

# Basics

- Interpretation
- Hardy-Weinberg equations
- Random Match Probability
- Likelihood Ratio
- Substructure

# Three Types of DNA Forensic Issues

- **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

- Interpretation of a result:
  - 1. Non-match - exclusion
  - 2. Inconclusive - no decision
  - 3. Match - estimate frequency

# What is an Exclusion?

**Single Source:** DNA profiles of the evidence and reference samples differ from each other at one or more loci; i.e., barring sample mix-up and/or false identity of samples, reference individual is not the source of DNA found in the evidence sample

**DNA Mixture:** Reference DNA profile contains alleles (definitely) not observed in the evidence sample for one or more loci; i.e., reference individual is excluded as a part contributor of the mixture DNA of the evidence sample

**Kinship:** Allele sharing among evidence and reference samples disagrees with the Mendelian rules of transmission of alleles with the stated relationship being tested

# What is an Inclusion?

**Single Source:** DNA profiles of the evidence and reference samples are identical at each interpretable locus (also called **DNA Match**); i.e., reference individual may be the source of DNA in the evidence sample

**DNA Mixture:** Alleles found in the reference sample are all present in the mixture; i.e., reference individual can not be excluded as a part contributor of DNA in the evidence sample

**Kinship:** Allele sharing among evidence and reference samples is consistent with Mendelian rules of transmission of alleles with the stated relationship being tested; i.e., the stated biological relationship cannot be rejected

# When is the Observation at a Locus Inconclusive?

- Compromised nature of samples tested failed to definitively exclude or include reference individuals
- May occur for one or more loci, while other loci typed may lead to unequivocal definite inclusion/exclusion conclusions
- Caused often by DNA degradation (resulting in allele drop out), and/or low concentration of DNA (resulting in alleles with low peak height and/or area) for the evidence sample

# Statistical Assessment of DNA Evidence

- Needed most frequently with an inclusion
- (Apparent) exclusionary cases may also be sometimes subjected to statistical assessment, particularly for kinship determination because of genetic events such as mutation, recombination, etc.
- Loci providing inconclusive results are often excluded from statistical considerations
- Even if one or more loci show inconclusive results, inclusionary observations of the other typed loci can be subjected to statistical assessment

# Exclusion vs Match

- Exclusion – numbers are not needed
- Match - requires a numerical estimate (weight of evidence)



# Statistical Analysis

About Evidence sample “Q”

- “K” matches “Q”
- Who else could match “Q”
- Who is in suspect population?
- partial profile, mixtures

# Estimate genotype frequency

1. Frequency at each locus

Hardy-Weinberg Equilibrium

2. Frequency across all loci

Linkage Equilibrium (multiply)

# Human Beings

23 different chromosomes

2 sets of chromosomes (from mom and dad) – two copies of each marker

Each genetic marker on different chromosome

Thus, each marker treated like coin toss – two possibilities

# Hardy - Weinberg Equilibrium

$$\frac{A_1A_1}{p_1^2} \quad \frac{A_1A_2}{2p_1p_2} \quad \frac{A_2A_2}{p_2^2}$$

$$\text{freq}(A_1) = p_1$$

$$\text{freq}(A_2) = p_2$$

	A <sub>1</sub>	A <sub>2</sub>
A <sub>1</sub>	$p_1^2$ A <sub>1</sub> A <sub>1</sub>	$p_1p_2$ A <sub>1</sub> A <sub>2</sub>
A <sub>2</sub>	$p_1p_2$ A <sub>1</sub> A <sub>2</sub>	$p_2^2$ A <sub>2</sub> A <sub>2</sub>

$$(p_1 + p_2)^2 = p_1^2 + 2p_1p_2 + p_2^2$$

# Alleles in populations – The Hardy-Weinberg Theory

Basis: Allele frequencies are inherited in a Mendelian fashion and frequencies of occurrence follow a predictable pattern of probability

# A Hardy-Weinberg Population

- LARGE POPULATION
- NO NATURAL SELECTION
- NO MUTATION
- NO IMMIGRATION / EMIGRATION
- RANDOM MATING

# A Hardy-Weinberg Population

We don't care these about criteria!

Only concerned about alleles...

The *Hardy-Weinberg* principle states:  
that single-locus genotype frequencies after  
one generation of random mating can be  
represented by a binomial (with two alleles)  
or multinomial (with multiple alleles)  
function of the alleles frequencies



# Hardy - Weinberg Equilibrium

## Two Allele System

$$\text{freq}(A_1) = p_1 \quad \text{freq}(A_2) = p_2$$

$$p_1 + p_2 = 1$$

$$(p_1 + p_2)^2 = 1^2$$

$$P_1^2 + 2p_1p_2 + P_2^2 = 1$$

$$A_1A_1 \quad A_1A_2 \quad A_2A_2$$



# Probability is ...

- frequency of an event in a large number of trials
- “frequentist”
- subjective degree of belief
- “Bayesian”

# Approaches for Statistical Assessment of DNA Evidence

**Frequentist Approach:** indicating the coincidental chance of the event observed

**Likelihood Approach:** indicating relative support of the event observed under two contrasting (mutually exclusive) stipulations regarding the source of the evidence sample

**Bayesian Approach:** providing a posterior probability regarding the source, when data in hand is considered with a prior probability of the knowledge of the source (latter is not generally provided by the DNA profiles being considered for statistical assessment)

# People vs Collins (1968, California)

- Partly yellow car  $1/10$
- Man with mustache  $1/4$
- Girl with ponytail  $1/10$
- Girl with blond hair  $1/3$
- Black man with beard  $1/10$
- Interracial couple in car  $1/1000$
- Estimate  $1/12,000,000$

# Frequentist Approach of Statistical Assessment for Transfer Evidence

- When the evidence sample DNA profile matches that of the reference sample, one or more of the following questions are asked:
- How often a random person would provide such a DNA match? Equivalently, what is the expected frequency of the profile observed in the evidence sample? – also called Random Match Probability, complement of which is the Exclusion Probability
- What is the expected frequency of the profile seen in the evidence sample, given that it is observed in another person (namely in the reference sample) – also called Conditional Match Probability
- What would be the expected frequency of the profile seen in the evidence sample in a relative (of specified kinship) of the reference individual, given the DNA match of the reference and evidence samples – also called the Match Probability in Relatives

## Bayes formula (odds form):

$$\left( \frac{P(H_1 | E)}{P(H_2 | E)} \right) = \left( \frac{P(E | H_1)}{P(E | H_2)} \right) \times \left( \frac{P(H_1)}{P(H_2)} \right)$$

posterior odds = likelihood ratio x prior odds

E = DNA evidence

H<sub>1</sub> = alleged father is biological father

H<sub>2</sub> = alleged father is not biological father

# Likelihood Ratio

$$\text{LR} = \frac{P(E | H_1)}{P(E | H_2)}$$

$E$  = DNA evidence

$H_1$  = Suspect is the source of the DNA

$H_2$  = Suspect is not the source of the DNA



# Random Match Probability

- Estimate frequencies of genotype at a locus
- Use product rule
- Correct for departures due to inbreeding ( $\theta$ / $F_{st}$ )
- Multiply estimated genotype frequency of each locus assuming independence among loci (biological basis)
- Correct for sampling (10 fold rule)

		Mom	
		15 (p)	17 (q)
Dad	15 (p)	15,15	15,17
	17 (q)	15,17	17,17

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

Remember this is based on the relationship between allele and genotype frequencies

# Population

Database samples are typically "convenience" samples that have been obtained from blood banks, parentage labs, sometime even Convicted Felon database samples

A major characteristic of these samples is self-declaration regarding "population affinity" ...  
i.e. Caucasian, Asian, Hispanic, African, etc.

Databases may also be defined based on region...country, state, city, etc.

# Population database

- Look up how often each allele occurs at the locus in a population (or populations)
- looking up the “allele” frequency

### Profiler Plus

Item	D3S1358	vWA	FGA	D8S1179	D21S11	D18S51	D5S818	D13S317	D7S820
Q1	16,16	15,17	21,22	13,13	29,30	16,20	8,12	12,12	8,11

### CoFiler

Item	D3S1358	D16S539	TH01	TPOX	CSF1P0	D7S820
Q1	16,16	10,12	8,9.3	9,10	12,12	8,11

D3S1358 = 16, 16 (homozygote)

Frequency of 16 allele = ??

TABLE 1—Observed allele distributions (as %) for 13 STR loci in six population groups.

D3S1358	African American (N=210)	Bahamian (N=157)	Jamaican (N=194)	Trinidad (N=80)	Caucasian (N=203)	Hispanic (N=209)
<12	0.476	0.000	0.000	0.000	0.000	0.000
12	0.238	0.000	0.515	0.000	0.000	0.000
13	1.190	0.000	1.546	0.000	0.246	0.239
14	12.143	7.643	6.701	5.625	14.039	7.895
15	29.048	31.847	33.763	31.250	24.631	42.584
15.2	0.000	0.318	0.258	0.000	0.000	0.000
16	30.714	33.758	30.670	31.875	23.153	26.555
17	20.000	19.745	21.134	20.000	21.182	12.679
18	5.476	6.369	4.639	11.250	16.256	8.373
19	0.476	0.318	0.773	0.000	0.493	1.435
>19	0.238	0.000	0.000	0.000	0.000	0.239
Homozygosity (Obs.)	21.4%	25.5%	27.8%	16.3%	19.2%	26.3%
Homozygosity (Exp.)	23.5%	26.2%	25.8%	25.0%	20.3%	28.0%
(p)	0.482	0.838	0.513	0.070	0.691	0.595
Exact Test	0.797	0.758	0.270	0.222	0.084	0.333
PD	0.903	0.885	0.886	0.878	0.920	0.880
PE	0.543	0.499	0.508	0.511	0.589	0.492

D3S1358 = 16, 16 (homozygote)

Frequency of 16 allele = 0.3071

When same allele:

Genotype Frequency =  $p^2$   
(for now!)

Genotype freq =  $0.3071 \times 0.3071 = 0.0943$



VWA = 15, 17 (heterozygote)

Frequency of 15 allele = ??

Frequency of 17 allele = ??

VWA	African American (N=180)	Bahamian (N=162)	Jamaican (N=244)	Trinidad (N=85)	Caucasian (N=196)	Hispanic (N=203)
11	0.278	0.926	0.410	0.588	0.000	0.246
13	0.556	2.778	0.820	0.588	0.510	0.000
14	6.667	6.173	7.377	8.824	10.204	6.158
15	23.611	15.123	22.746	14.118	11.224	7.635
16	26.944	26.235	29.098	29.412	20.153	35.961
17	18.333	20.679	18.238	26.471	26.276	22.167
18	13.611	18.210	13.115	13.529	22.194	19.458
19	7.222	7.099	5.328	4.706	8.418	7.143
20	2.778	2.778	2.254	1.765	1.020	1.232
21	0.000	0.000	0.615	0.000	0.000	0.000
Homozygosity (Obs.)	11.7%	17.3%	20.9%	20.0%	22.4%	24.6%
Homozygosity (Exp.)	18.9%	17.6%	19.4%	20.0%	18.7%	22.9%
(p)	0.014	0.928	0.557	0.991	0.179	0.564
Exact Test	0.328	0.790	0.655	0.229	0.063	0.928
PD	0.926	0.942	0.933	0.917	0.932	0.914
PE	0.624	0.648	0.617	0.602	0.625	0.563

VWA = 15, 17 (heterozygote)

Frequency of 15 allele = 0.2361

Frequency of 17 allele = 0.1833

When heterozygous:

Frequency = 2 X allele 1 freq X allele 2 freq  
( $2pq$ )

Genotype freq = 2 x 0.2361 x 0.1833 = 0.0866

Ideally, we should know the frequency of every genotype that might be encountered

Do we?

**Caucasian Database for Locus yWA**

**N = 196 Individuals**

	11	12	13	14	15	16	17	18	19	20	21
11											
12											
13						1			1		
14				1	4	10	10	10	4		
15					3	7	14	8	4	1	
16						11	27	7	3	1	
17							11	23	8	1	
18								16	6		
19									3	1	
20											
21											

**66 Possible Genotypes (N)(N+1)/2**  
**27 Genotypes Seen In Caucasians**

# Minimal Allele Frequency

Requires a minimum of 5 copies of an allele before the allele frequency can be used for calculation of genotype frequency

5

---

Total number of alleles at locus

For the 13 allele at vWA:

$$\text{Actual Freq} = 2 / 392 = 0.0051$$

$$\text{Minimal Freq} = 5 / 392 = 0.0128$$

# 5/2N

- N            min allele freq
- 100            2.50 %
- 150            1.67 %
- 200            1.25 %
- 250            1.00 %
- 300            0.83 %



# Minimum allele frequency

- Weir, B.S. 1992 & Nelson 1965.  
 $\text{minfreq} = 1 - \frac{1}{2N}$
- Budowle, B., K. Monson, R. Chakraborty, 1996.  
 $\text{minfreq} = 1 - [1 - (1 - \frac{1}{2N})^{1/C}]^{1/2N}$
- NRC II, 1996 & Budowle et al 1991.  
 $\text{minfreq} = \frac{1}{2N}$

# Minimum allele frequency comparisons ( $\alpha = .05$ $c=8$ )

• <u>N</u>	<u>Weir</u>	<u>Budowle</u>	<u>5/2N</u>
• 100	1.48%	2.49%	2.50 %
• 150	0.99%	1.67%	1.67 %
• 200	0.75%	1.26%	1.25 %
• 300	0.49%	0.84%	0.83 %
• 400	0.37%	0.63%	0.66 %

# Minimum allele frequency comparisons ( $\alpha = .05$ $c=16$ )

• <u>N</u>	<u>Weir</u>	<u>Budowle</u>	<u>5/2N</u>
• 100	1.48%	2.83%	2.50 %
• 150	0.99%	1.90%	1.67 %
• 200	0.75%	1.43%	1.25 %
• 300	0.49%	0.95%	0.83 %
• 400	0.37%	0.72%	0.66 %

# What do minimal allele frequencies provide?

- Sampling error correction
- Minimize population substructure effects

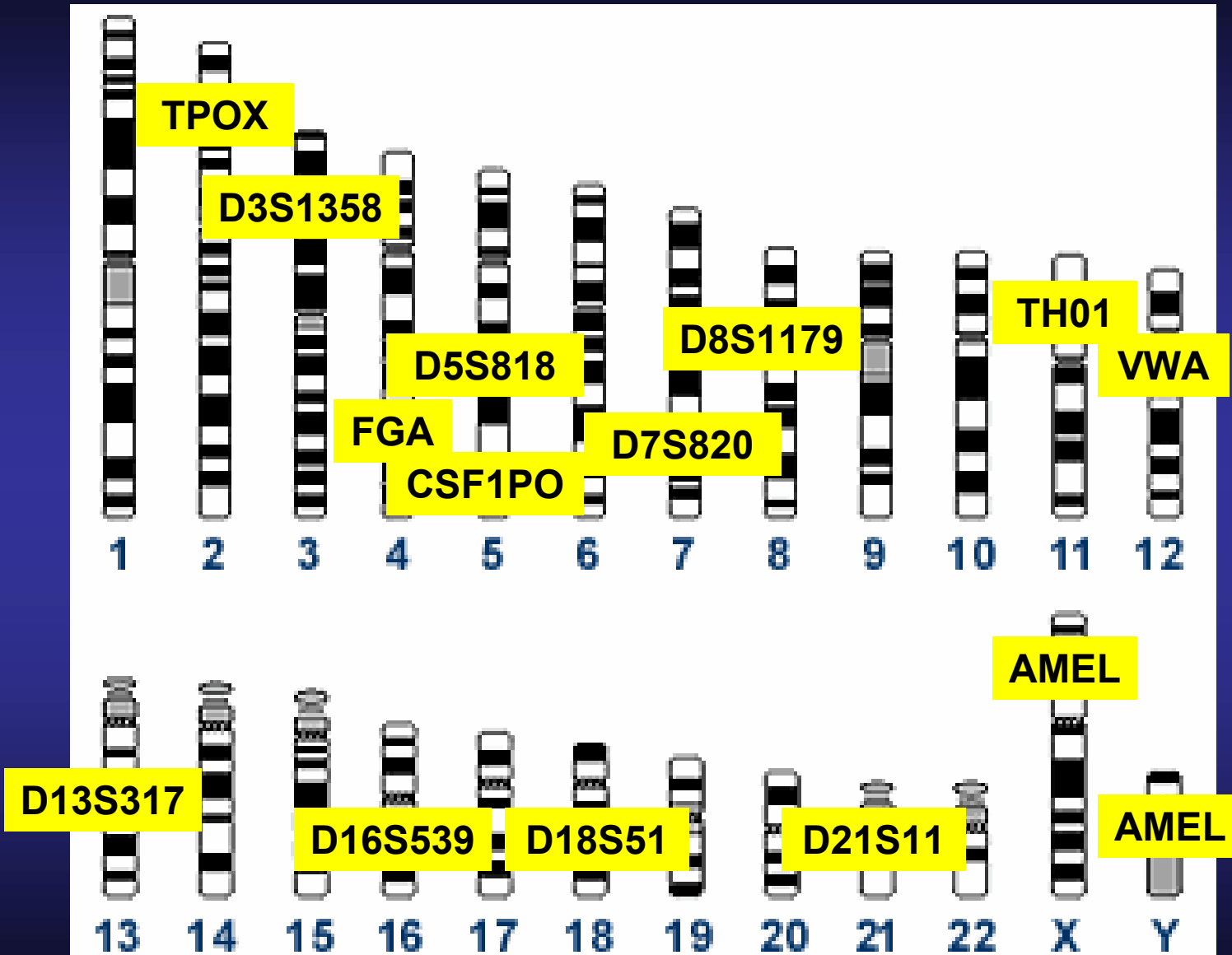
Everything done so far pertains to  
a single locus

Linkage equilibrium between two loci means that a genotype/allele at one locus is not associated with a genotype/allele at another locus

Linkage disequilibrium occurs routinely for Y chromosome loci and also mtDNA data

Linkage disequilibrium can exist because of population substructure or because of physical linkage

# 13 CODIS Core STR Loci with Chromosomal Positions



Biological Basis

# Profile Frequency Estimates Across Multiple Loci

Employ the PRODUCT RULE



# Product Rule

The frequency of a multi-locus STR profile is the product of the genotype frequencies at the individual loci

$$f_{\text{locus}_1} \times f_{\text{locus}_2} \times f_{\text{locus}_n} = f_{\text{combined}}$$

# Criteria for Use of Product Rule

Inheritance of alleles at one locus have no effect on alleles inherited at other loci

Loci are in linkage equilibrium

Overall profile frequency =

Frequency D3S1358 X Frequency vWA

$$0.0943 \times 0.0866 = 0.00817$$

## Steps – Single Source Target Profile

- Identify alleles of target profile
  - Look up allele frequencies for all loci in all appropriate populations
  - Determine if homozygous or heterozygous profile at each locus
  - Calculate genotype frequency at each locus
  - Calculate profile frequency with product rule
- 
- **Correct for sampling error – 10 fold less**

Random match probability = .000001

---

Random match probability = 1/1,000,000

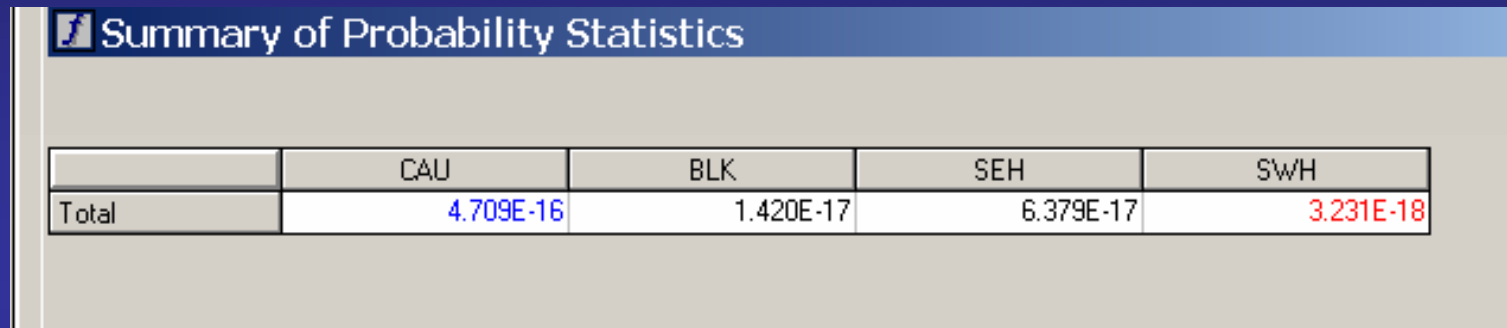
---

Exclusion probability = .999999

Exclusion probability = 99.9999%

# What do these numbers mean?

## Random Match Probability



	CAU	BLK	SEH	SWH
Total	4.709E-16	1.420E-17	6.379E-17	3.231E-18

This is the actual probability of seeing profile/genotype in the metapopulation

(Given that the databases provide a reasonable representation of the population)

13 CODIS loci typically yield  
extraordinarily small probabilities

---

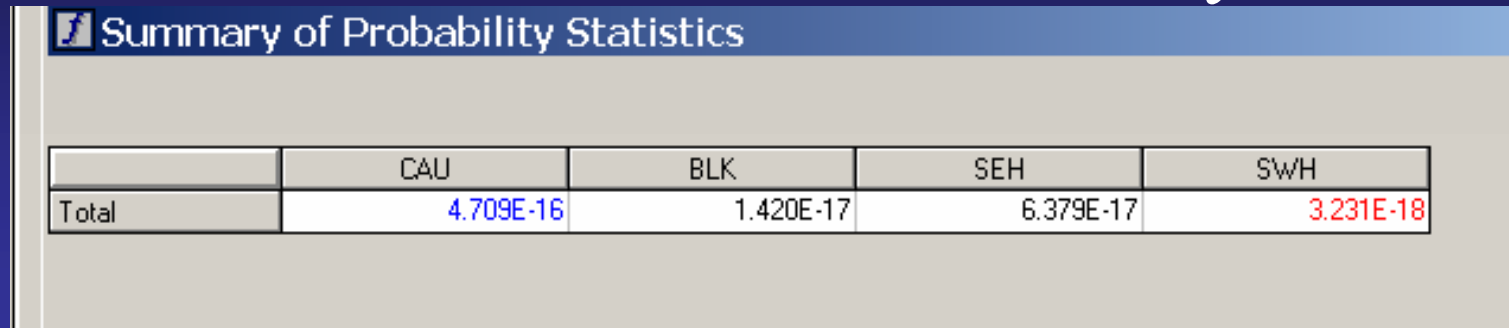
0.000000000000000000154

or

1 in 60,000,000,000,000,000 persons

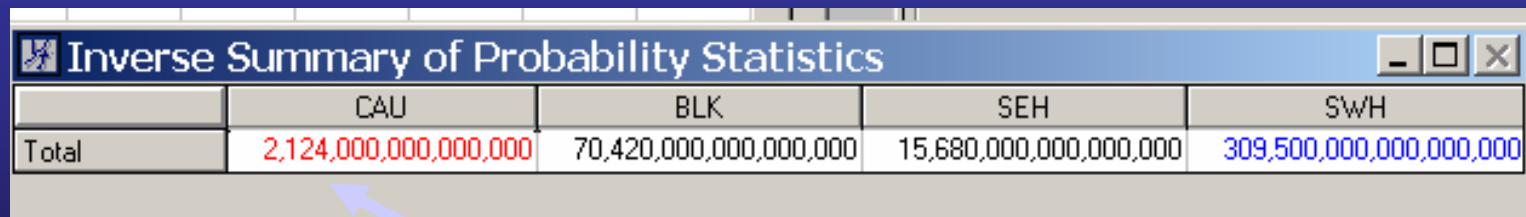
# What do these numbers mean?

## Random Match Probability



	CAU	BLK	SEH	SWH
Total	4.709E-16	1.420E-17	6.379E-17	3.231E-18

And then you have this...



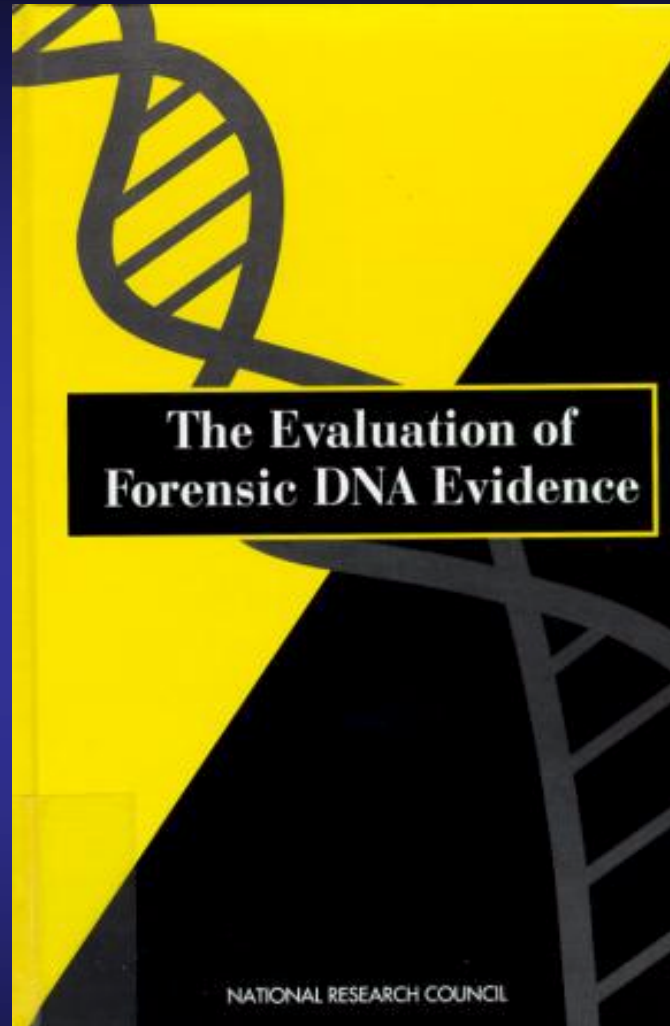
	CAU	BLK	SEH	SWH
Total	2,124,000,000,000,000	70,420,000,000,000,000	15,680,000,000,000,000	309,500,000,000,000,000

quadrillions???

And how many people are there??



# National Research Council Report II



National Academy of Sciences  
May 1996

# Population Structure

Racial, ethnic subgroups

Excess of homozygotes

What is “theta”  $\theta$

Modify only homozygote calculation?

NRC Formula 4.1 vs 4.4 vs 4.10

# Population Sub-Structure

Racial/ethnic group composed of distinct sub-groups within the sample population

Only a concern if sub-groups differ substantially at allele frequencies at the loci

# Problems created by population subdivision

Genotype frequencies calculated from  
population average allele

frequencies **could** lead to:

– Wrong estimates!

# Human Genetic Variation

- Individual
- Among population within a major population group
- Among major population groups

# Employ a Theta ( $\theta$ ) Correction

$\theta$  is used as a measure of the effects of population subdivision (inbreeding)

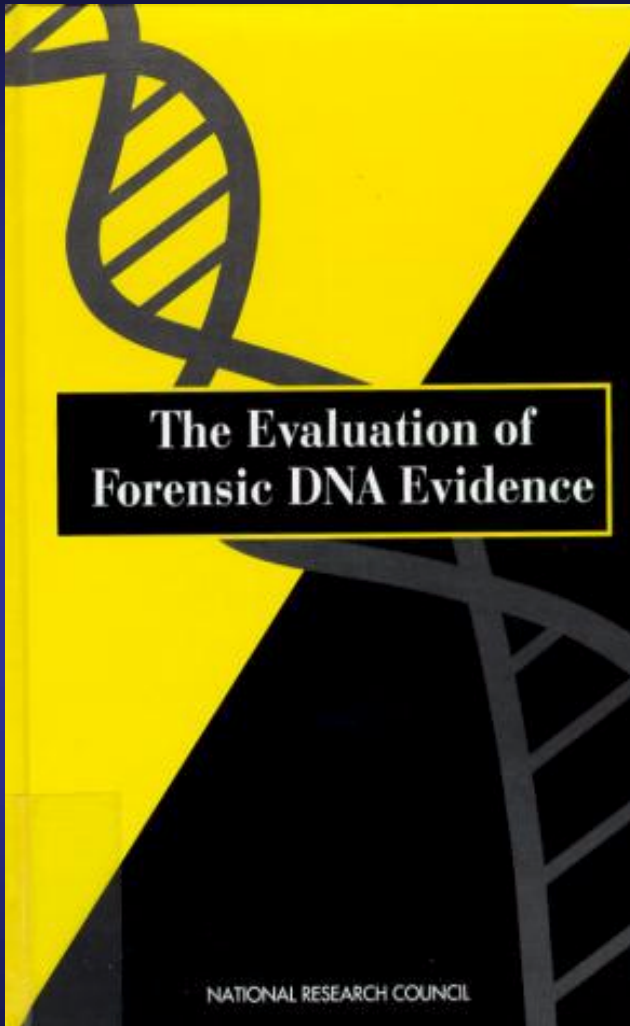
How many Great, Great, Great, Great, Great, Great, Great... Grandparents do you have?

In place of  $F_{ST}$  the parameter  $\theta$  was introduced (Weir and Cockerham, 1984) to clarify some of the nomenclature surrounding population evaluations and address some of the issues not carried by the F-Statistic model of Wright

By definition,  $\theta$  represents the correlation of genes of different individuals within the same population

$\theta$  is affected by population size and history, but unaffected by allele number, sample size, or number of populations

# National Research Council Report II



The significance of this theta or  $F_{ST}$   
is

That Hardy-Weinberg expectations  
are assumed not to be met



HWE:  $p^2$

NRC II, 4.4a:  $p^2 + p(1 - p)\theta$

NRC II, 4.10a: 
$$\frac{[2\theta + (1 - \theta)p_i][3\theta + (1 - \theta)p_i]}{(1 + \theta)(1 + 2\theta)}$$

This last formula addresses a conditional probability of the suspect genotype, given that of the perpetrator,  $P(A_i A_i | A_i A_i)$ , considering the person contributing the evidence and the suspect are from the same subgroup

Takes into account the assumption that the person contributing the evidence and the suspect are from the same subgroup

A conditional probability of the suspect genotype given that we have already seen that genotype in the perpetrator

Example... use if the suspect and all possible perpetrators are from the same small isolated population

Although **CAN** correct the heterozygote genotype estimate...it is **not** generally necessary

HWE:  $2pq$

NRC II, 4.4a:  $2pq(1 - \theta)$

NRC II, 4.10b: 
$$\frac{2[\theta + (1 - \theta)p_i][\theta + (1 - \theta)p_j]}{(1 + \theta)(1 + 2\theta)}$$

$$P(A_i A_j | A_i A_j)$$

# Theta Values Commonly Employed

- 0.01 for Cau, Afr Amer, SEH, and SWH
- 0.03 for Native American groups

Conservative Values

## Do the CODIS Loci Satisfy the Conditions for HWE and LE?

- As the loci being used in DNA forensics reside on regions of DNA with no effect on phenotypes that dictate mate choice, fertility, or viability, there is no evidence suggesting violations
- Population substructure exists, irrespective of definition of populations, but with the rate of mutation applicable for these loci, inter-population genetic variation in relation to within population variation at these loci ( $F_{ST}$  or  $\theta$ ) is not very large
- Population Studies support that the loci meet expectations quite well
- However, does this matter?

# Inbreeding Coefficient ( $F_{ST}$ )

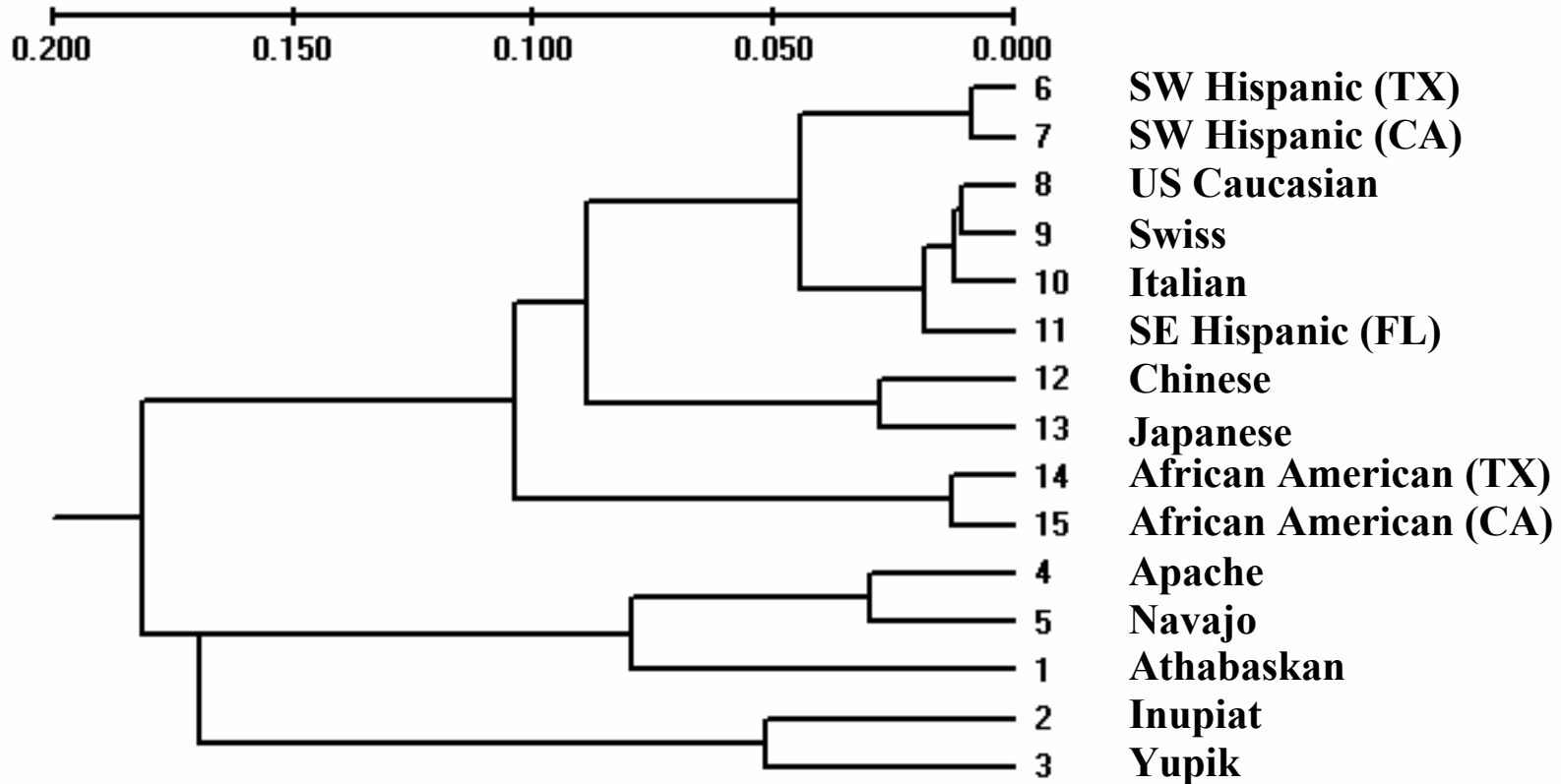
	Caucasian	African American	Hispanic	Asian	Native American
CSF1PO	-0.0007	-0.0009	-0.0003	-0.0012	0.0244
D13S317	-0.0008	0.0029	0.0047	0.0071	0.0157
D18S51	0.0001	0.0012	0.0011	0.0046	0.0268
D21S11	0.0008	0.0005	0.0013	0.0056	0.0371
D3S1358	-0.0009	-0.0009	0.0010	0.0035	0.0764
D5S818	-0.0001	0.0010	0.0010	0.0028	0.0656
D7S820	-0.0005	0.0000	0.0010	0.0039	0.0201

# Inbreeding Coefficient ( $F_{ST}$ )

	Caucasian	African American	Hispanic	Asian	Native American
D8S1179	0.0000	-0.0001	0.0005	0.0025	0.0125
FGA	-0.0004	0.0004	0.0008	0.0029	0.0168
THO1	-0.0012	0.0015	0.0041	0.0058	0.0356
TPOX	-0.0015	0.0021	0.0024	0.0100	0.0164
VWA	-0.0011	0.0011	0.0029	0.0027	0.0172
Average	<b>-0.0005</b>	<b>0.0006</b>	<b>0.0021</b>	<b>0.0039</b>	<b>0.0282</b>

# UPGMA

Bootstrap 1000 reps





# Inbreeding Coefficient ( $F_{ST}$ )

For Inupiat, Yupik

Average – 0.0167

# Inbreeding Coefficient ( $F_{ST}$ )

For Athabaskans, Apache, Navajo

Average – 0.0180

