Forensic Science International: Genetics 5 (2011) 222–225

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Short Communication

mtGenome reference population databases and the future of forensic mtDNA analysis

Jodi A. Irwin [a,*], Walther Parson [b], Michael D. Coble [a,1], Rebecca S. Just [a]

[a] Armed Forces DNA Identification Laboratory, Armed Forces Institute of Pathology, 1413 Research Blvd., Rockville, MD 20850, USA
[b] Institute of Legal Medicine, Innsbruck Medical University, Müllerstrasse 44, Austria

ARTICLE INFO

Article history:
Received 20 January 2010
Accepted 24 February 2010

Keywords:
Mitochondrial DNA
Coding region
Database

ABSTRACT

Mitochondrial DNA (mtDNA) testing in the forensic context requires appropriate, high quality population databases for estimating the rarity of questioned haplotypes. Currently, however, available forensic mtDNA reference databases only include information from the mtDNA control region. While this information is obviously strengthening the foundation upon which current mtDNA identification efforts are based, these data do not adequately prepare the field for recent and rapid advancements in mtDNA typing technologies. Novel tools that quickly and easily permit access to mtDNA coding region data for increased discrimination are now available in the form of single nucleotide polymorphism assays, sequence specific oligonucleotide probes, mass spectrometry instrumentation and next generation sequencing technologies. However, the randomly sampled entire mtGenome reference population data required for statistical interpretation of coding region data are lacking. As a result, in the near future, it seems that routine use of mtDNA coding region data in forensic case work will depend more upon the availability of high-quality entire mtGenome population reference data than the ease with which coding region data can be generated from evidence specimens. Until mtGenome reference databases are available, the utility of novel mtDNA typing technologies and the benefits of recovering mtDNA coding region information from forensic specimens will be limited. Thus, future mtDNA databasing efforts are needed for the development of entire mtDNA genome reference population data suitable for forensic comparisons.

Published by Elsevier Ireland Ltd.

Mitochondrial DNA (mtDNA) typing in forensic case work has historically focused on the two hypervariable segments (HVS) of the non-coding control region (CR) [1–5]. These approximately 600 bases have the highest average substitution rate in the mitochondrial genome (mtGenome), and thus present the greatest opportunity for inter-individual differentiation while minimizing data generation effort. It is the case, however, that examination of these 600 bases alone limits the power of forensic mtDNA testing in general, leading to situations in which HVS-I and HVS-II data do not provide sufficient discriminatory information to resolve distinct maternal lineages. Further resolution is often obtained by increasing the range of data analyzed to additional portions of the CR (e.g. with a sample of Austrians, analysis of the entire control region reduces the random match probability from 0.011 to 0.008) [6–8]. Yet, many individuals will remain indistinguishable despite complete CR data. In those cases, variation in the mtDNA coding region is often targeted [9–13].

It has been shown that mtDNA coding region data can be useful in a number of situations. For instance, it has been valuable in: resolving multiple casualty cases where more than one reference family shared the same mtDNA CR haplotype [14,15]; sorting and re-association of commingled remains [15]; increasing statistical support when exclusionary references are unavailable [16]; mtDNA haplogroup typing for rapid screening of casework specimens [17–19]; and assessing maternal bio-geographic ancestry as an investigative tool [20,21]. Additionally, coding region information has been strategically targeted in cases for which extremely limited evidentiary material is available following standard and, in these situations, non-distinguishing CR testing. In order to preserve the little remaining evidence for analyses likely to provide resolution, coding regions from the relevant reference samples were first investigated to identify sites that distinguished the reference lineages. These *case-specific* discriminatory sites were then directly typed on the remaining evidence material to ultimately establish identity [22].

Still, even in these very specific forensic scenarios, it is generally impractical to sequence large portions of the mtGenome. The cost

* Corresponding author. Tel.: +1 301 319 0244; fax: +1 301 295 5932.
E-mail address: jodi.a.irwin@us.army.mil (J.A. Irwin).
[1] Current address: National Institute of Standards and Technology, 100 Bureau Dr., M/S 8311, Gaithersburg, MD, 20899, USA.

1872-4973/$ – see front matter. Published by Elsevier Ireland Ltd.
doi:10.1016/j.fsigen.2010.02.008

and effort required to obtain even partial CR profiles from case specimens is substantial (especially in comparison to standard short tandem repeat (STR) typing), in part because mtDNA sequence data is usually sought when the genetic material is severely limited and/or compromised. Numerous short amplicons with adequate overlap among them, significant sequence coverage over each amplicon to ensure sufficient data quality, and highly redundant data analysis and review are required to produce CR haplotypes. Generation of coding region data for resolution of specific cases has therefore been not only prohibitively laborious for most practicing forensic laboratories, but also limited by the availability of sufficient evidentiary material. As a result, the forensic methods to access coding region data have typically involved either the optimization of a published assay or in-house development of sequencing or single nucleotide polymorphism (SNP) typing protocols that minimize effort and sample consumption ([23–25], for example).

Until recently there have been few commercial off-the-shelf products available for the generation of coding region data. Those that have been evaluated for forensic use have limited utility due to sample quantity requirements and other issues related to data quality standards required for forensic application [26,27]. However, a batch of new (or newly commercialized) technologies are emerging that will facilitate access to entire mtDNA genome data with relative ease and will likely make their way into forensic practice within the next few years. These include a coding region version of sequence-specific oligonucleotide arrays [28], coding region multiplexes for mass spectrometry [29–32] and so-called Next Generation Sequencing (NGS) technologies. The massively parallel sequencing enabled by NGS is revolutionizing genetic data generation and, in the not-too-distant future, is likely to make the development of entire mtDNA genome profiles from even highly degraded specimens relatively straight-forward and cost-effective [33–35]. Looking ahead then, it seems that the application of mtDNA coding region data in routine forensic casework will be dictated less by the quantity of specimen and/or effort required to produce the data than by the availability of large high-quality entire mtGenome population databases that can be used to determine the rarity of mtGenome haplotypes.

The lack of high-quality population databases covering the entire mtDNA coding region precludes a complete, empirically-based understanding of the additional discriminatory value that mtDNA coding region data may provide from randomly sampled individuals. Currently, GenBank is the only repository of complete mtGenomes that is regularly updated with new information. Although it contains a growing number of complete sequences, the available data are an imperfect substitute for a forensic reference database. Most of the sequences available in GenBank have not been produced as randomly sampled, unrelated individuals that are representative of particular population groups. For those populations that are represented, the datasets tend to be inconsistent in terms of the associated metadata required for their use in the forensic context. Further, because GenBank data are neither curated nor quality control checked, many sequences contain errors that may not only obscure precise estimates of mtDNA substitution rates (as required for likelihood calculations; [36]), but, more importantly, may also confound estimates of mtDNA haplotype frequencies. Finally, the tools available for GenBank searches are not the most useful for practical case work application. Search parameters that are specific to forensic mtDNA queries, including specific reference populations, inclusion/exclusion of polycytosine indels, and pre-defined sequence ranges, are unavailable and difficult to accommodate in the BLAST interface. Even novel tools that support the access and handling of GenBank mtDNA sequence data (e.g. MitoVariome [37]) fail to address specific alignment issues in length variant regions that are relevant to sequence comparisons in forensic casework [38].

Efforts are underway to improve and expand publicly available forensic mtDNA CR data sets: more than 5000 new sequences representing more than 30 populations will soon be available in the newest update of the EMPOP database (http://www.empop.org; [39]). While these data are substantially strengthening the foundation upon which *current* mtDNA identification efforts are based, they do not adequately prepare us for the recent and rapid advancements in mtDNA typing technologies that will soon facilitate access to coding region information in the most difficult forensic specimens.

Thus, future mtDNA databasing efforts are needed for the development of entire mtDNA genome reference population data suitable for forensic comparisons and which adhere to the same data quality standards already established for forensic control region reference population databases [40–42].

We should emphasize at this point that it is not our intention to advise on the precise coding region data to be utilized for forensic purposes, where the principal concern is detection of primary pathogenic mtDNA mutations. Although these variants, by their very nature, do not persist in the matriline, they arise spontaneously from time to time (and are therefore nearly always found in a heteroplasmic state), and are directly causal to disease phenotype when present in high enough proportion. In an effort to avoid this information, Coble et al. advocated a conservative strategy that targets information at synonymous sites only, suggesting that "This [targeting of synonymous variation] retains essentially an equal footing with accessing variation in the D-loop, which has yet not presented any problems" [43]. Although this statement is still valid, Mitomap [44] now lists 405 non-synonymous and structural RNA mutations; six synonymous and eight control region mutations with possible disease association. Although skepticism surrounds many of these reported associations [45–47], it is likely that our increasing understanding of mtDNA genomics, mitochondrial function and epigenetics may lead to the identification of additional pathogenic mutations. Mutations currently believed to be of no pathological significance (even those in non-coding regions) may be shown to be disease-associated in the future. But this is true for any genetic marker, including those routinely used in forensic testing (e.g. STRs). These and other pertinent medico-legal-ethics issues deserve further in-depth discussion as already begun in Coble et al. [12], Budowle et al. [48], and Coble et al. [43].

As a first step to employing coding region information in the forensic context, and in full accordance with appropriate Institutional Review Board (IRB) guidelines, the strategies of Brandstätter et al. [17], Lutz-Bonengel et al. [11], and Coble et al. [12], which target either silent mutations or sites with no presently known medically relevant mutations, are currently being employed in the authors' respective laboratories. In nearly every case encountered to date, the acquired coding region data have adequately resolved the question at hand. Instead, the primary limitation has been the lack of suitable population databases to assess the strength of the coding region evidence [22]. Appropriate mtGenome reference data are needed, so that they are readily available when specific laboratory, scientific working group or legislative guidelines are established for the use of coding region data.

The generation of high-quality entire mtGenome population reference datasets is clearly no small undertaking, particularly when considering that Sanger sequencing is the method currently used in most laboratories. New higher throughput technologies, such as mass spectrometry, may be preferred for their lower cost and higher capacity. However, this platform would produce population data specific to mass spectrometry applications. As a result, and until next generation sequencing methods are optimized and employed by more laboratories, the near-term effort will have to rely on technologies and protocols already used

to generate high-quality mtGenome data [12,49]. Such an undertaking will clearly require significant time, effort, funding and resources before even a few datasets of comparable size and quality to current control region databases are available. Yet, the long-term return on this investment will be novel high-quality entire mtGenome data that both positions the forensic community for the future of mtDNA testing and serves as a valuable resource for further characterization of mtDNA population genetics and molecular evolution as they relate to DNA evidence interpretation (e.g. mtDNA haplotype distributions, mtDNA substitution rates). With the large-scale availability of high-quality entire mtGenome data, forensic mtDNA interpretation guidelines can be greatly improved and the full potential of mtDNA testing can ultimately be realized.

Acknowledgements

We are grateful to Thomas J. Parsons, Hans-Jürgen Bandelt, Peter Gill and Frederick R. Bieber for their helpful suggestions. We also thank Odile Loreille for final manuscript review.

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 Armed Forces Institute of Pathology, the U.S. Department of Defense or the U.S. Department of the Army. In no case does specification of commercial equipment, instruments or materials imply a recommendation or endorsement by the Armed Forces Institute of Pathology, the U.S. Department of Defense or the U.S. Department of the Army, nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

References

[1] M.M. Holland, D.L. Fisher, L.G. Mitchell, W.C. Rodriquez, J.J. Canik, C.R. Merril, et al., Mitochondrial DNA sequence analysis of human skeletal remains: identification of remains from the Vietnam War, J. Forensic Sci. 38 (1993) 542–553.
[2] J.M. Butler, B.C. Levin, Forensic applications of mitochondrial DNA, Trends Biotechnol. 16 (1998) 158–162.
[3] M.M. Holland, T.J. Parsons, Mitochondrial DNA sequence analysis—validation and use for forensic casework, Forensic Sci. Rev. 11 (1999) 21–50.
[4] A. Carracedo, W. Bar, P. Lincoln, W. Mayr, N. Morling, B. Olaisen, et al., DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing, Forensic Sci. Int. 110 (2000) 79–85.
[5] T. Melton, K. Nelson, Forensic mitochondrial DNA analysis: two years of commercial casework experience in the United States, Croatian Med. J. 42 (2001) 298–303.
[6] S. Lutz, H.-J. Weisser, J. Heizmann, S. Pollak, Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany, Int. J. Legal Med. 111 (1998) 67–77.
[7] S. Lutz, H. Wittig, H.-J. Weisser, J. Heizmann, A. Junge, N. Dimo-Simonin, et al., 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 (2000) 97–101.
[8] A. Brandstätter, H. Niederstätter, M. Pavlic, P. Grubwieser, W. Parson, 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 (2007) 164–175.
[9] S.D. Lee, Y.S. Lee, J.B. Lee, Polymorphism in the mitochondrial cytochrome B gene in Koreans. An additional marker for individual identification, Int. J. Legal Med. 116 (2002) 74–78.
[10] H. Andreasson, A. Asp, A. Alderborn, U. Gyllensten, M. Allen, Mitochondrial sequence analysis for forensic identification using pyrosequencing technology, BioTechniques 32 (2002) 124–133.
[11] S. Lutz-Bonengel, U. Schmidt, T. Schmitt, S. Pollak, Sequence polymorphisms within the human mitochondrial genes MTATP6, MTATP8, and MTND4, Int. J. Legal Med. 117 (2003) 133–142.
[12] M.D. Coble, R.S. Just, J.E. O'Callaghan, I.H. Letmanyi, C.T. Peterson, J.A. Irwin, et al., Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians, Int. J. Legal Med. 118 (2004) 137–146.
[13] B. Quintáns, V. Álvarez-Iglesias, A. Salas, C. Phillips, M.V. Lareu, A. Carracedo, Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing, Forensic Sci. Int. 140 (2004) 251–257.
[14] K.A. Sturk, M.D. Coble, S.M. Barritt, T.J. Parsons, R.S. Just, The application of mtDNA SNPs to a forensic case, Forensic Sci. Int. Genet. Supplement Series 1 (2008) 295–297.
[15] R.S. Just, M.D. Leney, S.M. Barritt, C.W. Los, B.C. Smith, T.D. Holland, et al., The Use of mitochondrial DNA single nucleotide polymorphisms to assist in the resolution of three challenging forensic cases, J. Forensic Sci. 54 (2009) 887–891.
[16] J.A. Irwin, S.M. Edson, O. Loreille, R.S. Just, S.M. Barritt, D.A. Lee, et al., DNA identification of "Earthquake McGoon" 50 years postmortem, J. Forensic Sci. 52 (2007) 1115–1118.
[17] A. Brandstätter, T.J. Parsons, W. Parson, Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups, Int. J. Legal Med. 117 (2003) 291–298.
[18] W. Parson, A. Brandstätter, H. Niederstätter, P. Grubwieser, R. Scheithauer, Unraveling the mystery of Nanga Parbat, Int. J. Legal Med. 121 (2007) 309–310.
[19] V. Álvarez-Iglesias, J.C. Jaime, Á. Carracedo, A. Salas, Coding region mitochondrial coding region SNPs: Targeting East Asian and Native American haplogroups, Forensic Sci. Int. Genet. 1 (2007) 44–55.
[20] T.M. Nelson, R.S. Just, O. Loreille, M.S. Schanfield, D. Podini, Development of a multiplex single base extension assay for mitochondrial DNA haplogroup typing, Croatian Med. J. 48 (2007) 460–472.
[21] S. Köhnemann, H. Pfeiffer, Application of mtDNA SNP analysis in forensic casework, Forensic Sci. Int. Genet. 5 (2011) 216–221.
[22] R.S. Just, O.M. Loreille, J.E. Molto, D.A. Merriwether, S.R. Woodward, C. Matheson, et al., Titanic's unknown child: the critical role of the mitochondrial DNA coding region in a re-identification effort, Forensic Sci. Int. Genet. 5 (2011) 231–235.
[23] M. Allen, H. Andreasson, Mitochondrial d-loop and coding sequence analysis using pyrosequencing, Methods Mol. Biol. 297 (2004) 179–196.
[24] P.M. Vallone, R.S. Just, M.D. Coble, J.M. Butler, T.J. Parsons, A multiplex allele-specific primer extension assay for forensically informative SNPs distributed throughout the mitochondrial genome, Int. J. Legal Med. 118 (2004) 147–157.
[25] A. Salas, B. Quintáns, V. Álvarez-Iglesias, SNaPshot typing of mitochondrial DNA coding region variants, in: A. Carracedo (Ed.), Forensic DNA Typing Protocols, Methods in Molecular Biology, Humana Press, Totowa, New Jersey, 2005 pp. 197–208.
[26] P.M. Vallone, J.P. Jakupciak, M.D. Coble, Forensic application of the Affymetrix human mitochondrial resequencing array, Forensic Sci. Int. Genet. 1 (2007) 196–198.
[27] R.S. Just, A.M. Lehrmann, K.E. Harris, M.D. Coble, Comparison of the AB mitoSEQr resequencing sets to standard mtDNA sequencing protocols, in: Presented at the English Speaking Working Group Meeting of the International Society for Forensic Genetics, Sinaia, Romania, 2008.
[28] C. Calloway, S. Stuart, H. Erlich, Development of a multiplex PCR and linear array probe assay targeting informative polymorphisms within the entire mitochondrial genome, 2009. http://www.ncjrs.gov/pdffiles1/nij/grants/228279.pdf.
[29] T.A. Hall, B. Budowle, Y. Jiang, L. Blyn, M. Eshoo, K. Sannes-Lowery, et al., 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 (2005) 53–69.
[30] H. Oberacher, H. Niederstätter, F. Pitterl, W. Parson, Profiling 627 mitochondrial nucleotides via the analysis of a 23-plex polymerase chain reaction by liquid chromatography–electrospray ionization time-of-flight mass spectrometry, Anal. Chem. 78 (2006) 7816–7827.
[31] T.A. Hall, K.A. Sannes-Lowery, L.D. McCurdy, C. Fisher, T. Anderson, A. Henthorne, et al., Base composition profiling of human mitochondrial DNA using polymerase chain reaction and direct automated electrospray ionization mass spectrometry, Anal. Chem. 81 (2009) 7515–7526.
[32] M. Cerezo, V. Černý, A. Carracedo, A. Salas, Applications of MALDI-TOF MS to largescale human mtDNA population-based studies, Electrophoresis 30 (2009) 3665–3673.
[33] E.I. Rogaev, A.P. Grigorenko, G. Faskhutdinova, E.L.W. Kittler, Y.K. Moliaka, Genotype analysis identifies the cause of the "royal disease", Science 326 (2009) 817.
[34] B. Brenig, J. Beck, E. Schütz, Shotgun metagenomics of biological stains using ultra-deep DNA sequencing, Forensic Sci. Int. Genet., (2009), doi:10.1016/j.fsigen.2009.10.001.
[35] J. Krause, A.W. Briggs, M. Kircher, T. Maricic, N. Zwyns, A. Derevianko, et al., A complete mtDNA genome of an early modern human from Kostenki, Russia, Curr. Biol. 20 (2010) 231–236.
[36] Y.-G. Yao, A. Salas, I. Logan, H.-J. Bandelt, MtDNA data mining in GenBank needs surveying, Am. J. Hum. Genet. 85 (2009) 929–933.
[37] Y.S. Lee, W.Y. Kim, M. Ji, J.H. Kim, J. Bhak, MitoVariome: a variome database of human mitochondrial DNA, BMC Genomics (2009) S12.
[38] H.-J. Bandelt, W. Parson, Consistent treatment of length variants in the human mtDNA control region: a reappraisal, Int. J. Legal Med. 122 (2008) 11–12.
[39] W. Parson, A. Dür, EMOP—a forensic mtDNA database, Forensic Sci. Int. Genet. 1 (2007) 88–92.
[40] W. Parson, H.-J. Bandelt, Extended guidelines for mtDNA typing of population data in forensic science, Forensic Sci. Int. Genet. 1 (2007) 13–17.
[41] J. Irwin, J. Saunier, K. Strouss, K. Sturk, T. Diegoli, R. Just, M. Coble, W. Parson, T. Parsons, Development and expansion of high quality control region databases to improve forensic mtDNA evidence interpretation, Forensic Sci. Int. Genet. 1 (2007) 154–157.
[42] A. Carracedo, J.M. Butler, L. Gusmão, W. Parson, L. Roewer, P.M. Schneider, Publication of population data for forensic purposes, Forensic Sci. Int. Genet. 4 (2010) 145–147.

[43] M.D. Coble, P.M. Vallone, R.S. Just, T.M. Diegoli, B.C. Smith, T.J. Parsons, Effective strategies for forensic analysis in the mitochondrial DNA coding region, Int. J. Legal Med. 120 (2006) 27–32.
[44] MITOMAP: A human mitochondrial genome database. http://www.mitomap.org, 2009.
[45] H.-J. Bandelt, A. Salas, C.M. Bravi, What is a 'novel' mtDNA mutation—and does 'novelty' really matter? J. Hum. Genet. 51 (2006) 1073–1082.
[46] H.-J. Bandelt, Y.-G. Yao, A. Salas, The search of 'novel' mtDNA mutations in hypertrophic cardiomyopathy: MITOMAPping as a risk factor, Int. J. Cardiol. 126 (2008) 439–442.
[47] H.-J. Bandelt, A. Salas, R.W. Taylor, Y.-G. Yao, Exaggerated status of "novel" and "pathogenic" mtDNA sequence variants due to inadequate database searches, Hum. Mutat. 30 (2009) 191–196.
[48] B. Budowle, U. Gyllensten, R. Chakraborty, M. Allen, Forensic analysis of the mitochondrial coding region and association to disease, Int. J. Legal Med. 119 (2005) 314–315.
[49] L. Fendt, B. Zimmermann, M. Daniaux, W. Parson, Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences, BMC Genomics 10 (2009) 139.