Human DNA Sequences: More Variation and Less Race

Jeffrey C. Long,[1]* Jie Li,[1] and Meghan E. Healy[2]

[1]*Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109-5618*
[2]*Department of Anthropology, University of New Mexico, Albuquerque, NM 87131*

KEY WORDS race; DNA sequence; short tandem repeat; diversity; hierachical models

ABSTRACT Interest in genetic diversity within and between human populations as a way to answer questions about race has intensified in light of recent advances in genome technology. The purpose of this article is to apply a method of generalized hierarchical modeling to two DNA data sets. The first data set consists of a small sample of individuals ($n = 32$ total, from eight populations) who have been fully resequenced for 63 loci that encode a total of 38,534 base pairs. The second data set consists of a large sample of individuals ($n = 928$ total, from 46 populations) who have been genotyped at 580 loci that encode short tandem repeats. The results are clear and somewhat surprising. We see that populations differ in the amount of diversity that they harbor. The pattern of DNA diversity is one of nested subsets, such that the diversity in non-Sub-Saharan African populations is essentially a subset of the diversity found in Sub-Saharan African populations. The actual pattern of DNA diversity creates some unsettling problems for using race as meaningful genetic categories. For example, the pattern of DNA diversity implies that some populations belong to more than one race (e.g., Europeans), whereas other populations do not belong to any race at all (e.g., Sub-Saharan Africans). As Frank Livingstone noted long ago, the Linnean classification system cannot accommodate this pattern because within the system a population cannot belong to more than one named group within a taxonomic level. Am J Phys Anthropol 139:23–34, 2009. © 2009 Wiley-Liss, Inc.

Richard Lewontin was the first researcher to apply measures of genetic variability within and between human populations to questions about human races (Lewontin, 1972). He found in an analysis of blood group and protein allele frequencies that relative to the total diversity for our species, the within groups component is 85.4%, the component for populations within races is 8.3%, and the interracial component is only 6.3%. On this basis, Lewontin concluded that human races are of virtually no genetic or taxonomic significance. His findings have had lasting impact, and many later studies produced similar results using a similar model of population structure with different sources of data and different statistical methods (Michalakis and Excoffier, 1996; Barbujani et al., 1997; Jorde et al., 2000; Romualdi et al., 2002; Li et al., 2008).

Geneticists have renewed their interest in the variability within and between groups in light of advances in genome technology (Bamshad et al., 2003; Burchard et al., 2003; Cooper et al., 2003). Our current methods in genomics have several advantages relative to the methods of the 1970s. They reveal exact DNA sequence changes, can yield data in amounts that were unimaginable even a few decades ago, and allow us to make comparisons across related species. There is no ascertainment bias associated with the patterns of diversity detected in DNA sequencing studies (Clark et al., 2005; Keinan et al., 2007), whereas the patterns of diversity in single nucleotide polymorphism (SNP) marker studies reflect the methods of SNP discovery as well as evolutionary differences between populations. Studies at the DNA level are providing new insights into genetic diversity within and between human populations.

Rosenberg and colleagues analyzed 377 short tandem repeat (STR) DNA loci in a sample of 1,052 people (Rosenberg et al., 2002). These individuals came from 52 populations with locations throughout the world. This study found that the component of allelic diversity between major geographic regions accounted for only 4.3% of the total. However, the study's findings present us with a paradox because it was possible to use these genotypes to classify individuals back to their regional populations. Others have argued that Rosenberg and colleagues should have measured STR diversity using a different statistic, and offer that their favored statistic yields a higher component of diversity between major geographic regions, that is, 9.2% (Excoffier and Hamilton, 2003). However, neither 4.3% nor 9.2% is very different from Lewontin's 6.3%, and the high classification success in Rosenberg's study is counter intuitive in comparison with all such measures of diversity. We must now judge taxonomic significance on a different basis than the component of diversity between populations.

We know less about the pattern of diversity for nonrepeated DNA sequences; however, studies reporting on DNA sequence diversity in noncoding autosomal genomic

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: The National Science Foundation; Grant number: NSF 0321610; Grant sponsor: The University of Michigan Endowment for Biological Sciences.

*Correspondence to: Jeffrey C. Long, Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA.
E-mail: longjc@umich.edu

Received 31 March 2008; accepted 2 December 2008

DOI 10.1002/ajpa.21011
Published online 18 February 2009 in Wiley InterScience (www.interscience.wiley.com).

regions are now emerging (Yu et al., 2002; Fischer et al., 2006; Wall et al., 2008). These studies tend to utilize small samples of individuals that are limited in their geographic coverage. Nonetheless, these studies have provided provocative results.

Yu and colleagues collected DNA sequences comprising 25,000 nucleotides for each member in a sample of 30 individuals (Yu et al., 2002). These 25,000 nucleotides represent 50 autosomal loci each composed of ~500 contiguous base pairs. Ten of these individuals were African, 10 Asian, and 10 European. The investigators selected individuals from each continent from different local groups residing in widely spaced regions, but only a single person represented each local group. This sampling scheme extricates the between region component of nucleotide diversity but it confounds the diversity within populations with that between populations within the same region. The unexpected result from this study is that nucleotide diversity is greater between two African DNA sequences than between an African DNA sequence and a European DNA sequence, or an African DNA sequence and an Asian DNA sequence. Thus, in comparing the African sample to the European sample or the Asian sample, the estimated between groups diversity component is negative.

Fischer and colleagues (Fischer et al., 2006) analyzed DNA sequences comprising 22,401 nucleotides collected by Voight and colleagues (Voight et al., 2005) for each member in a sample of 45 individuals. These 22,401 nucleotides represent 26 autosomal loci each composed of ~860 contiguous base pairs. This sample includes 15 individuals from each of three human populations: Hausa, Italian, and Han Chinese. This sampling scheme provides a valid estimate of nucleotide diversity within populations, but it confounds the diversity for populations within geographic regions with the diversity between geographic regions. These researchers found that the diversity between populations (including both components) ranged from 9 to 15%, depending on the pair of populations being compared. The interesting feature of this study is that Fischer and colleagues went on to sequence the homologous loci in multiple samples from two species of great apes. They found that the diversity between human populations exceeds that estimated between Eastern (*Pan troglodytes schweinfurthii*) and Central (*Pan troglodytes troglodytes*) subspecies of Chimpanzee. In this light, we must either reassess the evidence for subspecies of Chimpanzee, or else judge taxonomic significance on a different basis than the component of diversity between populations.

The last four decades have also seen advances in population genetics theory and analytical methods. These advances are providing a better framework for drawing inferences about evolution from diversity within and between populations.

Hedrick showed that the between groups component of diversity was inversely proportional to the within groups component of diversity (Hedrick, 1999). Therefore, the between groups component cannot be high for systems such as STRs that are highly polymorphic within groups. Edwards gave an example of how, by using enough markers, classification is possible even when the between populations variance is low for all markers (Edwards, 2003). These two findings taken in combination explain why Rosenberg and colleagues were able to accurately classify individuals when the between groups component of diversity for their markers is so low.

Long and colleagues showed that partitioning diversity into within and between groups components is sensitive to a host of a priori assumptions (Urbanek et al., 1996; Long and Kittles, 2003). First, the expected diversity is the same within all populations sampled. Second, the expected diversity between any pair of populations within a region is the same, regardless of the region. Third, the expected diversity between any pair of populations in different regions is the same, regardless of the pair of regions. When these assumptions are violated, the total diversity is underestimated, the component of diversity within groups is over-estimated, and the component of diversity between groups is underestimated. Long and Kittles found that some human groups are far more diverged than would be implied by standard components of genetic diversity, whereas other groups are much less diverged (Long and Kittles, 2003).

The present study revisits the question of genetic diversity within and between populations. We provide new data from direct DNA sequencing of 63 noncoding loci. The total sequence length summed over all loci is 38,534 base pairs. We analyze a total sample of 32 people, which despite its small size, has multiple individuals from eight local populations and multiple local populations from Africa, Europe, and Asia. We also analyze a large set of STR data for 46 populations from Africa, Europe, and Asia. These STR data are from the CEPH diversity panel (Rosenberg et al., 2002), with the addition of a sample from the Gujarati of India (Rosenberg et al., 2006). Our strategy is as follows. First, for our DNA sequence data, we estimate components of diversity at the same levels of hierarchical population structure that Lewontin did (Lewontin, 1972). Second, we use an expanded hierarchical arrangement of populations to see i) if this arrangement achieves a significantly better fit to the data than does Lewontin's arrangement, and ii) to see whether this arrangement alters Lewontin's conclusions about the apportionment of diversity within and between human populations. Third, we repeat the first two steps of the analysis using the STR loci. The important question is—Are the patterns of nucleotide diversity that we observe for DNA sequences in eight populations changed by adding more populations assayed for a different kind of DNA polymorphism? We note that the STR data set includes seven of the eight populations for which we collected DNA sequences, and it includes a population that neighbors our eighth population.

MATERIALS AND METHODS

Samples and genetic systems

DNA sequencing. Our sample has the following composition: from Africa, (n = 4) Biaka, (n = 4) Yoruba, and (n = 4) Luhya from Kenya; from Europe, (n = 4) Iberian, (n = 4) Russian from Moscow; from Asia, (n = 4) Gujarati, (n = 4) Han, and (n = 4) Japanese. We purchased these DNA samples from Coriell Institute for Biomedical Research Human Diversity and HapMap Collections. The ID numbers and population affiliations for these samples are as follows NA10469 (Biaka), NA10470 (Biaka), NA10471 (Biaka), NA10472 (Biaka), NA18856 (Yoruba), NA19222 (Yoruba), NA19093 (Yoruba), NA19204 (Yoruba), NA19314 (Luhya), NA19351 (Luhya), NA19338 (Luhya), NA19376 (Luhya), NA17091 (Iberian), NA17092 (Iberian), NA17094 (Iberian), NA17097 (Iberian), NA13820 (Moscow), NA13838 (Moscow), NA13852 (Moscow), NA13876 (Moscow), NA20854 (Gujarati),

TABLE 1. Unbiased nucleotide diversity (×100) and d_A *genetic distance (×100) estimates*[a]

	Biaka	Yoruba	Luhya	Iberian	Moscow	Gujarati	Han	Japanese
Biaka	0.086	0.096	0.090	0.098	0.094	0.092	0.099	0.093
Yoruba	0.014	0.092	0.090	0.094	0.088	0.090	0.091	0.090
Luhya	0.012	0.005	0.082	0.085	0.081	0.081	0.085	0.085
Iberian	0.037	0.023	0.016	0.073	0.064	0.071	0.071	0.075
Moscow	0.046	0.026	0.024	0.000	0.057	0.065	0.066	0.069
Gujarati	0.034	0.024	0.016	0.005	0.009	0.064	0.067	0.065
Han	0.056	0.034	0.032	0.013	0.020	0.014	0.055	0.059
Japanese	0.040	0.029	0.030	0.017	0.021	0.007	0.002	0.060
N	4	4	4	4	4	4	4	4

[a] Within population nucleotide diversity estimates appear on the diagonal, between population nucleotide diversity estimates appear above the diagonal, genetic distance estimates appear below the diagonal.

TABLE 2. Treeness tests and model comparison statistics

Data set/model	Λ	d*f*	AIC	*P*-Value	$F_{df1,df2}$	*P*-Value
DNA sequences						
Two-level island model	184.2	33	250.2	0.000		
Expanded hierarchical model	49.4	21	91.4	0.000	2.4	0.020
$\Delta = AIC_{TLIM} - AIC_{EHM}$			158.8			
580 STRs						
Two-level island model	44,325.8	1,078	46,481.8	0.000		
Expanded hierarchical model	6,149.3	993	8,135.3	0.000	6.6	0.000
$\Delta = AIC_{TLIM} - AIC_{EHM}$			38,346.5			

NA20861 (Gujarati), NA20901 (Gujarati), NA20906 (Gujarati), NA16654 (Han Chinese), NA16688 (Han Chinese), NA17016 (Han Chinese), NA17020 (Han Chinese), NA17051 (Japanese), NA17053 (Japanese) NA17057 (Japanese), and NA17060 (Japanese). For clarification, the $n = 32$ individuals in our total sample contribute $2n = 64$ copies of each locus.

We sequenced 49 of the loci that Yu et al. sequenced (Yu et al., 2002), five loci from the W-H Li laboratory that have not appeared in their publications, and nine new loci that we identified using the procedure from Chen and Li (Venter et al., 2001). A single primer pair amplifies each locus. We designed PCR and sequencing primers using the human reference sequence, as made available on the UCSC genome browser (http://genome.ucsc.edu/). Supporting Information Table 1 provides the physical map positions and PCR primer sequences for all 63 loci. The University of Michigan DNA sequencing core facility performed dye-terminator sequencing for us using Applied Biosystems Big-Dye reagents on an Applied Biosystems automated sequencer. To ensure accuracy, we determined DNA sequence for both strands for all amplicons. We used the Seqman module in the DNASTAR package to analyze chromatograms and perform alignments; then, we used an original computer program to identify polymorphic sites, calculate allele frequencies, and tabulate statistics by population.

STR data. We also analyzed a set of autosomal STR polymorphisms that is publicly available for a large set of populations. We focused on a set of 46 populations that includes all 45 African, European, and Asian populations from the 52 worldwide populations that Rosenberg and colleagues analyzed, plus a sample from the Gujarati of India (Rosenberg et al., 2002, 2006). The genotypes are available from the Marshfield Clinic Genotyping Service (http://research.marshfieldclinic.org/genetics/home/index.asp). The genotypes for the non-Gujarati samples come from directly from the diversity collection. The genotypes for the Gujarati are a subset of the Marshfield India collection. Drs. Pragna Patel and Noah Rosenberg kindly provided us with information about which samples in the India collection belong to the Gujarati ethnic group. We analyzed 580 autosomal STR loci. This comprises the largest set of loci for which all members of all available samples from the CEPH diversity panel and the Gujarati were genotyped using the same amplification primers. The 46 STR genotyped populations include the Biaka, Yoruba, Kenya (Bantu speaking), Moscow, Gujarati, Han, and Japanese. Our DNA sequencing sample includes these seven populations, plus Iberians for whom there are no STR data. For some analyses, we focus on the STR data for the seven shared populations augmented by the French for whom there are STR data. We consider the French to be a useful proxy for Iberians on the intercontinental geographic scale considered in our analyses. Supporting Information Table 2 identifies the STR loci and gives sample sizes and allele frequencies for all 46 populations (including Gujarati).

Statistical analysis

Descriptive statistics. Our basic measure for analysis for DNA sequences is nucleotide diversity, which Nei defines as the number of differences per site between two copies of a locus (Nei, 1987). We speak of the nucleotide diversity within a population when both copies of the locus derive from the same local population. Similarly, we speak of the nucleotide diversity between two populations when the two copies derive from different local populations. For STR loci, we use gene diversity as our basic unit of analysis. Nei defines gene diversity as the probability that two randomly drawn copies of a locus differ in state, that is, are different alleles (Nei, 1987). We compute gene diversity within and between populations using definitions that parallel the definitions for nucleotide diversity.

For comparing populations, we constructed matrices denoted ${}_{\mathrm{R}}\hat{\Pi}$ and ${}_{\mathrm{R}}\hat{\mathrm{H}}$ with rows and columns equal to the number of local populations. The matrix ${}_{\mathrm{R}}\hat{\Pi}$ contains unbiased estimates of average nucleotide diversity and the matrix ${}_{\mathrm{R}}\hat{\mathrm{H}}$ contains unbiased estimates of average gene diversity. Each diagonal element of a diversity matrix contains the estimated diversity between sequences within the i^{th} population, and each off-diagonal element of a diversity matrix contains the diversity between sequences, the first from population i and the second from population j. The left-subscript R on these matrices indicates that they contain "raw" estimates that we make directly from the data without using a model of population relationships.

To investigate clustering of populations we convert the elements of ${}_{\mathrm{R}}\hat{\Pi}$ and ${}_{\mathrm{R}}\hat{\mathrm{H}}$ into genetic distance measures defined by Nei (Nei, 1987). For nucleotide sequences, we use the distance statistic $d_{\mathrm{A}} = \pi_{ij} - (\pi_{ii} + \pi_{jj})/2$ and for STRs we use the distance statistic $D_{\mathrm{m}} = h_{ij} - (h_{ii} + h_{jj})/2$.

Models and model fitting. Our strategy consists of fitting hierarchical models to our nucleotide diversity and STR gene diversity matrices. We use tree diagrams and terminology to display our models and explain our results. For brevity, we use the general term diversity when describing the statistical steps because the model fitting procedure is the same for nucleotide diversity and STR gene diversity. We fit two hierarchical models to the diversity matrices. Model 1 embodies two levels of stratification. At the first level, an allele exists within a local population. At the second level, a local population exists within a geographic region. This model places three constraints on the diversity within and between populations. i) The expected diversity is the same within all local populations, regardless of region. ii) The expected diversity is same between all populations within the same region, regardless of region. iii) The expected diversity is same between all populations in different regions, regardless of which regions the populations reside. Hereafter, we refer to Model 1 as the two-level island model, and denote it TLIM. The TLIM is the model of population structure used by Lewontin in his apportionment of diversity paper (Lewontin, 1972). We note that Lewontin used the word race to denote the level of population structure that we call geographic region. Model 2 embodies multiple levels of stratification. We construct Model 2 from a neighbor-joining tree (Saitou and Nei, 1987) calculated from the nucleotide sequence or STR genetic distance matrix and then rooted at the position that maximizes the likelihood of the tree after restricting all branches to non-negative values. This model places no constraints of equality on the diversity coefficients at different positions in a tree. We call Model 2 the expanded hierarchical model, and denote it EHM.

We use the system of equations developed by Anderson to fit our models to the data (Anderson, 1973). This procedure provides approximate maximum likelihood solutions. Several papers give more details on the application of this system of equations to genetic data (Cavalli-Sforza and Piazza, 1975; Urbanek et al., 1996; Lewis and Long, 2008). Ultimately, the method produces a new estimate of a nucleotide diversity or gene diversity matrix that is contingent on the hierarchical model. We denote the model-based diversity matrices by ${}_{\mathrm{M}}\hat{\Pi}$ and ${}_{\mathrm{M}}\hat{\mathrm{H}}$, where the left-subscript M indicates that the estimate is contingent on the model of population relationships.

Fixation indices. The TLIM estimates the following three coefficients of nucleotide diversity: π_{W} within local populations, π_{B} between local populations in the same geographic region, and π_{T} for local populations in different geographic regions (or equivalently, the total population). Three fixation indices completely summarize diversity according to the TLIM: $F_{\mathrm{SR}} = (\pi_{\mathrm{R}} - \pi_{\mathrm{w}})/\pi_{\mathrm{R}}$, $F_{\mathrm{RT}} = (\pi_{\mathrm{T}} - \pi_{\mathrm{R}})/\pi_{\mathrm{T}}$, and $F_{\mathrm{ST}} = (\pi_{\mathrm{T}} - \pi_{\mathrm{W}})/\pi_{\mathrm{T}}$, where F_{SR} quantifies diversity among local populations in the same region, F_{RT} quantifies diversity among regions, and F_{ST} quantifies diversity among local populations in different regions. These are the same fixation indices used by others (Excoffier et al., 1992; Hudson et al., 1992). The TLIM provides parallel results for STR gene diversity. The equations are the same, except that gene diversity values (h) replace nucleotide diversity (π) values. Although, our estimation procedure is slightly different, our gene diversity and fixation index parameters are those defined by Weir and Cockerham (Weir and Cockerham, 1984). The EHM estimates diversity coefficients at all internal and external nodes of the tree. This situation does not lend itself to a simple summarization of diversity at different levels of the hierarchy by a few fixation indices, but it is instructive to compare the diversity in each local population relative to the total diversity, which is estimated by the diversity between populations that span the root of the tree. We use the method of Long and Kittles (Long and Kittles, 2003) to calculate the population-specific F_{ST} coefficients. Finally, because fixation indices are inversely correlated with the diversity within local populations, we standardize our estimated fixation indices relative to their theoretical maximum values (Hedrick, 1999; Long and Kittles, 2003). The standardized values facilitate comparing fixation indices calculated from DNA sequences with those calculated from STRs.

Testing goodness-of-fit. We use Cavalli-Sforza and Piazza's treeness statistic (Cavalli-Sforza and Piazza, 1975) to test the fundamental hypothesis that the observed matrix ${}_{\mathrm{R}}\hat{\Pi}$ or ${}_{\mathrm{R}}\hat{\mathrm{H}}$ deviates from the model estimated matrix ${}_{\mathrm{M}}\hat{\Pi}$ or ${}_{\mathrm{M}}\hat{\mathrm{H}}$ by no more than would be expected from genetic and statistical sampling. The treeness statistic (denoted, Λ) approaches a chi-squared distribution under the ideal circumstances of a large number of independent alleles at different loci, each with equal heterozygosity. The number of degrees of freedom associated with this test is $r(r+1)/2 - p$, where r is the number of populations analyzed and p is the number of parameters in the model. We note that it is unusual to achieve the asymptotic properties of the test. The probable consequence of violating these assumptions is to reject the null hypothesis falsely more often than the chosen type I error rate, α. In addition to the problem of violating assumptions, we note that the test may reject a model because of minor deviations from treeness even though the tree predicts the data reasonably well. In light of these considerations, we take several measures to assess how well a hypothesized tree fits to a data set.

We begin our assessment of model fit by comparing observed and expected values. As noted above, the model fitting procedure creates an estimate of a diversity matrix contingent on the tree model. This enables us to plot the raw diversity estimates against the model-generated diversity estimates. We are also able to plot the raw genetic distance estimates against the model-generated genetic distance estimates.

There are also formal methods for comparing the treeness between different models that have been fitted to the same diversity matrix. We use two of these methods. To implement the first method of model comparison, we compute Akiake's Information Criterion (AIC) for each of the competing models (Akaike, 1974). With respect to our tree models, AIC = $\Lambda + 2p$, where Λ is the treeness statistic and p is the number of parameters in the model. Then we compute $\Delta = \text{AIC}_1 - \text{AIC}_2$, where AIC_2 denotes the lower of the two AIC values. Burnham and Anderson (Burnham and Anderson, 2002) give the following guidelines for interpreting Δ: for $\Delta \leq 2$ there is nearly equal support for both models, for $2 < \Delta < 10$ there is less support for the worse fitting model, for $\Delta \geq 10$ there is virtually no support for the worse fitting model. The second method of model comparison uses an F-test that is valid when the test statistics for both models deviate from the chi-squared distribution by a multiplier, and should be conservative when the chi-squared approximation to the worse fitting model is worse than the chi-squared approximation to the better fitting model. For this test, $F = (\Lambda_1/\mathrm{d}f_1)/(\Lambda_2/\mathrm{d}f_2)$, where the subscript 2 denotes the better fitting model (Lewis and Long, 2008). Both Δ and F take into account differences in the number of parameters between the competing models. Neither test is likely to favor a model simply because the model is parameter rich.

RESULTS

DNA sequence descriptive results

The 63 loci range in size from 436 to 985 base pairs (bp), with a median of 604 bp. In total, we sequenced 38,534 bp for each individual. All of our data are publicly available in GenBank. We found 204 sites that carried alternative alleles defined by a nucleotide substitution. At 80 of the substitution sites, the minor allele occurred as a singleton. At 26 of the substitution sites, the minor allele occurred as a doubleton, and at 12 sites, the minor allele occurred as a tripleton. According to Fu's neutral theory for stable Wright-Fisher populations (Fu, 1995), there should be about twice as many singletons as doubletons, and thrice as many singletons as tripletons. The clear excess of singletons is a signature of population growth (Yu et al., 2002). At the remaining 86 substitution sites, we observed more than three copies of the minor allele.

The Sub-Saharan African populations harbor the most diversity. This is easy to see from the allele frequency spectrum. To illustrate this, we combined the Moscow, Gujarati, and Han samples into a Eurasian group, and contrasted it with Sub-Saharan Africans, as represented by the combined Biaka, Yoruba, and Luhya. This results in samples such that Eurasians and Sub-Saharan-Africans are both represented by $n = 12$ individuals (see Fig. 1). In these sequences, 194 sites are variable. For the 113 variable sites with rare alleles (frequencies less than 3/48 in the two groups combined), 73 minor alleles are found only in the Sub-Saharan African populations, 37 are found only in Eurasians, and only three are shared between the two groups. For the 44 variable sites with moderate frequency minor alleles (frequencies between 4/48 and 8/48 in the two groups combined), 21 are found only in the Sub-Saharan African populations, seven are found only in Eurasians, and 16 are shared between the two groups. For the 40 variable sites with common minor alleles (frequencies between 9/48 and 24/48 in the two groups combined), just one appears only in the Sub-Saharan Africans, just one appears only in Eurasians, and 38 appear in both groups. These data make three things clear. First, Sub-Saharan Africans harbor many more rare variants than do Eurasians. Second, the chance that a moderate frequency allele that appears in Eurasians also appears in Sub-Saharan Africans (P = 16/(7 + 16) = 0.70) is substantially greater than the chance that a moderate frequency allele that appears in Sub-Saharan Africans also appears in Eurasians (P = 16/(21 + 16) = 0.43). Third, a common variant that appears in either Sub-Saharan Africans or Eurasians is likely to appear in the other group.

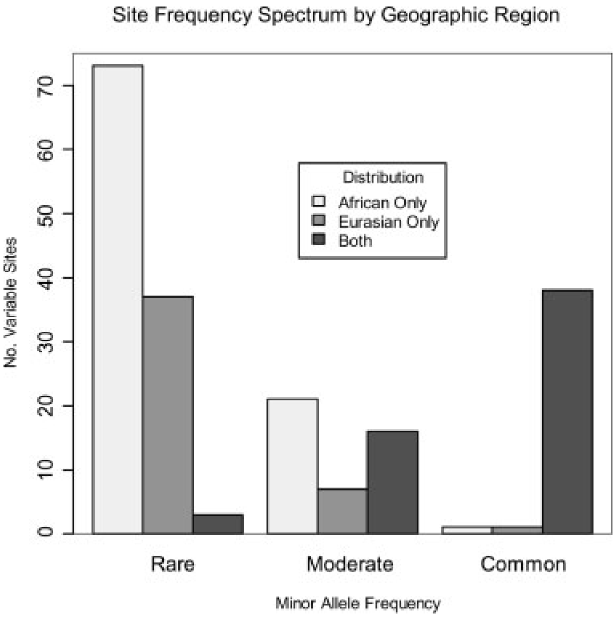

Fig. 1. Site frequency spectrum for $n = 12$ Sub-Saharan Africans and $n = 12$ Eurasians. Rare alleles have frequencies of 3/48, or less. Moderate frequency alleles have frequencies in the range 4/48 to 8/48. Common alleles have frequencies of 9/48 or greater.

Diversity matrices

Table 1 gives the estimated nucleotide diversity and genetic distances for the DNA sequence data set. Estimates of nucleotide diversity within populations appear on the major diagonal. The values in the upper triangle portion of Table 1 give the nucleotide diversity estimates between all pairs of populations. The values in the lower triangle of Table 1 give the genetic distance estimates between all pairs of populations. Supporting Information Table 3 gives the estimates of STR gene diversity and genetic distances for the complete 46-population data set.

Hierarchical models

Figure 2 presents the TLIM and EHM fit to the nucleotide diversity matrix. Table 2 shows that the data reject both models; however, both tests for comparing models confirm that the EHM fits significantly better fit than does the TLIM. The plot of raw- versus TLIM-based diversity coefficients in Figure 3A reveals the nature of lack-of-fit for the TLIM. Specifically, we see that the

TABLE 3. Diversity components and fixation indices from two-level island model

Level	100*π	Component	% Total	Index	Value	$F/F_{(\max)}$
DNA sequences						
Within population	0.070	0.070	84.1	F_{SR}	0.051	0.051
In same region	0.073	0.004	4.5	F_{RT}	0.114	0.114
Total population	0.083	0.009	11.4	F_{ST}	0.159	0.159
Level	H	Component	% Total	Index	Value	$F/F_{(\max)}$
580 STRs						
Within population	0.715	0.715	94.3	F_{SR}	0.027	0.093
In same region	0.735	0.020	2.6	F_{RT}	0.031	0.110
Total population	0.759	0.024	3.1	F_{ST}	0.057	0.201

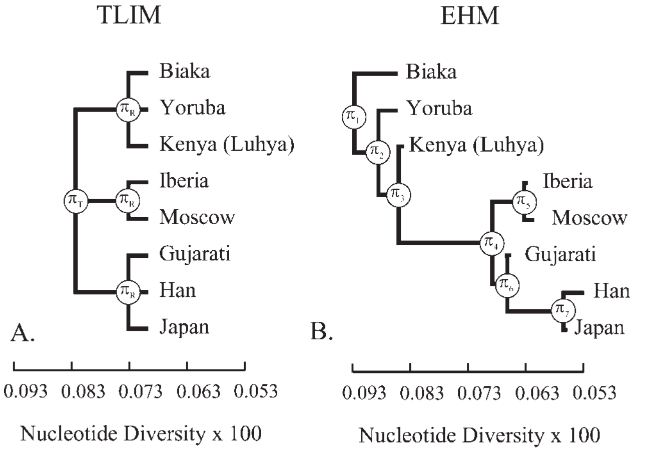

Fig. 2. Diagrams of hierarchical models fit to the DNA sequences. **A** and **B** are the TLIM and the EHM, respectively. Both graphs are calibrated to the nucleotide diversity scale below. Numerical values for the nodes and branch lengths appear in Tables 3 and 4. The circled symbols indicate the diversity coefficient estimated for the internal nodes. For the TLIM, the letter T denotes total population, and the letter R denotes region, and all external branches terminate at the within populations diversity. For the EHM each external branch terminates at diversity within a particular population.

TLIM i) underestimates diversity within African populations and overestimates diversity within European and Asia populations, ii) underestimates diversity between African populations and overestimates diversity between European and Asian populations, and iii) underestimates diversity between Africans and Asians, underestimates diversity between Africans and Europeans, and overestimates diversity between Europeans and Asians. The plot of raw genetic distances versus TLIM-generated genetic distances reveals an analogous pattern of discrepancy (Fig. 3C).

Comparison of the raw- versus EHM-based nucleotide diversity coefficients in Figure 3B reveals the improved fit of the EHM. We see a tight clustering of raw diversity coefficients about the EHM-based estimates. It is difficult to discern any particular pattern in the lack-of-fit. The plot of raw genetic distances versus EHM-generated genetic distances in Figure 3D reveals that the EHM predicts genetic distances extraordinarily well. The genetic distances are tightly clustered (r^2 = 0.94), although there is more dispersion among raw and EHM-generated genetic distances for larger distances than for smaller distances.

Figure 4 presents the TLIM and EHM fit to the STR gene diversity matrix. Table 2 shows that these data also reject both models; however, both tests for comparing models confirm that the EHM fits significantly better fit than does the TLIM. Figure 5 presents graphs of raw and model-generated statistics for the subset of eight populations that match the populations in the DNA sequence analysis (as noted, the French serve as proxies for the Iberians). The plot of raw- versus TLIM-based gene diversity coefficients in Figure 5A reveals the same nature of lack-of-fit for the TLIM that we observed in the analysis of DNA sequences. That is, the TLIM i) underestimates diversity within African populations and overestimates diversity within European and Asia populations, ii) underestimates diversity between African populations and overestimates diversity between European and Asian populations, and iii) underestimates diversity between Africans and Asians, underestimates diversity between Africans and Europeans, and overestimates diversity between Europeans and Asians. The plot of raw genetic distances versus TLIM-generated genetic

Fig. 3. Nucleotide diversity and genetic distance plots for the DNA sequences. **A** and **B** show "raw" nucleotide diversity coefficients (×100) plotted against the values predicted by the TLIM and EHM, respectively. **C** and **D** show "raw" genetic distance coefficients plotted against the values predicted by TLIM and EHM, respectively. For the TLIM, T denotes total population, R denotes region, and W denotes within. The TLIM values for both models are jittered to avoid eclipsing. To simplify visually, we plot only the diversity coefficients for internal nodes of the EHM. Tables 3 and 4 give all numerical values for these plots. The key at the bottom gives the color-coding for all four panels. The circles denote model-generated expectations for nucleotide diversity coefficients.

Fig. 5. Gene diversity and genetic distance plots for the STRs. Gray circles present all points. To simplify visually, colored squares show values for the populations for which we obtained DNA sequences. **A** and **B** show "raw" gene diversity coefficients plotted against the values predicted by the TLIM and EHM, respectively. **C** and **D** show "raw" genetic distance coefficients plotted against the values predicted by TLIM and EHM, respectively. The TLIM values for both models are jittered to avoid eclipsing. For the TLIM, T denotes total population, R denotes region, and W denotes within. We plot only the diversity coefficients for internal nodes of the EHM in order to reduce the size of the graphs. Tables 3 and 4 give the numerical values for these plots. The key at the bottom gives the color-coding for all four panels. The circles denote model-generated expectations for gene diversity coefficients. Supporting Information Table 3 gives all raw and estimated gene diversity and genetic distance coefficients for the 46 population STR analyses.

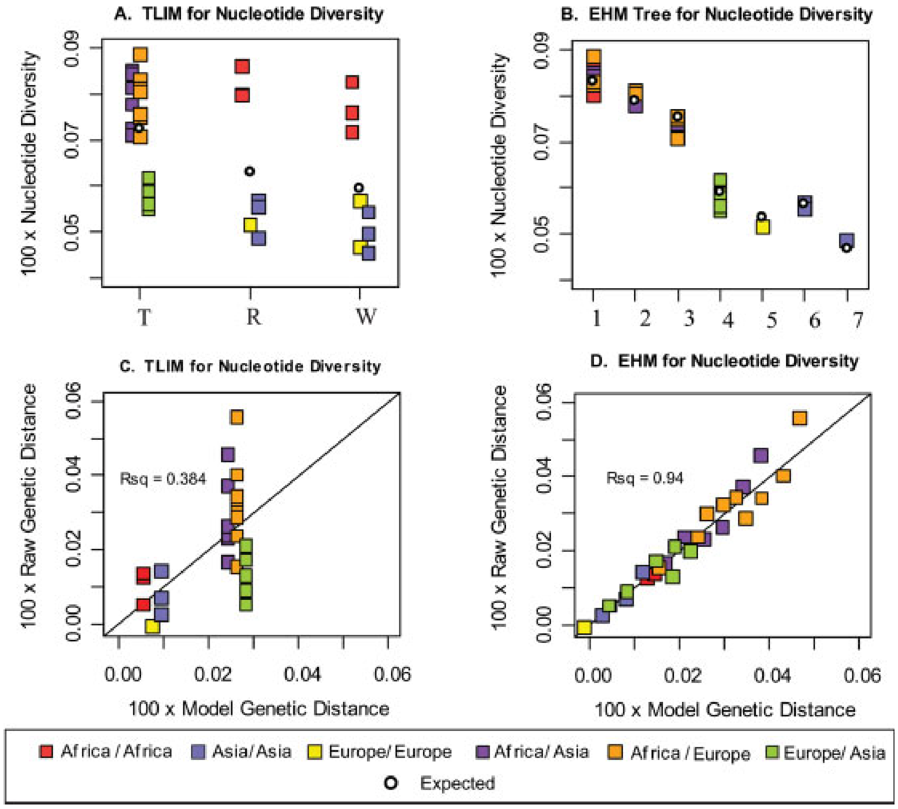
A. TLIM for Nucleotide Diversity
100 x Nucleotide Diversity
0.05
0.07
0.09
T
R
W
B. EHM Tree for Nucleotide Diversity
100 x Nucleotide Diversity
0.05
0.07
0.09
1
2
3
4
5
6
7
C. TLIM for Nucleotide Diversity
100 x Raw Genetic Distance
0.00
0.02
0.04
0.06
Rsq = 0.384
0.00
0.02
0.04
0.06
100 x Model Genetic Distance
D. EHM for Nucleotide Diversity
100 x Raw Genetic Distance
0.00
0.02
0.04
0.06
Rsq = 0.94
0.00
0.02
0.04
0.06
100 x Model Genetic Distance
Africa / Africa
Asia/Asia
Europe/Europe
Africa/Asia
Africa / Europe
Europe/Asia
Expected

Fig. 3.

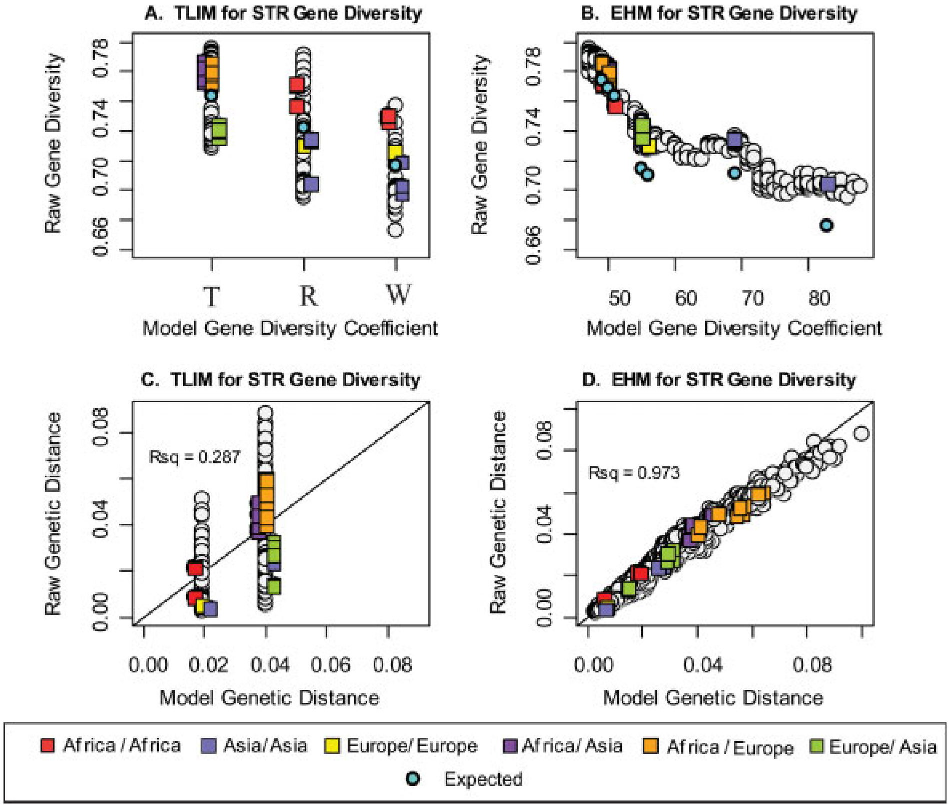
A. TLIM for STR Gene Diversity
Raw Gene Diversity
0.66
0.70
0.74
0.78
T
R
W
Model Gene Diversity Coefficient
B. EHM for STR Gene Diversity
Raw Gene Diversity
0.66
0.70
0.74
0.78
50
60
70
80
Model Gene Diversity Coefficient
C. TLIM for STR Gene Diversity
Raw Genetic Distance
0.00
0.04
0.08
Rsq = 0.287
0.00
0.02
0.04
0.06
0.08
Model Genetic Distance
D. EHM for STR Gene Diversity
Raw Genetic Distance
0.00
0.04
0.08
Rsq = 0.973
0.00
0.04
0.08
Model Genetic Distance
Africa / Africa
Asia/Asia
Europe/Europe
Africa/Asia
Africa / Europe
Europe/Asia
Expected

Fig. 5.

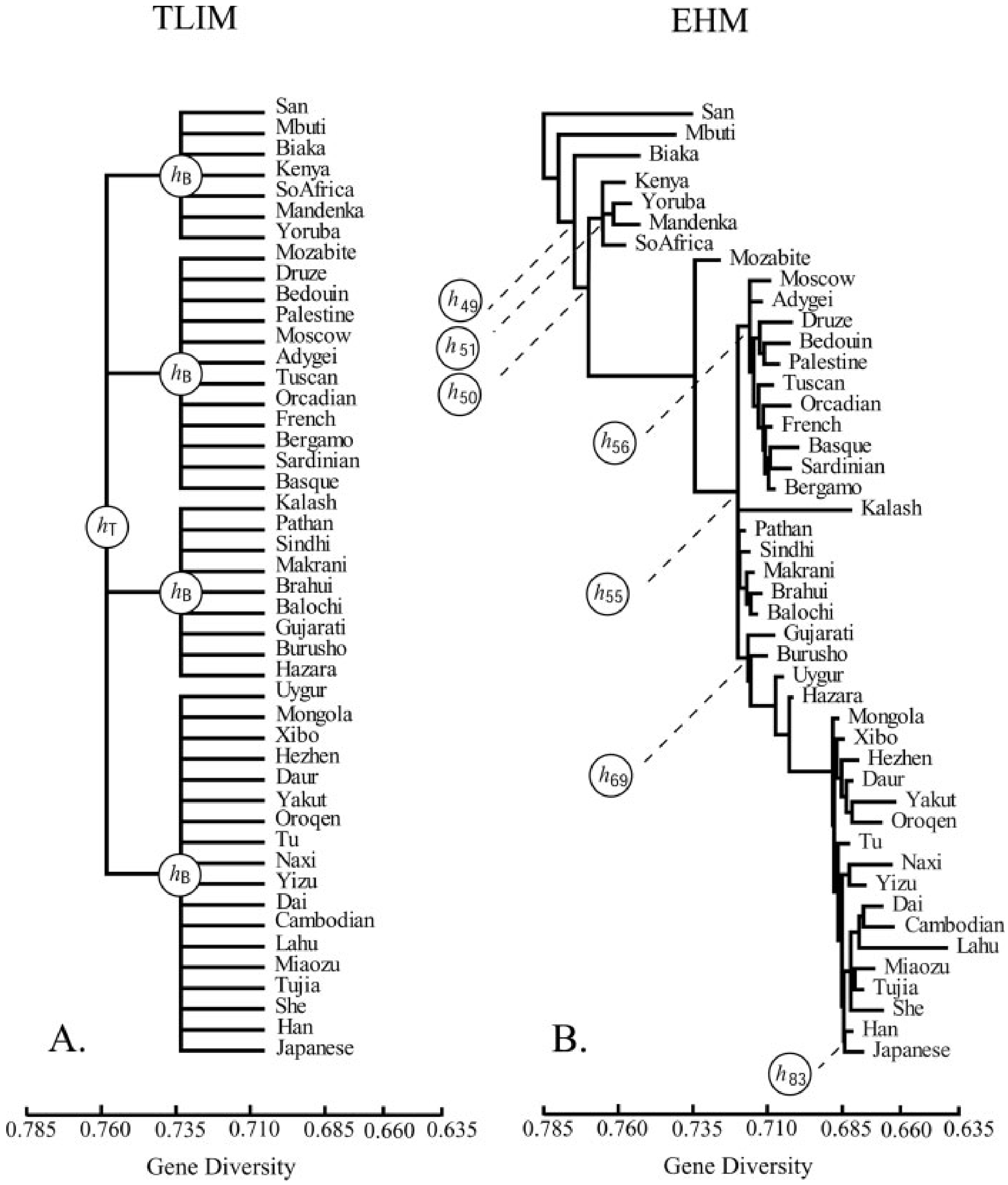

Fig. 4. Diagrams of hierarchical models fit to the STRs. **A** and **B** are the TLIM and the EHM, respectively. Both graphs are calibrated to the gene diversity scale below. Numerical values for selected nodes and branch lengths appear in Tables 3 and 4. The circled symbols indicate the diversity coefficient estimated for the internal nodes. For the TLIM, the letter T denotes total population, and the letter R denotes region, and all external branches terminate at the within populations diversity. For the EHM, we have labeled the nodes that link the eight populations from the DNA sequence analysis and each external branch terminates at diversity within a particular population. Supporting Information Table 3 has the complete results for all 46 populations.

distances reveals an analogous pattern of discrepancy (Fig. 5C).

The situation is more complicated for the EHM fit to the 46 population STR data. For the nodes corresponding to African populations, we see a close correspondence and tight clustering of raw diversity coefficients about the EHM-generated estimates. By contrast, the model overestimates diversity for the nodes pertaining to non-African populations. Nevertheless, Figure 5D shows that the raw and EHM-generated genetic distance estimates cluster tightly ($r^2 = 0.973$). The good fit of the raw and EHM-generated genetic distances is enigmatic because of the biased EHM-generated gene diversity estimates in non-Africans. Supporting Information Table 3 provides the complete model-generated gene diversity and genetic distance matrices for both the TLIM and EHM.

Although the measures nucleotide diversity and gene diversity are on vastly different numerical scales, Fig-

TABLE 4. Results from expanded hierarchical model[a]

Node	Ancestor[b]	100*π	Length[c]	% Total[d]	$F_{ST(k)}$	$F_{ST(k)}/F_{(max)}$
DNA sequences						
Biaka	1	0.086	0.007	92.2	0.078	0.086
Yoruba	2	0.086	0.003	92.3	0.077	0.085
Luhya	3	0.087	0.000	93.9	0.061	0.067
Iberian	5	0.066	0.000	71.2	0.288	0.308
Moscow	5	0.062	0.001	66.8	0.332	0.354
Gujarati	6	0.068	0.000	72.7	0.273	0.292
Han	7	0.054	0.003	57.5	0.425	0.449
Japanese	7	0.057	0.000	61.4	0.386	0.410
1		0.093	–	100.0	–	–
2	1	0.089	0.004	95.4	–	–
3	2	0.085	0.004	91.6	–	–
4	3	0.069	0.016	74.2	–	–
5	4	0.064	0.006	68.3	–	–
6	4	0.067	0.003	71.5	–	–
7	6	0.057	0.010	61.1	–	–
Node[e]	Ancestor[b]	h	Length[f]	% Total[d]	$F_{ST(k)}$	$F_{ST(k)}/F_{(max)}$
STRs						
Biaka	49	0.751	0.023	97.0	0.030	0.121
Kenya	51	0.756	0.008	97.6	0.024	0.097
Yoruba	51	0.756	0.008	97.7	0.023	0.096
French	56	0.702	0.008	90.7	0.093	0.312
Moscow	56	0.703	0.007	90.8	0.092	0.309
Gujarati	69	0.706	0.005	91.2	0.088	0.298
Han	83	0.673	0.003	87.0	0.130	0.399
Japanese	83	0.669	0.006	86.5	0.135	0.409
49	–	0.774	–	100.0	–	–
50	49	0.769	0.005	99.3	–	–
51	50	0.764	0.005	98.7	–	–
55	50	0.715	0.054	92.3	–	–
56	55	0.710	0.004	91.7	–	–
69	55	0.711	0.004	91.8	–	–
83	69	0.676	0.035	87.3	–	–

[a] Nodes are numbered as in Figures 2 and 4.
[b] Ancestor of node in first column.
[c] Length of branch from ancestor to node in first column.
[d] Diversity at node as a percent of estimated total diversity.
[e] Nodes selected from the analysis of 46 populations for comparison to the DNA sequence results.
[f] Total length summed over all branches between the given nodes.

ures 2A and 4A show that the two data sets provide qualitatively similar TLIM trees. The numerical values for the fixation indices F_{SR}, F_{RT}, and F_{ST} as obtained from the TLIM differ between the DNA sequence and STR data (Table 3). The DNA sequences provide higher fixation indices than do the STRs. However, based on Hedrick's and Long's derivations (Hedrick, 1999; Long and Kittles, 2003), we should expect this difference because diversity is low at the per nucleotide level, whereas it is high at the per STR locus level. Standardizing these fixation indices relative to their theoretical maxima reveals the qualitative similarity between the DNA sequence and STR results. $F_{ST}/F_{(max)} = 0.159$ for the DNA sequences and $F_{ST}/F_{(max)} = 0.201$ for the STRs. It is of interest that F_{ST} estimated from our DNA sequences (0.159) is nearly the same as the value (0.158) recently published for another DNA sequence dataset (Wall et al., 2008), and the fixation indices estimated from the STRs are close to previously published estimates for these data (Rosenberg et al., 2002).

Figures 2B and 4B show the qualitative similarity between the EHM trees for the two data sets. The standardized population-specific fixation indices (Table 4) illustrate this proportionality. For both data sets, the African populations show modest divergence from the basal node ($F_{ST(k)}$/max(F_{ST}) (0.10), the European populations and Gujarati show substantial divergence from the basal node ($F_{ST(k)}$/max(F_{ST}) (0.30), and the East Asian populations show great divergence from the basal node ($F_{ST(k)}$/max(F_{ST}) (0.40).

DISCUSSION

Despite the small sample size, this is the first study to collect complete DNA sequences from autosomal loci in a manner that makes it possible to evaluate diversity at more than one level of population hierarchy. In addition, we use a large publicly available data set that includes STR genotypes from 580 loci from many individuals from 46 populations to confirm and evaluate patterns of diversity in our DNA sequence data.

We began by fitting a simple TLIM that embodies a partition of diversity into components attributable to within populations, between populations within regions, and between different regions (Figs. 2A and 4A). Using this model, we find a higher percentage of diversity between regions for DNA sequences than for STR polymorphisms. This difference is likely due to the scale effect that results from the fact that polymorphism is low in DNA sequences but high in STRs (Hedrick, 1999;

Long and Kittles, 2003). The similarity between the normalized fixation indices from DNA sequences and STRs confirms this interpretation (Tables 3 and 4). A more important finding is that the TLIM produces biases that are consistent between the DNA sequences and STRs (Figs. 3 and 5).

We next turned to models with multiple levels of nesting and variable branch lengths (Figs. 2B and 4B). We found that the DNA sequences and STRs converge on similar topological relationships and similar patterns of diversity within and between populations. For both the DNA sequences and STRs, the treeness statistics show that the EHM fits the data substantially better than does the TLIM (Table 2). A summary of the major differences between the TLIM and the EHM follows. First, the TLIM estimates less diversity for the species as a whole than does the EHM. Second, the TLIM estimates an excess diversity within non-Sub-Saharan African populations, but it estimates a deficit of diversity within Sub-Saharan African populations. Third, the TLIM forces all continental populations to diverge equally from the deepest node, whereas the expanded hierarchy splits the Biaka from all other populations at the deepest node. Fourth, in the TLIM, the European and Asian populations diverge from African populations independently, but in the EHM, the European and East Asian population diverge together from African populations.

We see that both data types (DNA sequences and STRs) reject both population structure models (TLIM and EHM). Nevertheless, the EHM does a significantly better job at representing the data. Two statistical methods confirm the superiority of the EHM (Table 2), but our comparisons of observed and model-based diversity and genetic distance values demonstrate the ways in which the EHM out-performs the TLIM (Figs. 3 and 5). It is of note that both the DNA sequences and STRs show the same pattern of improvements with the EHM over the TLIM. An aphorism from the statistician George Box (Box and Draper, 1987) puts our results into perspective—*Essentially, all models are wrong, but some are useful*.

It is interesting to see how our findings compare with those from other recent studies of human variability in DNA. A recent study of 640,000 SNPs genotyped in a largely overlapping data set reproduces our key finding that the deepest node of the hierarchy splits the Biaka from all other populations (Li et al., 2008). In fact, the SNP data produces a tree with the same major topological relationships that we find using the 46-population STR data set. As we presented earlier, there is considerable overlap between the loci that we have sequenced and those that Yu and colleagues have sequenced (Yu et al., 2002). The deep population structure that we find in Sub-Saharan Africa explains their seemingly impossible finding that nucleotide diversity is greater between two African DNA sequences than between an African DNA sequence and a non-African DNA sequence. The reason for their result is that, in their sample, the percentage of DNA sequence comparisons that spanned the deepest African nodes was higher for African x African DNA sequence comparisons than for African x non-African sequence comparisons. We should not expect to see Yu and colleagues' exact findings reproduced in other studies because the apparent diversity within African samples will be a complicated function of which groups contribute to the sample, and how many individuals represent each group. We should also like to point out that our DNA sequencing findings closely parallel three findings from a recent study of the SNP allele frequency spectrum (Keinan et al., 2007). First, East Asian and Northern European ancestors shared the same population bottleneck in their migration Out of Africa. Second, both East Asians and Northern Europeans have drifted independently after the Out of Africa migration bottleneck. Third, East Asians have drifted more since the bottleneck than Europeans. It is easy to read all three of these findings from branch sequence and branch lengths in the graphs of our EHM applied to both data sets.

We also wish to point out that despite the fact that there are large differences in the level of diversity within populations; every population sampled harbors ample diversity such that every copy of the genome is unique, and every individual is unique (excepting identical twins). Every person is likely to be heterozygous at millions of nucleotide positions when the whole genome is considered. With this level of variability, population membership is unlikely to be a precise indicator of an individual's genotype at any particular nucleotide position.

We now turn to how these estimates and analyses of genetic diversity within and between populations effect the assessment of human races in our species. Lewontin's argument against race is historically important and interesting (Lewontin, 1972). He was not the first to argue against race taxonomy using genetics, but his argument was unique. He confronted race by trying to show that classical racial groupings account for too little of the total diversity to be worth further concern. Our results show that race, as represented in the TLIM, fits both data sets poorly. Comparisons between raw and model-generated diversity and genetic distance estimates reveal that the TLIM indeed misrepresents both the pattern and amount of diversity within and between populations. A strong message from our findings is that the model used in an analysis biases the outcome measurements. We agree entirely with Lewontin that classical race taxonomy is a poor reflection of human diversity. However, we do not believe that the diversity components that he estimated using this model reflect an intrinsic property of human genetic structure as some scientists have suggested (Templeton, 1999, 2007; Brown and Armelagos, 2001).

The pattern of DNA sequence diversity also creates some unsettling problems for applying to humans the definition of races as groups of populations within which the individuals are more related to each other than they are to members of other such groups (Hartl and Clark, 1997). This definition essentially encompasses Templeton's evolutionary lineage definition of race (Templeton, 1999) and Dobzhansky's gene frequency definition of race (Dobzhansky, 1970). Although it is logically consistent to group populations by relationship, the nested pattern of genetic diversity in the EHM disagrees with the traditional anthropological classifications that placed continental populations at the same level of classification (i.e., race). A classification that takes into account evolutionary relationships and the nested pattern of diversity would require that Sub-Saharan Africans are not a race because the most exclusive group that includes all Sub-Saharan African populations also includes every non-Sub-Saharan African population (Figs. 2B and 4B). Moreover, the Out-of-Africa branch would place all Eurasians in the same race, but this would necessitate placing Europeans and Asians in sub-races. Several sub-sub-races would be necessary to account for the population groups throughout the world. We see no need for such a classification in light of the fact that our evolutionary history gives good guidance for understanding the structure of human diversity.

Some biologists define races based purely on correct assignment of individuals to groups. The best known version of this approach is the seventy-five percent correct classification rule (Amadon, 1949; Mayr, 1969). Edwards has explained how accurate classification will be achieved when multiple polymorphic loci are considered (Edwards, 2003), and we see empirically that there are applications to human data that satisfy the seventy-five percent criterion (Rosenberg et al., 2002; Bamshad et al., 2003). However, the clustering methods in popular use produce human population groups that have a simpler structure than even the TLIM (Pritchard et al., 2000; Falush et al., 2003). This structure is clearly a weak description of the true human population structure, because it does not capture the complete nested arrangement of populations. We do not expect that such a classification will serve any application better than the full nested structure of populations.

In summary, we find for our own data and for a large published data set, that human populations have much diversity when DNA sequences are considered. We show that simple partitions of diversity are biased and that they hide the true extent of diversity. The pattern of diversity that we reveal is richer and worthy of study as it sheds light on the peopling of the world, ancestry and natural selection, and disease patterns (Ramachandran et al., 2005; Rosenberg et al., 2005; Lohmueller et al., 2008).

ACKNOWLEDGMENTS

The authors thank Cris Erin Hughes, Keith Hunley, Kaanan Shah, and Jennifer Wagner for comments on earlier drafts of this article. We thank Pragna Patel (University of Southern California) and Noah Rosenberg (University of Michigan) for information about which individuals in the Marshfield India collection belong to the Gujarati ethnic group. We thank Wen-Hsung Li (University of Chicago) for sending us information about laboratory methods for the loci that his group has sequenced. The authors claim full responsibility for all interpretations and any errors of omission or commission.

LITERATURE CITED

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans Automatic Control AC 19:716–723.

Amadon C. 1949. The seventy-five per cent rule for subspecies. Condor 51:250–258.

Anderson TW. 1973. Asymptotically efficient estimation of covariance matrices with linear structure. Ann Stat 1:79–95.

Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB. 2003. Human population genetic structure and inference of group membership. Am J Hum Genet 72:578–589.

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. 1997. An apportionment of human DNA diversity. Proc Natl Acad Sci USA 94:4516–4519.

Box GEP, Draper NR. 1987. Empirical model-building and response surfaces. New York: Wiley.

Brown RA, Armelagos GJ. 2001. Apportionment of racial diversity: a review. Evol Anthropol 10:34–40.

Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, Mountain JL, Perez-Stable EJ, Sheppard D, Risch N. 2003. The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med 348:1170–1175.

Burnham KP, Anderson DR. 2002. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer.

Cavalli-Sforza LL, Piazza A. 1975. Analysis of evolution: evolutionary rates, independence and treeness. Theor Popul Biol 8:127–165.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496–1502.

Cooper RS, Kaufman JS, Ward R. 2003. Race and genomics. N Engl J Med 348:1166–1170.

Dobzhansky T. 1970. Genetics of the evolutionary process. New York: Columbia University Press.

Edwards AW. 2003. Human genetic diversity: Lewontin's fallacy. Bioessays 25:798–801.

Excoffier L, Hamilton G. 2003. Comment on "Genetic structure of human populations." Science 300:1877; author reply 1877.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587.

Fischer A, Pollack J, Thalmann O, Nickel B, Paabo S. 2006. Demographic history and genetic differentiation in apes. Curr Biol 16:1133–1138.

Fu YX. 1995. Statistical properties of segregating sites. Theor Popul Biol 48:172–197.

Hartl DL, Clark AG. 1997. Principles of population genetics. Sunderland, MA: Sinauer and Assoc.

Hedrick PW. 1999. Highly variable loci and their interpretation in evolution and conservation. Evolution 53:313–318.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583–589.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979–988.

Keinan A, Mullikin JC, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet 39:1251–1255.

Lewis CM Jr, Long JC. 2008. Native South American genetic structure and prehistory inferred from hierarchical modeling of mtDNA. Mol Biol Evol 25:478–486.

Lewontin RC. 1972. The apportionment of human diversity. Evol Biol 6:381–398.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, Bustamante CD. 2008. Proportionally more deleterious genetic variation in European than in African populations. Nature 451:994–997.

Long JC, Kittles RA. 2003. Human genetic diversity and the nonexistence of biological races. Hum Biol 75:449–471.

Mayr E. 1969. Principles of systematic zoology. New York: McGraw-Hill.

Michalakis Y, Excoffier L. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061–1064.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA 102:15942–15947.

Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G. 2002. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res 12:602–612.

Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MG, Nino-Rosales L, Ninis V, Das P, Hegde M, Molinari L, Zapata G, Weber JL, Belmont JW, Patel PI. 2006. Low levels of genetic divergence across geographically and linguistically diverse populations from India. PLoS Genet 2:e215.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1:e70.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. Science 298:2381–2385.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425.

Templeton AR. 1999. Human races: a genetic and evolutionary perspective. Am Anthropol 100:632–650.

Templeton AR. 2007. Genetics and recent human evolution. Evol Int J Org Evol 61:1507–1519.

Urbanek M, Goldman D, Long JC. 1996. The apportionment of dinucleotide repeat diversity in Native Americans and Europeans: a new approach to measuring gene identity reveals asymmetric patterns of divergence. Mol Biol Evol 13:943–953.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. 2001. The sequence of the human genome. Science 291:1304–1351.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci USA 102:18508–18513.

Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, Hammer MF. 2008. A novel DNA sequence database for analyzing human demographic history. Genome Res 18:1354–1361.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.

Yu N, Chen FC, Ota S, Jorde LB, Pamilo P, Patthy L, Ramsay M, Jenkins T, Shyue SK, Li WH. 2002. Larger genetic differences within Africans than between Africans and Eurasians. Genetics 161:269–274.