Letter to the Editor

Gene Density and Human Nucleotide Polymorphism

Bret A. Payseur and Michael W. Nachman

Department of Ecology and Evolutionary Biology, Biosciences West Building, University of Arizona

Population genetics theory indicates that natural selection will affect levels and patterns of genetic variation at closely linked loci. Background selection (Charlesworth, Morgan, and Charlesworth 1993) proposes that the removal of recurrent deleterious mutations and associated neutral variants will cause a reduction of nucleotide variation in low-recombination regions. The strength of background selection depends on the deleterious mutation rate, the magnitude of selection and dominance, and the recombination rate. Genetic hitchhiking (Maynard Smith and Haigh 1974), the fixation of advantageous alleles and the associated fixation of linked neutral alleles, can also decrease nucleotide diversity in low-recombination regions. The extent of genetic hitchhiking depends on the strength of selection and the rate of recombination. Therefore, under both background selection and genetic hitchhiking, theory predicts that genomic regions that rarely recombine may be subject to reductions in nucleotide diversity. Furthermore, if the rate of deleterious mutation or selective sweeps (or both) is sufficiently high, background selection (Hudson and Kaplan 1995) and genetic hitchhiking (Wiehe and Stephan 1993) models predict an overall positive correlation between nucleotide polymorphism and recombination rate.

Empirical investigations of nucleotide variation support these predictions. In *Drosophila melanogaster*, regions of the genome with little recombination show reduced heterozygosity (Aguade, Miyashita, and Langley 1989; Begun and Aquadro 1991; Berry, Ajioka, and Kreitman 1991). Furthermore, there is evidence that nucleotide variation and recombination rate are positively correlated in several taxa, including fruit flies (Begun and Aquadro 1992), house mice (Nachman 1997), goatgrasses (Dvorak, Luo, and Yang 1998), sea beets (Kraft et al. 1998), tomatoes (Stephan and Langley 1998), humans (Nachman et al. 1998; Przeworski, Hudson, and Di Rienzo 2000; Nachman 2001), and maize (Tenaillon et al. 2001). The combination of theoretical and empirical results indicates that selection acting at linked sites is likely to be a major force shaping genomic patterns of nucleotide variation.

The documented relationship between nucleotide variation and recombination rate raises the question of whether other measurable variables can explain additional variation in nucleotide polymorphism in the context of selection at linked sites. We predict that the effects of selection at linked sites will depend on local gene density. If selection acts primarily on genes, genomic regions with high gene density will harbor more potential selective targets than genomic regions with low gene density. This prediction should be valid irrespective of whether positive or purifying selection is driving observed patterns. Humans provide a good system in which this prediction can be tested, for two reasons. First, gene density varies substantially across the genome (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). For example, sequence data suggest that chromosome 19 has an average of 23 genes per Mbp, whereas chromosome 4 averages only 6 genes per Mbp (Venter et al. 2001). Second, estimates of nucleotide polymorphism assessed using reasonable sample sizes are available for multiple loci across the human genome. Here, we demonstrate that nucleotide diversity and gene density are negatively correlated in humans. This result provides further evidence for the importance of selection at linked sites and suggests that the number of genes in a genomic region is a reasonable indicator of selective intensity.

We assessed the relationship between nucleotide polymorphism (measured by Watterson's $\hat{\theta}$ [1975]) and gene density using data from sequence-based studies of variation that sampled more than 10 chromosomes (table 1). The variance in $\hat{\theta}$ can be quite large with sample sizes smaller than 10 (Pluzhnikov and Donnelly 1996).

For X-linked loci, nucleotide diversity was multiplied by ⅘ to account for the fact that the effective population size of the X chromosome is ¾ of that of the autosomes (assuming a sex ratio of 1). Sequence-based maps of the human genome (http://www.ncbi.nlm.nih.gov/genome/guide/human/, June 2001 Version) were used to estimate the base pair position of each locus. Gene density was estimated by counting the number of genes in a window, including 1 Mbp of sequence on either side of each locus (http://www.ncbi.nlm.nih.gov/genome/guide/human/). Gene density estimates based on a 10-Mbp window gave similar results. Recombination rates were taken from Payseur and Nachman (2000), who compared the genetic and physical positions of microsatellites spaced at approximately 2-Mbp intervals. Recombination rates for X-linked loci were multiplied by ⅔ to correct for differences in population recombination rates. All variables were approximately normally distributed (visual inspection of histograms; Shapiro-Wilks goodness-of-fit test; $P > 0.05$). Additionally, the residuals from the regression of $\hat{\theta}$ on all variables were normally distributed ($P > 0.05$). All analyses were done using least-squares linear regression.

Nucleotide polymorphism and recombination rate are strongly, positively correlated ($R^2 = 0.63$; $P = 0.0002$; fig. 1*a*) for these data, despite no evidence for a positive relationship between divergence and recom-

Address for correspondence and reprints: Bret A. Payseur, Department of Ecology and Evolutionary Biology, Biosciences West Building, University of Arizona, Tucson, Arizona 85721.
E-mail: payseur@email.arizona.edu.

Mol. Biol. Evol. 19(3):336–340. 2002

Table 1
Data used for analyses in this study

Locus[a]	Chromosome	Band[b]	Sequence[c]	N[d]	L (bp)[e]	θ (%)[f]	cM/Mbp[g]	Genes/Mbp	Divergence (%)[h]	GC (%)	Reference
β-globin . . .	11	p15.5	Intron	349	2670	0.110	1.84	10.5	1.34	39.9	Harding et al. 1997
Lp1.	8	p22	Intron	142	9734	0.147	2.26	2.0	1.48	40.8	Clark et al. 1998
Hox B6 . . .	17	q21.3	Intergenic	210	1000	0.068	0.97	11.5	[i]	50.8	Deinard and Kidd 1999
Mc1r.	16	q24.3	Coding	242	951	0.104	1.92	11.0	1.58	63.3	Rana et al. 1999
Ace.	17	q23.1	Intron	22	24070	0.089	0.79	8.5	[i]	58.7	Rieder et al. 1999
Apoe.	19	q13.2	Intron	192	5491	0.069	1.37	17.0	1.18	59.5	Fullerton et al. 2000
OR	17	p13.3	Intron	66	4535	0.097	1.59	13.5	2.34	53.3	Gilad et al. 2000
Duffy	1	q22	Intron	34	2931	0.108	1.39	8.5	[i]	54.6	Hamblin and Di Rienzo 2000
22q11.2 . . .	22	q11.2	Intergenic	128	9901	0.132	1.04	6.0	1.35	46.0	Zhao et al. 2000
1q24.	1	q24	Intergenic	122	8991	0.060	1.54	5.0	0.62	31.5	Yu et al. 2001
Pdha1.	X	p22.2–22.1	Intron	35	4153	0.195	2.95	4.5	1.10	45.9	Harris and Hey 1999
Xq13.3. . . .	X	q13.3	Intergenic	69	10163	0.091	0.45	2.5	0.92	37.8	Kaessmann et al. 1999
Zfx	X	p21.3	Intron	336	1089	0.192	2.38	2.5	1.17	39.1	Jaruzelska et al. 1999
Dmd I44 . .	X	p21.2	Intron	41	3000	0.197	2.26	4.5	0.90	37.4	Nachman and Crowell 2000
Dmd I7 . . .	X	p21.2	Intron	41	2389	0.117	2.26	4.5	1.63	33.1	Nachman and Crowell 2000
Msn	X	q11.2–12	Intron	41	4622	0.060	0.38	3.5	0.80	42.3	[j]
Alas	X	p11.2	Intron	41	5125	0.043	0.43	6.5	0.63	42.5	[j]

[a] Loci are arranged chronologically and then alphabetically by reference. All studies of nucleotide diversity include individuals from sub-Saharan Africa, except Duffy and OR. Two loci for which nucleotide polymorphism estimates have been published were excluded from our analyses. The F9 locus (Harris and Hey 2001) is a statistical outlier in all analyses. As noted by Harris and Hey (2001), this locus may have experienced strong, recent selection, and hence may not be expected to be representative of loci with its recombination rate and gene density. A second locus (Alonso and Armour 2001), in chromosomal region 16p13.3, lies at the tip of the chromosome, making it difficult to reliably estimate recombination rate. Although we excluded this locus from our analyses, it appears to be the most polymorphic locus among those surveyed at the nucleotide level in humans and it is situated adjacent to a recombinational hotspot, with recombination rates more than ten times higher than the genomic average (Badge et al. 2000). Hence, data from this locus conform to patterns expected under models of selection at linked sites.

[b] Band indicates band position of each locus on the cytogenetic map (http://www.ncbi.nlm.nih.gov/genome/guide/human/).

[c] Sequence indicates whether the majority of surveyed sequence is from intergenic regions or introns. In some studies, a small amount of sequence from exons was also surveyed. Only MC1r includes a large percentage of coding sequence (100%).

[d] N is the number of chromosomes surveyed.

[e] L is the number of base pairs surveyed.

[f] Nucleotide diversity for X-linked loci was multiplied by 4/3 to correct for differences in effective population size.

[g] Recombination rate for X-linked loci was multiplied by 2/3 to correct for differences in population recombination rate.

[h] Divergence was estimated by comparing human and chimpanzee sequences.

[i] Divergence estimates were not available for these loci.

[j] M. W. Nachman et al. (unpublished data).

bination rate ($P > 0.05$). Comparing the residuals of the regression of nucleotide polymorphism on recombination rate with gene density reveals a significant negative association ($R^2 = 0.25$; $P = 0.04$; fig. 1*b*). As predicted, nucleotide polymorphism is reduced in regions with higher gene density, once recombination rate variation is taken into account. A model including both recombination rate and gene density as independent variables explains 68% (adjusted R^2; $P = 0.0001$; recombination rate: $P = 0.0001$; gene density: $P = 0.05$) of the variation in nucleotide polymorphism. There is weak evidence for a negative association between nucleotide polymorphism and gene density alone ($R^2 = 0.17$; $P = 0.10$). There is no evidence of a statistical interaction between recombination rate and gene density, although such an interaction would be difficult to detect with our small sample size. We also asked whether an alternative measure of nucleotide variation, the average pairwise divergence between sequences, $\hat{\pi}$ (Nei and Li 1979), is associated with gene density. There is a slight trend toward a negative relationship, but it is not statistically significant ($P > 0.05$ in all tests). $\hat{\theta}$ and $\hat{\pi}$ incorporate

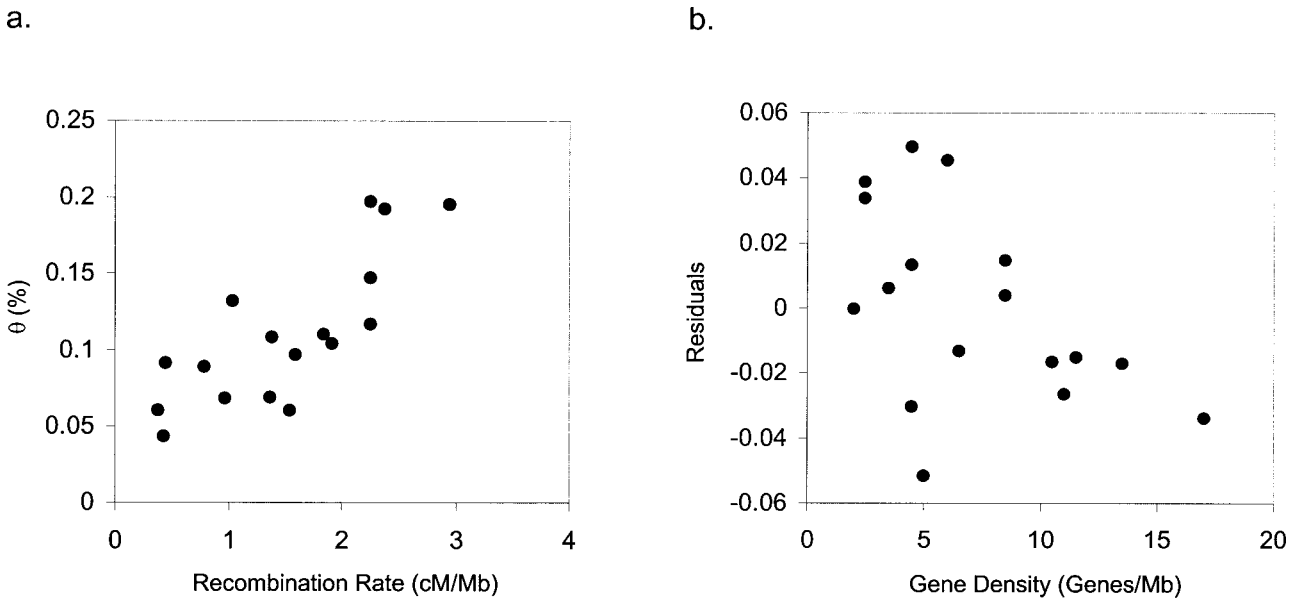

FIG. 1.—(*a*) The relationship between nucleotide polymorphism and recombination rate. Scatterplot of nucleotide polymorphism ($\hat{\theta}$, expressed as a percentage) versus recombination rate (cM/Mb) for 17 loci surveyed in humans. $R^2 = 0.63$; $P = 0.0002$. (*b*) The relationship between nucleotide polymorphism, corrected for variation in recombination rate, and gene density. Scatterplot of the residuals from a regression of nucleotide polymorphism on recombination rate versus gene density. $R^2 = 0.25$; $P = 0.04$.

different aspects of the data in their estimates of nucleotide diversity. Whereas $\hat{\theta}$ is estimated by counting the number of segregating sites in the total sample, $\hat{\pi}$ is estimated by comparing all the pairwise sequence combinations and calculating the average number of differences. As a result, $\hat{\pi}$ contains information about allele frequencies and $\hat{\theta}$ does not. However, $\hat{\theta}$ has a lower sampling variance than $\hat{\pi}$. Using the average number of sampled chromosomes ($n = 124$) and the average $\hat{\theta}$ or $\hat{\pi}$ value (approximately 0.1%) for the studies included in our analysis, under an infinite sites model with no recombination, the sampling variance of $\hat{\pi}$ (0.034%) is nearly twice that of $\hat{\theta}$ (0.019%). Although this effect may be ameliorated by recombination (Pluzhnikov and Donnelly 1996), the increased statistical difficulty in estimating $\hat{\pi}$ may contribute to our failure to detect an association between gene density and $\hat{\pi}$.

An alternative interpretation of our results is that nucleotide polymorphism is shaped by other variables that are correlated with gene density or recombination rate. GC content is positively correlated with both gene density (International Human Genome Sequencing Consortium 2001) and recombination rate (Fullerton, Bernardo Carvalho, and Clark 2001) in humans. Consequently, we asked whether GC content was associated with nucleotide polymorphism alone or once gene density and recombination rate had been taken into account. There is no evidence that GC content affects levels of polymorphism in these data ($P > 0.05$, bivariate and multiple linear regression analyses), although a weak correlation between SNP (single nucleotide polymorphism) heterozygosity and GC content in humans has been reported (International SNP Map Working Group 2001). This discrepancy may be because of the relatively small number of loci used in our study.

Several conclusions follow from these results. First, natural selection at the molecular level has a pronounced effect on the levels of nucleotide heterozygosity in humans. Even if the total number of sites under selection is relatively modest, it is clear that the effects on linked, neutral variation can be substantial. It remains to be seen whether the patterns depicted in figure 1 are driven mainly by positive selection and associated genetic hitchhiking, purifying selection, or some combination of both. Background selection and genetic hitchhiking are not mutually exclusive, and it seems likely that both processes may be contributing to observed patterns (Kim and Stephan 2000). Second, these results suggest that the density of genes is a reasonable indicator of the potential for selection and that genes are likely the targets of selection in many cases. However, the high degree of sequence conservation between human and mouse in intergenic regions suggests that many of these intergenic regions may also be functional, possibly containing important *cis*-regulatory elements (Shabalina et al. 2001). The degree to which the densities of genes and *cis*-regulatory elements covary is therefore an interesting question for further investigation. Finally, our results indicate that levels of human nucleotide polymorphism can be predicted with reasonable precision, given the knowledge about local recombination rate and gene density. Because recombination rate and gene density can now be measured throughout the human genome, this predictive ability could assist efforts to map genes underlying complex diseases.

Acknowledgments

We thank Bruce Walsh for helpful discussions. We also acknowledge the useful comments of Adam Eyre-Walker and two anonymous reviewers.

LITERATURE CITED

AGUADE, M., N. MIYASHITA, and C. H. LANGLEY. 1989. Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics **122**: 607–615.

ALONSO, S., and J. A. L. ARMOUR. 2001. A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. Proc. Natl. Acad. Sci. USA **98**:864–869.

BADGE, R. M., J. YARDLEY, A. J. JEFFREYS, and J. A. L. ARMOUR. 2000. Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. Hum. Mol. Genet. **9**:1239–1244.

BEGUN, D. J., and C. F. AQUADRO. 1991. Molecular population genetics of the distal portion of the X chromosome in *Drosophila*: evidence for genetic hitchhiking of the *yellow-achaete-scute* region. Genetics **129**:1147–1158.

———. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356**:519–520.

BERRY, A. J., W. AJIOKA, and M. KREITMAN. 1991. Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. Genetics **129**:1111–1117.

CHARLESWORTH, B., M. T. MORGAN, and D. CHARLESWORTH. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics **134**:1289–1303.

CLARK, A. G., K. M. WEISS, D. A. NICKERSON et al. (11 coauthors). 1998. Haplotype structure and population genetic inferences from nucleotide sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. **63**:595–612.

DEINARD, A., and K. KIDD. 1999. Evolution of a HOXB6 intergenic region within the great apes and humans. J. Hum. Evol. **36**:687–703.

DVORAK, J., M. C. LUO, and Z. L. YANG. 1998. Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing *Aegilops* species. Genetics **148**: 423–434.

FULLERTON, S. M., A. BERNARDO CARVALHO, and A. G. CLARK. 2001. Local rates of recombination are positively correlated with GC content in the human genome. Mol. Biol. Evol. **18**:1139–1142.

FULLERTON, S. M., K. M. WEISS, A. G. CLARK et al. (11 coauthors). 2000. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am. J. Hum. Genet. **67**: 881–900.

GILAD, Y., D. SEGRE, K. SKORECKI, M. W. NACHMAN, D. LANCET, and D. SHARON. 2000. Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nat. Genet. **26**:221–224.

HAMBLIN, M. T., and A. DI RIENZO. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am. J. Hum. Genet. **66**:1669–1679.

HARDING, R. M., S. M. FULLERTON, R. C. GRIFFITHS, J. BOND, M. J. COX, J. A. SCHNEIDER, D. S. MOULIN, and J. B. CLEGG. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60**:772–789.

HARRIS, E. E., and J. HEY. 1999. X chromosome evidence for ancient human histories. Proc. Natl. Acad. Sci. USA **96**: 3320–3324.

———. 2001. Human populations show reduced DNA sequence variation at the factor IX locus. Curr. Biol. **11**:774–778.

HUDSON, R. R., and N. L. KAPLAN. 1995. Deleterious background selection and recombination. Genetics **141**:1605–1617.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409**:928–933.

JARUZELSKA, J., E. ZIETKIEWICZ, M. BATZER, D. E. C. COLE, J. P. MOISAN, R. SCOZZARI, S. TAVARE, and D. LABUDA. 1999. Spatial and temporal distribution of the neutral polymorphisms in the last Zfx intron: analysis of haplotype structure and genealogy. Genetics **152**:1091–1101.

KAESSMANN, H., F. HEIBIG, A. VON HAESELER, and S. PAABO. 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat. Genet. **22**:78–81.

KIM, Y., and W. STEPHAN. 2000. Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics **155**:1,415–1,427.

KRAFT, T., T. SALL, I. MAGNUSSON-RADING, N. O. NILSSON, and C. HALDEN. 1998. Positive correlation between recombination rates and levels of genetic variation in natural populations of sea beet (*Beta vulgaris* subsp. *maritima*). Genetics **150**:1239–1244.

MAYNARD SMITH, J., and J. HAIGH. 1974. The hitch-hiking effect of a favourable gene. Genet. Res. **23**:23–35.

NACHMAN, M. W. 1997. Patterns of DNA variability at X-linked loci in *Mus domesticus*. Genetics **147**:1303–1316.

———. 2001. Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. **17**:481–485.

NACHMAN, M. W., V. L. BAUER, S. L. CROWELL, and C. F. AQUADRO. 1998. DNA variability and recombination rates at X-linked loci in humans. Genetics **147**:1133–1141.

NACHMAN, M. W., and S. L. CROWELL. 2000. Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy locus, *Dmd*, in humans. Genetics **155**:1855–1864.

NEI, M., and W.-H. LI. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76**:5269–5273.

PAYSEUR, B. A., and M. W. NACHMAN. 2000. Microsatellite variation and recombination rate in the human genome. Genetics **156**:1285–1298.

PLUZHNIKOV, A., and P. DONNELLY. 1996. Optimal sequencing strategies for surveying molecular genetic diversity. Genetics **144**:1247–1262.

PRZEWORSKI, M., R. R. HUDSON, and A. DI RIENZO. 2000. Adjusting the focus on human variation. Trends Genet. **16**: 296–302.

RANA, B. K., D. HEWETT-EMMETT, L. JIN et al. (12 co-authors). 1999. High polymorphism at the human melanocortin 1 receptor locus. Genetics **151**:1547–1557.

RIEDER, M. J., S. L. TAYLOR, A. G. CLARK, and D. A. NICKERSON. 1999. Sequence variation in the human angiotensin converting enzyme. Nat. Genet. **22**:59–62.

SHABALINA, A. S., A. Y. OGURTSOV, V. A. KONDRASHOV, and A. S. KONDRASHOV. 2001. Selective constraint in intergenic regions of human and mouse genomes. Trends Genet. **17**: 373–376.

STEPHAN, W., and C. H. LANGLEY. 1998. DNA polymorphism in *Lycopersicon* and crossing-over per physical length. Genetics **150**:1585–1593.

TENAILLON, M. I., M. C. SAWKINS, A. D. LONG, R. L. GAUT, J. F. DOEBLEY, and B. S. GAUT. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea*

mays ssp. *Mays* L.). Proc. Natl. Acad. Sci. USA **98**:9161–9166.

VENTER, J. C., M. D. ADAMS, E. W. MYERS et al. 2001. (274 co-authors). The sequence of the human genome. Science **291**:1304–1351.

WATTERSON, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**:256–276.

WIEHE, T. H. E., and W. STEPHAN. 1993. Analysis of a genetic hitchhiking model and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol. Biol. Evol. **10**:842–854.

YU, N., Y.-X. FU, N. SAMBUUGHIN, M. RAMSAY, T. JENKINS, E. LESKINEN, L. PATTHY, L. B. JORDE, T. KUROMORI, and W.-H. LI. 2001. Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. **18**:214–222.

ZHAO, Z., L. JIN, Y.-X. FU et al. (13 co-authors). 2000. Worldwide DNA sequence variation in a 10-kb noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA **97**:11354–11358.

ADAM EYRE-WALKER, reviewing editor

Accepted October 8, 2001