Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International: Genetics 1 (2007) 154–157
FSI GENETICS
www.elsevier.com/locate/fsig

Development and expansion of high-quality control region databases to improve forensic mtDNA evidence interpretation

Jodi A. Irwin [a], Jessica L. Saunier [a], Katharine M. Strouss [a], Kimberly A. Sturk [a], Toni M. Diegoli [a], Rebecca S. Just [a], Michael D. Coble [a,*], Walther Parson [b], Thomas J. Parsons [a,1]

[a] *Armed Forces DNA Identification Laboratory (AFDIL), 1413 Research Blvd., Rockville, MD 20850, USA*
[b] *Institute of Legal Medicine, Innsbruck Medical University, Müllerstrasse 44, 6020 Innsbruck, Austria*

Received 22 January 2007; accepted 27 January 2007

Abstract

In an effort to increase the quantity, breadth and availability of mtDNA databases suitable for forensic comparisons, we have developed a high-throughput process to generate approximately 5000 control region sequences per year from regional US populations, global populations from which the current US population is derived and global populations currently under-represented in available forensic databases. The system utilizes robotic instrumentation for all laboratory steps from pre-extraction through sequence detection, and a rigorous eight-step, multi-laboratory data review process with entirely electronic data transfer. Over the past 3 years, nearly 10,000 control region sequences have been generated using this approach. These data are being made publicly available and should further address the need for consistent, high-quality mtDNA databases for forensic testing.
Published by Elsevier B.V.

Keywords: Mitochondrial DNA; mtDNA databases; Robotics; Control region

1. Introduction

Mitochondrial DNA testing in the forensic context requires appropriate, high-quality population databases for estimating the rarity of questioned haplotypes. However, large forensic mtDNA databases, which adhere to strict guidelines in terms of their generation and maintenance, are not widely available for many regional populations of the United States or most global populations outside of the United States and Western Europe. The mtDNA databases that are available for these populations have primarily been generated for independent anthropological studies and, as a result, are by and large inconsistent in terms of the regions sequenced, the nomenclature used, and the overall data quality. Furthermore, both forensic and anthropological mtDNA databases have recently been shown to harbour a rather high rate of so-called ''phantom mutations'' that have been detected via phylogenetic analysis [1–6]. These erroneous polymorphisms, resulting from errors in the interpretation of data, inadvertent sample switches, or manual data transcription have highlighted the need for high-quality mitochondrial data, particularly in the forensic arena.

In order to address this issue, the Armed Forces DNA Identification Lab (AFDIL) has undertaken a high-throughput control region databasing effort. The system is currently designed to generate upwards of 4000 sequences per year. Over the next year and one-half, using a robust approach for both data generation and analysis, we plan to sequence over 7500 individuals. Global populations that are currently under-represented in available forensic mtDNA databases will comprise approximately 25% of the total number of samples. The remaining individuals will represent regional samples of various U.S. populations and global populations that contribute to the overall mtDNA diversity of the U.S. The high-quality mtDNA data generated from these efforts will be publicly available to permit examination of regional mtDNA sub-structure and admixture, and ultimately to improve our ability to interpret mtDNA evidence.

* Corresponding author. Tel.: +1 301 319 0268; fax: +1 301 295 5932.
E-mail address: michael.coble@afip.osd.mil (M.D. Coble).
[1] Current address: International Commission on Missing Persons, Alipašina 45 A, 71000 Sarajevo, Bosnia and Herzegovina.

1872-4973/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.fsigen.2007.01.019

2. Materials and methods

A robust system has been developed to process samples through the laboratory, analyze sequence data, and reliably transfer data to a master database. The system accommodates various sample substrates, but is primarily designed for buccal swabs and blood stained cards. Laboratory processes have been completely automated with the use of a Wallac DBS hole puncher, a Qiagen 9604 for extraction, a Corbett CAS-1200 for amplification, a Tecan Genesis workstation for sequencing, and Applied Biosystems 3130*xl*s for electrophoresis. In addition, multiple scientists are present at key laboratory steps in order to eliminate sample processing errors.

The amplification and sequencing strategy of Brandstätter et al. [7], which includes significant primer redundancy and base coverage, has been modified slightly to incorporate more efficient primers that minimize sequencing re-runs. Sequencing primers R274, R16175, F15 and R16400 have been replaced with R285 (GTTATGATGTCTGTGTGGAA), R16410 (GAGGATGGTGGTCAAGGGA), R484 (TGAGATTAGTAGTATGGGAG) and F34 (GGGAGCTCTCCATGCATTTGGTA).

Automated laboratory processing is followed by a rigorous data review process, which involves independent, redundant data analysis by multiple individuals in different laboratories. At least four scientists review the raw data for every sample—two scientists at the AFDIL and another two scientists at the Institute of Legal Medicine in Innsbruck, Austria who confirm the AFDIL analysis for inclusion of the sequences into the EMPOP database. To ensure the proper transfer of data to production databases available for haplotype comparison, an automated export is used to transfer reviewed and confirmed sequence data to a secured, master environment. Once imported, data are finalized with two additional reviews that evaluate the imported polymorphisms against the original electropherogram reviews. Finalized data are being made publicly available through publication and submission to Genbank, SWGDAM and EMPOP.

3. Results and discussion

During the past 3 years, we have generated over 7000 control region sequences from global populations (Table 1). At least 27 population samples are currently represented. When regional sub-populations are considered, approximately 50 different populations and sub-populations have been typed. As databases have been completed by both AFDIL and EMPOP, the data have been (and are being) made publicly available through Genbank, EMPOP and publication in peer-reviewed journals (e.g. [7,8]).

Thus far, both the laboratory processing and data review procedures have proven to be extremely robust. In the laboratory, multiple safeguards are in place at key steps: multiple scientists are present to witness initial sample placement and cherry picking for re-dos; robotics enforce the standard placement of samples, reagent blanks and negative controls, thereby eliminating the potential for sample switches at every step; and a highly redundant sequencing strategy provides at least double strand coverage over the entire control region (with the exception of short stretches adjacent to length heteroplasmic C-stretches, where confirmation is only possible with multiple reads from a single strand). In terms of data output, the most limiting factor has proven to be initial sample quality. The system is, of course, designed primarily for pristine, high-copy number samples. However, from blood and buccal swabs submitted by diverse collaborators, we regularly encounter degraded samples that do not yield a 1 kb amplicon. For these situations, we have designed additional robotic and biochemical protocols (three overlapping amplicons, revised sequencing strategy [AFDIL, unpublished]) to accommodate the lower sample quality. The safeguards already implemented for the highest quality samples are particularly helpful for these more challenging specimens, because they generally require more work and, as a result, introduce more opportunity for errors. For instance, the standard placement of samples on plates guards against artificial recombination between amplicons from the same individual. Likewise, pipetting errors and inadvertent sample switches are minimized by the presence of multiple scientists when samples are cherry-picked from plates. Thus, despite the fact that variability in sample quality sometimes limits our overall productivity, the integrity of the final sequence data is abetted by the overall system design.

The high data-quality standards that we require, as well as the sheer redundancy in sequence data and base coverage, facilitate the post-laboratory data review and analysis steps.

Table 1
Global and U.S. populations databased by AFDIL for the entire mitochondrial control region since 2004

Database	Completed samples
Afghanistan	98
African American (U.S.)	484
Asian (U.S.)	43
Bahrain	218
Caucasian (U.S.)	557
China	383
Cyprus	91
United Arab Emirates	191
Egypt	278
Greece	319
Hispanic (U.S.)	1039
Hungary: Baranya County Roma	205
Hungary: Budapest Caucasian	211
Indonesia	279
Iraq	189
Jordan	210
Kazakhstan	256
Kyrgyzstan	249
Lebanon	170
Kenya: Nairobi	100
Native Americans (U.S.)	120
Puerto Rico	209
Russia	151
Tajikistan	244
Turkmenistan	249
Uzbekistan	331
Vietnam	187
Total	7061

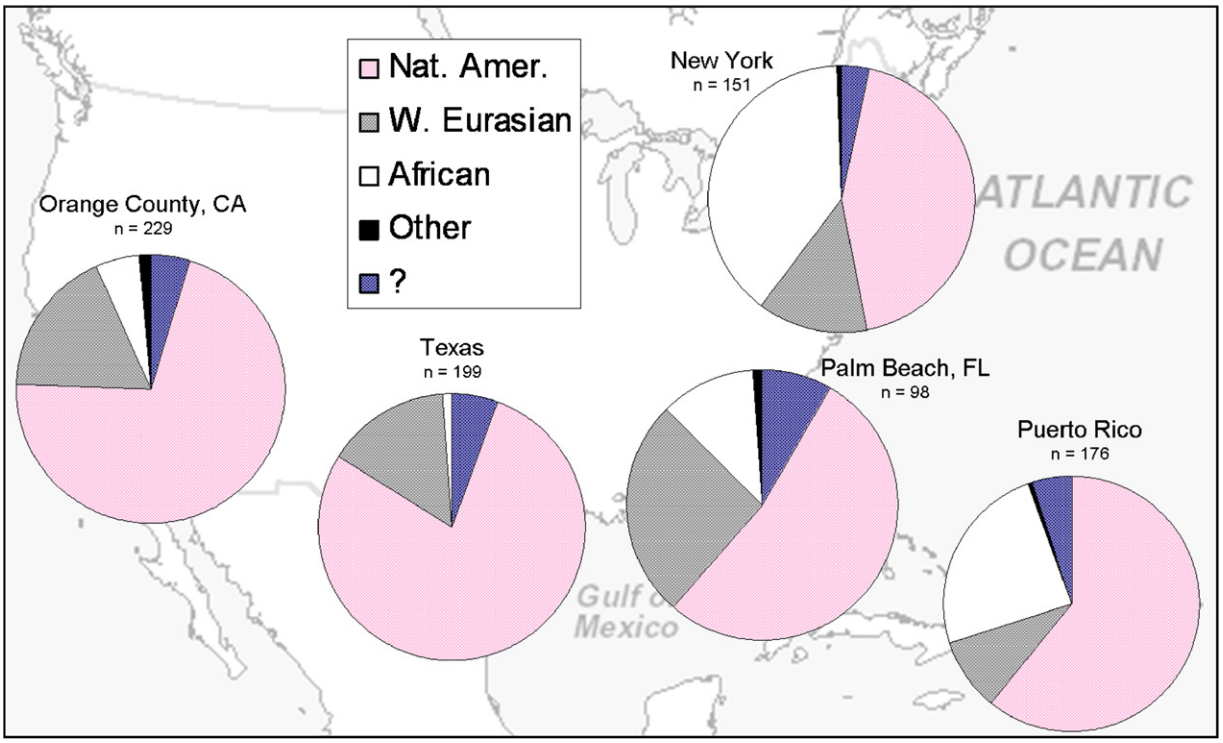

Fig. 1. Mitochondrial DNA haplogroup distribution among 853 regional United States "Hispanics". All inter-population pairwise F_{st} values are significant at the 0.05 level.

With the raw data for every sample reviewed by four scientists, all data electronically transferred and a final phylogenetic check of the data by EMPOP, no phantom mutations or other artifactual errors have been detected at the final data check step over the past 3 years.

In terms of the inter-laboratory analyses, only minor interpretational differences of the data have been encountered. The discrepancies have all related to either the interpretation and nomenclature of unusual length variants or the designation of point and length heteroplasmies. While the heteroplasmy issues have been easy to resolve, the nomenclature of novel length variants continues to be a significant challenge in developing inter-laboratory data consistency. This is a serious consideration when novel length variants are encountered, as different nomenclature schemes from various laboratories can potentially skew database searches and underestimate the frequency of length variant haplotypes. It is true that phylogenetic information provides useful data upon which to base nomenclature [9]. However, we at AFDIL find ourselves constrained by nomenclature guidelines that have been in use in our laboratory for a number of years and are now reflected in thousands of in-house control region sequences. The forensic mitochondrial DNA community would benefit from additional discussions of length variant nomenclature and further investigation of potential solutions to this increasingly intractable issue.

In addition to providing high-quality reference data for the forensic community, we also hope to use these data to better understand the magnitude and significance of mtDNA variation among regional sub-populations. Because diverse population groups may differ significantly in their mtDNA haplotype distributions, the maintenance of forensic mtDNA databases should reflect this differentiation in order to provide the best estimate of haplotype frequencies [10]. However, for many populations, the extent of inter-population, or regional sub-population, differentiation is not well-understood simply because it has not yet been thoroughly examined. With diverse regional population sampling, particularly of component U.S. populations, we hope to study this issue in order to better understand the level at which separate mtDNA reference databases should be maintained for U.S. populations.

A preliminary analysis of 853 U.S. "Hispanics", representing five regional sub-samples sequenced over the past 2 years, shows significant substructure (Fig. 1; all pairwise F_{st} values are significant at the 0.05 level). While 39% of New York "Hispanics" exhibit African mtDNA haplotypes, only 5% and 1% of "Hispanics" from California (Orange County) and Texas, respectively, reflect African-derived lineages. Likewise, the proportion of Native American haplotypes differs dramatically between regional sub-populations of "Hispanics". Forty-four percent of the New York sample comprises Native American haplotypes, as opposed to 70% and 78% of California and Texas samples, respectively. Obviously, the term "Hispanic" does not refer to a real population in the biological sense. Rather it is a poorly defined historical term relating to Spanish/Portuguese speaking people with at least partial cultural and/or genetic heritage from Central or South America. The differentially admixed nature of individuals labeled as "Hispanic" is obvious from our data. Additional regional "Hispanic" samples, as well as diverse samples of "Caucasian" and African American

populations, will offer a better picture of the degree of geographic differentiation among U.S. populations and provide databases that better represent any regional substructure that may exist.

4. Conclusion

The Armed Forces DNA Identification Laboratory, in collaboration with EMPOP, has implemented a high-throughput control region databasing system in an effort to improve the quantity and quality of mtDNA data available to the international forensic community. These data, for both U.S. and global populations, are not only providing a framework within which to interpret forensic mtDNA data, but are also providing a comprehensive picture of the variability of global mtDNA types. AFDIL's databasing efforts are being made available for public use and will be the basis for further, in-depth analyses of mitochondrial DNA diversity and its forensic implications.

Acknowledgements

The authors wish to thank Jennifer O'Callaghan, Carla Paintner, Kimberly Watson, Heather Williams, Melissa Scheible, Leslie Mounkes, and Naila Bahtri (AFDIL) for data generation, as well as Anita Brandstätter (EMPOP, Innsbruck Medical University, Austria) for assistance with database confirmation. We also thank James Canik, Kevin Carroll, Brion Smith, Louis Finelli and James Ross (AFDIL) for logistical, administrative, and computer support. A large portion of this work was funded by United States National Institute of Justice grants 2000-IJ-CX-K010 and 2003-DN-R-086 to TJP, as well as an inter-agency agreement 2005-91821-DC-IJ with the Armed Forces Institute of Pathology. 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, or the U.S. Department of the Army.

References

[1] H.J. Bandelt, P. Lahermo, M. Richards, V. Macaulay, Detecting errors in mtDNA data by phylogenetic analysis, Int. J. Legal Med. 115 (2001) 64–69.
[2] H.J. Bandelt, L. Quintana-Murci, A. Salas, V. Macaulay, The fingerprint of phantom mutations in mitochondrial DNA data, Am. J. Hum. Genet. 71 (2002) 1150–1160.
[3] H.J. Bandelt, A. Salas, S. Lutz-Bonengel, Artificial recombination in forensic mtDNA population databases, Int. J. Legal Med. 118 (2004) 267–273.
[4] H.J. Bandelt, A. Salas, C. Bravi, Problems in FBI mtDNA database, Science 305 (2004) 1402–1404.
[5] P. Forster, To err is human, Ann. Hum. Genet. 67 (2003) 2–4.
[6] C. Herrnstadt, G. Preston, N. Howell, Errors, phantoms and otherwise, in human mtDNA sequences, Am. J. Hum. Genet. 72 (2003) 1585–1586.
[7] A. Brandstätter, C. Peterson, J. Irwin, S. Mpoke, D. Koech, W. Parson, T.J. Parsons, Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic and population genetic parameters for the establishment of a forensic database, Int. J. Legal Med. 118 (2004) 294–306.
[8] J. Irwin, B. Egyed, J. Saunier, G. Azamosi, J. O'Callaghan, Z. Padar, T. Parsons, Hungarian mtDNA population databases from Budapest and the Baranya County Roma, Int. J. Legal Med., in press.
[9] H.J. Bandelt, W. Parson, Consistent treatment of length variants in the human mtDNA control region: a reappraisal, Int. J. Legal Med., in press.
[10] M. Holland, T. Parsons, Mitochondrial DNA sequence analysis—validation and use for forensic casework, For. Sci. Rev. 11 (1999) 22–50.