Forensic Mitochondrial DNA analysis, phylogenetics, and the potential for mt-Whole Genome Sequencing

Mark R. Wilson, Ph.D.
Director
Forensic Science Program
Western Carolina University
Cullowhee, NC
November 14, 2008
13 CODIS Core STR Loci

- TPOX
- D3S1358
- D2S1338
- D5S818
- FGA
- CSF1PO
- D7S820
- D8S1179
- TH01
- VWA

Sex-typing
- Penta E
- D19S433
- D13S317
- D16S539
- D18S51
- D21S11
- Penta D

AMEL
Short Tandem Repeats (STR)

Tetranucleotide Repeat Units

Allele

5

AATG AATG AATG AATG AATG

6

AATG AATG AATG AATG AATG AATG AATG

7

AATG AATG AATG AATG AATG AATG AATG AATG

Flanking regions of unique sequence
Short Tandem Repeat (STR) Typing
Product Rule

The frequency of a multi-locus STR profile is the product of the genotype frequencies at the individual loci.

\[ f_{\text{locus}_1} \times f_{\text{locus}_2} \times f_{\text{locus}_n} = f_{\text{combined}} \]
13 CODIS loci typically yield extraordinarily small match probabilities

e.g., 0.000000000000000000154

OR

1 in 60,000,000,000,000,000,000 persons
Human mitochondrial DNA (mtDNA)
Mitochondrial DNA

- Inherited from mother
- Can use maternal relative for reference sample
- Good discrimination factor
- Resists degradation

Nuclear DNA

- Inherited from mother and father
- High discrimination factor up to identity
- Degrades faster than mtDNA

Inherited from mother
Control region (D-loop)

1/16,569

cyt b
ND5
ND6
ND4
ND4L
ND3
COIII
ATP6
ATP8
COII
COI

16S rRNA
12S rRNA
1/16,569

9-bp deletion

16024 16365 1 73 340

HV1
HV2

Control region (D-loop)

22 tRNAs
2 rRNAs
13 genes

CONTROL REGION PRIMERS

• 8 FOR AMPLIFICATION
• 10 FOR SEQUENCING
CONTROL REGION AMPLIFICATION PRIMERS
CONTROL REGION AMPLIFICATION PRIMERS
CONTROL REGION AMPLIFICATION PRIMERS
To infer rarity, compare case DNA sequence to database of DNA sequences
Phylogenetics:

The use of genetic data to infer or estimate evolutionary relationships of entities related by descent over time.
Phylogenetic Methods

• Distance (clustering, phenetic, i.e. neighbor joining).
• Parsimony (shortest tree (MPT), requiring the fewest assumptions or lowest score).
• Maximum likelihood (probability of a tree given the data set and individual character change probabilities).
Utility of Phylogenetics in Forensics

- Database validation through Haplogroup analysis.
- SNP identification.
- Relative mutation rate interpretation ("hot spot" identification.)
All methods begin with a data matrix:
# Dissimilarity Matrix

<table>
<thead>
<tr>
<th></th>
<th>UK</th>
<th>JAP</th>
<th>AFAM</th>
<th>AFCR</th>
<th>AFRORI</th>
<th>AFR</th>
<th>ASNORI</th>
<th>AUS</th>
<th>CAUORI</th>
<th>HISPORI</th>
<th>FR</th>
<th>KOR</th>
<th>USC</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK</td>
<td>0.11</td>
<td>0.20</td>
<td>0.19</td>
<td>0.20</td>
<td>0.20</td>
<td>0.13</td>
<td>0.03</td>
<td>0.02</td>
<td>0.19</td>
<td>0.05</td>
<td>0.17</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>JAP</td>
<td>0.11</td>
<td>0.16</td>
<td>0.12</td>
<td>0.15</td>
<td>0.16</td>
<td>0.03</td>
<td>0.09</td>
<td>0.10</td>
<td>0.17</td>
<td>0.13</td>
<td>0.07</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>AFAM</td>
<td>0.20</td>
<td>0.16</td>
<td>0.06</td>
<td>0.03</td>
<td>0.06</td>
<td>0.15</td>
<td>0.19</td>
<td>0.21</td>
<td>0.07</td>
<td>0.23</td>
<td>0.16</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>AFCR</td>
<td>0.19</td>
<td>0.12</td>
<td>0.06</td>
<td>0.04</td>
<td>0.07</td>
<td>0.11</td>
<td>0.18</td>
<td>0.20</td>
<td>0.08</td>
<td>0.22</td>
<td>0.11</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>AFR</td>
<td>0.20</td>
<td>0.15</td>
<td>0.02</td>
<td>0.04</td>
<td>0.05</td>
<td>0.14</td>
<td>0.18</td>
<td>0.21</td>
<td>0.06</td>
<td>0.23</td>
<td>0.14</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>ASNORI</td>
<td>0.13</td>
<td>0.03</td>
<td>0.15</td>
<td>0.11</td>
<td>0.14</td>
<td>0.16</td>
<td>0.12</td>
<td>0.13</td>
<td>0.16</td>
<td>0.16</td>
<td>0.04</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>AUS</td>
<td>0.03</td>
<td>0.09</td>
<td>0.19</td>
<td>0.18</td>
<td>0.18</td>
<td>0.19</td>
<td>0.12</td>
<td>0.03</td>
<td>0.19</td>
<td>0.05</td>
<td>0.15</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>CAUORI</td>
<td>0.02</td>
<td>0.10</td>
<td>0.21</td>
<td>0.20</td>
<td>0.21</td>
<td>0.13</td>
<td>0.03</td>
<td>0.21</td>
<td>0.04</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>HISP</td>
<td>0.19</td>
<td>0.17</td>
<td>0.07</td>
<td>0.08</td>
<td>0.06</td>
<td>0.06</td>
<td>0.16</td>
<td>0.19</td>
<td>0.21</td>
<td>0.23</td>
<td>0.17</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>FR</td>
<td>0.05</td>
<td>0.13</td>
<td>0.23</td>
<td>0.22</td>
<td>0.23</td>
<td>0.23</td>
<td>0.16</td>
<td>0.05</td>
<td>0.04</td>
<td>0.23</td>
<td>0.19</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>KOR</td>
<td>0.17</td>
<td>0.07</td>
<td>0.16</td>
<td>0.11</td>
<td>0.14</td>
<td>0.04</td>
<td>0.15</td>
<td>0.17</td>
<td>0.17</td>
<td>0.19</td>
<td>0.16</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>USC</td>
<td>0.03</td>
<td>0.10</td>
<td>0.22</td>
<td>0.20</td>
<td>0.21</td>
<td>0.22</td>
<td>0.13</td>
<td>0.04</td>
<td>0.01</td>
<td>0.21</td>
<td>0.05</td>
<td>0.16</td>
<td></td>
</tr>
</tbody>
</table>
Parsimony Analysis

• Search for the tree or cladogram that explains the data with the fewest assumptions or changes (MPT).
• The actual character changes (nucleotide sites) are mapped on the tree, so information about the characters themselves is retained.
• Information is obtained about the behavior of each character.
CHARACTERS

<table>
<thead>
<tr>
<th>Taxa</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>B</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>C</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>D</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
</tbody>
</table>

1. A -> B -> C -> D
2. A -> C -> D -> B
3. A -> D -> C -> B
CHARACTERS

<table>
<thead>
<tr>
<th>Taxa</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>B</td>
<td>•</td>
<td>o</td>
<td>o</td>
<td>•</td>
<td>o</td>
<td>•</td>
</tr>
<tr>
<td>C</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>o</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>D</td>
<td>•</td>
<td>o</td>
<td>•</td>
<td>o</td>
<td>•</td>
<td>o</td>
</tr>
</tbody>
</table>

1. A, B, C, D
2. A, C, D, B
3. A, B, C, D
### CHARACTERS

<table>
<thead>
<tr>
<th>Taxa</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>•</td>
<td>0</td>
<td>0</td>
<td>•</td>
<td>0</td>
<td>•</td>
</tr>
<tr>
<td>C</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>0</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>D</td>
<td>•</td>
<td>0</td>
<td>•</td>
<td>0</td>
<td>•</td>
<td>0</td>
</tr>
</tbody>
</table>

**7 steps**

**9 steps**

**8 steps**
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>g</td>
<td>d</td>
<td>e</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>h</td>
<td>i</td>
<td>f</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>j</td>
<td>k</td>
<td>l</td>
<td></td>
</tr>
</tbody>
</table>
Fast changing sites provide resolution for closely related organisms.
Simplified mtDNA lineages

All branch markers have the rCRS nucleotide on the left side of the position number. The defining polymorphism for each haplogroup as it diverges from the rCRS is on the right side of the position number. In some cases arrows are used to clarify directionality.
Human mtDNA Migrations

http://www.gen.emory.edu/MITOMAP/WorldMigrations.pdf
Copyright © Emory University, Atlanta, GA 30322

Mutation rate = 2.2 - 2.9 % / MYR
Time estimates are YBP
The structure of human mtDNA variation
Whole Genome mtDNA analysis

• Currently use less than 4% of genome.
• Need methods to rapidly and efficiently generate DNA sequence information at high coverage depth.
• Database is needed to support rarity estimates.
Mitochondrial genome variation and the origin of modern humans

Max Ingman*, Henrik Kaessmann†, Svante Pääbo‡ & Ulf Gyllensten*  

*Department of Genetics and Pathology, Section of Medical Genetics, Redbrick Laboratory, University of Uppsala, S-751 85 Uppsala, Sweden  
†Max Planck Institute for Evolutionary Anthropology, Inselstrasse 22, D-04103 Leipzig, Germany

The analysis of mitochondrial DNA (mtDNA) has been a potent tool in our understanding of human evolution, owing to characteristics such as high copy number, apparent lack of recombination, high substitution rate and maternal mode of inheritance. However, almost all studies of human evolution based on mtDNA sequencing have been confined to the control region, which constitutes less than 7% of the mitochondrial genome.

Figure 2 Neighbour-joining phylogram based on complete mtDNA genome sequences (excluding the D-loop). Data was constructed using PAUP*4.0 Beta (Sinauer Associates) and bootstrapped with 1,000 replicates (bootstrap values shown on nodes). The population origin of the individual is given at the talus. Branches have been colour coded as in Fig. 4. Individuals of African descent are found below the dashed line and non-Africans above. The node marked with an asterisk refers to the MRCA of the youngest clad containing both African and non-African individuals.
Sequence mixture due to heteroplasmy (intra-individual variation)
But...fluorescent noise complicates Sanger dideoxy sequence analysis
Roche/454 Sequencing by Synthesis

Genome Sequencer
FLX System
PCR template generation for Roche/454 sequencing

Figure 1: Schematic representation of an amplification product generated by the Amplicon Library preparation procedure. The composite primers each comprise a 20-25 bp target-specific sequence region at their 3' end and a 19 bp region (Primer A or Primer B) that will be used in subsequent clonal amplification and sequencing reactions, at their 5' end.
DNA extraction
Multiple Displacement Amplification (MDA)

Generates additional template molecules for PCR amplification
GS FLX Technology
Sequencing-by-synthesis

Simultaneous sequencing of the entire genome in hundreds of thousands of picoliter-size wells.

Pyrophosphate signal generation upon complimentary nucleotide incorporation — dark otherwise.

• Polymerase adds nucleotide (dATP)
• Pyrophosphate is released (PPi)
• Sulfurylase creates ATP from PPi
• Luciferase hydrolyses ATP and uses luciferin to make light

DNA capture bead containing millions of copies of a single clonal fragment

www.roche-applied-science.com
Amplicon Variant Analysis (AVA)  \( A \rightarrow G \)

Kozal et al.
• Commercial ABI kit will amplify entire human mtDNA genome with 46 sets of M13 tagged primers.
Multiplex Amplification Detection - Goal is even efficiency across markers
GS FLX Technology
Sequencing instrument
The End

Questions?