

The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes

Michael F Hammer^{1,2}, August E Woerner¹, Fernando L Mendez², Joseph C Watkins³, Murray P Cox¹ & Jeffrey D Wall⁴

The ratio of X-linked to autosomal diversity was estimated from an analysis of six human genome sequences and found to deviate from the expected value of 0.75. However, the direction of this deviation depends on whether a particular sequence is close to or far from the nearest gene. This pattern may be explained by stronger locally acting selection on X-linked genes compared with autosomal genes, combined with larger effective population sizes for females than for males.

Researchers in two recent studies found that the ratio of X-linked to autosomal nucleotide diversity departed from the expected value under neutral equilibrium. Keinan *et al.*¹ reported ratios of human X chromosome to autosomal nucleotide diversity that were lower than the expected value of 0.75 in their analysis of publicly available genome sequences. This deviation was apparent after the X chromosome and autosomal diversity (π) values were normalized by human-macaque divergence (D) to account for differences in mutation rates between these two parts of the genome. These researchers found the ratio of π/D for the X chromosome to π/D for the autosomes—which serves as a simple proxy for relative effective population size (defined as N_{eX}/N_{eA})—to be 0.64, 0.61 and 0.76 in their sample of five North Europeans, four East Asians and five West Africans, respectively. They concluded that the lower-than-expected N_{eX}/N_{eA} ratio in non-Africans was the result of a *smaller* effective number of mating females than males. Based on a sequence survey of 40 X-linked and autosomal intergenic regions, Hammer *et al.*² found that the N_{eX}/N_{eA} ratio was higher than 0.75 in all six populations they examined; the ratio reached a maximum value of 1.08 in their sample of French Basque. They concluded that the observed increase in X-linked diversity was the result of an excess of breeding females.

In an attempt to reconcile these dramatically different results, Bustamante and Ramachandran³ pointed out that these two studies did not disagree substantially in their estimates of autosomal nucleotide diversity (π_A) and that most of the discrepancy could be traced to the 40–50% higher estimates of X chromosome diversity (π_X) inferred

by Hammer *et al.*². Beyond the use of different outgroups (Hammer *et al.*² measured divergence to orangutan instead of macaque), each study dealt with the potential confounding effects of natural selection in different ways. To minimize the possibility of linkage to sites affected by natural selection, Hammer *et al.*² chose to sequence a set of loci that are located far from genes and that lie within regions that have moderately high rates of crossing over. Keinan *et al.*¹ analyzed genomic-scale data, with the presumed assumption that a large number of loci should reflect the demographic history of the population rather than the localized effects of selection. To test this assumption, Keinan *et al.*¹ examined regions at increasing physical distances from genes and found that the N_{eX}/N_{eA} ratios remained similar. However, it is genetic distance, not physical distance, from a gene that is expected to correlate with the probability that a neutral site is affected by natural selection^{4,5}. To further investigate the underlying causes of this discrepancy, we repeated the analyses of Hammer *et al.*² on an expanded dataset and explored the relationship between π/D and genetic distance from genes on the X chromosome and on autosomes in genomic-scale sequence data.

Our updated resequencing dataset consisted of 30 loci on the X chromosome and 61 loci on the autosomes, representing ~13.6 Mb of X-linked sequence data and ~70.0 Mb of autosomal sequence data (**Supplementary Methods**). Although our estimates of π_A are still similar to those reported by Keinan *et al.*¹, our estimates of π_X are now 23% higher for Africans and 35% higher for non-Africans (**Supplementary Table 1**). As in our earlier study, the N_{eX}/N_{eA} ratio estimate was still greater than 0.75 in all six populations and was significantly greater than 0.75 in three populations according to the likelihood ratio test (**Supplementary Fig. 1**).

Next, we examined π/D at varying genetic distances from functional elements in six publicly available human genomes from individuals of European descent (**Supplementary Methods**). We used two approaches (involving binned and continuous distances) to cull sequences from the full genome data. In the binned approach, each nongenic region was divided into bins of length 0.1 centimorgan (cM), whereas in the continuous approach, we used the medial 0.1-cM portion of each nongenic region (**Supplementary Fig. 2**). For the binned approach, we compared coverage for different outgroups, aligned human and orangutan sequences (**Supplementary Table 2**) and computed values of π and D based on a total of 19,568 autosomal regions and 831 X-linked aligned regions and obtained an N_{eX}/N_{eA} estimate of 0.71 (**Supplementary Table 3**). However, reduced diversity on the X chromosome was most pronounced in regions closely linked to genes (**Fig. 1**). This pattern disappeared—and was even reversed—as increasing genetic distances from genic regions (for example, N_{eX}/N_{eA} ratios increased from 0.67 to 1.11 in bins 1 through 4). This result also

¹Arizona Research Laboratories, Division of Biotechnology, University of Arizona, Tucson, Arizona, USA. ²Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, USA. ³Department of Mathematics, University of Arizona, Tucson, Arizona, USA. ⁴Institute for Human Genetics, University of California San Francisco, San Francisco, California, USA. Correspondence should be addressed to M.F.H. (mfh@email.arizona.edu).

Received 15 March; accepted 29 July; published online 29 August 2010; doi:10.1038/ng.651

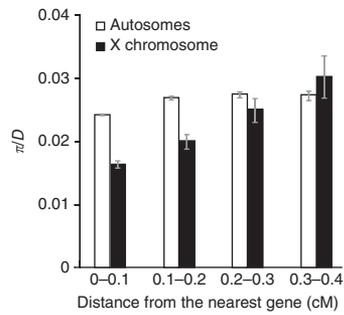


Figure 1 Diversity (π/D) as a function of genetic distance from genes. The values shown are means \pm s.e.m. See **Supplementary Figure 2** and **Supplementary Methods** for a summary of the strategy for sampling genomes and **Supplementary Figure 3** for a comparison with diversity as a function of physical distance from genes¹.

held when we used different outgroups (**Supplementary Table 4**). In contrast, when we measured diversity as a function of physical distance from genes at the attenuated scale used by Keinan *et al.*¹, we reproduced their results (**Supplementary Fig. 3**). Because the vast majority of the regions compared were found in bins that are closest to genes (**Supplementary Fig. 4**), the average π/D across the X chromosome was substantially lower than in bins far from genes. We then plotted π/D values for regions at varying genetic distances from genes using the continuous approach and found that the slope in the plot for the X chromosome (note scaled axis) was greater than that for the autosomes (**Fig. 2**) using a *t* test based on an iterated weighted least squares regression (Student's *t* = 3.19, *P* = 0.0007).

What factors could account for these observed patterns? First, we must consider whether our results reflect biases in the ways that SNPs are called or genes are identified on the X chromosome as compared with the autosomes. Still, it remains difficult to explain why these biases would lead to an overestimate of the amount of diversity on the X chromosome and, especially, why this effect would be a function of genetic distance from genes. Similarly, demographic factors affecting the entire genome, such as population bottlenecks or expansions⁶, are unlikely to preferentially increase X-linked diversity with distance from genes. On the other hand, there are theoretical reasons to believe that the joint effects of selection and linkage might contribute to this pattern. For example, background selection⁴ and selective sweeps⁵ can both result in reduced nucleotide diversity at linked sites. The X chromosome does not cross over with the Y chromosome in males except in the pseudoautosomal regions, and therefore the X chromosome has a lower evolutionarily effective recombination rate than the

autosomes (~1/3 less) in regions with comparable recombination rates on the X chromosome and on the autosomes. This could enhance both background selection and selective sweep effects⁷. Moreover, the role of selection at linked sites is influenced by the different numbers of copies of X chromosomes and autosomes between males and females. For instance, hemizygosity in males means that novel recessive beneficial mutations on the X chromosome will be exposed to positive selection sooner than mutations on the autosomes, whereas recessive deleterious mutations will be removed more efficiently from the X chromosome than from the autosomes^{8,9}. Indeed, the X chromosome differs from the autosomes in that it has unusually high linkage disequilibrium and a higher rate of nonsynonymous substitutions, features that have been interpreted as the consequence of more effective positive selection^{10–12} in accord with population genetics theory¹³.

These results lead us to posit that the shift from a deficit to an excess of X-linked diversity at increasing distances from genes can only be explained by the local effects of diversity-reducing selection near genes in combination with the global effects of an excess of breeding females. Because gene density and mean length are lower on the X chromosome than on most autosomes¹⁴, hemizygosity in males, rather than an excess of selective targets on the X chromosome, may be driving the observed trends¹⁵. If this hypothesis is correct, multiple evolutionary processes may confound inferences based on wholesale comparisons of full genome sequence data. If we wish to disentangle the history of selection, recombination and demography, a targeted set of carefully chosen regions at sufficient genetic distances from functional elements is needed. Intriguingly, at least for the human X chromosome, the signature left solely by demographic history may be hidden in the small fraction of selectively neutral polymorphisms that reside far from genes.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank O. Savina for excellent computational support and M. Nachman for comments on the manuscript.

AUTHOR CONTRIBUTIONS

M.F.H., A.E.W. and J.D.W. conceived and designed the study and wrote the paper. F.L.M., M.P.C. and J.C.W. designed the statistical analyses. A.E.W. performed bioinformatics and statistical analyses. All authors discussed the results and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturegenetics/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

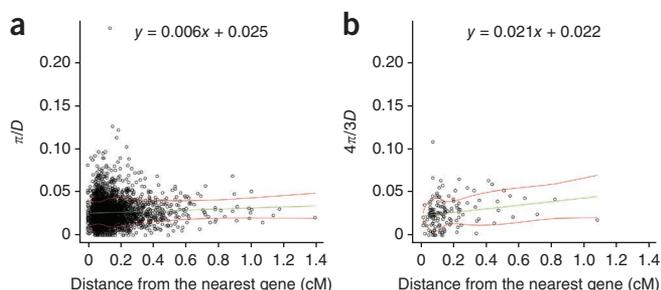


Figure 2 Scatterplot of π/D versus genetic distance from genes. (a) Autosomes. (b) X chromosome. Each point represents a 0.1 cM region. The regression line was determined by iterated weighted least-squares regression (**Supplementary Methods**). Red curves show local regression estimates for one standard deviation.

- Keinan, A., Mullikin, J.C., Patterson, N. & Reich, D. *Nat. Genet.* **41**, 66–70 (2009).
- Hammer, M.F., Mendez, F.L., Cox, M.P., Woerner, A.E. & Wall, J.D. *PLoS Genet.* **4**, e1000202 (2008).
- Bustamante, C.D. & Ramachandran, S. *Nat. Genet.* **41**, 8–10 (2009).
- Charlesworth, B., Morgan, M.T. & Charlesworth, D. *Genetics* **134**, 1289–1303 (1993).
- Smith, J.M. & Haigh, J. *Genet. Res.* **23**, 23–35 (1974).
- Pool, J.E. & Nielsen, R. *Evolution* **61**, 3001–3006 (2007).
- Vicoso, B. & Charlesworth, B. *Genetics* **181**, 1699–1701 (2009).
- Charlesworth, B. *Genet. Res.* **68**, 131–149 (1996).
- Hudson, R.R. & Kaplan, N.L. *Genetics* **141**, 1605–1617 (1995).
- Lu, J. & Wu, C.I. *Proc. Natl. Acad. Sci. USA* **102**, 4063–4067 (2005).
- Tapper, W. *et al. Proc. Natl. Acad. Sci. USA* **102**, 11835–11839 (2005).
- Wang, E.T., Kodama, G., Baldi, P. & Moyzis, R.K. *Proc. Natl. Acad. Sci. USA* **103**, 135–140 (2006).
- Betancourt, A.J., Kim, Y. & Orr, H.A. *Genetics* **168**, 2261–2269 (2004).
- Ross, M.T. *et al. Nature* **434**, 325–337 (2005).
- Payseur, B.A. & Nachman, M.W. *Mol. Biol. Evol.* **19**, 336–340 (2002).