Random fluctuation of selection coefficients and the extent of nucleotide variation in human populations

Sayaka Miura, Zhenguo Zhang, and Masatoshi Nei[1]

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, University Park, PA 16802

Contributed by Masatoshi Nei, May 8, 2013 (sent for review April 4, 2013)

It is well known that the selection coefficient of a mutant allele varies from generation to generation, and the effect of this factor on genetic variation has been studied by many theoreticians. However, no consensus has been reached. One group of investigators believes that fluctuating selection has an effect of enhancing genetic variation, whereas the other group contends that it has a diversity-reducing effect. In recent years, it has become possible to study this problem by using single nucleotide polymorphisms (SNPs) as well as exome sequence data. In this article we present the theoretical distributions of mutant nucleotide frequencies for the two models of fluctuating selection and then compare the distributions with the empirical distributions obtained from SNP and exome sequence data in human populations. Interestingly, both SNP and exome sequence data showed that the neutral mutation model fits the empirical distribution quite well. Furthermore, the mathematical models with diversity-enhancing and diversity-reducing effects also fit the empirical distribution reasonably well. This result implies that there is no need of distinguishing between the diversity-enhancing and diversity-reducing models of fluctuating selection and the nucleotide polymorphism in human populations can be explained largely by neutral mutations when long-term evolution is considered.

allele frequency distribution | fluctuating selection coefficient | heterozygosity

The importance of fluctuating selection in the study of allele frequency changes was first emphasized by Fisher and Ford (1). However, the potential significance of fluctuating selection in enhancing genetic variation was recognized when mathematical studies (2–4) showed that certain types of fluctuation of selection coefficients may generate stable genetic polymorphism in large populations. These studies were then extended to the case of finite populations using diffusion approximations, and the frequency distribution of mutant alleles was studied under the assumption of stabilizing selection (5–7). These studies suggested that the mean ($M_{\delta x}$) and variance ($V_{\delta x}$) of allele frequency changes per generation used by Wright (8) and Kimura (9) were incorrect. For this reason, the diffusion approximations with the stabilizing selection model are currently used by most investigators (10, 11).

However, Nei and Yokoyama (12) derived different $M_{\delta x}$ and $V_{\delta x}$ values using a competitive selection model (13, 14). In this model, the carrying capacity of individuals in the environment is considered, and therefore the mean fitness ($\overline{w}$) remains to be one every generation, whereas in the stabilizing selection model $\overline{w}$ varies from generation to generation. In this competitive selection model, the fluctuation of selection coefficient has an effect to reduce genetic variability. However, it was difficult to decide which model is more appropriate for predicting the genetic change of a population, because there was no empirical data to support any model.

In recent years, however, SNP data and exome sequence data have become available in human populations, and these data can be used for testing the adequacy of the two models for studying genetic variation. Experimental data about the magnitude of variance (V_s) of s have also accumulated recently for various genes from insects (1), snail (15), Daphnia (16), and many other organisms (17). These data indicate that the fluctuation of s is quite large. Huerta-Sanchez et al. (18) published a likelihood method for testing the effect of fluctuating selection, but it is difficult to use this method for real data analysis, as will be mentioned later. In this article we develop a different statistical method that can be used for testing the two competing models and then study the agreement between the theoretical and empirical distributions. Here, we consider the case of genic (semidominant) selection in a diploid population of effective size N.

Theoretical Studies

Theoretical Distributions of Mutant Allele Frequencies and Heterozygosities. We first study the distribution of mutant allele frequencies under the assumption that there are infinitely many nucleotide sites (infinite-site model) (19) in the genome. We also assume that at each nucleotide site a new mutation occurs with a probability v per generation and the mutant allele (A_2) has a selective advantage or disadvantage of s over the original allele (A_1), s varying at random in different generations (Table S1). If this process of mutation, selection, and genetic drift continues to operate every generation, we will eventually have an equilibrium distribution (spectrum) of mutant frequency x (20). This distribution can be derived by using the diffusion method (21) (ref. 22, p. 121), and it becomes:

$$\Phi(x) = \frac{2v}{V_{\delta x} G(x)} \int_x^1 G(z)\,dz \Big/ \int_0^1 G(z)\,dz, \qquad [1]$$

where $G(x) = \exp(-\int (2M_{\delta x}/V_{\delta x})dx)$. Here $M_{\delta x}$ and $V_{\delta x}$ represent the mean and variance of allele frequency change per generation in diffusion approximations.

In the absence of fluctuation of s (20), the above formula reduces to:

$$\Phi(x) = \frac{4Nv}{x(1-x)} \frac{1-e^{-4Ns(1-x)}}{1-e^{-4Ns}}. \qquad [2]$$

Furthermore, if there is no selection (s = 0), it becomes $\Phi(x) = 4Nv/x$. Another quantity of interest in data analysis is the distribution of heterozygosities in relation to allele frequency x—that is, the number of nucleotide sites with heterozygosity of $2x(1-x)$. This quantity is given by:

$$H(x) = 2x\,(1-x)\,\Phi(x). \qquad [3]$$

In the case of neutral mutations, $H(x)$ is known to be given by $8Nv(1-x)$ (20). If we consider the allele frequency range between q = 1/2N to 0.5, disregarding the ancestral and derived allelic status, $H(x)$ can be written as $H_F(x) = 8Nv(1-x) + 8Nvx = 8\ Nv$ for $q \leq x \leq 0.5$. Therefore, this folded distribution of heterozygosities

Author contributions: S.M. and M.N. designed research; S.M., Z.Z., and M.N. performed research; S.M. and Z.Z. analyzed data; and S.M. and M.N. wrote the paper.

The authors declare no conflict of interest.

[1] To whom correspondence should be addressed. E-mail: nxm2@psu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1308462110/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1308462110

shows a uniform distribution. This property is useful for some data analysis, as will be discussed later.

In the present case we are interested in the effect of random fluctuation of selection coefficient s, and it is difficult to obtain an explicit formula for $\Phi(x)$ or $H(x)$ analytically. However, they can be obtained by numerical integration of Eq. **1**. In this case we first need to evaluate the values of $M_{\delta x}$ and $V_{\delta x}$ for the stabilizing selection model (S) and the competitive selection model (C). The $M_{\delta x}$ and $V_{\delta x}$ values for model S have already been given by Avery (7), whereas for model C they are given by Nei and Yokoyama (12).

Avery used the relative fitness of $1 - s$, 1, and $1 + s$ for genotypes A_1A_1, A_1A_2, and A_2A_2, respectively, and assumed that s fluctuates randomly every generation. He then obtained $M_{\delta x} = V_s x(1-x)(1-2x) + \bar{s}x(1-x)$, where $\bar{s}$ and V_s are the mean and variance of s over generations (Table S1). This equation is interesting because $M_{\delta x}$ is not 0 even for $\bar{s} = 0$. Rather it has a stabilizing selection effect as long as $V_s > 0$. In other words, $M_{\delta x}$ is positive when $x < 0.5$, whereas it becomes negative when $x > 0.5$. In the absence of fluctuating selection, the relative fitnesses of A_1A_1, A_1A_2, and A_2A_2 can also be written as 1, $1 + s$, and $1 + 2s$, respectively (model S_A) or by $1 - 2s$, $1 - s$, and 1, respectively (model S_B), without having significant differences. However, when s varies with generation, the diffusion method gives different $M_{\delta x}$ values in Avery's approach (models S, S_A, and S_B in Table S1) (12). That is, $M_{\delta x}$ generates positive selection for model S_A, but it causes negative selection for model S_B. Therefore, we examined the distributions of mutant allele frequencies for these models under the assumption of $\bar{s} = 0$ and $NV_s = 1$.

Fig. 1*A* shows the distribution (spectrum) of mutant alleles with frequency x. It is seen that models S and S_A indeed give a higher value of $\Phi(x)$ than that for neutral mutations even if a small value of $NV_s = 1$ is used. This pattern is seen more clearly for the distribution of heterozygosities (Fig. 1*B*). By contrast, model S_B shows a lower value of $\Phi(x)$ than in the case of neutral mutations. These results indicate that the Avery approach gives quite different results depending on the model used. We believe this is a deficiency of Avery's modeling of fluctuating selection because we are considering the same biological situation. In practice, most investigators in the past have used model S. We will therefore consider only this model in the following.

In model C the carrying capacity of individuals (adult size of N) is assumed to be constant, and therefore the mean fitness ($\overline{w}$) is equal to 1 every generation (12). For this reason, the fitnesses of A_1A_1, A_1A_2, and A_2A_2 become a function of allele frequency x (Table S1). However, the mean allele frequency change per generation is given by $\bar{s}x(1-x)$, as expected for genic selection (14). In the present case $M_{\delta x} = 0$ when $\bar{s} = 0$, and the allele frequency change is dictated by the term of $V_{\delta x}$, which is usually greater than the genetic drift term $[x(1-x)/2N]$ for neutral mutations. Therefore, model C is expected to have a diversity-reducing effect.

Fig. 1*C* shows the theoretical distributions of mutant alleles for the stabilizing (S) and competitive (C) selection models in comparison with the distribution for neutral mutations, where the same $4Nv$ value is used. It is clear that the stabilizing selection model (S) enhances the number of high frequency alleles considerably compared with the model of neutral mutations particularly when NV_s is relatively high (e.g., $NV_s = 10$). By contrast, the competitive selection model reduces the number of polymorphic alleles substantially compared with the neutral model. This is also clear from Fig. 1*D*, where the expected distribution of heterozygosities is presented.

Normalized Distributions of Mutant Frequencies Required for Data Analysis. In real data analysis, Eqs. **1** and **3** are not very useful, because $\Phi(x)$ gives the absolute number of sites with allele frequencies x and this makes it difficult to compare the theoretical distributions with the observed distribution. Also, the distributions in Fig. 1 are for the region between $x = 1/(2N)$ and $x = 1 - 1/(2N)$, where N is the population size. However, N is usually unknown in real populations. Furthermore, the observed number of polymorphic alleles would be affected by the number of genomes (n) sampled for identifying polymorphic sites.

For these reasons, we examined the normalized distributions of polymorphic nucleotide sites for different intervals of allelic frequencies. The expected number of nucleotide sites with discrete mutant frequency i is given by:

$$f(i) = \int_{q}^{1-q} \binom{n}{i} x^{i}(1-x)^{n-i}\Phi(x)\,dx, \qquad \textbf{[4]}$$

where $q = 1/(2N)$ and n is the genome sample size (23, 24). The relative nucleotide site number, $F(x_1, x_2)$, of allele frequencies between x_1 and x_2 is given by:

$$F(x_1, x_2) = \sum_{i=x_1}^{i=x_2} f(i)/T, \qquad \textbf{[5]}$$

where T is the total number of polymorphic sites in the sample $\left[T = \sum_{1}^{n-1} f(i)\right]$. This expected frequency, $F(x_1, x_2)$, is convenient for comparing the theoretical distribution with the observed distribution, and we call $F(x_1, x_2)$ the normalized frequency distribution. However, when the sample size is large and we consider the relatively high allele frequency classes, this expected frequency is practically equal to the following population frequency:

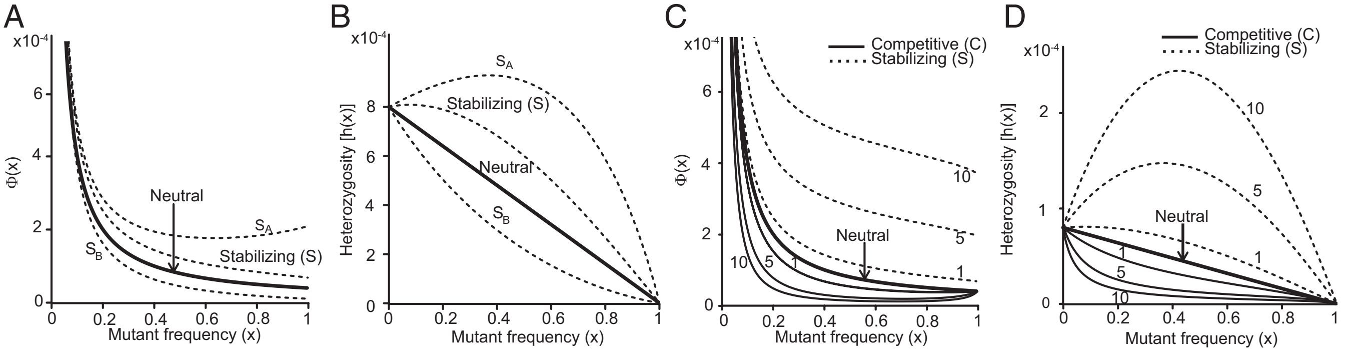

Fig. 1. Theoretical distributions (spectra) of mutant allele frequencies (x) for various cases of random fluctuation of selection coefficient. Neutral, neutral mutations. S, S_A, S_B, and C refer to the models presented in Table S1. In *C* and *D* the number given for each curve stands for the NV_s value. Here $N\bar{s} = 0$ is assumed.

$$F(x_1,x_2)=\int_{x_1}^{x_2}\Phi(x)\,dx\bigg/\int_{q}^{1-q}\Phi(x)\,dx. \qquad [6]$$

Note that the normalized distribution is independent of mutation rate (v), because the v value in Eq. **1** is cancelled out.

In the case of sample size n = 700 genomes with neutral mutations, the expected frequency of x is somewhat different from the population frequency of x (Fig. S1*A*). However, if we consider only the frequency region of x = 0.05–0.95, the normalized distribution obtained by Eq. **5** is virtually identical with the distribution obtained by Eq. **6** (Fig. S1*B*). In practice, the computation of Eq. **6** is much simpler than that of Eq. **5**. We therefore use Eq. **6** when n is large.

Normalized Allele Frequency Distributions for a Few Different Types of Selection. Let us now compute a few examples of theoretical distributions to examine whether the normalized distributions of x can be used for distinguishing between different selection models when n is large. When there is no fluctuating selection with NV_s = 0, the distributions for the cases of neutral (Ns = 0), advantageous (Ns = 10), and deleterious mutations (Ns = –10) are easily distinguishable even if the absolute values of s are very small (Fig. 2*A*). (If n = 10,000 and Ns = 10, s will be 0.001.) In the case of deleterious mutations, the relative frequency of rare alleles is very high, but the frequency rarely increases more than x = 0.2. By contrast, advantageous mutations (genic selection) reduce the alleles with intermediate frequencies and show a mild U-shaped distribution. Neutral mutations show an L-shaped distribution, as expected from Fig. 1*A*. Similarly, the normalized distributions of heterozygosities are distinguishable (Fig. 2*B*).

For the case of fluctuating selection with $N\bar{s}=0$, the competitive selection and stabilizing selection models with NV_s = 10 and the neutral model are all distinguishable either by the allele frequency distribution or by the distribution of heterozygosities (Fig. 2 *C* and *D*). These results indicate that the normalized frequency distributions of mutant alleles and heterozygosities are useful for distinguishing between the competitive and stabilizing selection models. We will therefore use this method in the following data analysis.

Analysis of SNP and SNV Data

Problems Arising in Data Analysis. In recent years several research groups have published data on SNPs or single nucleotide variants (SNVs) for exome sequences from various human populations (25–28). However, many of the datasets published cannot be used for our purpose for several reasons. First, the mathematical theories presented above are for a population of which the effective size has remained more or less constant for a long time, and therefore the allelic frequency distribution has reached the steady state. Previous studies have suggested that the long-term effective size (N) of the human population is of the order of 10,000 for allozyme data that do not contain low-frequency alleles (29). According to the Out-of-Africa theory of human evolution (30, 31), non-Africans are recent migrants from Africa during the last 150,000 y. Therefore, the distribution of mutant nucleotide frequencies may be deviated from the steady-state distribution. However, African populations seem to have maintained a relatively constant effective population size for hundreds of thousands of years (32). We have therefore decided to use only African populations in our study. More specifically, we used populations inhabiting central Africa, because their population size appears to have been stable until recently. Of course, rare alleles may have increased by a moderate increase of population size in recent years (27), but high-frequency alleles are unlikely to be affected by recent population growth. For this reason, we decided to use combined SNP data obtained from three central African populations (Table S1).

The SNP data were obtained from the HapMap database (33). In some genomes SNP sites were not clearly identifiable, and these sites were eliminated. The total number of genomes used was n = 689, and each genome contained about 1,546,052 SNP sites when all SNP sites were aligned with the reference human genome (Table S2). Note that because SNP sites were identified by microarray analysis and there were other technical problems, most of low-frequency alleles were not correctly identified (33). We therefore eliminated low-frequency alleles, as will be mentioned below. In the case of DNA sequences (exome data) of the protein-coding regions, however, there is almost always an excess of rare alleles caused by deleterious mutations. In our study we are not interested in deleterious mutations, which are eventually eliminated from the population, and therefore these rare alleles should also be discarded from our data analysis. However, it is difficult to establish a method to eliminate deleterious mutations, and it is required to use some subjective judgment, as will be mentioned below.

In our data analysis we considered protein-coding regions and intergenic regions of DNA separately, because they may be under different selection pressures. In the case of protein-coding region, we extracted all of the exons of a protein-coding gene, excluding the noncoding elements such as UTRs and other regulatory elements (e.g., *cis*-regulatory regions). By contrast, the intergenic regions included all of the DNA regions except for the protein-coding regions and the 5 kb regions adjacent to the protein-coding region. Therefore, the intron regions were not used.

Raw Allele Frequency Data. To examine the quality of our raw data, we first studied the folded allele frequency distribution for the interval of 1/n to 0.5, because in this case we cannot determine the ancestral and derived alleles (Fig. 3). Fig. 3 *A* and *C* shows that a large proportion of SNP sites (HapMap data) with low mutant frequencies are clearly deficient because any theoretical distributions show a large proportion of rare alleles (Fig. 1 *A* and *C*). To obtain a rough extent of the deficiency, we computed the frequency distribution expected for neutral mutations

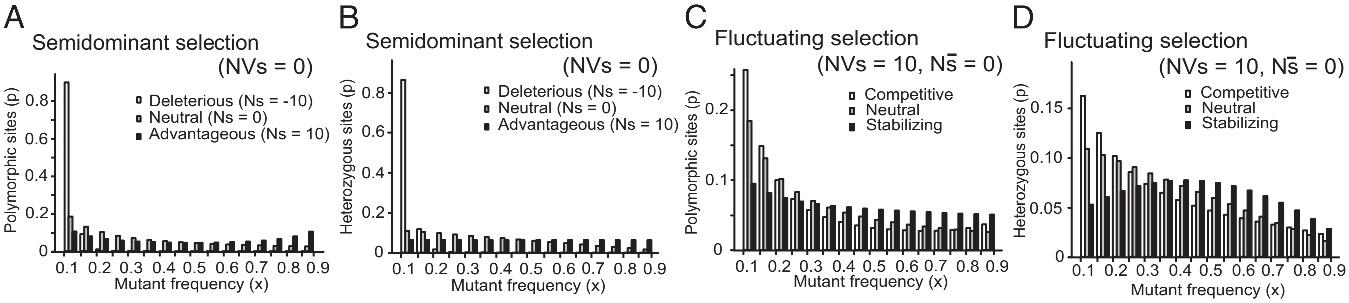

Fig. 2. Normalized distributions of mutant allele frequency (x) for various types of selection. p indicates the proportional frequency of polymorphic sites or heterozygosities rather than the absolute number.

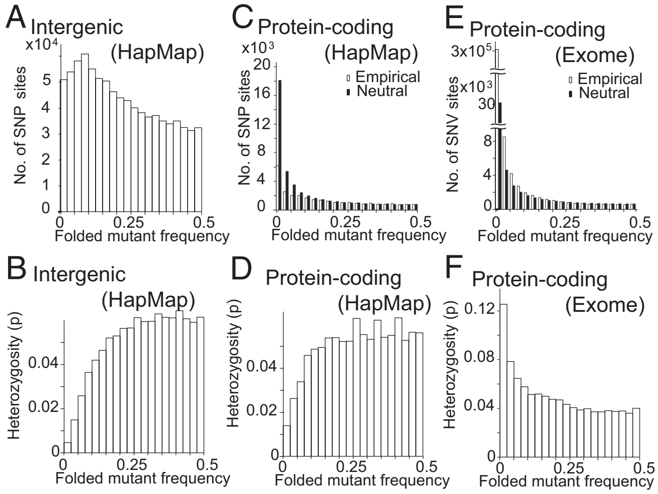

Fig. 3. Folded empirical distributions of mutant frequency (*x*) and heterozygosity for an African population (HapMap data) and an African American population (exome data). The allele frequencies considered are from 1/*n* to 0.5, where *n* is the sample size. *p*, proportion of heterozygous sites.

for the case of the protein-coding region by using the following formula: $K4Nv\left(\int_{x_1}^{x_2}\frac{dx}{x}+\int_{1-x_2}^{1-x_1}\frac{dx}{x}\right)=K4Nv\ln[x_2(1-x_1)/(x_1/(1-x_2))]$, where K is the total number of nucleotide sites examined. K was 34,255,626 for the protein-coding region. The neutral expectation was not obtainable for the intergenic region, because this region contained many nucleotide deletions/insertions. For the protein-coding region, we estimated $4Nv$ by fitting the neutral distribution to the observed for the regions of x = 0.2–0.5, where the distribution appeared to be close to the neutral one (see below). The $4Nv$ value was estimated to be 0.00021 (Table S3).

Fig. 3C clearly indicates that the observed number of SNP sites with low frequencies is much lower than the neutral expectation. In this case it is possible to use Huerta-Sanchez et al.'s likelihood method of testing the neutral distribution. However, because the deficiency of rare alleles in comparison with the neutral expectation is so obvious and the number of nucleotide sites examined is so large, it was not really necessary to use it for testing the appropriateness of the neutral theory for the entire frequency region of x. The shortage of rare alleles is also observed in the distributions of heterozygosities for both intergenic (Fig. 3*B*) and protein-coding regions (Fig. 3*D*). According to the neutral expectation, the distribution of heterozygosities for different allele frequency classes is expected to be rectangular. However, the frequencies of rare allele classes are substantially lower than those for high-frequency alleles. This shortage of rare alleles was quite unexpected because natural populations usually harbor many deleterious mutations. Therefore, the shortage must be caused by the technical problems mentioned above. This interpretation is supported by the allele frequency distributions for the DNA regions used in the ENCODE project (33). These DNA regions were completely sequenced to find all rare alleles as well as common alleles. In this dataset the number of rare alleles is much larger than the neutral expectation (Fig. S2).

Inspection of Fig. 3 *B* and *D* suggests that the neutral theory would fit the observed data better in the frequency region of x = 0.25–0.75 than in the region of x = 0.2–0.8, because the folded allele frequency distribution is slightly more uniform in the former region than in the latter. However, to avoid the bias toward the neutral theory, we considered both regions.

Comparison of Empirical Allele Frequency Distribution with Theoretical Distribution. The main purpose of this article is to examine the agreement of the stabilizing selection and the competitive selection models of fluctuating selection with the observed distribution of mutant nucleotide frequencies. We therefore identified which of the two nucleotides at each SNP site is the derived mutant by using chimpanzee and macaque genome sequences. Suppose that there are nucleotides A and T at a given SNP site and the chimpanzee and macaque genome sequences show only nucleotide A at the homologous site. In this case we can infer by parsimony that nucleotide T is the derived mutant in the human population. Similarly, if the chimpanzee and macaque both have nucleotide T, we infer that nucleotide A is the derived allele. However, if the chimpanzee and macaque have any other nucleotides (e.g., C or G, or A and G, respectively), we cannot infer the mutant allele in the human population. These SNP sites were eliminated from our data analysis. Using this approach, we could identify mutant nucleotides at each SNP site. Essentially the same results were obtained when only the chimpanzee was used as the outgroup species.

The theoretical distributions of the competitive selection (C), stabilizing selection (S), and neutral models were obtained by fitting the expected distribution to the observed distribution using the least-squares method. In the least squares method, we estimated the parameters $N\bar{s}$ and NV_s by minimizing the least squares residue $LSR=\left[\sum_{i=1}^{m}\left((o_i-e_i)/e_i\right)^2\Big/m\right]^{1/2}$, where o_i and e_i are the observed number of SNP sites for the frequency class i and the expected number of sites for the same frequency class i, respectively, and m is the total number of frequency classes used. In practice, we minimized LSR by choosing appropriate values of $N\bar{s}$ and NV_s numerically. The $N\bar{s}$ and NV_s values thus obtained are called the least squares estimates. The minimum value of LSR that gives these estimates is denoted by LSR^*.

First, the observed distribution of mutant nucleotide frequencies for the SNP data of the intergenic regions was compared with the theoretical distributions for neutral mutations and the competitive (C) and stabilizing (S) selection models for the region of x = 0.2–0.8 (Fig. 4*A*). It is clear that the observed distribution agrees with the neutral expectation quite well, the LSR^* being only 0.042 (Table S4). The expected distribution for models C and S also fit the observed distribution very well, the LSR^* values being 0.037 and 0.036, respectively. The least squares estimates of $N\bar{s}$ and NV_s are all small and close to 0 (Table S4). Essentially the same results are obtained from the analysis of the distribution of heterozygosities (Fig. 4*B*, Table S4). These results indicate that the SNP data for the intergenic regions can be explained largely by neutral mutations. Inspection of Fig. 4 *C* and *D* indicates that the agreement between the empirical and theoretical distributions is less satisfactory for the protein-coding region than for the intergenic regions. This can also be seen from the LSR^* values. However, the estimates of $N\bar{s}$ and NV_s are nearly the same as those for the intergenic regions.

As mentioned earlier, we also considered the allele frequency region of x = 0.25–0.75 to study the distributions of mutant nucleotides and heterozygosities. In this case the agreement between the neutral and the empirical distributions was slightly better than that for the case of x = 0.2–0.8 (Fig. S3, Table S4). This result suggests that if we consider the middle allele frequency region, the distribution of mutant allele frequencies are largely explained by neutral mutations. However, because $N\bar{s}$ was always negative, some extent of purifying selection appears to be operating (Table S4).

Analysis of Exome SNV Data. In recent years great effort has been made to identify rare alleles related to human diseases by sequencing the exon regions of genes in the whole genome (exome). In these studies the genomes of individuals with some complex diseases such as lung and heart diseases are sequenced with the intention of identifying disease-causing genes (27). However, the single nucleotide polymorphic variants (SNVs)

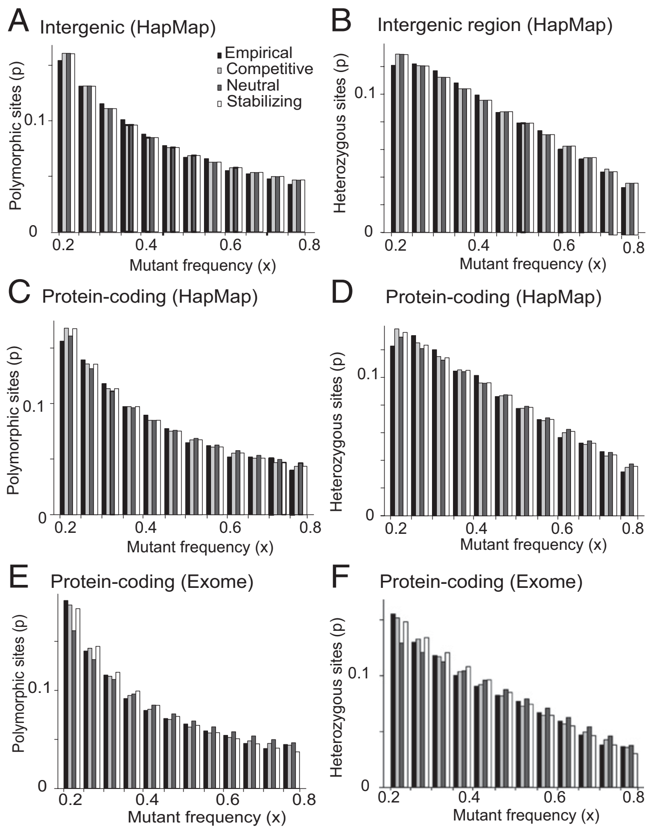

Fig. 4. Theoretical and empirical distributions of mutant nucleotide frequency (*x*) and heterozygosity for the intergenic and protein-coding regions of HapMap data and the protein-coding region of exome data from human populations. The ordinate represents the number of polymorphic sites or heterozygous sites in proportion of the total number of sites. Here the allele frequency region of x = 0.2–0.8 is considered. *p*, proportional frequency rather than the absolute number.

identified in this way are useful for studying the evolutionary dynamics of mutant alleles if we eliminate rare alleles as we did in the above analysis. Such SNV data are now available from the database provided by the National Heart, Lung and Blood Institute (NHLBI) exome sequencing project (27). At present, there are no SNV data for Africans in the NHLBI database, but the data for African Americans are available. African Americans (primarily derived from central Africa) are known to contain many Caucasian genes because of the past genetic admixture, and this admixture might have increased the frequency of rare alleles in the population (27). Furthermore, because the exomes sequenced were obtained from disease-inflicted individuals, this biased sampling might also have increased rare alleles. The number of genomes used for African Americans in the NHLBI project was 4,406, and the number of polymorphic sites obtained was 559,974 (Table S2). As expected, this dataset indicates that the African American population contains a large number of SNV sites with low allele frequencies (Fig. 3*F*). The number of SNV sites with low frequencies is much higher than the neutral expectation. Even in the frequency classes of $x \sim 0.2$, the observed frequency is slightly higher than the neutral expectation (Fig. 3*E*). The excess of rare alleles is also observed from the number of heterozygous sites (Fig. 3*F*). However, if we consider the frequency classes of 0.2–0.5, the excess of rare alleles does not seem to affect our analysis seriously. We therefore decided to use these sites for our study. As in the case of HapMap data, we will also consider the x = 0.25–0.5 region separately.

We identified the derived mutant nucleotides using the same method as that used for SNP data. The agreement between the observed and the theoretical distributions is less satisfactory than in the case of SNP data (Fig. 4*E*). The observed number of SNV sites (proportional frequency) is higher than the neutral expectation when the mutant frequency (x) is small but tends to be lower when x is large (Fig. 4*E*). In fact, the LSR^* value for this comparison is considerably higher than that for the SNP data (Table S4). Interestingly, however, LSR^* for the competitive selection model is considerably lower than that for the neutral and the stabilizing selection model, suggesting that the competitive model fits the data better than the other two models, although it is difficult to study the statistical significance of the difference for the reasons mentioned in the *Discussion*. The same conclusion can be obtained from the study of heterozygosities (Fig. 4*F*, Table S4). When the allele frequency region of x = 0.25–0.75 was considered, however, LSR^* was virtually the same for models C and S, and the values were smaller than those for the neutral model (Fig. S3, Table S4).

Discussion

There are many SNP and SNV datasets available now, but most of them were obtained for medical purposes and are not always appropriate for evolutionary studies, because rare alleles are sometimes deficient and sometimes overrepresented. However, the data with intermediate mutant frequencies with x = 0.2–0.8 are useful for evolutionary studies. These alleles have stayed in the population for a long time and appear to have reached the steady-state distribution. This judgment is certainly subjective, but the results of our data analyses consistently support this view, the estimate of $4Nv$ being about 0.0002 for all of the datasets.

Furthermore, it should be noted that this type of study is intended to obtain a crude and general conclusion. Strictly speaking, our assumption that all SNP or SNV loci are subject to the same evolutionary force must be incorrect, because the patterns of selection and mutation rate cannot be identical for all polymorphic sites. We are interested only in the general evolutionary pattern for all of the sites. If we study the evolutionary pattern of a specific SNP or SNV site, we may find a very different mechanism. However, this is not of our interest.

A similar cautionary note should be stated about our statistical analysis. We used the least-squares method for fitting a theoretical distribution of mutant frequencies to empirical data, and we obtained the conclusion that the empirical distributions can be explained approximately by the neutral model. The standard statistical method for testing the agreement between the two distributions is the χ^2 test. In the present case, however, this test gives the conclusion that any two distributions are different with a high probability, because the number of SNP or SNV sites used is very large. For this reason, we used LSR^* as an indicator of the discrepancies between data and theory. However, note that this indicator cannot be subject to Fisherian small-sample tests because the samples used are very large and heterogeneous. For these reasons, our tests are not completely objective, but for our purpose they are sufficient. Because so many heterogeneous factors are involved in any evolutionary process, there is a limit for knowing the real truth.

In the case of fluctuating selection many authors have stated that it has a diversity-enhancing effect similar to overdominant selection without analyzing any empirical data. This was particularly emphasized by Gillespie (10). The present study does not support this view and the neutral mutation model can explain the observed distribution reasonably well, and there is no need of considering the effect of fluctuation selection seriously. We concluded that the alternative model of fluctuating selection, which has a diversity-reducing effect, can also explain the observed data quite well. Actually, there seems to be a small extent of purifying selection operating.

As mentioned above, our results suggest that the fluctuation of selection coefficients does not enhance genetic variability, but rather it has a tendency to reduce it. However, its significance is minor, and the extent of genetic polymorphism can be explained largely by the neutral mutation theory. This does not mean that there is no selection. It is quite possible that natural selection occurs mostly on a short-term basis in either positive or negative direction, but the effects of positive and negative selection are cancelled out when long-term evolution is considered, as in the case of industrial melanism of the peppered moth *Biston betularia* (34). Because it takes a long time for an allelic substitution to take place, a short-time observation of natural selection may not be so important for long-term evolution. Note also that the selection coefficient for a particular mutation can be very small, because most amino acid changes and silent nucleotide substitutions are largely neutral (34). Whatever the reasons are, our study suggests that most nucleotide substitutions in the evolutionary process are more or less neutral when a large number of SNP or SNV sites are considered.

In recent years a large number of authors have reported signature of natural selection by studying nucleotide polymorphisms (35–38). However, Nei et al. (39) and Nei (34) have indicated that many of these studies are based on questionable statistical methods or assumptions, and therefore their conclusions are not very reliable. Actually, conducting an extensive literature survey, Nei (34) proposed that the evolutionary changes of phenotypic characters as well as DNA and protein sequences occur primarily by mutation, whether the mutation is due to nucleotide substitution, gene duplication, or polyploidization and that evolution is intrinsically unpredictable. He also concluded that phenotypic evolution is caused by a relatively small proportion of major-effect mutations and the remaining portion of mutations is largely neutral. The results in this article are more or less consistent with this theory.

ACKNOWLEDGMENTS. We thank Naoyuki Takahata, Ranajit Chakraborty, and Shozo Yokoyama for their useful comments.

1. Fisher RA, Ford EB (1947) The spread of a gene in natural conditions in a colony of the moth Panaxia dominula L. *Heredity* 1(2):143–174.
2. Haldane J, Jayakar S (1963) Polymorphism due to selection of varying direction. *J Genet* 58:237–242.
3. Gillespie JH (1973) Natural selection with varying selection coefficients—A haploid model. *Genet Res* 21:115–120.
4. Hartl DL, Cook RD (1973) Balanced polymorphisms of quasineutral alleles. *Theor Popul Biol* 4:163–172.
5. Jensen L (1973) Random selective advantages of genes and their probabilities of fixation. *Genet Res* 21(3):215–219.
6. Karlin S, Levikson B (1974) Temporal fluctuations in selection intensities: Case of small population size. *Theor Popul Biol* 6:383–412.
7. Avery PJ (1977) The effect of random selection coefficients on populations of finite size-some particular models. *Genet Res* 29(2):97–112.
8. Wright S (1948) On the roles of directed and random changes in gene frequency in the genetics of populations. *Evolution* 2(4):279–294.
9. Kimura M (1954) Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. *Genetics* 39(3):280–295.
10. Gillespie JH (1991) *The Causes of Molecular Evolution* (Oxford Univ Press, New York).
11. Mustonen V, Lässig M (2007) Adaptations to fluctuating selection in Drosophila. *Proc Natl Acad Sci USA* 104(7):2277–2282.
12. Nei M, Yokoyama S (1976) Effects of random fluctuation of selection intensity of genetic variability in a finite population. *Jpn J Genet* 51:355–369.
13. Mather K (1969) Selection through competition. *Heredity (Edinb)* 24(4):529–540.
14. Nei M (1971) Fertility excess necessary for gene substitution in regulated populations. *Genetics* 68(1):169–184.
15. Cain AJ, Cook LM, Currey JD (1990) Population size and morph frequency in a long-term study of Cepaea nemoralis. *P Roy Soc Lond B Bio* 240:231–250.
16. Lynch M (1987) The consequences of fluctuating selection for isozyme polymorphisms in Daphnia. *Genetics* 115(4):657–669.
17. Bell G (2010) Fluctuating selection: The perpetual renewal of adaptation in variable environments. *Philos Trans R Soc Lond B Biol Sci* 365(1537):87–97.
18. Huerta-Sanchez E, Durrett R, Bustamante CD (2008) Population genetics of polymorphism and divergence under fluctuating selection. *Genetics* 178(1):325–337.
19. Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge, UK).
20. Wright S (1938) The distribution of gene frequencies under irreversible mutation. *Proc Natl Acad Sci USA* 24(7):253–259.
21. Kimura M (1964) Diffusion models in population genetics. *J Appl Probab* 1:177–232.
22. Nei M (1975) *Molecular Population Genetics and Evolution* (North–Holland, Oxford, UK).
23. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. *Genetics* 132(4):1161–1176.
24. Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. *Genetics* 159(4):1779–1788.
25. Altshuler DM, et al.; International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311):52–58.
26. Ku CS, Naidoo N, Pawitan Y (2011) Revisiting Mendelian disorders through exome sequencing. *Hum Genet* 129(4):351–370.
27. Tennessen JA, et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64–69.
28. Li Y, et al. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* 42(11):969–972.
29. Nei M, Graur D (1984) Extent of protein polymorphism and the neutral mutation theory. *Evol Biol* 17:73–118.
30. Nei M, Roychoudhury AK (1982) Genetic relationship and evolution of human races. *Evol Biol* 14:1–59.
31. Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature* 325(6099):31–36.
32. Kimmel M, et al. (1998) Signatures of population expansion in microsatellite repeat data. *Genetics* 148(4):1921–1930.
33. The International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58.
34. Nei M (2013) *Mutation-Driven Evolution* (Oxford Univ Press, Oxford, UK).
35. Begun DJ, et al. (2007) Population genomics: Whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol* 5(11):e310.
36. Sawyer SA, Parsch J, Zhang Z, Hartl DL (2007) Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila. *Proc Natl Acad Sci USA* 104(16):6504–6510.
37. Sabeti PC, et al.; International HapMap Consortium (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449(7164):913–918.
38. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res* 16(8):980–989.
39. Nei M, Suzuki Y, Nozawa M (2010) The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet* 11:265–289.