



Mixtures and Interpretation Software

Amy Decker
National Institute of Standards and Technology



Forensic Biology Workshop:
Advanced Forensic DNA Concepts
May 5, 2009



Outline

- Biostatistical calculations for mixed stains
- NIST involvement with DNA mixtures
- Description/features of GMID-X Mixture Analysis Tool
- GMID-X with Mix05 Case #1: 3:1 mixture
- DNA_DataAnalysis with Mix05 Case #1: 3:1 mixture
- Summary of AB's GMID-X and USACIL's DNA_DataAnalysis

Calculating Statistics for Mixtures

There are various statistical approaches that can be used for reporting mixture results:

- Probability of exclusion (PE)/Probability of inclusion (PI)
- Random Match Probability (RMP)
- Likelihood Ratio (LR)

Probability of Exclusion/Inclusion

Also known as the Combined Probability of Exclusion/Inclusion (CPE/CPI)

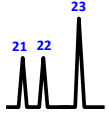
Prob. of Inclusion (PI) is the combined frequency of all combinations of genotypes that **CANNOT BE EXCLUDED** from contributing to the mixture

- Makes no assumptions about # of contributors
- aka random man not excluded (RMNE)

Prob. of Exclusion (PE) is the probability of **EXCLUDING** a randomly selected person

Probability of Exclusion/Inclusion

Suppose the following scenario:



From X population database

| Allele | Frequency |
|--------|-----------|
| 21 | 0.187 |
| 22 | 0.182 |
| 23 | 0.156 |

$$PI = (P_A + P_B + P_C)^2$$

$$= (0.187 + 0.182 + 0.156)^2$$

$$= (0.525)^2$$

$$= 0.276$$

Thus it is expected that 28% of a group of randomly selected persons will not be excluded as contributors or 1 out of 4 randomly selected persons (RMNE)

$$PE = 1 - PI = 1 - 0.276 = 0.724$$

Thus it is expected that 72% of a group of randomly selected persons will be excluded as contributors

Random Match Probability

Random Match Probability (RMP) is the probability of obtaining a match between two distinct and unrelated individuals

RMP is calculated by taking the inverse of the genotype frequency for a marker or a full profile

For example,

| Locus | Allele 1 | Allele 2 | Allele 1 Freq (p) | Allele 2 Freq (q) | Calc | Genotype Freq |
|---------|----------|----------|-------------------|-------------------|------|---------------|
| D13S317 | 11 | 12 | 0.3394 | 0.24834 | 2pq | 0.1686 |
| D16S539 | 12 | 11 | 0.32616 | 0.32219 | 2pq | 0.2095 |
| | | | Combined Freq | | 1 in | 0.0953217 |
| | | | | | | 28.311268 |

RMP = 1 in 28

Random Match Probability

What does this mean? **RMP = 1 in 28**

This is the theoretical chance that if one person is pulled at random from a population, they will have this particular profile. Obviously in this case, there are only 2 loci therefore the chance is relatively high.

-It does **NOT** mean the chance that someone else is guilty
-It is **NOT** the chance that the defendant is not guilty

| Locus | Allele 1 | Allele 2 | Allele 1 Freq (p) | Allele 2 Freq (q) | Calc | Genotype Freq |
|---------|----------|----------|-------------------|-------------------|------|------------------------|
| D1S5317 | 11 | 12 | 0.3394 | 0.24834 | 2pq | 0.1686 |
| D16S539 | 12 | 11 | 0.32616 | 0.32119 | 2pq | 0.2095 |
| | | | Combined Freq | | 1 in | 0.0353217 28.311208 |

Likelihood Ratio

Likelihood Ratio is based on defined hypotheses as to the origin of the mixture. This calculation compares the probabilities of the evidence as 2 alternatives.

- The **prosecution hypothesis** (H_p)
- The **defense hypothesis** (H_d)

Typically, the **prosecution's hypothesis** is that DNA profile generated from the crime scene originates from **the victim and the suspect**. The **defense's hypothesis** is that the evidence originates from **the victim and an unknown person**.

LR provides a numerical value that indicates how many more times likely the observed DNA profile originated from H_p than H_d

Likelihood Ratio

- Likelihood Ratio requires a description of the scenario
- Hypotheses must clearly state who contributed to the stain
- Hypotheses must state how many **unknown** contributors are assumed

This allows the evidential value of a stain to be calculated with reference to a specific person involved in the case (i.e. the accused stain donor)

Note: This information is from Schneider PM, Fimmers R, Keil W, Molsberger G, Patzelt D, Pflug W, Rothämel T, Schmitter H, Schneider H, Brinkmann B. [The German Stain Commission: recommendations for the interpretation of mixed stains](#), Int J Legal Med. 2009 Jan;123(1):1-5.

Likelihood Ratio

We must define the scenario:

Two contributors, unambiguous DNA profile

The hypothesis of the prosecution is that the victim and the defendant contributed to the mixture; $H_p = 1$ (or 100% probability)

However, the defense claims the victim and an **unknown person** contributed to the mixture.

Example

The victim's genotype is 21,23. The suspect's genotype is 22,23. The defense hypothesis **must explain the 22 allele** and would include the following possible combinations:
(21,22) (22,22) (22,23)

| Allele | Frequency |
|--------|-----------|
| 21 (a) | 0.187 |
| 22 (b) | 0.182 |
| 23 (c) | 0.156 |

$$LR = 1/(2ab + b^2 + 2bc)$$

$$LR = 1/[(2(0.187)(0.182) + 0.182^2 + 2(0.182)(0.156))]$$

$$LR = 6.33$$

Likelihood Ratio

LR = 6.33

The result can be described as follows:

It is **6.33** times more likely to observe the DNA profile **if** the mixed stain originated from the victim and the suspect than if it originated from the victim and an unknown person in **X** population.

Note: This information is from Schneider PM, Fimmers R, Keil W, Molsberger G, Patzelt D, Pflug W, Rothämel T, Schmitter H, Schneider H, Brinkmann B. [The German Stain Commission: recommendations for the interpretation of mixed stains](#), Int J Legal Med. 2009 Jan;123(1):1-5.

NIST involvement with DNA Mixtures

- NIST interlaboratory studies
- Mixture Case Summaries
- Mixture Data Sets
- Evaluate Software for Mixture Deconvolution

- AB's GMID-X Mixture Interpretation Tool
- USACIL's DNA_DataAnalysis
- i-STReam portion of FSS-i3 v4.1.3 (Promega Corporation)
- the Web-based Least Squares Deconvolution (Web-LSD)
- Genoproof Mixture (Qualitytype AG)

See poster from 19th International Symposium in Human Identification available from STRBase: http://www.cstl.nist.gov/biotech/strbase/pub_pres/Promega2008poster.pdf for more information

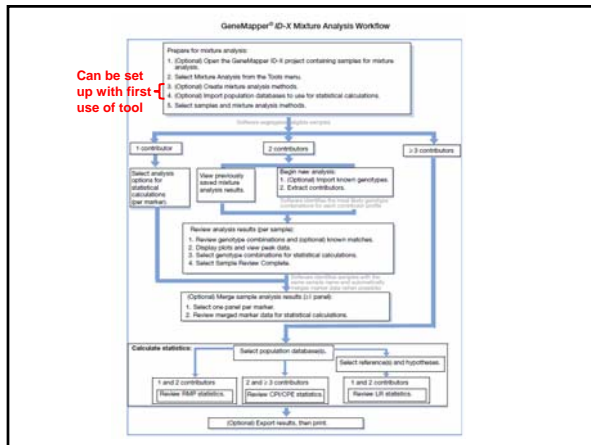
**NIST MIXTURE INTERPRETATION
INTERLABORATORY STUDY
MIX05**

- **Case #1:** Sample S- “Evidence” mixture, **3** parts female: **1** part male
Sample P- Female “victim”
- Case #2:** Sample B- “Evidence” mixture, **1** part female: **3** parts male
Sample A- Female “victim”
- Case #3:** Sample M- “Evidence” mixture, **1** part female: **1** part male
Sample K- Female “victim”
- Case #4:** Sample G- “Evidence” mixture, **7** part female: **1** part male
Sample F- Female “victim”

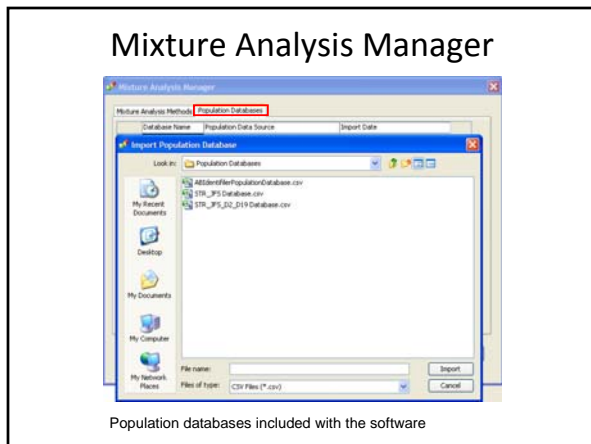
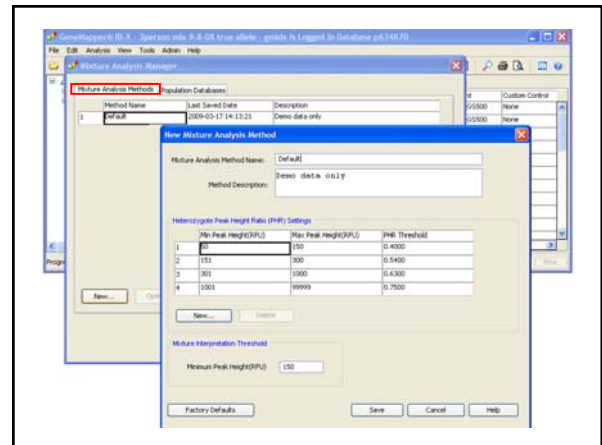
Participants were to summarize the perpetrator’s alleles in each case and provide appropriate statistic.

<http://www.cstl.nist.gov/biotech/strbase/interlab/MIX05/MIX05poster.pdf>

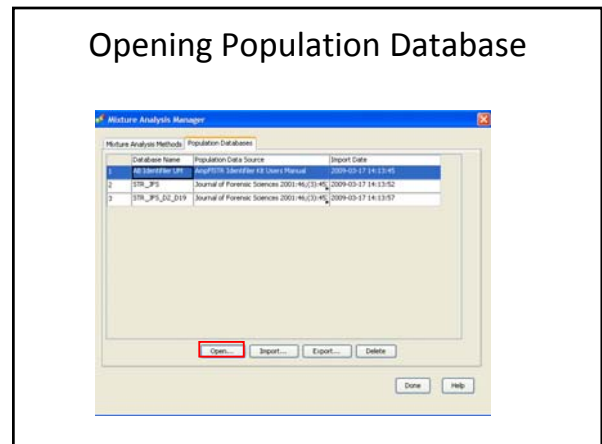
**MIX05 Data in
GeneMapper ID-X Version 1.1
Mixture Analysis Tool**



Can be set up with first use of tool

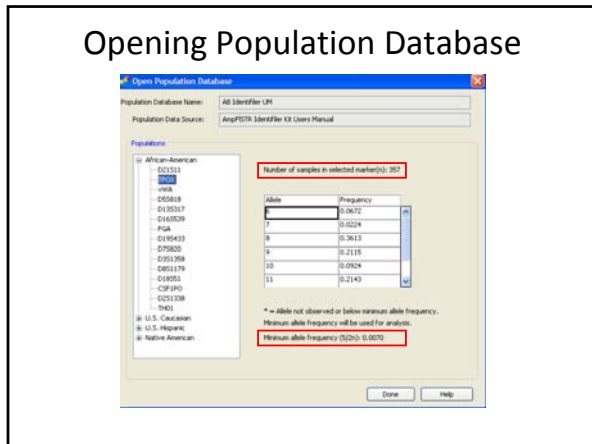


Population databases included with the software

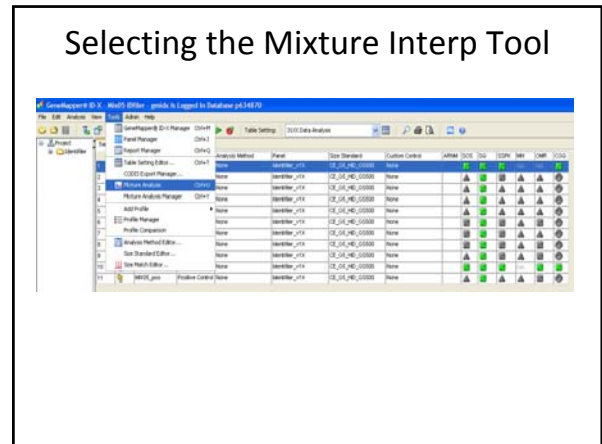


Opening Population Database

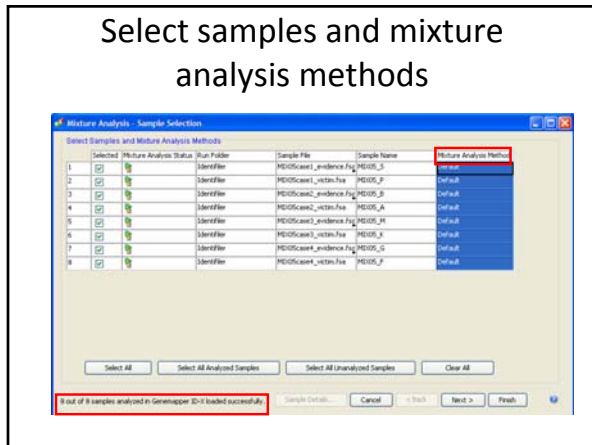
Opening Population Database



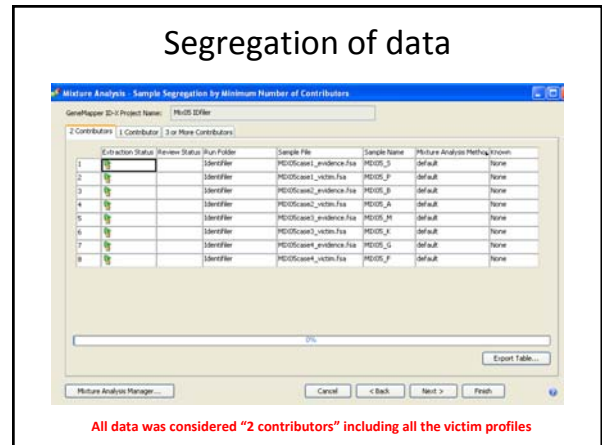
Selecting the Mixture Interp Tool



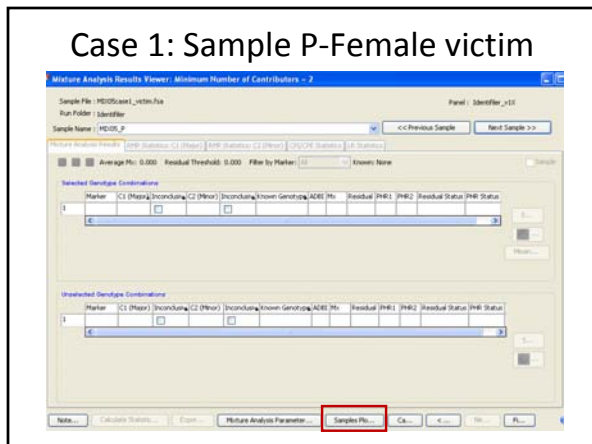
Select samples and mixture analysis methods



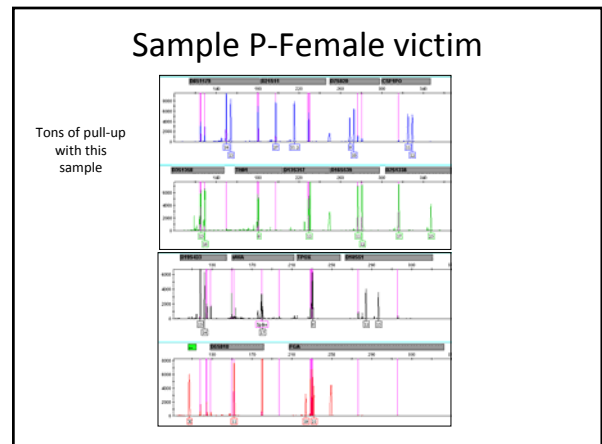
Segregation of data



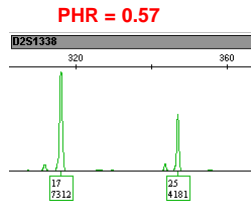
Case 1: Sample P-Female victim



Sample P-Female victim



Case 1: Sample P (single-source victim)



This PHR is below the set threshold of 0.75 and therefore the entire sample was placed under the 2 contributor tab

Sample Segregation Rules

- 1 Contributor (single source)
 - ALL markers must pass the PHR thresholds
 - AND
 - contain max of 1 marker with 3 alleles
- 2 Contributors
 - 1 or more 2 peak markers failing PHR thresholds
 - 3 or more alleles at 2 or more markers (max 4 alleles)
- 3 Contributors
 - 1 or more markers with more than 4 alleles

Analyzing Mix05 Case 1

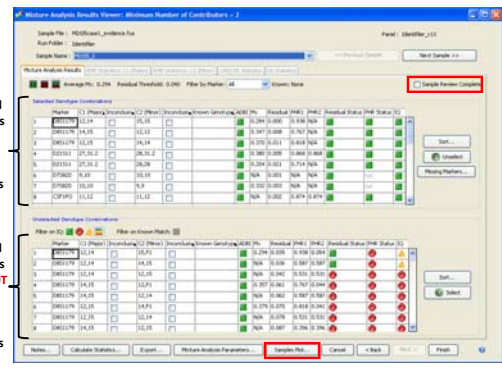
See handout

- “Evidence” is a 2 contributor mixture with 1 known contributor (“victim”)
- Must first review the results and mark the samples as reviewed before assigning known genotypes.
- Assuming the victim is present in the evidence (intimate sample)
- The known contributor will be “subtracted” from the selected genotype combinations table by the software

Mix05 Case 1: evidence_S- 1st pass

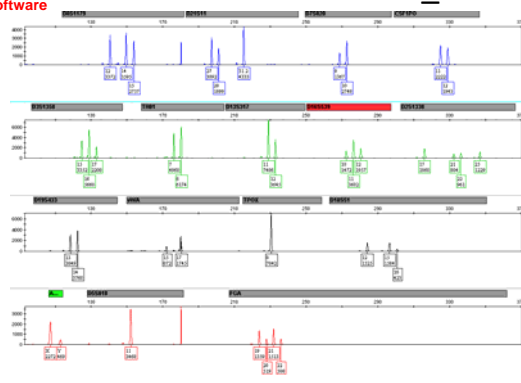
Includes all possibilities that meet all mixture analysis parameters

Includes all possibilities that do NOT meet all mixture analysis parameters



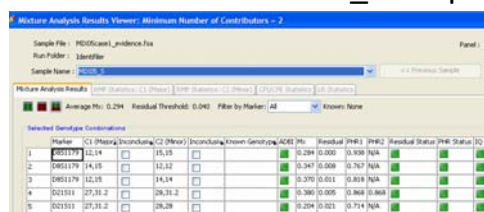
View in MA software

Mix05 Case 1: evidence_S

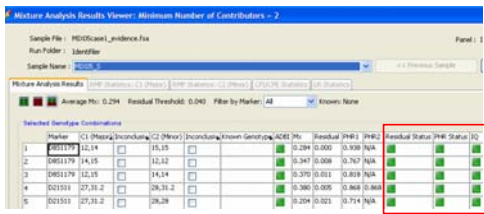


Mix05 Case 1: evidence_S- 1st pass

- ADBI- above detection, below interpretation
- Mx- mixture proportion
- PHR1 – Peak Height Ratio of C1 (Major)
- PHR2 – Peak Height Ratio of C2 (Minor)
- IQ-Inclusion Quality. When this is green, both the residual status and PHR status thresholds have been met.



Mix05 Case 1: evidence_S- 1st pass



- 1. Residual Status**
-above MIT (150 RFU)
-compares expected peak proportions (using average Mx) to observed peak proportions (using residual calc). It is a measure of how close the minor contributor proportion for a particular genotype combination is to the expected minor contributor proportion.
- 2. Peak Height Ratio Status**
-based on PHR thresholds set in Mixture Analysis method

Mixture Analysis Parameters

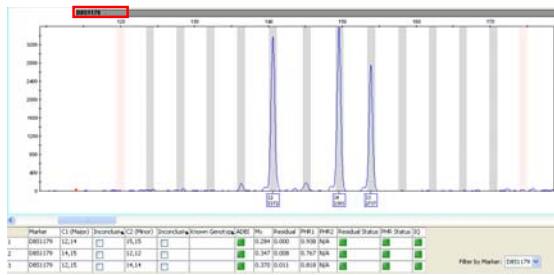
These parameters can be changed. The default method was used for this analysis which includes the following:

Mixture Analysis Method Name: default
Mixture Interpretation Threshold: 150

Heterozygote Peak Height Ratio (PHR) Settings:

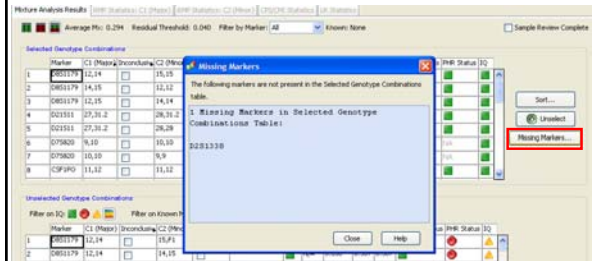
| Min Peak Height (RFU) | Max Peak Height (RFU) | PHR Threshold |
|-----------------------|-----------------------|---------------|
| 50 | 150 | 0.4 |
| 151 | 300 | 0.54 |
| 301 | 1000 | 0.63 |
| 1001 | 99999 | 0.75 |

Mix05 Case 1: evidence_S



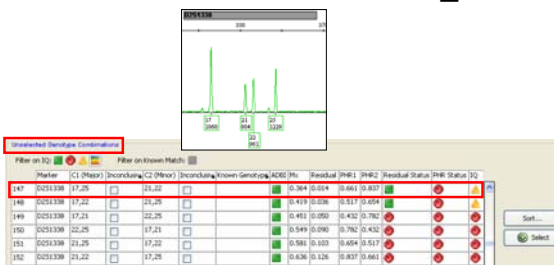
For **D8S1179**, there are 3 combinations listed in the "selected genotype combinations" table (above). Genotypes in this table are automatically included in the statistical calculations.

Mix05 Case 1: evidence_S



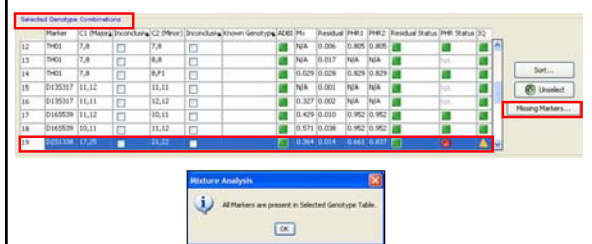
All markers are represented in the "selected genotype combinations" table **except D2S1338**.

Mix05 Case 1: evidence_S



All markers are represented in the "selected genotype combinations" table **except D2S1338**. This is because no combinations passed the PHR thresholds for D2. If a combination is determined to be included in the "selected" table, it can easily be moved. The combination boxed in red is the most likely combination (see notes) and this will be moved to the "selected" table.

Mix05 Case 1: evidence_S



Mix05 Case 1: evidence_S

Mix05 Case 1: evidence_S

At this point, the data can be exported as a .csv file. This file contains all the results in each of the tables and all the possible genotypes for the major and minor contributors based on the "selected genotypes" table. **OR**, the statistics can be calculated first and then all data can be exported.

Mix05 Case 1: evidence_S- 1st pass

Statistical Calculations

| | 1 Contributor (Single-source) | 2 Contributors | 3 or More Contributors |
|---|-------------------------------|----------------|------------------------|
| Random Match Probability (RMP) | X | X | |
| Combined Probability of Inclusion/Exclusion (CPI/CPE) | | X | X |
| Likelihood Ratio (LR) | X | X | |

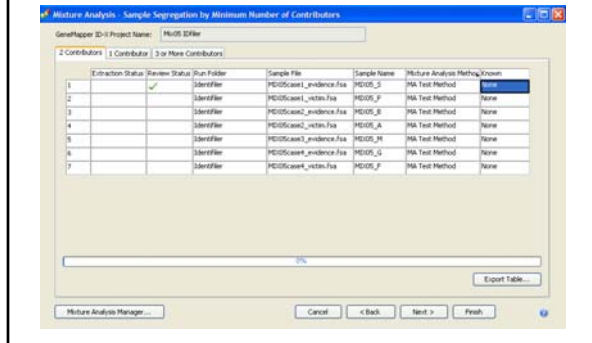
Statistics-Evidence without known applied -RMP (Major)

The RMP for the major and the CPI/CPE are calculated for the evidence, but do not include D2S1338.

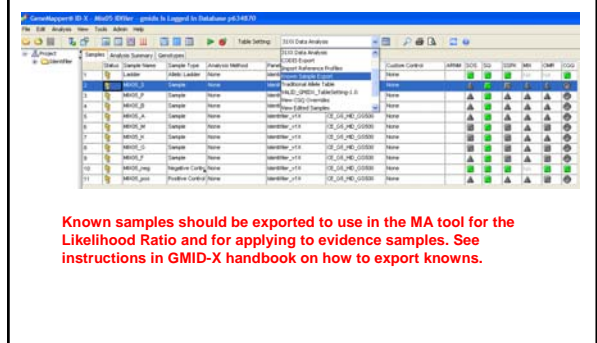
Case #1- LR calculation

In this case, we only know the victim's profile (sample P), which is applied as reference 1 (red box). Need to deconvolute Sample S to determine a profile for a "suspect"

Mix05 Case 1: evidence_S
Assign Known Genotype

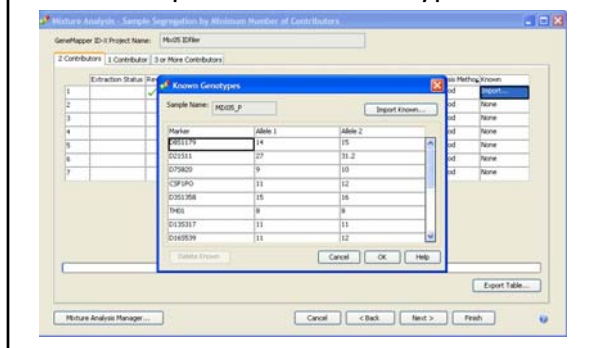


Mix05 Case 1: evidence_S
Known Sample Export

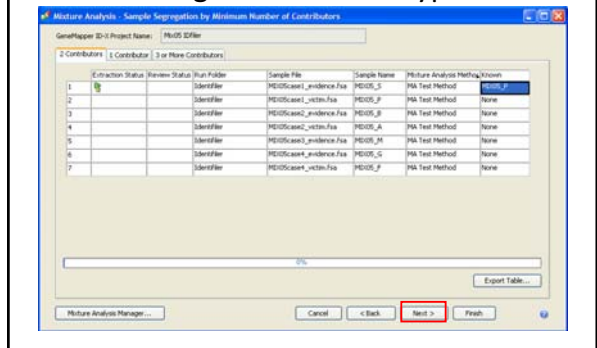


Known samples should be exported to use in the MA tool for the Likelihood Ratio and for applying to evidence samples. See instructions in GMID-X handbook on how to export knows.

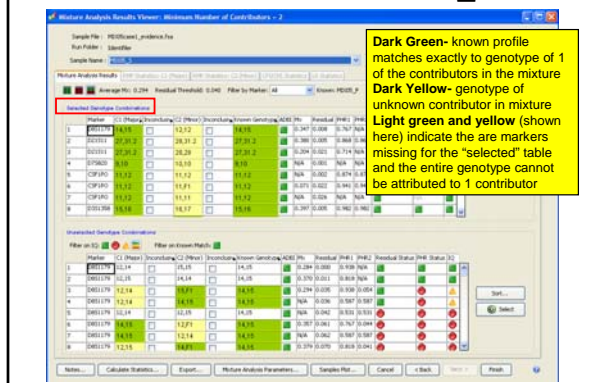
Mix05 Case 1: evidence_S
Import Known Genotype



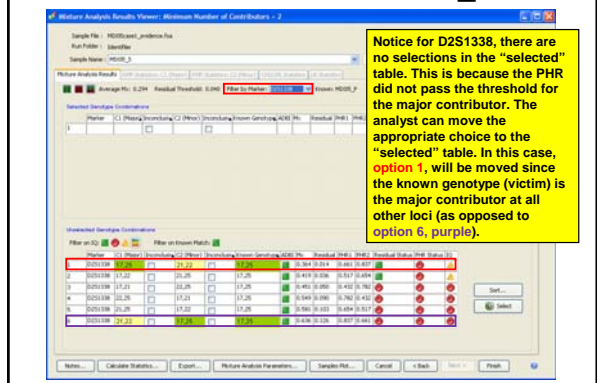
Mix05 Case 1: evidence_S
Assign Known Genotype



Mix05 Case 1: evidence_S



Mix05 Case 1: evidence_S



Mix05 Case 1: evidence_S

Because of D2S1338, no statistics applying the known are calculated. Unfortunately, even if D2 is not included in the evidence, the software still will **not** allow for statistics to be calculated.

MIX05 Data in DNA_DataAnalysis v2.1.3

DATA SET: 4 CASES WITH IDENTIFILER

Data Analysis:

- all OL alleles and stutter peaks removed, therefore all alleles included in data set are the “true” types
- Since all 4 cases are sexual assault cases, will assume the evidence is “intimate sample” and therefore will apply the victim

Mixture Interpretation Settings:

- mPH** (minimum peak height for interpretation) = 150 RFU
- PHr** (peak height ratio) = variable based on overall RFU values of the alleles at each locus but generally between 0.5 – 0.6
- mP** (minimum proportion) = 0

RULE: Determine proportion for all 4 allele loci (& 3 allele loci if possible). Deconvolute all other loci using the proportion range and high PHR.

Case data 1-4 in software (Analyst page)

Case data #1 – Evidence decon (victim applied) – D2S1338 ~ 2.9:1 victim: unknown

The purple box displays the mixture tool for D2S1338. The red box displays the deconvoluted profile after applying the victim's sample. The only possible combination for the 4 alleles with a phr threshold = 0.6 and mPH = 150 is 17,25 (victim) 21,22 (unknown contributor).

Comparing deconced contributor profile to “true” contributor

| Sample ID | D8S1179 | D21S11 | D7S820 | CSF1PO | D3S1358 | TH01 | D13S317 | D16S539 |
|---------------------|---------|----------|--------|--------|---------|------|---------|---------|
| Case1 Suspect decon | 12, 12 | 26, 31.2 | 10, 10 | 11, 12 | 16, 17 | 7, 7 | 12, Any | 10, 11 |
| True Contributor | 12, 12 | 26, 31.2 | 10, 10 | 11, 12 | 16, 17 | 7, 7 | 12, 12 | 10, 11 |

| Sample ID | D2S1338 | D19S433 | VWA | TPOX | D18S51 | AmeI | D5S818 | FGA |
|---------------------|---------|---------|---------|------|---------|------|--------|--------|
| Case1 Suspect decon | 21, 22 | 13, 14 | 15, Any | 8, 8 | 16, Any | X, Y | 11, 11 | 20, 22 |
| True Contributor | 21, 22 | 13, 14 | 15, 17 | 8, 8 | 15, 16 | X, Y | 11, 11 | 20, 22 |

The 3 loci boxed in red (D13, VWA & D18) were the loci that had more than one possible combination that were too close to call. In each case, an obligate allele was determined but the “Any” refers to any other possible allele for that locus.

For the 4 loci boxed in blue, the answer was chosen based on the p-values closest to the average (.65 major, .35 minor) as well as the highest phr. All answers matched the true contributor.

1 Case data #1 – Evidence decon (victim applied) – D13S317 ~2.9:1 victim: unknown

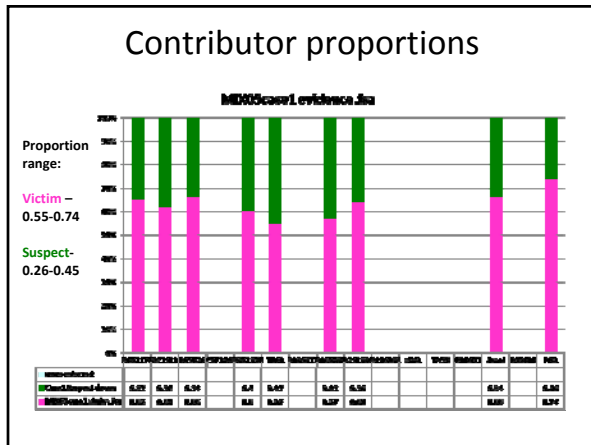
Red box displays the deconvoluted profile after applying the victim's sample. The purple box displays the mixture tool for D13S317. There are 2 possible combinations for the 2 alleles that meet the thresholds. Both combos boxed in blue have very similar p-values.

| | | | | | |
|--------|--------|------|------|------|------|
| Ref1 | Ref2 | Ref3 | Ref4 | Ref5 | Ref6 |
| 11, 11 | 11, 11 | | | | |
| 11, 11 | 12, 12 | | | | |
| 11, 11 | 11, 12 | | | | |

2 Case data #1 – Evidence decon (victim applied)-D21S11 ~2.9:1 victim: unknown

Red box displays the deconvoluted profile after applying the victim's sample. The purple box displays the mixture tool for D21S11. There are 2 possible combinations for the 3 alleles that meet the thresholds. The combo boxed in blue is closest to the average p-value of 0.65 and has a higher pfr.

| | | | | | |
|----------|----------|----------|------|------|------|
| Ref1 | Ref2 | Ref3 | Ref4 | Ref5 | Ref6 |
| 27, 31.2 | 27, 31.2 | | | | |
| 27, 31.2 | 27, 31.2 | 28, 28 | | | |
| 27, 31.2 | 27, 31.2 | 27, 31.2 | | | |



Case #1 Stats- CPE/CPI

| Group | Allele | Frequency | Scientific notation | In words |
|-------|----------|-----------|---------------------|----------|
| Cauc | 1 in ... | 4.18e+08 | 4.18 billion | |
| Black | 1 in ... | 1.205e+08 | 120 million | |
| Hisp | 1 in ... | 3.634e+07 | 36.3 million | |

Combined Probability of Inclusion from entire profile of Case #1 using the same values found in PopStats from the FBI population database for Caucasians, African Americans and Hispanics.

Case #1 Stats- LR

| Group | Allele | Frequency | Scientific notation | In words |
|-------|----------|------------|---------------------|----------|
| Cauc | 1 in ... | 8.5708E+11 | 857 billion | |
| Black | 1 in ... | 3.7063E+08 | 370 million | |
| Hisp | 1 in ... | 2.9653E+12 | 2.96 trillion | |

Likelihood Ratio includes all obligate alleles for each locus. Those loci NOT included in the calculation (no obligate alleles) are D7, CSF, D19, TPOX, Amel & D5. Allele frequencies used in the calculation are found in PopStats from the FBI population database for Caucasians, African Americans and Hispanics.

Summary of Mixture Interp. Tool

PROS:

- Seamless transition of data (no export tables needed like in USACIL or FSS3)
- Automatic separation into 1, 2 or 3 or more contributor tabs. The only samples to deconvolute are the 2 person mixtures.
- Stats calculated by the software including Random Match Probability (RMP), Combined Probability of Inclusion/Exclusion (CPI/CPE) and Likelihood Ratio (LR) using population databases stored in the Mixture Analysis Manager. Results are easy to export. (LR calc simple to use!)
- Table of "most likely" possibilities (helps prioritize possible genotype combinations)

NOTE: Only analyzed samples with green or yellow sizing quality (SQ) flags and no off ladder (OL) labels are eligible for mixture analysis.

Summary of Mixture Interp. Tool

CONS:

- Data from GMID v3.2 cannot be run in Mixture Interp Tool. **Must REANALYZE samples in GMID-X**
- Mixture samples are **automatically separated into 1,2 or 3** or more contributor tabs. In data sets evaluated so far, some single source samples have been incorrectly placed under the “2 contributor” tab
- When calculating LR, **single source reference samples** must be **exported and saved individually** for import into LR calculation. When applying a known profile to evidence sample, **ALL LOCI** in reference sample must match evidence in order to perform statistics
- Can **NOT** handle Y-STR mixtures

Summary of DNA_DataAnalysis

PROS:

- All possible genotype combinations with **PHR and proportion** calc
- Graph of contributor percentages** for deconvoluted loci
- Performs statistics including PE/PI, LR

CONS:

- Putting data in correct format** for import into software
- Saving and exporting.** Only certain pages can be saved and reopened in the software. Some pages default to save incorrectly and must be changed manually.

In Summary...

- Numerous methods and software programs for solving mixtures.
- GMID-X provides **seamless transition**
 - Faster
 - No typing or calculation error
- COST/BENEFIT**
 - DNA_DataAnalysis & other software and programs- MUCH CHEAPER upfront costs BUT require more user interface
 - Uniformity of software programs
- FUTURE WORK**
 - Analyze more samples with other data sets
 - Compare statistical results with various software

Thanks To:



John Butler
Group Leader



Jan Redman



Margaret Kline



Michelle Burns
(summer 2008+)



Angie Dolph
Spessard
(summer 2007)

Collaborators

- Timothy Kalafut (USACIL)
- Ann Gross (MN BCA)
- Todd Bille (ATF)
- Amy Christian and Rhonda Roby (NEST Project Team)

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