Y-Chromosome and Mitochondrial DNA Analysis

# mitochondrial DNA

NEAFS 2006 Workshop
Rye Brook, NY
November 1, 2006

Northeastern Association of Forensic Scientists

**Dr. John M. Butler**
**Dr. Michael D. Coble**

john.butler@nist.gov
Michael.Coble@afip.osd.mil

## SNPs for Forensic Discrimination

- 59 SNPs – that met our criteria (neutral, shared, non-redundant).
  - 49 – Protein coding (silent)
    - 8 – Control Region (outside HV1/2)
    - 1 – Non-coding spacer region
    - 1 – 16S rRNA*

* 3010 G-A

## SNPs for Forensic Discrimination

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| 477 | 477 | 72 | 482 | 4808 | 64 | 3826 | 64 |
| 3010 | 3010 | 513 | 5198 | 5147 | 4745 | 3834 | 4688 |
| 4580 | 3915 | 4580 | 6260 | 9380 | 10211 | 4688 | 11377 |
| 4793 | 5004 | 5250 | 9548 | 9899 | 10394 | 6293 | 12795 |
| 5004 | 6776 | 11719 | 9635 | 11914 | 10685 | 7891 | 13293 |
| 7028 | 8592 | 12438 | 11485 | 15067 | 11377 | 11533 | 14305 |
| 7202 | 10394 | 12810 | 11914 | 16519 | 14470 | 12007 | 16519 |
| 10211 | 10754 | 14770 | 15355 | | 14560 | 12795 | |
| 12858 | 11864 | 15833 | 15884 | | 16390 | 15043 | |
| 14470 | 15340 | 15884 | 16368 | | 14869 | 16390 | |
| 16519 | 16519 | 16519 | | | | 16519 | |
| | | | | | | | |
| H1 | H2 H3 H6 | V1 H5 | J1 J2 K2 K3 | J4 T2 T3 H4 | V1 H1 H2 H3 | J1 J3 T1 | K1 |

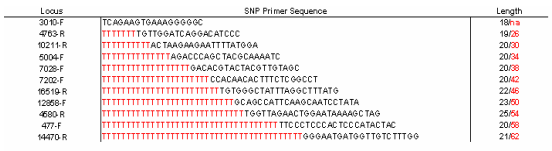
## SNPs for Forensic Discrimination

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| 477 | 477 | 72 | 482 | 4808 | 64 | 3826 | 64 |
| 3010 | 3010 | 513 | 5198 | 5147 | 4745 | 3834 | 4688 |
| 4580 | 3915 | 4580 | 6260 | 9380 | 10211 | 4688 | 11377 |
| 4793 | 5004 | 5250 | 9548 | 9899 | 10394 | 6293 | 12795 |
| 5004 | 6776 | 11719 | 9635 | 11914 | 10685 | 7891 | 13293 |
| 7028 | 8592 | 12438 | 11485 | 15067 | 11377 | 11533 | 14305 |
| 7202 | 10394 | 12810 | 11914 | 16519 | 14470 | 12007 | 16519 |
| 10211 | 10754 | 14770 | 15355 | | 14560 | 12795 | |
| 12858 | 11864 | 15833 | 15884 | | 16390 | 15043 | |
| 14470 | 15340 | 15884 | 16368 | | 14869 | 16390 | |
| 16519 | 16519 | 16519 | | | | 16519 | |
| | | | | | | | |
| H1 | H2 H3 H6 | V1 H5 | J1 J2 K2 K3 | J4 T2 T3 H4 | V1 H1 H2 H3 | J1 J3 T1 | K1 |

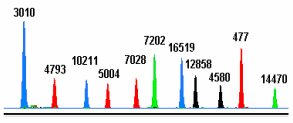
## SNPs for Forensic Discrimination

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| 477 | 477 | 72 | 482 | 4808 | 64 | 3826 | 64 |
| 3010 | 3010 | 513 | 5198 | 5147 | 4745 | 3834 | 4688 |
| 4580 | 3915 | 4580 | 6260 | 9380 | 10211 | 4688 | 11377 |
| 4793 | 5004 | 5250 | 9548 | 9899 | 10394 | 6293 | 12795 |
| 5004 | 6776 | 11719 | 9635 | 11914 | 10685 | 7891 | 13293 |
| 7028 | 8592 | 12438 | 11485 | 15067 | 11377 | 11533 | 14305 |
| 7202 | 10394 | 12810 | 11914 | 16519 | 14470 | 12007 | 16519 |
| 10211 | 10754 | 14770 | 15355 | | 14560 | 12795 | |
| 12858 | 11864 | 15833 | 15884 | | 16390 | 15043 | |
| 14470 | 15340 | 15884 | 16368 | | 14869 | 16390 | |
| 16519 | 16519 | 16519 | | | | 16519 | |
| | | | | | | | |
| H1 | H2 H3 H6 | V1 H5 | J1 J2 K2 K3 | J4 T2 T3 H4 | V1 H1 H2 H3 | J1 J3 T1 | K1 |

## The SNaPShot™ Platform




Rebecca Just
Dr. Peter Vallone




Vallone *et al.* (2004) *IJLM* **118:** 147- 157.

## SNPs for Forensic Discrimination

18 common HV1/HV2 types, 241 individuals

| +8 Multiplexes (59 SNPs) | +8 Multiplexes (with AC indel) |
|---|---|
| 105 types (55 "unique") | 112 types (64 "unique") |

6-fold improvement!

## The Nature of the SNPs

- Are resolving SNPs **slow and rare?** Did these SNPs arise once during the evolution of a haplogroup?

OR…

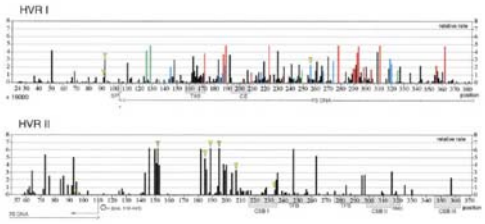
- Are resolving SNPs "universally" **fast hot spots**, useful for all haplogroups (L, M, N)?

## SNPs for Forensic Discrimination

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| 477 | 477 | 72 | 482 | 4808 | 64 | 3826 | 64 |
| 3010 | 3010 | 513 | 5198 | 5147 | 4745 | 3834 | 4688 |
| 4580 | 3915 | 4580 | 6260 | 9380 | 10211 | 4688 | 11377 |
| 4793 | 5004 | 5250 | 9548 | 9899 | 10394 | 6293 | 12795 |
| 5004 | 6776 | 11719 | 9635 | 11914 | 10685 | 7891 | 13293 |
| 7028 | 8592 | 12438 | 11485 | 15067 | 11377 | 11533 | 14305 |
| 7202 | 10394 | 12810 | 11914 | 16519 | 14470 | 12007 | 16519 |
| 10211 | 10754 | 14770 | 15355 | | 14560 | 12795 | |
| 12858 | 11864 | 15833 | 15884 | | 16390 | 15043 | |
| 14470 | 15340 | 15884 | 16368 | | 14869 | 16390 | |
| 16519 | 16519 | 16519 | | | | 16519 | |
| | | | | | | | |
| H1 | H2 H3 H6 | V1 H5 | J1 J2 K2 K3 | J4 T2 T3 H4 | V1 H1 H2 H3 | J1 J3 T1 | K1 |

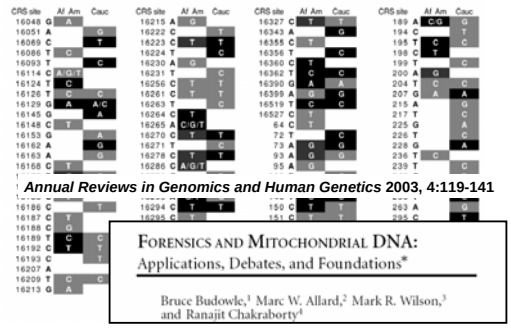
## Mutation Rate Analysis in the mtDNA Control Region

Mutation rate heterogeneity – the variation of mutation rates among sites.




Meyer *et al.* (1999) *Genetics*

## Mutation Rate Analysis in the mtDNA Control Region



***Annual Reviews in Genomics and Human Genetics* 2003, 4:119-141**

FORENSICS AND MITOCHONDRIAL DNA:
Applications, Debates, and Foundations*

Bruce Budowle,[1] Marc W. Allard,[2] Mark R. Wilson,[3] and Ranajit Chakraborty[4]

## Mutation Rate Analysis in the mtDNA Coding Region

Previous Assumptions (I)

Adam Eyre-Walker *et al.* (1999) *Proc. R. Soc. Lond B*. Using partial DNA sequences of the human mtDNA genome (filled with errors), this group observed a significant amount of recurrent mutations (homoplasy) in their data.

Conclusion – **Recombination!** (between paternal and maternal mtDNA)

## Mutation Rate Analysis in the mtDNA Coding Region

- Eyre-Walker *et al.* assume mutation rate ***Homo***geneity…
- "There is no evidence of variation in the mutation rate."
- (Mostly discredited for their poor data choice and method of calculating LD)

## Mutation Rate Analysis in the mtDNA Coding Region

Previous Assumptions (II)

*Am. J. Hum. Genet.* 70:1152–1171, 2002

**Reduced-Median-Network Analysis of Complete Mitochondrial DNA Coding-Region Sequences for the Major African, Asian, and European Haplogroups**

Corinna Herrnstadt,[1] Joanna L. Elson,[2] Eoin Fahy,[1] Gwen Preston,[1] Douglass M. Turnbull,[3] Christen Anderson,[1] Soumitra S. Ghosh,[1] Jerrold M. Olefsky,[2] M. Flint Beal,[4,*] Robert E. Davis,[1,4] and Neil Howell[1,5]

## Mutation Rate Analysis in the mtDNA Coding Region

Table 2

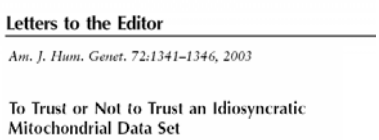
Polymorphisms That Are Associated with More than One Haplogroup

| Nucleotide Position[a] | No. of Polymorphisms in Haplogroup[b] | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A (25) | B (18) | C (13) | D (9) | E (3) | H (226) | I (14) | J (33) | K (47) | L1 (13) | L2 (23) | L3 (20) | M (1) | T (46) | U (42) | V (8) | W (8) | X (11) |
| 593 | | | | 1 | | | | 1 | 2 | | | | | | | | | |
| 709 | | 2 | | | | 1 | | | 9 | 7 | | 1 | | 46 | 1 | | 8 | |
| 750 | 25 | 18 | 13 | 9 | 3 | 218 | 14 | 33 | 47 | 13 | 23 | 17 | 1 | 46 | 42 | 8 | 8 | 11 |
| 769 | | | | | | | | | | 13 | 23 | | | | | | | |
| 930 | 1 | | | | | | | | | | 1 | | | 16 | | | | |
| 1018 | | | | | | | | | | 13 | 23 | | | | | | | |
| 1438 | 25 | 18 | 13 | 9 | 3 | 208 | 12 | 33 | 47 | 6 | 23 | 20 | 1 | 46 | 42 | 8 | 8 | 11 |
| 1598 | 2 | | | | | | 1 | | | | | | | 1 | | | | |
| 1719 | | | | | | 2 | 14 | 1 | | | | | | 1 | | | | 11 |
| 1811 | | | | | | 1 | | | 46 | | | | | | 15 | | | |

"One important result to emerge from these studies is the ***relatively large number of sites*** at which ***homoplasic events*** have occurred."

(see our Table 2)

## Mutation Rate Analysis in the mtDNA Coding Region



Letters to the Editor

*Am. J. Hum. Genet.* 72:1341–1346, 2003

To Trust or Not to Trust an Idiosyncratic Mitochondrial Data Set

- "***Homoplasy*** in the coding region is ***much less*** than in the control region and may have ***only a few*** hot spots (see, e.g., table 2 of Herrnstadt *et al.* [2002])"

## Mutation Rate Analysis in the mtDNA Coding Region

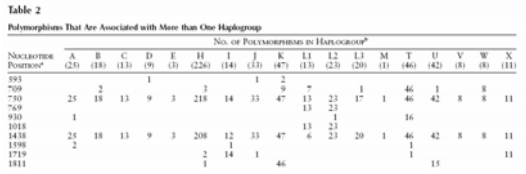


Table 2

Polymorphisms That Are Associated with More than One Haplogroup

| Nucleotide Position* | No. of Polymorphisms in Haplogroup* | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A (25) | B (18) | C (13) | D (9) | E (3) | H (226) | I (14) | J (13) | K (47) | L1 (13) | L2 (23) | L3 (20) | M (1) | T (46) | U (42) | V (8) | W (8) | X (11) |
| 593 | | | | 1 | | | | 1 | 2 | | | | | | | | | |
| 709 | | 2 | | | | 1 | | | 9 | 7 | | 1 | | 46 | 1 | | 8 | |
| 750 | 25 | 18 | 13 | 9 | 3 | 218 | 14 | 33 | 47 | 13 | 23 | 17 | 1 | 46 | 42 | 8 | 8 | 11 |
| 769 | | | | | | | | | | 13 | 23 | | | | | | | |
| 930 | 1 | | | | | | | | | | 1 | | | 16 | | | | |
| 1018 | | | | | | | | | | 13 | 23 | | | | | | | |
| 1438 | 25 | 18 | 13 | 9 | 3 | 208 | 12 | 33 | 47 | 6 | 23 | 20 | 1 | 46 | 42 | 8 | 8 | 11 |
| 1598 | 2 | | | | | | 1 | | | | | | | 1 | | | | |
| 1719 | | | | | | 2 | 14 | 1 | | | | | | 1 | | | | 11 |
| 1811 | | | | | | 1 | | | 46 | | | | | | 15 | | | |

Herrnstadt – LARGE NUMBER!!  Yao – Only a few…

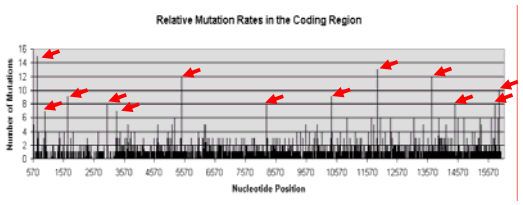
## Our Results…

- Analysis of 646 coding region genomes

| | Parsimony | | NJ | |
|---|---|---|---|---|
| Data Set (# genomes) | Tree Length | α estimation | Tree Length | α estimation |
| Ingman HV1 (53) | 144 | 0.2091 | 144 | 0.2081 |
| Ingman Control Region (53) | 273 | 0.0038 | 281 | 0.0036 |
| Ingman Coding Region (53) | 588 | 0.0075 | 588 | 0.0074 |
| Ingman Full Data (53) | 873 | 0.0050 | 876 | 0.0067 |
| Total Coding Data (646) | 2352 | 0.0086 | 2353 | 0.0083 |

**Extreme** rate variation exists in the coding region

## Relative Mutation Rates




## The Mutation Rate Spectrum

- How does the estimated mutation rate spectrum compare to the forensically informative SNPs?
- Are all of the forensic SNPs mutational "hot-spots?"

## Mutation Rates and the 8 Multiplex SNP Panels

| Length | Character | Gene | codon | 241 Caucasians |
|---|---|---|---|---|
| 15 | 709 | 12S | * | Yes |
| 13 | 11914 | ND4 | 3 | Yes-SNP |
| 12 | 5460 | ND2 | 1 | Yes |
| 12 | 13708 | ND5 | 1 | Yes |
| 10 | 15924 | tRNA(thr) | * | Yes |
| 9 | 1719 | 16S | * | Yes |
| 9 | 10398 | ND3 | 1 | Yes |
| 8 | 3010 | 16S | * | Yes-SNP |
| 8 | 8251 | COII | 3 | |
| 8 | 14470 | ND6 | 3 | Yes-SNP |
| 8 | 15784 | CYTB | 3 | |
| 7 | 961 | 12S | * | |
| 7 | 3316 | ND1 | 1 | |
| 6 | 5237 | ND2 | 3 | Yes |
| 6 | 10915 | ND4 | 3 | Yes |
| 6 | 11719 | ND4 | 3 | Yes-SNP |
| 6 | 12007 | ND4 | 3 | Yes-SNP |
| 6 | 12346 | ND5 | 1 | |
| 6 | 13105 | ND5 | 1 | Yes |
| 6 | 13928 | ND5 | 2 | |
| 6 | 14569 | ND6 | 3 | |
| 6 | 14766 | CYTB | 2 | |
| 6 | 15301 | CYTB | 3 | |
| 6 | 15670 | CYTB | 3 | |
| 6 | 15884 | CYTB | nc | Yes-SNP |

Only 6 of the 59 SNPs are among the "fastest" sites

So…

Most of the SNPs that we identified as being very useful for discrimination were SLOW….

## A Case Example

Skeletal remains - "H1" in the HV1/HV2 region.

Thought to belong to one of two individuals…

(Smith or Jones)

Family references for Smith and Jones were obtained.

| Smith Family | Jones Family |
|---|---|
| 263 A-G | 263 A-G |
| 315.1 C | 315.1 C |

## A Case Example

Skeletal remains - "H1" in the HV1/HV2 region.

Thought to belong to one of two individuals…

(Smith or Jones)

Remains tested for VR region: 477 T-C and 16519 T-C

| Smith Family | Jones Family |
|---|---|
| 263 A-G | 263 A-G |
| 315.1 C | 315.1 C |
| 477 T-C | |
| 16519 T-C | 16519 T-C |

## Question….

Can the Smith Family be excluded as a possible family reference for the skeletal remains?

| Smith Family | Jones Family |
|---|---|
| 263 A-G | 263 A-G |
| 315.1 C | 315.1 C |
| 477 T-C | |
| 16519 T-C | 16519 T-C |

## Question….

# NO!

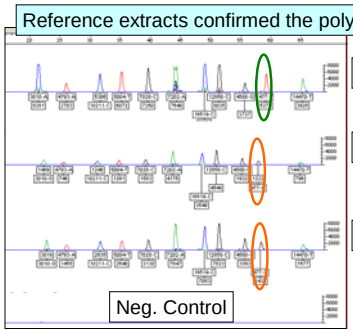
Only one mutation differs between the two families…

| Smith Family | Jones Family |
|---|---|
| 263 A-G | 263 A-G |
| 315.1 C | 315.1 C |
| 477 T-C | |
| 16519 T-C | 16519 T-C |

**INCONCLUSIVE**

## A Case Example

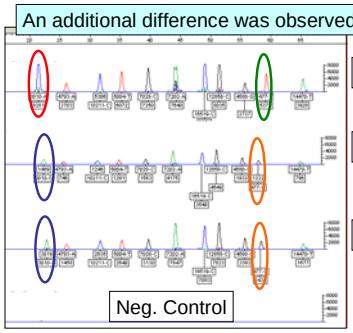
Reference extracts confirmed the polymorphism at 477




(R. Just, AFDIL)

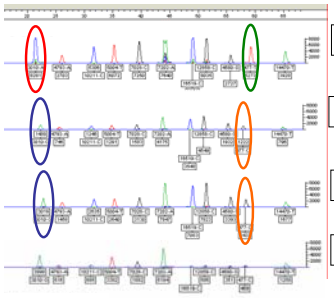
## A Case Example

An additional difference was observed at position 3010




(R. Just, AFDIL)

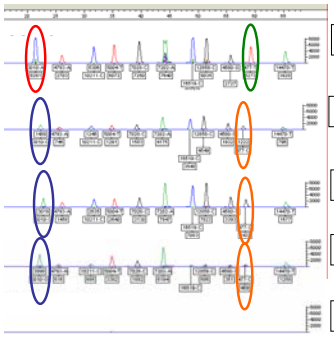
## A Case Example




15uL Reaction; 0.07Units/uL Taq; 31 cycles --- 100 RFU cutoff

(R. Just, AFDIL)

## A Case Example




(R. Just, AFDIL)

## A Case Example

| Smith Family | Skeletal Remains | Jones Family |
|---|---|---|
| 263 A-G | 263 A-G | 263 A-G |
| 315.1 C | 315.1 C | 315.1 C |
| 477 T-C | 477 T-C | 16519 T-C |
| 3010 A-G | 3010 A-G | |
| 16519 T-C | 16519 T-C | |

Remains – match exactly the Smith family, now **2 differences** from the Jones family – can be *excluded*.

## Summary

- Purpose – Maximize Discrimination.
- A supplement to current HV1/HV2 testing.
- When the Forensic Scientist encounters a common type, select the most discriminating SNP panel.

## Summary

- AFDIL – focused on sites that are not associated with the potential for phenotypic change.
- Most of the informative sites are rare, slow polymorphisms that are useful for discrimination in a particular common type.
- A few SNP sites may be useful for resolving common HV1/HV2 types from various backgrounds.
- Evaluation of non-synonymous sites that are not associated with diseases may also be useful for forensic discrimination… site-by-site evaluation (e.g. 3010 is very useful among HgH.

## Publications

Michael D. Coble · Rebecca S. Just
Jennifer E. O'Callaghan · Ilona H. Letmanyi
Christine T. Peterson · Jodi A. Irwin · Thomas J. Parsons

**Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians**

*IJLM* (2004) **118:** 137-146

Peter M. Vallone · Rebecca S. Just · Michael D. Coble
John M. Butler · Thomas J. Parsons

**A multiplex allele-specific primer extension assay for forensically informative SNPs distributed throughout the mitochondrial genome**

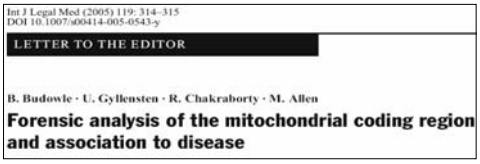
*IJLM* (2004) **118:** 147- 157

http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm

## Efforts with Coding Region Sequencing Applied to Human mtDNA Testing

- Tzen *et al.* (2001) *Forensic Sci. Int.* 120:204-209
  - Portions of mtATP6, mtATP8 among 119 Chinese individuals
- Andreasson *et al.* (2002) *Biotechniques* 32:124-133
  - Highly variable regions of mtDB among 190 Swedish individuals
- Lee *et al.* (2002) *Int. J. Legal Med.* 116:74-78
  - mtCyt B among 98 Korean individuals
- Lutz-Bonengel *et al.* (2003) *Int. J. Legal Med.* 117:133-142
  - mtATP6, mtATP8, mtND4 among 109 German individuals
- Poetsch *et al.* (2003) *Mitochondrion* 3:133-137
  - portions of tRNA K, ATP6, ATP8 among 180 German individuals
- Coble *et al.* (2004) *Int. J. Legal Med.*, 118:137-146
  - 241 complete mtGenomes among 18 common Cauc. HV1/HV2 types

## Criticisms of Synonymous SNPs for Discrimination


Int J Legal Med (2005) 119: 314–315
DOI 10.1007/s00414-005-0543-y

LETTER TO THE EDITOR

B. Budowle · U. Gyllensten · R. Chakraborty · M. Allen

**Forensic analysis of the mitochondrial coding region and association to disease**

## Budowle *et al.* (2005)

- [Coble and Vallone] have proposed that forensic analyses of the coding region [should] be restricted to synonymous substitutions [and] suggest that sequencing strategies for forensic analyses of the coding region of the mtDNA genome should be avoided [and] that only SNP-based systems should be employed.
- We disagree with this proposition [would] severely hamper the use of mtDNA in forensic testing.

## Budowle *et al.* (2005)

- "by limiting the analysis only to synonymous polymorphisms that cannot have any phenotypic effect, ***a large part of the polymorphic positions (and thus forensically informative) would be excluded."***

## An Evaluation of Coding Region SNPs to Sequencing

- Data

How well will multiplex A perform among with random samples belonging to the most common haplotype?

→ 24 Samples H:1 (263 A-G; 315.1C)

What is the general utility of the SNPs for HgH proposed by Coble *et al.* (2004)?

→ 54 hgH Sequences (Achilli et al. 2004)

## An Evaluation of Coding Region SNPs to Sequencing

- Methods

| | |
|---|---|
| G3010A | A7202G |
| G4580A | C10211T |
| A4793G | C12858T |
| T5004C | T14470C |
| C7028T | |

9 SNPs

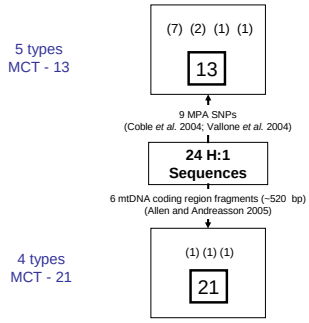
Coble *et al.* (2004) *Int J Legal Med* 118:137–146.
Vallone *et al.* (2004) *Int J Legal Med* 118:147–157.

| Region (np) | Variable Position(s) |
|---|---|
| 2782 - (2662-2762) | A2706G |
| 4275 - (4303-4363) | T4336C |
| 8665 - (8689-8780) | G8697A |
| | T8705C |
| 10362 - (10385-10484) | A10398G |
| | T10463C |
| 12673 - (12694-12784) | C12705T |
| 15758 - (15777-15873) | C15833T |

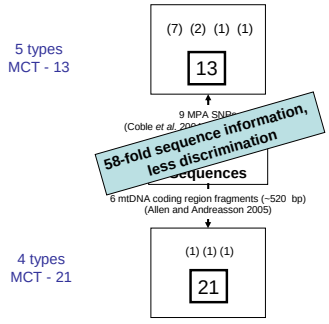
~520 bp

Allen and Andreasson (2005) *Methods Mol Biol* 297:179–196.

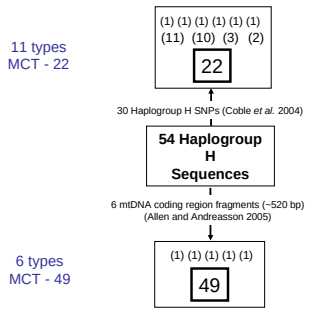
## Results



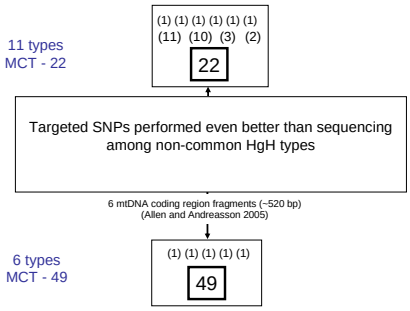

## Results




## Results



## Results




## Why Did 9 SNPs Outperform Sequencing?

| Region (np) | Variable Position(s) |
|---|---|
| 2782 - (2662-2762) | A2706G |
| 4275 - (4303-4363) | T4336C |
| 8665 - (8689-8780) | G8697A |
| | T8705C |
| 10362 - (10385-10484) | A10398G |
| | T10463C |
| 12673 - (12694-12784) | C12705T |
| 15758 - (15777-15873) | C15833T |

**Diagnostic Haplogroup T SNPs**

Allen and Andreasson (2005) *Methods Mol Biol* 297:179–196.

## Why Did 9 SNPs Outperform Sequencing?

| Region (np) | Variable Position(s) |
|---|---|
| 2782 - (2662-2762) | A2706G |
| 4275 - (4303-4363) | T4336C |
| 8665 - (8689-8780) | G8697A |
| | T8705C |
| 10362 - (10385-10484) | A10398G |
| | T10463C |
| 12673 - (12694-12784) | C12705T |
| 15758 - (15777-15873) | C15833T |

**Diagnostic for SuperHaplogroup R**

Allen and Andreasson (2005) *Methods Mol Biol* 297:179–196.

**All Caucasians have the rCRS variant**

## How Much Information is Lost?

African-derived Sequence

Haplogroup L0a1

"Hausa" (Ingman et al. 2000)

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| C | 64 | T | A | 750 | G | C | 7028 | T | G | 11719 | A |
| A | 93 | G | G | 769 | A | A | 7146 | G | G | 11914 | A |
| C | 150 | T | T | 825 | A | C | 7256 | T | G | 12007 | A |
| G | 185 | A | G | 1018 | A | G | 7521 | A | C | 12705 | T |
| A | 189 | G | C | 1048 | T | C | 8428 | T | A | 12720 | G |
| A | 200 | G | A | 1438 | G | C | 8468 | T | A | 13105 | G |
| T | 236 | C | A | 2245 | C | A | 8566 | G | A | 13276 | G |
| G | 247 | A | A | 2706 | G | C | 8655 | T | C | 13506 | T |
| A | 263 | G | G | 2758 | A | A | 8701 | G | C | 13650 | T |
| C | 522 | : | T | 2885 | C | A | 8860 | G | T | 14308 | C |
| A | 523 | : | C | 3107 | : | C | 9042 | T | C | 14766 | T |
| G | 16129 | A | C | 3516 | A | A | 9347 | G | C | 15136 | T |
| C | 16148 | T | C | 3594 | T | T | 9540 | C | A | 15326 | G |
| C | 16168 | T | T | 3866 | C | G | 9755 | A | G | 15431 | A |
| T | 16172 | C | A | 4104 | G | C | 9818 | T | | | |
| C | 16187 | T | C | 4312 | T | A | 10398 | G | | | |
| C | 16188 | G | T | 4586 | C | G | 10589 | A | | | |
| T | 16189 | C | A | 4769 | G | C | 10664 | T | | | |
| C | 16223 | T | T | 5096 | C | G | 10688 | A | | | |
| A | 16230 | G | G | 5231 | A | T | 10810 | C | | | |
| T | 16311 | C | T | 5442 | C | T | 10873 | C | | | |
| C | 16320 | T | G | 5460 | A | T | 10915 | C | | | |
| T | 16362 | C | C | 5603 | T | G | 11176 | A | | | |
| T | 16519 | C | T | 6185 | C | A | 11641 | G | | | |

## How Much Information is Lost?

27 non-synon/RNA mutations

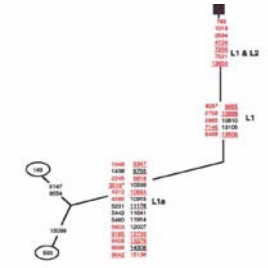
35 synon mutations

**"a large part of the polymorphic positions (and thus forensically informative) would be excluded."**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 750 | G | C | 7028 | T | G | 11719 | A |
| G | 769 | A | A | 7146 | G | G | 11914 | A |
| T | 825 | A | C | 7256 | T | G | 12007 | A |
| G | 1018 | A | G | 7521 | A | C | 12705 | T |
| C | 1048 | T | C | 8428 | T | A | 12720 | G |
| A | 1438 | G | C | 8468 | T | A | 13105 | G |
| A | 2245 | C | A | 8566 | G | A | 13276 | G |
| A | 2706 | G | C | 8655 | T | C | 13506 | T |
| G | 2758 | A | A | 8701 | G | C | 13650 | T |
| T | 2885 | C | A | 8860 | G | T | 14308 | C |
| C | 3107 | : | C | 9042 | T | C | 14766 | T |
| C | 3516 | A | A | 9347 | G | C | 15136 | T |
| C | 3594 | T | T | 9540 | C | A | 15326 | G |
| T | 3866 | C | G | 9755 | A | G | 15431 | A |
| A | 4104 | G | C | 9818 | T | | | |
| C | 4312 | T | A | 10398 | G | | | |
| T | 4586 | C | G | 10589 | A | | | |
| A | 4769 | G | C | 10664 | T | | | |
| T | 5096 | C | G | 10688 | A | | | |
| G | 5231 | A | T | 10810 | C | | | |
| T | 5442 | C | T | 10873 | C | | | |
| G | 5460 | A | T | 10915 | C | | | |
| C | 5603 | T | G | 11176 | A | | | |
| T | 6185 | C | A | 11641 | G | | | |

Coding Region

## How Much Information is Lost?




High frequency is not necessarily a reliable indicator of "informativeness" in the coding region.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 750 | G | C | 7028 | T | G | 11719 | A |
| G | 769 | A | A | 7146 | G | G | 11914 | A |
| T | 825 | A | C | 7256 | T | G | 12007 | A |
| G | 1018 | A | G | 7521 | A | C | 12705 | T |
| C | 1048 | T | C | 8428 | T | A | 12720 | G |
| A | 1438 | G | C | 8468 | T | A | 13105 | G |
| A | 2245 | C | A | 8566 | G | A | 13276 | G |
| A | 2706 | G | C | 8655 | T | C | 13506 | T |
| G | 2758 | A | A | 8701 | G | C | 13650 | T |
| T | 2885 | C | A | 8860 | G | T | 14308 | C |
| C | 3107 | : | C | 9042 | T | C | 14766 | T |
| C | 3516 | A | A | 9347 | G | C | 15136 | T |
| C | 3594 | T | T | 9540 | C | A | 15326 | G |
| T | 3866 | C | G | 9755 | A | G | 15431 | A |
| A | 4104 | G | C | 9818 | T | | | |
| C | 4312 | T | A | 10398 | G | | | |
| T | 4586 | C | G | 10589 | A | | | |
| A | 4769 | G | C | 10664 | T | | | |
| T | 5096 | C | G | 10688 | A | | | |
| G | 5231 | A | T | 10810 | C | | | |
| T | 5442 | C | T | 10873 | C | | | |
| G | 5460 | A | T | 10915 | C | | | |
| C | 5603 | T | G | 11176 | A | | | |
| T | 6185 | C | A | 11641 | G | | | |

Coding Region

## Conclusions

- A selected SNP method out-performed a random sequencing protocol for increased discrimination.
- This method was developed to avoid additional sequencing, as often, the casework at AFDIL involves challenging cases where the quantity and quality of extract would prohibit an extensive post-HV1/HV2 sequencing strategy.

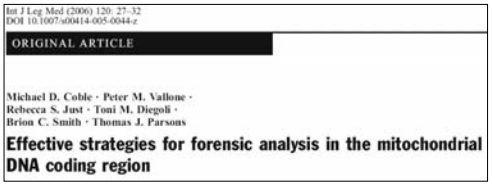
## Conclusions

- Budowle *et al.* (2005) make several valid points about the usefulness of non-synonymous sites for discrimination, and we have made a careful evaluation about the potential use of these sites.
- However, many cases processed by AFDIL are publicly visible and involves large segments of the general population. The US military now has a policy of compulsory submission of a blood sample retained solely for the purposes of DNA identification, which is necessary in the face of military casualty.

## Conclusions

- A conservative approach was developed, and this may or may not meet the needs of other forensic laboratories
- Some countries, such as Germany, have strict regulations the use of forensic testing that may reveal medical information… this has resulted in the call for disqualification of certain markers (e.g. X chromosome – see Szibor *et al.* 2005 *IJLM*).
- Need to weigh the costs and benefits for developing effective strategies to increase mtDNA discrimination.

## More Information


Int J Leg Med (2006) 120: 27–32
DOI 10.1007/s00414-005-0044-z

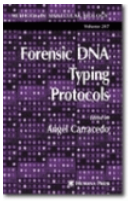
ORIGINAL ARTICLE

Michael D. Coble · Peter M. Vallone · Rebecca S. Just · Toni M. Diegoli · Brion C. Smith · Thomas J. Parsons

**Effective strategies for forensic analysis in the mitochondrial DNA coding region**

http://www.cstl.nist.gov/biotech/strbase/pub_pres/Coble_IJLM_coding_mtSNPs.pdf

## Other SNP Assays




Methods in Molecular Biology
**Volume: 297**
**(2005)**

**SNaPshot Typing of Mitochondrial DNA Coding Region Variants**

Antonio Salas
Beatriz Quintáns
Vanesa Álvarez-Iglesias

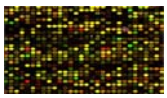
**Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing.**

Quintáns et al. (2004) *Forensic Sci Int.*

## Emerging mtDNA technologies

mtDNA micro-chip technology

## Steps in Running the Affymetrix Resequencing Array



Day 1
1. PCR
2. Desalting and Reaction Clean Up
3. Quantitation
4. Pooling
5. Fragment
6. Label
7. Hybridize

Day 2
8. Wash/Stain
9. Scan
10. Data Analysis

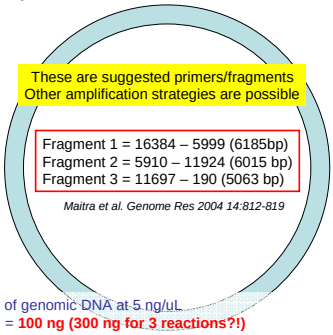

*Dr. Peter Vallone, NIST – Presentation at DNA in Forensics, Innsbruck, Austria (Sept. 06)*

## The mtGenome is Amplified in 3 Singleplex Reactions

PCR is performed using a
7 min extension time
94 2 min
94 15 sec
68 7 min {30 cycles}
68 12 min
Takes about 5 hours

TaKaRa LA Taq™ polymerase

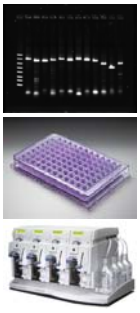
50 µL total volume
2.5 Units of Taq
400 µM dNTPs
Primer 0.2 µM F/R
2.5 mM $Mg^{++}$




Affymetrix suggests preparation of genomic DNA at 5 ng/uL and usage of 20 uL per reaction = **100 ng (300 ng for 3 reactions?!)**
A 7.5kb control template is also amplified and carried out through the process

## Post PCR Steps

| | |
|---|---|
| Presence of amplicon confirmed on agarose gel | 15 min |
| Amplicons are desalted (salts,PCR primers) Millipore Montage screen plate (96 well) | 30 min |
| The 3 amplicons are quantitated (UV, pico green) | 30 min |
| Amplicons are pooled in an equimolar ratio and fragmented (enzymatically) down to ~200 bp | 30 min |
| Labeled (with biotin) | 2 hours |
| Sample loaded onto GeneChip and incubated at 45°C | 16 hours |
| Arrays are washed and fluorescently labeled on fluidics station | (~1 hour) |



## GeneChip Version 2.0

Developed for mutation detection – disease association studies

Version 2.0 interrogates the entire ~16kb mitchondrial genome
(ver 1.0 coding region only)

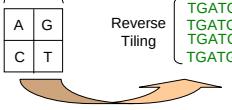
Contains common variants in HVI and HVII regions
Information from FBI database (~500 types)
www.fbi.gov/hq/lab/fsc/backissu/april2002/miller1.htm
HV-Types observed at least twice in the database
Plus 250 singletons

Array is tiled with 25 nt oligomers; the center (13th) base interrogates the sequence

Both forward and reverse strands of the sequence are probed

*Version 1.0 Cutler et al. Genome Res 2001 11:1913-1925*
*Version 2.0 Zhou et al. J Mol Diagn 2006 8: 476-482*

## Tiling of the Array




*Labeled PCR fragment*

Forward CGTACTACACGACACGTACTACGTTGTACA

Reverse Tiling
TGATGTGCTGTGCATGATGCAACAT
TGATGTGCTGTGTATGATGCAACAT
TGATGTGCTGTGAATGATGCAACAT
TGATGTGCTGTGGATGATGCAACAT

*25-mers with variable SNP in the 13th position*

## Tiling of the Array


Forward CGTACTACACGACACGTACTACGTTGTACA
X
TGATGTGCTGTGTATGATGCAACAT

Forward CGTACTACACGACACGTACTACGTTGTACA
X
TGATGTGCTGTGAATGATGCAACAT

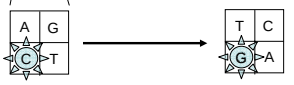
Forward CGTACTACACGACACGTACTACGTTGTACA
X
TGATGTGCTGTGGATGATGCAACAT

Forward CGTACTACACGACACGTACTACGTTGTACA
|||
TGATGTGCTGTGCATGATGCAACAT

## Tiling of the Array

Forward CGTACTACACGACACGTACTACGTTGTACA

TGATGTGCTGTGCATGATGCAACAT

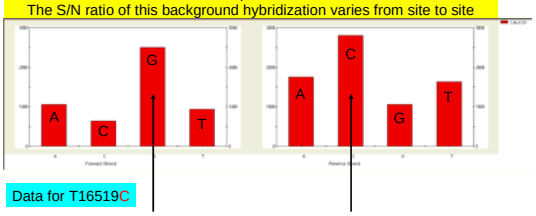



**Sequence Tiled** **Sequence Hybridized**

## GeneChip Version 2.0

Hybridization signal intensity can be viewed for each probe

Note that the mismatched probes bind to a lesser extent
The S/N ratio of this background hybridization varies from site to site




Data for T16519C

## GeneChip Version 2.0

Array is read in a fluorescent scanner (10 min)

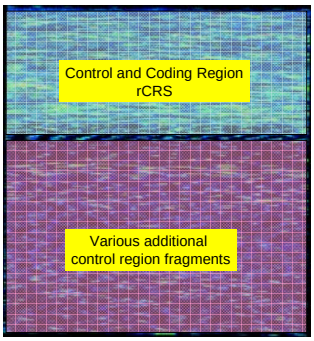
Software interprets the array scan

Probe intensities are tabulated

Sequence calls are made

Sequence analysis is performed in a 'batch mode' with at least 10-15 other arrays

File Size ~20MB
> 37k probe elements collected



## How are Base Calls Made?

Using the Affymetrix GSEQ software fluorescence cell intensity is evaluated (base calling algorithms)

Array data should be analyzed in a batch (> 10 samples)

A 'Score' value allows for varying degrees of stringency

The Score parameter ranges between 1 and 12
- 12 = conservative (more N calls)
- 1 = liberal (less N calls, but possible miscalls)

An N call is an ambiguous base call (A,G,C,T)

## Data Analysis

Calls made using Affymetrix GSEQ software
Fluorescent probe intensities are evaluated (base calling algorithms)

Scores of 1, 6, and 12 were used
Score = a base calling parameter allows for varying degrees of stringency
12 = conservative (more N calls)
1 = liberal (less N calls, but possible miscalls)

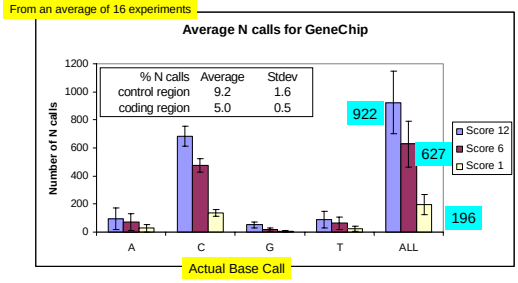
An N call is an ambiguous base call (A,G,C,T, no call)

Sequence files were exported (composite of all calls made on the array)

Data summary will focus on the expected sequence differences from rCRS as determines by fluorescent sequencing

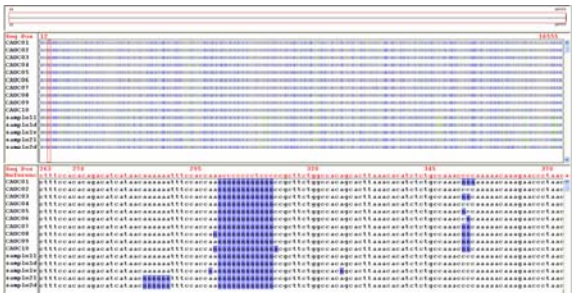
## Effect of Score on N calls

From an average of 16 experiments




C base calls are most commonly assigned as N
As the Score drops the number of N calls decreases significantly
From ~5.5 to 1.2 % of the mtGenome
On average, N calls are **not randomly** distributed throughout the mtGenome with the majority found in poly-C regions

## Location of N calls



On average N calls are **not randomly** distributed throughout the mtGenome
The majority are found in poly-C regions

## Samples

From our set of NIST U.S. population samples

AA01
- 63 differences from rCRS
- 14 in control region
- 49 in coding region
- Heteroplasmy at G1709R & C15978Y

Hisp01
- 46 differences from rCRS
- 14 in control region
- 32 in coding region

Full genome sequencing was performed for all 12 samples

AA01 and Hisp01 at AFDIL
Cauc 01-10 at NIST

Cauc 01- 10 Contain the identical control region sequence (16024-577)

## Sample AA01 – Control Region

**4 of 14** control region polymorphisms missed
(1) 146 T-C (results in poly-C stretch)
(1) 315.1 C insertion
(2) 523/524 AC deletion

Samples run in triplicate and analyzed for 3 Score values

| AA01 | | Score 1 | | | Score 6 | | | Score 12 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| position | | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 146 | T-C | C | C | N | C | N | C | N | N | N |
| 315.1 | C-ins. | missed | missed | missed | missed | missed | missed | missed | missed | missed |
| 523 | A-del. | missed | missed | missed | missed | missed | missed | N | N | N |
| 524 | C-del. | missed | missed | missed | missed | missed | missed | missed | missed | missed |

Note: The array failed at other positions (1 - 5% N calls), but we are focusing on comparing differences from rCRS

## Sample AA01 – Coding Region

**7 of 49** coding region polymorphisms missed
(2) Heteroplasmy not called for 1,709 & 15,978
(2) 2,416 & 10873 T-C results in poly-C stretch
(2) 11,719 &11,914 G-A not called?
(1) 13,650 C-T weak signal

| AA01 position | | Score 1 | | | Score 6 | | | Score 12 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 1709 | G-R | G | G | G | G | G | G | G | G | G |
| 2416 | T-C | N | N | N | N | N | N | N | N | N |
| 10873 | T-C | N | C | C | N | N | N | N | N | N |
| 11719 | G-A | G | G | G | N | N | N | N | N | N |
| 11914 | G-A | G | G | G | G | G | G | N | N | N |
| 13650 | C-T | T | T | T | N | N | N | N | N | N |
| 15978 | C-Y | C | C | C | N | N | N | N | N | N |

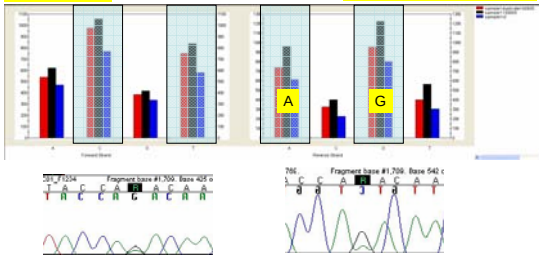
Samples run in triplicate and analyzed for 3 Score values

Note: (1 - 5% N calls)

## Heteroplasmy

AA01 G1709R

Data from triplicate analysis




After a more careful examination, the array data would seem to exhibit heteroplasmy
This would have not been found without prior knowledge (fl sequencing)

## Sample Hisp01 – Control Region

**6 of 14** control region polymorphisms missed
(3) 249, 290, 291 A deletion
(1) 315.1 C insertion
(2) 523/524 AC deletion

| Hisp01 position | | Score 1 | | | Score 6 | | | Score 12 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 249 | A-del. | missed | missed | N | N | N | N | N | N | N |
| 290 | A-del. | missed | missed | missed | N | missed | N | N | N | N |
| 291 | A-del. | missed | missed | missed | missed | missed | N | N | N | N |
| 315.1 | C-ins. | missed | missed | missed | missed | missed | missed | missed | missed | missed |
| 523 | A-del. | missed | missed | missed | N | missed | missed | N | N | N |
| 524 | C-del. | missed | missed | missed | missed | missed | missed | missed | missed | missed |

Samples run in triplicate and analyzed for 3 Score values

Note: (1 - 5% N calls)

## Sample Hisp01 – Coding Region

**7 of 32** coding region polymorphisms missed
(3) close proximity calls: 9,540 9,545 & 9,557
(1) 10,400 not called once (a polymorphism at 10,398 may effect this call)
(1) 10,873 T-C results in poly-C stretch
(2) 11,719 &11,914 G-A not called?

| Hisp01 position | | Score 1 | | | Score 6 | | | Score 12 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 9,540 | T-C | N | N | C | N | C | N | N | C | N |
| 9,545 | A-G | N | N | N | G | N | N | N | N | N |
| 9,557 | C-T | T | T | T | N | N | N | N | N | N |
| 10,400 | C-T | T | T | T | T | T | T | T | T | N |
| 10,873 | T-C | N | N | N | N | N | N | N | N | N |
| 11,719 | G-A | G | G | G | N | N | N | N | N | N |
| 11,914 | G-A | G | G | G | G | G | G | N | N | N |

Samples run in triplicate and analyzed for 3 Score values

Note: (1 - 5% N calls)

## Sample Hisp01

Sites
9540 T-C
9545 A-G
9557 C-T

Δ = 6 nt — Δ = 13 nt

rCRS TACCCCCCAATTAGGAGGGCACTGGCCCCCAACAGGCATCA
Hisp01 TACCCCCCAACTAGGGGGGCACTGGCCTCCAACAGGCATCA

Δ = 18 nt

Hisp01 array calls: NNNNNNNNNNNNNNNNNNNNCTGGNNNNNAACAGGCATCA

These three closely spaced polymorphisms in sample Hisp01 present a challenge for the array

Although the differences from rCRS are shown for Hisp01 - the array calls indicate that entire region is affected, not just the 3 sites

Cauc01 array calls: NNNNNCCCAATTAGGAGGGCACTGGNNNNNAACAGGCATCA

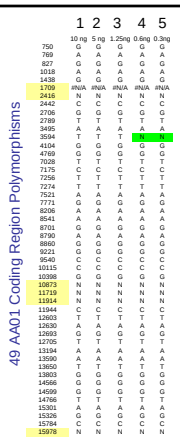
## Sensitivity

Recommended amount = 100 ng per reaction
We were using 25 ng (for 3 PCRs) = 75 ng
How low can we go?
Prepared a dilution series:

| | per reaction | Total Used | N calls |
|---|---|---|---|
| 1 | 10 ng | 30 | 831 |
| 2 | 5 ng | 15 | 836 |
| 3 | 1.25ng | 3.75 | 864 |
| 4 | 0.6ng | 1.8 | 1070 |
| 5 | 0.3ng | 0.9 | 992 |

Sample AA01
Version 1.0 array - coding region only

Results were identical down to 0.3 ng except for C3594T which gave an N at 0.6 and 0.3 ng



## Sensitivity

If the amplicons were detected and quantitated then a result was obtained on the GeneChip array

Using the 3 amplicon approach we were able to obtain results using 0.3 ng (0.9 ng total) of genomic DNA

Additional experiments will be needed to ascertain how this approach will work on degraded (casework) samples

- 'Degraded' primer sets?
- Focus on a subset of the mtGenome?
- Mito whole genome amplification?

## Resolving a Common HV-Type

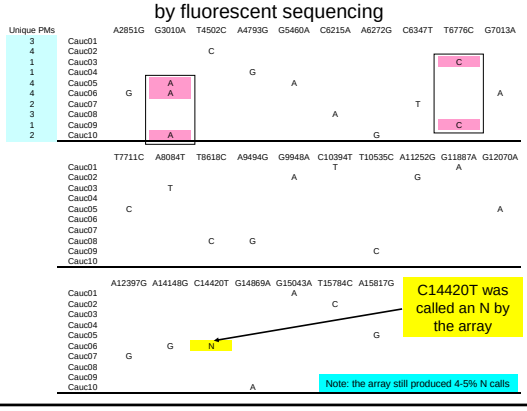
Ten samples (haplogroup H:1) sharing the most common HV-type found in Caucasians were tested
(Control region sequencing performed at AFDIL)

Samples were run on the array only once

15 ng (total) of genomic DNA were used (5 ng/PCR rxn)

How do array results compare with fluorescent sequencing for finding resolving SNPs?

## 27 polymorphisms were detected by fluorescent sequencing



## Conclusions

The array has difficulties with calling insertions, deletions, closely spaced polymorphisms and poly-C stretch regions

Based on the amount of DNA required by the array it may not be useful for casework

Won't take the place of fluorescent sequencing if 100% coverage is required, but it is a decent screening tool

**It may have utility of discovering polymorphisms in reference individuals - these polymorphisms could then be targeted by traditional mtDNA sequencing methods (or SNP assays)**

## Thank You!

**NIST**
Dr. John Butler
Dr. Peter Vallone
Margaret Kline
Jan Redman

**NIJ**
Dr. Lois Tully

**ILM**
Dr. Walther Parson
Dr. Anita Brandstatter

**Dept de Medicina Legal**
Dr. Cintia Friedman (Brasil)

**AFDIL**
Dr. Tom Parsons (now at the ICMP)
Jodi Irwin
Rebecca Just
Jessica Saunier
Kim Sturk
Toni Diegoli
Katie Strouss
Melissa Scheible
Suni Edson
Suzie Barritt-Ross
Col. Lou Finelli
Dr. Brion Smith
Ilona Letmanyi
Christine Peterson
AFDIL Scientists and Staff

Mike Coble – michael.coble@afip.osd.mil

## QUESTIONS?

http://www-medlib.med.utah.edu/WebPath/jpeg2/EM003.jpg