J Forensic Sci, September 2009, Vol. 54, No. 5
doi: 10.1111/j.1556-4029.2009.01105.x
Available online at: www.blackwell-synergy.com

Bruce Budowle,[1,2] *Ph.D.; Jianye Ge,*[3,4] *M.S.; Xavier G. Aranda,*[2] *M.S.; John V. Planz,*[2] *Ph.D.; Arthur J. Eisenberg,*[1,2] *Ph.D.; and Ranajit Chakraborty,*[4] *Ph.D.*

Texas Population Substructure and Its Impact on Estimating the Rarity of Y STR Haplotypes from DNA Evidence*

ABSTRACT: Three sampled populations of unrelated males—African American, Caucasian, and Hispanic, all from Texas—were typed for 16 Y short tandem repeat (STR) markers using the AmpF*l*STR® Yfiler™ kit. These samples also were typed previously for the 13 core CODIS autosomal STR loci. Most of the 16 marker haplotypes (2478 out of 2551 distinct haplotypes) were observed only once in the data sets. Haplotype diversities were 99.88%, 99.89%, and 99.87% for the African American, Caucasian, and Hispanic sample populations, respectively. F_{ST} values were very small when a haplotype comprised 10–16 markers. This suggests that inclusion of substructure correction is not required. However, haplotypes consisting of fewer loci may require the inclusion of F_{ST} corrections. The testing of independence of autosomal and Y STRs supports the proposition that the frequencies of autosomal and Y STR profiles can be combined using the product rule.

KEYWORDS: forensic science, DNA typing, autosomal STR loci, Y STR, Y chromosome, population substructure, F_{ST}, partial profile, independence, statistics, joint match probability and theta correction

The Y chromosome short tandem repeat (Y STR) markers have been used successfully for several years to analyze DNA from forensic biological evidence (1–10). While much work has been dedicated to develop robust assays (8,10), less effort has been devoted in developing a robust methodology for estimating the rarity of a Y STR profile. Because the Y STR loci that were selected for forensic applications reside on the nonrecombining part of the Y chromosome, Budowle et al. (1,11,12) recommended using the counting method for estimating the rarity of a Y STR haplotype. The counting method is conservative, simple, and has precedence in forensic analyses of lineage-based forensic systems (13). However, unless an upper confidence limit is invoked with the counting approach (12), it may not be sufficiently conservative. Population substructure may have an effect on the conservative nature of the counting method if the effects of population substructure are large. The lack of independence between loci and smaller effective population size may yield greater population substructure effects on a locus-by-locus basis than have been observed previously for the autosomal loci, as can be inferred from the study of allele frequency differences at seven Y STR markers in 20 world populations (14). Studies are needed on forensically relevant populations to determine empirically what population substructure effects exist at the haplotype level and what methods might be applicable to correct for those effects when estimating the conditional probability of a Y STR haplotype. In addition, as haplotype level population substructure effects may depend upon the number of loci encompassed in the haplotype, the dependency of haplotype-level substructure effects on the number of loci (and the locus compositions) is also an issue that needs further investigation.

In this paper, the three largest subpopulations from Texas, i.e., African Americans, Caucasians, and Hispanics, were typed for 16 Y STR markers and tested for the degree of substructure within them. In addition, as 13 core autosomal STR loci have been typed previously for these samples, a test for independence based on mismatch distribution was used to determine if the frequencies of autosomal and Y STR profiles can be multiplied under the assumption of independence. The results herein provide guidance for calculating the rarity of Y STR haplotypes.

Materials and Methods

Samples

DNA was obtained from unrelated male donors from paternity testing cases submitted to the DNA Identification Lab at the University of North Texas Health Science Center, Ft. Worth, Texas. Population affinity was ascribed by self-declaration. The samples were African Americans, $N = 950$; Caucasians, $N = 957$; and Hispanics, $N = 1005$.

Buccal swabs, sterile foam–tipped swab with a 6″ polypropylene shaft (Catalogue #: 25–1616 2PF, Puritan Hardwood Products Company, LLC, Guilford, MN) were used for sample collection. Each swab was removed from the packaging by grasping the plastic handle and the tip was placed directly into the individual's mouth. The swab was vigorously rubbed against the inner cheek for a minimum of 20 up and down strokes per swab. The swab was allowed to air dry.

[1]Institute of Investigative Genetics, University of North Texas Health Science Center at Ft. Worth, Ft. Worth, TX 76107.
[2]Department of Forensic and Investigative Genetics, University of North Texas Health Science Center at Ft. Worth, Ft. Worth, TX 76107.
[3]Department of Biomedical Engineering, University of Cincinnati, Cincinnati, OH 45221.
[4]Department of Environmental Health, Center for Genome Information, University of Cincinnati College of Medicine, Cincinnati, OH 45267.

*08–14 is publication number of the Laboratory Division of the Federal Bureau of Investigation. Names of commercial manufacturers are provided for identification only, and inclusion does not imply endorsement by the Federal Bureau of Investigation.

Received 17 May 2008; and in revised form 29 Sept. 2008; accepted 12 Oct. 2008.

Sample Preparation

The buccal cells were eluted from the swab into 1 mL Isotone® III buffer (Beckman, Fullerton, CA) contained in a 2.0 mL Dolphine (nipple) microcentrifuge tube (Costar, Cat no. 3213, Corning, Inc., Corning, NY). Cells were eluted by swirling and pelleted by centrifugation (1 min at 1500 × *g*). A 15 μL aliquot of Coomassie Brilliant Blue G250 (1 mg/mL in H_2O) was added to the eluted cells to aid visualization of the cell pellet for subsequent spotting on to the FTA matrix (Whatmann, Florham Park, NJ). The majority of the supernatant was decanted and the pellet was resuspended in the remaining residual buffer. A 15 μL aliquot was spotted directly onto the target circles on FTA® Micro Sheet™ (Whatmann). The FTA® Micro Sheet™ was air-dried for a minimum of 1 h prior to processing.

The FTA® Micro Sheet™ was placed into a Hybriboat and 35 mL of FTA® Purification Reagent containing 10 μg/mL of Proteinase K was added and incubated at 65°C for 15 min. The reagent was discarded and the FTA® Micro Sheet™ was washed according to the manufacturer's recommendations. The FTA® Micro Sheet™ was air-dried prior to storage and sampling for amplification.

Y STR Typing

The PCR amplification was performed using the AmpF*l*STR® Yfiler™ kit (Applied Biosystems, Foster City, CA) according to the manufacturer's instructions, except that a 12.5 μL reaction volume was employed with a 1.2 mm FTA punch serving as the template DNA. Amplification was performed in an ABI PRISM® GeneAmp® 9700 Gold-plated or Silver block Thermal Cycler (Applied Biosystems) using the 9600 emulation mode for 28 cycles. PCR products were separated and detected on an ABI PRISM® 3100 and 3130 xl Genetic Analyzer (Applied Biosystems) following the manufacturer's recommendations. Prior to electrophoresis, 1.5 μL of the amplified product or allelic ladder and 0.3 μL of GeneScan™-500 LIZ® size standard (Applied Biosystems) were added to 8.7 μL of deionized Hi-Di™ formamide (Applied Biosystems), denatured at 95°C for 5 min, and then chilled on ice for 5 min. Samples were injected for 15 sec at 3 kV in performance optimized polymer (POP-4™; Applied Biosystems) using the GeneScan 36vb_POP4 Dye Set G5 Module (Applied Biosystems) for both instruments and run time was 1500 sec. The data were collected using the ABI PRISM® 3100 Data Collection Software v1.1 (Applied Biosystems) and ABI PRISM® 3130 xl Genetic Analyzer Data Collection Software 3.0, respectively. Electrophoresis results were analyzed with GeneMapper® ID software v3.2 (Applied Biosystems). Allele peaks were called when the peak heights were greater than or equal to 50 relative fluorescence units.

Autosomal STR Typing

The PCR amplification was performed using the AmpF*l*STR® Profiler Plus™ *ID* and COfiler™ kits (Applied Biosystems) according to the manufacturer's instructions, except that a 12.5 μL reaction volume was employed with a 1.2 mm FTA punch serving as the template DNA. Typing was performed as described above for the Y STR markers.

Statistical Analyses

Gene diversity at each locus, the number of haplotypes, and haplotype diversity were calculated using a program developed by one of us (Jianye Ge). Power of discrimination (PD) (or diversity) was calculated as $PD = 1-\sum p_i^2$ where p_i is the haplotype frequency. Bias correction (i.e., multiplication of this expression by a factor of $N\,[N–1]^{-1}$) was not carried out because of the possibility that when each haplotype was observed once in a database of size *N*, bias-corrected PD would equal 1. The F_{ST} values were calculated according to Weir and Cockerham (15) as described by Weir (16) using haplotype data according to the logic described previously (11,12).

To assess autosomal loci and Y STR haplotype independence, within a sampled population, the Y STR profiles for each pairwise comparison of individuals were scored based on the number of mismatched loci (i.e., having a different allele-type at a locus and scored from 0 to 16 mismatched markers), a concept used in molecular evolution (17) that has been used in the context of Y-STR studies (18). The same was carried out for each pair of samples for the profiles based on 13 CODIS autosomal loci (i.e., having 0 to 13 loci with mismatched genotypes). The independence of the mismatch distribution between the Y STR haplotypes and 13-locus autosomal STR profiles was computed using the chi-squared statistic for each relevant r × c contingency table (14 × 17 for each population). Significance of the chi-squared statistic was determined by a permutation test (19), with 10,000 replications for each population. This procedure was different compared with the previously reported approaches to study independence between autosomal loci and Y-STR DNA profiles (see e.g., 20,21); but it had the benefit of avoiding problems that arose from degeneracy of the test statistic when most of the Y-STR haplotypes occurred in single copies in a sampled population.

Data Access

The haplotype profiles for the three Texas sample populations have been submitted to the U.S. YSTR Database (http://usystrdatabase.org).

Results and Discussion

Population Data

This study analyzed 2,912 males from three sample populations residing in Texas. Although there are technically 17 loci typed with the Y Filer kit, in this study, they will be referred to as 16 markers because the DYS385 marker comprised two loci resulting from a tandem duplication. The PD for the markers individually for African Americans, Caucasians, and Hispanics is listed in Table 1. The DYS385 and DYS458 markers are the two most

TABLE 1—*PD per Y STR marker per population.*

Locus	African American	Caucasian	Hispanic
DYS389I	0.5375	0.5238	0.5766
DYS389II	0.7438	0.6972	0.7296
DYS390	0.6480	0.6949	0.6294
DYS456	0.6344	0.7301	0.6806
DYS19	0.7427	0.5413	0.6635
DYS458	0.7501	0.7673	0.7937
DYS437	0.5301	0.5904	0.5670
DYS438	0.5441	0.5810	0.7067
DYS448	0.6962	0.6321	0.6953
GATA H4	0.5921	0.5947	0.5911
DYS391	0.4353	0.5451	0.5588
DYS392	0.4452	0.6022	0.7162
DYS393	0.5920	0.3229	0.4450
DYS439	0.6290	0.6528	0.6757
DYS635	0.7364	0.6251	0.7068
DYS385	0.9488	0.8252	0.9233

PD, power of discrimination; STR, short tandem repeat.

polymorphic (i.e., highest PD on per-marker basis) across all populations and, on average, the DYS393 marker is the least discriminating on a per-marker basis.

Haplotype diversity, however, is a better indicator of the power of the 16 marker system because of the lack of biological independence among Y STR markers. The 16 marker haplotype diversities were 0.9988, 0.9989, and 0.9987, for African Americans, Caucasians, and Hispanics, respectively (Table 2). A total of 2,551 distinct haplotypes was observed in the total data set. The majority of haplotypes was observed only once in all sample populations (2,478 out of a total of 2,551 distinct haplotypes in the total Texas data set). Consequently, only 73 distinct haplotypes were seen more than once in the total data set (Table 3). The number of shared haplotypes between African American and Caucasian sets was only 30, African American and Hispanic sets had only 10 haplotypes in common, and Caucasian and Hispanic sets had only 24 haplotypes in common. The maximimum and minimum PD values, given a number of markers comprising a haplotype, are displayed in Table 4. The five markers that contributed the most to the maximum haplotype PD in all three Texas populations were DYS389II, DYS456, DYS458, DYS439, and DYS385. The five markers contributing the least to the maximum haplotype PD varied among the three populations but the DYS438 and DYS392 markers had consistently low PDs in these sample populations. The DYS438 marker did not contribute substantially to haplotype diversity, as described previously (11,22). The high haplotype diversity in these results support that any 16 marker Y haplotype derived from evidentiary material will be infrequent in all three Texas population groups.

TABLE 2—*Haplotype data per population group.*

Population (Sample Size)	No. Distinct Haplotypes	No. Haplotypes Observed Once	Haplotype Diversity
African American (N = 950)	906	866	0.9988
Caucasian (N = 957)	925	896	0.9989
Hispanic (N = 1005)	904	826	0.9984

TABLE 3—*Most frequently observed haplotypes in Texas population samples.*

	African American	Caucasian	Caucasian	Caucasian	Hispanic
Sample Size	N = 950	N = 957	N = 957	N = 957	N = 1005
No. Most Frequent Haplotypes	n = 5	n = 3	n = 3	n = 3	n = 6
DYS389I	13	13	12	13	13, 14
DYS389II	30	30	28	29	29
DYS390	21	25	22	25	24
DYS456	16	16	14	17	15
DYS19	17	17	14	14	14
DYS458	16	16	15	17	15
DYS437	13	14	16	15	15
DYS438	11	11	10	12	12
DYS448	21	20	20	18	19
Y GATA H4	11	12	11	12	11
DYS391	10	10	10	11	11
DYS392	11	11	11	14	13
DYS393	13	13	14	13	13
DYS439	12	10	11	12	10, 11
DYS635	22	23	21	23	23
DYS385	18, 18	10, 14	13, 14	11, 13	13, 15

TABLE 4—*Maximum and minimum PD values for a specified number of markers in each population.**

No. Markers	African American		Caucasian		Hispanic	
	Maximum	Minimum	Maximum	Minimum	Maximum	Minimum
1	0.9488	0.4353	0.8252	0.3229	0.9233	0.4450
2	0.9840	0.6187	0.9523	0.6578	0.9776	0.7499
3	0.9934	0.7402	0.9822	0.7620	0.9909	0.8758
4	0.9966	0.8088	0.9917	0.8403	0.9957	0.9149
5	0.9979	0.8593	0.9955	0.8905	0.9975	0.9422
6	0.9984	0.9196	0.9973	0.9208	0.9981	0.9670
7	0.9986	0.9576	0.9980	0.9465	0.9984	0.9798
8	0.9987	0.9776	0.9984	0.9662	0.9985	0.9876
9	0.9988	0.9880	0.9986	0.9799	0.9986	0.9922
10	0.9988	0.9937	0.9987	0.9886	0.9987	0.9951
11	0.9988	0.9964	0.9988	0.9937	0.9987	0.9966
12	0.9988	0.9977	0.9988	0.9964	0.9987	0.9975
13	0.9988	0.9983	0.9988	0.9978	0.9987	0.9981
14	0.9988	0.9986	0.9989	0.9984	0.9987	0.9985
15	0.9988	0.9988	0.9989	0.9987	0.9987	0.9986
16	0.9988	0.9988	0.9989	0.9989	0.9987	0.9987

PD, power of discrimination.

*These (haplotype-based) PD values are based on the number of markers comprising a haplotype. The specific markers will vary obviously for the maximum and minimum PD values for the same number of markers and may vary among the population groups.

Population Structure for Forensic Analyses

The counting method is a conservative approach for estimating the rarity of a Y STR profile, as it inherently assumes that these markers are at perfect linkage disequilibria with each other. The Y STR loci reside in the nonrecombining region of the Y chromosome and thus are linked (although the high mutation rate at these loci may disrupt the linkage disequilibrium to some degree [11, 20]). Therefore, each haplotype is treated as an allele instead of each component of the haplotype being counted. In a database search, the number of haplotypes matching the evidence haplotype are counted, this number is divided by the number of samples in the database, and to correct for sampling error, a 95% upper confidence limit value is applied (12). However, in some scenarios the estimate might not be sufficiently conservative, such as merging multiple populations for the count estimate and/or when partial profiles are derived from limited quantity and low quality samples. Actually, haplotype-level population substructure effect is influenced by two counter-balancing factors. At the individual locus level, Y STRs are subject to a larger population substructure effect (than autosomal loci) due to small effective population size (resulting from a haploid nature and being present only in males) and uniparental transmission. In contrast, at the haplotype level, the greater number of loci, generally the higher was the PD, which was expected to produce a smaller substructure effect (23). Thus, the potential effects of population substructure which may impact estimates of the rarity of a Y STR profile should be examined for varying compositions of the markers (in terms of number as well as combinations) that define a haplotype.

Budowle et al. (11,12), based on the recommendations of the NRC II Report (24), have described the forensically based approach to derive the co-ancestry coefficient θ values to correct profile frequency estimates for possible population substructure effects. The variable θ is used as a generic term for the measure of population substructure (24). The probability of observing a specific haplotype in an unrelated male, given that it was seen in another male, was calculated using the formula: $p + \theta(1 - p)$, where p is the estimated haplotype frequency, and θ is the haplotype-level F_{ST}, which

is directly derived from the theory (applied at the haplotype level) as described in Balding and Nichols (25), also used in (11,12). Forensic applications assess comparisons of evidence and reference Y STR profiles in terms of identity by state as either a "match" (inclusion) or "nonmatch" (exclusion). Therefore, the effect of population substructure (i.e., θ) should be based on F_{ST} values (11,12). While estimating F_{ST} for each population group would yield a more meaningful estimate, an upper bound estimate of F_{ST} could be obtained by combining the major populations of Texas. The F_{ST} value was 0.00012 for using the three Texas sample populations for a 16 marker haplotype (Table 5). Generally, this value should be smaller for population specific estimates (which is the subject of a manuscript in preparation). Regardless, the F_{ST} value was so small that with the size of the current reference population data sets described herein, an F_{ST} correction had little or no effect on the upper bound of the Y STR count proportion. The F_{ST} results lend further support to the proposition that there is no need to employ a substructure correction for estimating the probability of observing a specific haplotype in an unrelated male given that it was seen in another male when the haplotype comprised 10 or more markers (Table 5).

This observation of a small F_{ST} value was consistent with the theory of population substructure and expectations for Y chromosome markers. On average, the F_{ST} values were larger for individual Y STRs than for the autosomal loci (Tables 6 and 7). However, the Y STRs were not treated as individual markers, but instead as part of a haplotype. As such, they were combined and represented as a single locus with many alleles (a haplotype essentially is an allele). The Y STR haplotype diversity was greater than observed for most autosomal loci used in forensic analyses. This high diversity combined with little haplotype sharing between populations support that F_{ST} values should be exceedingly small for situations involving a complete profile obtained from the validated, commercially available forensic Y STR kits (8,10).

Partial profile evidence may require a correction for the effects of population substructure. As the number of Y STR markers that comprise a haplotype decreased, the number of shared haplotypes within and between sample populations increased. Thus, the F_{ST} value was expected to increase for partial profiles from evidence samples. At some point, an F_{ST} correction will not be overwhelmed by the upper bound of the count proportion, and then a θ correction should be invoked in a statistical calculation.

Two ways are suggested to determine when an F_{ST} value should be invoked for partial profile calculations: (i) a simple pragmatic approach and (ii) a marker specific approach. The pragmatic approach uses the same maximum F_{ST} value for a specified number of markers comprising an evidence profile (Table 5), irrespective of the specific loci displayed in a profile. Under this approach, this maximum F_{ST} value is used and thus in most cases would add another dimension of conservatism to the rarity of the frequency estimate. For example, the maximum F_{ST} value for five markers was 0.0508 and was observed specifically for the markers DYS390, DYS437, DYS438, DYS391, and DYS392 (Table 5).

TABLE 5—*Maximum* F_{ST} *and accompanying PD values for various numbers of markers comprising a Y STR profile using the three Texas populations.*

Number of markers comprising haplotype	Marker combination with maximum F_{ST}*†	F_{ST}‡	PD
1	2	0.1776	0.7518
2	7, 11	0.1507	0.8041
3	6, 7, 11	0.0989	0.8820
4	2, 6, 7, 11	0.0721	0.9411
5	2, 6, 7, 10, 11	0.0508	0.9634
6	2, 6, 7, 8, 10, 11	0.0289	0.9764
7	0, 2, 6, 7, 8, 10, 11	0.0156	0.9859
8	0, 2, 6, 7, 8, 9, 10, 11	0.0083	0.9922
9	0, 1, 2, 6, 7, 8, 9, 10, 11	0.0045	0.9955
10	0, 1, 2, 6, 7, 8, 9, 10, 11, 12	0.0024	0.9969
11	0, 1, 2, 4, 6, 7, 8, 9, 10, 11, 12	0.0013	0.9977
12	0, 1, 2, 4, 6, 7, 8, 9, 10, 11, 12, 14	7.75E-04	0.9984
13	0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 14	4.61E-04	0.9991
14	0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14	2.88E-04	0.9994
15	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	1.80E-04	0.9995
16	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15	1.22E-04	0.9996

PD, power of discrimination; STR, short tandem repeat.

*0 = DYS389I, 1 = DYS389II, 2 = DYS390, 3 = DYS456, 4 = DYS19, 5 = DYS458, 6 = DYS437, 7 = DYS438, 8 = DYS448, 9 = Y GATA H4, 10 = DYS391, 11 = DYS392, 12 = DYS393, 13 = DYS439, 14 = DYS635, and 15 = DYS385.

†These loci were the combination of those that yielded the largest F_{ST} value; other combinations yielded smaller F_{ST} values.

‡Note that the F_{ST} value generally decreases with an increasing number of markers comprising the haplotypes. F_{ST} is calculated based on haplotype diversity. More markers comprising the haplotypes generally will result in more diversity and thus more haplotypes.

TABLE 6—F_{ST}* *on a per Y STR marker basis for the combined Texas populations.*

Marker	F_{ST}
DYS389I	0.0023
DYS389II	0.0347
DYS390	0.1775
DYS456	0.0164
DYS19	0.1007
DYS458	0.0067
DYS437	0.1038
DYS438	0.1544
DYS448	0.0940
Y GATA H4	0.0323
DYS391	0.0530
DYS392	0.1443
DYS393	0.0556
DYS439	0.0034
DYS635	0.0834
DYS385	0.0335

STR, short tandem repeat.

*F_{ST} calculations were based on alleles.

TABLE 7—F_{ST}* *on a per autosomal locus basis for the Texas populations.*

Marker	F_{ST}
CSF1PO	0.0082
D13S317	0.0356
D16S539	0.0144
D18S51	0.0125
D21S11	0.0173
D3S1358	0.0137
D5S818	0.0297
D7S820	0.0152
D8S1179	0.0170
FGA	0.0080
TH01	0.0450
TPOX	0.0308
vWA	0.0172
Average	0.0201

*F_{ST} calculations were based on alleles.

The conditional probability would be calculated with θ as 0.0508 for any combination of five markers. Now consider that the actual evidence partial profile was comprised instead of the five markers DYS389II, DYS456, DYS458, DYS439, and DYS385 (the ones that contribute more so to the maximum haplotype PD). The specific F_{ST} value for these latter five markers is 0.00091, which is 55 times smaller than the maximum F_{ST} value. Now consider that a sample is notably degraded; it is likely that those markers with the smallest size amplicons will have a greater success of typing than larger amplicon markers. The markers DYS391, DYS393, DYS456, DYS458, and Y GATA H4 have the smallest amplicons when using the AmpF*l*STR® Yfiler™ kit. The specific F_{ST} value for these small-sized amplicon five markers is 0.00272, which is more than 18 times smaller than the maximum F_{ST} value. The pragmatic approach is therefore conservative with the ancillary benefit of being simple; only a 16 row table is needed for capturing the necessary F_{ST} values to employ. The alternate approach would employ the F_{ST} value dependent on the specific markers observed in the profile, and in many practical cases (such as partial profiles due to DNA degradation) will be notably less than the maximum F_{ST} value. With today's computer power all F_{ST} computation values for all possible markers and their combinations could be archived and accessed. If the specific approach is used, there will be more situations where the F_{ST} value will be so small that it will be unnecessary to correct for population substructure. While the pragmatic and specific approaches generally will yield different values, this should not be construed as a conflict. Either estimate is a conservative estimate; the difference is in the degree of conservatism.

Independence of Y and Autosomal Markers

There may be evidence samples that have been typed for both autosomal and Y STRs. Certainly, the combined results of autosomal and Y markers are rarer than either an autosomal STR or Y STR profile frequency individually. In this study, we considered only independence of Y and autosomal markers for unrelated individuals; in the situation where paternal relatives are involved, the conditional probability for the Y haplotype (barring mutation) will be 1. Because the Y chromosome is biologically independent of the autosomal chromosomes, it would seem appropriate to use the product rule to combine the random match probability derived for an autosomal STR profile and the upper bound frequency estimate of a Y STR haplotype (with substructure corrections when deemed appropriate). However, this assumption should be tested. Prior to assessing the assumption of independence between the two marker systems, the quality of the autosomal STR loci data was evaluated. There were no more departures from Hardy-Weinberg and linkage equilibrium expectations than expected by chance for the autosomal loci in any of the three Texas sample populations (data not shown). In addition, the allele frequencies distributions were not significantly different from previously reported similar populations (26) (data not shown). Based on mismatch distributions (17), there was no evidence for departure from independence between the autosomal and Y STR markers for African Americans ($p = 0.354$), Caucasians ($p = 0.303$), and Hispanics ($p = 0.227$). These results were similar to those reported previously (11,20,21), obtained by using test statistics different from the one used here. Thus, the data herein also support that a single locus or multilocus autosomal STR profile frequency (adjusted for population substructure effect) can be multiplied by the upper bound Y STR haplotype frequency (or the one corrected for the effects of population substructure when appropriate).

Concordance

An ancillary benefit of this study is that a subset of these samples (48 African Americans, 162 Caucasians, and 128 Hispanics) were typed previously with the PowerPlex® Y System (Promega Corp., Madison, WI) (11). Therefore, there was an opportunity to compare the two commercial kit systems for typing concordance for the markers they have in common. These markers were DYS389I, DYS389II, DYS390, DYS19, DYS437, DYS438, DYS391, DYS392, DYS393, DYS439, and DYS385. There were no typing discrepancies between the two commercial kits. While the sampling is not large ($N = 338$), the data support that a comparison of Y STR profiles between laboratories using different kits should not be problematic. This observation was consistent with the concordance study by Gross et al. (27).

Conclusion

The Texas data further support that Y STR haplotypes are highly polymorphic and have a high power of discrimination in forensically relevant U.S. populations. The three major Texas populations provide data for the upper bound of the effect of population substructure. Analyses showed, for haplotypes that comprised at least 10–16 markers, the effects of population substructure were small, and there was little or no need to correct for population substructure when estimating the conditional probability of a Y STR haplotype using the counting method. The counting method with a correction for sampling error appeared to be sufficiently conservative. However, estimates of the conditional probability of partial profiles, depending on the number of markers and haplotype sharing, may require a correction for population substructure. A pragmatic maximum F_{ST} value approach or a specific marker F_{ST} value approach is suggested for conditional probability estimates for partial profiles where substructure correction is warranted. Both approaches are valid and conservative. A computer program evaluating the counterbalancing effects of the number of markers and combination of markers on F_{ST} and PD (both at the haplotype-level) and its impact on estimating the conditional probability of any target haplotype (partial or full) will be made available online in the near future (at the University of Cincinnati). Lastly, the data support that frequency estimates of autosomal and Y STR profiles can be combined by multiplying under the assumption of independence.

References

1. Budowle B, Sinha SK, Lee HS, Chakraborty R. Utility of Y-chromosome STR haplotypes in forensic applications. Forensic Sci Rev 2003;15:153–64.
2. Corach D, Risso LF, Marino F, Penacino G, Sala A. Routine Y-STR typing in forensic casework. Forensic Sci Int 2001;118:131–5.
3. Dekairelle AF, Hoste B. Application of Y-STR pentaplex PCR (DYS19, DYS389I and II, DYS390, and DYS393) to sexual assault cases. Forensic Sci Int 2001;118:122–5.
4. Honda K, Roewer L, de Knijff P. Male DNA typing from 25-year-old vaginal swabs using Y chromosomal STR polymorphisms in retrial request case. J Forensic Sci 1999;44:868–72.
5. Jobling MA, Pandya A, Tyler-Smith C. The Y chromosome in forensic analysis and paternity testing. Int J Legal Med 1997;110:118–24.
6. Kayser M, Cagliá A, Corach D, Fretwell N, Gehrig C, Graziosi G, et al. Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 1997;110:125–33.
7. Krenke BE, Fulmer PM, Driftmier Miller K, Sprecher CJ. The PowerPlex® Y system. Profiles DNA 2003;6:6–9.
8. Krenke BE, Viculis L, Richard ML, Prinz M, Milne SC, Ladd C, et al. Validation of a male-specific, 12-locus fluorescent Short Tandem Repeat (STR) multiplex. Forensic Sci Int 2005;148:1–14.

9. Prinz M, Ishii A, Coleman A, Baum HJ, Shaler RC. Validation and casework application of a Y chromosome specific STR multiplex. Forensic Sci Int 2001;120:177–88.
10. Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, et al. Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. J Forensic Sci 2006;51(1):64–75.
11. Budowle B, Adamowicz M, Aranda X, Barna C, Chakraborty R, Eisenberg AJ, et al. Twelve short tandem repeat loci Y chromosome haplotypes: genetic analysis on populations residing in North America. Forensic Sci Int 2005;150(1):1–15.
12. Budowle B, Ge J, Chakraborty R. Basic principles for estimating the rarity of Y-STR haplotypes derived from forensic evidence. Eighteenth International Symposium on Human Identification, Oct 1–4, 2007; Hollywood, California. Madison, WI: Promega Corporation, 2007. http://www.promega.com/geneticidproc/ussymp18proc/oralpresentations.htm. Accessed June 16, 2009.
13. Budowle B, Allard MW, Wilson MR, Chakraborty R. Forensics and mitochondrial DNA: applications, debates, and foundations. Ann Rev Genomics Hum Genet 2003;4:119–41.
14. Kayser M, Krawczak M, Excoffier L, Dieltjes P, Corach D, Pascali V, et al. An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. Am J Hum Genet 2001;68:990–1018.
15. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution 1984;38:1358–70.
16. Weir BS Genetic data analysis II. Sunderland, MA: Sinauer Associates, 1996; 170–6.
17. Rogers AR, Harpending H. Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 1992;9:552–69.
18. Pereira L, Prata MJ, Amorim A. Mismatch distribution analysis of Y-STR haplotypes as a tool for the evaluation of identity-by-state proportions and significance of matches—the European picture. Forensic Sci Int 2002;130:147–55.
19. Roff DA, Bentzen P. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small populations. Mol Biol Evol 1989;6:539–45.
20. Sinha SK, Budowle B, Chakraborty R, Paunovic A, Guidry RD, Larsen C, et al. 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 2004;49:691–700.
21. Walsh B, Redd A, Hammer M. Joint match probabilities for Y chromosomal and autosomal markers. Forensic Sci Int 2008;174:234–8.
22. Beleza S, Alves C, Gonzalez-Neira A, Lareu M, Amorim A, Carracedo A, et al. Extending STR markers in Y chromosome haplotypes. Int J Legal Med 2003;117:27–33.
23. Jin L, Chakraborty R. Population substructure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics. Heredity 1995;74:274–85.
24. National Research Council. The evaluation of forensic DNA evidence. Washington, DC: National Academy Press, 1996.
25. Balding DJ, Nichols RA. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 1994;64:125–40.
26. Budowle B, Shea B, Niezgoda S, Chakraborty R. CODIS STR loci data from 41 sample populations. J Forensic Sci 2001;46(3):453–89.
27. Gross AM, Berdos P, Ballantyne J. Y-STR concordance study between Y-Plex5, Y-Plex6, Y-Plex12, PowerplexY, Y-Filer, MPI, and MPII. J Forensic Sci 2007;51(6):1423–8.

Additional information and reprint requests:
Bruce Budowle
Institute of Investigative Genetics
Department of Forensic and Investigative Genetics
University of North Texas Health Science Center at Ft. Worth
3500 Camp Bowie Blvd
Ft. Worth, TX 76107
E-mail: bbudowle@hsc.unt.edu