Int. J. Appl. Math. Comput. Sci., 2003, Vol. 13, No. 3, 385–394

SAMPLING PROPERTIES OF ESTIMATORS OF NUCLEOTIDE DIVERSITY AT DISCOVERED SNP SITES

ALEXANDER RENWICK*, PENELOPE E. BONNEN**, DIMITRA TRIKKA**, DAVID L. NELSON**
RANAJIT CHAKRABORTY***, MAREK KIMMEL*

* Department of Statistics, Rice University
6100 Main Street, Mail Stop 138
Houston, TX 77005, USA
e-mail: renwick@stat.rice.edu, kimmel@rice.edu

**Department of Molecular and Human Genetics
Baylor College of Medicine, Houston, TX, USA
e-mail: nelson@bcm.tmc.edu

***Center for Genome Information
University of Cincinnati, Cincinnati, OH, USA
e-mail: ranajit.chakraborty@uc.edu

SNP sites are generally discovered by sequencing regions of the human genome in a limited number of individuals. This may leave SNP sites present in the region, but containing rare mutant nucleotides, undetected. Consequently, estimates of nucleotide diversity obtained from assays of detected SNP sites are biased. In this research we present a statistical model of the SNP discovery process, which is used to evaluate the extent of this bias. This model involves the symmetric Beta distribution of variant frequencies at SNP sites, with an additional probability that there is no SNP at any given site. Under this model of allele frequency distributions at SNP sites, we show that nucleotide diversity is always underestimated. However, the extent of bias does not seem to exceed 10–15% for the analyzed data. We find that our model of allele frequency distributions at SNP sites is consistent with SNP statistics derived based on new SNP data at ATM, BLM, RQL and WRN gene regions. The application of the theory to these new SNP data as well as to the literature data at the LPL gene region indicates that in spite of ascertainment biases, the observed differences of nucleotide diversity across these gene regions are real. This provides interesting evidence concerning the heterogeneity of the rates of nucleotide substitution across the genome.

Keywords: single nucleotide polymorphisms, ascertainment bias, nucleotide diversity, molecular evolution

1. Introduction

Single Nucleotide Polymorphisms (SNPs) are variants of the DNA sequence arising when one nucleotide has been substituted for another. They typically are observed to be bi-allelic, which is frequently interpreted as suggesting that each SNP in the population may result from a unique mutational event (Clark *et al.*, 1998). Recent sequencing projects find SNPs to be ubiquitous throughout the human genome, so that SNPs may become a powerful tool in helping to locate genetic loci responsible for complex phenotypes (Cargill *et al.*, 1999; Halushka *et al.*, 1998).

In applications, the most useful SNPs are those in which the minor (rarer) allele is relatively common. While the discovery of such SNPs is the aim of many studies, the data these studies generate allow employing and testing theories concerning the interpretation of the genetic diversity of DNA sequences. Two measures of the genetic diversity of particular interest are π, the average heterozygosity, and s, the number of segregating (polymorphic) sites divided by a normalizing factor $a_n = \sum_{i=1}^{n-1} (1/i)$, where n is the number of chromosomes in the sample. For our purposes, it seems more convenient to further normalize π and s, by dividing them by the number l of DNA sites in the sequence.

Models of SNP evolution. Under the Fisher-Wright model of the drift and the Infinite Sites Model of mutation, the latter assuming that each site of a sequence undergoes at most one mutation in the history of the population, the expected values of π and s are equal to θ, here defined

as four times the product of the effective population size and the average mutation rate per site (Ewens, 1979), i.e.,

$$E(\pi) = E(s) = \theta.$$

The Infinite Sites Model is questionnable if there is a possibility of recurrent and/or reversible mutations at SNP sites. It can serve as a useful approximation. Such an approach was taken by Eberle and Kruglyak (2000). However, it also is possible to use a parametric model of the distribution of SNP frequencies, without a reference to a particular evolutionary mechanism, provided that such a model explains the data. If a sufficiently satisfactory fit is obtained, it is then possible to speculate what assumptions may lead to the parametric model. We will proceed in this way in this paper. From a comparison with data, it seems that the distribution we use provides a satisfactory fit.

Biased sampling. The primary motivation for SNP identification is its use in genetic mapping. To be useful for this purpose, the minor (rarer) allele must be of a sufficiently high frequency. This motivation, combined with the high cost of fully sequencing large chromosomal regions, leads some investigators (Bonnen *et al.*, 2000, Trikka *et al.*, 2002) to the following scheme for SNP discovery:

1. Obtain a small screening sample of (k) chromosomes.
2. Sequence the region of interest (of length l) in each of the k screening sample chromosomes.
3. Identify sites exhibiting polymorphism (of which there are m).
4. Obtain a large test sample of (n) chromosomes.
5. Probe each of the n test chromosomes only at the m sites found to be polymorphic in the initial screening sample.

This method is likely to find SNPs where the less frequent allele is sufficiently common to be useful. However, it introduces a bias, which reduces the apparent frequency of sites where the minor allele is rare. The question we address is: What effect does this biased sampling scheme have on the sample statistics π and s?

2. Modeling the SNP Statistics and the Ascertainment Bias

Statistics of nucleotide diversity. In keeping with the literature, we treat π and s as normalized for the sequence length. We can represent π and s as summations across sites. This helps to clarify the effect of limited screening and simplifies the calculation of expectations:

$$\pi = \frac{1}{l}\frac{2n}{n-1}\sum_{i=1}^{l} x_i(1-x_i) = \sum_{i=1}^{l}\pi_i,$$

and

$$s = \frac{\text{Number of polymorphic sites}}{a_n l} = \sum_{i=1}^{l} s_i,$$

where

$$\pi_i = \frac{1}{l}\frac{2n}{n-1}x_i(1-x_i),$$

x_i denoting the fraction of the wild-type variant in the population, and

$$s_i = \begin{cases} \dfrac{1}{l}\dfrac{1}{a_n} & \text{if the } i\text{-th site is polymorphic,} \\ 0 & \text{if the } i\text{-th site is monomorphic.} \end{cases}$$

We model $\{\pi_i, i = 1,\ldots,l\}$ and $\{s_i, i = 1,\ldots,l\}$ as sequences of identically distributed, although not independent, random variables.

Indices of the bias. As has been mentioned in Introduction, the usual method of SNP discovery results in under-representation of rare SNP variants, which leads to underestimation of the measures of nucleotide diversity π and s. One remedy to this problem is to determine the factor by which the diversity existing in a sample of size n will be decreased by using a screening sample of size k. This factor will be called the index of bias. In this section, we will define two indices of bias, $B(\pi\,|\,k)$ and $B(s\,|\,k)$, for π and s, respectively. The bias index $B(\pi\,|\,k)$ is equal to the ratio of the expected π value (whether or not the SNP is discovered) to the expected value of π conditional on the SNP being discovered. The bias index $B(s\,|\,k)$ is equal to the ratio of the expected s value (whether or not the SNP is discovered) to the expected value of s conditional on the SNP being discovered. Therefore, multiplication by the bias index $B(\pi\,|\,k)$ (respectively $B(s\,|\,k)$) of the estimate of π (resp. s) conditional on the SNP being discovered results in an estimate corrected for the discovery bias.

For the purpose of bias correction, we will assume that at each site two variants occur, one with frequency X, the other with frequency $1-X$, where X is a random variable with distribution $F(x)$. Using the frequent convention, we will denote random variables using capital letters, and their values (realizations) using lower-case letters. If most mass of $F(x)$ is concentrated close to $x = 0$ or $x = 1$, then, with a high probability, only one variant is observed. We will assume that $F(x)$ is symmetric, essentially meaning that $F(x) = 1 - F(1-x)$. Symmetry arises from the inability to distinguish the wild type and mutant alleles at an SNP locus. Sometimes, this loss of information can be remedied by using data from an outgroup and assuming the variant present in the outgroup to be the wild type. We do not consider outgroup data here. We will assume that each site can be monomorphic with

probability p, so the distribution can be written as

$$F(x) = \frac{p}{2} + (1-p)\int_0^x f(u)\,\mathrm{d}u, \quad x \in [0,1),$$

where $f(x)$ is a normed density. The parameter p plays an important role in the theory, in that it defines the proportion of sites that will not turn out polymorphic even in an "infinite" sample.

Suppose that SNPs are discovered by sequencing a sample of k chromosomes. With probability $X^k + (1-X)^k$, given X, only one variant is observed in this sample. Therefore, conditional on observing a single variant at a given site in the discovery sample, i.e., conditionally on not discovering an SNP, the density of X at this site is equal to

$$\begin{aligned} f_0(x \mid k) &= \frac{\frac{p}{2}\left[\delta(x)+\delta(1-x)\right]+(1-p)\left[x^k+(1-x)^k\right] f(x)}{p+(1-p)\int_0^1 \left[x^k+(1-x)^k\right] f(x)\,\mathrm{d}x} \\ &= \frac{\frac{p}{2}\left[\delta(x)+\delta(1-x)\right]+(1-p)\left[x^k+(1-x)^k\right] f(x)}{E\left[X^k+(1-X)^k\right]}, \end{aligned}$$

where $\delta(x)$ is the Dirac pseudo-function. Analogously, conditional on observing more than one variant, i.e., on discovering an SNP, the density of X at the site is equal to

$$f_1(x \mid k) = \frac{\left[1-x^k-(1-x)^k\right] f(x)}{E\left[1-X^k-(1-X)^k\right]}.$$

If we calculate sample estimates of π and s using the m sites (out of the total of l sites in the sequence locus) at which SNPs were discovered based on k chromosomes, the expected values of such statistics will be equal to

$$\begin{aligned} E_1(\pi \mid k) &= \frac{m}{l}\frac{2n}{n-1}\int_0^1 x(1-x) f_1(x \mid k)\,\mathrm{d}x \\ &= \frac{m}{l}\frac{2n}{n-1}\frac{E\left\{X(1-X)\left[1-X^k-(1-X)^k\right]\right\}}{E\left[1-X^k-(1-X)^k\right]}, \end{aligned}$$

$$\begin{aligned} E_1(s \mid k) &= \frac{m}{l}\frac{1}{a_n}\int_0^1 \left[1-x^n-(1-x)^n\right] f_1(x \mid k)\,\mathrm{d}x \\ &= \frac{m}{l}\frac{1}{a_n}\frac{E\{[1-X^n-(1-X)^n][1-X^k-(1-X)^k]\}}{E\left[1-X^k-(1-X)^k\right]}. \end{aligned}$$

Similarly, expected π and s from sites, at which no SNPs were discovered, are equal to

$$\begin{aligned} E_0(\pi \mid k) &= \frac{l-m}{l}\frac{2n}{n-1} \\ &\quad \times \frac{E\{X(1-X)[X^k+(1-X)^k]\}}{E[X^k+(1-X)^k]}, \\ E_0(s \mid k) &= \frac{l-m}{l}\frac{1}{a_n} \\ &\quad \times \frac{E\{[1-X^n-(1-X)^n][X^k+(1-X)^k]\}}{E[X^k+(1-X)^k]}. \end{aligned}$$

Despite being defined "per site", as indicated by the factor l in the denominators, symbols $E_0(\cdot)$ and $E_1(\cdot)$ are additive, i.e., they concern two disjoint classes of sites, Class 0, in which no polymorphic SNPs were discovered, and Class 1, in which polymorphic SNPs were discovered. Therefore, aggregate expected estimates $E(\pi \mid k) = E_0(\pi \mid k) + E_1(\pi \mid k)$ and $E(s \mid k) = E_0(s \mid k) + E_1(s \mid k)$ may be computed if $F(x)$ is known. Then the biased estimates of π and s, based on the discovery sample, can be multiplied by bias indices

$$B(\pi \mid k) = 1 + \frac{E_0(\pi \mid k)}{E_1(\pi \mid k)},$$

$$B(s \mid k) = 1 + \frac{E_0(s \mid k)}{E_1(s \mid k)},$$

resulting in unbiased estimates. However, since $F(x)$ is not known, we will consider bounds on $B(\pi \mid k)$ and $B(s \mid k)$ over $F(x)$ belonging to a plausible family of distributions.

Let us notice that the aggregate expectations $E(\pi \mid k)$ and $E(s \mid k)$ are not exactly equal to the unconditional expectations $E(\pi)$ and $E(s)$. The reason is that the observed number m of SNPs is used in $B(\pi \mid k)$ and $B(s \mid k)$, However, if the sample value m is replaced by its "infinite sample" expectation $E(m) = lE[1-X^k-(1-X)^k]$, we obtain $E(\pi \mid k) = E(\pi)$ and $E(s \mid k) = E(s)$.

Model: Symmetric Beta distribution. We assume that the frequency spectrum from which allele frequencies are drawn is a symmetric beta distribution, modified by allowing a site to be monomorphic with probability p,

$$\begin{aligned} f(x) &= \frac{p}{2}\left[\delta(x)+\delta(1-x)\right] \\ &\quad + (1-p)\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}[x(1-x)]^{\alpha-1}, \quad \alpha > 0, \end{aligned}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\exp(-x)\,\mathrm{d}x$ is the Euler Gamma function. We are using a flexible family like Beta, since in this way we can flexibly shape the frequency spectra. We are using a parametric model since, arguably, there is currently no other simple theory that would allow deriving the frequency spectrum from first principles. The Infinite Sites Model (Clark *et al.*, 1998], advocated for SNPs, seems insufficient to explain SNP diversity, as demonstrated in Fig. 14 of Venter *et al.* (2001).

The continuous part of this distribution is U-shaped if $\alpha \in (0,1)$, uniform if $\alpha = 1$, and unimodal if $\alpha > 1$. The expected values of diversities π and s are equal to

$$E(\pi) = (1-p)\frac{n}{n-1}\frac{\alpha}{2\alpha+1},$$

$$E(s) = (1-p)\frac{1}{a_n}\left[1 - 2\frac{\Gamma(2\alpha)\Gamma(\alpha+n)}{\Gamma(2\alpha+n)\Gamma(\alpha)}\right],$$

respectively.

The symmetric Beta model can be treated as arising when forward and backward mutations between two variants, with identical rates, are possible at each site (Ewens, 1979, p. 139) if the site can mutate at all. This mechanism is different from the Infinite Sites Model. We will return to this subject in Discussion.

We can determine the expression for the cumulative distribution $F_1(x\,|\,k)$ of X given SNP discovery, by integrating the density $f_1(x\,|\,k)$. Furthermore, we can determine the cumulative distribution $G_1(y\,|\,k)$ of the random variable Y, equal to the frequency of the less frequent allele, from the formula $G_1(y\,|\,k) = 1 + F_1(y\,|\,k) - F_1(1-y\,|\,k)$, $y \in [0, 1/2]$. This is useful, since in SNP data it is frequently not known which variant is the wild type and which is the mutant. Figure 1 depicts numerically obtained plots of $G_1(y\,|\,k)$ for $k = 10$ and three values of α. The plot is concave for $\alpha \in (0,1)$, almost linear for $\alpha = 1$, and convex for $\alpha > 1$.

Expressions for the conditional expectations of π and s given in Table 1 result from the Beta distribution model.

Figure 2 depicts the impact of a limited discovery sample on biases of diversity, for a range of values $\alpha \in [0, 10]$ and for $p = 0.9, 0.99, 0.999$ in panels (a), (b), and (c), respectively. Other parameters were kept at values corresponding to those true for the data from our observations (see next section), $l = 13,000$, $m = 10$, $k = 10$,

Table 1. Conditional expectations of π and s.

$$E_0(\pi\,|\,k) = \frac{l-m}{l}\frac{2n}{n-1}\frac{2(1-p)\alpha\dfrac{\Gamma(2\alpha)\Gamma(\alpha+k+1)}{\Gamma(\alpha)\Gamma(2\alpha+k+2)}}{p + 2(1-p)\dfrac{\Gamma(2\alpha)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(2\alpha+k)}},$$

$$E_0(s\,|\,k) = \frac{l-m}{l}\frac{1}{a_n}\frac{2(1-p)\dfrac{\Gamma(2\alpha)}{\Gamma(\alpha)}\left[\dfrac{\Gamma(\alpha+k)}{\Gamma(2\alpha+k)} - \dfrac{\Gamma(\alpha+k)\Gamma(\alpha+n)+\Gamma(\alpha)\Gamma(\alpha+n+k)}{\Gamma(\alpha)\Gamma(2\alpha+k+n)}\right]}{p + 2(1-p)\dfrac{\Gamma(2\alpha)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(2\alpha+k)}},$$

$$E_1(\pi\,|\,k) = \frac{m}{l}\frac{2n}{n-1}\frac{\alpha\left[\dfrac{1}{2(2\alpha+1)} - 2\dfrac{\Gamma(2\alpha)\Gamma(\alpha+k+1)}{\Gamma(\alpha)\Gamma(2\alpha+k+2)}\right]}{1 - 2\dfrac{\Gamma(2\alpha)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(2\alpha+k)}},$$

$$E_1(s\,|\,k) = \frac{m}{l}\frac{1}{a_n}\frac{1 - 2\dfrac{\Gamma(2\alpha)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(2\alpha+k)} - 2\dfrac{\Gamma(2\alpha)\Gamma(\alpha+n)}{\Gamma(\alpha)\Gamma(2\alpha+n)} + 2\dfrac{\Gamma(2\alpha)\Gamma(\alpha+k)\Gamma(\alpha+n)+\Gamma(2\alpha)\Gamma(\alpha)\Gamma(\alpha+n+k)}{\Gamma(\alpha)^2\Gamma(2\alpha+n+k)}}{1 - 2\dfrac{\Gamma(2\alpha)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(2\alpha+k)}}.$$

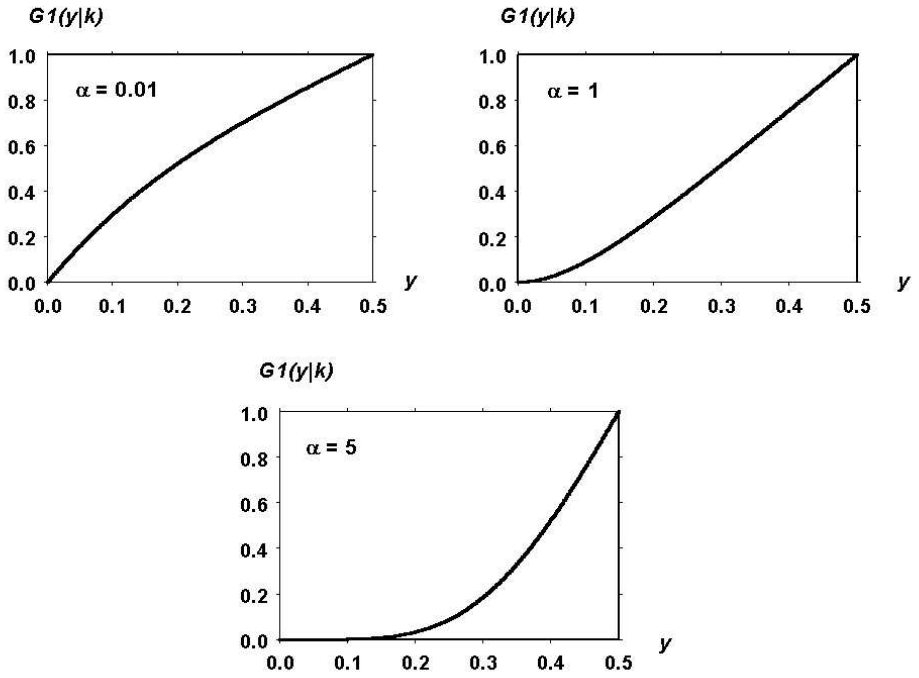

Fig. 1. Numerically obtained empirical distribution functions of the frequency $G_1(y \mid k)$ of the less frequent variant at an SNP site ($k = 10$ and $\alpha = 0.01, 1, 5$).

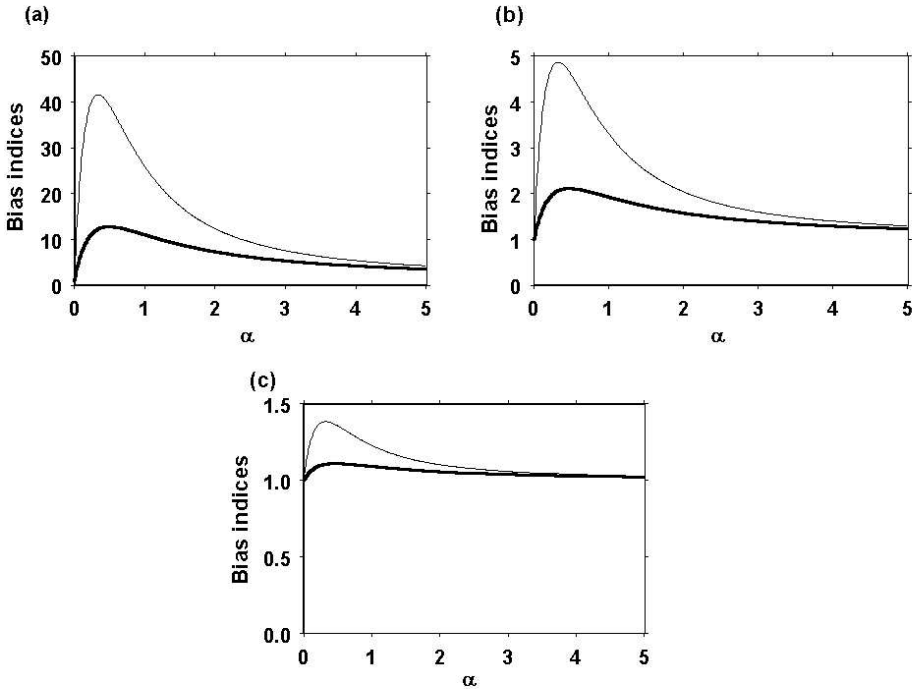

Fig. 2. Ascertainment bias corrections for diversity estimates $\hat{\pi}$ (thick lines) and $\hat{s}$ (thin lines), calculated for a range of values of α and $p = 0.9, 0.99, 0.999$ (panels (a), (b) and (c), respectively). The remaining parameters are $l = 13,000$, $m = 10$, $k = 10$, and $n = 150$.

except for $n = 150$, which is less than our total sample size (see below), but the largest value at which numerics do not diverge. Note that bias $B(\pi \mid k)$ does not depend on n. Also, we demonstrated computationally that bias $B(s \mid k)$ is insensitive to changes in $n > 100$. In general, s is more sensitive to the effect of limited screening than π. Decreasing p below 0.99 also leads to a considerable $B(\pi \mid k)$ bias.

3. Estimates of the Parameters of the Model and of Diversity in SNP Data

Data. We use estimates of nucleotide diversity from a study of SNPs at cancer-related loci, conducted by ourselves (Bonnen *et al.*, 2000; Trikka *et al.*, 2002). In addition, we identified two literature sources, which report estimates of genetic diversity (Halushka *et al.*, 1999; Nickerson *et al.*, 1998).

SNPs in our studies (Bonnen *et al.*, 2000; Trikka *et al.*, 2002) were discovered by sequencing regions of $k = 10$ chromosomes, belonging to 5 Caucasian individuals. For each SNP discovered in this way, a molecular probe was developed and the SNP was typed in almost all cases in 295 individuals (71 African American, 39 Asian American, 77 Caucasian American, 73 Hispanic and 35 CEPH Caucasian). This gives $n = 590$. Sequenced regions involved non-coding regions of 4 genes, with potential impact in familial cancers: Ataxia telangiectasia (ATM), Bloom's syndrome (BLM), RecQL (RQL) and Werner's syndrome (WRN). For ATM, 14 SNPs were discovered in regions of total length 13.5 kb. For BLM, 8 SNPs were discovered in regions of total length 13.7 kb. For RQL, 11 SNPs were discovered in regions of total length 12.6 kb. For WRN, 12 SNPs were discovered in regions of total length 14.0 kb (Table 2).

SNPs in (Halushka *et al.*, 1999) were discovered by sequencing DNA samples from 40 Africans and 34 Americans of Northern European descent. This gives $n = k = 148$. In total, 190 kb of DNA was analyzed, covering coding sequences, introns and 3' and 5' untranslated regions of 75 candidate genes for blood pressure homeostasis and hypertension. Total of 874 SNPs were discovered, 387 of them in the coding sequence.

SNPs in (Nickerson *et al.*, 1998) were discovered by sequencing DNA samples from 24 African-Americans from Jackson, MS (USA), 24 Europeans from North Karelia (Finland), and 23 European-Americans from Rochester, MN (USA). This gives $n = k = 142$. In total, 9.7 kb of DNA was analyzed, covering a fraction of the lipoprotein lipase (LPL) gene, a candidate gene for the susceptibility to the cardiovascular disease. A total of 88 SNPs were discovered (including 9 insertion/deletion variants).

Table 2. Numbers of SNP discovered in the BLM, WRN, RQL and ATM gene regions (m), the corresponding sequence lengths (l), and estimates of diversity $\hat{\pi}$ and $\hat{s}$, and those of model parameters $\hat{\alpha}$ and $\hat{p}(\hat{\alpha})$.

Gene	Number of discovered SNPs (m)	Sequence length (l)	$\hat{\pi} \times 10^4$	$\hat{s} \times 10^4$	$\hat{\alpha}$	$\hat{p}(\hat{\alpha})$
BLM	8	13.7 kb	2.4	1.4	1.8	0.9994
WRN	12	14.0 kb	2.5	2.1	0.01	0.9990
RQL	11	12.6 kb	3.5	1.7	1.7	0.9990
ATM	14	13.5 kb	4.7	2.0	5.0	0.9994

Fitting the symmetric Beta model. The described procedure is based on the observation that expected diversities π and s, given discovery, i.e., $E_1(\pi \mid k)$ and $E_1(s \mid k)$ do not depend on parameter p (the probability that the site is invariant). Therefore, the diversities $\hat{\pi}$ and $\hat{s}$, estimated based on SNPs discovered in the sample of $k = 10$ chromosomes (Table 2) can be fitted to their theoretical counterparts. The value of α at which the best fit is achieved, is denoted by $\hat{\alpha}$. Figure 3 depicts the quality of the fit. The model reproduces the diversities estimated from data quite well. However, fitting results in very diversified estimates $\hat{\alpha}$ (Table 2). They vary from $\hat{\alpha} = 0.01$ (practically, equal to 0) for the WRN locus, through $\hat{\alpha} = 1.7$ for the RQL locus, and $\hat{\alpha} = 1.8$ for the BLM locus, to $\hat{\alpha} = 5$ for the ATM locus. Unfortunately, in view of the flatness of the theoretical curves, more data seem to be needed to establish confidence intervals for the estimates $\hat{\alpha}$.

To verify the above results, we computed the empirical cumulative distributions of the frequency of the less frequent SNP variant at the four loci. These are presented in Fig. 4. Let us notice that the WRN graph is similar to the low-α concave theoretical curve in Fig. 1(a), while the RQL and ATM graphs are similar to the high-α convex theoretical curve in Fig. 1(c). This is consistent with the estimated $\hat{\alpha}$-values at these loci. The BLM graph does not conform to either pattern.

Furthermore, for each locus it is possible to calculate the expected number of the SNP discovered, given the symmetric Beta model

$$\begin{aligned} E(\#SNPs) &= lE[1 - X^k - (1-X)^k] \\ &= l(1-p)\left[1 - 2\frac{\Gamma(2\alpha)\Gamma(\alpha+n)}{\Gamma(2\alpha+n)\Gamma(\alpha)}\right] \\ &= l(1-p)\varphi(\alpha). \end{aligned}$$

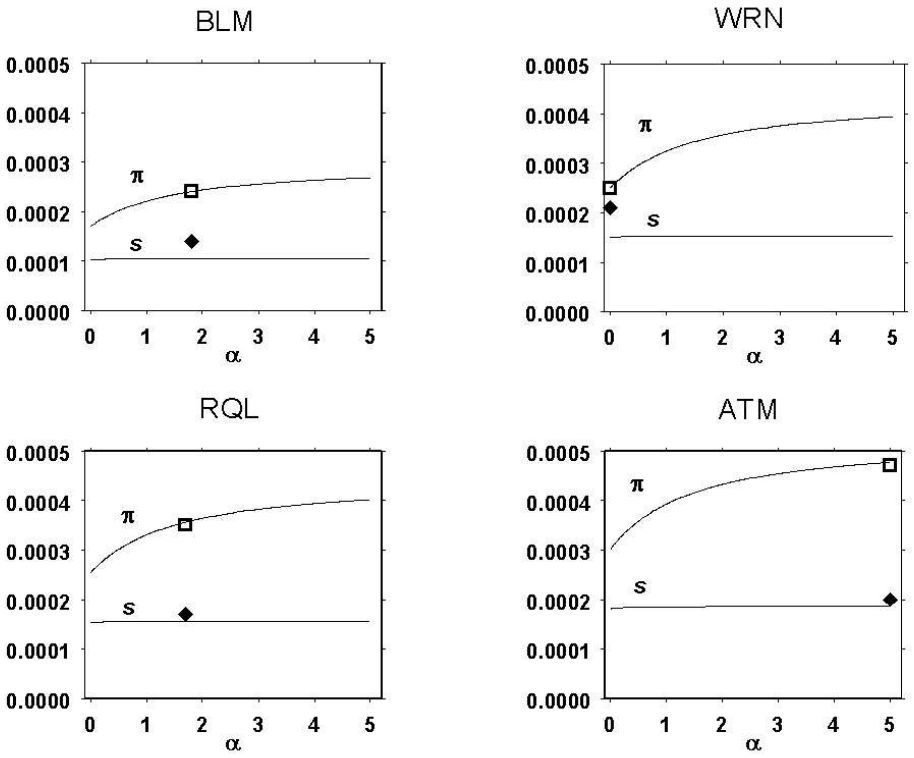

Fig. 3. Fitting data to the symmetric Beta model. Theoretical curves of $E_1(\pi \mid k)$ and $E_1(s \mid k)$ drawn superimposed on the estimated diversities $\hat{\pi}$ and $\hat{s}$ (square and diamond symbols, respectively). The symbols are drawn at the value of α at which the best fit is achieved.

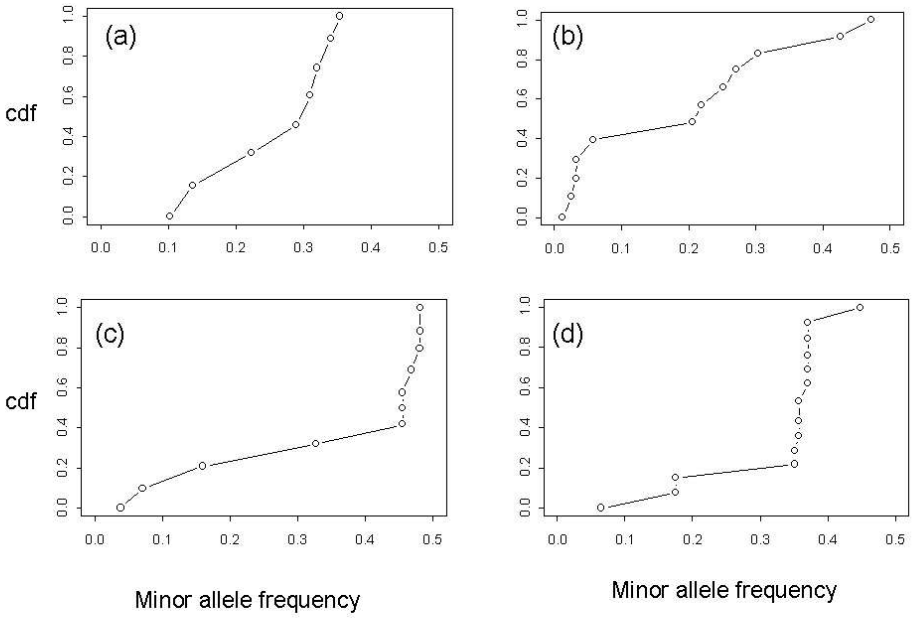

Fig. 4. Empirical cumulative distributions of the frequency of the less frequent SNP variant at the BLM, WRN, RQL and ATM loci (panels (a)–(d), respectively).

Since this latter is estimated by m, the number of the SNPs discovered, we obtain an estimate of p of the form

$$\hat{p}(\hat{\alpha}) = 1 - \frac{m}{l\varphi(\hat{\alpha})}.$$

These estimates (Table 2) have values ranging from 0.999 to 0.9994.

Projections of the ascertainment bias of diversity estimates. Figure 5 depicts the values of π and s estimated from our data, those reported by Nickerson *et al.*

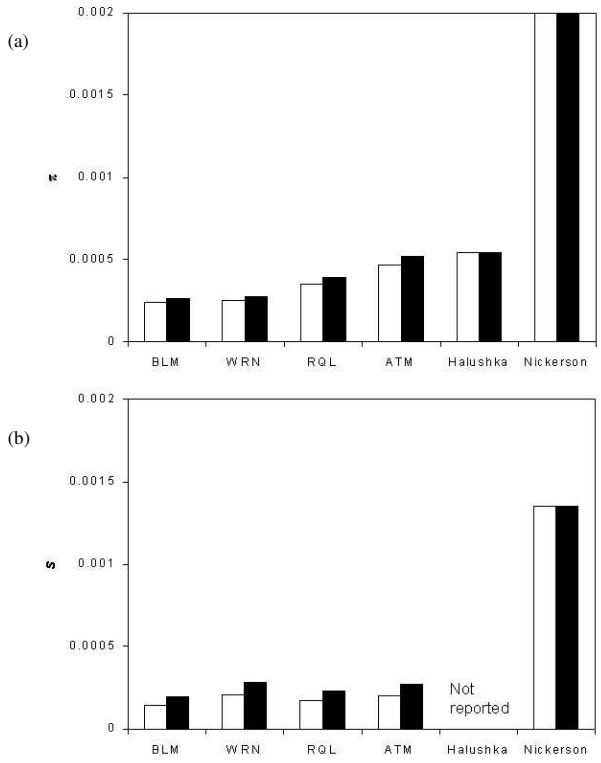

Fig. 5. Estimates of diversities $\hat{\pi}$ (a) and $\hat{s}$ (b), at the BLM, WRN, RQL and ATM loci, not corrected (lightly shaded bars) and corrected (heavily shaded bars) for the ascertainment bias. For comparison, estimates from (Nickerson *et al.*, 1998; Halushka *et al.*, 1999), based on large discovery samples, are also provided. Note that the estimate $\hat{s}$ is not available for the data of (Halushka *et al.*, 1999).

(1998), as well as the estimated value of s, reported by Halushka *et al.* (1999) (π was not estimated in the latter paper). Extension bars in our data represent the maximal bias expected with a given screening sample size, computed from the $B(\pi \,|\, k)$ and $B(s \,|\, k)$ expression, assuming $p = 0.999$, $l = 13,000$, $m = 10$, $k = 10$, and $n = 150$. The biases were computed at the "worst-case" values of α, at which they assume maximum values, so they should be treated as upper bounds. We used one set of the values of l and m at these loci, since the impact of small variations in these parameters is negligible. Observe that the bias corrections are quite small.

Data from (Halushka *et al.*, 1999; Nickerson *et al.*, 1998) are considered unbiased since all individuals' DNA was analyzed to discover SNPs.

4. Discussion

We developed a method to estimate the ascertainment bias of the genetic diversity estimates of DNA sequences in populations, due to a limited number of chromosomes used in SNP discovery. The method is based on fitting a parametric model of the distribution of the SNP frequency along the sequence to data, and making projections based on this model. The model assumes that with probability p only one variant is possible at a nucleotide site and if there are two variants, which happens with probability $1 - p$, the frequency X of variant 1 is a random variable with a given distribution. The distribution we use is symmetric Beta. The fit is obtained for p-values close to 0.999 and for a wide range of the shape coefficient α of symmetric Beta, from α close to 0 to α close to 5.

The symmetric Beta distribution arises in genetic theory in the mutation-drift equilibrium of the Wright-Fisher model with symmetric reversible mutation between two variants (Ewens, 1979, p. 155). In this setup, the shape coefficient has the interpretation of

$$\alpha = 4N_e\nu,$$

where N_e is the effective population size, while ν is the mutation rate. However, in our setup, we take into account also sites which stay invariant, and so the aggregate mutation rate will not be equal to ν (see below). For this reason, we use the notation ν instead of the previously used μ. As an example, let us assume $\alpha = 4$, and an effective population size of $N_e = 10^5$. This results in $\nu = 10^{-5}$ per generation. The average mutation rate per site is equal to $(1-p)\nu = 10^{-8}$, which seems quite realistic, particularly as an upper bound. Therefore, the model seems to be in a reasonable agreement with the rate at which mutation processes are thought to occur (Li 1997).

The symmetric Beta model is not identical with the Infinite Sites Model, which has been invoked for the SNP loci. In particular, its form suggests a symmetric mutation mechanism, contradictory to the Infinite Sites Model. Unfortunately, it is difficult to resolve the question of the symmetry of the distribution of SNP frequencies, mainly because it is difficult to determine which variant is ancestral. Some help may come from using the chimpanzee as the outgroup. However, even then, it has to be remembered that humans did not descend from the chimpanzee, but that both species had a common ancestor. It is quite possible that asymmetric Beta is closer to the truth. In

such a case, the distribution of the frequency of the less frequent variant would be still difficult to distinguish from the symmetric Beta. Asymmetric Beta would be closer to the Infinite Sites Model. Differences are not likely to be identified if frequencies of individual SNPs are examined.

It is worth noting that there are some differences of our modeling of the ascertainment (sampling) of the SNP sites, as compared to other approaches seen in the literature on the subject. The formulation of sampling described in Introduction is prompted by the study design of (Bonnen *et al.*, 2000; Trikka *et al.*, 2002). It is different from the sampling scheme described in (Wang *et al.*, 1998) and from the approach in (Eberle and Kruglyak, 2000). Wang *et al.* (1998) used a two-step ascertainment procedure for retaining an SNP site for a more detailed study. In their sampling scheme, first an SNP site had to exhibit a variant allele in a screening set of 3 individuals (i.e., satisfy Step 3 of our formulation with $k = 6$). Further, in a larger screening set of 20 chromosomes, they retained SNP sites that showed allele frequencies exceeding 30% for both alleles. Clearly, this two-stage ascertainment leaves some SNP sites unexamined that would have satisfied their second stage even when the first three individuals did not exhibit the variant allele.

In contrast, Eberle and Kruglyak (2000) describe a single-stage ascertainment scheme, called the $S(n, k)$ strategy in their notation, where the screening set (n chromosomes) is the ultimate sample itself. They retain the SNP sites that exhibit frequencies (the number of copies) of the minor allele between k and $n - k$.

A technical comment concerning the formulation of Eberly and Kruglyak (2000) is also important in this context. Since the expected relative frequency x of the minor allele is unknown, it is assumed by these authors that it follows the "probability distribution" (unnormed) proportional to $x^{-1}(1 - x)^{-1}$. This fact could be viewed as a consequence of the Infinite Sites Model. The unnormed "density" $x^{-1}(1 - x)^{-1}$ might be treated as a limiting case of the symmetric Beta model as $\alpha \to 0$. Assuming the symmetric Beta model, the probability of discovering an SNP using their $S(n, k)$ strategy is equal to

$$\int_0^1 \sum_{i=k}^{n-k} \binom{n}{i} \frac{\Gamma(2\alpha)}{\Gamma(\alpha)\Gamma(\alpha)} x^{i+\alpha-1}(1-x)^{n-i+\alpha-1}\,\mathrm{d}x$$

$$= \sum_{i=k}^{n-k} \frac{\Gamma(2\alpha)}{\Gamma(\alpha)\Gamma(\alpha)} \frac{\Gamma(i+\alpha)\Gamma(n-i+\alpha)}{\Gamma(n+2\alpha)}.$$

As $\alpha \to 0$, the term

$$\frac{\Gamma(2\alpha)}{\Gamma(\alpha)\Gamma(\alpha)} \frac{\Gamma(i+\alpha)\Gamma(n-i+\alpha)}{\Gamma(n+2\alpha)}$$

tends to zero if $i = 1, \ldots, n-1$, and it tends to $1/2$ if $i = 0$ or $i = n$. Eberly and Kruglyak (2000) sample simply from the product of their Equations (1) and (2), which does not seem justified.

However, like Eberle and Kruglyak's model, our model does not take into account the dependence between SNP sites. It is not needed if expected values only are considered. However, such a dependence is of importance for the efficiency of estimators we used. Further work is needed in this direction.

The ascertainment bias has been considered for genetic data (see, e.g., Rogers and Jorde, 1996) and as a more general sampling bias (Chakraborty and Rao, 2000). Rogers and Jorde (2000) consider the conventional practice, which ensures that only most variable loci are most likely to be discovered and thus included in the sample of loci. Consequently, estimates of average heterozygosity are biased upward. They describe a model of this bias. A different but related type of bias arises when data discovered in Population 1 are typed in Population 2. Then, the estimated heterozygosity will be higher in Population 1 than that in Population 2.

In the discovery setup considered in this paper, we do not consider the between-population bias (which may well exist). Instead, we focus on a new type of ascertainment bias, which tends to decrease the observed genetic diversity. The number of segregating sites, s, is more sensitive to this than the average number of pairwise differences, π. The presence of a large number of sites with a rare variant (high α in the symmetric Beta distribution model) tends to increase bias, especially in s. Both the biases are particularly sensitive to the parameter p, equal to the proportion of completely monomorphic sites. The decrease in p below 0.99 leads to a dramatic increase in both biases. The bias we observe in our data seems to be limited, due to the estimated p being not less than 0.999.

Taking into account the limited influence of the ascertainment bias on our estimates of genetic diversity, we conclude that the diversity of DNA sequences at the BLM, WRN, RQL and ATM loci is indeed much lower than the diversity reported by Halushka *et al.* (1999) and Nickerson *et al.* (1998) at the blood-pressure homeostasis and lipoprotein lipase loci. Our estimates and corrections provide an idea of the extent of the heterogeneity of genetic diversity over the human genome.

Acknowledgments

The research was supported by US Public Health Service Research grants GM 41399 and CA 75432 from the National Institutes of Health and by an NSF graduate fellowship granted to A. Renwick.

References

Bonnen P.E., Story M.D., Ashorn C.L., Buchholz T.A., Weil M.A. and Nelson D. (2000): *Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium*. — Am. J. Hum. Genet., Vol. 67, No. 6, pp. 1437–1451.

Cargill M., Altshuler D., Ireland J., Sklar P., Ardlie K., Patil N., Shaw N., Lane C.R., Lim E.P., Kalyanaraman N., Nemesh J., Ziaugra L., Friedland L., Rolfe A., Warrington J., Lipshutz R., Daley G.Q. and Lander E.S. (1999): *Characterization of single-nucleotide polymorphisms in coding regions of human genes*. — Nat. Genet., Vol. 22, No. 3, pp. 231–238.

Chakraborty R. and Rao C.R. (2000): *Selection biases of samples and their resolution*, In: Handbook of Statistics (C.R. Rao, P.K. Sen, Eds.). — Amsterdam: Elsevier Science.

Clark A.G., Weiss K.M., Nickerson D.A., Taylor S.L., Buchanan A., Stengard J., Salomaa V., Vartiainen E., Perola M., Boerwinkle E., Sing C.F. (1998): *Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase*. — Am. J. Hum. Genet., Vol. 63, No. 2, pp. 595–612.

Eberle M. and Kruglyak L. (2000): *An analysis of strategies for discovery of single-nucleotide polymorphisms*. — Genet. Epidem., Vol. 19, No. S1, pp. S29–S35.

Ewens W.J. (1979): *Mathematical Population Genetics. Biomathematics, Vol. 9*. — Berlin: Springer.

Halushka M.K., Fan J.B., Bentley K., Hsie L., Shen N., Weder A., Cooper R., Lipshutz R. and Chakravarti A. (1999): *Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis*. — Nat. Genet., Vol. 22, No. 3, pp. 239–247.

Li W.-H. (1997): *Molecular Evolution*. — Sunderland, MA: Sinauer Associates.

Nickerson D.A., Taylor S.L., Weiss K.M., Clark A.G., Hutchinson R.G., Stengard J., Salomaa V., Vartiainen E., Boerwinkle E., Sing C.F. (1998): *DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene*. — Nat. Genet., Vol. 19, No. 3, pp. 233–240.

Rogers A.R., Jorde L.B. (1996): *Ascertainment bias in estimates of average heterozygosity*. — Am. J. Hum. Genet., Vol. 58, No. 5, pp. 1033–1041.

Trikka D., Fang Z., Renwick A., Jones S.H., Chakraborty R., Kimmel M., Nelson D.L. (2002): *Complex SNP-based haplotypes in three human helicases demonstrate the need for ethnically-matched control populations in association studies*. — Genome Res, Vol. 12, No. 4, pp. 627–639.

Venter J.C. et al. (2001): *The sequence of the human genome*. — Science, Vol. 291, No. 5507, pp. 1304–1351.

Wang D.G., Fan J.B., Siao C.J., Berno A., Young P., Sapolsky R., Ghandour G., Perkins N., Winchester E., Spencer J., Kruglyak L., Stein L., Hsie L., Topaloglou T., Hubbell E., Robinson E., Mittmann M., Morris M.S., Shen N., Kilburn D., Rioux J., Nusbaum C., Rozen S., Hudson T.J., Lander E.S. et al. (1998): *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome*. — Science, Vol. 280, No. 5366, pp. 1077–1082.