ALLELE MATCHING

Within-population Matching

The key forensic genetic issue is that of matching profiles. What is the probability that two people have the same STR profile?

We can get some empirical estimate of this 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:

Allele Counts in NIST Data for DYS390

	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 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. We can establish some theory that describes these proportions.

Partial Matching

For autosomal markers, two profiles may be:

Match:	AA, AA or AB, AB
Partially Match:	AA, AB or AB, AC
Mismatch:	AA, BB or AA, BC or AB, CD

How likely are each of these?

Database Matching

If every profile in a database is compared to every other profile, each pair can be characterized as matching, partially matching or mismatching without regard to the particular alleles. We find the probabilities of these events by adding over all allele types.

The probability P_2 that two profiles match (at two alleles) is

$$\begin{aligned} P_2 &= \sum_A \Pr(AA, AA) + \sum_{A \neq B} \Pr(AB, AB) \\ &= \frac{\sum_A p_A[\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A][3\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)} \\ &\quad + \frac{2\sum_{A \neq B}[\theta + (1-\theta)p_A][\theta + (1-\theta)p_B]}{(1+\theta)(1+2\theta)} \end{aligned}$$

Database Matching

This approach leads to probabilities P_2, P_1, P_0 of matching at 2,1,0 alleles:

$$\begin{aligned} P_2 &= \frac{1}{D}[6\theta^3 + \theta^2(1-\theta)(2+9S_2) + 2\theta(1-\theta)^2(2S_2+S_3) \\ &\quad + (1-\theta)^3(2S_2^2 - S_4)] \end{aligned}$$

$$\begin{aligned} P_1 &= \frac{1}{D}[8\theta^2(1-\theta)(1-S_2) + 4\theta(1-\theta)^2(1-S_3) \\ &\quad + 4(1-\theta)^3(S_2 - S_3 - S_2^2 + S_4)] \end{aligned}$$

$$\begin{aligned} P_0 &= \frac{1}{D}[\theta^2(1-\theta)(1-S_2) + 2\theta(1-\theta)^2(1-2S_2+S_3) \\ &\quad + (1-\theta)^3(1-4S_2+4S_3+2S_2^2-3S_4)] \end{aligned}$$

where $D = (1+\theta)(1+2\theta)$, $S_2 = \sum_A p_A^2$, $S_3 = \sum_A p_A^3$, $S_4 = \sum_A p_A^4$. For any value of θ we can predict the matching, partially matching and mismatching proportions in a database.

FBI Caucasian Matching Counts

One-locus matches in FBI Caucasian data (18,721 pairs of 13-locus profiles).

		θ				
Locus	Observed	.000	.001	.005	.010	.030
D3S1358	.077	.075	.075	.077	.079	.089
vWA	.063	.062	.063	.065	.067	.077
FGA	.036	.036	.036	.038	.040	.048
D8S1179	.063	.067	.068	.070	.072	.083
D21S11	.036	.038	.038	.040	.042	.051
D18S51	.027	.028	.029	.030	.032	.040
D5S818	.163	.158	.159	.161	.164	.175
D13S317	.076	.085	.085	.088	.090	.101
D7S820	.062	.065	.066	.068	.070	.080
CSF1PO	.122	.118	.119	.121	.123	.134
TPOX	.206	.195	.195	.198	.202	.216
THO1	.074	.081	.082	.084	.086	.096
D16S539	.086	.089	.089	.091	.094	.105

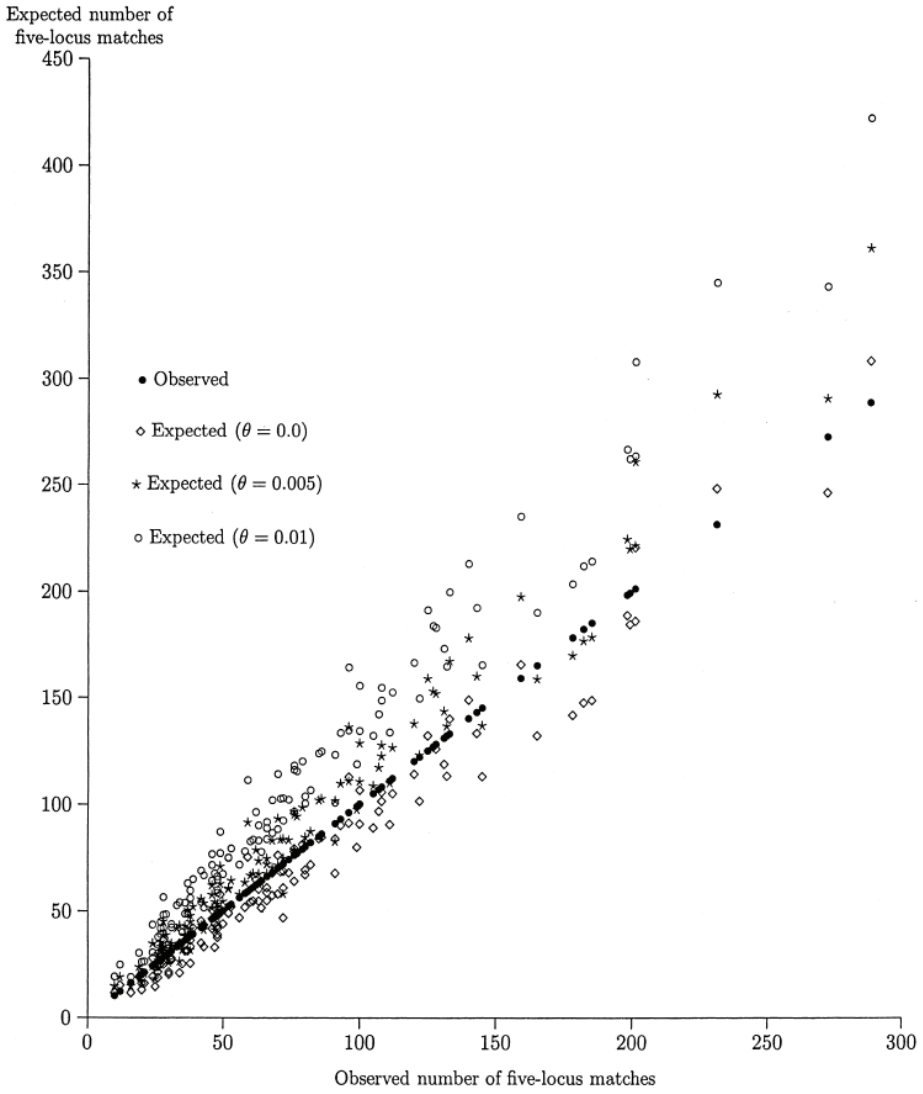
FBI Database Matching Counts

Matching loci	θ	Number of Partially Matching Loci												
		0	1	2	3	4	5	6	7	8	9	10	11	12
0	Obs.	0	3	18	92	249	624	1077	1363	1116	849	379	112	25
	.000	0	2	19	90	293	672	1129	1403	1290	868	415	134	26
	.010	0	2	14	70	236	566	992	1289	1241	875	439	148	30
1	Obs.	0	12	48	203	574	1133	1516	1596	1206	602	193	43	3
	.000	0	7	50	212	600	1192	1704	1768	1320	692	242	51	5
	.010	0	5	40	178	527	1094	1637	1779	1393	767	282	62	6
2	Obs.	0	7	61	203	539	836	942	807	471	187	35	2	
	.000	1	9	56	210	514	871	1040	877	511	196	45	5	
	.010	1	8	50	193	494	875	1096	969	593	239	57	6	
3	Obs.	0	6	33	124	215	320	259	196	92	16	1		
	.000	1	7	36	116	243	344	334	220	94	23	3		
	.010	0	6	35	117	256	380	387	268	120	32	4		
4	Obs.	1	5	17	29	54	82	67	16	6	0			
	.000	0	3	15	40	70	81	61	29	8	1			
	.010	0	3	15	44	81	98	78	40	12	1			
5	Obs.	0	1	2	6	12	14	6	5	0				
	.000	0	1	4	9	13	11	6	2	0				
	.010	0	1	4	11	16	15	9	3	0				
6	Obs.	0	1	0	2	2	0	0	0					
	.000	0	0	1	1	1	1	0	0					
	.010	0	0	1	2	2	1	1	0					

Predicted Matches when $n = 65,493$

Matching loci	Number of partially matching loci							
	0	1	2	3	4	5	6	7
6	4,059	37,707	148,751	322,963	416,733	319,532	134,784	24,125
7	980	7,659	24,714	42,129	40,005	20,061	4,150	
8	171	1,091	2,764	3,467	2,153	530		
9	21	106	198	163	50			
10	2	7	8	3				
11	0	0	0					
12	0	0						
13	0							

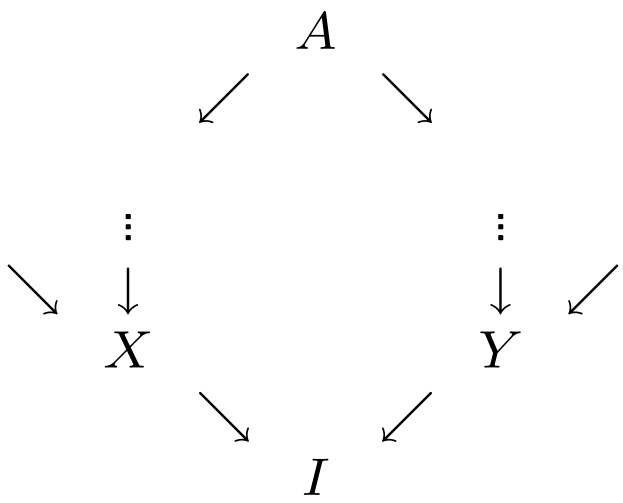
Multi-locus Matches



STR Survey: $\widehat{\beta}$ Values for Groups and Loci

	Geographic Region								
Locus	Africa	AusAb	Asian	Cauc	Hisp	IndPK	NatAm	Poly	Aver.
CSF1PO	0.003	0.002	0.008	0.008	0.002	0.007	0.055	0.026	0.011
D1S1656	0.000	0.000	0.000	0.002	0.003	0.000	0.000	0.000	0.011
D2S441	0.000	0.000	0.002	0.003	0.021	0.000	0.000	0.000	0.020
D2S1338	0.009	0.004	0.011	0.017	0.013	0.003	0.023	0.005	0.031
D3S1358	0.004	0.010	0.009	0.006	0.012	0.040	0.079	0.001	0.025
D5S818	0.002	0.013	0.009	0.008	0.014	0.018	0.044	0.007	0.029
D6S1043	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016
D7S820	0.004	0.021	0.010	0.007	0.007	0.046	0.030	0.005	0.026
D8S1179	0.003	0.007	0.012	0.006	0.002	0.031	0.020	0.008	0.019
D10S1248	0.000	0.000	0.000	0.002	0.004	0.000	0.000	0.000	0.007
D12S391	0.000	0.000	0.000	0.003	0.020	0.000	0.000	0.000	0.010
D13S317	0.015	0.016	0.013	0.008	0.014	0.025	0.050	0.014	0.038
D16S539	0.007	0.002	0.015	0.006	0.009	0.005	0.048	0.004	0.021
D18S51	0.011	0.012	0.014	0.006	0.004	0.010	0.033	0.003	0.018
D19S433	0.009	0.001	0.009	0.010	0.014	0.000	0.022	0.014	0.023
D21S11	0.014	0.012	0.013	0.007	0.006	0.023	0.067	0.018	0.021
D22S1045	0.000	0.000	0.007	0.001	0.000	0.000	0.000	0.000	0.015
FGA	0.002	0.009	0.012	0.004	0.007	0.016	0.021	0.006	0.013
PENTAD	0.008	0.000	0.012	0.012	0.002	0.017	0.000	0.000	0.022
PENTAE	0.002	0.000	0.017	0.006	0.003	0.012	0.000	0.000	0.020
SE33	0.000	0.000	0.012	0.001	0.000	0.000	0.000	0.000	0.004
TH01	0.022	0.001	0.022	0.016	0.018	0.014	0.071	0.017	0.071
TPOX	0.019	0.087	0.016	0.011	0.007	0.018	0.064	0.031	0.035
VWA	0.009	0.007	0.017	0.007	0.012	0.022	0.028	0.005	0.023
All Loci	0.006	0.014	0.010	0.007	0.008	0.018	0.043	0.011	0.022

Predicted Coancestry Values



Identify the path linking the parents X, Y of I to their common ancestor(s).

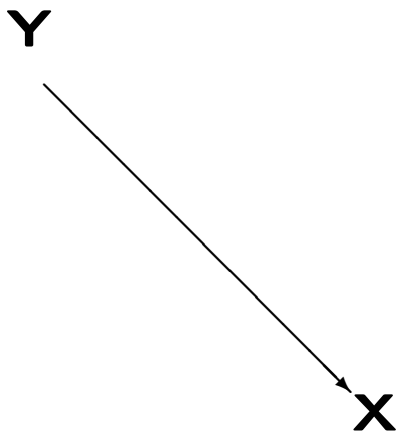
Path Counting

If the parents X, Y of an individual I have ancestor A in common, and if there are n individuals (including X, Y, I) in the path linking the parents through A, then the inbreeding coefficient of I, or the coancestry of X and Y, is

$$F_I = \theta_{XY} = \left(\frac{1}{2}\right)^n (1 + F_A)$$

If there are several ancestors, this expression is summed over all the ancestors.

Parent-Child



The common ancestor of parent X and child Y is X. The path linking X, Y to their common ancestor is YX and this has $n = 2$ individuals. Therefore

$$\theta_{XY} = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

Common Relatives

Relationship	Coancestry
Identical Twins	0.5
Parent Child	0.25
Full Sibs	0.25
Half Sibs	0.125
Double First Cousins	0.125
First Cousins	0.0625
Uncle Niece	0.0625
Unrelated	0

Comparing Hypothesized Relationships

Current practise is to compare the likelihoods of two profiles under alternative hypotheses about their degrees of relatedness.

On the verge now of being able to estimate the degree of relatedness.

Estimating Relatedness

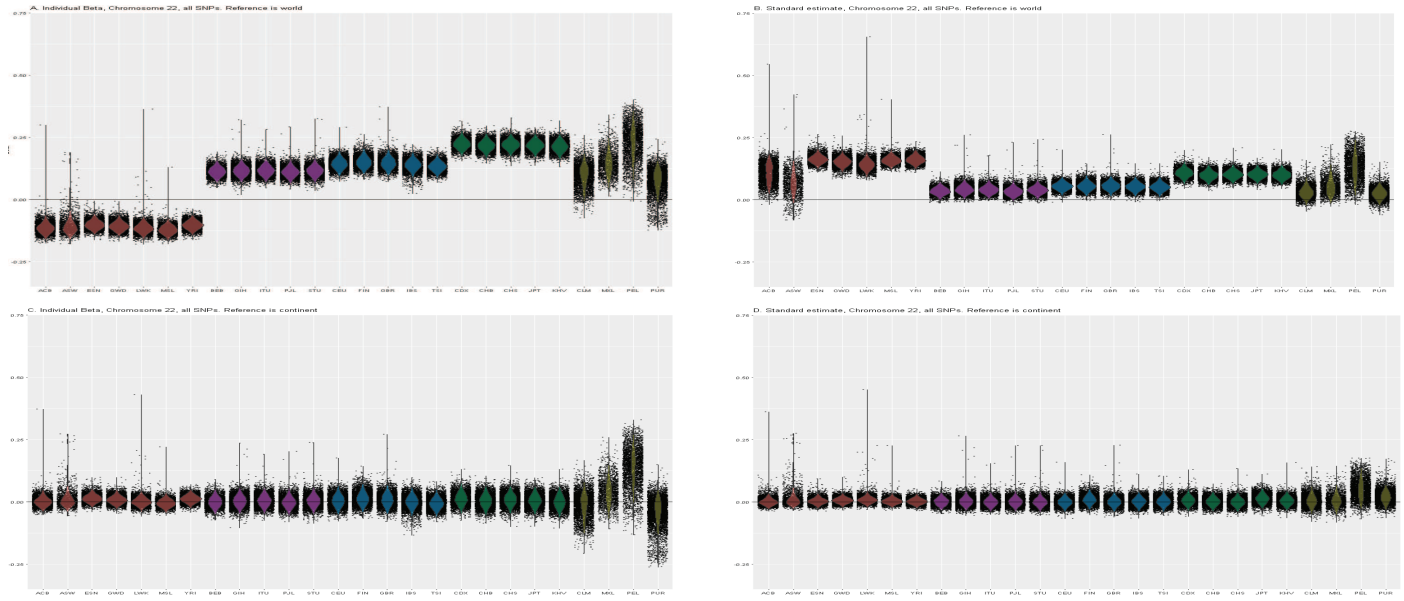
The proportion $\tilde{M}_{XY}$ of pairs of alleles, one from individual X and one from individual Y, that match is 0, 0.5 or 1:
Proportion=1: AA and AA
Proportion=0.5: AA and AB or AB and AB
Proportion=0: AA and BB or AA and BC or AB and CD

Averaging over all pairs of individuals, the observed proportion is $\tilde{M}_B$. The coancestry of individuals X, Y, relative to that of all individuals in the sample is

$$\hat{\theta}_{XY} = \frac{\tilde{M}_{XY} - \tilde{M}_B}{1 - \tilde{M}_B}$$

Coancestry is relative, not absolute

Top row: Whole world reference. Bottom row: Continental group reference.



Beta estimates | Standard estimates

Chromosome 22 data from 1000 Genomes.

Continents (left to right): AFR, SAS, EUR, EAS, AMR

Populations (l to r):**AFR**: ACB, ASW, ESN, GWD, LWK, MSL, YRI;
SAS: BEB, GIH, ITU, PJL, STU; **EUR**: CEU, FIN, GBR, IBS, TSI;
EAS: CDX, CHB, CHS, JPT; **AMR**: KHV, CLM, MXL, PEL, PUR