The Use of Statistical Analysis to Characterize Noise and Zygosity in Targeted Sequencing of Forensic STR Markers

Sarah Riman\textsuperscript{1}, PhD; Hari Iyer\textsuperscript{2}, PhD; Lisa Borsuk\textsuperscript{1}, MS; Peter M. Vallone\textsuperscript{1}, PhD

\textsuperscript{1}Applied Genetics Group
\textsuperscript{2}Statistical Design, Analysis, and Modeling Group
Disclaimer

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.
Overview

- Forensic DNA typing using the gold standard “Capillary Electrophoresis” (CE) technology vs. “Next Generation Sequencing” (NGS) technology

- Why implement NGS if you can accomplish DNA typing by CE?

- Characterization of single-source PowerSeq 46GY DNA profiles
Forensic DNA typing using the gold standard “Capillary Electrophoresis” (CE) technology vs. “Next Generation Sequencing” (NGS) technology
General Workflow for Next Generation Sequencing (NGS)

**CE Workflow**
- DNA Extraction
- DNA Quantitation
- Targeted DNA Amplification
- Library Construction
- Library Quantitation
- Library Pooling
- Clonal Amplification/Cluster Generation
- Cleanup of PCR products
- Cleanup of Library products

**Detection by CE**

**Data Analysis**
Targeted sequencing of STR markers relies on the PCR-amplification process
What is Library Construction?
The aim of library preparation is to flank amplified STR products with adapters on both ends.

Library preparation is essential for successful sequencing.
Data Analysis using CE technology vs. NGS technology
Data Analysis by CE

❖ Separation by size
❖ Length variation: 15
❖ Peak Height in RFU
Data Analysis by CE

- Separation by size
- Length variation: 15
- Peak Height in RFU

Data Analysis by NGS

- Separation by size
- Length variation: 15
- Peak Height in RFU
- Depth of coverage
- Sequence variation

Bioinformatics pipeline

ATCCTGCAGATGCATCC
GTCTGTGCTGTGCTGCTG
GTTATTTGAAGTCTCCTCC
GATTCCCTTTTAGTTCGC
TCTCATTTGCACCTGTT
CTGGGCCAACAAAGCA
CAGCAGTTTTCCTCCCTCC
TTTCTTTATGTTGCCTTG

vWA

coverage

[TAGA]11 [CAGA]3 TAGA
[TAGA]10 [CAGA]4 TAGA
Why implement NGS if you can accomplish DNA typing by CE?
Current Markers used in Forensic Genetics

- Autosomal STRs
- Y-STRs
- X-STRs
- Control region of mtDNA
- SNPs

NGS Sequencing Application and Markers

- Autosomal STRs
- Y STRs
- X STRs
- Autosomal SNPs
- Y-SNPs
- Ancestry SNPs
- Phenotype SNPs
- Pharmacogenetics SNPs
- Microbial Identification
- DNA methylation
- mRNA profiling
- Whole mt-Genome
- Microhaplotypes

- Examine one marker type at a time in one sample
- Multiplex samples
- Multiplex markers
- Distinguish between alleles identical by length but different in sequence content
Forensic labs are moving from threshold based systems towards fully continuous and probabilistic DNA interpretation systems.

Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles.

Bleka Ø¹, Storvik G², Gill P³.

What is DNA•VIEW®?
An integrated software package for DNA identification

Validating TrueAllele® DNA mixture interpretation.

Perlin MW¹, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW.
Current considerations of the CE probabilistic genotyping (PG) systems

PGs are based on models that describe the behavior of DNA profiles by:

- Understanding allele and stutter peak height variability
- Recognizing the limitations of the interpretation methodology
- Studying signal saturation
- Validating STR multiplexes
- Determining allele drop-out and drop-in
What do we need to understand to establish STR NGS interpretation systems?
We need to understand and analyze the STR NGS sequence data.

Characterization of Single-Source NGS DNA Profiles

- Allele drop-in/drop-out
- Heterozygote sequence balance
- Noise sequences
- Stutter sequences
- Zygosity
- Validate NGS STR multiplexes
- Effect of sample multiplexing
CE and NGS sensitivity experimental design
CE and NGS sensitivity experimental design

- **DNA Extraction**
  - 3 samples

  PowerPlex 46GY prototype
  - 500 pg, 250 pg, 125 pg, 60 pg, 30 pg, and 15 pg

- **Bead-based PCR Cleanup**
- **Library Construction**
  - TruSeq

  - 0.7X Bead-based Library Cleanup

- **Sequence**

  **Coverage ≥ 1**
  - STRait-Razor

- **Detection by CE**
  - RFU = 1

- **Three unique samples selected**
- **Run in triplicate**
  - Three unique amplifications of the serial dilutions
- **Dilution points**
  - 0.5 ng, 0.25 ng, 0.125 ng, 0.0625 ng, 0.03 ng, and 0.015 ng

  **Stochastic effects**
Noise Thresholds for CE Data

Maximizing allele detection: Effects of analytical threshold and DNA levels on rates of allele and locus drop-out.

Rakay CA¹, Bregu J, Grgicak CM.

Run-specific limits of detection and quantitation for STR-based DNA testing.

Gilder JR¹, Doom TE, Inman K, Krane DE.

Analytical thresholds and sensitivity: establishing RFU thresholds for forensic DNA analysis.

Bregu J¹, Conklin D, Coronado E, Terrill M, Cotton RW, Grgicak CM.
Analytical Threshold Most Commonly Determined by:

\[ AT_{M1} = \bar{Y}_{bl} + kS_{bl} \]

- \( k \) = Numerical factor (e.g. \( k=3 \))

Average RFU signal, STDEV of the signal

\[ AT = \mu + 3*\sigma \]
\[ AT = \mu + 10*\sigma \]

Noise Thresholds for NGS Data

AT level is set at 1.5% of total locus coverage.

Removal of general noise using thresholds created by fitting the distribution of general noise sequences.

\[ AT = c \times (\text{Max}_{\text{noise}} - \text{Min}_{\text{noise}}) \]

Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories.

Investigation of the STR loci noise distributions of PowerSeq™ Auto System.
Characterization of sequences in STR profiles generated on MiSeq platform using the PowerSeq 46GY prototype kit
We grouped the generated sequences into three categories:

S1 = Back stutter of the longest uninterrupted stretch of the basic repeat motifs within an allelic sequence
S2 = Back stutter sequences not attributed to S1
N = Noise sequences
Distribution of Known Allele, Stutter, and Noise Sequences

As expected, improved discrimination between known alleles (A) and the remainder of the sequences (N, S1, and S2) is observed as the amount of DNA template increases.
Observed Sequences and their coverage at a heterozygote D21S11 Locus

Locus Coverage = 6949

Y-axis in log scale

AT = \mu + 10*\sigma

AT = \% Coverage * Locus Coverage

AT = Range of Noise
At a percent coverage of 15%:
- **363 peaks are called**
  - **352 can be attributed to A**
  - **21 can be attributed to N**
  - **0 can be attributed to S**

Evaluating the tradeoff between the allelic (true positives), stutter, and noise sequences (false positives)
Evaluating the tradeoff between the allelic (true positives), stutter, and noise sequences (false positives)

A value of 15 % is **ONLY** used for illustrative purposes and not as a recommended threshold. Each lab should perform sensitivity experiments and establish a threshold for interpretational purposes.
**Summary**

- Understanding the behavior of STR NGS profiles can help in statistical modeling and probability distributions needed for establishing an STR NGS interpretation system.

- Future work will focus on analyzing more single source and mixture samples.

Presentation will be available for download from STRBase:
http://strbase.nist.gov/NISTpub.htm#Presentations

Acknowledgement

NIST

Pete Vallone

Hari Iyer (Statistical Design, Analysis, and Modeling Group)

Lisa Borsuk

Becky Steffen

Erica Romsos

Katherine Gettings

Kevin Kiesler

Margaret Kline

Megan Cleveland

Promega

Doug Storts

Spencer Hermanson

Funding

NIST Special Programs Office: Forensic DNA

FBI Biometrics Center of Excellence: Forensic DNA Typing as a Biometric tool.

All work presented has been reviewed and approved by the NIST Human Subjects Protections Office.

Contact: sarah.riman@nist.gov