Microsatellite Data Support an Early Population Expansion in Africa

Mark D. Shriver,[1,2] Li Jin,[3] Robert E. Ferrell,[1] and Ranjan Deka[1,4]

[1]Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania 15261; [3]Human Genetics Center, University of Texas Health Science Center, Houston, Texas 77225

We have developed a method for the analysis of microsatellite data that is useful in the elucidation of the demographic history of populations. This method, the P_K distribution method of pairwise comparisons, is analogous to the mismatch distribution of sequence comparisons developed for the analysis of mitochondrial sequence data by Rodgers and Harpending and is defined as the distribution of the number of repeat unit differences between alleles when each allele in a sample is compared with every other allele in the sample. Using computer simulations of microsatellite loci, we show that the shape of the distribution of P_K changes in a distinctive manner as a function either of time since population expansion or effective population size. Increases in both of these affect the P_K distribution in a similar fashion leading to a change from a steep distribution with a P_0 peak to one with a nonzero peak. Analysis of three data sets from surveys of microsatellite loci in ethnographically defined populations reveals that most (9/12) of the African populations analyzed, but none of the 30 non-African populations showed P_K distributions with nonzero peaks. These P_K distributions indicate either an earlier expansion or a larger effective population size for African populations. This observation is consistent with the hypothesized African origin of modern human.

Based on the topologies of mitochondrial and nuclear phylogenetic trees, it has been hypothesized that modern humans originated in Africa (Cann et al. 1987; Cavalli-Sforza et al. 1988; Nei and Roychoudhury 1993). If this were true, it would be expected that the genetic diversity of African populations would be greater than non-African populations as they would have expanded in size earlier. This increased genetic diversity of African populations, although clearly evident in the mitochondrial genome, has been more difficult to demonstrate for nuclear markers. Many classical polymorphic loci such as blood group, enzyme, and restriction fragment length polymorphic DNA markers demonstrate the highest levels of heterozygosity in European populations, reflecting the fact that these markers were first identified in European populations (Mountain and Cavalli-Sforza 1994; Rodgers and Jorde 1995). Because of their high degree of polymorphism, microsatellites are less affected by such ascertainment bias. Although large surveys of microsatellite loci in human populations have reported higher levels of heterozygosity in African populations (Bowcock et al. 1994; Deka et al. 1995a,b), most often these differences are not significant. We present a new method for analysis of microsatellite data that clearly shows an earlier expansion and/or a larger effective size of African populations.

Although the precise molecular mechanism of mutational change in repeat number of microsatellite loci has not been elucidated, it is evident from observed mutations and allele frequency distributions that they evolve via a forward–backward stepwise mutational process (Shriver et al. 1993; Weber and Wong 1993; DiRienzo et al. 1994). This observation implies that there will be evolutionary information in the allele size distribution, as alleles closer in size most likely share a more recent common ancestor than alleles with larger size differences. Several new measures of genetic distance have been developed that use the evolutionary information present in the size of microsatellite alleles (Goldstein et al. 1995a,b; Shriver et al. 1995; Slatkin 1995; Kimmel et al. 1996). We have developed a method for the analysis of microsatellite data that uses the evolutionary information inherent in microsatellite allele length to elucidate the magnitude of genetic diversity within populations. This method, the P_K distribution method of pairwise comparisons, is analogous to the mismatch distribution of sequence comparisons developed for the analysis of mito-

Present address: [2]Allegheny University of the Health Sciences, Pittsburgh, Pennsylvania.
[4]Corresponding author.
E-MAIL rdeka@helix.hgen.pitt.edu; FAX (412) 624-3020.

7:586–591 ©1997 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/97 $5.00

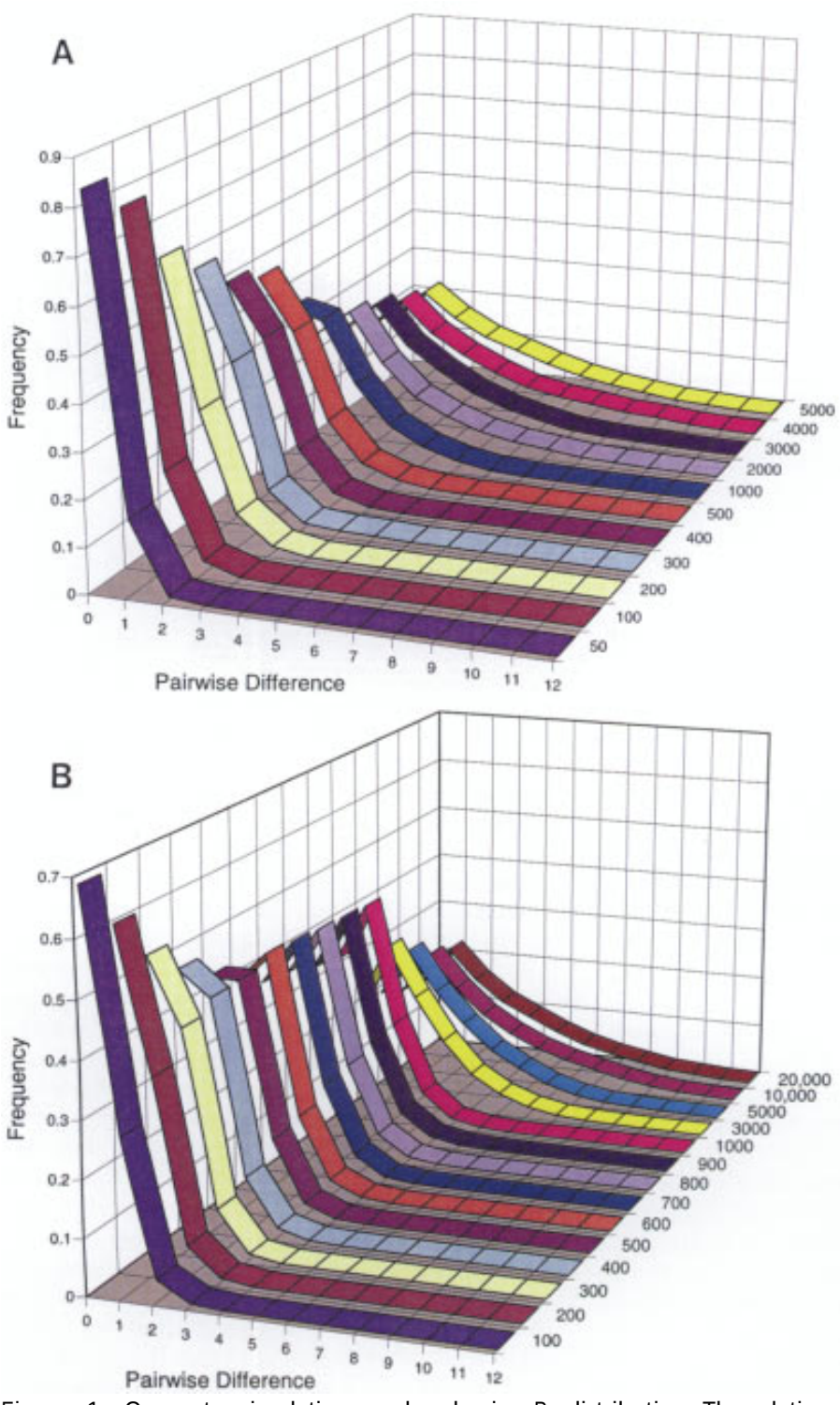

Figure 1 Computer simulations analyzed using P_K distribution. The relationship between effective population size and P_K distribution shape is shown in *A*. Steady-state P_K distributions are shown for 11 populations ranging in N_e from 50 to 5000. The relationship between the time since population expansion and the P_K distribution is shown in *B*. P_K distributions for 14 time points (in generations) after a 1000-fold increase in population size. Specifics for the simulations are given in Methods.

chondrial sequence data by Rodgers and Harpending (1992), and is defined as the distribution of the number of repeat unit differences between alleles when each allele in a sample is compared with every other allele in the sample.

RESULTS

We have applied the P_K method to the interpretation of several sets of simulated microsatellite data. Genetic loci subject to forward–backward mutational events, when unconstrained, do not reach equilibrium with respect to the identity of alleles or their frequency profiles (Moran 1975). However, after a number of generations (~$4N_e$, where N_e is the effective population size) a steady state is reached for certain summary statistics, including heterozygosity, number of alleles, and other measures of genetic diversity (Shriver et al. 1993). Similarly, the P_K distribution also reaches a steady state acquiring a shape that does not change after $4N_e$ generations at a given mutation rate and effective population size (Moran 1975). Figure 1A shows the shape of the P_K distribution at equilibrium for populations of different effective size. It is clear that the shape of the P_K distribution is related to N_e, with smaller populations having P_0 peaks (most pairwise comparisons showing no difference in size), larger populations have P_1 peaks (most pairwise comparisons differing by one repeat), and intermediate sized populations (e.g., $N_e = 1000$) having plateaus. When a

population is not in equilibrium (e.g., a population that has recently expanded in size), the P_K is also dependent on the demographic history (e.g., the magnitude of and time since population expansion). Figure 1B shows a series of P_K distributions at different time points in generations since a population expansion event. For these computer simulations, parameter levels were set to reasonable estimates for human populations, namely $\mu = 0.001$, final $N_e = 5000$, and magnitude of expansion = 1000 (see Methods for more detail). These simulations were performed as an instructive example, and we do not presume to accurately model human history with this simple design. Nonetheless, there is a clear relationship between time since expansion and the distribution of P_K. After a population has expanded, the distribution has a peak at P_0 from immediately following expansion to a point at which the peak shifts to P_1. This distribution then flattens progressively until a steady-state distribution is reached at equilibrium, $4N_e$ generations after expansion.

We have also compiled P_K distributions for three independent sets of microsatellite data [our own (Deka et al. 1995a,b) and two from the literature (Bowcock et al. 1994; Jorde et al. 1995)]. These results are shown in Figure 2. Figure 2A shows the results of data gathered in the laboratory of R.D., which includes 24 microsatellite loci in 15 human populations belonging to five major groups: three African, four Caucasian, three Asian Mongoloid, two Pacific Islander, and three Amerindian. Africans are the only populations with prominent nonzero peaks, whereas the Amerindian and Pacific Island populations show very steep P_K distributions, with the Caucasian and Asian populations having intermediate patterns. The steep slopes and high P_0 peaks observed in the Amerindians and Pacific Island populations are consistent with their recent settlement and population history. Two populations, Brazilian whites and Bramin, show plateaus in the distribution of P_K, where $P_0 \cong P_1$. Figure 2B shows the P_K distributions for a set of microsatellite frequency data on 30 loci in 13 populations (6 African, 2 European, and 5 Asian) reported by Jorde et al. (1995). Four of the six African populations again show distinct nonzero peaks, whereas for these loci the other two African, the two European, and one of the Asian populations show plateau P_K distributions. The P_K distributions of the other five Asian populations show steeper P_0 peaks. Figure 2C shows the P_K distributions for 30 microsatellite loci reported by Bowcock et al. (1994). This group studied a total of 14 populations, including three African, two European, three Asian, three Amerindian, and three Austronesian. For these data, two of the three African populations show nonzero peaks and all other populations show peaks at P_0. Overall, in three large microsatellite studies on ethnically and geographically well-defined populations, 9/12 African populations show P_K peaks at P_1, 2/12 show plateaus where $P_0 \cong P_1$, and 1/12 shows a P_0 peak; 0/8 Caucasian populations show a P_1 peak, 4/8 show P_0 peaks, and 4/8 show plateaus; 0/11 Asian populations show a P_1 peak, 10/11 show a P_0 peak, and 1/11 shows a plateau; 11/11 other populations have P_0 peaks and none show peaks at P_1 or plateaus.

DISCUSSION

There is good evidence from observed mutational events and the good fit of microsatellite loci to the stepwise mutation/drift model that the majority of di-, tri-, and tetranucleotide repeat loci evolve via a stepwise mutational mechanism, most likely replication slippage (Levinson and Gutman 1987; Shriver et al. 1993; Weber and Wong 1993; DiRienzo et al. 1994). It is also clear that there are differences in the mutational spectrum among specific microsatellite loci and classes of loci (Shriver et al. 1993; Weber and Wong 1993; DiRienzo et al. 1994; Chakraborty et al. 1997). Notwithstanding these interlocus differences, the P_K distribution approach is applicable when the same set of loci are analyzed in all populations being considered. In addition, because the P_K is the average distribution of pairwise differences at many loci, the effect of a rare locus that deviates from the stepwise mutational process or has an altered mutational spectrum in one or more populations will be diminished by the other loci in the survey.

The analysis of P_K distributions among ethnically and geographically well-defined populations shows that African populations have P_K distributions that are more similar in form to the simulated equilibrium distributions than the P_K distributions of non-African populations. The P_K distribution results can also be interpreted as African populations

Figure 2 P_K distributions for three sets of microsatellite data. Microsatellite data generated in the laboratory of R.D. on 24 loci in 15 populations were analyzed using P_K and are shown in *A.* Data for 30 loci in 13 populations presented in Jorde et al. (1995) are shown in *B. C* shows P_K distributions for data on 30 microsatellite loci on 14 populations by Bowcock et al. (1994).

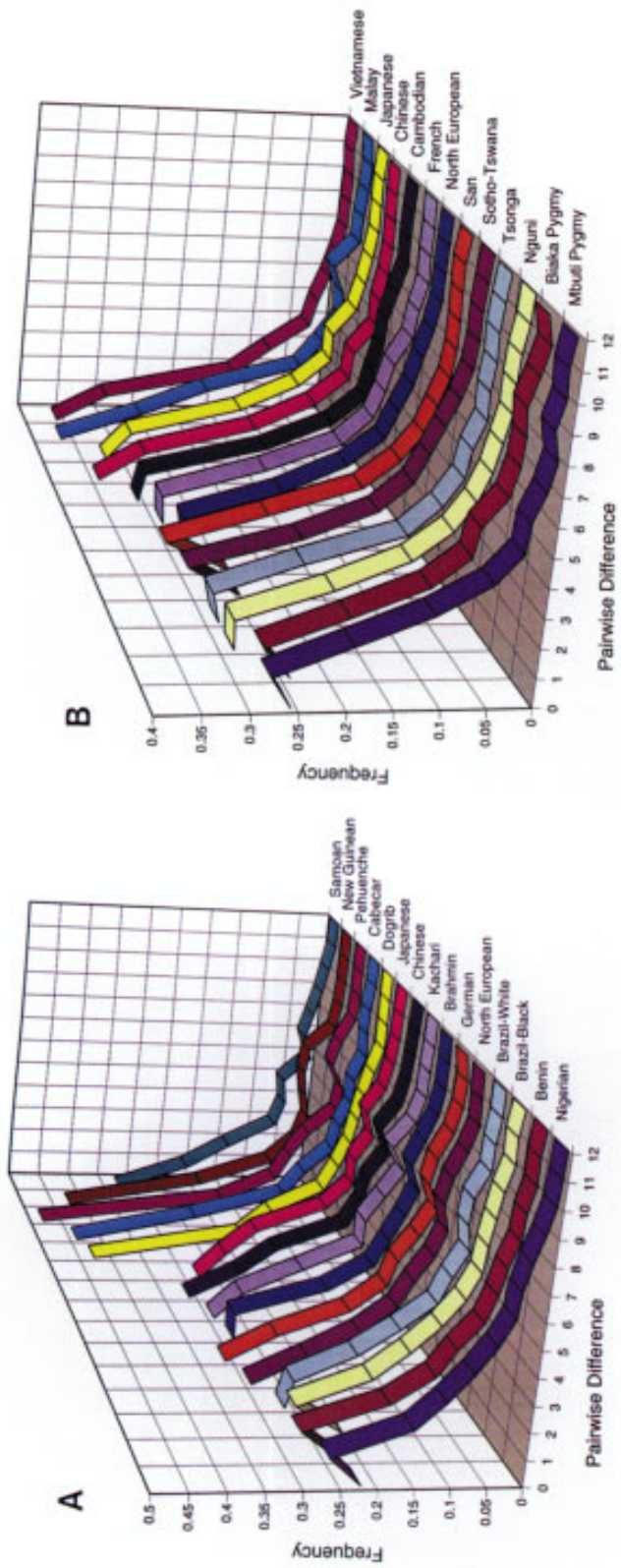

Figure 2 *(See facing page for legend.)*

having a larger effective population size (N_e) than non-African populations. Computer simulations of microsatellite loci show that both larger equilibrium N_e and longer time since population expansion can cause the P_1 peaks that are characteristic of African populations. Because N_e is a function of the population size for many generations in the past, and the contemporary census size of African populations is not larger than European or Asian populations, we can conclude that the ancestral African population was larger than the ancestral populations of contemporary inhabitants of the other continents. This could be so if African population expansion preceded non-African population expansion. Alternatively, the smaller N_e of contemporary non-African populations could be the result of a reduction in the population size, or population bottleneck, in the ancestral populations of non-Africans that Africans did not experience. Given other genetic and fossil evidence, it is likely that the actual course of history involved some combination of these two models. Population bottlenecks have been recognized to occur when migrating groups surmount geographical barriers as in the peopling of the Pacific and the Americas. It has been suggested that the movement of populations northward out of Africa would have resulted in a similar reduction in population size. We thus find that these P_K distribution results are most consistent with an African origin of modern humans followed by migrations of smaller groups out of Africa. Although these data do not exclude admixture between populations moving out of Africa and local archaic populations, they are inconsistent with a European or Asian origin of humans unless likely scenarios for population bottlenecks related to population movement within Europe and Asia but not Africa can be advanced.

METHODS

Computer simulations of the one-step stepwise mutation model were performed to investigate the effects of population expansion and effective population size on the distribution of P_K. The design of these simulations is based on previous software used to study the population dynamics of the stepwise mutation model (Shriver et al. 1993, 1995). Each generation consists of randomly drawing a number of alleles ($2N_e$ alleles) from the previous generation. Once an allele is selected from the last generation, a second random number is used to determine whether this allele will mutate and if the allele does mutate (this random number is <0.001, the mutation rate), a third random number is used to determine in which direction the allele will mutate, one step larger or one step smaller. To reach a steady-state P_K distribution, simulations were carried out for $4N_e$ generations. Previous simulation studies have shown that the stepwise mutation model reaches steady state by this point (Shriver et al. 1993). To simulate the effects of population expansion, we used equilibrium distributions for an $N_e = 5$ and increased the effective population size to $N_e = 5000$ in one generation. This model for population expansion is similar to that used by researchers studying the dynamics of sequence mismatch distributions (Rodgers and Harpending 1992). In addition to this instantaneous increase model, we simulated exponential expansion rates and expansion rates of 1% per year [a conservative estimate of the rate of human population expansion (Rodgers and Jorde 1995)]. Within this range, the rate of increase in population size had negligible effect on the shape of the P_K (data not shown). The simulated distributions show the results of sampling 50 individuals (100 chromosomes) for 100 independent microsatellite loci.

ACKNOWLEDGMENTS

We thank Dr. R. Chakraborty for helpful discussion and encouragement in the course of this study. This work was supported in part by grants to M.S. from the National Institute of Justice (95-IJ-CX-0008) and the Keck Foundation for Advanced Training in Computational Biology, and from the National Institutes of Health (NIH) to R.D. (GM-45861), and an NIH Training Grant (T32-GS08404) to L.J.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ''advertisement'' in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bowcock, A.M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, and L.L. Cavalli-Sforza. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368: 455–457.

Cann, R., M. Stoneking, and A. Wilson. 1987. Mitochondrial DNA and human evolution. *Nature* 325: 31–36.

Cavalli-Sforza, L.L., A. Piazza, and P. Menozzi. 1988. Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci.* 85: 6002–6006.

Chakraborty, R., M. Kimmel, D.N. Stivers, L.J. Davison, and R. Deka. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl. Acad. Sci.* 94: 1041–1046.

Deka, R., L. Jin, M.D. Shriver, L.M. Yu, S. DeCroo, J. Hundrieser, C.H. Bunker, R.E. Ferrell, and R. Chakraborty. 1995a. Population genetics of dinucleotide (dC-cA)/dG-dT) polymorphisms in world populations. *Am. J. Hum. Genet.* 56: 461–474.

Deka, R., M.D. Shriver, L.M. Yu, R.E. Ferrell, and R. Chakraborty. 1995b. Intra- and inter-population diversity at short tandem repeat loci in diverse populations of the world. *Electrophoresis* 16: 1659–1664.

DiRienzo, A., A.C. Peterson, J.C. Garza, A.M. Valdes, M. Slatkin, and N.B. Freimer. 1994. Mutational processes of

simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci.* 91: 3166–3170.

Goldstein, D.B., A. Ruiz-Linares, L.L. Cavalli-Sforza, and M.W. Feldman. 1995a. An evaluation of genetic distances for use with microsatellite loci. *Genetics* 139: 463–471.

———. 1995b. Genetic absolute dating based on microsatellites and the origins of modern humans. *Proc. Natl. Acad. Sci.* 92: 6723–6727.

Jorde, L., M.J. Bamshad, W.S. Watkins, R. Zenger, A.E. Fraley, P.A. Krakowiak, K.D. Carpenter, H. Soodyall, T. Jenkins, and A.R. Rodgers. 1995. Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. *Am. J. Hum. Genet.* 57: 523–538.

Kimmel, M., R. Chakraborty, D. Stivers, and R. Deka. 1996. Dynamics of repeat polymorphisms under a forward-backward mutation model: Within- and between-population variability at microsatellite loci. *Genetics* 143: 549–555.

Levinson, G. and G.A. Gutman. 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4: 203–221.

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

Moran, P.A.P. 1975. Wandering distributions and the electrophoretic profile. *Theor. Popul. Biol.* 8: 318–330.

Nei, M. and A. Roychoudhury. 1993. Evolutionary relationships of human populations on a global scale. *Mol. Biol. Evol.* 10: 927–943.

Rodgers, A. and H. Harpending. 1992. Population growth makes waves in the distribution of pairwise differences. *Mol. Biol. Evol.* 9: 552–569.

Rodgers, A. and L. Jorde. 1995. Genetic evidence on modern human origins. *Hum. Biol.* 67: 1–36.

Shriver, M.D., L. Jin, R. Chakraborty, and E. Boerwinkle. 1993. VNTR allele frequency distributions under a stepwise mutation model: A computer simulation approach. *Genetics* 134: 983–993.

Shriver, M.D., L. Jin, E. Boerwinkle, R. Deka, R.E. Ferrell, and R. Chakraborty. 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Genetics* 12: 914–920.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies *Genetics* 139: 457–462.

Weber, J.L. and C. Wong. 1993. Mutation of human short tandem repeats. *Hum. Mol. Genet.* 2: 1123–1128.

Received January 31, 1997; accepted in revised form April 25, 1997.