

Clustering for Forensic Mitotype Quality Analysis

Bobi K Den Hartog(1), John W Elling* (1), Deborah Polanskey (2), Constance L Fisher (2)

(1) MitoTech, LLC, Santa Fe, NM, United States

(2) FBI Laboratory, Quantico, VA, United States

Abstract

Errors in mitochondrial DNA (mtDNA) sequencing can result in unusual patterns of polymorphisms which can be detected by the lack of similarity to the sequences in a database. In this work, an approach to routine data quality review is described that employs cluster analysis to identify conserved sequence similarities. Comparison of new sequences to the clusters identifies unusual polymorphisms and sequence regions that may warrant further attention by the analyst. This approach is accessible, easily automatable and can be tailored specifically to targeted population groups in individual investigations.

Key Words

forensic science, mitochondrial DNA, phylogenetics, sequence error, SWGDAM database, EMPOP database

Introduction

Phylogenetic clustering analysis has been helpful in detecting errors in mtDNA data by identifying unusual occurrences of sequence variation [1², 3, 4]. Bandelt *et. al.* (2001) discuss and illustrate many different sources of experimental, data analysis, and clerical errors. Most error mechanisms will affect individual polymorphisms in the data - which could be overlooked, misdescribed, mislocated, or spurious. "Artificial recombination" errors are also observed when sequence regions from different sources are reported in the sample sequence [3]. The error rate in forensic databases is low and, with the increasing automation of sample preparation, sequencing, data analysis, difference reporting, and data basing, the risk of many of these errors has declined [4, 5]. However there remain numerous opportunities for experimental error and careful review of the data is always warranted.

Bandelt *et. al.* (2001) propose a method of reviewing mtDNA sequence variations that begins with haplogroup assignment of the sequence into a phylogeny to compare the sequence to "phylogenetic neighbors". However the use of haplogroup assignment presupposes both the availability of a phylogenetic network and coverage of every position in the sample that defines the relevant haplogroups. Forcing a haplogroup assignment can also result in a mistaken indication of an error [4]. Other statistical clustering techniques without this overhead are more amenable to investigating the

similarity of the observed polymorphisms from groups of previously-observed sequences in order to identify unusual variations and sequence regions. Here we describe a data review process based in part on standard k-means clustering that can be used as a general review tool and potentially tailored to specific population studies.

Method

Two well-curated public mitotype databases, the European Mitochondrial DNA Population database (EMPOP) [6] and SWGDAM[7] were used to develop and test these techniques. Only polymorphisms in the HV1 region (reference positions 16024 to 16365) and HV2 region (reference positions 73 to 340) were included in the analyses. In an intermediate release of the SWGDAM database there are 3552 unique mitotypes that occur in 6595 samples. In the EMPOP database 2622 unique mitotypes were observed in 5173 samples. The mitotype sequence descriptions in the EMPOP database use a phylogenetic alignment which in a very few cases differ from the description generated with standard hierarchical rules and in these cases the types were converted [8, 9]. Polymorphisms in these HV1 and HV2 regions defined 5212 unique mitotypes in the two databases.

One thousand and ten unique polymorphisms were found in 421 of the 640 sequence positions and 30 insertion positions across all 5212 mitotypes. A unique polymorphism is defined to be a unique base or deletion at a unique reference position. A binary descriptor vector was created to describe each mitotype in the databases, recording the presence or absence of each observed unique polymorphism in 1010 bits. Using this scheme, for example, the insertion of a T after the reference 191 position would be described by setting the 191.1T bit equal to 1 as well setting 191.1- equal to 0 (where the latter describes the lack of the reference's absence of an insertion after 191).

In this work, standard k-means clustering using a Euclidian distance calculation was used to partition the 5212 mitotypes into 200 clusters. The number of clusters was chosen to roughly avoid over fitting the data. Other methods of statistical clustering are equally valid and are being explored.

Data Review Process

The data review process proceeds in a three step process.

1. Novel Polymorphisms. Following sequencing and standard difference-from-reference reporting, the polymorphism in the mitotype of a new sample is first checked against the list of 1010 unique polymorphisms. The data supporting the discovery of any novel polymorphism should be carefully rechecked.
2. Cluster Fit Threshold. The new sequence is compared to the centroid of the closest cluster. The Euclidian distance from the sample mitotype to the closest cluster is compared to the size of the cluster. The cluster size is calculated to be the mean distance of the cluster members from the centroid and the standard deviation of these distances provides a threshold to identify a poor fit. If the mitotype being reviewed is further than the mean plus one standard deviation from the centroid, the poor fit is highlighted.

3. Cluster Comparison. The new sequence is compared to the centroid of the few closest clusters. Each bit in the descriptor vector is compared to the 1010 real value elements defining the cluster centroid and the significant differences are highlighted. Any sequence value that is significantly different than the cluster centroid value is highlighted so that the analyst can reconfirm the sequence position.

Case Studies

In an earlier version of the SWGDAM database, three sequences with suspected artificial recombination artifacts (USA.AFR.000063, USA.AFR.000074, and USA.AFR.000942) were identified and subsequently confirmed and corrected [3, 4]. Figure 1 shows a plot of the distance of the original and corrected USA.AFR.000942 sequence from the five nearest clusters along with a box plot showing the size of each cluster. The erroneous sequence is a clear outlier from all five of the nearest clusters and so is easily identified as suspicious. Once corrected, Figure 1 shows that the sequence is an excellent fit in the closest cluster.

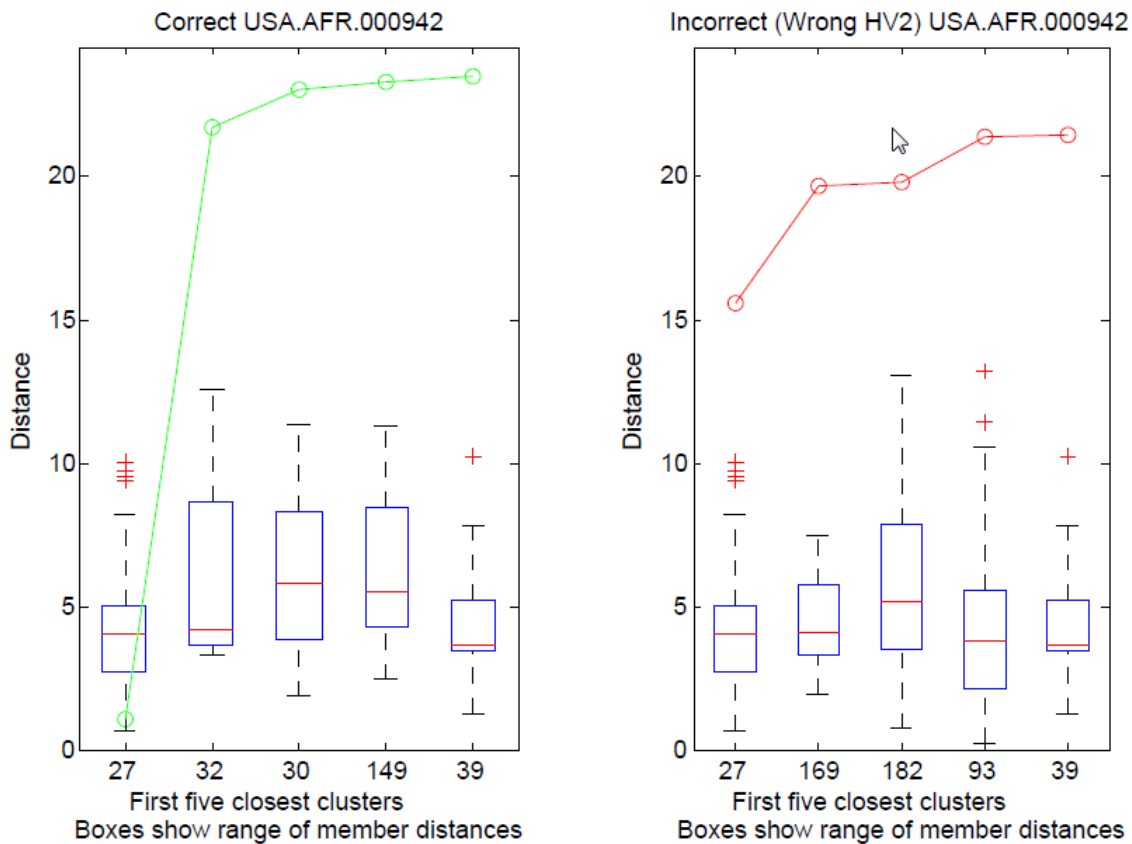


Figure 1 The original USA.AFT.000942 sequence with an artificial recombination error does not fall within any cluster of sequences in the database (right), indicating a potential problem. Once corrected, the sequence is a good fit within the closest cluster (left). In both plots the distance to the closest five clusters is shown along with a box plot of the distribution of members of each cluster as an indication of the cluster size. A new sample must fall within the cluster distribution (|--|--|) and ideally within the standard deviation box to be considered a member of that cluster.

Phylogenetic network analysis was used to identify eight suspect samples (USA.HIS.000093, USA.HIS.000100, USA.HIS.000110, USA.HIS.000204, USA.HIS.000267, USA.HIS.000274, USA.HIS.000552, and USA.HIS.000770) in the SWGDAM Hispanic data set based on the absence of the 235G mutation. Transcriptional errors in three samples, USA.HIS.000100, USA.HIS.000110, and USA.HIS.000204, were confirmed and the remaining five sequences were validated [4]. When analyzing the original suspect sequences in the review process described here, the “253G” transcription error in USA.HIS.000100 and the “236G” transcription error in HIS.000110 and both the “153C” and “235C” transcription errors in the original sequence for USA.HIS.000204. were immediately highlighted for review in step 1 as novel polymorphisms

Phylogenetic classification incorrectly led to suspicion of errors in the five other Hispanic sequences above due to the lack of 235G mutation despite their putative membership in the A2 haplogroup [3]. In the data review process described here, four of these sequences fell into clusters defined in part by a statistically significant presence of the 235G and so this position would have also been highlighted for further review.

Conclusion

A straightforward procedure for routine quality review is described that employs cluster membership and difference to identify unusual mitotypes and individual polymorphisms that warrant review by the analyst. The review process can be tailored to the specific investigation by clustering databases that are representative of the investigations. In this example, two diverse databases were combined for clustering which is appropriate when no target population for the sequence being reviewed can be predicted.

Role of Funding

Funding for this work was provided by the United States Department of Justice

Acknowledgements

None

Conflict of Interest

Two authors, Drs Den Hartog and Elling, are primarily employed by MitoTech LLC. The commercially-available Mitotyper(tm) software developed by MitoTech was employed in this research.

¹ Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, et al. (2007) The Genographic Project public participation mitochondrial DNA database. *PLoS Genet* 3(6): e104. doi:10.1371/journal.pgen.0030104 - see the bottom of page 1084

² 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.

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 Budowle B, Polanskey D, Allard MW, Chakraborty R, Addressing the use of phylogenetics for identification of sequences in error in the SWGDAM mitochondrial DNA database, *J. Forensic Sci.* Volume 49, Issue 6 (November 2004)

⁵ Budowle B, Polanskey D, Fisher CL, Den Hartog BK, Kpler RB, Elling JW (2010) Automated alignment and nomenclature for consistent treatment of polymorphisms in the human mitochondrial DNA control region. *J For Sci* in press

⁶ Parson W, Dur A. EMPOP - A forensic mtDNA database. *For Sci Int: Genetics* 1 (2007) 88-92

⁷ Monson KL, Miller KWP, Wilson MR, DiZinno JA, Budowle B. The mtDNA population database: an integrated software and database resource for forensic comparison. *For Sci Comm* 2002;4(2).

⁸ Bandelt HJ, Parson W. Consistent treatment of length variants in the human mtDNA control region: a reappraisal. *Int J Legal Med* 2008;122:11-21.

⁹ Polanskey D, Den Hartog BK, Elling JW, Fisher CL, Kepler RB, Budowle B (2010) J. Comparison of MitotyperTM and Phylogenetic-based mtDNA Nomenclature Systems. *J For Sci*, in press