Y-STR PROFILES

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

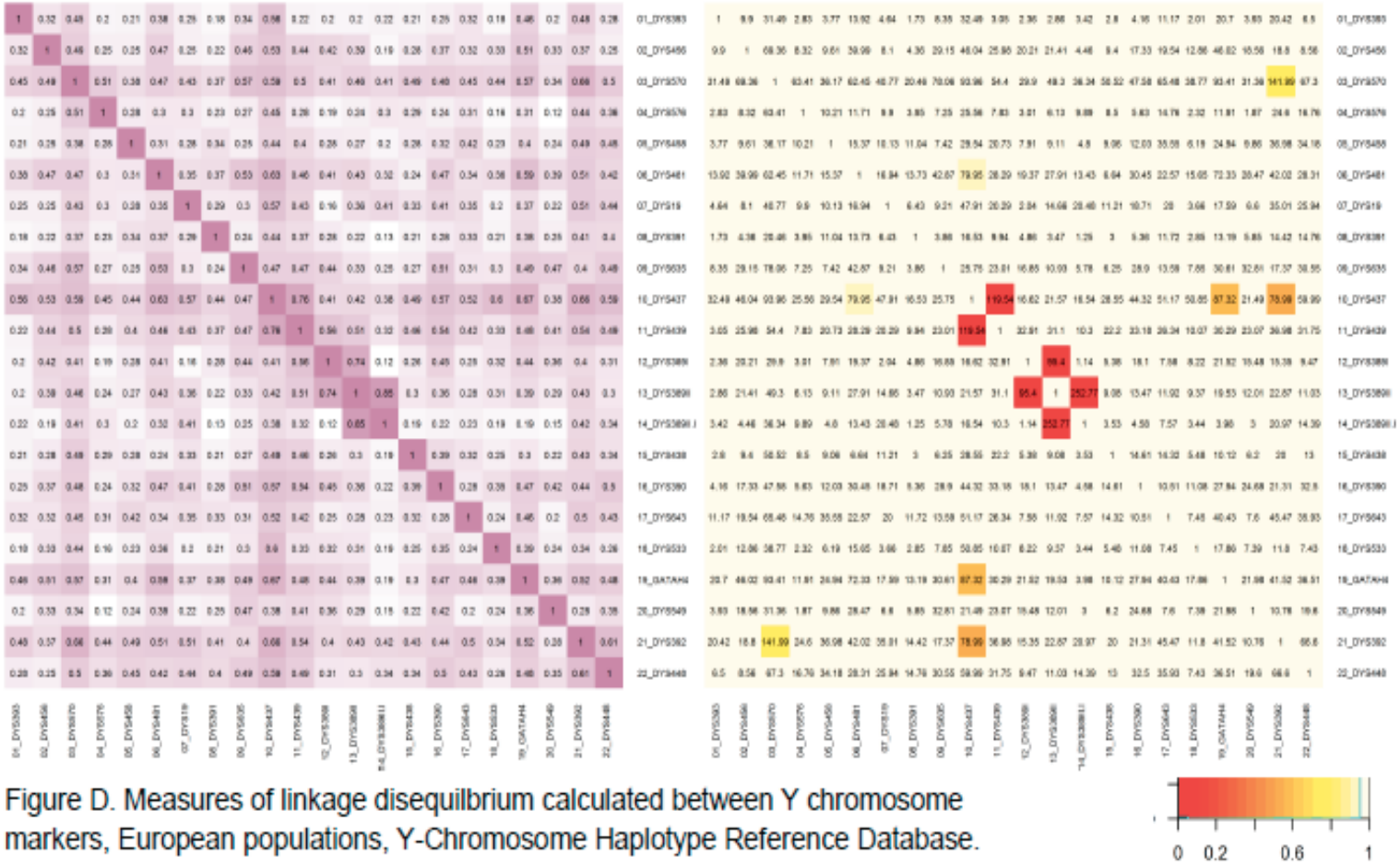


Figure D. Measures of linkage disequilbrium calculated between Y chromosome markers, European populations, Y-Chromosome Haplotype Reference Database.

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 $A_u B_v$ 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 Marker	Entropy		
	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.

Examples

YHRD											
Africa				Asia				Europe			
Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond	Marker order	Single	Combined	Cond
DYS385ab	4.750	4.750	4.750	DYS385ab	5.716	5.716	5.716	DYS385ab	4.100	4.100	4.100
DYS481	2.962	6.972	2.222	DYS570	2.769	8.115	2.399	DYS570	2.563	6.435	2.336
DYS570	2.554	8.447	1.474	DYS576	2.562	9.944	1.828	DYS576	2.381	8.475	2.040
DYS576	2.493	9.318	0.871	DYS458	2.598	10.998	1.055	DYS458	2.362	10.170	1.695
DYS458	2.220	9.741	0.423	DYS481	2.860	11.406	0.408	DYS481	2.842	11.360	1.190
DYS389II	2.329	9.906	0.165	DYS389II	2.319	11.582	0.176	DYS456	2.163	12.099	0.739
DYS549	1.719	9.999	0.093	DYS439	1.923	11.664	0.082	DYS389II	2.095	12.627	0.528
DYS635	2.136	10.052	0.053	DYS549	1.773	11.703	0.039	DYS549	1.792	12.964	0.337
DYS19	2.112	10.080	0.028	DYS635	2.465	11.728	0.024	DYS439	1.920	13.182	0.218
DYS439	1.637	10.104	0.024	GATAH4	1.727	11.744	0.016	DYS390	2.046	13.304	0.122
DYS533	1.433	10.114	0.010	DYS533	1.708	11.756	0.012	DYS635	2.001	13.372	0.068
DYS456	1.691	10.120	0.006	DYS456	1.775	11.765	0.009	GATAH4	1.569	13.420	0.049
GATAH4	1.512	10.124	0.005	DYS391	1.097	11.774	0.009	DYS391	1.279	13.454	0.033
DYS393	1.654	10.128	0.003	DYS448	2.299	11.778	0.005	DYS533	1.668	13.471	0.018
DYS448	1.858	10.130	0.002	DYS390	2.187	11.782	0.004	DYS19	1.837	13.484	0.013
DYS643	2.456	10.132	0.002	DYS437	1.212	11.786	0.003	DYS437	1.579	13.491	0.007
DYS390	1.844	10.134	0.002	DYS19	1.974	11.788	0.002	DYS393	1.218	13.497	0.006
DYS391	1.058	10.135	0.002	DYS643	2.267	11.790	0.002	DYS448	1.709	13.501	0.004
				DYS392	2.124	11.791	0.001	DYS643	1.885	13.504	0.003
				DYS393	1.754	11.791	0.001	DYS392	1.674	13.506	0.002
								DYS438	1.908	13.508	0.002
Max		10.284				11.859				13.581	
Selected set percent of max		0.986				0.994				0.995	

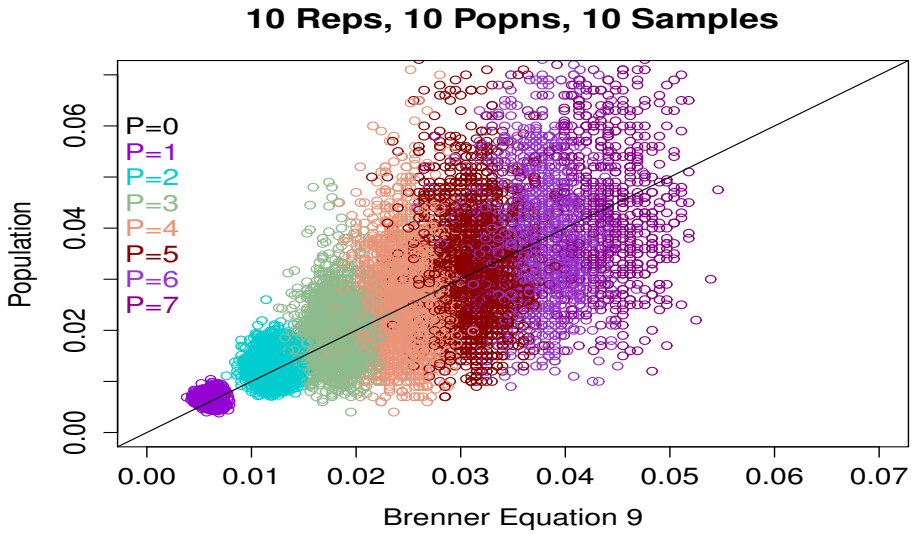
Brenner's Method

Brenner (2010) proposed the use of the proportion κ of profiles that occurred only once in a database that had been augmented by the evidentiary profile. His approach did not require a genetic model, although κ values can be predicted for some genetic models. The probability of a person taken randomly from a population would have the same profile as the evidentiary type when that type was not present in a sample of size $(n-1)$ (i.e. occurred once in the sample augmented by the evidentiary profile) was given by $(1-\kappa)/n$.

For profiles that occur p times in the augmented sample (those with "popularity" p), Brenner suggested a modification to $p(1-\kappa)/n$ that approaches the sample proportion $\tilde{p}$ when the proportion of singletons in the database becomes small.

Brenner's Method

Here we compare Brenner's 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. Brenner's 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.



Brenner's 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. In Table 1 we show 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 $1-(1-\mu)^L \approx L\mu$.

F_{ST} for 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)

Y-STR Matches

The chance of a random man having Y-STR haplotype A is written as p_A, the profile probability.

The chance that two men have haplotype A is written as P_{AA}.

The chance that a man has haplotype A given that another man has been seen to have that profile is $P_{A|A}$, the match probability. The three quantities are related by $P_{A|A} = P_{AA}/p_A$.

A major difficulty is that we generally do not have samples from the relevant (sub)population to give us estimates of p_A or P_{AA}. Instead we have a database of profiles that may represent a larger population.

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

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

A population genetic model introduces the quantity θ:

$$\Pr(AA) = \theta p_A + (1-\theta)p_A^2$$

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

where θ is the probability that two profiles are identical by descent.

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:

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

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

Use of Database Frequencies

If data (database) from the population of interest are available they should be used directly.

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.

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	$\tilde{M}_W$	$\tilde{M}_B$	$\widehat{\beta}_W$
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

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	$\hat{\theta}_Y$	$\hat{\theta}_{AY}$	$\hat{\theta}_A$	$\hat{\theta}_{A\|Y}$	$\hat{\theta}_{A\|Y} - \hat{\theta}_A$	Walsh	$\hat{\theta}_{AY}/(\hat{\theta}_A\hat{\theta}_Y)$
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.

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). "

An Example

YHRD RESULTS:

DYS456 15; DYS389I 12; DYS390 25; DYS389II 28; DYS458 16; DYS19 -
DYS393 13; DYS391 10 ; DYS439 - ; DYS635 22 ; DYS392 11; YGATAH4 11
DYS438 11; DYS448 19

Worldwide:
Found no match in 209,111 Haplotypes (95% UCI: 1 in 69,803).

An Example

USYSTR:

```
Found no match in 7,118 Haplotypes (95% UCI: 1 in 2,377)
  in United States (African American).
Found no match in 4,081 Haplotypes (95% UCI: 1 in 1,363)
  in United States (Asian).
Found no match in 8,483 Haplotypes (95% UCI: 1 in 2,832)
  in United States (Caucasian).
Found no match in 6,012 Haplotypes (95% UCI: 1 in 2,007)
  in United States (Hispanic).
Found no match in 3,581 Haplotypes (95% UCI: 1 in 1,196)
  in United States (Native American).
Found no match in 29,275 Haplotypes (95% UCI: 1 in 9,773)
  in United States (Overall).
```

An 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.

SWGDAM 2018 Notice

"The theta values provided in Appendix 1 of the 2014 SWGDAM Guidelines were calculated comparing average within-population match proportions to between population match proportions. As such, they address the issue of substructure between the major population groups within the total population, not between subpopulations within the major population groups. While they do not directly address within-population substructure, it was anticipated that they would provide conservative surrogates under the expectation that differences between major population groups would be larger than differences within major populations groups. In that spirit, U.S. Y-STR Database applied these theta estimates to population-specific match probabilities. YHRD instead, applies them only to the combined (Overall) database results for which these theta estimates are directly applicable."