Y-STR PROFILES

Y-STR Guiding Principles

- The strength of Y-STR evidence depends on the probability of the evidence under alternative hypotheses about the contributors to the evidence.

- The probability of the evidence under an hypothesis of an unknown contributor can depend on the observation of known men not listed in the hypothesis.

- Account is taken of all the alleles detected in an evidential profile.

- Account is taken of Y-STR profiles being shaped by evolutionary forces.

Current US Status

In November 2018, SWGDAM issued "Notice to U.S. Forensic Laboratories on the status of the U.S. Y-STR Database."

This notice said "the U.S. Y-STR Database haplotypes have been permanently transferred to the Y-Chromosome Haplotype Reference Database (YHRD, http://yhrd.org) for continuance of usage, and the U.S. Y-STR Database will be decommissioned (scheduled for June 30, 2019). "

The Counting Method

An evidentiary Y-STR profile is queried against a database of profiles. The largest database is at https://YHRD.org. In February 2022, this database had 283,483 profiles for the Y17 Dataset and PowerPlex Y23 Kit. US forensic scientists can choose to query only US data within YHRD with 29,207 profiles for the Y17 Dataset and PowerPlex Y23 Kit.

The website reports the number of database profiles that match the evidence profile. For example, an actual PowerPlex 23 partial profile with alleles at 19 loci:

```
Found no match in 7,120 Haplotypes (95\% UCI: 1 in 2,377) in United States (African American).
Found no match in 4,034 Haplotypes (95\% UCI: 1 in 1,347) in United States (Asian).
Found no match in 8,488 Haplotypes (95\% UCI: 1 in 2,834) in United States (Caucasian).
Found no match in 6,024 Haplotypes (95\% UCI: 1 in 2,011) in United States (Hispanic).
Found no match in 3,541 Haplotypes (95\% UCI: 1 in 1,183) in United States (Native American).
Found no match in 29,207 Haplotypes (95\% UCI: 1 in 9,750) in United States (Overall).
```

Confidence Limits

For haplotype A, the database proportion $\tilde{p}_A$ is unbiased for the population proportion p_A. A confidence interval can be constructed, using properties of the binomial distribution. The $100(1-\alpha)\%$ upper confidence limit p_U when a database of size n has x copies of the target haplotype satisfies

$$\sum_{k=0}^{x} \binom{n}{k} p_U^k (1-p_U)^{n-k} \ \geq \ \alpha$$

If $x = 0$, then $(1-p_U)^n \geq \alpha$ or $p_U \leq 1 - \alpha^{1/n}$ and this is 0.0295 if $n = 100, \alpha = 0.05$. More generally $p_U \approx 3/n$ when $x = 0$ is the upper 95% confidence limit.

Comments on The Counting Method

The counting method number does not address any of our guiding principles. The method gives a number that

- ... is an estimate of the profile probability, and ignores the fact that the profile is known to be carried by (at least) one man.

- ... does not depend on any hypotheses about the source of the profile.

- ... may not be based on all the loci in the evidentiary profile.

- ... does not consider the genetic nature of Y-STR profiles.

Comments on The Counting Method

The number produced by the counting method is very dependent on the size of the database.

The number of loci in the previous example was reduced by one locus at a time: no matches were found even for an 8-locus partial profile, giving the same numerical results as for the original 19 locus-profile.

Y-chromosome Profiles

[Work of Taryn Hall, University of Washington.]

The Y-chromosome has several STR markers that are useful in forensic science. In one respect, the profiles are easier to interpret as each man has only one allele at an STR locus. Otherwise interpretation is made more complicated by the lack of recombination on the Y chromosome, meaning that alleles at different loci are not independent. Or are they?

We expect that mutations act independently at different loci and this may counter the lack of recombination to some extent.

Y-STR Databases

There are three public databases of Y-STR profiles:

- Y-Chromosome Haplotype Reference Database (YHRD) predecessor. Purps et al. FSI: Genetics 12:12-23 (2014)

- Human Genome Diversity Project (HGDP) Science 296:262-262 (2002)

- Data published by Xu et al. (XU) Mol Genet Genomics 290:1451-150 (2014)

Two-locus LD for Y-STR Loci

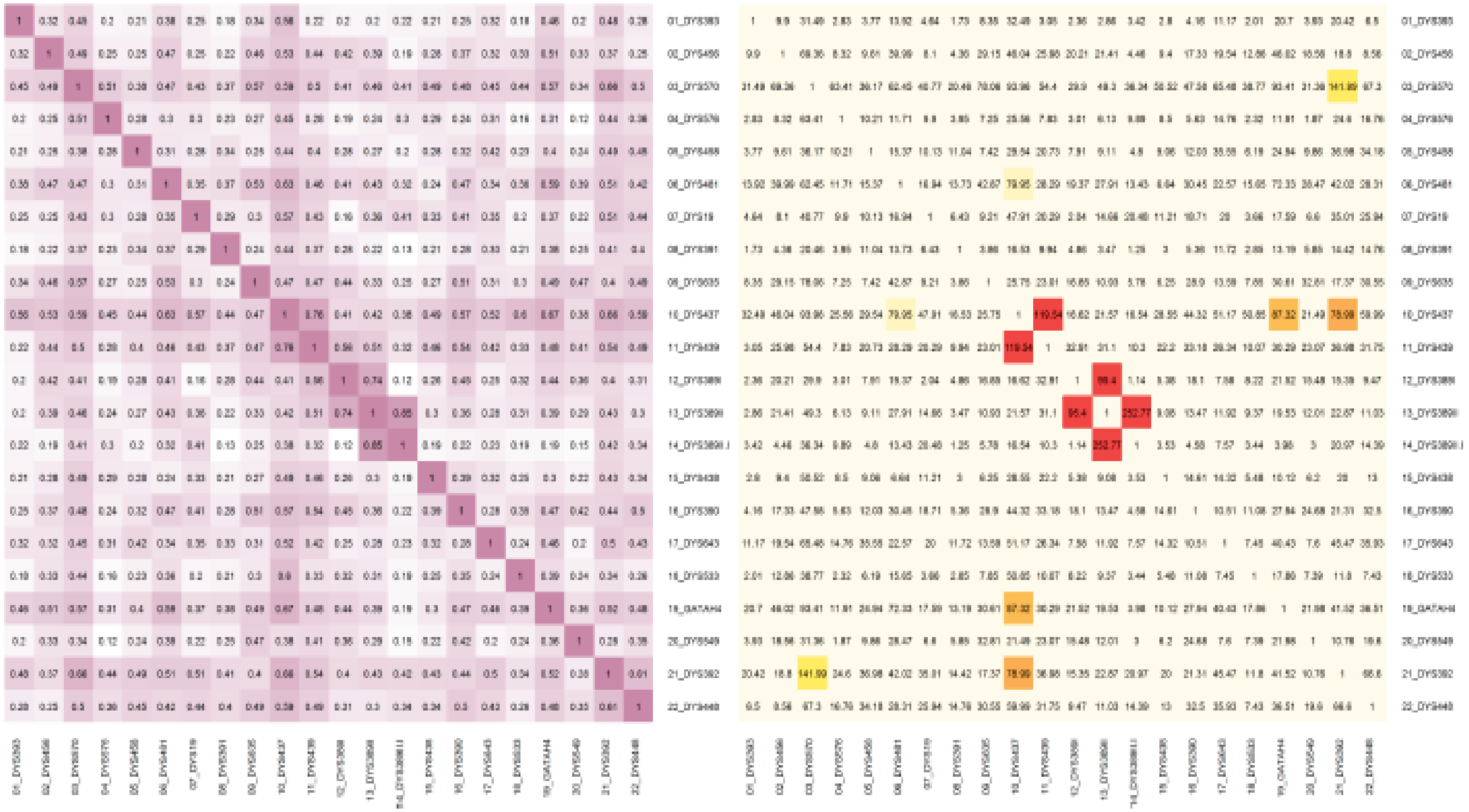


Figure D. Measures of linkage disequilbrium calculated between Y chromosome markers, European populations, Y-Chromosome Haplotype Reference Database.

Multi-locus Disequilibria: Entropy

It is difficult to describe associations among alleles at several loci. One approach is based on information theory.

For a locus with sample frequencies $\tilde{p}_u$ for alleles A_u the entropy is

$$H_A = -\sum_u \tilde{p}_u \ln(\tilde{p}_u)$$

For independent loci, entropies are additive: if haplotypes A_uB_v have sample frequencies $\tilde{P}_{uv}$ the two-locus entropy is

$$H_{AB} = -\sum_u \sum_v \tilde{P}_{uv} \ln(\tilde{P}_{uv}) = -\sum_u \sum_v \tilde{p}_u \tilde{p}_v [\ln(\tilde{p}_u) + \ln(\tilde{p}_v)] = H_A + H_B$$

so if $H_{AB} \neq H_A + H_B$ there is evidence of dependence. This extends to multiple loci.

Conditional Entropy

If the entropy for a multi-locus profile A is H_A then the conditional probability of another locus B, given A, is $H_{B|A} = H_{AB} - H_A$.

In performing meaningful calculations for Y-STR profiles, this suggests choosing a set of loci by an iterative procedure. First choose locus L_1 with the highest entropy. Then choose locus L_2 with the largest conditional entropy $H(L_2|L_1)$. Then choose L_3 with the highest conditional entropy with the haplotype L_1L_2, and so on.

Conditional Entropy: YHRD Data

Added	Entropy		
Marker	Single	Multi	Cond.
YS385ab	4.750	4.750	4.750
DYS481	2.962	6.972	2.222
DYS570	2.554	8.447	1.474
DYS576	2.493	9.318	0.871
DYS458	2.220	9.741	0.423
DYS389II	2.329	9.906	0.165
DYS549	1.719	9.999	0.093
DYS635	2.136	10.05	0.053
DYS19	2.112	10.08	0.028
DYS439	1.637	10.10	0.024
DYS533	1.433	10.11	0.010
DYS456	1.691	10.12	0.006
GATAH4	1.512	10.12	0.005
DYS393	1.654	10.13	0.003
DYS448	1.858	10.13	0.002
DYS643	2.456	10.13	0.002
DYS390	1.844	10.13	0.002
DYS391	1.058	10.13	0.002

This table shows that the most-discriminating loci may not contribute to the most-discriminating haplotypes. Furthermore, there is little additional discriminating power from Y-STR haplotypes beyond 10 loci.

The kappa Method

YHRD provides a number based on Brenner's κ, the proportion of profiles in a database (augmented by the evidentiary profile) that occur once only. The match probability for a profile not in the database is estimated as $(1-\kappa)/n$ where $(n-1)$ is the size of the database before augmentation. This is less than the counting method $1/n$.

Like the counting method, this does not make explicit use of the number of loci and is very dependent on the size of the database. Has not been extended to other pairs of hyotheses (e.g. mixtures or relatives).

The kappa Method

For the previous 19-locus PowerPlex 19-locus profile:

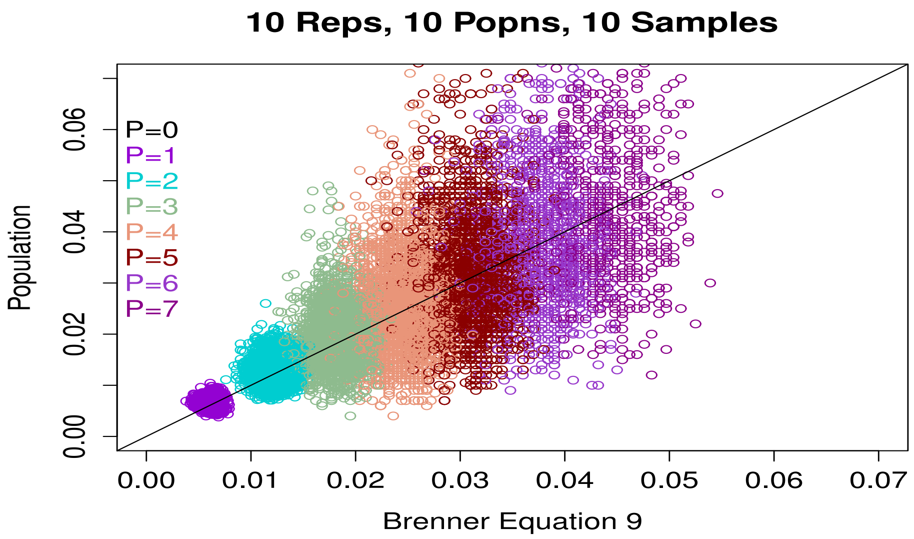
```
Expected Kappa

Approx. 1 match in 16,833 Haplotypes in United States (African American)
Approx. 1 match in 28,259 Haplotypes in United States (Asian)
Approx. 1 match in 20,765 Haplotypes in United States (Caucasian)
Approx. 1 match in 11,345 Haplotypes in United States (Hispanic)
Approx. 1 match in  8,024 Haplotypes in United States (Native American)
Approx. 1 match in 63,892 Haplotypes in United States (Overall)
```

These results remained the same as the number of loci was reduced from 19 to 8.

kappa Method

Here we compare kappa estimates for every profile in the augmented database with the proportion of profiles of that type in the population from which the sample was drawn. Kappa values appear better than the sample proportions for profiles not seen in the sample before it was augmented, as desired by Brenner. The quality decreases as the sample proportion of the evidentiary profile increases.



kappa Method

Brenner’s estimate uses only the number of times a profile occurs (“popularity”) in a database. It was not intended to do well for profiles that are seen more than a small number of times. Actual databases do have some profiles in high frequency. Here are the PPY23 haplotype counts for the Purps database.

Popul.	Count	Popul.	Count	Popul.	Count	Popul.	Count
1	9004	14	12	28	1	53	1
2	1254	15	4	29	1	54	1
3	416	16	5	30	2	57	1
4	196	17	2	33	2	58	3
5	105	18	7	35	1	61	1
6	85	19	4	36	1	62	1
7	50	20	3	37	2	68	1
8	41	21	3	38	1	91	1
9	34	22	2	41	3	118	1
10	24	24	4	42	3	126	1
11	28	25	4	43	2	170	1
12	16	26	1	45	1	242	1
13	9	27	2	48	2		

Genetic Model

A genetic approach can be built on the notion of identity by descent. For large numbers of loci, profiles of the same type are likely to match because they have a common ancestral haplotype. If θ_i is the probability of identity by descent of two random haplotypes in population i, the probability a random profile in population i is of type A given the evidentiary profile, also from population i, is that type is $\Pr(A|A)_i = \theta_i + (1-\theta_i)p_{Ai}$.

As profile proportions p_{Ai} become small the matching probabilities approach θ_i. These quantities, in turn, decrease as the number of loci increases. Kimura and Ohta (1968) showed that, for single-step mutations, STR loci have predicted θ values of $1/\sqrt{1+4N\mu}$. For L loci undergoing independent mutation we could replace μ by $1-(1-\mu)^L \approx L\mu$.

θ in Y-STR Literature

J Mol Evol 45:265-270 (1997)
Ann Hum Genet 64:395-412 (2000)
FSI 117:163-173 (2001)
FSI 149:99-107 (2005)
Legal Med 12:265-269 (2010)
Am J Phys Anthrop 143:591-600 (2010)
Legal Med 14:105-109 (2012)
Am J Hum Biol 25:313-317 (2013)
PLoS One 8:e64054 (2013)
FSI Genetics 19:255-262 (2015)
FSI Genetics Supp 5:E365-E367 (2015)

Interpreting Evidence

Two hypotheses for observed match between suspect and evidence:

H_P: Suspect is source of evidence.
H_D: Suspect is not source of evidence.

Then

$$\frac{\Pr(H_P|\text{Match})}{\Pr(H_D|\text{Match})} = \frac{\Pr(\text{Match}|H_P)}{\Pr(\text{Match}|H_D)} \times \frac{\Pr(H_P)}{\Pr(H_D)}$$

Interpreting Evidence

Suppose matching Y-STR profile is type A. The likelihood ratio reduces to

$$\begin{aligned}\frac{\mathrm{Pr}(\mathrm{Match}|H_P)}{\mathrm{Pr}(\mathrm{Match}|H_D)} &= \frac{\mathrm{Pr}(A|A,H_P)}{\mathrm{Pr}(A|A,H_D)} \\ &= \frac{1}{\mathrm{Pr}(A|A)}\end{aligned}$$

The Counting Method and the Discrete Laplace Method address $\mathrm{Pr}(A)$ rather than $\mathrm{Pr}(A|A)$.

Genetic Approaches

Population genetic theory for genetic markers (like Y-STR loci) subject to genetic drift and step-wise mutation was developed in the 1970's, for protein variants (allozymes) detected by gel electrophoresis. Migration distances on a gel depended on the net charge on a protein molecule, and charge changed in discrete units with amino acid substitutions in the protein sequence.

Moran (1975) showed that the distribution of allele frequencies "wandered" so that allele frequencies did not reach an equilibrium value and could not be predicted. Moments of the distribution, however, could be predicted. Ohta and Kimura (1975) showed that homozygosity at a single locus has an equilibrium value of $1/\sqrt{1+8N\mu}$ where N is the number of diploid individuals in a population, μ is the single-step mutation rate ($\mu/2$ in each direction), and the population undergoes random mating. This result reflects a balance between drift and mutation.

Genetic Approaches

The Ohta & Kimura result predicts that the logarithm of the probability two Y-STR haplotypes match is a linear function of the haplotype mutation rate. Matching decreases as mutation increases: this means that matching decreases as the number of loci increases.

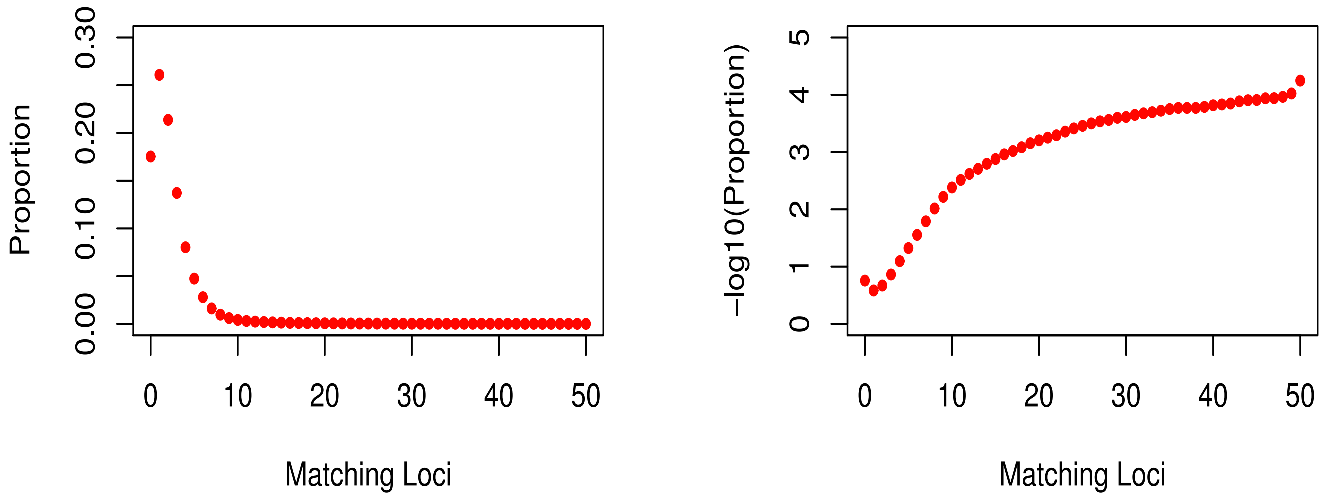
The matching proportion for haplotype pairs is essentially θ. It is greater than the profile proportions in a population but it does depend on the number of matching loci and very little on the size of a database.

The match probability for haplotype A is

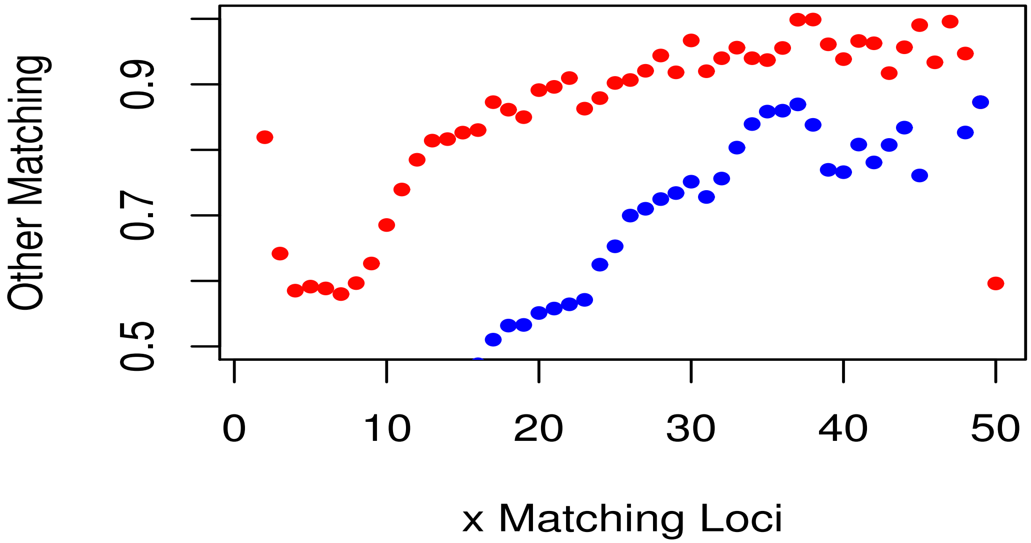
$$\Pr(A|A) \;=\; \theta + (1-\theta)\Pr(A) > \Pr(A)$$

Genetic Simulation

A population with a specified number of haplotypes each generation, independent mutations at each locus, no recombination among loci, and random choice of parent for each haplotype (Wright-Fisher model) can be simulated. The population proportions of pairs of 50-locus Y-STR profiles that matched at various numbers of loci, when $N = 10^4, \mu = 10^{-2}$ at each locus were:



Genetic Simulations



Plot of proportion of haplotype pairs matching at x loci that also match at $x+1$ loci (red) or $x+5$ loci (blue). Match probabilities are not independent among loci.

Within- and Between-population Matching

If the sample from population i has within-population matching proportion of $\tilde{M}_i$, the average over populations is:

$$\tilde{M}_W = \frac{1}{r}\sum_{i=1}^{r}\tilde{M}_i$$

If the sample between-population matching proportion for populations i and i' is $\tilde{M}_{ii'}$, the average over pairs of populations is:

$$\tilde{M}_B = \frac{1}{r(r-1)}\sum_{\substack{i=1 \\ i\neq i'}}^{r}\sum_{i'=1}^{r}\tilde{M}_{ij}$$

We estimate theta as $\beta_W = (\tilde{M}_W - \tilde{M}_B)/(1-\tilde{M}_B)$.

Use of θ-based Match Probabilities

If data are not available from the population of interest, but are available from a larger population (e.g. ethnic group), then the match-probability can be used with θ assigned or estimated from a set of subpopulations from the database population. The match probabilities use the database fequencies and β_W (for θ) and apply on average for any subpopulation.

θ for any subpopulation, or for the average over subpopulations, cannot be estimated from a single database. For example, a value for Native Americans cannot be estimated from a Native American database.

One-locus NIST Y-STR Estimates

Locus	$\tilde{M}_W$	$\tilde{M}_B$	$\hat{\beta}_W$
DYS19	0.32571062	0.24309148	0.10915340
DYS385a/b	0.07982377	0.04427420	0.03719640
DYS389I	0.41279418	0.38319082	0.04799436
DYS389II	0.26072434	0.23741323	0.03056847
DYS390	0.28981997	0.18813203	0.12525182
DYS391	0.52191425	0.48517426	0.07136392
DYS392	0.39961865	0.35168087	0.07394164
DYS393	0.50285122	0.48769253	0.02958906
DYS437	0.46400112	0.38595032	0.12710828
DYS438	0.36817530	0.23212655	0.17717601
DYS439	0.35507469	0.34990863	0.00794667
DYS448	0.30091326	0.22640195	0.09631787
DYS456	0.33444029	0.32578009	0.01284478
DYS458	0.21642167	0.19701369	0.02416976
DYS481	0.18867019	0.14121936	0.05525373
DYS533	0.39365769	0.37177174	0.03483757
DYS549	0.33976578	0.30691346	0.04740003
DYS570	0.21298105	0.20775666	0.00659442
DYS576	0.20955290	0.18125443	0.03456321
DYS635	0.27720127	0.20653182	0.08906400
DYS643	0.28394262	0.20058158	0.10427710
Y-GATA-H4	0.40667782	0.39899963	0.01277568

Multiple-locus US-YSTR Estimates

No. Loci	Added Locus	$\tilde{M}_W$	$\tilde{M}_B$	$\hat{\beta}_W$
1	DYS_438	0.37903281	0.27283973	0.14603806
2	DYS_392	0.22353526	0.10233258	0.13501958
3	DYS_19	0.11294942	0.05471374	0.06160639
4	DYS_390	0.05923470	0.02393636	0.03616398
5	DYS_643	0.04798422	0.02456341	0.02401059
6	YGATA_C4	0.03119210	0.01541060	0.01602851
7	DYS_533	0.01979150	0.00777794	0.01210774
8	DYS_393	0.01482393	0.00650531	0.00837309
9	DYS_456	0.01073170	0.00396487	0.00679377
10	DYS_438	0.00889934	0.00287761	0.00603912
11	DYS_549	0.00524369	0.00123093	0.00401770
12	DYS_481	0.00317518	0.00055413	0.00262250
13	DYS_389I	0.00240161	0.00031517	0.00208710
14	DYS_391	0.00200127	0.00017039	0.00183119
15	DYS_576	0.00106995	0.00005877	0.00101124
16	DYS_ 389II	0.00089896	0.00004205	0.00085695
17	DYS_385	0.00065020	0.00002729	0.00062293
18	YGATA_H4	0.00063652	0.00002427	0.00061227
19	DYS_448	0.00055062	0.00000713	0.00054349
20	DYS_458	0.00051100	0.00000423	0.00050677
21	DYS_570	0.00043010	0.00000423	0.00042587
22	DYS_439	0.00038612	0.00000423	0.00038189

YHRD Example

Theta-corrected Match Probability

Given a theta-value of 2.0 x 10-04 and a 95\% UCI of the combined Haplotype frequency of 1 in 8577 (no matches in 25694 Haplotypes at U.S. subpopulations without Native American), the corrected Match Probability is 1 in 3159.

Given a theta-value of 6.0 x 10-04 and a 95\% UCI of the combined Haplotype frequency of 1 in 9773 (no matches in 29275 Haplotypes at U.S. subpopulations with Native American), the corrected Match Probability is 1 in 1424.

Combining Y & Autosomal Match Probabilities

Although autosomal and Y STR loci are unlinked, matching at autosomal and Y loci are not independent (matching in one system implies some degree of kinship and therefore matching in the other system).

N	μ	$\widehat{\theta}_Y$	$\widehat{\theta}_{AY}$	$\widehat{\theta}_A$	$\widehat{\theta}_{A\|Y}$	$\widehat{\theta}_{A\|Y} - \widehat{\theta}_A$	Walsh	$\widehat{\theta}_{AY}/(\widehat{\theta}_A\widehat{\theta}_Y)$
10^4	10^{-2}	0.00040	0.00001270	0.00123	0.03143	0.03020	0.03025	25.5580
10^4	10^{-3}	0.00447	0.00007101	0.01233	0.01587	0.00355	0.00361	1.2878
10^4	10^{-4}	0.04343	0.00483898	0.11110	0.11142	0.00032	0.00038	1.0029
10^5	10^{-2}	0.00004	0.00000123	0.00012	0.03036	0.03024	0.03024	246.6184
10^5	10^{-3}	0.00045	0.00000217	0.00125	0.00483	0.00359	0.00359	3.8785
10^5	10^{-4}	0.00452	0.00005742	0.01234	0.01271	0.00036	0.00037	1.0293
10^6	10^{-2}	0.00000	0.00000012	0.00001	0.03025	0.03024	0.03024	2457.2222
10^6	10^{-3}	0.00004	0.00000017	0.00012	0.00372	0.00359	0.00359	29.7852
10^6	10^{-4}	0.00045	0.00000073	0.00125	0.00161	0.00037	0.00037	1.2928

Y-STR matching has little effect on autosomal coancestry when θ_A, θ_Y are large but the effects can be substantial when θ_A, θ_Y are small.

SWGDAM 2022 Guidelines

FAQ-26 in Supplementary Material:

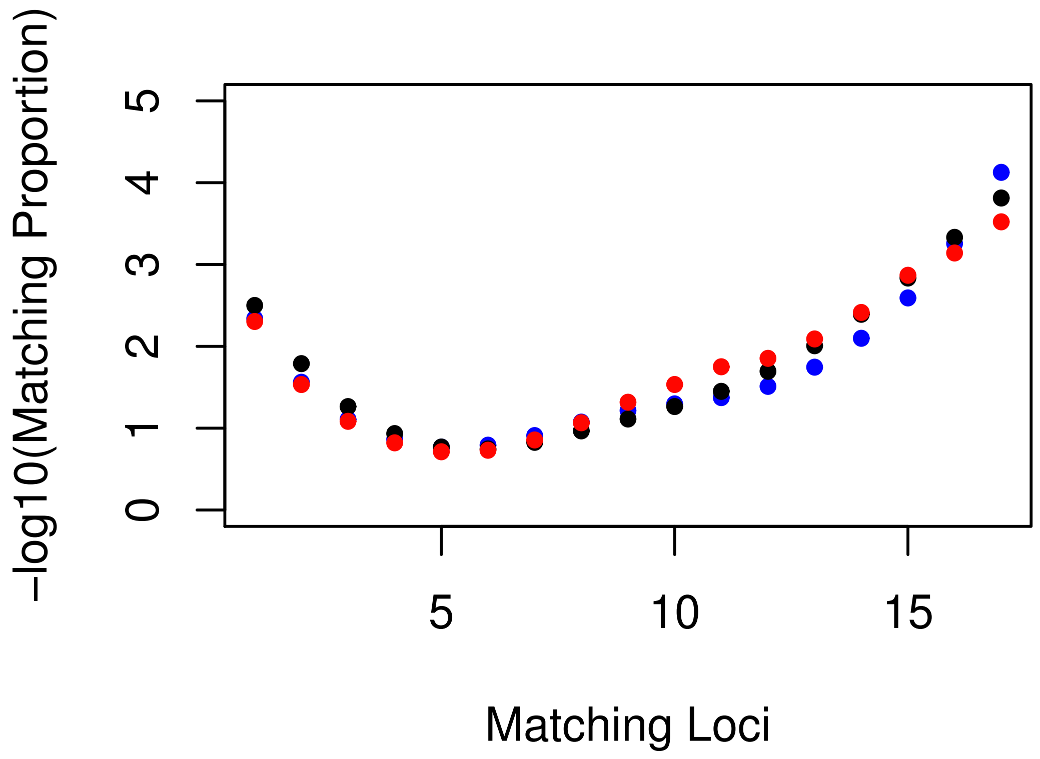
"There is not currently a publication for Y-STR θ values from a world-wide survey as there is for autosomal STRs (Buckleton et al. 2016). Such a publication is forthcoming. It is likely that values of 10^{-4} or less are appropriate for 15 or more Y-STR loci, and 10^{-5} or less are appropriate for 20 or more Y-STR loci."

Empirical Studies

European privacy laws prevent access to YHRD for numerical comparisons of different methods of assessing the evidential strength of Y-STR profiles. We have extracted data from over 150 publications, including about 100 of those cited by YHRD, to construct a database of over 50,000 profiles (1.25 billion pairs of profiles).

Data cleaning is in process. Data will allow estimation of θ for up to 22 loci for world-wide populations as was done for autosomal profiles by Buckleton et al. (FSI:Genetics, 2016).

Empirical Match Proportions



17-locus Y-STR matching proportions for 6,924 AFR profiles (black), 21,485 EUR profiles (blue) and 6,722 SAS profiles (red).