Section 3:
Allelic Independence and Matching

Testing for Allelic Independence

What is the probability a person has a particular DNA profile? What is the probability a person has a particular profile if it has already been seen once?

The first question is a little easier to think about, but difficult to answer in practice: it is very unlikely that a profile will be seen in any sample of profiles. Even for one STR locus with 10 alleles, there are 55 different genotypes and most of those will not occur in a sample of a few hundred profiles.

For locus D3S1358 in the African American population, the FBI frequency database shows that 31 of the 55 genotype counts are zero. Estimating the population frequencies for these 31 types as zero doesn't seem sensible.

D3S1358 Genotype Counts

Observed	<12	12	13	14	15	16	17	18	19	>19
<12	0									
12	0	0								
13	0	0	0							
14	0	0	0	2						
15	0	0	1	19	15					
16	1	1	1	15	39	19				
17	0	0	2	10	26	24	9			
18	1	0	1	2	6	10	3	0		
19	0	0	0	1	0	0	1	0	0	
>19	0	0	0	0	1	0	0	0	0	0

Hardy-Weinberg Law

A solution to the problem is to assume that the Hardy-Weinberg Law holds. For a random mating population, expect that genotype frequencies are products of allele frequencies.

For a locus with two alleles, A, a:

$$\begin{aligned} P_{AA} &= (p_A)^2 \\ P_{Aa} &= 2p_A p_a \\ P_{aa} &= (p_a)^2 \end{aligned}$$

For a locus with several alleles A_i:

$$\begin{aligned} P_{A_iA_i} &= (p_{A_i})^2 \\ P_{A_iA_j} &= 2p_{A_i} p_{A_j} \end{aligned}$$

D3S1358 Hardy-Weinberg Calculations

The allele counts for D3S1358 in the African-American sample are:

											Total
Allele	<12	12	13	14	15	16	17	18	19	>19	
Count	2	1	5	51	122	129	84	23	2	1	420

If the Hardy-Weinberg Law holds, then we would expect to see $n\tilde{p}_{13}^2 = 210 \times (5/420)^2 = 0.03$ individuals of type 13,13 in a sample of 210 individuals.

Also, we would expect to see $2n\tilde{p}_{13}\tilde{p}_{14} = 420 \times (5/420) \times (51/420) = 0.61$ individuals of type 13,14 in a sample of 210 individuals.

Other values are shown on the next slide.

D3S1358 Observed and Expected Counts

		<12	12	13	14	15	16	17	18	19	>19
<12	Obs.	0									
	Exp.	0.0									
12	Obs.	0	0								
	Exp.	0.0	0.0								
13	Obs.	0	0	0							
	Exp.	0.0	0.0	0.0							
14	Obs.	0	0	0	2						
	Exp.	0.2	0.1	0.6	3.1						
15	Obs.	0	0	1	19	15					
	Exp.	0.6	0.3	1.5	14.8	17.7					
16	Obs.	1	1	1	15	39	19				
	Exp.	0.6	0.3	1.5	15.7	37.5	19.8				
17	Obs.	0	0	2	10	26	24	9			
	Exp.	0.4	0.2	1.0	10.2	24.4	25.8	8.4			
18	Obs.	1	0	1	2	6	10	3	0		
	Exp.	0.1	0.1	0.3	2.8	6.7	7.1	4.6	0.6		
19	Obs.	0	0	0	1	0	0	1	0	0	
	Exp.	0.0	0.0	0.0	0.2	0.6	0.6	0.4	0.1	0.0	
>19	Obs.	0	0	0	0	1	0	0	0	0	0
	Exp.	0.0	0.0	0.0	0.1	0.3	0.3	0.2	0.1	0.0	0.0

Testing for Hardy-Weinberg Equilibrium

A test of the Hardy-Weinberg Law will somehow decide if the observed and expected numbers are sufficiently similar that we can proceed as though the law can be used.

In one of the first applications of Hardy-Weinberg testing in a US forensic setting:

> "To justify applying the classical formulas of population genetics in the Castro case the Hispanic population must be in Hardy-Weinberg equilibrium. Applying this test to the Hispanic sample, one finds spectacular deviations from Hardy-Weinberg equilibrium."

Lander ES. 1989. DNA fingerprinting on trial. Nature 339: 501-505.

VNTR “Coalescence”

Forensic DNA profiling initially used minisatellites, or VNTR loci, with large numbers of alleles. Heterozygotes would be scored as homozygotes if the two alleles were so similar in length that they coalesced into one band on an autoradiogram. Small alleles often not detected at all, and this is a likely cause of Lander’s finding (Devlin et al, Science 249:1416-1420.) .

Considerable debate in early 1990s on alternative “binning” strategies for reducing the number of alleles (Science 253:1037-1041, 1991).

Typing has moved to microsatellites with fewer and more easily distinguished alleles, but testing for Hardy-Weinberg equilibrium continues. There are still reasons why the law may not hold.

Population Structure can Cause Departure from HWE

If a population consists of a number of subpopulations, each in HWE but with different allele frequencies, there will be a departure from HWE at the population level. This is the Wahlund effect.

Suppose there are two equal-sized subpopulations, each in HWE but with different allele frequencies, then

	Subpopn 1	Subpopn 2	Total Popn
p_A	0.6	0.4	0.5
p_a	0.4	0.6	0.5
P_{AA}	0.36	0.16	$0.26 > (0.5)^2$
P_{Aa}	0.48	0.48	$0.48 < 2(0.5)(0.5)$
P_{aa}	0.16	0.36	$0.26 > (0.5)^2$

Population Structure

Effect of population structure taken into account with the "theta-correction." Matching probabilities allow for a variance in allele frequencies among subpopulations.

$$\Pr(AA|AA) \;=\; \frac{[3\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)}$$

where p_A is the average allele frequency over all subpopulations. We will come back to this expression.

Population Admixture

A population might represent the recent admixture of two parental populations. With the same two populations as before but now with 1/4 of marriages within population 1, 1/2 of marriages between populations 1 and 2, and 1/4 of marriages within population 2. If children with one or two parents in population 1 are considered as belonging to population 1, there is an excess of heterozygosity in the offspring population.

If the proportions of marriages within populations 1 and 2 are both 25% and the proportion between populations 1 and 2 is 50%, the next generation has

	Population 1	Population 2
P_{AA}	$0.09 + 0.12 = 0.21$	0.04
P_{Aa}	$0.12 + 0.26 = 0.38$	0.12
P_{aa}	$0.04 + 0.12 = 0.16$	0.09
	0.75	0.25

Exact HWE Test

The preferred test for HWE is an "exact" one. The test rests on the assumption that individuals are sampled randomly from a population so that genotype counts have a multinomial distribution:

$$\Pr(n_{AA}, n_{Aa}, n_{aa}) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(P_{AA})^{n_{AA}}(P_{Aa})^{n_{Aa}}(P_{aa})^{n_{aa}}$$

This equation is always true, and when there is HWE ($P_{AA} = p_A^2$ etc.) there is the additional result that the allele counts have a binomial distribution:

$$\Pr(n_A, n_a) = \frac{(2n)!}{n_A!n_a!}(p_A)^{n_A}(p_a)^{n_a}$$

Exact HWE Test

Putting these together gives the conditional probability of the genotypic data given the allelic data and given HWE:

$$\Pr(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a, \text{HWE}) = \frac{\frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}(p_A^2)^{n_{AA}}(2p_Ap_a)^{n_{Aa}}(p_a^2)^{n_{aa}}}{\frac{(2n)!}{n_A!n_a!}(p_A)^{n_A}(p_a)^{n_a}}$$

$$= \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!}\frac{2^{n_{Aa}}n_A!n_a!}{(2n)!}$$

Reject the Hardy-Weinberg hypothesis if this probability is unusually small.

Exact HWE Test Example

Reject the HWE hypothesis if the probability of the genotypic array, conditional on the allelic array, is among the smallest probabilities for all the possible sets of genotypic counts for those allele counts.

As an example, consider $(n_{AA} = 1, n_{Aa} = 0, n_{aa} = 49)$. The allele counts are $(n_A = 2, n_a = 98)$ and there are only two possible genotype arrays:

AA	Aa	aa	$\Pr(n_{AA}, n_{Aa}, n_{aa} \mid n_A, n_a, \text{HWE})$
1	0	49	$\frac{50!}{1!0!49!} \frac{2^0 2!98!}{100!} = \frac{1}{99}$
0	2	48	$\frac{50!}{0!2!48!} \frac{2^2 2!98!}{100!} = \frac{98}{99}$

Exact HWE Test

The probability of the data on the previous slide, conditional on the allele frequencies and on HWE, is $1/99 = 0.01$. This is less than the conventional 5% significance level.

In general, the p-value is the (conditional) probability of the data plus the probabilities of all the less-probable datasets. The probabilities are all calculated assuming HWE is true.

Exact HWE Test

Still in the two-allele case, for a sample of size $n = 100$ with minor allele frequency of 0.07, there are only 8 sets of genotype counts:

			Exact	
n_{AA}	n_{Aa}	n_{aa}	Prob.	p-value
93	0	7	0.0000	0.0000*
92	2	6	0.0000	0.0000*
91	4	5	0.0000	0.0000*
90	6	4	0.0002	0.0002*
89	8	3	**0.0051**	**0.0053***
88	10	2	0.0602	0.0654
87	12	1	0.3209	0.3863
86	14	0	0.6136	1.0000

So, for a nominal 5% significance level, the actual significance level is 0.0053 for an exact test that rejects when $n_{Aa} \leq 8$.

Permutation Test

For large sample sizes and many alleles per locus, there are too many genotypic arrays for a complete enumeration and a determination of which are the least probable 5% arrays.

A large number of the possible arrays is generated by permuting the alleles among genotypes, and calculating the proportion of these permuted genotypic arrays that have a smaller conditional probability than the original data. If this proportion is small, the Hardy-Weinberg hypothesis is rejected.

Permutation Test

Mark a set of five index cards to represent five genotypes:

Card 1:	A	A
Card 2:	A	A
Card 3:	A	A
Card 4:	a	a
Card 5:	a	a

Tear the cards in half to give a deck of 10 cards, each with one allele. Shuffle the deck and deal into 5 pairs, to give five genotypes.

Permutation Test

The permuted set of genotypes fall into one of four types:

AA	Aa	aa	Number of times
3	0	2	
2	2	1	
1	4	0	

Permutation Test

Check the following theoretical values for the proportions of each of the three types, from the expression:

$$\frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} \times \frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

AA	Aa	aa	Conditional Probability
3	0	2	$\frac{1}{21} = 0.048$
2	2	1	$\frac{12}{21} = 0.571$
1	4	0	$\frac{8}{21} = 0.381$

These should match the proportions found by repeating shufflings of the deck of 10 allele cards.

Permutation Test for D3S1358

For a STR locus, where $\{n_g\}$ are the genotype counts and $n = \sum_g n_g$ is the sample size, and $\{n_a\}$ are the alleles counts with $2n = \sum_a n_a$, the exact test statistic is

$$\Pr(\{n_g\}|\{n_a\}, \text{HWE}) = \frac{n!2^H \prod_a n_a!}{\prod_g n_g!(2n)!}$$

where H is the count of heterozygotes.

This probability for the African American genotypic counts at D3S1358 is 0.6163×10^{-13}, which is a very small number. But it is not unusually small if HWE holds: a proportion 0.81 of 1000 permutations have an even smaller probability. We do not reject the HWE hypothesis in this case.

Linkage Disequilibrium

This term is generally reserved for association between pairs of alleles – one at each of two loci. In the present context, it may simply mean some lack of independence of profile or match probabilities at different loci.

Unlinked loci are expected to be almost independent.

However, if two profiles match at several loci this may be because they are from the same, or related, people and so are likely to match at additional loci.

Allele Matching

Forensic genetics is concerned with matching of genetic profiles from evidence and from persons of interest. Profile match probabilities rest on the probabilities of matching among the alleles constituting the profiles.

Allele matching can refer to alleles within an individual (inbreeding), between individuals within a population (relatedness) and between populations (population structure). In all these cases there are parameters that describe profile match probabilities, and these parameters can be estimated by comparing the observed matching for a target set of alleles with that between a comparison set.

Allele Matching Within Individuals

The inbreeding coefficient for an individual is the probability it receives two alleles at a locus, one from each parent, that are *identical by descent.*

What can be observed, however, is identity in state. An individual is either homozygous or heterozygous at a locus: the two alleles either match or miss-match at that locus. The proportion of matching alleles at a locus is either zero or one, not a very informative statistic, but the proportion of an individual's loci that are homozygous may be informative for their inbreeding status.

There is still a need for a reference: for a locus such as a SNP with a small number of alleles many loci will be homozygous even for non-inbred individuals. Therefore we compare the proportion of loci with matching alleles for an individual with the matching proportion for pairs of alleles taken one from each of two individuals: is allele matching higher within than between individuals?

Inbreeding

If $\tilde{M}_j$ is the observed proportion of loci with matching alleles (i.e. homozygous) for individual j, and if $\tilde{M}_S$ is the observed proportion of matching alleles, one from each of two individuals in the population, then the within-population inbreeding coefficient f_j is estimated as

$$\hat{f}_j = \frac{\tilde{M}_j - \tilde{M}_S}{1 - \tilde{M}_S}$$

Note that this can be negative for individuals with high degrees of heterozygosity.

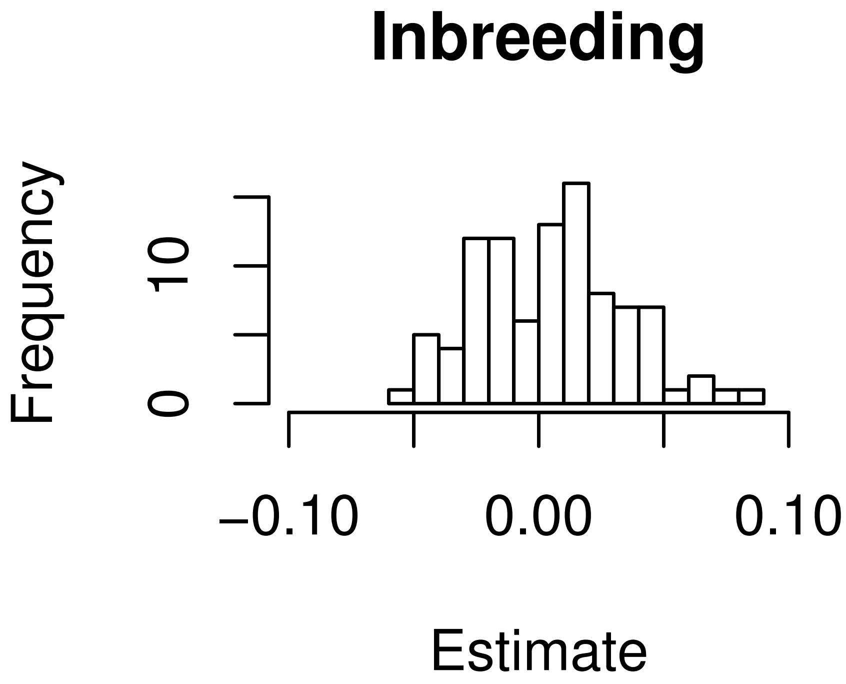
The average of these estimates over all the individuals in a sample from a population estimates the within-population inbreeding coefficient f:

$$\hat{f} = \frac{\tilde{M}_I - \tilde{M}_S}{1 - \tilde{M}_S}$$

where $\tilde{M}_I = \sum_{j=1}^{n} \tilde{M}_j / n$. Hardy-Weinberg equilibrium corresponds to $f = 0$.

SNP-based Inbreeding

From 400,000 SNPs on Chromosome 22 of the 1000 Genomes ACB populations (96 Afro-Caribbeans in Barbados);



Allele Matching Between Individuals

How can we tell if a pair of individuals has a high degree of allele matching? What does “high” mean?

We assess relatedness of individuals within a population by comparing their degree of allele matching with the degree for pairs of individuals with one from each of two different populations.

Allele Matching Between Individuals

If $\tilde{M}_{jj'}$ is the observed proportion of loci with matching alleles, one from each of individuals j and j', and if $\tilde{M}_S$ is the average of all the $\tilde{M}_{jj'}$'s, then the within-population kinship coefficient $beta_{jj'}$ is estimated as

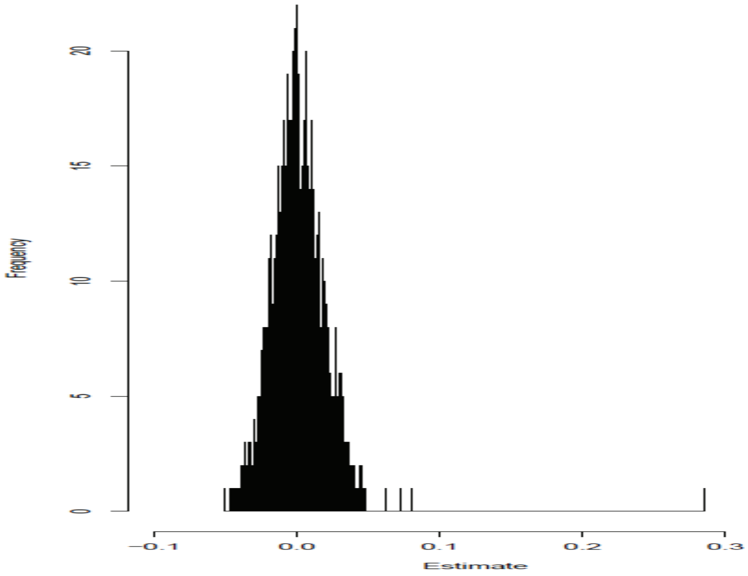
$$\hat{\beta}_{jj'} = \frac{\tilde{M}_{jj'} - \tilde{M}_S}{1 - \tilde{M}_S}$$

Note that this can be negative for pairs of individuals less related than the average pair-matching in the sample.

The average of these estimates over all pairs of individuals in a sample is zero, but this doesn't allow us to compare populations.

SNP-based Coancestry

From 400,000 SNPs on Chromosome 22 of the 1000 Genomes ACB populations (4560 pairs of Afro-Caribbeans in Barbados);



Allele Matching Between Populations

We calibrated allele matching within individuals by comparison with matching between pairs of individuals.

We calibrate the allele matching between pairs of individuals by comparison with matching between pairs of populations. If $\tilde{M}^{ii'}$ is the observed proportion of loci with matching alleles, one from each of populations i and i', and if $\tilde{M}_B$ is the average of all the $\tilde{M}^{ii'}$'s, then the total kinship coefficient $\beta_{jj'}$ is estimated as

$$\hat{\beta}_{jj'} = \frac{\tilde{M}_{jj'} - \tilde{M}^B}{1 - \tilde{M}^B}$$

The average of these estimates over all pairs of individuals in a sample from a population is

$$\hat{\beta} = \frac{\tilde{M}_S - \tilde{M}^B}{1 - \tilde{M}^B}$$

This is the "θ" needed for the "theta correction" discussed below.

Within-population Matching

We can get some empirical matching proportions when we have a set of profiles. To simplify this initial discussion, consider the following data for the Y-STR locus DYS390 from the NIST database:

	Population				
Allele	Afr.Am.	Cauc.	Hisp.	Asian	Total
20	4	1	1	0	6
21	176	4	17	1	198
22	43	45	14	17	119
23	36	116	50	17	219
24	56	145	129	21	351
25	23	46	21	36	126
26	3	2	2	4	11
27	0	0	2	0	2
Total	341	359	236	96	1032

Within- and Between-population Matching for DYS390

Within the African-American sample there are $341 \times 340 = 115,940$ pairs of profiles and the number of between individual-pair matches is

$$4 \times 3 + 176 \times 175 + 43 \times 42 + 36 \times 35 + 56 \times 55 + 23 \times 22 + 3 \times 2 = 37,470$$

so the within-population matching proportion is $37,470/115,940 = 0.323$.

Between the African-American and Caucasian samples, there are $341 \times 359 = 122,419$ pairs of profiles and the number of matches is

$$4 \times 1 + 176 \times 4 + 43 \times 45 + 36 \times 116 + 56 \times 145 + 23 \times 4 + 3 \times 2 = 12,403$$

so the between-population matching proportion is $12,403/122,419 = 0.101$.

Allele Counts in NIST Data for DYS391

	Population				
Allele	Afr.Am.	Cauc.	Hisp.	Asian	Total
7	0	0	1	0	1
8	0	1	0	1	2
9	2	12	16	3	33
10	238	162	128	79	607
11	93	175	89	13	370
12	7	9	2	0	18
13	1	0	0	0	1
Total	341	359	236	96	1032

The within-population matching proportion for the African-American sample is 65,006/115,940=0.561.

The between-population matching proportion for the African-American and Caucasian samples is 54,918/122,419=0.449.

Two-locus counts in NIST African-American Data for DYS390, DYS391

DYS390	DYS391	Count n_g	$n_g(n_g-1)$
22	10	34	1122
22	11	9	72
24	10	15	210
24	11	39	1482
24	12	1	0
24	9	1	0
23	10	19	342
23	11	14	182
23	12	3	6
21	10	157	24492
21	11	15	210
21	12	2	2
21	9	1	0
21	13	1	0
25	10	11	110
25	11	12	132
26	10	1	0
26	11	2	2
20	10	1	0
20	11	2	2
20	12	1	0

Two-locus counts in NIST Caucasian Data for DYS390, DYS391

DYS390	DYS391	Count n_g	$n_g(n_g-1)$
22	10	43	1806
22	11	1	0
22	9	1	0
24	10	48	2256
24	11	88	7656
24	12	4	12
24	9	5	20
23	10	50	2450
23	11	60	3540
23	12	2	2
23	9	3	6
23	8	1	0
21	10	3	6
21	11	1	0
25	10	18	306
25	11	22	462
25	12	3	6
25	9	3	6
26	11	2	2
20	11	1	0

Two-locus Matches

The within-population matching proportion for the African-American sample is 28,366/115,940=0.245.

The within-population matching proportion for the Caucasian sample is 18,536/128,522=0.144.

The between-population matching proportion for the African-American and Caucasian samples is 8,347/122,419=0.068.

There is a clear decrease in matching between populations from within populations.

Will match probabilities keep decreasing?

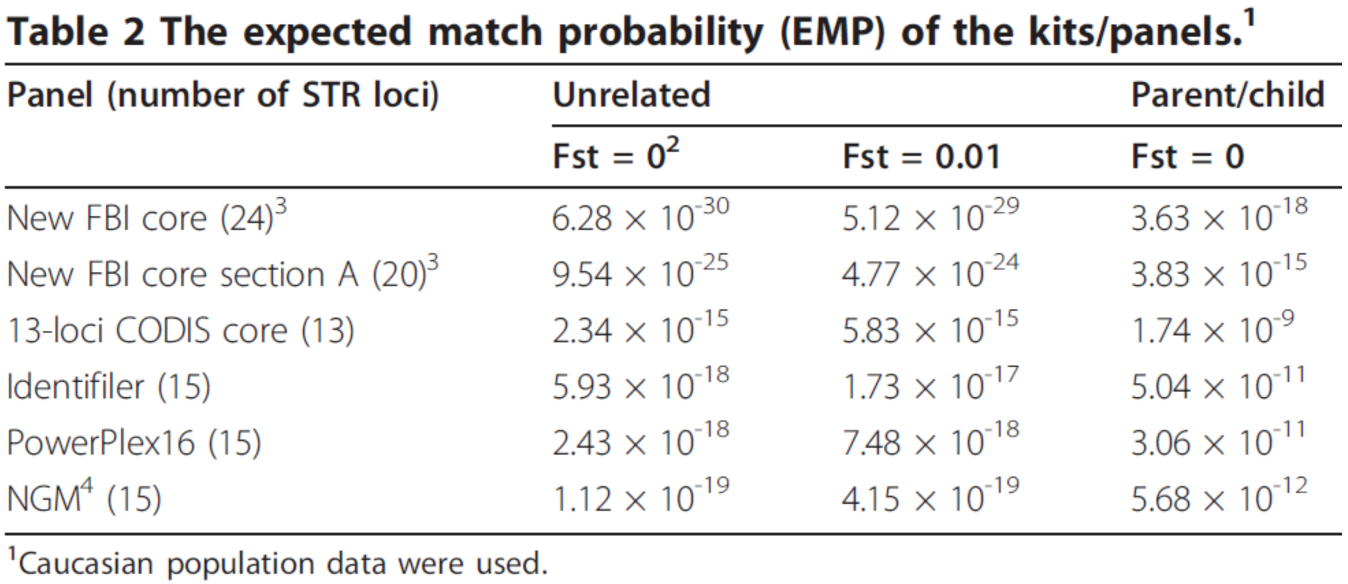


Table 2 The expected match probability (EMP) of the kits/panels.[1]

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.

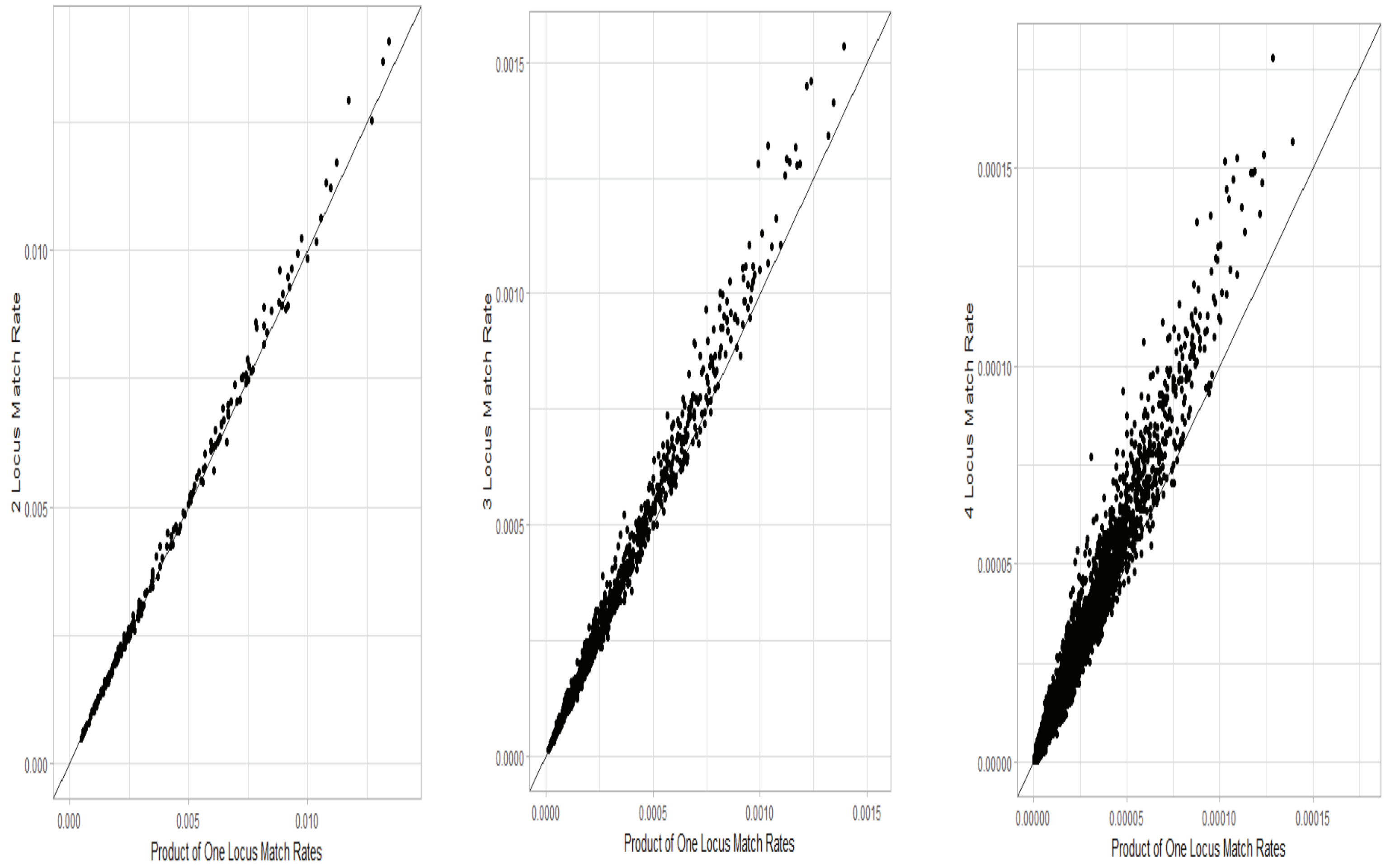
Likelihood ratios use match probabilities

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.

Empirical dependencies: 2849 20-locus profiles


2 Locus Match Rate
Product of One Locus Match Rates
3 Locus Match Rate
Product of One Locus Match Rates
4 Locus Match Rate
Product of One Locus Match Rates