Sequencing STRs in forensic mixtures:
Current perspective on the benefits and challenges

Green Mountain DNA Conference
Burlington, VT
August 2, 2016

Katherine Gettings PhD
Applied Genetics Group

Dr. Pete Vallone
Lisa Borsuk
Kevin Kiesler
Becky (Hill) Steffen

Disclaimer: Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

Information presented does not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Justice.

STR Sequence Applications

- One to one matching?
  - new core loci = higher CE stats
  - Partial profiles
- Kinship
- Degraded samples?
  - Smaller PCR amplicons relative to CE
- Mixtures
  - Resolve alleles identical by length, but differ by sequence
  - Separate stutter from low level contributors (based on sequence)

Sampling of forensic NGS kits

- Illumina ForenSeq (FGx)
  - 27 aSTR + 24 YSTR + 7 XSTR + Amel
  - 94 HID-SNPs + 56 ancestry SNPs + 22 pheno SNPs
  - Multiplex A – all markers
- Promega PowerSeq kits (Auto, Y, Mito)
  - 22 aSTR + DYS391 + Amel
  - 22 aSTR + 21 YSTR
- Thermo Fisher Precision ID (STR, SNP, Mixtures & Mito)
  - Mixture panel: STR, MN, TNAP
  - 29 aSTR + DYS391 + Amel
  - 128 Identity SNPs
  - 185 Ancestry SNPs
  - Mito control region or whole genome
STR Sequence Gains
- PowerSeq Auto (MiSeq)
- 183 samples (3 US pops)
- STRait Razor (& ExactID)

- Currently working sequencing NIST population samples (>1000)
- Sequencing nearly complete
- Bioinformatic checks
- CE data backfilling
- The goal is to provide the allele frequencies of the sequenced STRs

Recognition Site-Based Informatics for STRs

- Recognition site (≈10 nt)
- PCR primers
- STR repeat region

Strait Razor:
- A length-based forensic STR allele-calling tool for use with second generation sequencing data.
- 2 Strait Razor v2.0: the improved STR Allele Identification Tool—Razor.

- Additional trimming for sequences with known flanking sequence present
- Additional processing for loci which require specific handling—special rules and exceptions
Data analysis

- FASTQ files were analyzed by STRaitRazor
- STRaitRazor data was parsed with custom scripts
- Allele calling: majority coverage and/or het balance (0.4)
  - Data compared to CE calls; discordant allele calls addressed
- Allele sequences were counted for each locus

Forensic STR Sequence Diversity

- Additional alleles observed by sequencing
- STR sequences provided in paper

**DSS818 SEQUENCE HAPLOTYPES**

- rs514684105 (98 bp) (A/C)
- rs37708 (33 bp) (C/T)
- rs41272001 (52 bp) (G/A)
- G/T (4 bp) AGAG to AGAT (G/T)
TPOX: rs13422969 and two additional rare SNPs

• Well distributed across Africa, rare elsewhere

Flanking region SNP categories from paper

1. Multiple polymorphisms in haplotype (D5S818 example)
2. “Old” single polymorphisms
3. Population or allele specific polymorphisms (TPOX example)
4. Polymorphisms associated with STR sequence variants
5. Rare polymorphisms (<5%)
6. No polymorphisms

Associated with “9” allele
Allele Distribution by Length

Loci which gain the most from sequencing the repeat region

Allele Distribution by Length

Loci which gain the most from sequencing the repeat region & flanking region (PowerSeq)

D12S391 stutter by sequence

([AGAT]_{n} [AGAC]_{n})

[n-4]_{seq} = 12%

[n-8]_{seq} = 1%

([AGAT] [AGAC] [AGAT])

9% = [AGAT] [AGAC] [AGAT]

5% = [AGAT] [AGAC] [AGAT]

2% = [AGAT] [AGAC] [AGAT]

D12S391 stutter by sequence [AGAT]_{n} [AGAC]_{n}

([AGAT] [AGAC] [AGAT])

[n-4]_{seq} = 22%

[n-8]_{seq} = 5%

([AGAT] [AGAC] [AGAT])

16% = [AGAT] [AGAC] [AGAT]

6% = [AGAT] [AGAC] [AGAT]

4% = [AGAT] [AGAC] [AGAT]

1% = [AGAT] [AGAC] [AGAT]
D12S391 stutter by sequence

AGAT\(n\) trends "3% higher stutter than AGAC\(n\)"

\[ y = 0.014x - 0.0401 \]
\[ R^2 = 0.7526 \]

D18S51 n-4 Stutter by Kit

\[ y = 0.0121x - 0.0416 \]
\[ R^2 = 0.8211 \]
D18S51 n-4 Stutter by Kit

Stutter Rates by Kit at 8 Simple Repeat Loci

NGS Mixture Study

Are mixture ratios by NGS the same as mixture ratios by CE?

<table>
<thead>
<tr>
<th>Loci</th>
<th>CE</th>
<th>NGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PowerPlex Fusion</td>
<td>PowerSeq Fusion</td>
<td>PowerSeq, Auto + Y</td>
</tr>
<tr>
<td>Input DNA</td>
<td>2.5 ng each</td>
<td>2.5 ng total</td>
</tr>
<tr>
<td>Amp Parameters</td>
<td>30 cycles</td>
<td>30 cycles, same as PPF</td>
</tr>
<tr>
<td>Everything Else</td>
<td>TruSeq PCR Free Library Prep, MiSeq v3</td>
<td></td>
</tr>
</tbody>
</table>

NGS similar to CE

NGS >5% higher than CE

Analysis of NIST population samples (N=192), Average + 3 Standard Deviations


Parameters: 100 ng template DNA, PowerSeq Fusion, PPFusion, GlobalFiler, 30 cycles, same as PPFusion, TruSeq PCR Free Library Prep, MiSeq v3
NGS Mixture Study

Ideally, each locus contributes $\frac{1}{3}$ or 4.35% of the total. This shows the difference from that expected 4.35% contribution.

Are mixture ratios by NGS the same as mixture ratios by CE?

- **CE**: no alleles below 75 RFU
- **NGS**: no alleles below 75x coverage
- Average of 3 replicates
Expected vs Observed Mixture Contributions

NGS Mixture Study 1:9:9

D2S441  1:9:9

1 additional allele
**D8S1179  1:9:9**

1 additional allele

1. Additional allele helps with stutter

**D3S1358  1:9:9**

1. Additional allele

16 allele not distinguishable from 17 stutter

**D12S391  1:9:9**

1 additional allele

2 alleles help with stutter

**VWA  1:9:9**

1 allele

Help with stutter
NGS Implications for Mixtures

Conclusions

- Sequencing forensic STR loci can uncover underlying sequence variation in the repeat and flanking regions
- This will increase allelic diversity, thus increasing the ability to discriminate among individuals in a mixture
- Additionally, sequence specific stutter ratios may improve mixture models

NGS Implications for Mixtures

Conclusions

Prior to implementation:

- Sequence-based allele frequency databases
- Characterization of any NGS-specific effects on
  - peak height ratios
  - minor components/stutter
- Probabilistic genotyping software amenable to sequence data (and sequence-based stutter?)

Summary 1:9:9

<table>
<thead>
<tr>
<th>Locus</th>
<th>Additional Alleles</th>
<th>Help with Stutter</th>
</tr>
</thead>
<tbody>
<tr>
<td>vWA</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>D1S1656</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>D2S441</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>D8S1179</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D3S1358</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D12S391</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

NGS profile contains four additional alleles and improved stutter attribution for four alleles
For calculations we need allele frequencies.

Acknowledgments

• Rachel Aponte (GWU – now at Bode)

• Funding
  • NIST - Forensic DNA
  • FBI - DNA as a Biometric

Dr. Pete Vallone
Lisa Borsuk
Kevin Kiesler
Becky (Hill) Steffen
Dr. Mike Coble

NIST Disclaimer: Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or any implication that the materials, instruments or equipment identified are necessarily the best available for the purposes.

Information presented does not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Commerce.

katherine.gettings@nist.gov