The Numbers behind DNA Analysis:
How do you get 1 in a trillion
from only testing a few hundred people?

John M. Butler, Ph.D.
National Institute of Standards and Technology

Lockheed Martin BEACON Lecture
Rockville, MD
August 19, 2009

NIST Applied Genetics Group

Applied Genetics

Group Leader

Formally organized October 2008

John Butler | **Marcia Holden** | **Margaret Kline** | **Pete Vallone**

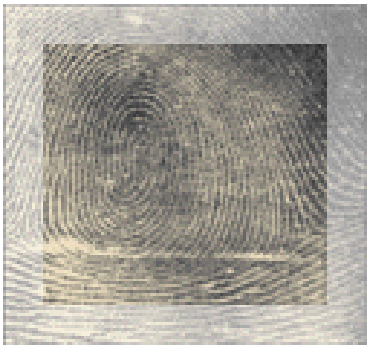
Amy Decker | **Ross Haynes** | **Becky Hill** | **Jan Redman**

Group Mission Statement

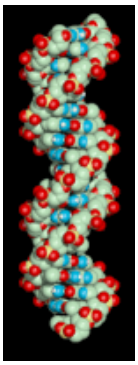
Advancing technology and traceability through quality genetic measurements to aid work in

- **forensic DNA testing,**
- **clinical genetics,**
- **agricultural biotechnology,** and
- **DNA biometrics.**

Methods for Human Identification/Biometrics



Fingerprints have been used since 1901



DNA since 1986

We are finding new ways to use DNA…

Steps in Forensic DNA Analysis

Usually 1-2 day process (a minimum of ~5 hours)

Steps Involved

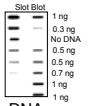
- Collection
- Specimen Storage
- Extraction
- Quantitation
- Multiplex PCR
- STR Typing
- Interpretation of Results
- Database Storage & Searching
- Calculation of Match Probability

Blood Stain | Buccal swab

Sample Collection & Storage

Biology

DNA Extraction



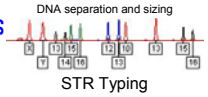
DNA Quantitation

Multiplex PCR Amplification

Genetics

If a match occurs, comparison of DNA profile to population allele frequencies to generate a case report with probability of a random match to an unrelated individual

Technology



STR Typing

Male: 13,14-15,16-12,13-10,13-15,16

Interpretation of Results

DNA Database Search

DNA Testing Requires a Reference Sample

A DNA profile by itself is fairly useless because it has no context…

DNA analysis for identity only works by comparison – **you need a reference sample**

Crime Scene Evidence compared to **Suspect(s)** (Forensic Case)
Child compared to **Alleged Father** (Paternity Case)
Victim's Remains compared to **Biological Relative** (Mass Disaster ID)
Soldier's Remains compared to **Direct Reference Sample** (Armed Forces ID)

Elements Going into the Calculation of a Rarity Estimate for a DNA Sample

2 Population allele frequencies

1 DNA Profile (with specific alleles)

Appropriate genetic formulas **3**

→ Rarity estimate of the specific DNA profile

There are different ways to express the profile rarity

Table 11.3 Random match probability for a 13-locus STR profile using the U.S. Caucasian allele frequencies found in Table 11.1.

	Allele 1	Allele 2	Allele 1 Frequency (p)	Allele 2 Frequency (q)	Formula	Expected Genotype Frequency
D13S317	11	14	0.33940	0.04801	2pq	0.0326
TH01	6	6	0.23179		p^2	0.0537
D18S51	14	16	0.13742	0.13907	2pq	0.0382
D21S11	28	30	0.15894	0.27815	2pq	0.0884
D3S1358	16	17	0.25331	0.21523	2pq	0.1090
D5S818	12	13	0.38411	0.14073	2pq	0.1081
D7S820	9	9	0.17715		p^2	0.0314
D8S1179	12	14	0.18543	0.16556	2pq	0.0614
CSF1PO	10	10	0.21689		p^2	0.0470
FGA	21	22	0.18543	0.21854	2pq	0.0810
D16S539	9	11	0.11258	0.32119	2pq	0.0723
TPOX	8	8	0.53477		p^2	0.2860
VWA	17	18	0.28146	0.20033	2pq	0.1128
AMEL	X	Y				
Product rule						1.20×10^{-15}
Combined frequency						$1 \text{ in } 8.37 \times 10^{14}$

1 in 837 trillion

John M. Butler (2009) Fundamentals of Forensic DNA Typing, Table 11.3

Generating a DNA Profile

Population allele frequencies

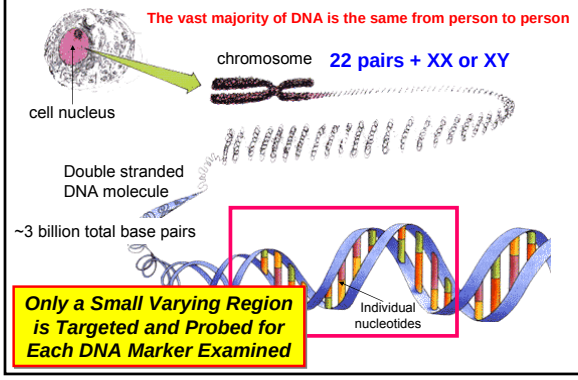
DNA Profile (with specific alleles)

Appropriate genetic formulas

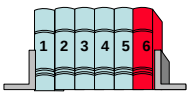
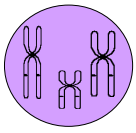
→ Rarity estimate of the specific DNA profile

DNA in the Cell

The vast majority of DNA is the same from person to person



Organization of Information

Printed	Genetic
	
Volumes in a Set of Encyclopedias	23 Pairs of Chromosomes in a Cell

Information Storage

You know that no two people share the same fingerprint, but did you know that the cells that make up your body also have a unique fingerprint unlike anyone else's? Your cells contain a complex molecule that we call DNA. Unless you have an identical twin, no one else has DNA just like yours.

Scientists can analyze DNA. If a criminal leaves DNA at a crime scene, police can use it to prove who committed the crime. At NIST, we help crime labs analyze DNA accurately. We make DNA standards so crime labs can tell if their results are right.

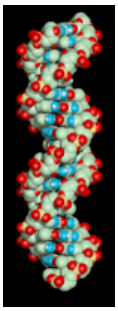
Text Storage is by the order of letters, words and paragraphs

DNA Storage is by the order of nucleotides, genes and chromosomes

Identification of Information

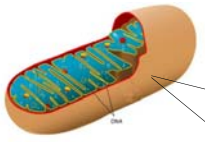
Printed Information	Genetic Information
Library	Body
Book	Cell
Chapter	Nucleus
Page Number	Chromosome
Line on Page	Locus (part of chromosome)
Word	Short DNA sequence
Letter	DNA nucleotides

Characteristics of DNA



- Each person has a unique DNA profile (except identical twins).
- Each person's DNA is the same in every cell.
- An individual's DNA profile remains the same throughout life.
- Half of your DNA comes from your mother and half from your father.

The Human DNA Genome within a Cell



The Nucleus = control center for the cell (one per cell)

Mitochondria = the power houses for the cell (hundreds per cell)

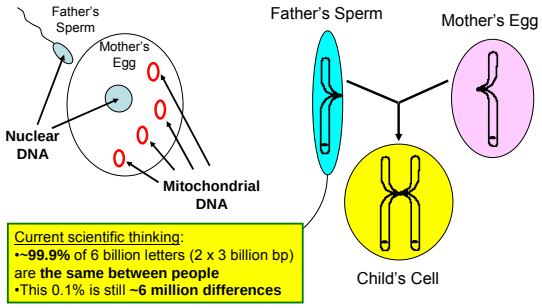
Mitochondrial DNA
(16,569 bp)

Inherited from only your mother

Nuclear DNA
(3.2 billion bp)

Inherited from both your mother and your father

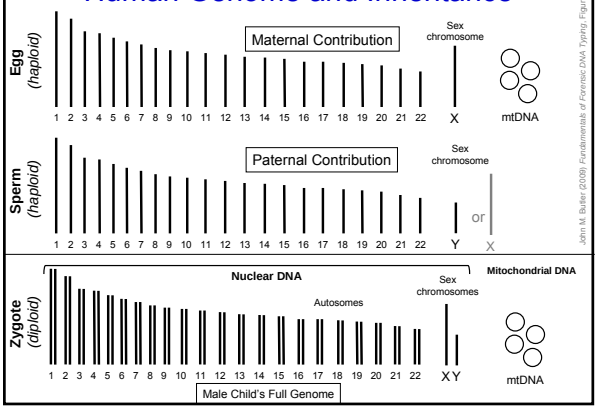
Genetic Inheritance



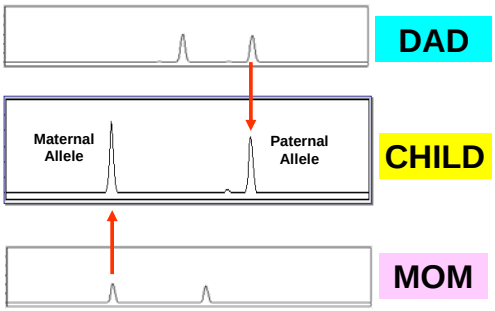
Current scientific thinking:
•~**99.9%** of 6 billion letters (2 x 3 billion bp) are **the same between people**
•This 0.1% is still **~6 million differences**

Father contributes: 22 autosomes (1 of each pair), X or **Y**
Mother contributes: 22 autosomes (1 of each pair), X and **mtDNA**

Human Genome and Inheritance

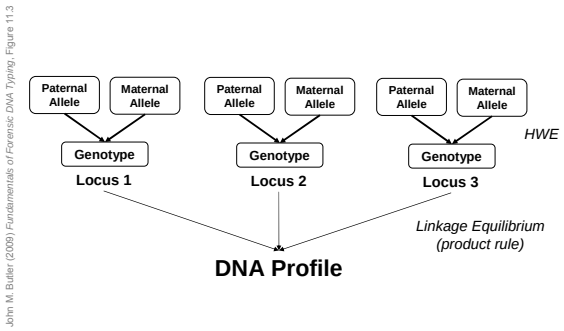


Inheritance Pattern of DNA Profiles



Result from a Single Locus (specific region of DNA)

Composition of a DNA Profile



Basis of DNA Profiling

The genome of **each individual is unique** (with the exception of identical twins) and **is inherited from parents**

Probe subsets of genetic variation in order to differentiate between individuals (**statistical probabilities of a random match are used**)

DNA typing must be **performed efficiently and reproducibly** (information must hold up in court)

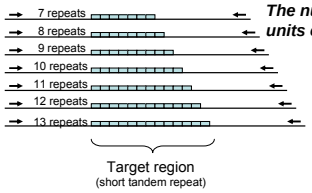
Current standard DNA tests **DO NOT look at genes** – little/no information about race, predisposal to disease, or phenotypical information (eye color, height, hair color) is obtained

Short Tandem Repeat (STR) Markers

An accordion-like DNA sequence that occurs between genes

TCCCAAGCTCTTCCTCTTCCCTAGATCAATACAGACAGAAGACA
GGTG**GATAGATAGATAGATAGATAGATAGATAGATAGATAGA**
TAGATATCATTGAAAGACAAAACAGAGATGGATGATAGATACAT
GCTTACAGATGCACAC

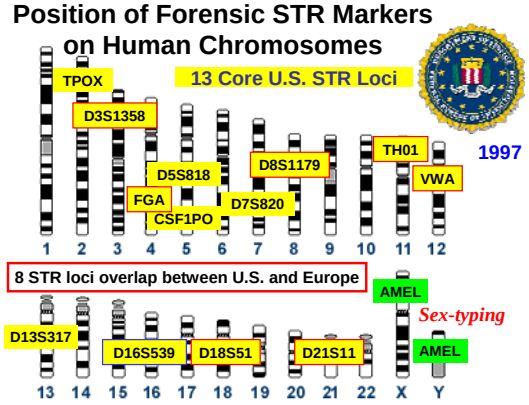
= 11 GATA repeats ("11" is all that is reported)



The number of consecutive repeat units can vary between people

The FBI has selected **13 core STR loci** that must be run in all DNA tests in order to provide a common currency with DNA profiles

Position of Forensic STR Markers on Human Chromosomes



DNA Marker Nomenclature

TH01

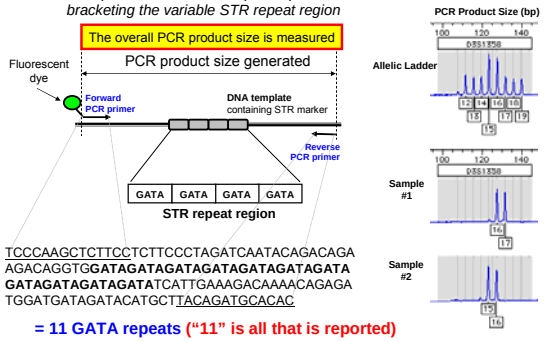
Tyrosine **H**ydoxylase gene, intron **01**

D16S539

D: **D**NA
16: chromosome **16**
S: **s**ingle copy sequence
539: **539**th locus described on chromosome 16

Short Tandem Repeat (STR) Markers

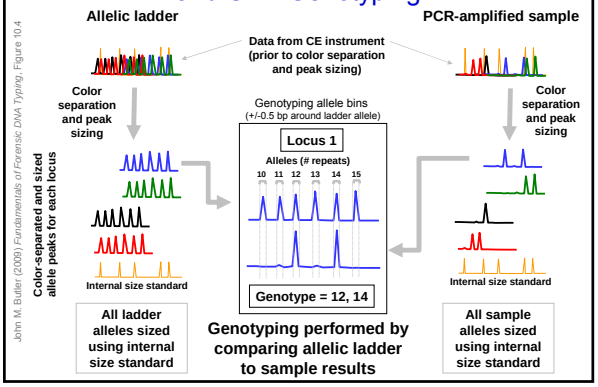
PCR primers anneal to unique sequences bracketing the variable STR repeat region



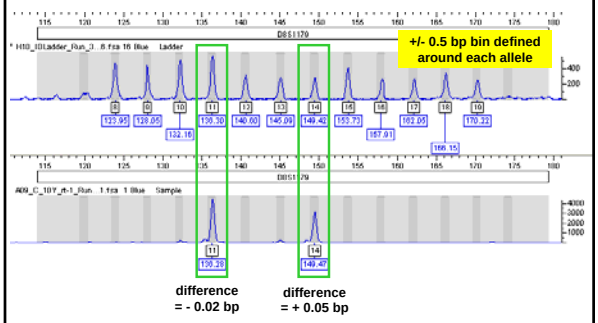
TCCCAAGCTCTTCCTCTTCCCTAGATCAATACAGACAGA
AGACAGGTG**GATAGATAGATAGATAGATAGATAGATA**
GATAGATAGATAGATATCATTGAAAGACAAAACAGAGA
TGGATGATAGATACATGCTTACAGATGCACAC

= 11 GATA repeats ("11" is all that is reported)

Peak Color Separation and Sizing and STR Genotyping

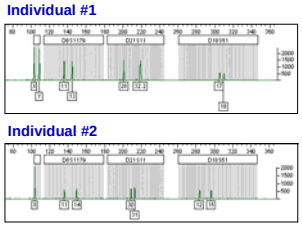


Comparison of Allelic Ladder to Samples to Convert Size into Allele Repeat Number



STR Results

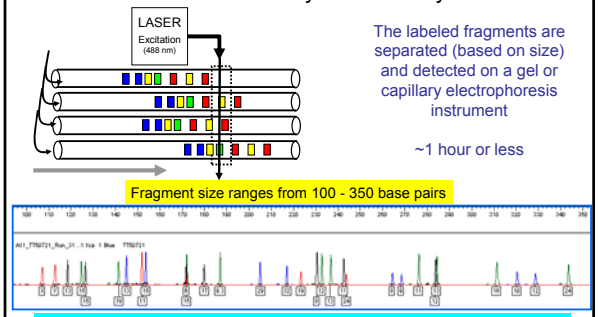
- Individuals will differ from one another in terms of their STR profile
- STR genotype can then be put into an alpha numeric form for search on a DNA database



What would be entered into a DNA database for searching:

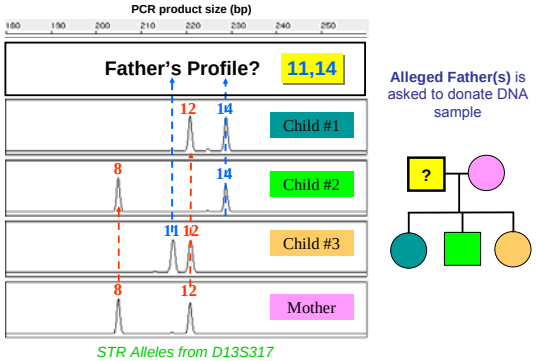
	AMEL	D8S1179	D21S11	D18S51
Individual #1	X,Y	11,13	28,32.2	17,18
Individual #2	X,X	11,14	30,31	12,15

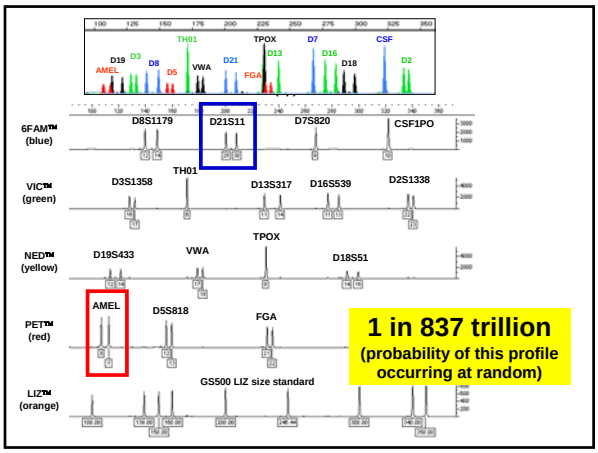
A DNA Profile is Produced by Separating DNA Molecules by Size and Dye Color



Peaks represent labeled DNA fragments separated by electrophoresis
This 'profile of peaks' is unique for an individual – a DNA type

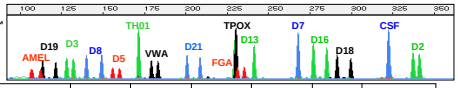
PATERNITY TESTING





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	value	allele	value	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	1.38×10^{9}
D7S820	9	*0.1772*			31.85	4.38×10^{10}
D16S539	9	*0.1126*	11	*0.3212*	13.8	6.05×10^{11}
THO1	6	*0.2318*			18.62	1.13×10^{13}
TPOX	8	*0.5348*			3.50	3.94×10^{13}
CSF1PO	10	*0.2169*			21.28	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})**

Population Allele Frequencies

Population allele frequencies

DNA Profile (with specific alleles)

Rarity estimate of the specific DNA profile

Appropriate genetic formulas

Determining the Frequency of Various STR Genotypes

Genotypes from 20 samples

11,12
8,11
11,11
11,12
13,13
12,12
12,12
11,11
12,13
12,13
11,14
8,8
11,13
11,12
11,12
10,13
11,11
8,11
9,11
13,14

Summary Count of Observed Genotypes

11,12 seen 4 times
8,11 seen 2 times
11,11 seen 3 times
13,13 seen 1 time
12,12 seen 2 times
12,13 seen 2 times
11,14 seen 1 time
8,8 seen 1 time
11,13 seen 1 time
10,13 seen 1 time
9,11 seen 1 time
13,14 seen 1 time

Summary Count of Observed Alleles

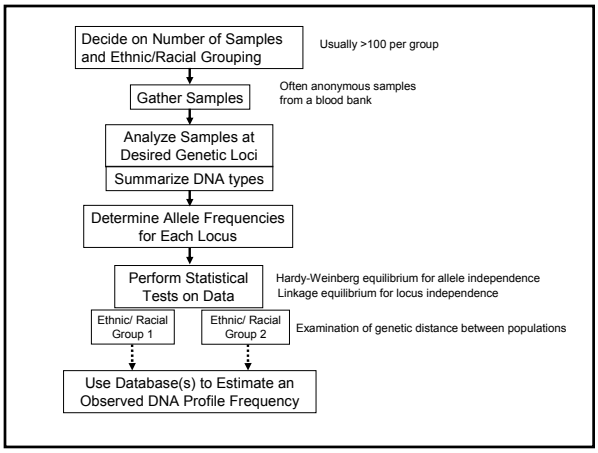
8 seen 4 times
9 seen 1 time
10 seen 1 time
11 seen 15 times
12 seen 10 times
13 seen 7 times
14 seen 2 times

Alleles Frequencies

8 = 4/40 = 0.10
9 = 1/40 = 0.025
10 = 1/40 = 0.025
11 = 15/40 = 0.375
12 = 10/40 = 0.25
13 = 7/40 = 0.175
14 = 2/40 = 0.05

***Allele 11 is the most common** and occurs 37.5% of the time in this sampling*

20 samples = 40 alleles



A Computer Program Used to Perform Statistical Analysis of STR Allele Frequencies

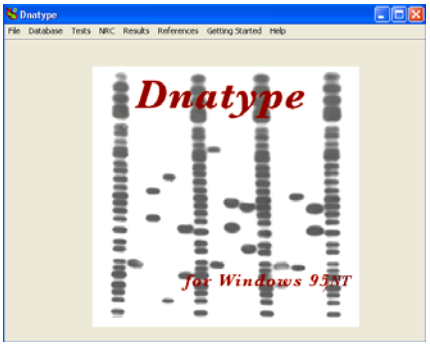


Figure 20.2, J.M. Butler (2005) *Forensic DNA Typing*, 2nd Edition © 2005 Elsevier Academic Press

Comparison of Allele Frequencies Measured with Different Studies

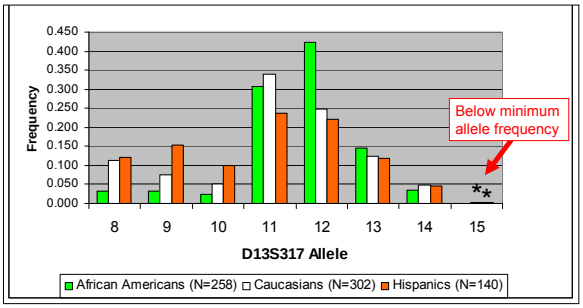
D13S317	African American	
Alleles	*N* = 7833	*N* = 258
7	0.0001	–
8	0.0260	0.0330
9	0.0218	0.0330
10	0.0273	0.0233
11	0.2940	0.3062
12	0.4290	0.4244
13	0.1520	0.1454
14	0.0486	0.0349
15	0.0010	–
16	0.0002	–
Minimum allele frequency (5/2N)	0.0003	0.0096

30 times more samples in the larger study yet the allele frequencies are fairly similar

Smaller Study
Butler *et al.* (2003) *JFS* 48(4):908-911

Larger Study
Einum *et al.* (2004) *JFS* 49(6): 1381-1385

Comparison of Allele Frequencies Between Different Population Groups



D13S317 Allele Frequencies from NIST U.S. Population Data

Allele	Caucasian *N* = 302	African-American *N* = 258	Hispanic *N* = 140
8	0.11258	0.03295	0.12143
9	0.07450	0.03295	0.15357
10	0.05132	0.02326	0.10000
11	**0.33940**	0.30620	**0.23571**
12	0.24834	**0.42442**	0.22143
13	0.12417	0.14535	0.11786
14	0.04801	0.03488	0.04643
15	0.00166*		0.00357*

5/(2x302) = 0.00828 5/(2x140) = 0.01786

Minimum Allele Frequency = **5**/2N

N = number of individuals tested
2N = number of chromosomes sampled (2 per person)

Want to sample at least 5 chromosomes to provide a somewhat reliable estimate of an allele's frequency in a population

Data behind FBI PopStats Program

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*

There was little evidence for departures from Hardy-Weinberg expectations (HWE) in any of the populations.

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.

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

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^{4}
D5S818	12,13	166	2.28×10^{3} to 4.22×10^{11}	4.51×10^{3}
D13S317	11,14	166	4.32×10^{4} to 1.69×10^{13}	1.38×10^{4}
D7S820	9,9	166	1.17×10^{10} to 2.98×10^{16}	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}

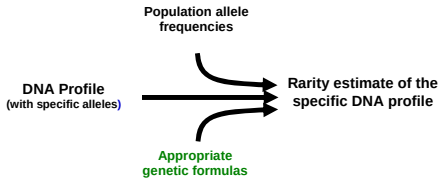
10^{14} to 10^{21}

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

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)

Applying Genetic Models and Formulas



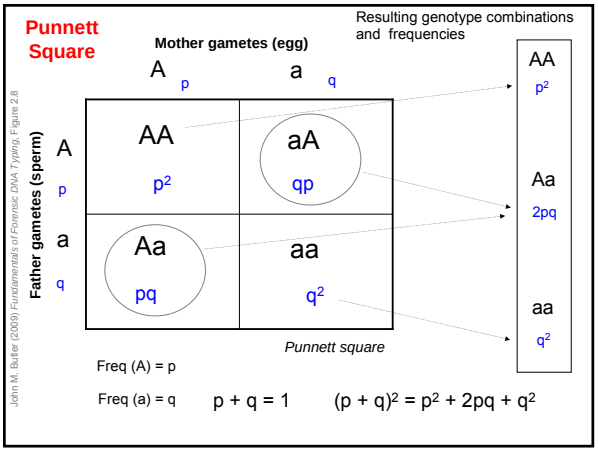

Godfrey Hardy

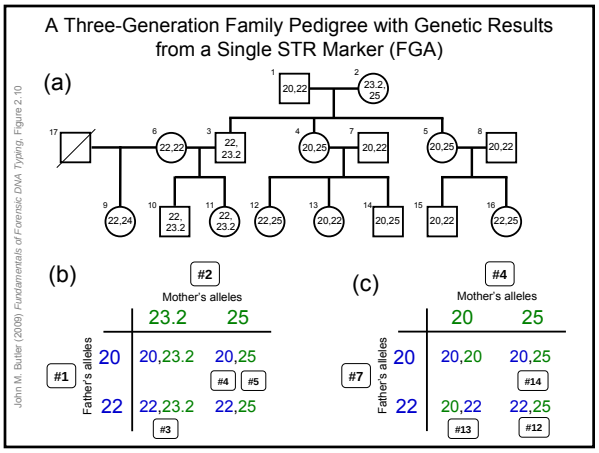
Hardy–Weinberg Equilibrium (HWE)

$p^2 + 2pq + q^2 = 1$

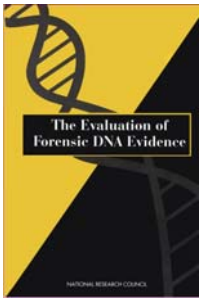

Wilhelm Weinberg

- Godfrey Hardy (1877–1947) and Wilhelm Weinberg (1862–1937) both independently discovered the mathematics for independent assortment that is now associated with their names as the Hardy–Weinberg principle.
- HWE proportions of genotype frequencies can be reached in a single generation of random mating. **HWE is simply a way to relate allele frequencies to genotype frequencies**.





The Second National Research Council Report (**NRC II**) Published in 1996



- Recommends various formulas to use to correct for inbreeding (subpopulation structure)
- Theta (θ) is a measure of the average level of co-ancestry (i.e., inbreeding)
  - Usually <0.01 with normal groups
  - Usually <0.03 with closed populations (e.g., Native American tribes)

"*Inbreeding* means mating of two persons who are more closely related than if they were chosen at random" (NRC II, p. 98).

Comparison of STR Genotype Frequencies with Different Correction Factors

Table 11.2 Comparison of statistical treatment for homozygotes and heterozygotes under different assumptions.

	Under HWE	Unconditional (NRC II Recommendation 4.1)	Conditional with Substructure Adjustment
			(NRC II 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** θ = **0.01**	$(0.23)^2$ = **0.053**	$(0.23)^2 + (0.23)(1-0.23)(0.01)$ = 0.053 + 0.0018 = **0.055**	[(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**
			(NRC II Recommendation 4.10b)
Heterozygote	$2p_i p_j$	$2p_i p_j$	$\frac{2[(p_i(1-\theta)+\theta][p_j(1-\theta)+\theta]}{(1+\theta)(1+2\theta)}$
D13S317 **11,14** p_i = **0.34** p_j = **0.05** θ = **0.01**	2(0.34)(0.05) = **0.0340**	2(0.34)(0.05) = **0.0340**	2[(0.34)(1 − 0.01) + 0.01][(0.05)(1 − 0.01) + 0.01] / (1 + 0.01)(1 + 2(0.01)) = 2(0.346)(0.0595)/(1.01)(1.02) = **0.0400**

Note: Allele frequency values (p_i, p_j) for the TH01 and D13S317 example data are from Table 11.1 (U.S. Caucasians). Note that if θ is zero, then unconditional and conditional formulas collapse to their Hardy–Weinberg equilibrium (HWE) functions.

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						Under HWE		NRCII Recommendation 4.1			NRCII Recommendation 4.10	
	A1	A2	Allele 1 freq (p)	Allele 2 freq (q)		Calc freq		θ = 0.01	θ = 0.03		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	0.0628	0.0821
D18S51	14	16	0.13742	0.12907	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
D3S1358	16	17	0.25331	0.21523	2pq	0.1090	2pq	0.1090	0.1090	eq. 4.10b	0.1129	0.1206
D5S818	12	13	0.38411	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	0.0456
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

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						Under HWE		NRCII Recommendation 4.4			
	A1	A2	Allele 1 freq (p)	Allele 2 freq (q)		Calc freq		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	0.0982	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 0.3745
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

How Are Such Large Numbers Generated with Random Match Probabilities?

- Each allele is sampled multiple times to produce a statistically stable allele frequency
- Using theoretical models from genetics, multiple loci are multiplied together to produce an estimate of the rarity of a particular DNA profile (combination of STR alleles based on individual allele frequencies)
- Remember that relatives will share genetic characteristics and thus have STR profiles that are more similar to one another than unrelated individuals
- We are not looking at every person on the planet nor are we looking at every nucleotide in the suspect's genome

Three DNA Forensic Categories Typically Faced

- **Single Source**: DNA profile of the evidence sample providing indications of it being of a single source origin
- **Mixture of DNA**: Evidence sample DNA profile suggests it being a mixture of DNA from multiple (more than one) individuals
- **Kinship Determination**: Evidence sample DNA profile compared with that of one or more reference profiles is to be used to determine the validity of stated biological relatedness among individuals

http://www.promega.com/geneticidproc/ussymp17proc/workshops/PromegaMixtureStats2006.pdf

The Three Possible Outcomes

of Evidence Examination (**Q-K Comparison**)

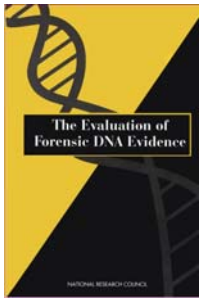
	"Suspect" Known (K) Sample	"Evidence" Question (Q) Sample
• **Exclusion** (no match)	11 12	13
• **Non-exclusion** – "Match" or "inclusion"	11 12	11 12

Must provide significance of this match

- **Inconclusive result** — 11 12 — No result (or a complex mixture)

(no decision as there is insufficient data to support a conclusion)

Inclusions (Matches) Require Statistics



- 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

The Statistic (Determining the Weight of the Evidence) Should Be **Calculated from the Evidence**

Evidence (partial profile):

	Type	Statistic
Locus 1	16,17	1 in 9
Locus 2	17,18	1 in 9
Locus 3	21,22	1 in 12
Locus 4	12,14	1 in 16
Locus 5	28,30	1 in 11

		Product = 1 in 171,000

Match Observed at All Loci that May Be Compared

Reference (full profile):

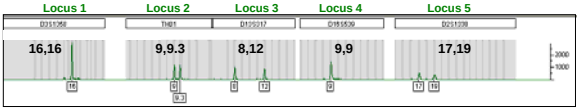
	Type	Statistic
Locus 1	16,17	1 in 9
Locus 2	17,18	1 in 9
Locus 3	21,22	1 in 12
Locus 4	12,14	1 in 16
Locus 5	28,30	1 in 11
Locus 6	14,16	1 in 26
Locus 7	12,13	1 in 9
Locus 8	11,14	1 in 31
Locus 9	9,9	1 in 32
Locus 10	9,11	1 in 14
Locus 11	6,6	1 in 19
Locus 12	8,8	1 in 3
Locus 13	10,10	1 in 21

		Product = 1 in 665 trillion

The reference sample is still a "match" – just not as much information is available from the evidence for comparison

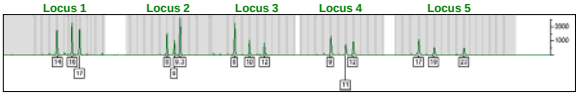
Single Source vs. Mixture Samples

Single Source Sample



One or two peaks observed at each locus (tested DNA region)

Mixture Sample



More than two peaks observed at more than two loci (tested DNA regions)

Different possible combinations could have given rise to the particular mixture observed

Thank you for your attention…

Our team publications and presentations are available at:
http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm

See also http://www.dna.gov/research/nist
http://www.cstl.nist.gov/biotech/strbase
john.butler@nist.gov