Section 9:
Profile and Match Probabilities; CPI/CPE

Balding's Sampling Formula

if we have examined n alleles, and have seen n_A of type A, what is the probability the next allele is type A?

$$\Pr(A|n_A, n) = \frac{n_A\theta + (1-\theta)p_A}{1 + (n-1)\theta}$$

This implies the result for seeing a previously-unseen allele type B:

$$\Pr(B|n_B = 0, n) = \frac{(1-\theta)p_B}{1 + (n-1)\theta}$$

Examples of Balding's Formula

n	n_A	$\Pr(A\|n_A, n)$
0	0	p_A
1	0	$(1-\theta)p_A$
	1	$\theta + (1-\theta)p_A$
2	0	$(1-\theta)p_A/(1+\theta)$
	1	$[\theta + (1-\theta)p_A]/(1+\theta)$
	2	$[2\theta + (1-\theta)p_A]/(1+\theta)$
3	0	$(1-\theta)p_A/(1+2\theta)$
	1	$[\theta + (1-\theta)p_A]/(1+2\theta)$
	2	$[2\theta + (1-\theta)p_A]/(1+2\theta)$
	3	$[3\theta + (1-\theta)p_A]/(1+2\theta)$

Match Probability

Balding's formula lets the genotype match probabilities be found very easily from the third law of probability:

$$\begin{aligned}\Pr(AA|AA) &= \Pr(A|AA)\Pr(A|AAA) \\ &= \frac{2\theta+(1-\theta)p_A}{1+\theta}\times\frac{3\theta+(1-\theta)p_A}{1+2\theta}\end{aligned}$$

$$\begin{aligned}\Pr(AB|AB) &= \Pr(B|AB)\Pr(A|ABB)+\Pr(A|AB)\Pr(B|AAB) \\ &= \frac{2[\theta+(1-\theta)p_A][\theta+(1-\theta)p_B]}{(1+\theta)(1+2\theta)}\end{aligned}$$

Paternity Calculation

Balding's formula also lets paternity calculations be done very easily. In the case where the mother, child and alleged father are all homozygous AA, the paternity index is

$$\begin{aligned}
\text{LR} &= \frac{\Pr(\text{M, C, AF}|\text{AF is father})}{\Pr(\text{M.C.AF}|\text{AF not father})} \\
&= \frac{\Pr(\text{C}|\text{M, AF})\Pr(\text{M, AF})}{\Pr(\text{C}|\text{M})\Pr(\text{M, AF})} \\
&= \frac{1}{\Pr(A|AAAA)} \\
&= \frac{(1+3\theta)}{4\theta+(1-\theta)p_A}
\end{aligned}$$

The paternal allele is A, and four A alleles have been seen already.

Profile Probabilities

For a single autosmal locus, the probability a random person has genotypes AA or AB is written as $\Pr(AA)$ or $\Pr(BB)$.

If Hardy-Weinberg Equilibrium is assumed (NRC 4.1a,b):

$$\begin{aligned} \Pr(AA) &= p_A^2 \\ \Pr(AB) &= 2p_Ap_B \end{aligned}$$

If a random individual has probability F of being inbred, then the probabilities become (NRC 4.2a,b):

$$\begin{aligned} \Pr(AA) &= p_A^2 + p_A(1-p_A)F \\ \Pr(AB) &= 2p_Ap_B - 2p_Ap_BF \end{aligned}$$

The probability of a homozygote is greater than the HWE value, and the probability of a heterozygote is less than the HWE value. Here F is the pedigree-value that follows from the path-counting method and it is greater than zero. p_A, p_B are the total population allele frequencies as can be estimated from a database.

NRC Equation

The National Research Council recommended using

$$\begin{aligned} \Pr(AA) &= p_A^2 + p_A(1-p_A)F \\ \Pr(AB) &= 2p_Ap_B \end{aligned}$$

in the interest of being conservative.

Single-allele Profile

An STR profile may show only one allele A at a locus. The true genotype may be homozygous AA or heterozygous AB where allele B is not detected or not called. The HWE probability for the profile allele is

$$\begin{aligned}
\Pr(A) &= \Pr(AA) + \sum_{B \neq A} \Pr(AB) \\
&= p_A^2 + \sum_{B \neq A} 2 p_A p_B \\
&= p_A^2 + 2p_A(1 - p_A) \\
&= 2p_A - p_A^2
\end{aligned}$$

The "2p" rule approximates this by the conservative value $2p_A$ (NRC Page 105).

Single-allele Profile

For inbred individuals, the value would be

$$\begin{aligned}
\Pr(A) &= \Pr(AA) + \sum_{B \neq A} \Pr(AB) \\
&= p_A^2 + p_A(1-p_A)F + \sum_{B \neq A} 2p_A p_B(1-F) \\
&= p_A^2 + p_A(1-p_A)F + 2p_A(1-p_A) - 2p_A(1-p_A)F \\
&= 2p_A - p_A^2 - p_A(1-p_A)F
\end{aligned}$$

which also has $2p_A$ as a (conservative) upper bound.

Match Probability for Relatives

For unilineal relatives, $k_2 = 0, k_1 > 0 : \theta = k_1/4$: (NRC 4.8a,b)

$$\begin{aligned}
\Pr(AA|AA) &= \frac{\Pr(AAAA)}{\Pr(AA)} = \frac{k_2 p_A^2 + k_1 p_A^3 + k_0 p_A^4}{p_A^2} \\
&= 4\theta p_A + (1 - 4\theta) p_A^2 \\
&= p_A^2 + 4 p_A (1 - p_A)\theta
\end{aligned}$$

$$\begin{aligned}
\Pr(AB|AB) &= \frac{\Pr(ABAB)}{\Pr(AB)} = \frac{2k_2 p_A p_B + k_1 p_A p_B (p_A + p_B) + 4k_0 p_A^2 p_B^2}{2 p_A p_B} \\
&= 2\theta(p_A + p_B) + 2(1 - 4\theta) p_A p_B \\
&= 2 p_A p_B + 2(p_A + p_B - 4 p_A p_B)\theta
\end{aligned}$$

Match Probability for Full Sibs

For full sibs, $k_2 = 1/4, k_1 = 1/2, k_0 = 1/4$: (NRC 4.9a,b)

$$\begin{aligned}
\Pr(AA|AA) &= \frac{k_2 p_A^2 + k_1 p_A^3 + k_0 p_A^4}{p_A^2} \\
&= \frac{1}{4}(1 + 2p_A + p_A^2) \\
\Pr(AB|AB) &= \frac{2k_2 p_A p_B + k_1 p_A p_B (p_A + p_B) + 4k_0 p_A^2 p_B^2}{2p_A p_B} \\
&= \frac{1}{4}(1 + p_A + p_B + 2p_A p_B)
\end{aligned}$$

Probability of Exclusion

The Principles of Evidence Interpretation, leading to the likelihood ratio for the probabilities of the evidence under alternative hypotheses, allow all situations to be addressed. The prosecution and defense perspectives are explicitly taken into account.

The probability of exclusion considers only the evidence profile and ignores prosecution and defense perspective. It does not inform the court.

For a single-contributor stain with genotype AA, anyone not of that type is excluded. The probability of exclusion is $(1 - p_A^2)$. For type AB the probability is $(1 - 2p_A p_B)$. If many loci are typed, the combined probability of exclusion is the probability a person is excluded for at least one locus - i.e. one minus the probability of no exclusions:

$$\text{CPI} = 1 - \prod_{\text{loci}\, l} [1 - \Pr(\text{Excluded at locus } l)]$$

Exclusion for Mixtures

The Probability of Exclusion understates the strength of mixture evidence. If a crime stain is observed to have four alleles A, B, C, D at a locus the probability of exclusion is $\Pr(AA) + \Pr(BB) + \Pr(CC) + \Pr(DD) + \Pr(AB) + \Pr(AC) + \Pr(AD) + \Pr(BS) + \Pr(BD) + \Pr(CD)$. This is $(p_A + p_B + p_C + p_D)^2$.

If the prosecution says the evidence represents the victim of type AB and the defendant of type CD then the evidence has probability of 1.

If the defense says the evidence (e.g. bedding) is not associated with either the victim or the defendant then the evidence has probability $\Pr(AB, CD) + \Pr(AC, BD) + \Pr(AD, BC) = 24 p_A p_B p_C p_D$.

If all allele frequencies are 0.1, the PE is $0.4^2 = 0.16$ ("1 in 6") and the LR is $1/0.0024 = 416$.

Will match probabilities keep decreasing?

Table 2 The expected match probability (EMP) of the kits/panels.[1]

Panel (number of STR loci)	Unrelated		Parent/child
	Fst = 0[2]	Fst = 0.01	Fst = 0
New FBI core (24)[3]	6.28×10^{-30}	5.12×10^{-29}	3.63×10^{-18}
New FBI core section A (20)[3]	9.54×10^{-25}	4.77×10^{-24}	3.83×10^{-15}
13-loci CODIS core (13)	2.34×10^{-15}	5.83×10^{-15}	1.74×10^{-9}
Identifiler (15)	5.93×10^{-18}	1.73×10^{-17}	5.04×10^{-11}
PowerPlex16 (15)	2.43×10^{-18}	7.48×10^{-18}	3.06×10^{-11}
NGM[4] (15)	1.12×10^{-19}	4.15×10^{-19}	5.68×10^{-12}

[1]Caucasian population data were used.

Ge et al, Investigative Genetics 3:1-14, 2012.

Will match probabilities keep decreasing?

How do these match probabilities address the observation of Donnelly:

> "after the observation of matches at some loci, it is relatively much more likely that the individuals involved are related (precisely because matches between unrelated individuals are unusual) in which case matches observed at subsequent loci will be less surprising. That is, knowledge of matches at some loci will increase the chances of matches at subsequent loci, in contrast to the independence assumption."

Donnelly P. 1995. Heredity 75:26-64.

Are match probabilities independent over loci?

Is the problem that we keep on multiplying match probabilities over loci under the assumption they are independent? Can we even test that assumption for 10 or more loci?

Or is our standard "random match probability" not the appropriate statistic to be reporting in casework? Is it actually appropriate to report statements such as

> The approximate incidence of this profile is 1 in 810 quintillion Caucasians, 1 in 4.9 sextillion African Americans and 1 in 410 quadrillion Hispanics.

Putting "match" back in "match probability"

Let's reserve "match" for a statement we make about two profiles and take "match probability" to mean the probability that *two profiles match*. This requires calculations about *pairs of profiles*.

If the source of an evidence profile is unknown (e.g. is not the person of interest), then the match probability is the probability this unknown person has the profile *already seen in the POI.* No two profiles are truly independent, and their dependence affects match probabilities across loci.

Likelihood ratios use match probabilities

As with many other issues on forensic genetics, the issue of multi-locus match probability dependencies is best addressed by comparing the probabilities of the evidence under alternative propositions:

H_p: the person of interest is the source of the evidence DNA profile.
H_d: an unknown person is the source of the evidence DNA profile.

Write the profiles of the POI and the source of the evidence as G_s and G_c. The evidence is the pair of profiles G_c, G_c.

Likelihood ratios use match probabilities

The likelihood ratio is

$$\begin{aligned} \text{LR} &= \frac{\Pr(E|H_p)}{\Pr(E|H_d)} \\ &= \frac{\Pr(G_c, G_s|H_p)}{\Pr(G_c, G_s|H_d)} \\ &= \frac{1}{\Pr(G_c|G_s, H_d)} \\ &= \frac{1}{\text{Match probability}} \end{aligned}$$

providing $G_c = G_s$ under H_p. The match probability is the chance an unknown person has the evidence profile given that the POI has the profile: this is not the profile probability.

Special Cases: Use of Sample Allele Frequencies

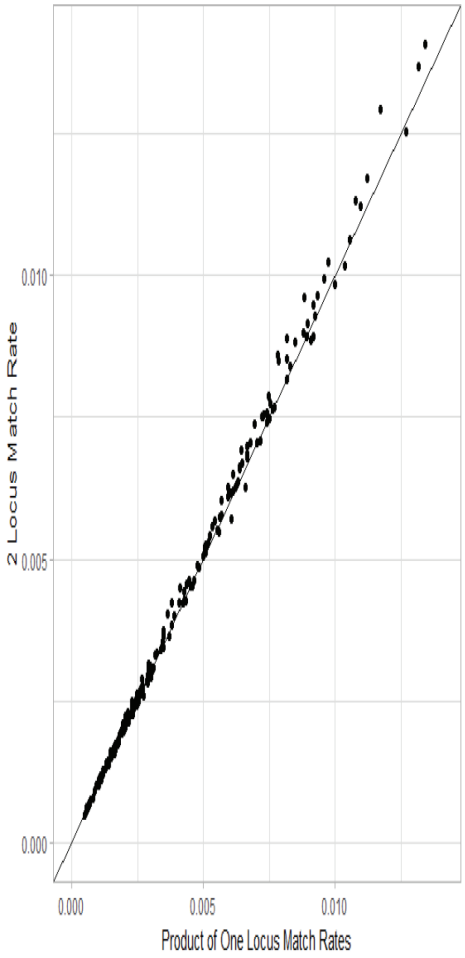
The match probability is usually estimated using allele frequencies from a database representing some broad class of people, such as "Caucasian" or "African American" or "Hispanic."

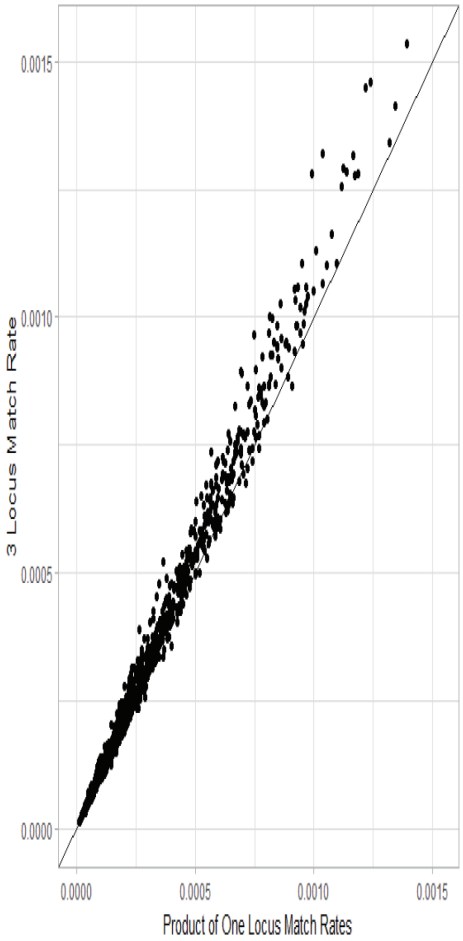
The population relevant for a particular crime may be a narrower class of people. There is population structure. If p are the allele frequencies in the database, the match probabilities are estimated as

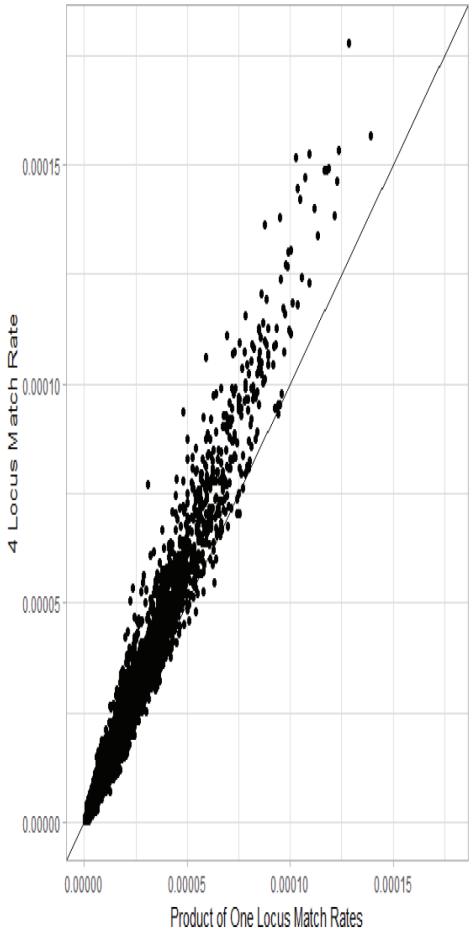
$$\begin{aligned}
\Pr(AA|AA) &= \frac{[3\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)} \\
\Pr(AB|AB) &= \frac{2[\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)}
\end{aligned}$$

Can these be multiplied over loci?

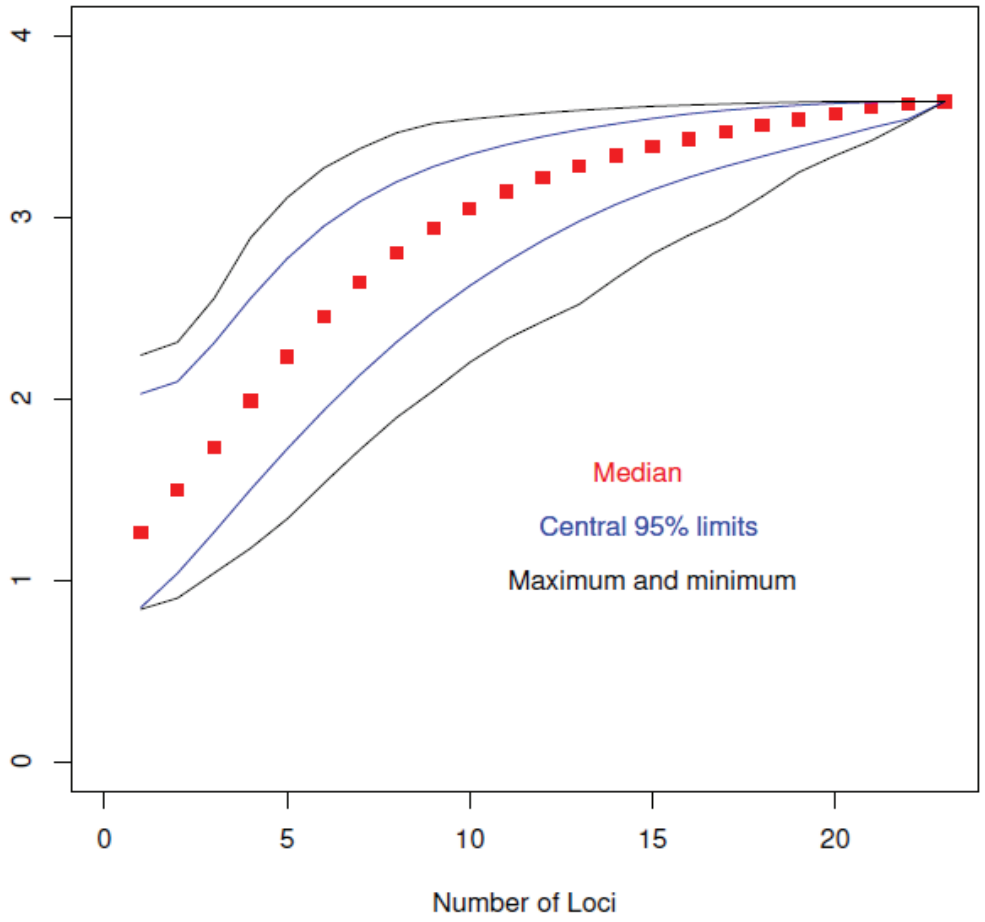
Empirical dependencies: 2849 20-locus profiles







Empirical dependencies: Y-STR profiles



Plot of negative log of match probabilities for Purps et al. database.

Theoretical dependencies: No mutation

The probability an individual is homozygous $AABB$ at loci **A,B** is

$$\begin{aligned} \Pr(AABB) &= \Pr(AA)\Pr(BB) + p_A(1-p_A)p_B(1-p_B)\eta \\ &\geq \Pr(AA)\Pr(BB) \end{aligned}$$

where η is the *identity disequilibrium*. It can non-zero even for pairs of loci that are unlinked and/or in linkage equilibrium.

> Sampling among parents or gametes and/or the inclusion of random elements in the uniting gametes leads to a correlation in identity by descent even between unlinked loci because genes at both loci are of necessity included in each gamete.

Weir & Cockerham, Genetics 63:711-742, 1969.

Theoretical dependencies: Mutation

Ratios of two-locus genotypic match probabilities to products of one-locus probabilities for unlinked loci with equal mutation rates

μ	$N = 10,000$	$N = 100,000$
1×10^{-1}	2.3535×10^{4}	2.3332×10^{6}
2.5×10^{-2}	1.4097×10^{2}	1.2559×10^{4}
1×10^{-2}	7.0211	3.6675×10^{2}
5×10^{-3}	1.8802	2.9326×10^{1}
1×10^{-3}	1.0276	1.3020
1×10^{-4}	1.0053	1.0029
1×10^{-5}	1.0044	1.0050
1×10^{-6}	1.0006	1.0044
1×10^{-7}	1.0001	1.0006
1×10^{-8}	1.0000	1.0001

Laurie CA, Weir BS. 2003. Theoretical Population Biology 63:207-219.

Theoretical dependencies: Mutation

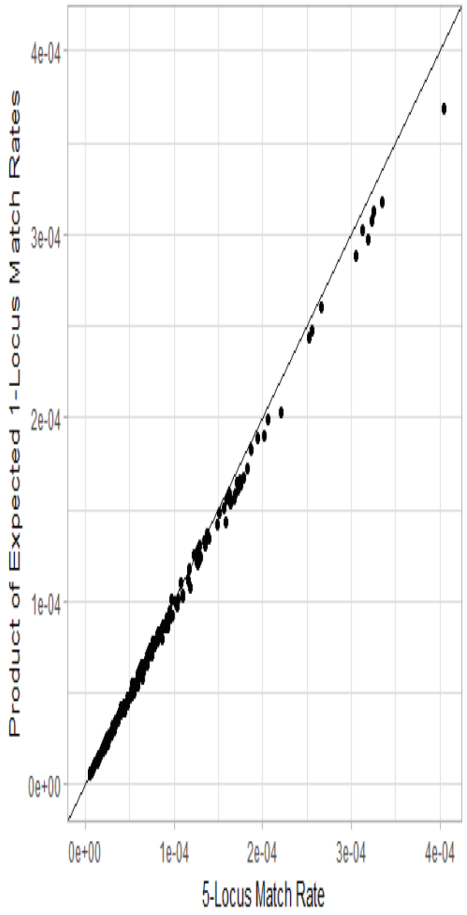
"Between-locus dependencies in finite populations can lead to under-estimates of genotypic match probabilities when using the product rule, even for unlinked loci.

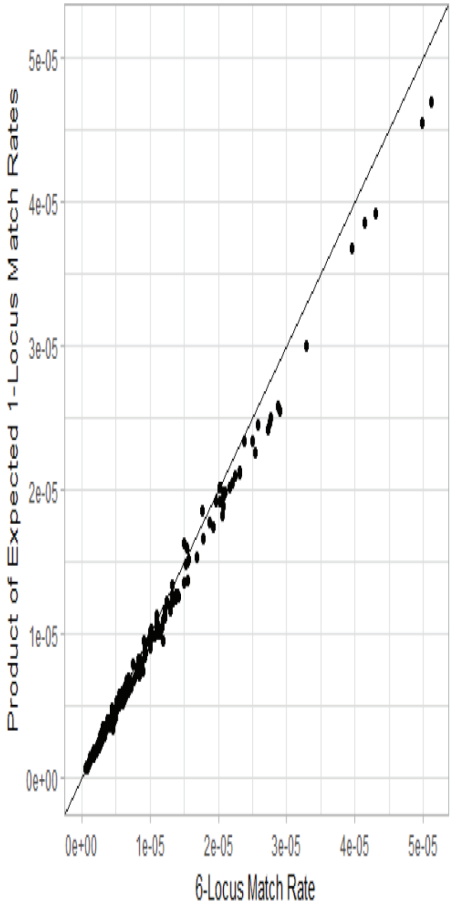
The three-locus ratio is greater than one and is greater than the corresponding two-locus ratio for large mutation rates. These results provide evidence that between-locus dependency effects are magnified when considering more loci.

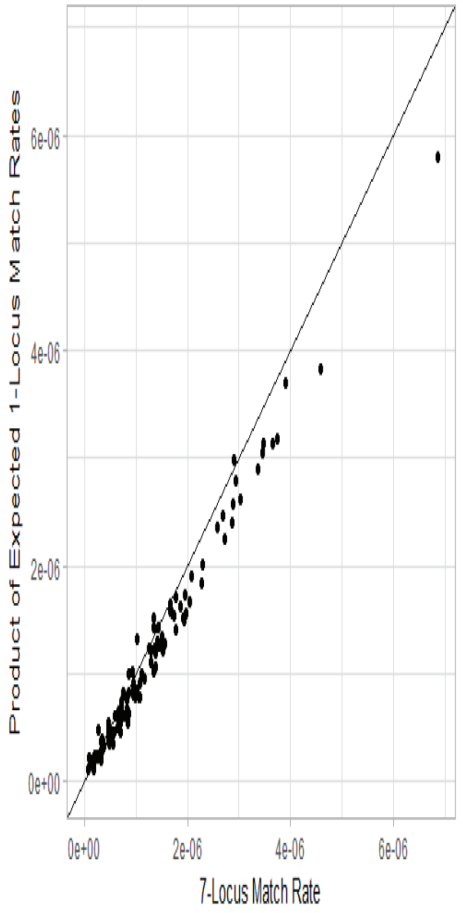
High mutation rates mean that specific mutants are likely to be recent and rare. Hence, if two individuals share alleles at one locus, they are more likely to be related through recent pedigree, and hence more likely to share alleles at a second locus."

Laurie CA, Weir BS. 2003. Theoretical Population Biology 63:207-219, 2003.

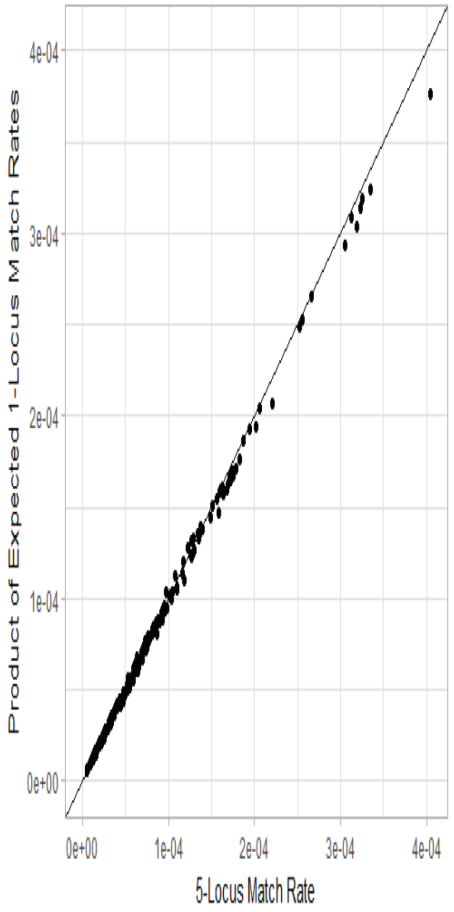
One population simulated data: $\theta = 0$

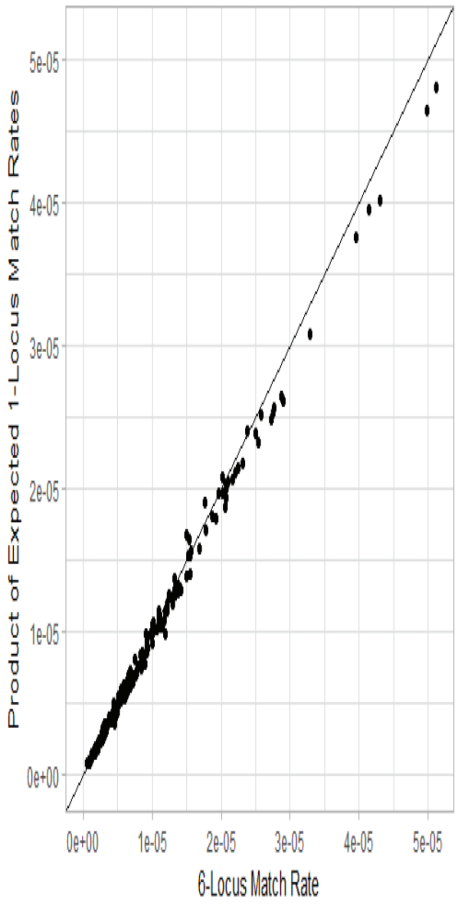


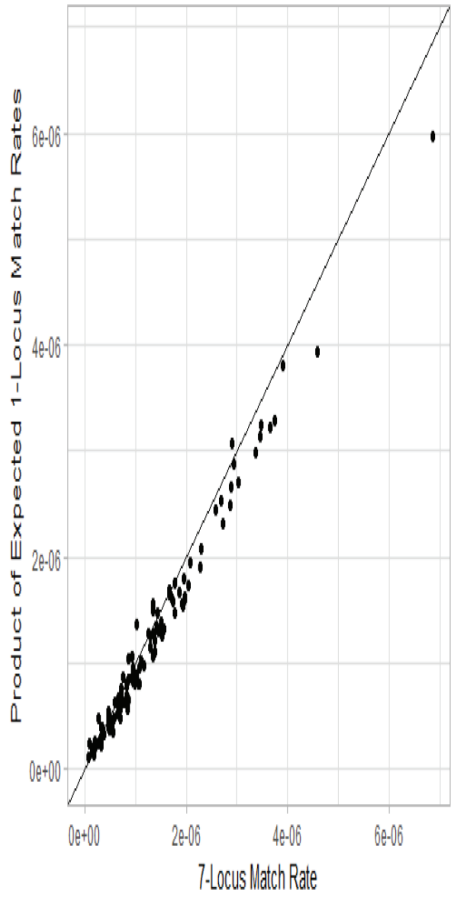




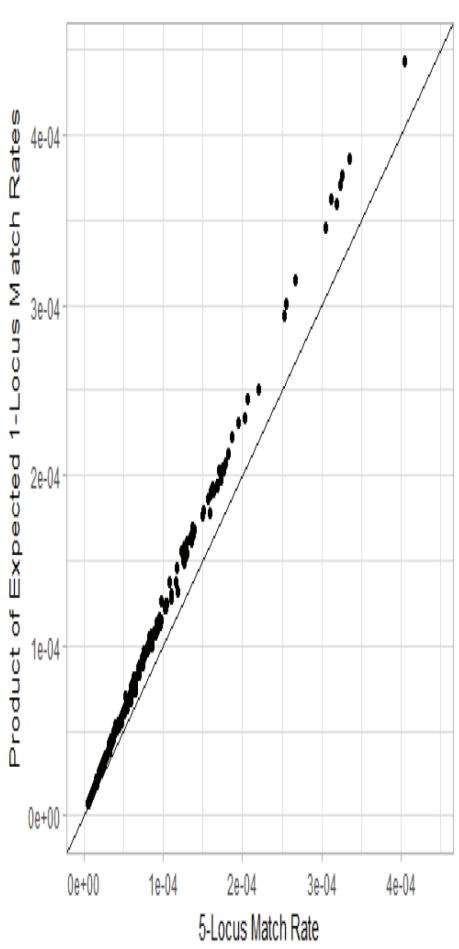
One population simulated data: $\theta = 0.001$

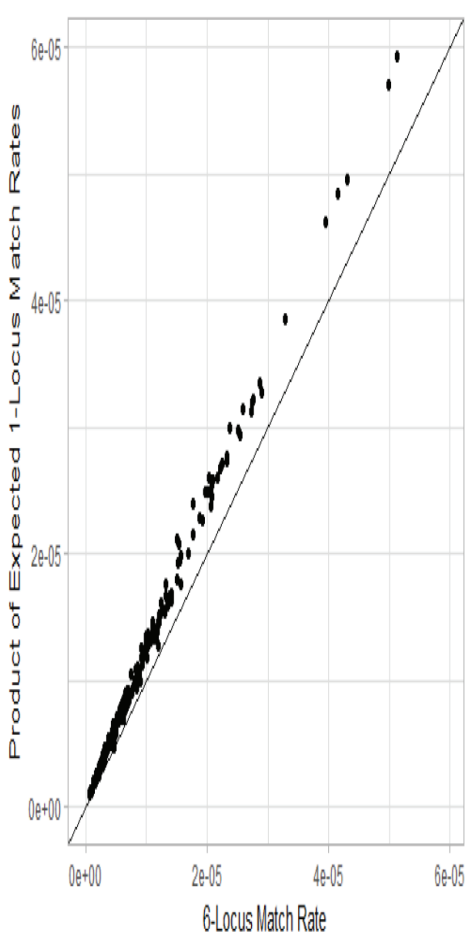


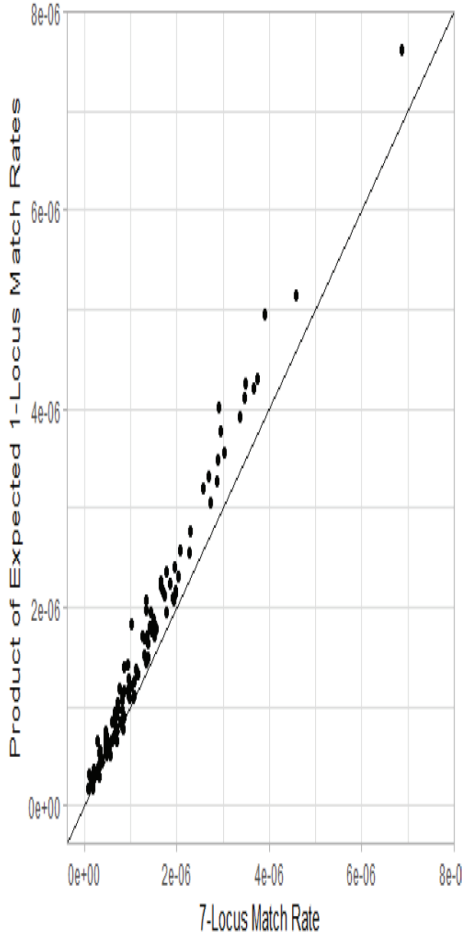




One population simulated data: $\theta = 0.01$

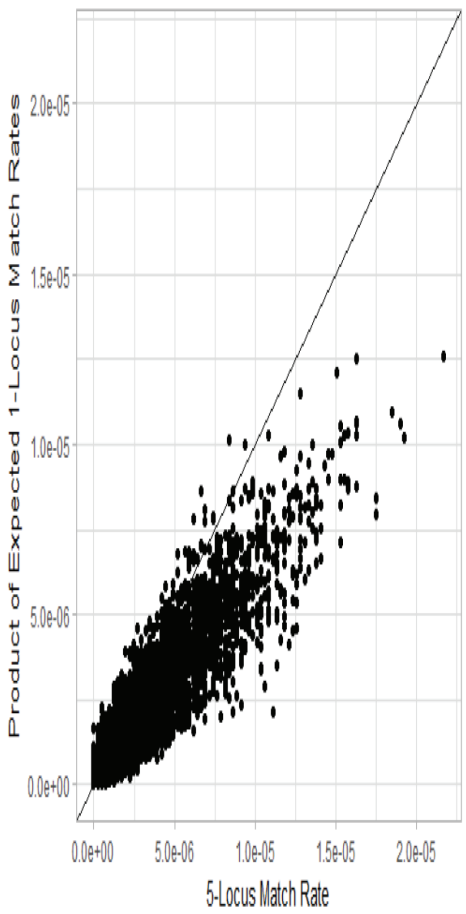




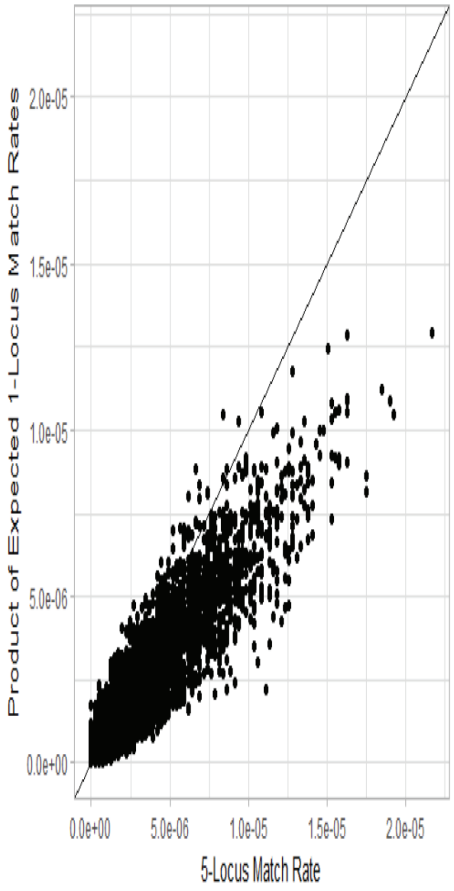


2849 US profiles

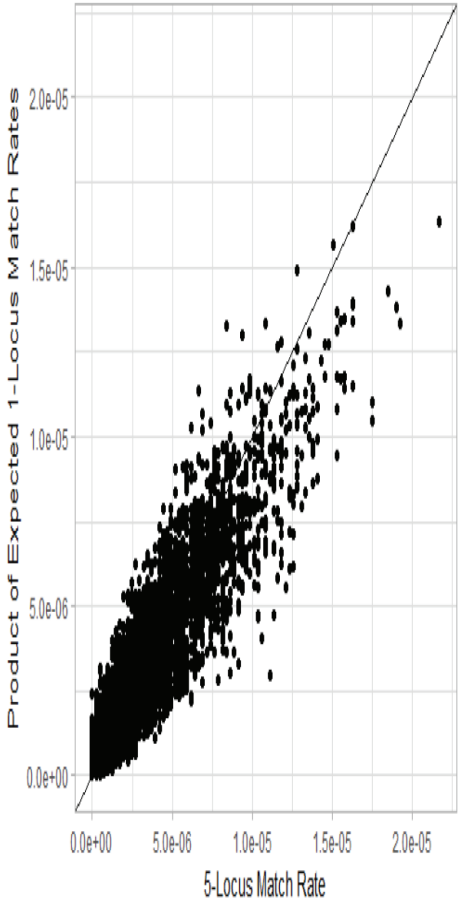
$\theta = 0$



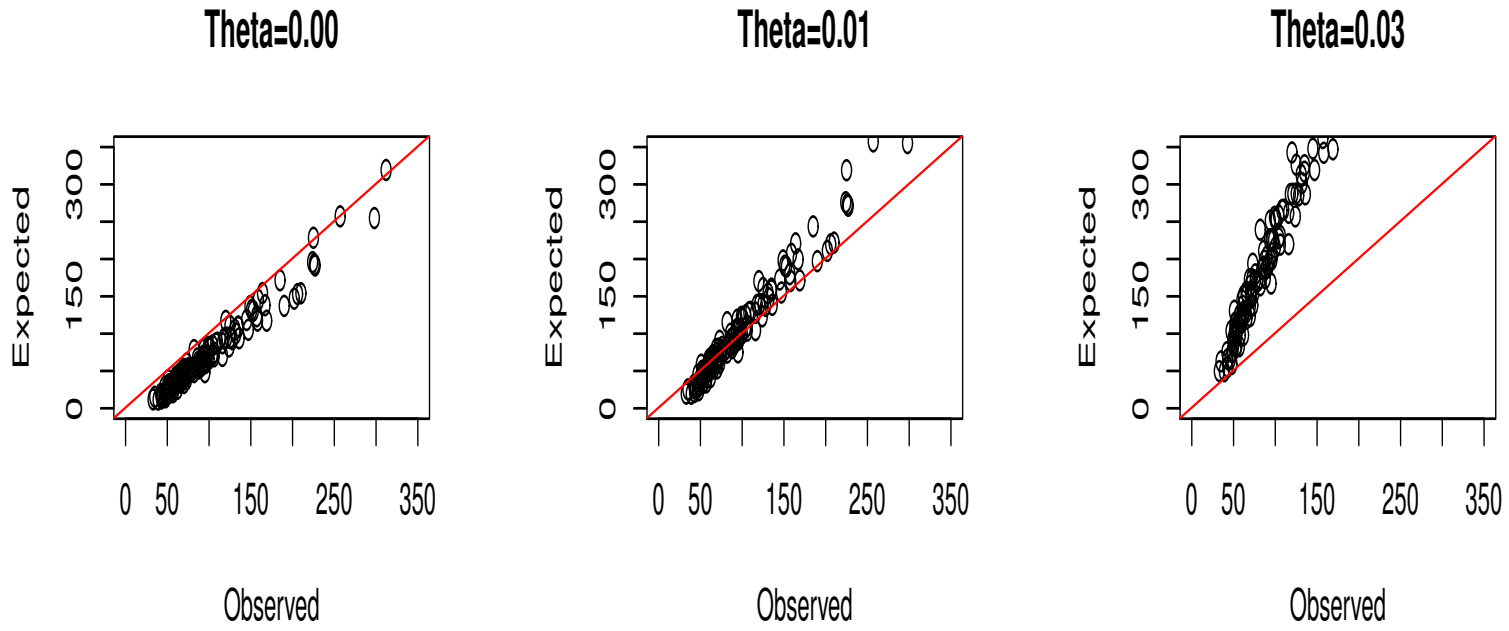
$\theta = 0.001$



$\theta = 0.01$



15,000 Australian Profiles



Numbers of five-locus matches among nine-locus profiles.

Weir BS. 2004. Journal of Forensic Sciences 49:1009-1014, 2004.

Conclusions

- Profile probabilities decrease at the same rate as number of loci increases.
- Match probabilities are not profile probabilities.
- Match probabilities decrease more slowly as number of loci increases.
- "Theta correction" may accommodate multi-locus dependencies.
- Empirical studies need much larger databases.