4-1-2007

Ancestral alleles and population origins: Inferences depend on mutation rate

Alan R. Rogers
The University of Utah

Stephen Wooding
The University of Utah

Chad D. Huff
The University of Utah

Mark A. Batzer
Louisiana State University

Lynn B. Jorde
The University of Utah

Recommended Citation

Rogers, A., Wooding, S., Huff, C., Batzer, M., & Jorde, L. (2007). Ancestral alleles and population origins: Inferences depend on mutation rate. *Molecular Biology and Evolution, 24* (4), 990-997. https://doi.org/10.1093/molbev/msm018

Ancestral Alleles and Population Origins: Inferences Depend on Mutation Rate

Alan R. Rogers, Stephen Wooding,† Chad D. Huff,* Mark A. Batzer,‡ and Lynn B. Jorde†*

*Department of Anthropology, University of Utah; †Department of Human Genetics, University of Utah School of Medicine; and‡Department of Biological Sciences, Biological Computation and Visualization Center, Center for BioModular Multi-Scale Systems, Louisiana State University

Previous studies have found that at most human loci, ancestral alleles are "African," in the sense that they reach their highest frequency there. Conventional wisdom holds that this reflects a recent African origin of modern humans.

This paper challenges that view by showing that the empirical pattern (of elevated allele frequencies within Africa) is not as pervasive as has been thought. We confirm this African bias in a set of mainly protein-coding loci, but find a smaller bias in *Alu* insertion polymorphisms, and an even smaller bias in noncoding loci. Thus, the strong bias that was originally observed must reflect some factor that varies among data sets—something other than population history. This factor may be the per-locus mutation rate: the African bias is most pronounced in loci where this rate is high.

The distribution of ancestral alleles among populations has been studied using 2 methods. One of these involves comparing the fractions of loci that reach maximal frequency in each population. The other compares the average frequencies of ancestral alleles. The first of these methods reflects history in a manner that depends on the mutation rate. When that rate is high, ancestral alleles at most loci reach their highest frequency in the ancestral population. When that rate is low, the reverse is true. The other method—comparing averages—is unresponsive. Average ancestral allele frequencies are affected neither by mutation rate nor by the history of population size and migration. In the absence of selection and ascertainment bias, they should be the same everywhere. This is true of one data set, but not of 2 others. This also suggests the action of some factor, such as selection or ascertainment bias, that varies among data sets.

Introduction

An interest in modern human origins has led many geneticists to compare the frequencies of "ancestral alleles" in human populations. In a genetic sample from a single locus, the ancestral allele is the allelic state of the last common ancestor (LCA). Previous work suggests that the vast majority of human ancestral alleles are "African" in the sense that they reach their highest frequency there. Geneticists disagree, however, about what this implies. According to some, it reflects a recent African origin of modern humans (Takahata et al. 2001; Excoffier 2002; Satta and Takahata 2002, 2004). According to others, it refutes this view (Templeton 2002, p. 50).

This disagreement reflects in part the lack of any well-developed theory. Takahata et al. (2001) and Satta and Takahata (2002, 2004) have explored the effect on ancestral alleles of various models of migration and population history. Here, we consider another factor: the mutation rate per locus.

In addition to this weakness of theory, our understanding is also clouded by a tendency to conflate similar concepts. In particular, the term "ancestral allele" is used in 2 senses. For some authors, the term refers to what we will call the "narrow-sense ancestral allele" (NAA)—the allele carried by the LCA of a sample. For others, the term refers to the allele within the modern sample that differs least from the NAA. We refer to this as the "broad-sense ancestral allele" (BAA). Although the distinction between them is usually blurred, we show that these 2 forms of ancestral allele can behave differently.

Finally, different authors use different methods for comparing frequencies of ancestral alleles. One method, which we call "comparing fractions," first calculates the fractions of loci at which the ancestral allele is most common in each subpopulation and then compares these fractions. For example, Takahata et al. (2001) found that ancestral alleles reached highest frequency in Africa at 9 of the 10 loci in their sample.

Others compare ancestral alleles using a different method, which we call "comparing averages." This method first averages ancestral allele frequencies across loci within each subpopulation and then compares these averages. For example, Watkins et al. (2003) studied a sample of 100 *Alu* loci. Within this sample, the mean ancestral allele frequency was about 20% higher in Africa than in non-African populations. Similar results were obtained by Mountain et al. (1992) (using classical loci) and by Mountain and Cavalli-Sforza (1994) (using restriction polymorphisms). Comparing averages is confusingly similar to comparing fractions, and the 2 methods yield similar results in these examples. Yet, we will show that they differ in important ways.

In what follows, we explore the effect of mutation rate on ancestral alleles, first by 2 simple theoretical arguments and then by computer simulation. In our exposition, "gene" will refer to a physical copy of some genetic locus, which may or may not code for protein, and "allele" will refer to any of the alternative forms that the genes at a given locus may take. In sequence data, "allele" is a synonym for "haplotype."

Three data sets

Data Set I

The first of our 3 data sets (table 1) was gleaned from published literature. It includes most of the loci studied by Takahata et al. (2001) and Satta and Takahata (2002, 2004), but adds additional loci and omits those with samples of fewer than 20 chromosomes. For the loci marked with a dagger (†), we reestimated allele frequencies and used DNAML (Felsenstein 2004) to infer the ancestral human state of each locus. At these loci, we discarded a few

Key words: ancestral allele, last common ancestor, human evolutionary history, replacement hypothesis, multiregional hypothesis, diffusion wave hypothesis.

E-mail: rogers@anthro.utah.edu.

Mol. Biol. Evol. 24(4):990–997. 2007
doi:10.1093/molbev/msm018
Advance Access publication January 30, 2007

Table 1
Data Set I: Counts and Frequencies of Ancestral Alleles

Locus	NAA	*S*	Africa ($x/n = p$)	Asia ($x/n = p$)	Europe ($x/n = p$)
MAOA[a]	*	5	1/40 = **0.025**	0/33 = 0	0/73 = 0
FIX[b]	*	6	11/18 = 0.611	13/13 = **1**	3/5 = 0.6
ECP[c]	*	7	4/42 = **0.095**	1/34 = 0.029	0/32 = 0
EDN[c]	*	9	3/40 = **0.075**	0/34 = 0	0/60 = 0
ZFX[d]	*	10	3/113 = **0.027**	0/129 = 0	0/93 = 0
RRM2P4[e,f]		13	1/10 = **0.1**	1/21 = 0.048	0/10 = 0
CCR5[g,h]	*	15	23/116 = **0.198**	11/108 = 0.102	5/48 = 0.104
MC1R[h,i]	*	16	70/148 = **0.473**	17/168 = 0.101	5/356 = 0.014
TNFSF5[f]		16	1/10 = **0.1**	0/21 = 0	0/10 = 0
AMELX[f]	*	17	2/10 = 0.2	2/21 = 0.095	3/10 = **0.3**
CYP1A2[h,j]	*	17	1/60 = 0.017	0/102 = 0	1/46 = **0.022**
PsGBA[h,k]		17	1/30 = **0.033**	0/40 = 0	0/19 = 0
APXL[f]	*	19	1/10 = **0.1**	2/21 = 0.095	0/10 = 0
β-Globin[l]	*	21	9/103 = **0.087**	0/200 = 0	0/46 = 0
PDHA1[m]		25	1/16 = **0.063**	0/13 = 0	0/6 = 0
Dmd[n]		28	1/10 = **0.1**	0/21 = 0	0/10 = 0
Xq13.3[h,o]	*	32	4/23 = **0.174**	3/31 = 0.097	0/11 = 0
HFE[p]		41	2/20 = **0.1**	0/20 = 0	0/20 = 0
Y[h,q]	*	43	25/358 = **0.07**	0/2150 = 0	0/316 = 0
MtDNA[h,r]	*	177	1/143 = **0.007**	0/319 = 0	0/99 = 0
apoB[s]	*		123/194 = **0.634**	41/448 = 0.092	98/442 = 0.222
apoE[t]	*		286/1294 = **0.221**	139/1538 = 0.09	323/2126 = 0.152

NOTE.—*, NAA; *S*, number of segregating sites (ignoring indels); *x*, number of copies of the AA; *n*, haploid sample size; and *p*, AA frequency (bold face indicates location of maximum AA frequency for each locus).

[a] Balciuniene et al. (2001).
[b] Harris and Hey (2001).
[c] Zhang and Rosenberg (2000).
[d] Jaruzelska et al. (1999).
[e] Garrigan et al. (2005) (treating alleles B and C as a single allele).
[f] Hammer (2004).
[g] Bamshad et al. (2002).
[h] Frequencies and human LCA reestimated.
[i] Harding et al. (2000) and Rana et al. (1999).
[j] Wooding et al. (2002).
[k] Martinez-Arias et al. (2001).
[l] Harding et al. (1997).
[m] Harris and Hey (1999).
[n] Nachman and Crowell (2000) (intron 7 only).
[o] Kaessmann et al. (1999).
[p] Toomajian and Kreitman (2002).
[q] Hammer et al. (2001).
[r] Vigilant et al. (1991).
[s] Breguet et al. (1990) and Rapacz et al. (1991).
[t] Hallman et al. (1991) and Zekraoui et al. (1997).

DNA sequences that were ambiguous or incomplete. This accounts for the minor differences between the values of *n* and *S* in table 1 and those in the original publications.

Data Set II

The second data set consists of 100 *Alu* insertion polymorphisms (Watkins et al. 2003). At each locus, the *Alu*-absent allele is the NAA.

Data Set III

The third data set consists of 38 loci, each comprising about 500 bp of noncoding DNA sequence (Yu et al. 2002). These loci are a subset of the 50 described by Yu et al. We excluded 8 loci with substantial missing data and then eliminated all nucleotide positions with missing values. Several of the resulting loci were monomorphic and were eliminated for consistency with data sets I and II. We then used PHASE 2.0.2 (Stephens et al. 2001) to infer haplotypes and inferred ancestral states as described above, using chimpanzee (Chen and Li 2001) as an outgroup.

In analysis of data sets I and III, we ignore the statistical error involved in determining which allele is ancestral. There is no such error with data set II. In analyzing these data, we encounter 2 sorts of tie. First, at some loci, 2 or more alleles differ from the NAA by the same minimum amount. In such cases, we treat all the tied alleles as a single allele. Second, the maximum frequency of the AA sometimes occurs in more than one population. When *k* populations are tied at a locus, we allocate a fraction $1/k$ to each tied population.

Comparing Fractions

Table 1 presents the raw data for data set I, one row for each locus. For each locus, it shows the type (narrow or broad) of AA and the number *S* of segregating sites. For each population, it shows the number *x* of copies of the AA, the haploid sample size *n*, and the frequency

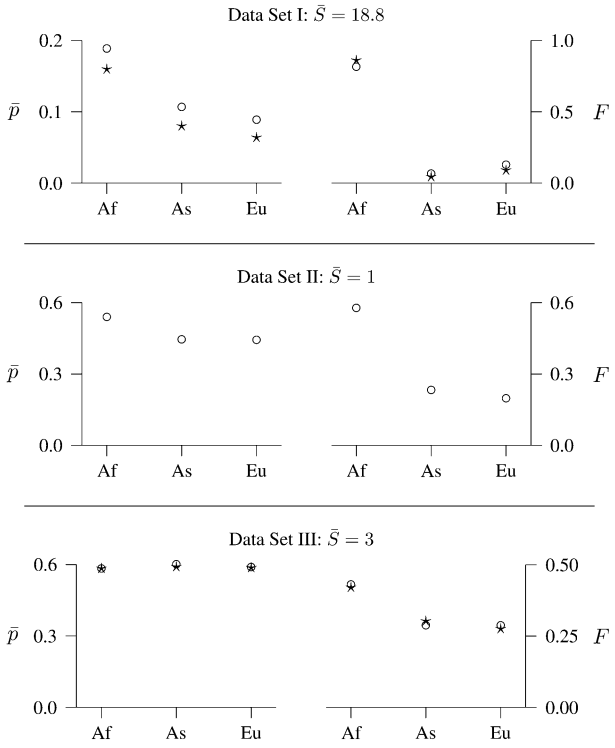

FIG. 1.—Comparing averages (left panels) and comparing fractions (right panels). Symbols: Af, Africa; As, Asia; Eu, Europe; °, NAA; and ★, BAA.

$p = x/n$. Within each row, the largest value of p is printed in bold type.

We used these data to calculate the fraction F of loci at which the ancestral allele is most common in each subpopulation. These values are shown in the upper-right panel of figure 1. With this data set, the African F is much higher than those of the other 2 populations. To evaluate the statistical significance of this pattern, we fit the data to a null hypothesis in which the ancestral allele is equally likely to reach maximal frequency in any population. Our statistical test is based on a goodness-of-fit statistic (Press et al. 1992, p. 665),

$$H = \sum_{i=1}^{K} \frac{(y_i - m_i)^2}{v_i},$$

where i indexes populations, $K = 3$ is the number of populations, y_i the number of loci whose ancestral allele reaches maximal frequency in population i, and m_i and v_i the mean and variance of y_i under the null hypothesis. In the absence of ties, y_i would be binomial with parameters $1/K$ and L (the number of loci). Thus, we use the binomial formulas $m_i = LK^{-1}$ and $v_i = LK^{-1}(1 - K^{-1})$.

These formulas for m_i and v_i are only approximations because our y_i are only approximately binomial. (As explained above, some loci exhibit ties between populations, and the y_i are not all integers.) Consequently, we cannot assume the usual chi-squared distribution. We generate the sampling distribution of H by computer simulation.

Let x_{ij} represent the number of copies of the ancestral allele of locus i in population j and n_{ij} the corresponding haploid sample size. Our null hypothesis assumes that x_{ij} is drawn from a binomial distribution with parameters z_i [the global allele frequency at locus i] and n_{ij}. We estimate each z_i from the allele frequency in the sample as a whole and then repeatedly generate data sets by sampling from the binomial distribution just described. Each simulated data set is used to calculate y_i and H. The tail probability P is estimated by the fraction of 10^6 replicates in which the simulated H is at least as large as the observed one.

With data set I, the African F is far the largest, and this African excess is highly significant ($P = 0.0002$ for NAA and $P = 1 \times 10^{-6}$ for BAA). Although our estimates differ slightly because of the loci included, this is essentially the result that led Takahata and Satta (Takahata et al. 2001; Satta and Takahata 2002, 2004) to support an African origin and Templeton (2002, p. 50) to reject it.

The other 2 panels on the right side of figure 1 repeat this analysis using data sets II and III. The results are similar but less extreme: the fraction of loci with African ancestral alleles falls to 0.58 in data set II and to 0.42 in data set III. The first of these values is a significant departure from the null hypothesis ($P = 1.1 \times 10^{-5}$), but the second is not. Thus, the African excess is progressively less pronounced in data sets II and III.

Before attempting to explain this difference, let us apply a different method: that of comparing averages.

Comparing Averages

Within each population, $\bar{p}$ will refer to the average across loci of ancestral allele frequencies. Several authors have used this statistic to compare populations (Mountain et al. 1992; Mountain and Cavalli-Sforza 1994; Watkins et al. 2003), a process that we call "comparing averages." We use it to compare populations in the panels on the left side of figure 1. In data set I (upper left panel), African values are nearly twice as large as Eurasian ones. This African excess falls to about 20% in data set II (middle left panel), and it disappears altogether in data set III (lower left panel). Thus, the 3 data sets present no consistent picture of the relationship between African and Eurasian values of $\bar{p}$.

To make sure that these differences are not statistical flukes, we used a randomization test to compare the African value of $\bar{p}$ with the average of European and Asian values. In each repetition, we generated new data vectors by randomly swapping the African and Eurasian values of p at each locus and then calculated the absolute difference between the means of the resulting data vectors. The tail probability, P, is the fraction of 10^6 randomized absolute differences that are at least as large as the observed difference. This method showed that African and Eurasian values of $\bar{p}$ differ significantly in data sets I and II but not in data set III ($P = 0.018$, 0.0008, and 3×10^{-5} for the NAA of data set I, the BAA of data set I, and the NAA of data set II, respectively). Thus, the African excess seen in data sets I and II is real. On the other hand, it is not equally strong in all data sets. This implies the action of some factor that varies among data sets. We consider here the possibility that ancestral allele frequencies respond to differences in mutation rate.

Although we have not measured the mutation rate, we have measured the number S of segregating (i.e.,

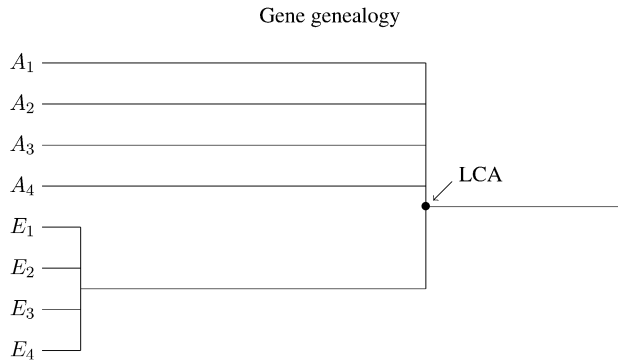

FIG. 2.—A hypothetical gene genealogy reflecting the effects of an African origin. The long branch leading to samples from population *E* reflects a bottleneck associated with the colonization by emigrants from population *A*.

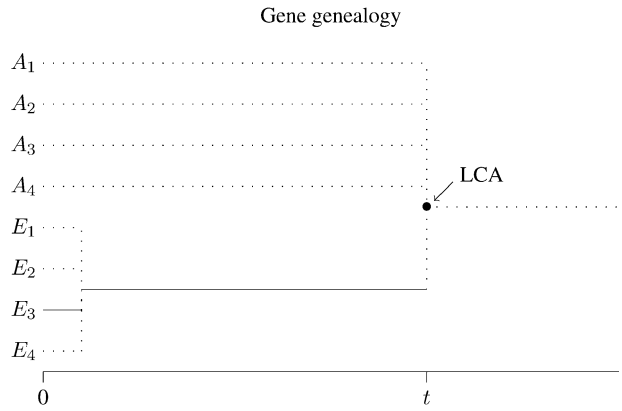

FIG. 3.—The probability that E_3 (or any other gene) carries the EAA depends on the length t of the branch back to the LCA. (This branch is emphasized.)

polymorphic) sites at each locus. And because S tends to increase with the per-locus mutation rate, it will serve as a useful proxy. In data set I, S ranges from 5 to 44 among autosomal loci and averages $\bar{S} = 18.8$. For the *Alu* loci (data set II), the analog of the mutation rate is the rate of retrotransposition. At any given site within the genome, this rate is very low, and each *Alu* locus represents a single retrotransposition. Thus, $\bar{S} = 1$ for data set II. The DNA sequences in data set III are short (~500 bp) and therefore represent a smaller mutational target than do the loci in data set I. Consequently, $\bar{S}$ is also low—about 3—in data set III. In short, the tendency for ancestral alleles to be African is greatest where the mutation rate per locus is highest—in data set I.

This does not prove that mutation rate underlies the observed differences, but it does suggest that we ought to take a closer look. We begin with simple theory.

Theory

In this section, we consider a single stylized gene genealogy (fig. 2), which relates 8 modern genes: 4 from population *A* (Africa) and 4 from population *E* (Eurasia). The Eurasian genes share a long branch back to the LCA, reflecting a bottleneck in their ancestry, and are therefore correlated with each other. By contrast, the African genes are uncorrelated. This captures, in exaggerated form, features that we think characterize the gene genealogies of Africa and Eurasia (Vigilant et al. 1991; Underhill et al. 2000).

Comparing Averages

If the number of loci is large, average allele frequencies will approximate expected allele frequencies. These are easy to calculate for the NAA.

Consider the event that gene E_3 in figure 3 is a copy of the NAA. Under the model of infinite sites, this event occurs if and only if there is no mutation on the path that separates the 2. This path is t generations long, so the probability that no mutation occurs is e^{-ut}, where u is the per-locus mutation rate. This is also the probability that E_3 is a copy of the NAA. Furthermore, the distance from every other gene to the LCA is also t. Consequently, every gene in the sample is a copy of the NAA with this same probability.

Within subpopulations, the expected frequency of the NAA equals the probability that a gene drawn at random is a copy of the NAA. But no matter which gene is chosen, this probability is e^{-ut}. Consequently, the expected frequency of the NAA is the same—e^{-ut}—within each subpopulation, no matter how these subpopulations are defined. Comparing averages thus tells us nothing about the location of the LCA.

This argument has assumed the model of infinite sites, but that is not essential. One need only assume that the same mutation rate prevails throughout the gene genealogy. The argument makes no assumption about sample size or the details of gene genealogy. Consequently, it holds for any sample size and for any gene genealogy. Population history affects samples of neutral genes only via its effect on gene genealogy. Because our argument is unaffected by gene genealogy, it is also unaffected by population history. It does, however, depend crucially on the absence of selection and ascertainment bias.

This theory predicts that the average frequency ($\bar{p}$) of the NAA will be the same in each subpopulation. With data set III (fig. 1), we see just that. Results for data sets I and II, on the other hand, are not as predicted. With those data sets, $\bar{p}$ is substantially greater in Africa than elsewhere. We suspect that these differences reflect ascertainment bias, an issue that we explore elsewhere (Rogers AR, Wooding S, Batzer MA, Jorde LB, unpublished observation).

Comparing Fractions

As explained above, comparing fractions involves asking what fraction of loci have ancestral alleles that reach highest frequency in each subpopulation. To understand how this method behaves, consider the special case illustrated in figure 4. This figure shows the same genealogy as before but with different branches highlighted. The Eurasian genes coalesce rapidly to a common ancestor. In the limiting case, this happens so fast that we can ignore mutations that fall in the dotted portion of the genealogy.

If the mutation rate per locus is high, the NAA will be absent from the sample, so inference must be based on the BAA—the modern allele that differs least from the NAA. If

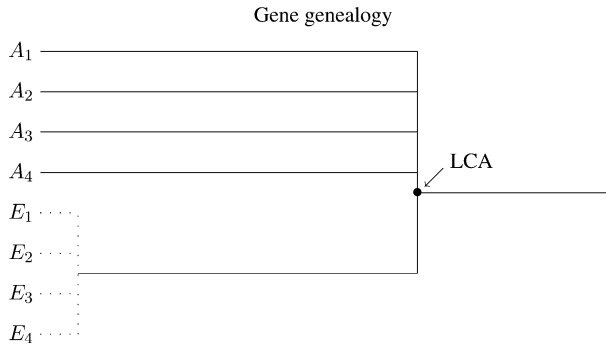

FIG. 4.—Another view of the same gene genealogy, with long branches highlighted.

the mutation rate is high enough, there will be no ties: each horizontal solid line in figure 4 will carry a different number of mutations. Consequently, each horizontal line has an equal chance of giving rise to the BAA. The BAA is more likely to fall in Africa (probability 4/5) than in Eurasia (probability 1/5). The method of comparing fractions would thus tell us that the ancestral allele is usually most common in Africa. These results support the approach of Satta and Takahata (Takahata et al. 2001; Satta and Takahata 2002, 2004), which involves comparing fractions.

On the other hand, the method of comparing averages would find no difference between the populations, for the expected frequency is the same—1/5—in both populations. (The expected frequency is $4/5 \times 1/4 + 1/5 \times 0 = 1/5$ in Africa and $4/5 \times 0 + 1/5 \times 1 = 1/5$ in Eurasia.) Thus, the 2 methods give very different answers in this example.

Consider now the case of low per-locus mutation rate. If this rate is very low, only one mutation will fall within the gene genealogy. As before, we assume that the dotted portion of the genealogy is short enough to neglect. The single mutation is thus equally likely to fall on each of the horizontal solid lines. With probability 4/5, its descendants are African. In that case, the frequency of the ancestral allele is less than 1 in Africa but equals 1 in Eurasia. Thus, most loci have Eurasian ancestral alleles. (The expected frequency, however, is still the same in both populations.)

In summary, the method of comparing averages tells us nothing about the location of the LCA, and the method of comparing fractions gives results that depend on the mutation rate. The ancestral allele will usually be most common in the ancestral population if the per-locus mutation rate is high, but in a descendant population if that rate is low. These results rely on extreme assumptions, so it is important to ask whether they hold more generally. To find out, we turn next to computer simulations.

Computer Simulations

Methods

We employ computer simulations that run backwards in time under the coalescent algorithm (Hudson 1990). Our version is based on Rogers (1997). In each iteration, the algorithm generates a gene genealogy with samples of 30 genes in each of several subpopulations. The algorithm divides history into a series of epochs within which all parameters are constant, as shown in table 2. Each row describes an epoch, and epochs are numbered backwards from the present. The length of an epoch is measured on a mutational time scale by $\tau = 2ut$, where u is the per-locus mutation rate and t is duration in generations. Thus, τ measures time in units of $1/2u$ generations. The size of subpopulation i is measured by $\theta_i = 4uN_i$, where N_i is the diploid subpopulation size during the epoch.

Table 2
Base Population History Parameters

Epoch	τ	θ_A	θ_E
0	1	1	0.01
1	∞	1	0.00

During the earliest epoch, 1, E was uninhabited and A had size $\theta_A = 1$. Thus, this model assumes that subpopulation A is ancestral. E becomes inhabited during the following epoch, 0. During this epoch, A retains its size but E is only 1/100 as large. There is no migration in either epoch.

Mutations obey the model of infinite sites (Kimura 1971). On a branch t generations in length, the number of mutations is Poisson with mean ut, where u is the mutation rate per locus. We ignore genealogies that receive no mutations. In the others, we count the mutations between each leaf (terminal node) and the LCA. Leaves that differ by 0 mutations are copies of the NAA. Those that differ by the minimum number of mutations are copies of the BAA. When the minimum equals 0, there is no difference between the 2 kinds of ancestral allele.

This model is not meant to be realistic. Instead, it generates gene genealogies similar to the one in figure 4. During epoch 0, all genes in E usually coalesce rapidly into a single line of descent, mimicking the long branch leading in figure 4 to the sample from E. A remains large in both epochs, so its coalescent events occur slowly. They do not all occur at once, as in the figure. Nonetheless, the simulated genealogies agree with the figure in that genes in A tend to be less correlated with each other than do genes in E. This is an exaggerated version of the pattern often seen in human data, where subpopulation A represents Africa and E Eurasia.

In table 2, the τ and θ parameters are proportional to the per-locus mutation rate. Thus, if the mutation rate were 10-fold higher or if we looked at 10 times as much DNA, these parameters would be 10 times as large. Consequently, we can investigate the behavior of the model under different mutation rates simply by rescaling τ and θ. In the simulations below, we examine mutation rates corresponding to values of S that span the range seen among nuclear loci in table 1 and figure 1. The low end of this range is $S = 1$, for *Alu*s, and the high end is $S = 41$, for locus HFE.

High per-locus mutation rate

For the case of a high mutation rate, we inflated the parameters in table 2 by a factor of 10. The results are shown in the upper 2 panels of figure 5. The panel on the left shows $\bar{p}$, the mean frequency of the ancestral allele across all simulated loci, and the panel on the right shows F, the fraction of these loci in which the ancestral allele is most common in the indicated subpopulation. The number of segregating sites averaged $\bar{S} = 50.2$, close to the highest

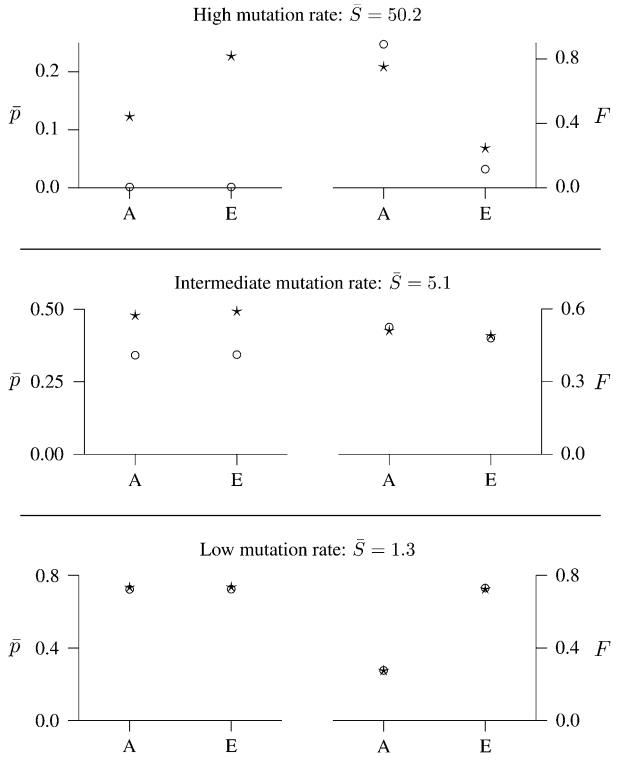

FIG. 5.—Simulation results. Symbols: A, Africa; E, Eurasia; °, NAA; and ★, BAA.

value of S seen among nuclear loci in the data. Thus, this simulation describes loci with mutation rates that are high but realistic.

Note first that for the NAA, the mean frequency ($\bar{p}$) is almost identical in subpopulations A and E, in agreement with the theoretical argument above. Although that argument makes no prediction about the mean frequency of the BAA, the simulation provides no support for the view that the BAA is on average more common in the ancestral population. In this simulation, A is the ancestral population, yet $\bar{p}$ is higher in population E.

The upper-right panel of figure 5 shows the values of F that we obtained from the same simulations. The ancestral allele is African at a large majority of simulated loci. This bias is strongest (89%) for the NAA but is also strong (75%) for the BAA. These simulation results are consistent with the theory, which predicted an African excess for the BAA but made no prediction for the NAA. They are also in rough agreement with data set I. The magnitude of the African excess for the NAA (89%) is about right, and this excess is apparent not only for the NAA but also for the BAA. However, the simulated $\bar{S}$ is high; we need to consider lower mutation rates.

Intermediate Per-Locus Mutation Rate

Reducing the mutation rate (and thus each τ and θ value) by an order of magnitude returns us to table 2. The middle pair of panels in figure 5 shows the results of simulations under this assumption. The mean number of segregating sites drops to $\bar{S} = 5.1$ (near the lower end of the values in table 1), and the difference between populations A and E disappears. It does not matter whether we look at $\bar{p}$ or F or whether we use the narrow or broad definition of the ancestral allele: the 2 populations have nearly the same value. With these mutation rates, ancestral alleles are devoid of information about the location of the LCA.

Low Per-Locus Mutation Rate

We argued above that under low mutation rates, ancestral alleles should be most common not in the ancestral population but in a descendant one. To evaluate this claim, we set each τ and θ parameter an order of magnitude lower than its value in table 2. Among loci simulated under this assumption, the number of segregating sites averaged $\bar{S} = 1.3$ near the low end of the range of values represented in our data.

The simulation results appear in the bottom pair of panels in figure 5 and are just as predicted. The value of F is higher in Eurasia rather than Africa at a majority (~73%) of simulated loci. Thus, in low-mutation systems, the conventional wisdom about ancestral alleles is precisely backwards. In spite of these differences in F, the populations have essentially identical values of $\bar{p}$, also in agreement with theory.

These results suggest a testable hypothesis. If modern humans originated in Africa, then ancestral alleles should usually be Eurasian at loci with low per-locus mutation rates and consequently few segregating sites (low S). They should usually be African at loci with high S. There is some support for this hypothesis. Satta and Takahata (2004) list a single locus at which the ancestral allele is most common outside Africa, and this is a locus with only a single segregating site. Our own table 1, however, excludes this locus because of its small sample size and adds several others. In our table, there is no obvious relationship between S and the tendency of ancestral alleles to be African.

Discussion and Conclusions

Frequencies of ancestral alleles are compared either by comparing fractions or by comparing averages. Each method can be implemented using either of 2 definitions (narrow and broad) of the ancestral allele. Until now, all such studies have agreed that human ancestral alleles are more common in Africa than in Europe or Asia (Mountain et al. 1992; Mountain and Cavalli-Sforza 1994; Takahata et al. 2001; Templeton 2002; Satta and Takahata 2002, 2004; Watkins et al. 2003). The present study reveals a more complex pattern. There is a tendency for ancestral alleles to reach their highest frequency in Africa, but this tendency is not consistent across data sets. It is pronounced in data set I but weak in data sets II and III. Ancestral alleles are apparently affected by some factor that differs among data sets. Because all data sets share the same population history, it is dangerous to interpret these data solely in terms of history.

The pattern in these data parallels that of mean mutation rate, which is higher in the first data set than in the latter 2. This led us to investigate the effect of per-locus mutation rate on the distribution of ancestral alleles. In our model, one subpopulation (called ancestral) has never experienced

a bottleneck, whereas the other (called descendant) passed through a bottleneck at the time the 2 populations separated. For convenience, we refer to the ancestral population as "African" and to the descendant one as "Eurasian."

We proceed from theoretical arguments with extreme assumptions to simulations in which these assumptions are relaxed. The theory is not offered as a model of nature. Instead, it provides a context within which it is easy to see how things work. The extreme assumptions of the theory are relaxed in the simulations, yet the qualitative results are the same. Thus, the intuition provided by the theory appears relevant in a broader context.

The method of comparing fractions calculates the fraction (F) of loci with African ancestral alleles. This fraction is sensitive to the per-locus mutation rate, both in theory and simulation. When the mutation rate is high, the African fraction is large, in agreement with conventional wisdom. Under low mutation rates, however, the conventional wisdom is incorrect: only a minority of loci have African ancestral alleles. This is consistent with the pattern in our data, where the fraction of loci with African ancestral alleles is highest in the data set with high mutation rate. On the other hand, there is no apparent relationship between mutation rate (as reflected by S) and the bias toward African ancestral alleles *within* data set I. This discrepancy suggests a role for some factor in addition to mutation rate.

These results do not apply to the mean ($\bar{p}$) of ancestral allele frequencies within subpopulations. For the NAA, $\bar{p}$ is affected neither by mutation rate nor by the history of population size and migration rate. In the absence of selection and ascertainment bias, there should be no difference in $\bar{p}$ between subpopulations. This conclusion rests on a model that is quite general, making no assumptions about population history and only weak assumptions about the mutational process.

It explains a pattern that Mountain and Cavalli-Sforza (1994, fig. 6) discovered in simulation results. In the absence of ascertainment bias, $\bar{p}$ was the same in each subpopulation. We saw the same pattern in real data with set III. A different pattern, however, appears in data sets I and II. There, $\bar{p}$ was much higher in Africa, contrary to prediction. This discrepancy provides a second reason to suspect that data sets I and II have been affected by ascertainment bias or selection.

BAA and NAA behave similarly except when the per-locus mutation rate is high. Then, the $\bar{p}$ of the BAA does respond to history, but not as is usually assumed. It tends to be highest in the descendant population rather than the ancestral one.

The 2 discrepancies noted above suggest that some additional factor is at work. In a separate publication (Rogers AR, Wooding S, Batzer MA, Jorde LB, unpublished observation), we investigate the role of ascertainment bias in the selection of loci. Yet even without that additional analysis, the present study has shown that ancestral alleles are sensitive to a factor—the mutation rate—that has nothing to do with population history. This does not prove that ancestral alleles are uninformative. Indeed, we have shown that ancestral alleles are usually most common in the ancestral population if the per-locus mutation rate is high. It does, however, show that these alleles respond to other influences, which must be understood before inferences about history can be reliable.

Acknowledgments

We are grateful for comments from Elizabeth Cashdan, Vinayak Eswaran, Joseph Felsenstein, Henry Harpending, Naoyuki Takahata, and David Witherspoon. This work was supported by National Institutes of Health RO1 GM59290 (L.B.J. and M.A.B.), National Science Foundation BCS-0218338 (M.A.B.), BCS-0218370 (L.B.J.), and EPS-0346411 (M.A.B.) as well as the State of Louisiana Board of Regents Support Fund (M.A.B.).

Literature Cited

Balciuniene J, Syvanen AC, McLeod HL, Pettersson U, Jazin EE. 2001. The geographic distribution of monoamine oxidase haplotypes supports a bottleneck during the dispersion of modern humans from Africa. J Mol Evol. 52:157–163.

Bamshad MJ, Mummidi S, Gonzalez E, et al. (11 co-authors). 2002. A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci USA. 99:10539–10544.

Breguet G, Butler R, Butler-Brunner E, Sanchez-Mazas A. 1990. A worldwide population study of the Ag-system haplotypes, a genetic polymorphism of human low-density lipoprotein. Am J Hum Genet. 46:502–517.

Chen F-C, Li W-H. 2001. Genomic divergences between human and other hominoids and the effective population size of the common ancestor of human and chimpanzee. Am J Hum Genet. 68:444–456.

Excoffier L. 2002. Human demographic history: refining the recent African origin model. Curr Opin Genet Dev. 12:675–682.

Felsenstein J. 2004. PHYLIP (phylogeny inference package). Version 3.6. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Garrigan D, Mobasher Z, Severson T, Wilder JA, Hammer MF. 2005. Evidence for archaic Asian ancestry on the human X chromosome. Mol Biol Evol. 22:189–192.

Hallman DM, Boerwinkle E, Saha N, Sandholzer C, Menzel HJ, Csár A, Utermann G. 1991. The apolipoprotein E polymorphism: a comparison of allele frequencies and effects in nine populations. Am J Hum Genet. 49:338–349.

Hammer MF. 2004. Heterogeneous patterns of variation among multiple human X-linked loci: the possible role of diversity-reducing selection in non-Africans. Genetics. 167:1841–1853.

Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL. 2001. Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol. 18:1189–1203.

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet. 60:722–789.

Harding RM, Healy E, Ray AJ, et al. (11 co-authors). 2000. Evidence for variable selective pressures ad MC1R. Am J Hum Genet. 66:1351–1361.

Harris EE, Hey J. 1999. X chromosome evidence for ancient human histories. Proc Natl Acad Sci USA. 96:3320–3324.

Harris EE, Hey J. 2001. Human populations show reduced DNA sequence variation at the factor IX locus. Hum Biol. 11:774–778.

Hudson RR. 1990. Gene genealogies and the coalescent process. In: Futuyma D, Antonovics J, editors. Gene genealogies and

the coalescent process, Vol. 7. Oxford: Oxford University Press. p. 1–44.
Jaruzelska J, Zietkiewicz E, Batzer M, Cole DEC, Moisan J-P, Scozzari R, Tavaré S, Labuda D. 1999. Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: analysis of the haplotype structure and genealogy. Genetics. 152:1091–1101.
Kaessmann H, Heißig F, von Haeseler A, Pääbo S. 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat Genet. 22:78–81.
Kimura M. 1971. Theoretical foundation of population genetics at the molecular level. Theor Popul Biol. 2:174–208.
Martinez-Arias R, Calafell F, Mateu E, Comas D, Andres A, Bertranpetit J. 2001. Sequence variability of a human pseudogene. Genome Res. 11:1071–1085.
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.
Mountain JL, Lin A, Bowcock A, Cavalli-Sforza LL. 1992. Evolution of modern humans: evidence from nuclear DNA polymorphism. Philos Trans R Soc Lond Ser B Biol Sci. 337:159–165.
Nachman MW, Crowell SL. 2000. Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, *Dmd*, in humans. Genetics. 155:1855–1864.
Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. Numerical recipes in C++: the art of scientific computing. 2nd ed. New York: Cambridge University Press.
Rana BK, Hewett-Emmett D, Jin L, et al. (12 co-authors). 1999. High polymorphism at the human melanocortin 1 receptor locus. Genetics. 151:1547–1557.
Rapacz J, Chen L, Butler-Brunner E, Wu MJ, Hasler-Rapacz JO, Butler R, Schumaker VN. 1991. Identification of the ancestral haplotype for apolipoprotein B suggests an African origin of Homo sapiens sapiens and traces their subsequent migration to Europe and the Pacific. Proc Natl Acad Sci USA. 88:1403–1406.
Rogers AR. 1997. Population structure and modern human origins. In: Donnelly PJ, Tavaré S, editors. Population structure and modern human origins. New York: Springer-Verlag. p. 55–79.
Satta Y, Takahata N. 2002. Out of Africa with regional interbreeding? Modern human origins. Bioessays. 24:871–875.
Satta Y, Takahata N. 2004. The distribution of the ancestral haplotype in finite stepping-stone models with population expansion. Mol Ecol. 13:877–886.
Stephens M, Smith N, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 68:978–989.
Takahata N, Lee SH, Satta Y. 2001. Testing multiregionality of modern human origins. Mol Biol Evol. 18:172–183.
Templeton A. 2002. Out of Africa again and again. Nature. 416:45–51.
Toomajian C, Kreitman M. 2002. Sequence variation and haplotype structure at the human HFE locus. Genetics. 161:1609–1623.
Underhill PA, Shen P, Lin AA, et al. (21 co-authors). 2000. Y chromosome sequence variation and the history of human populations. Nat Genet. 26:358–361.
Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. 1991. African populations and the evolution of human mitochondrian DNA. Science. 253:1503–1507.
Watkins WS, Rogers AR, Ostler CT, et al. (14 co-authors). 2003. Genetic variation in world populations: inferences from 100 *Alu* insertion polymorphisms. Genome Res. 13:1607–1618.
Wooding SP, Watkins WS, Bamshad MJ, Dunn DM, Weiss RB, Jorde LB. 2002. DNA sequence variation in a 3.7-kb noncoding sequence 5′ of the CYP1A2 gene: implications for human population history and natural selection. Am J Hum Genet. 71:528–542.
Yu N, Chen F-C, Ota S, Jorde LB, Pamilo P, Patthy L, Ramsay M, Jenkins T, Shyue S-K, Li W-H. 2002. Larger genetic differences within Africans than between Africans and Eurasians. Genetics. 161:269–274.
Zekraoui L, Lagarde JP, Raisonnier A, Gerard N, Aouizerate A, Lucotte G. 1997. High frequency of the apolipoprotein E*4 allele in African pygmies and most of the African populations in sub-Saharan Africa. Hum Biol. 69:575–581.
Zhang J, Rosenberg HF. 2000. Sequence variation at two eosinophil-associated ribonuclease loci in humans. Genetics. 156: 1949–1958.

Yoko Satta, Associate Editor

Accepted January 19, 2007