Sequencing and Standards for Characterization of the Mitochondrial Genome

Kevin Kiesler, M.S.
Research Biologist - NIST Applied Genetics Group
Forensics @NIST
November 7, 2018
Outline

• Sanger sequencing and the “next generation”

• Reference materials and mtDNA sequencing

• Population sequencing project
NIST Disclaimer

- **Disclaimer** - Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Commerce or the Department of Justice. 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 endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

- All work presented has been reviewed and approved by the NIST Human Subjects Protections Office.
Mitochondrial DNA Sequencing - Introduction

• Mitochondrial DNA advantages
  • High copy number
    • Very small quantity of evidence required
  • Single nucleotide variants (SNPs or SNVs) = profile
    • Small region of DNA analyzed
    • Short PCR amplicons
  • Applicable to challenging/degraded/limited samples
    • Mass disaster, missing persons
      • Example: World Trade Center victim identification – 44,000 mtDNA profiles

• Disadvantage
  • Low power of discrimination using “Control Region”
    • 26,127 CR genotypes in EMPOP database for matching
  • Can be improved with whole mtGenome analysis

Mitochondrial Genome Diagram
Image credit: Emmanuel Douzery
Sanger Sequencing Workflow - mtGenome

Lyons et al. BMC Genomics 2013, 14:881
Next Generation Sequencing Workflow

1. Extract DNA
2. Quantitate DNA
3. PCR Template for Sequencing
4. Library Preparation
5. QC & Quantify
6. Sequence
7. Data Analysis
8. Variant List

**Manual Procedure**

- Step 3: Amplify mtGenome with long PCR (2 reactions)
- Step 4: Library preparation
- Step 6: Sequence on MiSeq

**Image credit:**
- www.Illumina.com
- www.thermofisher.com

Fendt et al. BMC Genomics 2009, 10:139
Next Generation Sequencing Workflow

**Automated Procedure**

(Appplied Biosystems Precision ID mtDNA Whole Genome Panel)

1. Extract DNA
2. Quantitate DNA
3. Automated Library Preparation
4. Library Quantitation (manual)
5. Automated Template Preparation
6. Sequencing
7. Data Analysis
8. Variant List

Automated Reaction Setup

Step 3: Amplify mtGenome with 162 primer pairs in 2 reactions eight samples at a time on Ion Chef

Step 4: Prepare templated beads

Step 5: Sequence on Ion Torrent S5 or PGM

Step 6: Align NGS reads to reference genome (rCRS)

Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8

Human mtDNA 16,569 nts

Image credit: www.wright.edu/~oleg.paliy/NGS.html

Image credit: www.thermofisher.com
Sequencing Instruments at NIST

- Multiple platforms
- Orthogonal measurements
- Characterize Standard Reference Materials

- Applied Biosystems SOLiD
- Illumina MiSeq FGx
- Ion Torrent PGM
- Ion Torrent S5
NIST Mitochondrial Sequencing SRMs

• SRM 2392
  • Three components
    • Component A: DNA from cell line CHR
    • Component B: DNA from cell line 9947A
    • Component C: Cloned fragment from HV1 region of CHR containing C-stretch

• SRM 2392-I
  • One component
    • DNA from cell line HL60

• Characterized with Sanger methods
  • Released in 2001
Sanger-Based Sequence Agrees With NGS Values

Table 2. Certified Human mtDNA Sequence Differences from the Revised Cambridge Reference Sequence for SRM 2392 Component GM09947A

<table>
<thead>
<tr>
<th>Site</th>
<th>rCRS</th>
<th>GM09947A</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>93</td>
<td>A</td>
<td>G</td>
<td>X</td>
</tr>
<tr>
<td>195</td>
<td>T</td>
<td>C</td>
<td>X</td>
</tr>
<tr>
<td>214</td>
<td>A</td>
<td>G</td>
<td>X</td>
</tr>
<tr>
<td>263</td>
<td>A</td>
<td>G</td>
<td>X</td>
</tr>
<tr>
<td>309.1</td>
<td>C</td>
<td>X</td>
<td>insertion</td>
</tr>
<tr>
<td>309.2</td>
<td>C</td>
<td>X</td>
<td>insertion</td>
</tr>
<tr>
<td>315.1</td>
<td>C</td>
<td>X</td>
<td>insertion</td>
</tr>
<tr>
<td>750</td>
<td>A</td>
<td>G</td>
<td>X</td>
</tr>
<tr>
<td>1438</td>
<td>A</td>
<td>G</td>
<td>X</td>
</tr>
<tr>
<td>3107</td>
<td>ACT</td>
<td>C</td>
<td>X</td>
</tr>
<tr>
<td>4135</td>
<td>T</td>
<td>C</td>
<td>deletion</td>
</tr>
<tr>
<td>4769</td>
<td>A</td>
<td>G</td>
<td>X</td>
</tr>
<tr>
<td>7645</td>
<td>T</td>
<td>C</td>
<td>X</td>
</tr>
<tr>
<td>7861</td>
<td>T</td>
<td>C</td>
<td>X</td>
</tr>
<tr>
<td>8448</td>
<td>T</td>
<td>C</td>
<td>X</td>
</tr>
<tr>
<td>8860</td>
<td>A</td>
<td>G</td>
<td>X</td>
</tr>
<tr>
<td>9315</td>
<td>T</td>
<td>C</td>
<td>X</td>
</tr>
<tr>
<td>13572</td>
<td>T</td>
<td>C</td>
<td>X</td>
</tr>
<tr>
<td>13759</td>
<td>G</td>
<td>A</td>
<td>X</td>
</tr>
<tr>
<td>15326</td>
<td>A</td>
<td>G</td>
<td>X</td>
</tr>
<tr>
<td>16311</td>
<td>T</td>
<td>C</td>
<td>X</td>
</tr>
<tr>
<td>16519</td>
<td>T</td>
<td>C</td>
<td>X</td>
</tr>
</tbody>
</table>

Previously no software available that correctly handled these C-stretch insertions with forensic nomenclature.
NGS Can Detect Low Level Heteroplasmy

• Level is below what we can reliably see with Sanger methods

SRM 2392 Component A (CHR)
Position 1,393 R (17% A)

SRM 2392-I (HL-60)
Position 2,445 Y (7% C)
Multiple Orthogonal Measurements

- Great approach for certifying reference materials!

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Component A</th>
<th>Component B</th>
<th>SRM 2392</th>
<th>SRM 2392-I</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGM Edge</td>
<td>26.8</td>
<td>21.2</td>
<td>72.7</td>
<td>10.6</td>
</tr>
<tr>
<td>PGM NIST 1</td>
<td>24.3</td>
<td>15.6</td>
<td>45.0</td>
<td>7.5</td>
</tr>
<tr>
<td>PGM NIST 2</td>
<td>25.0</td>
<td>17.5</td>
<td>65.7</td>
<td>7.7</td>
</tr>
<tr>
<td>PGM NIST 3</td>
<td>29.7</td>
<td>16.5</td>
<td>59.4</td>
<td>7.7</td>
</tr>
<tr>
<td>PGM NIST HiQ</td>
<td>33.2</td>
<td>15.2</td>
<td>77.7</td>
<td>7.7</td>
</tr>
<tr>
<td>MiSeq Edge</td>
<td>33.0</td>
<td>19.0</td>
<td>87.4</td>
<td>10.7</td>
</tr>
<tr>
<td>MiSeq NIST</td>
<td>31.6</td>
<td>17.9</td>
<td>88.4</td>
<td>9.1</td>
</tr>
<tr>
<td>HiSeq BC</td>
<td>30.6</td>
<td>16.9</td>
<td>88.3</td>
<td>7.4</td>
</tr>
<tr>
<td>SOLiD NIST</td>
<td>29.0</td>
<td>16.7</td>
<td>87.3</td>
<td>7.3</td>
</tr>
<tr>
<td>Average</td>
<td>29.2</td>
<td>17.4</td>
<td>74.6</td>
<td>8.4</td>
</tr>
<tr>
<td>St. Dev.</td>
<td>3.1</td>
<td>1.7</td>
<td>14.5</td>
<td>1.3</td>
</tr>
</tbody>
</table>
Multiple Orthogonal Measurements

• Great approach for characterizing reference materials!

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Component A</th>
<th>Component B</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGM Edge</td>
<td>26.8</td>
<td>21.2</td>
</tr>
<tr>
<td>PGM NIST 1</td>
<td>24.3</td>
<td>15.6</td>
</tr>
<tr>
<td>PGM NIST 2</td>
<td>25.0</td>
<td>17.5</td>
</tr>
<tr>
<td>PGM NIST 3</td>
<td>29.7</td>
<td>16.5</td>
</tr>
<tr>
<td>PGM NIST HiQ</td>
<td>33.2</td>
<td>15.2</td>
</tr>
<tr>
<td>MiSeq Edge</td>
<td>33.0</td>
<td>19.0</td>
</tr>
<tr>
<td>MiSeq NIST</td>
<td>31.6</td>
<td>17.9</td>
</tr>
<tr>
<td>HiSeq BC</td>
<td>30.6</td>
<td>16.9</td>
</tr>
<tr>
<td>SOLiD NIST</td>
<td>29.0</td>
<td>16.7</td>
</tr>
<tr>
<td>Average</td>
<td>29.2</td>
<td>17.4</td>
</tr>
<tr>
<td>St. Dev.</td>
<td>3.1</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Average: 29.2
St. Dev.: 3.1

1. St. Dev. 87%
2. DEL: 222

It can also educate you about your technology.
Multiple Orthogonal Measurements

- Great approach for characterizing reference materials! SOFTWARE
Conclusions

• Reference materials
  • Can identify technical limitations/bias
    • Often need multiple measurements
    • Orthogonal techniques
  • Help to select best procedures
Population Scale Sequencing
Project Goals

• Submit forensic-quality whole mtGenome data to EMPOP
  • Database used for match statistics
  • Current version (V4, Release 11)
    • n = 26,127 control region sequences
    • n = \textbf{256 whole genome} sequences
• NIST population samples (n > 1,000)
  • African American, Asian, Caucasian, Hispanic
• Sequencing plan
  • Start with Caucasian population
  • \approx 440 mtGenomes
Project Plans

• What **instrument** do we use?
• What **protocol/chemistry** do we use?
• What **analysis** procedure do we use?
  • Software, data review, etc.
Project Planning: Instrument Selection

• Considerations
  • Cost
  • Time/labor
    • Automation

<table>
<thead>
<tr>
<th>Method</th>
<th>Cost Per 96 Samples (Approximate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S5XL Automated Library Prep</td>
<td>$25,000</td>
</tr>
<tr>
<td>S5XL Manual Library Prep</td>
<td>$15,000</td>
</tr>
<tr>
<td>MiSeq ILMN Method</td>
<td>$5,000</td>
</tr>
<tr>
<td>MiSeq AFDIL Method</td>
<td>$5,000</td>
</tr>
</tbody>
</table>
Project Planning: Protocol Selection

• **Options**
  - Illumina mtGenome procedure
  - Long PCR primers developed by Dr. Mark Wilson’s lab
  - TaKaRa LA Taq
  - Illumina Nextera XT library preparation
  - Illumina MiSeq v2 2x150 cartridge (per protocol)

• mtGenome procedure used by Armed Forces DNA Identification Lab (AFDIL)
  - Long PCR primers from Fendt *et al.*, BMC Genomics 2009, 10:139
  - TaKaRa LA Taq (GC Buffer & BSA)
  - Kapa HyperPlus Library Kit
  - Illumina V3 2x300 cartridge
More Consistent Coverage Depth with Kapa HyperPlus
AFDIL method allows higher multiplexing with less likelihood of dropout sites
## Haplogroup Estimation from EMPOP

- **No surprise haplogroups**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Missing Mutations</th>
<th>Private Mutations</th>
<th>Haplogroup</th>
<th>Continent</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT38086</td>
<td>none</td>
<td>T16189C</td>
<td>H1c21</td>
<td>Europe (H)</td>
</tr>
<tr>
<td>GT38087</td>
<td>none</td>
<td>none</td>
<td>T2b6b</td>
<td>Europe</td>
</tr>
<tr>
<td>GT38089</td>
<td>none</td>
<td>T16189C</td>
<td>J1c8a</td>
<td>Europe</td>
</tr>
<tr>
<td>GT38091</td>
<td>none</td>
<td>C10933T A15467G</td>
<td>V</td>
<td>Europe</td>
</tr>
<tr>
<td>GT38092</td>
<td>none</td>
<td>C198T A9327G A13801G T15670C</td>
<td>H5e1a1</td>
<td>Europe (H)</td>
</tr>
<tr>
<td>GT38093</td>
<td>none</td>
<td>none</td>
<td>H1o</td>
<td>Europe</td>
</tr>
<tr>
<td>GT38094</td>
<td>none</td>
<td>A11252G</td>
<td>H65</td>
<td>Europe (H)</td>
</tr>
<tr>
<td>GT38095</td>
<td>-573.1C</td>
<td>G709A C9727T</td>
<td>I1a1b</td>
<td>Europe (I1a1)</td>
</tr>
<tr>
<td>GT38097</td>
<td>none</td>
<td>G8027A G15301A</td>
<td>K1a1b1</td>
<td>Europe</td>
</tr>
<tr>
<td>GT38098</td>
<td>A16183C -309.1C -309.2C</td>
<td>C16111T T152C</td>
<td>H1b1</td>
<td>Europe</td>
</tr>
<tr>
<td>GT38100</td>
<td>none</td>
<td>-309.1C G4655A</td>
<td>H1</td>
<td>Europe</td>
</tr>
<tr>
<td>GT38106</td>
<td>none</td>
<td>T2416C A8817G</td>
<td>H2a2a</td>
<td>Europe</td>
</tr>
<tr>
<td>GT38107</td>
<td>none</td>
<td>C3388A C8788T</td>
<td>U5a1a1</td>
<td>Europe</td>
</tr>
<tr>
<td>GT38108</td>
<td>none</td>
<td>T4373C T15313C</td>
<td>H1+16189</td>
<td>Europe</td>
</tr>
<tr>
<td>JA44327</td>
<td>-573.1C</td>
<td>G16474T</td>
<td>I2</td>
<td>Europe (I)</td>
</tr>
<tr>
<td>JM28315</td>
<td>-309.1C</td>
<td>none</td>
<td>K2a6</td>
<td>Europe</td>
</tr>
<tr>
<td>JT52345</td>
<td>none</td>
<td>A16138G A73G</td>
<td>H5a1c1a</td>
<td>Europe</td>
</tr>
<tr>
<td>JT52346</td>
<td>none</td>
<td>none</td>
<td>H5g</td>
<td>Europe</td>
</tr>
<tr>
<td>MT97121</td>
<td>none</td>
<td>T16093C G7762A</td>
<td>J1c2</td>
<td>Europe</td>
</tr>
<tr>
<td>MT97122</td>
<td>none</td>
<td>none</td>
<td>T2b2b</td>
<td>Europe</td>
</tr>
<tr>
<td>MT97123</td>
<td>none</td>
<td>A16158G</td>
<td>T1a1</td>
<td>Europe (T1a)</td>
</tr>
<tr>
<td>MT97124</td>
<td>T16093C</td>
<td>-309.1C -524.3A -524.4C</td>
<td>K1a</td>
<td>Europe</td>
</tr>
<tr>
<td>MT97125</td>
<td>none</td>
<td>T152C</td>
<td>U5a2d1a</td>
<td>Europe (U5a)</td>
</tr>
</tbody>
</table>
Informatics

• Forensic mtDNA nomenclature is challenging!
• Commercial software now available
  • Softgenetics GeneMarker HTS
    • Compatible with forensic nomenclature
    • EMPOP formatted report
  • CLC Genomics Workbench
    • AFDIL / Qiagen – developed AQME Tool
  • ThermoFisher Scientific
    • Converge mtDNA Analysis (released October 2018)
Conclusions

• Mitochondrial Genome Protocol
  • Cost and data quality directed decision process
  • Selected AFDIL-developed procedure for reference-quality samples
    • Even coverage
    • Higher multiplexing
  • Degraded samples will need a different procedure
  • Analysis method must be high-throughput
    • High accuracy required for EMPOP submission
Thank You!
Questions?

Contact info:
Kevin.Kiesler@NIST.gov

• Funding
  • NIST Special Programs Office: Forensic DNA
  • FBI Biometrics Center of Excellence: Forensic DNA Typing as a Biometric tool.

√ Acknowledgements √

Armed Forces DNA Identification Laboratory (AFDIL)
Kim Andreaggi
Charla Marshall

NIST Applied Genetics Group NGS Team
Dr. Peter Vallone, Group leader
Lisa Borsuk
Sarah Riman
Becky Steffen
Katherine Gettings
Q&A Session 10:55 – 11:05
Break 11:05 – 11:20
• Digital & Trace Tour signups please report to registration booth @ 11:00