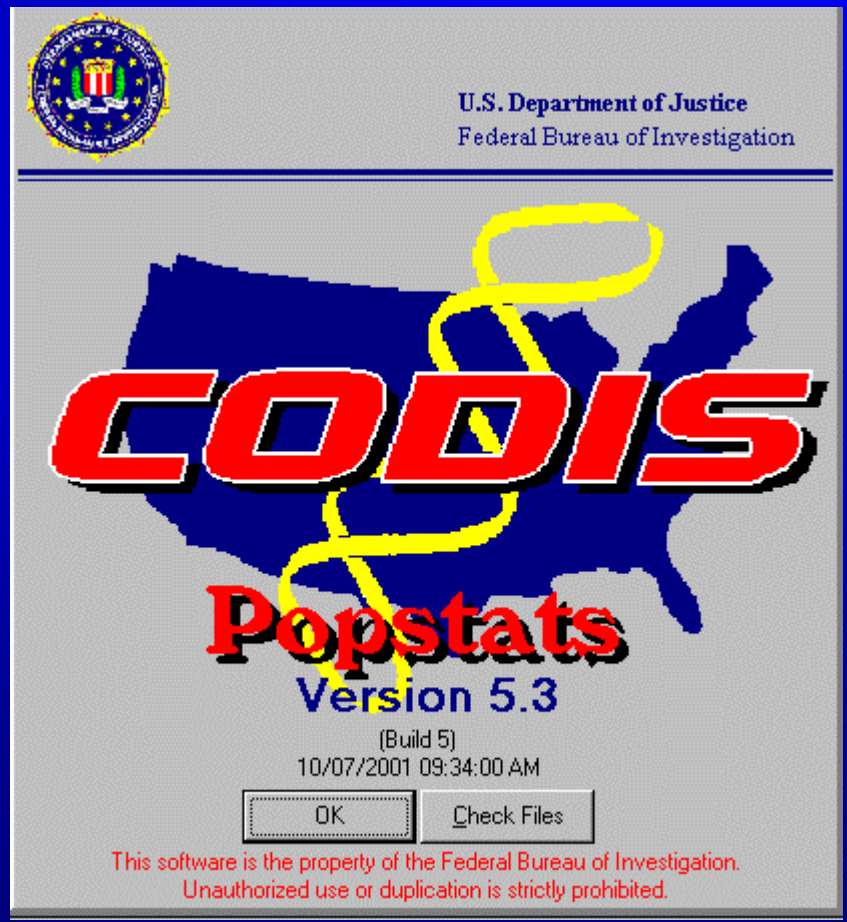
Popstats Unplugged

14th International Symposium on Human Identification



John V. Planz, Ph.D.
UNT Health Science Center at Fort Worth

Forensic Statistics

From the ground up…

Why so much attention to statistics?

Exclusions don’t require numbers

Matches do require statistics

Problem of verbal expression of numbers

Transfer evidence

Laboratory result

1. Non-match - **exclusion**
2. Inconclusive- **no decision**
3. Match - **estimate frequency**

Statistical Analysis

Focus on the question being asked…

About "Q" sample

"K" matches "Q"

Who else could match "Q"

partial profile, mixtures

Match – estimate frequency of:

Match to forensic evidence

NOT suspect DNA profile

Who is in suspect population?

So, what are we really after?

Quantitative statement that expresses the rarity of the DNA profile

Estimate genotype frequency

1. Frequency at each locus

   Hardy-Weinberg Equilibrium

2. Frequency across all loci

   Linkage Equilibrium

Terminology

Genetic marker variant = allele

DNA profile = genotype

Database = table that provides frequency of alleles in a population

Population = some assemblage of individuals based on some criteria for inclusion

Where Do We Get These Numbers?

POPULATION DATA
and
Statistics

DNA databases are needed for placing statistical weight on DNA profiles

vWA data (N=129)

	14	15	16	17	18	19	20	freq
14	9							75
15	3	0						6
16	19	1	1					46
17	23	1	14	9				72
18	6	0	3	10	4			31
19	6	1	7	3	2	2		23
20	0	0	0	3	2	0	0	5
								258

Because data are not available for every genotype possible,

We use allele frequencies instead of genotype frequencies to estimate rarity.

Estimate allele frequencies by "gene counting"

A_1	n_1
A_2	n_2
A_3	n_3
A_4	n_4
:	:
A_k	n_k

$$n_1 + n_2 + n_3 + n_4 + \ldots n_k = 2N$$

$$p_1 = n_1 / 2N$$

$$p_2 = n_2 / 2N$$

$$p_3 = n_3 / 2N$$

Population database

Look up how often each allele occurs at the locus in a population (or populations)

AKA looking up the “allele” frequency

TECHNICAL NOTE

Bruce Budowle,[1] Ph.D.; Tamyra R. Moretti,[1] Ph.D.; Anne L. Baumstark,[1] B.S.; Debra A. Defenbaugh,[1] B.S.; and Kathleen M. Keys,[1] B.S.

Population Data on the Thirteen CODIS Core Short Tandem Repeat Loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians*

REFERENCE: Budowle B, Moretti TR, Baumstark AL, Defenbaugh DA, Keys KM. Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. J Forensic Sci 1999;44(6):1277–1286.

markers are required, and all laboratories that contribute to the database should use the same genetic loci. Short tandem repeat (STR) loci are the most informative PCR-based genetic markers available to date for attempting to individualize biological material (2–5). The 13 STR loci CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818,

CSF1PO.SWH - Notepad

File Edit Format Help

Bin	Range (alleles)	Count	Fraction
1	<6- <6	0	0.0000
2	6- 6	0	0.0000
3	7- 7	1	0.0024
4	8- 8	0	0.0000
5	9- 9	3	0.0072
6	10- 10	106	0.2536
7	10.3- 10.3	0	0.0000
8	11- 11	111	0.2656
9	12- 12	164	0.3923
10	12.1- 12.1	0	0.0000
11	13- 13	27	0.0646
12	14- 14	4	0.0096
13	15- 15	2	0.0048
14	>15- >15	0	0.0000
	Totals	418	1.0001

Minimum allele frequency = 0.0120
Null allele frequency = 0.0120

Min Allele Frequency and Null Allele Frequency calculated as 5/2N where N = 209

Note: single-allele patterns are entered twice in database

Provided by B. Budowle, FBI Academy

FBI population Data

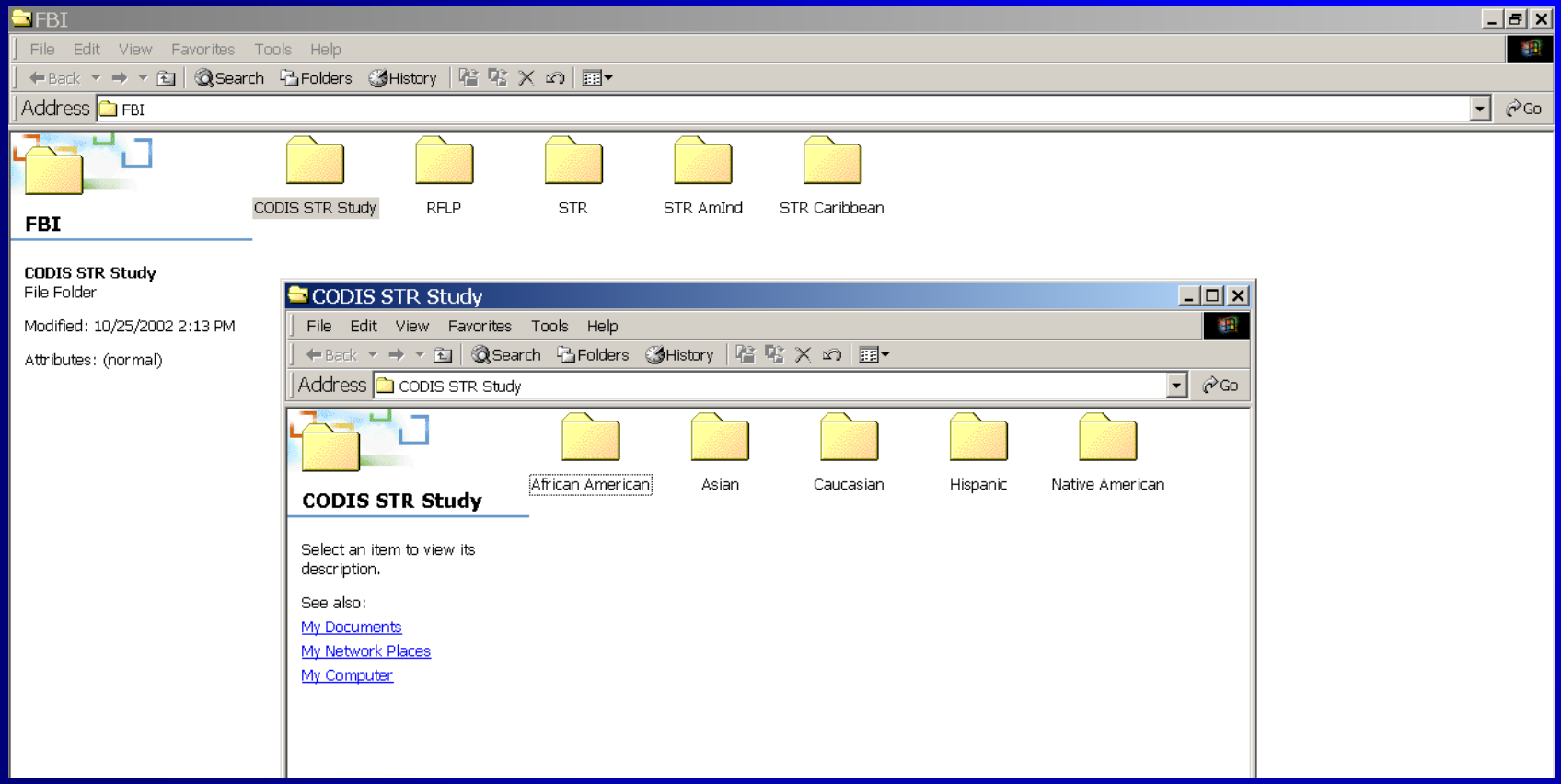
Bruce Budowle,[1] *Ph.D.; Brendan Shea,*[2] *M.S.; Stephen Niezgoda,*[2] *M.B.A.; and Ranajit Chakraborty,*[3] *Ph.D.*

CODIS STR Loci Data from 41 Sample Populations*

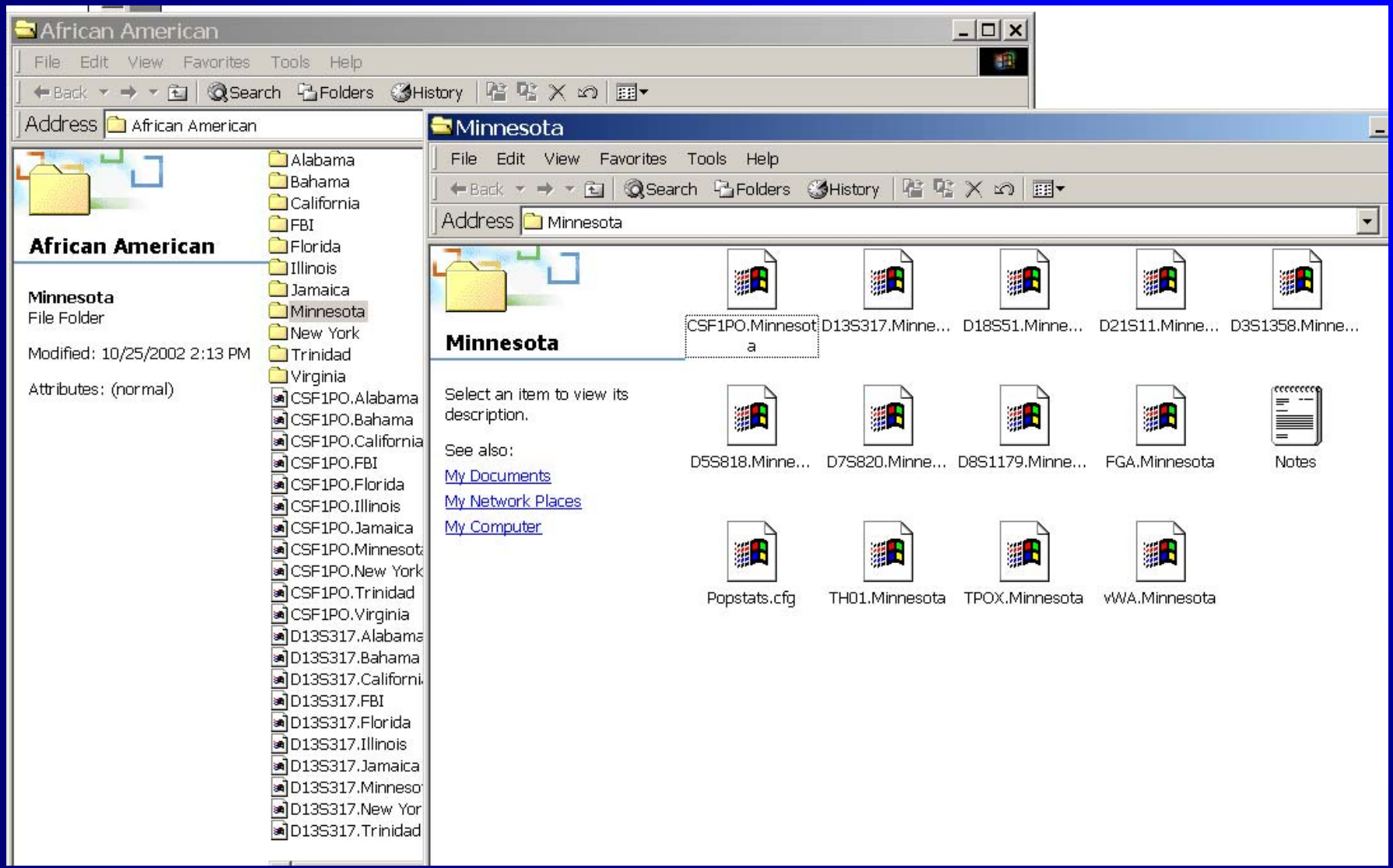
REFERENCE: Budowle B, Shea B, Niezgoda S, Chakraborty R. CODIS STR loci data from 41 sample populations. J Forensic Sci 2001:46;(3):453–489.

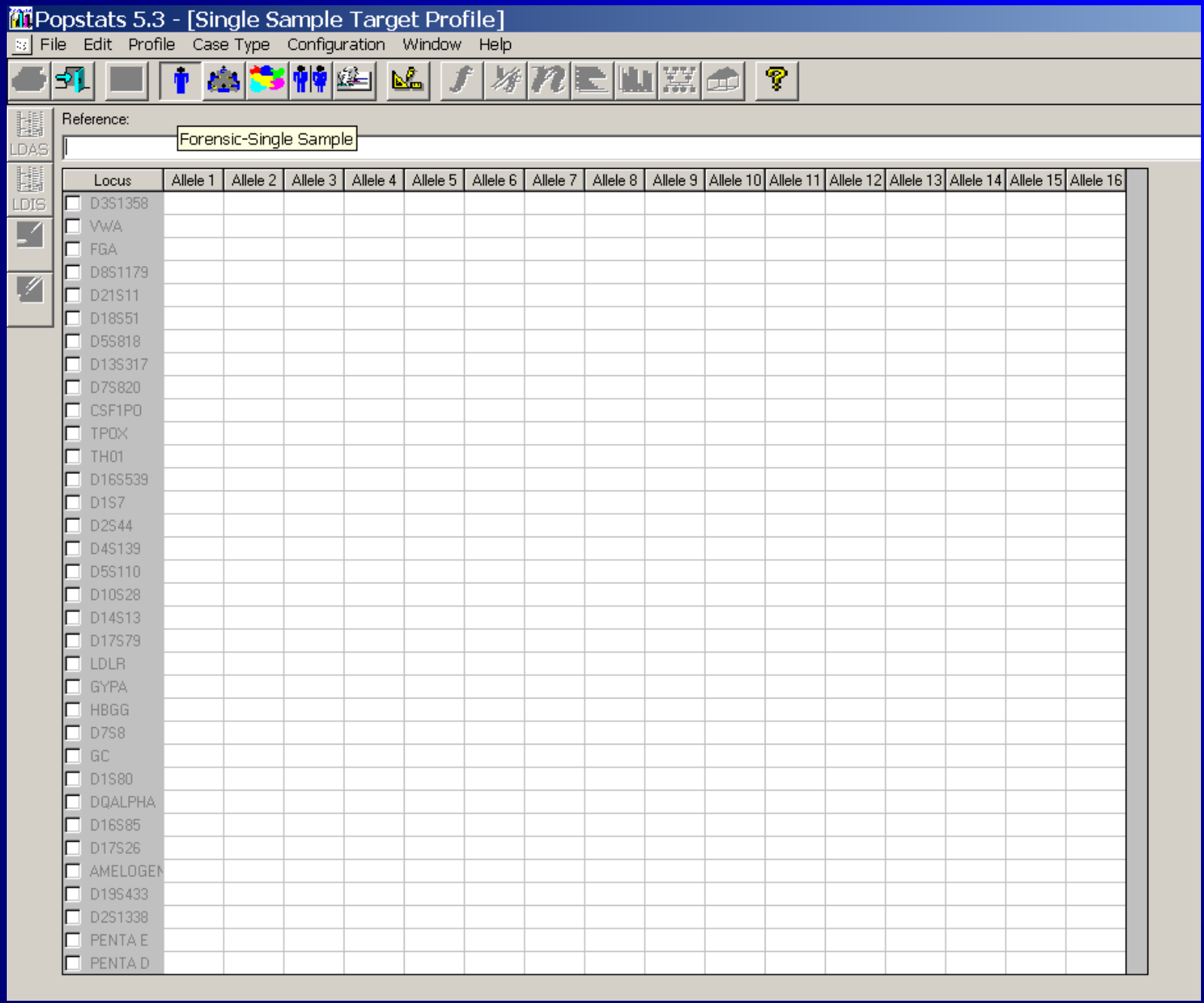
Materials and Methods

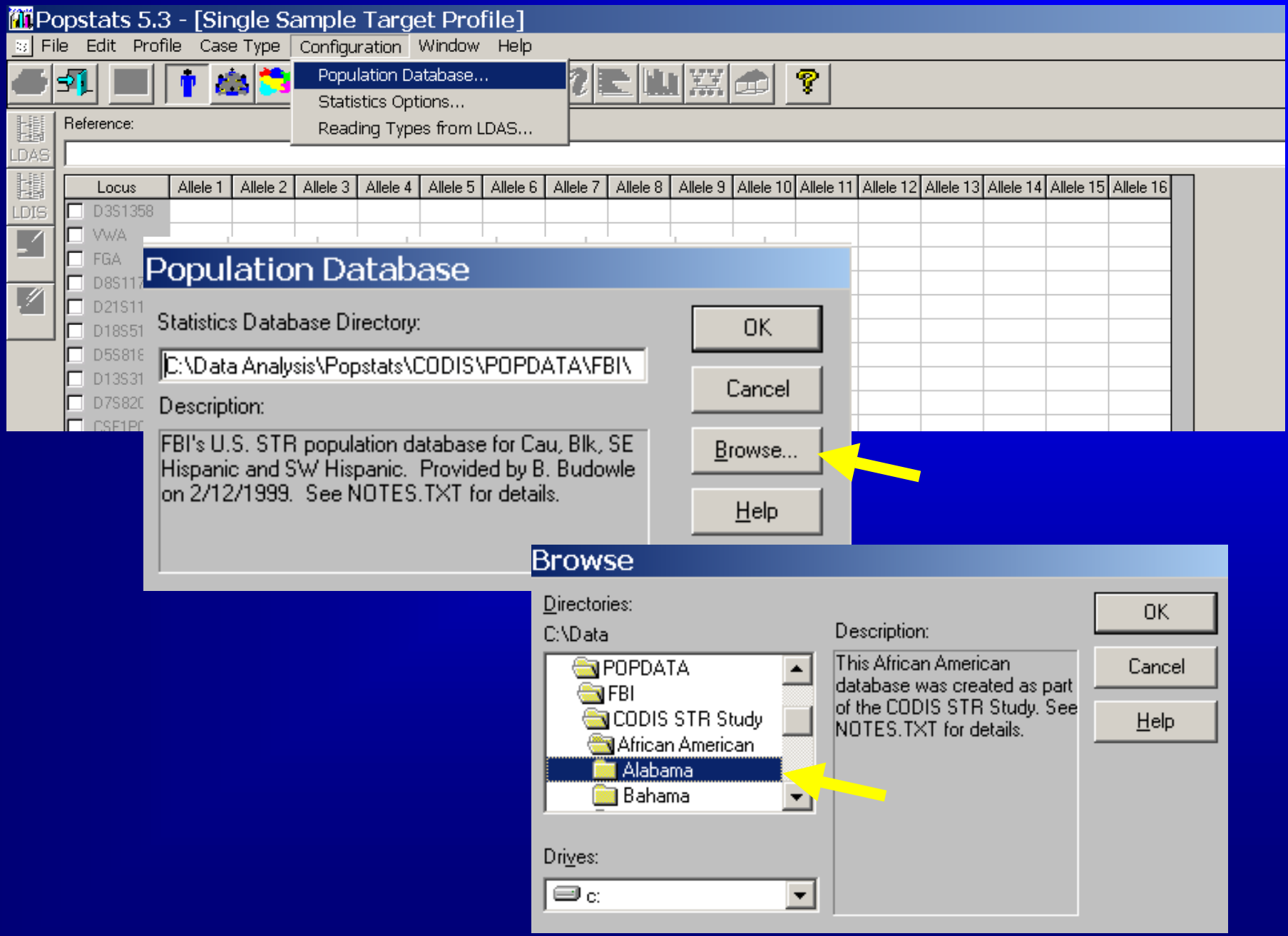
Samples



You can get other population databases to be used by Popstats if needed


African American
File Edit View Favorites Tools Help
Back Search Folders History
Address African American
African American
Minnesota
File Folder
Modified: 10/25/2002 2:13 PM
Attributes: (normal)
Alabama
Bahama
California
FBI
Florida
Illinois
Jamaica
Minnesota
New York
Trinidad
Virginia
CSF1PO.Alabama
CSF1PO.Bahama
CSF1PO.California
CSF1PO.FBI
CSF1PO.Florida
CSF1PO.Illinois
CSF1PO.Jamaica
CSF1PO.Minnesota
CSF1PO.New York
CSF1PO.Trinidad
CSF1PO.Virginia
D13S317.Alabama
D13S317.Bahama
D13S317.California
D13S317.FBI
D13S317.Florida
D13S317.Illinois
D13S317.Jamaica
D13S317.Minnesota
D13S317.New York
D13S317.Trinidad
Minnesota
File Edit View Favorites Tools Help
Back Search Folders History
Address Minnesota
Minnesota
Select an item to view its description.
See also:
My Documents
My Network Places
My Computer
CSF1PO.Minnesota
D13S317.Minne...
D18S51.Minne...
D21S11.Minne...
D3S1358.Minne...
D5S818.Minne...
D7S820.Minne...
D8S1179.Minne...
FGA.Minnesota
Notes
Popstats.cfg
TH01.Minnesota
TPOX.Minnesota
vWA.Minnesota


Popstats 5.3 - [Single Sample Target Profile]
File Edit Profile Case Type Configuration Window Help
LDAS
LDIS
Reference:
Forensic-Single Sample
Locus Allele 1 Allele 2 Allele 3 Allele 4 Allele 5 Allele 6 Allele 7 Allele 8 Allele 9 Allele 10 Allele 11 Allele 12 Allele 13 Allele 14 Allele 15 Allele 16
D3S1358
VWA
FGA
D8S1179
D21S11
D18S51
D5S818
D13S317
D7S820
CSF1PO
TPOX
TH01
D16S539
D1S7
D2S44
D4S139
D5S110
D10S28
D14S13
D17S79
LDLR
GYPA
HBGG
D7S8
GC
D1S80
DQALPHA
D16S85
D17S26
AMELOGEN
D19S433
D2S1338
PENTA E
PENTA D


Popstats 5.3 - [Single Sample Target Profile]
File Edit Profile Case Type Configuration Window Help
Population Database...
Statistics Options...
Reading Types from LDAS...
Reference:
LDAS
LDIS
Locus Allele 1 Allele 2 Allele 3 Allele 4 Allele 5 Allele 6 Allele 7 Allele 8 Allele 9 Allele 10 Allele 11 Allele 12 Allele 13 Allele 14 Allele 15 Allele 16
D3S1358
VWA
FGA
Population Database
Statistics Database Directory:
C:\Data Analysis\Popstats\CODIS\POPDATA\FBI\
Description:
FBI's U.S. STR population database for Cau, Blk, SE Hispanic and SW Hispanic. Provided by B. Budowle on 2/12/1999. See NOTES.TXT for details.
OK
Cancel
Browse...
Help
Browse
Directories:
C:\Data
POPDATA
FBI
CODIS STR Study
African American
Alabama
Bahama
Drives:
c:
Description:
This African American database was created as part of the CODIS STR Study. See NOTES.TXT for details.
OK
Cancel
Help

OK

So now we have our population databases

How do they work for us in Popstats??

Anchor principle

Analysis of genetic makeup in individuals is based on the *Genotype* at the locus being queried

To remove "individual variation" so that we can focus on population-wide variation we must meld all the genotypes into a pool…separated as alleles

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

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.

This mathematical model serves as our *Null Hypothesis* when comparing populations, demes etc.

Hardy - Weinberg Equilibrium

A_1A_1	A_1A_2	A_2A_2
p_1^2	$2p_1p_2$	p_2^2

freq(A_1) = p_1

freq(A_2) = p_2

	A_1	A_2
A_1	p_1^2 A_1A_1	p_1p_2 A_1A_2
A_2	p_1p_2 A_1A_2	p_2^2 A_2A_2

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

A Hardy-Weinberg Population

LARGE POPULATION

NO NATURAL SELECTION

NO MUTATION

NO IMMIGRATION / EMIGRATION

RANDOM MATING

Estimate genotype frequency:

1. **Frequency at each locus**

2. **Frequency across all loci**

Product Rule

Product Rule

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

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

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:

Frequency = genotype frequency (p^2)
(for now!)

Genotype freq = 0.3071 x 0.3071 = 0.0943

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

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

(2*pq*)

Genotype freq = 2 x 0.2361 x 0.18331 = 0.0866

Overall profile frequency =

Frequency D3S1358 X Frequency vWA

0.0943 x 0.0866 = 0.00817

This is basically what Popstats does for us in it's simplest task

Steps – Single Sample Target Profile

- enter alleles of target profile
- look up allele frequencies at all loci for all populations
- determine if homozygous or heterozygous at each locus
- calculate genotype frequency at each locus
- calculate profile frequency with product rule

But this doesn't address all of the issues!

What if…

We encounter alleles not represented in the population database…

…or alleles that are extremely rare in the database???

Minimum allele frequency

The first NRC report proposed a minimum allele frequency based on NO empirical data and without any statistical basis!

10 % or 0.1

What…you are surprised??

Ceiling Principle

Minimum allele frequency

Weir, B.S. 1992. $\text{minfreq} = 1 - \alpha^{1/2N}$

Budowle, B., K. Monson, R. Chakraborty, 1996. $\text{minfreq} = 1 - [\ 1 - (\ 1 - \alpha)^{1/C}\ ]^{1/2N}$

NRC II, 1996. $\text{minfreq} = 5/2N$

Minimum allele frequency

This method 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

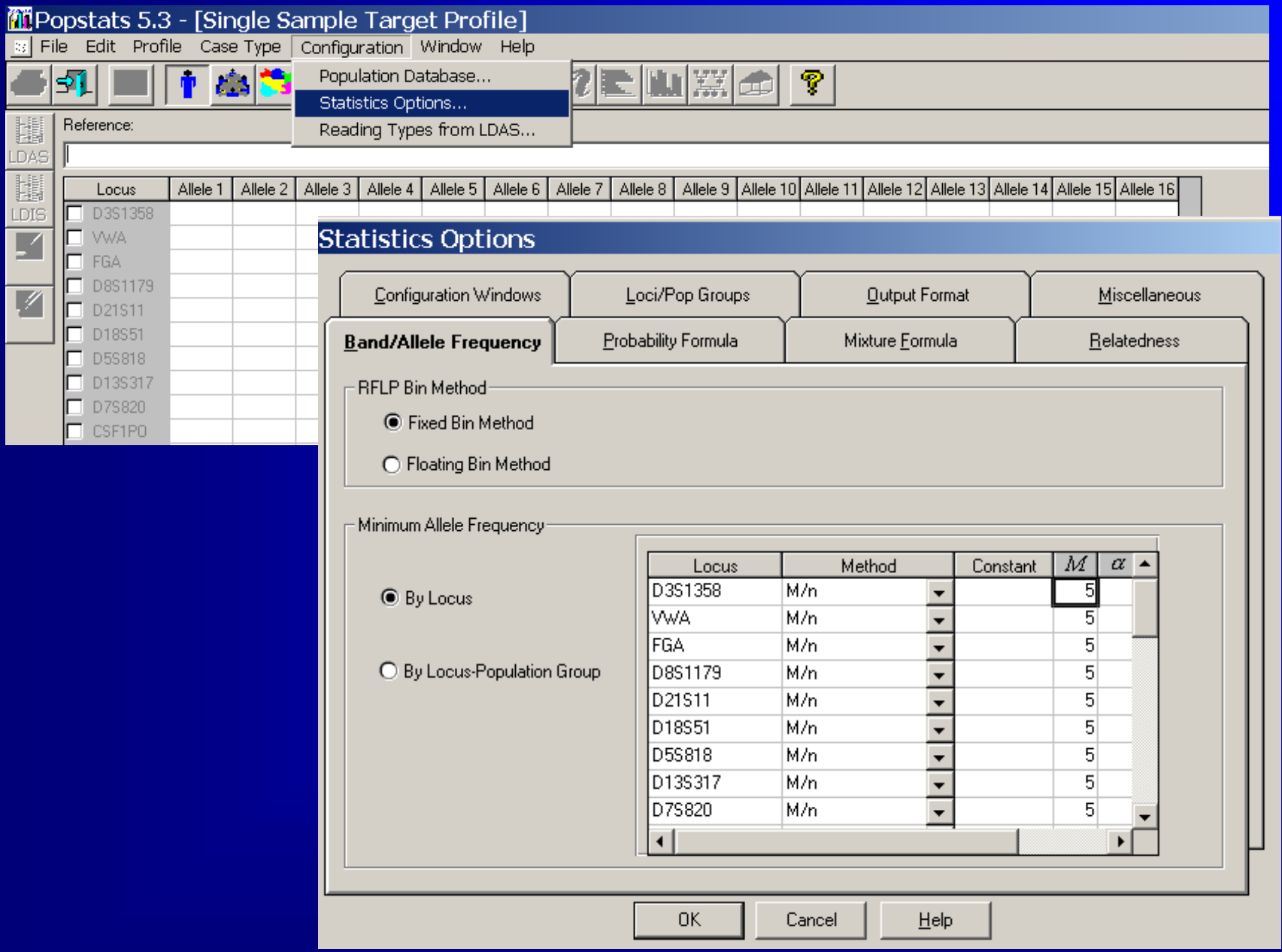
For the 13 allele at vWA:
Actual Freq = 2 / 392 = 0.0051
Minimal Freq = 5 / 392 = 0.0128

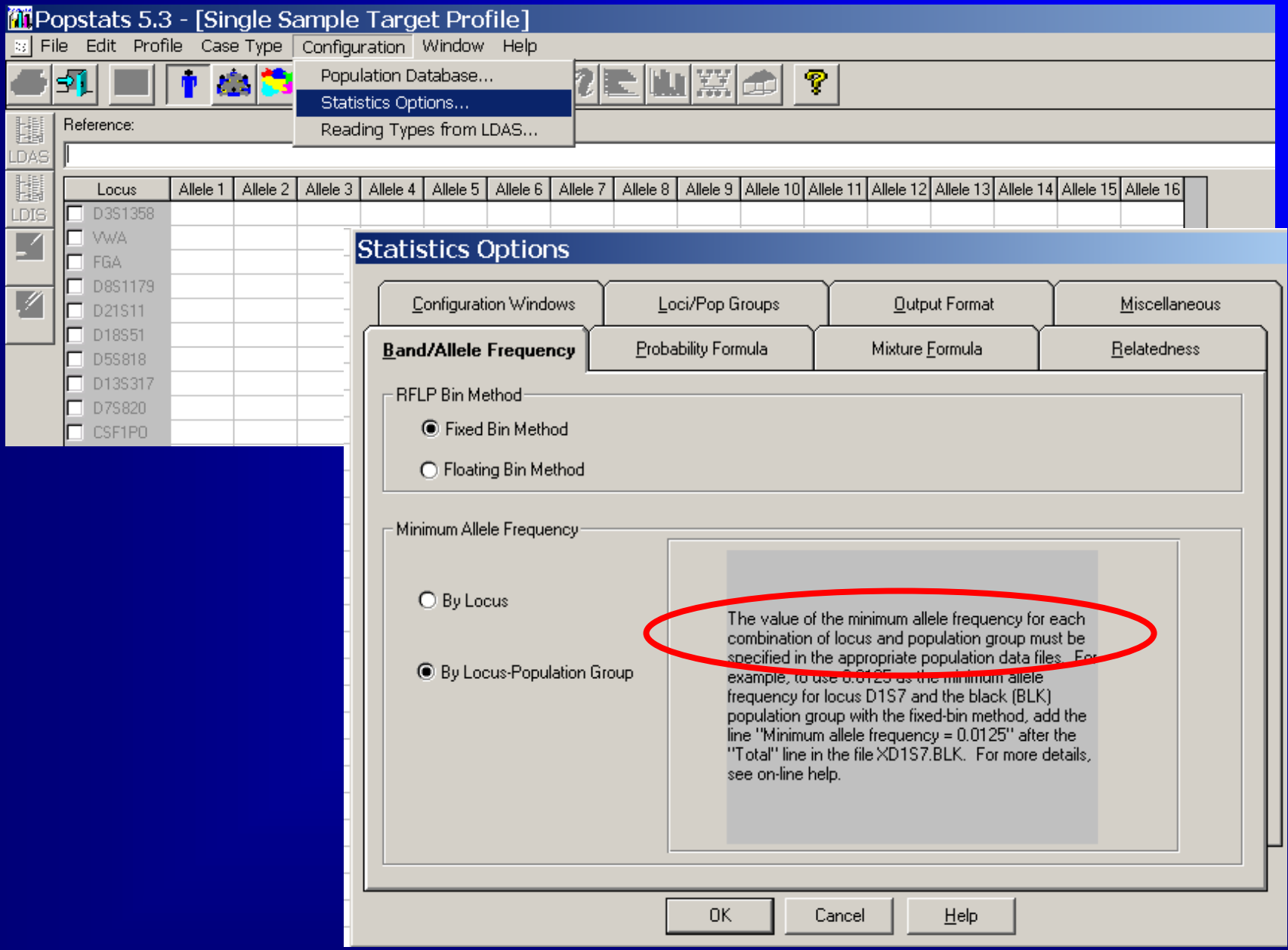
Conservatism & also addresses some substructure effects

This estimate is strictly driven by database size:

N	min allele freq
100	2.50 % (0.025)
150	1.67 % (0.0167)
200	1.25 % (0.0125)
250	1.00 % (0.01)
300	0.83 % (0.0083)

Where N is the number of individuals in database


Popstats 5.3 - [Single Sample Target Profile]
File Edit Profile Case Type Configuration Window Help
Population Database...
Statistics Options...
Reading Types from LDAS...
Reference:
LDAS
LDIS
Locus Allele 1 Allele 2 Allele 3 Allele 4 Allele 5 Allele 6 Allele 7 Allele 8 Allele 9 Allele 10 Allele 11 Allele 12 Allele 13 Allele 14 Allele 15 Allele 16
D3S1358
VWA
FGA
D8S1179
D21S11
D18S51
D5S818
D13S317
D7S820
CSF1PO
Statistics Options
Configuration Windows
Loci/Pop Groups
Output Format
Miscellaneous
Band/Allele Frequency
Probability Formula
Mixture Formula
Relatedness
RFLP Bin Method
Fixed Bin Method
Floating Bin Method
Minimum Allele Frequency
By Locus
By Locus-Population Group
Locus Method Constant M α
D3S1358 M/n 5
VWA M/n 5
FGA M/n 5
D8S1179 M/n 5
D21S11 M/n 5
D18S51 M/n 5
D5S818 M/n 5
D13S317 M/n 5
D7S820 M/n 5
OK
Cancel
Help


Popstats 5.3 - [Single Sample Target Profile]
File Edit Profile Case Type Configuration Window Help
Population Database...
Statistics Options...
Reading Types from LDAS...
Reference:
LDAS
LDIS
Locus Allele 1 Allele 2 Allele 3 Allele 4 Allele 5 Allele 6 Allele 7 Allele 8 Allele 9 Allele 10 Allele 11 Allele 12 Allele 13 Allele 14 Allele 15 Allele 16
D3S1358
VWA
FGA
D8S1179
D21S11
D18S51
D5S818
D13S317
D7S820
CSF1PO
Statistics Options
Configuration Windows
Loci/Pop Groups
Output Format
Miscellaneous
Band/Allele Frequency
Probability Formula
Mixture Formula
Relatedness
RFLP Bin Method
Fixed Bin Method
Floating Bin Method
Minimum Allele Frequency
By Locus
By Locus-Population Group
The value of the minimum allele frequency for each combination of locus and population group must be specified in the appropriate population data files. For example, to use 0.0125 as the minimum allele frequency for locus D1S7 and the black (BLK) population group with the fixed-bin method, add the line "Minimum allele frequency = 0.0125" after the "Total" line in the file XD1S7.BLK. For more details, see on-line help.
OK
Cancel
Help

CSF1PO.SWH - Notepad

File Edit Format Help

Bin	Range (alleles)	Count	Fraction
1	<6- <6	0	0.0000
2	6- 6	0	0.0000
3	7- 7	1	0.0024
4	8- 8	0	0.0000
5	9- 9	3	0.0072
6	10- 10	106	0.2536
7	10.3- 10.3	0	0.0000
8	11- 11	111	0.2656
9	12- 12	164	0.3923
10	12.1- 12.1	0	0.0000
11	13- 13	27	0.0646
12	14- 14	4	0.0096
13	15- 15	2	0.0048
14	>15- >15	0	0.0000
	Totals	418	1.0001

Minimum allele frequency = 0.0120
Null allele frequency = 0.0120

Min Allele Frequency and Null Allele Frequency calculated as 5/2N where N = 209

Note: single-allele patterns are entered twice in database

Provided by B. Budowle, FBI Academy

So the only other real thing left to consider regarding the NRC concerns is population subdivision.

Population Structure

- **Racial, ethnic subgroups**
- **Excess of homozygotes**
- **What is "theta" θ**
- **Why modify just homozygous calculation?**
- **NRC Formula 4.1 vs 4.4 vs 4.10**

Population Subdivision

We've always surmised…

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

Human Genetic Variation

between populations within racial groups …

between racial groups ……

within populations within racial groups ……

- Barbujani, Magagni, Minch, Cavalli-Sforza. 1997. An apportionment of human DNA diversity. *PNAS* 94:4516-4519.

Problems created by population subdivision

Genotype frequencies calculated from population average allele frequencies could lead to:

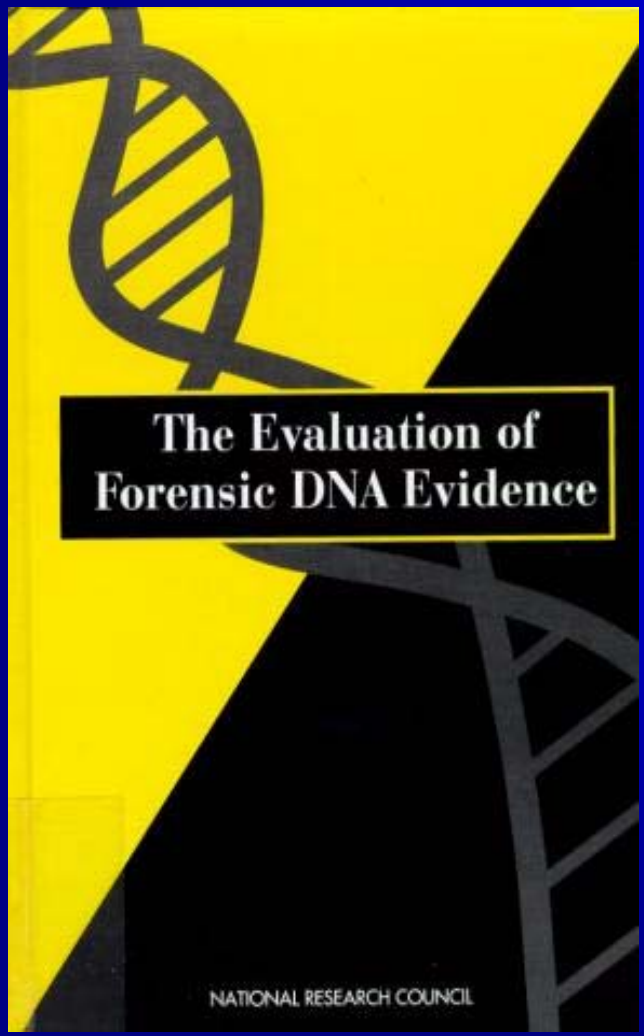
Wrong estimates!

Employ a Theta (θ) Correction

θ 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?

National Research Council Report II

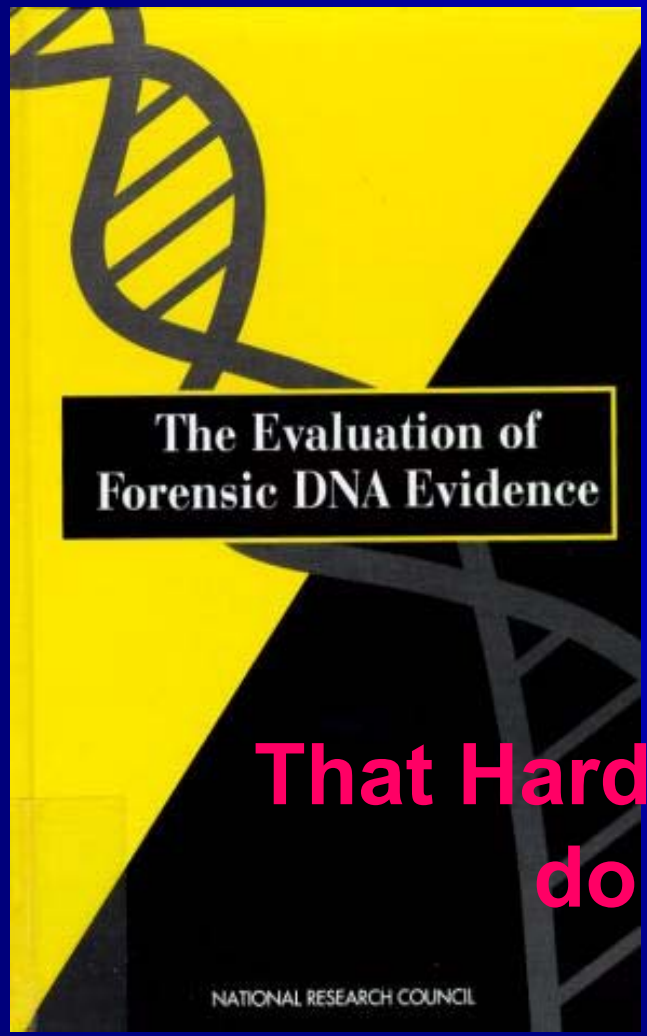


National Academy of Sciences

Data support the recommendation that F_{ST} of 0.01 is conservative

Issued in May 1996

National Research Council Report II



The significance of this F_{ST} is

That Hardy-Weinberg Expectations do not have to be met

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

Modifying the pro

Use correction factor f

$P^2 + P(1-P)\theta$

$\theta = 0.01$ 0.03 for Native populations

Intermediate to the F_{ST} that you would find in populations with 1st and 2nd cousin matings

use $2p_ip_j$ for heterozygotes (ie: no correction)

Really, this is more than ten fold more conservative

Modifying the product rule

- Formula 4.1 - HW
- Formula 4.4 - Simple subdivision
- Formula 4.10 - assumption of population
- Conditional vs Unconditional Probability

HWE: p^2

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

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

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

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

HWE: $2pq$

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

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

$$P(A_iA_j \mid A_iA_j)$$

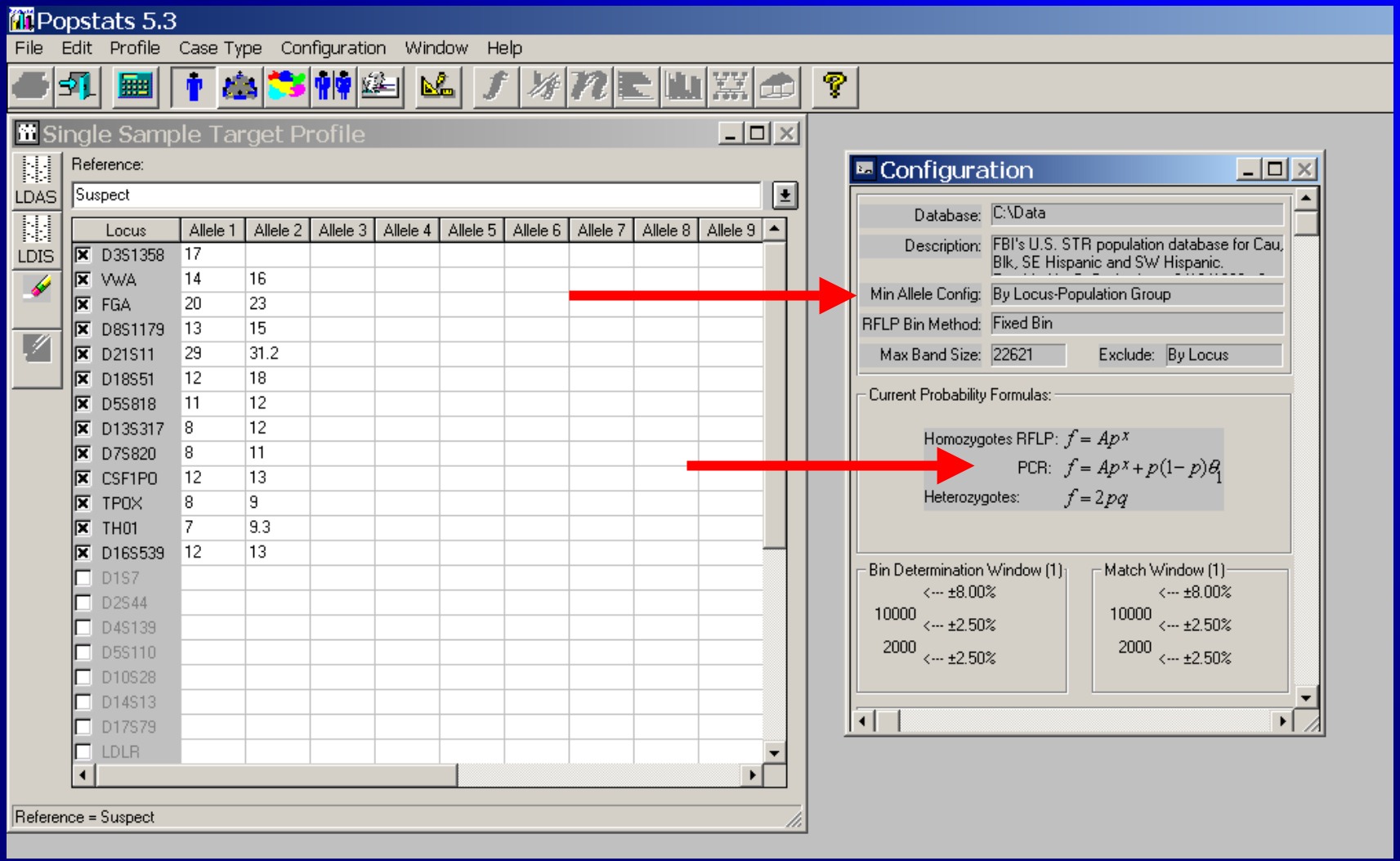
Theta Values Commonly Employed

0.01 for Cau, AA, SEH, and SWH

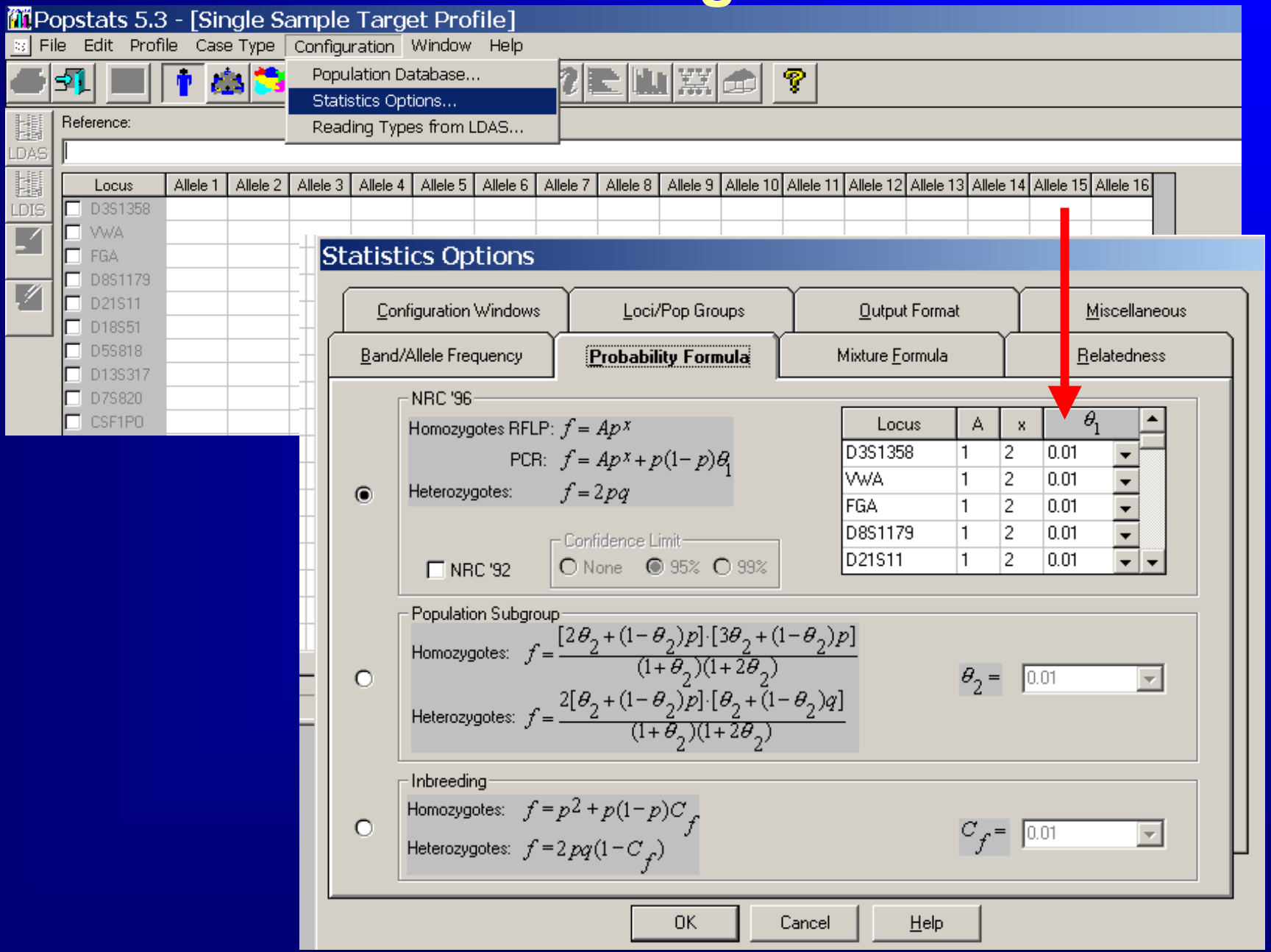
0.03 for Native American groups

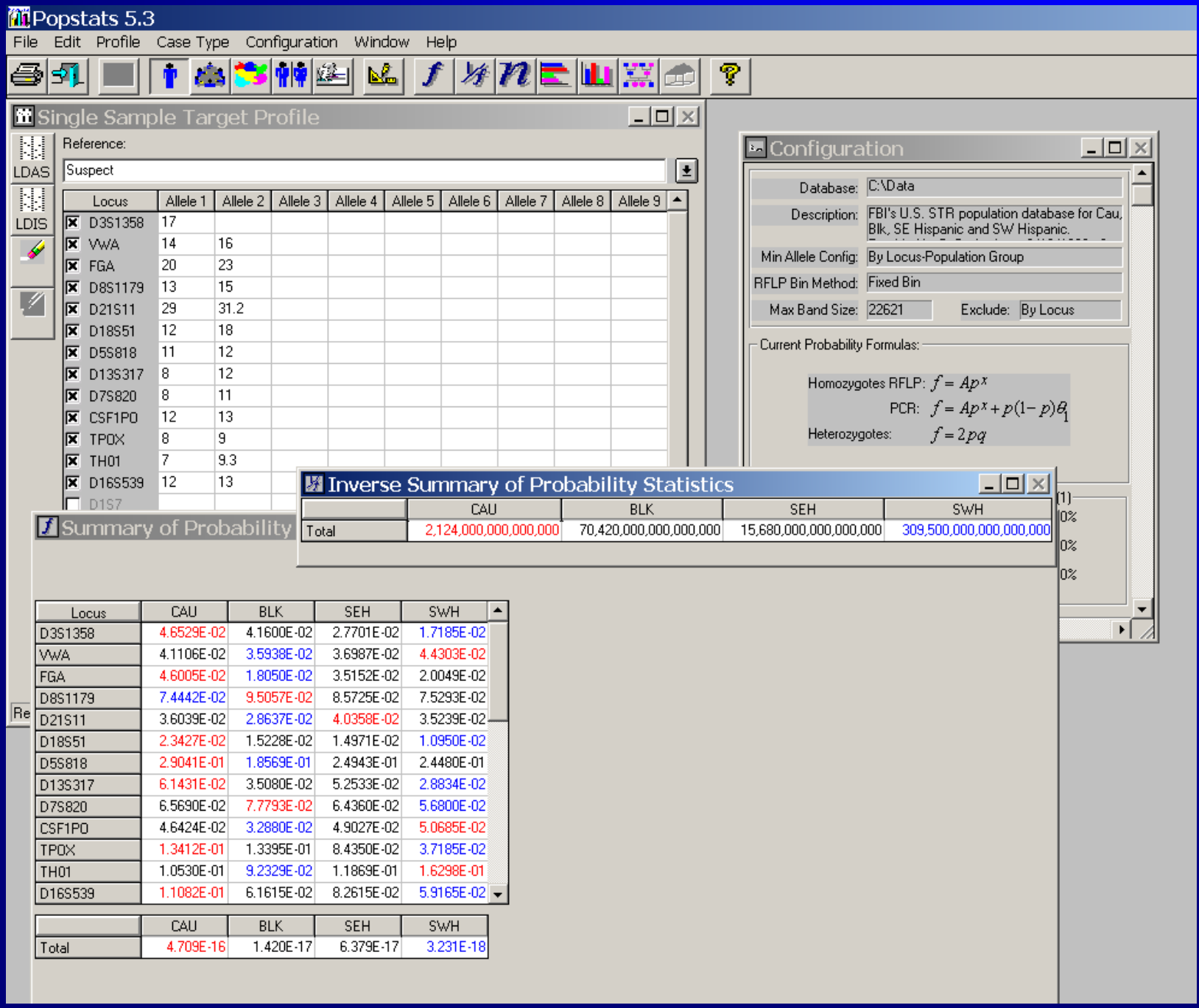
Conservative Values

So how do we deal with these concepts in PopStats??

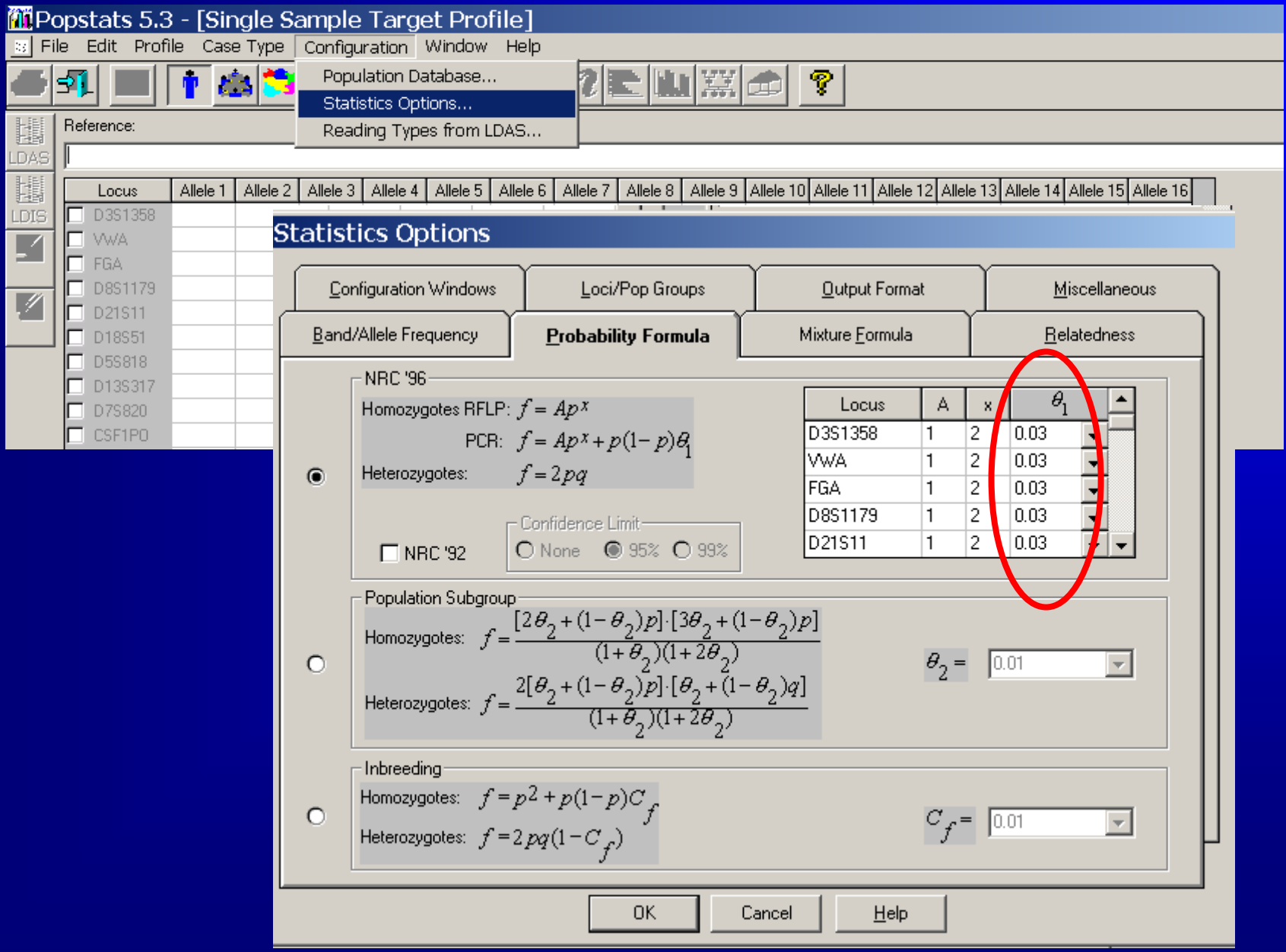


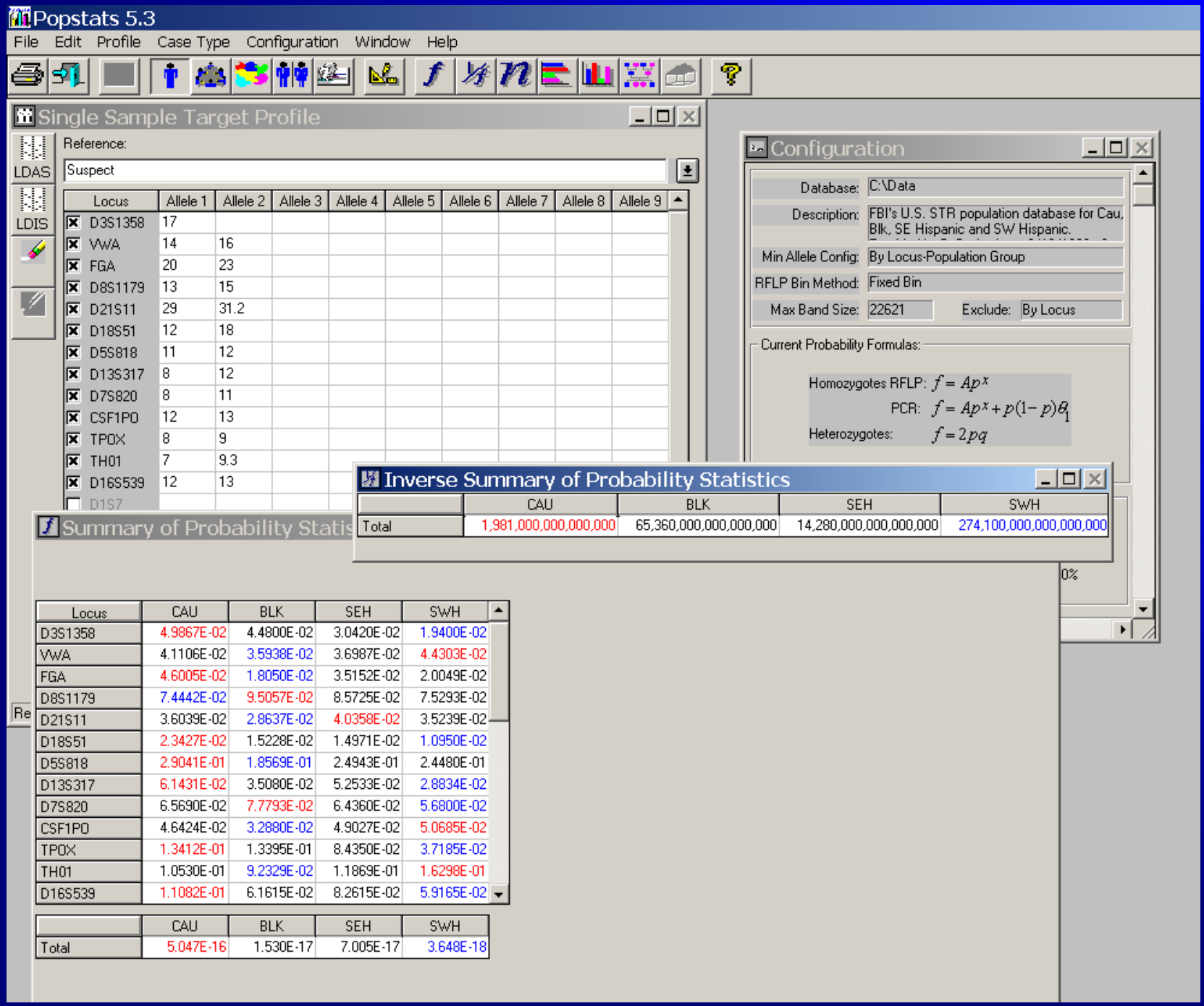
This is the default using 0.01 θ




Popstats 5.3
File Edit Profile Case Type Configuration Window Help
Single Sample Target Profile
LDAS
LDIS
Reference:
Suspect
Locus Allele 1 Allele 2 Allele 3 Allele 4 Allele 5 Allele 6 Allele 7 Allele 8 Allele 9
D3S1358 17
VWA 14 16
FGA 20 23
D8S1179 13 15
D21S11 29 31.2
D18S51 12 18
D5S818 11 12
D13S317 8 12
D7S820 8 11
CSF1PO 12 13
TPOX 8 9
TH01 7 9.3
D16S539 12 13
Configuration
Database: C:\Data
Description: FBI's U.S. STR population database for Cau, Blk, SE Hispanic and SW Hispanic.
Min Allele Config: By Locus-Population Group
RFLP Bin Method: Fixed Bin
Max Band Size: 22621
Exclude: By Locus
Current Probability Formulas:
Homozygotes RFLP: $f = Ap^x$
PCR: $f = Ap^x + p(1-p)\theta$
Heterozygotes: $f = 2pq$
Inverse Summary of Probability Statistics
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
Summary of Probability
Locus CAU BLK SEH SWH
D3S1358 4.6529E-02 4.1600E-02 2.7701E-02 1.7185E-02
VWA 4.1106E-02 3.5938E-02 3.6987E-02 4.4303E-02
FGA 4.6005E-02 1.8050E-02 3.5152E-02 2.0049E-02
D8S1179 7.4442E-02 9.5057E-02 8.5725E-02 7.5293E-02
D21S11 3.6039E-02 2.8637E-02 4.0358E-02 3.5239E-02
D18S51 2.3427E-02 1.5228E-02 1.4971E-02 1.0950E-02
D5S818 2.9041E-01 1.8569E-01 2.4943E-01 2.4480E-01
D13S317 6.1431E-02 3.5080E-02 5.2533E-02 2.8834E-02
D7S820 6.5690E-02 7.7793E-02 6.4360E-02 5.6800E-02
CSF1PO 4.6424E-02 3.2880E-02 4.9027E-02 5.0685E-02
TPOX 1.3412E-01 1.3395E-01 8.4350E-02 3.7185E-02
TH01 1.0530E-01 9.2329E-02 1.1869E-01 1.6298E-01
D16S539 1.1082E-01 6.1615E-02 8.2615E-02 5.9165E-02
CAU BLK SEH SWH
Total 4.709E-16 1.420E-17 6.379E-17 3.231E-18

If we want to use 0.03 θ




Popstats 5.3
File Edit Profile Case Type Configuration Window Help
Single Sample Target Profile
Reference:
Suspect
LDAS
LDIS
Locus Allele 1 Allele 2 Allele 3 Allele 4 Allele 5 Allele 6 Allele 7 Allele 8 Allele 9
D3S1358 17
VWA 14 16
FGA 20 23
D8S1179 13 15
D21S11 29 31.2
D18S51 12 18
D5S818 11 12
D13S317 8 12
D7S820 8 11
CSF1PO 12 13
TPOX 8 9
TH01 7 9.3
D16S539 12 13
D1S7
Configuration
Database: C:\Data
Description: FBI's U.S. STR population database for Cau, Blk, SE Hispanic and SW Hispanic.
Min Allele Config: By Locus-Population Group
RFLP Bin Method: Fixed Bin
Max Band Size: 22621
Exclude: By Locus
Current Probability Formulas:
Homozygotes RFLP: $f = Ap^x$
PCR: $f = Ap^x + p(1-p)\theta_1$
Heterozygotes: $f = 2pq$
Inverse Summary of Probability Statistics
CAU BLK SEH SWH
Total 1,981,000,000,000,000 65,360,000,000,000,000 14,280,000,000,000,000 274,100,000,000,000,000
Summary of Probability Statis
Locus CAU BLK SEH SWH
D3S1358 4.9867E-02 4.4800E-02 3.0420E-02 1.9400E-02
VWA 4.1106E-02 3.5938E-02 3.6987E-02 4.4303E-02
FGA 4.6005E-02 1.8050E-02 3.5152E-02 2.0049E-02
D8S1179 7.4442E-02 9.5057E-02 8.5725E-02 7.5293E-02
D21S11 3.6039E-02 2.8637E-02 4.0358E-02 3.5239E-02
D18S51 2.3427E-02 1.5228E-02 1.4971E-02 1.0950E-02
D5S818 2.9041E-01 1.8569E-01 2.4943E-01 2.4480E-01
D13S317 6.1431E-02 3.5080E-02 5.2533E-02 2.8834E-02
D7S820 6.5690E-02 7.7793E-02 6.4360E-02 5.6800E-02
CSF1PO 4.6424E-02 3.2880E-02 4.9027E-02 5.0685E-02
TPOX 1.3412E-01 1.3395E-01 8.4350E-02 3.7185E-02
TH01 1.0530E-01 9.2329E-02 1.1869E-01 1.6298E-01
D16S539 1.1082E-01 6.1615E-02 8.2615E-02 5.9165E-02
CAU BLK SEH SWH
Total 5.047E-16 1.530E-17 7.005E-17 3.648E-18

$\theta = 0.01$

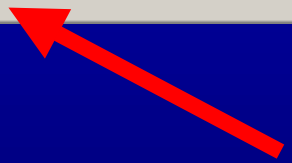
Inverse Summary of Probability Statistics

	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

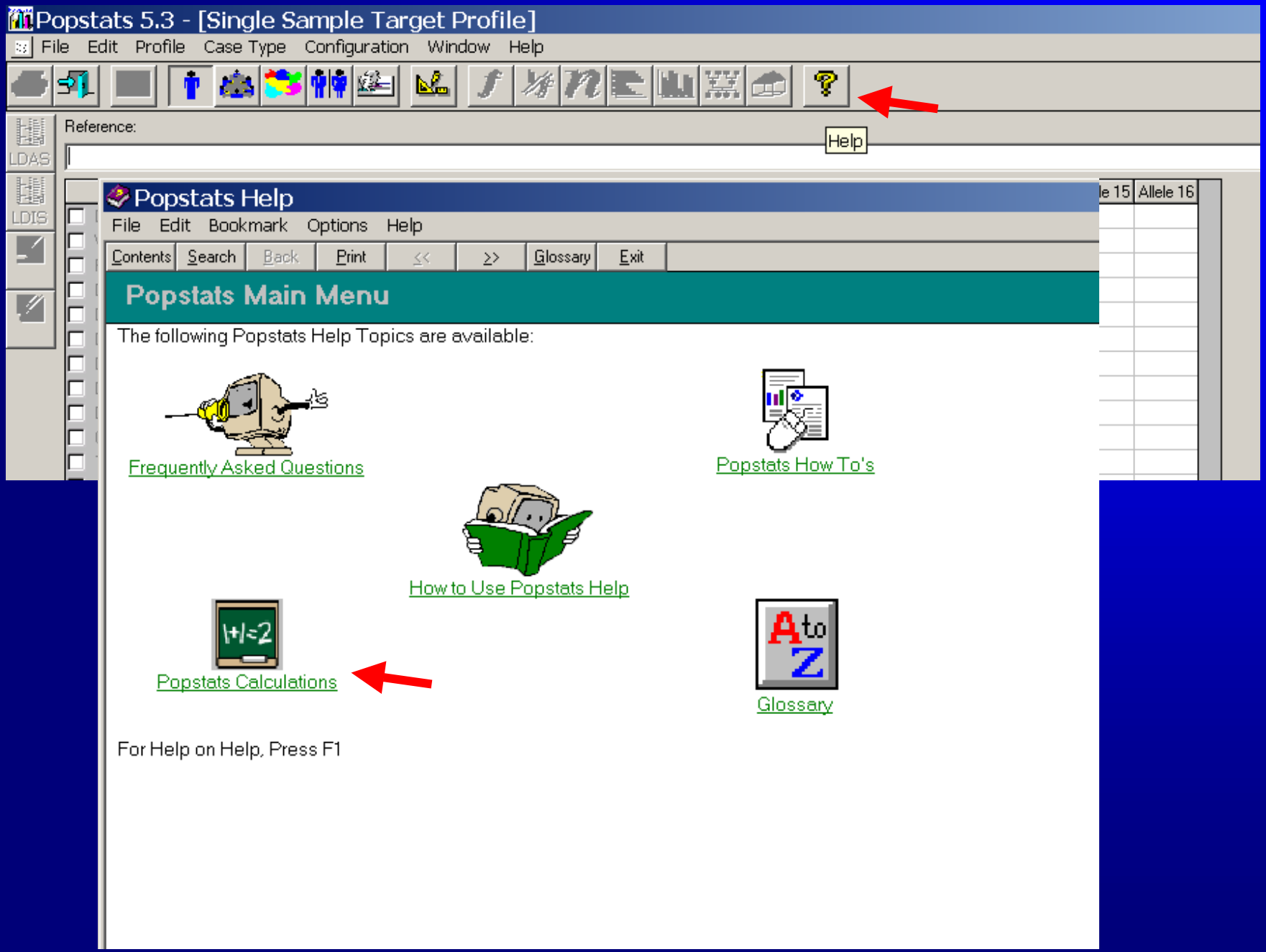
$\theta = 0.03$

Inverse Summary of Probability Statistics

	CAU	BLK	SEH	SWH
Total	1,981,000,000,000,000	65,360,000,000,000,000	14,280,000,000,000,000	274,100,000,000,000,000



WHERE CAN WE FIND INFORMATION ON WHAT POPSTATS CAN DO?


Popstats 5.3 - [Single Sample Target Profile]
File Edit Profile Case Type Configuration Window Help
Reference:
Help
LDAS
LDIS
Popstats Help
File Edit Bookmark Options Help
Contents Search Back Print << >> Glossary Exit
Popstats Main Menu
The following Popstats Help Topics are available:
Frequently Asked Questions
Popstats How To's
How to Use Popstats Help
Popstats Calculations
Glossary
For Help on Help, Press F1
Allele 15
Allele 16

Popstats Help

File Edit Bookmark Options Help

Contents | Search | Back | Print | << | >> | Glossary | Exit

Popstats Calculations

Additional Information

Probability of A Locus Genotype

Where:
p and q are band/allele occurrence frequencies
A and x are user-configurable constants for each locus.

For VNTR/RFLP loci, the default settings are:
$A = 2$
$x = 1$.

For PCR loci, the default settings are:
$A = 1$ (A must be positive)
$x = 2$ (x must be non-negative).

2. NRC '96 random-match probability formulae with θ adjustment for homozygotes of PCR loci:

homozygotes for RFLP loci: $f = Ap^x$

homozygotes of PCR loci: $f = Ap^x + p(1-p)\theta_1$

heterozygotes: $f = 2pq$

Where p, q, A, and x are defined as above. The parameter θ_1 is only available for PCR loci The user can enter any value for θ_1. The default value is $\theta_1 = 0.01$. For some small, isolated populations like the American Indians, a value of 0.03 is suggested for θ_1 by the *National Research Council (NRC) Report 1996*.

3 The conditional probability formulae with population subgroup correction:

homozygotes: $f = \dfrac{[2\theta_2 + (1-\theta_2)p]\cdot[3\theta_2 + (1-\theta_2)p]}{(1+\theta_2)(1+2\theta_2)}$

heterozygotes: $f = \dfrac{2[\theta_2 + (1-\theta_2)p]\cdot[\theta_2 + (1-\theta_2)q]}{(1+\theta_2)(1+2\theta_2)}$

These equations calculate the probability of one person having the DNA profile genotype under the condition that another person has that DNA profile genotype. θ_2 is a user-configurable parameter that represents the degree of population subdivision. Its value is configured globally for all loci and is equal to 0.01 by default.

4. The random-match probability formulae with inbreeding correction are

homozygotes: $f = p^2 + p(1-p)C_f$

heterozygotes: $f = 2pq(1-C_f)$

Where p and q are defined as above. The inbreeding coefficient (C_f) is a measure of the degree of inbreeding. The value of C_f is configured globally and can be 0.01, 0.03 or other decimal number. The default value for C_f is 0.01.

OK

Ok…so lets look at our other main option

NRC II, 4.10a:

$$\frac{[2\theta + (1-\theta)p_i][3\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)}$$

NRC II, 4.10b:

$$\frac{2[\theta + (1-\theta)p_i][\theta + (1-\theta)p_j]}{(1+\theta)(1+2\theta)}$$

Note: both homozygotes and heterozygotes are treated in this application

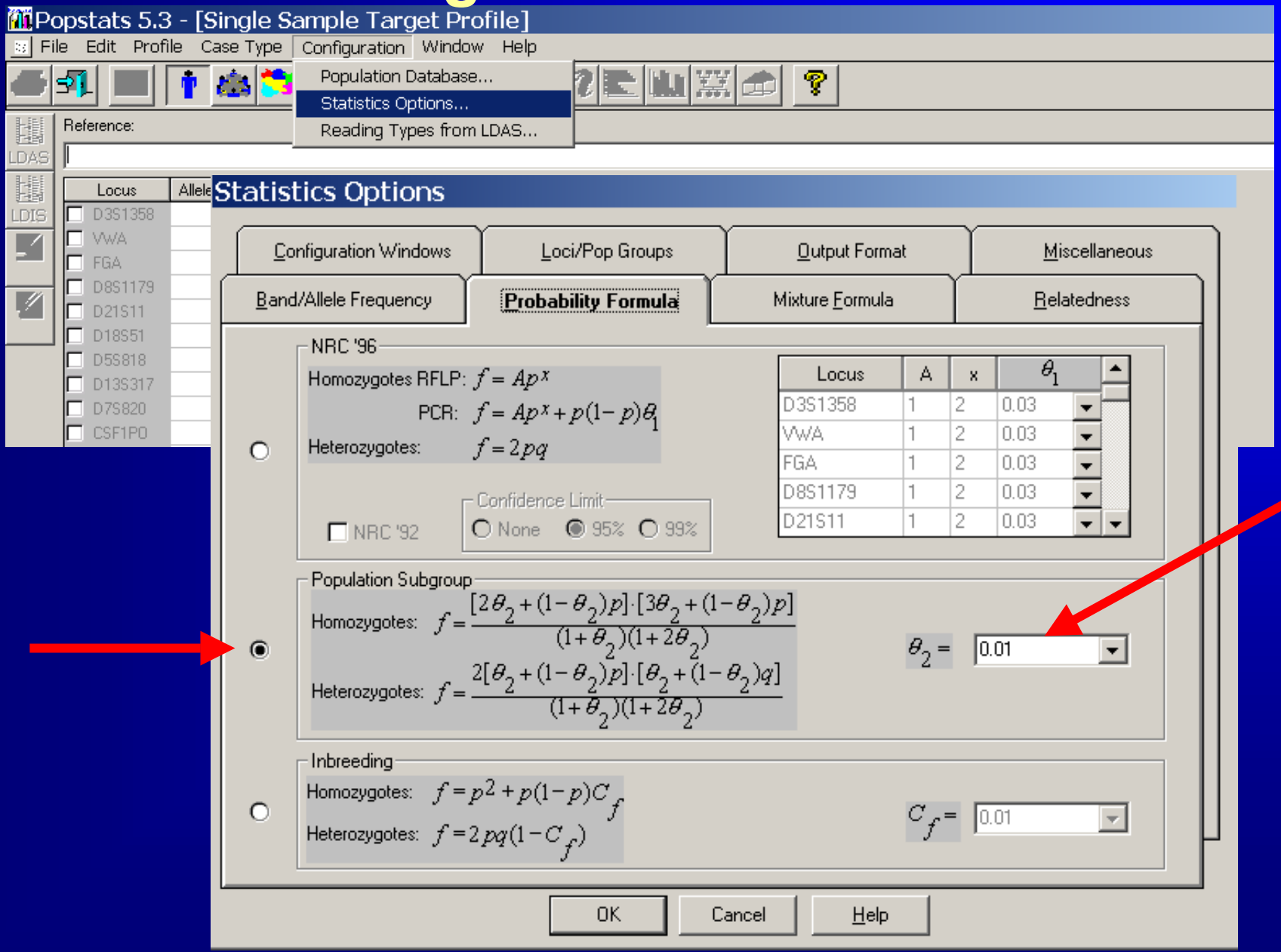
When and why should we consider this??

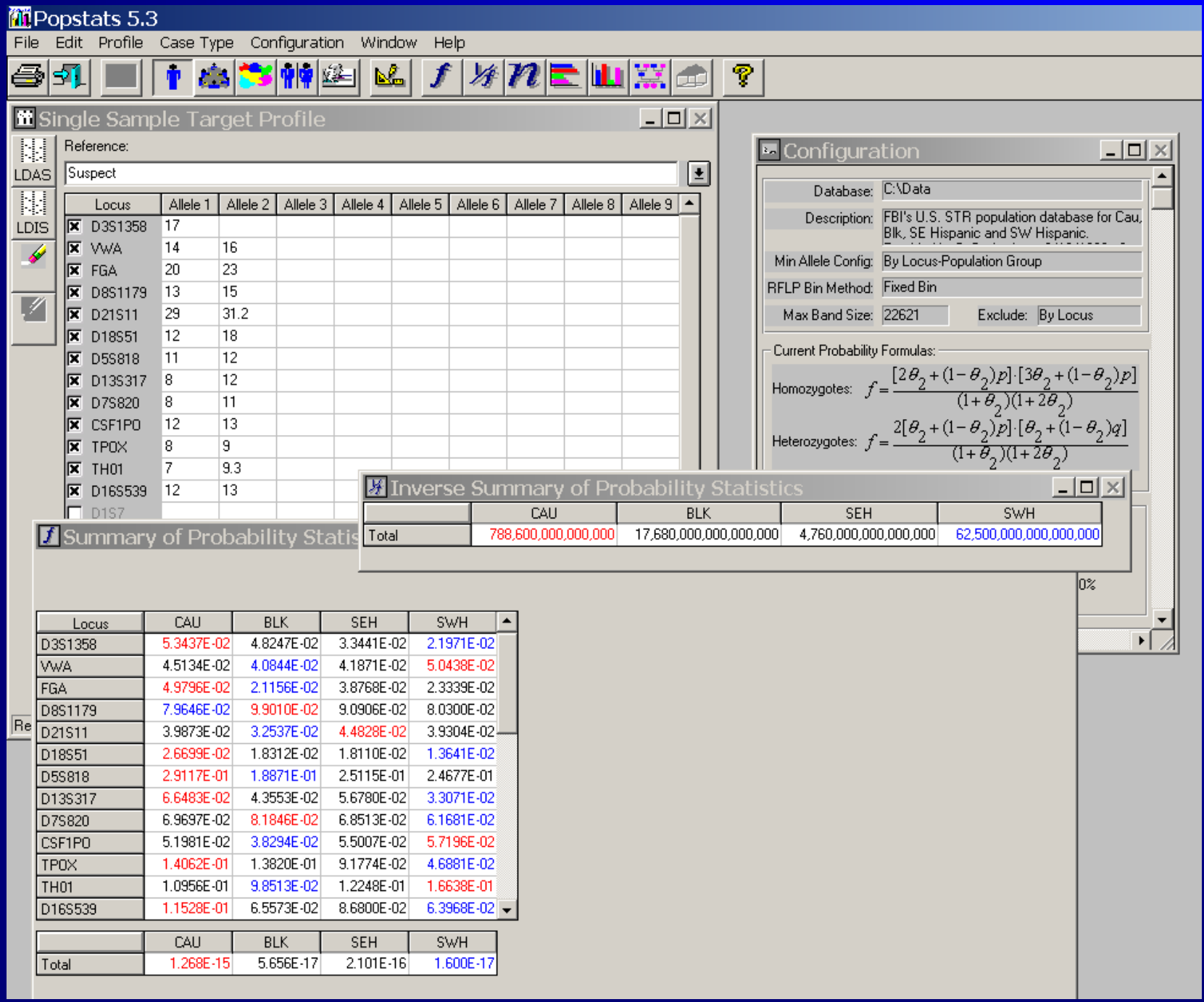
Takes into account the assumption that the person contributing the evidence and the suspect are from the same subgroup

What it gives us is a conditional probability of the suspect genotype given that of the perpetrator.

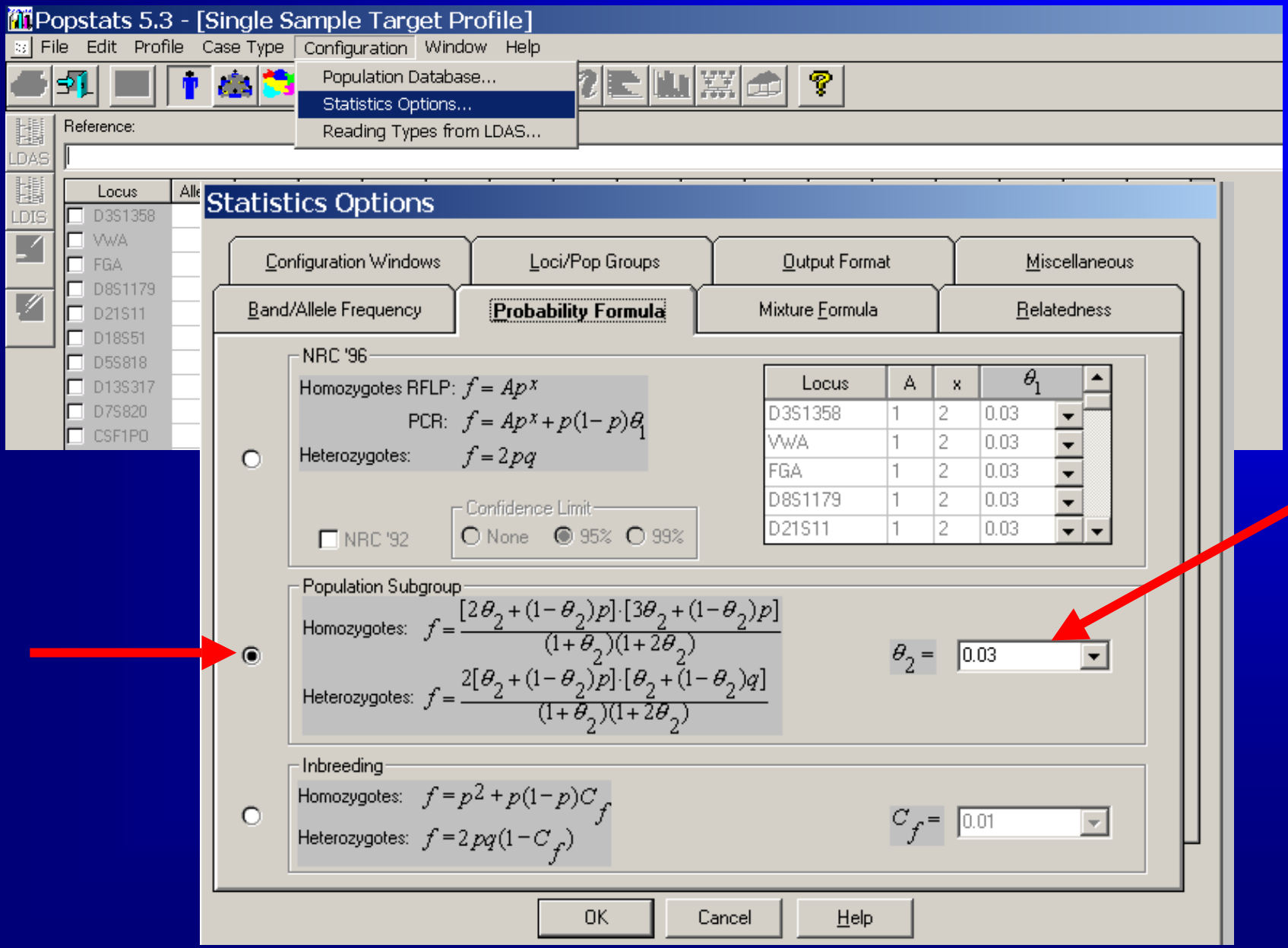
Example… use if the suspect and all possible perpetrators are from the same small isolated town i.e. religious sects, native communities

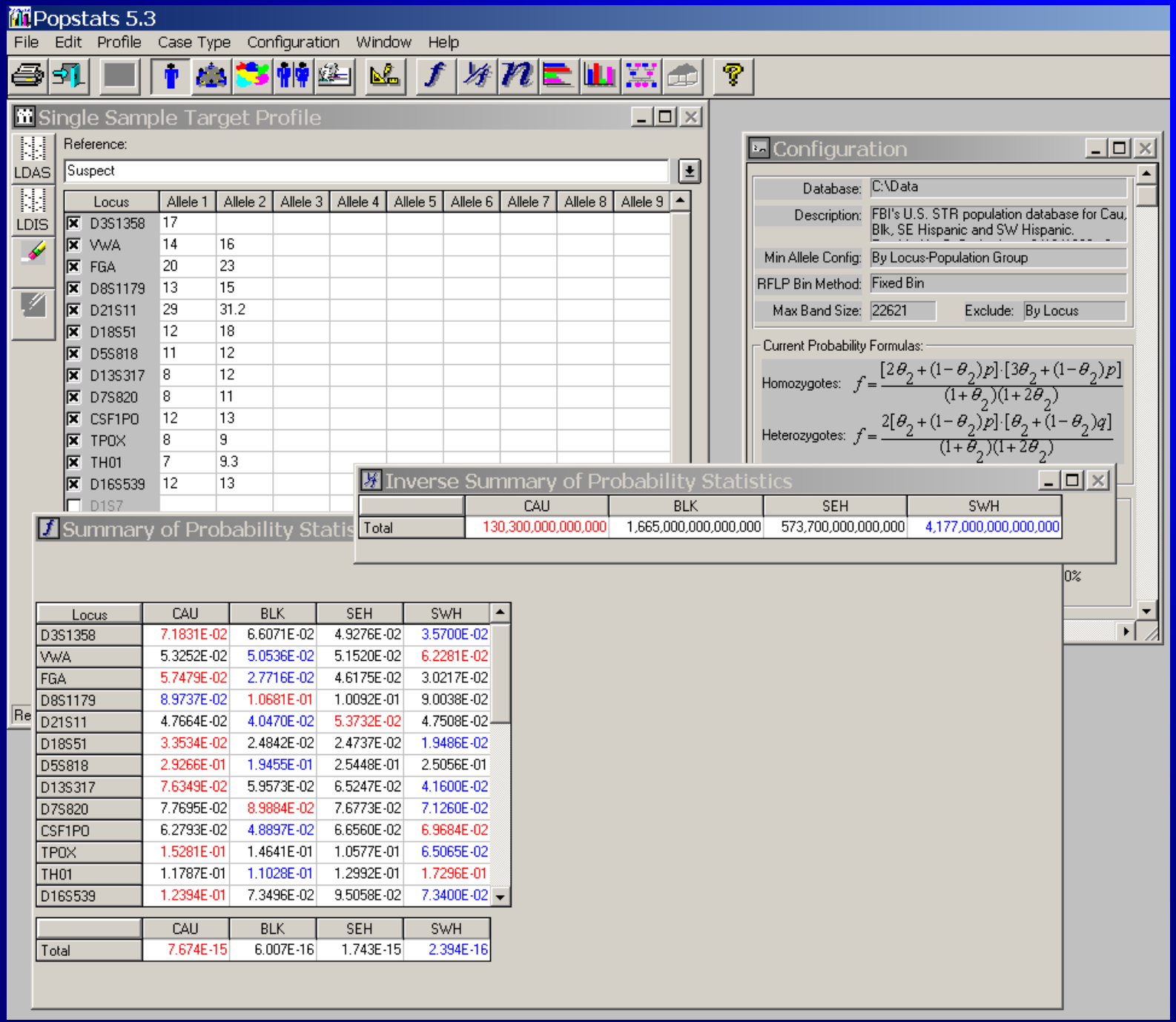
Here it is using the default of 0.01 θ




Popstats 5.3
File Edit Profile Case Type Configuration Window Help
Single Sample Target Profile
LDAS
LDIS
Reference:
Suspect
Locus Allele 1 Allele 2 Allele 3 Allele 4 Allele 5 Allele 6 Allele 7 Allele 8 Allele 9
D3S1358 17
VWA 14 16
FGA 20 23
D8S1179 13 15
D21S11 29 31.2
D18S51 12 18
D5S818 11 12
D13S317 8 12
D7S820 8 11
CSF1PO 12 13
TPOX 8 9
TH01 7 9.3
D16S539 12 13
D1S7
Configuration
Database: C:\Data
Description: FBI's U.S. STR population database for Cau, Blk, SE Hispanic and SW Hispanic.
Min Allele Config: By Locus-Population Group
RFLP Bin Method: Fixed Bin
Max Band Size: 22621
Exclude: By Locus
Current Probability Formulas:
Homozygotes: $f = \frac{[2\theta_2 + (1-\theta_2)p]\cdot[3\theta_2 + (1-\theta_2)p]}{(1+\theta_2)(1+2\theta_2)}$
Heterozygotes: $f = \frac{2[\theta_2 + (1-\theta_2)p]\cdot[\theta_2 + (1-\theta_2)q]}{(1+\theta_2)(1+2\theta_2)}$
Inverse Summary of Probability Statistics
CAU BLK SEH SWH
Total 788,600,000,000,000 17,680,000,000,000,000 4,760,000,000,000,000 62,500,000,000,000,000
Summary of Probability Statis
Locus CAU BLK SEH SWH
D3S1358 5.3437E-02 4.8247E-02 3.3441E-02 2.1971E-02
VWA 4.5134E-02 4.0844E-02 4.1871E-02 5.0438E-02
FGA 4.9796E-02 2.1156E-02 3.8768E-02 2.3339E-02
D8S1179 7.9646E-02 9.9010E-02 9.0906E-02 8.0300E-02
D21S11 3.9873E-02 3.2537E-02 4.4828E-02 3.9304E-02
D18S51 2.6699E-02 1.8312E-02 1.8110E-02 1.3641E-02
D5S818 2.9117E-01 1.8871E-01 2.5115E-01 2.4677E-01
D13S317 6.6483E-02 4.3553E-02 5.6780E-02 3.3071E-02
D7S820 6.9697E-02 8.1846E-02 6.8513E-02 6.1681E-02
CSF1PO 5.1981E-02 3.8294E-02 5.5007E-02 5.7196E-02
TPOX 1.4062E-01 1.3820E-01 9.1774E-02 4.6881E-02
TH01 1.0956E-01 9.8513E-02 1.2248E-01 1.6638E-01
D16S539 1.1528E-01 6.5573E-02 8.6800E-02 6.3968E-02
CAU BLK SEH SWH
Total 1.268E-15 5.656E-17 2.101E-16 1.600E-17

If we want to use 0.03 θ




Popstats 5.3
File Edit Profile Case Type Configuration Window Help
Single Sample Target Profile
LDAS
LDIS
Reference:
Suspect
Locus Allele 1 Allele 2 Allele 3 Allele 4 Allele 5 Allele 6 Allele 7 Allele 8 Allele 9
D3S1358 17
VWA 14 16
FGA 20 23
D8S1179 13 15
D21S11 29 31.2
D18S51 12 18
D5S818 11 12
D13S317 8 12
D7S820 8 11
CSF1PO 12 13
TPOX 8 9
TH01 7 9.3
D16S539 12 13
D1S7
Configuration
Database: C:\Data
Description: FBI's U.S. STR population database for Cau, Blk, SE Hispanic and SW Hispanic.
Min Allele Config: By Locus-Population Group
RFLP Bin Method: Fixed Bin
Max Band Size: 22621
Exclude: By Locus
Current Probability Formulas:
Homozygotes: $f = \frac{[2\theta_2 + (1-\theta_2)p]\cdot[3\theta_2 + (1-\theta_2)p]}{(1+\theta_2)(1+2\theta_2)}$
Heterozygotes: $f = \frac{2[\theta_2 + (1-\theta_2)p]\cdot[\theta_2 + (1-\theta_2)q]}{(1+\theta_2)(1+2\theta_2)}$
Inverse Summary of Probability Statistics
CAU BLK SEH SWH
Total 130,300,000,000,000 1,665,000,000,000,000 573,700,000,000,000 4,177,000,000,000,000
Summary of Probability Statis
0%
Locus CAU BLK SEH SWH
D3S1358 7.1831E-02 6.6071E-02 4.9276E-02 3.5700E-02
VWA 5.3252E-02 5.0536E-02 5.1520E-02 6.2281E-02
FGA 5.7479E-02 2.7716E-02 4.6175E-02 3.0217E-02
D8S1179 8.9737E-02 1.0681E-01 1.0092E-01 9.0038E-02
D21S11 4.7664E-02 4.0470E-02 5.3732E-02 4.7508E-02
D18S51 3.3534E-02 2.4842E-02 2.4737E-02 1.9486E-02
D5S818 2.9266E-01 1.9455E-01 2.5448E-01 2.5056E-01
D13S317 7.6349E-02 5.9573E-02 6.5247E-02 4.1600E-02
D7S820 7.7695E-02 8.9884E-02 7.6773E-02 7.1260E-02
CSF1PO 6.2793E-02 4.8897E-02 6.6560E-02 6.9684E-02
TPOX 1.5281E-01 1.4641E-01 1.0577E-01 6.5065E-02
TH01 1.1787E-01 1.1028E-01 1.2992E-01 1.7296E-01
D16S539 1.2394E-01 7.3496E-02 9.5058E-02 7.3400E-02
CAU BLK SEH SWH
Total 7.674E-15 6.007E-16 1.743E-15 2.394E-16

$\theta = 0.01$

Inverse Summary of Probability Statistics

	CAU	BLK	SEH	SWH
Total	788,600,000,000,000	17,680,000,000,000,000	4,760,000,000,000,000	62,500,000,000,000,000

$\theta = 0.03$

Inverse Summary of Probability Statistics

	CAU	BLK	SEH	SWH
Total	130,300,000,000,000	1,665,000,000,000,000	573,700,000,000,000	4,177,000,000,000,000

So,

PopStats has given us the numbers we desired…

What do we do with them???

Well,

We report them of course!

But we should consider what we are reporting and the information we are conveying in our "statistics"

Source attribution

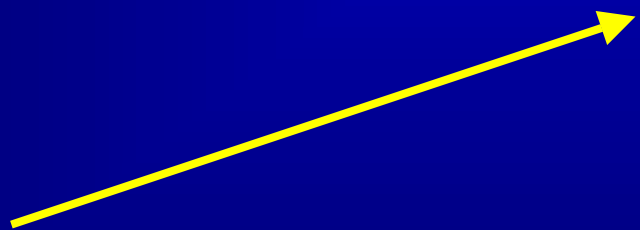
Hot topic for statistical debate

With the current panel of genetic markers available to forensic testing, it is not uncommon for the reciprocal of the random match probability determined for a genetic profile to exceed the worlds population several fold.

So, how do you want to express this fact in your reports and testimony?

What do these *numbers* mean to you?

the prosecutor?
the defense?
the judge?
the jury?



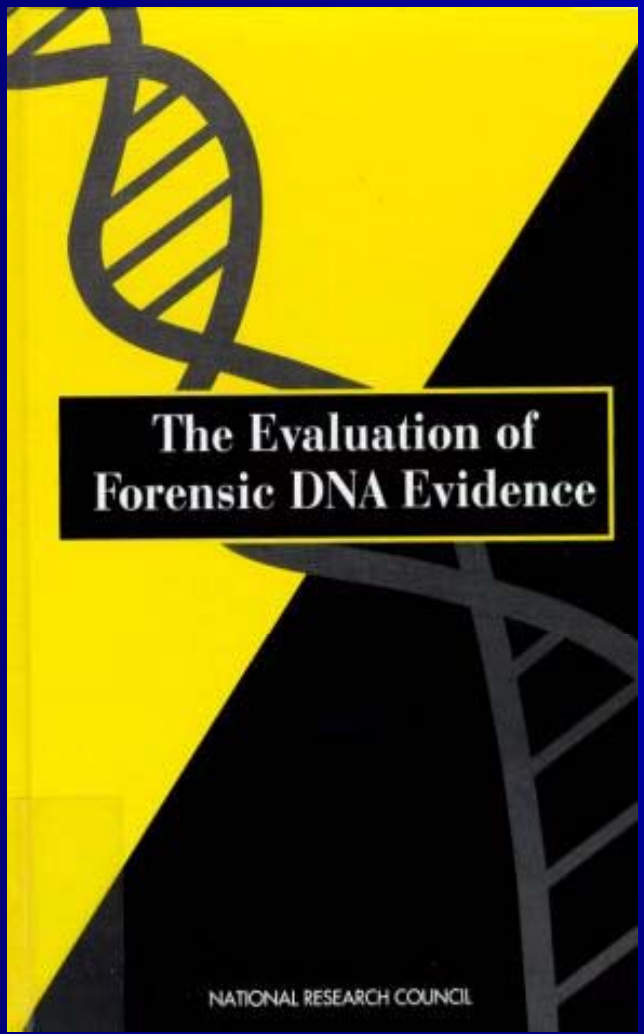
This is what really matters!!!

Big Number Names:

1,000,000	million
1,000,000,000	billion
1,000,000,000,000	trillion
1×10^{15}	quadrillion
1×10^{18}	quintillion
1×10^{21}	sextillion
1×10^{24}	septillion
1×10^{27}	octillion
1×10^{30}	nonillion
1×10^{33}	decillion

Even Bigger Number Names:

1×10^{36}	undecillion
1×10^{39}	duodecillion
1×10^{42}	tredecillion
1×10^{45}	quattordecillion
1×10^{48}	quindecillion
1×10^{51}	sexdecillion
1×10^{54}	septendecillion
1×10^{57}	octodecillion
1×10^{60}	novemdecillion
1×10^{63}	vigintillion



NRC II May 1996

"...that profile might be said to be unique if it is so rare that it becomes unreasonable to suppose that a second person in the population might have the same profile."

To address uniqueness we are back to the same old question…
population sample size

Here the population size differs from what we discussed when calculating allele frequencies…

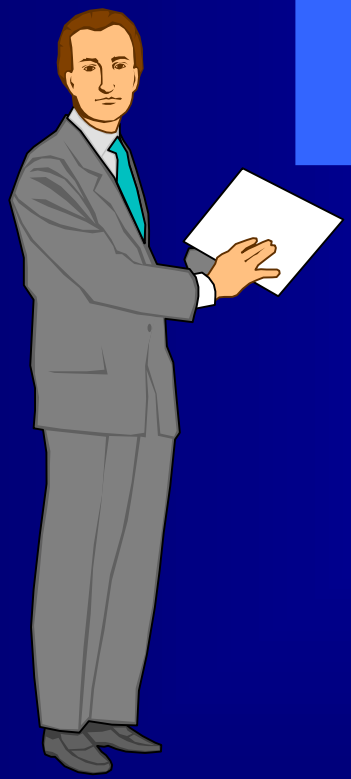
The relevant population is at issue here

Define the Question

(or at least make sure you know what question you are answering)

Define the Question

Estimates of the Rarity of
a DNA Profile:

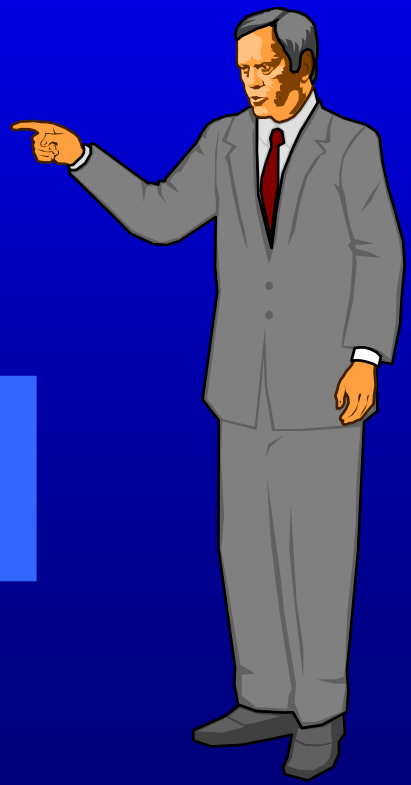


1 in 130 million

Based on
unrelated individuals

1 in 128

Based on
brothers



Uniqueness / Source Attribution
Webster's Definitions

only one

Unusual

Some [circumstance] that is the only one of it kind

Uniqueness / Source Attribution
Webster's Definitions

Attribution evaluated within context of case

Rarely is the world's population the appropriate context

Thus, a circumstance that is the only one of its kind is appropriate context

Uniqueness

A profile that exists in one person and no other (excluding identical twins)

Context?

- **Population of the world…maybe**
- **Population of the US….there is a thought!**
- **Population with access to a crime scene…**

Uniqueness

A profile that exists in one person and no other (excluding identical twins)

Actually we are interested in source attribution, not whether the profile is unique in the world

Is it reasonable to consider the profile to be so rare that one can opine about the source of the evidence?

Let the RMP of a given evidentiary profile X be p_x
(Calculate using NRC II Report Recommendations)

Then $(1-p_x)^N$

is the probability of not observing the profile in a population of N unrelated individuals

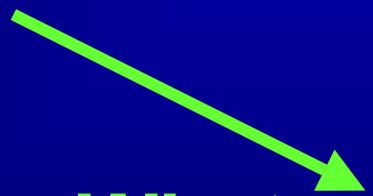
This probability should be greater than or equal to a 1- α confidence level

$$(1-p_x)^N \geq 1-\alpha$$

$$p_x \leq 1-(1-\alpha)^{1/N}$$

Source Attribution

- Specify (1- α) confidence level of 95% or 99% (uses an α of 0.05 or 0.01, respectively)
- Determine RMP threshold to assert with a specific degree of confidence that the particular evidence profile is unique with a population of N unrelated individuals



What population????

Source Attribution Values

Calculate p for major population groups

θ = 0.01or 0.03

Take the most common value for p
Increase p by factor of 10
Determine if $p \leq 1 - (1 - a)^{1/N}$

What N ??

The standard basis that is used here in the US is an estimate of US population of approximately 260 million people

So, taking this and if we accept an α of 0.01 (99% confidence level) with

$$p_x \leq 1-(1-\alpha)^{1/N}$$

A random match probability less than 3.9×10^{-11} would convey at least 99% confidence that the evidentiary profile is unique in the population

RMP thresholds for source attribution at various population sizes and confidence levels

SAMPLE SIZE N	CONFIDENCE LEVELS 0.90	0.95	0.99	0.999
2	5.1×10^{-2}	2.5×10^{-2}	5.0×10^{-3}	5.0×10^{-4}
3	3.5×10^{-2}	1.7×10^{-2}	3.3×10^{-3}	3.3×10^{-4}
4	2.6×10^{-2}	1.3×10^{-2}	2.5×10^{-3}	2.5×10^{-4}
5	2.1×10^{-2}	1.0×10^{-2}	2.0×10^{-3}	2.0×10^{-4}
6	1.7×10^{-2}	8.5×10^{-3}	1.7×10^{-3}	1.7×10^{-4}
7	1.5×10^{-2}	7.3×10^{-3}	1.4×10^{-3}	1.4×10^{-4}
8	1.3×10^{-2}	6.4×10^{-3}	1.3×10^{-3}	1.3×10^{-4}
9	1.2×10^{-2}	5.7×10^{-3}	1.1×10^{-3}	1.1×10^{-4}
10	1.1×10^{-2}	5.1×10^{-3}	1.0×10^{-3}	1.0×10^{-4}
25	4.2×10^{-3}	2.1×10^{-3}	4.0×10^{-4}	4.0×10^{-5}
50	2.1×10^{-3}	1.0×10^{-3}	2.0×10^{-4}	2.0×10^{-5}
100	1.1×10^{-3}	5.1×10^{-4}	1.0×10^{-4}	1.0×10^{-5}
1×10^{3}	1.1×10^{-4}	5.1×10^{-5}	1.0×10^{-5}	1.0×10^{-6}
1×10^{5}	1.1×10^{-6}	5.1×10^{-7}	1.0×10^{-7}	1.0×10^{-8}
1×10^{6}	1.1×10^{-7}	5.1×10^{-8}	1.0×10^{-8}	1.0×10^{-9}
1×10^{7}	1.1×10^{-8}	5.1×10^{-9}	1.0×10^{-9}	1.0×10^{-10}
5×10^{7}	2.1×10^{-9}	1.0×10^{-9}	2.0×10^{-10}	2.0×10^{-11}
2.6×10^{8}	4.1×10^{-10}	2.0×10^{-10}	3.9×10^{-11}	3.9×10^{-12}
1×10^{9}	1.1×10^{-10}	5.1×10^{-11}	1.0×10^{-11}	1.0×10^{-12}
5×10^{9}	2.1×10^{-11}	1.0×10^{-11}	2.0×10^{-12}	2.0×10^{-13}

So with our PopStats results obtained for a θ = 0.01

Locus	CAU	BLK	SEH	SWH
D3S1358	4.6529E-02	4.1600E-02	2.7701E-02	1.7185E-02
VWA	4.1106E-02	3.5938E-02	3.6987E-02	4.4303E-02
FGA	4.6005E-02	1.8050E-02	3.5152E-02	2.0049E-02
D8S1179	7.4442E-02	9.5057E-02	8.5725E-02	7.5293E-02
D21S11	3.6039E-02	2.8637E-02	4.0358E-02	3.5239E-02
D18S51	2.3427E-02	1.5228E-02	1.4971E-02	1.0950E-02
D5S818	2.9041E-01	1.8569E-01	2.4943E-01	2.4480E-01
D13S317	6.1431E-02	3.5080E-02	5.2533E-02	2.8834E-02
D7S820	6.5690E-02	7.7793E-02	6.4360E-02	5.6800E-02
CSF1PO	4.6424E-02	3.2880E-02	4.9027E-02	5.0685E-02
TPOX	1.3412E-01	1.3395E-01	8.4350E-02	3.7185E-02
TH01	1.0530E-01	9.2329E-02	1.1869E-01	1.6298E-01
D16S539	1.1082E-01	[illegible]	[illegible]	5.9165E-02

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

3.9 x 10^{-11}

Profile frequency is less than 99% threshold

So with our PopStats results obtained for a $\theta = 0.03$

Locus	CAU	BLK	SEH	SWH
D3S1358	4.9867E-02	4.4800E-02	3.0420E-02	1.9400E-02
VWA	4.1106E-02	3.5938E-02	3.6987E-02	4.4303E-02
FGA	4.6005E-02	1.8050E-02	3.5152E-02	2.0049E-02
D8S1179	7.4442E-02	9.5057E-02	8.5725E-02	7.5293E-02
D21S11	3.6039E-02	2.8637E-02	4.0358E-02	3.5239E-02
D18S51	2.3427E-02	1.5228E-02	1.4971E-02	1.0950E-02
D5S818	2.9041E-01	1.8569E-01	2.4943E-01	2.4480E-01
D13S317	6.1431E-02	3.5080E-02	5.2533E-02	2.8834E-02
D7S820	6.5690E-02	7.7793E-02	6.4360E-02	5.6800E-02
CSF1PO	4.6424E-02	3.2880E-02	4.9027E-02	5.0685E-02
TPOX	1.3412E-01	1.3395E-01	8.4350E-02	3.7185E-02
TH01	1.0530E-01	9.2329E-02	1.1869E-01	1.6298E-01
D16S539	1.1082E-01	8.1613E-02	8.2613E-02	5.9165E-02

	CAU	BLK	SEH	SWH
Total	5.047E-16	1.530E-17	7.005E-17	3.648E-18

3.9×10^{-11}

Profile frequency is less than 99% threshold

"To a reasonable degree of scientific certainty, _____is the source of the DNA in specimen Q2."

"I have a high degree of confidence, _____is the source of the DNA in specimen Q2."

"Based on an estimate of 260 million people resident in the population of the United States, there is 99% confidence that _____is the source of the DNA in specimen Q2."

We are not stating that _____ is the only person to possess that profile. We are stating that we would not expect to find it in a population of N individuals.

Source Attribution

- Method is simple
- Conservative because N is so large (260,000,000)
- If N = 260,000,000, then RMP threshold is 3.9×10^{-11}
- Most of the time the RMP is far less, so confidence is greater than 0.99

Source Attribution

- N can be configured to context of the case
- Two individuals to entire town, state, or whatever
- Laboratory policy to set N

Random match probability is NOT

Chance that someone else is guilty

Chance that someone else left the bloodstain

Chance of defendant not being guilty

"We are neither hunters nor gatherers. We are statisticians."

Intermission