Exploring DNA Interpretation Software Using
the PROVEDIt Dataset

Sarah Riman¹; Hari Iyer²; Peter M. Vallone¹

¹Applied Genetics Group, National Institute of Standards and Technology
²Statistical Design, Analysis, and Modeling Group, National Institute of Standards and Technology
Overview of PROVEDIt database

- Large publicly available database
- Contains over 25,000 STR profiles
- 1 to 5 person mixtures varying
  - contributor ratios
  - DNA quality
- Amplified with different STR kits
  - DNA quantity (0.007-1 ng)
- Analyzed with different CE instrument types and injection times
- Allows examination of probabilistic genotyping systems
- Examine effect of analytical thresholds and peak detection parameters on downstream analysis
- Assess approaches to evaluate STR signal (genotyping software packages and validation software)


https://lftdi.camden.rutgers.edu/provedit/
Objective of this study

Interpretation parameters used for each software

<table>
<thead>
<tr>
<th>Software</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>STRmix v2.6</td>
<td>- ( N-1, N-2 ) and ( N+1 ) stutter peaks were modeled</td>
</tr>
<tr>
<td></td>
<td>- <strong>Drop-in frequency</strong> = 0.0015 and maximum cap = 180 RFU</td>
</tr>
<tr>
<td></td>
<td>- <strong>Saturation threshold</strong> = 30,000 RFU</td>
</tr>
<tr>
<td></td>
<td>- <strong>MCMC settings</strong>: 8 chains of 100,000 burn-in accepts, 50,000 post burn-in accepts per chain</td>
</tr>
<tr>
<td></td>
<td>- Allelic, stutter, and locus-specific amplification efficiency variance were determined using <em>Model Maker</em> where over 300 single source profiles of varying quality and quantity were assessed</td>
</tr>
<tr>
<td></td>
<td>- The sub-source LR is reported</td>
</tr>
<tr>
<td>EuroForMix v2.1.0</td>
<td>- <strong>MLE</strong> (Maximum likelihood estimation) approach</td>
</tr>
<tr>
<td></td>
<td>- <strong>Degradation and stutter models</strong> jointly turned on</td>
</tr>
<tr>
<td></td>
<td>- Default parameters, except for a 35 RFU detection threshold, ( Pr(C) = 0.0015 ) and ( \lambda = 0.018 ).</td>
</tr>
<tr>
<td></td>
<td>- The MLE based method LR is reported</td>
</tr>
<tr>
<td>Both software</td>
<td>- Profiles were analyzed using the per dye ATs</td>
</tr>
<tr>
<td></td>
<td>- <strong>NIST 1036-Caucasian</strong> allele frequencies</td>
</tr>
<tr>
<td></td>
<td>- ( \theta ) correction was applied using an ( F_{st}(\theta) = 0.01 )</td>
</tr>
<tr>
<td></td>
<td>- True NOC and same propositions were used in both software</td>
</tr>
</tbody>
</table>


O. Bleka et al., EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic science international. Genetics 21 (2016) 35-44.
Dataset used in our study

<table>
<thead>
<tr>
<th>Kit (PCR cycle no.)</th>
<th>CE instrument (injection time)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GlobalFiler (29 cycles)</strong></td>
<td><strong>3500 (15 s)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of contributors</th>
<th>Mixture ratios</th>
<th>Pristine DNA</th>
<th>Degraded DNase I</th>
<th>Degraded Sonication</th>
<th>Damaged UV</th>
<th>Inhibited Humic Acid</th>
<th>Minor Contributor DNA amount (pg)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2P</strong> (16 unique individuals)</td>
<td>1:1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>15; 30; 62; 125</td>
</tr>
<tr>
<td></td>
<td>1:2</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>15; 30; 62; 125</td>
</tr>
<tr>
<td></td>
<td>1:4</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>15; 30; 62; 125</td>
</tr>
<tr>
<td></td>
<td>1:9</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>15; 30; 54; 62; 75</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td>88</td>
<td>228</td>
<td>44</td>
<td>104</td>
<td>108</td>
<td>572</td>
<td></td>
</tr>
</tbody>
</table>

| **3P** (21 unique individuals) | 1:1:1 | x | x | x | x | x | 15; 30; 62; 125 |
| | 1:2:1 | x | x |   | x |   | 15; 30; 62; 125 |
| | 1:2:2 | x | x |   | x |   | 15; 30; 62; 125 |
| | 1:4:1 | x | x | x | x | x | 15; 30; 62; 125 |
| | 1:4:4 | x | x | x | x | x | 15; 30; 62; 83 |
| | 1:9:1 | x | x |   | x |   | 15; 30; 45; 62 |
| | 1:9:9 | x | x |   | x |   | 15; 26; 30; 40 |
| **Sum** | 114 | 324 | 72 | 138 | 162 | 810 |
### Dataset used in our study

<table>
<thead>
<tr>
<th>Kit (PCR cycle no.)</th>
<th>CE instrument (injection time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GlobalFiler (29 cycles)</td>
<td>3500 (15 s)</td>
</tr>
</tbody>
</table>

#### Equation:

**2P**

\[ H_p = \text{POI} + U_1 \]
\[ H_d = U_1 + U_2 \]

**3P**

\[ H_p = \text{POI} + U_1 + U_2 \]
\[ H_d = U_1 + U_2 + U_3 \]

**POI**

- True contributors
- True non-contributors 🔄️ (random person from NIST 1036)
Analysis of 2P and 3P mixtures
\( \log_{10}(LR) \) for 2P and 3P mixtures

Log_{10}(LR) Distribution by Software, NOC, & Propositions

- **Hp true STRmix**
- **Hd true STRmix**
- **Hp true EFM**
- **Hd true EFM**

2 Person 3 Person
Receiver operating characteristic (ROC) plots

ROC Plots for 2 & 3 Person Mixtures (STRmix and EFM)

<table>
<thead>
<tr>
<th>Comparison Group</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2P (STRmix vs EFM)</td>
<td>0.74206</td>
</tr>
<tr>
<td>3P (STRmix vs EFM)</td>
<td>0.64155</td>
</tr>
<tr>
<td>STRmix (2P vs 3P)</td>
<td>0.02346</td>
</tr>
<tr>
<td>EFM (2P vs 3P)</td>
<td>0.04607</td>
</tr>
</tbody>
</table>
Log_{10}(LR) Distribution from 2P by software, contributor ratios and DNA treatments
Log_{10}(LR) Distribution by Software & Mixture Ratios

Log_{10}(LR) Distribution for 2P by Software & Proposition

Log_{10}(LR) Distribution by Software & Mixture Ratios (2P)

<table>
<thead>
<tr>
<th>Software &amp; Mixture Ratios</th>
<th>STRmix</th>
<th>EFM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\textbf{Log}_{10}(LR) Distribution by Software & Treatment}

\textbf{Log}_{10}(LR) Distribution for 2P by Software & Proposition

\textbf{Log}_{10}(LR) Distribution by Software & Treatment (2P)

- Pristine
- DNase I
- UV
- Sonication
- Inhibition

\textbf{Proposition}

- Hp true STRmix
- Hd true STRmix
- Hp true EFM
- Hd true EFM
Log$_{10}(LR)$ Distribution from 3P by software, contributor ratios and DNA treatments
Log$_{10}(LR)$ Distribution by Software & Mixture Ratios

Log$_{10}(LR)$ Distribution for 3P by Software & Proposition

1;1;1
1;2;1
1;2;2
1;4;1
1;4;4
1;9;1
1;9;9

H$^p$ true STRmix  H$^d$ true STRmix  H$^p$ true EFM  H$^d$ true EFM

STRmix  EFM

STRmix  EFM

STRmix  EFM

STRmix  EFM
Global overall profile $\text{Log}_{10}(LR)$ from each software

Note for the following plots: $\text{Log}_{10}(LR)$ from each software is shown ‘as is’ without further designation of the sample type, ratio, treatment, or software run diagnostics (this will be addressed in future work)
Global profile $\log_{10}(LR)$ from 2P and 3P

Software-A vs Software-B

- Hp true
- Hd true

2P

- Factor of 10\(^2\)
- Factor of 10\(^4\)
- Factor of 10\(^6\)

Software-A Log$\log_{10}(LR)$

Software-A Log$\log_{10}(LR)$

Software-A vs Software-B

3P

- Hp true

Software-B Log$\log_{10}(LR)$

Software-B Log$\log_{10}(LR)$

Note: $\log_{10}(LR)$ for Hd true tests with values of $-\infty$ from either software are not shown in these graphs
Global profile $\log_{10}(LR)$ from 2P and 3P for $Hp$ true

**Software-A vs Software-B**

- Factor of $10^2$
- Factor of $10^4$
- Factor of $10^6$

**Software-A vs Software-B**

- Factor of $10^2$
- Factor of $10^4$
- Factor of $10^6$
Conclusions

- The publicly available PROVEDIt database is a useful resource to understand probabilistic genotyping software.

- The effects of software (STRmix and EuroForMix), NOC, mixture ratios, and DNA treatments on LR assessment were examined.

- As expected, both software showed high degree of discrimination between Hp TRUE and Hd TRUE distributions across different ratios and treatments for 2 and 3 contributor samples.

- When it came to sample to sample profile comparisons the degree of agreement between the two software varied.
Future work

- Further investigation is needed to understand the source(s) behind the LR differences (e.g. MCMC settings, diagnostics, number of iterations, stutter models on/off, seed number)
- Analyze additional samples at different mixture ratios, treatments, and DNA amounts
- Explore the 4P mixtures
- Study deconvolution analysis of major and minor contributors in both software
- Examine the reported LR values at a per-locus level
Points of view in this presentation are mine and do not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Commerce.

NIST Disclaimer Certain commercial products and instruments are identified in order to specify experimental procedures as completely as possible. In no case does such an identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of these products are necessarily the best available for the purpose.
Acknowledgements

- Pete Vallone (NIST)
- Hari Iyer (NIST)
- Øyvind Bleka (EFM author)
- Zane Kerr (STRmix)
- Judi Morawitz (STRMix)
- Corina Benschop (Netherlands Forensic Institute)

Contact: sarah.riman@nist.gov