181956

97-LB-VX-0010 Final Report

Final Progress Report of Award No. 97-LB-VX-0010
"Development of the Human Y Chromosome as a Forensic Tool"
Principal Investigators: Michael F. Hammer and Susan D. Narveson

I. Summary

This grant was funded to support a collaborative effort between the laboratories of Dr. Michael Hammer at the University of Arizona (Laboratory of Molecular Systematics and Evolution-LMSE) and Dr. Susan Narveson at the Phoenix Police Department Laboratory Services Bureau (PPDLSB) to develop a set of male-specific markers for use in forensic typing laboratories. The main goals were to 1) identify of a set of polymorphic markers mapping to the non-recombining portion of the Y chromosome (NRY) that are robust in forensic analysis, 2) develop detailed protocols for high throughput, fluorescence-based typing of these markers, and 3) establish a NRY data base for US population groups. Our first priority was to identify a set of microsatellites (Y-STRs) that behave optimally with high quality samples and that are easily typed using both the ABI 373/377 and 310 systems. Towards this goal, we identified several tri-, tetra- and penta-nucleotide repeats that exhibited robust amplification without artifactual banding and that did not produce high frequency alleles in all populations. Towards the goal of establishing a NRY polymorphism database we genotyped 5 Y-STRs in a panel of 1141 individuals representing five US populations (Southwest Hispanic, Caucasian, African American, Native American, East Asian) and 15 populations from Asia, Europe, and the Americas. Six additional Y-STRs were genotyped in a subset (n=397) of the aforementioned US population groups. All 1141 individuals were also genotyped at 31 biallelic polymorphisms on the NRY (Y-SNPs). These data were used to 1) compare the relative utility of Y-STRs, Y-SNPs, and combination haplotypes (e.g., constructed from a combination of Y-STR and Y-SNP information) for forensic work, and 2) make the first estimates of levels of population substructure within each of the major US groups. Y-SNPs were useful for identifying population-specific Y-chromosome haplotypes, while Y-STRs and combination haplotypes provided a high degree of individualization among male lineages within populations. These results also demonstrated the importance of considering the potential impact of both population structure and admixture among US groups on the statistical analysis of Y-chromosome forensic data. Thus, the forensic implications of NRY variation in US population groups may be quite different from those resulting from the analysis of autosomal data. Most of the goals of the original proposal were achieved; some data are still being analyzed. In the following, we submit a report to the NIJ on the work that has been accomplished.

II. Markers and Samples Analyzed

A. Y-STRs

Our first priority was to identify a set of Y-STRs that behave optimally with high quality samples and that are easily typed using the ABI 373, 377 and 310 platforms. We made use of published Y-STRs and began to screen for novel Y-STRs that exhibit Y-specificity, high heterozygosity, and clear amplification products without artifactual banding. A list of all published Y-STRs, as well as a few Y-STRs that were discovered in our lab (and that have not yet been published) is shown in **Table 1.** In the course of this study we focused only on those STRs with tri-, tetra-, and penta-nucleotide repeat structures (bolded STR loci in **Table 1**).

Table 1. STRs mapping to the non-recombining portion of the Y-chromosome (NRY).

locus	Sequence of variable repeats	Number of variable repeats	Number of alleles	Size range of alleles (bp)	multiplex PCR reaction
dimeric					
DYS288	$(CA)_n$	-[b]	2	119-121	-
YCAII	$(CA)_n$	-[b]	28	147-165	-
YCAIII	$(CA)_n$	19-25	15	192-204	-
2D6	$(CA)_n$	-[b]	5	205-213	-
trimeric					
DYS388	$(ATA)_n$	-[b]	5	126-138	pentaplex 2
DYS392	$(ATT)_n$	7-16	8	236-263	pentaplex 2
DYS425	$(AAC)_n$	-[b]	3	104-110	pentaplex 2
DYS426	$(AAC)_n$	-[b]	2	94-97	-
DYF371 [a]	$(AAC)_n$	-[b]	9[c]	195-213	-
tetrameric					
DYS19	$(CTAT)_n$	10-19	10	174-210	pentaplex 1
DYS385 [a]	$(GAAA)_n$	9-22	49[c]	360-412	pentaplex 1
DYS389I	$(CTAT)_n/(CTGT)_n$	7-13	7	239-263	pentaplex 2
DYS389II	$(CTAT)_n/(CTGT)_n$	23-31	9	353-385	pentaplex 2
DYS390	$(CTAT)_n$	18-27	8	191-227	pentaplex 1
DYS391	$(CTAT)_n$	8-13	6	275-295	pentaplex 1
DYS393	$(GATA)_n$	9-15	6	108-132	pentaplex 1
A7.1	$(GATA)_n$	11-15	5	161-181	-
A7.2	$(GATA)_n$	10-14	5	174-190	-
A10	$(GATA)_n$	12-15	4	160-172	-
C4	$(GATA)_n$	10-15	6	251-271	-
H4	$(GATA)_n$	10-13	4	362-374	-
GGAA10 [a]	$(GGAA)_n$	-[b]	39[c]	244-272	-
TAGA13	$(TAGA)_n$	14-18	5	140-160	pentaplex 2
G09411	NR	8-11	4	NR	-
G10123	NR	10-13	4	NR	-
pentameric					
DXYS156Y	$(TAAAA)_n$	8-15	8	145-180	-

[a] multiple copies on NRY, [b] alleles not yet fully sequenced; [c] number of banding patterns; NR, not reported.

B. Y-SNPs

We also focused our attention on selected biallelic polymorphisms that exhibited population specificity. These biallelic polymorphisms which included single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels), most likely represent unique mutational events in human population history (we refer to both kinds of biallelic polymorphisms as SNPs). The 30 Y-SNPs studied here gave rise to 31 NRY haplotypes (**Figure 1**). These 31 haplotypes were further divided into 8 "Haplogroups" (A-H) based on the topology of the phylogenetic tree and the geographic distribution of the haplotypes.

Our reasons for typing a set of population-specific SNPs were two-fold. First, by constructing STR/SNP "combination" haplotypes, it was possible in many cases to distinguish among Y chromosomes that have common STR alleles. Point mutations were useful for determining whether or not STR alleles of the same length in different populations are identical by descent or by coincidence. Second, by examining SNPs with low mutation rates and STRs with high mutation rates on the same chromosomes, it was possible to identify both population-specific markers and markers that will eventually lead to the individualization of nearly all Y

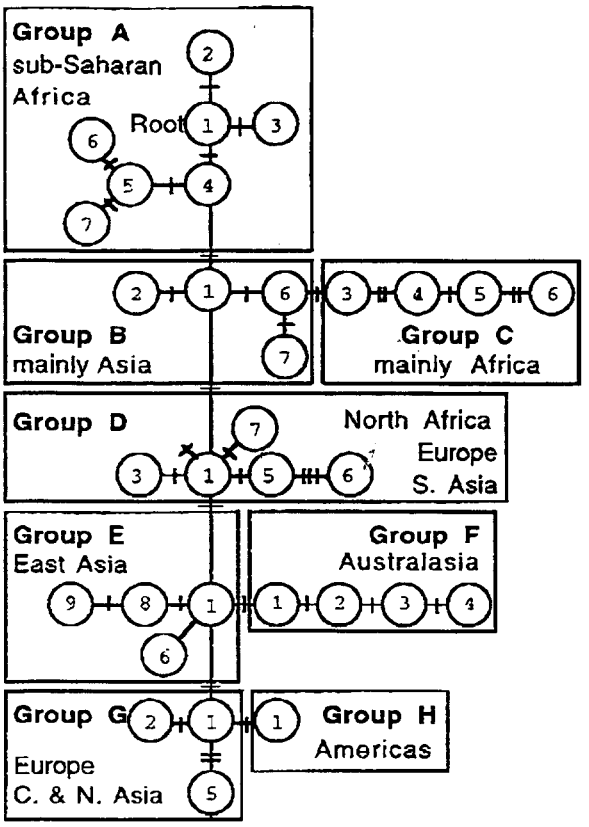

Figure 1. Evolutionary network for 31 haplotypes constructed from 30 Y-SNPs.

chromosomes within a population. We identified several SNPs that are specific to particular populations and began searching for Y-STRs that are variable in populations so far found to possess several common STR alleles and haplotypes. Then we developed high throughput systems for genotyping Y-STRs and Y-SNPs. Towards this goal we a) optimized two "pentaplex" assays for Y-STRs for genotyping samples on an ABI 373 platform, b) discovered novel Y-SNPS using current mutation detection methods, and c) devised PCR genotyping assays based on allele-specific PCR.

C. Populations

We analyzed DNA samples from three US population groups from the southwest including African Americans, Caucasian Americans, Southwest Hispanics, and two Southwest Native American populations (Apache and Navajo), as well as a composite sample of four East Asian populations (**Table 2a**). We also analyzed DNA samples from four other Native American populations, four European populations, and eight North Asian populations (**Table 2b**) (included as a reference for Native Americans because they have a similar "tribal" population structure).

Table 2. DNA samples analyzed (N=1141).

a. US Population groups (n=529)	n
African Americans	95
Caucasian Americans	94
Southwest Hispanics	97
Southwest Native Americans	118
Apache	
Navajo	
Pima	38
East Asians	
Chinese	37
Japanese	34
Koreans	35
Taiwanese	19

b. Other World populations (n=612)	n
Native Americans	209
Inuit (Greenland)	56
Mixtec (Mexico)	23
Panamanians	48
Europeans	108
Germans	34
Italians	24
Russians	29
British	21
North Asians	257
Altai	27
Buryats	47
Evenks	32
Forest Nentsi	34
Kets	29
Mongolians	30
Selkup	27
Tibetans	31

c. YCC Panel (n=74)	n
Africans	25
Europeans/West Asians	18
East/South Asians	15
Native Americans	13
Australasians	3

III. Results

A. Diversity Values for 18 Y-STRs

Table 3. Y-STR diversity in YCC panel.

Marker	#alleles	Haplotype diversity
AAGG10[a]	39[b]	0.975±0.007
DYS385[a]	36[b]	0.972±0.007
DYF371[a]	19[b]	0.887±0.021
DYS390	9	0.786±0.026
DYS392	8	0.771±0.037
DYS389B	6	0.747±0.032
DYS19	6	0.742±0.032
TAGA13	5	0.718±0.026
DYS393	5	0.659±0.042
DYS389D	5	0.641±0.041
DXYS156Y	5	0.545±0.046
DYS391	4	0.536±0.043
DYS426	4	0.525±0.043
DYS389A	4	0.515±0.056
G09411	4	0.512±0.055
DYS425	4	0.458±0.070
DYS388	7	0.355±0.069
G10123	4	0.155±0.056
ALL	71[b]	0.999±0.003

[a] banding patterns
[b] STR haplotypes

Table 3 lists heterozygosities for 18 Y-STR systems genotyped in the YCC panel—a repository of 74 lymphoblastoid cell-lines established from indigenous males originating in different parts of the world (**Table 2c**). The YCC panel represents an important resource because it provides a nearly inexhaustible supply of DNA that can be used for assessing the information content of forensic markers and for the establishment allelic and DNA standards. The three hypervariable STRs at the top of the table exhibited the highest diversity values (h = 0.975-0.887). The forensic utility of *DYS385* has been recognized (Caglia et al. 1998, Schneider et al. 1999); however, broad geographical surveys of *DYF371* have not appeared in the literature (Jobling et al. 1996). The AAGG10 STR, discovered in the course of this research, was the marker with the highest number of banding patterns and diversity. We plan to further investigate the forensic utility of both AAGG10 and *DYF371* in this research proposal.

Six of the seven core STRs recommended by Kayser et al. (1997) had reasonably high diversity values in the YCC panel. *DYS390* had the highest value (0.786±0.026) while *DYS388* exhibited a fairly low value (0.355±0.069). Another STR discovered in our lab—TAGA13, had an intermediate diversity value of 0.718±0.026. The five markers (*DYS19*, *DYS390*, *DYS391*, *DYS393*, and *DYS385*) chosen for our first pentaplex assay (see next section) produced 59 haplotypes with a combined haplotype diversity of 0.992±0.004. When all 18 STRs were considered together, 71 of the 74 Y-chromosomes in the YCC panel were individualized (discrimination capacity = 96%; h = 0.999±0.003). Interestingly, the three pairs of chromosomes that remained undifferentiated were from isolated populations (e.g., Yakuts, Biaka Pygmies, and Tsumkwe San). It is quite likely that these three pairs of Y-chromosomes were sampled from patrilineal male relatives.

B. STR Variation in Population Samples from Major US Groups

Table 4 summarizes results from our survey of five Y-STRs in a sample of 1141 Y-chromosomes from 28 population groups. Calculations of gene diversity by STR revealed values that were comparable to those estimated for *DYS19* (0.73), *DYS390* (0.73), *DYS391* (0.50), and *DYS393* (0.55) in a global survey of 986 males (de Knijff, unpublished). Our diversity value for *DYS385* (0.96), is higher than that reported by Kayser et al. (1997) for a German sample and similar to Caglia et al.'s (1998) estimate for an Italian sample. When all five STRs were considered together, a total of 530 haplotypes were observed in our sample of 1141 Y-chromosomes. However, the discrimination capacity (D.C.) of these Y-STR haplotypes varied among population groups: while most of the Y-chromosomes were differentiated in five of the seven groups (D.C. values ranged from 75% to 85%), less than 40% of the Y-chromosomes in our Native American and North/Central Asian samples were distinguished (**Table 4**). These

groups are known to be composed of relatively isolated sub-populations or demes that may contain high frequencies of paternally related lineages. This was also reflected in lower estimates of haplotype diversity (*h*) in the North/Central Asian group relative to other populations. Interestingly, Native American Y-haplotype diversity was not significantly lower than that estimated for Caucasian Americans.

Table 4. Summary statistics of Y-STR database (pentaplex I)

Population Sample	N	k[a]	D.C.[b] (%)	# alleles *DYS19*	*DYS390*	*DYS391*	*DYS393*	*DYS385*[c]	Haplotype diversity
African Americans	95	73	76.8	5	5	3	4	29	0.991±0.004
Caucasian Americans	94	70	74.5	6	5	4	4	29	0.981±0.008
SW Hispanics	97	78	80.4	5	6	3	3	29	0.990±0.005
Native Americans	365	135	37.0	4	5	4	4	37	0.981±0.002
East Asians	125	106	84.8	5	5	3	6	38	0.996±0.002
Europeans	108	83	76.9	5	6	4	4	28	0.991±0.004
North/Central Asians	257	91	35.4	6	6	4	4	31	0.959±0.005
Total	1141	530	46.5	7	7	5	6	66	0.994±0.001
Gene diversity:				0.747	0.698	0.465	0.511	0.961	

[a] number of STR haplotypes; [b] discrimination capacity; [c] number of banding patterns

Table 5. Locus-specific variation in US populations groups[a]

Locus	number of alleles	gene diversity	variance in repeat size
Pentaplex I			
DYS19	7	0.751	1.49
DYS390	6	0.732	1.38
DYS391	4	0.437	0.27
DYS393	5	0.484	0.40
DYS385 (A/B)	55[b]	0.951	na
average	15.4	0.671	0.88
Pentaplex 2			
DYS388	8	0.432	1.11
DYS389-1	6	0.700	0.93
DYS389-2	5	0.578	0.46
DYS392	8	0.754	2.48
DYS426	5	0.562	0.39
TAGA	6	0.708	1.49
average	6.3	0.621	1.15

[a]N=402; [b] number of patterns

Table 5 compares levels of variation at each of the Y-STRS in pentaplex 1 and pentaplex 2 in a sample of 402 males from five US population groups. Average levels of variability were slightly higher for pentaplex 1; however, when the hypervariable Y-STR—*DYS385*—was not considered, average gene diversity and variance were slightly higher for pentaplex 2 (e.g., H=.601for pentaplex 1). *DYS392* and a new Y-STR discovered in the course of this research (TAGA) demonstrated particularly high levels of variability. Pentaplex 2 also yielded interesting results when diversity statistics were compared across US populations. For example, Native Americans exhibited the highest allele size variance (1.20) and gene diversity (0.602) values of any of the five US population groups. For comparison, African Americans had values of 0.61 and 0.482, respectively.

C. SNP Variation in Major US and World Population Groups.

Figure 2 summarizes the frequencies of 18 Y-SNP haplotypes observed in six population groups. The haplotypes are color-coded according to their probable source (e.g., where they are found in indigenous populations or where they are hypothesized to have originated). For

example, nearly 70% of our African- American sample is composed of haplotypes (A and C-green) that are almost entirely limited to sub-Saharan African populations. However, this sample also has several haplotypes that are typical of European populations (D-orange and G-blue). This is most likely the result of paternal gene flow from Caucasian-American males to the African-American population (e.g., admixture in the US). The degree of admixture between Caucasian-Americans and African Americans is known to vary in different parts of the US (Parra et al. 1998). Native Americans are characterized by a high frequency of haplotypes that originated in both Asia (G1, red) and the Americas (H1, magenta). However, a low level of Caucasian-American and African-American Y-haplotypes are present in our Native American population sample (also see Karafet et al. 1999). A completely different set of haplotypes (Haplogroup E, red) is found in East Asian populations. It is interesting to note that European, Caucasian-American, and Southwest Hispanics have almost the same set of haplotypes at very similar frequencies. However, some degree of paternal gene flow from Native American males into the Caucasian-American and SW Hispanic populations is apparent by the presence of haplotype H1 (magenta) at low frequencies.

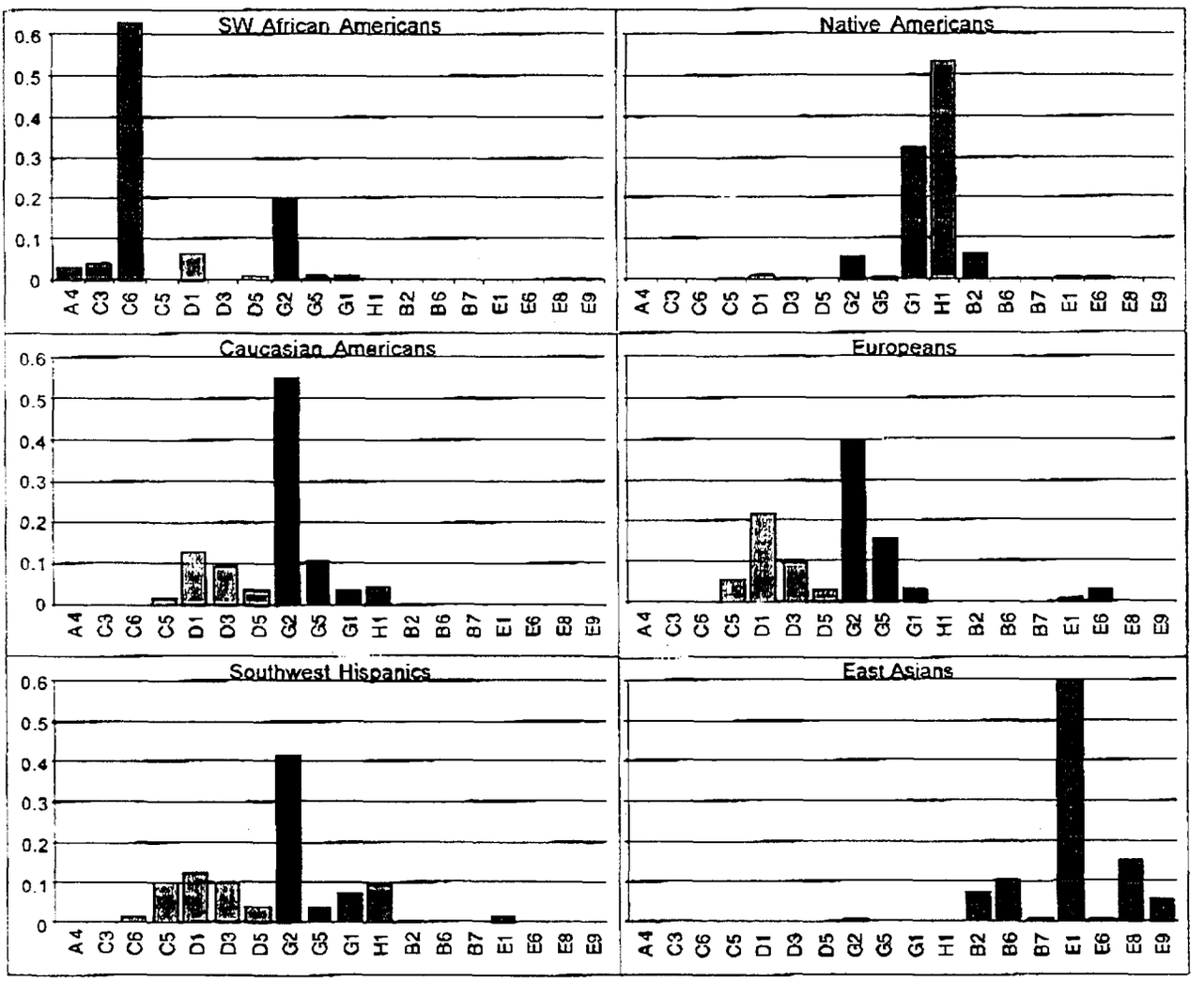

Figure 2. Haplotype frequencies in four US and two world population groups. Haplotype designations are the same as in Figure 2. Colors within each vertical bar indicate the inferred geographic origin of each haplotype: sub-Saharan Africa = green; North Africa/Middle East = orange; Europe = blue; Asia = red; and the Americas = magenta.

Table 6. Fst values for six population groupings.

Population grouping	STR haplotypes	SNP haplotypes	combination haplotypes
All populations (n=28)	0.069	0.356	0.068
US groups (n=5)	0.011	0.313	0.010
Native Americans (n=9)	0.053	0.105	0.054
North/Central Asians (n=8)	0.192	0.402	0.189
East Asians (n=4)	0.007	0.254	0.007
Europeans (n=4)	0.009	0.083	0.009

The high degree of population specificity of Y-SNP haplotypes is also reflected in their high Fst values. (**Table 6**). The Fst values for all 28 populations in this study was 0.356. This means that 36% of the total variance of Y-haplotypes is attributable to differences between populations. This SNP haplotype Fst value compares with an Fst of 0.069 for pentaplex I Y-STR haplotypes in the same set of 28 population samples. The lower Fst for Y-STR haplotypes is a function of the convergent mode of STR mutation between populations, as well as the greater degree of haplotype individualization within populations. When the five major US population groups were considered, the SNP haplotype Fst value remained high (0.313) while the STR haplotype Fst value was further reduced (0.011).

In sum, these results point to the utility of Y-SNPs for identifying population-specific Y-chromosome haplotypes, and to the importance of considering admixture between US groups in forensic analysis. In the next section we explore the value of combining information from both SNP and STR genetic systems.

D. Combination Haplotypes, Population Structure, and Admixture

When we combined information from 18 Y-SNPs and five pentaplex I STRs surveyed in our sample of 1141 chromosomes, there was a 10.6% increase in the number of haplotypes. In particular, several Native American haplotypes that were indistinguishable based on STR information alone became separate combination haplotypes on G1 or H1 haplotype backgrounds (see **Figure 2**). Fst values based on combination haplotypes were generally as low (or slightly lower than) those based on STR haplotypes (**Table 6**). Therefore, combination haplotypes reveal the individualization properties of Y-STRs while retaining the geographic and evolutionary information of SNPs.

A comparison of the frequency distribution of combination haplotypes is shown in **Figure 3**. Of the 586 STR haplotypes, ~75% occurred only in a single individual (unique), 17% occurred more than once in only one population (population-specific), and 8% were shared among populations (shared). A similar frequency distribution was observed in East Asian (highest frequency of unique haplotypes) and African-American populations. Interestingly, Caucasian-Americans, Europeans, and Hispanics displayed very similar frequencies distributions: they had the highest frequencies of shared haplotypes (~35%) and the lowest frequencies of population-specific haplotypes (~1%). North/Central Asian and Native Americans populations displayed very similar frequency distributions, with the highest frequency of population-specific haplotypes (27% and 33%, respectively).

When we examined which populations were sharing combination haplotypes, it was clear that European, Caucasian- American, and SW Hispanics were mainly sharing haplotypes with each other and to a lesser extent with African-Americans (**Table 7**). As discussed above, the sharing of haplotypes between Caucasian-American and African-Americans most likely reflects male-mediated admixture. There was also some sharing of combination haplotypes between Caucasian-Americans/Europeans/Hispanics and Native Americans. In pairwise population

haplotype permutation tests, Caucasian-Americans, Europeans, and Hispanics were not differentiated. Interestingly, African-Americans were only marginally differentiated from Europeans and SW Hispanics, whereas Native Americans were significantly differentiated from African-, Caucasian-, and Hispanic-Americans (**Table 7**, above the diagonal). The only group that did not share any haplotypes with other populations was our East Asian sample.

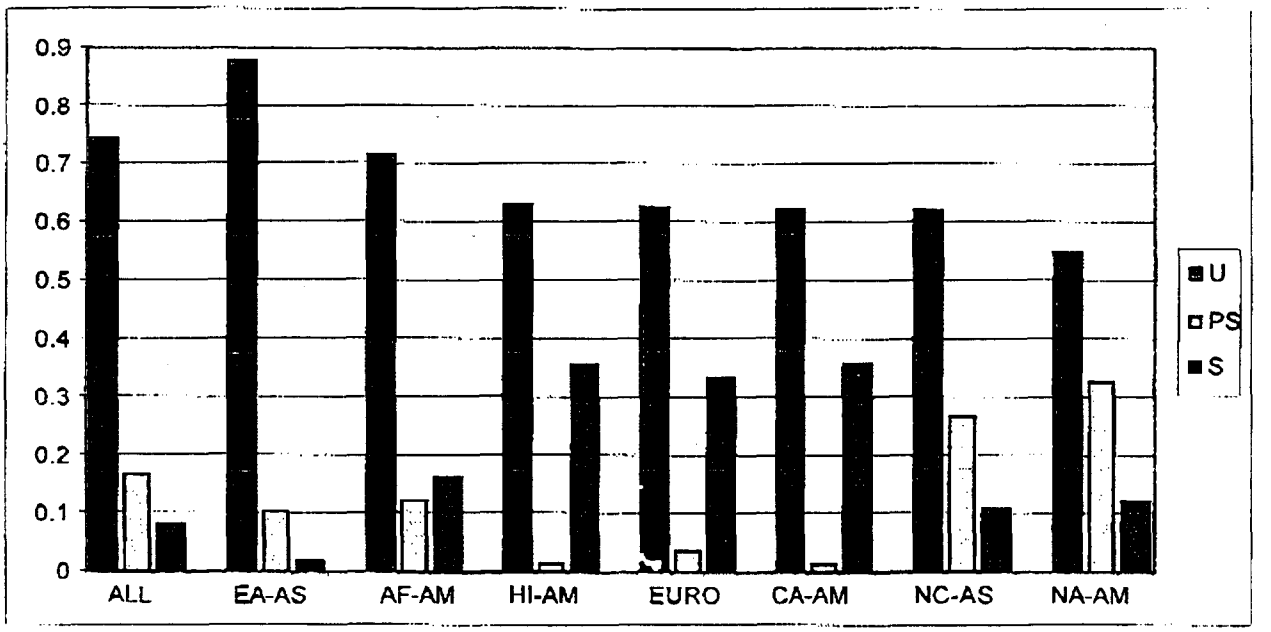

Figure 3. Frequency distribution of unique (blue-filled bars), population-specific (yellow-filled bars), and shared (red-filled bars) combination haplotypes in global and regional population groups: EA-AS = East Asians, AF-AM = African Americans, HI-AM = Southwest Hispanics, EURO = Europeans, CA-AM = Caucasian Americans, NC-AS = North/Central Asians, and NA-AM = Native Americans.

E. Further Considerations

Table 7. Number of haplotypes shared between population groups (below diagonal) and p value of population differentiation test (above diagonal).

	AF-AM	CA-AM	EURO	HI-AM	EA-AS	NC-AS	NA-AM
AF-AM	-	*	0.06	0.08	*	*	*
CA-AM	8	-	0.79	0.56	*	*	*
EURO	8	15	-	0.93	*	*	*
HI-AM	5	14	18	-	*	*	*
EA-AS	0	0	0	0	-	*	*
NC-AS	0	2	5	3	2	-	*
NA-AM	5	11	8	10	0	3	-

* p <0.05, significantly differentiated.

We note that Fst values for US population groups as a whole and those for East Asians and Europeans are ≤1%. In contrast, the percent of Y-chromosome combination haplotype variance attributable to differences between Native American populations is >5%, and for North/Central Asian populations it is almost 20%. These two population groups also had the highest levels of population-specific haplotypes. Continued analysis will reveal if the elevated Fst values could be the result of heterogeneity of particular population–specific haplotypes among sub-populations or tribal groups. In any case, we can conclude that the amount of variation among Native American populations is 6-8 times higher than among East Asians and European, respectively, and comparable to the among-group component of the 28 world populations considered in this study. Our continued analyses of these data, as well as

continued efforts to establish a more detailed Y chromosome database of well-defined subpopulations, should allow a better assessment of how both population structure and admixture in US population groups will impact the statistical analysis of Y-chromosome forensic data.

Literature Cited

Caglia A, Dobosz M, Boschi I, d'Aloja E, Pascali VL (1998) Increased forensic efficiency of a STR-based Y-specific haplotype by addition of the highly polymorphic DYS385 locus. Int J Legal Med 111:142-6

Jobling MA, Samara V, Pandya A, Fretwell N, Bernasconi B, Mitchell RJ, Gerelsaikhan T, et al (1996) Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. Hum Mol Genet 5:1767-75

Karafet T, Osipova L, Posukh O, Weibe V, Hammer MF (1998b) Y chromosome microsatellite haplotypes and the history of Samoyed-speaking populations in Northwest Siberia. In: Goldstein DB, Schlötterer C (eds) Microsatellites: Evolution and Applications. Oxford University Press, Oxford, pp in press

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, et al (1999) Ancestral Asian Source(s) of New World Y-Chromosome Founder Haplotypes. Am. J. Hum. Genet. 64:817-831

Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, et al (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110:125-33

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, et al (1998) Estimating African American admixture proportions by use of population- specific alleles. Am J Hum Genet 63:1839-51

Schneider PM, Meuser S, Waiyawuth W, Seo Y, Rittner C (1998) Tandem repeat structure of the duplicated Y-chromosomal STR locus DYS385 and frequency studies in the German and three Asian populations. Forensic Sci Int 97:61-70

White PS, Tatum OL, Deaven LL, Longmire JL (1999) New, male-specific microsatellite markers from the human Y chromosome. Genomics 57:433-7