Am. J. Hum. Genet. 58:1033–1041, 1996

Ascertainment Bias in Estimates of Average Heterozygosity

Alan R. Rogers[1] and Lynn B. Jorde[2]

Departments of [1]Anthropology and [2]Human Genetics, University of Utah, Salt Lake City

Summary

Population geneticists work with a nonrandom sample of the human genome. Conventional practice ensures that unusually variable loci are most likely to be discovered and thus included in the sample of loci. Consequently, estimates of average heterozygosity are biased upward. In what follows we describe a model of this bias. When the mutation rate varies among loci, bias is increased. This effect is only moderate, however, so that a model of invariant mutation rates provides a reasonable approximation. Bias is pronounced when estimated heterozygosity is $<\sim 35\%$. Consequently, it probably affects estimates from classical polymorphisms as well as from restriction-site polymorphisms. Estimates from short-tandem-repeat polymorphisms have negligible bias, because of their high heterozygosity. Bias should vary not only among categories of polymorphism but also among populations. It should be largest in European populations, since these are the populations in which most polymorphisms were discovered. As this argument predicts, European estimates exceed those of Africa and Asia at systems with large bias. The magnitude of this European excess is consistent with the version of our model in which mutation rates vary across loci.

The Problem

Students of human population genetics are seldom lucky enough to work with loci drawn at random from the genome. More often, we work with loci chosen for their variability. Our sample of loci is therefore unusually variable, and estimates of heterozygosity are biased upward. This bias interferes with inference in various ways. It confounds comparisons of human heterozygosity with that of other species; it also confounds comparisons among human populations. Biased estimates of average heterozygosity also generate biased estimates of effective population size.

Received March 10, 1995; accepted for publication January 29, 1996.

Address for correspondence and reprints: Dr. Alan R. Rogers, Department of Anthropology, University of Utah, 102 Stewart Hall, Salt Lake City, UT 84112. E-mail: rogers@anthro.utah.edu

0002-9297/96/5805-0016$02.00

Several mechanisms have introduced bias into the sample of human polymorphisms. Early work relied on blood groups, which are recognized by antigen-antibody reactions. Since reactions occur only between individuals who carry different alleles, polymorphic loci are most likely to be discovered and therefore included in the sample of loci. This inclusion introduces an ascertainment bias, which inflates estimates of heterozygosity. Lewontin (1967) pointed out that this bias would have been largest in the earliest studies, since they compared only limited numbers of individuals: "Rare variants will be seen only as the number of bloods examined becomes larger and larger, so that at any particular time the sample of loci is biased toward polymorphic loci; but this bias will grow smaller as the number of bloods examined grows larger. Eventually, when all antigen-specifying loci are known, the bias would disappear" (Lewontin 1967, p. 681). Lewontin used this argument to interpret the data in figure 1. There, "cumulative heterozygosity" in year x refers to the average heterozygosity over loci that had been discovered by year x. Cumulative heterozygosity declines with time, as Lewontin observed. Although the graph is nearly flat just before 1962, subsequent years saw a continued decline (Nei and Roychoudhury 1974, 1982).

Ascertainment bias is also a problem in table 1, which uses various categories of data to compare heterozygosity estimates from different human populations. The columns are arranged from left to right in order of increasing European heterozygosity. Since these loci were nearly all ascertained using European subjects, the European estimates should have the largest bias (Bowcock et al. 1991; Cavalli-Sforza et al. 1994, pp. 141–42). That may explain the high European values in columns a and c–e. It is intriguing that the heterozygosity estimate for protein systems (column b) is also highest among Europeans, since many of these systems were ascertained not on the basis of variability but rather in an effort to estimate the overall heterozygosity level in humans (Harris and Hopkinson 1972). Indeed, the data set includes 18 monomorphic loci. However, Nei and Roychoudhury suggest that "it is also possible that monomorphic loci have not been reported as often as have polymorphic loci in recent research, the reason being that many investigators are primarily interested in polymor-

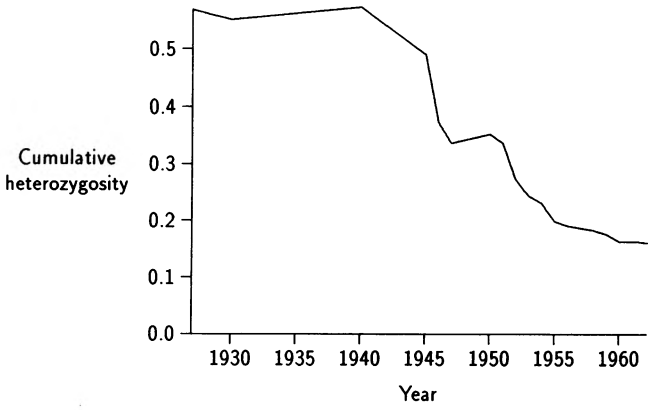

Figure 1 Cumulative heterozygosity as a function of time (Lewontin 1967).

phism" (Nei and Roychoudhury 1982, p. 8). If so, then ascertainment bias may account for the elevated European value in column b as well as those in columns a and c–e.

The European excess disappears in columns f–h. To understand why, one must consider two opposing effects. First, there is sample size. Lewontin's argument implies that by the 1950s ascertainment of classical polymorphisms had come to involve large samples. Modern molecular polymorphisms, on the other hand, are ascertained using small samples. This distinction reflects their primary function—mapping disease genes. Since highly polymorphic loci are most useful in gene mapping, these polymorphisms are ascertained using a small number of subjects—usually no more than eight. Loci are ascertained as polymorphic only if there is some polymorphism in these small samples (Mountain and Cavalli-Sforza 1994). These considerations suggest that ascertainment samples were larger for classical polymorphisms (columns a–c of table 1) than for molecular polymorphisms (columns d–h). Since bias is greatest when ascertainment samples are small, we might expect the greatest bias in molecular polymorphisms—columns d–h in the table—and predict a pattern unlike that in the table. If bias were most pronounced in molecular polymorphisms, the European estimates should be large in columns d–h rather than in columns a–e.

In addition to this sample-size effect, there is also an effect of heterozygosity. Bias results when loci with low heterozygosity are excluded. But short-tandem-repeat (STR) loci are so extremely variable that few loci may be excluded. If so, ascertainment bias should be weak in STR loci. This argument is consistent with the pattern in table 1. It suggests that the high European heterozygosity seen in columns a–e reflects ascertainment bias, which is important in those columns because of their relatively low heterozygosity. Mountain and Cavalli-Sforza (1994) used computer simulation to show that this idea is plausible.

Yet several questions remain. First, it is not yet clear that the effect of heterozygosity on bias outweighs the effect of sample size—Mountain and Cavalli-Sforza did not consider the two effects separately. Neither is it clear that the crossover from high European values to high African values occurs at the right level of heterozygosity. After all, the RFLP and RSP loci (RFLPs consisting solely of restriction-site polymorphisms) in table 1 are much more heterozygous than the classical polymorphisms in columns a–c. Perhaps the heterozygosity effect would predict a crossover between columns c and d rather than between columns e and f. To answer such questions, we need a model relating heterozygosity to bias and to the size of samples used in ascertainment. In what follows,

Table 1

Average Heterozygosity

Population	Blood Group[a]	Protein[b]	Classical[c]	RFLP[d]	RSP[e]	STR-4[f]	STR-2[g]	STR-3[h]
Africa	.164	.179	.163	.297	.322	.769	.807	.850
Asia	.145	.164	.189	.327	.377	.681	.685	.820
Europe	.179	.186	.202	.379	.432	.724	.730	.807

NOTE.—Largest entry in each column is underlined. Columns are in order of increasing European heterozygosity.

[a] 32 blood groups (Nei et al. 1993).

[b] 80 protein polymorphisms (Nei et al. 1993).

[c] 110 classical polymorphisms (Bowcock et al. 1994).

[d] 79 RFLPs (Bowcock et al. 1994).

[e] 30 RFLPs consisting solely of restriction site polymorphisms (Jorde et al. 1995*a*).

[f] 30 tetranucleotide STRs (Jorde et al. 1995*a*).

[g] 30 dinucleotide STRs. The difference between Africa and Europe is significant (Bowcock et al. 1994).

[h] 5 trinucleotide STRs (Watkins et al. 1995).

we describe such a model and apply it to the data of figure 1 and table 1.

In building such a model, one must assume something about the statistical distribution from which mutation rates are drawn. Our model will assume that selective neutrality and stationary population size have prevailed long enough for the population to reach a mutation-drift equilibrium at each locus.

Model

We imagine that research proceeds in two stages. In stage I, the ascertainment stage, a small number of subjects are typed at a large number of loci to determine which loci are polymorphic. In stage II, a large number of subjects are typed at the polymorphic loci to estimate heterozygosity. Bias arises if the loci studied in stage II are more heterozygous than randomly chosen loci would have been. We refer to the sample of stage I as the "ascertainment sample." In stage II, we calculate only the expected value of the estimate of heterozygosity. This step makes it unnecessary to deal explicitly with the sample size in stage II.

In stage I, we assume that loci are ascertained as polymorphic by typing a sample of z statistically independent individuals (or $2z$ independent genes). If the $2z$ genes are identical, then the locus is deemed to be monomorphic and is discarded. Otherwise, the locus is ascertained as polymorphic. We denote by $\mathcal{A}$ the event that a given locus was ascertained as polymorphic by this method. Our assumption accepts a locus as polymorphic if even a single variant gene is found in the sample. Procedures that require more variants than this will induce a larger bias. Thus, our assumption provides a lower bound on the bias for samples of a given size. In addition to providing a lower bound, our assumption is also a fair description of recent practice. It provides only a crude approximation, however, to the procedures by which older polymorphisms were ascertained. In those cases it provides only a lower bound on the bias.

We assume that each locus has K alleles and denote the vector of allele frequencies by $\mathbf{x} = (x_1, x_2, \ldots, x_K)$. We also assume that the mutational process is symmetric, so that each allele is equally likely to mutate to each of the $K - 1$ other alleles. These assumptions imply that the probability density p of $\mathbf{x}$ is symmetric—the density of $\mathbf{x}$ is equal to that of every permutation of $\mathbf{x}$. This symmetry applies not only to p, but also to the conditional density $p_{|\mathcal{A}}$ of $\mathbf{x}$, given $\mathcal{A}$. Because of this symmetry, the conditional heterozygosity given $\mathcal{A}$ can be written as

$$h_{|\mathcal{A}} \equiv 1 - E\left[\sum_{i=1}^{K} x_i^2 \,|\, \mathcal{A}\right] = 1 - KE[x_1^2 \,|\, \mathcal{A}], \quad (1)$$

where E denotes the expectation operator. The first section of the appendix shows that the expectation in this equation equals

$$E[x_1^2 \,|\, \mathcal{A}] = \frac{E[x_1^2] - E[x_1^{2z+2}] - (K-1)E[x_1^2 x_2^{2z}]}{1 - KE[x_1^{2z}]}. \quad (2)$$

To proceed further, it is necessary to specify the probability distribution of $\mathbf{x}$, and we rely for this purpose on the assumption of mutation-drift equilibrium. This assumption implies that $\mathbf{x}$ has a Dirichlet distribution with density (Ewens 1979, Eq. [5.108])

$$p(\mathbf{x}) = \left(\frac{\Gamma(K\alpha)}{[\Gamma(\alpha)]^K}\right)\left(\prod_{i=1}^{K} x_i\right)^{\alpha-1}, \quad (3)$$

where Γ is the Gamma function (Abramowitz and Stegun 1964),

$$\alpha \equiv \theta/(K-1),$$

$$\theta \equiv 4Nu.$$

Here, u is the mutation rate and N the effective population size. Conventionally, population geneticists have treated u as a constant. We employ this assumption below in model A and then relax it in developing model B.

Model A: Fixed u

We assume for the moment that all loci have the same mutation rate, u, and consequently have the same values of $\theta = 4Nu$ and of $\alpha = \theta/(K-1)$. This assumption implies that each of the K marginal distributions are Beta distributions with parameters α and $(K-1)\alpha$ and with mean $1/K$. When α is small, most alleles have frequencies near 0 or 1, and heterozygosity is low. When α is large, most allele frequencies are near $1/K$ and heterozygosity is high, approaching $1 - 1/K$ as $\alpha \to \infty$.

Substituting equations (11) and (13) (from the appendix) into equations (2) and (1) leads to the conditional heterozygosity,

$$h_{|\mathcal{A}} = 1 - K \times \frac{\frac{\Gamma(\alpha+2)}{\Gamma(K\alpha+2)} - \frac{\Gamma(\alpha+2z+2)}{\Gamma(K\alpha+2z+2)} - (K-1)\frac{\Gamma(\alpha+2)\Gamma(\alpha+2z)}{\Gamma(\alpha)\Gamma(K\alpha+2z+2)}}{\frac{\Gamma(\alpha)}{\Gamma(K\alpha)} - K\frac{\Gamma(\alpha+2z)}{\Gamma(K\alpha+2z)}}. \quad (4)$$

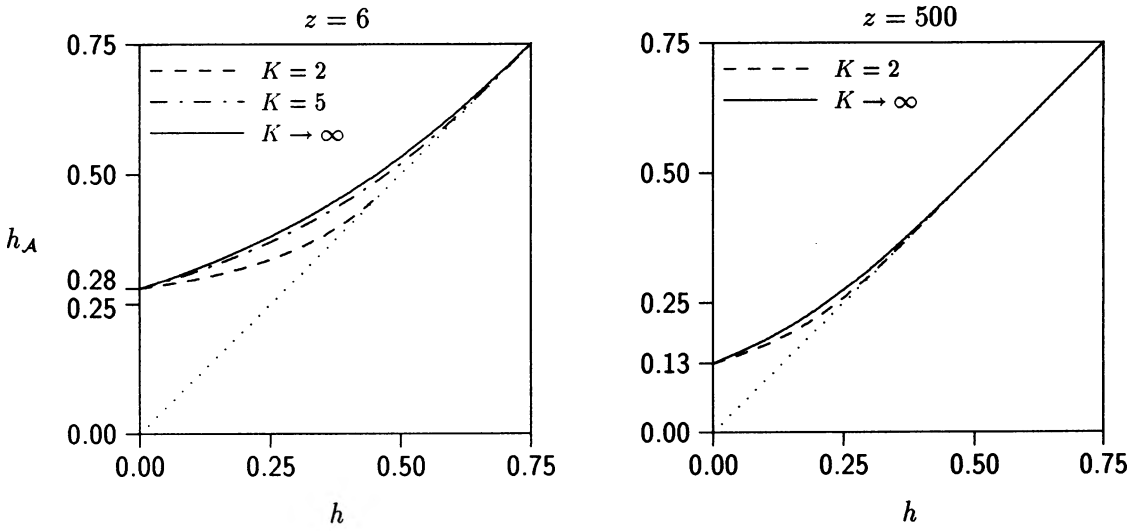

Figure 2 Biased-against-unbiased heterozygosity under model A. The left and right panels show the bias when ascertainment samples are ($z = 6$) and ($z = 500$), respectively.

Meanwhile, unconditional heterozygosity is

$$h \equiv 1 - KE[x_1^2] = 1 - (\alpha + 1)/(K\alpha + 1)\,, \tag{5}$$

as shown in the last section of the appendix (Ewens 1979, eq. [5.118]).

In the limit as $K \to \infty$ these become

$$h_{\mathcal{A}} = 1 - \frac{\frac{1}{\theta+1} - \theta\beta(2z+2,\theta) - \frac{\theta}{\theta+1}\beta(2z,\theta+2)}{1 - \theta\beta(2z,\theta)} \tag{6}$$

$$h = \frac{\theta}{\theta+1}. \tag{7}$$

Here, $\beta(a, b) \equiv \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function (Abramowitz and Stegun 1964).

These formulas are illustrated in figure 2, where the two panels plot biased heterozygosity ($h_{\mathcal{A}}$) against unbiased heterozygosity (h), assuming ascertainment samples of $z = 6$ and $z = 500$, respectively. The left side of each panel, where heterozygosity is low, corresponds to small values of θ, whereas the right side corresponds to large θ. Ascertainment bias is measured by the vertical distance between $h_{\mathcal{A}}$ and the dotted 45° line. Clearly, bias is pronounced when heterozygosity is low and declines as heterozygosity increases. This finding is as expected, since nearly all loci are ascertained as polymorphic when heterozygosity is high. Bias results from discarding loci and is therefore negligible when few loci are discarded. A similar result was obtained by Mountain (1994, p. 119), who found that bias decreases when the mutation rate is high.

Model B: Gamma-Distributed *u*

To add realism, we now relax the assumption that the mutation rate is constant across loci. In the past, the distribution of mutation rates has been fit using both the lognormal distribution (Cavalli-Sforza and Bodmer 1971, p. 105) and the gamma distribution (Nei et al. 1976). The latter approach has good empirical support (Chakraborty et al. 1980) and will therefore be used here. We assume that the mutation rate, u, at each locus is drawn independently from a gamma distribution with density

$$g(u) = u^{c-1}(c/\bar{u})^c e^{-uc/\bar{u}}/\Gamma(c)\,. \tag{8}$$

Here, $\bar{u}$ is the mean mutation rate, and c is a "shape parameter," which equals the reciprocal of the square of the coefficient of variation of u. As $c \to \infty$, the gamma model converges to the model of constant mutation rate. Under model B, equations (11) and (13) must be replaced by their expectations with respect to $g(u)$. We then obtain h from equation (5) and $h_{\mathcal{A}}$ from equations (2) and (1). These calculations are performed numerically, since we have been unable to get analytical results.

Nei et al. (1976) estimated that the shape parameter, c, lay somewhere between 1 and 2, so we study values in this range. Figure 3 compares models A and B. The dashed line is identical to the dashed line in the left panel of figure 2. The open circles show results for model B

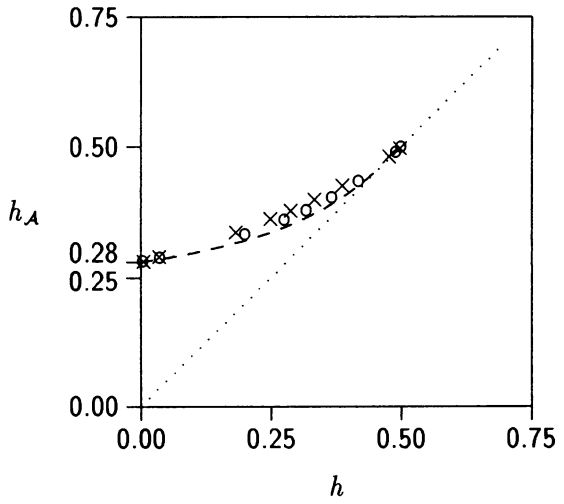

Figure 3 Model B compared with model A. Dashed line illustrates model A and is identical to the dashed line in the left panel of figure 2. Open circles show comparable results from model B, in which mutation rates follow a Gamma distribution with $c = 2$ and $\bar{u}$ varying from 10^{-7} to 10^{-2}. Results for model B assume a population of size 10,000. In both models, $K = 2$ and $z = 6$.

with $c = 2$; the crosses show results for $c = 1$. These symbols do not fall on the dashed line, so it is clear that variation has an effect. This effect is larger when c is small (i.e., when the coefficient of variation is large). Yet none of the symbols fall far from the dashed line. If these parameter values are realistic, then model A provides a reasonable approximation to model B. This approximation is especially accurate at the extreme left and extreme right.

Discussion

Our assumption of mutation-drift equilibrium may be violated either because of natural selection or because of changes in population size. These possibilities are of special concern with human data, since mtDNA provides evidence for a substantial population increase in the late Pleistocene (Harpending et al. 1993; Harpending 1994; Sherry et al. 1994; Rogers 1995). Thus, the results should be applied to real populations only with caution. We should also point out that our results do *not* imply bias at any particular locus. Instead, the results refer to a bias that arises because the sample of loci is not drawn at random from the genome. The sample of loci yields only a biased estimate of mean heterozygosity, even though the estimate at each locus may be unbiased.

The magnitude of this bias is remarkable. The left panel of figure 2 shows that when $K = 2$ and h_I is in the neighborhood of .3, the bias under model A may be several times as large as the unbiased value, h. With more than two alleles, the bias is even larger. Clearly, estimates in the neighborhood of .3 should be regarded with suspicion. On the other hand, the relative bias is negligible when h_I exceeds ~.35 (with two alleles) or .5 (with multiple alleles). Bias decreases rapidly as h_I increases from .3 to .5. As true heterozygosity approaches 0, biased heterozygosity approaches a limit that is well above 0 and is independent both of the number of alleles and of variation in mutation rates. As the figures indicate, this limit equals .2802 when $z = 6$, and .1333 when $z = 500$.

For a given value of unbiased heterozygosity, bias increases both with the number of alleles and also with the level of variation in u. For example, if $z = 6$ and the unbiased value is $h = .3$, then the biased values are as seen in table 2. In both columns, bias increases rapidly as one moves from two to five alleles but eventually levels off. An infinite number of alleles yields essentially the same bias as five. Bias is greater under model B (in which u varies) than under model A (in which u is fixed).

Return now to figure 1 and note that there was little change in mean heterozygosity in the years immediately preceding 1962. This leveling off intrigued Lewontin (1967, p. 681). If ascertainment samples increased steadily with time, then such a leveling off would be expected only when "all antigen-specifying loci were known" and the bias disappears. Thus, Lewontin suggested that the graph's final values were nearly unbiased. We evaluate this conjecture in figure 4, which shows that biased heterozygosity does indeed decline as the size, z, of ascertainment samples increases. (The figure uses model A with an infinite number of alleles and $\theta = .1$, but other assumptions yield qualitatively similar results.) Yet the slope is near zero when h_I is still much larger than h. If sampling error were added to the points along this graph, we might easily conclude that h_I were approaching an asymptote with a value near .16. Yet the unbiased value ($h \approx .09$) is still far away. Clearly, we are not justified in inferring an absence of bias from an apparent asymptote such as that in figure 1.

We now return to table 1. There, the crossover be-

Table 2

Biased Values

	h_I	
ALLELES	Model A	Model B ($c = 1$)
2	.3557	.3844
5	.3939	.4469
25	.4039	.4660
50	.4049	.4680
∞	.4059	...

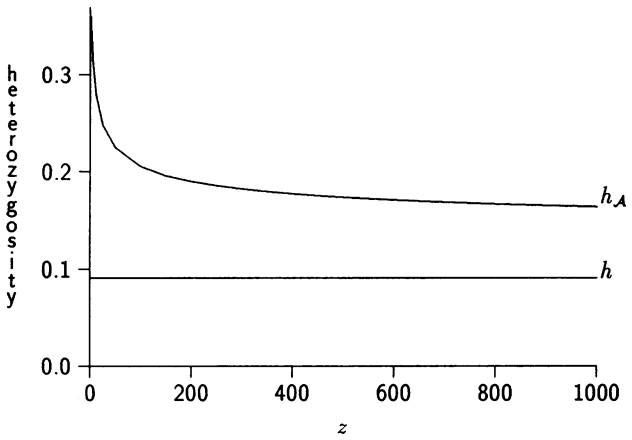

Figure 4 Heterozygosity as a function of the number of subjects used in ascertainment. h_{ℓ} is biased heterozygosity; h is unbiased heterozygosity; and z is the number of subjects used in ascertainment. Calculations use model A, with $K \to \infty$ and with $\theta = .1$.

tween high European values and high African values occurs when estimates of European heterozygosity exceed ~.4. This value is slightly above the point at which bias becomes small in the two-allele model, but slightly below the analogous point in the five-allele model. Thus, the crossover is in reasonable agreement with model A when $z = 6$.

The numbers in the first three columns of table 1 are too low for the model with $z = 6$, suggesting that these polymorphisms were ascertained using much larger samples. (Incidentally, this suggestion does not imply an absence of bias, for bias can be substantial even when z is large.)

In table 1, columns RFLP and RSP show greatest heterozygosity in Europe. Here, heterozygosity is high enough to be consistent with the model in which $z = 6$. This consistency is comforting, since these polymorphisms were ascertained more or less as the model assumes, and using small samples (Mountain and Cavalli-Sforza 1994). Since our assumptions about ascertainment are fairly accurate here, the model provides quantitative as well as qualitative information. We can use it to ask whether Europe's excess heterozygosity was caused by ascertainment bias.

It is important to realize that ascertainment bias may contaminate all three estimates, not just that of Europe. Loci that are polymorphic in Europe are also likely to be polymorphic in Africa and Asia. Thus, African and Asian estimates may also be biased upward. The three biases, however, are probably unequal. Ascertainment using European subjects should provide more information about polymorphism in Europe than elsewhere. Consequently, the upward bias should be largest in Europe (Mountain and Cavalli-Sforza 1994). The difference in these upward biases may account for the excess European heterozygosity in table 1.

We have no theory describing the difference between African and European biases. But since all three biases are in the same direction, the absolute differences between them are necessarily smaller than the largest bias—that of Europe. And we do have a theory for that. By placing a bound on the European bias, we also place a bound on the difference between African and European biases.

There are two alleles at each restriction site, so we set $K = 2$. These loci were ascertained using samples of approximately eight individuals (Mountain and Cavalli-Sforza 1994), so we assume that $z = 6$ in order to be conservative. We set h_{ℓ} equal to the estimate of European heterozygosity and solve equation (4) for θ. Substitution into equation (5) then yields an estimate of h, the unbiased heterozygosity. The resulting estimates are shown in table 3. The first two rows are copied from table 1 and show the uncorrected estimates of European and African heterozygosity. The bottom two rows show corrected estimates of European heterozygosity, using models A (fixed u) and B (gamma-distributed u). In the row for model A, both values are larger than the corresponding uncorrected estimate of African h_{ℓ}. Thus, ascertainment bias cannot explain the data under model A. But model B yields smaller estimates of h, one of which is smaller than the corresponding African h_{ℓ}. And h would be smaller still if we assumed that $c < 1$. Consequently, we cannot exclude the possibility that the excess European heterozygosity is caused entirely by ascertainment bias.

These results appear inconsistent with those of Mountain and Cavalli-Sforza (1994), who conclude that the observed European excess is too large to have been caused by ascertainment bias. But their model differs from ours in several ways:

1) They allow for the subdivision of an ancestral population into several isolated continental populations. This enables them to compare the effect of ascertainment bias on the European and African populations separately. Our approach, on the other hand, places only an upper bound on the effect of ascertainment bias and is

Table 3

Corrected Heterozygosity

	RFLP	RSP
European h_{ℓ}	.379	.432
African h_{ℓ}	.297	.322
European h (model A)	.348	.427
European h (model B with $c = 1$)	.288	.397

therefore less likely to reject the hypothesis of ascertainment bias.

2) They assume that the mutation rate does not vary across sites. This assumption is also made by our model A, which cannot explain the European excess either. Thus, our results are consistent with theirs when we make the same assumption about mutations. Our model B, however, shows that an invariant mutation rate can give misleading results.

3) They use a different model of the ascertainment process: They calculate allele frequencies from the entire population and accept loci as polymorphic if at least two alleles have frequencies >1%. This procedure, together with the assumptions that $N = 10{,}000$ and $u = 10^{-7}$ (Mountain 1994, pp. 117–19) led in their simulations to a biased European heterozygosity of .379 ± .015 (Mountain and Cavalli-Sforza 1994, p. 6517). Our procedure, on the other hand, looks at a small "ascertainment sample" and accepts loci if at least two alleles are found within this sample. To compare these two procedures, we used their simulation parameters (see above) to set θ and then used our model to calculate biased heterozygosity, $h_{.1}$, under various assumptions about the number of alleles and the size of the ascertainment sample. In no case was our $h_{.1}$ as large as their estimate. The maximal value under our model is obtained with an ascertainment sample of one individual under the infinite-alleles model: $h_{.1} = .3349$. This value is not far below the lower bound, .3496, of the confidence interval surrounding their estimate. Thus, there is no strong evidence that the two procedures produce different biases. There is a weak indication, however, that their procedure introduces a greater bias, which would have reduced their chances of rejecting the hypothesis of ascertainment bias. Since they did reject this hypothesis, the difference between our results must reflect assumptions 1 and/or 2.

We turn finally to the heterozygosity estimates from STR loci (see table 1). These loci differ from all others in suggesting that heterozygosity is highest in Africa rather than Europe. Heterozygosity is extremely high in these data (>70%) and figure 2 shows that this eliminates nearly all ascertainment bias. These results may still be artifacts of sampling error, for the high African value is significant in only one of the three columns, and that one significant result may be spurious. (It treats linked loci as statistically independent [Bowcock et al. 1994].) But if our model is even approximately correct, then the STR loci are probably not affected much by ascertainment bias.

It is interesting that STR-3 loci yield estimates so similar to the other STRs, since each of the STR-3 loci can cause genetic disease (Jorde et al. 1995*b*). These loci also imply a pattern of population relationships that is consistent with that implied by other sets of loci (Watkins et al. 1995). Thus, although selection has certainly affected these loci, it has produced no obvious distortion in genetic distances or in average heterozygosity.

The high African values at STR loci cast doubt on the suggestion (Mountain and Cavalli-Sforza 1994) that European heterozygosity is elevated in RFLP loci because the European population is admixed. Admixture should elevate heterozygosity at STR loci, too, yet the data show no evidence of this. It seems likely that much of the European excess results from the ascertainment of polymorphisms in European populations.

On the other hand, other factors may also be at work:

1) We have not studied the effect of variation in mutation rates on correlations between the bias observed in different groups. When mutation rates vary among loci, the loci that are ascertained as polymorphic will tend to have high mutation rates, inflating heterozygosity estimates in *all* groups. When $h_{.1}$ is inflated in Europe, it will tend also to be inflated in Africa and Asia. When we account for this effect, it may turn out that ascertainment bias cannot account for the observed group differences.

2) Our analysis is conservative in using the European bias to place an upper bound on the difference between African and European biases. If we could calculate this difference directly, as Mountain and Cavalli-Sforza do (1994), we might reject the hypothesis of ascertainment bias.

Conclusions

The sample of human genetic loci is biased in favor of polymorphic loci, and estimates of average heterozygosity are therefore biased upward. The apparent asymptote in the graph of average heterozygosity against time (fig. 1) does not imply that classical polymorphisms yield unbiased estimates of average heterozygosity. Because the procedure used in ascertaining modern molecular polymorphisms is fairly well described, one can calculate the bias that it introduces into estimates of heterozygosity. When estimated heterozygosity is below ~.3, bias is large. As estimated heterozygosity increases, bias decreases and eventually becomes negligible. The point at which this occurs varies among models. With two alleles and a fixed mutation rate, bias is negligible when estimated heterozygosity exceeds ~.35. Because of their high heterozygosity, STR loci are essentially free of ascertainment bias. These loci are therefore uniquely useful for comparing populations. Race differences in estimated heterozygosity are larger than predicted by the version of our model that assumes all loci to have equal rates of mutation. When varying mutation rates are allowed, however, the magnitude of bias is consistent with observed race differences.

Acknowledgments

We thank L. Luca Cavalli-Sforza, Henry Harpending, Li Jin, and Joanna Mountain for comments, and Scott Watkins for providing the data in the STR-3 column in table 1. Wen-Hsiung Li pointed out to us that the geographic structure of a sample can affect estimates of heterozygosity. This research was supported in part by National Science Foundation grant DBS-9310105.

Appendix

Derivation of Expression for $E[x_1^2|\mathcal{A}]$

Bayes's rule allows the conditional density of $\mathbf{x}$ to be written as

$$p_{\mathcal{A}}(\mathbf{x}) = \Pr(\mathcal{A}|\mathbf{x})p(\mathbf{x})/\Pr(\mathcal{A}) . \quad (9)$$

Given $\mathbf{x}$, the conditional probability of $\mathcal{A}$ is

$$\Pr(\mathcal{A}|\mathbf{x}) = 1 - \sum_{i=1}^{K} x_i^{2z} .$$

The unconditional probability of $\mathcal{A}$ is therefore

$$\begin{aligned}\Pr(\mathcal{A}) &= \int_0^1 \Pr(\mathcal{A}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= 1 - KE[x_1^{2z}]\end{aligned} ,$$

where the second line follows from the symmetry of p. Substituting into equation (9) yields the conditional density of x. The conditional expectation of x_1^2 is

$$\begin{aligned}E[x_1^2|\mathcal{A}] &\equiv \int_0^1 x_1^2 p_{\mathcal{A}}(\mathbf{x})d\mathbf{x} \\ &= \frac{E[x_1^2] - E[x_1^{2z+2}] - (K-1)E[x_1^2x_2^{2z}]}{1 - KE[x_1^{2z}]} ,\end{aligned}$$

where we have once again made use of the symmetry of p. This verifies equation (2).

Properties of the Dirichlet Distribution

Moments are expectations of powers and products of powers of the K allele frequencies. We will need formulas for such moments as $E[x_i]$, $E[x_i^2]$, and $E[x_i^2x_j^{2z}]$, where the expectation is taken with respect to the Dirichlet density defined in equation (3). These moments are all special cases of the general moment,

$$m(\mathbf{s}) \equiv E[x_1^{s_1}x_2^{s_2} \ldots x_K^{s_K}] ,$$

where $\mathbf{s} \equiv (s_1, s_2, \ldots, s_K)$ is a vector of integers. For example, $E[x_i]$ is obtained by setting $s_i = 1$ and setting $s_j = 0$ for all $j \neq i$; $E[x_i^2x_j^{2z}]$ is obtained by setting $s_i = 2$, $s_j = 2z$, and setting all the other s_k equal to zero. For the Dirichlet distribution, the general moment is (Wilks 1962, eq. [7.7.6] on p. 179)

$$m(\mathbf{s}) = \left(\frac{\Gamma(K\alpha)}{\Gamma(K\alpha + \sum_{i=1}^{K} s_i)}\right) \prod_{i=1}^{K} \left(\frac{\Gamma(\alpha + s_i)}{\Gamma(\alpha)}\right) . \quad (10)$$

This gives

$$E[x_i^r] = \frac{\Gamma(K\alpha)\Gamma(\alpha + r)}{\Gamma(K\alpha + r)\Gamma(\alpha)} \quad (11)$$

$$E[x_i] = \frac{1}{K} \quad (12)$$

$$E[x_i^r x_j^s] = \frac{\Gamma(K\alpha)\Gamma(\alpha + r)\Gamma(\alpha + s)}{\Gamma(K\alpha + r + s)\Gamma(\alpha)^2} . \quad (13)$$

The marginal distribution of x_1, obtained by integrating over $x_2, \ldots, x_{K-1}$, is

$$p_1(x_1) = \frac{x_1^{\alpha-1}(1 - x_1)^{(K-1)\alpha-1}}{\beta(\alpha, (K-1)\alpha)} , \quad (14)$$

which is a Beta density with parameters α and $(K-1)\alpha$ (Johnson and Kotz 1970).

The unconditional heterozygosity is

$$\begin{aligned}h &\equiv 1 - KE[x_1^2] \\ &= 1 - K\left(\frac{\Gamma(K\alpha)\Gamma(\alpha + 2)}{\Gamma(K\alpha + 2)\Gamma(\alpha)}\right) .\end{aligned}$$

Equation (5) is obtained by substituting the identity

$$\Gamma(x + i) \equiv \Gamma(x)x(x + 1) \ldots (x + i - 1) ,$$

which holds for any x and any integer i.

References

Abramowitz M, Stegun IA (1964) Handbook of mathematical functions with formulas, graphs, and tables. National Bureau of Standards, Washington, DC

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK, Cavalli-Sforza LL (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci USA 88:839–843

Bowcock A, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd J, Cavalli-Sforza LL (1994) High resolution of human evolu-

tionary trees with polymorphic microsatellites. Nature 368: 455–457

Cavalli-Sforza LL, Bodmer WF (1971) The genetics of human populations. WH Freeman, San Francisco

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

Chakraborty R, Fuerst PA, Nei M (1980) Statistical studies on protein polymorphism in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. Genetics 94:1039–1063

Ewens WJ (1979) Mathematical population genetics. Springer, New York

Harpending H (1994) Signature of ancient population growth in a low resolution mitochondrial DNA mismatch distribution. Hum Biol 66:591–600

Harpending HC, Sherry ST, Rogers AR, Stoneking M (1993) The genetic structure of ancient human populations. Curr Anthropol 34:483–496

Harris H, Hopkinson D (1972) Average heterozygosity in man. Ann Hum Genet 36:9–20

Johnson NL, Kotz S (1970) Distributions in statistics: continuous univariate distributions—2. Wiley, New York

Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, et al (1995*a*) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am J Hum Genet 57:523–538

Jorde LB, Carey JC, White RL (1995*b*) Medical genetics. Mosby, St Louis

Lewontin RC (1967) An estimate of average heterozygosity in man. Am J Hum Genet 19:681–685

Mountain JL (1994) Inferring human evolutionary history from mitochondrial DNA sequences and nuclear DNA allele frequencies. PhD thesis, Stanford University, Stanford

Mountain J, Cavalli-Sforza L (1994) Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. Proc Natl Acad Sci USA 91:6515–6519

Nei M, Chakraborty R, Fuerst PA (1976) Infinite allele model with varying mutation rate. Proc Natl Acad Sci USA 73: 4164–4168

Nei M, Livshits G, Ota T (1993) Genetic variation and evolution of human populations. In: Sing CF, Hanis CL (eds) Genetics of cellular, individual, family, and population variability. Oxford University Press, Oxford, pp 239–252

Nei M, Roychoudhury A (1974) Genetic variation within and between the three major races of Man, Caucasoids, Negroids, and Mongoloids. Am J Hum Genet 26:421–443

——— (1982) Genetic relationship and evolution of human races. In: Hecht MK, Wallace B, Prance CT (eds) Evolutionary biology. Vol 14. Plenum, New York, pp 1–59

Rogers AR (1995) Genetic evidence for a Pleistocene population explosion. Evolution 49:608–615

Sherry S, Rogers AR, Harpending HC, Soodyall H, Jenkins T, Stoneking M (1994) Mismatch distributions of mtDNA reveal recent human population expansions. Hum Biol 66: 761–775

Watkins WS, Bamshad MJ, Fraley AE, Jorde LB (1995) Population genetics of trinucleotide repeat polymorphisms. Hum Mol Genet 4:1485–1491

Wilks SS (1962) Mathematical statistics. Wiley, New York