# Evaluating self-declared ancestry of U.S. Americans using autosomal, Y-chromosomal and mitochondrial DNA

Oscar Lao, Peter M Vallone, Michael D Coble, Toni M. Diegoli, Mannis van Oven, Kristiaan J van Der Gaag, Peter de Knijff, Manfred Kayser

Human Mutation

# Evaluating self-declared ancestry of U.S. Americans using autosomal, Y-chromosomal and mitochondrial DNA



| | |
|---|---|
| Journal: | *Human Mutation* |
| Manuscript ID: | humu-2010-0254.R1 |
| Wiley - Manuscript type: | Mutation in Brief |
| Date Submitted by the Author: | 24-Aug-2010 |
| Complete List of Authors: | Lao, Oscar; ErasmusMC, Forensic Molecular Biology<br>Vallone, Peter; National Institute of Standards and Technology, Biochemical Science Division<br>Coble, Michael; Research Section, The Armed Forces DNA Identification Laboratory<br>Diegoli, Toni; Research Section, The Armed Forces DNA Identification Laboratory<br>van Oven, Mannis; ErasmusMC, Forensic Molecular Biology<br>van der Gaag, Kristiaan; Leiden University Medical Center, Department of Human and Clinical Genetics<br>de Knijff, Peter; Leiden University Medical Center, Department of Human and Clinical Genetics<br>Kayser, Manfred; ErasmusMC, Forensic Molecular Biology |
| Key Words: | U.S. Americans, genetic ancestry, self-declared ancestry, ASM, AIM, Y-chromosome, NRY, mtDNA |
| | |

HUMAN MUTATION Mutation in Brief #____ (20XX) Online

MUTATION IN BRIEF

HUMAN MUTATION

# Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA

Oscar Lao[1,+], Peter M. Vallone[2,+], Michael D. Coble[3], Toni M. Diegoli[3], Mannis van Oven[1], Kristiaan J. van der Gaag[4], Peter de Knijff[4], and Manfred Kayser[1*]

*1- Department of Forensic Molecular Biology, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands*
*2- National Institute of Standards and Technology, Biochemical Science Division, 100 Bureau Drive, Mail Stop 8311, Gaithersburg, MD 20899-8311, United States of America*
*3- Research Section, The Armed Forces DNA Identification Laboratory, Rockville, Maryland, United States of America*
*4- Department of Human and Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands*
+ These authors contributed equally to this work

*Correspondence to Prof. Dr. Manfred Kayser, Department of Forensic Molecular Biology, Erasmus University Medical Center Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands; E-mail: m.kayser@erasmusmc.nl.

Contract grant sponsor: Netherlands Forensic Institute; Netherlands Genomics Initiative (NGI) / Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands (FGCN).

Short Title: Bio-geographic ancestry estimation in U.S. Americans.

Communicated by <Please don't enter>

ABSTRACT: The current U.S. population represents an amalgam of individuals originating mainly from four continental bio-geographic ancestries (Africa, Europe, Asia and America). We have analyzed paternally, maternally and bi-parentally inherited DNA markers sensitive for indicating continental genetic ancestry in all four major U.S. American groups. We found that self-declared U.S. Hispanics and U.S. African Americans tend to show variable degrees of continental genetic admixture among the three genetic systems, with evidence for a marked sex-biased admixture history. Moreover, we observed significant regional variation across the country in genetic admixture. In contrast, self-declared U.S. European and U.S. Asian Americans were genetically more homogeneous at the continental ancestry level. Two autosomal ancestry-sensitive markers located in skin pigmentation candidate genes showed significant differences in self-declared U.S. African Americans or U.S. European Americans, relative to their assumed parental populations from Africa or Europe. This provides genetic support for the importance of skin color in the complex process of ancestry identification. 

Received <date>; accepted revised manuscript <date>.

KEY WORDS: U.S. Americans, genetic ancestry, self-declared ancestry, ASM, AIM, Y-chromosome, NRY, mtDNA

## INTRODUCTION

The current U.S. American population is particularly interesting for studying bio-geographic ancestry, as it represents an amalgam of individuals who originate from at least four major continental regions that (at least potentially) started to admix at different time scales starting with the European colonization of North America. The four most frequently self-assigned clusters by U.S. Americans according to the U.S. Census Bureau (2008) are white (U.S. European), black (U.S. African), Asian (U.S. Asian) and Hispanic / Latinos (U.S. Hispanic). It should be noticed, however, that such classification mixes bio-geographic ancestry, sociological and cultural, respectively linguistic variables. For example, individuals self-defined as U.S. Hispanics share cultural aspects, such as the Spanish mother tongue, but can be of different bio-geographic ancestry reflecting the more than 500 years of admixture history between Native Americans, Europeans and Africans in the Americas (Salazano and Bortolini, 2002). Similarly, self-declared U.S. Africans generally carry some degree of European genetic ancestry which in particular cases can reach more than 80% of the total ancestry (Sinha, et al., 2006). Finally, additional sub-continental population substructure can also be detected within self-identified groups, such as within self-declared U.S. European (Campbell, et al., 2005), U.S. African (Tishkoff, et al., 2009; Zakharia, et al., 2009) and U.S. Hispanic (Wang, et al., 2008), as genetic heterogeneity within the respective parental populations has also been observed (Jakobsson, et al., 2008; Lao, et al., 2008; Li, et al., 2008; Novembre, et al., 2008). In the present study we have analyzed the bio-geographic ancestry of U.S. Americans with self-declared African, European, Asian and Hispanic ancestry, respectively, using uniparental non-recombining part of the human Y-chromosome (NRY) and mtDNA single nucleotide polymorphisms (SNPs) as well as carefully ascertained autosomal SNPs. All genetic markers used were ascertained to be sensitive for indicating bio-geographic ancestry on the level of the four continental regions (Africa, Europe, Asia, and America) expected to have contributed to the current U.S. population. Very few previous studies have analyzed all three genetic systems in at least one of these U.S. groups (Parra, et al., 1998; Lind, et al., 2007; Stefflova, et al., 2009). As far as we know, our study represents the first of its kind combining suitable ancestry-sensitive markers from all three genetic systems to detect separately patrilineal, matrilineal and bi-parental genetic ancestry in all four major U.S. American groups.

## MATERIALS AND METHODS

### Samples

Anonymous liquid blood or buccal swab samples from a total of 664 U.S. individuals were obtained from Interstate Blood Bank, Inc. (Memphis, TN), Millennium Biotech, Inc. (Ft. Lauderdale, FL) and DNA Diagnostics Center (Fairfield, OH). Among them, 246 were self-declared U.S. African Americans, 127 were self-declared U.S. Hispanic Americans, and 245 were self-declared U.S. European Americans from Temple and Killeen, TX, Louisville, KY, Baltimore, MD, Philadelphia, PA, Memphis, TN and Miami, FL and 46 were self-declared U.S. Asian Americans from the Fairfield, OH source. Each sample was examined with 15 autosomal short tandem repeats and the amelogenin sex-typing marker using the AmpFlSTR Identifiler kit (Applied Biosystems, Foster City, CA) to verify that each sample was unique (Butler, et al., 2003; Decker, et al., 2008). In addition to the U.S. American samples, autosomal markers were also genotyped in the Human Genome Diversity Project- Centre d'Etude du Polymorphisme Humain (HGDP-CEPH) samples (Cann, et al., 2002). From those, four groups i.e. i) Sub-Saharan Africans (Bantu, Biaka Pygmies, Mandenka, Mbuti Pygmies, San, Yoruba); ii) East Asians (Cambodian, Dai, Daur, Han, Hazara, Hezhen, Japanese, Lahu, Miaozu, Mongola, Naxi, Oroqen, She, Tu, Tujia, Uygur, Xibo, Yakut, Yizu); iii) Eurasians (Adygei, Basque, Bergamo, French, Orcadian, Russian, Sardinian, Tuscan); and iv) Native Americans (Colombian, Karitiana, Maya, Pima, Surui) were used as parental groups in some of the statistical analyses.

### Autosomal DNA analysis

Tweny four autosomal single nucleotide polymorphisms (SNPs): rs1876482, rs2179967, rs1048610, rs1371048, rs1478785, rs1369290, rs952718, rs1405467, rs1344870, rs1391681, rs1461227, rs1907702, rs2052760, rs714857,

rs721352, rs722869, rs926774, rs1448484, rs1667751, rs1858465, rs1465648, rs16891982, rs1808089, rs3843776 were genotyped via two SNaPshot multiplex reactions as described in detail in the supplementary material 1. These SNPs were ascertained to be ancestry-sensitive on the continental level as described in detail elsewhere (Lao, et al., 2006; Lao, et al., 2007; Kersbergen, et al., 2009; Corach, et al., 2010). In brief, Affymetrix 10K SNP data in 76 human individuals from 21 worldwide sampling localities from the Y-Chromosome Consortium (YCC) panel were analyzed using the informativeness of ancestry statistic ($I_n$; (Rosenberg, et al., 2003)) and applying a genetic algorithm to select a minimal set of markers that maximized the amount of ancestry information for differentiating four continental populations (Sub-Saharan Africa, Eurasia, East Asia and America) (Lao, et al., 2006). In parallel, a single population $F_{ST}$ (Weir and Cockerham, 1984) strategy was applied to ascertain markers that differentiate each population (Kersbergen, et al., 2009). In addition, SNPs were added from 3 genes associated with variation in skin pigmentation showing large frequency differences between Europeans, Africans and East Asian ancestry and for which evidence of positive selection was established (Lao, et al., 2007). The current set of 24 ancestry-sensitive markers (ASMs) was obtained by ascertaining from the pooled data the set of SNPs that maximizes the $I_n$ statistic considering four continental groups.

#### Mitochondrial DNA analysis

The entire mtDNA control region [range 16024-576] was amplified and sequenced using an automated, high-throughput, redundant sequencing and review strategy as described in (Irwin, et al., 2007). Sequence assembly and confirmation was performed independently by two different analysts, and followed by electronic data transfer to a secured laboratory information management system (LIMS) for sequence verification. The raw data was then exported to a second laboratory (the European DNA Profiling Group (EDNAP) mtDNA Population Database (EMPOP); (Parson and Dur, 2007)) for additional review and quality control examination. Control region haplotypes for the self-described African American (Diegoli, et al., 2009) and Hispanic (Saunier, et al., 2008) samples have been published previously, and the sequences, along with those generated here for European Americans and Asian Americans have all been deposited in GenBank under accession numbers: DQ906460–DQ906701 and DQ906703–DQ906708 (African Americans), DQ906175-DQ906459 (European Americans), EU014897–EU015024 (Hispanics), and HM214959-HM215005 (Asian Americans). MtDNA haplogroup assignment of the samples was conducted using a multitude of references found within the reference section of (Diegoli, et al., 2009) for the African American samples, (Saunier, et al., 2008) for the Hispanic samples, (Irwin, et al., 2008) for the European American samples, and (Irwin, et al., 2009) for the Asian American samples, and checked against the most recent human mtDNA tree at http://www.phylotree.org (van Oven and Kayser, 2009). In those cases where haplogroup assignment based upon sequence polymorphisms in the control region was ambiguous, additional sequencing of coding region SNPs was performed as described elsewhere (Just, et al., 2008). The continental region of geographic origin of the haplogroups was assumed from published mtDNA data (Richards, et al., 1998; Macaulay, et al., 1999; Finnila, et al., 2001; Kivisild, et al., 2006; Kong, et al., 2006; Achilli, et al., 2008; Behar, et al., 2008), and is provided for all mtDNA haplogroups observed in this study in the supplementary material 4.

#### Y-chromosomal DNA analysis

Variation of the NRY was identified by means of 42 NRY-SNPs in total. Twenty four NRY-SNPs were genotyped in all samples (including: SRY 1532, M91, M168, M145, M174, 12f2, M96, M213, M201, M69, M52, M170, M172, M9, M20, M106, M214, Tat, M175, M45, MEH2, M207, M269, and M124).

Aiming to maximize continental differentiation of haplogroup origins we additionally genotyped 18 additional SNPs among samples identified as belonging to haplogroup E (M33, P2, M2, M154, M191, M215, M35, M78, V12, M224, V32, V13, V22, M81, M123, M281, V6, and M75). A single multiplex PCR and SNaPshot assay using the principle of primer extension was designed for the core set of 24 NRY-SNPs as described elsewhere (Corach, et al., 2010). Genotyping of the additional 18 NRY SNPs for subtyping of haplogroup E was performed in a multiplex, designed in a similar way as described for the core set of 24 NRY-SNPs, the only exception being a final MgCl2-concentration of 3mM in the multiplex PCR. PCR-product sizes ranged from 76-150 bp. Sequences and concentrations of the primers used in the monoplex and multiplex PCR and extension reactions are provided in supplementary material 2 and a phylogenetic tree of the NRY-SNPs used is in the supplementary material 3. NRY haplogroups were derived from genotyping of Y-SNPs using the marker phylogeny as described elsewhere (Karafet, et al., 2008). The continental region of geographic origin of the haplogroups was assumed from

published NRY data (Semino, et al., 2000; Bortolini, et al., 2003; Jobling and Tyler-Smith, 2003; Luis, et al., 2004; Cruciani, et al., 2007), and is provided for all NRY haplogroups observed in this study in the supplementary material 5.

### Statistical analyses

Suitability of the 24 ascertained SNPs to recover continental ancestry was checked by means of performing a STRUCTURE analysis (Pritchard, et al., 2000) in the HGDP-CEPH panel. We increased the number of groups from K=2 to K=6 under the Admixture model with a burn-in of 100,000 simulations and retaining the next 100,000. Five runs were performed for each K. For the estimation of the parental ancestry of the U.S. samples, a STRUCTURE analysis considering four parental populations (Native Americans, East Asians, Eurasians, and Sub-Saharan Africans from HGDP-CEPH) based on expected continental ancestry was used. Ten thousand simulations were used as burn-in and the next 10,000 simulations retained for admixture estimates. Reproducibility of results was checked by repeating 10 times the same analyses, obtaining in all cases similar values of admixture from the parental populations. Bar plot was performed from the STRUCTURE estimations with Distruct software 1.1 (Rosenberg, 2004). Differences in the amount of ancestry were tested in regions with more than 10 sampled individuals by means of a Kruskal-Wallis test. In particular, it was computed for the African component in U.S. Africans (regions = Baltimore (n = 34), Louisville (n = 21), Memphis (n = 41), Miami (n = 25), Philadelphia (n = 104), Temple (n = 17)) and for the Native American component in U.S. Hispanics (regions = Miami (n=61), Temple (n=29), Killeen (n=17), Philadelphia (n=13)). Additionally, we compared the genetic clustering of U.S. individuals with self-identified ethnicity by means of a STRUCTURE analysis assuming no admixture between the inferred clusters and 4 populations (Tang, et al., 2005). An identical by state distance matrix between all pairs of individuals including parental HGDP-CEPH populations was computed considering the 24 SNPs and was used to compute a non parametric multidimensional scaling (MDS) (Kruskal and Wish, 1990) with the package isoMDS of the R software (R Development Core Team, 2006) specifying 3 dimensions. When the distance between two individuals was 0, a small quantity of 0.001 was added.

$I_n$ statistic was computed for each of the 24 ASMs using as populations: self-declared U.S. European and the Sub-Saharan African HGDP-CEPH population cluster (set A), self-declared U.S. African and HGDP-CEPH European group (set B), and Sub-Saharan African HGDP-CEPH population cluster and HGDP-CEPH European group (set C). A linear regression was performed with SPSS (SPSS, 2003) between set A and C, and between set B and C; the SNPs falling out of the prediction with a 99% confidence estimation in any of the two linear regressions were recovered. Analysis of Molecular Variance (AMOVA; (Excoffier, et al., 1992)) was conducted in Arlequin 3.0 software (Excoffier, et al., 2005) assuming self-identified ancestry.

## RESULTS

### Autosomal DNA

The ancestry information provided by the 24 autosomal ASMs was first tested by performing a STRUCTURE analysis with the HGDP-CEPH samples assuming no prior knowledge of the ancestral groups. After K=4 the estimated loglikelihood of the data given the model (-19135) did not substantially change anymore. The four clusters detected at K=4 broadly match the four geographic regions: America, Sub-Saharan Africa, East Asia, and Eurasia (including Europe / Middle East / South Asia / Central Asia) (Figure 1). Only a small percentage of misclassified individuals was observed i.e., 0.47% Sub-Saharan Africans, 4.2% of Eurasians, 4.6% of Native American individuals, and 6.2% of East Asians (the latter was mainly in the Eurasian cluster with 3.6%). We concluded that these genetic markers are suitable for inferring bio-geographic ancestry in U.S. Americans since the four geographic regions identified with the 24 ASMs represent the putative parental populations of the four major groups of U.S. Americans.

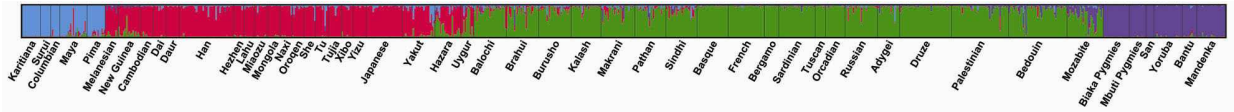



Figure 1. Genetic ancestry per individual in the global HGDP-CEPH panel as estimated by STRUCTURE using 24 autosomal ancestry-sensitive SNPs (K=4).

Next we used the Native Americans, East Asians, Eurasians, and Sub-Sahara Africans from HGDP-CEPH as parental groups of the U.S. Americans (the genotype data of the 24 autosomal SNPs can be found in the supplementary material 6) in a STRUCTURE analysis. Self-declared U.S. Europeans showed on average 93.2% of European ancestry (95% CI from 73.23% to 98.09%); self-declared U.S. Asians carried on average 89.5% of East Asian ancestry (95% CI from 37.43% to 97.46%); self-declared U.S. Africans revealed on average 86.2 % Sub-Sahara African ancestry (95% CI from 47.82% to 98.5%) (Figure 2). For these three U.S. groups rather small (between 0.8 and 8.1% on average) components of continental ancestries other than self-declared were detected (Figure 2). In contrast, self-declared U.S. Hispanics carried on average 61.2% European ancestry (95% CI from 8.33% to 95.75%), 14.9% Native American (95% CI from 1.21% to 55.54%), 10.8% East Asian (95% CI from 1.12% to 56.35%), and 11.6%, Sub-Saharan African ancestries (95% CI from 0.41% to 58.49%) (Figure 2). Furthermore, we observed for self-declared U.S. Africans statistically significant heterogeneity in the amount of African genetic ancestry depending on the geographic sampling region (Kruskal-Wallis test p-value=0.0042), as well as for self-declared U.S. Hispanics in the amount of Native American genetic ancestry (Kruskal-Wallis p-value = 1.48e-07). An AMOVA grouping individuals based on self-declared ancestry explained 34.2% (two tail p value <0.0005) of the total genetic variation suggesting strong genetic differentiation between self-declared ancestry groups of U.S. Americans.

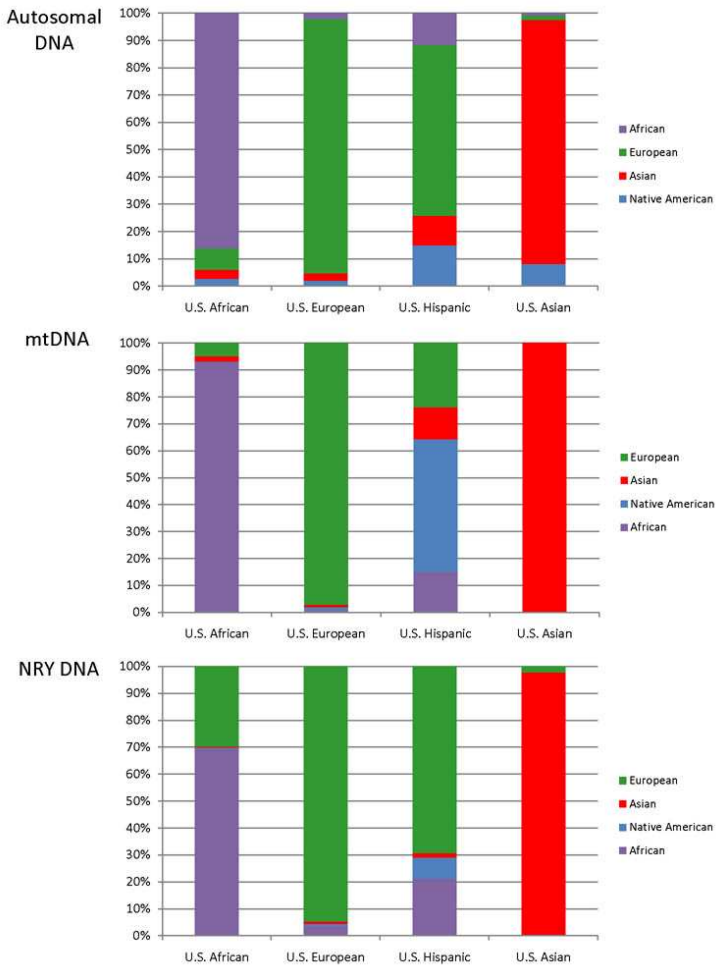



Figure 2. Proportions of average continental genetic ancestry in four U.S. American groups of self-declared ancestry based on autosomal DNA, mtDNA and NRY DNA.

We performed an additional STRUCTURE analysis considering only U.S. samples with K=4 and assuming no admixture (loglikelihood of the data given the model = -16287.9), which showed that the majority of U.S. Africans was in one of the four clusters (K4), and almost all U.S. Asians were in another cluster (K1) (see Table 1). In contrast, 15% of self-declared U.S. Hispanic samples were classified in the main cluster of U.S. Europeans (K3), and 19% of self-declared U.S. Europeans were clustered in the main cluster of self-declared U.S. Hispanics (K2).

Table 1. Correspondence between self-declared ancestry and STRUCTURE-based genetic ancestry inferred from 24 autosomal ASMs in four major U.S. American self-declared groups.

| | Clusters from STRUCTURE | | | |
|---|---|---|---|---|
| Self-declared ancestry | K1 | K2 | K3 | K4 |
| U.S. African | 0% | 2.2% | 1.0% | 96.8% |
| U.S. European | 0% | 19.0% | 80.6% | 0.4% |
| U.S. Hispanic | 2.4% | 77.8% | 15.7% | 4.0% |
| U.S. Asian | 99.9% | 0.1% | 0% | 0% |

From the MDS plot (Figure 3) it is evident that self-declared U.S. Europeans, Africans and Asians form rather discrete data clouds without strong overlaps between these groups, and tend to cluster close to their respective HGDP parental population. Self-declared U.S. Hispanics, however, did not cluster separately but either overlapped with U.S. European / HGDP European or appear between the U.S. European / HGDP European cluster and the U.S. Asian / HGDP East Asian cluster. Some U.S. Hispanics overlapped with the U.S. African / HGDP African cluster or appeared between the U.S. African / HGDP African and the U.S. European / HGDP European clusters.

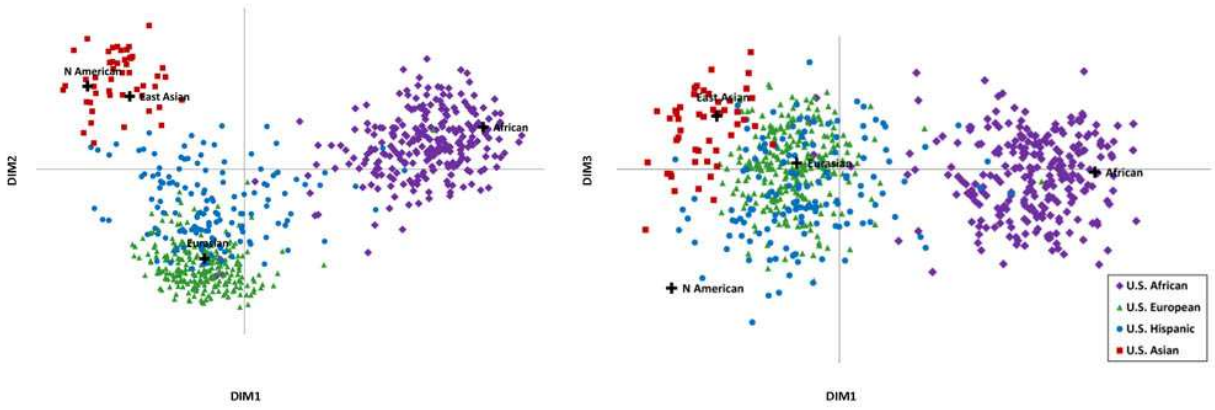



Figure 3. Two-dimensional plots of the first dimension, second dimension and third dimension obtained from a MDS analysis (stress = 0.13) performed with an Identical By State (IBS) matrix computed between pairs of individuals using. Centroids of the four parental populations from HGDP-CEPH are marked by crosses.

We also tested whether any of the 24 ASMs were more or less informative proportionally to the amount of information of the other markers for self-identification of U.S. Africans and U.S. Europeans. The lineal regression between the $I_n$ values computed for each SNP using U.S. European and continental African versus continental African and continental European (see methods for definition of continental populations) was highly statistically significant (R-squared = 0.98, two tail p-value = 3.91e-020; slope = 1.07, p value different from one = 0.0375). The $I_n$ value observed for rs16891982 when considering U.S. Europeans and continental Africans was significantly higher (falling out of the 99% predicted interval) than the one predicted by the linear regression using the 24 markers. In a similar way, comparison of the $I_n$ values computed between U.S. African and continental European versus these computed considering continental African and continental European was also statistically significant (R-squared = 0.97, two tail p-value = 1.85e-018; slope = 0.67, p value that the slope is different from 1 = 3.04e-12). Rs1448484 showed a larger $I_n$ value and rs16891982 smaller for the comparison between U.S. Africans and continental Europeans than predicted by the linear regression.

### NRY-DNA and mtDNA

The values of ancestry provided by uni-parentally inherited markers (Figure 2) were similar to the autosomal SNPs in the case of self-declared U.S. Europeans (estimated European ancestry for NRY: 94.7% and mtDNA: 96.7%; Fisher exact test value of the hypothesis of equal proportion of ancestry components between NRY and mtDNA = 4.85, two tail p value = 0.19) and for U.S. Asians (estimated East Asian ancestry for NRY: 97.8% for NRY and mtDNA; Fisher exact test value = 1.40, two tail p value = 1). In contrast, self-declared U.S. Africans showed discrepancies between the three genetic systems: 69.5% of NRY-DNA but 92.7% of mtDNA were of African ancestry; the second largest NRY ancestry component was European with 29.7%. The differences in the ancestry proportions between the two types of uniparental markers in U.S. Africans were highly statistically significant (Fisher exact test value = 58.80, two tail p value = 6.00e-014). In contrast to autosomal ASMs, we did not detect any statistically significant geographic substructure in the NRY and mtDNA ancestry data within self-declared U.S. Africans (Fisher statistic for NRY = 22.82, two tail p-value = 0.45 and Fisher statistic for mtDNA = 19.56, two tail p-value = 0.39). Self-declared U.S. Hispanics showed the most complex pattern of all the U.S. American groups studied also for uniparental markers. NRY ancestry was 69.3% European, 21.3% African and only 7.9% Native American, whereas the East Asian component was 1.6%. MtDNA ancestry was 48.8% Native American, 23.6% European and 11.8% East Asian. Differences on ancestry proportions in U.S. Hispanics between the two uni-parentally inherited markers were statistically significant (Fisher exact test value = 82.41, two tail p value = 3.11e-018). In contrast to autosomal ASMs, there was no significant NRY differentiation between self-declared U.S. Hispanics from the different sampled regions across the country (Fisher statistic for NRY = 11.69, two tail p-value = 0.14), whereas mtDNA data revealed statistically significant differences (Fisher statistic for mtDNA = 23.3, two tail p-value = 0.0024) as autosomal ASMs did. AMOVA analyses performed on the NRY and mtDNA data separately and considering self-declared ancestry grouping explained 27.65% (two tail p value < 0.000005) and 7.6% (two tail p value < 0.000005) of the total genetic diversity, respectively. AMOVA using the autosomal ASM data and considering groupings based on NRY ancestry and separately on mtDNA ancestry revealed 23.3% (two tail p-value <0.0005) and 30.2% (two tail p-value <0.0005) of the total genetic diversity, respectively. The NRY and mtDNA haplogroups for all individual samples included can be found in the supplementary material 6.

## DISCUSSION

The current U.S. population represents a mixture of groups with different bio-geographic ancestries, mainly from Europe, Sub-Saharan Africa, East Asia and the Americas. We have shown in the HGPD-CEPH samples that the ascertained ASMs are informative for detecting the ancestry of them. Our results based on these markers tend to corroborate previous findings performed in self-identified U.S. Europeans (Halder, et al., 2008; Halder, et al., 2009; Kosoy, et al., 2009) and U.S. Asians (Kosoy, et al., 2009), although usually many more markers were applied before. However, we observed discrepancies between our data and previous studies for self-declared U.S. Africans and U.S. Hispanics. For U.S. Africans we found a slightly larger percentage of African ancestry and a slighty lower percentage of European ancestry relative to previous reports (Tian, et al., 2006; Halder, et al., 2008; Halder, et al., 2009; Kosoy, et al., 2009; Zakharia, et al., 2009). For U.S. Hispanics, the Native American component tends to be rather low compared to previous studies (Price, et al., 2007; Halder, et al., 2009; Kosoy, et al., 2009). Differences in the admixture histories in different regions of the U.S. as reported elsewhere (Salazano and Bortolini, 2002; Kittles and Weiss, 2003; Zakharia, et al., 2009) are likely to explain such discrepancies. This view also is supported by the considerable heterogeneity in continental genetic ancestry depending on the geographic origin of the sampling region within the U.S. we observed for these two U.S. American groups. An alternative explanation in the case of U.S. Hispanics could be a lack of power of the set of autosomal ASMs we applied to distinguish Native American from East Asian ancestry (also explaining the apparent small Native American ancestry component in U.S. Asians). Native Americans and East Asians show a general genetic proximity due to their shared population history (Jakobsson, et al., 2008; Li, et al., 2008). Repeating the STRUCTURE analysis for U.S. Hispanics without considering East Asians as parental population raised the Native American ancestry component up to 27.44%, which is more comparable to previous studies. However, the fact that some of the self-declared U.S. Hispanic individuals carried NRY haplogroups typical for East Asians, and because a previous study also detected Asian ancestry in U.S. Hispanics (Guthery, et al., 2007), indicate that excluding East Asian admixture a priory would be incorrect for estimating genetic ancestry in U.S. Hispanics.

Overall, STRUCTURE, MDS and AMOVA analyses indicate that in U.S. Americans self-declared ancestry serves on average as a good proxy of the underlying autosomal genetic diversity, especially of European, African and Asian Americans. Our STRUCTURE results are in line with an earlier study reporting that ancestry self-identification corresponded well with STRUCTURE-based predictions for U.S. Americans (Tang, et al., 2005). Ancestry estimations obtained here with uni-parentally inherited markers are in good agreement with previous studies for U.S. Europeans, U.S. Africans and U.S. Hispanics for NRY (Kayser, et al., 2003; Hammer, et al., 2006; Lind, et al., 2007) and mtDNA (Allard, et al., 2002; Allard, et al., 2004; Allard, et al., 2005). In contrast, the percentage of Native American mtDNA ancestry estimated in the U.S. Hispanics studied here appears smaller than that of other studies (ranging from ~70% to ~85.11%) (Merriwether, et al., 1997; Allard, et al., 2006), although differences between U.S. Hispanic groups from different U.S. regions were observed, which may explain the discrepancies

Combining the ancestry information of patrilineal, matrilineal and biparental markers, a special quality of our study, offers the possibility to study the patterns of admixture at different levels of complexity. We observed the same degree of ancestry homogeneity in the three types of genetic markers for self-identified U.S. Europeans and U.S. Asians, which suggests relatively low genetic admixture with other ancestry groups than the one indicated by self-identification. Noticeably, this finding for U.S. Europeans contrasts with common observation for self-declared European Americans from South America (Goncalves, et al., 2007; Corach, et al., 2010). In those South American groups European ancestry signals are usually high for NRY-DNA, intermediate for autosomal DNA, but low for mtDNA, whereas Native American genetic ancestry signals are reverse, indicating sex-bias admixture between European males and Native American females (Goncalves, et al., 2007; Corach, et al., 2010). This discrepancy has been explained in terms of local differences in social practices (Goncalves, et al., 2007). However, it could also be explained if the concept of self-identification had different meanings depending on the country of origin. This is supported by the fact that genetic admixture proportions of self-identified U.S. Hispanics of our study resemble those from self-declared European Americans in some South American countries with similar evidence for sex-biased admixture history. Our data also indicate sex-biased admixture for U.S. Africans with considerably more European NRY than mtDNA ancestry, and autosomal DNA estimates in-between. Previous studies analyzing NRY and mtDNA ancestry in U.S. Africans have reported similar results (Kayser, et al., 2003; Lind, et al., 2007), (see (Stefflova, et al., 2009) for a review), which we complement here with agreeing autosomal DNA evidence.

Why did we (and others) not detect similarly strong signals of genetic admixture in U.S. Europeans, in contrast to U.S. Africans and U.S. Hispanics? One explanation may be that admixed individuals traditionally self-classify in a biased way and towards only one of the parental groups involved in the admixture process. Ancestry self-identification is the result of both visible traits (with a biological basis) such as skin color combined with cultural/sociological aspects (Bamshad and Guthery, 2007). In the present study rs1448484 appeared to be more informative and rs16891982 less informative for differentiating U.S. Africans from continental Europeans than continental Africans from continental Europeans. In contrast, rs16891982 was more informative for differentiating U.S. Europeans from continental Africans than continental Europeans from continental Africans. rs1448484 is within the *OCA2* gene, which when mutated can lead to oculocutaneous albinism type II (OMIM entry no. 203200); in addition, it has been previously associated with differences in pigmentation using pooled U.S. African / African-Caribbean population and U.S. European individuals (Shriver, et al., 2003); however, there is no evidence thus far that rs1448484 in *OCA2* is directly involved in pigmentation variation, although it could be in LD with a functional variant. In contrast, rs16891982 represents a non-synonymous amino acid change (F374L) in *SLC45A2*, and this gene, if mutated, leads to oculocutaneous albinism type IV (OMIM entry no. 606574). Notably, the *SLC45A2*-374 F allele of rs16891982 is almost fixed in the Europeans (Soejima and Koda, 2007), and affects the amount of pigmentation (Stokowski, et al., 2007). Individuals carrying the genotypes *SLC45A2*-374L/L or *SLC45A2*-374L/F tend to show a darker skin color than *SLC45A2*-374F/F individuals (Cook, et al., 2009). Here we hypothesize that within the self-identified as U.S. Europeans or Africans, individuals with the L/L or F/L genotypes would tend to declare themselves as U.S. African whereas individuals F/F would as U.S. European. In that case, the presence of heterozygotes in U.S. Africans would decrease the $I_n$ statistic more than expected with continental Europeans and increase it between U.S. Europeans and continental Africans, as observed by our data. Although these data provide genetic evidence for the role of skin color in the complex process of ancestry self-identification, it would be extremely simplistic to reduce ancestry self- identification only to the type of analysis performed here.

## REFERENCES

U.S. Census Bureau. 2008. Annual Estimates of the Population by Sex, Race, and Hispanic Origin for the United States:April 1,2000 to July 1, 2007 (NC-EST2007-03). Release date: May, 12008.

Achilli A, Perego UA, Bravi CM, Coble MD, Kong QP, Woodward SR, Salas A, Torroni A, Bandelt HJ. 2008. The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. PLoS One 3(3):e1764.

Allard MW, Miller K, Wilson M, Monson K, Budowle B. 2002. Characterization of the Caucasian haplogroups present in the SWGDAM forensic mtDNA dataset for 1771 human control region sequences. Scientific Working Group on DNA Analysis Methods. J Forensic Sci 47(6):1215-23.

Allard MW, Polanskey D, Miller K, Wilson MR, Monson KL, Budowle B. 2005. Characterization of human control region sequences of the African American SWGDAM forensic mtDNA data set. Forensic Sci Int 148(2-3):169-79.

Allard MW, Polanskey D, Wilson MR, Monson KL, Budowle B. 2006. Evaluation of variation in control region sequences for Hispanic individuals in the SWGDAM mtDNA data set. J Forensic Sci 51(3):566-73.

Allard MW, Wilson MR, Monson KL, Budowle B. 2004. Control region sequences for East Asian individuals in the Scientific Working Group on DNA Analysis Methods forensic mtDNA data set. Leg Med (Tokyo) 6(1):11-24.

Bamshad M, Guthery SL. 2007. Race, genetics and medicine: does the color of a leopard's spots matter? Curr Opin Pediatr 19(6):613-8.

Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, Bertranpetit J, Quintana-Murci L, Tyler-Smith C, Wells RS, Rosset S. 2008. The dawn of human matrilineal diversity. Am J Hum Genet 82(5):1130-40.

Bortolini MC, Salzano FM, Thomas MG, Stuart S, Nasanen SP, Bau CH, Hutz MH, Layrisse Z, Petzl-Erler ML, Tsuneto LT, Hill K, Hurtado AM, Castro-de-Guerra D, Torres MM, Groot H, Michalski R, Nymadawa P, Bedoya G, Bradman N, Labuda D, Ruiz-Linares A. 2003. Y-chromosome evidence for differing ancient demographic histories in the Americas. Am J Hum Genet 73(3):524-39.

Butler JM, Schoske R, Vallone PM, Redman JW, Kline MC. 2003. Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. J Forensic Sci 48(4):908-11.

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN. 2005. Demonstrating stratification in a European American population. Nat Genet 37(8):868-72.

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(5566):261-2.

Cook AL, Chen W, Thurber AE, Smit DJ, Smith AG, Bladen TG, Brown DL, Duffy DL, Pastorino L, Bianchi-Scarra G, Leonard JH, Stow JL, Sturm RA. 2009. Analysis of cultured human melanocytes based on polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P loci. J Invest Dermatol 129(2):392-405.

Corach D, Lao O, Bobillo C, van Der Gaag K, Zuniga S, Vermeulen M, van Duijn K, Goedbloed M, Vallone PM, Parson W, de Knijff P, Kayser M. 2010. Inferring continental ancestry of argentineans from Autosomal, Y-chromosomal and mitochondrial DNA. Ann Hum Genet 74(1):65-76.

Cruciani F, La Fratta R, Trombetta B, Santolamazza P, Sellitto D, Colomb EB, Dugoujon JM, Crivellaro F, Benincasa T, Pascone R, Moral P, Watson E, Melegh B, Barbujani G, Fuselli S, Vona G, Zagradisnik B, Assum G, Brdicka R, Kozlov AI, Efremov GD, Coppa A, Novelletto A, Scozzari R. 2007. Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. Mol Biol Evol 24(6):1300-11.

Decker AE, Kline MC, Redman JW, Reid TM, Butler JM. 2008. Analysis of mutations in father-son pairs with 17 Y-STR loci. Forensic Sci Int Genet 2(3):e31-5.

Diegoli TM, Irwin JA, Just RS, Saunier JL, O'Callaghan JE, Parsons TJ. 2009. Mitochondrial control region sequences from an African American population sample. Forensic Sci Int Genet 4(1):e45-52.

Excoffier L, Laval G, Schneider S. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47-50.

Excoffier L, Smouse PE, Quattro JMV. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131(2):479-91.

Finnila S, Lehtonen MS, Majamaa K. 2001. Phylogenetic network for European mtDNA. Am J Hum Genet 68(6):1475-84.

Goncalves VF, Prosdocimi F, Santos LS, Ortega JM, Pena SD. 2007. Sex-biased gene flow in African Americans but not in American Caucasians. Genet Mol Res 6(2):256-61.

Guthery SL, Salisbury BA, Pungliya MS, Stephens JC, Bamshad M. 2007. The structure of common genetic variation in United States populations. Am J Hum Genet 81(6):1221-31.

Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. 2008. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat 29(5):648-58.

Halder I, Yang BZ, Kranzler HR, Stein MB, Shriver MD, Gelernter J. 2009. Measurement of admixture proportions and description of admixture structure in different U.S. populations. Hum Mutat 30(9):1299-309.

Hammer MF, Chamberlain VF, Kearney VF, Stover D, Zhang G, Karafet T, Walsh B, Redd AJ. 2006. Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. Forensic Sci Int 164(1):45-55.

Irwin J, Saunier J, Strouss K, Paintner C, Diegoli T, Sturk K, Kovatsi L, Brandstatter A, Cariolou MA, Parson W, Parsons TJ. 2008. Mitochondrial control region sequences from northern Greece and Greek Cypriots. Int J Legal Med 122(1):87-9.

Irwin JA, Saunier JL, Beh P, Strouss KM, Paintner CD, Parsons TJ. 2009. Mitochondrial DNA control region variation in a population sample from Hong Kong, China. Forensic Sci Int Genet 3(4):e119-25.

Irwin JA, Saunier JL, Strouss KM, Sturk KA, Diegoli TM, Just RS, Coble MD, Parson W, Parsons TJ. 2007. Development and expansion of high-quality control region databases to improve forensic mtDNA evidence interpretation. Forensic Sci Int Genet 1(2):154-7.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451(7181):998-1003.

Jobling MA, Tyler-Smith C. 2003. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4(8):598-612.

Just RS, Diegoli TM, Saunier JL, Irwin JA, Parsons TJ. 2008. Complete mitochondrial genome sequences for 265 African American and U.S. "Hispanic" individuals. Forensic Sci Int Genet 2(3):e45-8.

Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res 18(5):830-8.

Kayser M, Brauer S, Schadlich H, Prinz M, Batzer MA, Zimmerman PA, Boatin BA, Stoneking M. 2003. Y chromosome STR haplotypes and the genetic structure of U.S. populations of African, European, and Hispanic ancestry. Genome Res 13(4):624-34.

Kersbergen P, van Duijn K, Kloosterman AD, den Dunnen JT, Kayser M, de Knijff P. 2009. Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. BMC Genet 10:69.

Kittles RA, Weiss KMV. 2003. Race, ancestry, and genes: implications for defining disease risk. Annu Rev Genomics Hum Genet 4:33-67.

Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knijff P, Feldman M, Cavalli-Sforza LL, Oefner PJ. 2006. The role of selection in the evolution of human mitochondrial genomes. Genetics 172(1):373-87.

Kong QP, Bandelt HJ, Sun C, Yao YG, Salas A, Achilli A, Wang CY, Zhong L, Zhu CL, Wu SF, Torroni A, Zhang YP. 2006. Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet 15(13):2076-86.

Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF. 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat 30(1):69-78.

Kruskal J, Wish M. 1990. Multidimensional Scaling (Quantitative Applications in the Social Sciences). California: Newbury Park.

Lao O, de Gruijter JM, van Duijn K, Navarro A, Kayser M. 2007. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet 71(Pt 3):354-69.

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Ruther A, Schreiber S, Becker C, Nurnberg P, Nelson MR, Krawczak M, Kayser M. 2008. Correlation between genetic and geographic structure in Europe. Curr Biol 18(16):1241-8.

Lao O, van Duijn K, Kersbergen P, de Knijff P, Kayser M. 2006. Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. Am J Hum Genet 78(4):680-90.

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(5866):1100-4.

Lind JM, Hutcheson-Dilks HB, Williams SM, Moore JH, Essex M, Ruiz-Pesini E, Wallace DC, Tishkoff SA, O'Brien SJ, Smith MW. 2007. Elevated male European and female African contributions to the genomes of African American individuals. Hum Genet 120(5):713-22.

Luis JR, Rowold DJ, Regueiro M, Caeiro B, Cinnioglu C, Roseman C, Underhill PA, Cavalli-Sforza LL, Herrera RJ. 2004. The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. Am J Hum Genet 74(3):532-44.

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonne-Tamir B, Sykes B, Torroni A. 1999. The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 64(1):232-49.

Merriwether DA, Huston S, Iyengar S, Hamman R, Norris JM, Shetterly SM, Kamboh MI, Ferrell RE. 1997. Mitochondrial versus nuclear admixture estimates demonstrate a past history of directional mating. Am J Phys Anthropol 102(2):153-9.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD. 2008. Genes mirror geography within Europe. Nature 456(7218):98-101.

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD. 1998. Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63(6):1839-51.

Parson W, Dur A. 2007. EMPOP--a forensic mtDNA database. Forensic Sci Int Genet 1(2):88-92.

Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, Tandon A, Schirmer C, Neubauer J, Bedoya G, Duque C, Villegas A, Bortolini MC, Salzano FM, Gallo C, Mazzotti G, Tello-Ruiz M, Riba L, Aguilar-Salinas CA, Canizales-Quinteros S, Menjivar M, Klitz W, Henderson B, Haiman CA, Winkler C, Tusie-Luna T, Ruiz-Linares A, Reich D. 2007. A genomewide admixture map for Latino populations. Am J Hum Genet 80(6):1024-36.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155(2):945-59.

R Development Core Team. 2006. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Richards MB, Macaulay VA, Bandelt HJ, Sykes BC. 1998. Phylogeography of mitochondrial DNA in western Europe. Ann Hum Genet 62(Pt 3):241-60.

Rosenberg NA. 2004. Distruct: a program for the graphical display of population structure. Molecular Ecology Notes 4:137-138.

Rosenberg NA, Li LM, Ward R, Pritchard JK. 2003. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73(6):1402-22.

Salazano FM, Bortolini MC. 2002. The evolution and genetics of Latin American populations. Cambridge, United Kingdom: Cambridge University Press.

Saunier JL, Irwin JA, Just RS, O'Callaghan J, Parsons TJ. 2008. Mitochondrial control region sequences from a U.S. "Hispanic" population sample. Forensic Sci Int Genet 2(2):e19-23.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza

LL, Underhill PAV. 2000. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290(5494):1155-9.

Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N, Baron A, Jackson T, Argyropoulos G, Jin L, Hoggart CJ, McKeigue PM, Kittles RA. 2003. Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 112(4):387-99.

Sinha M, Larkin EK, Elston RC, Redline S. 2006. Self-reported race and genetic admixture. N Engl J Med 354(4):421-2.

Soejima M, Koda Y. 2007. Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. Int J Legal Med 121(1):36-9.

SPSS. 2003. SPSS for Windows. Version 12.0.

Stefflova K, Dulik MC, Pai AA, Walker AH, Zeigler-Johnson CM, Gueye SM, Schurr TG, Rebbeck TR. 2009. Evaluation of group genetic ancestry of populations from Philadelphia and Dakar in the context of sex-biased admixture in the Americas. PLoS One 4(11):e7842.

Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, Jarman C, Filsell W, Ginger RS, Green MR, van der Ouderaa FJ, Cox DR. 2007. A genomewide association study of skin pigmentation in a South Asian population. Am J Hum Genet 81(6):1119-32.

Tang H, Quertermous T, Rodriguez B, Kardia SL, 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(2):268-75.

Tian C, Hinds DA, Shigeta R, Kittles R, Ballinger DG, Seldin MF. 2006. A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. Am J Hum Genet 79(4):640-9.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM. 2009. The genetic structure and history of Africans and African Americans. Science 324(5930):1035-44.

Vallone PM, Butler JM. 2004. AutoDimer: a screening tool for primer-dimer and hairpin structures. Biotechniques 37(2):226-31.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30(2):E386-94.

Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, Poletti G, Mazzotti G, Hill K, Hurtado AM, Camrena B, Nicolini H, Klitz W, Barrantes R, Molina JA, Freimer NB, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Dipierri JE, Alfaro EL, Bailliet G, Bianchi NO, Llop E, Rothhammer F, Excoffier L, Ruiz-Linares A. 2008. Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet 4(3):e1000037.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38(6):1358-1370.

Zakharia F, Basu A, Absher D, Assimes TL, Go AS, Hlatky MA, Iribarren C, Knowles JW, Li J, Narasimhan B, Sidney S, Southwick A, Myers RM, Quertermous T, Risch N, Tang H. 2009. Characterizing the admixed African ancestry of African Americans. Genome Biol 10(12):R141.

## ONLINE SUPPLEMENTARY MATERIAL

### Supplementary Material 1: Genotyping information autosomal SNPs

#### *PCR and extension primer design*

*PCR and extension primer design*

The 24 autosomal ASM SNPs were genotyped using two 12plex SNaPshot assays based on the principle of primer extension. 24 PCR primer pairs were selected using the commercially available primer selection software Visual OMP (DNA Software, Inc., Ann Arbor, MI). Template sequences consisting of approximately 500 base pairs up- and downstream from each SNP site were input into the Visual OMP program. Regions 30 bases up- and downstream from the SNP site were excluded from being selected as PCR primer binding sites. The size of each amplicon was kept under 150 base pairs to increase success when typing degraded samples e.g. in future forensic analyses. Each primer pair was selected independently (i.e. singleplex primer design). The final set of 24 PCR primer pairs were screened using AutoDimer for potential secondary structures such as primer-dimer and hairpin interactions (Vallone and Butler, 2004). Compatible primer pairs were divided into two separate PCR multiplexes containing 12 loci (see Table 1). The 24 extension primers were selected using the software module 'ASPE tool' (http://yellow.nist.gov:8444/dnaAnalysis/aspeToolsPage.do) present in the web-based AutoDimer software package (http://yellow.nist.gov:8444/dnaAnalysis/index.do). The user input consisted of the PCR amplicon sequences containing the corresponding SNP sites. Design parameter variables consisted of the desired length and predicted $T_m$ of an extension primer. Primer sequences up- and downstream adjacent to the SNP site were selected that had the appropriate length and $T_m$ characteristics. Extension primers were selected that had a predicted $T_m$ of approximately 60 $^{o}$C. Extension primers were subsequently screened for hairpin and primer-dimer interactions as described for the multiplex PCR primers. Poly-T tails of various lengths were added to the 5' end of extension primers to allow sufficient fragment separation on a capillary electrophoresis system (see Table 1). All oligonucleotides were purchased from Qiagen Operon (Alameda, CA). Oligonucleotides were delivered lyophilized and desalted and stock solutions of 100 µM were prepared by adding in the appropriate volumes of a low salt buffer (10 mM TrisHCl and 0.1 mM EDTA pH 7.2)

#### *Multiplex PCR*

PCR conditions for each of the two 12plex amplification reactions were identical. Multiplex amplifications were carried out in a total volume of 15 µL. Approximately 1 ng of human template (genomic) DNA was present in the multiplex PCR amplifications. Final PCR reagent concentrations were: 1 unit of AmpliTaqGold® DNA polymerase (Applied Biosystems), 1x Taq Gold PCR buffer, 250 µM dNTPs (Promega Corp., Madison, WI), 2 mM Mg++, 0.16 mg/mL bovine serum albumin (BSA) fraction V (Sigma, St. Louis, MO), 0.4 µM of each amplification primer pair (24 primers per multiplex). Thermal cycling for PCR and SNaPshot assays was carried out using the GeneAmp 9700 (Applied Biosystems) running in 9600-emulation mode (i.e. ramp speeds of 1 $^{o}$C/s). Note that for locus rs1344870 the final primer pair concentration was increased to 0.8 µM to reach balanced signals. The multiplex PCR thermal cycling conditions were as follows: 95$^{o}$C for 10 min followed by 32 cycles of #95$^{o}$C for 30 s, 55$^{o}$C for 35 s, 72$^{o}$C for 30 s# and a final step of 72$^{o}$C for 7 min (afterwards incubated at 4$^{o}$C). A combination of Exonuclease I (Exo I) and Shrimp Alkaline Phosphatase (SAP) (USB Corp., Cleveland, OH) was used to remove excess PCR primers and degrade unincorporated dNTPs. A mix of 1.4 µL of Exo I (1 µL = 10 units) and 2.6 µL SAP (1 µL = 1 unit) per sample was prepared and mixed. Four µL of the cocktail was added to each PCR reaction. The samples were incubated at 37$^{o}$C for 90 min followed by 80$^{o}$C for 20 min. The extensive incubation time ensured that the PCR primers were completely digested.

#### *Multiplex primer extension reaction*

Multiplex primer extension reactions were carried out in a total volume of 10 µL. Reaction components were: 2.5 µL of ABI Prism® SNaPshot® multiplex kit mix (Applied Biosystems), 0.5 µL of 10X AmpliTaqGold® PCR buffer, 3 µL of multiplex PCR products, 2.5 µL of deionized water, and 1.5 µL of a stock solution of extension primers (an unbalanced stock solution contained ~5 µM of each extension primer, see Table 1 for the exact values). Thermal cycling conditions for extension reactions were carried out as described in the SNaPshot multiplex kit manual: 25 cycles of 96$^{o}$C for 10 s, 50$^{o}$C for 5 s, 60$^{o}$C for 30s. Excess fluorescently labeled ddNTPs were inactivated by addition of 1 unit of Shrimp Alkaline Phosphatase (SAP). Reactions were mixed briefly and incubated at 37$^{o}$C for 30 min then 80$^{o}$C for 20 min. The ABI PRISM® 3130XL Genetic Analyzer was used for

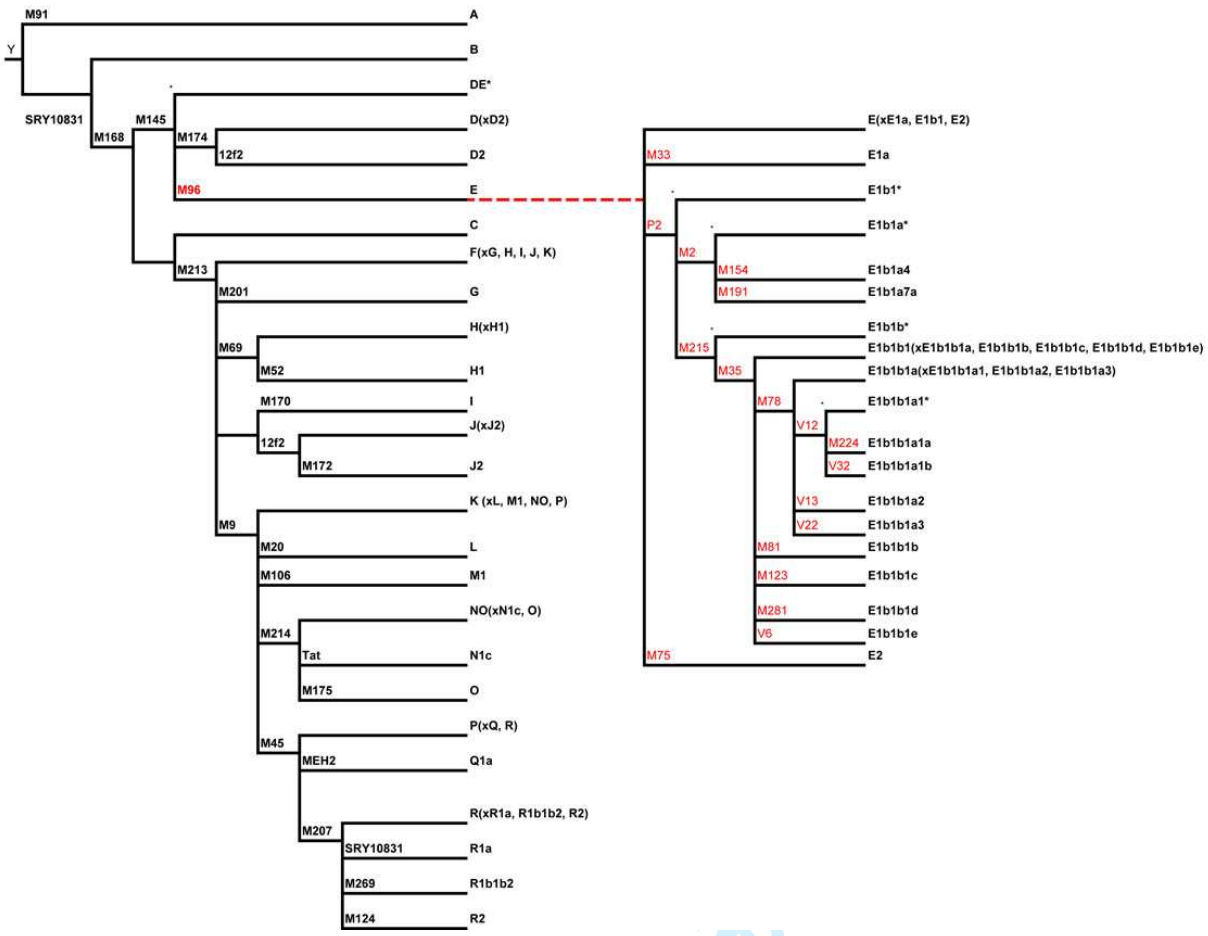
capillary electrophoresis (CE) with filter set E5 from the 5 dyes dR110, dR6G, dTAMRA™, dROX™, and LIZ™ after an appropriate spectral matrix had been created using materials from the matrix standard set DS-02 (Applied Biosystems). Fluorescently labeled extension reactions were prepared for CE analysis by mixing 14 µL of Hi-Di formamide™ (Applied Biosystems), 0.4 µL of the LIZ-120 internal sizing standard (Applied Biosystems), and 0.9 µL of SAP treated extension reaction. A 36 cm capillary array filled with denaturing POP6 performance optimized polymer (Applied Biosystems) was utilized for DNA fragment separation. A.C.E.™ (Ameresco, Solon, OH) capillary electrophoresis running buffer was used in 1 x concentration. Typical run module parameters were: Run temp = 60 °C, Capillary fill volume = 184 steps, Pre run voltage = 15 kV, Pre run time = 60 sec, Injection Voltage = 1kV, Injection time = 13 sec, Run Voltage = 15 kV, Data Delay = 200 sec, and Run time = 1200. Data analysis was performed using GeneMapperIDv3.2 software (Applied Biosystems). Bins and panels for the SNPs in each multiplex were developed based on fragment size and dye color for automated allele calling and are made available via the STRbase website http://www.cstl.nist.gov/biotech/strbase/SNP.htm.

| Multiplex A | | PCR Primers | SBE Primers | | Length | µM |
|---|---|---|---|---|---|---|
| rs1048610 | F | AGGCAGGTCTCAGAACAATCC | GTGTGCTGCAGGGACCTTTC | F | 20 | 5 |
| | R | GTTCAGCATCGACATAGGGC | | | | |
| rs1876482 | F | GAGCTGTTGATAGAGCTTTTGTGG | ttttttGGCTGTACCCTCACTATTGGTG | R | 28 | 5 |
| | R | ACGTGACACATAAAGAAAATGCCAT | | | | |
| rs2179967 | F | AAGAGTGTGTTGTATGCTTTGGAAA | ttttttCTTTGGAAATGGGTGTGCAACA | F | 28 | 6 |
| | R | TCCTTCCAGCCCGACTAGAAC | | | | |
| rs1858465 | F | GATTTCAAAAAGTCTACAGATTTGG | ttttttACTTCCTCTTTAATACTTCAACTGAGT | R | 32 | 7 |
| | R | TGACTTTGTCAAACTTCCTCTTTAA | | | | |
| rs1371048 | F | CTTAAATAGCCAAATAGCTCTAACT | tttttttttATTTGAGTATGCTCTGTAGATGCTTC | R | 36 | 5 |
| | R | ACAAACGAAATATTTGAGTATGCT | | | | |
| rs1369290 | F | GAGGCCCTACATGACCTGTC | ttttttttttttttACCACAGGCTCTTGATAAAGTGTCT | F | 40 | 5 |
| | R | GGGCTCCTCTTTCGCTCA | | | | |
| rs1465648 | F | ACCAGAAGGAAAGAGAAAAAGCAC | ttttttttttttttttGAAAAAGCACAGTATCAAGTTTGACTT | F | 44 | 6 |
| | R | AACAAACTACAGCAACAGAATCTTT | | | | |
| rs1391681 | F | GAGTAGTTGCTCATGAAGCTGAAAA | ttttttttttttttttttttTGTCACCCTTTACAAAACAGTTTGCA | F | 48 | 5 |
| | R | GGGCAGCCAAAAATAAAACAAAACA | | | | |
| rs1461227 | F | ACTGGGAAATTCTCACTGCAACT | ttttttttttttttttttttttttAACTACAACTAGCCCTAGGCTAATCTA | F | 52 | 5 |
| | R | TTGACAGATGGAGACACTGAAGC | | | | |
| rs1907702 | F | CCAACTCCTAATCAAGGCCTAC | ttttttttttttttttttttttttttCCTAATCAAGGCCTACAGAGACCTTC | F | 56 | 5 |
| | R | AGGAACATAAAGGAGGCCAGT | | | | |
| rs2052760 | F | ATTCAGAAAAGTGCATGCAGAAATT | ttttttttttttttttttttttttttttttttttATTATCAATGGGTTATTTTTGCCTCA | F | 60 | 5 |
| | R | GAGAGAGAGGAGTGAGAAAGGC | | | | |
| rs1667751 | F | CTGGTTCTTTTCCATCCAGCCTTTA | ttttttttttttttttttttttttttttttttttttttttCTTTACAAGCTACAAGACTTACGCCT | F | 64 | 5 |
| | R | GAGATCACCAAGGGAGTAAGTACAG | | | | |
| Multiplex B | | PCR Primers | SBE Primers | | Length | µM |
| rs1448484 | F | TCTCCTTCCAAGCCTTCTGAAAAAT | tATGAGAGCTGGCAGCTTCC | F | 20 | 6 |
| | R | GCAACCACACAGAACACAGC | | | | |
| rs714857 | F | GAAACTTCCCTAATGGGTCTTGTGA | tttCTTGTGAACCTTGGCTCCCTG | F | 24 | 6 |
| | R | CCTCCCTCACACATAAAACTTCTCA | | | | |
| rs16891982 | F | ATCCAAGTTGTGCTAGACCAGAA | tttttGAGGAAAACACGGAGTTGATGCA | F | 29 | 5 |
| | R | AGAGGAGTCGAGGTTGGATG | | | | |
| rs1808089 | F | TGTCAGGCCTTACCACTGCATAAGA | ttttttttACAAATGAGTAATGCCGTGGTGG | R | 31 | 5 |
| | R | AAACAACTCAGCGGCACAAA | | | | |
| rs1478785 | F | TCCTGGAGGCTTGAGGGCTA | ttttttttAGGGATGTTCATTTAAAATAACATCGC | F | 36 | 5 |
| | R | GGCTTGCTGGCTTTTTCTAGAT | | | | |
| rs952718 | F | GAGCCTAGATCCTGACTTCCTTG | ttttttttttttAAAATGCAAATTTCACCTTCTTCAAAT | R | 40 | 5 |
| | R | CTGTCACTGGAGATGTCATCTCAT | | | | |
| rs1405467 | F | AATTTGCAACAAAGAGGAAGGGGA | ttttttttttttttttttAAGTAGTCAGCTGAACTCACCTGAT | F | 43 | 5 |
| | R | GAGCAATAAGAGTGACTATGTCTGC | | | | |
| rs1344870 | F | CAATCTCAGTTTTAATTGCCATGT | ttttttttttttttttttttTCGCTCTTAAGTATGTTTTCTTGGTC | F | 48 | 5 |
| | R | AGGATGTATTGGGGCCTTTC | | | | |
| rs3843776 | F | AGGCCACTGTTGTGGTTTATG | ttttttttttttttttttttttttttTGTTGTGGTTTATGTTTCACTTCGAC | F | 53 | 6 |
| | R | TGAGGGCTCTACAACACTGC | | | | |
| rs721352 | F | TCTGTGCCCAGATGCAAATCCTTA | ttttttttttttttttttttttttttttTGCTTGATGGCTCCACCTATCA | R | 51 | 6 |
| | R | GACCCAGAACTGTGCAGG | | | | |
| rs722869 | F | CCTTCTGCACTTGGGCATATT | ttttttttttttttttttttttttttttttCAAATCCTTCATTTCACAAATGAAGCT | R | 60 | 5 |
| | R | AGGTAGAGATCTAACAAACCACAGT | | | | |
| rs926774 | F | AATCAAGTTCAGACTTTTGCCTCAT | ttttttttttttttttttttttttttttttttttAAGCTATTGTAGTGAGGAAGGCTAGA | R | 63 | 7 |

## Supplementary Material 2: Genotyping information NRY SNPs

| Additional | Haplogroup | SNP | Bibliographical source | GenBank | dbSNPs accession (if known) | Position Y-chromosome | Forward Amplification primer (5' --> 3') | Reverse Amplification primer (5' --> 3') | concentration in PCR (µM) | Amplicon size (bp) | Minisequencing primers (target-specific sequence in capitals) | Orientation | concentration in minis eq-reaction (µM) | Primer size (nt) | Mutation: Wildtype/ Mutant** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| hg E | E | M96 | Refs | AC010889 | rs9306841 | 20238386 | GCCAGCCAAGA ATGAAGAGA | TGAGCTGTGAT GTGTAACTTGG | 0.1 | 143 | GGAAAACAGGTCTC TCATAATA | R | 0.04 | 22 | G/C |
| hg E | E1a | M33 | 2 | AC009977 | | 20199838 | CCGTCATAGGCT GAGACAAGA | CCCCAAGAGAG ACAACTGAC | 0.15 | 150 | ccacgtcgtgaaagtctgaca aCAGTTACAAAAGTA TAATATGTCTGAGAT | R | 0.06 | 51 | C/G |
| hg E | E1b1 | P2 | 3 | AC010137 | | 20070219 | GAGAATCAGCTC CAGCCATC | TTTTGGATCTTC ATGCTGGTT | 0.03 | 100 | gacaaAGGTGCCCCT AGGAGGAGAA | F | 0.2 | 25 | T/C |
| hg E | E1b1a | M2 | 6 | AC011302 | rs3893 | 12606580 | ACGGAAGGAGT TCTAAAATTCAG G | AAAATACAGCT CCCCCTTTATC CT | 0.1 | 147 | cacgtcgtgaaagtctgacaa TTCATTGTTAACAAA AGTCC | R | 0.06 | 41 | G/A |
| hg E | E1b1a4 | M154 | 2 | AC010889 | | 20352065 | AGGCTACAAATT AGTGCGACA | GAGGCACAGAT ACTTAAACCATT G | 0.06 | 77 | acaaGTTACATGGCC TATAATATTCAGTAC A | R | 0.03 | 31 | G/A |
| hg E | E1b1a7 | M191 | 2 | AC004474 | rs2032590 | 13529007 | AAAAATGGAGTG TTTATCAGAGCT T | CCCAGACACAC CAAAATATCTC | 0.3 | 122 | gaaagtctgacaaAAAAT ATCTCATATTTTCAT | R | 0.25 | 33 | A/G |
| hg E | E1b1b | M215 | 2 | AC006376 | rs2032654 | 13977218 | TCAAACTGTTGG TAAATTTTAGAG AAA | CAGAAGCATCA GCTGGAACA | 0.25 | 97 | gtcgtgaaagtctgacaaCA GCTGGAACAGTTAG AAAG | R | 0.15 | 38 | C/T |
| hg E | E1b1b1 | M35 | 2 | AC009977 | rs1179188 | 20201091 | AGGGCATGGTC CCTTTCTAT | TCCATGCAGAC TTTCGGAGT | 0.2 | 96 | actgactaaactaggtgccac gtcgtgaaagtctgacaaTC GGAGTCTCTGCCTG TGTC | R | 0.06 | 59 | G/A |
| hg E | E1b1b1a | M78* | 2 | AC010889 | | 20352691 | TGCATTACTCCG TATGTTCGAC | TGGAAGCTTAC CATCTTTTTATG A | 0.05* | 132 | aagtctgacaaCTTATTT TGAAATATTTGGAAG GGC | R | 0.02 | 36 | A/C |
| hg E | E1b1b1a1 | V12 | 7 | AC012068 | | 6883099 | CTGAGTTGGATT GTTTTAAGTTGA | TTGGTCTCTCTT CATGTGCTG | 0.15 | 150 | acaaTTGTGTAGATAA TTCAAAGT | R | 0.25 | 24 | C/T |
| hg E | E1b1b1a1a | M224* | 2 | AC010889 | | 20352687 | TGCATTACTCCG TATGTTCGAC | TGGAAGCTTAC CATCTTTTTATG A | 0.05* | 132 | cgtgaaagtctgacaaAAT TGATACACTTAACAA AGATACTTC | F | 0.15 | 43 | A/G |
| hg E | E1b1b1a1b | V32 | 7 | AC012068 | | 6992821 | GCAAATGTTCCA TGAATGGTG | CCAGCCAGAGA GGCACTTTA | 0.4 | 111 | CCCaactgactaaactaggt gccacgtcgtgaaagtctgac aaCACACATGTATAT ACACACC | R | 0.25 | 63 | C/G |
| hg E | E1b1b1a2 | V13 | 7 | AC012068 | | 6902263 | CAACAGTGGAG GACAAAGCA | AAGACCAGCCT GACCAACAT | 0.15 | 106 | cgtcgtgaaagtctgacaaG CTCAAACTTCCCTTG | R | 0.15 | 35 | A/G |
| hg E | E1b1b1a3 | V22 | 7 | AC012068 | | 6919957 | TGGCAATGCCTC AACTTACA | ATTCCCCAAGG TTTCAGAGG | 0.15 | 110 | Caactgactaaactaggtgc cacgtcgtgaaagtctgacaa CCAAGGTTTCAGAG GTC | R | 0.15 | 58 | C/G |
| hg E | E1b1b1b | M81 | 2 | AC010889 | rs2032640 | 20351960 | GCACTATCATAC TCAGCTACACAT CTC | TTGTTTCTTCTT GGTTTGTGTGA | 0.03 | 99 | acaaCTTGGTTTGTGT GAGTATACTCTATGA C | R | 0.03 | 32 | G/A |
| hg E | E1b1b1c | M123 | 2 | AC010889 | | 20223974 | GTTGCCCAGGA ATTTGCAT | CACAGAGCAAG TGACTCTCAAA G | 0.15 | 89 | taaactaggtgccacgtcgtg aaagtctgacaaCATTTC TAGGTATTCAGGCG ATG | F | 0.1 | 56 | T/G |
| hg E | E1b1b1d | M281 | 4 | AC010889 | rs1344737 0 | 20223888 | AGCAAAGTTGAG GTTGCACA | TGGGCAACACC AGAATCTAA | 0.15 | 93 | gtgccacgtcgtgaaagtctg acaaGCACAAACTCA GTATTATTAAAC | F | 0.06 | 48 | T/C |
| hg E | E1b1b1e | V6 | 3 | AC012068 | | 6992007 | GATGGCACAGT GTTCGACAG | CTTCTCTCCAAA TGCCTGCT | 0.4 | 102 | taggtgccacgtcgtgaaagt ctgacaaCCTGCTGCC GCATCTGCA | R | 0.02 | 46 | T/C |
| hg E | E2 | M75 | 2 | AC010889 | rs2032639 | 20349565 | TGACTTGTCAAA AGCCAAAACA | TTGAACAGAGG CATTTGTGA | 0.1 | 123 | taggtgccacgtcgtgaaagt ctgacaaGAAAAGACA ATTATCAAACCACAT CC | F | 0.1 | 54 | C/T |

**Supplementary Material 3: Phylogenetic tree of NRY SNPs**



**Supplementary Material 4: MtDNA haplogroups observed among U.S. Americans and their assumed geographic region of origin**

| | Assumed continental origin | | | |
|---|---|---|---|---|
| **mtDNA haplogroup** | **Asian** | **Eurasian** | **African** | **Native American** |
| A | 1 | | | |
| A2 | | | | 1 |
| A5 | 1 | | | |
| B2 | | | | 1 |
| B4a | 1 | | | |
| B4b1 | 1 | | | |
| B4c | 1 | | | |
| B5b | 1 | | | |
| C1 | | | | 1 |
| D/E/G | 1 | | | |
| D/G | 1 | | | |
| D1 | | | | 1 |
| D4a | 1 | | | |
| D4e | 1 | | | |
| D4i | 1 | | | |
| D4k | 1 | | | |
| D5b | 1 | | | |
| E2 | 1 | | | |
| F1a | 1 | | | |
| F1b | 1 | | | |
| F2a | 1 | | | |
| F3b | 1 | | | |
| G | 1 | | | |
| H | | 1 | | |
| H11 | | 1 | | |
| H13a | | 1 | | |
| H1a | | 1 | | |
| H1b | | 1 | | |
| H1c | | 1 | | |
| H3a | | 1 | | |
| H5 | | 1 | | |
| H6 | | 1 | | |
| HV0 | | 1 | | |
| I | | 1 | | |
| J1b | | 1 | | |
| J1c | | 1 | | |
| J2a | | 1 | | |
| K | | 1 | | |
| L0a | | | 1 | |
| L1b | | | 1 | |
| L1c | | | 1 | |
| L2a1 | | | 1 | |
| L2b | | | 1 | |
| L2c | | | 1 | |
| L2d | | | 1 | |
| L3 | | | 1 | |
| L3a | | | 1 | |
| L3b | | | 1 | |
| L3d | | | 1 | |
| L3e1 | | | 1 | |
| L3e2 | | | 1 | |
| L3e3 | | | 1 | |
| L3e4 | | | 1 | |
| L3f | | | 1 | |
| L3h | | | 1 | |
| M10 | 1 | | | |
| M35 | | 1 | | |
| M7a | 1 | | | |
| M7b | 1 | | | |
| M8a | 1 | | | |
| M9a | 1 | | | |
| N1a | | 1 | | |
| N1b | | 1 | | |
| N9 | 1 | | | |
| R* | 0.5 | 0.5 | | |
| T1 | | 1 | | |
| T2 | | 1 | | |
| U2 | | 1 | | |

| | | | | |
|---|---|---|---|---|
| U3 | | 1 | | |
| U4 | | 1 | | |
| U5a | | 1 | | |
| U5b | | 1 | | |
| U6a | | | 1 | |
| U8a | | 1 | | |
| W | | 1 | | |
| X2 | | 1 | | |
| X2a | | | | 1 |

**Supplementary Material 5: NRY DNA haplogroups observed among U.S. Americans and their assumed geographic region of origin**

| | Assumed continental origin | | | |
|---|---|---|---|---|
| **NRY haplogroup** | **Asian** | **Eurasian** | **African** | **Native American** |
| A | | | 1 | |
| B | | | 1 | |
| C | 1 | | | |
| D | 1 | | | |
| E1a | | | 1 | |
| E1b1a*(xE1b1a4,E1b1a7) | | | 1 | |
| E1b1a7 | | | 1 | |
| E1b1b1*(xE1b1b1a,E1b1b1b,E1b1b1c,E1b1b1d,E1b1b1e) | | 0.5 | 0.5 | |
| E1b1b1a*(xE1b1b1a1,E1b1b1a2,E1b1b1a3) | | 0.5 | 0.5 | |
| E1b1b1a1*(xE1b1b1a1a,E1b1b1a1b) | | 0.5 | 0.5 | |
| E1b1b1a2 | | 1 | | |
| E1b1b1a3 | | 0.5 | 0.5 | |
| E1b1b1b | | 1 | | |
| E1b1b1c | | 0.8 | 0.2 | |
| E2 | | | 1 | |
| G | | 1 | | |
| I | | 1 | | |
| J*(xJ2) | | 1 | | |
| J2 | | 1 | | |
| K*(xL,M1,NO,P) | 0.333 | 0.333 | 0.333 | |
| N1c | | 1 | | |
| O | 1 | | | |
| Q1a | | | | 1 |
| R1a | | 1 | | |
| R1b1b2 | | 1 | | |
| R2 | | 1 | | |

**Supplementary Material 6: Genotype data for 24 ancestry-sensitive SNPs together with NRY and mtDNA haplogroup data for U.S. Americans**

See extra excel file Supp Mat 6