Positive and Negative Selection on the Human Genome

Justin C. Fay,* Gerald J. Wyckoff*,[1] and Chung-I Wu*,†

*Committee on Genetics, †Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

Manuscript received December 10, 2000
Accepted for publication April 5, 2001

ABSTRACT

The distinction between deleterious, neutral, and adaptive mutations is a fundamental problem in the study of molecular evolution. Two significant quantities are the fraction of DNA variation in natural populations that is deleterious and destined to be eliminated and the fraction of fixed differences between species driven by positive Darwinian selection. We estimate these quantities using the large number of human genes for which there are polymorphism and divergence data. The fraction of amino acid mutations that is neutral is estimated to be 0.20 from the ratio of common amino acid (*A*) to synonymous (*S*) single nucleotide polymorphisms (SNPs) at frequencies of $\geq$15%. Among the 80% of amino acid mutations that are deleterious at least 20% of them are only slightly deleterious and often attain frequencies of 1–10%. We estimate that these slightly deleterious mutations comprise at least 3% of amino acid SNPs in the average individual or at least 300 per diploid genome. This estimate is not sensitive to human population history. The *A*/*S* ratio of fixed differences is greater than that of common SNPs and suggests that a large fraction of protein divergence is adaptive and driven by positive Darwinian selection.

WHILE the fixation of adaptive mutations may be viewed as the crux of Darwinian evolution, it has long been argued that the majority of DNA changes that accumulate over time are not adaptive but neutral, fixed by stochastic fluctuations in a finite population (Kimura 1983). Indeed, except for a few proteins with extremely high rates of evolution, evidence for adaptive evolution at the molecular level has been elusive (Nei 1987). Much more is known about deleterious mutations and H. J. Muller argued in an influential address that the reduction in mean population fitness due to these mutations may constitute a considerable human health concern (Muller 1950).

The proportion of mutations that are deleterious has been estimated from both allozyme and DNA divergence data. Negative selection prevents deleterious mutation from reaching common frequencies and so should produce an excess of rare variation. In humans, the number of rare (<0.5%) allozyme alleles is much greater than expected under neutrality in an equilibrium population (Kimura 1983). The difference in the number of rare *vs.* common alleles was used to estimate that 79–85% of amino acid-altering mutations are deleterious (Kimura 1983). However, a recent increase in human population size can also account for the excess of rare variants. Negative selection also lowers the ratio of amino acid to synonymous divergence between populations and this ratio can be used to estimate the proportion of amino acid-altering mutations that are deleterious. Divergence of 46 genes among hominid species was used to estimate that 38% of amino acid-altering mutations are deleterious (Eyre-Walker and Keightley 1999). However, this may be an underestimate if positive selection has increased the rate of amino acid divergence.

The fitness effects of deleterious mutations are pertinent to public health as they determine the frequency of the deleterious mutations and consequently the number of individuals affected by a particular mutation. In the extreme case, if all deleterious mutations cause embryonic lethality and spontaneous abortion, then mutations that survive to be polymorphic in a population are either neutral or advantageous. Standing polymorphism, therefore, would not be of health concern. The increased rate of stillbirths and prereproductive mortality with inbreeding reveals substantial effects of partially recessive deleterious mutations segregating in the population (Morton and Crow 1956). However, it is not known whether most deleterious mutations an individual carries are rare and unique to one or a few individuals or quite common in the general population. This distinction is relevant to association studies and our understanding of complex human genetic diseases.

The proportions of neutral, deleterious, and adaptive mutations and their selection coefficients can be estimated by various methods from variation within and between species (Kimura 1983; Watterson 1987; McDonald and Kreitman 1991; Sawyer and Hartl 1992; Templeton 1996; Akashi 1999). If all detectable amino acid and synonymous polymorphism is neutral, the ratio of amino acid to synonymous variation within species should remain constant regardless of a population's demo-

Corresponding author: Chung-I Wu, Department of Ecology and Evolution, University of Chicago, Zoology Bldg., 1101 E. 57th St., Chicago, IL 60637. E-mail: ciwu@uchicago.edu

[1] *Present address:* Department of Human Genetics, University of Chicago, Chicago, IL 60637.

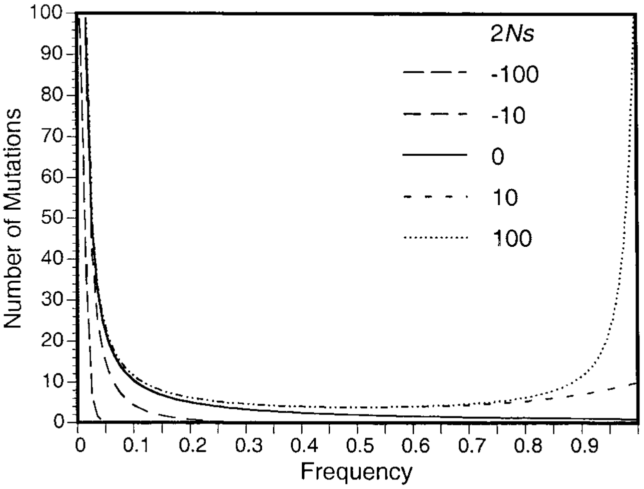

Figure 1.—The expected number of mutations found in a population as a function of frequency, where each mutation is either under positive ($2Ns > 1$), negative ($2Ns < -1$), or no selection ($2Ns = 0$). The distribution is obtained from an infinite-site model with free recombination, no dominance, $\theta = 4N\mu = 1$, where N is the effective population size, s is the selection coefficient, and μ is the mutation rate (materials and methods).

graphic history. However, this ratio should change as a function of frequency if some amino acid variation is under positive or negative selection (Figure 1). At allele frequencies on the order of the reciprocal of the effective population size, selection is ineffective and the relative proportions of neutral, deleterious, and adaptive variants correspond to their production by mutation. The relative number of deleterious to neutral mutations declines as a function of frequency in the population. Advantageous mutations, on the other hand, become enriched relative to neutral mutations in the high frequency portion of the distribution and in fixed differences between species. The effects of positive selection can be distinguished from negative selection only if an outgroup is used to infer whether a mutation is at, say, 5 or 95%. Many polymorphism studies do not make this distinction and so an excess of rare (low and high) compared to common single nucleotide polymorphisms (SNPs) cannot be attributed to just positive or negative selection. To estimate the fraction of DNA variation within and between species that has been under positive and negative selection we compare amino acid and synonymous polymorphism from two recent surveys of human DNA variation (Cargill *et al.* 1999; Halushka *et al.* 1999) and from divergence between humans and old world monkeys.

MATERIALS AND METHODS

Data: SNP-I is of 106 genes (~136 kb) surveyed in an average of 114 chromosomes (Cargill *et al.* 1999), and SNP-II is of 75 genes (~87 kb) surveyed in an average of 148 chromosomes (Halushka *et al.* 1999). The minor alleles of the SNP-I data were classified by the investigators as <5, 5–15, and 15–50% and correspond to occurrences of 1–5, 6–17, and 18–57 times out of 114 chromosomes. The SNP-II data were categorized by the original investigators first into population-specific and shared SNPs and then by the frequency of the minor allele in the population in which each SNP is found. To categorize the SNP-II data by total population frequency, African-specific, European-American-specific, and SNPs shared by both populations were split into low, moderate, and common frequency classes that roughly correspond to the three SNP-I frequency classes. For instance, the African-specific SNPs in the 1.25–6.25% class (1–5 occurrences in a sample of 80 African chromosomes) are placed into the low frequency class, which is 0.7–3.4% in the total population. Low, moderate, and common frequency classes respectively include the following: African-specific SNPs that occur 1–5, 6–9, and 10–40 times in a sample of 80; non-African-specific SNPs that occur 1–5, 6–8, and 9–34 times in a sample of 68; and shared SNPs that occur 0, 1–11, and 12–74 times in the total sample of 148. A portion of the low frequency class was removed (49 amino acid, 16 synonymous, and 38 noncoding SNPs) as compensation for false positive SNPs, estimated to be 17% of all SNPs identified (Halushka *et al.* 1999). No compensation was made for SNP-I because every SNP was confirmed.

The divergence data are from 182 orthologous human and old world monkey genes (~183 kb). The number of amino acid and synonymous differences between orthologs was corrected for multiple hits using Kimura's two-parameter model (Li 1997). The proportion of amino acid-altering sites in coding regions was calculated from simulation, and for the divergence data is 0.725 (Wyckoff *et al.* 2000). The divergence data can be obtained at http://ecolevol-mac2.uchicago.edu/supp.html. The proportion of amino acid-altering sites in the two SNP surveys was calculated as 1 minus the number of fourfold degenerate sites plus one-half the number of twofold degenerate sites divided by the total number of sites and is 0.714 and 0.757 for the SNP-I and SNP-II surveys, respectively. The slight differences between these numbers do not affect any conclusions.

Deleterious mutations model: We assume all amino acid mutations are either neutral, slightly deleterious, or strongly deleterious. The neutral class, f_0, is estimated from common SNPs to be 0.20 (see results). The slightly deleterious class, f_1, must account for the excess of low frequency amino acid SNPs (Table 1). The strongly deleterious class, f_2, is not found in a sample. The excess of low and moderate frequency amino acid SNPs is fit to a deleterious mutation model. The model assumes a randomly mating population of constant size, that each new mutation occurs in a monomorphic site (infinite-site model), free recombination (independence between all sites), and that each SNP is deleterious and reduces an individual's fitness by s in the homozygous state and hs in the heterozygous state, where s and h are the selection and dominance coefficients, respectively. The expected frequency distribution of deleterious SNPs in a population, or the frequency spectrum, is

$$\phi(x) = \frac{f_1\theta\int_x^1 \psi(y)\,dy}{x(1-x)\psi(x)\int_0^1 \psi(y)\,dy} \tag{1}$$

$$\psi(y) = \exp[2Ns(2h-1)y^2 - 2hy] \tag{2}$$

(Equation 4.24; Ewens 1979), where x is the frequency of a deleterious SNP, $\theta = 4N\mu$, N is the effective population size, μ is the mutation rate per generation, and f_1 is the fraction of amino acid SNPs that are slightly deleterious. The number of slightly deleterious SNPs in a population is a fraction, f_1, of the total number of amino acid SNPs expected in the absence of negative selection and can range from zero, in which case all amino acid SNPs are either neutral or strongly deleterious, to 0.80, in which case 20% of amino acid SNPs are neutral and 80% are slightly deleterious. θ was estimated from the number of synonymous SNPs at low and moderate frequency to be 87.10 using the formula

$$\theta = S\left(\sum_{i=1}^{17}\frac{1}{i} + \sum_{i=97}^{113}\frac{1}{i}\right)^{-1}\alpha, \qquad (3)$$

where S is the number of low and moderate frequency synonymous SNPs, α is the number of amino acid-altering sites per synonymous site, and the denominator is similar to that of Watterson's estimator of θ, except that it accounts for the fact that the estimate is based on segregating sites where the minor allele occurs 1–17 times in a sample (Watterson 1975; Fu 1995).

The expected number of deleterious SNPs in a sample was found by taking the integral of the product of the frequency spectrum, $\phi(x)$, and the sampling formula,

$$a(i,\ n) = \frac{n!}{(n - i)!\,i!}x^{i}(1 - x)^{n - i} \qquad (4)$$

over all possible values of x, where i is the number of occurrences of an allele in a sample of n chromosomes. We integrated from $1/2N$ to $1 - 1/2N$, where N was assumed to be 10^6. The results are the same for $N = 10^4$ or 10^7. We also consider a model where each deleterious SNP has a different selection coefficient and these selection coefficients are exponentially distributed. Under this model the expected number of deleterious SNPs in a sample is

$$\int_{0}^{\infty}\int_{1/2N}^{1 - 1/2N}\phi(x)\,a(i,\ n)\lambda e^{-\lambda s}dxds, \qquad (5)$$

where λ is $1/\bar{s}$, $\bar{s}$ is the average selection coefficient, and there is no dominance ($h = 0.5$).

To assess the fit of the model, the numbers of deleterious SNPs predicted to be found 1–5 times (<5%), 6–17 times (5–15%), and >17 times (>15%) in the sample were generated using the appropriate sampling formula. The number of sites for which an individual is expected to be heterozygous and homozygous for a deleterious mutation was found by numerical integration of $\phi(x)x^2$ and $\phi(x)2x(1 - x)$, respectively.

RESULTS

Detection of deleterious mutations: The numbers of amino acid (A) and synonymous (S) single nucleotide polymorphisms (SNPs) found in two published surveys of human DNA variation are shown in Table 1, hereafter referred to as SNP-I (Cargill *et al.* 1999) and SNP-II (Halushka *et al.* 1999). The two surveys differ in a number of respects and so are analyzed separately (materials and methods). In the SNP-I survey, there is a significant difference in the ratio of amino acid to synonymous SNPs in the low (<5%) compared to common (>15%) frequency class ($\chi^2 = 7.39$, $P < 0.01$). In the SNP-II survey, there is a significant difference in the ratio of amino acid to synonymous SNPs in the low ($\chi^2 = 8.97$, $P < 0.01$) and moderate ($\chi^2 = 4.74$, $P < 0.05$) compared to common frequency class. The different ratios of amino acid to synonymous SNPs between frequency classes suggest that a large fraction of the low frequency amino acid SNPs are deleterious (Cargill *et al.* 1999; Halushka *et al.* 1999). The low ratio of noncoding (NC) to synonymous SNPs indicates 50% of the noncoding sites surveyed are under constraint (Table 1). However, noncoding SNPs appear neutral in that their distribution is similar to that of synonymous variation. This suggests that amino acid and noncoding mutations differ in their distribution of selection coefficients (Sunyaev *et al.* 2000).

The difference in the per site ratio of amino acid to synonymous SNPs at low compared to common frequency is an estimate of the fraction of low frequency SNPs that are slightly deleterious. To make this estimate, amino acid mutations are assumed to fall into three classes: neutral, f_0; slightly deleterious, f_1; and strongly deleterious, f_2. Neutral amino acid mutations are responsible for all common SNPs as well as a proportion of low and moderate frequency SNPs. Slightly deleterious mutations account for the excess of low frequency amino acid variation and are prevalent in the general population. Strongly deleterious mutations rarely rise to detectable frequency and may be associated with simple genetic diseases. Adaptive mutations are assumed to be rare in polymorphism (but not divergence). The observed distribution of amino acid SNPs is a composite of these three classes. Although synonymous and common amino acid SNPs may not all be neutral, this would result in an underestimate of the deleterious fraction. In addition, this measure is independent of any demographic effects, which should influence both the distribution of neutral amino acid and synonymous SNPs and should thus have no effect on their ratio.

The ratio of amino acid to synonymous SNPs is shown in Table 1 as A^*/S^* for each frequency class and survey (* denotes the number of SNPs per site). The combined A^*/S^* ratio of common SNPs is 0.20, indicating that the majority (80%) of amino acid SNPs are deleterious. The A^*/S^* ratio of low frequency SNPs is higher than that of common SNPs and indicates a large fraction of the amino acid SNPs are slightly deleterious and reach only low frequencies. For SNP-I, the ratio increases from 0.23 to 0.46, which suggests that 23% of amino acid mutations are slightly deleterious. The other 54% of mutations are strongly deleterious and not found in the population sample. The excess number of amino acid SNPs at low frequency is the observed minus expected number of amino acid SNPs, calculated as the product of the A/S ratio of common SNPs and the number of synonymous SNPs at low or moderate frequency (Table 1).

We are interested in estimates of the number of deleterious mutations an individual is expected to carry. Such an estimate requires knowledge of the frequency of these mutations in the population. A conservative estimate is to assume each deleterious mutation is present only once in the sample. This estimate is not dependent on any population genetic model and is conservative because some of the deleterious mutations may be

TABLE 1

Summary of two DNA polymorphism surveys in humans and divergence from old world monkeys

Data	Class	*A*	*S*	*A**/*S**	Excess	NC	NC*/*S**
SNP-I	Low	93	81	0.46	47	52	0.41
	Moderate	26	38	0.27	4	35	0.59
	Common	36	63	0.23	0	49	0.50
	Total	155	182	0.34	51	136	0.48
SNP-II	Low	85	66	0.41	47	142	0.44
	Moderate	40	35	0.37	20	77	0.45
	Common	34	59	0.19	0	179	0.62
	Total	159	160	0.32	67	398	0.51
Combined	Common	70	122	0.20	0	228	0.69
Divergence	Total	3660	4151	0.34	1278		

A, amino acid SNP; *S*, synonymous SNP; *A**/*S**, the ratio of amino acid to synonymous SNPs per amino acid-altering and synonymous site, respectively; excess, observed minus expected number of amino acid SNPs, where expected is the product of *S* (low or moderate) and *A*/*S* (common); NC, noncoding SNP; and NC*/*S**, the per site ratio of noncoding to synonymous SNPs. Low, moderate, and common classes of the minor allele frequency are <5%, 5–15%, and >15%, respectively, for the SNP-I survey (see MATERIALS AND METHODS for a description of the SNP-II survey).

present more than once in the sample. Indeed, in both surveys the excess is not limited to the lowest frequency class. The conservative estimate is the excess divided by the number of individuals sampled and is 0.89 deleterious mutations per individual for both the SNP-I and SNP-II surveys (Table 2). These deleterious SNPs constitute 3.4 and 2.5% of the average number of amino acid SNPs found in an individual in the SNP-I and SNP-II surveys, respectively (Cargill *et al.* 1999; Halushka *et al.* 1999). Assuming there are 5 × 10^7 coding sites in the genome (~1.5% of the genome), the SNP-I and SNP-II surveys represent 1/368 and 1/575 of coding DNA in the genome and the conservative estimates of the average number of deleterious mutations an individual is expected to carry are 329 and 513, respectively (Table 2).

The frequency distribution (or frequency spectrum) of deleterious mutations in a population and the expected number an individual carries can also be assessed by comparing shared and population-specific SNPs. Deleterious mutations are kept at low frequency and should not easily spread to multiple populations. Figure 2 shows that amino acid SNPs are much more likely to be population specific than synonymous SNPs. In SNP-II, the *A*/*S* ratio of African- and European-American-specific SNPs is 135/111 and is significantly greater than 24/49, the *A*/*S* ratio of SNPs shared between the two populations ($\chi^2 = 10.9$, $P < 0.001$). If all amino acid SNPs were neutral the ratio of amino acid to synonymous SNPs should be the same for population-specific compared to shared SNPs, regardless of population history. Based on the ratio of shared amino acid and synonymous SNPs and the number of population-specific synonymous SNPs, the excess of population-specific amino acid SNPs is 80 (135 − 111 × 24/49). Using the most conservative estimate, which assumes that the deleterious mutations

TABLE 2

Conservative estimates of the number of deleterious SNPs in an individual's genome

Data	*D* (sample)	*T*	*D* (genome)	*D*/*T* (%)
SNP-I	0.89	26.62	329	3.4
SNP-II[a]	0.89	36.17	513	2.5
SNP-II[b]	1.09	36.17	627	3.0

D is the number of deleterious SNPs carried by an individual in the genes surveyed and is calculated as the excess of amino acid SNPs in the sample divided by the number of individuals sampled. *T* is total number of amino acid SNPs carried by an individual in the genes surveyed (Cargill *et al.* 1999; Halushka *et al.* 1999).

[a] The excess is calculated from frequency classes.

[b] The excess is calculated from population classes.

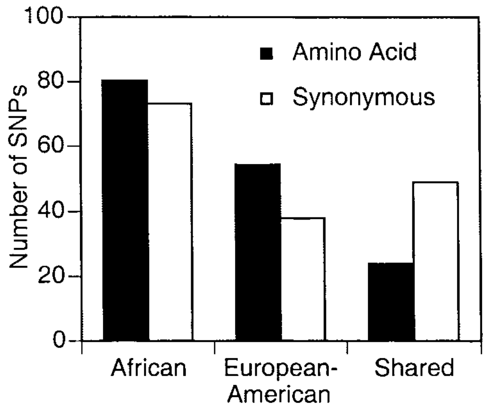

Figure 2.—The numbers of amino acid and synonymous SNPs that are population specific and shared between the European-American and African sample.

are found only once in the sample, an individual is expected to carry 627 deleterious mutations in his genome. This estimate is quite similar to that based on the frequency analysis, 513, and indicates most deleterious SNPs are population specific. However, both estimates are conservative in that all the deleterious SNPs were assumed to be found only once in the sample. A better estimate can be made by knowing the frequency distribution of deleterious SNPs in the population.

Fit to a deleterious mutation model: To estimate the frequency spectrum of deleterious SNPs in the population, the SNP-I survey is fit to an infinite-site deleterious mutation model (Materials and Methods). The SNP-II survey is not analyzed as it is categorized primarily by population rather than frequency. The model is fit to the excess of low and moderate frequency amino acid SNPs inferred to be in the f_1 or slightly deleterious class. The neutral fraction, f_0, is responsible for all common SNPs and is 0.20 from the combined A^*/S^* ratio of common SNPs (Table 1). Each mutation in the slightly deleterious class reduces fitness by s in the homozygous state and hs in the heterozygous state, where s and h are the coefficients of selection and dominance, respectively. The number of deleterious SNPs in a sample is determined by the strength of selection and the fraction of amino acid mutations that are slightly deleterious, f_1. As the strength of selection increases, fewer amino acid SNPs are found at a detectable frequency (>1%), so the proportion of mutations that are slightly deleterious, f_1, must increase to produce the same number of deleterious SNPs observed at low frequency. As selection becomes weaker, deleterious SNPs are found at higher frequencies and the excess of low frequency compared to common amino acid SNPs is lost. The model is fit such that 51 deleterious SNPs (the excess) are expected in a sample of 114 with three constraints: (i) the slightly deleterious fraction, f_1, cannot exceed 0.80 since 20% of mutations were estimated to be neutral; (ii) the predicted number of common deleterious SNPs (>15%) is not more than one SNP; and (iii) there is at least one slightly deleterious SNP predicted at moderate frequencies (5–15%).

The range of parameter values that can explain the observed excess of amino acid variation is presented in Table 3. The average number of deleterious mutations per individual predicted by the model lies in the narrow range of 500–1200 regardless of the strength of selection, degree of dominance, or proportion of mutations that are slightly deleterious, f_1. In the absence of the third constraint the estimated number of deleterious mutations approaches the conservative estimate of 329. The slightly deleterious fraction (f_1, fifth column) suggests that 20–45% of amino acid mutations are slightly deleterious and reach appreciable frequencies in the population. The fitted distribution shows that most (>60%) of the deleterious SNPs an individual carries are expected to be at a frequency of >1% in the population even though the majority (>80%) of all slightly deleterious SNPs in the population are expected to be at a frequency of <1% (Materials and Methods).

TABLE 3

Fit of a deleterious mutation model to the excess of amino acid SNPs

Parameters		Fit			Mutations	
h	$2Ns$	<5%	5–15%	f_1	Ho	He
0.5	−20	44	7	0.29	27	972
	−50	49	2	0.46	6	624
0.01	−100	41	10	0.21	33	1139
	−200	44	7	0.25	18	906
	−500	48	3	0.33	8	692
	−1000	50	1	0.42	5	579
0.5	−300[a]	46	4	0.80	40	847
Conservative		51	0		1	329

h, dominance coefficient; $2Ns$, the scaled selection coefficient where s is the selection coefficient and N is the effective population size; <5% and 5–15%, the number of deleterious mutations the model predicts at low and moderate frequency, respectively; f_1, the estimated proportion of amino acid mutations that are slightly deleterious; Ho and He, the estimated number of sites homozygous and heterozygous, respectively, for a deleterious mutation in an individual.

[a] The average strength of selection when s is exponentially distributed.

Given the large range of selection coefficients that can explain the deleterious excess, it may be appropriate to fit the excess to a model that assumes each deleterious SNP has a different selection coefficient. To do this, the excess is fit to an infinite-site deleterious mutation model, which assumes selection coefficients are exponentially distributed and there is no dominance (Materials and Methods). In this model there are two classes of amino acid mutations: a neutral class of amino acid mutations, f_0, and a deleterious class, f_1, which is fixed at 0.80 since the class must account for all deleterious amino acid mutations. An average scaled selection coefficient ($2Ns$) of −300 provides a good fit to the observed number of low and moderate frequency amino acid SNPs (Table 3).

Detection of advantageous mutations: The other major focus of this study is on how often advantageous mutations occur. Under the neutral theory, the small number of advantageous substitutions is overwhelmed by the large number of neutral substitutions between species. Positive selection is expected to increase the number of high frequency compared to common SNPs (Figure 1). This effect should be stronger for A than for S; S can also rise to high frequency by hitchhiking with mutations under positive selection (Fay and Wu 2000). For SNP-I, low *vs.* high frequency SNPs were distinguished using an outgroup. The ratio of common (15–85%) amino acid to synonymous SNPs (28/38) is lower than the ratio (7/7) at high frequency (>85%).

While many more high frequency SNPs would be needed to show a significant increase in the A/S ratio, analysis of flanking variation for a hitchhiking effect would provide direct evidence for positive selection (Fay and Wu 2000).

The A/S ratio of divergence is also inflated by positive selection. A common test for positive selection is a comparison of the A/S ratio of polymorphism and divergence (McDonald and Kreitman 1991; Sawyer and Hartl 1992; Templeton 1996). Since mutations under positive selection spread through a population quickly, they are not well represented in polymorphism but should have a cumulative effect on divergence. The A/S ratio from divergence is estimated from 182 orthologous human and old world monkey genes (Table 1). To avoid the confounding effects of deleterious mutations, which do not contribute to divergence but do make a significant contribution to polymorphism, the A/S ratio from divergence is compared to that of common SNPs (Table 1). The difference in the A/S ratio of common SNPs combined from both surveys compared to divergence is significant ($\chi^2 = 8.14$, $P < 0.01$) and can be explained by positive selection, assuming the average constraint on the divergence and polymorphism genes is the same.

The large number of amino acid substitutions suggests a high rate of adaptive evolution in primates. The expected number of amino acid substitutions is 2382 ($4151 \times 70/122$) based on the A/S ratio of common polymorphism and the excess is 1278. Therefore, a large proportion, 35%, of amino acid substitutions between humans and old world monkeys are estimated to have been driven by positive selection. Extrapolating this proportion to the total amount of coding DNA in the genome ($\sim 5 \times 10^7$ bp) yields an estimate of up to 1 advantageous substitution every ~200 years since humans separated from old world monkeys 30 million years ago (Li 1997).

If the proportion of deleterious, neutral, and adaptive mutations differs among common and low frequency variation and fixed differences, this may be reflected in the types of amino acid changes that occur. Amino acid substitutions were classified into conservative, moderate, and radical changes on the basis of their physicochemical properties (Grantham 1974; Wyckoff *et al.* 2000). As expected, radical changes show the largest decrease from low frequency to common SNPs and so are more likely deleterious (Figure 3). Interestingly, moderate changes show the largest increase from the common SNP class to fixed differences and so are more likely to have been influenced by positive selection.

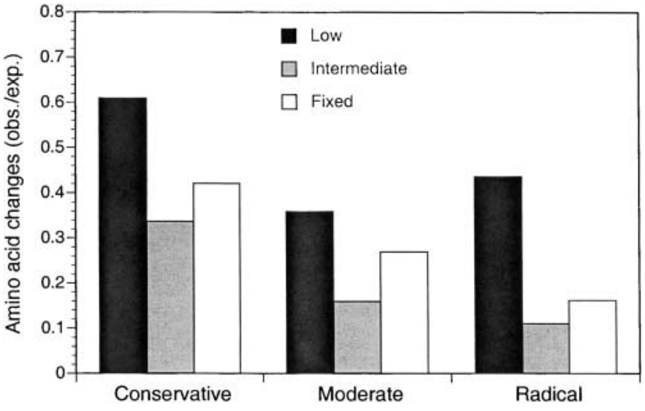

Figure 3.—The observed to expected ratio of conservative, moderate, and radical amino acid-altering mutations found at low and common frequency and fixed between species.

DISCUSSION

Segregating amino acid mutations with a strong phenotype can be detected through studies of linkage and association. In the absence of a strong phenotype it is difficult to determine which amino acid SNPs in a population are deleterious or adaptive and which are neutral. However, weak selection should have a significant impact on the distribution of amino acid variation within and between species and this can be used to estimate the fractions of amino acid variation that are deleterious, neutral, and adaptive. To estimate these quantities we make a number of assumptions that, as we discuss below, are conservative and should not affect our conclusions.

The fraction of amino acid mutations that are neutral is estimated from the A^*/S^* ratio of common SNPs by assuming synonymous mutations and common amino acid SNPs are neutral. If violated, this assumption leads to an overestimate of the neutral fraction and an underestimate of the slightly deleterious and advantageous fraction. For example, when $2Ns$ is -5, 14% of the deleterious SNPs in a sample of 114 are expected to be found at a frequency of >15% in the sample (integration of Equation 1). Evidence for selection on synonymous sites has been inferred from a lower frequency of mutations from G or C to A or T at synonymous sites compared to mutations from A or T to G or C (Eyre-Walker 1999).

The fraction of amino acid mutations that are deleterious is estimated to be 0.80 from the neutral fraction and is not much affected by advantageous mutations, which if under strong selection rarely affect polymorphism but accumulate as divergence. A previous estimate of the deleterious fraction was 0.38 based on divergence of 46 genes (Eyre-Walker and Keightley 1999). The difference between these two estimates is likely due to two factors: (i) the assumption that positive selection does not increase the rate of protein divergence and (ii) the relatively small number of genes available for use in the previous analysis.

The estimated number of deleterious mutations an individual carries is not likely inflated by a complex demographic history. The conservative estimate is not affected by population history but may be overly conservative as it assumes all deleterious amino acid SNPs are found only once in the population. The deleterious mutation model assumes an equilibrium population but the estimate is not likely inflated for the following reasons. First, the number of deleterious mutations an individual carries depends most on those mutations >1% in frequency and so should not be affected much by the distribution of deleterious mutations below 1%. While the distribution below 1% may be affected by a recent increase in population size, the distributions of synonymous SNPs in the low, moderate, and common classes are 81, 38, and 63, and are very close to those expected in a neutral equilibrium population: 80, 44, and 58, respectively (Fu 1995). Second, the estimated number of deleterious mutations per individual depends most on the excess to which it is fit and not the actual shape of the distribution. This is shown by the similar estimates obtained for different values of *s* and *h*, which determine the shape of the distribution (Table 3).

The deleterious mutation model may underestimate the strength of selection. Changes in population size, interference between selected mutations, and fluctuating selection coefficients reduce the efficacy of selection (Gillespie 1991; Barton 1995; McVean and Charlesworth 2000). These processes enable a deleterious mutation to attain a higher frequency than in their absence.

The fraction of amino acid substitutions that were driven by positive selection is based on the assumption that the average level of constraint on the genes in the polymorphism and divergence data is the same. Further work is needed to test whether this assumption is robust. Regardless, it is clear that an excess of amino acid divergence can be obscured by an increase in the *A*/*S* ratio of polymorphism due to slightly deleterious amino acid SNPs. In this study, the *A*/*S* ratio of total polymorphism is inflated to that of divergence and so no evidence of positive selection would have been detected using the standard McDonald-Kreitman test, which compares the *A*/*S* ratio of total polymorphism to divergence (McDonald and Kreitman 1991).

The number of deleterious mutations an individual carries and their individual selection coefficients cannot be used in a straightforward estimate of fitness reduction or genetic load. A genomic deleterious mutation rate greater than one suggests epistatic interactions between deleterious mutations in their translation to fitness (Crow 1970). The genomic deleterious mutation rate in humans was previously estimated to be at least 1.6 on the basis of an estimate that 38% of amino acid mutations are deleterious. The genomic deleterious mutation rate is likely much larger given our estimate that 80% of amino acid mutations are deleterious and given that it does not include deleterious mutations in noncoding regions, which may be quite common (Shabalina and Kondrashov 1999). The combined NC*/*S** ratio of common SNPs from both surveys indicates 50% of the noncoding sites are constrained and must serve some function (Table 1). Because an equal number of noncoding and amino acid-altering sites were surveyed, noncoding mutation should contribute at least 60% (0.50/0.80) as much as coding mutations to the total genomic deleterious mutation rate.

A contentious issue is whether most common human genetic diseases are caused by a few common frequency SNPs or the combined effects of many low frequency SNPs (Cargill *et al.* 1999; Halushka *et al.* 1999). Although the large number of low frequency slightly deleterious SNPs is suggestive of numerous factors of small effect, their contribution to human genetic diseases or health ailments is not necessary. Many of these deleterious mutations may no longer affect fitness due to environmental changes, and fitness in the evolutionary sense does not necessarily equate with common human genetic diseases. The fitness cost in the evolutionary sense may be manifested through parental care or sexual selection (Wyckoff *et al.* 2000) or may be through resistance to a previous pathogen or environmental condition. Nevertheless, the increase in the occurrence of human genetic diseases with inbreeding suggests there are many segregating deleterious mutations of health concern (Bittles and Neel 1994). Whether slightly or strongly deleterious mutations account for these observations has yet to be determined.

We thank N. Cox, J. Crow, S. Dorus, A. Kondrashov, T. Nagylaki, and K. Thornton for comments and discussions. We also thank the reviewers for their helpful comments. This work was supported by a National Science Foundation and National Institutes of Health grant to C.-I Wu and a National Institutes of Health Training Grant to J. C. Fay.

LITERATURE CITED

Akashi, H., 1999 Within- and between-species DNA sequence variation and the "footprint" of natural selection. Gene **238:** 39–51.

Barton, N. H., 1995 Linkage and the limits to natural selection. Genetics **140:** 821–841.

Bittles, A. H., and J. V. Neel, 1994 The costs of human inbreeding and their implications for variations at the DNA level. Nat. Genet. **8:** 117–121.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie *et al.*, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22:** 231–238.

Crow, J. F., 1970 Genetic loads and the cost of natural selection, pp. 128–177 in *Mathematical Topics in Population Genetics*, edited by K.-I. Kojima. Springer-Verlag, New York.

Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.

Eyre-Walker, A., 1999 Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics **152:** 675–683.

Eyre-Walker, A., and P. D. Keightley, 1999 High genomic deleterious mutation rates in hominids. Nature **397:** 344–347.

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Fu, Y. X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. **48:** 172–197.

Gillespie, J. H., 1991 *The Causes of Molecular Evolution*. Oxford University Press, Oxford.

Grantham, R., 1974 Amino acid difference formula to help explain protein evolution. Science **185:** 862–864.

Halushka, M. K., J. B. Fan, K. Bentley, L. Hsie, N. Shen *et al.*, 1999 Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22:** 239–247.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Li, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the *Adh* locus in Drosophila. Nature **351:** 652–654.

McVean, G. A., and B. Charlesworth, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929–944.

Morton, N. E., and J. F. Crow, 1956 An estimate of the mutational damage in man from data on consanguineous marriages. Proc. Natl. Acad. Sci. USA **42:** 855–863.

Muller, H. J., 1950 Our load of mutations. Am. J. Hum. Genet. **2:** 111–176.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161–1176.

Shabalina, S. A., and A. S. Kondrashov, 1999 Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. Genet. Res. **74:** 23–30.

Sunyaev, S. R., W. C. Lathe III, V. E. Ramensky and I. Bork, 2000 SNP frequencies in human genes an excess of rare alleles and differing modes of selection. Trends Genet. **16:** 335–337.

Templeton, A. R., 1996 Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial *cytochrome oxidase II* gene in the hominoid primates. Genetics **144:** 1263–1270.

Watterson, G. A., 1975 On the number of segregating sites. Theor. Popul. Biol. **7:** 256–276.

Watterson, G. A., 1987 Estimating the proportion of neutral mutants. Genet. Res. **50:** 155–163.

Wyckoff, G. J., W. Wang and C.-I Wu, 2000 Rapid evolution of male reproductive genes in the descent of man. Nature **403:** 304–309.

Communicating editor: Y.-X. Fu