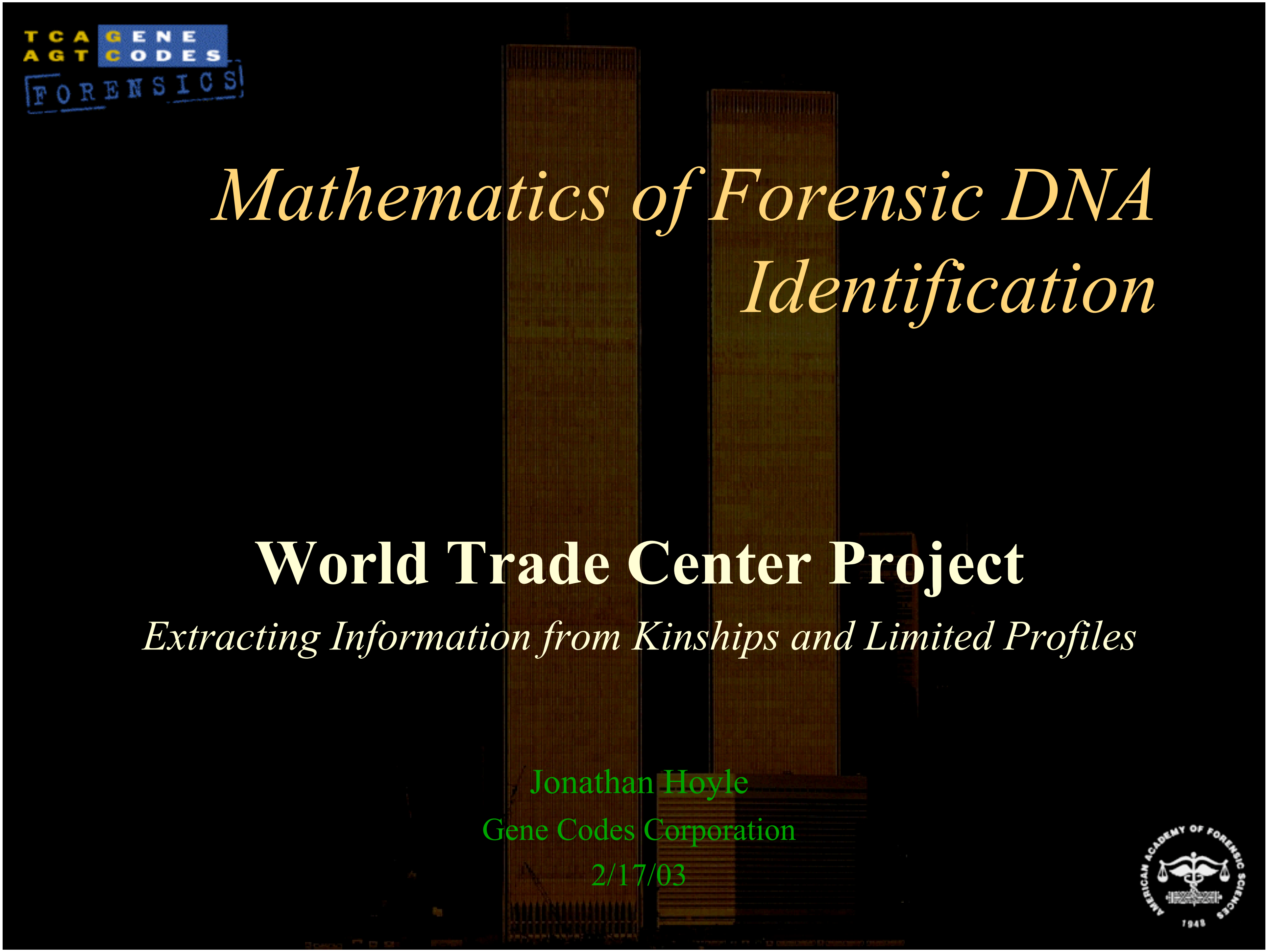


TCA GENE
AGT CODES
FORENSICS

Mathematics of Forensic DNA Identification

World Trade Center Project

Extracting Information from Kinships and Limited Profiles

Jonathan Hoyle
Gene Codes Corporation

2/17/03

AMERICAN ACADEMY OF FORENSIC SCIENCES
1948

TCA GENE
AGT CODES
FORENSICS

Introduction

- 2,795 people were killed in the World Trade Center attacks on September 11, 2001.
- 20,000 remains were recovered, the vast majority of which would require DNA matching for identification.
- Existing software tools for DNA identification proved wholly inadequate for the scope and magnitude of this project.

Timeline

- September 17: Armed Forces DNA Identification Lab [AFDIL] asks Gene Codes to update *Sequencher*™ for the Pentagon and Shanksville crashes.
- September 28: Office of the Chief Medical Examiner [OCME] in New York City contacts us for new software.
- October 15: Using the *Extreme Programming* [XP] methodology, software development is underway.
- December 13: *M-FISys* (**M**ass-**F**atality **I**dentification **Sys**tem) has its first release to the OCME.
- Since: Weekly releases personally delivered to the OCME, to accommodate rapidly changing requirements.

Identification Technologies

- Technologies used for Identification
  - STR
  - mtDNA
  - SNP
- Methods used:
  - Direct Match to a Personal Effect
  - Kinship Analysis

STR: Short Tandem Repeats

- A repeat of a short sequence of bases (4 or 5)
- For example, at locus position D7S280, it is the four base sequence `gata` we look for:

`...gatagatagatagatagatagatagatgtttatctc...`

- In the above example, `gata` is repeated 6 times with a 3-base partial repeat.
- "6.3" is therefore assigned for this allele.
- Being diploid, we have two alleles per locus, thus (up to) two values are stored, e.g. 6.3/8.

STR Frequency

- In 1997, the FBI standardized on 13 STR loci used in the national database, *CoDIS*.
- Frequency data for each locus/allele value is available for various races. For example:

Locus:	D16S539	TPOX	D3S1358	FGA	D7S820	vWA	D13S317	TH01
Allele:	11/13	8	15.2	21/13.2	10/11	15.2	11/12	9.3
Freq:	8.55%	39.4%	0.099%	0.796%	14.6%	0.099%	18.2%	9.21%

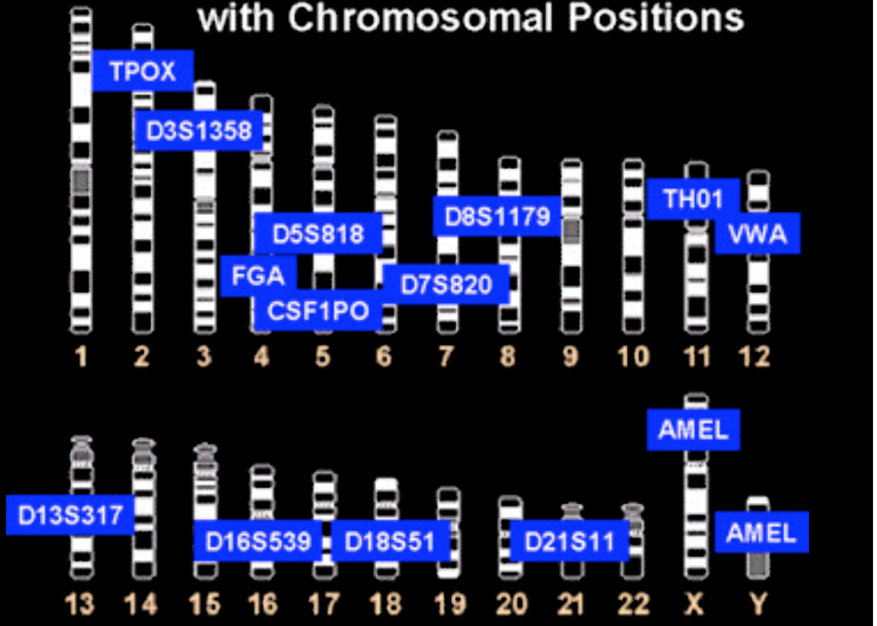
- Since STR loci are independent, these frequencies can be multiplied: 5.6×10^{-13}
- Likelihood = 1 / Frequency = 1.8×10^{12}

STR Profiles

- *M-FISys* STR profile contains 16 elements:
  - Amelogenin (Gender)
  - 13 CoDIS Core Loci
  - 2 PowerPlex Loci:
    - Penta D
    - Penta E
- Minimum Likelihood:

  7.6×10^{15}

13 CODIS Core STR Loci with Chromosomal Positions



STR Likelihood Threshold

- OCME wants a minimum likelihood for identification which ensures a chance of a mismatch to be less than 1 in a million.
- Assuming a population of 5000, what is the smallest *n* such that a 10^n min likelihood yields a mismatch prob $< 10^{-6}$?
- Since likelihood is the inverse of probability, $p = 1 / 10^n$
- The probability of no mismatch is $q = 1 - p = 1 - 1 / 10^n$
- The prob. of no mismatch in $5000 = 1 - q^{5000} = 1\text{-}(1\text{-}1/10^n)^{5000}$
- Thus we have the inequality:

$$1 - (1 - 1 / 10^n)^{5000} < 1 / 1{,}000{,}000$$

- Solving for *n* we get:

$$n > -\log_{10}(1 - (1 - 10^{-6})^{1/5000}) = 9.699$$

- Therefore we set $n = 10$.

TCA GENE
AGT CODES
FORENSICS

Direct STR Identification

- A victim remain (called a disaster sample) can be identified by direct match if its profile is either:
  - complete and matches Personal Effects (2 modalities)
  - partial with no mismatches, with a likelihood ≥ 10^{10} amongst common loci
- A sample was further investigated if its STR profile likelihood ≥ 10^{10} and with either:
  - a single mismatch only, supported by Kinship
  - mismatches due only to allelic dropout

Partial Profiles

- All STR profiles containing at least 11 CoDIS markers or more will have likelihoods $\geq 10^{10}$
- 70% of the victim samples yielded partial profiles (missing at least one CoDIS marker)
- 25% of these partial profiles had likelihood values $\geq 10^{10}$
- Leaving half of victim samples which cannot be identified through STR means alone (using these parameters).

STR Likelihood: Locus Probability

- Likelihood = 1 / Probability Frequency
- OCME has locus-allele frequency data
- Locus Probability can be first approximated by ignoring population structure and using the *Hardy-Weinberg proportions*:

  p^2 for homozygous alleles: p = frequency of allele

  $2pq$ for heterozygous alleles: p,q = frequency of each allele
- Above assumes an infinite population with random mating

STR Likelihood: θ

- Because the population is finite, we introduce the inbreeding coefficient θ
- Factoring this into the H-W equations:

$p^2 + p(1-p)\theta$	for homozygous alleles
$2pq(1-\theta)$	for heterozygous alleles

- Because θ is very small, 1-θ is close to 1, we round it to remain conservative:

$p^2 + p(1-p)\theta$	for homozygous alleles
$2pq$	for heterozygous alleles

- OCME chooses the standard θ = 0.03

STR Likelihood: Profiles

- Once we have calculated the probability frequency for each locus, we can calculate the likelihood of the entire profile:
- If P_k (A_k) is the probability of allele A at locus k, we can define the likelihood of STR profile S as:

$$L(S) = \prod_{k \in \text{Alleles}} 1 / P_k (A_k)$$

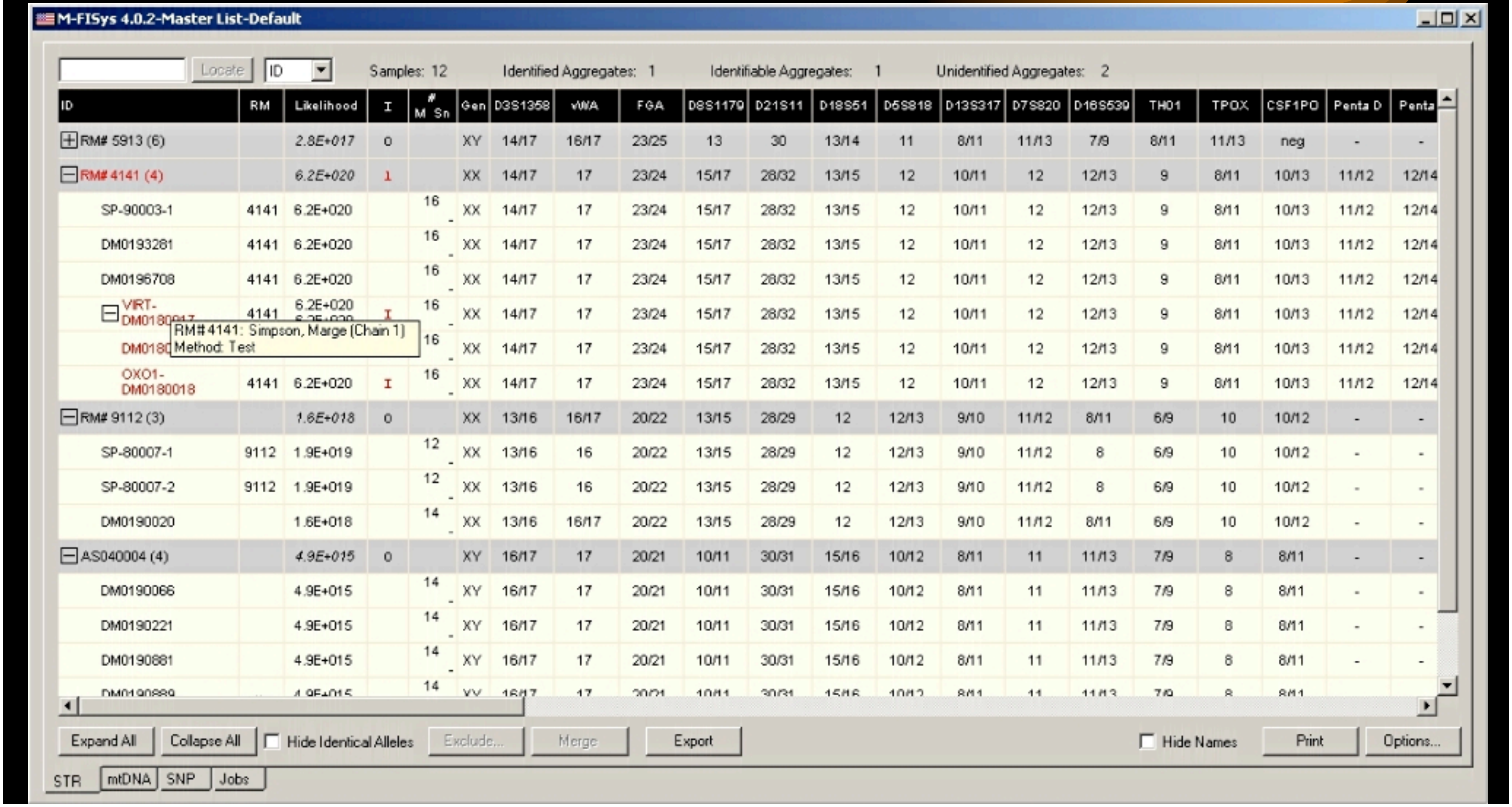
- Note that this works even for partial profiles

STR Likelihood: Race

- OCME has frequency values for four population groups: Asian, Black, Caucasian & Hispanic
- Cannot always rely on reported race, and the race is unknown for a disaster sample
- *M-FISys* computes the Likelihood value across all four races and chooses the lowest value, just to be on the safer, more conservative side.

TCA GENE
AGT CODES
FORENSICS

M-FISys STR Master List



M-FISys 4.0.2-Master List-Default

Locate | ID ▼ | Samples: 12 | Identified Aggregates: 1 | Identifiable Aggregates: 1 | Unidentified Aggregates: 2

ID	RM	Likelihood	I	# M Sn	Gen	D3S1358	vWA	FGA	D8S1179	D21S11	D18S51	D5S818	D13S317	D7S820	D16S539	TH01	TPOX	CSF1PO	Penta D	Penta
⊞ RM# 5913 (6)		*2.8E+017*	0		XY	14/17	16/17	23/25	13	30	13/14	11	8/11	11/13	7/9	8/11	11/13	neg	-	-
⊟ RM# 4141 (4)		*6.2E+020*	1		XX	14/17	17	23/24	15/17	28/32	13/15	12	10/11	12	12/13	9	8/11	10/13	11/12	12/14
SP-90003-1	4141	6.2E+020		16 -	XX	14/17	17	23/24	15/17	28/32	13/15	12	10/11	12	12/13	9	8/11	10/13	11/12	12/14
DM0193281	4141	6.2E+020		16 -	XX	14/17	17	23/24	15/17	28/32	13/15	12	10/11	12	12/13	9	8/11	10/13	11/12	12/14
DM0196708	4141	6.2E+020		16 -	XX	14/17	17	23/24	15/17	28/32	13/15	12	10/11	12	12/13	9	8/11	10/13	11/12	12/14
⊟ VIRT-DM0180[illegible]	4141	6.2E+020 [illegible]	I	16 -	XX	14/17	17	23/24	15/17	28/32	13/15	12	10/11	12	12/13	9	8/11	10/13	11/12	12/14
DM0180[illegible]	[illegible]	[illegible]	[illegible]	16 -	XX	14/17	17	23/24	15/17	28/32	13/15	12	10/11	12	12/13	9	8/11	10/13	11/12	12/14
OXO1-DM0180018	4141	6.2E+020	I	16 -	XX	14/17	17	23/24	15/17	28/32	13/15	12	10/11	12	12/13	9	8/11	10/13	11/12	12/14
⊟ RM# 9112 (3)		*1.6E+018*	0		XX	13/16	16/17	20/22	13/15	28/29	12	12/13	9/10	11/12	8/11	6/9	10	10/12	-	-
SP-80007-1	9112	1.9E+019		12 -	XX	13/16	16	20/22	13/15	28/29	12	12/13	9/10	11/12	8	6/9	10	10/12	-	-
SP-80007-2	9112	1.9E+019		12 -	XX	13/16	16	20/22	13/15	28/29	12	12/13	9/10	11/12	8	6/9	10	10/12	-	-
DM0190020		1.6E+018		14 -	XX	13/16	16/17	20/22	13/15	28/29	12	12/13	9/10	11/12	8/11	6/9	10	10/12	-	-
⊟ AS040004 (4)		*4.9E+015*	0		XY	16/17	17	20/21	10/11	30/31	15/16	10/12	8/11	11	11/13	7/9	8	8/11	-	-
DM0190066		4.9E+015		14 -	XY	16/17	17	20/21	10/11	30/31	15/16	10/12	8/11	11	11/13	7/9	8	8/11	-	-
DM0190221		4.9E+015		14 -	XY	16/17	17	20/21	10/11	30/31	15/16	10/12	8/11	11	11/13	7/9	8	8/11	-	-
DM0190881		4.9E+015		14 -	XY	16/17	17	20/21	10/11	30/31	15/16	10/12	8/11	11	11/13	7/9	8	8/11	-	-
DM0190889		4.9E+015		14	XY	16/17	17	20/21	10/11	30/31	15/16	10/12	8/11	11	11/13	7/9	8	8/11		

RM# 4141: Simpson, Marge (Chain 1)
Method: Test

STR: Kinship Analysis

- Many times there was not sufficient data to perform an STR direct match.
- Cheek swabs from family members of missing persons are taken, and a pedigree tree in *M-FISys* can be generated.
- Likelihoods are calculated on victim samples to determine to which pedigree(s) they belong.
- Kinship Analysis was not performed if more than one relative was in the victim list.

Kinship Analysis: Likelihood

- As with direct STR, Kinship Likelihood is:
  - the product of Locus Likelihoods over common loci
  - the Likelihood Ratio ≥ 10^6
  - calculated across all four races, using the lowest, most conservative value
  - uses frequency data from the OCME
- Analysis was performed for these relations: Parent-Child, Full Sibling, Half Sibling

Kinship Algorithm

- *M-FISys* uses the Kinship algorithm as implemented by Dr. George Carmody of *Carleton University*
- Kinship Locus Likelihood defined as:

$$k = r_2x_2 + r_1x_1 + r_0x_0$$

- where the r_i's are relationship proportions:

Parent-Child:	$r_2 = 0$	$r_1 = 1$	$r_0 = 0$
Full Sibling:	$r_2 = 1/4$	$r_1 = 1/2$	$r_0 = 1/4$
Half Sibling:	$r_2 = 1/2$	$r_1 = 1/2$	$r_0 = 0$
First Cousin:	$r_2 = 3/4$	$r_1 = 1/4$	$r_0 = 0$

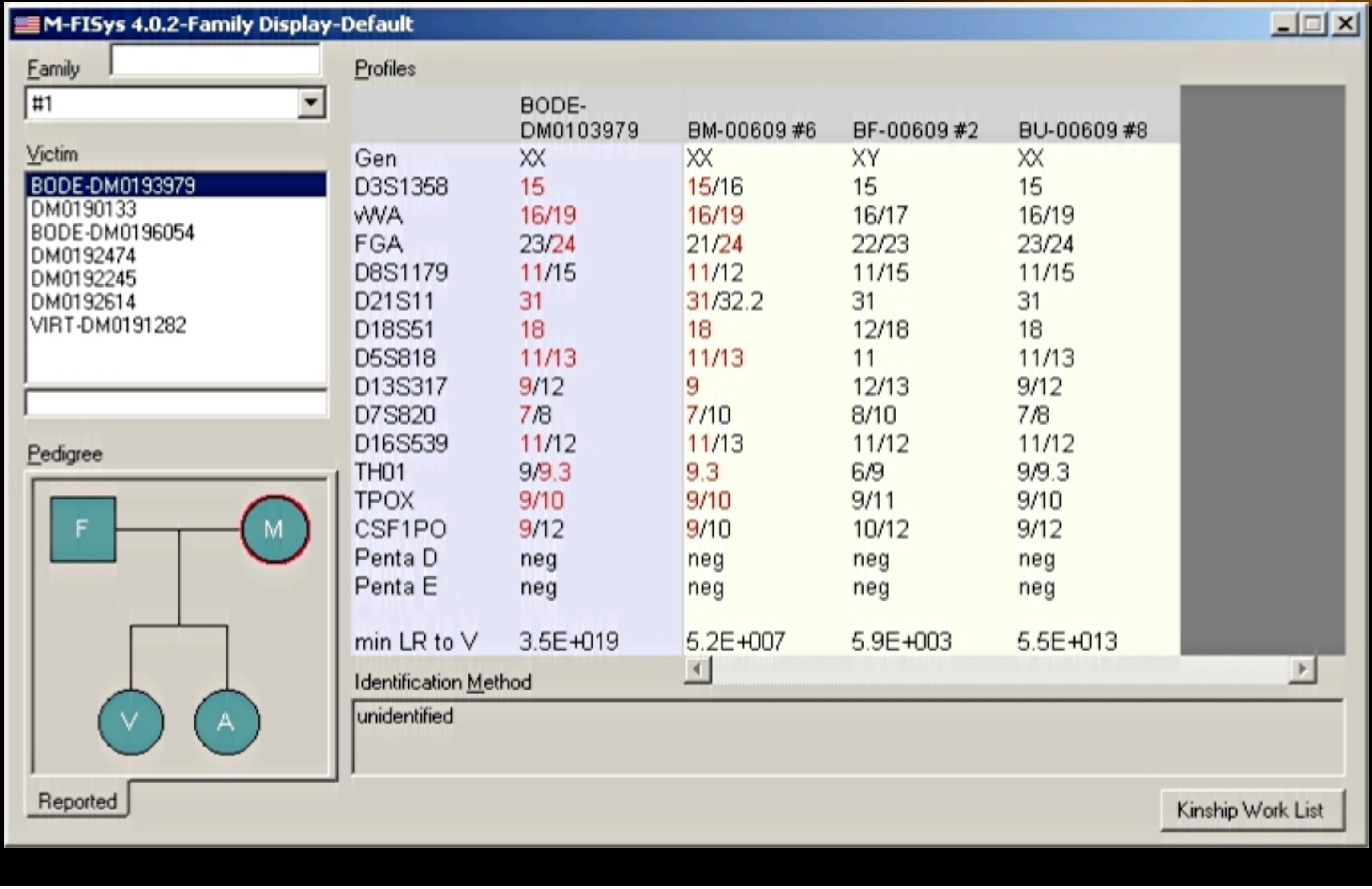
Kinship Algorithm

- and with p & q the frequencies of the high & low alleles resp., the x_i's are defined as:

X_2	$= p^2$	if victim is homozygous and matches an allele
	$= 2pq$	otherwise
X_1	$= 0$	if relative & victim share no common allele
	$= p$	if relative homozygous & shares low allele
	$= q$	if relative homozygous & shares high allele
	$= p/2$	if relative heterozygous & shares low allele
	$= q/2$	if relative heterozygous & shares high allele
	$= (p+q)/2$	if relative & victim are identical
X_0	$= 1$	if relative & victim alleles are identical
	$= 0$	otherwise

TCA GENE
AGT CODES
FORENSICS

M-FISys Kinship Form



M-FISys 4.0.2-Family Display-Default

Family: #1

Victim:
- BODE-DM0193979
- DM0190133
- BODE-DM0196054
- DM0192474
- DM0192245
- DM0192614
- VIRT-DM0191282

Pedigree

Profiles

	BODE-DM0103979	BM-00609 #6	BF-00609 #2	BU-00609 #8
Gen	XX	XX	XY	XX
D3S1358	15	15/16	15	15
vWA	16/19	16/19	16/17	16/19
FGA	23/24	21/24	22/23	23/24
D8S1179	11/15	11/12	11/15	11/15
D21S11	31	31/32.2	31	31
D18S51	18	18	12/18	18
D5S818	11/13	11/13	11	11/13
D13S317	9/12	9	12/13	9/12
D7S820	7/8	7/10	8/10	7/8
D16S539	11/12	11/13	11/12	11/12
TH01	9/9.3	9.3	6/9	9/9.3
TPOX	9/10	9/10	9/11	9/10
CSF1PO	9/12	9/10	10/12	9/12
Penta D	neg	neg	neg	neg
Penta E	neg	neg	neg	neg
min LR to V	3.5E+019	5.2E+007	5.9E+003	5.5E+013

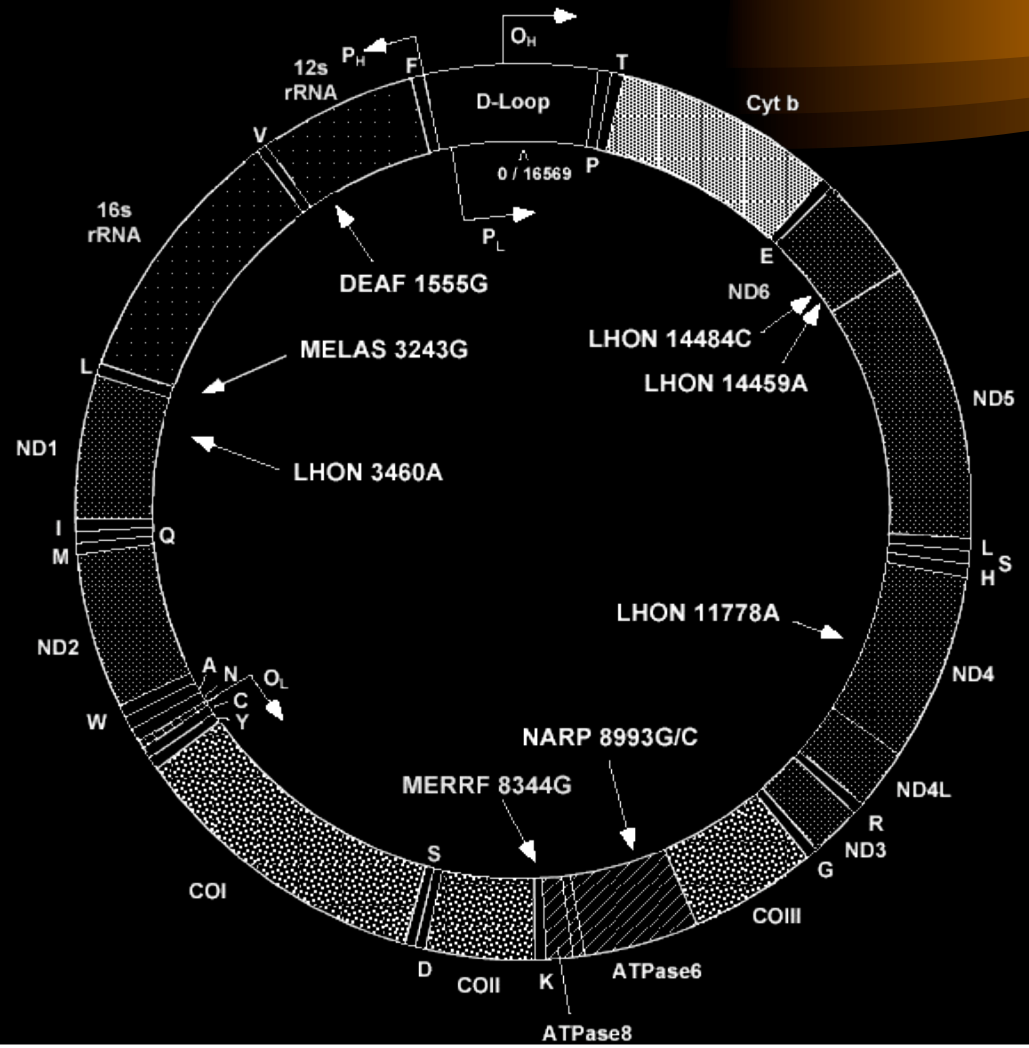
Identification Method: unidentified

TCA GENE
AGT CODES
FORENSICS

Mitochondrial DNA Analysis

- Some victim samples were so degraded that sufficient STR data was not available for either direct STR match or Kinship analysis.
- mtDNA is hardier material, surviving under conditions which nuclear DNA degrades
- mtDNA is a 16,569-based circular genome.
- It is maternally inherited, and thus not unique.
- 5% of the Caucasian population share the same common mitotype.

mtDNA Map


O_H
P_H
F
12s rRNA
D-Loop
T
Cyt b
0 / 16569
P
P_L
V
16s rRNA
DEAF 1555G
E
ND6
LHON 14484C
LHON 14459A
MELAS 3243G
L
ND5
ND1
LHON 3460A
I
Q
M
L
S
H
LHON 11778A
ND2
ND4
A
N
O_L
C
Y
W
NARP 8993G/C
MERRF 8344G
ND4L
R
ND3
G
COIII
COI
S
D
COII
K
ATPase6
ATPase8

TCA GENE
AGT CODES
FORENSICS

mtDNA Analysis

- Mito-typing involves direct sequencing of two highly variable regions of mtDNA.
- The two areas used for mitotyping (HV1 & HV2) are not in a coding region.
- Only a sample's differences from *the Anderson Sequence* (an internationally accepted standard) need be tracked.
- However, 25% of the WTC victims had no maternally-related kin samples.

Mito Likelihood

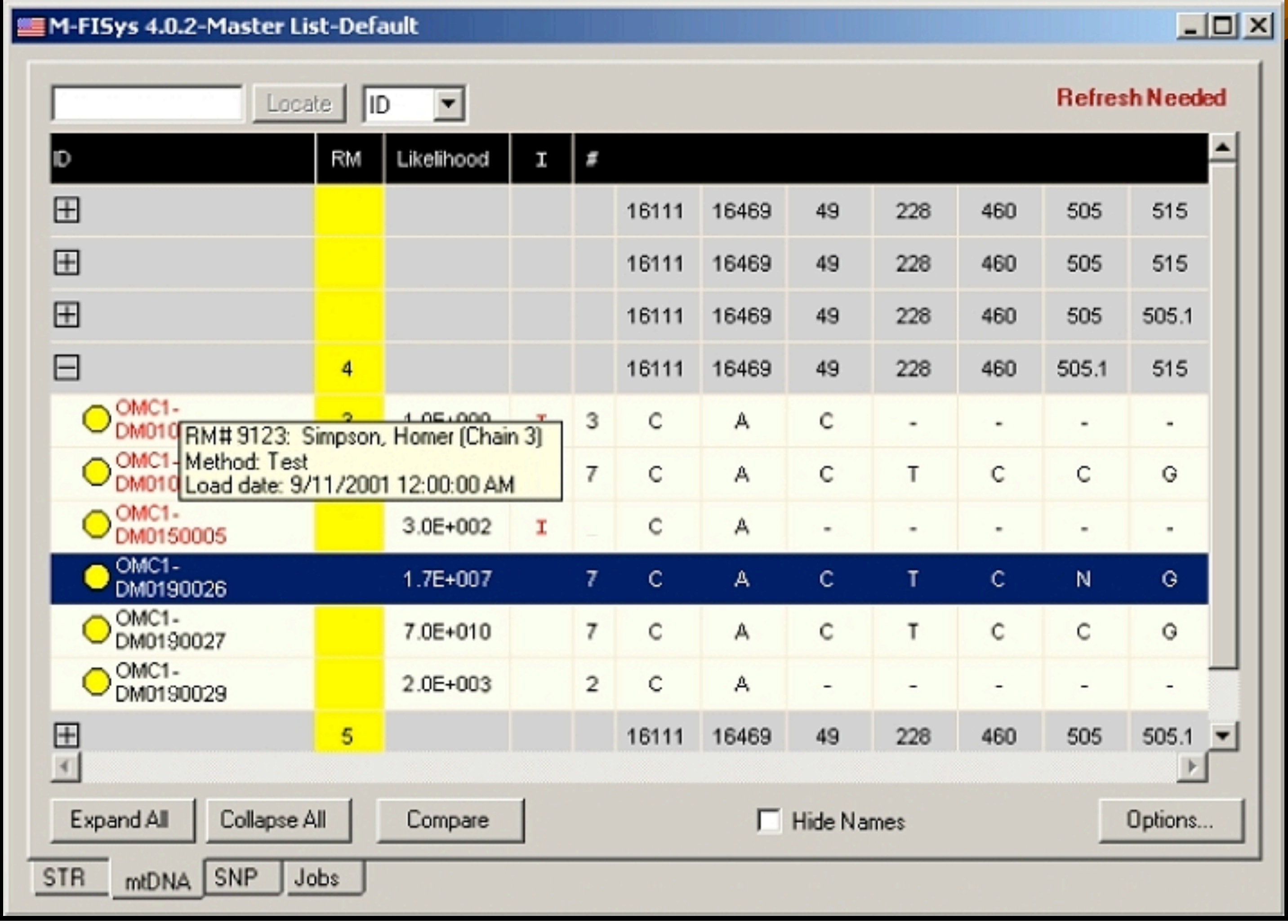
- To determine likelihood for a given mitotype, we begin by counting its frequency *x* in the FBI mtCoDIS data of size *n*.
- The 95% confidence interval for a population proportion with Binomial distribution is estimated by the formula:

  $$[\ \mu - 1.96\sigma/\sqrt{n},\ \mu + 1.96\sigma/\sqrt{n}\ ]$$

  where μ is the mean and σ is the standard deviation.
- Since the probability *p* is just the number database hits, we set $p = x/n$, and so we have $\mu = p$ and $\sigma = \sqrt{p(1-p)}$.
- Thus we have as the upper bound: $x/n + \sqrt{x\,(n-x)/n}$.
- If there are no database entries, we use: $1 - \alpha^{1/n}$ with $\alpha = 0.05$
- Likelihood = 1 / Frequency

TCA GENE
AGT CODES
FORENSICS

M-FISys mtDNA Form



ID	RM	Likelihood	I	#							
⊞					16111	16469	49	228	460	505	515
⊞					16111	16469	49	228	460	505	515
⊞					16111	16469	49	228	460	505	505.1
⊟	4				16111	16469	49	228	460	505.1	515
● OMC1-DM010	3	1.0E+000	I	3	C	A	C	-	-	-	-
● OMC1-DM010				7	C	A	C	T	C	C	G
● OMC1-DM0150005		3.0E+002	I		C	A	-	-	-	-	-
● OMC1-DM0190026		1.7E+007		7	C	A	C	T	C	N	G
● OMC1-DM0190027		7.0E+010		7	C	A	C	T	C	C	G
● OMC1-DM0190029		2.0E+003		2	C	A	-	-	-	-	-
⊞	5				16111	16469	49	228	460	505	505.1

RM# 9123: Simpson, Homer (Chain 3)
Method: Test
Load date: 9/11/2001 12:00:00 AM

Introduction to SNP's

- Single Nucleotide Polymorphisms
- Represents single base differences
- Work pioneered by the GeneScreen division of Orchid Biosciences
- By being able to collect data from very short sequences, this technology offers a great deal of hope for the identification of badly degraded samples

SNP Selection

- SNP’s occur on average every 100-300 bases within the human genome.
- 2 out of every 3 SNP’s involve replacing a C with a T.
- Of these, there is a panel of 70 which are chosen, specifically those in which C and T are equally likely.

SNP Likelihood

- A complete profile of 70 SNP's each with an *independent* probability of 1/2 would yield a likelihood of match at $2^{70} \approx 10^{21}$.
- The probabilities are independent if the SNP's are *unlinked*, which we define to be at least 50MB apart.
- Unfortunately, it is **not** possible to have 70 SNP's 50MB apart in a 3GB genome.

SNP Independence

- A study by Dr. Ranajit Chakraborty of the *Center for Genome Information* concluded:
  - – Allelic dependence is very low: 5.71% as compared to 5% expected by chance alone
  - – Average heterozygosity of 46% across three population groups: Causian, Black, Hispanic
  - – Despite lack of theoretically independent loci, his study supports the use of this 70 SNP panel for identification purposes

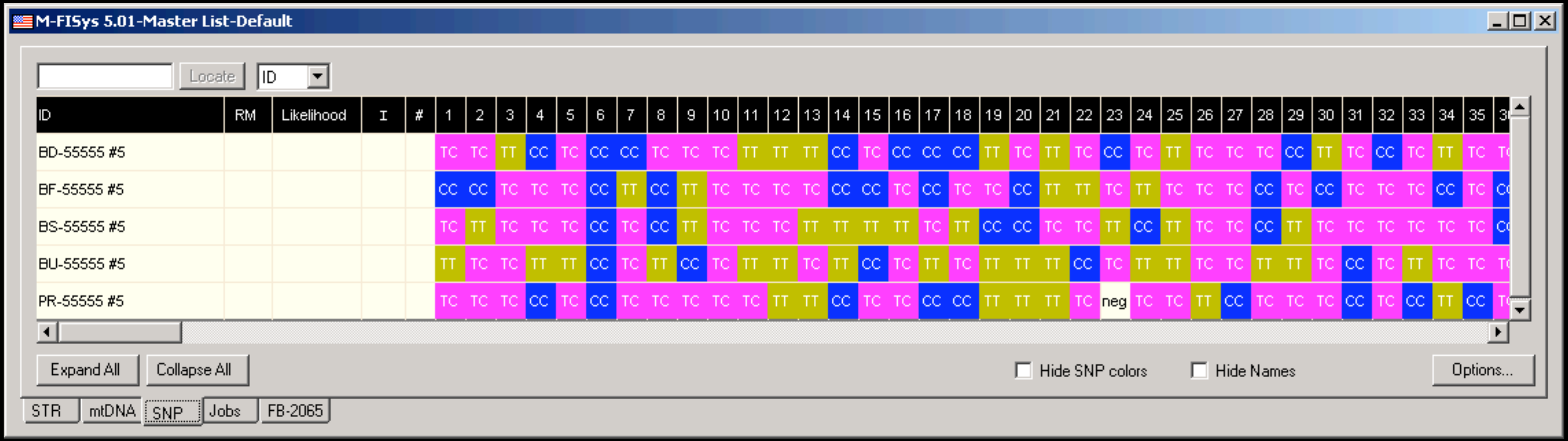
Non-Equiprobable SNP's

- Conservative likelihoods can be calculated even without the assumption of equi-probability.
- All bi-allelic heterozygous alleles have a minimum likelihood of 2, regardless of frequency:

$$f = 2pq = 2p(1-p) \leq 0.5 \ \forall p \in [0,1]; \ \therefore L = 1/f \geq 2$$

- The minimum likelihood of a SNP profile containing *n* heterozygous alleles is thus 2^n.
- As Forensic Mathematician Charles Brenner notes, even if the SNP frequencies were 0.1 and 0.9, 99% of cases will have 10 heterozygous loci out of 100.

M-FISys SNP Form



ID	RM	Likelihood	I	#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	3
BD-55555 #5					TC	TC	TT	CC	TC	CC	CC	TC	TC	TC	TT	TT	TT	CC	TC	CC	CC	CC	TT	TC	TT	TC	CC	TC	TT	TC	TC	TC	CC	TT	TC	CC	TC	TT	TC	TC
BF-55555 #5					CC	CC	TC	TC	TC	CC	TT	CC	TT	TC	TC	TC	TC	CC	CC	TC	CC	TC	TC	CC	TT	TT	TC	TT	TC	TC	TC	CC	TC	CC	TC	TC	TC	CC	TC	CC
BS-55555 #5					TC	TT	TC	TC	TC	CC	TC	CC	TT	TC	TC	TC	TT	TT	TT	TT	TC	TT	CC	CC	TC	TC	TT	CC	TT	TC	TC	CC	TT	TC	TC	TC	TC	TC	TC	CC
BU-55555 #5					TT	TC	TC	TT	TT	CC	TC	TT	CC	TC	TT	TT	TC	TT	CC	TC	TT	TC	TT	TT	TT	CC	TC	TT	TT	TC	TC	TT	TT	TC	CC	TC	TT	TC	TC	TC
PR-55555 #5					TC	TC	TC	CC	TC	CC	TC	TC	TC	TC	TC	TT	TT	CC	TC	TC	CC	CC	TT	TT	TT	TC	neg	TC	TC	TT	CC	TC	TC	TC	CC	TC	CC	TT	CC	TC

☐ Hide SNP colors

☐ Hide Names

Combining Technologies for Partial Profiles

- The *M-FISys* software package is designed for rapid cross-pollination of STR, Kinship, mtDNA and SNP data of DNA samples.
- Consistent or conflicting data in one technology can help determine experimental errors resulting in another technology.
- *M-FISys* also generates Quality Control reports for finding such inconsistencies.

Combining SNP's & STR's

- By selectively choosing SNP's which are unlinked to each other and existing STR loci, independent likelihoods can be multiplied.
- With the exception of CSF1PO & D5S818, all STR loci are on different chromosomes.
- Thus any unlinked SNP's on an unused chromosome can be included in likelihood calculations.
- STR profiles below threshold are missing ≥ 3 loci
- Even if only 10 SNP's are used, the likelihood can be increased by 3 orders of magnitude! ($2^{10} \approx 10^{+3}$)

More Information

Gene Codes Forensics
775 Technology Drive, Suite 100A
Ann Arbor, MI 48108
(734) 769-7249
http://www.genecodes.com

Updated Slides:

http://www.jonhoyle.com/GeneCodes

Thank You!