Available online at www.sciencedirect.com
SCIENCE DIRECT®
Forensic Science International xxx (2005) xxx–xxx
Forensic Science International
www.elsevier.com/locate/forsciint

Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases

Michael F. Hammer [a,b,*], Veronica F. Chamberlain [a], Veronica F. Kearney [a], Daryn Stover [a], Gina Zhang [a], Tatiana Karafet [a], Bruce Walsh [b], Alan J. Redd [a,1]

[a] *Division of Biotechnology, University of Arizona, Tucson, AZ 85721, USA*
[b] *Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA*

Received 17 August 2005; received in revised form 8 November 2005; accepted 8 November 2005

Abstract

A set of 61 Y chromosome single-nucleotide-polymorphisms (Y-SNPs) is typed in a sample of 2517 individuals from 38 populations to infer the geographic origins of Y chromosomes in the United States and to test for paternal admixture among African-, European-, Hispanic-, Asian-, and Native-Americans. All of the samples were previously typed with the 11 core U.S. Y chromosome short tandem repeats (Y-STRs) recommended by SWGDAM, which revealed high levels of among ethnic group variation and low levels of among-population-within-ethnic-group variation. Admixture estimates vary greatly among populations and ethnic groups. The frequencies of non-European (3.4%) and non-Asian (4.5%) Y chromosomes are generally low in European–American and Asian–American populations, respectively. The frequencies of European Y chromosomes in Native-American populations range widely (i.e., 7–89%) and follow a West to East gradient, whereas they are relatively consistent in African–American populations (26.4 ± 8.9%) from different locations. The European (77.8 ± 9.3%) and Native-American (13.7 ± 7.4%) components of the Hispanic paternal gene pool are also relatively constant among geographic regions; however, the African contribution is much higher in the Northeast (10.5 ± 6.4%) than in the Southwest (1.5 ± 0.9%) or Midwest (0%). To test for the effects of inter-ethnic admixture on the structure of Y-STR diversity in the U.S., we perform subtraction analyses in which Y chromosomes inferred to be admixed by Y-SNP analysis are removed from the database and pairwise population differentiation tests are implemented on the remaining Y-STR haplotypes. Results show that low levels of heterogeneity previously observed between pairs of Hispanic-American populations disappear when African-derived chromosomes are removed from the analysis. This is not the case for an unusual sample of European–Americans from New York City when its African-derived chromosomes are removed, or for Native-American populations when European-derived chromosomes are removed. We infer that both inter-ethnic admixture and population structure in ancestral source populations may contribute to fine scale Y-STR heterogeneity within U.S. ethnic groups.

Keywords: Forensic science; DNA typing; Single nucleotide polymorphisms (SNPs); Y chromosome; United States; Y-STR databases; Ethnic groups; Population structure; Admixture

* Corresponding author. Tel.: +1 520 621 9828;
fax: +1 520 621 9247.
E-mail addresses: mfh@u.arizona.edu, mfh@email.arizona.edu (M.F. Hammer).
[1] Present address: Department of Anthropology, University of Kansas, Lawrence, KS 66045, USA.

0379-0738/$ – see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2005.11.013

1. Introduction

Genetic markers on the non-recombining portion of the Y chromosome are becoming an important tool for the forensic scientist [1–5]. Haplotypes based on genotypes at multiple Y-linked short tandem repeats (Y-STRs) are useful for characterizing male DNA in material from sexual assault and forcible rape cases [6–9]. Y chromosome haplotype data are particularly useful when it is difficult to separate victim and perpetrator cell types, when signal from victim DNA overwhelms that from the perpetrator(s), or in cases when the perpetrator is vasectomized or azoospermic [6,8,9]. Because one cannot use the product rule to obtain estimated Y-STR haplotype frequencies, and because of the very high diversity of Y-STR haplotypes in most populations [10–12], large databases of complete Y-STR haplotypes are needed as a source for estimating haplotype frequencies. To assess the reliability of such databases as fair representations of actual population haplotype frequencies, the extent of structure among populations also needs to be considered. This is particularly true for markers on the Y chromosome because its haploid and paternal mode of inheritance makes it more sensitive to genetic drift than the autosomes [13]. Therefore, populations are more likely to differ substantially in Y-STR haplotype frequencies.

The population of the United States has been characterized as a cultural melting pot, with over 500 ancestries reporting in the 2000 Census [14]. However, for forensic purposes U.S. populations are typically categorized as African–American, Caucasian (European–American), Hispanic, Asian–American, and Native-American (we note that the U.S. census considers the ''Hispanic'' category separate from other ethnic groups because Hispanics are a heterogeneous group with multiple origins, see Section 4). Genetic studies indicate significant differentiation among these ethnic groups; a major outstanding question concerns the level of population structure within ethnic groups [15], especially for the Y chromosome. The extent to which we expect genetic heterogeneity among populations within an ethnic group in the U.S. depends mainly on four factors: levels of subdivision in ancestral source populations (e.g., Africa, Europe, etc.), the extent of non-random migration to the U.S., migration rates among geographic regions after arrival in the U.S., and the degree to which inter-ethnic admixture varies regionally [10,12]. Previously, we typed 11 ''core'' Y-STR loci (i.e., those that were recommended by the Scientific Working Group on DNA Analysis Methods) in a sample of 2517 individuals from 38 populations representing five U.S. ethnic groups and found surprisingly low levels of heterogeneity within African–American, European–American, and Hispanic groups in analyses of molecular variance (AMOVA) [12]. However, more stringent pair-wise population differentiation (PPD) tests [16] revealed heterogeneity between particular samples within our European–American and Hispanic groups. Kayser et al. [10] found similar results in their analysis of a separate U.S. Y-STR database and suggested that the observed heterogeneity was the result of chance. They concluded that the absence of significant broad-scale population structure within U.S. ethnic groups means that forensic DNA databases do not need to be constructed for separate geographic regions of the U.S.

In this report, we examine the role of regional variation in inter-ethnic admixture as a possible underlying cause of low levels of genetic heterogeneity within U.S. ethnic groups. Y chromosome single-nucleotide-polymorphisms (Y-SNPs) are valuable markers for quantifying admixture among U.S. populations. There is extensive knowledge regarding the geographic origins of Y-SNPs based on studies of global populations [17–19]. Because of the high geographic specificity of Y-SNPs [13,20], SNP haplogroups can be used to directly measure admixture among diverse populations without resorting to more complex models of admixture [21,22]. This global phylogeographic framework for inferring the origin of Y-SNP haplogroups is important because U.S. populations are composed of individuals with ancestry deriving from many parts of the world. In this paper we type a set of 61 binary markers in the same samples that were previously typed for 11 core Y-STRs, that is, a database of 2517 individuals representing five U.S. ethnic groups. We estimate the extent of population structure and proportions of multi-ethnic (African, European, Asian, and Native-American) paternal ancestry in all 38 populations (Table 1). We also assess the role of inter-ethnic admixture as an underlying cause of heterogeneity in the frequency of Y-STR haplotypes among U.S. populations.

2. Materials and methods

2.1. DNA samples

Samples for this study (Table 1) come from U.S. crime laboratories and have been described previously [12]. The population samples include individuals from five ethnic groups including: African–American (AA) ($n = 651$; 10 populations), European–American (EA) ($n = 927$; 10 populations), Hispanic-American (HA) ($n = 479$; 9 populations), Native-American (NA) ($n = 398$; 7 populations), and Asian–American (SA) ($n = 62$; 2 populations). Samples are reported to derive from individuals of self-described ancestry in all cases except those from New York City. These latter samples are from deceased individuals whose ethnicity was identified either by a family member or a medical examiner who made the determination of ethnicity based on the appearance of the decedent. All sampling protocols were approved by the Human Subjects Committee at the University of Arizona. Extraction and DNA quantification methods were previously described [11].

Table 1
Diversity statistics for Y chromosome haplogroups (Hg)

Ethnic group/population	Sample size	Number of Hgs	Discrimination capacity (%)	Haplogroup diversity (±S.E.)
African–Americans (AA)	651	24	3.7	0.585 ± 0.020
Arizona-Phoenix (AZ1)	76	8	10.5	0.564 ± 0.058
Arizona-Mesa (AZ2)	52	8	15.4	0.554 ± 0.076
Connecticut (CT)	89	13	14.6	0.514 ± 0.061
Florida (FL)	20	5	25.0	0.442 ± 0.133
North Carolina (NC)	84	10	11.9	0.595 ± 0.054
New York City (NYC)	42	5	11.9	0.440 ± 0.088
Ohio (OH)	103	13	12.6	0.671 ± 0.038
South Dakota (SD)	57	11	19.3	0.666 ± 0.066
Virginia (VA)	77	10	13.0	0.635 ± 0.050
Vermont (VT)	51	10	19.6	0.522 ± 0.083
European–Americans (EA)	927	30	3.2	0.637 ± 0.017
Arizona-Phoenix (AZ1)	56	12	21.4	0.688 ± 0.062
Arizona-Mesa (AZ2)	43	10	23.3	0.713 ± 0.067
Connecticut (CT)	85	13	15.3	0.578 ± 0.060
Florida (FL)	37	11	29.7	0.673 ± 0.085
North Carolina (NC)	87	12	13.8	0.568 ± 0.060
New York City (NYC)	42	13	31.0	0.818 ± 0.044
Ohio (OH)	99	15	15.2	0.660 ± 0.051
South Dakota (SD)	182	17	9.3	0.641 ± 0.036
Virginia (VA)	97	13	13.4	0.548 ± 0.058
Vermont (VT)	199	14	7.0	0.626 ± 0.037
Hispanic-Americans (HA)	479	27	5.6	0.786 ± 0.018
Arizona-Phoenix (AZ1)	109	15	13.8	0.792 ± 0.035
Arizona-Mesa (AZ2)	47	12	25.5	0.662 ± 0.076
Connecticut (CT)	90	19	21.1	0.792 ± 0.038
Florida (FL)	20	8	40.0	0.700 ± 0.109
New York City (NYC)	38	12	31.6	0.757 ± 0.067
Ohio (OH)	24	11	45.8	0.815 ± 0.072
South Dakota (SD)	42	13	31.0	0.812 ± 0.053
Virginia (VA)	92	20	21.7	0.817 ± 0.037
Vermont (VT)	17	9	52.9	0.868 ± 0.068
Native-Americans (NA)	398	18	4.5	0.775 ± 0.010
Apache	86	6	7.0	0.667 ± 0.032
Cheyenne	29	4	13.8	0.677 ± 0.069
Navajo	88	9	10.2	0.597 ± 0.031
Pima	19	3	15.8	0.608 ± 0.070
South Dakota	112	15	13.4	0.789 ± 0.025
South Dakota-Sioux	45	11	24.4	0.711 ± 0.058
Vermont	19	6	31.6	0.597 ± 0.122
Asian–Americans (SA)	62	12	19.4	0.848 ± 0.025
Arizona-Tucson	25	8	32.0	0.840 ± 0.038
New York City	37	11	29.7	0.857 ± 0.040

2.2. DNA typing

A set of 61 binary markers (SRY_{10831}, M91, M32, M6, M31, M13, M60, 50f2, M150, M152, P9, RPS4Y, M216, M217, P39, YAP, M174, M116, M125, SRY_{4064}, P2, M35, M78, P1, P14, M201, P15, M52, P19, P30, P37, p12f2a, M172, M12, M9, M70, M20, M5, M214, LLY22g, P43, Tat, M128, M175, M122, M134, M119, MSY2b, LINE1, P31, M95, M111, SRY_{465}, P27, P36, M3, M207, M173, M17, P25, M269) was typed using a previously described hierarchical protocol (e.g., [17,18,23,24]). Therefore, not every individual was typed for every marker. A tree of the 39 U.S. haplogroups observed in our sample is shown in Fig. 1 (note: haplogroups not present in our sample are not shown and,

thus, some of the above-mentioned markers are not represented in Fig. 1). Markers were typed using allele-specific PCR, restriction enzyme digest, or direct sequencing. Protocols and primer sequences for these assays were previously published [17,25]. We follow the terminological conventions recommended by the Y Chromosome Consortium [25] for naming NRY lineages. When no further downstream markers in the latest version of the YCC tree [19] were typed for this study, we considered the most derived marker that was typed to represent a haplogroup. Geographic origins of haplogroups were assigned based on inferences from published surveys [17–19,23,26–30].

2.3. Statistical analyses

Haplotype diversity was calculated following Nei [31]. We examined hierarchical genetic structuring based on an analysis of molecular variance (AMOVA) [32,33] as executed in ARLEQUIN (Version 2.0; [34]). AMOVA measures the partitioning of variance at different levels of population subdivision, giving rise to an analogue of F-statistics called Φ-statistics. Significance of the AMOVA analyses was assessed using 10,000 permutations. Populations were subdivided by self-described ethnicity (African-, European-, Hispanic-, Asian-, and Native-American) and by geographic location, including the Southwest (Arizona and New Mexico), Midwest (South Dakota and Ohio), Northeast (Vermont, Connecticut, and New York), and South (Virginia, North Carolina, and Florida). A pairwise genetic distance matrix that was generated using ARLEQUIN was used as input for an non-metric MDS analysis [35] using NTSYS (Exeter software). Subtraction analyses were performed on Y-STR haplotypes in the same database [12]. In these analyses, Y chromosomes inferred to have entered a particular population as a result of inter-ethnic admixture were removed from the dataset. Tests of pairwise population differentiation [16] using the R_{ST} statistic [36], which is calculated based on variance in allele frequencies and allele lengths, were performed also using ARLEQUIN. Significance of these tests was assessed using 100,000 permutations as described by Redd et al. [12].

3. Results

3.1. U.S. Y chromosome haplogroups

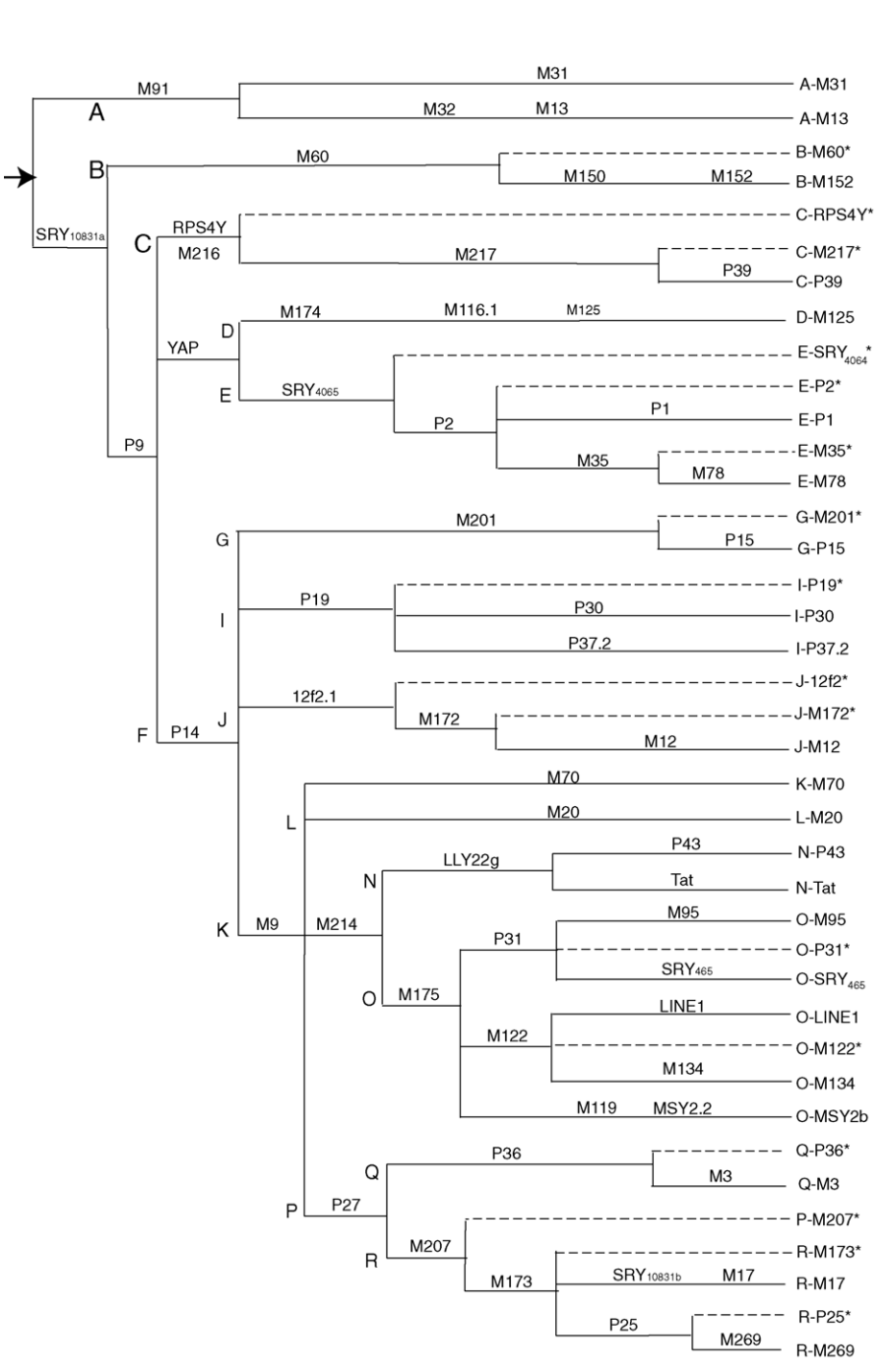
Fig. 1 depicts the evolutionary relationships among the 39 haplogroups found in our sample of 2517 individuals their frequencies in five U.S. ethnic groups. Haplogroups ranged in frequency from 0.04% to 37.8%. Only two haplogroups were present at a frequency greater than 10%, 13 were present at frequencies between 1 and 10%, and 24 were present at a frequency less than 1%. The most common haplogroup is R-M269 (37.8%), which is found in all of the ethnic groups. This haplogroup predominates in Western European populations [23]. E-P1, the second most frequent haplogroup in the U.S. (17.7%), is the most common haplogroup in West African populations [24]. It is found at high frequencies in our AA samples, and at lower frequencies in HA samples from the Eastern U.S. (Fig. 1). Three haplogroups that originate in Northern and Western European populations include I-P30 (6.1%), the third most common haplogroup in our U.S. sample, I-P19 (2.8%) and I-P37 (1.6%). Haplogroups that likely originate in Eastern and Southern European populations are also present in our U.S. database, including R-M17 (3.4%), E-M78 (2.4%), G-P15 (2.4%), and J-M172 (1.5%). The fourth and fifth most frequent haplogroups in our database, Q-P36 (5.9%) and Q-M3 (5.8%), along with C-P39 (1.5%), are founding Native-American Y chromosomes [30]. These haplogroups are frequent in our NA and HA samples, and are found at low frequency in our AA, EA, and SA samples. Asian-derived chromosomes, primarily in haplogroups O and N, are extremely rare in all but our SA sample.

3.2. Haplogroup diversity

The discrimination capacity of Y-SNPs is low, with only 39 haplogroups defined by the 61 SNPs surveyed here in the total sample of 2517 individuals (1.6%) (Table 1). Discrimination capacities generally run a little higher within ethnic groups (i.e., 3.7–19.4%), depending mostly on sample size. Haplogroup diversity ranges from a low of 0.440 in the NYC AA sample up to 0.868 in the VT HA sample (Table 1). The descending rank order of haplogroup diversity in the pooled ethnic groups is as follows: SA (0.848), HA (0.786), NA (0.775), EA (0.637), and AA (0.585). Native-American populations typically have low levels of haplogroup diversity [30], but here they are rather diverse, while the reverse is true for our AA sample. These results may be affected by ascertainment bias, by the choice of Y-SNPs, sample size differences, and/or admixture among ethnic groups.

3.3. MDS plot

Variation among the 38 U.S. populations can be seen in the MDS plot in Fig. 2. The low stress (0.09) and the close fit between the original distance matrix and a genetic distance matrix derived from the plot ($r = 0.98$) indicate that the MDS plot is a very good representation of the genetic distance matrix. Dotted circles are placed around populations from each ethnic group to illustrate that the populations cluster in groups that correspond with the five ethnic groups. The AA populations cluster to the top right of the plot, well separated from the other populations. Within the AA cluster, the FL sample is on the far right, while the OH sample is on the far left. The HA populations are found close to the center in the plot. Within the



Haplogroup Frequency (%)	U.S.	AA	EA	HA	NA	SA
A-M31	0.08	0.3				
A-M13	0.32	1.1	0.1			
B-M60*	0.20	0.8				
B-M152	0.44	1.5	0.1			
C-RPS4Y*	0.08			0.4		
C-M217*	0.12					4.8
C-P39	1.47		0.1		9.0	
D-M125	0.04					1.6
E-SRY4064*	1.63	5.1	0.2	1.0	0.3	
E-P2*	0.64	0.6	0.1	2.3		
E-P1	17.66	62.0	0.9	5.8	1.3	1.6
E-M35*	1.11	0.9	0.8	2.9	0.3	
E-M78	2.43	1.1	2.9	4.8	1.0	
G-M201*	0.08		0.2			
G-P15	2.43	0.9	3.6	4.4	0.3	
I-P19*	2.82	1.9	4.9	2.1	1.0	
I-P30	6.13	2.9	11.7	3.3	2.8	
I-P37.2	1.59	0.5	2.7	1.9	0.8	
J-12f2*	1.43	0.3	1.0	4.4	0.8	1.6
J-M172*	1.51	0.2	1.6	4.2	0.5	
J-M12	0.64	0.3	0.8	1.3	0.3	
K-M70	0.48	0.3	0.5	1.0		
L-M20	0.12		0.1	0.4		
N-P43	0.04		0.1			
N-Tat	0.08		0.1		0.3	
O-M95	0.28	0.3		0.2		6.5
O-P31*	0.12		0.1	0.2		1.6
O-SRY465	0.20					8.1
O-LINE1	0.40					16.1
O-M122*	0.48		0.3	0.6		9.7
O-M134	0.76			0.2		29.0
O-MSY2b	0.56		0.1	0.4		17.7
Q-P36*	5.93	0.2	0.6	3.8	31.2	
Q-M3	5.81		0.1	7.9	26.9	
P-M207*	0.16	0.3	0.1	0.2		
R-M173*	0.08		0.1	0.2		
R-M17	3.38	1.1	7.2	1.0	1.5	
R-P25*	0.60	0.5	0.6	1.0	0.3	
R-M269	37.79	17.3	58.3	43.8	21.9	1.6

Fig. 1. Maximum-parsimony tree of 39 Y chromosome haplogroups present in this survey along with their frequencies in five ethnic groups from the U.S. The root of the tree is denoted by an arrow. Major clades (i.e., A–R) are labeled with large capital letters to the left of each clade. Mutation names are given along the branches. The length of each branch is not proportional to the number of mutations or the age of the mutation. Dotted lines refer to internal nodes not defined by downstream markers (i.e., paragroups). The names of the 39 haplogroups observed in the present study are shown to the right of the branches. Haplogroup frequencies are shown on the far right for the total sample ($n = 2517$), African–Americans (AA, $n = 651$), European–Americans (EA, $n = 927$), Hispanic-Americans (HA, $n = 479$), Native-Americans (NA, $n = 398$), and Asian–Americans (SA, $n = 62$).

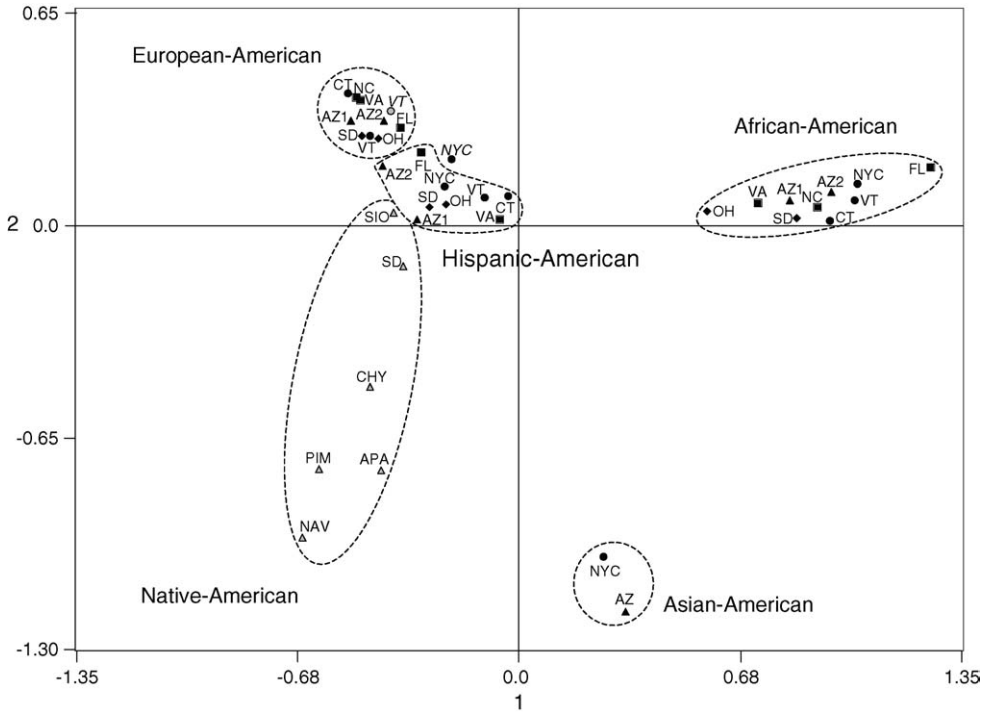


Fig. 2. MDS plot of 38 populations based on Φ_{ST} genetic distances. Population codes are the same as in Table 1. Symbols refer to the Southwest (solid triangles), Midwest (solid diamonds), Northeast (solid circles), and South (solid squares). Native-American samples are represented as gray triangles and a gray circle in the case of VT. Note that the NYC and the VT samples are outliers with respect to the dotted circles placed around the European–American and Native-American samples, respectively (shown in italics).

HA cluster, the AZ2 (Mesa) and FL samples are closer to the EA cluster. The EA populations form a tight cluster in the upper right of the plot adjacent to the HA cluster, with the exception of the NYC sample which is positioned closer to the HA cluster. The two SA populations cluster closely in the lower right quadrant of the plot. In contrast, the NA populations are found across a large area of the MDS plot, they transect both the upper and lower left quadrants of the plot. In fact, the SD and SIO samples are placed very close to the HA cluster, and the VT sample falls directly within the EA cluster.

3.4. AMOVA

When populations are divided into five ethnic groups most of the genetic variance (65.8%) is found within populations; a notable amount (32.3%) is found among ethnic groups; while only a small amount (2.0%) is found among populations within ethnic groups (Table 2). Separate AMOVA analyses within each of the ethnic groups show that only the NA group contains significant among-population-within-group variation (16.0%; $P < 0.001$). When populations are grouped into four geographic locations (i.e., Southwest, Midwest, Northeast, and South), we find similar results (Table 2).

3.5. Proportional African, European, Asian, and Native-American ancestry

The histograms in Fig. 3 illustrate dramatic variation in patterns of paternal ancestry among populations and ethnic groups. AA populations are composed mostly of African-derived Y chromosomes (mean ± S.D., 73.2 ± 8.7%); however, there is a consistent presence of European-derived Y chromosomes at frequencies ranging from 10% to 43% (26.4 ± 8.9%). European Y chromosomes are present in NA populations at a mean frequency of 35.3 ± 32.8%; however, there is a wide range of admixture rates, which show a West to East gradient. For example, European Y chromosomes are present in western populations at frequencies as low as 7% in the Apache and as high as 89% in VT. Hispanic-Americans are composed mostly of European Y chromosomes (77.8 ± 9.3%), but also have indigenous Native-American Y chromosomes (13.7 ± 7.4%) and African-derived Y chromosomes (6.6 ± 7.0%). Native-American Y chromosome frequencies range from 3% in CT to 18% in Mesa, AZ. The contribution of African Y chromosomes to HA populations varies by region, with the African component on the East Coast being higher (10.5 ± 6.4%) than that in the Southwest/Midwest (0.8 ± 1.0%) (t-test, $P = 0.019$). Admixture in European–Americans is notably low in all

Table 2
Analysis of molecular variance

Populations	Number of chromosomes	Number of populations	Number of groups	% Variance		
				Among groups	Among populations within groups	Within populations
African–Americans	651	10	1	–	1.4	98.6
		10	4[a]	0.4	1.0	98.6
European–Americans	927	10	1	–	0.2	99.8
		10	4[a]	–0.1	0.3	99.9
Hispanic-Americans	479	9	1	–	1.0	99.0
		9	4[a]	1.0	0.1	98.9
Native-Americans	398	7	1	–	**16.0**	**84.0**
		7	3[b]	17.9	**2.9**	**79.2**
Asian–Americans	62	2	1	–	0.0	100.0
Five ethnic groups	2517	38	5	**32.3**	**2.0**	**65.8**

Bolded numbers, $P < 0.001$.
[a] Southwest, Midwest, Northeast, and South.
[b] Southwest, Midwest, East (VT).

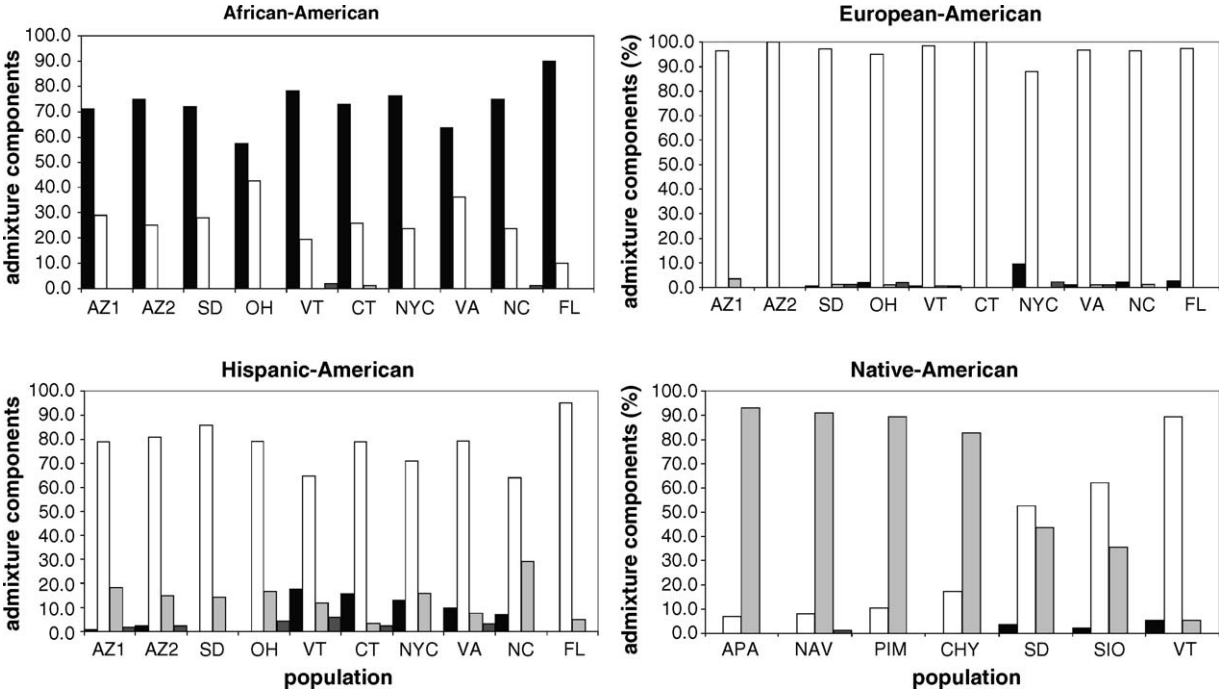


Fig. 3. Bar chart showing the relative proportions of Y chromosomes with African (black bar), European (open bar), Native-American (light gray bar), and Asian ancestry (dark gray bar). (A) African–Americans, (B) European–Americans, (C) Hispanic-Americans, and (D) Native-Americans. Population codes are the same as in Table 1. A small sample of Hispanics from North Carolina ($n = 15$) that was not included in other analyses is shown here.

samples surveyed except in our NYC sample, which has 10% African-derived Y chromosomes. Only two SA samples are examined here and both have very low frequencies of non-Asian haplogroups. The AZ sample is composed of 100% Asian Y chromosomes while the NYC sample has 2.7% African and 5.4% European Y chromosomes.

4. Discussion

These results support the conclusion that Y chromosomes are significantly differentiated among U.S. ethnic groupings, but not among populations within ethnic groups from different geographic regions within the U.S. Hence,

with the exception of Native-Americans, geographic origin of samples within a U.S. ethnic database is not critical. There is general correspondence between estimates of population structure parameters based on these 61 Y-SNPs and 11 core Y-STRs typed in the same samples [12]. However, the Y-SNP results indicate a greater proportion of total variation partitioned among ethnic groups (32.3%) than for Y-STRs (24.8%). This may be due to higher geographic specificity of Y-SNPs [20] and higher mutation rates of Y-STRs [37,38], which lead to much higher discrimination capacities and measures of Y chromosome diversity for Y-STR haplotypes [12] compared with Y-SNP haplogroups (Table 1). Similar to the case for Y-STRs [12], separate AMOVA analyses within each ethnic group show that only Native-Americans contain high levels of among population SNP haplogroup variation (Table 2). In contrast, $\leq$1% (n.s.) of the total SNP variation is partitioned among-populations-within-groups when considering only AA, EA, HA, and SA samples (data not shown).

4.1. Variation in paternal ancestry within and among U.S. ethnic groups

The set of 61 Y-SNPs employed here mark all 18 major haplogroups (A-R) on the Y chromosome haplogroup tree, as well as several sub-lineages providing information on the continental origins of Y chromosomes (Fig. 1). The geographic specificity of Y-SNP haplogroups allows direct estimates of the proportion of paternal genetic ancestry or admixture rates deriving from multiple source populations. We find that the proportion of chromosomes with African, European, Asian, and Native-American ancestry varies among populations within groups (Fig. 2). Regional variation in the proportion of European Y chromosomes in AA populations is apparent in the MDS plot in Fig. 2 (with OH as the most admixed population on the far left and FL as the least admixed population on the far right), as is regional variation in the frequency of African Y chromosomes in HA populations (with VA, CT, VT being placed on the right side of the HA cluster closest to the African–Americans). Native-Americans exhibit the largest regional variation in admixture rates, with European-derived Y chromosomes in Southwestern, Midwestern, and Eastern (VT) populations at frequencies of $8.5 \pm 1.8\%$, $44.1 \pm 23.7\%$, and 89.5%, respectively. The finding of high frequencies of European Y chromosomes in the VT, SD, and SIO Native-American samples helps to explain their position on the MDS plot.

It is interesting that the European and Native-American paternal contribution to HA populations is so consistent given that the term Hispanic does not refer to a defined geographic region, but can refer to individuals of Mexican, Puerto Rican, Cuban, Central/South American, or other Spanish culture ancestry. In fact, HA populations are known to have differing degrees of Spanish, Native-American, and African ancestry in different regions of the U.S. [10,39]. The higher frequency of African-derived Y chromosomes in the East is consistent with a greater contribution of Puerto Rican and Cuban Hispanics to East Coast U.S. populations, compared with a higher Mexican presence in the West [14]. In contrast to these Y chromosome results, both mtDNA and autosomal systems point to a much higher frequency of Native-American maternal lineages in HA populations, especially in Mexican Americans, and higher frequencies of African maternal lineages in Puerto Ricans and Cubans [10,39–41]. The larger European paternal contribution to HA populations likely reflects sex-specific biases in admixture rates for Hispanics, not necessarily while in the U.S. but in their source populations (e.g., [42]). Despite this regional variation, there were low levels of Hispanic Y-STR haplotype heterogeneity in our previous survey [12], as well as in the surveys of Kayser et al. [10] and Budowle et al. [43]. Thus, geographic origin of samples is not a critical factor in the construction of U.S. Hispanic Y-STR databases.

4.2. Effects of variation in paternal admixture on the structure of Y-STR haplotype diversity

One of our main objectives is to examine the extent to which variation in inter-ethnic admixture contributes to observed heterogeneity in Y-STR haplotype frequencies. As noted above, regional variation in the proportion of paternal ancestry may not always be due to local differences in rates of admixture (i.e., gene flow between ethnic groups after their arrival in the U.S.), but to different rates of inter-ethnic admixture in ancestral source populations, or to ancestral population structure in combination with non-random migration to the U.S. While previous studies revealed very little heterogeneity in Y-STR haplotype frequencies among populations within ethnic groups [10,12,43], cases of statistically significant differences between particular pairs of populations were observed in pairwise population differentiation (PPD) tests [10,12]. For example, in our Y-STR database [12], 3 of 45 comparisons between pairs of EA samples, and 2 of 36 comparisons between pairs of HA samples, were statistically significant. All three EA comparisons involved our sample from New York City, which differed from our Connecticut, Virginia, and North Carolina EA samples. Both HA comparisons involved our Mesa (Arizona) sample, which differed from our Connecticut and Virginia HA samples. Similarly, PPD tests performed by Kayser et al. [10] on their database of 1705 haplotypes based on nine Y-STRs revealed heterogeneity between their Texas European–American sample and other European–Americans, and between their Texas Hispanic-American sample and other Hispanic-Americans. They concluded that the significant heterogeneity involving these two samples reflected chance rather than any true biological differences.

We ask whether the statistically significant PPD tests involving Y-STR haplotypes in our Hispanic samples [12]

can be explained by variable frequencies of African Y chromosomes. When we remove the 36 African Y chromosomes identified by Y-SNPs from our HA Y-STR database and repeat the PPD tests, we find no significant difference between any of the 36 pairs of HA samples (data not shown). This suggests that regional variation in settlement patterns of Hispanics, for example, from the Caribbean or from Mexico, could cause regional heterogeneity in frequencies of Y-STR haplotypes. However, current data reveal only minor effects on Y chromosome variation [10,12,43].

Next, we test whether variable frequencies of African Y chromosomes in EA populations leads to the significant heterogeneity in Y-STR haplotypes in our NYC, CT, NC, and VA samples [12]. Upon removal of the 13 African-derived chromosomes from our EA database, we still find that our NYC sample is different from NC and VA (but not CT). We note that NYC has the lowest frequency of Northwestern European signature haplogroups R-M269 (35.7% for NYC versus $58.7 \pm 5.0\%$ for other EA samples) and I-P30 (4.8% for NYC versus $11.3 \pm 2.8\%$ for other EA samples), and the highest frequency of the Eastern European signature haplogroup, R-M17 (23.8% for NYC versus $7.4 \pm 4.7\%$ for other EA samples). Thus, we hypothesize that descent from a structured European source population (with non-random migration to the U.S.) underlies the observed Y-STR heterogeneity. To address this hypothesis, we analyze four western (England, Ireland, France, and Germany) and three eastern (Poland, Hungary, and Russia) European population samples that are potential sources for the EA population. We find statistically significant population structure in Europe, with 10.9% of the total Y-STR haplotype variance partitioned between Western and Eastern European samples (data not shown). Interestingly, our NYC sample itself is differentiated from three of the four Western European samples (England, Ireland, and France) and not from any of the three Eastern European samples. Therefore, we conclude that population structure in Europe is a potential factor leading to heterogeneity among European–Americans.

Finally, we wanted to know whether regional variation in admixture among NA populations plays an important role in structuring Native-American Y chromosome variation. When we remove the 124 European-derived Y chromosomes from our NA database, we find that AMOVA still results in significant differences in Y-STR haplotype frequencies among Western, Midwestern, and Eastern NA populations (data not shown). The percent of among group variance (9.4%) is only slightly lower than in the case when all (i.e., admixed and indigenous) Y-STR haplotypes are included in the analysis (11.1%). We conclude that NA Y chromosomes are differentiated with respect to geography and/or tribal affiliation, regardless of the degree of admixture with European–American males. This is consistent with a long history of genetic drift as a result of small effective population sizes of Native-American tribal groups, endogamy, isolation, and founder effects [30].

5. Conclusion

The population of the U.S. is comprised of people with ancestry tracing to Africa, Europe, Asia, the Pacific, and the Americas. The results presented here indicate that continental origin rather than current location in the U.S. determines major patterns of Y chromosome variation for most ethnic groups. Apparently, intermarriage among groups has not eliminated inter-ethnic genetic structure. Despite the potential for admixture to erode population structure, the 2000 U.S. Census revealed that only 2.4% of all respondents reported being derived from two or more "racial" groups (excluding Hispanics; [14]). Nonetheless, there is a body of literature indicating substantial mixing among some U.S. ethnic groups [39,44–48]. Continuing migration and admixture among ethnic groups may eventually reduce population structure to the point where we will no longer need to construct separate forensic databases in the U.S. In the meantime, more research is needed for several reasons. While simple methods for adjusting for minor levels of population structure among Hispanic populations should be sufficient for correcting haplotype frequency estimates [15], the finding of significant differences in frequencies of African-derived Y chromosomes among Hispanic samples raises potential concerns for the proper construction of Hispanic databases. In addition, more analyses of European–American samples from various parts of the U.S. that are known to have different ethnic compositions will help to determine how frequent we expect outliers (such as NYC) in Y chromosome databases. While AA populations seems to show the least amount of among population within group variation of any ethnic group surveyed, additional research should help to understand the underlying causes for the apparent homogeneity [10]. Additional Y-STR surveys of putative African source populations will help to determine whether a lack of structure in the putative source population, along with similar admixture rates in the U.S., can explain the observed homogeneity among AA subpopulations. Finally, more work is needed to construct appropriate Y-STR databases of Native-American populations.

Acknowledgements

We thank John Butler for comments on an earlier version of the manuscript. This research was supported by the National Institute of Justice Grant 2000-IJ-CX-K006 to M.F.H. We thank our colleagues from crime laboratories who supplied the U.S. population samples.

References

[1] M. Kayser, A. Caglia, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, K. Honda, M. Jobling, M. Krawczak, K. Leim, S. Meuser, E.

Meyer, W. Oesterreich, A. Pandya, W. Parson, G. Penacino, A. Perez-Lezaun, A. Piccinini, M. Prinz, C. Schmitt, L. Roewer, et al., Evaluation of Y-chromosomal STRs: a multicenter study, Int. J. Legal Med. 110 (1997) 125–133.

[2] V. Pascali, M. Dobosz, B. Brinkmann, Coordinating Y-chromosomal STR research for the courts, Int. J. Legal Med. 112 (1998) 1.

[3] L. Roewer, M. Kayser, P. de Knijff, K. Anslinger, A. Betz, A. Caglia, D. Corach, S. Furedi, L. Henke, M. Hidding, H.J. Kargel, R. Lessig, M. Nagy, V.L. Pascali, W. Parson, B. Rolf, C. Schmitt, R. Szibor, J. Teifel-Greding, M. Krawczak, A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males, Forensic Sci. Int. 114 (2000) 31–43.

[4] D. Corach, L. Filgueira Risso, M. Marino, G. Penacino, A. Sala, Routine Y-STR typing in forensic casework, Forensic Sci. Int. 118 (2001) 131–135.

[5] L. Roewer, M. Krawczak, S. Willuweit, M. Nagy, C. Alves, A. Amorim, K. Anslinger, C. Augustin, A. Betz, E. Bosch, A. Caglia, A. Carracedo, D. Corach, A.F. Dekairelle, T. Dobosz, B.M. Dupuy, S. Furedi, C. Gehrig, L. Gusmao, J. Henke, L. Henke, M. Hidding, C. Hohoff, B. Hoste, M.A. Jobling, H.J. Kargel, P. de Knijff, R. Lessig, E. Liebeherr, M. Lorente, B. Martinez-Jarreta, P. Nievas, M. Nowak, W. Parson, V.L. Pascali, G. Penacino, R. Ploski, B. Rolf, A. Sala, U. Schmidt, C. Schmitt, P.M. Schneider, R. Szibor, J. Teifel-Greding, M. Kayser, Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes, Forensic Sci. Int. 118 (2001) 106–113.

[6] A. Betz, G. Bassler, G. Dietl, X. Steil, G. Weyermann, W. Pflug, DYS STR analysis with epithelial cells in a rape case, Forensic Sci. Int. 118 (2001) 126–130.

[7] K. Honda, L. Roewer, P. de Knijff, Male DNA typing from 25-year-old vaginal swabs using Y chromosomal STR polymorphisms in a retrial request case, J. Forensic Sci. 44 (1999) 868–872.

[8] M. Prinz, A. Ishii, A. Coleman, H.J. Baum, R.C. Shaler, Validation and casework application of a Y chromosome specific STR multiplex, Forensic Sci. Int. 120 (2001) 177–188.

[9] A.F. Dekairelle, B. Hoste, Application of a Y-STR-pentaplex PCR (DYS19, DYS389I and II DYS390 and DYS393) to sexual assault cases, Forensic Sci. Int. 118 (2001) 122–125.

[10] M. Kayser, S. Brauer, H. Schadlich, M. Prinz, M. Batzer, P. Zimmerman, B. Boatin, M. Stoneking, Y chromosome STR haplotypes and the genetic structure of U.S. populations of African, European, and Hispanic ancestry, Genome Res. 13 (2003) 624–634.

[11] A. Redd, A. Agellon, V. Kearney, V. Contreras, T. Karafet, H. Park, P. de Knijff, J. Butler, M. Hammer, Forensic value of 14 novel STRs on the human Y chromosome, Forensic Sci. Int. 130 (2002) 97–111.

[12] A.J. Redd, V.F. Chamberlain, V.C. Kearney, D. Stover, T.M. Karafet, K. Calderon, M.F. Hammer, Genetic structure among 38 populations from the United States based on 11 U.S. core Y-chromosome STRs, J. Forensic Sci. (2006), in press.

[13] M.F. Hammer, S.L. Zegura, The role of the Y chromosome in human evolutionary studies, Evol. Anthropol. 5 (1996) 116–134.

[14] U.S.C. Bureau, 2000. http://www.census.gov/main/www/cen2000.html.

[15] NRC, The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, DC, 1996.

[16] M. Raymond, F. Rousset, An exact test for population differentiation, Evolution 49 (1995) 1280–1283.

[17] M.F. Hammer, T.M. Karafet, A.J. Redd, H. Jarjanazi, S. Santachiara-Benerecetti, H. Soodyall, S.L. Zegura, Hierarchical patterns of global human Y-chromosome diversity, Mol. Biol. Evol. 18 (2001) 1189–1203.

[18] P.A. Underhill, G. Passarino, A.A. Lin, P. Shen, M. Mirazon Lahr, R.A. Foley, P.J. Oefner, L.L. Cavalli-Sforza, The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations, Ann. Hum. Genet. 65 (2001) 43–62.

[19] M.A. Jobling, C. Tyler-Smith, The human Y chromosome: an evolutionary marker comes of age, Nat. Rev. Genet. 4 (2003) 598–612.

[20] M.A. Jobling, C. Tyler-Smith, Fathers and sons: the Y chromosome and human evolution, Trends Genet. 11 (1995) 449–456.

[21] G. Bertorelle, L. Excoffier, Inferring admixture proportions from molecular data, Mol. Biol. Evol. 15 (1998) 1298–1311.

[22] D. Paetkau, W. Calvert, I. Stirling, C. Strobeck, Microsatellite analysis of population structure in Canadian polar bears, Mol. Ecol. 4 (1995) 347–354.

[23] F. Cruciani, P. Santolamazza, P.D. Shen, V. Macaulay, P. Moral, A. Olckers, D. Modiano, S. Holmes, G. Destro-Bisol, V. Coia, D.C. Wallace, P.J. Oefner, A. Torroni, L.L. Cavalli-Sforza, R. Scozzari, P.A. Underhill, A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes, Am. J. Hum. Genet. 70 (2002) 1197–1214.

[24] E.T. Wood, D.A. Stover, C. Ehret, G. Destro-Bisol, G. Spedini, H. McLeod, L. Louie, M. Bamshad, B.I. Strassmann, H. Soodyall, M.F. Hammer, Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes, Eur. J. Hum. Genet. 13 (2005) 867–876.

[25] YCC, A nomenclature system for the tree of human Y-chromosomal binary haplogroups, Genome Res. (2002) 339–348.

[26] C. Cinnioglu, R. King, T. Kivisild, E. Kalfoglu, S. Atasoy, G.L. Cavalleri, A.S. Lillie, C.C. Roseman, A.A. Lin, K. Prince, P.J. Oefner, P. Shen, O. Semino, L.L. Cavalli-Sforza, P.A. Underhill, Excavating Y-chromosome haplotype strata in Anatolia, Hum. Genet. 114 (2004) 127–148.

[27] S. Rootsi, C. Magri, T. Kivisild, G. Benuzzi, H. Help, M. Bermisheva, I. Kutuev, L. Barac, M. Pericic, O. Balanovsky, A. Pshenichnov, D. Dion, M. Grobei, L.A. Zhivotovsky, V. Battaglia, A. Achilli, N. Al-Zahery, J. Parik, R. King, C. Cinnioglu, E. Khusnutdinova, P. Rudan, E. Balanovska, W. Scheffrahn, M. Simonescu, A. Brehm, R. Goncalves, A. Rosa, J.P. Moisan, A. Chaventre, V. Ferak, S. Furedi, P.J. Oefner, P. Shen, L. Beckman, I. Mikerezi, R. Terzic, D. Primorac, A. Cambon-Thomsen, A. Krumina, A. Torroni, P.A. Underhill, A.S. Santachiara-Benerecetti, R. Villems, O. Semino, Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe, Am. J. Hum. Genet. 75 (2004) 128–137.

[28] O. Semino, C. Magri, G. Benuzzi, A.A. Lin, N. Al-Zahery, V. Battaglia, L. Maccioni, C. Triantaphyllidis, P. Shen, P.J. Oefner, L.A. Zhivotovsky, R. King, A. Torroni, L.L. Cavalli-

Sforza, P.A. Underhill, A.S. Santachiara-Benerecetti, Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area, Am. J. Hum. Genet. 74 (2004) 1023–1034.

[29] O. Semino, G. Passarino, P.J. Oefner, A.A. Lin, S. Arbuzova, L.E. Beckman, G. De Benedictis, P. Francalacci, A. Kouvatsi, S. Limborska, M. Marcikiae, A. Mika, B. Mika, D. Primorac, A.S. Santachiara-Benerecetti, L.L. Cavalli-Sforza, P.A. Underhill, The genetic legacy of Paleolithic Homo sapiens in extant Europeans: a Y chromosome perspective, Science 290 (2000) 1155–1159.

[30] S.L. Zegura, T.M. Karafet, L.A. Zhivotovsky, M.F. Hammer, High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas, Mol. Biol. Evol. 21 (2004) 164–175.

[31] M. Nei, Molecular Evolutionary Genetics, Columbia University Press, New York, 1987.

[32] L. Excoffier, P.E. Smouse, J.M. Quattro, Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data, Genetics 131 (1992) 479–491.

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

[34] S. Schneider, J.-M. Kueffer, D. Roessli, L. Excoffier, ARLEQUIN: A Software for Population Genetic Analysis, Genetics and Biometry Laboratory, Stanford University, Geneva, 2000.

[35] J.B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964) 1–27.

[36] M. Slatkin, A measure of population subdivision based on microsatellite allele frequencies, Genetics 139 (1995) 457–462.

[37] B. Bonne-Tamir, M. Korostishevsky, A.J. Redd, Y. Pel-Or, M.E. Kaplan, M.F. Hammer, Maternal and paternal lineages of the Samaritan isolate: mutation rates and time to most recent common male ancestor, Ann. Hum. Genet. 67 (2003) 153–164.

[38] L. Jin, R. Chakraborty, Population structure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics, Heredity 74 (Pt 3) (1995) 274–285.

[39] D. Merriwether, S. Huston, S. Iyengar, R. Hamman, J. Norris, S. Shetterly, M. Kamboh, R. Ferrell, Mitochondrial versus nuclear admixture estimates demonstrate a past history of directional mating, Am. J. Phys. Anthropol. 102 (1997) 153–159.

[40] C.L. Hanis, D. Hewett-Emmett, T.K. Bertin, W.J. Schull, Origins of U.S. Hispanics. Implications for diabetes, Diabetes Care 14 (1991) 618–627.

[41] B.M. Chakraborty, M.E. Fernandez-Esquer, R. Chakraborty, Is being Hispanic a risk factor for non-insulin dependent diabetes mellitus (NIDDM)? Ethn. Dis. 9 (1999) 278–283.

[42] N.R. Mesa, M.C. Mondragon, I.D. Soto, M.V. Parra, C. Duque, D. Ortiz-Barrientos, L.F. Garcia, I.D. Velez, M.L. Bravo, J.G. Munera, G. Bedoya, M.C. Bortolini, A. Ruiz-Linares, Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: pre- and post-Columbian patterns of gene flow in South America, Am. J. Hum. Genet. 67 (2000) 1277–1286.

[43] B. Budowle, M. Adamowicz, X.G. Aranda, C. Barna, R. Chakraborty, D. Cheswick, B. Dafoe, A. Eisenberg, R. Frappier, A.M. Gross, C. Ladd, H.S. Lee, S.C. Milne, C. Meyers, M. Prinz, M.L. Richard, G. Saldanha, A.A. Tierney, L. Viculis, B.E. Krenke, Twelve short tandem repeat loci Y chromosome haplotypes: genetic analysis on populations residing in North America, Forensic Sci. Int. 150 (2005) 1–15.

[44] E. Parra, R. Kittles, G. Argyropoulos, C. Pfaff, K. Hiester, C. Bonilla, N. Sylvester, D. Parrish-Gause, W. Garvey, L. Jin, P. McKeigue, M. Kamboh, R. Ferrell, W. Pollitzer, M. Shriver, Ancestral proportions and admixture dynamics in geographically defined African–Americans living in South Carolina, Am. J. Phys. Anthropol. 114 (2001) 18–29.

[45] E. Parra, A. Marcini, J. Akey, J. Martinson, M. Batzer, R. Cooper, T. Forrester, D. Allison, R. Deka, R. Ferrell, M. Shriver, Estimating African–American admixture proportions by use of population-specific alleles, Am. J. Hum. Genet. 63 (1998) 1839–1851.

[46] E. Foster, M. Jobling, P. Taylor, P. Donnelly, P. de Knijff, R. Mieremet, T. Zerjal, C. Tyler-Smith, Jefferson fathered slave's last child, Nature 396 (1998) 27–28.

[47] C. Bonilla, E.J. Parra, C.L. Pfaff, S. Dios, J.A. Marshall, R.F. Hamman, R.E. Ferrell, C.L. Hoggart, P.M. McKeigue, M.D. Shriver, Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping, Ann. Hum. Genet. 68 (2004) 139–153.

[48] C. Bonilla, M. Shriver, E. Parra, A. Jones, J. Fernandez, Ancestral proportions and their association with skin pigmentation and bone mineral density in Puerto Rican women from New York City, Hum. Genet. 115 (2004) 57–68.