provides a unique opportunity to test a hypothesis of male mutation bias for CpG transitions and test whether these mutations are replication-dependent. First, the low divergence between the two species allows one to perform a reliable analysis of CpG mutations. Indeed, recent computer

simulations indicated that such analysis becomes inaccurate when divergence is >0.1 (Subramanian & Kumar 2003). Second, the sequences of the two organisms are close enough to avoid ambiguities in alignments. Thus, here we analyze transitions at CpG dinucleotides from noncoding sites in human-chimpanzee alignments at X, Y, and autosomes and ask whether these mutations (1) exhibit male mutation bias and (2) occur during DNA replication.

METHODS

The available chimpanzee genomic sequences (a total of 1,392 Mb, including 65 Mb on X and 14 Mb on Y) were aligned to the human genomic sequences (hg16 assembly of the human genome) with the use of BLASTZ (Schwartz et al. 2003). Two data sets were analyzed. To construct the first data set, all known genes (as annotated in the human genome at the UCSC Genome Web Browser) were excluded from the analysis. Sequences located 5 kb upstream and downstream from the genes were also excluded as such sequences have a high probability of possessing gene expression regulatory elements. Furthermore, CpG sites at CpG islands (as annotated in the human genome at the UCSC Genome Web Browser) were excluded as they (unlike other CpG sites) are usually unmethylated and are not hypermutable (Bird 1999). The second data set consisted of ancestral interspersed repeats (ARs) predating primate-rodent divergence. While the first data set might still be affected by selection at some sites, the ARs are considered to evolve neutrally.

To calculate the transition rate at CpG dinucleotides, we divided the number of transitions at CpG sites by the total number of CpG sites averaged between the two species (and additionally by two as there are two nucleotides at each CpG site). A correction for multiple hits according to Kimura's 2-parameter model is not expected to change our results significantly since the divergence between the two species is very low.

RESULTS

The resulting rates of CpG transitions are the lowest at chromosome X, the intermediate at the autosomes, and the highest at chromosome Y (Table 1) for both data sets. This is consistent with male mutation bias for transitions at CpG sites. The average transition rates at CpG sites are similar to those obtained by Ebersberger et al. (2002), who also studied human-chimpanzee alignments.

Table 1. Transition rates at CpG sites from human-chimpanzee alignments.

Chromosome	Autosomes	X	Y
Noncoding set	0.153 (11.3 Mb)	0.133 (525 kb)	0.216 (125 kb)
Ancestral repeats	0.157 (3.95 Mb)	0.136 (150 kb)	0.219 (17 kb)

Remarkably, $\alpha \approx 2.3$ when calculated from the three different comparisons (Y/X, X/A, and Y/A; Table 2). This suggests that replication might play an important role in generation of these mutations. However, the resulting α is lower than the ratio of the numbers of cell divisions between male and female germlines (c), which is equal to ~5-6 for higher primates.

DISCUSSION

Here we have shown that in higher primates transitions at CpG sites originate more frequently in males than in females. This is consistent with the hypothesis that DNA replication errors are the major source of these mutations. Our data sets as well as the data set in Anagnostopoulos et al. (1999) are substantially larger than the ones analyzed by Nachman and Crowell (2000) and by Smith and Hurst (1999). This might potentially explain the differences in the results among these studies.

How can we explain our observation that the male-to-female mutation rate at CpG transitions is lower than the male-to-female ratio in the number of germline cell divisions? Since human and chimpanzee are closely related, α can be underestimated because of ancient polymorphism. To correct for this, estimates of nucleotide diversity at CpG sites are required (Makova & Li 2002). This information is currently unavailable and we plan to derive it from the SNP Consortium data. Additionally, we will calculate α from the human-macaque alignments after the macaque genome sequences become available. Ancient genetic polymorphism is not expected to bias these α estimates significantly since human and macaque are not closely related. However, if α estimates for transitions at CpG sites are still low, other explanations should be considered. First, α might indeed be low for transitions at CpG sites and replication-independent factors might be important for these mutations. Second, sex bias for point mutations at both CpG sites and non-CpG sites might be lower than c for higher primates. This will be suggestive of the role of replication-independent factors in mutagenesis of all point mutations in primates. A comparison between α estimates for CpG sites vs. non-CpG sites is needed to distinguish between the two latter alternatives.

Additionally, we intend to calculate substitution rates employing a context-dependent maximum likelihood model recently developed by Siepel and Haussler (2004). This model allows parameter estimation based on sequence

Table 2. Male-to-female mutation rate ratio at CpG dinucleotide sites from human-chimpanzee alignments.

Comparison	Y/X	X/Autosomes	Y/Autosomes
α from noncoding set	2.35	2.29	2.39
α from ancestral repeats	2.32	2.35	2.31

context, and has been successfully used by the authors (Siepel & Haussler 2004) to study mutation rates at CpG dinucleotides from mammalian alignments. Although this method will provide more accurate estimates of mutation rates, this will unlikely change our primary conclusions.

Here we assume that the mutation rate at CpG sites is influenced by methylation. Indeed, 80% of CpG sites located outside of CpG islands are methylated in mammalian genomes (Bird 1999). However, currently, we do not have information about the present or ancestral methylation status of the sites we propose to study. Furthermore, methylation levels at CpGs outside of CpG islands might differ among chromosomes (similar to methylation of CpG islands on inactivated X chromosome). However, in this case, mutation rate is not expected to be proportional to the number of germline cell divisions. Genome-wide scans of DNA methylation (Strichman-Almashanu et al. 2002) will allow more precise tests of the role of replication in point mutations at CpG sites.

Mutations provide material for molecular evolution and are the source of genetic variation in natural populations. Thus, unraveling the mechanisms of mutagenesis is critical for our understanding of the evolutionary process. However, direct studies of mutagenesis are limited by the low frequency of mutations. Here we infer the origin of mutations at CpG dinucleotides based on the comparative genomic analysis of DNA sequences.

The higher number of cell divisions in the male than in the female germline, as observed in mammals (Vogel & Motulsky 1997), can be utilized to investigate whether mutations result primarily from replication errors. In males, the germline cell divisions are continuous throughout reproductive life. In females, almost all (all except one) germline cell divisions are completed before birth. As a result, male germline cells undergo more cell divisions and can accumulate more replication errors than female germline cells do. If the male-to-female ratio in mutation rates (α) is similar to the male-to-female ratio in the number of germline cell divisions (c), then mutations result primarily from errors in DNA replication. If α is smaller than c, then the role of replication-independent factors (e.g., free radicals) is significant.

REFERENCES

Anagnostopoulos T, Green PM and Rowley G (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. 64:508-17.

Bird A. DNA methylation de novo. (1999) Science 286(5448):2287-8.

Ebersberger, I., D. Metzler, C. Schwarz and S. Paabo (2002). Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet 70(6): 1490-7.

Krawczak M, Ball EV, Cooper DN. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet. 1998 Aug;63(2):474-88.

Li, W.-H., S. Yi, and K. D. Makova (2002) Male-driven evolution. Curr. Opin. Genet. Devel. 12:650-656.

Makova, K. D., and W.-H. Li (2002) Strong male-driven evolution of DNA sequences in humans and apes. Nature 416:624-6.

Makova, K. D., S. Yang, F. Chiaromonte (2004). Indels are male-biased too: a whole-genome analysis in rodents. Genome Res (in press)

Miyata, T., H. Hayashida, K. Kuma, K. Mitsuyasu and T. Yasunaga (1987). Male-driven molecular evolution: a model and nucleotide sequence analysis. Cold Spring Harb Symp Quant Biol 52: 863-7.

Monk M, Boubelik M, Lehnert S. (1987) Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development. 99(3):371-82.

Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156:297-304.

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(1):103-7.

Siepel A, Haussler D. (2004) Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood. Mol Biol Evol. 21(3):468-88.

Strichman-Almashanu, L. Z., R. S. Lee, P. O. Onyango, E. Perlman, F. Flam, M. B. Frieman and A. P. Feinberg (2002). A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. Genome Res 12(4): 543-54.

Subramanian S, Kumar S. (2003) Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. 13(5):838-44.

Vogel, F., and A. G. Motulsky (1997) Human genetics: problems and approaches. Springer, Verlag.