Received 22 October 1999, accepted 2 April 2001

Variance of protein heterozygosity in different species of mammals with respect to the number of loci studied

ANASTASSIA M. MAKARIEVA

Theoretical Biology Group, Coltegium Budapest, Institute for Advanced Study, Szentháromsaǵ utca 2, H-1014, Budapest, Hungary

Analysis of published data on protein heterozygosity of 321 species of mammals shows that it varies from 0 up to 22%, an average species being heterozygous at 5% of its protein-coding loci. Many attempts have been made to explain the observed differences in protein heterozygosity, relating its value to various species-, population-, or environment-specific parameters. In this work it is shown that the wide scatter of protein heterozygosity in different species of mammals can be explained by the small numbers of loci studied (usually 20–30). It is shown that with an increasing number of studied loci, the mean of the heterozygosity does not change, while its variance among different species decreases in accordance with a Poisson distribution. The true heterozygosity of the whole protein-coding region of the mammalian genome is thus characterized by a narrow spread around the mean. This means that the true heterozygosity of the protein-coding region is similar in all mammalian species. Its value can be viewed as the threshold level of variability of the protein-coding region of mammals, which characterizes the permissible level of erosion of genetic information of species and is maintained by stabilizing selection in natural ecological niches.

Keywords: mammals, Poisson distribution, polymorphic loci, protein heterozygosity, stabilizing selection, variance.

Introduction

The question of the nature and maintenance of intraspecific genetic variability is widely discussed in the literature (Nei, 1984; Nevo *et al.*, 1984; Kimura, 1991; Ayala & Fitch, 1997; Amos & Harwood, 1998). The observations show that intraspecific genetic variability is remarkably different even among evolutionarily close species (Nevo *et al.*, 1984). Much work has been done to reveal possible correlations between genetic variability and such characteristics as genome size (Pierce & Mitton, 1980; Larson, 1981), body mass (Wooten & Smith, 1985; Gorshkov & Makarieva, 1997), mating system (Appolonio & Hartl, 1993), ploidy level (Graur, 1985; Hofker *et al.*, 1986; Crespi, 1991; Burrows & Ryder, 1997), population size, structure and history (Nei & Graur, 1984; O'Brien, 1987; Carson, 1990) and other ecological or behavioural parameters such as, for example, environmental stress (Imasheva, 1999), host–parasite interactions (Thompson & Lymbery, 1996) or habitat fragmentation (Gaines *et al.*, 1997), but in spite of many significant findings, the general pattern remains unclear.

Abundant evidence on the levels of heterozygosity of the protein-coding region of the genome has been accumulated for thousands of species (Nevo *et al.*, 1984). At present, due to the explosive development of new measuring techniques, most studies assess the DNA variability directly (e.g. by DNA sequencing) rather than by studying proteins encoded by it. However, the development of various DNA techniques during the last several years has not yet led to creation of a comparable dataset of DNA variability for the coding region of *different species*, especially taking into account the growing bias in molecular studies towards medicine, and, consequently, the single species *Homo sapiens*. Thus, the protein variability dataset still remains unique due to its extensiveness with respect to the number of species encompassed (Butlin & Tregenza, 1998). In these data, mammals represent one of the best-studied taxa.

In this study the statistical aspect of the problem of intraspecific genetic variability is investigated. Published

Correspondence and present address: Petersburg Nuclear Physics Institute, 188300, Gatchina, St-Petersburg, Russia.
E-mail: elba@infopro.spb.su

data on the expected protein heterozygosity of 321 natural mammalian species were collected. It is shown that the different observed values of protein heterozygosity in different species and the wide spread of the observed values of protein heterozygosity around the mean can be explained by the small number of loci studied. With increasing number of studied loci the scatter of heterozygosity values decreases in accordance with a Poisson distribution while the mean does not change. This suggests that all mammals are characterized by nearly the same level of protein heterozygosity.

Rationale

The coding region of the mammalian genome is known to contain about 30–40 000 genes. This number of genes cannot be studied in any population genetics experiment. In most studies, sets of 20–30 loci are investigated in different species. The number of studied loci is determined by the available buffer systems, techniques developed, manpower and money.

The expected heterozygosity h of a given locus (i.e. the relative frequency of heterozygous state in a random-mating population) is determined as:

$$h = 1 - \sum_j q_j^2, \qquad (1)$$

where q_j is the relative frequency of the j-th allele.

Let l be the number of polymorphic loci (i.e. loci with non-zero h) in a study of L loci in a population. The average expected heterozygosity of individuals in a population is given by the expression

$$H_L = \frac{\sum_{i=1}^{l} h_i}{L} = \frac{h_l l}{L} \quad \text{if } l > 0 \quad \text{and} \quad H_L = 0 \quad \text{if } l = 0, \qquad (2)$$

where L is the total number of studied loci, h_i is the expected heterozygosity of the i-th polymorphic locus, see (eqn 1), h_l is the average heterozygosity of the l polymorphic loci, i.e. $h_l \equiv \sum_{i=1}^{l} h_i / l (l > 0)$.

Let us denote by $h_{l_{\text{tot}}}$ the average heterozygosity of all l_{tot} polymorphic loci existing in the protein-coding region of a given species, $h_{l_{\text{tot}}} \equiv \sum_{i=1}^{l_{\text{tot}}} h_i / l_{\text{tot}}$. The greater the number L of studied loci, the better the approximation of H_L that is given by the expression

$$H_L \approx \frac{h_{l_{\text{tot}}} l}{L} \qquad (3)$$

It will be shown below that under certain assumptions, a very accurate estimate of $h_{l_{\text{tot}}}$ can be obtained from data on heterozygosity of polymorphic loci in different species of mammals. Introduction of a constant value of $h_{l_{\text{tot}}}$ (eqn 3) instead of the random variable h_l (eqn 2) leads to a substantial simplification of all the formulae and allows one to analyse the variance of heterozygosity H_L as a function of variance of the number of polymorphic loci l alone, provided that the inaccuracy of the approximate formula (3) is not very large.

The random variable h_l defined in (eqn 2) assumes different values when different sets of L loci are studied in the same species. If in all sets of L loci the same number l of polymorphic loci were observed, the mean and variance of h_l were given by the expressions $\bar{h}_l \equiv \sum_{k=1}^{K_{\text{tot}}} h_l^k / K_{\text{tot}}$ and $\sigma_{h_l}^2 \equiv \sum_{k=1}^{K_{\text{tot}}} (h_l^k - \bar{h}_l)^2 / K_{\text{tot}}$, where h_l^k is the average heterozygosity of polymorphic loci in the k-th set of L loci, K_{tot} is the total number of sets of L loci that cover the whole protein-coding region of the species genome, $K_{\text{tot}} = L_{\text{tot}}/L$, where L_{tot} is the total number of loci in the protein-coding region. Note that $\bar{h}_l = h_{l_{\text{tot}}}$ by definition. Thus, $\sigma_{h_l}^2$ characterizes deviation of h_l from $h_{l_{\text{tot}}}$ and, consequently, deviation of the approximate formula (eqn 3) from the exact formula (eqn 2). It can be easily shown from the definitions of $\bar{h}_l$ and $\sigma_{h_l}^2$ that

$$\sigma_{h_l}^2 = \frac{\sigma_h^2}{l} \qquad (4)$$

where σ_h^2 is the variance of the random variable h, which assumes values of expected heterozygosity of all polymorphic loci l_{tot} of this species, $\sigma_h^2 \equiv \sum_{i=1}^{l_{\text{tot}}} (h_i - h_{l_{\text{tot}}})^2 / l_{\text{tot}}, l_{\text{tot}} = K_{\text{tot}} l$. Variance is a characteristic of the squared value of spread around the mean. Equality (eqn 4) essentially represents the law of large numbers and reflects the fact that the more polymorphic loci are studied, the more accurate is the estimate of their mean heterozygosity that is obtained. The average number of polymorphic loci observed in a given set of loci is proportional to the total number of loci studied in this set. It will be shown below that at most commonly used L the average inaccuracy of (eqn 3) does not exceed 30%, see (eqn 12). Thus, in further calculations we use (eqn 3) as an acceptable approximation of (eqn 2).

Let us call the hypothetical value of heterozygosity that could be obtained when studying all protein-coding loci in a species for true heterozygosity H:

$$H \equiv H_{L_{\text{tot}}} = h_{l_{\text{tot}}} \frac{l_{\text{tot}}}{L_{\text{tot}}} = h_{l_{\text{tot}}} P, \qquad (5)$$

where $P \equiv l_{\text{tot}}/L_{\text{tot}}$ is the true polymorphism. If H is equal in all species (the assumption that we will be further testing), it is likely that both $h_{l_{\text{tot}}}$ and P are equal simultaneously, rather than that there exists a certain inverse relationship between them. There are two independent factors that influence the level of heterozygosity of a particular locus in a particular species. Firstly, for a locus to become polymorphic it has to be affected by a mutation. It is natural to expect that the mutation-affected loci are randomly located in genomes of different species. Secondly, the exact value of heterozygosity h_i of a mutation-affected locus depends on how many chances the heterozygous genotypes have for spreading in the population. This may depend on the properties of the locus itself (Ward *et al.*, 1992). Indeed, certain groups of loci tend to show elevated values of heterozygosity in most species, whereas others are generally less variable (O'Brien *et al.*, 1980). However, if no mutation has affected the locus in the population, it will remain monomorphic with $h = 0$, irrespective of how high its heterozygosity could be were such a mutation to occur. Thus, we assume below that the equality of both independent factors, P and $h_{l_{\text{tot}}}$, in all species is both necessary and sufficient for equality of H in all species.

If determined by the random character of the process of mutagenesis, polymorphic loci should be distributed randomly over the coding region of genome. Then, given the low relative frequency of polymorphic loci as compared to monomorphic ones (as is usually the case) the probability $p(l)$ of finding l polymorphic loci among L studied loci will be determined by a Poisson distribution. (It has been tested that applying a binomial distribution does not change in any significant way any results obtained below).

If the true heterozygosity H, and, consequently, the true polymorphism P (eqn 5) are equal in all species, then the number l of polymorphic loci observed in sets of L loci in *different species* will follow the same Poisson distribution as the number of polymorphic loci in sets of L loci chosen in different parts of the coding region of the *same species*. Then the probability $p(l)$ of finding l polymorphic loci in a random set of L loci in a population of a given species is equal to

$$p(l) = \exp(-\bar{l})\frac{\bar{l}^l}{l!} \tag{6}$$

where $\bar{l}$ is the average number of polymorphic loci that are found in different species, $\bar{l} \equiv \sum_{k=1}^{N_L} l_k/N_L$, where l_k is the number of polymorphic loci in the k-th species and N_L is the number of species where L loci were studied. Note that $\bar{l} = LP$. Everywhere below we assume that only one population represents each species, so averaging over different populations is equivalent to averaging over different species.

Note that if different species are characterized by different levels of genetic variability determined by various species-specific parameters, there are no grounds to expect that the number of polymorphic loci found in different species will follow a Poisson distribution, see also (ii) below.

The Poisson distribution is characterized by the well-known relation between mean $\bar{l}$ and variance $\sigma_l^2 \equiv \sum_{k=1}^{N_L}(l_k - \bar{l})^2/N_L$:

$$\sigma_l^2 = \bar{l} \tag{7}$$

Using eqns (3) and (7) and the definition of mean and variance of H_L, $\bar{H}_L \equiv \sum_{k=1}^{N_L} H_L^k/N_L$ and $\sigma_{H_L}^2 \equiv \sum_{k=1}^{N_L}(H_L^k - \bar{H}_L)^2/N_L$, where H_L^k is the expected heterozygosity of the k-th species, we obtain for $\sigma_{H_L}^2$ at fixed L

$$\sigma_{H_L}^2 = \left(\frac{h_{l_{\text{tot}}}}{L}\right)\sigma_l^2 = h_{l_{\text{tot}}}\frac{\bar{H}_L}{L} \tag{8}$$

Note that in (eqn 8) we applied our assumption about equal values of $h_{l_{\text{tot}}}$ in different species, while the assumption about equal values of P in different species was implicitly used in (eqn 6).

Generally, the relationship between $\sigma_{H_L}^2$ and L can be written in the logarithmic form (decimal logarithms were chosen just for reasons of convenience):

$$\log \sigma_{H_L}^2 = a \log \frac{\bar{H}_L}{L} + b \tag{9}$$

Thus, if $h_{l_{\text{tot}}}$, P and, consequently, H are equal in all species, the following equalities should be true according to (eqn 8):

$$a = 1; \quad b = \log h_{l_{\text{tot}}} \tag{10}$$

These predictions were tested with the available empirical data.

Results

(i) Decreasing variance of heterozygosity with increasing number of studied loci

Published data on expected protein heterozygosity (eqn 2) of 321 mammalian species were collected (Fig. 1). In those cases when allele frequencies of polymorphic loci were available, heterozygosity h was calculated for each polymorphic locus according to (eqn 1) (99% criterion of polymorphism). In total, 2003 polymorphic loci were considered. The total number of all loci considered (including monomorphic ones) was 10 296.

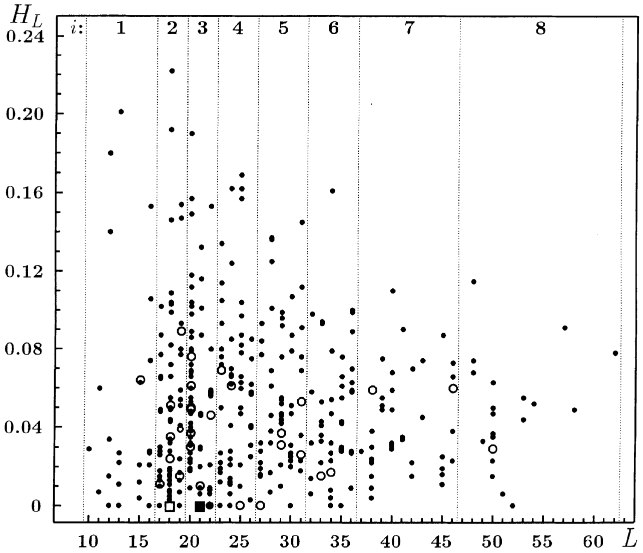

Fig. 1 Protein heterozygosity H_L with respect to the number of studied loci L. Small filled circles represent values for one population each. Large empty circles, large filled circles, the large open square and the large filled square represent randomly coinciding values of heterozygosity for two, three, four and five populations, respectively. Dotted lines divide the total range of L-values into eight intervals. The complete dataset used in this study is available at http://www.private.peterlink.ru/elba.

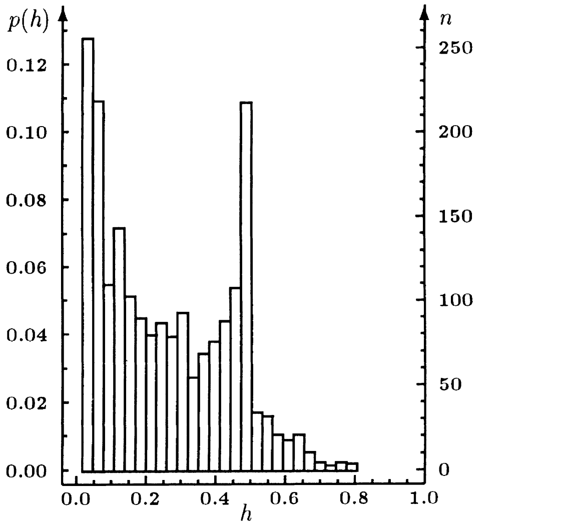

Fig. 2 Probability distribution of heterozygosity h of polymorphic loci in mammals (99% criterion polymorphism). $p(h)$ is the observed relative frequency of polymorphic loci with value of h falling within a given interval of the histogram; n is the absolute number of loci observed in each interval of h-values. $p(h)$ can be interpreted as the probability that a randomly chosen polymorphic locus will have heterozygosity value h within the respective interval. Note that only polymorphic loci are considered. To construct the general distribution including monomorphic loci it is necessary to add one more bar at $h = 0$. This bar will be about 30 times higher than the highest bar of the present histogram. A similar bimodal distribution of h with a maximum at high h-values was obtained theoretically by Fuerst *et al.* (1977) under assumptions of the neutral theory, whereas Altukhov & Dubrova (1981) attributed this maximum to the existence of a special type of gene locus where heterozygosity was maintained by heterosis. The most simple explanation of the observed maximum seems to be that it is an artefact caused by a bias towards choosing more variable loci in heterozygosity studies; see also end of Sect. (i).

The 2003 values obtained for h were used to construct the probability distribution $p(h)$, Fig. 2. Under our assumption of equal values $h_{l_{\mathrm{tot}}}$ of in all species, the mean of this distribution gives a rather accurate estimate of $h_{l_{\mathrm{tot}}}$ due to the great number of studied loci. It is equal to $h_{l_{\mathrm{tot}}} = \sum_{i=1}^{2003} h_i/2003 = 0.261$, where h_i is the expected heterozygosity (eqn 1) of the i-th polymorphic locus. Using this value we obtain an estimate of coefficient b in (eqn 10):

$$b = \log 0.261 = -0.58 \tag{11}$$

Let us now show that the average inaccuracy of the approximate formula (eqn 3), which was used in the Rationale instead of (eqn 2), is not more than 30%. The distribution $p(h)$ is characterized by variance $\sigma_h^2 = 0.034$. The dataset considered (Fig. 1) is characterized by the average number of polymorphic loci l_{av} of about $l_{\mathrm{av}} \approx 5.4$. Expression (4) that characterizes inaccuracy of (3) was obtained under the assumption that the number of polymorphic loci l is the same in all sets of loci studied. It can be shown, however, that for the dataset considered (Fig. 1) where l is a random variable, expression (4) remains adequate as well, if the average number of polymorphic loci l_{av} is used in expression (4) instead of l. Thus, according to expression (4) the standard deviation of h_l at $l \approx l_{\mathrm{av}}$ is equal to

$$\sigma_{h_l} \equiv \sqrt{\sigma^2_{h_l}} = \sqrt{\sigma^2_h/l_{\mathrm{av}}} = \sqrt{0.034/5.4} = 0.079, \tag{12}$$

which constitutes 30% of the average value $h_{l_{\mathrm{tot}}} = 0.261$. Thus, the average deviation of formula (3) from formula (2) is about 30%, as stated above.

The minimum number of studied loci L in a given population was 11, the maximum was 62 (Fig. 1). The range (11, 62) was divided into eight intervals (Fig. 1), each of them containing not less than 20 values of heterozygosity H_L. The division chosen is arbitrary. It was chosen to minimize differences between lengths of intervals and at the same time between numbers of heterozygosity values observed in each interval. It can be shown, however, that the results of the study do not depend on different ways of choosing intervals.

For each interval the average number of studied loci $\bar{L}$, the average heterozygosity $\bar{H}_L$, and variance of heterozygosity $\sigma^2_{H_L}$ were calculated, Table 1. It should be noted that though some species are represented by two or more populations (for 321 species a total of 411 populations were studied), none of the eight intervals of L contains two or more populations of the same species. Thus, in each interval all calculations are made for populations of different species.

Figure 3 shows the results of testing relationship (8) in the logarithmic form (9) with the available data. Variables $\sigma^2_{H_L}$, $\bar{H}_L$ and L in (9) assumed values $\sigma^2_{H_Li}$, $\bar{H}_{Li}$ and $\bar{L}_i$, respectively (i changed from 1 to 8, Table 1).

Linear regression of $\log \sigma^2_{H_L}$ on $\log \bar{H}_L/L$ (eqn 10) gave the following results (Fig. 3):

$$\begin{aligned} &a = 1.10 \pm 0.11; b = 0.17 \pm 0.31; \\ &r = 0.97 \text{ with 6 d.f.}; \; P < 0.0001 \end{aligned} \tag{13}$$

Uncertainty in (13) represents standard errors of the respective values.

The obtained value of a agrees very well with the corresponding relationship in (eqn 10). The high correlation coefficient and low probability level testifies for the statement that variance of heterozygosity really decreases inversely proportionally to L. This means that with growing L heterozygosity values observed in different species converge to a common mean.

The obtained estimate of b corresponds to $h_{l_{\text{tot}}} = 10^{0.17} = 1.19$ (which makes no sense, as $h_{l_{\text{tot}}}$ must be less than unity by definition), exceeds considerably the expected value $b = -0.58$ (eqn 11) and is characterized by high uncertainty. This is an indication of a departure of the distribution of the number of polymorphic loci in real samples from the proposed Poisson distribution (eqn 6). A possible reason for that may be nonrandom sampling of the loci studied. For example, there are many proteins coded by two or more loci. If a given protein is studied, all its loci are studied as a rule. If these loci tend to display similar levels of heterozygosity (which is often the case), they cannot be considered as random samples. This will result in a distortion of the Poisson distribution.

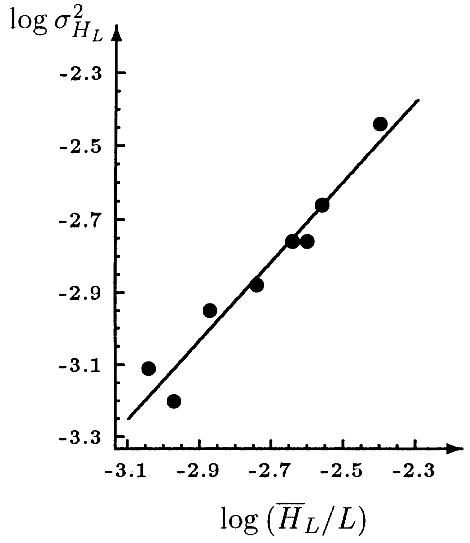

Fig. 3 Dependence of variance $\sigma^2_{H_L}$ of heterozygosity H_L of different species of mammals on the number of studied loci L in the logarithmic form (eqn 9). See eqn 13 for parameters of the regression.

There can be other sources of distortions as well. One of them is an unconscious bias towards more variable loci when choosing a set of loci to be studied. I considered the relationship between protein heterozygosity and the number of studied loci in 30 species of *Drosophila*. Data were taken from Nevo *et al.* (1984). It proved that heterozygosity decreased significantly with the increasing number of studied loci ($r = -0.58$; $P = 0.0002$). This confirms — at least for *Drosophila* studies — the statement that investigators tend to choose more variable loci first (Harris & Hopkinson, 1972; Graur, 1985). One of possible manifestations of this effect can be that papers where higher levels of heterozygosity are described may have more chances to be published than those where no polymorphism is discovered. Then the initial Poisson distribution of the number of polymorphic loci would be distorted due to overly frequent observations of large numbers of polymorphic loci.

Table 1 Numerical data for the eight intervals of Fig. 1

Interval	1	2	3	4	5	6	7	8
N_L	26	76	78	61	62	47	40	21
$\bar{L}$	13.8	18.1	20.6	24.4	29.0	34.2	41.3	51.6
$\bar{H}_L$	0.0554	0.0496	0.0521	0.0564	0.0524	0.0458	0.0440	0.0473
$\sigma^2_{H_L}$	3.63×10^{-3}	2.19×10^{-3}	1.75×10^{-3}	1.75×10^{-3}	1.33×10^{-3}	1.11×10^{-3}	0.63×10^{-3}	0.78×10^{-3}

N_L, the number of species studied in each interval; see text for other designations.

(ii) Testing the assumption of equal heterozygosity in different species of mammals

It can be shown that the observed decrease of variance of heterozygosity with increasing number of studied loci is not compatible with the assumption that different mammalian species have significantly different values of true heterozygosity H, see (eqn 5). Let us assume for simplicity that there are only two types of mammals with different true heterozygosities H_1 and H_2. This means that if we study all protein-coding loci in all mammalian species of one of the two types, we always get the value of either H_1 or H_2. However, if we study randomly chosen small sets of L loci, we will observe two probability distributions of heterozygosity $p_1(H_L)$ and $p_2(H_L)$ with nonzero variances $\sigma_1^2(L)$ and $\sigma_2^2(L)$ and means H_1 and H_2. Let the relative frequency of species of the two types be γ_1 and γ_2, $\gamma_1 + \gamma_2 = 1$. Then for the probability distribution of heterozygosity of all mammalian species $p(H_L)$ we can write

$$p(H_L) = \gamma_1 p_1(H_L) + \gamma_2 p_2(H_L) \tag{14}$$

Expression (14) reflects the fact that in γ_1 cases mammals of the first type are studied and the heterozygosity value follows distribution $p_1(H_L)$, while in the remaining γ_2 cases mammals of the second type are studied and the heterozygosity follows $p_2(H_L)$ distribution.

It can be easily shown in such a case that the variance $\sigma_{H_L}^2$ characterizing probability distribution $p(H_L)$ (eqn 14) is equal to

$$\sigma_{H_L}^2 = \gamma_1 \sigma_1^2(L) + \gamma_2 \sigma_2^2(L) + \gamma_1 \gamma_2 (H_1 - H_2)^2 \tag{15}$$

(To get expression (15) one needs to use the definitions of mean and variance of heterozygosity, $\bar{H}_L \equiv \int_0^1 H_L p(H_L) dH_L$ and $\sigma_{H_L}^2 \equiv \int_0^1 (H_L - \bar{H}_L)^2 p(H_L) dH_L$, and the well-known relation between the variance and mean of any random variable x, $\sigma_x^2 = \overline{x^2} - \bar{x}^2$).

With increasing L the first two terms in (eqn 15) decrease inversely proportionally to L, similarly to (eqn 8). This reflects the fact that with increasing number of studied loci more accurate estimates of H_1 and H_2 are obtained inside the two groups of species. The third term is determined by the difference between true heterozygosities of the two groups of species and does not depend on L. Thus, if the true heterozygosities of the two types of mammals are significantly different (i.e. the third term in eqn 15 is very large as compared to the first two terms), whatever large number of loci L is studied, it will not lead to any considerable decrease of $\sigma_{H_L}^2$. This is because whatever large number of loci is studied in one and the same species, it gives more information only about the mean value of heterozygosity for this particular species. It cannot make the estimate of the mean of all species more accurate if different species have different heterozygosities.

More generally, assuming that all species of mammals have different values of true heterozygosity, one can write for $\sigma_{H_L}^2$

$$\sigma_{H_L}^2 = \lambda \frac{\bar{H}_L}{L} + \sigma_0^2, \tag{16}$$

where λ is an arbitrary constant. Note also that under our assumptions $\bar{H}_L$ does not depend on L either. σ_0^2 is the constant term, which corresponds to the third term in eqn 15 and does not depend on L. Its value reflects the differences between values of true heterozygosity in different species. The greater are the differences, the greater is the value of σ_0^2.

Linear regression of $\sigma_{H_L}^2$ on $\bar{H}_L/L$ (eqn 16) gives the following results:

$$\begin{aligned} &\lambda = 0.91 \pm 0.07;\ \sigma_0^2 = -0.0002 \pm 0.0001; \\ &r = 0.98 \text{ with 6 d.f.};\ P < 0.0001 \end{aligned} \tag{17}$$

The analysed set of heterozygosity values for the 411 populations of mammals (Fig. 1) is characterized by variance σ_H^2 equal to 0.0017. The obtained value of σ_0^2 is negative, differs insignificantly from zero and is nearly an order of magnitude smaller by absolute value compared to σ_H^2. The linear regression giving results (17) is characterized by a high correlation coefficient and is highly significant ($P < 0.0001$). Taken altogether these results allow us to conclude that the available data on the decrease of $\sigma_{H_L}^2$ with increasing L cannot be explained under the assumption of a large value of σ_0^2 and therefore not by significant differences between true heterozygosities of different mammalian species.

Note that in (eqn 9) we used a log-scale when studying the dependence of $\sigma_{H_L}^2$ on $\bar{H}_L/L$, whereas in (eqn 16) a normal scale is used. The fact that, in both cases, linear regression fits the data well (eqn 13, 17) is not a contradiction. The coefficient a in (eqn 9) that determines the power of the $\bar{H}_L/L$ term and might cause deviations from linearity in (eqn 16) is very close to unity, see (eqn 13), while the coefficient σ_0^2 in (eqn 18) that might distort the log-linearity in (eqn 9) is negligibly small, see (eqn 17). Results 13 and 17 actually support our prediction that is $\sigma_{H_L}^2$ directly proportional to $\bar{H}_L/L$ (eqn 8), because only such dependence can be described by linear regression in both normal and log-scales.

It is possible to get an idea of the real differences in values of true heterozygosity in different mammalian species using (eqn 15) and the obtained absolute value of σ_0^2 (eqn 17). If a half of all mammalian species had true heterozygosity H_1 and the other half H_2 (i.e. $\gamma_1 = \gamma_2 = 0.5$) then the difference $H_1 - H_2$ would be equal to

$$H_1 - H_2 = \sqrt{\sigma_0^2/0.5^2} = \sqrt{(|-0.0002|/0.25)} = 0.028$$

Thus, given the average heterozygosity $\bar{H} = 0.051$, values of true heterozygosity of the two types of mammals will differ from the average by not more than 30% ((0.028/2)/0.051 × 100% ≈ 30%).

Note that the result obtained imposes constraints on the variance of the heterozygosity values in the whole class of mammals. It does not exclude the possibility of existence of a small number of mammalian species with heterozygosities noticeably differing from the average. In other words, when either γ_1 or γ_2 is very small, see (eqn 15), the difference $H_1 - H_2$ may become noticeable. In particular, the result obtained does not contradict the statement that large mammals may have noticeably lower values of true heterozygosity compared to the average for the whole class (Wooten & Smith, 1985; Gorshkov & Makarieva, 1997), because species of large mammals constitute but a small part of all mammalian species. Mammals with their body size exceeding 1 m are hardly responsible for more than a few percent of the total number of mammalian species known (Chislenko, 1981; Eisenberg, 1981).

(iii) Poisson distribution of polymorphic loci

It is possible to demonstrate the Poisson distribution of numbers of polymorphic loci (eqn 6) in different species directly. The only obstacle here is the small amount of data obtained at fixed values of L. The number of studied species is maximum when $L = 20$. Twenty loci were studied in 45 populations of different species, which constitutes somewhat more than 10% of the total number of populations considered (411). In order to enlarge the dataset, species with 19 and 21 studied loci were added, so that the three most commonly used values of L (19, 20 and 21) were considered.

The total number of species studied was 82. For each species the total number of studied loci L, the total number of observed polymorphic loci l_{sum}, and number of polymorphic loci belonging to four groups according to their heterozygosity values h were considered, Fig. 4.

Loci with the frequency of the most common allele equalling or exceeding 0.9 were not considered. The number of such loci strongly depends on the number of individuals studied. For example, the probability of not discovering the minor allele of a diallelic polymorphic locus with frequency of the minor allele q when sampling n diploid individuals is equal to $\delta_q = (1 - q)^{2n}$, which at $n = 15$ gives $\delta_{0.05} = 0.21$, $\delta_{0.01} = 0.74$. It means that in samples of 15 individuals the number of polymorphic loci with allele frequencies of 0.05 and 0.01 are underestimated by 21 and 74%, respectively. Meanwhile, in large samples all these loci can be discovered. To minimize possible distortions caused by this effect such weakly polymorphic loci were excluded from the analysis.

Figure 4 shows the results of approximation of the observed distributions of polymorphic loci belonging to different intervals of h by Poisson distribution. The goodness of approximation was estimated by the χ^2-test. Figure 4 makes it clear that the null hypothesis of Poisson distribution cannot be rejected in any of the five cases.

Discussion

In the present study it is shown that the variance of protein heterozygosity between different species of mammals is inversely proportional to the number of the studied loci. This fact suggests that there exists a certain value of protein heterozygosity that is common to most species of mammals.

According to the traditional approach, one would expect intraspecific genetic variability to vary greatly between mammals, since it is determined by numerous factors such as differences in mutation rates between loci within species, global mutation rate differences between species, differences in effective population or actual population size, differences in population histories (bottlenecks), differences in the intensity or mode of selection, etc. It is highly unlikely that such different factors would readjust in such a manner that they yielded the same degree of heterozygosity in all species. Hence, if all these factors are indeed significant for determination of intraspecific variability, then the observation of equal heterozygosity values in different species essentially means that all the above parameters are the same in all species. However, too many factors have to coincide for such an explanation to be considered as reliable.

It seems to be a much more likely explanation that the above factors (some of which are definitely different between many populations of mammals, see below) do not actually affect the levels of variability as strongly as predicted by the majority of selection/mutation/drift models. This point can be illustrated by the following two simple figures displaying different selection profiles, Fig. 5(a,b). Variable n stands for the number of mutational substitutions in the genome, which can be easily related to heterozygosity.

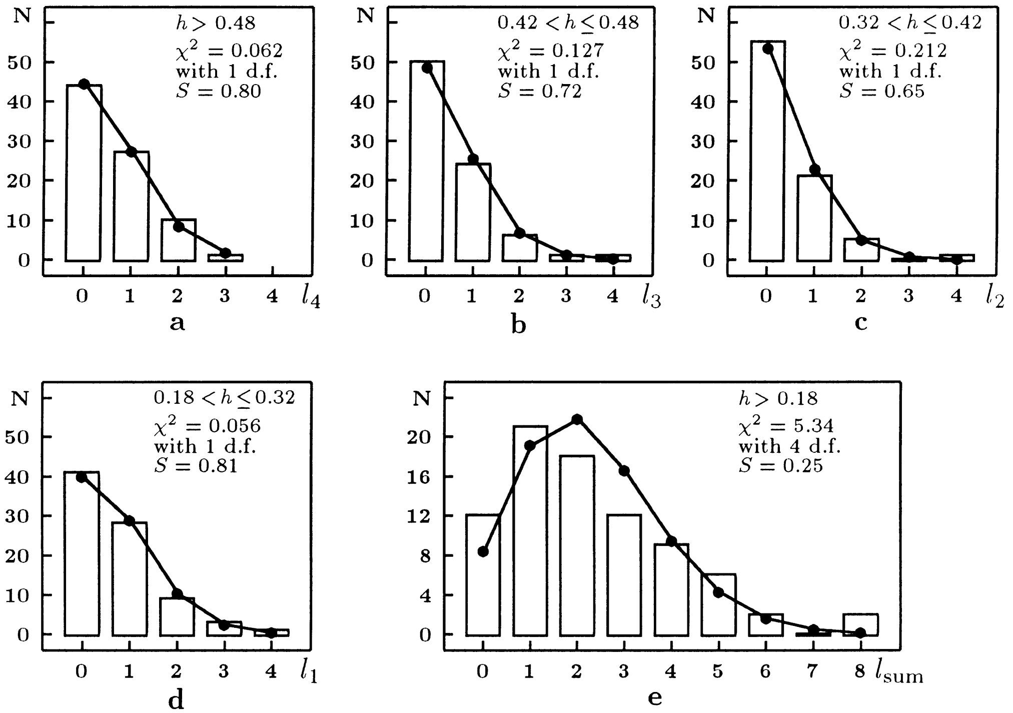

Fig. 4 Approximation of the observed distributions (bars) of numbers of polymorphic loci by Poisson distribution (curves with dots).Variables l_1, l_2, l_3, l_4 and l_{sum} stand for numbers of polymorphic loci with heterozygosity values falling within intervals shown in the upper parts of the diagrams, $l_{sum} = l_1 + l_2 + l_3 + l_4$. S is the significance level of the χ^2-test. N is the number of species with a given number of polymorphic loci. For example, the first bar in Fig. 4e shows that in the studied set of 82 species there are 12 species that do not have polymorphic loci ($l_{sum} = 0$). Meanwhile according to Poisson distribution this number should have been lower (about eight species).

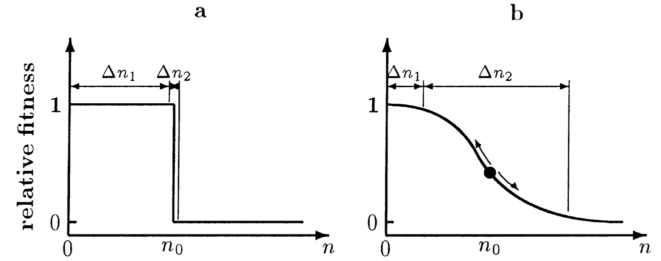

Fig. 5 Different modes of selection. Variable n stands for the number of mutational substitutions in the genome, n_0 represents its equilibrium value.

Figure 5(a) presents a selection profile that is often referred to as 'truncated selection'. The selection profile displayed in Fig. 5(b) is a more common profile, its particular properties in different areas of n-values depending on the mode of selection. In a real situation the mode of selection itself can be a function of n.

In a population where the selection profile conforms to Fig. 5(a), the equilibrium value of intraspecific genetic variability will be determined by $n_0 = \Delta n_1$ irrespective of the actual values of mutation rate and effective population number. On the contrary, in a population conforming to Fig. 5b, the equilibrium value of genetic variability n_0 may vary significantly within a broad corridor $\Delta n_1 < n_0 < \Delta n_1 + \Delta n_2$ depending — in the simplest case — on values of mutation rate and effective population number.

The critical difference between the selection profiles 5a and 5b lies in the ratio of the width of the area where fitness changes significantly, Δn_2, to the width of the area where changes of fitness are relatively small, Δn_1. In Fig. 5a this ratio is equal to zero (fitness changes abruptly) $\Delta n_2/\Delta n_1 = 0$ and $\Delta n_2 \ll \Delta n_1$. On the contrary, for profile 5b $\Delta n_2 \gg \Delta n_1$ and $\Delta n_2/\Delta n_1 \gg 1$.

If we begin to increase the Δn_1 part of profile 5b keeping the width of the Δn_2 part intact, we can finally arrive at a situation where the corridor of possible changes of the equilibrium value of n_0, $\Delta n_1 < n_0 < \Delta n_1 + \Delta n_2$, becomes *relatively* narrow as compared to the absolute value of n_0, *although the absolute width of this corridor may remain arbitrarily large*. In such a situation, factors determining the mutation/selection/drift balance will have only limited impact on the value of n_0, and, consequently, on the equilibrium value of intraspecific variability. This situation will effectively resemble case 5a, although if we again restrict our consideration to the $\Delta n_1 < n_0 < \Delta n_1 + \Delta n_2$ interval we will still detect the influence of the mutation/selection/drift factors. But on a larger scale their impact will be lost. In such a situation the equilibrium value of genetic variability will be determined by the value of Δn_1 with an accuracy of $\Delta n_2/\Delta n_1$.

An important point is that, due to the unprecedented complexity of organization of living objects and their interactions, it is virtually impossible to predict a real selection profile from a mathematical model alone, however complex the latter may be. Parameters of any models can only be deduced from empirical evidence. The study performed here suggests that the real selection profile in mammals is characterized by a relatively wide plateau and a relatively narrow corridor of significant changes of fitness. (In a number of studies a significant influence of newly acquired slightly deleterious mutations on organisms' fitness has been observed (see, e.g. de Visser *et al.*, 1997). This is consistent with the view that the equilibrium value of intraspecific genetic variability is located within a corridor where fitness changes sharply, see Fig. 5(b). Meanwhile, the fact that addition of a tiny number of slightly deleterious mutations (as compared to the number of mutational substitutions already present in the genome) results in a relatively sharp change of fitness, testifies that this corridor is indeed very narrow.)

The next question is what are the fundamental factors that determine the plateau's width and make it nearly the same in all species of mammals? The existence of a common value of protein heterozygosity in most species of mammals can be explained if one considers intraspecific genetic variability as a result of a mutational process that erases meaningful genetic information of a species. The process of information erosion is limited by natural stabilizing selection when the amount of information lost reaches a certain value that is determined by the sensitivity of selection. This critical value is a fundamental characteristic of a species and represents the permissible level of information erosion in natural populations. Evolutionarily close groups of organisms may share this fundamental characteristic, which may be the reason for all mammalian species being characterized by similar values of heterozygosity, close to the average heterozygosity of the whole class, $H = 0.051$.

In its essence, selection is a process of measurement and comparison of certain phenotypic traits of individuals that compete with each other within a population. The existence of a limit of sensitivity of a certain process of measurement is a very general phenomenon. For example, it is impossible to weigh small loads (e.g. milligrams) using scales that are calibrated in kilograms. Such scales do not 'discern' differences between small loads. Similarly, if the process of selection is 'calibrated' in thousands of mutational substitutions, selection will not be able to tell apart individuals differing from each other by two or three hundreds of substitutions. Due to the continuous process of mutagenesis, all individuals will finally accumulate a number of substitutions of the order of the value of selection sensitivity. (Note that sensitivity of selection cannot be quantified as a simple number of mutational substitutions tolerated in a genome. Rather, each substitution should be weighted according to the degree it affects normal functioning of the organism. But it is possible to speak about an average number of mutational substitutions of average 'deleteriousness').

The sensitivity of any process of measurement is determined by the properties of the applied instrument (e.g. scales) and the process itself. Similarly, sensitivity of selection is a fundamental property of the organization of life and cannot be quantified *a priori*. The present work provides an insight into the problem of quantitative analysis of this important characteristic of living organisms.

In addition to the general picture outlined above (Fig. 5a, b), it is possible to perform a more detailed quantitative analysis of one factor that has been traditionally considered to be one of the decisive ones in determining the levels of intraspecific variability, namely the effective population size, N_e (Amos & Harwood, 1998). Due to large difficulties in estimating effective population sizes of various species, information about this parameter is derived by comparing actual population sizes of species, see, e.g. (Nei & Graur, 1984).

The overwhelming majority of terrestrial mammalian species are characterized by average body masses of the order of $m \approx 1$ kg or less, which corresponds to a linear body size of the order of $x \approx (m/\rho)^{1/3} \approx 0.1$ m, where $\rho \approx 10^3$ kg m^{-3} is the living mass density, see Fig. 6, which presents data of Chislenko (1981) and Eisenberg (1981). Note the logarithmic scale of the density of species numbers $N(y)$.

Thus, an average terrestrial mammal has a linear body size of the order of $x \approx 10$ cm and is characterized by protein heterozygosity $H \approx 0.051$. Let us for simplicity restrict our consideration to purely neutral variability using the well-known formula $H = 4v_gN_e/(1 + 4v_gN_e)$, where v_g is the mutation rate per gene per generation. At $H \ll 1$ (which is the case) the heterozygosity is simply proportional to v_gN_e.

It is known that for herbivorous mammals the cumulative species biomass B grows approximately proportionally to body size x (Damuth, 1981), $B \propto x$. As far as body mass of an individual animal is proportional to the third power of its linear body size, $m \propto x^3$, this means that a species' population number

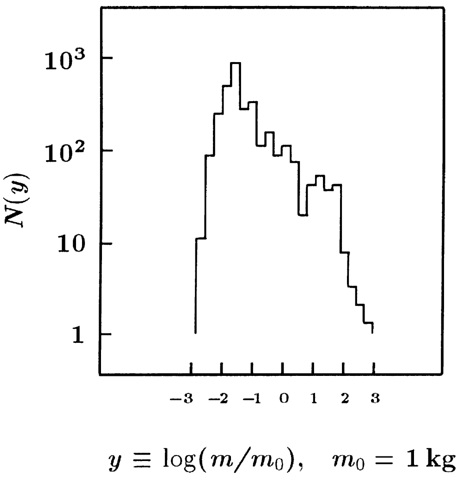

Fig. 6 Size spectrum of terrestrial herbivorous mammalian species. $N(y)$ is the density of the number of species per unit of the relative body mass interval Δy. $y \equiv \log(m/m_0)$, where m is body mass and $m_0 = 1$ kg.

decreases with body size of mammals as $N = B/m \propto x^{-2}$. Thus, population numbers of large herbivorous mammals with $x \approx 1$ m would be a hundred times smaller than those of average mammals with $x \approx 0.1$ m. When we turn to carnivores who reside at the top of the trophic pyramid (Odum, 1983), we conclude that population numbers of large carnivores would be about 10 times lower than population numbers of herbivores of the same body size. (This is due to the fact that the energy content of production of animal biomass does not generally exceed 10% of the energy content of the consumed food). Thus, the difference in population numbers between large carnivores and average mammals (which are small and herbivorous) constitutes three orders of magnitude. Note that this difference, caused by fundamental biochemical and ecological regularities, is unlikely to decrease during bottleneck events.

Were the effective population sizes to change proportionally to real population sizes, then, according to Nei's formula, we would expect heterozygosity values in carnivores to be of the order of $0.051 \times 10^{-3} = 5.1 \times 10^{-5}$. In the sample of 411 populations considered in this study carnivores accounted for 24 populations with an average heterozygosity value of $H_{\mathrm{carn}} = 0.032$, which corresponds to 63% of the global average $\bar{H} = 0.051$.

Whatever the real dependence between the effective and actual population sizes could be, it is very unlikely that a three orders of magnitude change in real population numbers of mammalian species scales to a 37% reduction of the effective population number. This calls for looking for different reasons that could account for the observed slight decrease of heterozygosity values in a small number of large mammalian species. Note also that this slight decrease does not create a distortion from the first order effect of constant heterozygosity observed in this study.

In conclusion, I would like to add that although the present study has focused on allozyme variability in mammals, the proposed approach can be readily applied to analysis of significance of the observed interspecific differences in values of genetic variability at the nucleotide or even microsatellite level as well, provided a sufficiently large dataset for different species has been created.

Acknowledgements

I am extremely grateful to Dr Victor Gorshkov for the problem's setting, discussions and inspiration; to Dr Laurence Hurst and Dr Jos van Damme for encouragement and invaluable comments on an early version of the paper; to Dr Maria Filipucci and Dr Shiro Wada for providing me with sources of data unavailable in my country. Special thanks are due to two anonymous referees and Dr John Brookfield, whose detailed critical analysis of the paper has greatly assisted the author in clarifying many important issues. The work was supported by the Ministry of Natural Resources of the Russian Federation, the Research Support Scheme of the Open Society Support Foundation, grant No. 800/2000, and the European Science Foundation Scientific Programme on Theoretical Biology of Adaptation.

References

ALTUKHOV, I. U. P. AND DUBROVA, I. U. E. 1981. Biochemical polymorphism of populations and its biological significance. *Prog. Modern Biol. (Uspekhi Sovremennoj Biologii)*, **91**, 467–480 (in Russian).

AMOS, W. AND HARWOOD, J. 1998. Factors affecting levels of genetic diversity in natural populations. *Phil. Trans. R. Soc. B*, **353**, 177–186.

APPOLONIO, M. AND HARTL, G. B. 1993. Are biochemical-genetic variation and mating systems related in large mammals? *Acta Theriol.*, **38** (Suppl. 2), 175–185.

AYALA, F. J. AND FITCH, W. M. 1997. Genetics and the origin of species: an introduction. *Proc. Nat. Acad. Sci. U.S.A.*, **94**, 7691–7697.

BURROWS, W. AND RYDER, O. A. 1997. Y-chromosome variation in great apes. *Nature*, **385**, 125–126.

BUTLIN, R. K. AND TREGENZA, T. 1998. Levels of genetic polymorphism: marker loci versus quantitative traits. *Phil. Trans. R. Soc. B*, **353**, 187–198.

CARSON, H. L. 1990. Increased genetic variation after a population bottleneck. *Trends Ecol. Evol.*, **5**, 228–230.

CHISLENKO, L. L. 1981. *Structure of Flora and Fauna as Related to Body Size of Organisms*. Moscow University Press, Moscow (in Russian).

CRESPI, B. J. 1991. Heterozygosity in the haplodiploid Thysanoptera. *Evolution*, **45**, 458–464.

DAMUTH, J. 1981. Population density and body size in mammals. *Nature*, **290**, 699–700.

EISENBERG, J. 1981. *The Mammalian Radiations: An Analysis of Trends in Evolution, Adaptation and Behaviour*. Athlone Press, London.

FUERST, P. A., CHAKRABORTY, R. AND NEI, M. 1977. Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. *Genetics*, **86**, 455–483.

GAINES, M. S., DIFFENDORFER, J. E., TAMARIN, R. H. AND WHITTAM, T. S. 1997. The effects of habitat fragmentation on the genetic structure of small mammal populations. *J. Hered.*, **88**, 294–304.

GORSHKOV, V. G. AND MAKARIEVA, A. M. 1997. Dependence of heterozygosity on body weight in mammals. *Proc. Russian Acad. Sci. (Dokl. Akad. Nauk)*, **355**, 418–421 (in Russian).

GRAUR, D. 1985. Gene diversity in Hymenoptera. *Evolution*, **39**, 190–199.

HARRIS, H. AND HOPKINSON, D. A. 1972. Average heterozygosity per locus in man: an estimate based on the incidence of enzyme polymorphisms. *Ann. Hum. Genet.*, **36**, 9–19.

HOFKER, M. H., SCRAASTAD, M. I., BERGEN, A. A. B., WAPENAAR, M. C. *ET AL*. 1986. The X chromosome shows less genetic variation at restriction sites than the autosomes. *Am. J. Hum. Genet.*, **39**, 438–451.

IMASHEVA, A. G. 1999. Environmental stress and genetic variation in animal populations. *Genetika*, **35**, 421–431 (in Russian).

KIMURA, M. 1991. The neutral theory of molecular evolution: a review of recent evidence. *Jap. J. Genet.*, **66**, 367–386.

LARSON, A. 1981. A re-evaluation of the relationship between genome size and genetic variation. *Am. Nat.*, **118**, 119–125.

NEI, M. 1984. Genetic polymorphism and neomutationism. *Lecture Notes Biomath.*, **53**, 214–241.

NEI, M. AND GRAUR, D. 1984. Extent of protein polymorphism and the neutral mutation theory. *Evol. Biol.*, **17**, 73–118.

NEVO, E., BEILES, A. AND BEN-SHLOMO, R. 1984. The evolutionary significance of genetic diversity: ecological, demographic and life history correlates. *Lecture Notes Biomath.*, **53**, 13–213.

O'BRIEN, S. J., GAIL, M. H. AND LEVIN, D. L. 1980. Correlative genetic variation in natural populations of cats, mice and men. *Nature*, **288**, 580–583.

O'BRIEN, S. J., WILDT, D. E., BUSH, M., CARO, T. M., FITZGIBBON, C., AGGUNDEY, I. AND LEAKEY, R. E. 1987. East African Cheetahs: evidence for two population bottlenecks? *Proc. Natl. Acad. Sci.*, **84**, 508–511.

ODUM, E. P. 1983. *Basic Ecology*. Saunders College Publications, New York.

PIERCE, B. A. AND MITTON, J. B. 1980. The relationship between genome size and genetic variation. *Am. Nat.*, **116**, 850–861.

THOMPSON, R. C. AND LYMBERY, A. J. 1996. Genetic variability in parasites and host–parasite interactions. *Parasitology*, **112** (Suppl.), S7–S22.

DE VISSER, J. A., HOEKSTRA, R. F. AND VAN DEN ENDE, H. 1997. An experimental test for synergistic epistasis and its application in *Chlamydomonas*. *Genetics*, **145**, 815–819.

WARD, R. D., SKIDINSKI, O. F. AND WOODMARK, M. 1992. Protein heterozygosity, protein structure and taxonomic differentiation. *Evol. Biol.*, **26**, 73–159.

WOOTEN, M. C. AND SMITH, M. H. 1985. Large mammals are genetically less variable? *Evolution*, **39**, 210–212.