

Strong male-driven evolution of DNA sequences in humans and apes

Kateryna D. Makova & Wen-Hsiung Li

Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA

Studies of human genetic diseases have suggested a higher mutation rate in males than in females¹ and the male-to-female ratio (α) of mutation rate has been estimated from DNA sequence and microsatellite data to be about 4–6 in higher primates^{2–5}. Two recent studies, however, claim that α is only about 2 in humans^{6,7}. This is even smaller than the estimates ($\alpha > 4$) for carnivores and birds^{8,9}; humans should have a higher α than carnivores and birds because of a longer generation time and a larger sex difference in the number of germ cell cycles. To resolve this issue, we sequenced a noncoding fragment on Y of about 10.4 kilobases (kb) and a homologous region on chromosome 3 in humans, greater apes, and lesser apes. Here we show that our estimate of α from the internal branches of the phylogeny is 5.25 (95% confidence interval (CI) 2.44 to ∞), similar to the previous estimates^{2–5}, but significantly higher than the two recent ones^{6,7}. In contrast, for the external (short, species-specific) branches, α is only 2.23 (95% CI: 1.47–3.84). We suggest that closely related species are not suitable for estimating α , because of ancient polymorphism and other factors. Moreover, we provide an explanation for the small estimate of α in a previous study⁶. Our study reinstates a high α in hominoids and supports the view that DNA replication errors are the primary source of germline evolution.

An effective way to estimate α is to use highly similar noncoding sequences on different types of chromosomes¹⁰. Noting that the *DAZ* locus was translocated from chromosome 3 to Y after the split between Old and New World monkeys¹¹, we sequenced three noncoding segments (~10.4 kb total) on chromosomes Y and 3 in the human, bonobo, gorilla, siamang and gibbon.

When we estimate the α -values separately for the external and the

internal branches of the tree (Fig. 1) an intriguing pattern emerges (Table 1). The α -values vary enormously among the external branches, but their average is low. The sum of the external branch lengths is 4.43% for the Y sequences (Y) and 3.22% for the chromosome 3 sequences (A), leading to $Y/A = 1.38$. According to a previous study¹²: $Y/A = 2\alpha/(1 + \alpha)$, so $\alpha = 2.23$ (95% CI: 1.47–3.84), not significantly different from the estimate (1.55) in ref. 6. In contrast, the sum of the internal branch lengths is 3.77% for the Y sequences and 2.25% for the chromosome 3 sequences, leading to $Y/A = 1.68$ and $\alpha = 5.25$ (95% CI: 2.44 to ∞). This is similar to the earlier estimates of α in primates^{2–5}, but significantly higher than those in refs 6 and 7. Thus, estimates of α obtained from closely related species tend to be lower than those obtained from more distantly related species.

We wondered how to explain the above discrepancy. The level of polymorphism on chromosome Y is usually extremely low because of selective sweep and background selection^{13,14} and because the effective population size of Y is only one-quarter of that of an autosome. The average divergence between two species is equal to $\pi + 2ut$, where t is the divergence time, u is the mutation rate, and π is the average divergence between two sequences in the ancestral population (ancient nucleotide diversity) at the time of speciation¹⁵. The ratio of Y divergence (Y) to autosome divergence (A) is

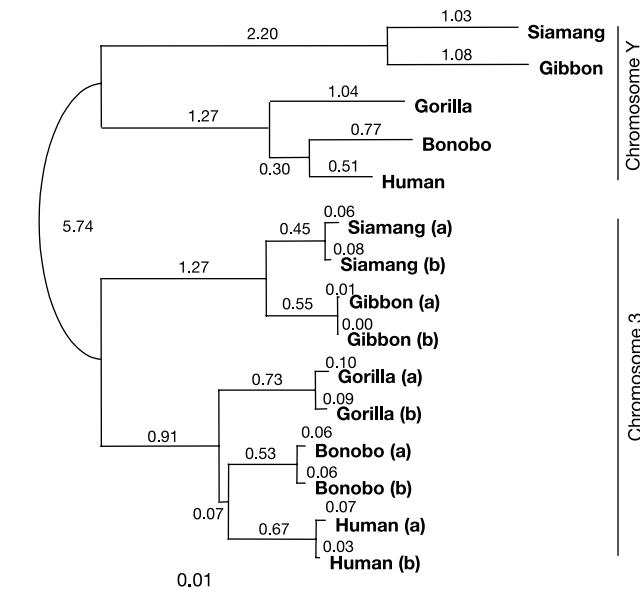


Figure 1 Phylogenetic tree of nucleotide sequences. Branch lengths were estimated from the pairwise evolutionary distances (substitutions per 100 sites). The pseudosequences of the two alleles at the chromosome 3 locus in each species are labelled a and b.

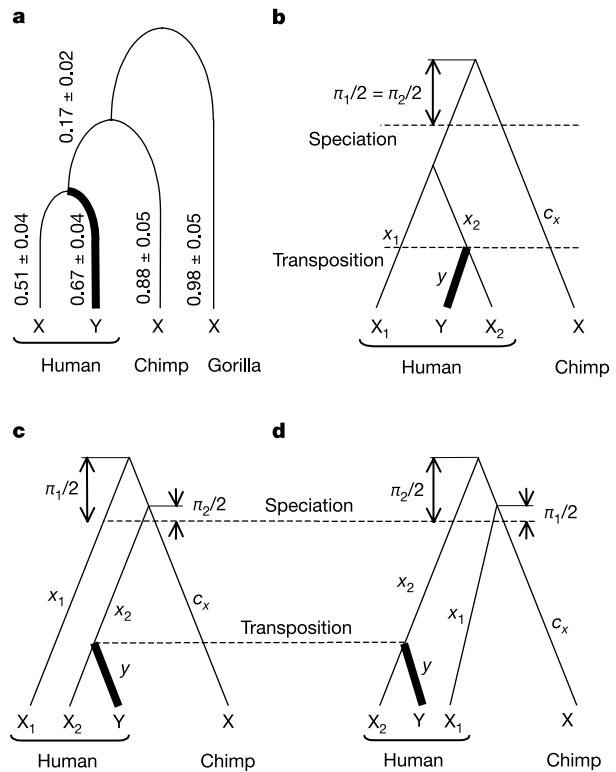


Figure 2 Phylogenetic schemes for the sequences of ref. 6. **a**, The original scheme (modified from ref. 6). **b–d**, Three possible schemes: one of them is likely to be true. π_1 , the ancient nucleotide diversity between the ancestral human X_1 and chimpanzee X chromosomes; π_2 , the ancient nucleotide diversity between the ancestral human X_2 and chimpanzee X chromosomes; x_1 , the number of nucleotide substitutions on human X_1 since the common ancestor of X_1 and X_2 or since the speciation; x_2 , the number of nucleotide substitutions on human X_2 between the common ancestor of X_1 and X_2 (or the speciation) and the transposition; c_x , the number of nucleotide substitutions on chimpanzee X chromosome; y , the number of nucleotide substitutions on the branch for the human Y chromosome locus from the time of transposition to present. Note that **a** assumes $X_2 = X_1$.

Table 1 Male-to-female ratio (α) of mutation rate

Branch	Y*	A*	Y/A	α	95% CI
External branches					
Human	0.51	0.72	0.71	0.55	0.31–0.98
Bonobo	0.77	0.59	1.31	1.90	0.87–13.29
Gorilla	1.04	0.83	1.25	1.67	0.90–5.50
Siamang	1.03	0.52	1.98	99.00	2.40– ∞
Gibbon	1.08	0.56	1.93	27.57	2.40– ∞
Total	4.43	3.22	1.38	2.23	1.47–3.84
Internal branches					
	3.77	2.25	1.68	5.25	2.44– ∞

Ratios were estimated for external (species-specific) and internal branches from a ~10.4-kb segment.

* Branch length in terms of the number of substitutions per 100 sites.

$Y/A = (y + \pi_y)/(a + \pi_a)$, where y is the number of mutations on Y, a is the number of mutations on an autosome, and π_y and π_a are the ancient nucleotide diversities at the point of speciation on Y and an autosome, respectively. As π_y is usually lower than π_a , $(y + \pi_y)/(a + \pi_a)$ can be substantially lower than y/a for closely related species. For instance, π_a between the two gorilla alleles at the chromosome 3 locus is 0.19% (Fig. 1). If we assume $\pi_a = 0.19\%$ and $\pi_y \approx 0$ in the common ancestor of the bonobo and the gorilla, then Y/A between the bonobo and the gorilla increases from $2.11\%/1.49\% = 1.42$ to $2.11\%/(1.49\% - 0.19\%) = 1.62$, and α increases from 2.45 to 4.26. Clearly, ancient polymorphism can considerably reduce the estimate of α for closely related species. If A is small, even a small error in the estimate of A can have a strong effect on Y/A .

A more serious problem in ref. 6 is that their phylogeny (Fig. 2a) assumes that the Y-linked sequence, which was derived from the transposition of an X-linked sequence in an ancestral human population, has always evolved as a Y-linked sequence since its separation from the X-linked sequence sampled in their study. This is unlikely. Rather, one of the schemes in Fig. 2b–d should be true. In Fig. 2b, when chimpanzee X is used as a reference, we have $X = x_1$ and $Y = x_2 + y$. If $y < x_2$, Y can be close to X and α can be seriously underestimated. In Fig. 2c, $X = \pi_1 + x_1$, $Y = \pi_2 + x_2 + y$ and $\pi_1 > \pi_2$, so if y is very small, Y can even be smaller than X; that is, both Y/X and α can be smaller than 1. In Fig. 2d, $\pi_1 < \pi_2$, so Y/X is greater than 1. However, Y/X can be close to 1, if y is small and if π_2 is close to π_1 . Among the three schemes, Fig. 2b is most probable because the observed human X–human Y divergence (1.18%) is smaller than the observed human X–chimpanzee X divergence (1.56%)⁶. In any case, α is underestimated if Fig. 2a is used as the phylogeny.

As to the estimate of $\alpha = 2.1$ in ref. 7, corrections for multiple substitutions were not made in the data analysis. This could have underestimated α for old repetitive elements. For young elements, the low estimate of α could be in part due to low polymorphism on Y. One additional problem was the false assumption that the repetitive elements of the same subfamily were inserted into the genome at the same time¹⁶, introducing errors into the estimation of α .

We have provided a resolution for the controversy on the magnitude of α in primates. Most previous studies compared X- and Y-linked sequences (see ref. 2), and some have argued that the high α -values could be due to a reduction in mutation rate in X rather than an elevated rate in Y¹⁷. To avoid this possibility, we compared an autosomal and a Y-linked locus. The fact that this study and all previous studies, with the exception of comparisons between closely related species, give consistent estimates of $\alpha \approx 4$ –6 provides conclusive evidence for strong male-driven evolution in hominoids. □

Methods

DNA sequencing

We studied three homologous fragments on Y and chromosome 3 that correspond to nucleotides 36,887–40,238, 49,037–54,082, and 55,171–57,328 of human Y contig

AC006983.4 and nucleotides 59,378–56,044, 47,226–42,245, and 40,284–38,159 of human chromosome 3 contig AC010139.4. The 5' end of each fragment is located 1.5, 13.7 and 20.6 kb, respectively, downstream from the 3' untranslated region of the *DAZ* gene and they have no homology with an expressed sequence tags and contain no known or predicted gene. Interspersed and simple repeats were identified using RepeatMasker (<http://ftp.genome.washington.edu/RM/RepeatMasker.html>). Searches employing GenScan and BLAST failed to identify any genes or exons within these fragments.

Genomic DNAs of a female and a male human, bonobo (*Pan paniscus*), gorilla (*Gorilla gorilla*), gibbon (*Hylobates lar*) and siamang (*Hylobates syndactylus*) were used separately as templates in polymerase chain reaction (PCR). The Expand High Fidelity PCR system (Roche) was used to minimize errors in PCR amplification. The PCR primers were designed from the human sequence (available as Supplementary Information). Female DNA was used to amplify the chromosome 3 locus and male DNA to amplify the chromosome Y locus. The sequences were deposited in GenBank under accession numbers AF483550–AF483579.

Statistical analyses

The sequences of the fragments studied were concatenated and aligned using the MegAlign module of DNASTar (Lasergene), and the alignment was adjusted manually. As the divergence between the sequences studied is low (at most 11%), the alignment was rather simple. We found evidence of at least two copies of the *DAZ* locus on chromosome Y in human (there were five 'polymorphic' sites in a 10.4-kb sequence), gorilla (two polymorphic sites), and siamang (four polymorphic sites). Thus, the copies were almost identical, suggesting that the duplications happened recently.

To take the presence of polymorphisms on chromosome 3 into account, two pseudosequences were generated for each of the individuals studied. For example, if the site was heterozygous for A and G, one of the pseudosequences was assigned 'A' and the other was assigned 'G'. This assignment was done randomly by DAMBE¹⁸. Insertion/deletion polymorphisms (indels) in the alignment, as well as heterozygous indel sites, were not included in the calculations. These indels included one deletion of 106 base pairs (bp) in gibbon Y and one insertion of 33 bp in human Y, although most of the others were short indels, many of which were in mononucleotide microsatellites. One segment of 121 bp in gorilla chromosome 3 was difficult to sequence and was excluded from analysis. So, in total, 1,148 nucleotide sites were excluded from analysis. Tajima and Nei's genetic distances¹⁹ were estimated using DAMBE¹⁸. The distances were apportioned among the branches of the phylogeny using the FITCH module of PHYLIP²⁰. The ratio of substitution rates (Y/A) was calculated from the sum of the branches on Y (Y) and chromosome 3 (A). The average of the two values for the chromosome 3 pseudosequences was used in the calculations. This procedure was also used for treating the 'polymorphic' sites in Y-linked sequences. The formula from ref. 12, $Y/A = 2\alpha/(1 + \alpha)$, was used to calculate α . To calculate the 95% CI for α , we used two methods. First, we derived the variance of Y/A . If L is the length of the sequence, $V(Y) = Y(1 - Y)/[L(1 - 4Y/3)^2]$, $V(A) = A(1 - A)/[L(1 - 4A/3)^2]$, and $V(Y/A) = V(Y)/E(A)^2 + E(Y)^2V(A)/E(A)^4$. Second, we used the bootstrap. The results for the two methods were similar. The second method was used in the text.

Received 8 October 2001; accepted 25 January 2002.

- Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nature Rev. Genet.* **1**, 40–47 (2000).
- Huang, W., Chang, B. H.-J., Gu, X., Hewett-Emmett, D. & Li, W.-H. Sex differences in mutation rate in higher primates estimated from AMG intron sequences. *J. Mol. Evol.* **44**, 463–465 (1997).
- Agulnik, A. I. *et al.* Evolution of the *DAZ* gene family suggests that Y-linked *DAZ* plays little, or a limited, role in spermatogenesis but underlines a recent African origin for human populations. *Hum. Mol. Genet.* **7**, 1371–1377 (1998).
- Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
- Ellegren, H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genet.* **24**, 400–402 (2000).
- Bohossian, H. B., Skaletsky, H. & Page, D. C. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* **406**, 622–625 (2000).
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2000).
- Pecon Slattery, J. & O'Brien, S. J. Patterns of Y and X chromosome DNA sequence divergence during the Felidae radiation. *Genetics* **148**, 1245–1255 (1998).
- Ellegren, H. & Fridolfsson, A. K. Male-driven evolution of DNA sequences in birds. *Nature Genet.* **17**, 182–184 (1997).
- Shimmin, L. C., Chang, B. H.-J., Hewett-Emmett, D. & Li, W.-H. Potential problems in estimating the male-to-female mutation rate ratio from DNA sequence data. *J. Mol. Evol.* **37**, 160–166 (1993).
- Saxena, R. *et al.* Four *DAZ* genes in two clusters found in the *AZFc* region of the human Y chromosome. *Genomics* **67**, 256–267 (2000).
- Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K. & Yasunaga, T. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 863–867 (1987).
- Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520 (1992).
- Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Phil. Trans. R. Soc. Lond. B* **355**, 1563–1572 (2000).
- Li, W.-H. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics* **85**, 331–337 (1977).
- Erlandsson, R., Wilson, J. F. & Paabo, S. Sex chromosomal transposable element accumulation and male-driven substitutional evolution in humans. *Mol. Biol. Evol.* **17**, 804–812 (2000).
- McVean, G. T. & Hurst, L. D. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature* **386**, 388–392 (1997).

18. Xia, X. *Data Analysis in Molecular Biology and Evolution* (Kluwer Academic, Boston, 2000).
 19. Tajima, F. & Nei, M. Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* **1**, 269–285 (1984).
 20. Felsenstein, J. PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* **5**, 164–166 (1989).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

The DNA samples were purchased from San Diego Zoological Society and the gibbon sample was given by M. Jensen-Seaman. We thank J. Crow and D. Page for comments. This study was supported by NIH grants.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.-H.L. (e-mail: whli@uchicago.edu).

Future projections for Mexican faunas under global climate change scenarios

A. Townsend Peterson*, Miguel A. Ortega-Huerta†, Jeremy Bartley‡, Victor Sánchez-Cordero§, Jorge Soberón||, Robert H. Buddemeier‡ & David R. B. Stockwell¶

* Natural History Museum; † Department of Geography and Kansas Applied Remote Sensing Program; ‡ Kansas Geological Survey and Department of Geography, The University of Kansas, Lawrence, Kansas 66045, USA
 § Departamento de Zoología, Instituto de Biología, Apartado Postal 70-153; || Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, D.F. 04510, México
 ¶ San Diego Supercomputer Center, University of California, 9500 Gilman Drive, La Jolla, California 92093, USA

Global climates are changing rapidly, with unexpected consequences¹. Because elements of biodiversity respond intimately to climate as an important driving force of distributional limitation², distributional shifts and biodiversity losses are expected^{3,4}. Nevertheless, in spite of modelling efforts focused on single species² or entire ecosystems⁵, a few preliminary surveys of fauna-wide effects^{6,7}, and evidence of climate change-mediated shifts in several species^{8,9}, the likely effects of climate change on species' distributions remain little known, and fauna-wide or community-level effects are almost completely unexplored⁶. Here, using a genetic algorithm and museum specimen occurrence data, we develop ecological niche models for 1,870 species occurring in Mexico and project them onto two climate surfaces modelled for 2055. Although extinctions and drastic range reductions are predicted to be relatively few, species turnover in some local communities is predicted to be high (>40% of species), suggesting that severe ecological perturbations may result.

We present a fauna-wide suite of predictions of the biodiversity consequences of global climate change. Taking advantage of the enormous biodiversity data resources accumulated by Mexico's Comisión para el Conocimiento y Uso de la Biodiversidad (CONABIO), as well as new tools in biodiversity informatics and quantitative geography¹⁰, we develop predictions of the effects of global climate change on distributions of 1,870 species (all 1,179 birds, all 416 mammals and all 175 butterflies of the families Papilionidae and Pieridae in Mexico) under two scenarios of global climate change (one conservative, one liberal) and three assump-

tions about dispersal ability¹¹. We use the Genetic Algorithm for Rule-set Prediction (GARP), a machine-learning system that has shown excellent predictive ability in delineating species' ecological niches and predicting geographic distributions¹². This first large-scale application of species-by-species models to the challenge of understanding the biodiversity consequences of global climate change provides a first view of the consequences of climate-change processes across a biodiversity-rich region.

To provide a detailed example of model results for a species, we examine the distribution of the west Mexican chachalaca (*Ortalis poliocephala*), a bird species endemic to tropical southwestern Mexico. Its present geographic distribution was summarized well by the ecological niche model, with correct prediction of 90% of independent test points, indicating that the model of the species' ecological niche was adequate. Under both scenarios of climate change, although the coastal portion of the species' distribution remained intact, the interior portion became less habitable, and a narrow band in the foothills of the coastal mountain ranges (Sierra Madre del Sur) more habitable (Fig. 1). Assuming universal dispersal abilities (not a good assumption), the species would encounter overall somewhat less (22.7% decline under the conservative climate change scenario) or more (75.1% increase under the liberal scenario) habitable area. However, under more realistic limited dispersal assumptions (dispersal into contiguous areas or no dispersal), the two scenarios agree more closely (conservative scenario 33.7% decrease, liberal scenario 29.7% decrease).

Examining patterns of change in distributional area across all 1,870 species, two climate-change scenarios, and three distributional assumptions analysed, consistent patterns emerge: species'

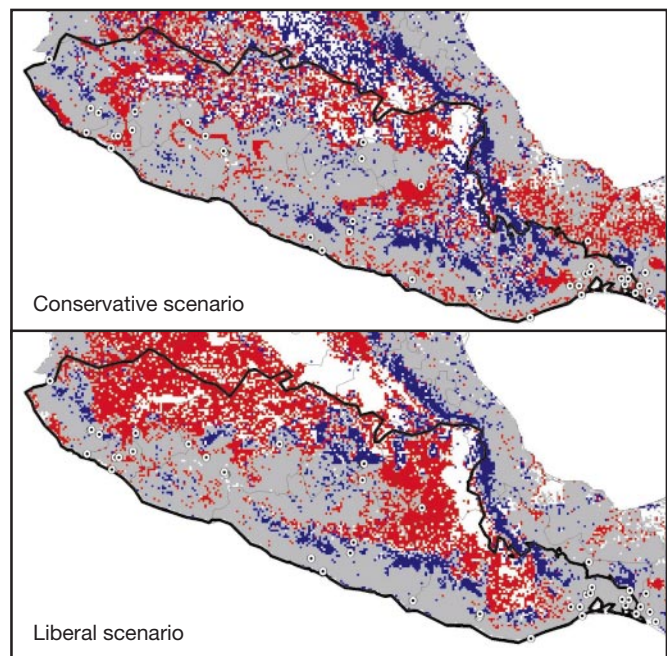


Figure 1 Example of analyses of effects of global climate change on a species' (*Ortalis poliocephala*) potential geographic distribution. Top, changes expected under the conservative scenario of the Hadley simulation; bottom, changes expected under the liberal scenario. Grey areas indicate conditions appropriate for the species both at present and under the scenario of change; white areas indicate conditions inappropriate for the species both at present and under the scenario of change; red areas are at present appropriate, but are predicted to become uninhabitable by the species; blue areas are at present not appropriate, but are predicted to become appropriate for the species; dotted circles represent known occurrence points; the black line indicates present geographic limits to species' distribution. Areas outside the black line correspond to the distributions of other *Ortalis* species, and represent areas of overprediction owing to historical factors²¹.