Poster B92 at AAFS, Seattle, WA, February 24, 2006

An Evaluation of Mitochondrial DNA Variation: From Linear Arrays to Whole Genome Sequencing

Michael D. Coble

National Institute of Standards and Technology, Biochemical Science Division, 100 Bureau Drive MS 8311, Gaithersburg, MD 20899-8311

Email: michael.coble@nist.gov

Phone: 301-975-4330

References

1) Kline MC, Vallone PM, Redman JW, Duewer DL, Calloway CD, Butler JM. (2005) Mitochondrial DNA typing screens with control region and coding region SNPs. *J Forensic Sci.* 50(2):377–385.

2) Coble MD, Just RS, O'Callaghan JE, Letmanyi IH, Peterson CT, Irwin JA, Parsons TJ. (2004) Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. *Int J Legal Med.* 118:137-46.

3) Vallone PM, Just RS, Coble MD, Butler JM, Parsons TJ. A multiplex allele-specific primer extension assay for forensically informative SNPs distributed throughout the mitochondrial genome. *Int J Legal Med.* 118:147-57.

4) Budowle, B., Gyllensten, U., Chakraborty, R., and Allen, M. (2005) Forensic analysis of the mitochondrial coding region and association to disease. *Int. J. Leg. Med.* 119:314-315.

5) Coble MD, Vallone PM, Just RS, Diegoli TM, Smith BC, Parsons TJ. (2006) Effective strategies for forensic analysis in the mitochondrial DNA coding region. *Int J Legal Med.* 120:27-32.

6) Allen, M. and Andreasson, H. (2005) Mitochondrial D-loop and coding sequence analysis using pyrosequencing. *Methods Mol. Biol.* 297: 179-196.

7) Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knijff P, Feldman M, Cavalli-Sforza LL, Oefner PJ. (2006) The role of selection in the evolution of human mitochondrial genomes. *Genetics.* 172:373-387.

8) Achilli, A., Rengo, C., Magri, C., Battaglia, V., Olivieri, A., Scozzari, R., Cruciani, F., Zeviani, M., Briem, E., Carelli, V., Moral, P., Dugoujon, J. M., Roostalu, U., Loogvali, E. L., Kivisild, T., Bandelt, H. J., Richards, M., Villems, R., Santachiara-Benerecetti, A. S., Semino, O., and Torroni, A. (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *Am. J. Hum. Genet.* 75:910-918.

9) Brandstaller A, Peterson CT, Irwin JA, Mpoke S, Koech DK, Parson W, Parsons TJ. Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. *Int J Legal Med* 2004;118:294-306.

10) Ingman M, Kaessmann H, Paabo S, Gyllensten U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* 408:708-713.

Forensic mtDNA analysis of highly degraded materials, or samples lacking sufficient quantity of nuclear DNA (e.g. shed hairs), has found an important niche in DNA testing. Recent mtDNA research has focused on two important limitations for mtDNA testing: (a) the cost of generating a mtDNA sequence profile, and (b) the low power of discrimination associated with common mtDNA types. To overcome the cost prohibition of mtDNA testing, Linear Arrays have been evaluated as a screening tool (Kline *et al.* 2005). To increase the power of discrimination for individuals sharing one of the few common mtDNA types, strategies to identify resolving polymorphisms in the coding region through sequencing short (~100 bp) fragments of (Allen and Andreasson 2005) or through the identification of SNPs (Coble *et al.* 2004; Vallone *et al.* 2004) have been proposed.

To access the amount of variation gained by entire control region sequencing compared to Linear Array mitotyping, a comparison of discrimination from Linear Array – HV1 – HV1/HV2 – Control Region was determined among 666 population samples. Further discrimination for the most common haplotype, MCH (A263G; 315.1C), was determined by coding region SNP analysis (Coble *et al.* 2004; Vallone *et al.* 2004). In addition, a survey of the underlying source of null alleles (blanks) in Linear Arrays, as determined by sequence information was performed.

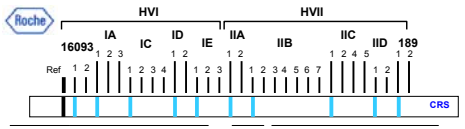
Finally, an evaluation of coding region variation in both a global dataset of mtDNA genomes, and among the most common HV1/HV2 haplotype in Caucasians (A263G; 315.1C) was also examined to determine if sequencing strategies for identifying mtDNA variation are more effective than targeting synonymous SNPs (Budowle *et al.* 2005; Coble *et al.* 2006).

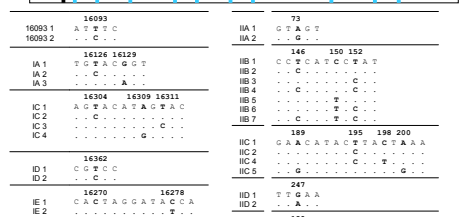
Evaluation of mtDNA Sequence Data to the Linear Array

Materials and Methods

Linear Array mitotypes for 666 samples representing African American, Caucasian, and Hispanic populations were generated using the manufacturer's protocol and the results were published in Kline *et al.* (2005). Complete control region sequence data was generated at the Armed Forces DNA Identification Laboratory (AFDIL) using the established protocol and strategy described in Brandstatter *et al.* (2004). Haplogroup associations were determined from Kivisild *et al.* (2006) and references within.

Schematic representation of mtDNA Linear Array HVI/HVII probe regions





Analysis of Null Alleles Comparing Sequence Information

A

HVII Array Locus (rCRS position)	African American	Hispanic	Caucasian	Haplogroup Associated Polymorphisms	Percentage of Null Alleles from Haplogroup Polymorphisms
16093	8	8	7	--	--
HVIA (16126; 16124)	24	2	7	16124C - L3b and L3d	27/33 (82%)
HVIC (16304; 16309; 16311)	28	38	10	16320T - L3e2 16290T and 16319A - A (Asian)	69/76 (91%)
HVID (16362)	29	3	2	16360T - L1c	29/34 (85%)
HVIE (16270; 16278)	40	11	8	16270T and 16278T - L1b 16264T - L3e4 16265T - L3e3 16265C - L1c2	33/59 (56%)
	129	62	34	225 total	158/225 (70%)

B

HVII Array Locus (rCRS position)	African American	Hispanic	Caucasian	Haplogroup Associated Polymorphisms	Percentage of Null Alleles from Haplogroup Polymorphisms
HVIIA (73)	1	0	2	72T-C - preV	1/3 (33%)
HVIIB (146; 150; 152)	46	36	15	151C-T - L1c 143G-A - L2a* 153A-G - A2; X*	88/97 (91%)
HVIIC (189; 195; 198; 200)	66	19	37	186 C-A - L1c 185G-T - L1b 189A-C - L2b/c 185G-A; 189A-G; and 200A-G - L3e* 194C-T - D*/D4b2 199T-C and 204T-C - I	78/122 (64%)
HVIID (247)	5	27	7	249A-del - CZ 242C-T - J1* 250T-C - I2*	32/39 (82%)
HVIIE (189)	103	18	32	182C-T and 195T-C L1*/L2* 185G-T - L1b 195T-C and 198C-T - L2a* 189A-C - L2b/c 185G-A and 200A-G - L3e1a 185G-A - J1* 194C-T - D*/D4b2	123/153 (80%)
	221	100	93	414 total	322/414 (78%)

Null alleles (or "blanks") are produced by a failure of PCR product to hybridize to the immobilized SSO probe of the Linear Array. We compared the Linear Array mitotypes for 666 samples (Kline *et al.* 2005) to the underlying polymorphisms from sequence data (A. HV1; B. HV2) to characterize the reasons for null alleles. We determined that a considerably large number of null alleles produced by the Linear Array were caused by the existence of haplogroup-associated polymorphisms (70% in HV1 and 78% in HV2). Haplogroups noted with an asterisk indicate instances where most, but not necessarily all, individuals within the haplogroup carry the polymorphism.

Breakdown of Discrimination among 666 population samples

A **Linear Array (HV1&HV2)** Kline *et al.* (2005) *J. Forensic Sci.* 50(2): 377-385.

Frequency: Times type observed	Frequency: % of individuals tested	Mitotypes: Number of types observed	Mitotypes: % of total mtDNA types observed
1	27.9	185	27.9
2	13.6	45	6.8
3	8.1	18	2.7
4	2.4	4	0.6
5	3.0	4	0.6
6	2.7	3	0.5
7	1.1	1	0.2
8	10.9	9	1.4
9	2.7	2	0.3
10	6.0	4	0.6
11	1.7	1	0.2
12	1.8	1	0.2
18	2.7	1	0.2
23	3.5	1	0.2
28	4.2	1	0.2
51	7.7	1	0.2
		281	

B **HV1 only (Sequence data)**

Frequency: Times type observed	Frequency: % of individuals tested	Mitotypes: Number of types observe	Mitotypes: % of total mtDNA types observed
1	50.4	334	50.4
2	11.5	38	5.7
3	5.0	11	1.7
4	4.2	7	1.1
5	3.0	4	0.6
6	4.5	5	0.8
7	7.4	7	1.1
9	1.4	1	0.2
10	1.5	1	0.2
11	3.3	2	0.3
12	1.8	1	0.2
40	6.0	1	0.2
		412	

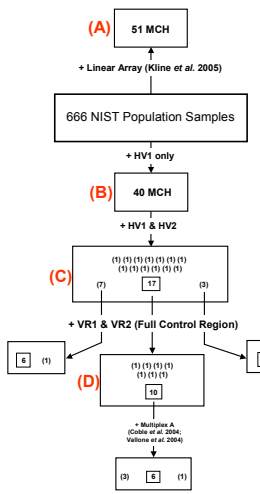
C **HV1& HV2 (Sequence data)**

Frequency: Times type observed	Frequency: % of individuals tested	Mitotypes: Number of types observed	Mitotypes: % of total mtDNA types observed
1	68.5	454	68.5
2	10.9	36	5.4
3	5.0	11	1.7
4	2.4	4	0.6
5	4.5	6	0.9
6	1.8	2	0.3
7	3.2	3	0.5
8	1.2	1	0.2
17	2.6	1	0.2
		518	

An increase in discrimination was observed when comparing the number of types assayed with the Linear Array (A) compared to sequence data generated from HV1 (B), HV1&HV2 (C), and the entire control region (D). Using the Linear Array (A) resolved the population samples into 281 different haplotypes. About 28% of the samples were observed once (185 unique individuals among the 666 population samples), with one common haplotype shared among 51 individuals (7.7% of the individuals tested). Sequence information from HV1(B), HV1&HV2 (C), and the entire control region (D) gave an increasing number of haplotypes and a decreasing number of common types (see also the schematic representation to the right).

D **Control Region (Sequence data)**

Frequency: Times type observed	Frequency: % of individuals tested	Mitotypes: Number of types observe	Mitotypes: % of total mtDNA types observed
1	75.7	502	75.7
2	11.2	37	5.6
3	4.5	10	1.5
4	0.6	1	0.2
5	3.8	5	0.8
6	2.7	3	0.5
10	1.5	1	0.2
		559	



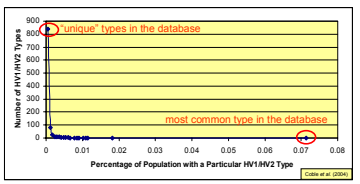
Schematic representation of the most common haplotype observed at each level of discrimination. (A). Fifty-one individuals shared the most common haplotype (MCH) using the Linear Array (Kline *et al.* 2005). (B) HV1 sequence data produced 40 individuals having the MCH. (C) HV2 Sequence information produced 13 types observed only once, one type shared by 3 individuals, one type shared by 7 individuals, and the MCH shared by 17 individuals. (D) the full control region resolved the 17 MCH from HV1/HV2 into 8 types: 7 unique types and one common type shared by 10 individuals. Applying the 11 "Multiplex A" mtDNA Single Nucleotide Polymorphisms (SNPs) of Coble *et al.* (2004) and Vallone *et al.* (2004) resolved the 10 MCH individuals from control region sequencing into 3 types: 1 unique type, one type shared by 3 individuals, and the most common type shared by 6 individuals. These results show that increased information (including coding region SNPs) were useful for improved discrimination among commonly shared mtDNA types.

An Evaluation of Coding Region mtDNA Sequence Data to Increase Forensic Discrimination – Effective Strategies

The Problem: One major disadvantage of mtDNA testing lies with the low power of discrimination for individuals that share one of the few common types. In Caucasians, the most common haplotype (MCH) occurs in ~7% of the population (H:1 – A263G; 315.1C).

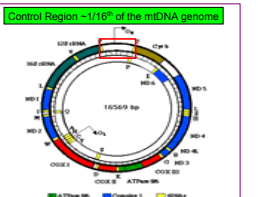
A Possible Solution: Forensic Scientists have started to examine variation in the mtDNA coding region to increase the discrimination among the common mtDNA haplotypes. Coble *et al.* (2004) and Vallone *et al.* (2004) proposed a targeted SNP approach to increase discrimination among common types. This approach considers the possibility that a limited amount of highly degraded template may only facilitate a small number of additional mtDNA tests.

Objective: Recently, Budowle *et al.* (2005) offered several criticisms to this approach (below). To evaluate these criticisms, we examined the performance of selected SNPs to a published method that sequences short mtDNA coding region fragments for increasing mtDNA discrimination (Coble *et al.* 2006).



mtDNA Population Distribution Among Caucasians (n=1665)

Control Region ~1/16th of the mtDNA genome



Criticisms of Budowle *et al.* (2005) using mtDNA coding region SNPs rather than sequencing short fragments of "informative" regions.

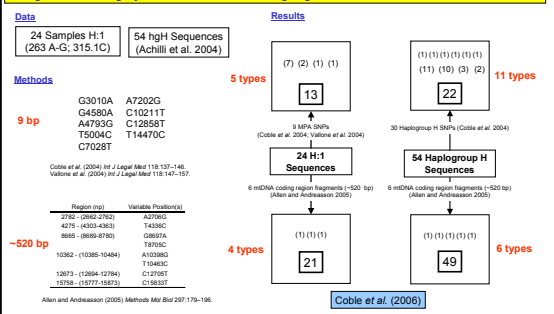
"[Coble and Vallone] have proposed that forensic analyses of the coding region [should] be restricted to synonymous substitutions [and] suggest that sequencing strategies for forensic analyses of the coding region of the mtDNA genome should be avoided [and] that only SNP-based systems should be employed."

"We disagree with this proposition as applying such a strict criterion *is not well thoughtout* and also would severely hamper the use of mtDNA in forensic testing."

"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.*

"even if only synonymous changes are being used, *since the mtDNA effectively is one locus*, such polymorphisms would still be linked to functional polymorphisms in other regions of the molecule."

An evaluation of the performance of selected SNPs compared to the sequencing of short fragments of "highly variable" mtDNA coding region DNA for increased discrimination



We sequenced 17 MCH Samples from the NIST population samples and included 7 MCH samples from Achilli *et al.* (2004). We also analyzed the haplogroup H samples from Achilli *et al.* (2004) to determine how well the 30 haplogroup H SNPs discovered by Coble et al. (2004) performed among individuals not selected for the MCH.

We found that the 9 coding region SNPs performed better as discrimination than sequencing ~520 additional bp in the coding region. Of the 24 individuals sharing the MCH, the 9 SNPs produced 5 types with 13 individuals still unresolved. The sequencing strategy produced 4 types with 21 individuals still unresolved. Similar results were observed with a set of 54 haplogroup H sequences, indicating the overall usefulness of SNPs discovered by Coble *et al.* (2004) even among sequences outside of the most common types (compared to a random sequencing strategy).

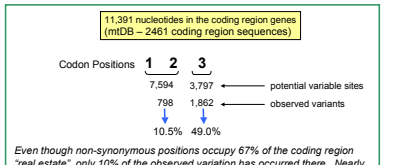
Why Did 9 SNPs Outperform Sequencing?

Region (np)*	Variable Position[b]	mtDB Frequency	Haplogroup Association	Comments
2782 - (2662-2762)[c]	A2706G	1631/2063 (79.1%)	Diagnostic H	Uninformative
4275 - (4303-4363)[d]	T4336C	24/2064 (1.2%)	"H:5"	Redundant for C456T
8665 - (8689-8780)	G8697A	109/2064 (5.3%)	Diagnostic T	Uninformative
	T8705C	15/2064 (0.7%)	Diagnostic X2	Uninformative
10362 - (10385-10484)	A10398G	890/2063 (43.1%)	Superhaplogroup N	Uninformative
	T10463C	110/2064 (5.3%)	Diagnostic T	Uninformative
12673 - (12694-12784)	C12705T	902/2063 (43.7%)	Diagnostic Superhaplogroup R	Uninformative
15758 - (15777-15873)	C15833T	20/2064 (1.0%)	"H:5"	

*Determined by dispensation order (5'-3') from Table 3 of Allen and Andreasson (2005)
[a]As determined in Table 1 of Allen and Andreasson (2005)
[b]Dispensation order for this fragment was not listed in Table 3 of Allen and Andreasson (2005), 100 bp were used as default
[c]Dispensation order in Table 3 of Allen and Andreasson (2005) did not cover the 4336 polymorphism, 60 bp were used as default

We found that the 6 "highly variable" coding regions identified by Allen and Andreasson (2005) contained several SNPs that, although observed at high frequencies, were uninformative for resolving individuals sharing the most common haplotype for haplogroup H. For example, 12705 C (rCRS) variant is shared among all Caucasians, and would offer no additional information from the sequence data generated by HV1/HV2 alone.

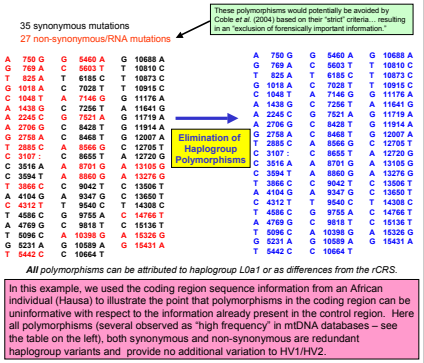
What is the Effect of Avoiding Non-Synonymous Polymorphisms?



Even though non-synonymous positions occupy 67% of the coding region "real estate", only 10% of the observed variation has occurred there. Nearly half of all potential synonymous positions in a global database of 2461 coding region sequences have been observed, making synonymous positions a better repository of genetic variation.

Using the information from mtDB (http://www.genpat.uu.se/mtDB/), we found that the overall percentage of variants observed at non-synonymous 1st and 2nd codon positions (among 2461 coding region sequences) to be significantly lesser than synonymous 3rd position codons – despite the fact that synonymous codons are outnumbered 2:1. Of the 65 synonymous polymorphisms occurring at a frequency of 1% or greater in the global database, **all** could be readily assigned to a haplogroup.

A mtDNA genome from an African-derived sequence (Hausa) belonging to haplogroup L0a1 (Ingman et al. 2000)



***All** polymorphisms can be attributed to haplogroup L0a1 or as differences from the rCRS.*

In this example, we used the coding region sequence information from an African individual (Hausa) to illustrate the point that polymorphisms in the coding region can be uninformative with respect to the information already present in the control region. Here all polymorphisms (several observed as "high frequency" in mtDNA databases – see the table on the left), both synonymous and non-synonymous are redundant haplogroup variants and provide no additional variation to HV1/HV2.

Summary

- Increased sequence information, especially complete control region data, provides increased discrimination.
- A knowledge of diagnostic haplogroup polymorphisms can be useful in understanding the underlying causes of null alleles in Linear Arrays. Rather than being thought of as "blanks", null alleles may be useful for determining biogeography ancestry based on haplogroup-associated polymorphisms.
- We found that high frequency polymorphisms are not necessarily a reliable indicator of "informativeness" when additional discrimination in the coding region is desired because many of these sites are associated with haplogroups, and therefore offer very little information beyond HV1/HV2.

For more information, see STRBase:
http://www.cstl.nist.gov/biotech/strbase/mtDNA.htm

Acknowledgments

•Funding from the National Institute of Justice through the NIST Office of Law Enforcement Standards
•Margaret Kline, Jan Redman, Peter Vallone, and John Butler (NIST) for assistance in Linear Array analysis.
•Jodi Irwin, Rebecca Just, Toni Diegoli, Brion Smith, Tom Parsons, and members of the AFDIL Research Section for the high quality sequence data used in this study.