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

### Example Calculation

Suppose the following scenario:

<table>
<thead>
<tr>
<th>Allele</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>0.187</td>
</tr>
<tr>
<td>22</td>
<td>0.182</td>
</tr>
<tr>
<td>23</td>
<td>0.156</td>
</tr>
</tbody>
</table>

**PI** = \((P_A + P_B + P_C)^2\)

\[
P_I = (0.187 + 0.182 + 0.156)^2 = (0.525)^2 = 0.276
\]

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.

**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:

<table>
<thead>
<tr>
<th>Locus</th>
<th>Allele 1</th>
<th>Allele 2</th>
<th>Allele 1 Freq</th>
<th>Allele 2 Freq</th>
<th>Genotype Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>D8S1172</td>
<td>14</td>
<td>12</td>
<td>0.0489</td>
<td>0.0489</td>
<td>0.1988</td>
</tr>
<tr>
<td>D8S1173</td>
<td>14</td>
<td>12</td>
<td>0.0242</td>
<td>0.0242</td>
<td>0.0578</td>
</tr>
<tr>
<td>D8S1173</td>
<td>12</td>
<td>12</td>
<td>0.0210</td>
<td>0.0210</td>
<td>0.0421</td>
</tr>
</tbody>
</table>

**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

Mixture & Interpretation Software – A. Decker

Likelihood Ratio

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

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 H1 than H0

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) (21,23) (22,23)

LR = \( \frac{1}{2(0.182)(0.187)} + 0.182^2 + 0.187^2 \)

LR = 6.33


Likelihood Ratio

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.


Likelihood Ratio

NIST involvement with DNA Mixtures

- NIST interlaboratory studies
- Mixture Case Summaries
- Mixture Data Sets
- Evaluate Software for Mixture Deconvolution
  - AL’s GMDM-X Mixture Interpretation Tool
  - USACIL’s DNA DataAnalysis
  - i-Stream portion of FSS-3 v4.1.3 (Promega Corporation)
  - the Web-based Least Squares Deconvolution (Web-LSD)
  - Genoproof Mixture (Qualigene AG)

See poster from 19th International Symposium in Human Identification available from STRBase:
http://www.cstl.nist.gov/biotech/strbase/pub_pres/promegaDNAposter.pdf for more information
Mixtures & Interpretation Software – A. Decker

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

Mixture Analysis Manager

Opening Population Database

Can be set up with first use of tool

Population databases included with the software
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)

PHR = 0.57

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
  - 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

• “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

Mix05 Case 1: evidence_S

View in MA software

ADBI - above detection, below interpretation
Mix - mixture proportion
PHR 1 - 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.
Mixtures & Interpretation Software – A. Decker

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:

<table>
<thead>
<tr>
<th>Min Peak Height (RFU)</th>
<th>Max Peak Height (RFU)</th>
<th>PHR Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>150</td>
<td>0.4</td>
</tr>
<tr>
<td>151</td>
<td>300</td>
<td>0.54</td>
</tr>
<tr>
<td>301</td>
<td>1000</td>
<td>0.63</td>
</tr>
<tr>
<td>1001</td>
<td>99999</td>
<td>0.75</td>
</tr>
</tbody>
</table>

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
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.

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

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”
Mixtures & Interpretation Software – A. Decker

Mix05 Case 1: evidence_S
Assign Known Genotype

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 knowns.

Mix05 Case 1: evidence_S
Import Known Genotype

Mix05 Case 1: evidence_S
Assign Known Genotype

Mix05 Case 1: evidence_S

Dark Green - known profile matches exactly to genotype of 1 of the contributors in the mixture
Dark Yellow - genotype of unknown contributor in mixture
Light green and yellow (shown here) indicate the are markers missing for the “selected” table and the entire genotype cannot be attributed to 1 contributor.

Notice for D2S1338, there are no selections in the “selected” table. This is because the PHR did not pass the threshold for the major contributor. The analyst can move the appropriate choice to the “selected” table. In this case, option 1, will be moved since the known genotype (victim) is the major contributor at all other loci (as opposed to option 6, purple).
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.

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.

Comparing deconed contributor profile to “true” contributor

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 (.45 major, .35 minor) as well as the highest PHr. All answers matched the true contributor.
Case data #1 – Evidence decon (victim applied) – D13S317

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.

Contributor proportions

Proportion range:
Victim = 0.55-0.74
Suspect = 0.26-0.45

Case data #1 – Evidence decon (victim applied)-D21S11

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 phr.

Case #1 Stats- CPE/CPI

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

Likelihood Ratio includes all obligates alleles for each locus. Those loci NOT included in the calculation (no obligate alleles) are D7, CSP, D19, TPOX, Amel & D8. 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 FSSi3)
- 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.
Mixtures & Interpretation Software – A. Decker

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 Spessard (summer 2007)

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

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