Advanced Topics in Forensic DNA Analysis

Statistics and Population Genetics

New Jersey State Police
Training Workshop

Hamilton, NJ
December 5-6, 2006

Dr. John M. Butler
National Institute of Standards and Technology

john.butler@nist.gov

Outline for This Section

- Result interpretation possibilities: exclusion, inconclusive, **match with frequency estimate**
- How allele frequency databases are generated
- Use of the product rule to determine RMP
- OmniPop program

Forensic DNA Typing, 2nd Edition:

Biology, Technology, and Genetics of STR Markers
(John M. Butler, Elsevier Science/Academic Press, 2005)

5 chapters on statistical issues

- **Basic Genetic Principles and Statistics**
- **STR Database Analyses**
- **Profile Frequency Estimates**
- **Approaches to Statistical Analysis of Mixtures**
- **Kinship and Paternity Testing**

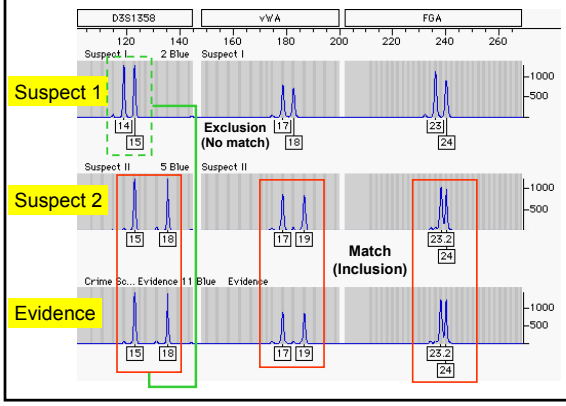
Examples are carefully worked through using the same U.S. population database to illustrate concepts

Three Possible Outcomes of a DNA Result

Butler, J.M. (2005) *Forensic DNA Typing*, 2nd Edition, p. 385

- **Exclusion (Non-match)** – The genotype comparison shows profile differences that can only be explained by the two samples originating from different sources.
- **Inconclusive** – The data does not support a conclusion as to whether the profiles match. This finding might be reported if two analysts remain in disagreement after review and discussion of the data and it is felt that insufficient information exists to support any conclusion.
- **Match (inclusion)** – Peaks between the compared STR profiles have the same genotypes and no unexplainable differences exist between the samples. Statistical evaluation of the significance of the match is usually reported with the match report.

Crime Scene STR Profile Compared to Two Suspects



Single Source Samples

Calculating a Random Match Probability (RMP)

Why Compute a Match Statistic?

- It would not be scientifically justifiable to speak of a match as proof of identity in the absence of underlying data that permit some reasonable estimate of how rare the matching characteristics actually are (NRC II, p. 192).
- Significance or weight of the evidence…

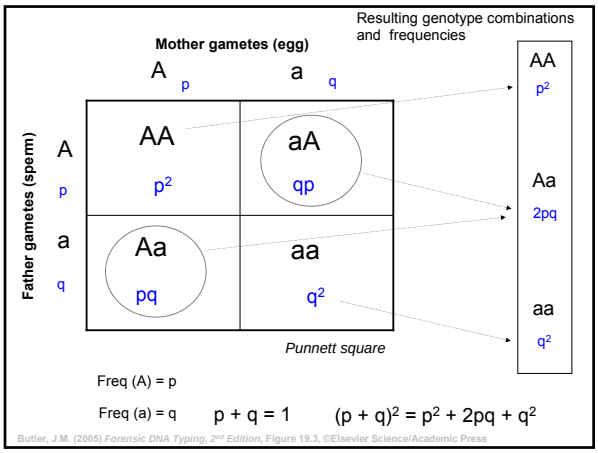
Population Genetics

- Population genetics seeks to understand genetic variation among individuals within and between population groups
- How can we estimate the frequency of a particular DNA profile?
- **Random match probability** - The probability that the DNA in a random sample from the population has the same profile as the DNA in the evidence sample. (Officers of the Court CD)

How Statistical Calculations are Made

- Generate data with set(s) of samples from desired population group(s)
  - Generally only 100-150 samples are needed to obtain reliable allele frequency estimates
- Determine allele frequencies at each locus
  - Count number of each allele seen
- Allele frequency information is used to estimate the rarity of a particular DNA profile
  - **Homozygotes (p^2), Heterozygotes (2pq)**
  - **Product rule used (multiply locus frequency estimates)**

For more information, see Chapters 20 and 21 in *Forensic DNA Typing*, 2nd Edition



Assumptions behind the Product Rule

- Independence between alleles (Hardy-Weinberg equilibrium)
  - permits correlation of allele frequency with genotype frequency
- Independence between loci (linkage equilibrium)
  - permits multiplication of genotype frequencies across all tested loci
- Typically only match probabilities for unrelated individuals are reported

Assumptions with Hardy-Weinberg Equilibrium

The Assumption	The Reason
Large population	Lots of possible allele combinations
No natural selection	No restriction on mating so all alleles have equal chance of becoming part of next generation
No mutation	No new alleles being introduced
No immigration/emigration	No new alleles being introduced or leaving
Random mating	Any allele combination is possible

None of these assumptions are really true…

Butler, J.M. (2005) Forensic DNA Typing, 2nd Edition, Table 20.6, ©Elsevier Science/Academic Press

Steps in Generating a Population Database

Decide on Number of Samples and Ethnic/Racial Grouping

Usually >100 samples per group

↓

Gather Samples

Often anonymous samples from a blood bank (self-declared ethnicities)

↓

Analyze Samples at Desired Genetic Loci

Summarize DNA types

↓

Determine Allele Frequencies for Each Locus

The number of times each allele is observed is counted and divided by the total number of alleles measured

↓

Perform Statistical Tests on Data

Hardy-Weinberg equilibrium for allele independence

Linkage equilibrium for locus independence

↓

Ethnic/ Racial Group 1 | Ethnic/ Racial Group 2

Examination of genetic distance between populations

↓

Use Database(s) to Estimate an Observed DNA Profile Frequency

PopStats calculation

From Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, Figure 20.1, ©Elsevier Science/Academic Press

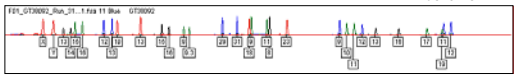
U.S. Population Study with Identifiler STR Kit

J. Forensic Sci. 48(4):908-911

FOR THE RECORD

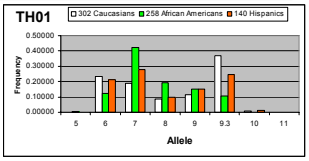
Allele Frequencies for 15 Autosomal STR Loci on U.S. Caucasian, African American, and Hispanic Populations*

Identifiler™



TH01 Allele Frequencies

	302 Caucasians	258 African Americans	140 Hispanics
5	0.00166	0.00388	
6	0.23179	0.12403	0.21429
7	0.19040	0.42054	0.27857
8	0.08444	0.19380	0.09643
9	0.11424	0.15116	0.15000
9.3	0.36755	0.10465	0.24643
10	0.00828	0.00194	0.01429
11	0.00166		



http://www.cstl.nist.gov/biotech/strbase/NISTpop.htm

Individual Genotypes Are Summarized and Converted into Allele Frequencies

Genotype Array	8	9	10	11	12	13	14	15		Allele Count	Observed Frequency
	8,8	8,9	8,10	8,11	8,12	8,13	8,14	8,15			
8	9	9	1	17	13	10	0	0	**8**	68	0.11258
		9,9	9,10	9,11	9,12	9,13	9,14	9,15			
9		1	2	15	10	4	3	0	**9**	45	0.07450
			10,10	10,11	10,12	10,13	10,14	10,15			
10			2	12	6	3	2	1	**10**	31	0.05132
				11,11	11,12	11,13	11,14	11,15			
11				37	54	21	12	0	**11**	205	0.33940
					12,12	12,13	12,14	12,15			
12					21	18	7	0	**12**	150	0.24834
						13,13	13,14	13,15			
13						7	5	0	**13**	75	0.12417
							14,14	14,15			
14							0	0	**14**	29	0.04801
								15,15			
15								0	**15**	**1**	**0.00166**
										604	

The **11,14** genotype was seen **12 times** in 302 samples (604 examined chromosomes)

Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, Table 20.2, ©Elsevier Science/Academic Press

Computer Programs for Performing Statistical Tests on Genetic Data

See Table 20.5 Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*

- PowerStats http://www.promega.com/geneticidtools/default.htm
- GDA http://lewis.eeb.uconn.edu/lewishome/software.html
- GENEPOP http://wbiomed.curtin.edu.au/genepop/index.html
- DNA-VIEW http://www.dna-view.com/ **(costs money)**
- DNATYPE ***Contact Ranajit Chakraborty about availability***
- ARLEQUIN http://lgb.unige.ch/arlequin/
- PowerMarker http://www.powermarker.net
- PopStats ***part of the FBI's CODIS system (not publicly available)***
- TFPGA http://bioweb.usu.edu/mpmbio/tfpga.asp

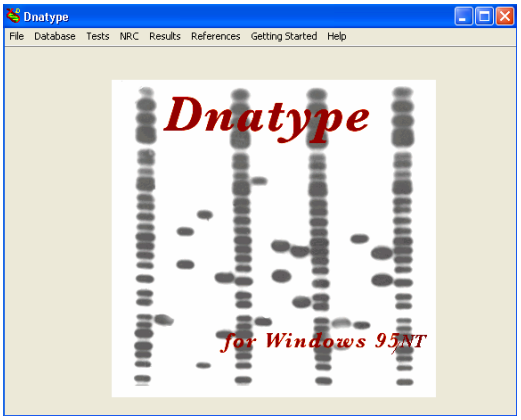


Figure 20.2, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

Testing for Independence within a Locus

- Hardy-Weinberg equilibrium (HWE) predicts stability of allele and genotype frequencies from one generation to the next
- Small p values ($p < 0.05$) cast doubt on the validity of the null hypothesis

DNA Statistics

For heterozygous loci

P = 2pq

P = probability; p and q are frequencies of allele in a given population

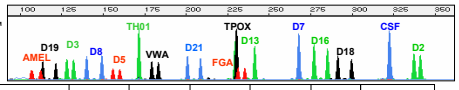
Example: For the locus D3S1358 and individual is 16,17 with frequencies of 0.2533 and 0.2152 respectively

P = 2(0.2533)(0.2152) = 0.1090 or **1 in 9.17**

For independent loci, the genotype frequencies can be combined through multiplication…
Profile Probability = (P1)(P2)…(Pn)
= *1 in a very large number…*

DNA Profile Frequency with all 13 CODIS STR loci

AmpFISTR® Identifiler™ (Applied Biosystems)

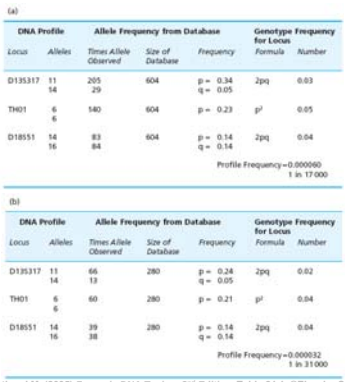


	Locus	allele 1	frequency	allele 2	frequency	1 in	Combined
2pq	D3S1358	16	0.2533	17	0.2152		9.17
2pq	VWA	17		18			81
2pq	FGA	21		22			1005
2pq	D8S1179	12		14			16,364
2pq	D21S11	28		30			185,073
2pq	D18S51	14		16			4,845,217
2pq	D5S818	12		13			44,818,259
2pq	D13S317	11		14			1.38×10^9
p^2	D7S820	9					4.38×10^{10}
2pq	D16S539	9		11			6.05×10^{11}
p^2	THO1	6					1.13×10^{13}
p^2	TPOX	8					3.94×10^{13}
p^2	CSF1PO	10					8.37×10^{14}

PRODUCT RULE

The Random Match Probability for this profile in the U.S. Caucasian population is **1 in 837 trillion (10^{12})**

Comparison of Results from Different Population Groups



(a)

DNA Profile		Allele Frequency from Database			Genotype Frequency for Locus	
Locus	Alleles	Times Allele Observed	Size of Database	Frequency	Formula	Number
D13S317	11	205	604	p = 0.34	2pq	0.03
	14	29		q = 0.05		
THO1	6	140	604	p = 0.23	p^2	0.05
	6					
D18S51	14	83	604	p = 0.14	2pq	0.04
	16	84		q = 0.14		

Profile Frequency = 0.000060
1 in 17,000

U.S. Caucasians (N = 302)

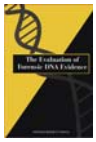
(b)

DNA Profile		Allele Frequency from Database			Genotype Frequency for Locus	
Locus	Alleles	Times Allele Observed	Size of Database	Frequency	Formula	Number
D13S317	11	66	280	p = 0.24	2pq	0.02
	14	13		q = 0.05		
THO1	6	60	280	p = 0.21	p^2	0.04
	6					
D18S51	14	39	280	p = 0.14	2pq	0.04
	16	38		q = 0.14		

Profile Frequency = 0.000032
1 in 31,000

U.S. Hispanics (N = 140)

Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, Table 21.1, ©Elsevier Science/Academic Press



NRC II Recommendations for Estimating Random-Match Probabilities

- Recommendation 4.1
  - Use the product rule to calculate profile frequency
  - If perpetrator's race is unknown, report calculations on racial groups for all possible suspects
  - For *heterozygotes*: use $\mathbf{2p_ip_j}$ or $2p_ip_j(1-\theta)$ (eq. 4.4b)
  - For *homozygotes*: use $\mathbf{p^2 + p(1-p)\theta}$ instead of p^2
  - With US population, use $\boldsymbol{\theta}$=0.01
  - With small, isolated populations, use $\boldsymbol{\theta}$=0.03

Why a Theta ($\boldsymbol{\theta}$) Correction?

- Used as a measure of the effects of population subdivision; due to co-ancestry (inbreeding) of alleles
  - Is essentially an attempt to correct for the degree of relatedness of alleles that have a common ancestry
- Basis in fixation indices (F-statistics) described by Sewall Wright in 1951 – F_{ST}, F_{IT}, F_{IS}
  - If the subpopulations are distinct and in HW proportions, then $\boldsymbol{\theta = F_{ST}}$
- Calculations typically performed as described by Weir and Cockerham (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358-1370

With US population groups (African Americans, Caucasians, etc.), use θ = 0.01
With small, isolated populations (Native Americans), use θ = 0.03

Empirical Measurements of Theta ($\boldsymbol{\theta}$)

- Budowle *et al.* (2001) CODIS STR Loci Data from 41 Sample Populations. *J. Forensic Sci.* 46(3): 453-489

TABLE 6—*F_{ST} values for the thirteen CODIS core STR loci.*

Locus	African American	Caucasian	Hispanic	Asian	Native American
CSF1PO	−0.0009	−0.0007	−0.0003	−0.0012	0.0244
D3S1358	−0.0005	−0.0009	0.0014	0.0035	0.0764
D5S818	0.0010	−0.0001	0.0010	0.0028	0.0656
D7S820	0.0000	−0.0005	0.0010	0.0039	0.0201
D8S1179	−0.0001	0.0000	0.0005	0.0025	0.0125
D13S317	0.0029	−0.0008	0.0047	0.0071	0.0157
D16S539	−0.0013	−0.0005	0.0067	0.0017	0.0132
D18S51	0.0012	0.0001	0.0011	0.0046	0.0268
D21S11	0.0005	0.0008	0.0013	0.0056	0.0371
FGA	0.0004	−0.0004	0.0008	0.0029	0.0168
TH01	0.0015	−0.0012	0.0041	0.0058	0.0356
TPOX	0.0021	−0.0015	0.0024	0.0100	0.0164
vWA	0.0011	−0.0011	0.0029	0.0027	0.0172
F_{ST} over all loci	0.0006	−0.0005	0.0021	0.0039	0.0282

θ < 0.01 **θ < 0.03**

Empirical Measurements of Theta (θ)

- Budowle and Chakraborty (2001) Population variation at the CODIS core short tandem repeat loci in Europeans. *Legal Med.* 3: 29-33
- "Because of the low value for theta, whether independence is assumed or an adjustment for substructure is employed, **there is little practical consequence for forensic purposes for estimating the frequency of a multiple locus DNA profile**. If theta is used, a value of 0.01 is very conservative for Europeans."
- **F_{ST} over all loci = 0.0028**

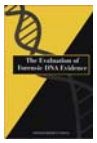
Basis of PopStats Calculations

Budowle *et al.* (2001) *J. Forensic Sci.* 46(3): 453-489

Bruce Budowle,[1] Ph.D.; Brendan Shea,[2] M.S.; Stephen Niezgoda,[2] M.B.A.; and Ranajit Chakraborty,[3] Ph.D.

CODIS STR Loci Data from 41 Sample Populations* **Data collected by 20 different forensic laboratories**

The F_{ST} estimates over all thirteen STR loci are
0.0006 for African Americans,
-0.0005 for Caucasians,
0.0021 for Hispanics,
0.0039 for Asians,
and 0.0282 for Native Americans.



NRC II Recommendations for Estimating Random-Match Probabilities **When Sub-Group Data Are Not Available**

Recommendation 4.2. If the particular subpopulation from which the evidence sample came is known, the allele frequencies for the specific subgroup should be used as described in Recommendation 4.1. If allele frequencies for the subgroup are not available, although data for the full population are, then the calculations should use the population-structure equations 4.10 for each locus, and the resulting values should then be multiplied.

$$\text{Homozygote: } P(A_iA_i|A_iA_i) = \frac{[2\theta + (1-\theta)p_i][3\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)} \quad (4.10a)$$

$$\text{Heterozygote: } P(A_iA_j|A_iA_j) = \frac{2[\theta + (1-\theta)p_i][\theta + (1-\theta)p_j]}{(1+\theta)(1+2\theta)} \quad (4.10b)$$

Butler, J.M. (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*, 2nd Edition, Elsevier: New York, Appendix VI, pp.623-625

Comparison of Results Obtained with Various NRC II Formula

	Under HWE	Unconditional (NRCII Recommendation 4.1)	Conditional with Substructure Adjustment
			(NRCII recommendation 4.10a)
Homozygote	p_i^2	$p_i^2+p_i(1-p_i)\theta$	$\frac{[p_i(1-\theta)+2\theta][p_i(1-\theta)+3\theta]}{(1+\theta)(1+2\theta)}$
TH01 6,6 $p_i = 0.23$ $\theta = 0.01$	$(0.23)^2$ $= 0.053$ 1 in 18.9	$(0.23)^2+(0.23)(1-0.23)(0.01)$ $= 0.053 + 0.0018$ $= 0.055$ 1 in 18.2	$\frac{[(0.23)(1-0.01)+2(0.01)][(0.23)(1-0.01)+3(0.01)]}{(1+0.01)(1+2(0.01))}$ $= (0.2477)(0.2577)/(1.01)(1.02)$ $= 0.062$ 1 in 16.1
			(NRCII recommendation 4.10b)
Heterozygote	$2p_ip_j$	$2p_ip_j(1-\theta)$ eq. (4.4b); NRCII, p.102	$\frac{2[p_i(1-\theta)+2\theta][p_j(1-\theta)+\theta]}{(1+\theta)(1+2\theta)}$
D13S317 11,14 $p_i = 0.34$ $p_j = 0.05$ $\theta = 0.01$	$2(0.34)(0.05)$ $= 0.0340$ 1 in 29.4	$2(0.34)(0.05)(1-0.01)$ $= 0.0337$ 1 in 29.7	$\frac{2[(0.34)(1-0.01)+0.01][(0.05)(1-0.01)+0.01]}{(1+0.01)(1+2(0.01))}$ $= 2(0.3466)(0.0595)/(1.01)(1.02)$ $= 0.0400$ 1 in 25.0

Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, Table 21.4, ©Elsevier Science/Academic Press

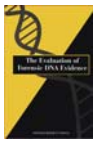
Example Calculations with Population Substructure Adjustments

Table 21.5
Example calculations with NRC II recommendations for population substructure adjustments (see Appendix VI). Scenarios with theta equal to 0.01 and 0.03 are examined.

From U.S. Caucasian (N = 302): Appendix II - sample in database

	A1	A2	Allele 1 freq (p)	Allele 2 freq (q)	Under HWE	Calc freq	NRCII Recommendation 4.1	θ = 0.01	θ = 0.03	NRCII Recommendation 4.10	θ = 0.01	θ = 0.03
D13S317	11	14	0.33940	0.04801	2pq	0.0326	2pq	0.0326	0.0326	eq. 4.10b	0.0386	0.0504
TH01	6	6	0.23179	—	p^2	0.0537	$p^2 + p(1-p)\theta$	0.0555	0.0591	eq. 4.10a	0.0428	0.0621
D16S51	14	16	0.13742	0.13907	2pq	0.0382	2pq	0.0382	0.0382	eq. 4.10b	0.0419	0.0493
D21S11	28	30	0.15894	0.27815	2pq	0.0884	2pq	0.0884	0.0884	eq. 4.10b	0.0927	0.1011
D5S1358	16	17	0.25331	0.21523	2pq	0.1090	2pq	0.1090	0.1090	eq. 4.10b	0.1129	0.1206
D5S818	12	13	0.36411	0.14073	2pq	0.1081	2pq	0.1081	0.1081	eq. 4.10b	0.1131	0.1228
D7S820	9	9	0.17715	—	p^2	0.0314	$p^2 + p(1-p)\theta$	0.0328	0.0358	eq. 4.10a	0.0390	0.0556
D8S1179	12	14	0.18543	0.16556	2pq	0.0614	2pq	0.0614	0.0614	eq. 4.10b	0.0654	0.0733
CSF1PO	10	10	0.21689	—	p^2	0.0470	$p^2 + p(1-p)\theta$	0.0487	0.0521	eq. 4.10a	0.0558	0.0744
FGA	21	22	0.18543	0.21854	2pq	0.0810	2pq	0.0810	0.0810	eq. 4.10b	0.0851	0.0930
D16S539	9	11	0.11258	0.32119	2pq	0.0723	2pq	0.0723	0.0723	eq. 4.10b	0.0773	0.0871
TPOX	8	8	0.53477	—	p^2	0.2860	$p^2 + p(1-p)\theta$	0.2885	0.2934	eq. 4.10a	0.2983	0.3227
VWA	17	18	0.28146	0.20033	2pq	0.1128	2pq	0.1128	0.1128	eq. 4.10b	0.1167	0.1245
AMEL	X	Y										
						1.20E-15		1.35E-15	1.70E-15		3.02E-15	3.02E-14

Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, Table 21.5, ©Elsevier Science/Academic Press



NRC II Recommendations for Estimating Random-Match Probabilities

- **Recommendation 4.3**. If the person who contributed the evidence sample is from a group or tribe for which no adequate database exists, **data from several other groups or tribes thought to be closely related to it should be used.** The profile frequency should be calculated as described in Recommendation 4.1 for each group or tribe.
  - For *heterozygotes*: use $\mathbf{2p_ip_j}$
  - For *homozygotes*: use $\mathbf{p^2 + p(1-p)\theta}$ instead of p^2

Allele Frequency Tables

Allele frequencies denoted with an asterisk (*) are **below the 5/2N minimum allele threshold** recommended by the National Research Council report (NRCII) *The Evaluation of Forensic DNA Evidence* published in 1996.

D3S1358

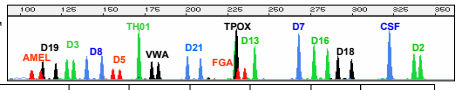
Allele	Butler *et al.* (2003) *JFS* 48(4):908-911 Caucasian N= 302	Einum *et al.* (2004) *JFS* 49(6) Caucasian N= 7,636	Allele	African American N=258	African American N= 7,602
11	0.0017*	0.0009	11	--	0.0003*
12	0.0017*	0.0007	12	--	0.0045
13	--	0.0031	13	0.0019*	0.0077
14	0.1027	0.1240	14	0.0892	0.0905
15	**0.2616**	**0.2690**	15	0.3023	0.2920
15.2	--	--	15.2	0.0019*	0.0010
16	0.2533	0.2430	16	**0.3353**	**0.3300**
17	0.2152	0.2000	17	0.2054	0.2070
18	0.15232	0.1460	18	0.0601	0.0630
19	0.01160	0.0125	19	0.0039*	0.0048
20	0.0017*	0.0001*	20		

Most common allele

>25X number of samples

DNA Profile Frequency with all 13 CODIS STR loci

AmpFISTR® Identifiler™ (Applied Biosystems)



What would be entered into a DNA database for searching: 16,17-17,18-21,22-12,14-28,30-14,16-12,13-11,14-9,9-9,11-6,6-8,8-10,10

Locus	allele	frequency	allele	frequency	1 in	Combined
D3S1358	**16**	*0.2533*	**17**	*0.2152*	**9.17**	**9.17**
VWA	**17**	*0.2815*	**18**	*0.2003*	**8.87**	**81**
FGA	**21**	*0.1854*	**22**	*0.2185*	**12.35**	**1005**
D8S1179	**12**	*0.1854*	**14**	*0.1656*	**16.29**	**16,364**
D21S11	**28**	*0.1589*	**30**	*0.2782*	**11.31**	**185,073**
D18S51	**14**	*0.1374*	**16**	*0.1391*	**26.18**	**4,845,217**
D5S818	**12**	*0.3841*	**13**	*0.1407*	**9.25**	**44,818,259**
D13S317	**11**	*0.3394*	**14**	*0.0480*	**30.69**	**1.38 x 10^9**
D7S820	**9**	*0.1772*			**31.85**	**4.38 x 10^{10}**
D16S539	**9**	*0.1126*	**11**	*0.3212*	**13.8**	**6.05 x 10^{11}**
THO1	**6**	*0.2318*			**18.62**	**1.13 x 10^{13}**
TPOX	**8**	*0.5348*			**3.50**	**3.94 x 10^{13}**
CSF1PO	**10**	*0.2169*			**21.28**	**8.37 x 10^{14}**

PRODUCT RULE

The Random Match Probability for this profile in the U.S. Caucasian population is **1 in 837 trillion (10^{12})**

The Same 13 Locus STR Profile in Different Populations

1 in 837 trillion

1 in 0.84 quadrillion (10^{15}) in U.S. Caucasian population (NIST)
1 in 2.46 quadrillion (10^{15}) in U.S. Caucasian population (FBI)*
1 in 1.86 quadrillion (10^{15}) in Canadian Caucasian population*

1 in 16.6 quadrillion (10^{15}) in African American population (NIST)
1 in 17.6 quadrillion (10^{15}) in African American population (FBI)*

1 in 18.0 quadrillion (10^{15}) in U.S. Hispanic population (NIST)

These values are **for unrelated individuals** assuming no population substructure (using only p^2 and 2 pq)

NIST study: Butler, J.M., *et al.* (2003) Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. *J. Forensic Sci.* 48(4):908-911. (http://www.cstl.nist.gov/biotech/strbase/NISTpop.htm)

*http://www.csfs.ca/pplus/profiler.htm

Theoretical **Most** Common STR Type

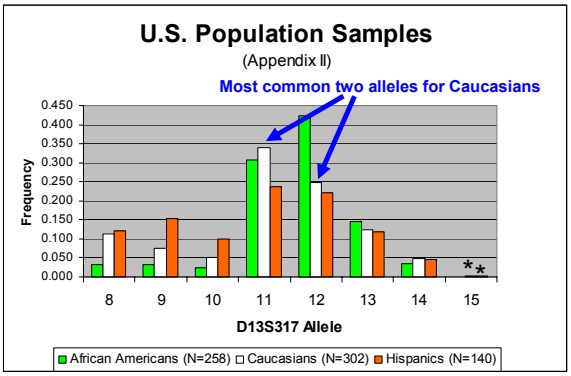
Locus	A1	A2	Allele 1 Freq (p)	Allele 2 Freq (q)		Most Common Genotype Frequency
D13S317	11	12	0.33940	0.24834	2pq	0.1686
D16S539	12	11	0.32816	0.32119	2pq	0.2095
D18S51	15	16	0.15894	0.13907	2pq	0.0442
D21S11	30	29	0.27815	0.19536	2pq	0.1087
D3S1358	15	16	0.26159	0.25331	2pq	0.1325
D5S818	12	11	0.38411	0.36093	2pq	0.2773
D7S820	10	11	0.24338	0.20695	2pq	0.1007
D8S1179	13	12	0.30464	0.18543	2pq	0.1130
CSF1PO	12	11	0.36093	0.30132	2pq	0.2175
FGA	22	21	0.21854	0.18543	2pq	0.0810
TH01	9.3	6	0.36755	0.23179	2pq	0.1704
TPOX	8	11	0.53477	0.24338	2pq	0.2603
VWA	17	18	0.28146	0.20033	2pq	0.1128
						6.26×10^{-12}
					1 in...	1.60×10^{11} (160 billion)

Calculations for the theoretically most common genotype frequencies and profile frequency based on two most common alleles found in a U.S. Caucasian allele frequency database

6.26×10^{-12} or 1 in 160 billion

Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, Table 20.9, ©Elsevier Science/Academic Press

Figure 20.3



Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, Figure 20.3, ©Elsevier Science/Academic Press

Theoretical **Least** Common STR Type

Using 5/2N minimum allele frequency rule

If N = 302, then 5/2N = 0.00828 = p = q

And $(2pq)^{13}$ = $[2 \times 0.00828 \times 0.00828]^{13}$

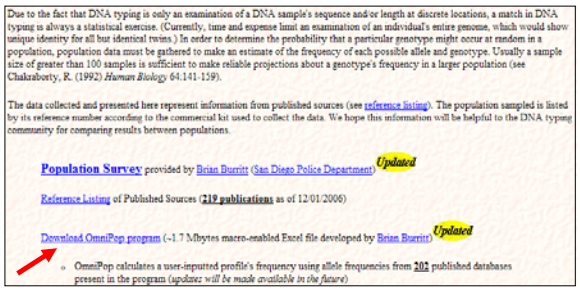
6.06×10^{-51}
or 1 in 1.65×10^{50}

Websites for Software Used in Match Probability Calculations

- **OmniPop** (**http://www.cstl.nist.gov/biotech/strbase/populationdata.htm**)
  - is an Excel-based program developed by a forensic scientist named Brian Burritt of the San Diego Police Department. OmniPop calculates a user-inputted STR profile's frequency using allele frequencies from 202 published databases. The program is freely available for download from the NIST STRBase website.
- **European Network of Forensic Science Institutes DNA Working Group STR Population Database** (**http://www.str-base.org/index.php**)
  - uses 5,699 samples from 24 European populations in order to make match probability calculation on user-inputted STR profiles containing the 10 STR loci present in the SGM Plus kit (Applied Biosystems) that is widely used in Europe.
- **Canadian Random Match Calculator** (**http://www.csfs.ca/pplus/profiler.htm**)
  - enables calculation of user-inputted STR profiles for the 13 U.S. core STR loci amplified by the Profiler Plus and COfiler kits sold by Applied Biosystems. This program enables comparison of results from limited FBI and Canadian collected allele frequencies.

OmniPop Program

Available from http://www.cstl.nist.gov/biotech/strbase/populationdata.htm


Due to the fact that DNA typing is only an examination of a DNA sample's sequence and/or length at discrete locations, a match in DNA typing is always a statistical exercise. (Currently, time and expense limit an examination of an individual's entire genome, which would show unique identity for all but identical twins.) In order to determine the probability that a particular genotype might occur at random in a population, population data must be gathered to make an estimate of the frequency of each possible allele and genotype. Usually a sample size of greater than 100 samples is sufficient to make reliable projections about a genotype's frequency in a larger population (see Chakraborty, R. (1992) *Human Biology* 64:141-159).

The data collected and presented here represent information from published sources (see reference listing). The population sampled is listed by its reference number according to the commercial kit used to collect the data. We hope this information will be helpful to the DNA typing community for comparing results between populations.

Population Survey provided by Brian Burritt (San Diego Police Department) *Updated*

Reference Listing of Published Sources (**219 publications** as of 12/01/2006)

Download OmniPop program (~1.7 Mbytes macro-enabled Excel file developed by Brian Burritt) *Updated*

- OmniPop calculates a user-inputted profile's frequency using allele frequencies from **202** published databases present in the program (*updates will be made available in the future*)

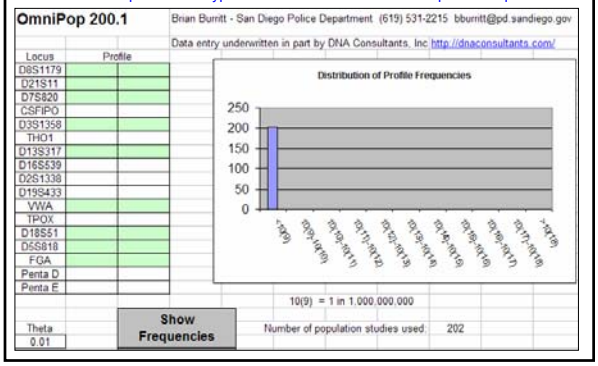
http://www.cstl.nist.gov/biotech/strbase/population/OmniPop200.1.xls

OmniPop 200.1

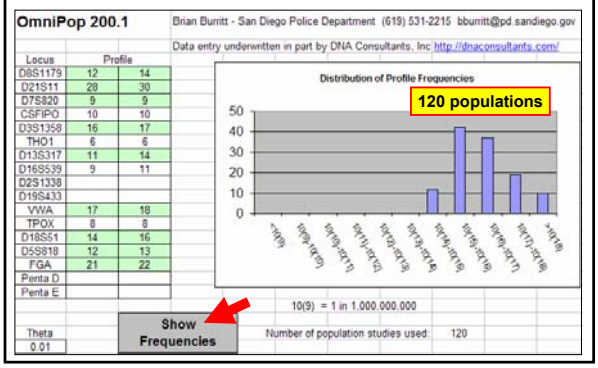
- Published allele frequencies
  - From 120 populations containing all 13 CODIS loci
  - From 202 populations with 9 loci (Profiler Plus)
- Based on 89 publications
- Available from Brian Buritt (San Diego Police Dept)
  - (619) 531-2215
  - bburritt@pd.sandiego.gov

OmniPop – Excel program

Takes user-inputted STR types and calculates distribution of profile frequencies



OmniPop Plots Profile Frequency Distributions



OmniPop Lists Frequency Calculations for All 120 Populations

Most Common Profile Frequencies		Least Common Profile Frequencies	
Serbian (157)	1.42E+14	Athabaskan (Alaska) (60)	9.65E+18
Portuguese (6)	3.43E+14	Inupiat (Alaska) (60)	1.71E+20
Belgian (99)	6.76E+14	Yupik (Alaska) (60)	2.02E+20
Caucasian (64)	**7.43E+14**	PC/BT-Asian (4)	3.77E+20
Swiss Caucasian (3)	7.59E+14	Canadian Aboriginal (56)	6.54E+20
Azores (82)	7.68E+14	Navajo (2)	7.09E+20
Scottish (11)	7.89E+14	Apache (2)	2.65E+21

OmniPop References

2-CODIS STR Loci Data from 41 Sample Populations, J Forensic Sci, 2001, 46(3), 453-489.

60 - Population studies on three Native Alaska population groups using STR loci, FSI, 2002, p51-57

64 - Allele Frequencies for 15 Autosomal STR Loci in U.S. Caucasian, African American, and Hispanic Populations, JFS, 2003, p908-911

157 - Allele frequencies of the 15 AmpFISTR Identifiler loci in the population of Vojvodina Province, Serbia and Montenegro, IJLM, 2004, 184-186

STR Cumulative Profile Frequency with Multiple Population Databases

STR Locus	Profile Computed	Number of Populations Used	Cumulative Profile Frequency Range (1 in ...)	Cumulative Profile Frequency against U.S. Caucasians (Appendix II)
D3S1358	16,17	166	5.24 to 62.6	9.19
VWA	17,18	166	37.6 to 1080	81.8
FGA	21,22	166	737 to 119 000	1010
D8S1179	12,14	166	8980 to 5 430 000	16 400
D21S11	28,30	166	165 000 to 248 000 000	186 000
D18S51	14,16	166	3.85×10^{4} to 2.68×10^{10}	4.88×10^{6}
D5S818	12,13	166	2.28×10^{3} to 4.22×10^{11}	4.51×10^{7}
D13S317	11,14	166	4.32×10^{4} to 1.69×10^{11}	1.38×10^{8}
D7S820	9,9	166	1.17×10^{10} to 2.98×10^{15}	4.22×10^{10}
D16S539	9,11	97	4.06×10^{11} to 1.11×10^{18}	5.82×10^{11}
TH01	6,6	97	9.30×10^{12} to 1.45×10^{19}	1.05×10^{13}
TPOX	8,8	97	3.33×10^{13} to 1.54×10^{20}	3.63×10^{13}
CSF1PO	10,10	97	3.43×10^{14} to 2.65×10^{21}	7.43×10^{14}

Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, D.N.A. Box 21.1, ©Elsevier Science/Academic Press

Theoretical RMP Range

6.26×10^{-12} or 1 in 160 billion

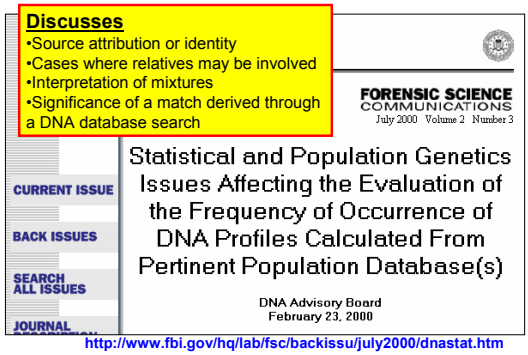
6.06×10^{-51} or 1 in 1.65×10^{50}

Observed RMP Range

10^{14} to 10^{21}

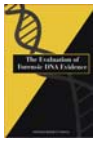
DNA Advisory Board Statistics Article

Discusses
- Source attribution or identity
- Cases where relatives may be involved
- Interpretation of mixtures
- Significance of a match derived through a DNA database search



http://www.fbi.gov/hq/lab/fsc/backissu/july2000/dnastat.htm

Impact of Relatedness on Match Probabilities



NRC II Recommendations for Estimating Random-Match Probabilities **When Relatives May Be Involved**

Recommendation 4.4. If the possible contributors of the evidence sample include relatives of the suspect, DNA profiles of those relatives should be obtained. If these profiles cannot be obtained, the probability of finding the evidence profile in those relatives should be calculated with Formulae 4.8 or 4.9.

Genotype of suspect	Probability of same genotype in a relative	
Homozygote: A_iA_i	$p_i^2 + 4p_i(1 - p_i)F$	(4.8a)
Heterozygote: A_iA_j	$2p_ip_j + 2(p_i + p_j - 4p_ip_j)F$	(4.8b)

For parent and offspring, F = 1/4; for half-siblings, 1/8; for uncle and nephew, 1/8; for first cousins, 1/16.

Full siblings, being bilineal rather than unilineal, require different formulae:

$$A_iA_i: \ (1 + 2p_i + p_i^2)/4 \qquad (4.9a)$$

$$A_iA_j: \ (1 + p_i + p_j + 2p_ip_j)/4 \qquad (4.9b)$$

Butler, J.M. (2005) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*, 2nd Edition, Elsevier: New York; Appendix VI, pp.623-625

Example Calculations with Corrections for Relatives

Table 21.6
Example calculations with corrections for relatives using the NRC II recommended formula.

From U.S. Caucasian (N = 302): Appendix II – sample in database

	A1	A2	Allele 1 freq (p)	Allele 2 freq (q)		Calc freq		NRCII Recommendation 4.4 F= 1/4 (parent)	F= 1/8 (half sib)	F= 1/16 (1st cousin)		Full sib
D13S317	11	14	0.33940	0.04801	2pq	0.0326	eq. 4.8b	0.1937	0.1131	0.0729	eq. 4.9b	0.3550
TH01	6	6	0.23179	—	p^2	0.0537	eq. 4.8a	0.2318	0.1428	0.0962	eq. 4.9a	0.3793
D16S539	9	11	0.11258	0.32119	2pq	0.0723	eq. 4.8b	0.2169	0.1446	0.1085	eq. 4.9b	0.3765
D18S51	14	16	0.13742	0.13907	2pq	0.0382	eq. 4.8b	0.1362	0.0862	0.0632	eq. 4.9b	0.3287
D21S11	28	30	0.15894	0.27815	2pq	0.0884	eq. 4.8b	0.2165	0.1535	0.1209	eq. 4.9b	0.3814
D3S1358	16	17	0.25331	0.21523	2pq	0.1090	eq. 4.8b	0.2343	0.1717	0.1403	eq. 4.9b	0.3944
D5S818	12	13	0.38411	0.14073	2pq	0.1081	eq. 4.8b	0.2624	0.1853	0.1467	eq. 4.9b	0.4082
D7S820	9	9	0.17715	—	p^2	0.0314	eq. 4.8a	0.1772	0.1043	0.0678	eq. 4.9a	0.3464
D8S1179	12	14	0.18543	0.16556	2pq	0.0614	eq. 4.8b	0.1795	0.1184	0.0899	eq. 4.9b	0.3531
CSF1PO	10	10	0.21689	—	p^2	0.0470	eq. 4.8a	0.2169	0.1320	0.0895	eq. 4.9a	0.3702
FGA	21	22	0.18543	0.21854	2pq	0.0810	eq. 4.8b	0.2020	0.1415	0.1113	eq. 4.9b	0.3713
TPOX	8	8	0.53477	—	p^2	0.2860	eq. 4.8a	0.5348	0.4104	0.3482	eq. 4.9a	0.5889
VWA	17	18	0.28146	0.20033	2pq	0.1128	eq. 4.8b	0.2409	0.1768	0.1448	eq. 4.9b	0.3986
AMEL	X	Y										
						1.20E-15		3.17E-09	1.68E-11	3.74E-13		4.04E-06 1 in 247,616

Effects of Family Relatedness on Match Probabilities

Relationship	Match probability formula		
	Homozygotes (A_iA_i)		Result with TH01 6,6 (p = 0.23179)
Full siblings	$\frac{(1+p)^2+(7+7p-2p^2)\theta+(16-9p+p^2)\theta^2}{4(1+\theta)(1+2\theta)}$		0.38921
Parent and child	$\frac{2\theta+(1-\theta)p}{(1+\theta)}$		0.24700
Half siblings	$\frac{[2\theta+(1-\theta)p][2+4\theta+(1-\theta)p]}{2(1+\theta)(1+2\theta)}$		0.27479
First cousins	$\frac{[2\theta+(1-\theta)p][2+11\theta+3(1-\theta)p]}{4(1+\theta)(1+2\theta)}$		0.10888
Unrelated	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{4(1+\theta)(1+2\theta)}$	(NRC II, 4.10a)	0.06283
			p^2=0.05373
	Heterozygotes (A_iA_j)		Result with D13 11,14 (p = 0.33940 q = 0.04801)
Full siblings	$\frac{(1+p_i+p_j+2p_ip_j)+(5+3p_i+3p_j-4p_ip_j)\theta+2(4-2p_i-2p_j+p_ip_j)\theta^2}{4(1+\theta)(1+2\theta)}$		0.35955
Parent and child	$\frac{2\theta+(1-\theta)(p_i+p_j)}{2(1+\theta)}$		0.19977
Half siblings	$\frac{(p_i+p_j+4p_ip_j)+(2+5p_i+5p_j+8p_ip_j)\theta+(8-6p_i-6p_j+4p_ip_j)\theta^2}{4(1+\theta)(1+2\theta)}$		0.11921
First cousins	$\frac{(p_i+p_j+12p_ip_j)+(2+13p_i+13p_j-24p_ip_j)\theta+2(8-7p_i-7p_j+6p_ip_j)\theta^2}{8(1+\theta)(1+2\theta)}$		0.07893
Unrelated	$\frac{2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	(NRC II, 4.10b)	0.03864 $2pq$= 0.03259

Formula from Weir B.S. (2003) Forensics. In Balding, D.J., Bishop, M. and Cannings, C. (eds) *Handbook of Statistical Genetics*, 2nd Edition, pp. 800-852. Hoboken, NJ.: John Wiley & Sons., p. 839

Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, Table 21.7, ©Elsevier Science/Academic Press

Obtain DNA from All Possible Suspects

- As expressed by NRC II Recommendation 4.4, "if possible contributors of the evidence sample include relatives of the suspect, **DNA profiles of those relatives should be obtained**."
- In other words, avoid the hypothetical and test the related individual in order to see if a direct match occurs between the evidence and the suspect…

Mixture Statistics

Approaches to Mixture Analysis and Statistics

Ladd *et al.* (2001) *Croatian Med. J.* 42(3): 244-246

- Qualitative Assessment (inclusion or exclusion of suspect)
- Deduction of Component Profiles followed by Calculation of Match Probabilities
- Probability of Exclusion (or Inclusion)
- Likelihood Ratio

ISFG Recommendations on Mixture Interpretation

July 13, 2006 issue of *Forensic Science International*

Available online at www.sciencedirect.com

SCIENCE DIRECT

Forensic Science International 160 (2006) 90–101

Forensic Science International

www.elsevier.com/locate/forsciint

DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures

P. Gill [a,*], C.H. Brenner [b], J.S. Buckleton [c], A. Carracedo [d], M. Krawczak [e], W.R. Mayr [f], N. Morling [g], M. Prinz [h], P.M. Schneider [i], B.S. Weir [j]

[a] Forensic Science Service, Trident Court, 2960 Solihull Parkway, Birmingham, UK
[b] Forensic Science Group, School of Public Health, University of California, Berkeley, CA 510-339-1911, USA
[c] ESR, Private Bag 92021, Auckland, New Zealand
[d] Institute of Legal Medicine, Faculty of Medicine, University of Santiago de Compostela, 15705 Santiago de Compostela, Spain
[e] Institute of Medical Informatics and Statistics, Kiel, Germany
[f] Division of Blood Group Serology, Medical University of Vienna, Austria
[g] Department of Forensic Genetics, Institute of Forensic Medicine, University of Copenhagen, Copenhagen, Denmark
[h] Office of the Chief Medical Examiner, Department of Forensic Biology, 520 First Avenue, New York, NY 10016, USA
[i] Institute of Legal Medicine University Clinic of Cologne, Melatengürtel, 60-62 D 30923 Köln, Germany
[j] University of Washington, Department of Biostatistics, Box 357232, Seattle, WA 98195, USA

Received 4 April 2006; accepted 10 April 2006

Discuss probability of exclusion and likelihood ratio methods

Statistical Calculations for Lineage Markers

Y-Chromosome and Mitochondrial DNA

Counting Method Typically Used for Lineage Markers

- Number of times that a particular DNA type occurs in a population database (frequency point estimate)
- Sampling corrections can be made with 95% confidence interval around the frequency point estimate