2000-IJ-CX-K006 Cumulative Technical Report

"Forensic Applications of Y Chromosome STRs and SNPs"

Principal Investigators: Michael Hammer and Alan J. Redd

Table of Contents

I. Abstract

We identified and characterized 20 novel STRs on the non-recombining portion of the Y chromosome (NRY) that are robust and informative for forensic casework. These Y-STRs, comprised of tetranucleotide, pentanucleotide, and hexanucleotide repeats, greatly improve resolution among paternal lineages above levels obtained with previously used Y-STRs. Multiplex protocols were optimized to amplify 41 Y-STRs in 5 PCR reactions (an additional 2 Y-STRs were typed in uniplex assays). A total of 38 Y-STRs was typed in a panel of 2,517 U.S. samples representing African-Americans, European-Americans, Hispanics, Native Americans, and Asian Americans, as well as a large worldwide database. The entire U.S. Y-STR database is available for online searches to estimate frequencies of Y chromosome haplotypes determined from crime scene material. Comparisons of commercially available kits revealed that Applied Biosystems Yfiler©, which contains three of our novel Y-STRs, is superior to others. The 11 "core" Y-STRs recommended by the Scientific Working Group on DNA Analysis Methods were analyzed to estimate the extent of population structure within and among ethnic groups in the U.S. The analyses support the creation of separate African-American, European-American, Hispanic-American, and Asian-American databases in which samples of the same ethnic group from different geographic regions within the U.S can be pooled. We recommend that separate databases be constructed for different Native American groups. A set of 61 Y chromosome single-nucleotide-polymorphisms (Y-SNPs) was also typed in the U.S. database to infer the geographic origins of Y chromosomes in the U.S. and to test for paternal admixture among U.S. ethnic groups. Admixture estimates vary greatly among populations and ethnic groups. A series of analyses was performed to test for the effects of inter-ethnic admixture on the structure of Y-STR diversity in the U.S. Results indicated that low levels of genetic heterogeneity 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 were removed, or for Native American populations when European-derived chromosomes were removed. We infer that both inter-ethnic admixture and population structure in ancestral source populations contributed to fine scale Y-STR heterogeneity within U.S. ethnic groups. Finally, empirical tests of association between Y-chromosome and autosomal markers are presented and a theoretical framework for determining a joint match probability is recommended. A conservative estimate of the joint probability is obtained by multiplying the largest value of the group autosomal match probabilities by the estimated matching probability for the Y chromosome.

II. Executive Summary

This award funded a project to develop DNA markers on the Y-chromosome for forensic applications. The specific aims of the grant were to: 1) develop a DNA typing system that targets the male-specific portion of the human genome (the non-recombining portion of the Y chromosome or NRY), 2) identify a set of informative NRY markers that are robust in forensic analysis (Y-SNPs and Y-STRs), 3) develop detailed protocols for PCR-based multiplex genotyping kits, and 4) construct a database of Y chromosome markers in U.S. populations. We identified and characterized 20 novel Y-STRs on the NRY (Note: some STRs are duplicated such that 15 STR sequences map to 20 locations; in this report we refer to each location as a Y-STR). These Y-STRs are comprised of tetranucleotide repeats (DYS449, DYS453, DYS454, DYS455, DYS456, DYS458, DYS459ab, DYS464abcd, and DYS724ab), pentanucleotide

repeats (DYS446, DYS447, DYS450, DYS452, and DYS463), and hexanucleotide repeats (DYS448). These novel Y-STRs greatly improve resolution among paternal lineages beyond levels obtained with previously used Y-STRs. After discussions with Applied Biosystems, three of our markers (DYS448, DYS456, DYS458) were incorporated in their Yfiler kit. Multiplex protocols were developed to amplify 41 Y-STRs in 5 PCR reactions (an additional 2 Y-STRs were typed in uniplex assays). A total of 38 Y-STRs was typed in a panel of 2,517 U.S. samples representing African-Americans, European-Americans, Hispanics, Native Americans, and Asian Americans. An additional 5 Y-STRs were typed in a subset (1,000) of these samples. The resolution of commercially available kits was compared using these markers. The Yfiler kit showed greater capacity to resolve paternal lineages than the standard U.S. core Y-STRs or the Promega PowerPlex Y kit in all ethnic groups, although the complete set of 38 Y-STRs had higher capacity to resolve paternal lineages, especially in Native American populations. The entire Y-STR database is available for online searches to estimate frequencies of Y chromosome haplotypes determined from crime scene material.

The 11 core U.S. loci recommended by the Scientific Working Group on DNA Analysis Methods were analyzed in our U.S. database of 2,517 individuals from 38 populations to estimate the extent of population structure within and among ethnic groups in the U.S. A multidimensional scaling (MDS) plot placed the populations into four discrete clusters (African-Americans, European-Americans, Hispanic-Americans, and Asian-Americans) and one dispersed cluster of Native Americans. An analysis of molecular variance (AMOVA) indicated that a large proportion of the total genetic variance is partitioned among ethnic groups (24.8%); while only a small amount (1.5%) is found among populations within ethnic groups. Separate AMOVA analyses within each ethnic group showed that only the sample of Native Americans contains statistically significant among-population variation. Pair-wise population differentiation tests did uncover heterogeneity among European-American and among Hispanic-American populations; however, this was due to only a single sample within each group. For example, only the NYC European-American sample and the Mesa Arizona Hispanic-American sample differed in frequency of Y-haplotypes when compared with a subset of populations within their respective ethnic groups.

In sum, analyses of Y-STRs indicated a lack of significant geographic structure among African-American and Asian-American populations, minor heterogeneity among European-American and Hispanic-American populations, and broad-scale subdivision among Native American populations. The extremely consistent patterns of genetic structure observed in this study and two others of similar scope (1, 2) lead us to make the following recommendations for the construction of U.S. databases of Y-STRs: 1) it is good policy to continue gathering more data from additional regional populations, especially those not represented in existing databases, 2) there is no evidence at present that pooling samples from different geographic regions will lead to strong biases in the estimation of Y-STR haplotype frequencies for African-American, European-American, Hispanic-American and Asian-American populations, and 3) separate, larger databases from Native American subpopulations (tribal groups) are needed. Methods to correct for very low levels of structure within European-American and Hispanic-Americans may need to be considered (3).

A set of 61 Y chromosome single-nucleotide-polymorphisms (Y-SNPs) was typed in the same database (i.e., 2,517 individuals from 38 populations) to infer the geographic origins of Y chromosomes and to test for paternal admixture among African-Americans, European-Americans, Hispanic-Americans, Asian-Americans, and Native-Americans. Admixture estimates varied greatly among populations and ethnic groups. The frequencies of non-European (3.4%) and non-Asian (4.5%) Y chromosomes were generally low in European-American and Asian-American populations, respectively. The frequencies of European Y chromosomes in Native-American populations ranged widely (i.e., 7-89%) and followed a West to East gradient, whereas they were 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 were also relatively constant among geographic regions; however, the African contribution was 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 performed "subtraction" analyses in which Y chromosomes inferred to be admixed by Y-SNP analysis were removed from the database and pairwise population differentiation (PPD) tests were implemented on the remaining Y-STR haplotypes. Results indicated that low levels of heterogeneity observed between pairs of Hispanic American populations (see above) disappeared when African-derived chromosomes were removed from the analysis. This was not the case for an unusual sample of European-Americans from New York City (see above) when its African-derived chromosomes were removed, or for Native American populations when European-derived chromosomes were removed.

The highly similar admixture rates among African-American samples from different locales is consistent with the absence of statistically significant PPD tests based on our Y-STR data. Likewise, the similarly high frequencies of European-derived Y chromosomes in our Hispanic-American samples may account for low levels of Hispanic Y-STR heterogeneity in this survey, as well as in the surveys of Kayser et al. (1) and Budowle et al. (2). The finding of frequencies of 6-18% African-derived Y chromosomes in Hispanic-American samples from the eastern seaboard suggests that broader regional patterns of African admixture in Hispanic-American populations should be investigated. The Y-SNP analysis provides insights into why Native American samples show such high levels of heterogeneity based on Y-STR AMOVA, MDS, and PPD tests. From West to East there is a pronounced admixture gradient, ranging from ~10%-90% European Y chromosomes, respectively. These results suggest that regional variation in inter-ethnic admixture, as well as population structure in ancestral European, Hispanic and Native American source populations, are important factors leading to population substructure within an ethnic group.

Empirical tests of association between Y-chromosome and autosomal markers are presented and a theoretical framework for determining a joint match probability is recommended. Statistical analyses of association were performed in sixteen U.S. populations between the autosomal genotypes from loci CSF1PO, FGA, THO1, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S512, D21S11 and Y chromosome haplotypes from loci DYS19, DYS385ab, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS438, and DYS439. The sample populations include individuals of European-American, African-American,

Hispanic-American, Native-American, and Asian-American ancestry. The results are consistent with independence of Y and autosomal markers. Hence, it is appropriate to compute joint-match probability by multiplying the Y haplotype frequency with the appropriately-corrected autosomal frequency. Since two individuals sharing the same Y haplotype are more recently related than two randomly-chosen individuals, the autosomal frequencies have to be corrected to account for this, and we develop this correction here. Likewise, one must also compute these probabilities for each appropriate subpopulation. A conservative estimate of the joint probability is obtained by multiplying the largest value of the group autosomal match probabilities by the estimated matching probability for the Y chromosome. Finally, we suggest an approach for presenting a single match probability (as opposed to separate values for each major ethnic group) that is based on weighting the major ethnic groups by the appropriate census sizes.

III. Overview

During the past ten years there has been a dramatic increase in the use and reliance on forensic DNA analysis. This has been especially true in sexual assault cases that routinely consist of evidentiary stains that are a mixture of body fluids from the victim and assailant. Current DNA differential analysis techniques permit the separation of the male and female components of these mixed stains (4); however, a complete separation is not always possible due to the size and condition of the stain and the percentage of each component present (5-10). In addition, mixed semen stains from vasectomized or azoospermic individuals cannot be relied upon to yield DNA information with current techniques. Because of the limitations of the separation process, forensic scientists are faced with the challenge of interpreting the significance of a mixed profile consisting of two or more contributors. This has become an increasingly important issue with the ability to analyze smaller samples through the use of Polymerase Chain Reaction (PCR) techniques. In many circumstances the application of PCR-based DNA typing methods results in the failure to amplify the minor (e.g., male) component of DNA mixtures due to competition with alleles from the major (e.g., female) component.

One approach to resolving this issue is to target male-specific polymorphisms on the non-recombining portion of the Y chromosome (NRY). By genotyping markers on the NRY, DNA typing will become more useful in cases involving 1) unsuccessful differential lysis separations of male and female cells as a result of the age and/or quality of semen stains, 2) azoospermic men, 3) the presence of other body fluid mixtures from victims and suspects of different sex, and 4) two or more male semen donors (5, 7, 9, 11, 12). Thus, the ability to target polymorphisms on the NRY expands the application of DNA technology to samples that were previously difficult to interpret and/or yielded no results.

Two classes of NRY markers are short tandem repeats (Y-STRs) and single nucleotide polymorphisms (Y-SNPs). The combination of alleles at multiple Y-STRs on a single Y chromosome defines a Y-STR haplotype, whereas the combination of alleles at multiple SNPs defines a NRY haplogroup (13). The high geographic specificity of many NRY haplogroups (14, 15) provides forensic scientists with clues to the source of the male genetic material left at a crime scene. The forensic utility of Y-STRs results from their high levels of polymorphism in human populations (16), their small size in base pairs (~100-400 bp), and the ability to type multiple Y-STRs in a single PCR reaction (5, 11, 17, 18). One limitation of Y-STRs in forensic and paternity applications is the lack of independence of these markers on the NRY (*i.e.*, the lack

of recombination). The difficulty in distinguishing paternally related males in a population means that Y-STR haplotypes will have reduced inclusion probabilities compared with autosomal STRs. However, Y-STRs provide a valuable addition to the forensic scientist's tool kit, especially as more variable markers are discovered and the potential to distinguish Y-chromosomes in a population increases.

The major questions to be addressed in this report are 1) what is the best way to construct a Y-STR database for forensic purposes, 2) what is the degree of Y chromosome population structure among U.S. ethnic groups and among populations within U.S. ethnic groups, 3) can population samples from different geographic regions within the U.S. be pooled in forensic databases, 4) how much admixture is there among U.S. ethnic groups and how do admixture rates vary regionally, 5) what is the effect of admixture on levels of population structure, and 6) how do we calculate a Y chromosome-autosome joint match probability? One must be wary of the possibility that different U.S. groups will have different levels of sub-structure (e.g., Native American populations that have a history of partial isolation). Because U.S. populations are composed of diverse immigrant groups it is also necessary to consider the effects of recent admixture. The degree to which these factors affect U.S. population groups and the impact they might have on statistical analyses of Y chromosome data can only be understood through the establishment of detailed databases of well-defined U.S. sub-populations and their parental populations in Europe, Africa, and Asia. Here we analyze both Y-STRs and Y-SNPs in a sample of 2,517 Y chromosomes from 5 U.S. ethnic groups: African-Americans (AA), European-Americans (EA), Hispanic-Americans (HA), Asian-Americans (SA), and Native Americans (NA).

IV. Main Findings

A. Forensic value of novel STRs on the human Y chromosome.

1. Background and Significance

At the time we began this work, approximately 35 STRs had been described on the Y chromosome. In total, these Y-STRs include: 5 dinucleotide (YCAI, YCAIIab, YCAII, and DYS288) 6 trinucleotide (DYF371, DYS388, DYS392, DYS425, DYS426, DYS436), 24 tetranucleotide (DYS19, DYS385ab, DYS389I, DYS389AB, DYS390, DYS391, DYS393, DYS434, DYS435, DY437, DYS439, DYS441, DYS442, DYS443, DYS444, DYS445, DYS460, DYS461, DYS462, G10123, A10, C4, and H4), and 2 pentanucleotide repeats (DXYS156, and DYS438). Most of the Y-STR primer pairs amplify one PCR product, while some Y-STR primer pairs amplify two or more PCR products (YCAI, YCAII, YCAIII, DYF371, DYS385, G10123). With the exception of one multicenter-study using 13 Y-STRs (11), global population screens using the majority of the published Y-STRs had not been performed; although much progress had been made in establishing a large database of European populations using the following Y-STRs: DYS19, DYS385ab, DYS389I, DYS389II-I, DYS390, DYS391, DYS392, and DYS393. These Y-STRs define the "minimal haplotype", while the addition of YCAIIab to the these Y-STRs has been termed the "extended haplotype" (11, 19, 20).

The identification of additional Y-STRs was warranted for several reasons. First and foremost, increasing the number of highly polymorphic markers would improve the ability to distinguish paternal lineages. There are shared Y-STR haplotypes in populations because either males share identity by descent or because a particular set of Y-STRs does not distinguish closely related, but

different, paternal lineages. In a sample of 41 European populations the discrimination capacities were 52% (n = 4,688 individuals) and 71% (n = 1,957 individuals) using the minimal and extended sets of Y-STRs, respectively (20). The sharing of paternal lineages is likely to be more common in isolated populations where there is a higher degree of genetic drift, such as Native American populations. Second, there is a need to identify more Y-STRs that have longer repeats units. For example, YCAII is a polymorphic dinucleotide marker that suffers from "stutter" products during the PCR process due to polymerase slippage (21). Stutter products are pronounced in dinucleotide repeats. Stutter bands are often reduced in longer repeat motif STRs, these loci can provide additional resolution in sample mixtures of multiple-male DNA profiles (22). Third, a large pool of Y-STRs would provide a diverse sample of markers from which one can select tailored sets of STRs with distinct characteristics for multiplex design for particular applications. A small multiplex of the most informative Y-STRs could more efficiently distinguish Y-chromosome lineages than a set of a dozen or more less informative Y-STRs. Finally, increasing the number of Y-STRs will improve the estimation of the time to the most recent common ancestor (TMRCA). The TMRCA between two Y-STR haplotypes provides a natural metric to describe the relatedness between two individuals and could be used to make exclusions in forensics (23). By including more Y-STRs, estimates of the TMRCA become more precise (23, 24) and the ability to exclude paternal relatives increases.

2. Results

In our first publication (25), we identified and characterized 14 novel Y-STRs mapping to 18 locations on the NRY and typed them in two samples, a globally diverse panel of 73 cell lines, and 148 individuals from a European-American population. These Y-STRs include 12 tetranucleotide repeats (DYS449, DYS453, DYS454, DYS455, DYS456, DYS458, DYS459ab, and DYS464abcd), 5 pentanucleotide repeats (DYS446, DYS447, DYS450, DYS452, and DYS463), and 1 hexanucleotide repeat (DYS448). Sequence data were obtained to designate a repeat number nomenclature. Multiplex PCR assays were designed in our lab, as well in collaboration with John Butler's lab (18), incorporating both previously described and novel Y-STRs. Gene diversities of an additional 26 Y-STR locations, including the most commonly used in forensic databases, were directly compared in the cell line DNAs. Six of the 10 most polymorphic markers include the newly identified Y-STRs. Furthermore, these novel Y-STRs greatly improved the resolution of paternal lineages, above the level obtained with commonly used Y-STRs, in the European-American population.

More recently, we identified and characterized additional novel Y-STRs (DYS724ab) and extensively genotyped an additional four Y-STRs (DYS442, DYS570, DYS576, DYS607) (26, 27). We typed a total of 38 Y-STRs in a database of 2,718 U.S. individuals from 38 populations representing 5 U.S. ethnic groups (see below), and a global database of 2,500 samples from Europe, Africa, and Asia. We also typed a total of 43 Y-STRs (**Figure 1**) in a *subset* of this database (n = ~1,0000). Our multiplex assays are shown schematically in **Figure 2**. The rank order of diversities of single copy Y-STRs are given in **Table 1**.

3. Conclusions

The novel Y-STRs discovered and characterized in this research are extremely useful for forensic casework for two reasons: 1) they provide additional power to resolve Y chromosome haplotypes, and 2) they are all tetra-, penta- and hexa-nucleotide repeats. The novel Y-STRs

have more than doubled the number of known pentanucleotide markers and they include the first hexanucleotide repeat on the NRY (DYS448). These longer repeat motif STRs may be useful for improving the interpretation of sample mixtures (12, 28). Depending on the particular STR, the stutter peak heights of dinucleotide and trinucleotide repeats can be higher than 30% of their corresponding STR allele, while stutter products of tetranucleotides are approximately 15%, and pentanucleotide repeats may have stutter products of less than 1-2% (22). These novel Y-STRs with longer repeat motifs should be evaluated for use in mixture studies.

Interestingly, these results suggest that different Y-STRs may vary in their gene diversity depending on the particular population sample under investigation, and perhaps the length of the alleles. While most of the variation in diversity was Y-STR dependent, there was some variation among different geographic groups in the rank order of Y-STR diversity. Our comparative analysis of Y-STR diversity indicates that 5 of our novel Y-STRs (DYS449, DYS458, DYS463, DYS447, and DYS448) are among the 10 most polymorphic single-copy Y-STRs (**Table 1**), while one multicopy Y-STR (DYS464abcd) appears to be the most polymorphic Y-STR yet described (data not shown).

The addition of novel Y-STRs to the minimal haplotype improves the ability to distinguish Y chromosomes in European-American samples. The most common European haplotype, 14-(11,14)-13-16-24-11-13-13, is found in 3.1% of 9,972 individuals in the Y-STR HRDatabase (20). This haplotype was found at 5.4% in the European-American sample. DYS464 alone distinguishes 75% of these haplotypes in the European-American sample. The remaining shared haplotypes are distinguished using only three additional Y-STRs. Thus, these European-American Y chromosomes are not identical by descent, as the minimal haplotype suggested. When the Bayesian method of Walsh (23) was used to estimate the TMRCA between pairs of Y chromosomes that are not resolved with 36 Y-STRs, assuming a mutation rate of 2.0×10^{-3} we found a TMRCA of only 11.8 generations (95% C.I. = 1.7-39.2 generations). Future studies of Y-chromosomes from well-characterized pedigrees could test the potential to distinguish between patrilineal relatives of various degrees using highly variable Y-STRs.

B. A Novel, searchable, online database—*USAYSTR*—for estimating Y chromosome haplotype frequencies in the U.S.

1. Background and Significance

As in any DNA analysis, in the event of a match between the Y-STR haplotypes for a case sample and a suspect sample, it is desirable to have an estimate of the probability that a match would occur by chance. Multiplication of single locus allele frequencies to obtain estimated Y-STR haplotype frequencies is not appropriate since all STRs are from the NRY and hence are completely linked. As with mitochondrial DNA, databases of complete Y-STR haplotypes have to be generated as a source for estimating frequencies. Because of the much higher diversity of combined Y-STR haplotypes compared with single Y-STR loci, such a database has to be much larger in order to be able to serve as a reliable representation of the underlying population haplotype frequencies. Also, since the NRY is more sensitive to genetic drift as a result of its haploid and paternal mode of inheritance, populations are more likely to show statistically significant differences with regard to Y-STR haplotype frequencies (29). The potential for population structure must therefore be considered when generating Y-STR haplotype databases.

Recently, a large Y-STR haplotype reference database (YHRD) for U.S. populations was made available (3). However, this database contains only 9-locus Y-STR haplotypes, which are determined by what is commonly referred to as the minimal haplotype loci (20). Here we describe a novel U.S. Y chromosome database based on the 11 "core" U.S. Y-STR loci recommended by the Scientific Working Group on DNA Analysis Methods (SWGDAM) and now used in forensic casework in the U.S. We make this database available to the U.S. forensic DNA community, along with tools for obtaining Y-STR haplotype frequencies needed for calculating matching or paternity probabilities in cases of non-exclusions in forensic analysis and paternity testing.

2. Results

An online (http://amadeus.biosci.arizona.edu/~kcaldero/str.php) Y-chromosome DNA database called *USAYSTR* is constructed to aid forensic scientists estimate Y chromosome haplotype frequencies (**Figure 3**). *USAYSTR* consists of genotypes at 38 Y-STR markers in 2,517 individuals from 38 populations currently living in 10 states (**Figure 4**). The 38 Y-STRs include the 11 core U.S. Y-STRs (DYS19, DYS385ab, DYS389I, DYS389II, DYS390, DYS391, DY392, DYS393, DYS438, and DYS439), STRs found in commercial Y- chromosome multiplexes (DYS437, DYS448, DYS456, DYS458, H4, and C4), as well as many additional Y-STRs (DYS388, DYS426, DYS442, DYS446, DYS447, DYS449, DYS450, DYS452, DYS453, DYS454, DYS455, DYS459ab, DYS460, DYS463, DYS464a-d, DYS570, DYS576, DYS607, DYS724ab, YCAIIab) (**Table 2**). Features of *USAYSTR* include: (1) flexibility in sorting the match results alphabetically or numerically, (2) the ability to include or exclude populations, (3) haplotype frequency estimation with either a standard 95% confidence interval (CI) or a bootstrap 95% CI, (4) a printout of the results, and (5) flexibility in the choice of STR markers with the potential to query haplotypes composed of up to 43 Y-STRs.

Based on the 11 U.S. core loci, we find that the number of haploypes (corrected for sample size) varies by ethnic group (**Table 2**). Haplotype resolution is highest in Asian-Americans (98%) and African-Americans (87%), followed by Hispanic-Americans (81%) and European-Americans (72%). Haplotype resolution in pooled Native American samples is much lower (65%). The frequency of common haplotypes, as well as the number of haplotypes observed two or more times, also vary by ethnic group. The most common haplotype is found at the following frequencies for African-American, European-American, Hispanic-American, Asian-American, and pooled Native Americans: 0.7%, 2.8%, 2.3%, 3.2%, and 3.5%. When we included a total of 38 Y-STRs, haplotype resolution is near 100% for all ethnic groups, except Native Americans, which has a value of 90%. The Y-Filer kit (AB), which includes three of the loci discovered during this research (DYS448, DYS456, DYS458), produces higher haplotype resolution (average for all ethnic groups = 93.4%) than either the Power-Plex (Promega) (82.6%) or the 11 core U.S. loci (82.2%) (**Figure 5**).

3. Conclusions

USYSTR provides the DNA forensic community the ability to estimate Y-STR haplotype frequencies with associated 95% confidence intervals. Common Y chromosome haplotypes using the 11 U.S. core loci can be resolved by adding additional informative markers. Forensic labs that have validated the AB Y-Filer kit will achieve excellent resolution. For labs that have access to custom multiplex assays, we recommend using the following high-resolution markers

to be used as a supplement to the 11 core U.S. loci: DYS449, DYS464, DYS570, DYS576, DYS463, and DYS607.

It is important to note (see below) that we do not endorse pooling of Native American populations for population genetic analyses. When considering individual Native American populations, haplotype resolution tend to be much lower and the frequency of common haplotypes much higher. For example, in our Apache sample, haplotype resolution is only 44.8% and the frequency of the most common haplotype is 16.7% (data not shown). In the following sections we explore reasons for the observed patterns of Y chromosome diversity in Native American and other U.S. populations.

C. Genetic structure among 38 populations from the United States based on 11 U.S. core Y chromosome STRs.

1. Background and Significance

A key consideration for the proper scientific use of Y-STRs in DNA forensics is the creation of an appropriate population database. A large population database is necessary to help estimate the probability that two or more unrelated males share the same Y-STR haplotype (19). To obtain an accurate estimate of a haplotype's frequency the database should represent the range ethnic groups within a population (30). Otherwise, the frequency of a Y-STR haplotype of an individual whose ethnic group is not represented in the database is likely to be underestimated. Analyses of patterns of population subdivision can provide a means for determining the appropriate structure of a database. Given limitations in the number of individuals that can be sampled, it is important to assess whether data can be pooled for populations from the same ethnic group that have been collected from different geographic regions in the U.S. Because Y chromosome haplotypes have been shown to exhibit large frequency differences among populations from different regions of the world (31, 32), and because U.S. populations are composed of individuals with ancestry deriving from many parts of the world, empirical studies are required to measure the proportion of variation within and among populations of different ethnic groups (33). Here we estimate the extent of U.S. population structure with an analysis of 11 Y-STRs in a sample of 2,517 individuals representing 38 U.S. populations.

2. Results

Polymorphism data are collected from an analysis of the 11 Y-STRs recommended by SWGDAM for use in the U.S. The population samples include individuals of self-described ancestry from five ethnic groups (African-Americans, European-Americans, Hispanic-Americans, Asian-Americans, and Native Americans) currently living in 10 Western (AZ, NM), Midwestern (SD, OH), Northeastern (VT, CT, NY, VA), and Southern states (NC, FL) (**Figure 4**). Population structure is examined using multidimensional scaling (MDS), analysis of molecular variance (AMOVA), and pairwise-population differentiation (PPD) tests. A MDS plot places the populations into four discrete clusters (African-American, European-American, Hispanic-American, and Asian-American) and one dispersed cluster of Native American populations (**Figure 6**). AMOVA indicates that most of the genetic variance (73.7%) is found within populations; a notable amount (24.8%) is found between ethnic groups; while only a small amount (1.5%) is found among populations within ethnic groups (**Table 3**). Separate AMOVA analyses within each ethnic group show that only Native Americans contain statistically significant among population variation, while no statistically significant variation is

found among populations within other ethnic groups (**Table 3**). PPD tests uncover no statistically significant differences among the 45 comparisons of pairs of African-American samples (**Table 4A**). For European-American samples, 3 out of 45 comparisons (NYC-CT, NYC-VA, and NYC-NC) are found to be statistically significant at $\alpha = 0.01$ level (**Table 4B**). Similarly, 2 of 36 comparisons among pairs of Hispanic-American samples reject panmixia (Mesa-CT, Mesa-VA) (**Table 4C**). Notably, this heterogeneity is due to only a single sample within each group (NYC and Mesa). For Native American samples, 13 out of 21 comparisons are statistically significant (**Table 4D**). Because multiple comparisons are made with the PPD tests, caution should be exercised in the interpretation of the results without a multiple test correction.

3. Conclusions

This is the third large-scale study of the structure of Y-STR diversity in multiple U.S. populations (1, 2), and only the second that we are aware of that screens the 11 core loci recommended by SWDAM (2). We find that levels of diversity based on the 11 U.S. core Y-STR loci (**Table 2**) are similar to those in Budowle et al.'s (2) 12 Y-STR analysis and consistently higher than those found by Kayser et al. (1) based on 9 Y-STRs. For example, discrimination capacity for Kayser et al.'s (1) African-American, European-American, and Hispanic-American samples were on average 8% lower than for our samples. Our survey differs from the other two published studies in that it examines multiple Native American populations. The results indicate that Native American populations have lower levels of Y chromosome diversity than other U.S. ethnic groups. This is reflected in a higher percentage of shared haplotypes and higher random match probabilities, both of which are important to take into consideration in forensic casework.

Similar to the results of Kayser et al. (1) we find that Y chromosomes are significantly differentiated among U.S. ethnic groupings, but not among populations within ethnic groups from different geographic regions in the U.S. The proportion of among ethnic group variance reported here is almost identical to that found (24.9%) in the three ethnic groups (African-American, European-American, and Hispanic-American) sampled by Kayser et al. (1), and higher (15.4%) than that reported in the 5 ethnic groups (African-American, European-American, Hispanic-American, Asian-American, and Native American) sampled by Budowle et al. (2). Our AMOVA results differ somewhat from those in previous studies in that the proportion of variance among-populations-within- groups is not significant for African-Americans, European-Americans, or Hispanic-Americans, either when placing all geographic populations into a single ethnic grouping, or when subdividing populations within each ethnic group into geographic regions in the U.S. (**Table 3**). Kayser et al. (1) found that the very small proportion of among-populations-within-groups variance in both their European-American and Hispanic-American samples (e.g., 1.8% and 2.6%, respectively) was statistically significant. The among-populations-within-groups variance (1.6%) for Budowle et al.'s (2) European-American sample was statistically significant, but not for their Hispanic-American sample (0.9%). Here, the only ethnic group with a statistically significant proportion of among-populations-within-group variance is Native Americans, where we find 9.5% of the total variance partitioned among 7 populations (tribes) (**Table 3**). Budowle et al. (2) found 3.0% of the total variance partitioned between their Navajo and Apache samples. The average among-populations-within-groups variance in the three studies is 1.2%. When we remove Native Americans from our analysis, the

among-populations-within-groups variance is only 0.4% (not statistically significant; data not shown).

Despite the lack of significant differentiation among regional African-American, European-American, Hispanic-American, and Asian-American populations in AMOVA, when multiple differentiation tests are performed among all pairs of populations some comparisons between European-American and Hispanic-American populations are statistically significant. The question we face is whether these comparisons are significant by chance or as a result of true biological differences. Our results are very similar to those of Kayser et al. (1), who did not find different frequencies of Y-STR haplotypes among their African-American samples, but did find heterogeneity within their European-American and Hispanic-American samples in pair-wise population differentiation tests. As is the case here, this heterogeneity was attributed to a single sample within each group. Because they could not identify an obvious reason why either of the samples was an outlier, they concluded that their result reflected chance (1). We concur that a single outlier does not support a pattern of broad-scale geographic structuring (e.g., as observed for Native American populations) and the combined results provide no compelling evidence for incorporating geographic structure within African-American, European-American, and Hispanic-American Y-STR databases at present. One implication of these results is that independent databases can be combined for each of these ethnic groups. Still, it would be prudent to continue sampling from additional populations to further assess the structure of U.S. populations.

The extent to which we expect significant population structure within an ethnic group depends mainly on four factors: levels of subdivision in the ancestral source populations, 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. Kayser et al. (1) suggested that the lack of geographic heterogeneity among their African-American samples may be a by-product of extensive migration from rural to urban areas during and after World War I. However, not enough is yet known about the structure of Y-STR haplotype variation among African source populations, or the extent of mixing among source populations in the process of forced migration to the U.S. The finding of relatively high levels of population structure in Native Americans is not unexpected given a long history of small effective population sizes, endogamy, isolation, and founder effects (34). Perhaps it is more surprising that Hispanic-American populations do not show stronger geographic structure 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, Hispanic-American populations are known to have differing degrees of Spanish, Native American, and African ancestry in different U.S. regions (1, 35). For example, Eastern Hispanics are expected to have more Afro-Caribbean ancestry than Hispanic populations from the Southwest, which are expected to have more Native American ancestry (36). However, our Eastern Hispanic-American populations do not seem to cluster closer to the African-American populations than do the southwest Hispanic-American populations (**Figure 6**).

In conclusion, the extremely consistent patterns of genetic structure observed in this study and previous studies (1, 2) suggest that pooling samples from different geographic regions will not lead to strong biases in the estimation of Y-STR haplotype frequencies for African-American, European-American, Hispanic-American, and Asian-American populations. On the other hand,

separate larger databases from Native American subpopulations are needed to infer match probabilities for different tribal groups. Finally, the continued collection of core Y-STR data from additional populations is needed to ensure that we construct databases that most accurately reflect the structure of U.S. populations.

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

1. Background and Significance

The MDS and AMOVA analyses of Y-STRs indicate low levels of population structure within African-American, European-American, Hispanic-American, and Asian-American groups. However, the more sensitive PPD test detected heterogeneity in our European-American and Hispanic-American samples. The fact that a single sample within each of the European-American (NYC) and Hispanic-American (Mesa) ethnic groups accounts for all of the statistically significant PPD tests argues against a broad-scale pattern of geographic structuring within these ethnic groups. One possibility is that the heterogeneity results from chance rather than any true biological differentiation (1). Additional inquiry into the ways that populations were sampled, the extent of admixture in each sample, and the gathering of additional samples from these and other regions may help to pinpoint the underlying causes of these observations.

A major component of population structure in the U.S. may be determined by admixture among ethnic groups. Because the largest proportion of among-group genetic variation in the U.S. is partitioned between ethnic groups, and only a small proportion is found among populations within ethnic groups (1, 2, 37), it is imperative to examine the extent to which there is regional variation in the rates of inter-ethnic admixture. There is a body of literature indicating that substantial mixing among some U.S. ethnic groups has occurred (35, 38-43). If inter-ethnic admixture rates do, in fact, vary regionally, then there could be a need for regional forensic databases.

In this study we employ a set of 61 Y-SNPs to estimate levels of population structure and rates of admixture among U.S. populations. Our estimates of population heterogeneity are then compared with similar estimates based on Y-STRs. The advantage of Y-SNPs for estimating admixture rates is that they permit direct estimates of admixture deriving from *multiple* parental source populations. This is due to the high geographic specificity of Y-SNPs, which makes these markers a powerful tool for inferring the ancestral source population(s) of paternal lineages. Through use of a large database of SNP haplogroup frequencies in samples collected from many regions of the world, we are able to infer the geographic provenance of Y chromosomes in the U.S. population; i.e., whether they originated in African, European, Asian, or Native American source populations.

2. Results

A set of 61 Y-SNPs is typed in 2,517 individuals from 38 populations in the U.S. The 61 Y-SNPs mark all 18 major haplogroups (A-R) on the Y chromosome tree, as well as several sub-lineages providing key information on the continental origins of Y chromosomes (**Figure 7**). While there is general correspondence between estimates of population structure based on Y-STRs and Y-SNPs, AMOVA indicates a greater proportion of SNP variation partitioned among

ethnic groups (32.3%) than for Y-STRs (24.8%). Most of the genetic variance (65.8%) is found within populations and only a small amount (2.0%) is found among populations within ethnic groups (**Table 5**). Separate AMOVA analyses within each ethnic group show that only the Native Americans contain high levels of among population variation (17.9%), while ≤1% of the total variation is partitioned among populations within other ethnic groups (**Table 5**). A MDS plot reflects these patterns, placing African-Americans, European-Americans, Hispanic-Americans, and Asian-Americans into discrete clusters, with Native Americans being more dispersed. Patterns of admixture vary dramatically across ethnic groups (**Figure 8**). All European-American (**Figure 9**) and Asian-American (data not shown) samples have very low levels of admixture. Approximately 30% of African-American Y chromosomes have European ancestry, and this proportion is relatively constant across the sampled geographic regions. Hispanic-American Y chromosomes descend mainly from European ancestors (79%). The proportion of Native American Y chromosomes is surprisingly consistent among Hispanic-American samples (averaging ~12%), while the African contribution (~7%) is clearly higher on the eastern seaboard (~16%) (**Figure 9**). Native Americans exhibit the largest regional variation in admixture rates, with European-derived Y chromosomes in Southwestern, Midwestern, and Eastern populations at frequencies of ~9%, 44%, and 90%, respectively.

3. Conclusions

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 (37). 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 (14) and higher mutation rates of Y-STRs (44, 45), which lead to much higher discrimination capacities and measures of Y chromosome diversity for Y-STR haplotypes (12) compared with Y-SNP haplogroups (**Table 4**). Similar to the case for Y-STRs (37), separate AMOVA analyses within each ethnic group show that only Native Americans contain high levels of among population SNP haplogroup variation (**Table 5**). In contrast, ≤1% (n.s.) of the total SNP variation is partitioned among-populations-within-groups when considering only African-American, European-American, Hispanic-American, and Asian-American samples (data not shown).

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 (**Figure 7**). 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 (**Figure 9**). Regional variation in the proportion of European Y chromosomes in African-American populations is apparent in the MDS plot in **Figure 8** (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 Hispanic-American populations (with

VA, CT, VT being placed on the right side of the Hispanic-American 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% ± 1.8%, 44.1% ± 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 Hispanic-American 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, Hispanic-American populations are known to have differing degrees of Spanish, Native American, and African ancestry in different regions of the U.S. (1, 35). 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 (46). In contrast to these Y chromosome results, both mtDNA and autosomal systems point to a much higher frequency of Native American maternal lineages in Hispanic-American populations, especially in Mexican Americans, and higher frequencies of African maternal lineages in Puerto Ricans and Cubans (1, 35, 36, 47). The larger European paternal contribution to Hispanic-American 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., (48)). Despite this regional variation, there were low levels of Hispanic Y-STR haplotype heterogeneity in our previous survey (37), as well as in the surveys of Kayser et al. (1) and Budowle et al. (2). Thus, geographic origin of samples is not a critical factor in the construction of U.S. Hispanic Y-STR databases.

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 (1, 2, 37), cases of statistically significant differences between particular pairs of populations were observed in pairwise population differentiation (PPD) tests (1, 37). For example, in our Y-STR database (37), 3 of 45 comparisons between pairs of European-American samples, and 2 of 36 comparisons between pairs of Hispanic-American samples, were statistically significant. All three European-American comparisons involved our sample from New York City, which differed from our Connecticut, Virginia, and North Carolina European-American samples. Both Hispanic-American comparisons involved our Mesa (Arizona) sample, which differed from our Connecticut and Virginia Hispanic-American samples. Similarly, PPD tests performed by Kayser et al. (1) on their database of 1,705 haplotypes based on 9 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 test whether the statistically significant PPD tests involving Y-STR haplotypes in our Hispanic samples (37) can be explained by variable frequencies of African Y chromosomes. When we remove the 36 African Y chromosomes identified by Y-SNPs from our Hispanic-American Y-STR database and repeat the PPD tests, we find no significant difference between any of the 36 pairs of Hispanic-American 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 (1, 2, 37).

Next, we test whether variable frequencies of African Y chromosomes in European-American populations leads to the significant heterogeneity in Y-STR haplotypes in our NYC, CT, NC, and VA samples (37). Upon removal of t(1, 37) European-American samples) and I-P30 (4.8% for NYC *versus* 11.3% ± 2.8% for other European-American samples), and the highest frequency of the Eastern European signature haplogroup, R-M17 (23.8% for NYC *versus* 7.4% ± 4.7% for other European-American 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 European-American 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 Native American populations plays an important role in structuring Native American Y chromosome variation. When we remove the 124 European-derived Y chromosomes from our Native American database, we find that AMOVA still results in significant differences in Y-STR haplotype frequencies among Western, Midwestern, and Eastern Native American 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 Native American 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 (34).

E. Joint Match Probabilities for Y Chromosome and Autosomal Markers.

1. Background and Significance

In several forensic settings, one may need to obtain a joint match probability for a set of autosomal and Y chromosome markers. The ability to combine data from the autosomes and the Y chromosome is particularly important in cases where the signal from the male autosomal markers is largely obscured by the overwhelming amount of the victim's DNA, such that only a subset of the tested autosomal markers can be detected in the resulting mixture (10, 49). In these cases Y-linked markers might be easily amplified, resulting in an *n*-locus haplotype for the

perpetuator's Y chromosome. How can we estimate the joint match probabilities for autosomal and Y chromosome markers with a random draw from the reference population?

Following the 1996 National Research Council recommendations (3), the product rule is typically used to compute the match probability for autosomal markers, multiplying the single locus genotype frequencies (suitably corrected for population structure and differences in allele frequencies among distinct subpopulations) of the *m* scored markers to obtain the *m*-locus genotype frequency. The product rule does not apply for obtaining the frequency of a particular Y haplotype from estimates of individual marker allele frequencies. Y-linked markers (i.e., from the non-recombining region) are typically in strong linkage disequilibrium. Thus, the frequency of the particular haplotype must be estimated from a reference database. Here we test for a Y-autosomal association and provide some theoretical recommendations for obtaining a joint match probability.

Because the Y chromosome and autosomes are unlinked, this might suggest using the product of the autosomal and Y match probabilities for the joint matching probability. There are some concerns with this assumption. First, disequilibrium between the mitochondrial DNA and the autosomes is known to exist in many hybrid zones in natural populations of various organisms (50, 51). Y-autosomal disequilibria should be empirically quantified in population samples rather than assumed. Sinha et al. (52) tested for associations between seven Y-STRs and thirteen autosomal STRs in an African-American and a European-American sample from Louisiana. We suggest that the appropriate test is between Y haplotypes and autosomal genotypes. Second, Y chromosome haplotypes can be highly informative as to which subpopulation an individual belongs, and this in turn potentially changes the autosomal allele frequencies used to compute the autosomal match probabilities.

2. Results

Empirical tests of association between Y chromosome and autosomal markers are presented and a theoretical framework for determining a joint match probability is recommended. Statistical analyses of association are performed in sixteen U.S. populations between the autosomal genotypes from loci CSF1PO, FGA, THO1, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S512, D21S11 and Y chromosome haplotypes determined from loci DYS19, DYS385ab, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS438, and DYS439. The sample populations include individuals of African-American, European-American, Hispanic-American, Asian-American, and Native American ancestry. The results are consistent with independence of Y and autosomal markers (**Figure 10**). A more complete account of our approach to computing the joint Y-autosomal matching probability is described in **APPENDIX A**. This approach includes weighting the major ethnic groups by the appropriate census sizes and a correction for population structure. A conservative estimate of the joint probability is obtained by multiplying the largest value of the group autosomal match probabilities by the estimated matching probability for the Y chromosome (see **APPENDIX A**).

3. Conclusions

Analysis of existing data show that the product rule can be safely applied in computing joint autosomal genotype/Y haplotype match probabilities. A conservative estimate of the joint probability is obtained by multiplying the largest value of the group autosomal match

probabilities by the estimated matching probability for the Y chromosome. We also suggest that an alternative approach to the current reporting of match probabilities by ethnic group is (given a defined reference population) to weight the ethnic match probabilities by the census values for ethnic groups, providing a much more natural probability.

V. Final Thoughts and Recommendations

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. Presently, 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; (46)). Nonetheless, there is a body of literature indicating substantial mixing among some U.S. ethnic groups (35, 38-42). 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 (3), 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 African-American 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 (1). 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 African-American subpopulations. Finally, more work is needed to construct appropriate Y-STR databases of Native American populations.

Summary of notable results that impact forensic casework:

1) An increase in the number of highly informative markers on the Y chromosome that are useful for forensic casework. Some of these markers are already included in a commercially available Y-typing kit (Applied Biosystems Yfiler), while many others may eventually be available in particular labs after testing on non-probative samples in forensic crime labs.

2) An online searchable database of these and many previously available markers already in use in forensic casework. This database includes 38 Y-STRs typed in over 2,500 samples representing the major U.S. population groups. An additional 5 Y-STRs have been typed in a subset of these samples, and 61 Y-SNPs have been typed in all samples. This database can be used by forensic laboratories to establish match statistics.

3) Demonstration of independence between Y haplotypes and autosomal markers for most surveyed populations and recommendations for a statistical approach on how to combine Y haplotype and autosomal marker frequencies to estimate joint match probabilities.

4) Demonstration that data from different regional areas in the U.S. can be pooled without leading to strong biases in the estimation of Y-STR haplotype frequencies. While this is true for four of the major U.S. ethnic groups, larger databases of Native Americans are needed for proper estimation of Native American Y chromosome haplotypes.

Publications:

Hammer MF, TM Karafet, AJ Redd, H. Jarjanazi H, S. Santachiara-Benerecetti, H. Soodyall, and S.L. Zegura. 2001. Hierarchical patterns of global human Y chromosome diversity. *Molecular Biology and Evolution* 18:1189-1203.

Redd AJ, AB Agellon, VA Kearney, VA Contreras, T Karafet, H Park, P de Knijff, JM Butler, and MF Hammer. 2002. Forensic value of fourteen novel STRs on the human Y chromosome. *Forensic Science International*. 130:97-111.

Butler JM, R Schoske, PM Vallone, MC Kline, AJ Redd, and MF Hammer. 2002. A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. *Forensic Science International* 129:10-24.

Redd AJ, J Roberts-Thomson, T Karafet, M Bamshad, LB Jorde, JM Naidu, B Walsh, and MF Hammer. 2002. Gene flow from the Indian subcontinent to Australia: evidence from the Y chromosome. *Current Biology* 12:673-677.

The Y chromosome Consortium: M Hammer , ME Hurles, MA Jobling, T Karafet, TE King, P de Knijff, A Pandya, A Redd, FR Santos, C Tyler-Smith, P Underhill, E Wood, M Thomas, L Cavalli-Sforza, N Ellis, T Jenkins, J Kidd, K Kidd, P Forster, S Zegura, and M Kaplan. 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Research* 12:339-348.

Bonne-Tamir B, M Korostishevsky, AJ Redd, Y Pel-Or, ME Kaplan, MF Hammer. 2003. Maternal and paternal lineages of the samaritan isolate: mutation rates and time to most recent common male ancestor. *Annals of Human Genetetics* 67:153-64.

Redd AJ, VA Contreras, VA Kearney, D Stover, T Karafet, and MF Hammer. 2006. Genetic structure among 38 populations from the United States based on 11 U.S. core Y chromosome STRs. *Journal of Forensic Science*: in press

Hammer MF, VA Contreras, VA Kearney, D Stover, G Zhang, T Karafet, and Redd AJ. 2006. Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. *Journal of Forensic Science*: in press.

In progress:

Walsh JB, AJ Redd, MF Hammer. Joint match probabilities for Y and autosomal markers. To be submitted to *Int. J. Leg. Med.*

Calderon K, AJ Redd, and MF Hammer. New, searchable, online database—*USAYSTR*—for estimating haplotype frequencies in the U.S. To be submitted to *Genomics*.

References

[1] M. Kayser, S. Brauer, H. Schadlich, M. Prinz, M. Batzer, P. Zimmerman, B. Boatin, and 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.

[2] 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, and 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.

[3] NRC. *The Evaluation of Forensic DNA Evidence*. 1996. Washington, D.C.: National Academy Press.

[4] P. Gill, A. J. Jeffreys, and D. J. Werrett, Forensic application of DNA 'fingerprints'. Nature 318 (1985) 577-9.

[5] M. Prinz, K. Boll, H. Baum, and B. Shaler, Multiplexing of Y chromosome specific STRs and performance for mixed samples. Forensic Sci Int 85 (1997) 209-18.

[6] K. Honda, L. Roewer, and 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) 868-72.

[7] A. Betz, G. Bassler, G. Dietl, X. Steil, G. Weyermann, and W. Pflug, DYS STR analysis with epithelial cells in a rape case. Forensic Sci Int 118 (2001) 126-30.

[8] M. Prinz and M. Sansone, Y chromosome-specific short tandem repeats in forensic casework. Croat Med J 42 (2001) 288-91.

[9] M. Prinz, A. Ishii, A. Coleman, H. J. Baum, and R. C. Shaler, Validation and casework application of a Y chromosome specific STR multiplex. Forensic Sci Int 120 (2001) 177-88.

[10] W. Parson, H. Niederstatter, S. Kochl, M. Steinlechner, and B. Berger, When autosomal short tandem repeats fail: optimized primer and reaction design for Y-chromosome short tandem repeat analysis in forensic casework. Croat Med J 42 (2001) 285-7.

[11] 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, and et al., Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110 (1997) 125-33.

[12] P. Gill, R. Sparkes, R. Pinchin, T. Clayton, J. Whitaker, and J. Buckleton, Interpreting simple STR mixtures using allele peak areas. Forensic Sci Int 91 (1998) 41-53.

[13] P. de Knijff, Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. Am J Hum Genet 67 (2000) 1055-61.

[14] M. A. Jobling and C. Tyler-Smith, Fathers and sons: the Y chromosome and human evolution. Trends Genet 11 (1995) 449-56.

[15] M. F. Hammer, T. M. Karafet, A. J. Redd, H. Jarjanazi, S. Santachiara-Benerecetti, H. Soodyall, and S. L. Zegura, Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol 18 (2001) 1189-203.

[16] L. Roewer, M. Kayser, P. Dieltjes, M. Nagy, E. Bakker, M. Krawczak, and P. de Knijff, Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations. Hum Mol Genet 5 (1996) 1029-33.

[17] A. J. Redd, S. L. Clifford, and M. Stoneking, Multiplex DNA typing of short-tandem-repeat loci on the Y chromosome. Biol Chem 378 (1997) 923-7.

[18] J. Butler, R. Schoske, P. Vallone, M. Kline, A. Redd, and M. Hammer, A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. Forensic Sci Int 129 (2002) 10-24.

[19] P. de Knijff, 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, L. Roewer, and et al., Chromosome Y microsatellites: population genetic and evolutionary aspects. Int J Legal Med 110 (1997) 134-49.

[20] 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, and M. Kayser, Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. Forensic Sci Int 118 (2001) 106-13.

[21] P. S. Walsh, N. J. Fildes, and R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. Nucleic Acids Res 24 (1996) 2807-12.

[22] J. M. Butler, Forensic DNA Typing. Academic Press, San Diego (2001).

[23] B. Walsh, Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. Genetics 158 (2001) 897-912.

[24] M. P. Stumpf and D. B. Goldstein, Genealogical and evolutionary inference with the human Y chromosome. Science 291 (2001) 1738-42.

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

[26] M. Kayser, R. Kittler, A. Erler, M. Hedman, A. C. Lee, A. Mohyuddin, S. Q. Mehdi, Z. Rosser, M. Stoneking, M. A. Jobling, A. Sajantila, and C. Tyler-Smith, A comprehensive survey of human Y-chromosomal microsatellites. Am J Hum Genet 74 (2004) 1183-97.

[27] Y. Ogata, T. Yamaba, K. Sawazaki, R. Iida, E. Tsubota, H. Takatsuka, T. Matsuki, T. Yasuda, and K. Kishi, Improvement of a multiplex PCR system for DYS441, DYS442, DYS443, DYS444 and DYS445, and a population study in 340 Japanese males. Leg Med (Tokyo) 7 (2005) 183-9.

[28] T. M. Clayton, J. P. Whitaker, R. Sparkes, and P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling. Forensic Sci Int 91 (1998) 55-70.

[29] M. Kayser, M. Krawczak, L. Excoffier, P. Dieltjes, D. Corach, V. Pascali, C. Gehrig, L. F. Bernini, J. Jespersen, E. Bakker, L. Roewer, and P. de Knijff, An extensive analysis of

Y-chromosomal microsatellite haplotypes in globally dispersed human populations. Am J Hum Genet 68 (2001) 990-1018.
[30] R. Lewontin and D. Hartl, Population genetics in forensic DNA typing. Science 20 (1991) 1745-1750.
[31] J. Wilder, S. Kingan, Z. Mobasher, M. Pilkington, and M. Hammer, Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nat Genet 36 (2004) 1122-1125.
[32] M. F. Hammer and S. L. Zegura, The role of the Y chromosome in human evolutionary studies. Evol Anthropol 5 (1996) 116-134.
[33] L. Zhivotovsky, S. Ahmed, W. Wang, and A. Bittles, The forensic DNA implications of genetic differentiation between endogamous communities. Forensic Sci Int 119 (2001) 269-272.
[34] S. L. Zegura, T. M. Karafet, L. A. Zhivotovsky, and 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) 164-75.
[35] D. Merriwether, S. Huston, S. Iyengar, R. Hamman, J. Norris, S. Shetterly, M. Kamboh, and R. Ferrell, Mitochondrial versus nuclear admixture estimates demonstrate a past history of directional mating. Am J Phys Anthropol. 102 (1997) 153-159.
[36] B. M. Chakraborty, M. E. Fernandez-Esquer, and R. Chakraborty, Is being Hispanic a risk factor for non-insulin dependent diabetes mellitus (NIDDM)? Ethn Dis 9 (1999) 278-83.
[37] A. J. Redd, V. F. Chamberlain, V. C. Kearney, D. Stover, T. M. Karafet, K. Calderon, and M. F. Hammer, Genetic structure among 38 populations from the United States based on 11 U.S. core Y-chromosome STRs. J Foren Sci in press (2006).
[38] E. Foster, M. Jobling, P. Taylor, P. Donnelly, P. de Knijff, R. Mieremet, T. Zerjal , and C. Tyler-Smith, Jefferson fathered slave's last child. Nature 396 (1998) 27-8.
[39] 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, and M. Shriver, Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. Am J Phys Anthropol 114 (2001) 18-29.
[40] E. Parra, A. Marcini, J. Akey, J. Martinson, M. Batzer, R. Cooper, T. Forrester, D. Allison, R. Deka, R. Ferrell, and M. Shriver, Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63 (1998) 1839-1851.
[41] 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, and 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) 139-53.
[42] C. Bonilla, M. Shriver, E. Parra, A. Jones, and 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.
[43] E. Lander and J. Ellis, Founding father. Nature 396 (1998) 13-14.
[44] B. Bonne-Tamir, M. Korostishevsky, A. J. Redd, Y. Pel-Or, M. E. Kaplan, and 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) 153-64.
[45] L. Jin and R. Chakraborty, Population structure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics. Heredity 74 (Pt 3) (1995) 274-85.

[46] U. S. C. Bureau.2000, http://www.census.gov/main/www/cen2000.html.
[47] C. L. Hanis, D. Hewett-Emmett, T. K. Bertin, and W. J. Schull, Origins of U.S. Hispanics. Implications for diabetes. Diabetes Care 14 (1991) 618-27.
[48] 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, and 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) 1277-86.
[49] N. Cerri, U. Ricci, I. Sani, A. Verzeletti, and F. De Ferrari, Mixed stains from sexual assault cases: autosomal or Y-chromosome short tandem repeats? Croat Med J 44 (2003) 289-92.
[50] Z. Redenbach and E. B. Taylor, Evidence for bimodal hybrid zones between two species of char (Pisces: Salvelinus) in northwestern North America. J Evol Biol 16 (2003) 1135-48.
[51] A. L. Roca, N. Georgiadis, and S. J. O'Brien, Cytonuclear genomic dissociation in African elephant species. Nat Genet 37 (2005) 96-100.
[52] S. K. Sinha, B. Budowle, R. Chakraborty, A. Paunovic, R. D. Guidry, C. Larsen, A. Lal, M. Shaffer, G. Pineda, E. Schneida, H. Nasir, and J. G. Shewale, Utility of the Y-STR typing systems Y-PLEX 6 and Y-PLEX 5 in forensic casework and 11 Y-STR haplotype database for three major population groups in the United States. J Forensic Sci 49 (2004) 691-700.

Table 1. Diversity rank and Rst in U.S. populations of Y-STRs*

DYS	#Alleles	Het	Rst
449	13	0.842	0.089
570	12	0.794	0.043
576	12	0.791	0.081
458	9	0.772	0.036
463	12	0.765	0.289
607	9	0.764	0.221
390	8	0.760	0.335
447	13	0.754	0.068
448	9	0.714	0.172
19	8	0.709	0.190
446	13	0.702	0.030
438	7	0.698	0.067
932	9	0.676	0.319
439	7	0.664	0.101
456	9	0.660	0.012
389II	7	0.659	0.125
452	10	0.635	0.056
442	8	0.617	0.103
H4	6	0.600	0.041
460	7	0.588	0.071
389I	5	0.587	0.030
437	6	0.582	0.132
391	5	0.528	0.029
426	4	0.520	0.232
393	5	0.503	0.098
450	5	0.467	0.287
388	9	0.416	0.026
453	6	0.238	0.059
455	6	0.198	0.024
454	5	0.173	0.024

*not including multicopy STRs; U.S. Core loci are bolded.

Table 2. Populations and haplotype diversity based on 11 U.S. core Y-STRs.

Ethnic Group/population	sample size	number of haplotypes	haplotype resolution	haplotype diversity (± SE)
African-Americans (AA)				
Arizona-Phoenix (AZ 1)	76	71	93.4	0.9982 ± 0.0025
Arizona-Mesa (AZ 2)	52	49	94.2	0.9977 ± 0.0043
Connecticut (CT)	89	89	100.0	1.0000 ± 0.0017
Florida (FL)	20	20	100.0	1.0000 ± 0.0158
North Carolina (NC)	84	83	98.8	0.9997 ± 0.0019
New York City (NYC)	42	41	97.6	0.9988 ± 0.0055
Ohio (OH)	103	99	96.1	0.9992 ± 0.0015
South Dakota (SD)	57	57	100.0	1.0000 ± 0.0033
Virginia (VA)	77	71	92.2	0.9976 ± 0.0027
Vermont (VT)	51	49	96.1	0.9984 ± 0.0043
Total	**651**	**564**	**86.6**	0.9994 ± 0.0002
European-Americans (EA)				
Arizona-Phoenix (AZ 1)	56	54	96.4	0.9987 ± 0.0037
Arizona-Mesa (AZ 2)	43	41	95.3	0.9978 ± 0.0056
Connecticut (CT)	85	76	89.4	0.9955 ± 0.0032
Florida (FL)	37	36	97.3	0.9985 ± 0.0067
North Carolina (NC)	87	81	93.1	0.9984 ± 0.0020
New York City (NYC)	42	42	100.0	1.0000 ± 0.0052
Ohio (OH)	99	87	87.9	0.9965 ± 0.0023
South Dakota (SD)	182	149	81.9	0.9968 ± 0.0012
Virginia (VA)	97	87	89.7	0.9970 ± 0.0022
Vermont (VT)	199	163	81.9	0.9958 ± 0.0015
Total	**927**	**664**	**71.6**	0.9972 ± 0.0004
Hispanic-Americans (HA)				
Arizona-Phoenix (AZ 1)	109	104	95.4	0.9992 ± 0.0014
Arizona-Mesa (AZ 2)	47	44	93.6	0.9972 ± 0.0051
Connecticut (CT)	90	80	88.9	0.9973 ± 0.0022
Florida (FL)	20	18	90.0	0.9895 ± 0.0193
New York City (NYC)	38	32	84.2	0.9986 ± 0.0065
Ohio (OH)	24	24	100.0	1.0000 ± 0.0120
South Dakota (SD)	42	38	90.5	0.9954 ± 0.0063

Table 2. Continued

Virginia (VA)	92	86	93.5	0.9981 ± 0.0021
Vermont (VT)	17	17	100.0	1.0000 ± 0.0202
Total	**479**	**386**	**80.6**	0.9981 ± 0.0004
Native Americans (NA)				
Apache	86	43	50.0	0.9436 ± 0.0141
Cheyenne	29	27	93.1	0.9951 ± 0.0106
Navajo	88	56	63.6	0.9804 ± 0.0059
Pima	19	17	89.5	0.9883 ± 0.0210
South Dakota	112	91	81.3	0.9924 ± 0.0035
South Dakota-Sioux	45	39	86.7	0.9909 ± 0.0080
Vermont	19	19	100.0	1.0000 ± 0.0171
	398	**259**	**65.1**	0.9938 ± 0.0010
Asian-Americans (SA)				
Arizona-Tucson (AZ)	25	24	96.0	0.9967 ± 0.0125
New York City (NYC)	37	37	100.0	1.0000 ± 0.0063
Total	**62**	**61**	**98.4**	0.9995 ± 0.0030

*DYS19,DYS385ab,DYS389I,DYS389II,DYS390,DYS391,DYS392,DYS393,DYS438,DYS439

Table 3. Y-STR AMOVA

Comparison	number of populations	number of groups**	among groups	among populations within groups	within populations
African-American (AA)	10	1	-	0.7	99.3
	10	4	0.2	0.5	99.3
European-American (EA)	10	1	-	0.2	99.8
	10	4	0	0.2	99.8
Hispanic-American (HA)	9	1	-	0.8	99.2
	9	4	0.4	0.4	99.2
Native-American (NA)	7	1	-	9.5	90.5*
	7	3	11.1*	1.8*	87.1*
Asian-American (SA)	2	na	-	0	100
All groups except NA	31	4	27.55*	0.42	72.03*
All groups	38	5	24.8*	1.5*	73.7*

* $P < 0.01$

** SW, MW, NE, & S

Table 4A. P values resulting from pairwise population differentiation tests on African-American samples.

AA	Phoenix	Mesa	CT	FL	NC	NYC	OH	SD	VA	VT
Phoenix	*									
Mesa	0.829	*								
CT	0.789	0.546	*							
FL	0.088	0.122	0.106	*						
NC	0.792	0.602	0.639	0.095	*					
NYC	0.554	0.382	0.875	0.173	0.373	*				
OH	0.213	0.112	0.048	0.011	0.125	0.056	*			
SD	0.542	0.705	0.333	0.065	0.211	0.271	0.103	*		
VA	0.481	0.273	0.140	0.019	0.269	0.119	0.868	0.218	*	
VT	0.360	0.544	0.562	0.343	0.500	0.644	0.028	0.145	0.067	*

Table 4B. P values resulting from pairwise population differentiation tests on European-American samples.

EA	Phoenix	Mesa	CT	FL	NC	NYC	OH	SD	VA	VT
Phoenix	*									
Mesa	0.968	*								
CT	0.541	0.805	*							
FL	0.797	0.813	0.260	*						
NC	0.436	0.636	0.943	0.474	*					
NYC	0.159	0.155	**0.004**	0.176	**0.004**	*				
OH	0.504	0.547	0.091	0.522	0.170	0.018	*			
SD	0.939	0.963	0.279	0.574	0.270	0.011	0.633	*		
VA	0.343	0.458	0.384	0.589	0.896	**0.007**	0.335	0.197	*	
VT	0.532	0.879	0.208	0.773	0.340	0.017	0.736	0.817	0.337	*

Bolded P values are ≤0.01.

Table 4C. P values resulting from pairwise population differentiation tests on Hispanic-American samples.

HA	Phoenix	Mesa	CT	FL	NYC	OH	SD	VA	VT
Phoenix	*								
Mesa	0.116	*							
CT	0.036	**0.006**	*						
FL	0.738	0.817	0.165	*					
NYC	0.479	0.142	0.219	0.479	*				
OH	0.804	0.084	0.282	0.512	0.630	*			
SD	0.657	0.070	0.256	0.399	0.370	0.467	*		
VA	0.138	**0.007**	0.686	0.215	0.271	0.693	0.333	*	
VT	0.234	0.030	0.734	0.313	0.417	0.645	0.263	0.743	*

Table 4D. P values resulting from pairwise population differentiation tests on Native American samples.

NA	APA	CHY	NAV	PIM	SD	SIO	VT
APA	*						
CHY	0.547	*					
NAV	0.047	0.104	*				
PIM	**0.001**	**0.001**	**0.006**	*			
SD	**0.000**	0.022	**0.000**	**0.000**	*		
SIO	**0.001**	0.021	**0.000**	**0.000**	0.714	*	
VT	**0.000**	**0.000**	**0.000**	**0.000**	0.018	0.038	*

Bolded P values are ≤0.01.

Table 5. Diverity statistics for Y chromosome haplogroups (Hg).

Ethnic Group/population	Sample size	Number of Hgs	Discrimination capacity (%)	Haplogroup diversity (+/- SE)
African-Americans (AA)	**651**	**24**	**3.7**	**0.585 ± 0.020**
Arizona-Phoenix	76	8	10.5	0.564 ± 0.058
Arizona-Mesa	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	56	12	21.4	0.688 ± 0.062
Arizona-Mesa	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	109	15	13.8	0.792 ± 0.035
Arizona-Mesa	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

Table 5. Continued

Ethnic Group/population	Sample size	Number of Hgs	Discrimination capacity (%)	Haplogroup diversity (+/- SE)
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

Table 6. Y-SNP AMOVA

				% variance		
Populations	number of chromosomes	number of populations	number of groups	among groups	among populations within groups	within populations
African-Americans	651	10	1	-	1.4	98.6
		10	4a	0.4	1.0	98.6
European-Americans	927	10	1	-	0.2	99.8
		10	4a	-0.1	0.3	99.9
Hispanic-Americans	479	9	1	-	1.0	99.0
		9	4a	1.0	0.1	98.9
Native-Americans	398	7	1	-	**16.0**	**84.0**
		7	3b	17.9	**2.9**	**79.2**
Asian-Americans	62	2	1	-	0.0	100.0
five ethnic groups	2718	38	5	**32.3**	**2.0**	**65.8**

bolded numbers, $P < 0.001$

a Southwest, Midwest, Northeast, and South

b Southwest, Midwest, East

Figure Legends

Figure 1. Locations of 43 Y-STRs typed in the Hammer lab. Approximate positions of Y-STR on megabase (Mb) scale shown above schematic of Y chromosome.

Figure 2. Size and color schemes for 43 Y-STRs in 5 multiplex reactions.

Figure 3. An example of a view of the online USAYSTR database.

Figure 4. Map showing the approximate geographic positions of populations sampled in this study. The populations are grouped by ethnicity (African America, AA; European American, EA; Hispanic American, HA; Native American, NA; and Asian American, SA) and by geography (dotted circles surround Southwest, Midwest, Northeast and Southern samples). The Cheyenne sample is not shown.

Figure 5. Comparison of haplotype resolution in 5 ethnic groups when using 11 U.S. core loci, 2 commercially available kits, and 38 Y-STRs.

Figure 6. MDS plot of 38 populations based on R_{ST} genetic distances. Population codes and symbols are the same as in **Table 2**.

Figure 7. 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 = 2,517), African Americans (AA, n = 651), European Americans (European-American, n = 927), Hispanic Americans (HA, n = 479), Native Americans (NA, n = 398), and Asian Americans (SA, n = 62).

Figure 8. MDS plot of 38 populations based on Φ_{ST} genetic distances. Population codes are the same as in Table 2. 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.

Figure 9. Bar chart showing the relative proportions of Y chromosomes with African (green bar), European (blue bar), Native American (pink bar), and Asian ancestry (orange bar). (A) African Americans, (B) European Americans, (C) Hispanic Americans, and (D) Native Americans. Population codes are the same as in Table 2. A small sample of Hispanics from North Carolina (n = 15) that was not included in other analyses is shown here.

Figure 10. Exact test of non-random association between Y haplotype and CODIS loci. A minus sign means not statistically significant.

Figure 1

Locations of 43 Y-STRs Typed in Hammer Lab

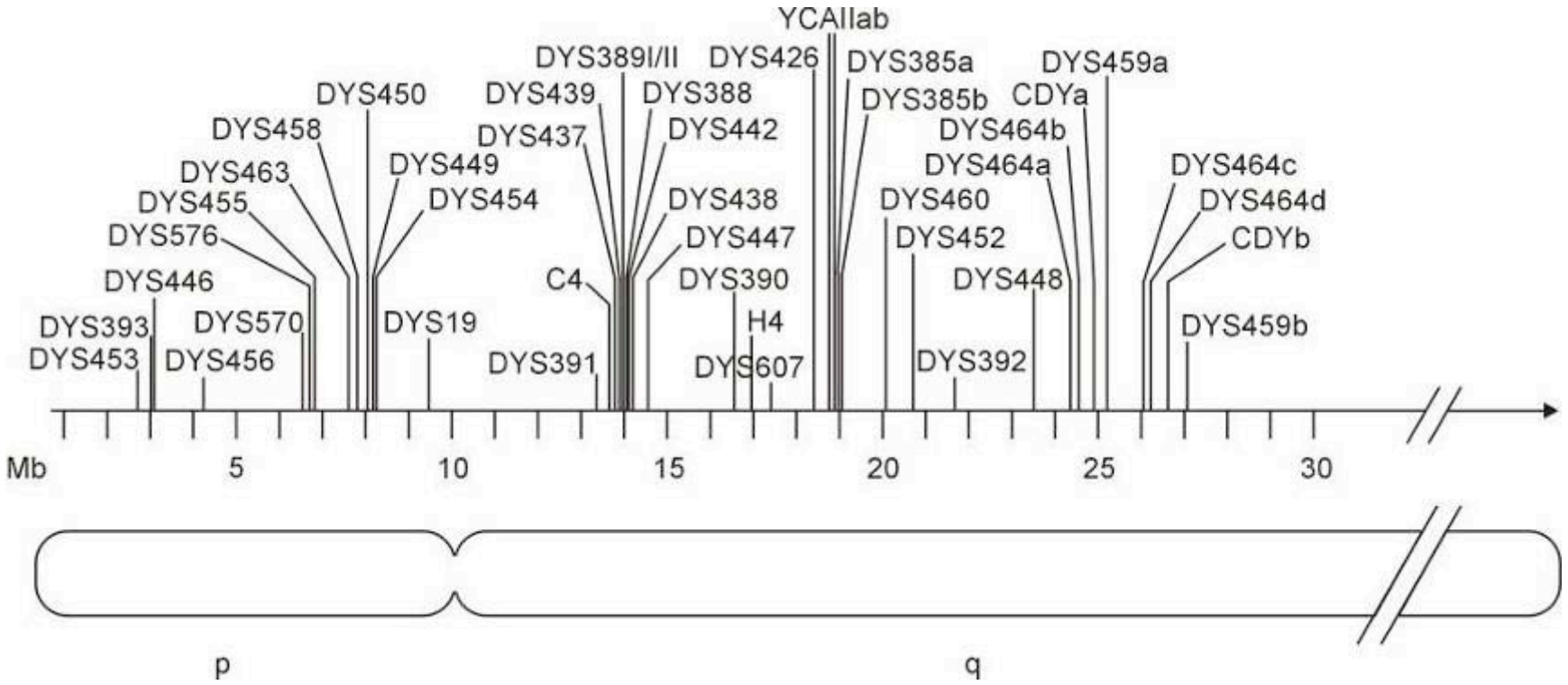


Figure 2

Y-STR Muliplex Reactions

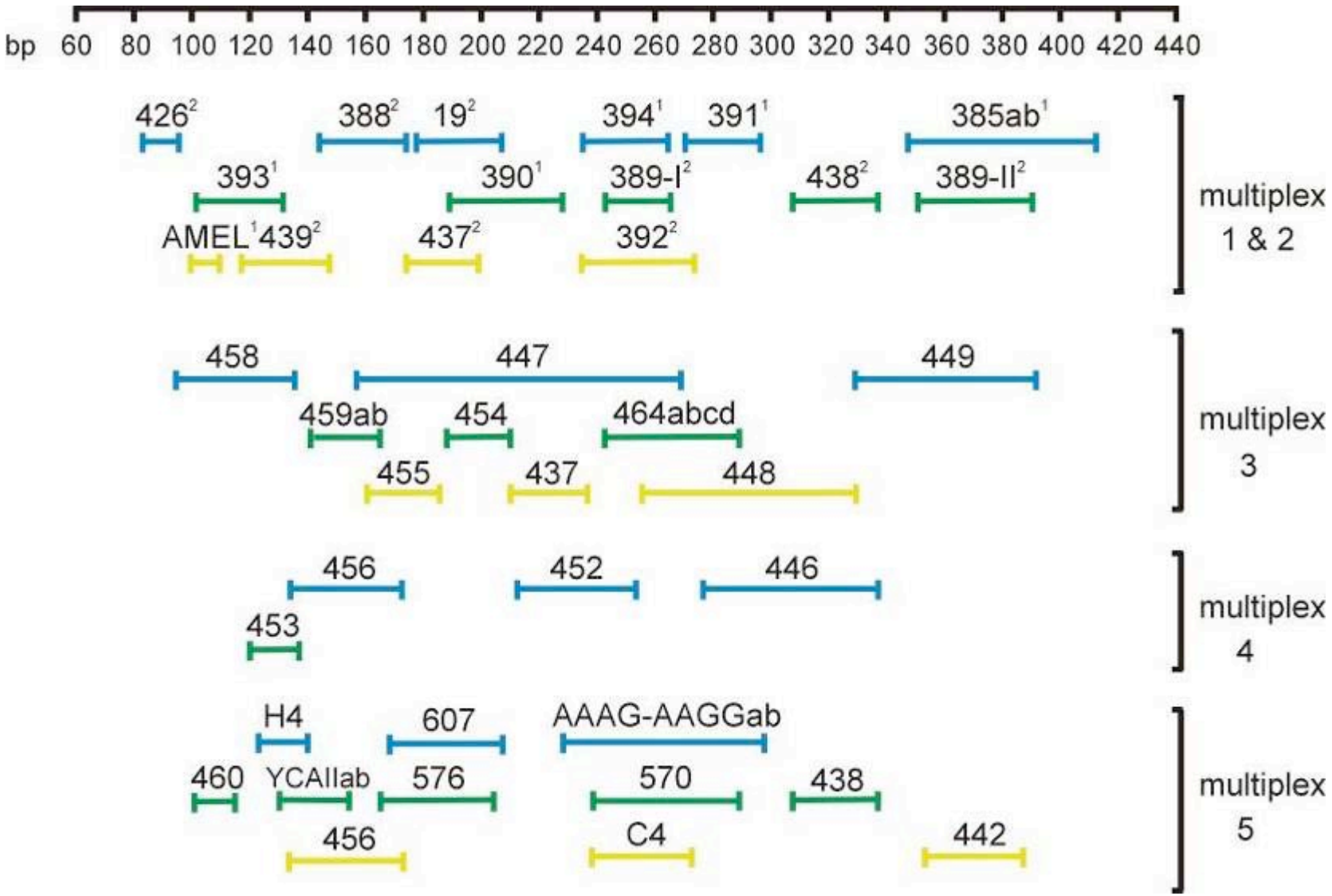


Figure 3

USAYSTR database

Our purpose is to provide a web-based, open access, searchable Y-STR database for estimating haplotype frequencies in the USA.

Search for a Y-haplotype

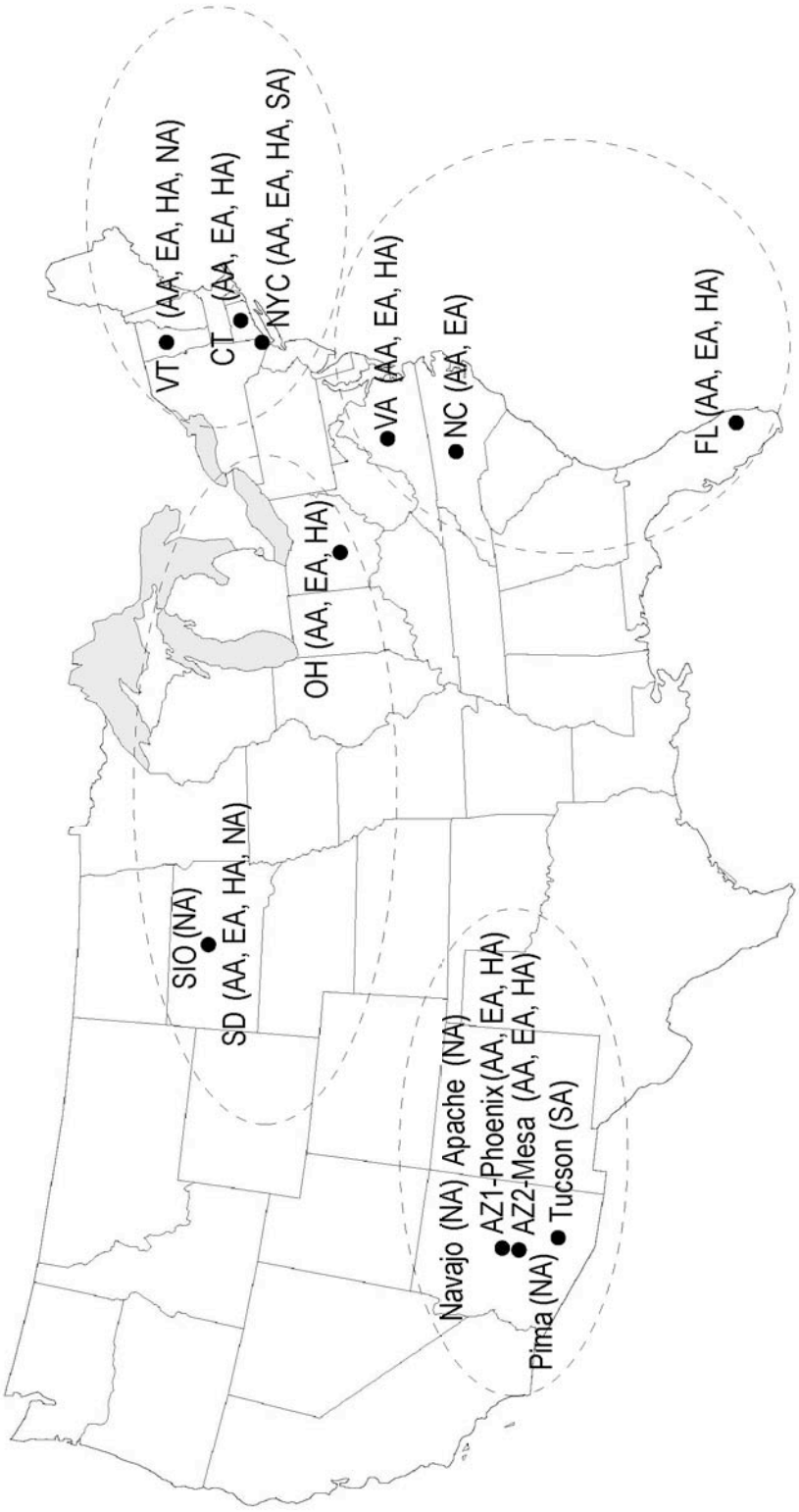


Figure 4

Figure 5

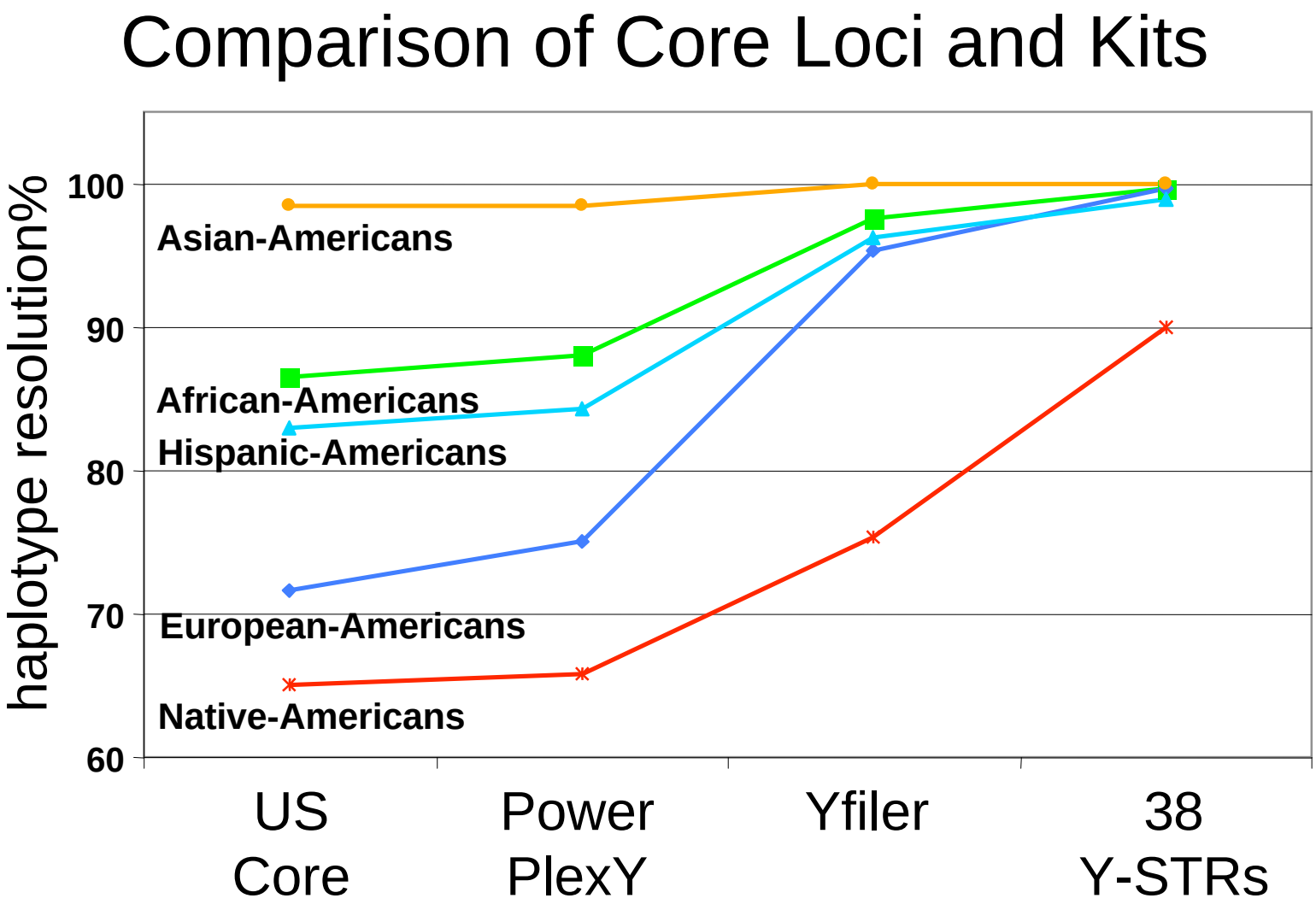


Figure 6

Y-STR MDS population plot

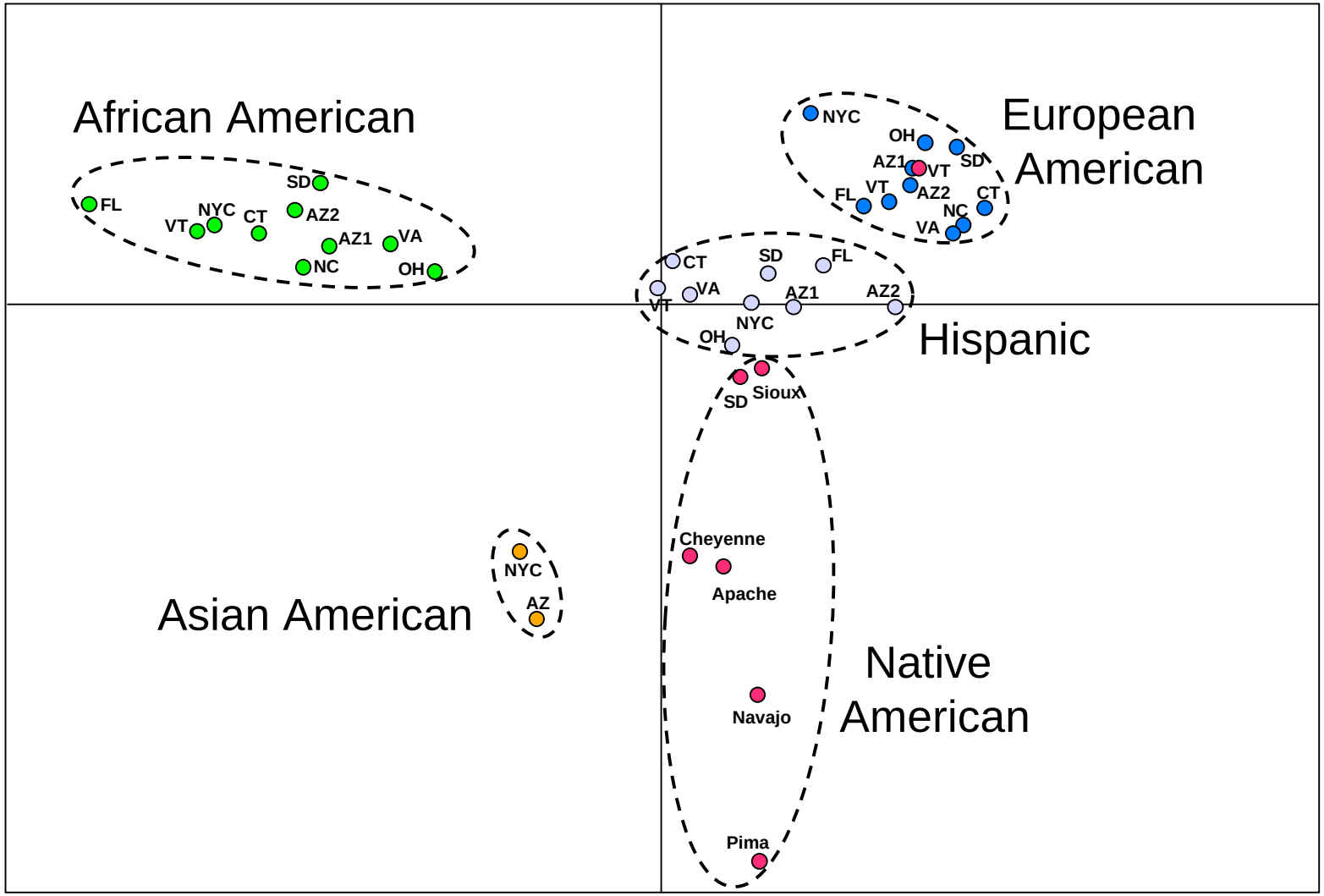
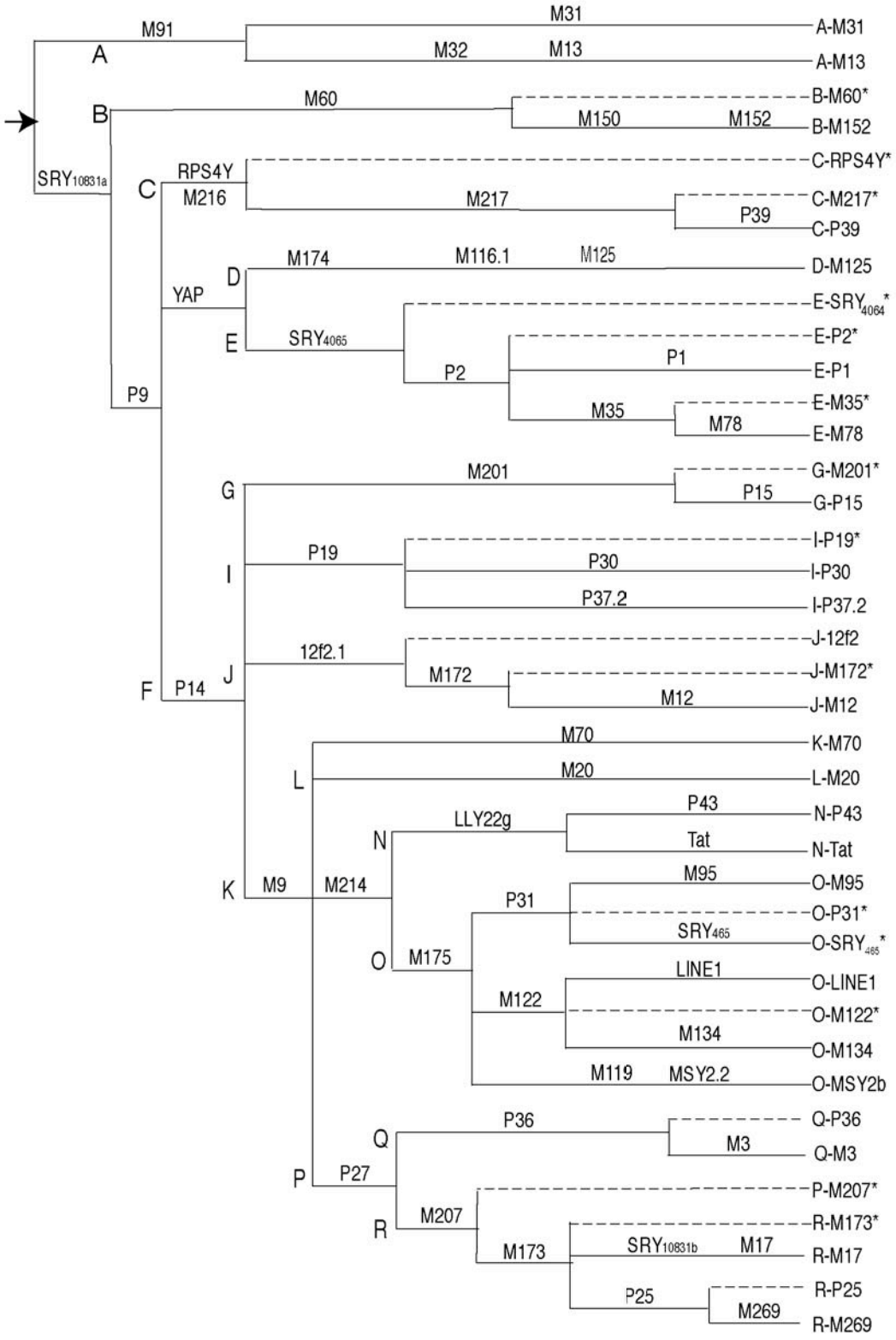


Figure 7



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

Figure 8

Y-SNP MDS population plot

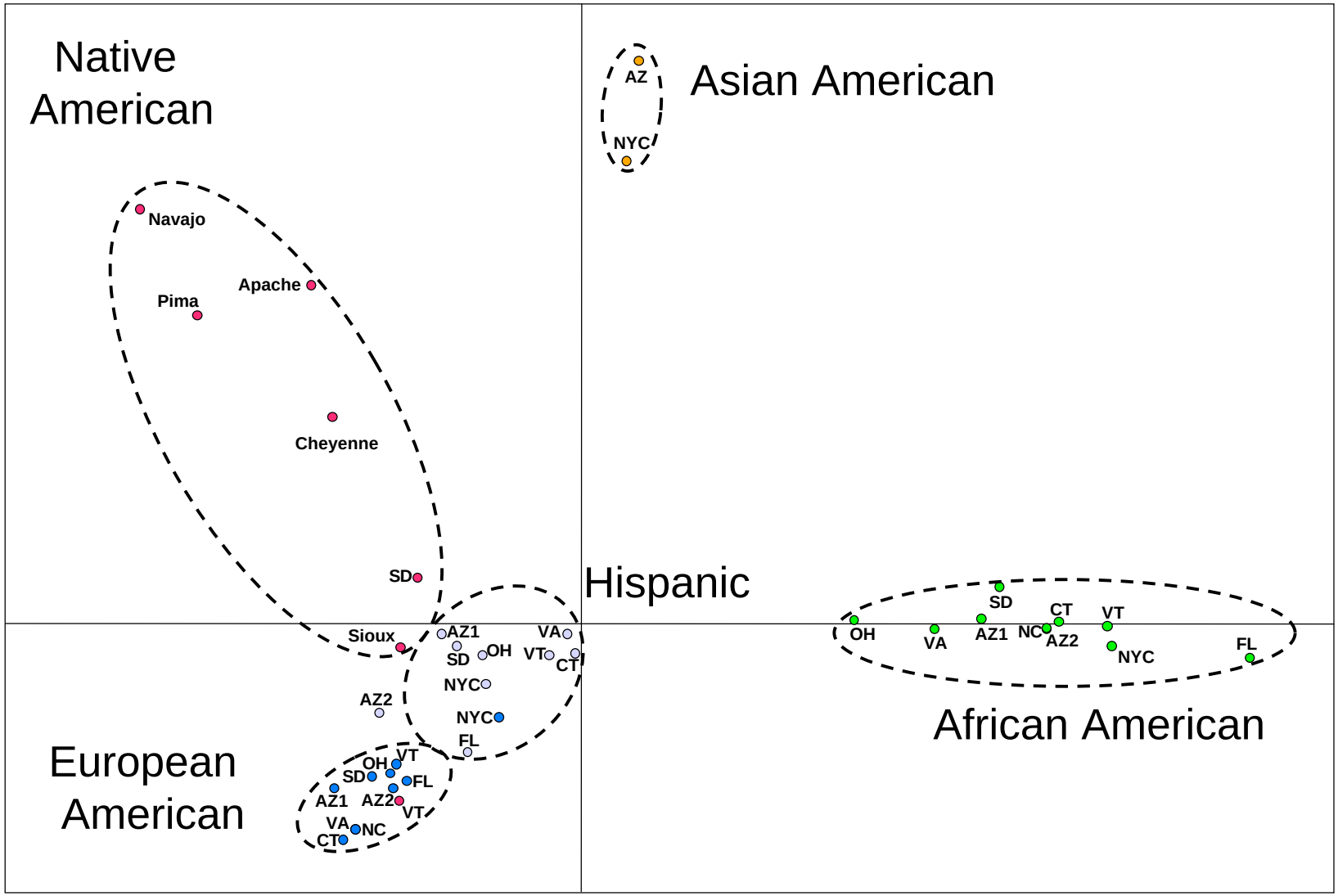


Figure 9

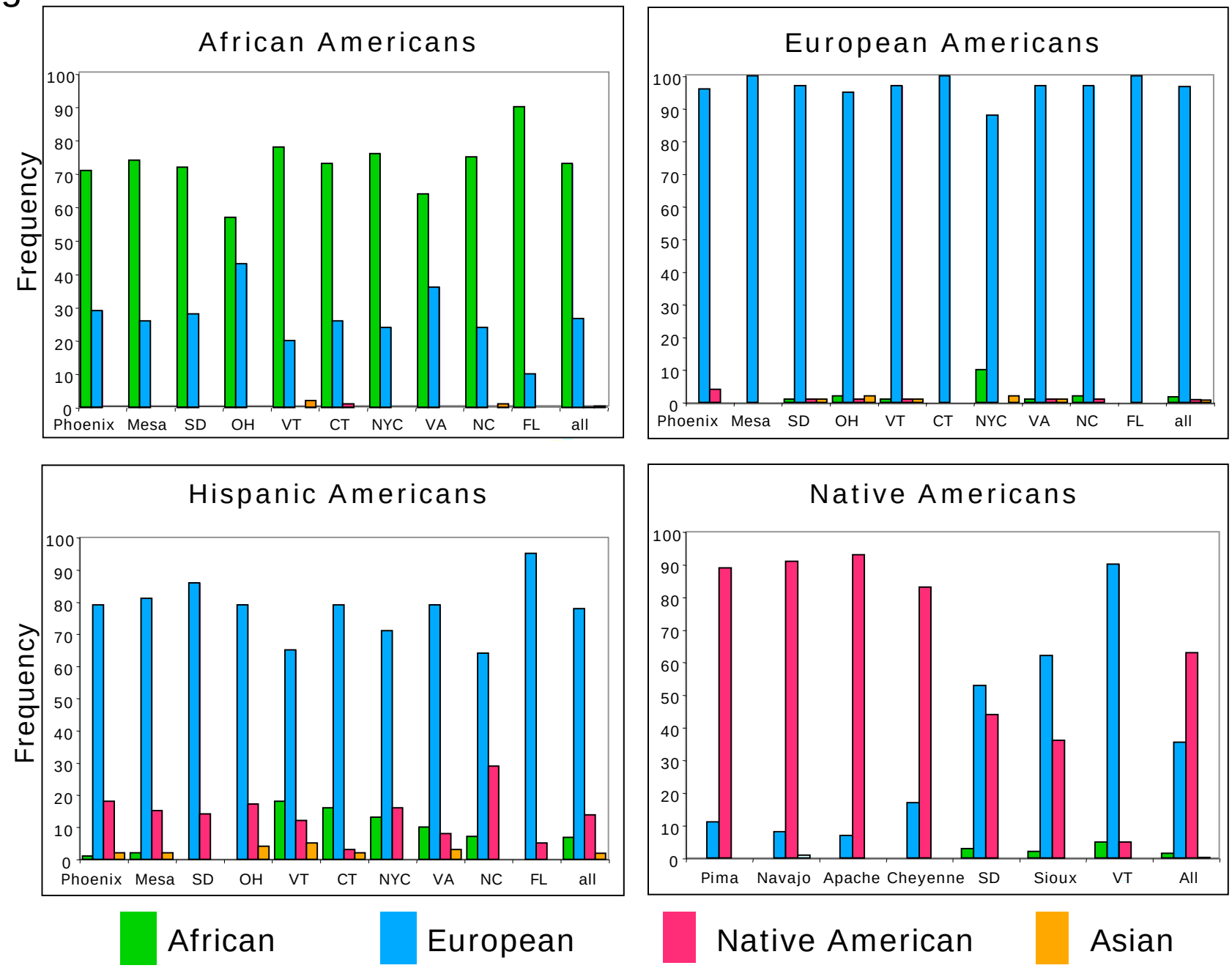


Figure 10

Exact test* of non-random association between Y haplotype and CODIS loci

	N		CSF1PO	FGA	THO1	TPOX	vWA	D3S1358	D5S818	D7S820	D8S1179	D13S317	D16S539	D18S512	D21S11
			1	2	3	4	5	6	7	8	9	10	11	12	13
NYC NC	137	Yhap	-	-	-	-	-	-	-	-	-	-	-	-	-
	43	Yhap	-	-	-	-	-	-	-	-	-	-	-	-	-
	94	Yhap	-	-	-	-	-	-	-	-	-	-	-	-	-
NYC NC	137	Yhap	-	-	-	-	-	-	-	-	-	-	-	-	-
	41	Yhap	-	-	-	-	-	-	-	-	-	-	-	-	-
	96	Yhap	-	-	-	-	-	-	-	-	-	-	-	-	-
NYC NC	53	Yhap	-	-	-	-	-	-	-	-	-	-	-	-	-
	38	Yhap	-	-	-	-	-	-	-	-	-	-	-	-	-
	15	Yhap	-	-	-	-	-	-	-	-	-	-	-	-	-
Apache Navajo	168	Yhap	-	-	-	-	-	+	-	-	-	-	-	-	-
	85	Yhap	-	-	-	-	-	-	-	-	-	-	-	-	-
	83	Yhap	-	-	-	-	-	-	-	-	-	-	-	-	-

* P < 0.01

APPENDIX A

As shown in **Figure 10**, the results are consistent with independence of Y and autosomal markers within each subpopulation. Thus our concern of lack of independence within a subpopulation does not appear to be an issue, at least for the populations examined here. Thus, within a particular subpopulation the joint match probability is simply obtained by multiplying the subpopulation-specific autosomal and Y match probabilities. However, this leads to a second potential concern, namely that a Y chromosome haplotype may be highly informative as to which subpopulation

an individual belongs, and this in turn potentially changes the autosomal allele frequencies used to compute the autosomal match probabilities. To formally see this, express the joint match probability by conditioning on the observed Y haplotype y,

$$\Pr(y, \text{ autosomal genotype}) = \Pr(\text{autosomal genotype}|y) \cdot \Pr(y) \tag{1}$$

As we detail below, there are several approaches for obtaining the autosomal match probabilities conditional on the Y haplotype. We start with two simple approaches before moving to a more complex approach that allows for population substructure conditioned on the Y haplotype.

For purposes of discussion, suppose we have the following match probabilities for the Y and autosomes using appropriate databases for various ethic populations:

	Caucasian	African	Hispanic	Native America
Y match	3/1,500	1/1,500	5/800	0/300
Y match freq	0.0020	0.0007	0.00625	0
Autosomal match	1/40,000	1/20,000	1/30,000	1/50,000

Further, suppose that the appropriate male reference population from which we sample a random individual consists of 45% Caucasian, 30% Africans, 20% Hispanics, and 5% Native Americans. While one might use as our estimate of the Y chromosome match probability the overall frequency of the haplotype in the database ($9/4100 = 0.0022$), it is more appropriate to weight by the actual fraction of ethnicity on the reference population, as opposed to the sample. In this case, the match probability for the Y haplotype becomes

$$0.45 \cdot 0.002 + 0.30 \cdot 0.0007 + 0.20 \cdot 0.00625 + 0.05 \cdot 0.00 = 0.00235$$

Clearly this Y haplotype is more frequent in Hispanic (0.625%, as opposed to 0.20% in Caucasian and 0.070% in individuals of African extraction), and this suggests that the appropriate autosomal conditional match probability to use is that for Hispanics, giving a match probability of

$$0.00235 \cdot (1/30,000) = 7.83 \times 10^{-8}$$

However, this same Y haplotype is also (albeit more rarely) found in individuals of Caucasian and Africa extraction. A more careful approach would be to use a weighted match probability, conditioning on autosomal matches over the probability of the y haplotype indicating membership in different ethnic groups,

$$\begin{aligned} &\Pr(\text{autosomal genotype}|\text{Y haplotype}) \\ &= \sum_i \Pr(\text{autosomal genotype}|\text{Group } i) * \Pr(\text{Group } i \,|\text{Y haplotype}) \end{aligned} \tag{2}$$

From Bayes' theorem, we can estimate the group probabilities as follows:

$$\Pr(\text{Group } i \,|\text{Y haplotype}) = \frac{\Pr(\text{Y haplotype} \mid \text{Group } i)\Pr(\text{Group } i)}{\sum_i \Pr(\text{ Y haplotype} \mid \text{Group } i)\Pr(\text{Group } i)} \tag{3}$$

For example, for our above hypothetical match probabilities,

$$\Pr(\text{Caucasian} \,|\text{Y haplotype}) = \frac{0.002 \cdot 0.45}{0.002 \cdot 0.45 + 0.0007 \cdot 0.3 + 0.00625 \cdot 0.20} = 0.383,$$

while the resulting weights for Africans and Hispanics are 0.085 and 0.532, respectively. The resulting match probability using Equation (2) becomes

$$0.00235 \cdot [\, 0.383(1/40000) + 0.085(1/20000) + 0.532(1/30000) \,] = 7.42 \times 10^{-8}$$

Finally, the most conservative approach is to chose the largest autosomal match probabilities over ethnic groups within which the haplotype occurs. In our example, this occurs in individuals of African extraction, giving a conservative match probability of

$$0.00235 \cdot (1/20000) = 1.18 \times 10^{-7}$$

Accounting for Autosomal subpopulations

It remains to specify how to compute the probability of two matching autosomal genotypes given that two individuals share the identical Y haplotype. As mentioned above, the Y haplotype may be rather informative as to ethnic group, which in turn suggests the appropriate group from which to compute allele frequencies. However, an additional correction may be in order beyond simply computing these probabilities for the different ethnic groups. Two individuals with identical Y haplotypes are likely more related to each other than are two random individuals. In a population-genetics framework, we expect the time back to the common (Y-contributing) ancestor of two individuals with identical Y haplotypes to be more recent than the time back to a common ancestor for two randomly-chosen individuals. In essence, the common Y haplotype imparts a subgroup structure and we need to account for this. Balding and Nichols (1994) show that the correct single-locus expressions for match probabilities when individuals come from the same subpopulation is a function of the average coefficient of coancestry θ, with

$$\Pr(A_iA_i \mid A_iA_i) = \frac{[2\theta + (1-\theta)p_i]\,[3\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)} \tag{4a}$$

$$\Pr(A_iA_j \mid A_iA_j) = \frac{2[\theta + (1-\theta)p_i]\,[\theta + (1-\theta)p_j]}{(1+\theta)(1+2\theta)} \tag{4b}$$

How do we compute the expected value of θ given two individuals are identical at all n tested Y chromosome markers? We do this by conditioning on the distribution of time back to a common (Y) ancestor given the number of identical markers scored on the Y. This uses the results of Walsh (2001), who developed Bayesian estimators for the distribution of the time back to this most recent common ancestor (MRCA) given the marker differences between two Y haplotypes.

If the mutation rate per marker is μ, then the chance that two haplotypes are identical at n markers given they last shared a common (Y) ancestor t generations ago is just

$$\Pr(n \text{ marker match} \mid t \text{ generations to MRCA}) = (1-u)^{2nt} \tag{5}$$

This is the likelihood function for t given the mutation rate and number of markers. A more detailed analysis under the stepwise mutation model in Walsh (2001) shows that, when all n markers match, that this simple expression based on the infinite alleles model is essentially equivalent to the more exact analysis under a stepwise model that allows for back mutations.

Population genetics theory provides the prior distribution for the time to MRCA for two randomly-drawn Y chromosomes in the absence of any marker information (Walsh 2001). The time back to MRCA follows a geometric distribution with parameter $\lambda = 1/N_e$, the reciprocal of the effective population size (since the Y chromosome is haploid). Hence, the prior becomes

$$\Pr(\tau \text{ generations to MRCA}) = \lambda(1-\lambda)^{\tau-1} \tag{6}$$

The resulting posterior distribution (from Equations 5 and 6) becomes

$$\Pr(\tau \text{ generations to MRCA} \mid \text{ n-marker match}) = C(1-u)^{2n\tau} \cdot \lambda(1-\lambda)^{\tau-1} \tag{7}$$

where the constant C assures that the probabilities sum to one.

In particular, the posterior distribution for the time to MRCA given two individuals exactly match at n markers is

$$p(t \mid t \geq k) = \frac{(1-u)^{2nt} \cdot (1-\lambda)^{t-1}}{\sum_{\tau=k}^{\infty} (1-u)^{2n\tau} \cdot (1-\lambda)^{\tau-1}} \tag{8}$$

We have invoked the conditioning of the MRCA being at least a certain number k of generations because good practice with Y chromosome data would involve testing both the father/son and any paternal sibs (full or half) of a suspect, and hence these individuals are removed from consideration.

Since the coefficient of coancestry for two individuals that shared a common ancestor τ generations ago is just $(1/2)^{2\tau+1}$, the expected coefficient of coancestry θ thus becomes

$$\begin{aligned} E[\theta \,|\, t \geq k] &= \sum_{t=k}^{\infty} \left(\frac{1}{2^{2t+1}} \right) p(t \,|\, t \geq k) \\ &= \frac{\sum_{t=k}^{\infty} (1/2)^{2t+1} (1-u)^{2nt} \cdot (1-\lambda)^{t-1}}{\sum_{t=k}^{\infty} (1-u)^{2nt} \cdot (1-\lambda)^{t-1}} \\ &= \frac{2^{1-2k} [\, 1 - (1-\lambda)(1-\mu)^{2n} \,]}{4 - (1-\lambda)(1-\mu)^{2n}} \end{aligned} \tag{9}$$

For $k = 2$ (the typical case in Forensic settings), this simplifies to

$$E[\theta \,|\, t \geq 2] = \frac{1 - (1-\lambda)(1-\mu)^{2n}}{32 - 8(1-\lambda)(1-\mu)^{2n}} \tag{10a}$$

$$\simeq \frac{\lambda + 2n\mu}{24 - 8(\lambda + 2n\mu)} \tag{10b}$$

For a test of $n = 11$ STR markers, Equation (10a) gives the following values for $E[\theta]$ for different mutation rates μ,

N_e	$\mu = 0.001$	$\mu = 0.002$	$\mu = 0.004$
∞	0.00090	0.00177	0.003426
5,000	0.00091	0.00178	0.00343
500	0.00098	0.00185	0.00349
100	0.00130	0.00216	0.00378
50	0.00170	0.00254	0.00414

Given that we except the effective population sizes from most ethnic groups to exceed 100, and most estimates of STR mutation rates are around 0.002, taking $E(\theta) = 0.002$ is expected to be conservative. With allele frequencies of 0.02 or higher, this value for θ in Equations (4a) and (4b) essentially recovers Hardy-Weinberg proportions (p_i^2 for homozygotes, $2p_ip_j$ for heterozygotes).

Recommendation

A conservative approach to computing the joint Y-autosomal matching probability is thus as follows. For each of the major ethnic groups, compute the autosomal match probability using the product of single-locus genotype frequencies obtained from using Equations (4a) and (4b) with $\theta = 0.002$

and the allele frequencies appropriate for each group. A consevative estimate of the joint probability is obtained by multipling the largest value of these group autosomal match probabilities by the estimated matching probability for the Y.

Finally, as to the issue of estimating the y match probability from a database, we favor a conservative modification of the simple counting method,

$$\widehat{\Pr}(y) = \frac{i+2}{n+2}$$

where there are i matches in an initial data base of n haplotypes. Making the converative (favoring a suspect) assumption that a crime sample came from a different individual than the suspect, we have two more samples (hence $n+2$ total samples) and matches for the suspect, the crime sample and i matches from the databases (hence $i+2$ matches).