AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 139:35–46 (2009)

The Global Pattern of Gene Identity Variation Reveals a History of Long-Range Migrations, Bottlenecks, and Local Mate Exchange: Implications for Biological Race

Keith L. Hunley,[1]* Meghan E. Healy,[1] and Jeffrey C. Long[2]

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

KEY WORDS biological race; coalescent simulations; gene identity

ABSTRACT Several recent studies have argued that human genetic variation conforms to a model of isolation by distance, whereas others see a predominant role for long-range migrations and bottlenecks. It is unclear whether either of these views fully describes the global pattern of human genetic variation. In this article, we use a coalescent-based simulation approach to compare the pattern of neutral genetic variation predicted by these views to the observed pattern estimated from neutral autosomal microsatellites assayed in 1,032 individuals from 53 globally-distributed populations. We find that neither view predicts every aspect of the observed pattern of variation on its own, but that a combination of the two does. Specifically, we demonstrate that the observed pattern of global gene identity variation is consistent with a history of serial population fissions, bottlenecks and long-range migrations associated with the peopling of major geographic regions, and gene flow between local populations. This history has produced a nested pattern of genetic structure that is inconsistent with the existence of independently evolving biological races. We consider the implications of our findings for methods that apportion variation into within- and between-group components and for medical genetics. Am J Phys Anthropol 139:35–46, 2009.

Frank Livingstone was one of the first anthropologists to argue in favor of abandoning the use of race in the study of human variation. In an exchange with Dobzhansky in 1962, he argued that while the species might be divided into subspecies (races) based on the frequency of one genetic character, different characters produced different and contradictory divisions. As a result, after more than a few characters were considered, there would be no grouping larger than a population.

Ten years later, Lewontin (1972) also rejected the use of race in the study of human variation. His rejection was based on the finding that differences between populations accounted for less than 15% of the total variation measured from 16 blood group proteins, serum proteins, and red blood cell enzymes. Moreover, the portion of variation between populations within seven operationally defined races was only 8.3% and the portion between the races was only 6.3%. He concluded that the amount of variation between races was so small that there was "virtually no genetic or taxonomic significance" to human race (1972, p 397). Subsequent studies of neutral genetic variation have replicated these seminal results (Nei and Roychoudhury, 1982; Barbujani et al., 1997; Jorde et al., 2000; Rosenberg et al., 2002), and the findings have been a key factor in the rejection of race in anthropology (Brown and Armelagos, 2001). Interestingly, all of these studies show that the portion of variation between groups is statistically significant, but concur with Lewontin that the percentage is too low for taxonomic significance.

More recent studies of human biological variation have turned to analysis of the geographic pattern of neutral genetic variation (Serre and Pääbo, 2004; Manica et al., 2005; Prugnolle et al., 2005; Ramachandran et al., 2005; Handley et al., 2007). The studies identified a strong positive correlation between global genetic and geographic distances. The correlation has been interpreted in several of these studies as being consistent with a model of isolation by distance in which there are no major geographic discontinuities in the pattern of neutral genetic variation. Serre and Pääbo (2004), for example, concluded that the pattern of global genetic variation was better described by clines than by race.

Today, anti-race views dominate instruction in biological anthropology (Jurmain et al., 2005; Boyd and Silk, 2006; Fuentes, 2007). However, there are potential problems with each view that call into question their validity. First, Livingstone provided no empirical evidence for his position. In the 1962 exchange, Dobzhansky agreed with Livingstone that discordant gene frequency patterns were common among human populations but also recognized that physical and social impediments to gene flow had regularly produced larger discontinuities and concordant allele frequency patterns between human groups. It is also possible that the opposing clines envisioned by Livingstone might form in a purely phyletic process through random character loss.

*Correspondence to: Keith L. Hunley, Department of Anthropology, MSC01-1040, Anthropology, 1 University of New Mexico, Albuquerque, NM 87131, USA. E-mail: khunley@unm.edu

Received 26 March 2008; accepted 6 August 2008

DOI 10.1002/ajpa.20932
Published online 18 February 2009 in Wiley InterScience (www.interscience.wiley.com).

Second, though any cutoff is arbitrary, it seems reasonable that if only 6.3% of the variation at neutral genetic loci is apportioned between races, then they have no taxonomic significance. However, Long and Kittles (2003) showed that when chimpanzees were added to a sample of human populations from five major geographic regions, the between-group portion of genetic variation increased only marginally from 11.9 to 18.3%. This finding calls into question genetic and taxonomic interpretations of the low values reported by Lewontin and others. Further, apportionments of variation are meaningful only if there is 1) evolutionary independence of populations within the same geographic region, 2) evolutionary independence between groups of populations in different geographic regions, and 3) levels of variation are the same in all local populations and in all geographic regions. Long and Kittles showed that human populations do not conform to this pattern of variation.

Third, Ramachandran et al. (2005) argued that the strong global correlation between genetic and geographic distances is not the result of isolation by distance, but is instead the outcome of a process of nested serial population fissions, migration, and isolation. This distinction between isolation by distance and the "serial fissions" model of Ramachandran et al. is important because isolation by distance is inconsistent with the existence of taxonomic units, while serial fissions is not.

It is important to address these issues for several reasons. First and foremost, racial classification and its attendant behavioral stereotypes have social consequence, which is why many anthropologists devote so much attention to debunking the typological views of race which some scientists and much of the lay public still apply. Second, knowledge of the pattern of global genetic variation and its evolutionary processes are essential for reconstructing the evolutionary history of our species (Harpending and Rogers, 2000; Excoffier, 2002). Third, knowledge of this pattern and process are crucial for identifying the genetic basis of multifactorial disease and potentially for the development of effective disease treatments (Pritchard, 2001; Reich and Goldstein, 2001; Wilson et al., 2001; Risch et al., 2002; Burchard et al., 2003).

To address these issues, we use a coalescent-based simulation approach to compare the pattern of neutral genetic variation predicted by several of the views outlined above to the observed pattern estimated from neutral autosomal microsatellites assayed in 1,032 individuals from 53 globally-distributed populations. We find that none of the views predict every aspect of the observed pattern of variation on their own, but that a combination of several of them does. We also observe a nested pattern of neutral genetic variation at the level of geographic regions and conclude that this nested structure makes it impossible to classify major population groups into categories at the same taxonomic level.

MATERIALS AND METHODS

Our methodological approach consisted of the following steps. We first used computer simulations to identify the global geographic pattern of neutral genetic variation predicted by several of the views of human variation described above, i.e., independent evolution of regions, isolation by distance, and serial fissions. This "expected" geographic pattern of neutral genetic variation was examined visually using several simple graphical methods. A simple visual comparative approach was chosen because the different models make very different predictions about the global pattern of gene identity variation at different hierarchical levels of population structure. The same graphical methods were also used to examine the "observed" pattern of variation estimated directly from neutral genetic data collected from a large sample of globally distributed human populations. Each simulated pattern was then compared to the observed pattern to determine which, if any, of the models best captured the observed pattern of human genetic variation.

Genetic data

The genetic data consist of the allele sizes for 783 autosomal microsatellite loci that were typed in 1,032 individuals from 53 globally distributed local populations located in seven geographic regions (Cann et al., 2002; Rosenberg et al., 2005). The seven regions are: Africa, the Middle East, Europe, South Central (SC) Asia, Oceania, East Asia, and the Americas. The population locations are shown in Figure 1, and the sample sizes and geographic coordinates are listed in Table 1. In several analyses, we refer to the combined Middle Eastern, European and SC Asian regions as Western Eurasia.

Our basic unit of analysis is gene identity, which is the probability that two randomly drawn copies of a locus are identical by state. The two copies of the locus may be sampled from within the same local population, or they may be sampled from different local populations. An advantage to using gene identity is that each of the four models described below makes clear predictions about the geographic pattern of gene identity variation at the within- and between-population level. Additional advantages to using gene identity are that it provides a direct measure of within-population genetic variation and that it does not confound within- and between-population variation as some measures of genetic distance do (e.g., Nei's (1987) minimum genetic distance), which aids in the comparison of simulated and observed patterns of gene identity variation.

Models and predictions

The first three models were suggested by views of human variation outlined in the introduction. Each model makes clear predictions for gene identity at three different levels: within local populations, between populations within geographic regions, and between populations in different regions. The models are as follows:

i. Independent regions (see Fig. 2A): The independent regions model is consistent with Lewontin's interpretation of human variation. It assumes that geographic regions have been evolving independently and at equal rates. The model predicts that gene identities will be highest within local populations, lower between populations within regions, and lowest between populations in different regions. It also predicts that gene identity at each of these three levels will be uniform throughout the geographic range of the species.
ii. Isolation by distance (see Fig. 2B): Isolation by distance is the only model tested that is inconsistent with any taxonomic concept of race. It results from the tendency of individuals to choose mates from adjacent populations that are evenly distributed across

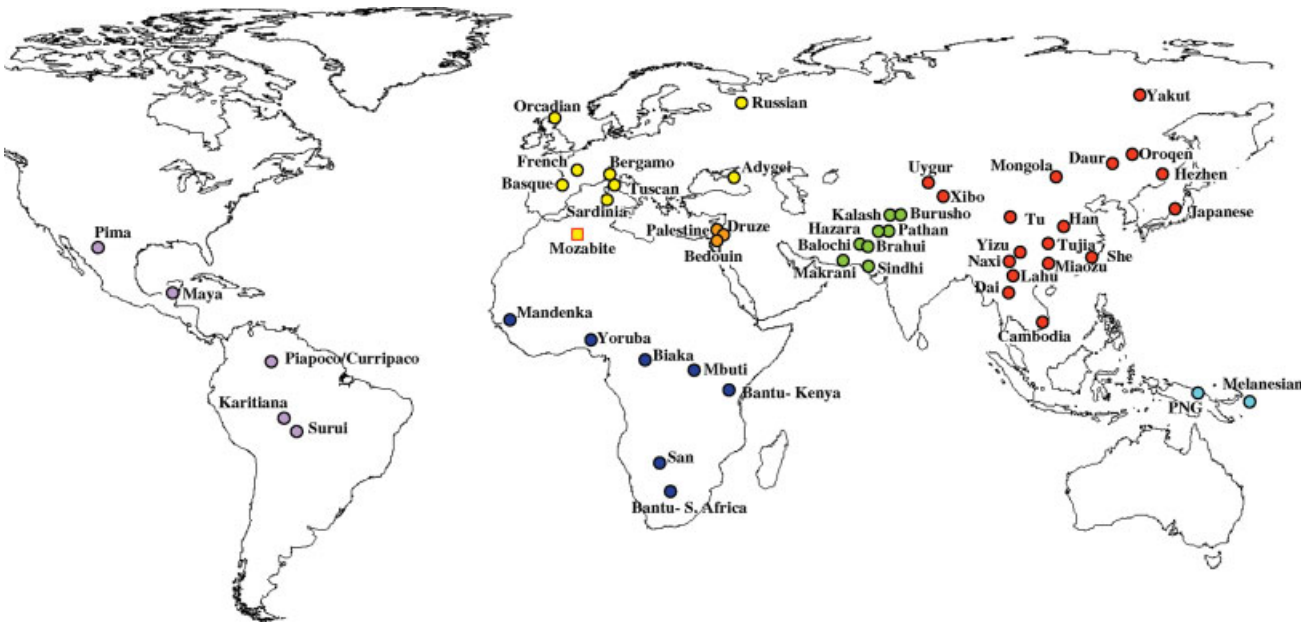

Fig. 1. CEPH population locations.

a landscape (Malécot, 1948; Kimura and Weiss, 1964). Under these conditions, individuals living in adjacent populations are likely to share more recent common ancestors than individuals living in more distant populations. This pattern of local mating leads to a monotonic decay in gene identity with increasing geographic distance between populations. Gene identities will be highest and uniform within local populations, and will decay monotonically with increasing geographic distance between populations.

iii. Serial fissions (see Fig. 2C): In the serial fissions process (Ramachandran et al., 2005), a single ancestral population grows in size and then splits. The parental population persists and the new daughter population grows and splits. This serial fissions process is repeated as populations spread out from the founding population. At the within-population level, the model predicts that gene identity will be lowest in the original ancestral population and that it will increase steadily as each new population forms, attaining its highest value in the most recently formed population. It also predicts a layered pattern of gene identity variation between populations whether they are in the same or different regions, i.e., the gene identity between the original ancestral population and all other populations will have a single low value, the gene identity between the first descendant population and all other populations will have a single slightly higher value, and so on. The serial fissions model is potentially consistent with the existence of taxonomic units both above and at the population level.

iv. Nested regions: We chose to include one additional model that has been suggested by recent studies of human population history (Bowcock et al., 1991, 1994; Cavalli-Sforza et al., 1994; Tishkoff et al., 1996; Jorde et al., 1997; Perez-Lezaun et al., 1997; Calafell et al., 1998; Ingman et al., 2000; Long and Kittles, 2003; Zhivotovsky et al., 2003). These studies suggest that relatively large bottlenecks occurred when humans initially colonized the world's major geographic regions. In a sense, this nested regions model is a version of the serial fissions model that predicts larger bottlenecks between populations in different regions than between populations within a region. Nested regions predicts multiple, uniform gene identity strata, one for each set of between-region comparisons (see Fig. 2D). For example, the comparisons of the oldest region to all other regions will form the lowest stratum. The comparisons of the next oldest region to all younger regions will form the second stratum, and so on. In each case, the stratum will be composed of a single gene identity value that is uniform throughout the geographic range of the species. In our version of the model, the populations within regions still form through a serial fissions process, so at the between-population, within-region level, the model predicts a layered pattern of gene identity (see predictions of the serial fissions model above). At the within-population level, nested regions predicts a steady increase in within-population gene identity: lowest in the founding population and highest in the most recently formed population. Because of the nesting of regions, it also predicts a stepwise increase in within-population gene identity variation between the regions. The nested regions model is also potentially consistent with the existence of taxonomic units both above and at the population level.

Comparison of simulated and microsatellite data

We used coalescent-based computer simulations to identify the predicted patterns of gene identity variation for the four models. The simulations took into account several factors that could affect the comparisons of simulated and observed patterns of gene identity variation. The factors include the relative youth of the species, recent possible changes in the pattern and magnitude of

TABLE 1. CEPH Sample

Geographic location	Coordinates	*N*
SUB-SAHARAN AFRICA		123
Biaka Pygmies	4N, 17E	32
Mbuti Pygmies	1N, 29E	15
Mandenka	12N, 12W	24
Yoruba	8N, 5E	25
San	21S, 20E	7
Bantu Kenya	3S, 37E	12
Bantu South Africa	26S, 24E	8
NORTH AFRICA		30
Mozabite	32N, 3E	30
MIDDLE EAST		145
Bedouin	31N, 35E	48
Druze	32N, 35E	46
Palestinian	32N, 35E	51
EUROPE		160
French (various regions)	46N, 2E	29
Basque	43N, 0	24
Sardinian	40N, 9E	28
Bergamo	46N, 10E	13
Tuscan	43N, 11E	8
Orcadian	59N, 3W	16
Adygei	44N, 39E	17
Russian	61N, 40E	25
SOUTH CENTRAL ASIA		193
Brahui	30N, 67E	25
Balochi	30N, 67E	25
Hazara	33N, 70E	24
Makrani	26N, 64E	25
Sindhi	25N, 69E	25
Pathan	33N, 71E	22
Kalash	36N, 72E	24
Burusho	36N, 74E	23
EAST ASIA		242
Han 1	32N, 114E	34
Han 2	32N, 114E	10
Tujia	29N, 109E	10
Yizu	28N, 103E	10
Miaozu	28N, 109E	10
Oroqen	50N, 127E	9
Daur	48N, 124E	10
Mongola	45N, 119E	10
Hezhen	47N, 134E	9
Xibo	43N, 82E	8
Uygur	44N, 81E	10
Dai	21N, 100E	10
Lahu	22N, 100E	10
She	27N, 119E	10
Naxi	26N, 100E	10
Tu	36N, 101E	9
Yakut	63N, 130E	24
Japanese	38N, 138E	28
Cambodian	12N, 105E	11
OCEANIA		35
Papuan	4S, 143E	17
NAN Melanesian	6S, 155E	18
AMERICA		104
Pima	29N, 108W	25
Maya	19N, 91W	25
Piapoco/others	3N, 68W	13
Karitiana	10S, 63W	24
Surui	11S, 62W	17
	Total	1,032

genetic exchange, the limited population sizes and numbers of the CEPH sample, and the evolutionary properties of autosomal microsatellites. We then visually compared the simulated gene identity patterns for each simulated model to the observed pattern estimated from the CEPH sample using two graphical methods. For the first graphical method, we constructed two sets of gene identity plots. The first set depicts within- and between-population gene identity *versus* geographic distance. Geographic distances for the microsatellite data are great circle distances estimated from the geographic coordinates listed in Table 1 using the haversine (Sinnott, 1984). The distances were estimated through waypoints on land, not across bodies of water. For the simulated data, populations were arrayed end-to-end and one unit of geographic distance was assigned between geographically contiguous populations (following Slatkin, 1993). The second set of plots depicts the within-population gene identity for each population.

For the second graphical method, we used an R-script (http://cran.r-project.org/) to construct contour maps of the gene identity matrices. In the contour maps, different levels of gene identity are represented by different colors. The colors permit easy visualization of patterns of within- and between-population gene identity variation. The diagonals of the maps are the within-population gene identities, and the off-diagonals are the between-population gene identities. For the simulated data maps, the populations were ordered from Population 1 in the bottom left corner to Population 49 on the right and at the top (see Fig. 2). For the microsatellite data, populations were also arranged by regions, beginning with Sub-Saharan Africa at the bottom left, and ending with the Americas to the right and on the top. To the extent possible, populations were arranged west to east in the order of their geographic distance from East Africa.

Simulations

The coalescent approach simulates the molecular diversity of genes sampled from a set of populations that have experienced a particular demographic history, e.g., isolation by distance. Unlike conventional forward-

Fig. 2. Gene identity plots and contour maps. The figure contains five sets of three panels. The first four sets (**A**–**D**) show the results for the simulations, and the last set (**E**) shows the results for the CEPH sample. The panels on the left show the within- and between-population gene identity *versus* geographic distance. For the simulations, the between-population geographic distances range from a value of 1 between all geographically contiguous populations to a value of 48 between populations 1 and 49. The middle panels show the within-population gene identities. For the simulations, the populations of Region 1 are on the left, the populations of Region 2 are next, and so on. For the CEPH data, the populations are plotted from left to right in order of their gene identity (lowest to highest), which corresponds well with their geographic distance from East Africa. The panels on the right show the contour maps. The maps are color-coded versions of the simulated and observed within- and between-population gene identity matrices. The gene identity matrices are square in form (49×49 for the simulations; 53×53 for the CEPH data), and so the colors above the diagonals are mirror images of the colors below the diagonals. The diagonals of the map show the within-population gene identities. The off-diagonals show the between-population gene identities. The specific gene identity levels are shown on the color-scale at the bottom of each map. (A) Independent regions. (B) Isolation by distance. (C) Serial fissions. Populations 1, 25, and 40 are color-coded in order to more clearly illustrate the gene identity pattern produced by the serial fissions process. (D) Nested regions. (E) CEPH sample. AF, Africa; ME, Middle East; EU, Europe; SC, South Central Asia; WE, Western Eurasia; OC, Oceania; EA, East Asia; AM, Americas.

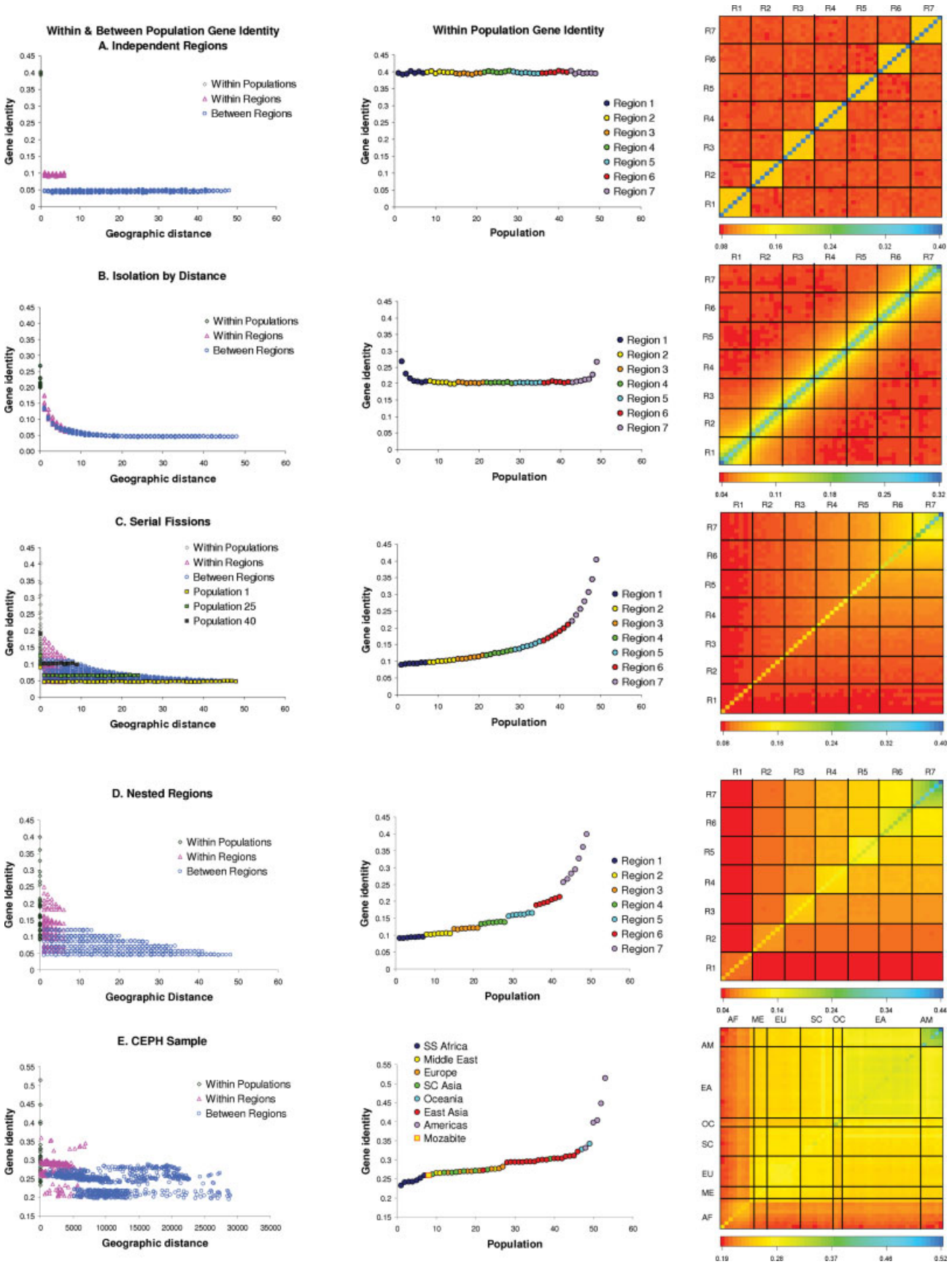
Within & Between Population Gene Identity
A. Independent Regions
Within Populations
Within Regions
Between Regions
Gene identity
Geographic distance
Within Population Gene Identity
Region 1
Region 2
Region 3
Region 4
Region 5
Region 6
Region 7
Population
B. Isolation by Distance
C. Serial Fissions
Population 1
Population 25
Population 40
D. Nested Regions
Geographic Distance
E. CEPH Sample
SS Africa
Middle East
Europe
SC Asia
Oceania
East Asia
Americas
Mozabite
R1 R2 R3 R4 R5 R6 R7
AF ME EU SC OC EA AM

Fig. 2.

in-time simulation methods, the coalescent approach is retrospective, simulating the coalescence of genes backwards in time (Kingman, 1982). This retrospective approach permits the rapid simulation of a large number of replicates of a given demographic history. For our analyses, we used the computer package Simcoal (v2.1, Laval and Excoffier, 2004) because it allowed us to simulate the different models and because it allowed us to easily vary important parameters like the number of populations, the number of genomes per population, the number of unlinked microsatellite loci per genome, and the microsatellite mutation rates and mutation model.

The coalescent simulations for each model shared the following features: 1) each simulation included seven regions, each comprised of seven populations. The number of populations and regions was chosen to roughly match the number of populations and geographic distribution of the CEPH sample. We opted to include equal numbers of populations per region even though the numbers differ for the CEPH sample. We felt this was useful for interpreting the results of several of the models, particularly isolation by distance; 2) the 49 populations were arrayed end to end, and one unit of geographic distance was assigned between contiguous populations. The edge Populations 1 and 49 were not connected, leading to a geographic distance between them of 48 units. The lack of a connection between the edge populations is equivalent to assuming that humans evolved in Africa, subsequently spread out of Africa, and arrived most recently in the Americas; 3) each population was comprised of multiple diploid genomes, and each genome was comprised of 50 unlinked microsatellite loci. The number of loci was limited to 50 to minimize the time required to complete each set of simulations, which allowed us to run many more simulations of each model using different parameters. Increasing the number of loci had no effect on the results; 4) the microsatellite loci accumulated mutations under a stepwise model (Slatkin, 1995) at a rate of 0.0075 mutations per generation (Brinkmann et al., 1998).

Parameters that varied in simulations of the different models included the effective population sizes (number of diploid genomes/population), the between-population migration rates, the pattern of migration, the population growth rates, and the order and timing of population splits. Each model was simulated using a host of different values for these parameters. For example, in different simulations, the timing of various region- and population-level splits ranged from 10,000 to 25 generations before the present (200,000–500 years assuming a generation length of 20 years). These values encompass ranges reported in the anthropological literature for the age of the species and the timing of colonization of major world regions (Cavalli-Sforza et al., 1994; Ingman et al., 2000; Zhivotovsky et al., 2000, 2003; Marth et al., 2003). Regardless of the particular parameters chosen, each simulation produced the distinctive gene identity patterns reported in the results.

The key parameters for the reported simulation results for each model are as follows:

i. Independent regions: In the independent regions simulations, starting from the present and moving backwards in time, seven populations fused into a single ancestral regional population in each of seven different regions 2,500 generation before the present. The seven ancestral regional populations then fused into a single ancestral population 5,000 generations before the present.
ii. Isolation by distance: In the isolation by distance simulations, the 49 populations fused into a single ancestral population 5,000 generations before the present. Prior to fusion, each population exchanged a constant portion of its genomes each generation with its immediate geographic neighbors. Exchange was not permitted between edge Populations 1 and 49. We also constructed a version of the model in which populations exchanged genomes with geographic neighbors in two dimensions. The results were qualitatively the same for both versions, and we report the results for the linear version of the model only.
iii. Serial fissions: The coalescent simulations of the serial fissions model involved the serial fusion of populations backwards in time, the first fusion occurring 300 generations before the present and the last occurring 5,000 generations before the present. One fusion occurred approximately every 100 generations. Exponential population decline occurred on the branch following each fusion (equivalent to population growth when moving forward in time). In keeping with Ramachandran's et al. (2005) version of the model, migration was not permitted between any of the populations.
iv. Nested regions: In the nested regions simulations, serial population fusions occurred within each region between 75 and 700 generations before the present, and the regions fused in a nested fashion, Region 7 into Region 6, Region 6 into Region 5, and so on, between 300 and 1,500 generations before the present.

We found that 50 simulations were sufficient to capture the pattern of gene identity variation associated with each model. Limiting the number of simulations to 50 also allowed us to run more simulations that varied the shared and model-specific parameters. At the end of each simulation, 20 diploid genomes were sampled from each population. Within- and between-population gene identities were estimated from the sampled genomes. The gene identities were then averaged for the 50 simulations and the estimates plotted using the graphical methods described above.

RESULTS

Figure 2 contains five sets of three panels. The first four sets (A–D) show the results for the simulations, and the last set (E) shows the results for the CEPH sample. The leftmost of the three panels shows gene identity *versus* geographic distance. The data points plotted at zero geographic distance are the within-population gene identities. The middle panel shows just these within-population gene identities plotted in the geographic order described in the figure legend. The rightmost panels show the contour maps.

Independent regions

Figure 2A shows the results of the independent regions simulations. As predicted, the gene identities are highest within local populations (dark green diamonds plotted at zero geographic distance in the leftmost panel), lower between populations in the same region (pink triangles), and lowest between populations in different regions (blue circles). The gene identities are uni-

form at each level of the three-tiered hierarchy, i.e., there is no tendency for gene identity to change with geographic distance. The middle panel confirms that the within-population gene identities are almost identical for all of the populations. The uniformity is also reflected by the presence of only three colors in the contour map (right panel in Fig. 2A), blue within populations, orange between populations within regions, and red between populations in different regions.

Simulations of the independent regions model using a host of different parameter values indicated that the results are not an artifact of the particular parameters chosen, but instead reflect the general pattern of gene identity variation that would have been produced had human evolution occurred in the manner envisioned by the model. This is true for each of the following models as well.

Isolation by distance

The simulations of isolation by distance reveal a smooth monotonic decrease in population pairwise gene identity with increasing geographic distance (Fig. 2B, left panel). The decay is reflected in the contour map by the steady change in color extending at right angles from the diagonals. The color starts out as blue on the diagonals and then progresses through green, yellow, and orange before quickly becoming red, which represents the asymptote in gene identity at about 0.04 seen in the left panel of Fig. 2B.

The within-population gene identity plot (Fig. 2B, middle panel) shows that the simulated gene identities are fairly uniform in the populations located in the middle of the one-dimensional geographic range of the 49 populations, and that the identities increase towards the edges. This "edge effect" is also evident in the increasing intensity of the blue color towards the opposing ends of the diagonals of the contour map (Fig. 2B, right panel). The edge effect occurs because genetic exchange was not permitted between Populations 1 and 49 (equivalent of assuming no gene flow between African and Native American populations). Otherwise the simulated pattern shows that the prediction of isolation by distance of uniform within-population gene identities is met.

Serial fissions

Figure 2C shows the results of the serial fissions simulations. The between-population gene identities are represented by 48 layers of gene identity values (Fig. 2C, left panel). Populations 1, 25, and 40 are color coded to more clearly illustrate this layered pattern. The lowest layer consists of the gene identities between Population 1 and the other 48 populations. The gene identity values for this layer are uniform, i.e., they do not change with geographic distance. The second layer consists of the gene identities between Population 2 and Populations 3–49, which are also uniform across geographic distance. The third layers consists of the gene identities between Population 3 and Populations 4–49, which are also uniform, and so on. This nested pattern of gene identity variation is reflected in the contour map by uniform layers of color proceeding rightwards (and upwards) from the diagonal and by a gradual change in color from red at the bottom and left of the map to light green at the top right.

The within-population gene identity plot (Fig. 2C, middle panel) shows that gene identities are lowest in Population 1 and increase steadily through Population 49. The steady increase is reflected in the contour map as a gradual change in color from orange in the lower left diagonal to blue in the upper right diagonal.

Nested regions

As predicted by the nested regions model, at the between-population in different regions level, there is a single gene identity stratum for each region (blue circles in Fig. 2D, left panel). The lowest stratum is formed by the comparisons of populations in Region 1 to populations in all other regions. The next stratum is the comparisons of Region 2 to the remaining 5 regions, and so on. In the contour map, these between-region comparisons are represented by separate strata, each with a distinct uniform color, e.g., red for the Region 1 *versus* all other region comparisons, dark orange for the Region 2 *versus* the other five regions, etc.

In the left panel, within-regions, there are seven groupings of pink triangles, and each is essentially a miniversion of the serial fissions plot. The lowest layer in each regional grouping is the comparisons of the oldest population (founding population) to all other populations in the region. The next layer is the comparisons of the second oldest population to other five populations, and so on.

The within-population gene identity pattern (Fig. 2D, middle panel and diagonals of the contour map) is similar to that of the serial fissions model, but there is more of a gap between the regions than between the populations within a region. The gap reflects the large regional-level bottlenecks of the nested regions model.

CEPH microsatellites

The gene identities for the CEPH microsatellites are shown in Figure 2E. The most salient feature of the left panel is the distinctive strata of gene identity at the within- (pink triangles) and between-region (blue circles) levels. The distinctive strata are not consistent with the independent regions model, which predicts a single between-region stratum. The distinctive strata are also not consistent with isolation by distance, which predicts a monotonic decay in gene identity with increasing geographic distance. The strata seem more consistent with the simulated nested regions results, though the observed strata are not as clear cut as they are in the simulated plots. Other features of the CEPH plot, such as a subtle decay in some regions in gene identity with geographic distance (discussed below), are not consistent with the nested region's predictions of uniform gene identity.

To aid in interpretation of the CEPH gene identity pattern, the regional strata are examined in more detail in Figure 3. Figure 3A highlights the gene identities between populations in different regions. Sub-Saharan African population comparisons form the bottom stratum, followed by separate strata for Western Eurasia, Oceania, and East Asia (vs. the Americas). Gene identity does not decay with increasing geographic distance on most of the strata. The existence of multiple, uniform regional strata are consistent with the predictions of the nested regions model.

Figure 3C highlights just the comparisons of Sub-Saharan African populations to all of the populations in each of the other regions. The plot confirms that the

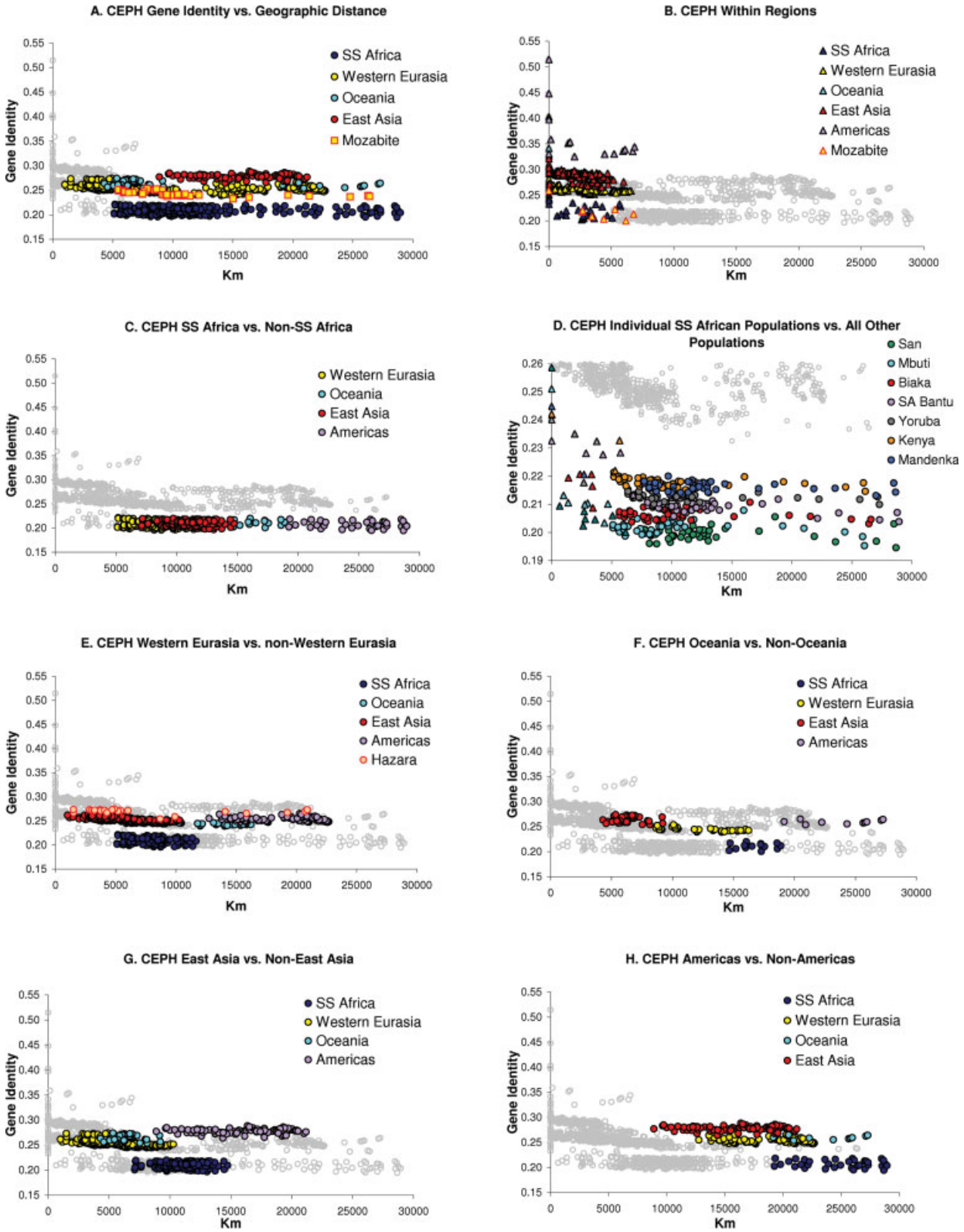

Fig. 3. Within- and between-region gene identity strata for the CEPH data. (**A**) Highlights all of the between-region gene identity strata. (**B**) Highlights all of the within-region gene identity strata. (**C**–**H**) Highlight individual between-region strata. Plot D compares individual Sub-Saharan African populations to each other (triangles) and to populations in other regions (circles).

gene identity between Sub-Saharan African and non-Sub-Saharan African populations is uniform for each region, and that there is no decay in gene identity with increasing geographic distance. Figure 3D highlights the individual Sub-Saharan African populations. The *Y*-axis scale is expanded to allow better visualization of the individual Sub-Saharan African population patterns. The triangle shapes are the comparisons between populations within Sub-Saharan Africa, and the circles are the comparisons between Sub-Saharan African and non-Sub-Saharan populations. The plot shows that there are separate layers of gene identity for most of the Sub-Saharan African populations. The separate strata are consistent with the predictions of the serial fissions model within Sub-Saharan Africa. The separate strata are evident even between populations within Africa, even though there is decay in gene identity with increasing geographic distance within the region. The decay suggests that some genetic exchange has occurred between populations within Sub-Saharan Africa. It is difficult to assess how long the gene flow has been taking place. For example, it may be the result of recent changes in subsistence in Sub-Saharan Africa (Destro-Bisol et al., 2004; Wilkins and Marlowe, 2006). However long it has been occurring, it has not yet erased the strata produced initially through serial fissions. This result is important because it shows that despite the decay with geographic distance, the pattern of genetic identity within Africa does not yet conform to the predictions of isolation by distance.

The Sub-Saharan African results have important implications for race. For Sub-Saharan Africans to belong to a single race, all Sub-Saharan African populations would have to cluster together on a single branch of a larger species tree, which means that each population would have about the same level of gene identity with all non-Sub-Saharan African populations, i.e., the individual population strata in Figure 3D wouldn't exist. These separate population strata indicate that the Sub-Saharan African populations do not cluster together, and, therefore, that there is no Sub-Saharan African race in any taxonomic sense.

The other regional strata are shown in Figure 3E–H. At the between-region level, with one important exception, gene identity is uniform across geographic distance on each stratum. The identities are particularly uniform across thousands of kilometers of geographic distance for the Western Eurasia *versus* Native American (Fig. 3E), Oceanic *versus* Native American (Fig. 3F) and East Asian *versus* Native American (Fig. 3G) strata. In contrast to the African pattern, the individual populations in the different regions do not form their own individual gene identity strata, but are more evenly mixed (results not shown). These results are highly consistent with the nested regions model.

The exception to the trend of uniform gene identity occurs for the Western Eurasia-East Asian comparisons shown in Figure 3E. Here, though it is difficult to make out in the figure (red circles), there is a slight but statistically significant decay in gene identity with increasing geographic distance. The trend suggests that genetic exchange has been occurring between the populations of Western Eurasia and East Asia long enough to have affected the geographic pattern of gene identity variation (see also Li et al., 2008). There is also a slight but statistically significant decay in gene identity between populations within each of the regions (Fig. 3B). This decay is also probably the result of genetic exchange between populations within each region.

It is important to stress that this genetic exchange has not erased evidence of a previous history of region- and population-level fissions. This conclusion is attested not only by the distinct regional gene identity strata but also by the geographic pattern of within-population gene identity variation shown in the middle panel of Figure 2E. The figure shows a steady increase in within-population gene identity with increasing geographic distance from Africa for all populations. It also shows a subtle stepwise increase in within-population gene identity between the regions. These results are consistent with nested regional bottlenecks and serial population fissions within regions. We conclude that the global pattern of gene identity variation is the outcome of nested serial population fissions within regions, successive, nested large-scale regional-level bottlenecks out of Africa, and gene flow between nearby populations.

Additional simulations

To confirm that this combination of serial fissions, regional bottlenecks, and gene flow is consistent with the observed pattern of gene identity variation, we ran one additional set of simulations. In this set, we repeated the nested regions simulations but also allowed 5% migration (genomes/generation) between geographically contiguous populations following each regional founder event. In other words, migration was permitted between adjacent populations after each region was colonized. We also matched the number of populations per simulated region with the actual numbers in the CEPH sample and treated Western Eurasia as a single region. The results of the revised nested regions simulations are shown in Figure 4B. Although the strata are more compressed for the simulated data in Figure 4B than they are for the CEPH data in Figure 4A (left panels), the patterns of gene identity variation are similar at all three levels (within-population, between-population within regions, and between-population in different regions). The left panel of Figure 4B also shows that the revised simulations capture the subtle decays in gene identity within regions seen for the CEPH sample. The similarities between the middle and right panels of Figure 4A and B are also striking. The revised simulations confirm that the observed pattern of global gene identity variation is consistent with a human prehistory of nested serial fissions within regions, nested regional bottlenecks, and genetic exchange between local populations.

Outlier populations

There are several outlier populations in the sample. Within Africa, the Mozabite layer stands out (Figs. 3A and 4A). They are the topmost population in the African section of the contour map in Figure 4A (right panel). Their unique gene identity pattern might reflect a bottleneck that separated Sub-Saharan African populations from other populations in Africa. It might also reflect the large geographic distance between the Mozabite and the Sub-Saharan African populations. If populations were sampled uniformly geographically within Africa, the Mozabite might not be so distinctive. In Western Eurasia, the Hazara stand out (Figs. 3E and 4A). In the Western Eurasian section of the contour map in Figure 4A, they form the yellow band amidst the sea of orange color.

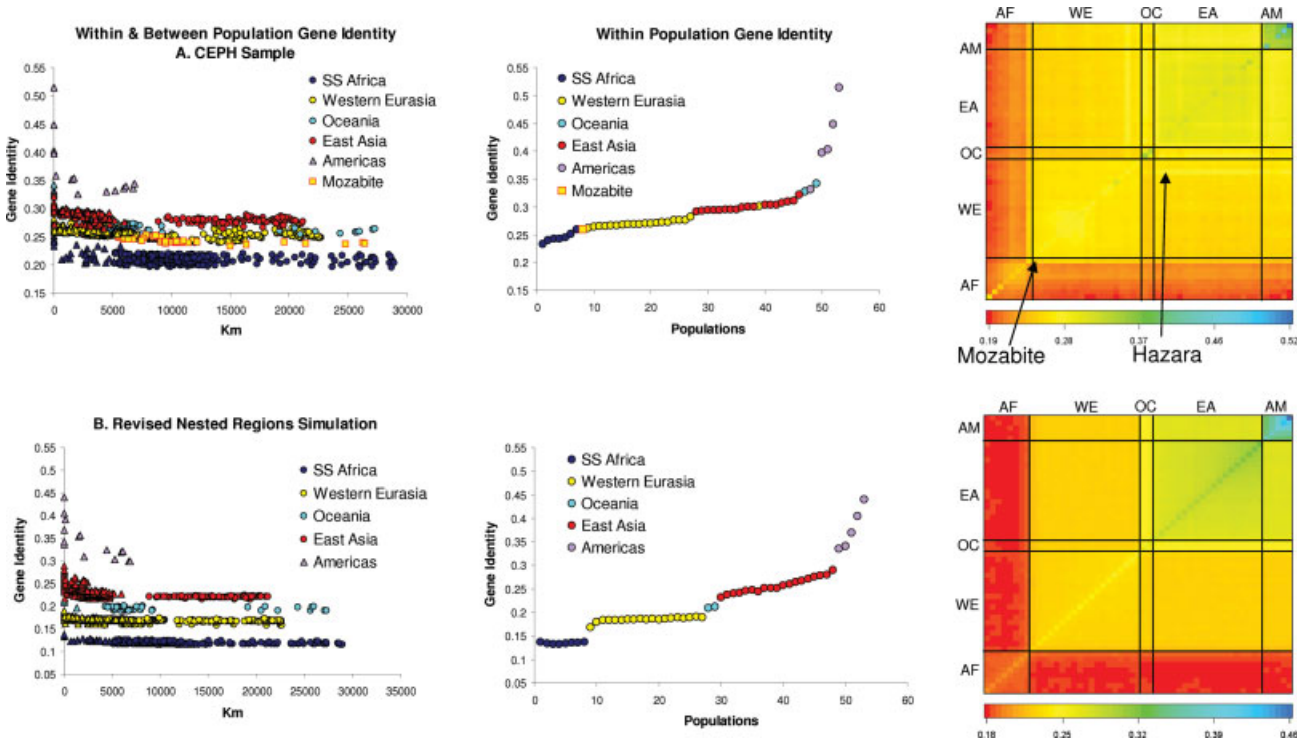

Fig. 4. (**A**) CEPH sample. The plots contain the same data as in Figures 2E and 3, but with modified color coding. (**B**) Revised Nested Regions Simulations.

Their gene identity to East Asia is higher than it is for the other Western Eurasia populations. This result might reflect an East Asian ancestry for the Hazara, as suggested in a recent ethnographic study (Mousavi, 1997). The Oceanic populations also stand out in having higher within-population gene identity than predicted by their position in the nested regional hierarchy. The distinctive within-population gene identities may have something to do with the complex history of secondary colonizations and population movements in the region (Spriggs, 1997). Since only two Oceanic populations are included in the CEPH sample, it is difficult to draw any firm conclusions about the region.

DISCUSSION

Our goal was to compare the pattern of within- and between-population gene identity predicted by several models to the observed pattern measured from neutral genetic variation assayed in 53 globally distributed populations. The models were based on several well-known descriptions of human biological variation (Lewontin, 1972; Ramachandran et al., 2005; Handley et al., 2007). The results indicate that no single model explains all details of the observed pattern of gene identity variation.

The results are inconsistent with the independent regions model, for example, which requires 1) independent evolution between local populations in the same region, and between population groups in different regions, and 2) equal diversity in all local populations, and in all regions. These requirements are necessary for methods that apportion genetic variation into within and between group components to be meaningful. Because the requirements are violated and groups are instead nested, we question the validity of results produced by such endeavors.

The results are also inconsistent with the predictions of isolation by distance. The relationship between gene identity and geographic distance forms multiple tiers rather than a monotonic decay. This result suggests that the highly debated results of Rosenberg et al. (2002, 2005) are not an artifact of uneven geographic sampling, as has been suggested (Serre and Pääbo, 2004). There is no sampling scheme or transformation of the data that can make the multi-tier gene identity pattern monotonic with respect to geographic distance. In addition, for several of the inter-regional comparisons that comprise the multiple tiers of gene identity variation, there is no decay in gene identity with increasing geographic distance. For example, East Asian *versus* American gene identities are uniform across several thousand kilometers of geographic distance. This geographic pattern is consistent with a bottleneck and long-range migration from East Asia into the Americas. Bottlenecks and long-range migrations of this type have been important factors in shaping human genetic variation (Cavalli-Sforza et al., 1988; Bowcock et al., 1991, 1994; Nei and Roychoudhury, 1993; Cavalli-Sforza et al., 1994; Jorde et al., 1997; Long and Kittles, 2003; Zhivotovsky et al., 2003) and have left a clear genetic signature in the pattern of inter-regional gene identity variation.

Though the pattern of gene identity variation does not conform to the predictions of isolation by distance, our results suggest that genetic exchange has led to some decay in gene identity with increasing geographic distance within regions and between Western Eurasia and East Asia. The magnitude of the decay is small, but it is statistically significant. The results suggest that the pattern of local genetic exchange that characterizes isolation by distance has been occurring within regions and between regions in Eurasia, and it is slowly erasing the

genetic signatures of regional bottlenecks and population fissions that occurred in the past. In other words, isolation by distance characterizes recent human population history in many locations, but it will take some time before the global gene identity pattern conforms to the predictions of the model. Simulations of the revised nested regions model that extend the pattern of local migration for several thousand generations confirm this prediction (results not shown).

With respect to the two remaining models, serial fissions and nested regions, the serial fissions model accounts for some of the within-region pattern of gene identity variation and for the geographic pattern of within-population gene identity variation, but it does not adequately take into account the comparative magnitude of the genetic effects of regional-level bottlenecks. The nested regions model does account for these inter-regional effects.

We conclude that the observed pattern of global gene identity variation was produced by a combination of serial population fissions, bottlenecks and long-range migrations associated with the peopling of major geographic regions, and subsequent gene flow between local populations.

Implications for race

If the independent regions model was correct, then individuals in the same geographic region would on average be more closely related to each other genetically than would be individuals in different geographic regions. Even in this case, the problem of finding a threshold level of gene identity for declaring taxonomic significance would remain unsolved.

In reality, the between-population gene identity pattern is nested. Because the between-population pattern is nested in Sub-Saharan Africa, and because Sub-Saharan African populations straddle the root of the species-wide population tree (e.g., Li et al., 2008), there can be no Sub-Saharan African race under the shared genetic relationship criterion. The first division in the population hierarchy that coincides with continental locations separates non-African populations from African populations. This division is consistent with the existence of a non-African race, but because of the root, the Africans would still not constitute a race. Another major division along continental lines separates East Asian and Native American populations from all others. However, classifiers would need to put East Asians and Native Americans into a sub-race, because they would already be members of the non-African race. Thus, we see that nested pattern of variation would require that the geographic groups that anthropologists traditionally considered races could not be assigned to the same level of hierarchical classification.

Other implications

The implications of our results for human prehistory and race are fairly straightforward. The medical genetics implications are less clear. Some argue that race is a reasonable proxy for genetic structure and that genetic structure has important implications for understanding health disparity (Risch et al., 2002; Burchard et al., 2003; Sankar et al., 2004; Tang et al., 2005). Others argue that race is a poor proxy for genetic structure and should not be used in medical diagnosis and research (King, 2000; Wilson et al., 2001; Cooper et al., 2003). Wilson et al. (2001), for example, found that variation at drug metabolizing enzyme (DME) loci corresponded to genetic clusters, but that the genetic clusters did not correspond closely to ethnic labels used in medicine, like Black, Caucasian and Asian. However, their genetic clusters did correspond to major geographic regions, implying that one reason for the poor correspondence between ethnicity and genetic clusters was the lack of geographic resolution of the ethnic labels (see also Risch et al., 2002). Our findings confirm that broad ethnic categories employed in medical genetic research might not adequately take into account the complex geographic pattern of genetic structure in the species, but for the same reason, neither may continental ancestry. This is because our results also indicate that substantial, potentially medically important genetic differences may exist between populations within regions. The pronounced genetic structure in Africa may have particular salience in medical genetic research in the United States.

CONCLUDING COMMENTS

The nested pattern of genetic diversity is at odds with the pattern of diversity that evolutionary independence of regions would produce. It also complicates any formal taxonomic system (e.g. race or subspecies) for human populations on different continents. Whereas traditional anthropological classifications placed human populations that reside on different continents at the same level of classification (i.e. race), a classification that takes into account evolutionary relationships and the nested cascade of diversity would require that Sub-Saharan Africans are not a race, and that nested sub-races would be necessary to account for non-Sub-Saharan Africans. We see no need for such a classification system in light of the fact that our evolutionary history gives good guidance for understanding the structure of human diversity.

LITERATURE CITED

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

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK, Cavalli-Sforza LL. 1991. Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci USA 88:839–843.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457.

Boyd R, Silk J. 2006. How humans evolved. New York: W. W. Norton & Company.

Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B. 1998. Mutation rate in human microsatellites: influences of the structure and length of the tandem repeat. Am J Hum Genet 62:1408–1415.

Brown R, Armelagos G. 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.

Calafell F, Shuster A, Speed W, Kidd JR, Kidd KK. 1998. Short tandem repeat polymorphism evolution in humans. Eur J Hum Genet 6:38–49.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L,

Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL. 2002. A human genome diversity cell line panel. Science 296:261–262.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton: Princeton University Press. 413 p.

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J. 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc Natl Acad Sci USA 85:6002–6006.

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

Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglia A, Tofanelli S, Spedini G, Capelli C. 2004. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. Mol Biol Evol 21:1673–1682.

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

Fuentes A. 2007. Core concepts in biological anthropology. Boston: McGraw Hill.

Handley LJ, Manica A, Goudet J, Balloux F. 2007. Going the distance: human population genetics in a clinal world. Trends Genet 23:432–439.

Harpending H, Rogers A. 2000. Genetic perspectives on human origins and differentiation. Annu Rev Genomics Hum Genet 1:361–385.

Ingman M, Kaessmann H, Pääbo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708–713.

Jorde L, Watkins W, Bamshad M, Dixon M, Ricker C, Seielstad M, Batzer M. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979–988.

Jorde LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, Kere J, Harpending HC. 1997. Microsatellite diversity and the demographic history of modern humans. Proc Natl Acad Sci USA 94:3100–3103.

Jurmain R, Kilgore L, Trevathan W. 2005. Introduction to physical anthropology. Toronto: Thomson Wadsworth.

Kimura M, Weiss G. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49:561–576.

King R. 2000. Racialization, recognition, and rights: lumping and splitting multiracial Asian Americans in the 2000 Census. J Asian Am Stud 3.2:191–217.

Kingman J. 1982. The coalescent. Stoch Proc Appl 13:235–248.

Laval G, Excoffier L. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20:2485–2487.

Lewontin R. 1972. The apportionment of human genetic diversity. In: Dobzhansky T, Hecht M, Steere W, editors. Evolutionary biology, Vol. 6. New York: Appleton-Century-Crofts.

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.

Livingstone F. 1962. On the non-existence of human races. Curr Anthropol 3:279–281.

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

Malécot G. 1948. Les Mathématiques de l'hérédité. Paris: Masson et Cie.

Manica A, Prugnolle F, Balloux F. 2005. Geography is a better determinant of human genetic differentiation than ethnicity. Hum Genet 118:366–371.

Marth G, Schuler G, Yeh R, Davenport R, Agarwala R, Church D, Wheelan S, Baker J, Ward M, Kholodov M, Phan L, Czabarka E, Murvai J, Cutler D, Wooding S, Rogers A, Chakravarti A, Harpending HC, Kwok PY, Sherry ST. 2003. Sequence variations in the public human genome data reflect a bottlenecked population history. Proc Natl Acad Sci USA 100:376–381.

Mousavi S. 1997. The Hazaras of Afganistan: an historical, cultural, economic and political study. New York: St. Martin's Press.

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

Nei M, Roychoudhury A. 1982. Evolutionary relationships and evolution of human races. Evol Biol 14:1–59.

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

Perez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Bertranpetit J. 1997. Microsatellite variation and the differentiation of modern humans. Hum Genet 99:1–7.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137.

Prugnolle F, Manica A, Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. Curr Biol 15:R159–R160.

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.

Reich DE, Goldstein DB. 2001. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20:4–16.

Risch N, Burchard E, Ziv E, Tang H. 2002. Categorization of humans in biomedical research: genes, race and disease. Genome Biol 3: comment 2007.

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.

Sankar P, Cho M, Condit C, Hunt L, Koenig B, Marshall P, Lee S, Spicer P. 2004. Genetic research and health disparities. J Am Med Assoc 291:2985–2989.

Serre D, Pääbo S. 2004. Evidence for gradients of human genetic diversity within and among continents. Genome Res 14:1679–1685.

Sinnott RW. 1984. Virtues of the Haversine. Sky and telescope 68:159–161.

Slatkin M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47:264–279.

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

Spriggs M. 1997. The Island Melanesians. Cambridge: Blackwell Publishers.

Tang H, Quertermous T, Rodriguez B, Kardia S, Zhu X, Brown A, Pankow JS, Province MA, Hunt SC, Boerwinkle E, Schork NJ, Risch NJ. 2005. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet 76:268–275.

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387.

Wilkins JF, Marlowe FW. 2006. Sex-biased migration in humans: what should we expect from genetic data? Bioessays 28:290–300.

Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, Bradman N, Goldstein DB. 2001. Population genetic structure of variable drug response. Nat Genet 29:265–269.

Zhivotovsky LA, Bennett L, Bowcock AM, Feldman MW. 2000. Human population expansion and microsatellite variation. Mol Biol Evol 17:757–767.

Zhivotovsky LA, Rosenberg NA, Feldman MW. 2003. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72:1171–1186.