DOI: 10.1534/genetics.107.080432

Testing for Archaic Hominin Admixture on the X Chromosome: Model Likelihoods for the Modern Human *RRM2P4* Region From Summaries of Genealogical Topology Under the Structured Coalescent

Murray P. Cox,* Fernando L. Mendez,† Tatiana M. Karafet,* Maya Metni Pilkington,‡ Sarah B. Kingan,* Giovanni Destro-Bisol,§ Beverly I. Strassmann and Michael F. Hammer†,‡,**,[1]**

**ARL Division of Biotechnology, †Department of Ecology and Evolutionary Biology and ‡Department of Anthropology, University of Arizona, Tucson, Arizona 85721, §Department of Animal and Human Biology, University of Rome "La Sapienza," and Istituto Italiano di Antropologia, 00185 Rome, Italy and **Department of Anthropology, University of Michigan, Ann Arbor, Michigan 48109*

Manuscript received August 13, 2007
Accepted for publication November 11, 2007

ABSTRACT

A 2.4-kb stretch within the *RRM2P4* region of the X chromosome, previously sequenced in a sample of 41 globally distributed humans, displayed both an ancient time to the most recent common ancestor (*e.g.*, a TMRCA of ~2 million years) and a basal clade composed entirely of Asian sequences. This pattern was interpreted to reflect a history of introgressive hybridization from archaic hominins (most likely Asian *Homo erectus*) into the anatomically modern human genome. Here, we address this hypothesis by resequencing the 2.4-kb *RRM2P4* region in 131 African and 122 non-African individuals and by extending the length of sequence in a window of 16.5 kb encompassing the *RRM2P4* pseudogene in a subset of 90 individuals. We find that both the ancient TMRCA and the skew in non-African representation in one of the basal clades are essentially limited to the central 2.4-kb region. We define a new summary statistic called the minimum clade proportion (p_{mc}), which quantifies the proportion of individuals from a specified geographic region in each of the two basal clades of a binary gene tree, and then employ coalescent simulations to assess the likelihood of the observed central *RRM2P4* genealogy under two alternative views of human evolutionary history: recent African replacement (RAR) and archaic admixture (AA). A molecular-clock-based TMRCA estimate of 2.33 million years is a statistical outlier under the RAR model; however, the large variance associated with this estimate makes it difficult to distinguish the predictions of the human origins models tested here. The p_{mc} summary statistic, which has improved power with larger samples of chromosomes, yields values that are significantly unlikely under the RAR model and fit expectations better under a range of archaic admixture scenarios.

FOSSIL, archaeological, and genetic data all lend support to the hypothesis that *Homo sapiens* originated in Africa (McBrearty and Brooks 2000; McDougall *et al.* 2005; Garrigan and Hammer 2006). With the acceptance of the role of Africa in our species' origin, there is now increasing interest in the question of how the ancestral population that gave rise to anatomically modern humans (AMH) was structured. Did AMH emerge from a single, isolated African deme or from a subdivided ancestral population with gene flow among subpopulations? A related question is whether the expanding AMH population completely replaced or interbred with then contemporaneous archaic populations such as Neanderthals and *H. erectus* (Eswaran 2002; Templeton 2002; Garrigan *et al.* 2005a,b; Plagnol and Wall 2006; Wall and Hammer 2006). Early studies of nonrecombining regions such as mtDNA and the Y chromosome were consistent with the hypothesis of a single origin followed by complete replacement, sometimes referred to as the recent African replacement (RAR) model. While many of the more recently published DNA sequencing studies of X-linked and autosomal loci are also concordant with this RAR model, a growing number are not (Evans *et al.* 2006; Garrigan and Hammer 2006).

Garrigan *et al.* (2005b) published one of the first studies to posit recent admixture between AMH and an archaic human population. A resequencing study of 2.4 kb of the ribonucleotide reductase M2 pseudogene 4 (*RRM2P4*) in a sample of 41 globally diverse humans identified an unusual pattern of nucleotide polymorphism compared with most of the human genome. The reconstructed gene tree revealed two clades of allelic sequences that were estimated to have diverged ~2

[1]*Corresponding author:* ARL Division of Biotechnology, Life Sciences South, University of Arizona, Tucson, AZ 85721.
E-mail: mfh@u.arizona.edu

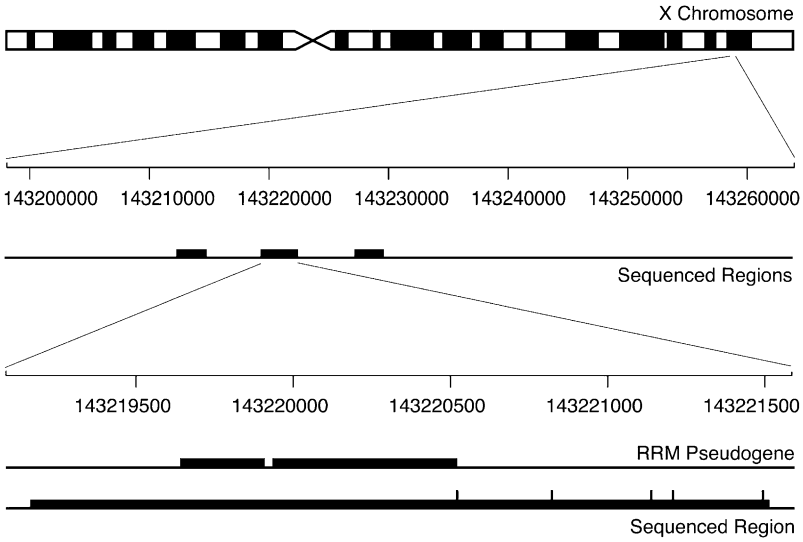

FIGURE 1.—Location of the *RRM2P4* locus on the X chromosome (top bar), positions of the three sequenced regions (middle bar), and placement of the *RRM2P4* pseudogene fragments relative to the *RRM2P4* central region (bottom bars). Vertical bars in the central region indicate SNPs defining the two basal clades. All coordinates correspond to the human genome UCSC March 2006 build.

million years ago (MYA). One clade with very little sequence variation was specific to Asians, while the other more diverse clade of *RRM2P4* sequences resembled a pattern typical of human variation and was globally distributed. By genotyping a diagnostic SNP in a large sample of humans, the divergent "Asian clade" was shown to be frequent in Southeast Asians and nearly absent in sub-Saharan Africans. The greater genealogical depth in Asia led to the hypothesis that *RRM2P4* is a genomic remnant of introgressive hybridization from an Asian archaic population (*H. erectus*) into AMH groups expanding from Africa. However, despite the increasing availability of genomic data, the uniqueness of such deep times to the most recent common ancestor (TMRCA) and of gene trees with non-African basal clades remains unclear. Importantly, we still do not understand the likelihood of such genealogies under either the RAR model or models involving admixture with archaic humans.

Garrigan *et al.*'s original *RRM2P4* study was limited by a paucity of sequence data and primarily qualitative analyses. Here, we extend earlier research by sequencing the 2.4-kb *RRM2P4* region in 131 African and 122 non-African individuals. We also determine the extent of sequence that exhibits the unusual pattern of polymorphism by selectively resequencing three DNA fragments totaling 5.6 kb within a 16.5-kb window flanking *RRM2P4*. We infer the likelihood of the observed *RRM2P4* genealogy using a suite of summary statistics and Monte Carlo coalescent simulations under the RAR model and a range of archaic admixture (AA) models (NORDBORG 2000; PLAGNOL and WALL 2006). The RAR models are parameterized by approximate Bayesian computation (ABC) conditioned on resequence data from an additional data set of 19 neutral, unlinked X-chromosomal loci. This resequence data set represents 12 Mb from 19 regions of the X chromosome, which are unlinked from genes (and each other) by medium to high rates of recombination ($\geq$1.0 cM/Mb). These loci, which represent selectively neutral X-chromosomal diversity, will be published elsewhere. However, using these data, we show here that simple RAR models often produce TMRCA values that are similar to that of *RRM2P4*, while genealogies with skewing of basal clade membership toward non-Africans remain statistical outliers.

SUBJECTS AND METHODS

Regions sequenced: Resequence data for the *RRM2P4* locus were generated within a "trio" design (Figure 1), which is an economical approach to jointly ascertain detailed polymorphism patterns and larger-scale linkage-disequilibrium profiles (GARRIGAN *et al.* 2005b). We sequenced three genomic segments of 1725 bp [University of California, Santa Cruz (UCSC) March 2006 genome coordinates 143,212,138–143,213,863], 2341 bp (coordinates 143,219,154–143,221,495), and 1601 bp (coordinates 143,227,057–143,228,658), which were separated by unsequenced regions of 5291 and 5562 bp, respectively. Our central subregion is the same as that described by GARRIGAN *et al.* (2005b).

Sampling: The *RRM2P4* central region was sequenced in 131 African and 122 non-African individuals (panel A). These samples were chosen without prior information about *RRM2P4* lineage status. DNA samples representing Mandenka from Senegal ($n = 16$), Biaka Pygmies from the Central African Republic ($n = 16$), Khoisan from Namibia ($n = 9$), French Basque ($n = 16$), Han Chinese ($n = 16$), and Nasioi from Bougainville ($n = 16$) were purchased from the Centre d'Etude du Polymorphisme Humain (CANN *et al.* 2002). Samples of Baka Pygmies from Cameroon ($n = 23$) were provided by Giovanni Destro-Bisol and the Dogon from Mali ($n = 32$) were provided by Beverly Strassmann. Samples from the Dinka of southern Sudan ($n = 21$)

were collected in Tucson, Arizona, with informed consent. Samples from three Siberian populations, the Selkups ($n = 32$), Forest Nentsi ($n = 28$), and Tundra Nentsi ($n = 3$), were described previously (Karafet *et al.* 2002). Non-population-based samples ($n = 25$) from the Y chromosome consortium cell lines (Y Chromosome Consortium 2002) were also included in this panel. Resequence data for the full trio were generated in a second panel of 42 African and 48 non-African individuals (panel B) from three African groups (Khoisan, Mandenka, and Biaka) and three non-African groups (French Basque, Han, and Nasioi) (samples as described above). All sampling protocols were approved by the Human Subjects Committee at the University of Arizona and by the institutions of all collaborators who provided DNA samples.

Recombination analysis and genetree dating: Rates of linkage disequilibrium across the sequence were determined using LDhat (McVean and Spencer 2006). The *RRM2P4* central region shows only limited signs of recombination, and a most parsimonious tree was reconstructed by breaking low-frequency reticulations. The TMRCA of the tree and the age of its polymorphisms were estimated with Genetree (Griffiths 2007). Genetree employs a full maximum-likelihood method that is based on the standard coalescent (Kingman 1982) and assumes an infinite-sites mutational model. Likelihood surfaces for the population mutation rate, θ, and the population growth rate, β, were generated under a panmictic single-deme model and an island model of population structure. TMRCA values were inferred from maximum-likelihood parameterizations under both models.

Demographic models: Summaries of the *RRM2P4* central region were compared with values obtained from simulations under a structured coalescent (Nordborg 1997; Hudson 2002) to determine the likelihood of the observed genealogy. We employed a framework for human demography similar to that developed by Plagnol and Wall (2006), but modified (see below) to yield the number of segregating sites, *S*, consistent with an independent data set of 19 X chromosome loci sequenced in the same individuals (panel B). We present results from two demographic models: a two-deme RAR model and an AA scenario similar to Nordborg's (2000) isolation and admixture model (Figure 2). Alternative RAR models with varying levels of recent population subdivision (*e.g.*, one-deme and six-deme models) did not differ significantly from the two-deme RAR model reported here (our unpublished data). These models are not meant to represent the true history of human populations, but they do let us explore the effects of expansion and replacement *vs.* archaic admixture on patterns of genomic variation. Coalescent dates (scaled by $3N_e$ generations) were translated to chronological time using a 28-year mean intergeneration interval (Fenner 2005). The use of a lower estimate of the intergeneration interval (*e.g.*, a 25-year mean intergeneration interval) does not alter our conclusions (our unpublished data).

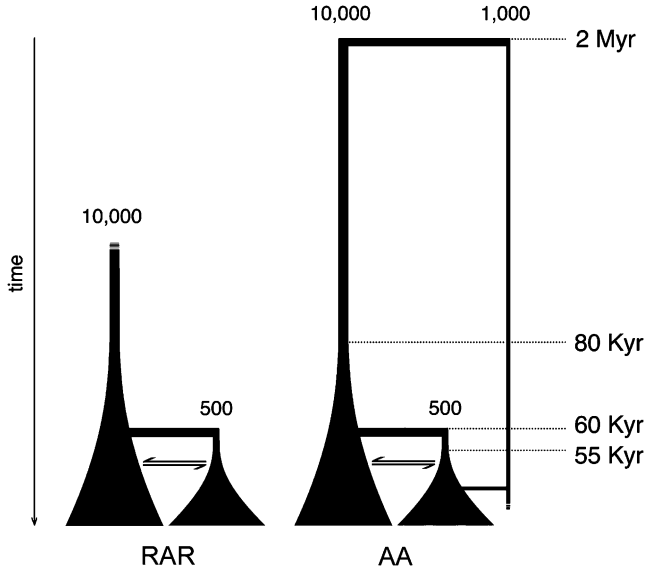

Figure 2.—Schematics of two demographic models. The recent African replacement (RAR) model grows exponentially from a 10^4 effective population size at 80 KYA, and 500 individuals found the non-African population at 60 KYA, experience a constant-sized bottleneck for 5 KYA, and grow exponentially from 55 KYA. The ancient admixture (AA) model includes introgression from an ancestral hominin source with constant population size of 10^3; the population leading to that of anatomically modern humans and ancestral hominins diverged 2 MYA. Modern effective sizes and the intercontinental migration rate were inferred by approximate Bayesian computation.

The RAR model depicts two panmictic, exponentially growing Wright–Fisher demes representing African and non-African populations. Growth begins in the African deme 80,000 years ago from a single ancestral population ($N_e = 10^4$) and continues until it reaches its current effective size. A small group ($N_e = 500$) splits from the African deme 60,000 years ago to form the non-African deme. This subgroup experiences a bottleneck for 5000 years before expanding exponentially to its current effective size. Modern effective sizes and the intercontinental migration rate were parameterized by approximate Bayesian computation (see below).

The AA model incorporates instantaneous admixture from an ancestral hominin source into the RAR model described above. An ancestral hominin population with constant effective population size of 10^3 splits from the ancestors of modern humans at 2 MYA. We considered a range of admixture rates (0–5%) from ancestral hominins to modern humans in Asia, coupled with a series of admixture times (10–55 KY before present). We could not infer an optimized AA model because the paucity of candidate loci is not sufficient both for model training and for subsequent statistical testing.

Model fitting via approximate Bayesian computation: We parameterized the RAR model to reflect what is known about deep human demography and, consequently, to produce simulated data sets that mimic real genomic data sets. Parameter inference rapidly becomes computationally intractable at high-dimensional state spaces, such as those associated with complex demographic models. Therefore, we fixed some demographic parameters that have been inferred elsewhere, *e.g.*, the time of onset of population growth and non-African bottleneck size (Plagnol and Wall 2006, Figure 2). However, a lower-dimensional state space of modern population size, N_0, and the intercontinental migration rate per generation, m, were inferred explicitly by ABC (Beaumont *et al.* 2002).

Essentially, we generated 10^4 coalescent simulations for each of 10^5 different sets of demographic parameters, $\Theta = \{N_0, m\}$, that were drawn randomly from two uniform distributions, $N_0 \in U[10^4, 10^5]$ and $m \in U[10^{-11}, 10^{-8}]$. Each set of coalescent simulations was compared with resequence data from 19 noncoding regions on the X chromosome (*i.e.*, panel B, 42 African and 48 non-African individuals; data to be published elsewhere). We estimated θ ($= 3N_0\mu$) for each demographic parameter set using the average mutation rate of these 19 loci, 8.3×10^{-10}/bp/year (range 4.8×10^{-10}–1.6×10^{-9}), inferred from sequence divergence (assuming a human/chimpanzee divergence time of 6 MYA). Furthermore, we chose to condition our ABC inference on the mean number of segregating sites, S, that was observed across the additional data set of 19 X chromosome loci ($\bar{S} = 26$). This statistic was chosen because it varies primarily with the mutation rate (a known parameter) and tree depth (a parameter we wished to infer). We selected the 0.01% of random demographic parameter sets, Θ, whose coalescent data sets produced an average number of segregating sites closest to the observed value ($\bar{S} = 26$). Mean values for N_0 and m were drawn from this subset of demographic parameter sets. These values represent best-fit estimates for the modern effective size, N_0, and migration rate, m, of the real-world demography underlying our observed 19 X chromosome neutral-locus genomic data set.

Test statistics: We calculated the approximate likelihoods of two summary statistics: the TMRCA and the minimum proportion of Africans in one of the two basal clades. Values were inferred from 10^5 replicates under all demographic models. We note that these tests are conservative, because both summaries are determined directly from coalescent genealogies. Resolution of the underlying genealogy is constrained by S for real data sets (*cf.* Nordborg 2000). The distribution of TMRCA values was extracted from the output of Hudson's (2002) ms using custom software (code available on request).

Here, we also define a new summary statistic, the minimum clade proportion (p_{mc}), which characterizes the minimum quantity of individuals from a specified geographic region or ethnic group (such as Africans) in each of the two basal clades of a binary gene tree (Figure 3). (This statistic is described fully in the appendix.) The p_{mc} statistic has an intuitive interpretation. Consider a binary tree of African and non-African sequences (Figure 3); clades (C_1, C_2) can be defined as the two basal branches that diverge from the coalescent of all sampled individuals. Here, the number of African chromosome copies is ($k_1 = 1$, $k_2 = 6$) and the total number of chromosome copies is ($n_1 = 4$, $n_2 = 10$). From Equation A1, the minimum clade proportion is the lesser of ($k_1/n_1 = \frac{1}{4}$, $k_2/n_2 = \frac{6}{10}$); here, $p_{\mathrm{mc}} = \frac{1}{4}$. The p_{mc} statistic was calculated directly from coalescent genealogies using custom software (code available on request).

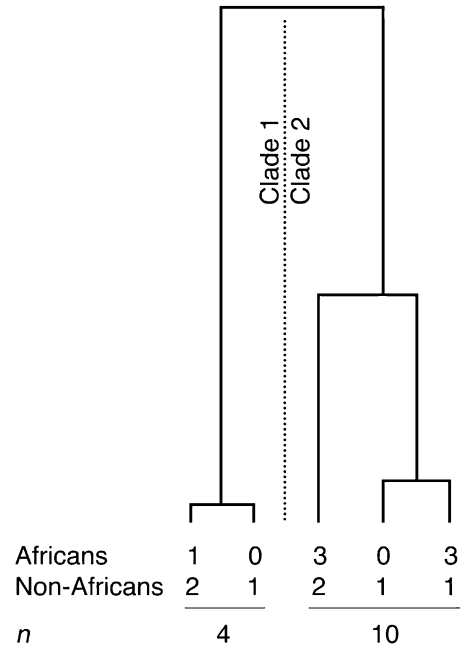

Figure 3.—Schematic of the minimum clade-proportion statistic, p_{mc}. See text for details.

RESULTS

Patterns of DNA sequence variation within and around *RRM2P4*: We sequenced 2.4 kb of the central *RRM2P4* segment, which encompasses the processed pseudogene (Figure 1), in 131 African and 122 non-African individuals. A total of 22 segregating sites were identified, producing 23 unique haplotypes (supplemental Table 1 at http://www.genetics.org/supplemental/). Similar to the previous study of Garrigan *et al.* (2005b), levels of nucleotide diversity are statistically higher in non-Africans ($\theta_\pi = 0.136$) than in Africans ($\theta_\pi = 0.0768$; $P \ll 0.001$; Table 1). Also, the central subregion still shows minimal evidence of recombination despite a sixfold increase in the size of the data set (supplemental Figure 1 at http://www.genetics.org/supplemental/).

TABLE 1

Estimates of summary statistics for the *RRM2P4* central and flanking regions

Region	Deme	n	Length	S	θ_w/bp (%)	Lower 5% quantile	θ_π/bp (%)	95% confidence interval
5′	Global	83	1727	5	0.058	—	0.068	(0.052, 0.081)
	African	37	1727	5	0.069	—	0.084	(0.059, 0.098)
	Non-African	46	1727	3	0.040	—	0.056	(0.034, 0.073)
Central	Global	253	2320	22	0.155	0.120	0.110	(0.093, 0.126)
	African	131	2320	16	0.126	0.071	0.077	(0.066, 0.087)
	Non-African	122	2320	15	0.120	0.104	0.136	(0.105, 0.163)
3′	Global	83	1612	7	0.088	—	0.061	(0.037, 0.081)
	African	37	1612	4	0.060	—	0.081	(0.044, 0.108)
	Non-African	46	1612	6	0.085	—	0.041	(0.014, 0.067)

Statistical confidence was determined from 10^4 bootstrap replicates. SNP diversity, θ_W, is significantly lower in the flanking regions than in the central region ($\alpha < 0.05$). African θ_π-values do not differ significantly among regions ($\alpha > 0.05$), but non-African θ_π- and θ_W-values are significantly higher in the central region. Within sequenced regions, non-Africans have significantly higher θ_π relative to Africans in the central region only ($P \ll 0.001$, underlined). This effect is not observed in the 5′ ($P = 0.966$) or 3′ regions ($P = 0.957$).

As before, this low level of recombination permitted partial reconstruction of a single nonreticulating gene tree for the central region (Figure 4). This genealogy has the same two unusual characteristics described by Garrigan *et al.* (2005b): a deep TMRCA and a basal clade composed almost entirely of Asian sequences. For example, of the 253 individuals resequenced for the central region, 21 are in the leftmost divergent clade in Figure 4 ("clade A"), and 20 of these are from East Asia and Oceania. A single Dogon from Mali was the only African member of this clade. This haplotype carries a SNP (1639) that occurs in both basal clades. An origin of this haplotype through recombination rather than homoplasy is more likely given the low mutation rate of the *RRM2P4* locus together with its moderate rate of recombination. Because the parental clade A form was found only in Asia and this haplotype is shared with a Melanesian, we suggest that this recombinant lineage may have originated in Asia and migrated recently to Africa.

We used both molecular-clock and coalescent approaches to estimate the TMCRA of the *RRM2P4* central subregion gene tree. Outgroup comparisons reveal an average of 22 nucleotide substitutions between all human and chimpanzee central *RRM2P4* region sequences. Given an average of 8.53 nucleotide differences observed between the two human *RRM2P4* lineages (*i.e.*, the average number of mutations between sequences across the base of the human gene tree), we estimate that the two deepest human clades diverged ~2.33 MYA (assuming a 6-MYA human–chimpanzee divergence time). We also inferred the TMRCA of the *RRM2P4* central subregion using a full maximum-likelihood method under both a panmictic and an island model. These models yielded TMRCA values of 1.24 and 2.88 MYA, which bracket the molecular-clock date (data not shown).

To determine the length of sequence within and around the *RRM2P4* locus that shows the unusual genealogical features, we sequenced two additional fragments of 1725 and 1601 bp that flank the central region (Figure 1) in a subset of African and non-African individuals (panel B). Levels of nucleotide diversity in African *vs.* non-African populations were more similar to average patterns for the genome, with African values ($\theta_\pi = 0.0836$ and 0.0805) greater than non-African values ($\theta_\pi = 0.0556$ and 0.0412) in the 5′- and 3′-flanking regions, respectively (Table 1). Recombination rates are low (0.43 cM/Mb) for the 5′ and central *RRM2P4* subregions, but are substantially elevated (16 cM/Mb) between the central and 3′ subregions (supplemental Figure 1 at http://www.genetics.org/supplemental/). This recombination hotspot effectively unlinks the central and 3′ subregions, which therefore have largely independent evolutionary histories. A network representing sequences of the entire 16.4-kb region illustrates the decoupling of these two genomic regions (supplemental Figure 2 at http://www.genetics.org/supplemental/). The unusual genealogy is less apparent in the 5′ fragment despite some linkage disequilibrium with the central region. (Recombination rates in these regions are about one-third the X chromosome average; supplemental Figure 1.) However, only five segregating sites were identified in the 5′-flanking region, a reduction of 63% over the central region diversity (Table 1), and no derived polymorphisms were identified in the 5′ region on chromosomes carrying central region clade A haplotypes (supplemental Table 2 at http://www.genetics.org/supplemental/). Because the unusual pattern of polymorphism is most apparent in the central region, we focus further analyses solely on this portion of the sequenced region.

RAR model parameters: The RAR model was parameterized by ABC. Because the model cannot be parameterized

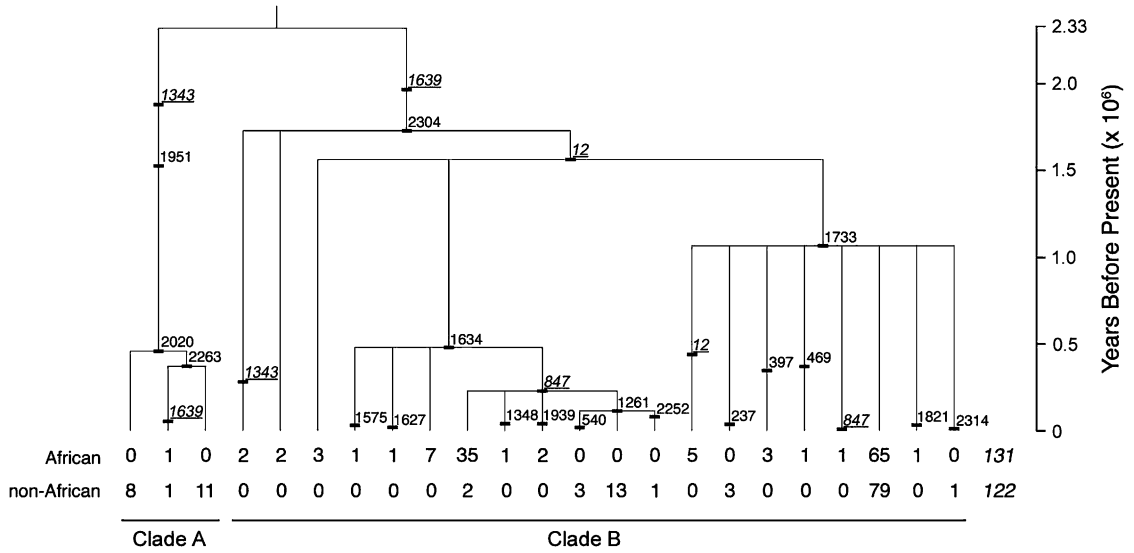

FIGURE 4.—Time-scaled gene tree of the *RRM2P4* central region. Polymorphisms are labeled with the nomenclature of GARRIGAN *et al.* (2005b) and are equivalent to that publication's Figure 1. Italicized and underlined polymorphisms indicate probable reticulation points. The proportions of African and non-African individuals carrying each lineage are indicated beneath the tree. Note the overrepresentation of non-Africans in the leftmost basal clade ("clade A"). Molecular-clock estimates of the TMRCA and mutation ages are shown in millions of years on the vertical axis.

on the test locus (here, *RRM2P4*), we conditioned the model on a separate training data set of 19 independent X chromosome noncoding regions. These resequence data are too extensive for detailed description here and are the subject of a separate publication. However, basic summaries of these 19 loci relevant to the current analysis are presented in supplemental Table 3 at http://www.genetics.org/supplemental/. The two-deme RAR model was parameterized by generating 10^5 random N_0 and m values (see full description in SUBJECTS AND METHODS) and accepting the 0.01% that best matched the number of segregating sites observed in 19 independent X chromosome noncoding loci (supplemental Figure 3 at http://www.genetics.org/supplemental/). The optimal parameters were inferred as a modern effective size, N_0, of 12,300 (range 12,000–12,500) and an intercontinental migration rate per generation, m, of 3.62×10^{-9} (range 6.75×10^{-10}–8.24×10^{-9}). Using parameter values at the extremes of these ranges had little effect on the following statistical analyses (our unpublished data). Importantly, the parameterized RAR model produces simulated data sets that yield summaries (such as growth rates and effective sizes) that are consistent with other demographic inferences (*e.g.*, from Genetree). Although there is strictly no way to assess the true history of our samples, this best-fit RAR model is a reasonable first approximation for human demographic history, as reconstructed from the X chromosome.

Assessing the uniqueness of the *RRM2P4* genealogy: The uniqueness of the *RRM2P4* central region was estimated using a simulation-based summary-likelihood approach under the RAR and AA models (Figure 2). Both the distribution of TMRCA values and the proportion of African individuals in one of the two basal clades were inferred under both models. The mean TMRCA under the two-deme model was 1.15 MYA, with dates exceeding 2.12 MYA being statistical outliers ($\alpha = 0.05$). While the maximum-likelihood method assuming panmixia returned a mean TMRCA value similar to that produced under the RAR model [*i.e.*, $P(\text{TMRCA} > 1.24 \times 10^6 \mid \text{RAR}) = 0.350$], both the molecular clock [$P(\text{TMRCA} > 2.33 \times 10^6 \mid \text{RAR}) = 0.030$] and the maximum-likelihood method assuming an island model [$P(\text{TMRCA} > 2.88 \times 10^6 \mid \text{RAR}) = 0.008$] yielded TMRCA values that are unusually old. Under the AA model, the TMRCA distribution is shifted deeper in time relative to RAR models, and genealogies exceeding 3 MYA would be expected under our archaic admixture scenario (Figure 5, a and b). The molecular-clock-based estimate of the true TMRCA is not an outlier under the AA model as specified here [$P(\text{TMRCA} > 2.33 \times 10^6 \mid \text{AA}) = 0.140$].

The geographical distribution of lineages on the central *RRM2P4* genealogy is also skewed: one of the two basal clades (Figure 4, clade A) is found infrequently in Africans (0.048) but commonly in non-Africans (0.952). This pattern can also be compared with the empirical distribution of p_{mc}, as determined from the additional data set of 19 X chromosome loci, among which the smallest minimum African clade proportion is 0.214. The likelihood of observing the same, or a more extreme, proportion of Africans (*i.e.*, $\frac{1}{21}$) in a basal clade was statistically significant under the RAR model [$P(p_{\text{mc}} \leq 0.048 \mid \text{RAR}) = 0.031$], but not so under the AA model [$P(p_{\text{mc}} \leq 0.048 \mid \text{AA}) = 0.24$]. To

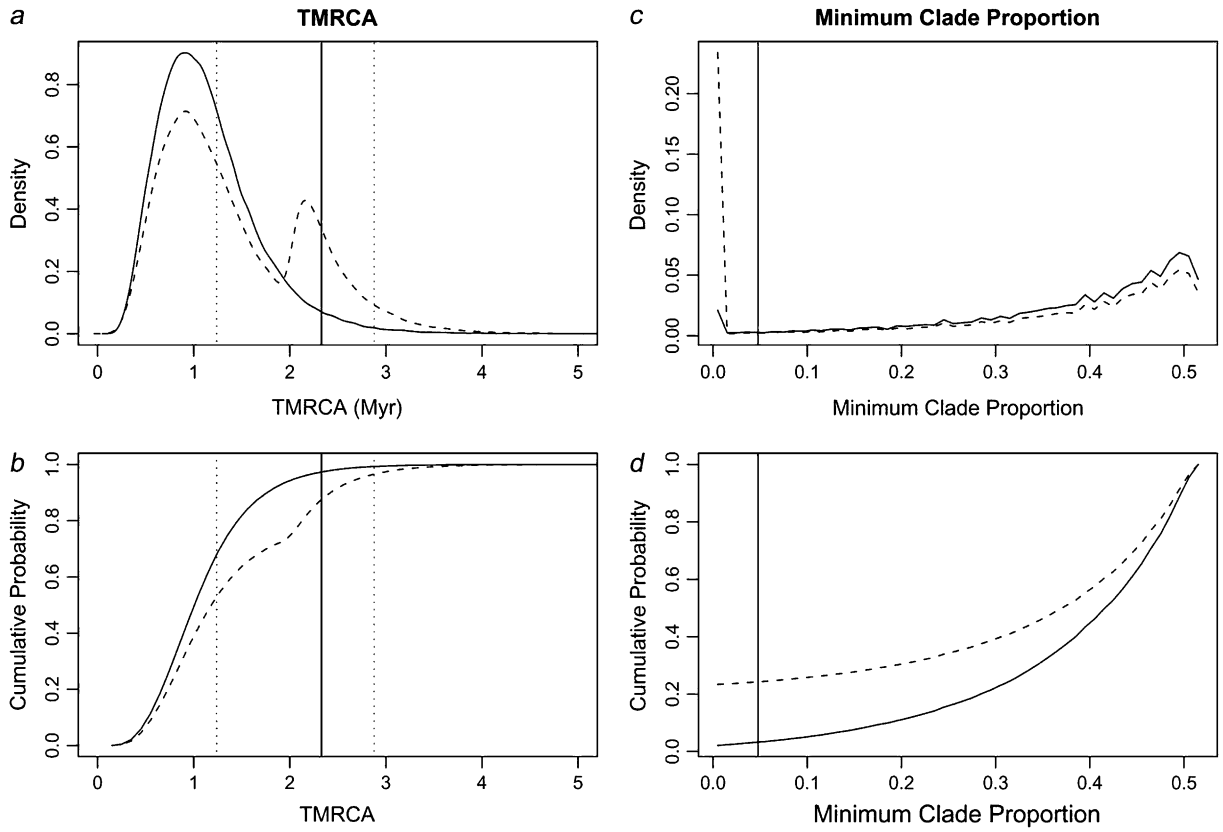

FIGURE 5.—Distribution and cumulative probability of (a and b) TMRCAs and (c and d) minimum clade proportions under the optimal two-deme RAR model (solid curve) and the corresponding model with 5% introgression from ancestral hominins at 50 KYA (dashed curve). The molecular-clock date is shown by a solid vertical line in a and b; the panmictic and island model dates are shown by dotted vertical lines (to the left and right, respectively). The admixture peak would increase as admixture occurred more frequently and shift right with deeper structure between admixing demes. The observed p_{mc} is illustrated by a solid vertical line in c and d.

further explore the effects of archaic admixture on the minimum clade proportion, we took the optimized RAR model and incorporated variable rates of admixture (0.5–5%) occurring at variable times (10–55 KY before present) and ran 10^5 simulations at each point on a 10×10 grid of parameter space (Figure 6). Basal clades dominated by non-Africans are observed more often on average as archaic admixture becomes more recent and more frequent. Higher rates of admixture make it more likely that one of the two basal clades derives entirely from the descendants of Asian ancestral hominins ($p_{mc} \rightarrow 0$), and more recent admixture makes it less likely that migrants will carry admixed lineages into African populations (also, $p_{mc} \rightarrow 0$).

DISCUSSION

GARRIGAN *et al.* (2005b) described a 2.4-kb region on the X chromosome with unusual genealogical structure in a sample of 41 humans: a deep TMRCA and a basal clade composed entirely of Asian ($n = 3$) sequences. This differs from most genealogies observed to date, where African individuals dominate at least one of the two basal clades (LABUDA *et al.* 2000; TAKAHATA *et al.* 2001; SATTA and TAKAHATA 2004). GARRIGAN *et al.* (2005b) also genotyped a single diagnostic SNP to test for the presence of clade A in a larger number of samples ($n = 570$ from 17 globally distributed populations). They discovered a decreasing frequency gradient centered on southern China (where the clade A is present >50%) and extremely low frequencies of the "Asian" divergent lineage in Europe, the Middle East, and Africa. (See Figure 1 in GARRIGAN *et al.* 2005b.) To explain the prevalence of basal *RRM2P4* lineages in East Asia they favored a model of recent admixture between divergent AMH and *H. erectus* populations; although they could not rule out founder effects leading to the loss of one of the two divergent lineages in Africa. Here, we resequence this locus in a much larger sample of humans, extend the length of the sequenced region, and test unusual aspects of the genealogy statistically, using a model-based coalescent simulation framework.

We generated resequencing data 5′ and 3′ of the central 2.4-kb region in a panel of 90 individuals to see whether the pattern originally described by GARRIGAN *et al.* (2005b) extended farther along the X chromosome. We found that a strong recombination hotspot almost completely decouples the central and 3′ regions,

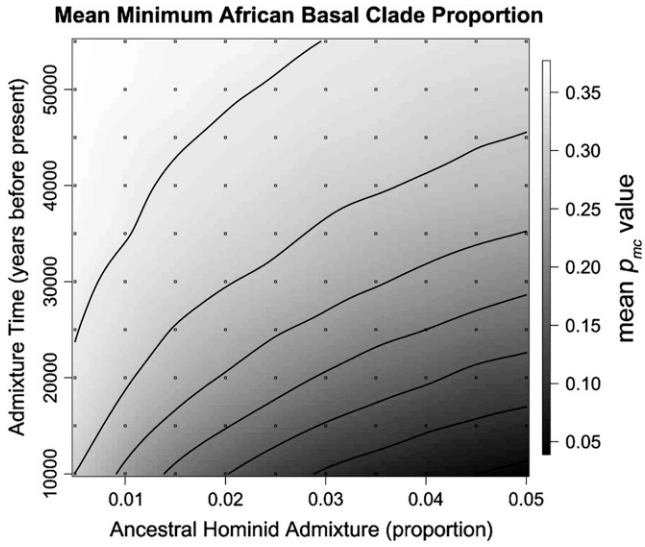

FIGURE 6.—Effect of admixture parameters on the minimum clade proportion. Fewer Africans are found in one basal clade as archaic Asian admixture occurs more frequently and more recently. The likelihood surface is generated by interpolation from mean p_{mc} values of 10^5 simulations at each point in a 10×10 grid (dots) covering the parameter space of admixture time and admixture proportion.

and despite linkage disequilibrium with the 5′ end of the sequence, the pattern of higher Asian diversity for the central region was not found for the other two regions. While it is possible that this reduced diversity may simply reflect the stochastic nature of the mutational process given the relatively short length of sequence examined (~2 kb), we suggest that the lineage history of the central region has been at least partly decoupled from those of the 5′ and 3′ regions through recombination. BLAST results show that the *RRM2P4* processed pseudogene has sequence conservation to the rhesus macaque (*Macaca mulatta*). Therefore, the high diversity of the central region does not result from increased mutation at the time of pseudogene insertion because human variation traces back only to the Homo lineage (*i.e.*, the TMRCA of human haplotypes is much more recent than the insertion time). Central region diversity is also unlikely to reflect paralogous gene conversion because the *RRM2P4* sequence has no close matches to other regions in the human genome. Finally, increased central region diversity is unlikely to reflect linkage to a gene under long-term balancing selection. The nearest 5′ gene is *SPANX-N2*, which encodes a protein of uncharacterized function. *SPANX-N2* is 580 kb, or ~1.8 cM, distant from the *RRM2P4* pseudogene, and the intervening region contains at least three hotspots of ~16 cM/Mb (each similar to the hotspot 3′ of the *RRM2P4* locus). While we cannot exclude the possibility that an unknown 5′ functional variant may be linked to the *RRM2P4* locus, the disparity in diversity between the *RRM2P4* 5′ and central regions makes linkage with a third selected locus located farther 5′ unlikely. In sum, it appears that polymorphisms that clearly define the unusual *RRM2P4* genealogy are found only within the stretch of noncoding DNA associated with the pseudogene (Figure 1) and that these sites are unlikely to be affected by recent positive selection.

Our sixfold larger database of central region sequences does not result in a substantial change in the topology of the gene tree described by GARRIGAN *et al.* (2005b) or in the partitioning of its two deepest branches. However, we did observe several new lineages at low frequency and discovered some novel evidence of recombination. Of the 21 individuals identified with the less frequent basal lineage, only one was from Africa. This Dogon individual from Mali was the same African individual identified as carrying a clade A lineage on the basis of the SNP-based genotyping assay in GARRIGAN *et al.* (2005b). Our extended sequence database identified only a single new polymorphism in clade A and 11 new polymorphisms in the more diverse globally distributed clade B (compare Figure 1 of GARRIGAN *et al.* 2005b with our Figure 4). This brings the total number of haplotypes in the divergent Asian clade A to 3, compared with a total of 20 haplotypes in the other basal clade. Interestingly, the new clade A haplotype was identified only in two individuals: the aforementioned Dogon individual and an individual from Bougainville Island, Melanesia. This is consistent with the possibility that recent migration carried this rare lineage between continents. Although we cannot exclude the possibility that unsampled African populations carry this clade at higher frequency than we observe here, our geographical coverage of the African continent is still quite extensive.

Our simulations show that the 2.33-million-year molecular-clock-based TMRCA estimate is a statistical outlier under the RAR model (Figure 5, a and b), but not under an AA model with ancestral structure dating from 2 MYA. However, maximum-likelihood estimates of the TMRCA inferred under different models of human population structure span a wide range of times. The large variance in TMRCA estimates is also expected because the short length of the central subregion sequence limits the number of sites segregating between the two basal clades. Unfortunately, the variance of TMRCA is always large relative to the mean and does not decrease appreciably with increased sample size (GRIFFITHS and TAVARÉ 1994; TANG *et al.* 2002; BASU and MAJUMDAR 2003). Indeed, as we increased the sample size from 41 (GARRIGAN *et al.* 2005b) to 253, we found no new polymorphic sites on the basal branches of the gene tree, which suggests that we have sampled sufficiently to observe the basal node of the genealogy for the entire global population. There is only a remote probability that we have not observed the deepest split of the *RRM2P4* tree in our data set of 253 individuals ($P \approx 0.0079$) (SAUNDERS *et al.* 1984; KLIMAN and HEY 1993). Our TMRCA estimates, while consistent with the AA

model specified here, are unlikely to be improved by sampling additional individuals.

We also considered the observation of Garrigan *et al.* (2005b) that Asian samples are overrepresented in one of the two basal clades of the *RRM2P4* tree. This pattern continued to hold even after increasing the size of our DNA sequence data set. To assess how unusual this aspect of the *RRM2P4* genealogy is under alternative models of human evolutionary history, we defined a new summary statistic, p_{mc}, which quantifies the skew in the proportion of individuals from two populations among the two basal clades of a gene tree. This summary statistic is applicable to *RRM2P4* because the central region is essentially tree-like. We observed $p_{mc} = 0.048$ for the central region of *RRM2P4*, which is significantly unlikely under the RAR model ($P = 0.031$, Figure 5, c and d). To examine the sensitivity of the p_{mc} statistic under a range of archaic admixture parameters, we simulated coalescent genealogies and varied both the admixture proportion and the timing of introgression (Figure 6). Although genealogies with small p_{mc} values are more common as the admixture proportion increases (*i.e.*, up to ~5%) and introgression begins more recently (*i.e.*, as recently as ~10 KYA), the *RRM2P4* genealogy is not a significant outlier under any of these AA model parameterizations. This further supports our conclusion that the *RRM2P4* genealogy fits expectations better under a scenario of archaic admixture.

While there are limitations with both TMRCA and p_{mc} for distinguishing predictions of the RAR and AA models, they do represent independent summaries of the data and, thus, complement one another. As already mentioned, the power of these two test statistics depends on different aspects of sampling. Variance in the estimate of the TMRCA is improved by longer sequences of the region with tree-like ancestry (in the case of *RRM2P4* this is limited by the small central region and flanking recombination), but only slightly by increasing the sample size. On the other hand, estimates of p_{mc} can be improved by increasing the sample size, because the variance of p_{mc} decreases approximately as the inverse of the sample size (analyses not shown). If the p_{mc} is genuinely an outlier under the RAR model, increasing the number of individuals sampled increases the power to reject RAR. Indeed, when we use the SNP data of Garrigan *et al.* (2005b), which included 177 Africans and 393 non-Africans, we reject the RAR model with greater confidence ($p_{mc} = 0.0189$, $P = 0.014$).

Further evidence in support of an archaic admixture model awaits analysis of additional loci exhibiting genealogical properties similar to the central *RRM2P4* region. Several candidates have already been identified (Harding *et al.* 1997; Ziętkiewicz *et al.* 2003; Stefansson *et al.* 2005; Shimada *et al.* 2007), but most lack rigorous statistical analyses under a range of demographic models, including ancient admixture alternatives. The frequency at which we expect to find introgressed regions depends largely on the amount of admixture between the two archaic populations (Wall 2000). Moreover, unless admixture was recent and involved highly divergent populations, the power to detect archaic admixture is low (Nordborg 2000). In the case of *RRM2P4*, divergence may have started at the time of separation of *H. ergaster*/*H. erectus* populations in Africa ~2 MYA (Anton and Swisher 2004). Yet, the length of the divergent sequence is short, possibly as a result of the nearby recombination hotspot or because admixture occurred in the more distant past and recombination has subsequently broken down the admixed chromosome. In any case, identifying longer sequences with greater divergence would allow for more sophisticated tests of archaic admixture. For example, Wall (2000) suggested a number of summary statistics that are based on both the level of divergence between two clades and the amount of recombination between them.

Recent population structure is another factor that may affect the probability of sampling a locus with a genealogy showing signs of archaic admixture. As pointed out by Nordborg (2000), population structure may actually increase the power to detect archaic admixture if we sample sufficiently among demes, because we would expect the introgressed alleles to still be present in the area of the world where admixture took place. In the case of *RRM2P4*, individuals carrying the less frequent divergent lineage are concentrated in East Asia, suggesting that admixture may have occurred at the Asian end of the global distribution of human populations. On the other hand, it is important to point out that current population structure is unlikely to reflect ancient patterns directly. Following a demic expansion, what was once subdivision between two African populations may now appear as structure between African and non-African populations. For loci with more ancient TMRCA, there is an increase in power to detect archaic admixture even if it occurred at more ancient times (Nordborg 2000). This means that for *RRM2P4*, which has an ancient TMRCA, we cannot be confident about where archaic admixture may have occurred geographically. In this regard, it is interesting to note that a growing number of loci have been discovered with two deeply divergent lineages where both the major and the minor types are present only in African populations (Barreiro *et al.* 2005; Garrigan *et al.* 2005a; Hayakawa *et al.* 2006). This supports models in which anatomically modern humans descend from a structured ancestral African population (Garrigan and Hammer 2006). We find some support that elevated admixture among highly divergent African subpopulations just prior to the recent African expansion could explain the pattern of polymorphism at *RRM2P4* (supplemental Figure 4 at http://www.genetics.org/supplemental/), but note that this model has little power to explain why *RRM2P4* clade A lineages are geographically restricted to East Asia today. For now, this locus represents

a genealogical history that is most consistent with recent admixture from an archaic hominin population in Asia.

We thank Zahra Mobasher (University of Arizona) for excellent technical assistance and David Morales (University of Arizona) for helpful discussion. This research forms part of the HOMINID project, a genomic resequence study funded by National Science Foundation grant BCS-0423670.

LITERATURE CITED

Anton, S. C., and C. C. Swisher, 2004 Early dispersal of Homo from Africa. Annu. Rev. Anthropol. **33:** 271–296.

Barreiro, L. B., E. Patin, O. Neyrolles, H. M. Cann, B. Gicquel *et al.*, 2005 The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. Am. J. Hum. Genet. **77:** 869–886.

Basu, A., and P. Majumdar, 2003 A comparison of two popular statistical methods for estimating the time to most recent common ancestor (TMRCA) from a sample of DNA sequences. J. Genet. **82:** 7–12.

Beaumont, M. A., W. Zhang and D. J. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162:** 2025–2035.

Cann, H. M., C. de Toma, L. Cazes, M. F. Legrand, V. Morel *et al.*, 2002 A human genome diversity cell line panel. Science **296:** 261–262.

Eswaran, V., 2002 A diffusion wave out of Africa: the mechanism of the modern human revolution? Curr. Anthropol. **43:** 749–774.

Evans, P. D., N. Mekel-Bobrov, E. J. Vallender, R. R. Hudson and B. T. Lahn, 2006 Evidence that the adaptive allele of the brain size gene *microcephalin* introgressed into *Homo sapiens* from an archaic *Homo* lineage. Proc. Natl. Acad. Sci. USA **103:** 18178–18183.

Fenner, J. N., 2005 Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. **128:** 415–423.

Garrigan, D., and M. F. Hammer, 2006 Reconstructing human origins in the genomic era. Nat. Rev. Genet. **7:** 669–680.

Garrigan, D., Z. Mobasher, S. B. Kingan, J. A. Wilder and M. F. Hammer, 2005a Deep haplotype divergence and long-range linkage disequilibrium at Xp21.1 provide evidence that humans descend from a structured ancestral population. Genetics **170:** 1849–1856.

Garrigan, D., Z. Mobasher, T. Severson, J. A. Wilder and M. F. Hammer, 2005b Evidence for archaic Asian ancestry on the human X chromosome. Mol. Biol. Evol. **22:** 189–192.

Griffiths, R. C., 2007 Genetree v. 9.0. http://www.stats.ox.ac.uk/~griff/software.html.

Griffiths, R. C., and S. Tavaré, 1994 Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46:** 131–159.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox *et al.*, 1997 Archaic African *and* Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60:** 772–789.

Hayakawa, T., I. Aki, A. Varki, Y. Satta and N. Takahata, 2006 Fixation of the human-specific CMP-*N*-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. Genetics **172:** 1139–1146.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.

Karafet, T. M., L. P. Osipova, M. A. Gubina, O. L. Posukh, S. L. Zegura *et al.*, 2002 High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. Hum. Biol. **74:** 761–789.

Kingman, J. F. C. (Editor), 1982 *On the Genealogy of Large Populations.* Applied Probability Trust, Sheffield, UK.

Kliman, R. M., and J. Hey, 1993 DNA sequence variation at the *period* locus within and among species at the *Drosophila melangaster* complex. Genetics **133:** 375–387.

Labuda, D., E. Ziętkiewicz and V. Yotova, 2000 Archaic lineages in the history of modern humans. Genetics **156:** 799–808.

McBrearty, S., and A. S. Brooks, 2000 The revolution that wasn't: a new interpretation of the origin of modern human behavior. J. Hum. Evol. **39:** 453–563.

McDougall, I., F. H. Brown and J. G. Fleagle, 2005 Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature **433:** 733–736.

McVean, G., and C. C. Spencer, 2006 Scanning the human genome for signals of selection. Curr. Opin. Genet. Dev. **16:** 624–629.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics **146:** 1501–1514.

Nordborg, M., 2000 On detecting ancient admixture, pp. 123–136 in *Genes, Fossils and Behaviour: An Integrated Approach to Human Evolution*, edited by P. Donnelly. IOS Press, Amsterdam.

Plagnol, V., and J. D. Wall, 2006 Possible ancestral structure in human populations. PLoS Genet. **2:** e105.

Satta, Y., and N. Takahata, 2004 The distribution of the ancestral haplotype in finite stepping-stone models with population expansion. Mol. Ecol. **13:** 877–886.

Saunders, I. W., S. Tavaré and G. A. Watterson, 1984 On the genealogy of nested subsamples from a haploid population. Adv. Appl. Probab. **16:** 471–491.

Shimada, M. K., K. Panchapakesan, S. Tishkoff, A. Q. Nato, Jr. and J. Hey, 2007 Divergent haplotypes and human history as revealed in a worldwide survey of X-linked DNA sequence variation. Mol. Biol. Evol. **24:** 687–698.

Stefansson, H., A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson *et al.*, 2005 A common inversion under selection in Europeans. Nat. Genet. **37:** 129–137.

Takahata, N., S. H. Lee and Y. Satta, 2001 Testing multiregionality of modern human origins. Mol. Biol. Evol. **18:** 172–183.

Tang, H., D. O. Siegmund, P. Shen, P. J. Oefner and M. W. Feldman, 2002 Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. Genetics **161:** 447–459.

Templeton, A., 2002 Out of Africa again and again. Nature **416:** 45–51.

Wall, J. D., 2000 Detecting ancient admixture in humans using sequence polymorphism data. Genetics **154:** 1271–1279.

Wall, J. D., and M. F. Hammer, 2006 Archaic admixture in the human genome. Curr. Opin. Genet. Dev. **16:** 606–610.

Y Chromosome Consortium, 2002 A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. **12:** 339–348.

Ziętkiewicz, E., V. Yotova, D. Gehl, T. Wambach, I. Arrieta *et al.*, 2003 Haplotypes in the Dystrophin DNA segment point to a mosaic origin of modern human diversity. Am. J. Hum. Genet. **73:** 994–1015.

Communicating editor: N. Takahata

APPENDIX

Here, we define a new summary statistic, the minimum clade proportion (p_{mc}). In its simplest form, this statistic characterizes the proportion of individuals from a specified group (*e.g.*, Africans) in each of the two basal clades of a binary gene tree. The minimum clade proportion can thus be defined as

$$p_{\mathrm{mc}} = \min\left(\frac{k_1}{n_1}, \frac{k_2}{n_2}\right), \tag{A1}$$

where $n_{1,2}$ and $k_{1,2}$ are, respectively, the total number of individuals and the number of chromosome copies from the specified group in basal clades 1 and 2. Clades 1 and 2 are defined arbitrarily because, *a priori*, branch rotation has no effect on the topology of a gene tree. We also note that the p_{mc} statistic is not itself a proportion, although the quantity (k_i/n_i) is. Strictly, $p_{\mathrm{mc}} \in [0, 1)$; *i.e.*, the minimum clade proportion does not include unity.

The variance of p_{mc} decreases approximately as the inverse of the sample size; *i.e.*, increasing the number of sampled individuals increases the power to reject the null model (our unpublished simulation results). Furthermore, the power of this test depends on sampling sizes as follows. Consider that we have N_1 individuals from the specified group, $N_1 + N_2$ total individuals, and an observed $p_{\mathrm{mc}} = \lambda$. Also, given an arbitrarily chosen clade, p_1 is that clade's frequency in the specific group, whereas p_2 is the frequency of that clade in the remaining individuals. For a given hypothesis H

$$P(p_{\mathrm{mc}} = \lambda \mid H, N_1, N_2) = \int_{p_1,p_2} P(p_{\mathrm{mc}} = \lambda \mid N_1, N_2, p_1, p_2) \cdot f(p_1, p_2 \mid H)\, dp_1\, dp_2, \tag{A2}$$

where $f(p_1, p_2 \mid H)$ is the probability density for the population frequencies given the hypothesis. Note that the first factor of the integrand in Equation A2 depends only on allele frequencies and sample sizes and not on the hypothesis, whereas the second factor depends only on the hypothesis. The conditioned probability distribution of p_{mc} therefore resembles the ratio of two nonindependent binomial distributions. Given p_1, p_2, N_1, and N_2 and taking $A \sim \mathrm{Binom}(p_1, N_1)$ and $B \sim \mathrm{Binom}\,(p_2, N_2)$, it follows that

$$p_{\mathrm{mc}} = \min\left(\frac{A}{A+B}, \frac{N_1 - A}{N_1 + N_2 - A - B}\right). \tag{A3}$$

If the sample size is sufficiently large and the frequencies p_1 and p_2 are not identical, the probability distribution of p_{mc} converges to the probability distribution of the ratio of random variables that has the smallest mean. As N_1 and N_2 increase, the values of A and B approach $N_1 p_1$ and $N_2 p_2$, respectively, whereas their standard errors grow on the order of $\sqrt{N_1}$ and $\sqrt{N_2}$. Consequently, it is perhaps not surprising that the variance of p_{mc} decreases at least as the inverse of the smallest sample size. In other words, for cases where population frequencies are unlikely under the null hypothesis, raising the sample size can increase the power to reject the model (*cf.* TMRCA).

We emphasize that the likelihood of an observed p_{mc} depends strongly on the demographic model underlying the null hypothesis. In practice, the probability that $p_{\mathrm{mc}} \leq \lambda$ must be determined by coalescent simulation.

Finally, we note that the p_{mc} statistic can be generalized to any subset of chromosome copies, K, and any number of clades, C (indexed by i), from a data set, N. Necessary conditions include $K \subset N$, $C_i \subset N$, and $0 < |K| \leq |N|$, where $|K|$ and $|N|$ are the cardinals of K and N, respectively. It follows that the number of chromosome copies from the specified group in clade C_i is $k_i = |(K \cap C_i)|$, and the total number of individuals in clade C_i is simply $n_i = |C_i|$. Consequently, the minimum clade proportion can be defined more generally for m clades as

$$p_{\mathrm{mc}} = \min\left(\frac{k_1}{n_1}, \ldots, \frac{k_m}{n_m}\right). \tag{A4}$$

Only coalescent simulations for which the relationship on the gene tree of the clades, m, is identical to that of the data should be used to determine the probability of this generalized p_{mc} statistic.