Int J Legal Med (2004) 118 : 137–146
DOI 10.1007/s00414-004-0427-6

ORIGINAL ARTICLE

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

Received: 25 August 2003 / Accepted: 5 January 2004 / Published online: 4 February 2004

Abstract We have sequenced the entire mtDNA genome (mtGenome) of 241 individuals who match 1 of 18 common European Caucasian HV1/HV2 types, to identify sites that permit additional forensic discrimination. We found that over the entire mtGenome even individuals with the same HV1/HV2 type rarely match. Restricting attention to sites that are neutral with respect to phenotypic expression, we have selected eight panels of single nucleotide polymorphism (SNP) sites that are useful for additional discrimination. These panels were selected to be suitable for multiplex SNP typing assays, with 7–11 sites per panel. The panels are specific for one or more of the common HV1/HV2 types (or closely related types), permitting a directed approach that conserves limiting case specimen extracts while providing a maximal chance for additional discrimination. Discrimination provided by the panels reduces the frequency of the most common type in the European Caucasian population from ~7% to ~2%, and the 18 common types we analyzed are resolved to 105 different types, 55 of which are seen only once.

Keywords Human mitochondrial DNA genome · Single nucleotide polymorphism · Forensic DNA testing · Increased forensic discrimination · mtDNA coding region

M. D. Coble · R. S. Just · J. E. O'Callaghan · I. H. Letmanyi
C. T. Peterson · J. A. Irwin · T. J. Parsons (✉)
The Armed Forces DNA Identification Laboratory, Building 101, 1413 Research Blvd., Rockville, Maryland 20850, USA
Tel.: +1-301-319-0268, Fax: +1-301-295-5932,
e-mail: parsons@afip.osd.mil

M. D. Coble
The George Washington University Graduate Program in Genetics, Washington, DC, USA

C. T. Peterson
The Institute for Genomic Research, Rockville, Maryland, USA

Introduction

The strengths of mitochondrial DNA (mtDNA) forensic testing are well established, primarily in the three following areas: 1) successful recovery from highly degraded sources where nuclear DNA typing may fail, 2) application to cases where the only reference samples available are from matrilineal relatives, and 3) application to samples where nuclear DNA is virtually absent (e.g. shed hairs) (Holland and Parsons 1999). Equally well established is mtDNA's greatest weakness: its relatively low power of discrimination. The maternal inheritance of mtDNA means that an individual will, barring mutation, match all maternal relatives, of which there may be many in the population. At a grand scale, from the standpoint of mtDNA, all people are maternal relatives, descended from a common ancestral mtDNA type that was present – along with other mtDNA lineages that have since become extinct – in the human population of sub-Saharan Africa, some ~150,000–200,000 years ago (reviewed in Wallace et al. 1999). It is the pattern of mutations that have accumulated over time along the transmission of this mtDNA line that give rise to the differences among us, and that are the basis for forensic discrimination. During this time, people and populations have migrated and expanded, leading to the geographic divergence of lineages, and chance increases of some mtDNA types within descendant populations due to the founder-effect. These processes are reflected in a quite uneven distribution of sequence types in various populations, with the result that mtDNA's greatest forensic limitation (low power of discrimination) is not uniformly applied to all. In populations studied to date, the distribution is such that some mtDNA "types" are quite common, while many are quite rare.

Currently mtDNA forensic testing consists primarily of sequence analysis of portions of the control region, most often targeting hypervariable regions one and two (HV1/HV2) for a total of 610 base pairs. Other regions within the control region are sometimes targeted (Lutz et al. 2000), and only recently has the suggestion been made that variation within the mtDNA coding region could also

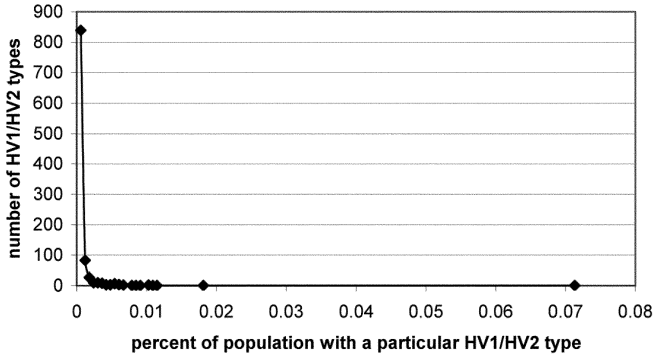

Fig. 1 Graph of HV1/HV2 type distribution in the Caucasian population. Data is taken from a pairwise comparison of a database of 1655 European Caucasians (Monson et al. 2002), and does not consider variation in the C-stretch region from bases 303–309

have forensic utility (Parsons and Coble 2001; Lee et al. 2002; Lutz-Bonengel et al. 2003). With reference to the HV1/HV2 sequencing that is most commonly performed, we can examine the random match probability within the Caucasian population. As suggested above, there is a highly skewed distribution, with the most common HV1/HV2 type occurring in ~7% of the population, while fully 50% of the population is "unique in the database" (Fig. 1). While most sequences are very rare and of correspondingly high evidentiary value, only a small portion of HV1/HV2 types are present at ~0.5% or greater in the population. It is in these common types, then, that the low power of discrimination of mitochondrial DNA is most strongly manifested.

We have undertaken to address the limitations of mtDNA discrimination by focusing specifically on those HV1/HV2 types where the limitation is most significant.

We report here the results of high throughput sequencing of the entire mtDNA genome (mtGenome) from multiple individuals who match 18 of the most common HV1/HV2 types in the Caucasian population, with the aim of identifying sites outside of HV1/HV2 that provide maximal discrimination among individuals who would otherwise match. These 18 common HV1/HV2 types comprise ~21% of the Caucasian population (Table 1). We will present panels of selected single nucleotide polymorphism (SNP) sites that are well suited for multiplex SNP assays and practical forensic applications. These SNP panels can be assayed in a directed manner to complement HV1/HV2 sequencing and greatly increase the power of mtDNA forensic testing in Caucasians.

Materials and Methods

Samples

Pairwise comparisons of Caucasian HV1/HV2 sequences from an mtDNA population database (Monson et al. 2002) posted on the Internet permitted a frequency distribution of types to be determined (Fig. 1). Throughout this paper, HV1/HV2 refers to the fol-

Table 1 Common HV1/HV2 type designations, sample sizes, population frequencies, and list of HV1/HV2 sequences targeted in this study

N	HV1/HV2 type	Frequency[a] (%)	HV1/HV2 sequence (+263G, 315.1C)
31	H:1	7.1	CRS
25	H:2	1.8	152 C
11	H:3	0.8	16129 A
8	H:4	0.5	16263 C
12	H:5	0.9	16304 C
11	H:6	0.9	73 G
7	H:7	0.7	16162 G 16209 C 73 G
15	J:1	1.1	16069 T 16126 C 73 G 185 A 228 A 295 T
8	J:2	0.6	16069 T 16126 C 73 G 228 A 295 T
13	J:3	0.7	16069 T 16126 C 73 G 185 A 188G 228 A 295 T
8	J:4	0.5	16069 T 16126 C 16145 A 16172 C 16222 T 16261 T 73 G 242 T 295 T
21	T:1	1.1	16126 C 16294 T 16296 T 16304 C 73 G
10	T:2	0.6	16126 C 16163 G 16186 T 16189 C 16294 T 73 G 152 C 195 C
8	T:3	0.5	16126 C 16294 T 16296 T 73 G
25	V:1	1.0	16298 C
14	K:1	1.0	16224 C 16311 C 73 G 146 C 152 C
7	K:2	0.5	16093 C 16224 C 16311 C 73 G
7	K:3	0.5	16224 C 16311 C 73 G
Total		20.8	

Sample size (*N*) refers to the number of individuals that were sequenced over the entire mtGenome.
Sequences are listed as differences from the Cambridge Reference Sequence, except the polymorphisms 263G and 315.1C are omitted as they are shared by all.
Length variation in the 303–309 C-stretch region is ignored.
HV1/HV2 refers to the ranges 16024–16365 and 73–340.
Haplogroup-associated polymorphisms are indicated in bold.
[a]Frequency in the European Caucasian population was determined from HV1/HV2 data for 1,655 Caucasians in a mtDNA population database (Monson et al. 2002).

lowing sequence ranges: HV1=16024–16365 and HV2=73–340. In determining "common" HV1/HV2 types to target for mtGenome sequencing, we chose a population frequency of 0.5% as the lower cut-off. Given the skewedness of the distribution (Fig. 1), this cut-off targets an appreciable proportion of the "problem" HV1/HV2 types. For our purposes, HV1/HV2 sequences were not distinguished on the basis of C-stretch length variants in the HV1 or HV2 regions. This follows accepted forensic practice of not using these unstable, often length-heteroplasmic characters for the purposes of exclusion (Holland and Parsons 1999; Stewart et al. 2001). Having identified common HV1/HV2 types, samples matching these types were accessed from blood samples collected by the Armed Forces DNA Identification Laboratory (AFDIL), for which appropriate consent was obtained. Unfortunately, multiple samples were not available from all the common HV1/HV2 types that we would have chosen to target, and our sampling was necessarily constrained by availability. The genetic privacy of individuals whose entire mtGenome was sequenced was protected by stripping the samples of all identifying information and randomizing them prior to mtGenome sequencing. The study was approved by the Armed Forces Institute of Pathology Institutional Review Board for the Use of Human Subjects.

Table 1 indicates the common Caucasian HV1/HV2 types that were targeted in this study. Based on the HV1/HV2 sequence, each type was categorized to an mtDNA haplogroup (Torroni et al. 1994, 1996; see also Macaulay et al. 1999). This assignment, however, was tentative because haplogroups cannot be unambiguously assigned by HV1/HV2 sequence alone. The 18 common HV1/HV2 types represent 5 Caucasian haplogroups: H, J, T, V and K (Table 1). Each of these HV1/HV2 types is designated by its haplogroup followed by a number (e.g. H:1). The colon is included in this convention to avoid confusion with accepted designations for mtDNA sub-haplogroups such as H1 etc. (e.g. Finnila et al. 2001).

Pairwise comparisons and population frequencies

Pairwise comparisons of Caucasian HV1/HV2 sequences from a forensic database (Monson et al. 2002), and entire mtGenome sequences determined in this study were performed using the AFDIL programs LookADat (J. Irwin, AFDIL, unpublished) and LISA (Laboratory Information Systems Applications), with C-stretch length differences ignored in the comparisons. Frequencies of HV1/HV2 types were obtained from the output of the pairwise comparisons; it is appropriate to exclude the 309 C-stretch region in distinguishing individuals because of its unstable hypermutability and frequent length heteroplasmy (Stewart et al. 2001).

DNA extraction

Total genomic DNA was extracted from bloodstains dried on paper cards (Fitzco, Minneapolis, MN). For 30 of the samples, 3 mm paper punches from the cards were extracted with a 5% (w/v) final concentration of Chelex 100 beads (BioRad Laboratories, Hercules, CA) following published protocols (Walsh et al. 1991). The remaining 211 samples were extracted robotically as follows: a Wallac DBS puncher (Perkin-Elmer Life Sciences, La Jolla, CA) was used to obtain single 3 mm blood stained paper punches, arrayed in a 96-well plate for extraction. These were extracted using the QIAmp 96-well DNA swab extraction kit (Qiagen, Gaithersburg, MD) on the Qiagen BioRobot 9604, following the manufacturers recommendations for liquid blood, but with 45 min lysis incubation instead of the suggested 15 min incubation.

MtGenome sequencing and data quality control

Amplification and cycle sequencing of the mtGenome was performed as previously described (the AFDIL protocol presented in Levin et al. 2003), with automated fluorescent sequencing on either the Applied Biosystems 377 or 3100 instruments. Sequence determination for any particular nucleotide position required clear confirmation from both strands, or in rare cases from a single strand if confirmed from multiple clear sequence reactions. Consensus sequences were generated using Sequencher software (GeneCodes, Ann Arbor, Michigan, USA). The magnitude of sequence data produced in this project required careful attention to avoid errors in sequence analysis, transcription, and manipulation, i.e. "phantom mutations" (Bandelt et al. 2002; see also Herrnstadt et al. 2003). Sequencher projects were completely reviewed by two individuals who independently tabulated a list of polymorphisms relative to the revised Cambridge Reference Sequence (Anderson et al. 1981; Andrews et al. 1999). Early in the project, entry of polymorphic differences into a master database was performed using a custom graphical user interface that required independent confirmation that the sequences had been correctly entered, after which the sequences in the database were locked to prevent subsequent corruption or overwriting. Later in the project, entry into the database was accomplished by direct export from Sequencher, still with redundant independent reviews of the Sequencher project and the exported list of polymorphisms. The position of polymorphisms within genes and proteins, as well as whether polymorphic variants were synonymous or non-synonymous, was determined using the web program MitoAnalyzer (Lee and Levin 2002; http://www.cstl.nist.gov/biotech/strbase/mitoanalyzer.html). The mtGenome sequences determined in this study are listed in GenBank, accession numbers AY495090-AY495330.

Results and discussion

MtGenome sequence variation

Our rather brute force strategy of sequencing the entire mtGenome for individuals who would otherwise match common HV1/HV2 types, in order to identify SNPs of maximal forensic utility, was based on two suppositions that have, in the end, proven to be correct: 1) that despite having an evolutionary rate ~4-fold lower than the non-coding control region (Aquadro and Greenberg 1983; Horai and Hayasaka 1990), the ~15-fold larger coding region should contain much additional sequence variation that can discriminate among individuals who match in the control region, and 2) that the coding region sites that permit such discrimination could not confidently be divined by inspection of coding region RFLP data, or recently published entire mtGenome data from diverse or random population samples (Ingman et al. 2000; Finnila et al. 2001; Maca-Meyer et al. 2001; Herrnstadt et al. 2002; Ingman and Gyllensten 2003). The latter is because coding region sites that distinguish among individuals who match in the rapidly evolving control region, often reflect fortuitous mutational events that are specific to the HV1/HV2 type (or closely related types) in question.

It is unusual for individuals to match over the entire mtGenome, even when they are identical in HV1/HV2. Within the 241 individuals in this study, representing 18 common HV1/HV2 types, a total of 209 haplotypes were observed when the entire mtGenomes were sequenced; only 32 of the 241 (13%) individuals matched one or more other individuals for the entire mtGenome (Table 2). The greatest number of differences seen within an HV1/HV2 type is 22, occurring between 2 individuals of type K:3 (16224C, 16311C, 263G, 315.1C), but the average number of differences between individuals who match in

Table 2 Entire mtGenome discrimination of common HV1/HV2 types

HV1/HV2 type	*N*[a]	No. of mtGenome Types[b]	Maximum no. of differences[c]	Average no. of differences[d]	Type distribution[f]
H:1	31	28	16	7.5	3,2,1(26)
H:2	25	24	19	7.5	2,1(23)
H:3	11	10	13	6.8	2,1(9)
H:4	8	5	2	1.2	4,1(4)
H:5	12	12	12	5.8	1(12)
H:6	11	9	19	8.8	3,1(8)
H:7	7	5	5	2.2	3,1(4)
J:1	15	14	13	6.5	2,1(13)
J:2	8	6	10	6.7	2(2),1(4)
J:3	13	10	10	4.4	4,1(9)
J:4	8	8	9	4.6	1(8)
K:1	14	10	14	6.6	4,2,1(8)
K:2	7	5	12	8.4	3,1(4)
K:3	7	6	22	8.9	2,1(5)
T:1	21	18	10	3.6	4,1(17)
T:2	10	9	6	3.0	2,1(8)
T:3	8	7	14	5.8	2,1(6)
V:1	25	23	17	6.2	3,1(22)
Total	241	209	22	6.2[e]	4(4),3(5),2(10), 1(190)

[a]Sample size of individuals sequenced.
[b]Number of distinct haplotypes observed after mtGenome sequencing.
[c]Maximum number of differences observed between any two sequences matching for HV1/HV2.
[d]Average number of differences between individuals for each HV1/HV2 type.
[e]Average number of mtGenome differences between individuals who match for HV1/HV2, within (but not between) all of the common HV1/HV2 types.
[f]Distribution of mtGenome haplotypes is indicated by the number of matching individuals, following numbers in parentheses indicate the number of different haplotypes involved, for example, 3,2, 1(26) indicates 1 haplotype with 3 matching individuals, 1 haplotype with 2 matching individuals, and 26 haplotypes represented by a single individual.

HV1/HV2 is 6.2. Different common HV1/HV2 types show variable degrees of diversity at the level of the entire mtGenome. For example, none of the 12 individuals with HV1/HV2 type H:5 matched one another, with an average number of ~6 differences between individuals, and a maximum number of 12 differences between individuals. This can be compared to the 8 H:4 individuals, where 4 individuals matched exactly, and the average number of differences between individuals was ~1, with a maximum of 2 differences. Given the sample sizes, these differences may be due at least in part to sampling effects, but such differences could also be expected to reflect differences in the time since divergence among the sequences that have been grouped according to HV1/HV2 type. It is worth noting that homoplasy (multiple independent mutations giving rise to the same character state at a nucleotide position) is common in HV1/HV2 (e.g. Meyer et al. 1999), so it is likely that mtDNA sequences grouped together by HV1/HV2 types will sometimes not all be close relatives with a single common ancestral sequence. In the terminology of evolutionary systematics, some of the HV1/HV2 groupings are likely to be either polyphyletic or paraphyletic. This could also contribute significantly to the differing amounts of mtGenome diversity seen within the HV1/HV2 types.

MtDNA variation within humans has been intensively studied, and phylogenetic analysis permits mtDNA types to be classified into defined haplogroups on the basis of evolutionary relatedness and the pattern of shared polymorphic variants that reflect shared evolutionary history (e.g. Wallace et al. 1999). Having the entire mtGenome sequences allows us to easily ascribe the sequences in our study to their correct haplogroup, according to defining constellations of polymorphisms (e.g. Herrnstadt et al. 2002). Our initial tentative haplogroup classification based on HV1/HV2 polymorphisms was largely accurate, with several exceptions. European haplogroup V is closely related to haplogroup H, and in HV1/HV2, haplogroup V is associated with the polymorphism 16298C. This is the only characteristic polymorphism within HV1/HV2 that distinguishes V from H, and homoplasious variation at this site caused us to misclassify (misname) haplogroups for three individuals within this study. Individual V:1-11 matches all the other V:1 individuals in HV1/HV2 (as per the conceptual definition of an HV1/HV2 type), all of which have the 16298T-C variant. However, V:1-11 lacks the polymorphisms 4580A, 7028T, and 15904T that are indicative of a true member of haplogroup V. V:1-11 instead shows all the coding region hallmarks of haplogroup H. This indicates that the V:1-11 lineage experienced a mutation at 16298, causing it to look like a V, while having arrived at that appearance independently of other (true) V's. Similarly, individuals H:1-20 and H:1-29 are in fact members of haplogroup V as seen by coding region polymorphisms, but have lost their characteristic 16298C due to homoplasious mutation (not shown).

The average number of sequence differences over the entire mtGenome among people who match for HV1/HV2 does not differ dramatically among the haplogroups (these values are: haplogroup H, 6.9; haplogroup J, 5.6; haplogroup K, 7.2; haplogroup T, 3.7; and haplogroup V, 6.2), in contrast to the situation for particular HV1/HV2 types within haplogroups described above. We can also compare the amount of variation observed within the various genes/regions of the mtGenome (Table 3). On a per-site basis, the least variable regions are the ribosomal RNA genes, followed by the genes encoding the 22 transfer RNAs. Outside of the control region, there are 91 bases scattered throughout the "coding region" that are short non-coding intergenic spacers. Taken together, these positions represent the most variable portion of the mtGenome outside of HV1/HV2, with variation observed at nearly 7% of these sites. The most variable gene in this study codes for ATP synthase 8, where nearly 5% of the sites are variable. Interestingly, the parts of the control region residing outside of HV1/HV2 do not have a dramatically higher proportion of discriminatory sites; the value of 4.1% is higher than almost all of the protein coding genes, but is less than twice

Table 3 Variation of discriminatory sites tabulated by gene/region

Gene/region	Length of region	Number of discriminatory sites	%[c]
CR outside of HV1/HV2	513	21	4.1
All tRNA's combined	1507	35	2.3
NC region outside of CR	91	6	6.6
12 s rRNA	954	13	1.4
16 s rRNA	1559	16	1.0
NADH dehydrogenase 1	957	22	2.3
NADH dehydrogenase 2	1044	37	3.5
Cytochrome C oxidase I	1542	38	2.5
Cytochrome C oxidase II	684	18	2.6
ATP synthase 8	207	10	4.8
ATP synthase 6	681	25	3.7
Cytochrome C oxidase III	783	18	2.3
NADH dehydrogenase 3	348	8	2.3
NADH dehydrogenase 4L	297	6	2.0
NADH dehydrogenase 4	1380	33	2.5
NADH dehydrogenase 5	1812	56	3.1
NADH dehydrogenase 6	525	15	2.9
Cytochrome b	1137	35	3.1
Total	15959[a]	410[b]	2.6[d]

Discriminatory sites are sites that vary within one or more common HVI/HVII types.

[a]The numbers for length of region do not sum to the value listed as Total because some of the genes in the mtGenome overlap for short stretches.

[b]The numbers for discriminatory sites do not sum to the value listed as Total because 2 of the discriminatory sites are in the aforementioned overlapping regions.

[c]Percent is taken as the number of sites that show discriminatory variation divided by the length of the gene/region ×100.

[d]Total percent is taken as the total number of sites that show variation outside the HVI/HVII regions divided by the total number of bases in the mtGenome outside HVI/HVII.

the average value of 2.5% for the entire mtGenome outside of HV1/HV2. This likely reflects evolutionary constraint on sequence variation in the control region in relation to its regulatory function (e.g. Saccone et al. 1991), even though it is non-coding.

Our results show that the control region outside of HV1/HV2 contains sites with the highest level of polymorphism both within and between the common HV1/HV2 types. The most highly variable SNP observed in this study is position 16519, lying between HV1 and HV2. Position 16519 varied in 9 of the 18 HV1/HV2 types, and therefore is of great utility in providing additional forensic resolution. A second highly polymorphic element of the control region is not a SNP, but a previously characterized dinucleotide repeat of AC occurring between bases 515 and 524 (Bodenteich et al. 1992). Insertion or deletion variants of the normal $(AC)_5$ repeat number are seen in 11 of the 18 common HV1/HV2 types. This repeat might be considered the mitochondrial version of an STR. This AC repeat and 16519 are evidently examples of mutational hotspots, where changes to and from alternate variants occur regularly across a range of haplotypic backgrounds.

Forensic utility

We have established that there is much sequence variation outside HV1/HV2 in the mtGenome that can distinguish between individuals that would otherwise match in standard forensic testing. However, a question of great importance is whether this variation is distributed in a manner consistent with development of practical forensic testing. Sequencing the entire mtGenome in forensic casework is obviously impractical. The ultimate goal of this project was to determine if we could identify particular combinations of SNP sites that vary in a manner that makes them especially useful, at least in selected circumstances, with subsequent development of forensically practical multiplex assays. To select sites to target for practical SNP assays, we employed the following basic criteria: 1) the sites selected must be neutral from the standpoint of genetic disease or other phenotypic expression, 2) the sites selected have to vary in multiple individuals, and 3) the sites selected are not redundant with each other in terms of forensic discrimination, and work in combination with other SNPs for maximum forensic discrimination given a minimum number of SNP sites.

The issue of phenotypic neutrality is not usually considered in mtDNA testing, but must be dealt with as we venture into the coding region, where sequence variation can have direct phenotypic manifestation (see also Lutz-Bonengel et al. 2003). All the genes in mtDNA are involved either in ATP production via oxidative respiration, or for the RNA genes, in the basic machinery of protein synthesis within the mitochondrion. Given the critical nature of mitochondrial oxidative respiration to cellular energy production, mutations that change the amino acid sequence in proteins or the secondary structure of transfer or ribosomal RNAs have potential to be of medical or physiological significance. In fact, many maternally transmitted diseases are known to be caused by specific mtDNA mutations (reviewed in Wallace 1999), and susceptibility to important widespread conditions such as diabetes (Cavelier et al. 2002), Parkinson's disease (Brown et al. 1996), and Alzheimer's disease (Lin et al. 2002) have also been associated with mtDNA sequence variation. Further details concerning mtDNA and disease are beyond the scope of this paper, but it can be said with certainty that not all disease associations with mitochondrial DNA mutations/polymorphisms have yet been discovered or cataloged (however, a growing list is maintained at the MitoMap web site: http://www.mitomap.org/). Therefore, unless an appropriate awareness is applied as we turn to the mtDNA coding region, it is quite possible that mtDNA forensic testing could inadvertently land investigators into serious situations involving the ethics of medical genetics. In a missing persons' case, for example, how is one to avoid discovering within a family reference sequence a polymorphism that is correlated with an increased risk of Alzheimer's disease? If discovered, is this information shared with the individual or family? Is there a practical and ethical way to ensure that such information would never be obtained by the family? Our conclusion is that any attempt

to access coding region variation should be restricted to particular SNP sites that have no potential for phenotypic expression or, in restricted instances, to sites where neutrality can be convincingly inferred. This eliminates standard sequencing as an approach for assaying the SNP sites, as well as other assays such as pyrosequencing (Andreasson et al. 2002) that generate substantial genetic information adjacent to the SNP of interest.

Application of the criterion of neutrality for target SNP sites, then, restricted our attention to 1) the control region outside of HV1/HV2, 2) short intergenic regions throughout the coding region, or 3) to synonymous substitutions in protein coding genes (however, see below for an exception). Synonymous substitutions occur within codons of protein coding genes at positions where, even though the DNA sequence has changed, the same amino acid will be incorporated into the protein product. The degeneracy of the mitochondrial genetic code is such that many third codon position and several first codon position substitutions are synonymous (or "silent").

Regarding our other criteria for inclusion as a target SNP site, the requirement that the site has been observed to vary in more than one individual simply seeks to avoid sites where the polymorphisms are "private," i.e. restricted to one or a very small number of individuals. The third criterion, that of eliminating redundant sites and incorporating those that work most advantageously in combination, was the most complex to implement. Because of the linkage of mtDNA mutations, multiple sites would frequently separate out the same set of individuals within a common HV1/HV2 type. To choose among those, comparisons were made to other HV1/HV2 types, and the most widely varying sites were selected. In our study we identified 59 SNP sites that satisfied our criteria for useful and appropriate forensic markers (Table 4). Of these 8 reside in the control region outside of HV1/HV2, 1 in a coding region intergenic spacer, 49 in silent codon positions in protein coding genes, and 1 in the 16S rRNA. The latter is position 3010, and is an exception to our criterion that sites do not reside within rRNA or tRNA genes because of the potential for phenotypic effects. The basis for the exception is: 1) position 3010G/A is a well characterized polymorphism, reported widely in the literature, and not suspected to be associated with disease or other manifestation (e.g. Mehta et al. 1989; Finnila et al. 2001), and 2) it is extremely useful in forensic discrimination, particularly in haplogroup H. Regarding 3010, we make reference to the neutral theory of molecular evolution (Kimura 1983) to reassure us of the neutrality of such a widespread, easily detectable polymorphism.

Table 4 lists the 59 forensic markers we selected, separated into 8 multiplex panels (A–H) that target particular HV1/HV2 types. This arrangement into separate panels reflects practical considerations that are not necessarily universal, but that we thought would provide general utility for readily available SNP assay methods. Clearly, for forensic purposes, multiplex assays are required to avoid excessive demands on sample extract volume that can be assumed to be limiting. On the other hand, multiplex assays can be challenging to develop, and an assay for all the sites together might be years in development and involve platforms unavailable to many laboratories. We have sought practical middle ground in batching the sites into panels of 7–11 sites, each to be applied for particular common HV1/HV2 types in question. This is in line with our experience with several SNP assay platforms that can be developed for approximately 10 sites with relative ease. One of these, an allele-specific primer extension (ASPE) assay using SNaPShot reagents (Applied Biosystems, Foster City, CA) has been optimized for multiplex A (see Vallone et al. 2004). The assay performs well from the standpoint of the following important practical forensic considerations: operation on standard forensic instrumentation, sensitivity, mixture and heteroplasmy detection, and robust performance on degraded casework samples.

The multiplex panels are designed to complement the results of HV1/HV2 testing for increased discrimination, permitting the most appropriate group of SNPs to be applied to a particular common HV1/HV2 type. For example, if the most common Caucasian type H:1 is deter-

Table 4 Multiplex panels A–H of SNP sites

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	
H:1	H:2,H:3,H:6	V:1,H:5	J:1,J:2,K:2,K:3	J:4,T:2,T:3,H:4	V:1,H:1,H:2,H:3	J:1,J:3,T:1	K:1

The panels of sites were assembled to provide maximal resolution for the common HV1/HV2 types (see Table 1) listed at the bottom of each column.

Table 5 Discrimination of 241 common HV1/HV1-type individuals resolved by the 8 multiplex panels (left) or the eight panels plus the 515–525 AC polymorphism (right)

8 Multiplexes		8 Multiplexes + AC indel polymorphism	
No. of types	No. of individuals/type	No. of types	No. of individuals/type
2	14	1	14
1	9	1	13
3	8	1	9
2	7	1	8
1	6	3	7
3	5	2	6
3	4	3	5
8	3	3	4
27	2	7	3
55	1	26	2
		64	1

Table 6 Resolution of 81 H:1 individuals with the application of multiplex panel A

Multiplex A	
No. of types	No. of individuals/type
1	22
1	20
1	8
1	7
3	4
1	3
3	2
3	1

mined from HV1/HV2, a single amplification with multiplex A provides the best opportunity for further discrimination. However, for H:1 and some other HV1/HV2 types, not all discriminatory SNPs were able to be placed in a single multiplex panel, so secondary panels are also available. In the case of H:1, if multiplex A does not provide sufficient resolution, a second amplification with multiplex F would target additional potentially discriminatory sites. Sites are sometimes included in multiple different multiplexes. This redundancy is included so that for any given HV1/HV2 type one could turn to a single multiplex that gives the greatest chance for resolution.

How much additional forensic discrimination would the multiplex panels provide? The 241 individuals in our study could, by standard HV1/HV2 sequencing, be divided into only 18 types; these types together comprise some 21% of the Caucasian population, and the most common type occurs in 7% (Table 1). If we apply the multiplex panels (Table 4) to these individuals, the result is 105 distinguishable types (Table 5) of which 55 are unique in the group, and therefore in most instances likely to be unique in the forensic database (N=1,655) from which they were sampled. We have applied the multiplex A ASPE assay (Vallone et al. 2004) to another 50 unrelated individuals matching H:1. With the addition of the 31 H:1's sequenced for the mtGenome we can evaluate the effect of multiplex A on 81 H:1 individuals (Table 6). This results in an average population frequency for the SNP-resolved H:1's of 0.5%, for a 16-fold improvement on average, with a ~4-fold decrease in frequency (22/81, see Table 6) of the new most common type. Roughly similar resolution is provided by the multiplex panels for the other common HV1/HV2 types, with the notable exception of H:7 that unfortunately remains a single type even after application of the SNP panels.

The multiplex panels in Table 4 do not include the extremely useful polymorphic AC repeat around 515–524 of the control region. While it is not actually a SNP, it nevertheless could be typed by any number of stand-alone assays, to include simply sequencing a short amplicon that spans the repeat (in the control region, we are absolved from issues of neutrality, at least to the same extent as current mtDNA testing). We are in the process of optimizing a modified primer extension assay that can be run together with the SNaPShot multiplexes. Doing so would be worthwhile: adding the AC indel polymorphism together with the multiplex panels takes the 241 individuals from 18 to 112 different types, 64 of which are unique (Table 5).

Due to limitations of sample availability, our study did not directly investigate all the common HV1/HV2 types in European Caucasians. However, one would predict that at least some of the same sites would discriminate within closely related HV1/HV2 types. This proved to be the case for many of the sites we have identified, as exemplified by the significant overlap of sites chosen for multiplexes A and B that target haplogroup H. Types H:1 and H:2 differ only at the extreme hotspot position 152 (Meyer et al. 1999; Allard et al. 2002; Malyarchuk et al. 2002), which is actually homoplasious within the H:1 and H:2 categories (not shown). The result is that largely the same SNPs distinguish within H:1 and H:2. The most significant common Caucasian type that we did not investigate is (146C, 263G, 315.1C), present in 0.6% of the population. However, this differs from H:1 only at 146, another hotspot (Meyer et al. 1999; Allard et al. 2002; Malyarchuk et al. 2002), and we would expect multiplexes A and B to be useful in that case as well.

In undertaking this study our aim was not to "reinvent the wheel" for mtDNA testing, but to provide data and approaches that would complement and strengthen current practices. For one thing, shifting completely to another set of markers would require establishment of new databases, which for mtDNA should be quite large. In fact, maximal use of the SNPs we present here also requires databasing for reporting evidentiary significance, and the AFDIL laboratory is in the process of adding these SNPs to current control region sequence databases. However, even before such databases become available, the ability to discriminate between matching HV1/HV2 types has many important applications. Among those are distinguishing among multiple matching suspects, distinguishing between victim and suspect, and distinguishing among multiple matching families in mass disasters or missing persons projects.

It would, in fact, be possible to develop an mtDNA testing system based solely on coding-region SNPs, or on a combination of control region and coding region SNPs. The sites we have identified would be poor choices for stand-alone assays in the absence of any control region data, but we are optimistic they will also be useful in combination with control region SNPs. At least two SNP-based mtDNA typing systems for HV1/HV2 polymorphisms have been developed (Gabriel et al. 2003; see also the commercially available kit associated with the Luminex 100 system http://www.marligen.com/products/signetmito.htm and Armstrong et al. 2000). The mtDNA types determined from these tests could be classified according to their corresponding common HV1/HV2 sequence types, and this classification used for the selection of the appropriate multiplex SNP panel. Even though control region SNP assays do not provide as much information as sequencing in HV1/HV2, the fact that discriminatory sites in our study are often useful within the same haplogroup will likely permit them to be of substantial utility when combined with control region SNPs. Likewise, our SNP panels could also complement partial control region sequencing, such as HV1 alone. In this regard, it is important to note that SNP assays such as ASPE can be substantially easier and faster to run than full sequencing (see Vallone et al. 2004). That said, the most powerful of the approaches we have discussed remains that of full HV1/HV2 sequencing together with the SNPs we report here.

Brandstätter et al. (2003) have recently reported a multiplex SNP system for categorizing European Caucasian mtDNAs to their correct haplogroup. The SNPs reported by Brandstätter et al. (2003) are from the coding region and are also typed using the SNaPShot ASPE assay. These haplogroup-associated SNPs have a stand-alone discriminatory power within the European Caucasian population that rivals the control region SNP system used in Gabriel et al. (2003). Again, given that our discriminatory SNP multiplexes have the potential to discriminate within multiple common types within a haplogroup, we are hopeful that they could powerfully complement the haplogroup typing system of Brandstätter et al. (2003). An added advantage of that combination is that typing would occur on a single, convenient capillary electrophoresis platform already present in many forensic laboratories that do not currently perform mtDNA testing.

The work reported here, together with other recent publications (Lee et al. 2002; Lutz-Bonengel et al. 2003), represents a strong beginning for efforts to tap information from the entire mtGenome for forensic purposes. Continued efforts to identify additional useful SNP sites for Caucasians, as well as to target other populations, will further add to the strength of mtDNA testing. The sites we have included for development of "value added" SNP multiplexes were selected with a specific operational approach in mind (as described above, to be used in combination with HV1/HV2, to resolve the most troublesome common types), and are based on the somewhat limited sample sizes that were available to us. In the future, it will be important to synthesize as much information as possible, both from directed studies such as ours, as well as general coding region population data coming from forensic (Tzen et al. 2001; Lee et al. 2002; Lutz-Bonengel et al. 2003) or academic laboratories (Ingman et al. 2000; Finnila et al. 2001; Maca-Meyer et al. 2001; Herrnstadt et al. 2002; Ingman and Gyllensten 2003). In this regard, it is reassuring to note that a number of sites identified as variable by Lutz-Bonengel et al. (2003) were also among our selected sites (e.g. 8592 in our multiplex panel B for various haplogroup H HV1/HV2 types, 11485 in multiplex panel D for various haplogroup J and K common types, and 11377 in multiplex panel F for various V and H haplogroup common types).

We caution, however, that general coding sequence population data must be evaluated carefully in relation to specific forensic goals. In this regard, it is important to note that, due to the evolutionary linkage of mtDNA polymorphisms, high population variability of a site does not automatically correlate with high additional discriminatory potential in relation to other sites. For example, position 11719 was identified by Lutz-Bonengel et al. (2003) as a highly variable SNP in a population sample of 109 individuals (39%A, 61%G), but assaying this site would only rarely provide further resolution when combined with any sort of control region typing. 11719G is a polymorphism essentially fixed in haplogroups H and V, with other Caucasian haplogroups and African and Asian populations so far appearing fixed for an alternate polymorphism 11719A (Finnila et al. 2001; Herrnstadt et al. 2002). The G/A polymorphic variants of 11719 mainly reflect a single evolutionary event that maps onto a major lineage split in European Caucasians: the information provided by typing 11719 would be largely superfluous to information already provided by HV1/HV2 testing. (That said, position 11719 is included in our multiplex panel C, because it is informative for distinguishing individuals of HV1/HV2 type H:5 – 16304C, 263G, 315.1C. This specific forensic utility is due to an apparent recent reversion mutation within this lineage that is coincidental to the larger population variation of 11719 that maps onto the H-V haplogroup split). We include this discussion of site 11719 to illustrate the type of detailed considerations we feel are required as we progress toward a maximally useful SNP-based approach for utilizing the entire mtGenome.

Conclusions

MtDNA population genetics and evolutionary biology has solidly entered the genomics era with recent publications of many whole mtGenome sequences (references above). It is inevitable and desirable that forensic applications will follow. As a result of technical developments in SNP assays, we can look forward to mtDNA testing that is ever more rapid and convenient, and, as we show here, has a higher power of discrimination when variation outside HV1/HV2 is accessed. We would like to add, in relation to issues we have heard aired regarding courtroom admis-

sibility of mtDNA testing, that such continued improvements of DNA testing systems do not detract from the reliability of the well validated systems that are currently in use. In the case of increased discrimination of mtDNA through SNP assays outside of HV1/HV2, the evidentiary significance of mtDNA testing may increase. Nonetheless, proper evaluation of the mtDNA evidence in relation to the chance of random match will always be necessary, and the new SNP testing will be neither more nor less valid than current HV1/HV2 testing, when evidence from the latter is properly interpreted and communicated.

Acknowledgements We thank interns Rachel Barry, Trina Bersola, Serena Filosa, Victoria Glynn, Carrie Guyan, William Ivory, Devon Pierce, and Jessica Saunier for assistance with sequence analysis and data tabulation; James Ross, Richard Coughlin, and Aaron Waldner for computer support; Jon Norris, Vinh Lam, and others from Future Technologies, Inc. for database development; Suzanne Barritt, Demris Lee, Tim McMahon, and James Thomas (AFDIL) for discussion; Walther Parson, Harrald Niederstätter, and Anita Brandstätter (ILM, Innsbruck) for discussion; John Butler and Pete Vallone (NIST) for discussion; Connie Fisher (FBI) for providing samples; Eliana Andrea and Michael Parry (American Registry of Pathology) for grant administration assistance; Mitchell Holland for early conceptual and administrative support; and Kevin (Scott) Carroll, James C. Canik, and Brion C. Smith (AFDIL) for logistical, administrative, and moral support. This work was supported by a National Institute of Justice grant 2000–1J-CX-K010 to T.J.P. The opinions and assertions contained herein are solely those of the authors and are not to be construed as official or as views of the U.S. Department of Defense, the U.S. Department of the Army, or the U.S. Department of Justice.

References

Allard MW, Miller K, Wilson M, Monson K, Budowle B (2002) Characterization of the Caucasian haplogroups present in the SWGDAM forensic mtDNA dataset for 1771 human control region sequences. Scientific Working Group on DNA Analysis Methods. J Forensic Sci 47:1215–1223

Anderson S, Bankier AT, Barrell BG et al. (1981) Sequence and organization of the human mitochondrial genome. Nature 290: 457–465

Andreasson H, Asp A, Alderborn A, Gyllensten U, Allen M (2002) Mitochondrial sequence analysis for forensic identification using pyrosequencing technology. Biotechniques 32:124–133

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147

Aquadro CF, Greenberg BD (1983) Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics 103:287–312

Armstrong B, Stewart M, Mazumder A (2000) Suspension arrays for high throughput, multiplexed single nucleotide polymorphism genotyping. Cytometry 40:102–108

Bandelt HJ, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71:1150–1160

Bodenteich A, Mitchell LG, Polymeropoulos MH, Merril CR (1992) Dinucleotide repeat in the human mitochondrial D-loop. Hum Mol Genet 1:140

Brandstätter A, Parsons TJ, Niederstätter H, Parson W (2003) Rapid screening of mtDNA coding region SNPs for the identification of Caucasian haplogroups. Int J Legal Med 117:291–298

Brown MD, Shoffner JM, Kim YL et al. (1996) Mitochondrial DNA sequence analysis of four Alzheimer's and Parkinson's disease patients. Am J Med Genet 61:283–289

Cavelier L, Erikson I, Tammi M et al. (2002) MtDNA mutations in maternally inherited diabetes: presence of the 3397 ND1 mutation previously associated with Alzheimer's and Parkinson's disease. J Neuropathol Exp Neurol 61:634–639

Finnila S, Lehtonen MS, Majamaa K (2001) Phylogenetic network for European mtDNA. Am J Hum Genet 68:1475–1484

Gabriel MN, Calloway CD, Reynolds RL, Primorac D (2003) Identification of human remains by immobilized sequence-specific oligonucleotide probe analysis of mtDNA hypervariable regions I and II. Croatian Med J 44:293–298

Herrnstadt C, Elson JL, Fahy E et al. (2002) Reduced-median-network analysis of complete mitochondrial DNA coding region sequences for the major African, Asian, and European haplogroups. Am J Hum Genet 70:1152–1171

Herrnstadt C, Preston G, Howell N (2003) Errors, phantoms and otherwise, in human mtDNA sequences. Am J Hum Genet 72: 1585–1586

Holland MM, Parsons TJ (1999) Mitochondrial DNA sequence analysis – Validation and use for forensic casework. Forensic Sci Rev 11:21–50

Horai S, Hayasaka K (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am J Hum Genet 46:828–842

Ingman M, Gyllensten U (2003) Mitochondrial genome variation and evolutionary history of Australian and New Guinean Aborigines. Genome Res 13:1600–1606

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

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Lee MS, Levin BC (2002) MitoAnalyzer, a computer program and interactive web site to determine the effects of single nucleotide polymorphisms and mutations in human mitochondrial DNA. Mitochondrion 1:321–326

Lee SD, Lee YS, Lee JB (2002) Polymorphism in the mitochondrial cytochrome B gene in Koreans. An additional marker for individual identification. Int J Legal Med 116:74–78

Levin BC, Holland KA, Hancock DK et al. (2003) Comparison of the complete mtDNA genome sequences of human cell lines – HL-60 and GM10742A – from individuals with pro-myelocytic leukemia and leber heredity optic neuropathy, respectively, and the inclusion of HL-60 in the NIST human mitochondrial DNA standard reference material – SRM 2392-I. Mitochondrion 2:387–400

Lin MT, Simon DK, Ahn CH, Kim LM, Beal MF (2002) High aggregate burden of somatic mtDNA point mutations in aging and Alzheimer's disease brain. Hum Mol Genet 11:133–145

Lutz S, Wittig H, Weisser HJ et al. (2000) Is it possible to differentiate mtDNA by means of HVIII in samples that cannot be distinguished by sequencing the HVI and HVII regions? Forensic Sci Int 113:97–101

Lutz-Bonengel S, Schmidt U, Schmitt T, Pollak S (2003) Sequence polymorphisms within the human mitochondrial genes MTATP6, MTATP8, and MTND4. Int J Legal Med 117:133–142

Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C, Cabrera VM (2001) Major genomic mitochondrial lineages delineate early human expansions. BMC Genet 2:13

Macaulay V, Richards M, Hickey E et al. (1999) The emerging tree of west Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 64:232–249

Malyarchuk BA, Rogozin IB, Berikov VB, Derenko MV (2002) Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region. Hum Genet 111: 46–53

Mehta AB, Vulliamy T, Gordon-Smith EC, Luzzatto L (1989) A new genetic polymorphism in the 16S ribosomal RNA gene of human mitochondrial DNA. Ann Hum Genet 53:303–310

Meyer S, Weiss G, Haeseler A von (1999) Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics 152:1103–1110

Monson KL, Miller KWP, Wilson MR, DiZinno JA, Budowle B (2002) The mtDNA population database: an integrated software and database resource for forensic comparison. Forensic Sci Comm 4:2. http://www.fbi.gov/hq/lab/fsc/backissu/april2002/miller1.htm

Parsons TJ, Coble MD (2001) Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome. Croatian Med J 42:304–309

Saccone C, Pesole G, Sbisa E (1991) The main regulatory region of mammalian mitochondrial DNA: structure-function model and evolutionary pattern. J Mol Evol 33:83–91

Stewart JE, Fisher CL, Aagaard PJ et al. (2001) Length variation in HV2 of the human mitochondrial DNA control region. J Forensic Sci 46:862–870

Torroni A, Lott MT, Cabell MF, Chen YS, Lavergne L, Wallace DC (1994) mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. Am J Hum Genet 55:760–776

Torroni A, Huoponen K, Francalacci P et al. (1996) Classification of European mtDNAs from an analysis of three European populations. Genetics 144:1835–1850

Tzen CY, Wu TY, Liu HF (2001) Sequence polymorphism in the coding region of mitochondrial genome encompassing position 8389–8865. Forensic Sci Int 120:204–209

Vallone PM, Just RS, Coble MD, Butler JM, Parsons TJ (2004) A multiplex allele-specific primer extension assay for forensically informative SNPs distributed throughout the mitochondrial genome. Int J Legal Med 118 (in press)

Wallace DC (1999) Mitochondrial diseases in man and mouse. Science 283:1482–1488

Wallace DC, Brown MD, Lott MT (1999) Mitochondrial DNA variation in human evolution and disease. Gene 238:211–230

Walsh PS, Metzger DA, Higuchi R (1991) Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. Biotechniques 10:506–513