Breaking News

The [US] President's Council of Advisors on Science and Technology (PCAST) is expected to issue a report tomorrow [Sept 1] afternoon which concludes that the following forensic disciplines lack sufficient scientific validation to be admissible as evidence in court:
Firearm Tool Mark with respect to marks on expended bullet casings
Shoe Print
Tire Tread
Multiple Source DNA
Bite Marks

It is anticipated that the report will urge prosecutors not to attempt to have such evidence admitted and judges not to admit such evidence. USDOJ was given a draft copy of the report earlier this week (under an embargo) and has strong objections.

How should we interpret Y-chromosome evidence?

Bruce Weir and Taryn Hall
University of Washington

Newton Institute for Mathematical Sciences
September 1, 2016

Y-STR Issues

- No recombination on Y chromosome suggests dependence among loci, offset by independent mutation events.
- Entropy/information theory metrics offer a way to characterize multi-locus dependencies.
- Matching evidence best described with the ratio of the probabilities of the evidence under alternative hypotheses.
- "Counting Method," "kappa method," "theta correction."
- "Discrete Laplace" method not covered here

Match Probability Calculations

- Purely statistical approach of estimating profile probability from database frequency not satisfactory.

- Method of Turing/Brenner uses proportion of singletons in a database and has good properties.

- Y-STR profiles are genetic and are shaped by evolutionary forces. Match probabilities depend on population and family structure.

YHRD Data

PPY23 data for 19,630 profiles available from Purps et al, FSI:Genetics 12(2014):12-23.

Five continental ancestry groups:

Group	No. of Profiles
African	1,294
Asian	3,976
European	12,585
Native American	558
Mixed American	1,217
Total	19,630

YHRD matching sets

Set size	No. of sets
1	18862
2	621
3	92
4	28
5	12
6	6
7	4
8	2
9	1
10	1
11	1

Capturing Multilocus Dependencies

Two-locus linkage disequilibrium is often mentioned in forensic genetic literature, but is not relevant for Y-STR profiles.

The real issue is the relationship, if any, between multi-locus profile probabilities and the product of single-locus probabilities.

The concept of entropy is useful in this context:
Caliebe et al, FSI:Genetics 15(2015):69-75;
Siegert et al, FSI:Genetics 16(2015):216-225.

Multiple Loci: Entropy

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

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

For a locus with only one allele, $\tilde{p}_1 = 1$, the entropy is zero. For a locus with m equally-frequent alleles the entropy is $\ln(m)$ and this increases with m.

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

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

so if $H_{AB} \neq H_A + H_B$ there is evidence of dependence.

Ordering loci by entropy

If the entropy for a multi-locus profile A is H_A then the conditional probability of another locus B, given A, is

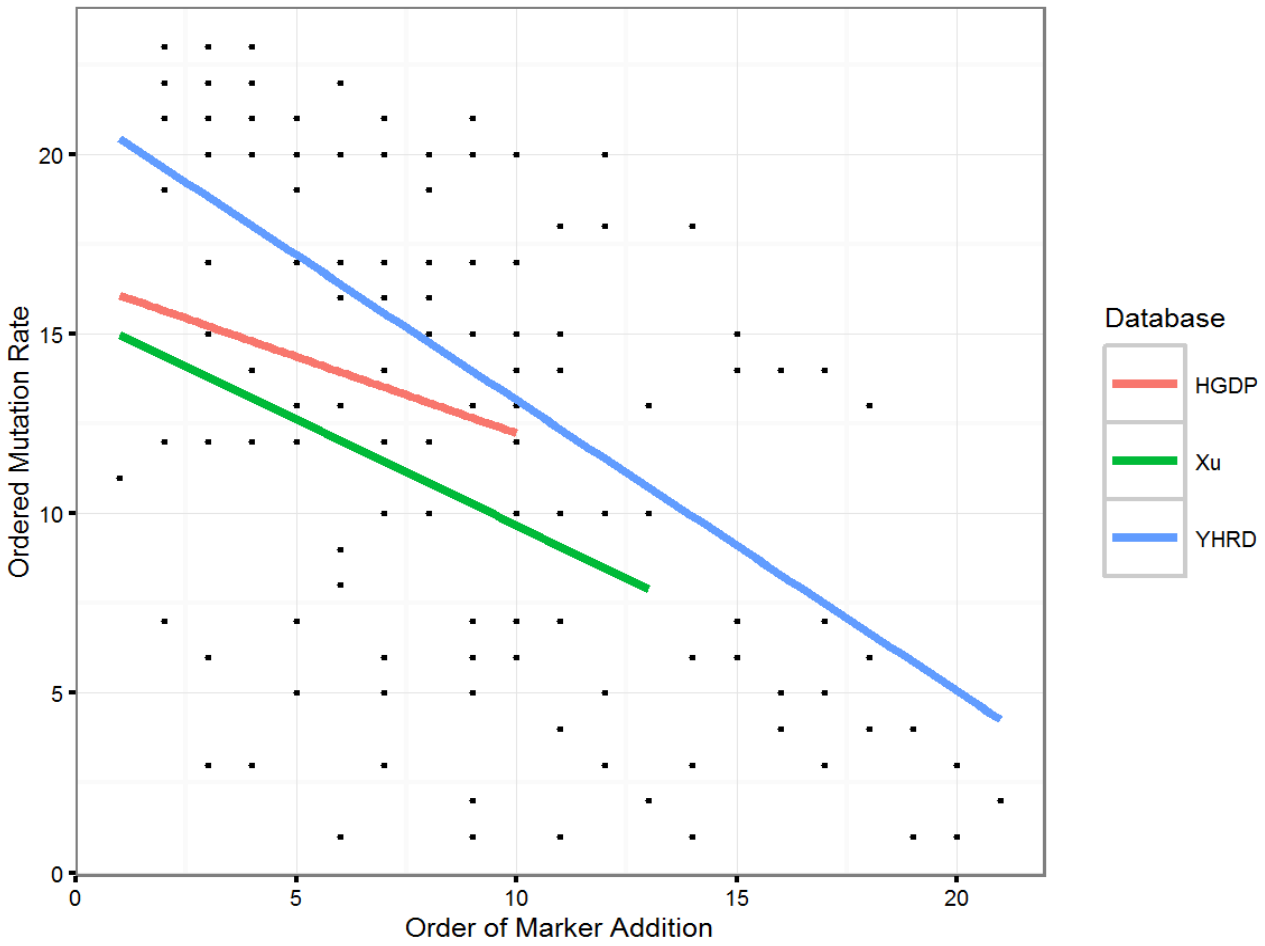
$$H_{B|A} = H_{AB} - H_A$$

This suggests choosing a B with the least dependence on A as the B with the largest conditional entropy $H_{B|A}$.

YHRD Entropies

Added Marker	Entropy		
	Single	Combined	Conditional
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

Entropy Decreases with Mutation Rate



Profile vs Match Probability

For profile (Y-STR haplotype) A:

- Profile probability $\Pr(A)$: the probability a randomly selected man has this profile. This is not of great forensic relevance.

- Match probability $\Pr(A|A)$: the probability a randomly selected man has this profile, given that the profile has already been seen. This is greater than the profile probability. Match probabilities provide the components of the LR: $\Pr(A|A, H_p)/\Pr(A|A, H_d)$.

Current Debate on Prosecutor's Fallacy

Current forensic scientist members of the US OSAC often believe that their experience and expert status allows them to commit the Prosector's Fallacy.

There also appears to be sympathy from other OSAC members for allowing statements about the probabilities of hypotheses given the evidence "if the expert accompanies them with data-driven statements of the accuracy of examiners who draw these conclusions."

US Supreme Court 2010

The US Supreme Court, in McDaniel v Brown, does not consider the Prosecutor's Fallacy to be a problem:

"... the [expert] Report provided no warrant for entirely excluding the DNA evidence or [the forensic scientist's] testimony from that court's consideration. The Report did not contest that the DNA evidence matched [the defendant]. That DNA evidence remains powerful inculpatory evidence even though the State concedes [the forensic scientist] overstated its probative value by failing to dispel the prosecutor's fallacy."

Counting Method

If profile A is seen n_A times in a sample of n profiles, then an unbiased estimate of the profile probability is n_A/n.

Most Y-STR profiles are not seen in a database (an L-locus profile has at least 10^L profiles). A 95% upper confidence limit for the population probability when $n_A = 0$ is close to $3/n$.

The problem: a 7-locus profile not seen in a database will also not be seen if it is part of a 17-locus profile. Changing n_A/n to $(n_A + 1)/(n + 1)$ or $(n_A + 2)/(n + 2)$ is still not estimating the match probability.

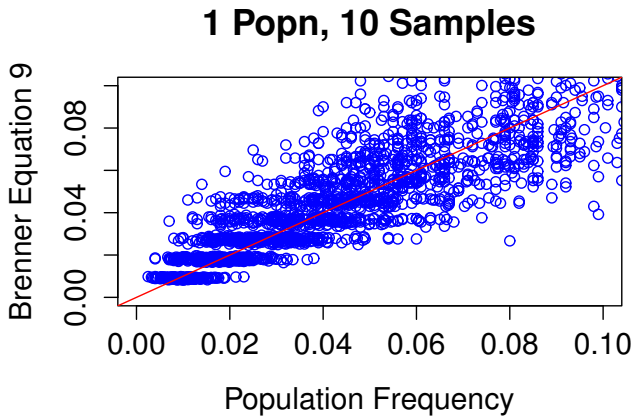
The kappa method

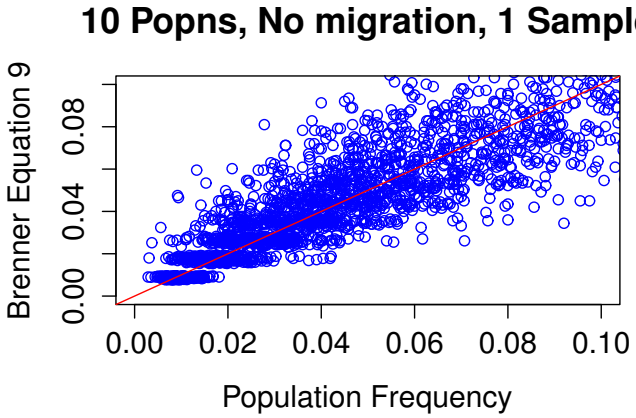
Brenner, FSI:Genetics 4(2010): 281-291, estimates the match probability of a profile not seen in a database of size $(n-1)$ as $(1-\kappa)/n$, where κ is the proportion of all n profiles that are singletons. This estimate is the same for every singleton, and it can be substantially less than the counting method estimate.

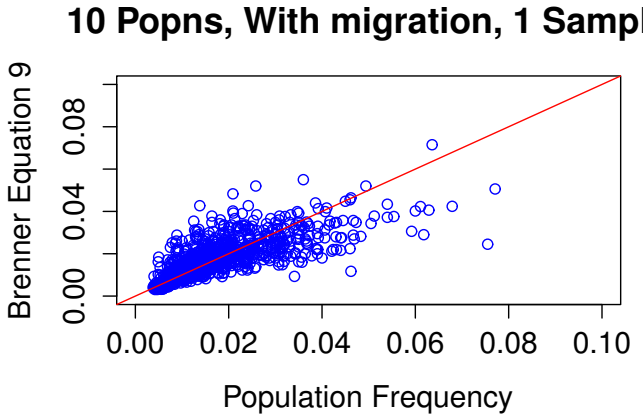
If the evidentiary profile has a "popularity" p in the augmented database, the estimate could be modified to $p(1-\kappa)/n$.

The problem remains that a previously-unseen profile will remain unseen if more loci are scored.

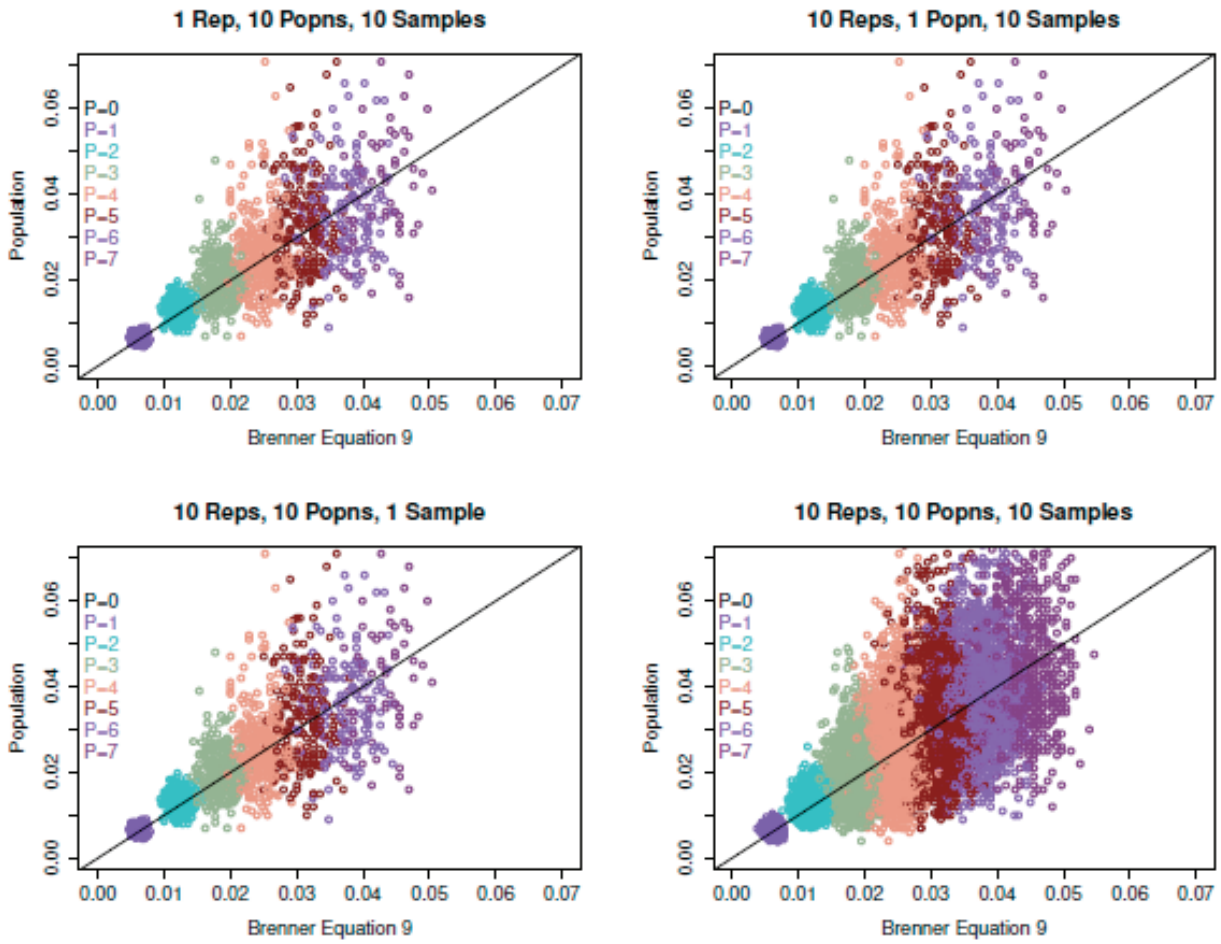
kappa method for every profile in augmented database







kappa method for every profile in population



YHRD Estimates

Counting methods for whole database:

$$\begin{aligned} 1/n &= 0.000,051 \\ 2/(n+2) &= 0.000,101 \\ 3/n &= 0.000,153 \end{aligned}$$

kappa method for whole database:

$$(1-\kappa)/n = 0.000,002$$

Theta correction method

If the suspect and offender (when different) are both in the same subpopulation i then the probability they both have haplotype A is p_{iA}^2. Taking averages over evolutionary replicates:

$$P_{AA_i} = \theta_i p_A + (1 - \theta_i) p_A^2$$

where p_A is the "total" haplotype frequency. The match probability within subpopulation i is

$$P_{A|A_i} = \theta_i + (1 - \theta_i) p_A$$

Average Match Probabilities

If the relevant subpopulation is not known, then the average over all subpopulations is:

$$P_{A|A_W} = \theta_W + (1 - \theta_W)p_u$$

since θ_W is the average of the θ_i's. Taking averages over all haplotypes gives the average within-subpopulation match probability

$$M_W = \theta_W + (1 - \theta_W)M_T$$

where $M_T = \sum_A p_A^2$ is the "total" match probability.

Y-STR Mutation

Population genetic theory predicts the value of θ for Y-STR haplotypes with step-wise mutation:

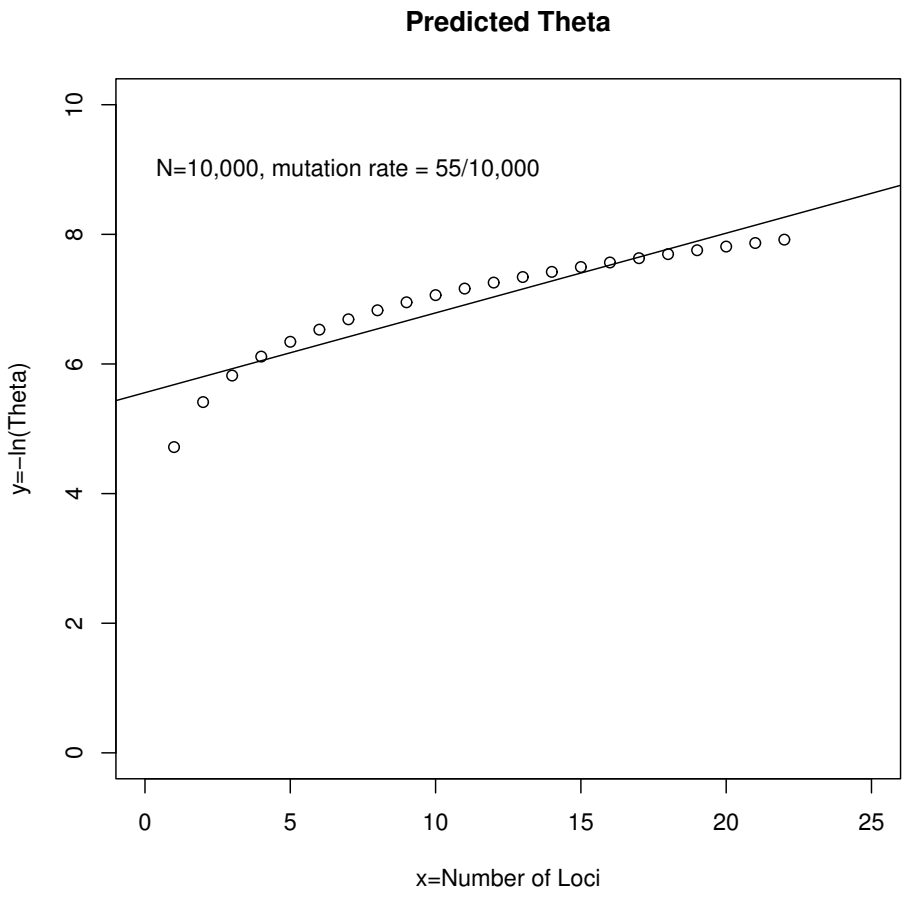
$$\theta \;=\; \frac{1}{\sqrt{1+8N\mu_Y}}$$

For L loci with independent and equal mutation rates μ: $\mu_Y = 1-(1-\mu)^L \approx L\mu$. On a log-scale

$$-\ln(\theta) \;=\; \ln(1+8N\mu_Y) \propto L\mu$$

This keeps on increasing with the number of loci.

Y-STR Mutation



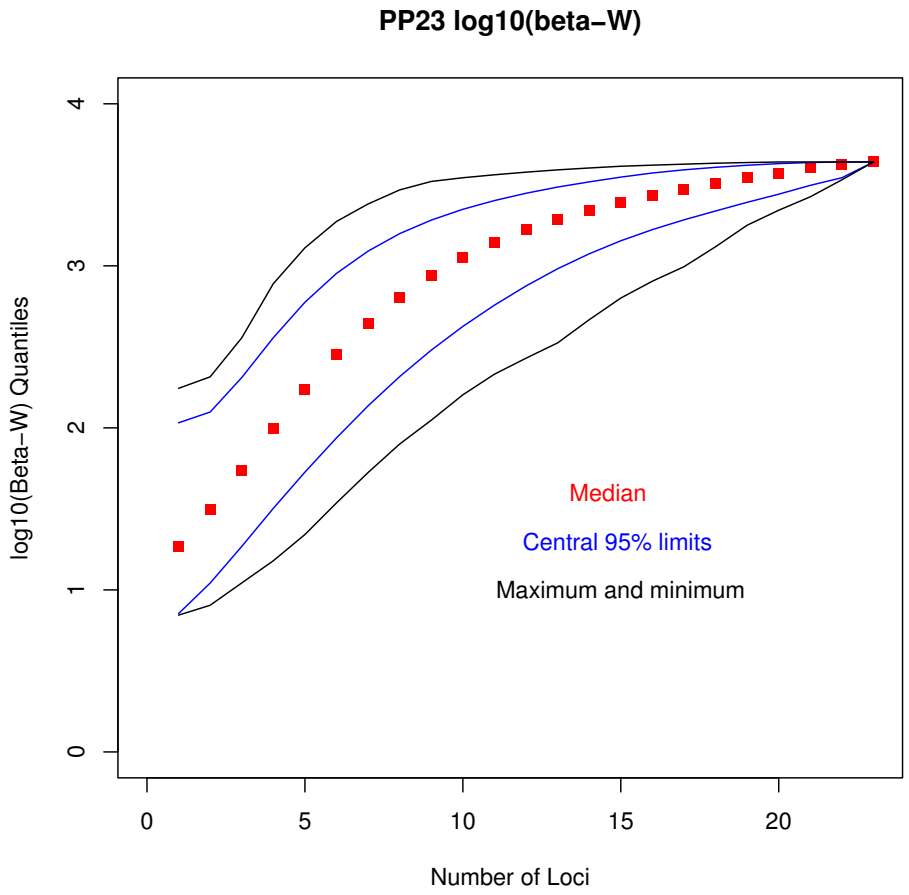
Estimating Theta

If there are data available from subpopulations, the average within-subpopulation θ_W is estimated by comparing actual haplotype-matching within subpopulations to actual haplotype matching between pairs of subpopulations.

NIST Y-STR Haplotype Estimates

No. Loci	Added Locus	Matching		$\widehat{\theta}_W$
		Within	Between	
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 $\hat{\theta}_W$ Values



YHRD Estimates

Counting methods for whole database:

$$\begin{aligned} 1/n &= 0.000,051 \\ 2/(n+2) &= 0.000,101 \\ 3/n &= 0.000,153 \end{aligned}$$

kappa method for whole database:

$$(1-\kappa)/n = 0.000,002$$

theta method for single subpopulation:

$$\begin{aligned} \theta_i &: 0.000,132 - 0.000,650 \\ \theta_W &= 0.000,381 \end{aligned}$$