DNA Future Trends Technical Review/Workshop

Michael D. Coble
NIST Applied Genetics Group
National Institute of Standards and Technology
Gaithersburg, Maryland

Presentation Outline
• Marker Systems – STRs
• Marker Systems – SNPs
• Low Template (Copy Number) Testing
• Second (Next) Generation Sequencing
• Mixture Interpretation

Types of Genetic Variation

Length Variation
• short tandem repeats (STRs)
CODIS Loci are STRs
CTAGTCGT(GATA)(GATA)(GATA)GGATCGGT

Sequence Variation
• insertions/deletions
• single nucleotide polymorphisms (SNPs)
GCTAGTCATGCTC(G/A)GCGTATGCT

The 11 STR Loci Beyond the CODIS 13

<table>
<thead>
<tr>
<th>STR Locus</th>
<th>Location</th>
<th>Repeat Motif</th>
<th>Allele Range*</th>
<th># Alleles*</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2S1338</td>
<td>2q35</td>
<td>TGCC/TTCC</td>
<td>10 to 31</td>
<td>40</td>
</tr>
<tr>
<td>D19S433</td>
<td>19q12</td>
<td>AAGG/AGGA</td>
<td>5.2 to 20</td>
<td>36</td>
</tr>
<tr>
<td>Penta D</td>
<td>21q22.3</td>
<td>AAGGA</td>
<td>1.1 to 19</td>
<td>50</td>
</tr>
<tr>
<td>Penta E</td>
<td>15q26.2</td>
<td>AAGGA</td>
<td>5 to 32</td>
<td>53</td>
</tr>
<tr>
<td>D1S1656</td>
<td>1q42</td>
<td>TAGA</td>
<td>8 to 20.3</td>
<td>25</td>
</tr>
<tr>
<td>D12S391</td>
<td>12p13.2</td>
<td>AGAG/AGAC</td>
<td>13 to 27.2</td>
<td>52</td>
</tr>
<tr>
<td>D2S441</td>
<td>2p14</td>
<td>TCTA/TCAA</td>
<td>8 to 17</td>
<td>22</td>
</tr>
<tr>
<td>D10S1248</td>
<td>10q26.3</td>
<td>GGAA</td>
<td>7 to 19</td>
<td>13</td>
</tr>
<tr>
<td>D22S1045</td>
<td>22q12.3</td>
<td>ATT</td>
<td>7 to 20</td>
<td>14</td>
</tr>
<tr>
<td>SE33</td>
<td>6q14</td>
<td>AAAG</td>
<td>3 to 49</td>
<td>178</td>
</tr>
<tr>
<td>D6S1043</td>
<td>6q15</td>
<td>AGAT/AGAC</td>
<td>8 to 25</td>
<td>25</td>
</tr>
</tbody>
</table>

*Allele range and number of observed alleles from Appendix 1, J.M. Butler (2011) Advanced Topics in Forensic DNA Typing: Methodology


New STR Loci Characterized


Characterization of New miniSTR Loci to Aid Analysis of Degraded DNA

Characterization of 26 MiniSTR Loci for Identification of Human Remains
SNP Typing

- Forensic scale SNP assays
  - ~10 - 50 SNP markers
  - Utility with low amounts of sample (< 1ng)

- High throughput sequencing and DNA/SNP microarrays
  - Thousands to millions of SNPs
  - Greater input amounts of DNA are required
  - Higher level computational methods are required (data storage and analysis)

Multiple approaches and technologies exist for SNP typing

SNP Classifications

- Individual Identification SNPs (IISNPs): SNPs that collectively give very low probabilities of two individuals having the same multi-locus genotype
- Ancestry Informative SNPs (AISNPs): SNPs that collectively give a high probability of an individual’s ancestry being from one part of the world or being derived from two or more areas of the world
- Phenotype Informative SNPs (PISNPs): SNPs that provide a high probability that the individual has particular phenotypes, such as a particular skin color, hair color, eye color, etc.
- Lineage Informative SNPs (LISNPs): Sets of tightly linked SNPs that function as multi-allelic markers that can serve to identify relatives with higher probabilities than simple bi-allelic SNPs

SNP Locus

- D5
- D2
- TPOX
- D16
- D7
- CSF1PO
- D21
- D18

Typing an Aged Blood Stain

Identifier genotyping result from a blood stain aged 15 years stored at room temperature. (stored on 903 paper, Chelex extracted)

The same sample extract as above typed by the 12-plex SNP assay.

11 different samples that gave partial profiles with Identifiler gave full profiles typed with the 12-plex assay.

Individual Identification SNPs

- Use for individual identification of a sample
- Power of Discrimination – how many SNPs are needed to match STRs?
- Can the assay amplify > 30 loci using a small amount of template DNA?
- Use on a degraded sample
- Issues with a mixture

General Criteria

- Low FST (not population specific)
- No linkage disequilibrium between SNPs or CODIS loci
- Amplicon size < 120 bp
- Minimum 30% heterozygosity
- Minimum distance of 100 kb between SNPs and neighboring genes

Challenges of Using IISNPs for Forensic Testing

- U.S. and international databases consist of STR profiles
  - Is there a benefit to changing the DNA typing technology for databanking and routine casework?
- Mixture analysis using genome-wide arrays
  - Detection is possible but interpretation is still ongoing
- SNPs in linkage disequilibrium: match probability calculations?
- Sensitivity
  - Genome-wide arrays require >500 ng DNA
  - Non-biased whole genome amplification is necessary but not here yet
- Cost
  - Approximately $500 per sample for genome-wide arrays
  - Arrays cost the same amount for typing 1,000 SNPs or 1 million SNPs

Ancestry Informative SNPs


Ongoing and active area of research

Screening for sets of autosomal SNPs that will estimate ancestry

Global perspective

Use of mitochondrial SNPs to estimate ancestry

Global Distribution of Y Haplogroups

- Use of Y-Chromosome SNPs to estimate ancestry

Y-SNPs have been primarily typed in world populations

### Biogeographical Ancestry

#### Sample 1
- **Mitochondrial Hg**: H
  - Europe
- **Y Chromosome Hg**: R1b1b2
  - Western Europe
- **Self Identified Ancestry**: Caucasian
- **24 autosomal SNPs**: 94% Caucasian, 6% (other)

#### Sample 2
- **Mitochondrial Hg**: L0a1a
  - Africa
- **Y Chromosome Hg**: E
  - Africa/Southern Europe
- **Self Identified Ancestry**: African American
- **24 autosomal SNPs**: 98% African Amer., 2% (other)

#### Sample 3
- **Mitochondrial Hg**: L3d1
  - Africa
- **Y Chromosome Hg**: R1b1b2
  - Western Europe
- **Self Identified Ancestry**: African American
- **24 autosomal SNPs**: 99% Afr. Amer., 1% (other)

#### Sample 4
- **Mitochondrial Hg**: A2
  - Americas
- **Y Chromosome Hg**: R1b1b2
  - Western Europe
- **Self Identified Ancestry**: Hispanic
- **24 autosomal SNPs**: 89% Hispanic, 7% Asian, 4% Caucasian

#### Sample 5
- **Mitochondrial Hg**: H
  - Europe
- **Y Chromosome Hg**: R1b1b2
  - Western Europe
- **Self Identified Ancestry**: Caucasian
- **24 autosomal SNPs**: 69% Hispanic, 28% Caucasian, 3% Asian

### Summary of Ancestry SNPs

<table>
<thead>
<tr>
<th>n</th>
<th>U.S. sampling</th>
<th>Mitochondrial</th>
<th>Y Chromosome</th>
<th>24 Autosomal</th>
</tr>
</thead>
<tbody>
<tr>
<td>259</td>
<td>African American</td>
<td>93%</td>
<td>69%</td>
<td>98%</td>
</tr>
<tr>
<td>262</td>
<td>Caucasian</td>
<td>97%</td>
<td>84%</td>
<td>81%</td>
</tr>
<tr>
<td>49</td>
<td>Asians</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>140</td>
<td>Hispanics</td>
<td>NA</td>
<td>NA</td>
<td>74%</td>
</tr>
</tbody>
</table>

Success estimating Self Identified Ancestry

The appropriate populations of interest need to be studied. Issues with ‘recent’ admixed populations (e.g. US Hispanics)
Phenotype Informative SNPs

• Predict an observable trait – definitively
  – Eye, hair, and skin color
  – Height, stature

• Some key pigmentation genes have been characterized
  – Wide range of pigmentation in humans
  – Multiple genes involved, complex phenotype

• Gene discovery and characterization is ongoing

• Should not be predictive of disease

Recent Work on PISNPs


Pigmentation Related Genes

<table>
<thead>
<tr>
<th>Eye</th>
<th>Hair</th>
<th>Skin</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASIP</td>
<td>ASIP</td>
<td>ASIP</td>
</tr>
<tr>
<td>DCT</td>
<td>DCT</td>
<td>DCT</td>
</tr>
<tr>
<td>HERC2</td>
<td>HERC2</td>
<td>HERC2</td>
</tr>
<tr>
<td>IRF4</td>
<td>IRF4</td>
<td>IRF4</td>
</tr>
<tr>
<td>KITLG</td>
<td>KITLG</td>
<td>KITLG</td>
</tr>
<tr>
<td>MATP</td>
<td>MATP</td>
<td>MATP</td>
</tr>
<tr>
<td>MC1R</td>
<td>MC1R</td>
<td>MC1R</td>
</tr>
<tr>
<td>OCA2</td>
<td>OCA2</td>
<td>OCA2</td>
</tr>
<tr>
<td>SLC24A4</td>
<td>SLC24A4</td>
<td>SLC24A4</td>
</tr>
<tr>
<td>SLC24A5</td>
<td>SLC24A5</td>
<td>SLC24A5</td>
</tr>
<tr>
<td>TYR</td>
<td>TYR</td>
<td>TYR</td>
</tr>
</tbody>
</table>

Potential for a panel of markers that could predict all three traits

Lineage Informative SNPs

• Much of the LISNP literature focused on Y-chromosome and mitochondrial DNA SNPs

• With genome-wide arrays, autosomal SNP typing for lineage analysis is possible

• Looking for blocks of DNA that have been transmitted unchanged from one generation to the next

• Useful in evolutionary studies and kinship analysis
Different Inheritance Patterns

**Lineage Markers**

- **Autosomal**: (passed on in part, from all ancestors)
- **Y-Chromosome**: (passed on complete, but only by sons)
- **Mitochondrial**: (passed on complete, but only by daughters)

Can infer a specific relationship

Cannot define a specific relationship, only paternal (Y) or maternal (mito) relatedness

---

Kinship Inference with Autosomal STRs or SNPs

<table>
<thead>
<tr>
<th>SNPs</th>
<th>Autosomal (5K-150K)</th>
<th>Autosomal (500K-1 million)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STRs</strong></td>
<td>CODIS 13</td>
<td>15-50</td>
</tr>
</tbody>
</table>

**Degree of relatedness**

- **1**: Parent/child
- **2**: Full siblings
- **3**: Half siblings
- **4**: Grandparent-grandchild
- **5**: First cousins
- **6**: Second cousins
- **7**: Unrelated

Low Template DNA Testing

Some Definitions of Low Template (LT) DNA

- Working with <100-200 pg genomic DNA
- Considered to be data below stochastic threshold level where PCR amplification is not as reliable (determined by each laboratory; typically 150-250 RFUs)
- Enhancing the sensitivity of detection (increasing PCR cycles, PCR product clean-up, increasing CE injection/voltage)
- Having too few copies of DNA template to ensure reliable PCR amplification (allelic or full locus drop-out)
- Can often be the minor component of mixture samples consisting of low level DNA template amounts

Stochastic (Random) Effects with LT-DNA When Combined with Higher Sensitivity Techniques

**Loss of True Signal (False Negative)**

- Heterozygote Peak Imbalance
  - Identifier: 50 pg DNA, 31 cycles
  - Identifiers: 20 pg DNA, 31 cycles

**Gain of False Signal (False Positive)**

- Allelic Drop-out
  - Higher Stutter
    - Identifier: 10 pg DNA, 31 cycles
    - Identifiers: 10 pg DNA, 31 cycles

  - Allelic Drop-in
    - 64% stutter
    - Identifiers: 10 pg DNA, 31 cycles

Suggestions for Optimal Results with LT-DNA

- Typically at least 2 – 3 PCR amplifications from the same DNA extract are performed to obtain consensus profiles
- An allele cannot be scored (considered real) unless it is present at least twice in replicate samples
- Extremely sterile environment is required for PCR setup to avoid contamination from laboratory personnel or other sources
Replicate Testing and Consensus Profiles

- Extract DNA from stain
- Quantify Amount of DNA Present
- Perform 2 or 3 Separate PCR Amplifications
- Interpret Alleles Present
- Develop a Consensus Profile (based on replicate consistent results)

Replicate LT-DNA Test Results from FSS


F' used to designate that allele drop-out of a second allele cannot be discounted when only a single allele is observed

Next Generation Sequencing

Trends and upcoming technologies
- Low-capacity sequencing systems
- Single molecule sequencing
- Bioinformatic analysis of sequence data
- Computer resources

“With the publication of this editorial, it is announced that the journal will not accept any further letters on this issue. However, publications are very welcome describing original scientific research contributing new data on the issues of validation and interpretation of results obtained from forensic genetic typing of DNA of low quantity and/or quality.”
NGS Studies with the Ychr

Xue and Tyler-Smith (2010)

NGS Studies with mtDNA
Heteroplasmic mitochondrial DNA mutations in normal and tumour cells

Because mtDNA template molecules are so numerous in comparison with nuclear DNA template molecules, they are also useful for forensic applications. Previous studies have shown variations in the length of mononucleotide tracts in mtDNA from hair roots compared with blood[28]. Our new results clearly show that heteroplasmies affect the entire mitochondrial genome, are common in normal individuals and vary markedly from tissue to tissue. Thus an individual, and perhaps even a single cell, does not have a single mtDNA genotype. Instead, tissues have a mixture of genotypes, a few of which may be maternally inherited and the remaining ones the result of somatic mutations. This suggests caution in excluding identity on the basis of a single or small number of mismatched alleles when the tissue in evidence (such as sperm) is not the same as the reference tissue of the suspect (such as blood or hair).
Before one can really set out to access to entire mtDNA genome data with relative ease for forensic purposes, one needs careful calibration studies under strict forensic conditions—or might have to wait for another generation."

"The pyrosequencing approach results in poor resolution of homopolymeric sequences, and PCR/sequencing artifacts require a filtering mechanism similar to that for STR stutter and spectral bleed through. In addition, chimeric sequences from jumping PCR must be addressed to make the method operational."
4.6.3. When using CPE/CPI (with no assumptions of number of contributors) to calculate the probability that a randomly selected person would be excluded/included as a contributor to the mixture, loci with alleles below the stochastic threshold may not be used for statistical purposes to support an inclusion.

Available Loci (CPI Stats)

- D8S1179  PI = 0.5927
- D3S1358  PI = 0.9704
- D2S1338  PI = 0.0658
- D5S818   PI = 0.5550
- CPI = 0.021

~2.1% of the Hispanic population can be included in this mixture
Expert Software for Mixture Analysis

Software Limitations…

The “16” allele is considered an artifact (stutter) peak from allele 17 and is ignored.

This marker is included in the CPI calculation.

True Allele Software

“Markov Chain Monte Carlo Testing”

Victim 15.2%

Suspect A 27.2%

Suspect B 57.6%

(1:2:4)

DNA input

D18S51

150 RFUs
Suspect A
LR = 34.2 Quintillion

Million – Billion – Trillion – Quadrillion - Quintrillion

Suspect A
LR = 2.45 Quintillion

Million – Billion – Trillion – Quadrillion - Quintrillion

Benefits of Increased Information

Manual Calculation (CPI)  Mixture Software (CPI)
1 in 2.1% included  1 in 2.6 million included

True Allele Software (LR)
2.45 Quintrillion and 34.2 Quintrillion

NIST Applied Genetics Group

Group Leader

John Butler  Marcia Holden  Margaret Kline  Pete Vallone  Mike Coble

Ross Haynes  Becky Hill  Erica Butts  O’Connor  Kevin Kiesler