Next Generation Sequencing Activities at NIST

NGS Workshop
Mid-Atlantic Association of Forensic Science
May 19, 2014

Katherine Butler Gettings, Ph.D.
Research Biologist, Applied Genetics Group
National Institute of Standards and Technology
Disclaimer

I will mention commercial STR kit names and information, but I am in no way attempting to endorse any specific products.

**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 it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

**Points of view are mine** and do not necessarily represent the official position of the National Institute of Standards and Technology or the U.S. Department of Justice. **Our group receives or has received funding from the FBI Laboratory and the National Institute of Justice.**
Outline

Background

NGS of Forensic DNA markers
  – STRs
  – mtDNA
  – Single Nucleotide Polymorphisms (SNPs)

NGS on the PGM- Ampliseq workflow

Experimental data
  – HID-Ion Ampliseq Identity Panel
  – HID-Ion Ampliseq Ancestry Panel
What’s in a name???

Massively parallel sequencing

NGS

Second-generation sequencing

Next-generation sequencing

Whole-genome sequencing

Third-generation sequencing

HIGH-THROUGHPUT SEQUENCING

Next-generation genomics
Parallel Sequencing

‘A million capillary Sanger sequencer’
Parallel Sequencing

‘A million capillary Sanger sequencer’

- Clonal vs population amplification
- Shorter reads (Range 75 to 400)
- Errors are more ‘detectable’
- High coverage 100 – 1000 - 10,000x
- **Rely more on informatics to assemble millions of short reads**
MOORE’S LAW

“Transistor density on integrated circuits doubles about every two years.”

1950s
Silicon Transistor

1960s
TTL Quad Gate

1970s
8-bit Microprocessor

1980s
32-bit Microprocessor

1990s
32-bit Microprocessor

2000s
64-bit Microprocessor

1 Transistor

16 Transistors

4500 Transistors

275,000 Transistors

3,100,000 Transistors

592,000,000 Transistors
If transistors were people

If the transistors in a microprocessor were represented by people, the following timeline gives an idea of the pace of Moore’s Law.

2,300
Average music hall capacity

134,000
Large stadium capacity

32 Million
Population of Tokyo

1.3 Billion
Population of China

1970
1980
1990
2000
2011

Intel 4004
Intel 286
Pentium III
Core i7 Extreme Edition

Now imagine that those 1.3 billion people could fit onstage in the original music hall. That’s the scale of Moore’s Law.
Forensic NGS Applications

• **Short Tandem Repeats (STRs)**
  – PCR fragment-length polymorphisms

• **Mitochondrial DNA (mtDNA)**
  – Sanger sequencing

• **Single Nucleotide Polymorphisms (SNPs)**

  Capillary electrophoresis electropherogram
NGS of Forensic STR Loci

- D8S1179
- D21S11
- D7S820
- CSF1PO
- D3S1358
- TH01
- D13S317
- D16S539
- D2S1338
- D19S433
- vWA
- TPOX
- D18S51
- A
- D5S818
- FGA
NGS of Forensic STR Loci

Sizes of largest observed alleles

*(not including primer binding/flanking region)*
NGS of Forensic STR Loci

Sizes of largest observed alleles

(not including primer binding/flanking region)
NGS of Forensic STR Loci

D21S11: Individual appears homozygous by CE but different sequencing composition shown with NGS.
NGS of Forensic STR Loci

D21S11: Individual appears homozygous by CE but different sequencing composition shown with NGS.
D21S11: Individuals appear homozygous by CE but different sequencing composition shown with NGS.
NGS of Forensic STR Loci

1. Reads containing both the leading and trailing flanking regions for a given locus are extracted from the raw sequence data. Reads with a user-defined number of allowable mismatches in the flanking regions, such as the substitutions denoted by asterisks (*), are detected, as well.
NGS of Forensic STR Loci

1. Reads containing both the leading and trailing flanking regions for a given locus are extracted from the raw sequence data. Reads with a user-defined number of allowable mismatches in the flanking regions, such as the substitutions denoted by asterisks (*), are detected, as well.

2. The surrounding sequence data, including the flanking regions, are "shaved" away, leaving the repeat regions themselves. The repeat regions are then filtered based on the presence of a small portion of the repeat motif. The number of bases in each filtered repeat region is then determined.
NGS of Forensic STR Loci

1. Reads containing both the leading and trailing flanking regions for a given locus are extracted from the raw sequence data. Reads with a user-defined number of allowable mismatches in the flanking regions, such as the substitutions denoted by asterisks (*), are detected, as well.

2. The surrounding sequence data, including the flanking regions, are "shaved" away, leaving the repeat regions themselves. The repeat regions are then filtered based on the presence of a small portion of the repeat motif. The number of bases in each filtered repeat region is then determined.

3. Allele determination is performed by comparing the repeat region lengths to the known repeat motif.

Allele = 9
NGS of Forensic STR Loci

STR sequence data from 2391c, Component A:

- Truseq Library Prep
- MiSeq sequencing
- STRait Razor data parsing
- R script (NIST) data viewer
Forensic DNA Markers

- **Short Tandem Repeats (STRs)**
  - PCR fragment-length polymorphisms

- **Mitochondrial DNA (mtDNA)**
  - Sanger sequencing

- **Single Nucleotide Polymorphisms (SNPs)**
mtDNA Information

- Increase in variants by whole genome analysis

Over ten times more variable
Three times more polymorphisms than HV alone

*(based on analysis of 3 SRM samples)*
mtDNA Information

Current Method
• Sequence based on chromatogram
• Consensus of one forward and one reverse

NGS
• Sequence based on thousands of individual reads
• Improved sensitivity:
  – Mixture detection
  – Low level heteroplasmy
mtDNA Information

Current Method
• Minor peaks may not be reproducible
• SRM 2392 9947a, 1393 G/A heteroplasmcy

NGS
• More consistent detection of minor genotypes
• Validation important
  – Variant calling thresholds
  – Characterizing noise
Characterization of SRM 2392 and 2392-I
Mitochondrial genome sequencing standard
Detection of low level heteroplasmy

Minor Allele Frequency

HL60 - 2445  C
HL60 - 5149  T
9947A - 1393  A
9947A - 7861  T
Characterization of SRM 2392 and 2392-I
Mitochondrial genome sequencing standard
Detection of low level heteroplasmy

Minor Allele Frequency

Coverage
Forensic DNA Markers

• Short Tandem Repeats (STRs)
  – PCR fragment-length polymorphisms
• Mitochondrial DNA (mtDNA)
  – Sanger sequencing
• Single Nucleotide Polymorphisms (SNPs)

Most methods are **low throughput** and/or **require a lot of DNA**

NGS method can analyze many SNPs for many samples in one run

http://portal.ccg.uni-koeln.de/
SNP Information

- IISNP-Individual
- AISNP-Ancestry
- LISNP-Lineage
- PISNP-Phenotype
SNP Information

• Individual Identification
  – Balancing has occurred in all populations
  – Low F statistics within ($F_{IS}$) and among ($F_{ST}$) populations
  – High heterozygosity
SNP Information

• Individual Identification
  
  Pakstis 2010, Kidd 2012
  
  – Panel of 45 unlinked SNPs
  – $F_{ST}$ below $\approx 0.07$
  – Avg het $> 0.4$
  – RMP $10^{-15}$ to $10^{-18}$
  in 44 populations
SNP Information

- HID-Ion Ampliseq Identity Panel *(version 2.3)*
  - 90 autosomal SNPs
  - 30 Y-chromosome SNPs
  - RMP $10^{-35}$
SNP Information

• Ancestry Information
  – High Fixation Index ($F_{ST}$)
  – Population specific fixation has occurred
  – Low heterozygosity

• Example
  – Malaria resistance SNPs in Sub-Saharan Africa
SNP Information

- HID Ancestry Panel
  - Beta version 3.0
  - Publicly available soon
  - 170 loci
  - Derived from
    - Kosoy et. al (2008): 128 SNPs
    - Kidd et. al (2014): 55 SNPs

**Research Article**

Ancestry Informative Marker Sets for Determining Continental Origin and Admixture Proportions in Common Populations in America

Roman Kosoy,1 Rami Nassir,1 Choo Tian,1 Phoebe A. White,2 Lesley M. Butler,2 Gabriel Silvo,4 Rick Kittles,5 Marta E. Alarcon-Riquelme,6 Peter K. Gregersen,7 John W. Belmont,8 Francisco M. De La Vega,7 and Michael F. Seldin1*
Life Tech - Ion Torrent - PGM

• Ion Torrent Personal Genome Machine (PGM)
  – Launched in 2010

• Ion Torrent sequencing:
  – Emulsion PCR for single copy reactors
  – Non-labeled nucleotide triphosphates
  – Flowed over a bead on a semiconductor surface

• Hydrogen Ion detection
  – pH change is detected
  – No optics
Ion Torrent PGM Workflow

http://www.youtube.com/watch?v=MxkYa9XCvBQ
The PGM Instrument at NIST
Ampliseq Workflow

1 ng DNA input

Ampliseq Primer Pool

PCR amplify

Chew back primers

Ligate adapters

Emulsion PCR

Sequencing

One template per bead/droplet

Ion Ampliseq Library Kit

Template Kit

Sequencing Kit

PGM “Ionogram”
Front-End: Multiplex PCR

• HID-Ion Ampliseq Identity Panel (IISNP)
  – 120 markers in a single PCR reaction
  – Amplified regions **33 bp to 192 bp long**

• HID-Ion Ampliseq Ancestry Panel (AISNP)
  – 170 markers in a single PCR reaction
  – Amplified regions **34 bp to 136 bp long**

• Small amplicons well suited to degraded or damaged DNA
Digest Primer Regions & Ligate Adaptors

- Enzymatic digestion removes ≈ 25 bp from ends of amplicons

- Universal sequencing adaptors are ligated to DNA
  - Adaptors termed P1 and A

- Barcoded sequencing adaptors can be used in this step
  - Sequence multiple samples in one PGM run
Prepare Ion Sphere Particles (ISPs)

- Libraries quantified by qPCR
  - Quantity of DNA going into emPCR is very important!
  - Goal: 10% to 30% template positive ISPs
    - Too much DNA → polyclonal ISPs (mixed read)

![Diagram of ideal and non-ideal ISPs](image)
Prepare Ion Sphere Particles (ISPs)

- Libraries quantified by qPCR
  - Quantity of DNA going into emPCR is very important!
  - Goal: 10% to 30% template positive ISPs
    - Too much DNA → polyclonal ISPs (mixed read)

- Emulsion PCR
  - Nanoliter droplets of PCR reagents in oil
  - Attaches sequencing template to the ISP

OneTouch 2
Prepare Ion Sphere Particles (ISPs)

- Libraries quantified by qPCR
  - Quantity of DNA going into emPCR is very important!
  - Goal: 10% to 30% template positive ISPs
    - Too much DNA → polyclonal ISPs (mixed read)

- Emulsion PCR
  - Nanoliter droplets of PCR reagents in oil
  - Attaches sequencing template to the ISP

- Enrich for positive ISPs
  - Liquid handler removes non-templated ISPs
  - Biotinylated primer/streptavidin beads
Sequencing & Data Analysis

- Library ISPs loaded onto chip
- PGM runs flows & detects pH
- Torrent Server & Torrent Suite Software
  - Processes pH signal into base calls
  - Displays run summary
  - Maps reads to reference genome
Data Analysis
HID SNP Genotyper Plugin

Allele coverage histogram

X-axis is refSNP I.D.

Normalized y-axis scale

Autosomal SNPs

Y-SNPs
Data Analysis
HID SNP Genotyper Plugin

<table>
<thead>
<tr>
<th>Sample Name</th>
<th>Barcode Id</th>
<th>Genotype String</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRAC_33-50</td>
<td>ionXpress_016</td>
<td>Details</td>
</tr>
<tr>
<td>FRAC_50-75</td>
<td>ionXpress_017</td>
<td>Details</td>
</tr>
<tr>
<td>FRAC_00-19</td>
<td>ionXpress_018</td>
<td>Details</td>
</tr>
<tr>
<td>FRAC_00-20</td>
<td>ionXpress_019</td>
<td>Details</td>
</tr>
<tr>
<td>FRAC_00-20</td>
<td>ionXpress_020</td>
<td>Details</td>
</tr>
</tbody>
</table>

**Total Coverage**

**Reads for Each Base**

**Coverage for Either Strand**

**Strand Bias**

**Quality Score**

**Major Allele Frequency**

**Genotype**

<table>
<thead>
<tr>
<th>Cov</th>
<th>A Reads</th>
<th>C Reads</th>
<th>G Reads</th>
<th>T Reads</th>
<th>Deletions</th>
<th>-Cov</th>
<th>Cov</th>
<th>% +Cov</th>
<th>Genotype</th>
<th>Qual</th>
<th>Maj. Allele Freq (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3235</td>
<td>1474</td>
<td>0</td>
<td>1716</td>
<td>0</td>
<td>45</td>
<td>1465</td>
<td>1785</td>
<td>44.0%</td>
<td>AC</td>
<td>747.64</td>
<td>53.04</td>
</tr>
<tr>
<td>2222</td>
<td>2214</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>873</td>
<td>1349</td>
<td>39.3%</td>
<td>AA</td>
<td>9788.11</td>
<td>99.64</td>
</tr>
<tr>
<td>1261</td>
<td>569</td>
<td>0</td>
<td>679</td>
<td>13</td>
<td>0</td>
<td>620</td>
<td>641</td>
<td>49.2%</td>
<td>AC</td>
<td>703.96</td>
<td>53.65</td>
</tr>
<tr>
<td>2331</td>
<td>2330</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1014</td>
<td>1317</td>
<td>43.5%</td>
<td>AA</td>
<td>181.32</td>
<td>99.96</td>
</tr>
<tr>
<td>3379</td>
<td>1544</td>
<td>1</td>
<td>1694</td>
<td>11</td>
<td>129</td>
<td>1223</td>
<td>2027</td>
<td>37.6%</td>
<td>AC</td>
<td>940.58</td>
<td>50.13</td>
</tr>
<tr>
<td>4655</td>
<td>3</td>
<td>3</td>
<td>4646</td>
<td>3</td>
<td>0</td>
<td>1581</td>
<td>3704</td>
<td>34.0%</td>
<td>GC</td>
<td>782.53</td>
<td>99.81</td>
</tr>
<tr>
<td>2182</td>
<td>0</td>
<td>986</td>
<td>2</td>
<td>1146</td>
<td>48</td>
<td>1161</td>
<td>573</td>
<td>54.4%</td>
<td>CT</td>
<td>683.67</td>
<td>52.52</td>
</tr>
<tr>
<td>592</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>586</td>
<td>1</td>
<td>325</td>
<td>262</td>
<td>55.7%</td>
<td>TT</td>
<td>3805.42</td>
<td>58.59</td>
</tr>
<tr>
<td>1664</td>
<td>1</td>
<td>1</td>
<td>1660</td>
<td>0</td>
<td>2</td>
<td>704</td>
<td>538</td>
<td>42.4%</td>
<td>CC</td>
<td>175.09</td>
<td>99.76</td>
</tr>
<tr>
<td>1135</td>
<td>1131</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>585</td>
<td>550</td>
<td>51.5%</td>
<td>AA</td>
<td>176.022</td>
<td>99.65</td>
</tr>
<tr>
<td>975</td>
<td>0</td>
<td>859</td>
<td>0</td>
<td>56</td>
<td>569</td>
<td>290</td>
<td>200</td>
<td>66.2%</td>
<td>CC</td>
<td>3669.93</td>
<td>92.88</td>
</tr>
<tr>
<td>1507</td>
<td>9</td>
<td>1019</td>
<td>932</td>
<td>2</td>
<td>5</td>
<td>907</td>
<td>1055</td>
<td>46.2%</td>
<td>GC</td>
<td>1046.97</td>
<td>51.8</td>
</tr>
<tr>
<td>1989</td>
<td>4</td>
<td>115</td>
<td>0</td>
<td>1018</td>
<td>35</td>
<td>969</td>
<td>985</td>
<td>49.6%</td>
<td>TT</td>
<td>3609.27</td>
<td>92.26</td>
</tr>
<tr>
<td>3770</td>
<td>26</td>
<td>1</td>
<td>1766</td>
<td>1975</td>
<td>2</td>
<td>2729</td>
<td>1039</td>
<td>72.4%</td>
<td>CT</td>
<td>805.29</td>
<td>52.39</td>
</tr>
<tr>
<td>5712</td>
<td>5543</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>165</td>
<td>2618</td>
<td>2929</td>
<td>47.2%</td>
<td>AA</td>
<td>3841.28</td>
<td>97.04</td>
</tr>
</tbody>
</table>
HID SNP Panel
Sensitivity Study

• Dynamic range of DNA input to PCR
  – 1 ng is recommended
  – 10 ng (1 data point) – no problems were observed
  – 1 ng
  – 0.5 ng
  – 0.1 ng
  – 0.05 ng

• Libraries were generated and pooled (n = 12)
• Sequenced on PGM 318 chip (11 M wells)
  – 200 bp read chemistry

3 Replicates
HID SNP Panel Sensitivity Study

90 Autosomal SNP loci, sorted from highest to lowest coverage

Thresholds: analytical = 50 RFU, stochastic = 300 RFU, PHR = 0.5

Thresholds: 50X analytical 300X stochastic 50% balance
HID SNP Panel Sensitivity Study

<table>
<thead>
<tr>
<th>Input DNA</th>
<th>0.05 ng</th>
<th>0.1 ng</th>
<th>0.5 ng</th>
</tr>
</thead>
</table>

**Identifiler® Plus amplification (29 cycle), 25 µl reaction, 3500x electrophoresis, 1.2 kV for 8 seconds**

Thresholds: analytical = 50 RFU, stochastic = 200 RFU, PHR = 0.5

All scaled to 4000 RFU
### HID SNP Panel Sensitivity Study

#### Identifiler Plus

<table>
<thead>
<tr>
<th>Locus</th>
<th>0.05 ng</th>
<th>0.1 ng</th>
<th>0.5 ng</th>
</tr>
</thead>
<tbody>
<tr>
<td>D8S1179</td>
<td>1.78E+01</td>
<td>1.17E+05</td>
<td>1.17E+06</td>
</tr>
<tr>
<td>D21S11</td>
<td>1.17E+06</td>
<td>6.19E+10</td>
<td>6.34E+11</td>
</tr>
<tr>
<td>D7S820</td>
<td>1.17E+06</td>
<td>2.88E+30</td>
<td>5.67E+14</td>
</tr>
<tr>
<td>CSF1PO</td>
<td>1.17E+06</td>
<td>6.34E+11</td>
<td>1.17E+06</td>
</tr>
<tr>
<td>D3S1358</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
</tr>
<tr>
<td>TH01</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
</tr>
<tr>
<td>D13S317</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
</tr>
<tr>
<td>D16S539</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
</tr>
<tr>
<td>D2S1338</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
</tr>
<tr>
<td>D19S433</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
</tr>
<tr>
<td>vWA</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
</tr>
<tr>
<td>TPOX</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
</tr>
<tr>
<td>D18S51</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
</tr>
<tr>
<td>D5S818</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
</tr>
<tr>
<td>FGA</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
<td>4.35E+03</td>
</tr>
</tbody>
</table>

#### PGM HID SNP Panel v2.3

<table>
<thead>
<tr>
<th>Locus</th>
<th>0.05 ng</th>
<th>0.1 ng</th>
<th>0.5 ng</th>
</tr>
</thead>
<tbody>
<tr>
<td>D8S1179</td>
<td>6.12E+15</td>
<td>3.62E+18</td>
<td>1.16E+35</td>
</tr>
<tr>
<td>D21S11</td>
<td>7.58E+27</td>
<td>7.58E+27</td>
<td>1.16E+35</td>
</tr>
<tr>
<td>D7S820</td>
<td>7.17E+21</td>
<td>2.88E+30</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>CSF1PO</td>
<td>8.87E+27</td>
<td>8.87E+27</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>D3S1358</td>
<td>7.17E+21</td>
<td>7.17E+21</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>TH01</td>
<td>7.17E+21</td>
<td>7.17E+21</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>D13S317</td>
<td>7.17E+21</td>
<td>7.17E+21</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>D16S539</td>
<td>7.17E+21</td>
<td>7.17E+21</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>D2S1338</td>
<td>7.17E+21</td>
<td>7.17E+21</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>D19S433</td>
<td>7.17E+21</td>
<td>7.17E+21</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>vWA</td>
<td>7.17E+21</td>
<td>7.17E+21</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>TPOX</td>
<td>7.17E+21</td>
<td>7.17E+21</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>D18S51</td>
<td>7.17E+21</td>
<td>7.17E+21</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>D5S818</td>
<td>7.17E+21</td>
<td>7.17E+21</td>
<td>2.31E+35</td>
</tr>
<tr>
<td>FGA</td>
<td>7.17E+21</td>
<td>7.17E+21</td>
<td>2.31E+35</td>
</tr>
</tbody>
</table>

*Just like STR loci, some SNPs are consistently less robust*
### Identifiler Plus

SNPs have more possible loci and better performance at low levels.

<table>
<thead>
<tr>
<th>SNPs</th>
<th>0.05 ng</th>
<th>0.1 ng</th>
<th>0.5 ng</th>
</tr>
</thead>
<tbody>
<tr>
<td>D8S1179</td>
<td>(1.78 \times 10^1)</td>
<td>(1.17 \times 10^5)</td>
<td>(1.17 \times 10^6)</td>
</tr>
<tr>
<td>D3S1358</td>
<td>(1.17 \times 10^5)</td>
<td>(4.35 \times 10^3)</td>
<td>(6.34 \times 10^1)</td>
</tr>
<tr>
<td>D2S1338</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
<tr>
<td>D19S433</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
<tr>
<td>vWA</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
<tr>
<td>TPOX</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
<tr>
<td>D18S51</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
<tr>
<td>FGA</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
<tr>
<td>TH01</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
<tr>
<td>D5S818</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
<tr>
<td>D7S820</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
<tr>
<td>D16S539</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
<tr>
<td>CSF1PO</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
<tr>
<td>D21S11</td>
<td>(6.19 \times 10^4)</td>
<td>(6.34 \times 10^3)</td>
<td>(5.67 \times 10^1)</td>
</tr>
</tbody>
</table>

### PGM HID SNP Panel v2.3

<table>
<thead>
<tr>
<th>SNPs</th>
<th>0.05 ng</th>
<th>0.1 ng</th>
<th>0.5 ng</th>
</tr>
</thead>
<tbody>
<tr>
<td>D8S1179</td>
<td>(6.12 \times 10^5)</td>
<td>(3.62 \times 10^8)</td>
<td>(7.17 \times 10^{11})</td>
</tr>
<tr>
<td>D3S1358</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>D2S1338</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>D19S433</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>vWA</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>TPOX</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>D18S51</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>FGA</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>TH01</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>D5S818</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>D7S820</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>D16S539</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>CSF1PO</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
<tr>
<td>D21S11</td>
<td>(3.62 \times 10^8)</td>
<td>(7.87 \times 10^{14})</td>
<td>(2.88 \times 10^{18})</td>
</tr>
</tbody>
</table>
HID SNP Panel Sensitivity Study

HID SNPs give better RMP with 50 pg than ID+ gives with 0.5 ng
HID SNP Panel Sensitivity Study Summary

• Higher RMPs are expected for SNP panel compared to STRs due to many more loci
• Under thresholds indicated, higher % SNPs produce results than STRs also
• Better STR assays (GlobalFiler or NGS-STR) may lessen the “gap”
• Validation needed for SNP thresholds
HID SNP Panel Degraded DNA Study

Sheared genomic DNA
→ Covaris S2 Focused Ultrasonicator

gDNA + Ultrasound = Sheared DNA
HID SNP Panel
Degraded DNA Study

Sheared DNA was fractionated by size range

Blue Pippin system (3% Gel)
Automated size selection

1) 50 bp to 200 bp
2) 50 bp to 150 bp
3) 50 bp to 100 bp
4) 50 bp to 75 bp
5) 35 bp to 50 bp

Five individual agarose columns
Size fractionated fragments collected into recovery wells
HID SNP Panel
Degraded DNA Study

Sheared DNA was fractionated by size range

- Agilent Bioanalyzer Trace
- Size selected sheared DNA
  - 50 bp to 200 bp
  - 50 bp to 150 bp
  - 50 bp to 100 bp
  - 50 bp to 75 bp
  - 35 bp to 50 bp

Input to HID Panel PCR
- 1 ng DNA
- Built libraries and sequenced
HID SNP Panel
Degraded DNA Study

90 autosomal IISNPs
HID SNP Panel Degraded DNA Study

<table>
<thead>
<tr>
<th>Fragmented, size selected &lt; 75 bp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fragmented, size selected &lt; 100 bp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fragmented, size selected &lt; 150 bp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fragmented, size selected &lt; 200 bp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fragmented, size selected &lt; 250 bp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fragmented, non-size selected</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Minifiler® amplification (30 cycle), 25 µl reaction, 3500x/ electrophoresis, 1.2 kV for 8 seconds
Thresholds: analytical = 100 RFU, PHR = 0.5; data scaled to 1000 RFU

Performed in triplicate
One rep shown
90 Autosomal SNPs, sorted from smallest to largest

Fragmented, size selected < 75 bp

Fragmented, size selected < 100 bp

Fragmented, size selected < 150 bp

Fragmented, size selected < 200 bp

Fragmented, size selected < 250 bp

Fragmented, non-size selected

Thresholds:
50X analytical
300X stochastic
50% balance

PGM 318 Chip, all scaled to 2000X coverage
HID SNP Panel
Degraded DNA Study

MiniFiler STRs (8)

PGM USNPs (90)
HID SNP Panel
Degraded DNA Study

SNPs have MANY more possible loci and better performance in degraded samples.
HID SNP Panel Degraded DNA Study

Random Match Probability

PGM HID SNP Panel

MiniFiler

<75  <100  <150  <200  <250  Fragmented
SNPs and STRs show expected performance in each fraction based on amplicon size.

Some SNPs can still amplify in degraded samples where STRs cannot.

Due to the high number of SNPs, very high RMPs are possible.

Better STR assays (GlobalFiler or NGS-STR) may lessen the “gap”.

Validation needed for SNP thresholds.
SNPs and Mixtures

At each SNP, a person is either homozygous or heterozygous.

Single source samples have 3 possible biallelic genotypes: AA, BB, or AB

% is coverage (like PH balance)

<table>
<thead>
<tr>
<th>1</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>AB</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>BB</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>
HID SNP Panel
Mixture Detection

One single source sample, major allele frequency plotted for 90 HID SNPs (in ascending order)

Akin to an imbalanced STR locus

Single source samples should be either
50% or 100%

HOMOZYGOTE
AA or BB

HETEROZYGOTE
AB
HID SNP Panel

Mixture Detection

One single source sample in triplicate, major allele frequency plotted for 90 HID SNPs (in ascending order)

3 SNPs give outlying values, less useful for mixtures

HOMOZYGOTE AA or BB

Single source samples should be either

50% or 100%

90 Autosomal SNPs
HID SNP Panel
Mixture Detection

Two single source samples in triplicate, major allele frequency plotted for 90 HID SNPs (in ascending order)

Single source samples should be either 50% or 100%
HID SNP Panel
Mixture Detection

Three single source samples in triplicate, major allele frequency plotted for 90 HID SNPs (in ascending order)

HOMOZYGOTE
AA or BB

HETEROZYGOTE
AB

Single source samples should be either
50% or 100%

90 Autosomal SNPs

Major Allele Frequency

50 55 60 65 70 75 80 85 90 95 100
HID SNP Panel
Mixture Detection

Four single source samples in triplicate, major allele frequency plotted for 90 HID SNPs (in ascending order).

HOMOZYGOTE
AA or BB

HETEROZYGOTE
AB

Single source samples should be either
50% or 100%
HID SNP Panel
Mixture Detection

Major Allele Frequency

Five single source samples in triplicate, major allele frequency plotted for 90 HID SNPs (in ascending order)

Assessing single source outliers will improve mixture model

Single source samples should be either 50% or 100%

HOMOZYGOTE AA or BB

HETEROZYGOTE AB

90 Autosomal SNPs
Two-person mixtures have 9 possible genotype combinations:
3 genotypes (Person 1) × 3 genotypes (Person 2)
One single source sample, major allele frequency plotted for 90 HID SNPs (in ascending order).

Genotype combinations in this bin are: AA:AB, AB:AA, AB:BB, BB:AB

A two-person mixture in a 1:1 ratio should have frequencies at: 50%, 75%, and 100%
SNPs in 2:1 Mixtures

Two-person mixtures have 9 possible genotype combinations:
3 genotypes (Person 1) × 3 genotypes (Person 2)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2A</th>
<th>1A</th>
<th>2B</th>
<th>1B</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>AA</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>AA</td>
<td>AB</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>83%</td>
<td>17%</td>
</tr>
<tr>
<td>AA</td>
<td>BB</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>67%</td>
<td>33%</td>
</tr>
<tr>
<td>AB</td>
<td>AA</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>67%</td>
<td>33%</td>
</tr>
<tr>
<td>AB</td>
<td>AB</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>AB</td>
<td>BB</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>33%</td>
<td>67%</td>
</tr>
<tr>
<td>BB</td>
<td>AA</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>33%</td>
<td>67%</td>
</tr>
<tr>
<td>BB</td>
<td>AB</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>17%</td>
<td>83%</td>
</tr>
<tr>
<td>BB</td>
<td>BB</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>
SNPs in 2:1 Mixtures

One single source sample, major allele frequency plotted for 90 HID SNPs (in ascending order)

A two-person mixture in a 1:1 ratio should have frequencies at: 50%, 67.5%, 82.5%, 100%
SNPs in 3:1 Mixtures

Two-person mixtures have 9 possible genotype combinations:
3 genotypes (Person 1) \times 3 genotypes (Person 2)
Variants Frequency

One single source sample, major allele frequency plotted for 90 HID SNPs (in ascending order)

A two-person mixture in a 3:1 ratio should have frequencies at: 50%, 62.5%, 75%, 87.5% and 100%
SNPs in 3:1 Mixtures

One single source sample and one 3:1 mixed sample, major allele frequency plotted for 90 HID SNPs (in ascending order)

A two-person mixture in a 3:1 ratio should have frequencies at: 50%, 62.5%, 75%, 87.5% and 100%
SNPs in 3:1 Mixtures

One single source sample and two 3:1 mixed samples, major allele frequency plotted for 90 HID SNPs (in ascending order).

A two-person mixture in a 3:1 ratio should have frequencies at: 50%, 62.5%, 75%, 87.5% and 100%.
SNPs in 3:1 Mixtures

A two-person mixture in a 3:1 ratio should have frequencies at: 50%, 62.5%, 75%, 87.5% and 100%

One single source sample and three 3:1 mixed samples, major allele frequency plotted for 90 HID SNPs (in ascending order)
HID SNP Panel Mixtures Summary

• Mixtures can be detected in SNP data based on the coverage levels at heterozygous loci
• It may be possible to determine two-person 1:1 or 2:1 mixtures (maaaybe 3:1)
• More than two contributors or greater than 3:1 mixtures will be difficult to distinguish
• Need to determine which SNPs “behave”
• Stay tuned!
PGM AIM Panel (beta testing)

- Ampliseq library prep
- 170 SNPs
- Seldin 128
- Kidd 55
- Analysis plug-in integrates FROGkb
AIM Panel
Ancestry Prediction – SRM 2391c

• Likelihood Ratio calculations
  – Four categories extant in both Kidd and Seldin studies
    • Europeans, African Americans, Maya, and Han Chinese
  – Allows comparison of SNP sets’ performance
  – Representative of major U.S. populations

<table>
<thead>
<tr>
<th>SRM 2391c Component</th>
<th>Gender</th>
<th>Ethnicity (self declared)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Female</td>
<td>Not listed</td>
</tr>
<tr>
<td>B</td>
<td>Male</td>
<td>Mexican-American</td>
</tr>
<tr>
<td>C</td>
<td>Male</td>
<td>Melanesian</td>
</tr>
<tr>
<td>D</td>
<td>Female:Male</td>
<td>Mixed sample</td>
</tr>
<tr>
<td>E</td>
<td>Female</td>
<td>Not listed</td>
</tr>
<tr>
<td>F</td>
<td>Male</td>
<td>Caucasian</td>
</tr>
</tbody>
</table>
HID SNP Genotyper Plugin (v4.1 Beta)
New Feature – Ancestry Map

- Heatmap of highest probability of origin
# Ancestry Prediction

**SRM 2391c Component A**

<table>
<thead>
<tr>
<th>SRM 2391c Component</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Kidd 55 Prediction</th>
<th>Seldin 128 Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Female</td>
<td>Not listed</td>
<td>European $1.02 \times 10^{33}$</td>
<td>European $6.32 \times 10^{66}$</td>
</tr>
</tbody>
</table>

**Kidd 55 SNPs**

**Seldin 128 SNPs**

[Maps of Kidd 55 SNPs and Seldin 128 SNPs showing genetic distributions across the world.]
## Ancestry Prediction

### SRM 2391c Component B

<table>
<thead>
<tr>
<th>SRM 2391c Component</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Kidd 55 Prediction</th>
<th>Seldin 128 Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Male</td>
<td>Mexican-American</td>
<td>$5.39 \times 10^{12}$</td>
<td>Han Chinese $1.48 \times 10^{19}$</td>
</tr>
</tbody>
</table>

**Kidd 55 SNPs**

**Seldin 128 SNPs**
# Ancestry Prediction

## SRM 2391c Component C

<table>
<thead>
<tr>
<th>SRM 2391c Component</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Kidd 55 Prediction</th>
<th>Seldin 128 Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>Male</td>
<td>Melanesian</td>
<td>Han Chinese $1.54 \times 10^{14}$</td>
<td>Han Chinese $6.67 \times 10^{28}$</td>
</tr>
</tbody>
</table>

*Maps show the distribution of Kidd 55 SNPs and Seldin 128 SNPs.*
## Ancestry Prediction

### SRM 2391c Component E

<table>
<thead>
<tr>
<th>SRM 2391c Component</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Kidd 55 Prediction</th>
<th>Seldin 128 Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>Female</td>
<td>Not listed</td>
<td>European $5.41 \times 10^{21}$</td>
<td>European $3.92 \times 10^{50}$</td>
</tr>
</tbody>
</table>

### Kidd 55 SNPs

### Seldin 128 SNPs

![World Map](image1.png)

![World Map](image2.png)
Ancestry Prediction
SRM 2391c Component F

<table>
<thead>
<tr>
<th>SRM 2391c Component</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Kidd 55 Prediction</th>
<th>Seldin 128 Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Male</td>
<td>Caucasian</td>
<td>European $2.35 \times 10^{31}$</td>
<td>European $1.16 \times 10^{55}$</td>
</tr>
</tbody>
</table>

Kidd 55 SNPs

Seldin 128 SNPs
HID SNP Panel Ancestry Summary

• 170 SNP panel containing two SNP sets that are suitable for use in U.S.
• Plug-in integrates FROG-kb ([http://frog.med.yale.edu/FrogKB/](http://frog.med.yale.edu/FrogKB/))
• Heat maps give quick overview
• Interpretation tools being developed
  – Combining loci
  – Choosing/combining populations
Conclusions

• NGS can give more information on currently used forensic markers
  – More STRs and STR sequence info
  – Whole genome mtDNA
• NGS facilitates genotyping of forensic SNPs
• SNPs may help with low level & degraded samples
• SNPs may provide ancestry (and phenotype?) information
• Forensic NGS kits/methods are being developed
• Many questions to answer prior to implementation
Acknowledgements

THANK YOU

Dr. Peter Vallone
Group Leader

Kevin Kiesler
Research Biologist

Funding from the
FBI Biometrics Center of Excellence
Forensic DNA Typing as a Biometric Tool

Thermo Fisher (Life Tech):
Nnamdi Ihuegbru
Robert Lagace
Thank you for your attention!

Contact Info:

katherine.gettings@nist.gov
301-975-6401