Letter to the Editor

Population Bottlenecks and Patterns of Human Polymorphism

Jody Hey and Eugene Harris

Department of Genetics, Rutgers University

Fay and Wu (1999) examine a historical model in which a population quickly shrinks, stays small for some time, and then suddenly recovers its original population size, at which it stays until samples are drawn from it. This is a bottleneck model of the kind traditionally used in population genetics (Nei, Maruyama, and Chakraborty 1975); however, it is not the model that is most commonly considered in the context of genetic evidence of human population size change. That model is one of population size expansion, and it pervades the literature on analysis of mitochondrial DNA (e.g., Rogers and Harpending 1992), as well as the more recent literature on microsatellites (e.g., Kimmel et al. 1998), some of which explicitly confounds ''expansion'' with ''bottleneck.''

Fay and Wu (1999) repeatedly cite Hey (1997), who showed that nuclear genes and mtDNA are not both consistent with simple historical models. However, Hey's paper explored only models of constant population size and models of expansion. Both of these models are considerably simpler than a bottleneck model, and they were the models that had received most of the attention in the literature. However, Hey's (1997) paper may have contributed to a confusion over models, for the final sentence did mention ''bottleneck.''

Fay and Wu (1999) bring up a valuable point, that the time over which the pattern of polymorphism responds to a change in population size will be greater for larger populations. This means that the polymorphism patterns of genes with inherently different population sizes (e.g., mtDNA and nuclear genes) may be out of phase with one another if the period of population size fluctuation is sufficiently short and the fluctuation has been recent. Fay and Wu's letter is also timely in light of the increasing amount of nuclear gene data that are emerging. Since Hey's (1997) paper, which focused mostly on a few small data sets, larger data sets from PDHA1 (Harris and Hey 1999), β-globin (Harding et al. 1997), dystrophin (Zietkiewicz et al. 1998), and lipoprotein lipase (Clark et al. 1998) have been published, and all reveal an abundance of intermediate frequency polymorphism and positive values for Tajima's *D* (Tajima 1989*b*). Also recently, small samples from a series of X-linked loci revealed that five out of six loci had positive values of *D* (although three were just slightly positive) (Nachman et al. 1998). Thus, the overall impression from single-copy nuclear genes is of a deviation from a constant-population-size model that is in the direction of population shrinkage (Tajima 1989*a*). Without exception, all of these more recent works reveal relatively ancient times for the most recent common ancestors of genes, with reported values consistently on the order of 1 Myr. Thus, the abundance of middle-frequency variants means that the bulk of human heterozygosity is due to mutations that arose over 200,000 years ago, as shown explicitly in the most detailed genealogical studies (Harding et al. 1997; Harris and Hey 1999). Yet population size reduction cannot have been the case in more recent times, certainly since agriculture arose, and a reduction model does not fit well with the patterns of increasingly widespread occurrence of modern human fossils and artifacts within the past 100,000 years. Thus, quite apart from contrasts among different categories of genes with different population sizes, there is good reason to consider a model in which human ancestral populations became smaller and then later became larger. When we turn to the mitochondria and the more recent Y chromosome data (Underhill et al. 1997; P. Underhill and L. Jin, personal communication), we see that they are consistent with a model of recent population expansion, as both reveal an excess of low-frequency polymorphisms and negative *D* values. Thus, in a general way, there is a good fit between several types of data when they are considered under a model of population reduction that is followed after some time by expansion.

In considering how best to find a model that explains all of the data, there are three concerns that deserve mention. First is the point that Tajima's *D* statistic is not ideal for fitting models that depart from a constant historical population size. Under such models, the expected value of *D* depends fairly strongly on sample size, as shown in figure 1, which recreates figure 1*A* of Fay and Wu (1999), with curves generated for samples of 50 (as Fay and Wu had done) and 10. The latter set of curves are consistently much closer to zero, especially during the bottleneck. Second, genetic bottleneck models are complex, and properly include at least six parameters (population sizes for before, after, and during the bottleneck; duration of the bottleneck; time since the bottleneck; and mutation). It is usually reasonable to reduce these to five parameters by scaling both mutation and time to the population sizes (as was done for fig. 1), but bottleneck models will still surpass the scope for inference that is available in most summarizations of most data sets. Third, the mtDNA and the Y chromosome lack recombination and are expected to be subject to effects of an interaction of linkage and selection that reduces effective population size and increases the number of low-frequency polymorphisms (Hill and Robertson 1966) –

Key words: population bottleneck, human origins, polymorphism, coalescent.

Address for correspondence and reprints: Jody Hey, Department of Genetics, Rutgers University, Nelson Biological Labs, 604 Allison Road, Piscataway, New Jersey 08854-8082. E-mail: jhey@mbcl.rutgers.edu.

Mol. Biol. Evol. 16(10):1423–1426. 1999

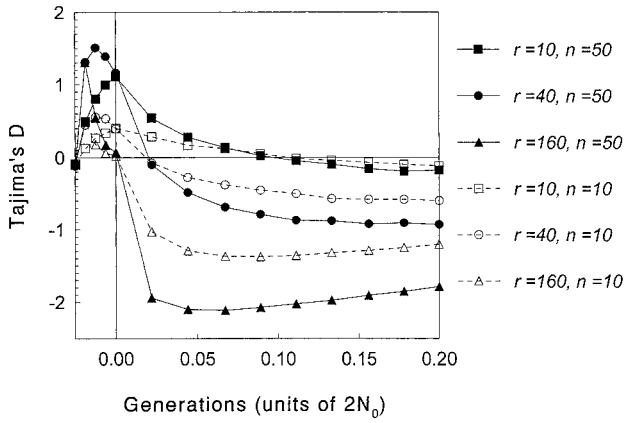

FIG. 1.—The effect of a bottleneck on Tajima's *D* as a function of sample size (*n*). Coalescent simulations were carried out as in Hey (1997) using parameter values given for figure 1 of Fay and Wu (1999). The starting population size is N_0, and bottleneck time and bottleneck population size scale in proportion to this quantity. The ratio of starting population size to the bottleneck population size (N_0/N_1) is given by *r*. Each point represents the mean of 5,000 independent simulations.

exactly the patterns that are reflected in the data for these genes. There is some evidence that these factors have shaped variation to some extent in the mtDNA (Nachman et al. 1996; Wise, Sraml, and Easteal 1998), but so far, there appears just a small signal of such an effect on the Y chromosome (Nachman 1998).

One way to mitigate all of these concerns is to work with a fuller description of polymorphism data, one that contains more information than does Tajima's *D*. In Hey's (1997) paper, *D* was used as a surrogate for the distribution of polymorphism frequencies. However, in fact, it is possible to use that distribution directly, as the expected values of the distribution are tractable under a variety of demographic models, including cases of population size change (Wakeley and Hey 1997). The expected values under models of selective neutrality also do not depend on recombination. While this does not bear on whether low-recombination genes have been shaped by natural selection, it is a feature of the analysis that adds considerable convenience when comparing data sets from genes with different recombinational histories.

If we assume that each polymorphism arose by not more than one mutation (Kimura 1969), then each is represented in a sample by an ancestral base and a mutant base (the latter can be identified by using an outgroup sequence). The frequency of the mutation in a sample of n sequences can vary between 1 and $n - 1$, and the expected number of observations in frequency class i, $E(s_i)$, under standard assumptions of neutrality and constant population size is known to be θ/i (Fu 1995), where θ is equal to twice the effective number of gene copies in the population times the neutral mutation rate. An expression for $E(s_i)$ is also known for a population that has undergone a sudden change in population size (Wakeley and Hey 1997). That expression was developed by considering the contribution to $E(s_i)$ from both before and after the time of change. This before-and-after approach can be compounded to consider multiple episodes of population size change. All that is needed, in effect, is to break up the period before the change into a full model that includes a new additional time, with additional before-and-after periods.

The strict bottleneck model represents a special case of a general model with two periods of population size change. The parameters are as follows: θ is as described above; N_0 and N_1 are the current and bottleneck population sizes, respectively; and T_0 and T_1 are the time to the bottleneck and the duration of the bottleneck in units of $2N_0$ and $2N_1$ generations, respectively. This notation is similar to that of Fay and Wu (1999). In the general model, N_1 can be greater or less than N_0, and we can allow the population size prior to the bottleneck to take on any value N_2. However, under a simple bottleneck model, $N_1 < N_0$, and $N_2 = N_0$. The only additional notation required is that of the number of ancestors of the n items that existed at times T_0 and T_1. Following the format that has been established so far, let $n_0 = n$, let n_1 be the number of ancestors at T_0, and let n_2 be the number at T_1.

When there has been just one period of population size change, the contribution to $E(s_i)$ from mutations that arose prior to the time of that change is

$$\sum_{n_1=2}^{n_0} P_{n_0 n_1}(T_0) \sum_{k=1}^{n_1-1} P(k \to i \,|\, n_0, n_1) \frac{\theta N_1/N_0}{k}, \qquad (1)$$

and the contribution from after the time of change is

$$\frac{\theta}{i} - \theta \sum_{n_1=2}^{n_0} P_{n_0 n_1}(T_0) \sum_{k=1}^{n_1-1} P(k \to i \,|\, n_0, n_1) \frac{1}{k}. \qquad (2)$$

These expressions follow equations (17) and (18) of Wakeley and Hey (1997), respectively. $P_{n_0 n_1}(T_0)$ is the probability that the sample of n_0 gene copies had n_1 ancestors at T_0 (Takahata and Nei 1985), and $P(k \to i \,|\, n_0, n_1)$ is the probability that a mutation of size k, when there are n_1 items, grows to size i when there are n_0 items (Wakeley and Hey 1997). The final component in equation (1), $\theta N_1/N_0/k$, is the expected number of mutations of size k at T_0, and it is this part of the overall expression that can be broken down into contributions from before and after an additional, more ancient, period of population size change (i.e., it is replaced by new expressions that resemble eq. 1 plus eq. 2). Carrying out this substitution within equation (1) and then summing with equation (2) yields

$$E(s_i) = \frac{\theta}{i} + \theta \sum_{n_1=2}^{n_0} P_{n_0 n_1}(T_0) \sum_{j=1}^{n_1-1} P(j \to i \,|\, n_0, n_1) \times \left(\frac{(N_1/N_0) - 1}{j} + \frac{(N_2 - N_1)}{N_0} \sum_{n_2=2}^{n_1} P_{n_1 n_2}(T_1) \times \sum_{k=1}^{n_2-1} \frac{P(k \to j \,|\, n_1, n_2)}{k} \right). \qquad (3)$$

Note that θ remains a scalar for the entire distribution. In other words, the bulk of equation (3) describes the relative height of the expected value of a frequency class, while the absolute height depends on θ. This is a

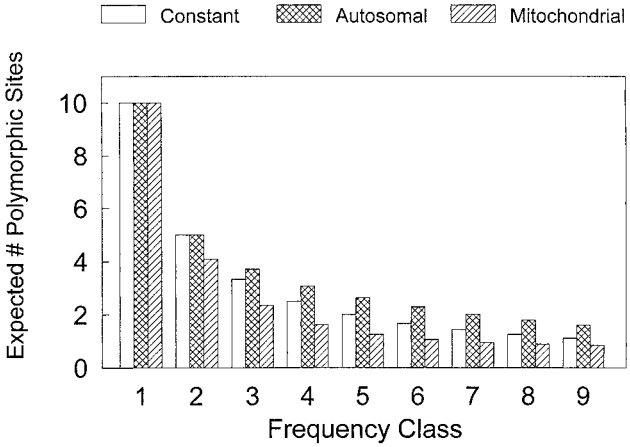

FIG. 2.—The effect of a bottleneck on the polymorphism frequency distribution for a sample of $n = 10$ DNA sequences. The parameter values for population size change were set to resemble those for figure 2 of Fay and Wu (1999): $N_0/N_1 = 20$; the duration of the bottleneck was 0.025 in units of N_0 generations (0.0125 in units of $2N_0$ generations) for the autosomal case and 0.1 for the mitochondrial case (i.e., four times that for the autosome because of the different mode of inheritance), while the time since the bottleneck was 0.1 for the autosome and 0.4 for the mitochondrial case. For example, if $N_0 = 30{,}000$, then $N_1 = 7{,}500$, the time to the bottleneck is 1,000 generations, and the duration of the bottleneck is 750 generations. The expected values under constant population size are simply θ/i (Fu 1995). To aid comparison, θ was set for each case so that $E(s_1) = 10$ (autosomal with bottleneck—$\theta = 14.5$, mitochondrial with bottleneck— $\theta = 12.7$; constant population size—$\theta = 10.0$).

useful point for considering the role of mutation in shaping the polymorphism distribution—under a simple mutation model, that role is limited to the overall height of the distribution and does not affect the expected relative values within it.

Expression (3) can be used to provide a different view of the bottleneck effect than is found with D in the simulations of Fay and Wu (1999). Figure 2 shows an example for the site frequency distribution for two genes that differ in their modes of inheritance and that have recently passed through a bottleneck. For comparison, the case for a constant population size is also shown, and in each case θ was set so that the height for class $i = 1$ was 10. The distribution for an autosomal gene and that for a mitochondrial gene are shaped differently, both with respect to each other and with respect to the constant-population-size case. The expected values of high- frequency polymorphisms are reduced for the mitochondrial curve (which causes a negative D), as expected under population expansion. However, for the autosomal example, the expected values of high-frequency polymorphisms are elevated (which causes a positive D). For the parameter values used in figure 2, the slower rate of drift for the autosome relative to the mitochondria causes the former to reflect the effects of the population reduction (the first step of the bottleneck) and causes the latter to reflect the effects of population expansion (the second step of the bottleneck). The bottleneck model, and the more general equation (3), have many parameters, and the scope for fitting all of them with any one data set will be limited. However, the effect described by Fay and Wu (1999) offers the hope that complex models of historical population size change can be fit to multilocus data sets. The contrasting pattern expected of genes with inherently differing effective population sizes can provide additional information beyond that contained within the pattern for any one gene.

LITERATURE CITED

CLARK, A., K. M. WEISS, D. A. NICKERSON et al. (11 coauthors). 1998. Haplotype structure and population genetic inferences from nucleotide sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. **63**:595–612.

FAY, J. C., and C. I. WU. 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol. Biol. Evol. **16**:1003–1005.

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

HARDING, R. M., S. M. FULLERTON, R. C. GRIFFITHS, J. BOND, M. J. COX, J. A. SCHNEIDER, D. S. MOULIN, and J. B. CLEGG. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60**:772–789.

HARRIS, E., and J. HEY. 1999. X chromosome evidence for ancient human histories. Proc. Natl. Acad. Sci. USA **96**: 3320–3324.

HEY, J. 1997. Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol. Biol. Evol. **14**:166–172.

HILL, W. G., and A. ROBERTSON. 1966. The effect of linkage on limits to artificial selection. Genet. Res. Camb. **8**:269–294.

KIMMEL, M., R. CHAKRABORTY, J. P. KING, M. BAMSHAD, W. S. WATKINS, and L. B. JORDE. 1998. Signatures of population expansion in microsatellite repeat data. Genetics **148**: 1921–1930.

KIMURA, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**:893–903.

NACHMAN, M. W. 1998. Y chromosome variation of mice and men. Mol. Biol. Evol. **15**:1744–1750.

NACHMAN, M. W., V. L. BAUER, S. L. CROWELL, and C. F. AQUADRO. 1998. DNA variability and recombination rates at X-linked loci in humans. Genetics **150**:1133–1141.

NACHMAN, M. W., W. M. BROWN, M. STONEKING, and C. F. AQUADRO. 1996. Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics **142**:953–963.

NEI, M., T. MARUYAMA, and R. CHAKRABORTY. 1975. The bottleneck effect and genetic variability in populations. Evolution **29**:1–10.

ROGERS, A. R., and H. HARPENDING. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9**:552–568.

TAJIMA, F. 1989*a*. The effect of change in population size on DNA polymorphism. Genetics **123**:597–601.

———. 1989*b*. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

TAKAHATA, N., and M. NEI. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110**:325–344.

UNDERHILL, P. A., L. JIN, A. A. LIN, S. Q. MEHDI, T. JENKINS, D. VOLLRATH, R. W. DAVIS, L. L. CAVALLI-SFORZA, and P. J. OEFNER. 1997. Detection of numerous Y chromosome

biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res. **7**:996–1005.

WAKELEY, J., and J. HEY. 1997. Estimating ancestral population parameters. Genetics **145**:847–855.

WISE, C. A., M. SRAML, and S. EASTEAL. 1998. Departure from neutrality at the mitochondrial NADH dehydrogenase subunit 2 gene in humans, but not in chimpanzees. Genetics **148**:409–421.

ZIETKIEWICZ, E., V. YOTOVA, M. JARNIK et al. (11 co-authors). 1998. Genetic structure of the ancestral populations of modern humans. J. Mol. Evol. **47**:146–155.

WOLFGANG STEPHAN, reviewing editor

Accepted June 30, 1999