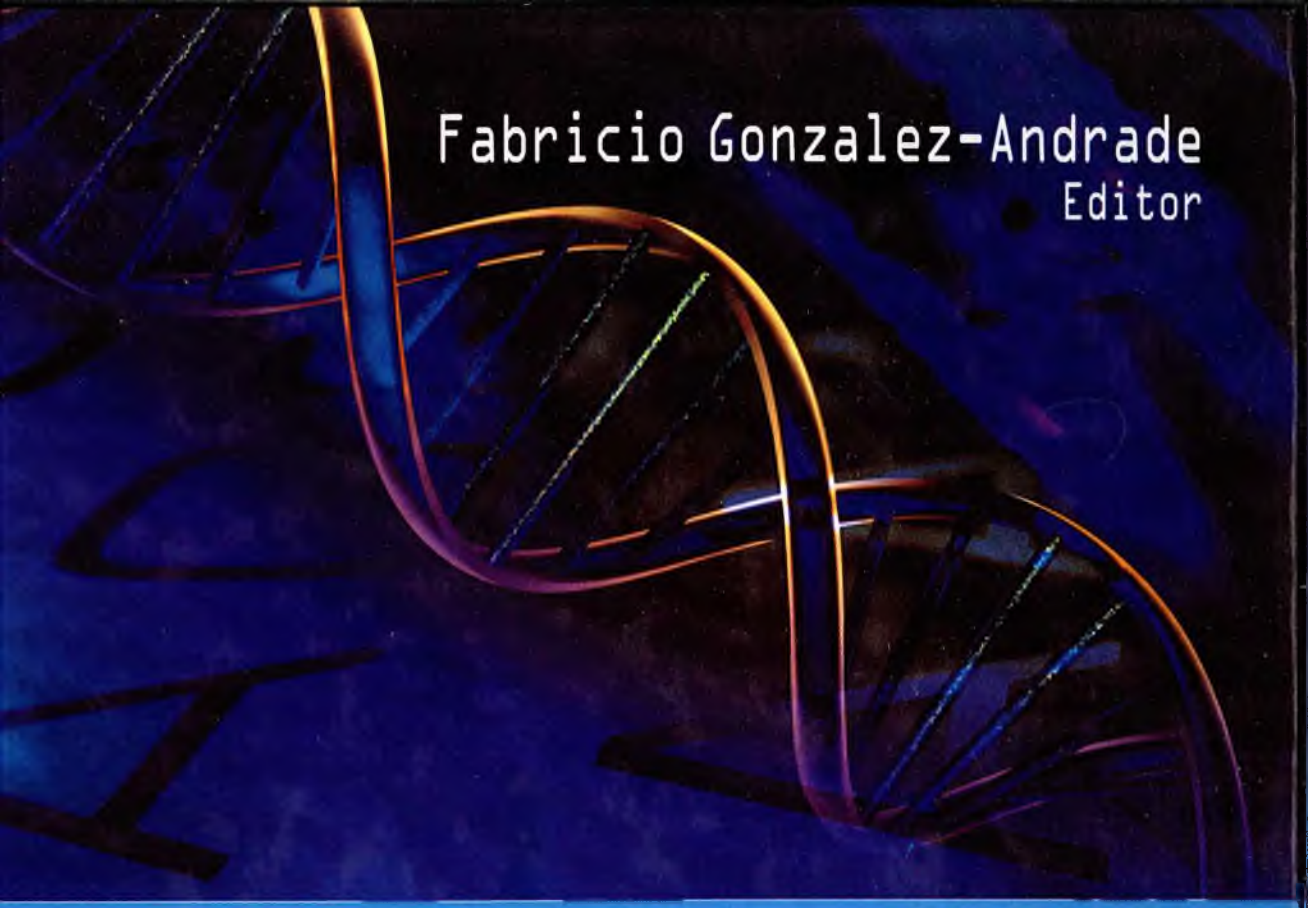


Fabricio Gonzalez-Andrade
Editor

Forensic Genetics Research Progress

NOVA

CONTENTS

In: Forensic Genetics Research Progress
Editor: F. Gonzalez-Andrade

ISBN: 978-1-60876-198-2

Chapter 9

FORENSIC MITOCHONDRIAL DNA ANALYSIS

***Luísa Pereira*[1,2,*], *Farida Alshamali*[3], *Fabricio González-Andrade*[4-5,†]**

[1]Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Portugal
[2]Faculdade de Medicina da Universidade do Porto (IPATIMUP), Portugal
[3]Dubai Police Forensic Administration, Dubai, UAE
[4]Department of Medicine, Metropolitan Hospital, Av. Mariana de Jesús Oe8, Quito, Ecuador
[5]Forensic Genetics Laboratory, Forensic Medicine Department, University of Zaragoza, Calle Domingo Miral s/n, Zaragoza 50.009, Spain

ABSTRACT

The introduction of mitochondrial DNA (mtDNA) investigation in forensic genetics allowed to obtaining results from ancient, residual and degraded samples, enlarging extensively the possibility of applying genetic analyses to difficult forensic cases. However, the particular characteristics of mtDNA brought some conceptual and statistical challenges to forensic genetics, namely: the uniparental (maternal) transmission implies lineage instead of individual characterization, so that mtDNA can be more informative in excluding rather than in including a suspect; the absence of recombination in mtDNA renders impossible to apply the product rule for estimation of match probabilities, so that evaluations are limited to the frequency of a certain haplotype in a database; most of mtDNA haplotypes are unique or very low frequent, implying that databases must have a considerable number of individuals in order to be informative; heterogeneity in mutation rates between mtDNA positions and heteroplasmy must be taken into account when evaluating if diverse haplotypes can come from the same individual. Typically, the mtDNA survey in forensic genetics is performed by sequencing two hypervariable regions in the control region or D-loop. Some databases, reporting

* E-mail: lpereira@ipatimup.pt
• E-mail: shamali@emirates.net.ae
† E-mail: fabriciogonzaleza@yahoo.es

haplotypes in diverse populations, are publically available for forensic purposes. Recently, information from other polymorphisms located in the coding region is being also added to forensic analyses, which allows to inferring more securely the haplogroup to which the haplotype belongs. This phylogenetic information can be very informative for quality purposes, helping in detecting possible mix-up of samples and in checking haplogroup defining polymorphisms. Lately, the mtDNA screening is being enlarged to the total control region (~1200bp), and in the near future to the complete molecule. Such amount of information, in such a short period of time, will challenge forensic genetics in maintaining its strict quality-control of sequences and in being efficient to updating online databases for match evaluation.

THE MTDNA MOLECULE

The mitochondrial DNA (mtDNA) is a circular genome localized inside mitochondria, in a variable number of copies. In humans (Figure 1) it is about 16,569bp long and composed of two main regions: the coding and the control region. The coding region, extending from position 577 to 16123, bears the genes for 22 tRNAs, 2 rRNAs and 13 proteins of the oxidative phosphorylation cycle, as well as the origin of replication for the light chain. The control region or D-loop, located in the remaining ~1,200bp, contains the control regions for replication of the heavy chain and for translation, being some of these controls located in two regions having a higher mutation rate than the rest of the molecule, receiving the designation of hypervariable region I and II (abbreviated, HVI, HVSI or HVRI and HVII, HVSII or HVRII).

MtDNA is uniparentaly transmitted, by the maternal side: mothers pass mtDNA to daughters and sons, but in the next generation, only the daughters will transmit the mtDNA. It is not well known why mtDNA present in the sperm is not maintained in the egg, but maybe a chemical inhibition is involved. So, all maternal related individuals will share the mtDNA lineage.

MtDNA does not undergo recombination, being transmitted in block. Recently, there were some claims in the opposite, but were refuted or not confirmed in other individuals. This absence of recombination allows an easy reconstruction and dating of the phylogeny, explaining the huge success of mtDNA in the population genetic field, but renders ineffective the product rule so familiar in forensic genetics. It is therefore impossible to estimate the frequency of composite genotypes (such as HVRI and HVRII) from their individual frequencies, as it is done for autosomal markers.

The first population studies using mtDNA (Brown 1980; Cann et al. 1987) screened a few single nucleotide polymorphisms (SNPs) along the molecule by Restriction Fragment Length Polymorphism Analysis (RFLP), but soon after, in the early 1990's, the advent of PCR led to an extensive sequencing of HVRI. In forensic genetics, both HVRI and HVRII began to be typed, by using mainly primers described by Vigilant et al. (1989) or by Wilson et al. (1995).

The characterization of HVRI and RFLP diversity in worldwide populations led to the reconstruction of the mtDNA phylogenetic tree and to the definition of haplogroup, being a monophyletic group of sequences, hence sharing the same ancestral and set of polymorphisms (Torroni et al. 1993). The first haplogroups to be defined were Asian ones being observed in

America, receiving the designations of A, B, C and D. Following these, the Eurasian and sub-Saharan haplogroups received the remaining letters of the alphabet. Hierarchy inside a certain haplogroup is named by a number following the letter and so on (letter, number, letter, number). Sub-Saharan haplogroups were shown to be at the root of the human mtDNA phylogenetic tree, favoring the hypothesis of a unique origin for the modern humans, with the further verification that this origin was a recent one, around 200,000 years ago (Cann et al. 1987; see revision in Torroni et al. 2006). The Out-of-Africa migration, occurring between 80,000-60,000 years ago, was responsible for the settlement of the World, being all non-African mtDNA haplogroups derived from a unique typical East African haplogroup, designated L3 (Figure 2). Thus, sub-Saharan haplogroups include the most diverse ones: L0, L1, L2, L3, L4, L5, L6 and L7. The L3 out-of-Africa gave rise to two macro-haplogroups: N, which is more frequent in Eurasia; and M, more frequent in East Asia. The macro-haplogroup N comprises clades N1, N2, X and R, being this last one split in R0, JT and U (including K). M is divided in a multitude of haplogroups observed throughout East Asia, Southeast Asia and America.

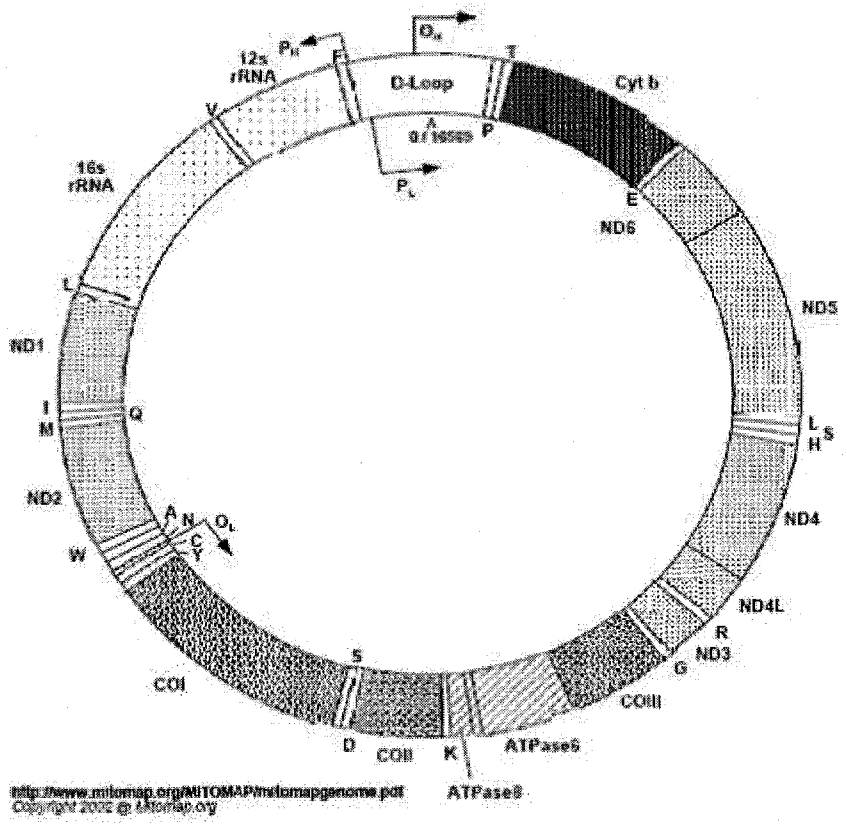


Figure 1. The map of the human mtDNA molecule, showing the D-Loop, the two rRNAs (12s rRNA and 16s rRNA), the 13 protein-coding genes (NADH dehydrogenases 1, 2, 3, 4, 4L, 5 and 6, cytochrome c oxidase I, II and III, cytochrome c oxidoreductase and ATP synthases 6 and 8), and the 22 tRNAs (represented by the first letter). The origins of replication of the two chains are represented (O_H and O_L) as well as the strand-promoters for both chains (P_H and P_L). Figure adapted from mitomap.org.

This high population structure, that is the existence of lineages characteristic from geographic regions, rendering the proportion of variance in diversity between population groups considerable, is in part due to the ¼ of mtDNA effective size when compared with autosomal markers. This lower effective size turns mtDNA much more sensitive to demographic phenomena such as founder effects, bottlenecks and genetic drift.

It would be tempting, in the forensic field, to assign an individual to a certain population group based on its mtDNA haplogroup. However, it should be stressed that mtDNA is only transmitted by the maternal side – while an individual can have a sub-Saharan mtDNA lineage, most of its nuclear genome can have a European constitution, and the individual present accordingly a European phenotype. For instance, the genetic composition just referred is observed in 10% of the autochthonous south Portuguese population (Pereira et al. 2004a).

MTDNA MUTATION RATE AND HETEROGENEITY

The mutation rate in the control region is around 10 times higher than the one in the coding region (Vigilant et al. 1991). In 1996, Forster et al. estimated, more precisely, a rate of 1 substitution every 20,180 years for HVRI, between nucleotide positions 16090-16383. This mutation rate was used since then for the estimation of the Time for the Most Recent Common Ancestor (TMRCA) in the various haplogroups.

It was also shown around this time that the heterogeneity in mutation rates per position was higher for the HVRII when compared with HVRI, bearing many positions which are highly recurrent (as 150, 152 and 189) in long stretches of almost invariable positions (Meyer et al. 1999; Schneider and Excoffier 1999). These fast-evolving positions can mutate back and can appear in several haplogroup backgrounds (a condition known as homoplasy), being phylogenetically uninformative.

These mutation rates were being estimated by applying phylogenetic methods, and were shown to be lower than the estimates obtained when using genealogical inferences (analyzing mutations occurring along large familiar pedigrees; Howell et al. 1996). This uncertainty in the estimation of the mutation rates led to claims that they did not allow a safe reconstruction of phylogeny (Howell et al. 1996); a discussion ensued soon after, with opposite claims of a secure phylogenetic reconstruction by using mtDNA lineages (Macaulay et al. 1997). One explanation for the higher mutation rates when using genealogical inferences was that these were catching mutations in fast-evolving positions, which led to the over-estimation of the mutation rates. Although some arguments continued around this issue, this genome has been the main genetic tool used for inferences related with human migrations in the past.

A further support for its use came from the characterization of the complete molecule (the first population study was published by Ingman et al. 2000, for 53 worldwide samples), which showed the robustness of previous phylogenetic inferences. Surprisingly, some of the nucleotide positions located in the coding region showed also to be highly recurrent, as for instance 709, 3010 and 15301 (Freitas and Pereira 2008).

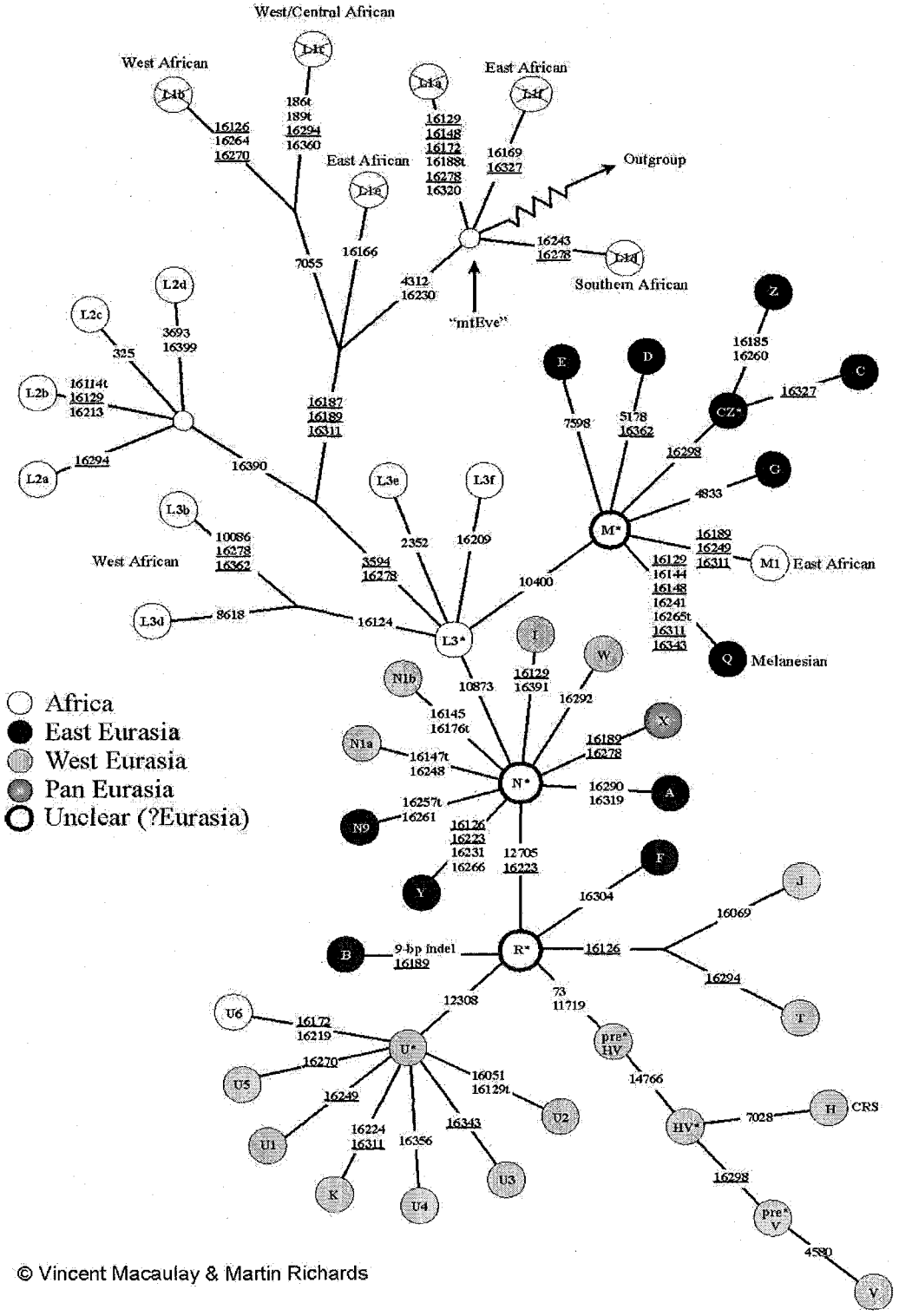


Figure 2. The human mtDNA worldwide phylogenetic tree based on HVRI and some RFLP diversities.

Forensic genetics must be, nonetheless, aware of this heterogeneity in mutation rates between positions, as they affect the evaluation of a match. Ultimately, the mutation rate for a certain position involved in a match evaluation should be taken into consideration. Unfortunately, there is still a lot of uncertainty around the estimation of mutation rates per position, for the human mtDNA. There are some tables reporting some values, but most are unapdated (see a sum up of these in Salas et al. 2007). The increasing publication of many complete human mtDNA sequences will soon resolve this situation.

ISSUES RELATED WITH NOMENCLATURE

The first complete human mtDNA sequence was published by Anderson et al. (1981), being known as Cambridge Reference Sequence or CRS. Later on, Andrews et al. (1999) revised the CRS, correcting some of the previous errors, except insertions and deletions which would imply an alteration of the numbering; this new revised CRS, or rCRS, should be the one used as the human mtDNA reference sequence, versus which all the other sequences must be compared to. This sequence is deposited in GenBank with the Accession Number NC_000021.

This comparison versus a unique sequence for each species is very important in phylogenetic and forensic fields. Otherwise, data from diverse publications, each using a different reference sequence, would not be directly comparable. This confounding comparison versus several reference sequences was occurring for the dog mtDNA (Pereira et al. 2004b), a species that has gained an increasing interest in the forensic field.

Other nomenclature issues are also very important for standardization of mtDNA sequence report, such as the edition of the alignment in a homopolymeric track (a stretch of the same base). Different nomenclature criteria could misleadingly create diverse sequences. For instance, a track of 5 C's can be derived from one of 4 C's by deletion of a C in the first position, or in the second and so on. A few nomenclature rules were established in order to standardize these substitutions (Wilson et al. 2002a,b) – in the case referred, it was established that the substitution should be considered at the 3' end of the track and insertions referred as 15534.1C if the base inserted is a C (or X.2C if there is insertion of 2 Cs), while deletion coded as so (e.g., 15938del).

Concerning the recording of substitutions, for simplicity reasons, a haplotype can be described as being 15627-15639T/A-15814 or 15627-15639A-15814, where numbers without letters denote traansitions (15627 refers the A to G transition and 15814 the C to T), while transversions (as in the case of 15639) are explicitly indicated (or by just the new base).

There are other cases where the alignment of a certain gap can be interpreted in different ways, conducting to potentially miscoded variation. Wilson et al. (2002a) recommend an alignment approach that is based on a phylogenetic context using differential weighting of transitions, transversions and indels. Basically, they proved that most variants could be characterized if the following three recommendations are followed:

(1) Characterise profiles using the least number of differences from the reference sequence.

(2) If there is more than one way to maintain the same number of differences relatively to the reference, differences should be prioritised in the following manner:
   (a) indels
   (b) transitions
   (c) transversions
(3) Indels should be placed 3' with respect to the light strand. Insertions and deletions should be combined in situations where the same number of differences to the reference sequence is maintained.

For instance, when aligning the following haplotype F1 versus the reference:

F1 AAACCCTCCCCCTATG
Ref AAACCCTTCTCCCCTCCCCTATG

One possible alignment is:

F1 AAACCCT--------CCCCCTATG
Ref AAACCCTTCTCCCCTCCCC–TATG

that is: 15523del-15524del-15525del-15526del-15527del-15528del-15529del-15530del-15534.1C, where the combination of the insertion with the deletions is supported by phylogeny, since all the remaining F haplotypes have this insertion in comparison with the reference. But according to the first of the above rules, the following alignment must be considered

F1 AAACCCTC-------CCCCTATG
Ref AAACCCTTCTCCCCTCCCCTATG

that is: 15523-15524del-15525del-15526del-15527del-15528del-15529del-15530del, being the transition at position 15523 also supported by phylogeny.

Bandelt and Parson (2008) argued that a binary comparison does not solve all the problems, being bound to produce artificial alignments. Instead, they suggest a phylogenetic approach for multiple alignment and resulting notation, indicating the following rules:

- (Phylogenetic law) Sequences should be aligned with regard to the current knowledge of the phylogeny. In the case of multiple equally plausible solutions, one should strive for maximum (weighted) parsimony. Variants flanking long C tracts, however, are subject to extra conventions in view of extensive length heteroplasmy.
- (C tract conventions) The long C tracts of HVRI and HVRII should always be scored with 16189C and 310C, respectively, so that phylogenetically subsequent interruptions by novel C to T changes are encoded by the corresponding transition. Length variation of the short A tract preceding 16184 should be notated in terms of transversions.
- (Indel scoring) Indels should be placed 3′ with respect to the light strand unless the phylogeny suggests otherwise.

HETEROPLASMY

As there are many copies of mtDNA per mitochondrion and many mitochondria per cell, this genome is most of the times the only one recovered from residual and degraded samples, very common on the forensic routine. This explains the success of mtDNA in forensic casework related with some recent human calamities, such as the Asian tsunami in 2004 (Deng et al. 2005).

However, not all the mtDNA copies present in an individual are perfectly equal. As the mtDNA mutation rate is high, along life the mtDNA molecules will accumulate mutations, differing from the main dominant inherited molecule. Individuals are constituted by populations of diverse mtDNA molecules. This condition is known as heteroplasmy. For tissues which have a high replacement rate, such as blood, this issue is not problematic, as there is no time for accumulation of mutations in the mtDNA; but for tissues which have a low turnover (as cardiac muscle) or do not replace at all (as brain), the population of heteroplasmic molecules will increase with age.

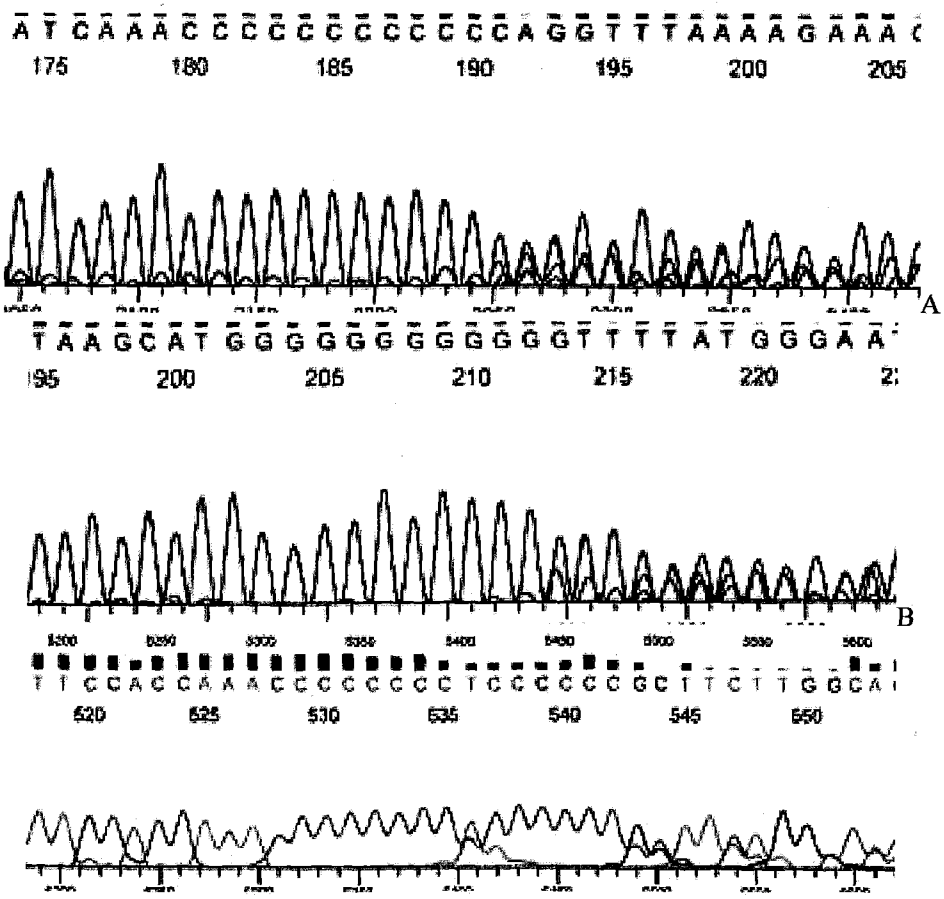


Figure 3. Length heteroplasmy in the region 16184-16193 in HVRI (A – forward and B- reverse senses) and 303-309 in HVRII.

In the forensic field, heteroplasmy is a very important issue when the samples consist in hairs collected in the crime scene, being this a very frequent occurrence. Each hair is a cell and its mitochondria passed through a severe bottleneck, so that the population of mitochondria present in one hair can be different from the other present in another hair. It was also shown that different shafts of the same hair can bear different populations of mtDNA (Brandstätter and Parson 2003).

Certain polymorphisms contribute extensively to heteroplasmy, as the indels in the homopolymeric tracks (for instance, the poly-C stretches in positions 303-309 and 310-315 in HVRII; and dinucleotides CAs between positions 514-523). These locations are highly prone to mutations due to slippage of the DNA polymerase, a condition known as length heteroplasmy (Figure 3).

When evaluating a match between the mtDNA from a suspect and a hair left in the crime scene, the possibility of heteroplasmy must be taken into account: it is possible that sequences differing in one position can still belong to the same individual, especially if it is a highly recurrent position. So, heterogeneity in mutation rates between positions is very important for the evaluation of heteroplasmy. Another issue which can contribute important information for the evaluation of heteroplasmy is phylogeny – heteroplasmy cannot erase one haplogroup and affiliate the sample in another haplogroup.

When Environment Mimics Mutation and Positions Prone to Lab Hotspotting

Some old and degraded samples can still bear organic material enabling its molecular analyses; most probably, the mtDNA is the best preserved molecule or at least, due to the high number of copies per cell, the most frequent.

By the end of the 90's, there was a huge boom of publications reporting ancient mtDNA sequences. GenBank displayed even the mtDNA sequence of a dinosaur, which latter was shown to be a fragment of human mtDNA inserted in the nuclear genome (NUMTs; to be explained bellow). However, a natural limitation for ancient mtDNA studies was reported, due to DNA being postmortem degraded: calculations of deamination and depurination kinetics for the four nucleotides led to the estimation that under physiological salt conditions, neutral pH, and an ambient temperature of 15°C, 100,000 years is a likely limit of time beyond which DNA will be un-retrieved (Lindahl 1993).

These observations with ancient DNA can be valuable in the field of forensic genetics, where in some cases the only material available consists in very badly preserved bones. The biochemical modifications occurring in bones are analogous to those seen *in vivo*, and act through both the cross-linking and fragmentation of the molecule's chemical backbone and the alteration of individual nucleotide bases, being subjected to the environmental conditions (reviewed in Gilbert 2006). DNA fragmentation can be due to the effect of radiation or to the hydrolytic cleavage of diester bonds in the phosphate sugar backbone or the glycosidic bonds joining the bases to the sugars. DNA fragmentation will render amplification by PCR very difficult, especially if the fragments to amplify have around 300bp, as is typical in routine casework. In order to deal with this, some mini-sets of primers have been designed for highly degraded samples (Alonso et al. 2003; Eichmann and Parson 2008). With respect to the point

base alterations, the most common damage-driven changes observed are the four transitions (C/T; G/A; T/C; A/G), mimicking the *in vivo* mutations and misleading the haplotyping identification.

Curiously, it was observed that lab processing of samples can also "induce" mutations. Brandstätter et al. (2005) analyzed 5,400 pairs of mtDNA control region electropherograms, from extant samples, for the light and heavy strands. These samples were typed with diverse chemistries and run in diverse automated sequencers. The authors were able to identify "phantom mutations", which are systematic artifacts generated in the course of the sequencing process, being the amount of these artifacts dependent on the sort of automated sequencer, the sequencing chemistry employed and other lab-specific factors. Further analysis of more than 30,000 published HVRI sequences confirmed some potential hotspots for phantom mutations, especially for variation at positions 16085 and 16197.

This propensity for hotspot under lab techniques and under postmortem modifications must also be taken into account in match evaluation and in phylogeny reconstructions.

NUMTs and Contamination

mtDNA only codes for 13 proteins of the oxidative phosporylation chain, which is composed of many more proteins. It is estimated that 80% of the mitochondrial proteins are coded by the nuclear DNA. Most probably, these proteins were once coded by the mtDNA, but genes coding them migrated, at some point, to the nucleus, being then lost from the mtDNA. These migrations of mtDNA fragments to the nuclear DNA do happen along time, and if they become successfully integrated in the nuclear DNA they receive the name of NUMTs – nuclear mitochondrial DNA sequences. When comparing the human and chimpanzee nuclear genomes, Mishmar et al. (2004) found a NUMT which is absent in the chimpanzee, showing that its insertion into the nuclear DNA occurred only in the line leading to *Homo sapiens*, being thus a very recent event. These authors accounted for a total of 247 NUMTs in the human genome.

Could these NUMTs be an additional source of contamination in routine forensic mtDNA analyses? In order to evaluate this possibility, Goios et al. (2006) analyzed the possibility of primer annealing for one of the most used HVRI and HVRII sets of primers in forensic genetics (the ones described by Wilson et al. 1995) in 19 of the 247 known human NUMTs bearing hypervariable regions. Their conclusion was that there is no possibility of HVRI and HVRII primer annealing in the NUMTs containing the hypervariable regions due to considerable mismatches between the primers and the complement region in the 19 NUMTs.

As it is now current to screen additional information from the coding region even in forensic investigations, these authors (Goios et al. 2008) conducted further analyses by focusing on a NUMT which bears a 97% homology to a segment of the mtDNA between positions 3914 and 9755. They designed two specific sets of primers for the mtDNA and for the nuclear DNA sequence by identifying regions with at least two mismatches between both genomes; these set of primers defined a segment of around 240bp which presented two SNPs allowing to distinguishing between mtDNA and nuclear segments. Amplifications with both sets of primers were performed along a range of annealing temperature and in several tissues (blood, hairs, buccal swabs and differential lyses of semen). Conclusions were that there was

no risk of routine mis-amplification of nuclear fragments, being the opposite true: mtDNA mis-amplification by using nuclear specific primers at low temperatures. Some caution must, nevertheless, be taken into account when analyzing samples with very low number of mtDNA copies (as in the sperm).

MTDNA DATABASES

mtDNA databases are essential tools for frequency estimations of mtDNA sequences, a basic step in the evaluation of a match.

When the first data on population mtDNA databases began to be analyzed, it was found that most haplotypes were unique. Pereira et al. (2004a) performed several empirical tests of the effect of sample size (n=50, 100, 200, 300 and 400) on the estimation of relevant parameters (such as haplotype diversity, number of different haplotypes, nucleotide diversity and number of polymorphic positions) in an enlarged mtDNA database (n=549) for the Portuguese population. While haplotype and nucleotide diversities did not vary significantly with sample size, the numbers of haplotypes and polymorphic positions raised continuously inside the tested interval. When using these data to extrapolate saturation curves (Figure 4), it was found that a sample size of 1,000 individuals is required for practical saturation of the number of haplotypes for HVRI (defined as the point where a sample size increase of 100 individuals corresponds to an increment in the diversity measure below 5%). For HVRII the same level is reached at n=900 and n=1,300 is needed when both regions are analyzed simultaneously. Consequently, the typical sample sizes of around 100 individuals are inadequate for both anthropological and forensic purposes.

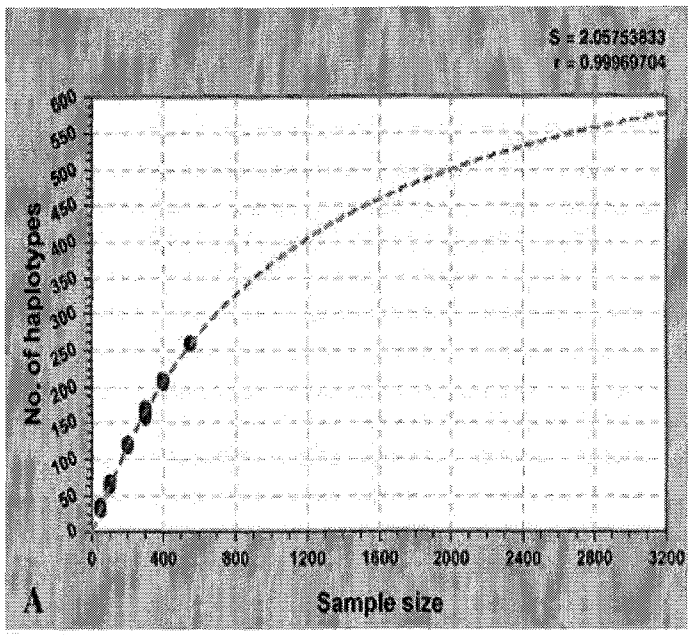
Besides the issue of considerable sample sizes, in order to have an informative database, there is the issue of the high population structure. As already referred, a considerable proportion of genetic diversity is observed between populations when analyzing mtDNA, in opposition to when screening nuclear markers. This demands that question sample and control population must be geographically matched, even at a micro-geographic scale, when evaluating a mtDNA match (Pereira et al. 2005). A statistical significant bias can be committed when the frequency of a haplotype detected in a French casework is calculated in a Spanish mtDNA database.

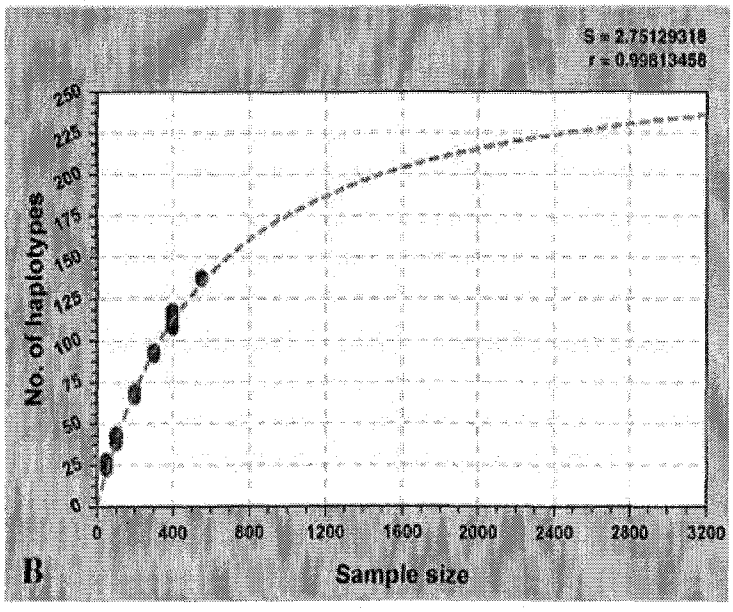
Although there are many HVRI datasets published for most worldwide populations, in population genetic surveys, these cannot be all directly used in forensic evaluations. In fact, quality criteria in forensic investigations are very strict, implying that databases should pass a strong-quality control before its use. Errors and phantom mutations were detected in many population and clinical genetic studies (e.g. Salas et al. 2005), rendering these datasets inadequate for forensic applications.

One of the largest first databases with forensic purposes was designed by the FBI (Monson et al. 2002). But even in this case, a few errors were reported (Bandelt et al. 2004), namely mix-up of samples between HVRI and HVRII haplotypes, which are usually typed independently. These errors were corrected afterwards (Budowle and Polanskey 2005).

Salas, Bandelt and co-authors have published many examples of errors in datasets, aiming to call the attention of the forensic and clinical genetic fields to the amount of errors committed when reporting mtDNA sequences and its implications when included in databases

(e.g. Salas et al. 2005). These authors also showed how phylogenetics can work as a quality control of mtDNA databases (Salas et al. 2007). Haplogroups are defined by motifs, or in other words, certain nucleotides at certain positions along the mtDNA molecule; hierarchy and absence of recombination renders that new mutations can appear additionally to the basic motif, but this cannot be erased, except for some defining mutations which can be located in fast-evolving positions (being prone to back-mutations). This phylogenetic analysis makes it possible to detect some mix-up of samples if, for instance, for a certain sample the indicated HVRI haplotype belongs to haplogroup J while its HVRII haplotype belongs to haplogroup X; most probably there were two different samples instead of one analyzed in the independent HVRI and HVRII screenings.





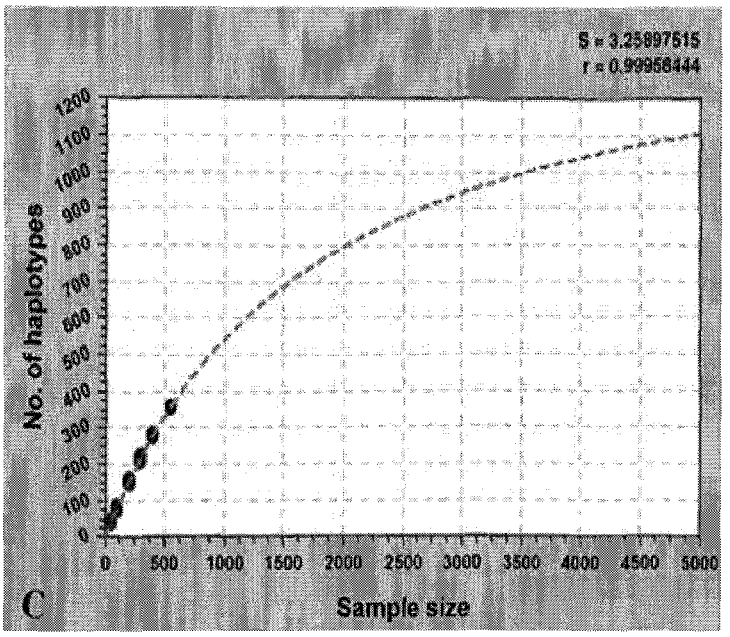


Figure 4. Sample sizes and number of haplotypes: regression curves for HVRI (*A*), HVRII (*B*) and HVRI+HVRII (**C**). *S* standard error, *r* correlation coefficient. All curves are of the form y=ax/(b+x), and coefficients were a=775.77 and b=1098.72 for HVRI, a=279.50 and b=588.64 for HVRII, a=1493.22 and b=1742.11 for HVRI+HVRII. (From Pereira et al. 2004a).

Taking these recommendations into account, several years were necessary for the development of a mtDNA database called EMPOP (EDNAP mtDNA Population Database), which meets forensic standards (Parson and Dür 2007). This database is available online at http://www.empop.org, and established a concept for mtDNA data generation, analysis, transfer and quality control (exemplified in Brandstätter et al. 2007). To help in the difficulty of detecting errors, the database displays software based on quasi-median network analysis for visualizing mtDNA data tables and thus pinpointing sequencing, interpretation and transcription errors. The first release launched on 16 October 2006 contained an effective of 5173. Most of these sequences were carefully screened for quality (a total of 4527), but the database also contains published databases (n=646).

Evaluation of a Match

All the issues described so far make evaluation of a match for mtDNA haplotypes not an easy task to perform. The Scientific Working Group on DNA Analysis Methods (SWGDAM) published some guidelines for mtDNA nucleotide sequence interpretation, which although simplistic can work as a point of departure (SWGDAM 2003). These guidelines are to report:

Exclusion - If there are two or more nucleotide differences between the questioned and known samples

Inconclusive - If there is one nucleotide difference between the questioned and known samples

Cannot Exclude - If the sequences from questioned and known samples under comparison have a common base at each position or a common length variant in the HVRII C-stretch

Weight of Evidence - The mtDNA profile of a reference sample and an evidence sample that cannot be excluded as potentially originating from the same source can be searched in a population database.

One should bear in mind that there are many additional issues which must be taken into account when evaluating a match, namely:

- type of tissue of questioned and known samples, which influences dramatically the occurrence of heteroplasmy
- if heteroplasmy is under consideration, relative proportions of the two bases in forward and reverse senses must be compared
- type of polymorphism, knowing that indels are more recurrent and hence more prone to heteroplasmy; then transitions are more frequent than transversions; and the variable relative mutation rates between positions
- haplogroup affiliation: questioned and known samples must belong to the same haplogroup
- database must be reliable in terms of quality control of sequences, effective size and geographic matching.

The issue of an objective approach to match evaluation for mtDNA haplotypes is far from being resolved. It will continue to be a field of intense debate between forensic experts, which, most dangerously, can be perceived as a weakness by judges and lawyers. No matter the difficulties in interpreting mtDNA evidences, for sure superior to the autosomal ones, its power of information for resolving difficult forensic casework is not under question.

THE PRESENT AND THE FUTURE

With the development of sequencing techniques, leading to faster, cheaper and easier typing, the sequencing of the hypervariable regions became complemented by screening some SNPs in the coding region. In particular, the description of minisequencing or SNaPshot allowed the design of multiplexes for several mtDNA SNPs. Some of the SNaPshots focused in sub-characterising the most frequent Eurasian haplogroup H (Quintáns et al. 2004; Brandstätter et al. 2006), other tried to sub-characterise all the Eurasian HVRI+HVRII haplotypes attaining a considerable frequency (Vallone et al. 2004), while still other aimed to help on the affiliation of samples in the main Eurasian haplogroups (Parson et al. 2008). As these SNPs are located in the coding region, a lower recurrence rate was expected. Nevertheless, several population data on complete mtDNA sequences and reconstructed phylogenies showed that some positions in the coding region can also be highly recurrent (as 3010, one of the positions defining a sub-haplogroup of H, called H1). But it is a fact that the

combination of HVRI+HVRII+SNaPshot can increase considerable the power of discrimination between haplotypes (see Pereira et al. 2006 for the increase in resolution when screening 8 informative SNPs for sub-characterisation of haplogroup H in an extensive European dataset).

Other authors are performing the typing of the control region (~1,200bp) in forensic applications. Besides the hypervariable regions I and II, this region contains the highly recurrent position 16519, totally uninformative phylogenetically but conferring a high capacity of discrimination between haplotypes. However, the gain with the extension from HVRI+HVRII to the total control region is not so considerable as the segment between both hypervariable segments is among the most conserved ones in the mtDNA genome. Nevertheless, there are technical protocols for its typing in the forensic field (Brandstätter et al. 2004; Alshamali et al. 2008) and worldwide databases are being developed (Irwin et al. 2007).

There is no doubt that the desirable situation would be to sequence the complete mtDNA molecule, in order to obtain the maximum resolution. Parsons and Coble (2001) reported that of 31 individuals with the most common HVRI+HVRII haplotype, only three still matched after complete mtDNA sequencing, with similar high discrimination being seen for other common haplotypes. Technically, the complete mtDNA sequencing is easily accessible for normal-quality samples, being currently applied in population genetic studies. The first population study based on complete mtDNA sequences was published for a worldwide sample composed of 53 individuals (Ingman et al. 2000). By the end of August 2008, there were 5,140 complete human mtDNA genomes published in GenBank (revised in Pereira et al. 2009), amounting to around 85,164,660bp. Such amount of information poses serious problems on the maintenance of the strict forensic quality control criteria: when will it be possible to launch a forensic online database for complete mtDNA genomes scrutinized by the high-quality criteria? How many haplotypes would be there? How long would it take to update it?

Most of the complete mtDNA genomes screened in population genetic studies are being obtained by sequencing only one strand of the molecule, which is not adequate to forensic applications. In order to deal with this necessity to sequencing both strands, Fendt et al. (2009) published a protocol for high quality and reliable sequencing of full mtDNA genomes, consisting in: (1) amplifying two overlapping PCR-fragments comprising each about 8500 bases in length; and (2) then performing sequencing reactions with a set of 96 primers that can be applied to a (manual) 96 well-based technology, which results in at least double strand sequence coverage of the entire coding region.

Other challenges relate with the application of novel technologies rather than the typical automatic sequencing and mini-sequencing, in order to resolve difficult cases. One example is the development of mass spectrometry assays for resolving mixtures of mtDNA sequences (Hall et al. 2005). This technology can also be used to resolve heteroplasmy quantification.

CONCLUSION

The inclusion of mtDNA analyses in forensic casework enlarged considerably the resolution of difficult cases through genetic evidence. These more than 20 years of

application have shown that there are some technological challenges, namely related with heteroplasmy, admixture of samples and postmortem alterations. However, the highest difficulties relate with interpretation of data and evaluation of a mtDNA match. Experience has proved that a forensic investigator, when evaluating mtDNA evidence for the resolution of a forensic casework, must be aware of population genetics discoveries on the field of mtDNA. To do otherwise is equivalent to apply a poor scientific evaluation of the genetic proofs. Knowledge on phylogeography is essential for quality control of sequences; consideration of heterogeneous mutation rates between positions is basic for match evaluation; a sense of high population structure is fundamental on the calculation of the frequency of a haplotype in a reference database.

In the impossibility of deciding if it is preferable to have big databases containing a few errors or a very limited one in terms of sample size but of high quality, forensic genetics will pursuit in applying its high quality-control standards. This is, in our opinion, the biggest challenge faced now by forensic genetics: construct and maintain reliable and updated databases for complete mtDNA genomes under the restrictive forensic quality criteria. A lesson can be taken from the high quality HVRI+HVRII/control region EMPOP database: it took many years to be released and it was not updated since 16 October 2006. We must be more efficient in terms of bioinformatics sustained on phylogeographic knowledge. Otherwise so much information available will only contribute noise to our capacity of resolving forensic casework through mtDNA genetic evidences.

COMPETING INTERESTS

The authors declare that they have no competing interests. The authors alone are responsible for the content and writing of the paper.

REFERENCES

Alonso A, Martin P, Albarrán C, García P, Primorac D, García O, Fernández de Simón L, García-Hirschfeld J, Sancho M, Fernández-Piqueras J (2003) Specific quantification of human genomes from low copy number DNA samples in forensic and ancient DNA studies. *Croat Med J.* 44:273-280.

Alshamali F, Brandstätter A, Zimmermann B, Parson W (2008) Mitochondrial DNA control region variation in Dubai, United Arab Emirates. *Forensic Sci Int Genet.* 2:e9-10.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. *Nature.* 290:457-465.

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.

Bandelt HJ, Salas A, Bravi C (2004) Problems in FBI mtDNA database. *Science.* 305:1402-1404.

Bandelt HJ, Parson W (2008) Consistent treatment of length variants in the human mtDNA control region: a reappraisal. *Int J Legal Med.* 122:11-21.

Brandstätter A, Parson W (2003) Mitochondrial DNA heteroplasmy or artefacts - a matter of the amplification strategy? *Int J Legal Med.* 117:180-184.

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

Brandstätter A, Sänger T, Lutz-Bonengel S, Parson W, Béraud-Colomb E, Wen B, Kong QP, Bravi CM, Bandelt HJ (2005) Phantom mutation hotspots in human mitochondrial DNA. *Electrophoresis.* 26:3414-3429.

Brandstätter A, Salas A, Niederstätter H, Gassner C, Carracedo A, Parson W (2006) Dissection of mitochondrial superhaplogroup H using coding region SNPs. *Electrophoresis.* 27:2541-2550.

Brandstätter A, Niederstätter H, Pavlic M, Grubwieser P, Parson W (2007) Generating population data for the EMPOP database - an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example. *Forensic Sci Int.* 166:164-175.

Brown WM (1980) Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc Natl Acad Sci U S A.* 77:3605-3609.

Budowle B, Polanskey D (2005) FBI mtDNA database: a cogent perspective. *Science.* 307:845-847.

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature.* 325:31-36.

Deng YJ, Li YZ, Yu XG, Li L, Wu DY, Zhou J, Man TY, Yang G, Yan JW, Cai DQ, Wang J, Yang HM, Li SB, Yu J (2005) Preliminary DNA identification for the tsunami victims in Thailand. *Genomics Proteomics Bioinformatics.* 3:143-157.

Eichmann C, Parson W (2008) 'Mitominis': multiplex PCR analysis of reduced size amplicons for compound sequence analysis of the entire mtDNA control region in highly degraded samples. *Int J Legal Med.* 122:385-388.

Fendt L, Zimmermann B, Daniaux M, Parson W (2009) Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences. *BMC Genomics.* 10:139.

Forster P, Harding R, Torroni A, Bandelt HJ (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. *Am J Hum Genet.* 59:935-945.

Freitas F, Pereira L (2008) Heterogeneity in coding mtDNA mutation rates: implications in forensic genetics. *Forensic Sci Int: Genetics. Supplementary Series* 1: 274-276.

Gilbert MTP (2006) Postmortem damage of mitochondrial DNA. In Bandelt H-J, Macaulay V, Richards M (Eds) Human mitochondrial DNA and the evolution of *Homo sapiens. Nucleic Acids and Molecular Biology.* 18: 91-115. Springer Verlag. Berlin; Heildelberg.

Goios A, Amorim A, Pereira L (2006) Mitochondrial DNA pseudogenes in the nuclear genome as possible sources of contamination. *International Congress Series 1288:697-*699.

Goios A, Prieto L, Amorim A, Pereira L (2008) Specificity of mtDNA-directed PCR-influence of NUclear MTDNA insertion (NUMT) contamination in routine samples and techniques. *Int J Legal Med.* 122:341-345.

Hall TA, Budowle B, Jiang Y, Blyn L, Eshoo M, Sannes-Lowery KA, Sampath R, Drader JJ, Hannis JC, Harrell P, Samant V, White N, Ecker DJ, Hofstadler SA (2005) Base composition analysis of human mitochondrial DNA using electrospray ionization mass spectrometry: a novel tool for the identification and differentiation of humans. *Anal Biochem.* 344:53-69.

Howell N, Kubacka I, Mackey DA (1996) How rapidly does the human mitochondrial genome evolve? *Am J Hum Genet.* 59:501-509.

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

Irwin JA, Saunier JL, Strouss KM, Sturk KA, Diegoli TM, Just RS, Coble MD, Parson W, Parsons TJ (2007) Development and expansion of high-quality control region databases to improve forensic mtDNA evidence interpretation. *Forensic Sci Int Genet* 1:154-157.

Lindahl T (1983) Instability and decay of the primary structure of DNA. Nature 362:709-715.

Macaulay VA, Richards MB, Forster P, Bendall KE, Watson E, Sykes B, Bandelt HJ (1997) mtDNA mutation rates--no need to panic. *Am J Hum Genet.* 61:983-990.

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

Mishmar D, Ruiz-Pesini E, Brandon M, Wallace DC (2004) Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. *Hum Mutat.* 23:125-133.

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 Sciences Communications.* 4.

Parson W, Dür A (2007) EMPOP-a forensic mtDNA database. *Forensic Sci Int Genet.* 1:88-92.

Parson W, Fendt L, Ballard D, Børsting C, Brinkmann B, Carracedo A, Carvalho M, Coble MD, Real FC, Desmyter S, Dupuy BM, Harrison C, Hohoff C, Just R, Krämer T, Morling N, Salas A, Schmitter H, Schneider PM, Sonntag ML, Vallone PM, Brandstätter A (2008) Identification of West Eurasian mitochondrial haplogroups by mtDNA SNP screening: results of the 2006-2007 EDNAP collaborative exercise. *Forensic Sci Int Genet.* 2:61-68.

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

Pereira L, Cunha C, Amorim A (2004a) Predicting sampling saturation of mtDNA haplotypes: an application to an enlarged Portuguese database. *Int J Legal Med.* 118:132-136.

Pereira L, Van Asch B, Amorim A (2004b) Standardisation of nomenclature for dog mtDNA D-loop: a prerequisite for launching a *Canis familiaris* database. *Forensic Sci Int.* 141:99-108.

Pereira L, Gonçalves J, Goios A, Rocha T, Amorim A (2005) Human mtDNA haplogroups and reduced male fertility: real association or hidden population substructuring. *Int J Androl.* 28:241-247.

Pereira L, Richards M, Goios A, Alonso A, Albarrán C, Garcia O, Behar DM, Gölge M, Hatina J, Al-Gazali L, Bradley DG, Macaulay V, Amorim A (2006) Evaluating the

forensic informativeness of mtDNA haplogroup H sub-typing on a Eurasian scale. *Forensic Sci Int.* 159:43-50.

Pereira L, Freitas F, Fernandes V, Pereira JB, Costa MD, Costa S, Máximo V, Macaulay V, Rocha R, Samuels DC (2009) The diversity present in 5140 human mitochondrial genomes. *Am J Hum Genet* 58:1-13.

Quintáns B, Alvarez-Iglesias V, Salas A, Phillips C, Lareu MV, Carracedo A (2004) Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. *Forensic Sci Int.* 140:251-257.

Salas A, Carracedo A, Macaulay V, Richards M, Bandelt HJ (2005) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. *Biochem Biophys Res Commun.* 335:891-899.

Salas A, Bandelt HJ, Macaulay V, Richards MB (2007) Phylogeographic investigations: the role of trees in forensic genetics. *Forensic Sci Int.* 168:1-13.

Schneider S, Excoffier L (1999) Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics.* 152:1079-1089.

SWGDAM (2003) Guidelines for mitochondrial DNA (mtDNA) nucleotide sequence interpretation. *Forensic Sci. Commun.* 5:2.

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, Vullo CM, Wallace DC. (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am J Hum Genet.* 53:563-590.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ. (2006) Harvesting the fruit of the human mtDNA tree. Trends Genet. 22:339-345.

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.* 118:147-157.

Vigilant L, Pennington R, Harpending H, Kocher TD, Wilson AC (1989) Mitochondrial DNA sequences in single hairs from a southern African population. *Proc Natl Acad Sci* U S A. 86:9350-9354.

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science.* 253:1503-1507.

Wilson MR, DiZinno JA, Polanskey D, Replogle J, Budowle B (1995) Validation of mitochondrial DNA sequencing for forensic casework analysis. *Int J Legal Med.* 108:68-74.

Wilson MR, Allard MW, Monson K, Miller KW, Budowle B (2002a) Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region. *Forensic Sci Int.* 129:35-42.

Wilson MR, Allard MW, Monson K, Miller KW, Budowle B (2002b) Further discussion of the consistent treatment of length variants in the human mitochondrial DNA control region. *Forensic Sci. Commun.* 4: 4.