mtDNA Data Mining in GenBank Needs Surveying

To the Editor: Since the first sequencing of the complete human mtDNA genome,[1] both the sequencing techniques and the quality of commercial kits have improved greatly. This has led to a growing number of reports for complete mtDNA sequences from the fields of molecular anthropology, medical genetics, and forensic science; and there are now over 6700 complete or near-complete mtDNA sequences available for study.[2] However, in comparison to the pioneer manual-sequencing efforts in the early nineties, the overall mtDNA data quality, especially in the medical field, is still far from satisfactory.[3] Sequencing errors and inadvertent mistakes in the reported mtDNA data are not infrequent.[4–10] Deficient mtDNA data sets of complete genomes can have important consequences for the conclusions achieved in many studies and may also pose problems for any subsequent reanalyses.

Most recently, Pereira and colleagues[11] discussed the overall picture of the mtDNA genome diversity in worldwide human populations with a comprehensive reanalysis of 5140 published complete or near-complete (lacking some control region information) mtDNA sequences. Their study represents an important advance in defining the effects of gene structures on limiting mtDNA diversity and may have valuable implications for mtDNA studies in the medical field.[11] However, all of the data used in the study by Pereira et al.[11] were directly retrieved from GenBank without any scrutiny for problematic or flawed data that should have been excluded. Many of the mtDNA sequences analyzed in their study[11] have in fact already been questioned in the literature or even corrected by their authors, but unfortunately, in several instances the new corrected versions of the sequences have not been made generally available or updated in GenBank.

In Table 1, we list some of the problematic data sets and single sequences used by Pereira et al. in their study.[11] Among them is the original data set of Herrnstadt et al.,[12] which was announced by the authors[13] as having been corrected, although the new sequences have never been entered into GenBank. Portions of those coding-region data (in either corrected or uncorrected form) were augmented by the associated control-region data and published in several papers; thus, none of these expanded data can be downloaded from GenBank but have to be retrieved from the figures in the corresponding articles. To cite a more recent example, the African mtDNA data set published by Gonder et al.[14] is of particularly poor quality. These sequences are incompletely recorded (as already mentioned by Behar et al.[15]); the most extreme instance of this is the haplogroup L0k1 sequence EF184609 that lacks as many as 25 expected variants scattered along the whole pathway from the haplogroup root to the revised Cambridge reference sequence (rCRS).[16] Also, several different phantom mutations appear throughout the data set; in particular, five sequences have been affected by phantom base changes to G within the short 9949–9978 stretch. We have annotated problems in 14 sequences by way of example, but nearly all sequences of Gonder et al.[14] may suffer from overlooked variants, except for the three sequences from the well-described West Eurasian haplogroups J1 and N1. Additional details are given in the Supplemental Data, available online.

Again, if one examines the ten Vietnamese complete mtDNA sequences that were submitted to GenBank by Phan et al. and used in the Pereira et al. study,[11] it is possible to see errors of many kinds. First, all sequences miss three expected variants (A263G, 315+C [or written as 315insC], and C14766T). Second, there are many phantom mutations that are not observed elsewhere. Third, several sequences are incomplete; e.g., the haplogroup M7b1 sequence DQ826448 lacks an additional nine expected variants by oversight or artefactual recombination. This sequence also has a base-shift error and harbors six suspicious transversions. Finally, the haplogroup N9a sequence (DQ834258) has a problem with artefactual recombination. Detailed annotations for these Vietnamese mitochondrial genomes and a few more GenBank complete mtDNA sequences with similar problems are listed in the Supplemental Data.

It is likely that most conclusions in the Pereira et al. study[11] would essentially remain unaltered after the flawed data sets or single problematic sequences were filtered out. Nonetheless, the results reported in their tables would benefit from a reanalysis using an improved version of the complete genome database. It depends on the particular aspect under study as to whether a small residue of errors would matter or not. A good example of where it would cause problems is with the estimation of the transition:transversion ratio, because transversions are relatively rare and flawed data are often enriched in transversions (see phantom mutations in the Supplemental Data). The number of artefactual transversions from some of the data sets does appear to be raised, in particular in the sequences from Gasparre et al.[17] (Table 1 and Supplemental Data). In addition, misalignment of seven sequences (DQ341085–DQ341090 and EU600343) in the Pereira et al. study[11] has produced at least another 21 artefactual transversions at positions 292, 296–299, 300, 302, and 303. Similarly, the insertion 5436insG in DQ246818 has been shifted by four base pairs and scored as C5437G 5440insC, so that a transversion is created artificially. Suboptimal alignment induced further artificial transversions: e.g., the two sequences AY922293 and AY922275 are identical in the 54–60 region (GTTATT versus GTATTTT in the rCRS) and yet the former was interpreted as 55insT-59delTT

Table 1. List of Some Flawed Data and Uncorrected Sequences Employed in the Study by Pereira Et Al.[11]

GenBank Data	Cause of Error	Reference	Errors Detected or Corrected
DQ156212, DQ156214	NUMT contamination	Montiel-Sosa et al.[27]	Yao et al.[28]
DQ112878	NUMT contamination	Kivisild et al.[29]	Yao et al.[28]
DQ112952	Missed mutation	Kivisild et al.[29]	this study
DQ341068.1	Artefactual recombination	Torroni et al.[30]	Behar et al.;[15] DQ341068.2 (updated May 5, 2009)
AP008259, AP008269, AP008278, AP008306, AP008552, AP008776, AP008777, AP008798, AP008799, AP008801, AP008803	Artefactual recombination	Tanaka et al.[23]	Kong et al.[21]
Various	Missed mutations	Maca-Meyer et al.[31]	Palanichamy et al.[32]
Various	Phantom mutations and documentation errors	Herrnstadt et al.[12]	Herrnstadt et al.;[13] Bandelt et al.[19]
Various	Missed mutations	Rajkumar et al.[33]	Sun et al.[34]
Various	Various	Gonder et al.[14]	Behar et al.;[15] this study
AY963586.1	Editing error in GenBank submission	Bandelt et al.[4]	AY963586.3 (updated June 29, 2009)
EF660912–EF661013	Phantom mutations and missed mutations	Gasparre et al.[17]	This study
AM260596–AM260597	Missed mutations	Annunen-Rasila et al.[35]	This study
AY289073	Missed mutations and recombination	Ingman and Gyllensten[36]	This study
AY195745, AY195756, AY195767, AY195775	Phantom mutations and missed mutations	Mishmar et al.[37]	Brandstätter et al.;[38] this study
EU095205, EU095208, EU095250	Phantom mutations and missed mutations	Fagundes et al.[39]	Perego et al.;[40] this study
AY339437, AY339463.2, AY339546, AY339549,AY339581.2, AY339582	Phantom mutations and missed mutations	Finnilä et al.[41]	This study
AF46968, AF346973, AF347006	Missed mutations, phantom mutations, and recombination	Ingman et al.[42]	Kong et al.;[21] this study
Various	Phantom indels and missed mutations	Kumar et al.[43]	This study
EU597580	Missed mutation	Hartmann et al.[44]	This study
DQ826448, DQ834253-DQ834261	Various	Phan et al. (unpubl. data)[a]	This study
DQ418488, DQ437577, DQ462232–DQ462234, DQ519035	Various	The State Key Laboratory of Forensic Sciences (unpubl. data) [a]	This study
DQ358973–DQ358977	Documentation errors (position 750)	Detjen et al. (unpubl. data)[a]	This study
EF446784, EF488201	Poor sequencing quality (artefactual heteroplasmy)	Noer et al. (unpubl. data)[a]	This study

[a] Unpublished data were released by GenBank, and detailed annotation of the potential errors is given in the Supplemental Data.

and the latter as 56T-57A-60delT in that region by Pereira et al.[11] Inconsistent alignment is also seen in the long C stretch in regions 16184–16193 and 303–315 in the Pereira et al. study.[11]

Another instance in which a small amount of error could have a significant influence involves the estimation of the positional rate spectrum along the molecule. For instance, the change C12705T (characteristic of non-R status) is a rare mutation but was overlooked by Gonder et al.[14] half a dozen of times, and the mutation T10810C (characteristic of non-L2′6 status) was overlooked an additional eight times.[14] Similarly, the estimated rate of any mutation scored between the roots of frequent haplogroups in the mtDNA phylogeny gets inflated by the use of incomplete or recombinant sequences. Thus, the incorporation of flawed data considerably distorts the estimation of rates for a number of positions. The same effect may be caused by systematic documentation errors, as in the case of the 14766 transition, which has often been misrecorded because of the discrepancy at 14766 between rCRS and a

partly corrected CRS (which was in use for a long time).[3,10] Moreover, for parts of the mtDNA phylogeny in which numerous mutations are missed in the data used, estimation of haplogroup coalescent times becomes distorted. The consequences of using wrong data can be dramatic under particular circumstances, as we have discussed before.[3–10,18–21] Fortunately, the standard and quality of sequencing from the large laboratories (with long-standing experience) has improved over the years, and the results from these laboratories are now setting the standard against which all smaller institutions should compare themselves. This does not preclude the possibility that single sequences from data sets released by large laboratories may have minor problems.

Bioinformatics-based projects are more and more popular, drawing conclusions from whatever can be retrieved from GenBank (e.g., Gonder et al.'s data[14] were also employed by Atkinson et al.[22]). However, the common practice of mining mtDNA data from GenBank or other genomic resources should be carried out with the necessary caution in order to avoid erroneous claims in future studies. For instance, one could foresee that the use of the original incorrect sequences by Tanaka et al.[23] would easily lead to erroneous signals of mtDNA recombination.[21] To eliminate errors in the published mtDNA data or at least to exclude the suspicious GenBank entries from any subsequent reanalyses, we call for a stringent scrutiny of reported data and a bookkeeping annotation of errors in the public databases, such as in Phylotree.org (maintained by Mannis van Oven)[2] and some personally owned websites (e.g. Ian Logan's website). For the benefit of science, submissions to GenBank should be revised as promptly as possible by the authors responsible for the data in question. And, importantly, when submitting a new paper for publication, authors should provide evidence that their data has been checked for the more common errors that come from poor sequencing technique and data handling, as well as for discrepancies between the actual submissions to GenBank and what has been shown or inferred in the paper. But instances will remain in which authors either do not react or claim that they did everything right (as in the prominent case analyzed by Bandelt and Kivisild[24] and Parson[25]). Therefore, when one plans to perform a cumulative reanalysis of mtDNA data, one cannot avoid making a substantiated, though partly subjective, decision as to which data are to be included and which are to be excluded, as has been exemplified in a recent paper by Soares et al.[26]

Yong-Gang Yao,[1] Antonio Salas,[2] Ian Logan,[3] and Hans-Jürgen Bandelt[4,*]
[1]Key Laboratory of Animal Models and Human Disease Mechanisms of Chinese Academy of Sciences & Yunnan Province, Kunming Institute of Zoology, Kunming 650223, China; [2]Unidade de Xenética, Instituto de Medicina Legal and Departamento de Anatomía Patolóxica e Ciencias Forenses, Facultade de Medicina, Universidade de Santiago de Compostela, Galicia 15782, Spain; [3]Exmouth, Devon, UK; [4]Department of Mathematics, University of Hamburg, 20146 Hamburg, Germany
*Correspondence: bandelt@math.uni-hamburg.de

Supplemental Data

Supplemental Data include one appendix and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

This work was supported by Yunnan Province (云南省高端人才计划) and the Chinese Academy of Sciences (百人计划), as well as from grants from National Natural Science Foundation of China (30925021), the Ministerio de Ciencia e Innovación (SAF2008-02971), and Fundación de Investigación Médica Mutua Madrileña (2008/CL444). We thank two anonymous reviewers for their helpful comments on the early version of the manuscript.

Web Resources

The URLs for data presented herein are as follows:

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/
Ian Logan's website, http://www.ianlogan.co.uk
PhyloTree.org, http://www.phylotree.org/

References

1. Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., et al. (1981). Sequence and organization of the human mitochondrial genome. Nature *290*, 457–465.
2. van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. *30*, E386–E394.
3. Bandelt, H.-J., Yao, Y.-G., Bravi, C.M., Salas, A., and Kivisild, T. (2009). Median network analysis of defectively sequenced entire mitochondrial genomes from early and contemporary disease studies. J. Hum. Genet. *54*, 174–181.
4. Bandelt, H.-J., Achilli, A., Kong, Q.-P., Salas, A., Lutz-Bonengel, S., Sun, C., Zhang, Y.-P., Torroni, A., and Yao, Y.-G. (2005). Low "penetrance" of phylogenetic knowledge in mitochondrial disease studies. Biochem. Biophys. Res. Commun. *333*, 122–130.
5. Bandelt, H.-J., Olivieri, A., Bravi, C., Yao, Y.-G., Torroni, A., and Salas, A. (2007). 'Distorted' mitochondrial DNA sequences in schizophrenic patients. Eur. J. Hum. Genet. *15*, 400–402.
6. Bandelt, H.-J., Yao, Y.-G., Salas, A., Kivisild, T., and Bravi, C.M. (2007). High penetrance of sequencing errors and interpretative shortcomings in mtDNA sequence analysis of LHON patients. Biochem. Biophys. Res. Commun. *352*, 283–291.
7. Salas, A., Carracedo, Á., Macaulay, V., Richards, M., and Bandelt, H.-J. (2005). A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem. Biophys. Res. Commun. *335*, 891–899.
8. Salas, A., Yao, Y.-G., Macaulay, V., Vega, A., Carracedo, Á., and Bandelt, H.-J. (2005). A critical reassessment of the role of mitochondria in tumorigenesis. PLoS Med. *2*, e296.

9. Yao, Y.-G., Macaulay, V., Kivisild, T., Zhang, Y.-P., and Bandelt, H.-J. (2003). To trust or not to trust an idiosyncratic mitochondrial data set. Am. J. Hum. Genet. *72*, 1341–1346.
10. Yao, Y.-G., Salas, A., Bravi, C.M., and Bandelt, H.-J. (2006). A reappraisal of complete mtDNA variation in East Asian families with hearing impairment. Hum. Genet. *119*, 505–515.
11. Pereira, L., Freitas, F., Fernandes, V., Pereira, J.B., Costa, M.D., Costa, S., Máximo, V., Macaulay, V., Rocha, R., and Samuels, D.C. (2009). The diversity present in 5140 human mitochondrial genomes. Am. J. Hum. Genet. *84*, 628–640.
12. Herrnstadt, C., Elson, J.L., Fahy, E., Preston, G., Turnbull, D.M., Anderson, C., Ghosh, S.S., Olefsky, J.M., Beal, M.F., Davis, R.E., et al. (2002). Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am. J. Hum. Genet. *70*, 1152–1171.
13. Herrnstadt, C., Preston, G., and Howell, N. (2003). Errors, phantoms and otherwise, in human mtDNA sequences. Am. J. Hum. Genet. *72*, 1585–1586.
14. Gonder, M.K., Mortensen, H.M., Reed, F.A., de Sousa, A., and Tishkoff, S.A. (2007). Whole-mtDNA genome sequence analysis of ancient African lineages. Mol. Biol. Evol. *24*, 757–768.
15. Behar, D.M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., Comas, D., et al. (2008). The dawn of human matrilineal diversity. Am. J. Hum. Genet. *82*, 1130–1140.
16. Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. *23*, 147.
17. Gasparre, G., Porcelli, A.M., Bonora, E., Pennisi, L.F., Toller, M., Iommarini, L., Ghelli, A., Moretti, M., Betts, C.M., Martinelli, G.N., et al. (2007). Disruptive mitochondrial DNA mutations in complex I subunits are markers of oncocytic phenotype in thyroid tumors. Proc. Natl. Acad. Sci. USA *104*, 9001–9006.
18. Bandelt, H.-J., Kong, Q.-P., Parson, W., and Salas, A. (2005). More evidence for non-maternal inheritance of mitochondrial DNA? J. Med. Genet. *42*, 957–960.
19. Bandelt, H.-J., Kong, Q.-P., Richards, M., and Macaulay, V. (2006). Estimation of mutation rates and coalescence times: some caveats. In Human Mitochondrial DNA and the Evolution of Homo sapiens, H.-J. Bandelt, V. Macaulay, and M. Richards, eds. (Berlin, Heidelberg: Springer-Verlag), pp. 47–90.
20. Bandelt, H.-J., and Salas, A. (2009). Contamination and sample mix-up can best explain some patterns of mtDNA instabilities in buccal cells and oral squamous cell carcinoma. BMC Cancer *9*, 113.
21. Kong, Q.-P., Salas, A., Sun, C., Fuku, N., Tanaka, M., Zhong, L., Wang, C.-Y., Yao, Y.-G., and Bandelt, H.-J. (2008). Distilling artificial recombinants from large sets of complete mtDNA genomes. PLoS ONE *3*, e3016.
22. Atkinson, Q.D., Gray, R.D., and Drummond, A.J. (2008). mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. Mol. Biol. Evol. *25*, 468–474.
23. Tanaka, M., Cabrera, V.M., González, A.M., Larruga, J.M., Takeyasu, T., Fuku, N., Guo, L.J., Hirose, R., Fujita, Y., Kurata, M., et al. (2004). Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res. *14*, 1832–1850.
24. Bandelt, H.-J., and Kivisild, T. (2006). Quality assessment of DNA sequence data: autopsy of a mis-sequenced mtDNA population sample. Ann. Hum. Genet. *70*, 314–326.
25. Parson, W. (2007). The art of reading sequence electropherograms. Ann. Hum. Genet. *71*, 276–278.
26. Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., Salas, A., Oppenheimer, S., Macaulay, V., and Richards, M.B. (2009). Correcting for purifying selection: an improved human mitochondrial molecular clock. Am. J. Hum. Genet. *84*, 740–759.
27. Montiel-Sosa, F., Ruiz-Pesini, E., Enríquez, J.A., Marcuello, A., Díez-Sánchez, C., Montoya, J., Wallace, D.C., and López-Pérez, M.J. (2006). Differences of sperm motility in mitochondrial DNA haplogroup U sublineages. Gene *368*, 21–27.
28. Yao, Y.-G., Kong, Q.-P., Salas, A., and Bandelt, H.-J. (2008). Pseudomitochondrial genome haunts disease studies. J. Med. Genet. *45*, 769–772.
29. Kivisild, T., Shen, P., Wall, D.P., Do, B., Sung, R., Davis, K., Passarino, G., Underhill, P.A., Scharfe, C., Torroni, A., et al. (2006). The role of selection in the evolution of human mitochondrial genomes. Genetics *172*, 373–387.
30. Torroni, A., Achilli, A., Macaulay, V., Richards, M., and Bandelt, H.-J. (2006). Harvesting the fruit of the human mtDNA tree. Trends Genet. *22*, 339–345.
31. Maca-Meyer, N., González, A.M., Larruga, J.M., Flores, C., and Cabrera, V.M. (2001). Major genomic mitochondrial lineages delineate early human expansions. BMC Genet. *2*, 13.
32. Palanichamy, M.G., Sun, C., Agrawal, S., Bandelt, H.-J., Kong, Q.-P., Khan, F., Wang, C.-Y., Chaudhuri, T.K., Palla, V., and Zhang, Y.-P. (2004). Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am. J. Hum. Genet. *75*, 966–978.
33. Rajkumar, R., Banerjee, J., Gunturi, H.B., Trivedi, R., and Kashyap, V.K. (2005). Phylogeny and antiquity of M macrohaplogroup inferred from complete mt DNA sequence of Indian specific lineages. BMC Evol. Biol. *5*, 26.
34. Sun, C., Kong, Q.-P., Palanichamy, M.G., Agrawal, S., Bandelt, H.-J., Yao, Y.-G., Khan, F., Zhu, C.-L., Chaudhuri, T.K., and Zhang, Y.-P. (2006). The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. Mol. Biol. Evol. *23*, 683–690.
35. Annunen-Rasila, J., Finnilä, S., Mykkänen, K., Pöyhönen, J.S., Veijola, J., Poyhonen, M., Viitanen, M., Kalimo, H., and Majamaa, K. (2006). Mitochondrial DNA sequence variation and mutation rate in patients with CADASIL. Neurogenetics *7*, 185–194.
36. Ingman, M., and Gyllensten, U. (2003). Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. Genome Res. *13*, 1600–1606.
37. Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., et al. (2003). Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. USA *100*, 171–176.
38. Brandstätter, A., Sänger, T., Lutz-Bonengel, S., Parson, W., Béraud-Colomb, E., Wen, B., Kong, Q.-P., Bravi, C.M., and Bandelt, H.-J. (2005). Phantom mutation hotspots in human mitochondrial DNA. Electrophoresis *26*, 3414–3429.
39. Fagundes, N.J., Kanitz, R., Eckert, R., Valls, A.C., Bogo, M.R., Salzano, F.M., Smith, D.G., Silva, W.A., Jr., Zago, M.A., Ribeiro-dos-Santos, A.K., et al. (2008). Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. Am. J. Hum. Genet. *82*, 583–592.
40. Perego, U.A., Achilli, A., Angerhofer, N., Accetturo, M., Pala, M., Olivieri, A., Kashani, B.H., Ritchie, K.H., Scozzari, R., Kong, Q.-P.,

et al. (2009). Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr. Biol. *19*, 1–8.
41. Finnilä, S., Lehtonen, M.S., and Majamaa, K. (2001). Phylogenetic network for European mtDNA. Am. J. Hum. Genet. *68*, 1475–1484.
42. Ingman, M., Kaessmann, H., Pääbo, S., and Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. Nature *408*, 708–713.
43. Kumar, S., Padmanabham, P.B., Ravuri, R.R., Uttaravalli, K., Koneru, P., Mukherjee, P.A., Das, B., Kotal, M., Xaviour, D., Saheb, S.Y., et al. (2008). The earliest settlers' antiquity and evolutionary history of Indian populations: evidence from M2 mtDNA lineage. BMC Evol. Biol. *8*, 230.
44. Hartmann, A., Thieme, M., Nanduri, L.K., Stempfl, T., Moehle, C., Kivisild, T., and Oefner, P.J. (2009). Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. Hum. Mutat. *30*, 115–122.

DOI 10.1016/j.ajhg.2009.10.023.

Response to Yao et al.

To the Editor: We are also concerned about errors in GenBank sequences, and that is why we took precautions to evaluate the effects of potential sequence errors.[1] But many of the potential errors reported by Yao et al. are highly subjective. They defined "phantom mutations" as (with exceptions) the exclusive presence of rare transversions in a specific data set. Although it is reasonable to be skeptical of such variations, surely such rare variations do actually occur without being errors. To deal with potential sequence errors, we took the step of doing the analysis twice; once for all reported variations and once for only variations present in more than 0.1% of the sequences. We made the latter choice to filter out sequencing errors, assuming that specific errors would not be repeated in many different sequences. This filtering process did remove 94% of their listed "phantom mutations." As Yao et al. acknowledge, the removal of these rare variations (some of which may be sequencing errors) had little effect on most of our results.

Yao et al. define "missing variants" as those variants expected to be seen in a particular haplogroup but not reported in a sequence assigned to it. The problem with this definition is that it presupposes that we already have a complete picture of mtDNA variation and that all deviations from it are errors. There are many examples of such "missing variants" being true variations. It was once thought that all L- sub-Saharan haplogroups had the substitution at position 16223, but later some lineages were characterized without it (L0d1a, L1c1a1, L2d, L3x2a). Also, the M1- defining substitution at position 16249 is absent in the branch M1a1a.

After the careful data mining of Yao et al., potential errors were found in < 200 of the 5140 sequences. So, ~96% of the sequences deposited in GenBank by the end of August 2008 did pass their extreme quality filter. Yao et al. list many cases in which errors in the original sequences have been acknowledged and corrected by authors but the GenBank sequence has not been updated. GenBank[2] allows the sequence depositor to update that sequence, but it depends on each depositor to carry out this procedure. Identifying these possible sequence errors is complex and is arguably highly subjective. To expect every author of a sequence data-mining project to carry out such a very subjective quality-control step is not reasonable, in our opinion.

Though we may disagree on specifics raised by Yao et al., we do share with them a concern about mtDNA sequence quality. Spirited discussions such as this one have been going on for the past decade. It is time to provide the mtDNA research community with analysis tools that allow them to efficiently check their sequences for potential problems, such as sequencing errors or unusual variations. We tried to go forward in this direction with our paper[1] by providing the mtDNA Gene-Syn software. Fortunately, others are also advancing along the same path.[3–5]

Luísa Pereira[1,2] and David C. Samuels[3]
[1]Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto 4200-465 Porto, Portugal; [2]Faculdade de Medicina da Universidade do Porto, 4200-465 Porto, Portugal; [3]Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN 37232, USA

References

1. Pereira, L., Freitas, F., Fernandes, V., Pereira, J.B., Costa, M.D., Costa, S., Maximo, V., Macaulay, V., Rocha, R., and Samuels, D.C. (2009). The Diversity Present in 5140 Human Mitochondrial Genomes. Am. J. Hum. Genet. *84*, 628–640.
2. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2009). GenBank. Nucleic Acids Res. *37*, D26–D31.
3. Brandon, M.C., Ruiz-Pesini, E., Mishmar, D., Procaccio, V., Lott, M.T., Nguyen, K.C., Spolim, S., Patil, U., Baldi, P., and Wallace, D.C. (2009). MITOMASTER: A Bioinformatics Tool for the Analysis of Mitochondrial DNA Sequences. Hum. Mutat. *30*, 1–6.
4. van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. *30*, E386–E394.
5. Lee, H.Y., Song, I., Ha, E., Cho, S.B., Yang, W.I., and Shin, K.J. (2008). mtDNAmanager: a Web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences. BMC Bioinformatics *9*, 483.

DOI 10.1016/j.ajhg.2009.10.022.

Supplemental Data
AJHG*, Volume *85

mtDNA Data Mining in GenBank Needs Surveying

Yong-Gang Yao, Antonio Salas, Ian Logan, and Hans-Jürgen Bandelt

Appendix S1. Annotations for Some Problematic mtDNA Genomes Released in GenBank

We have listed all the sequence variations in each mtDNA relative to the revised Cambridge reference sequence (rCRS),[1] but any 'C3107N' has been ignored since this refers to an inappropriate comparison with the rCRS [1] promoted by Mitomap (http://www.mitomap.org). We followed the usual convention in forensics to score the insertions and deletions (indels) in the sequence and recorded the indels at the last possible site. Each mtDNA was classified according to the available worldwide mtDNA phylogeny [2-6]. 'Missing variants' are those that are expected in a particular mtDNA haplotype according to its haplogroup status; the corresponding haplogroup or its complement (signified by the 'non-' prefix) that is supported by the mutation is listed in parentheses. Variants highlighted in italics are known to be highly recurrent and refer to the 67 hotspots listed by Soares et al. [7]. Phantom mutations are defined by the exclusive presence of the rare transversions in a specific data set, unless otherwise indicated. It is to be understood that any *a posteriori* evaluation can never be 100% correct and would be best confirmed by the original authors.

1. Sequences EF184580-EF184641

Reference: Gonder, M.K., Mortensen, H.M., Reed, F.A., de Sousa, A., and Tishkoff, S.A. (2007). Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol 24, 757-768.
Submitted to GenBank: 13-DEC-2006
Released in GenBank: 18-APR-2007
General comment: Sequences contain various kinds of error; the most salient one is the very high number of missing variants.
Fourteen sequences are annotated in detail as follows (where only some of the potential phantom transversions are highlighted):

EF184585 Haplogroup L0d3
A73G T146C C150T T152C T195C C198T T236C G247A 315insC G316A 537delC 551delA A567C A636T T721C A750G G769A T825A G1018A C1048T T1243C T2586G A2706G A2755G G2758A T2885C C3516A C3594T A4104G G4113A T4232C C4312T A4769G G4812A T5442C G5460A G5773A T6185C C6377T T6815C C7028T A7146G C7256T G7521A T8013G C8113A G8152A G8251A A8459G C8468T T8598C C8655T A8701G A8860G C9027T C9042T A9347G C9488T T9540C A10398G G10589A C10664T G10688A T10810C T10873C C11061T G11390C A11653G G11719A G11914A T12121C C12390T C12705T A12720G A13105G A13276G G13359A C13506T C13650T C13932T C14766T A15236G T15312C A15326G T15461C G15466A G15930A T15941C G16129A

C16187T T16189C C16223T A16230G T16243C C16278T C16290T A16300G T16311C T16362C A16399G T16505G T16506G C16507G T16519C
Comments: missing variants T10915C (L0), G12007A (non-L1'5); phantom mutations A567C, T16505G, T16506G, C16507G

EF184589 Haplogroup L0d3
A73G T146C C150T T195C C198T G247A 315insC G316A C679T T721C A750G G769A T825A G1018A C1048T T1243C A2706G A2755G G2758A T2885C C3516A C3594T A4104G G4113A T4232C C4312T A4769G G4812A T5442C G5460A G5773A T6185C C6377T T6815C C7028T A7146G C7256T G7521A C8113A G8152A G8251A A8459G C8468T T8598C C8655T A8701G A8860G C9027T C9042T A9347G C9488T T9540C A10398G G10589A C10664T G10688A C10727G T10873C T10915C C11061T A11653G G11719A G11914A G12007A T12121C A12172G C12390T C12705T A12720G A13105G A13276G G13359A C13487T C13506T C13650T C13932T C14766T A15236G T15312C A15326G T15461C G15466A T15586C G15930A T15941C G16129A C16148T C16187T T16189C C16223T A16230G T16243C C16278T C16290T A16300G T16311C A16399G T16519C
Comments: missing variants *C152T*, T10810C (non-L2'6); phantom mutation C10727G

EF184595 Haplogroup L0f
G143A T146C G185A A189G T236C G247A A263G 309insC 315insC C320T 523-524delAC A750G G769A T825A G1018A C1048T A1438G T1822C A2245G C2484G A2706G G2758A A2879G T2885C T3027C C3516A C3594T A4012G A4104G C4312T T4586C A4769G C4964T G5147A T5442C C5603T T6185C C7028T A7146G T7148C C7256T C7336G C7341G A7343G G7521A T7660C T8227C C8468T C8655T A8701G A8860G C9042T A9347G T9540C T9581C C9620T C9818T A10398G G10586A G10589A C10664T T10873C A11641G C11782T G11914A C12092A C12705T A12720G A12961G A13105G A13470G C14109T G14305A C14482T C14620T C14766T C15136T A15326G G15431A A15607G T15672C C15857A C16052T G16129A C16169T C16187T T16189C C16223T A16230G C16278T C16290T T16311C A16316G T16325C C16327T C16354T T16368C T16519C
Comments: missing variants *T152C*, *G207A*, G10688A (non-L2'6), T10810C (non-L2'6), T10915C (L0), G11719A (non-R0), G12007A (non-L1'5), G13145A (L0f), A13276G (L0), C13506T (non-L2'6), C13650T (non-L3'4), T15852C (L0f), *T16172C*; phantom mutations C7336G, C7341G, and possibly the transition A7343G as well

EF184596 Haplogroup L0f2a
A750G G769A T825A G1018A C1048T A1438G G1719A A2245G A2706G G2758A T2885C C3516A T3552C C3594T A4104G C4194T C4312T A4562G T4586C G4655A A4769G C4964T T5442C C5603T T6185C T6227C C7028T A7146G T7148C C7256T A7361G G7521A C8346. C8468T C8655T A8701G A8860G C9042T T9078C A9260G A9347G T9540C G9554A T9581C C9620T C9818T T9949G G9962A T9978G T10361C A10398G A10532G G10589A C10664T G10688A T10873C T11287C A11432G T11617C A11641G G11719A C11896A G12007A A12720G A13105G A13276G A13422G C13464A A13470G G13928C

C13938T C14109T C14620T C14766T C15136T G15221A A15326G G15431A T15852C G16129A C16169T T16172C C16187T T16189C C16223T A16230G C16278T T16311C T16325C C16327T C16354T T16368C
Comments: This sequence is incomplete, obviously lacking the entire variation in HVS-II&III; missing variants T10810C (non-L2'6), T10915C (L0), *G11914A* (non-L1'5), C12705T (non-R), C13506T (non-L2'6), C13650T (non-L4'6), C13680T (L0f2), *T16519C*; phantom mutations T9949G, T9978G

EF184598 Haplogroup L0f
A73G T152C G185A A189G G207A G247A A263G 315insC T391C 523-524delAC A750G T825A G1018A C1048T A1438G A2245G G2700C A2706G G2758A T2885C C3417T C3510A C3516A C3594T A4104G C4312T T4586C T4695C A4769G C4964T A5276G T5442C C5603T A5605G T6152C C6164T T6185C A6923G C7028T A7146G T7148C C7256T A7364G G7521A C8468T C8655T A8701G A8860G C9042T G9055A A9347G T9540C T9581C C9620T C9818T A10055G A10398G G10589A G10688A T10810C T10873C T10909C T10915C A11641G G11719A G11914A G12007A C12705Y A12720G T12903C A13105G A13276G A13470G C13506T C13650T C14109T A14409R C15136T A15326G G15431A T15852C C16169T T16172C C16187T T16189C C16223T A16230G A16265C C16270T T16311C C16327T T16368C 16460delC C16465T T16519C
Comments: This sequence shares 17 variants (underlined) with EF184600 (see below) beyond a partial L0f motif and thus defines a novel deepest branch ("L0f3") of L0f which does not share G13145A and C16354T with L0f1'2 (by descent or oversight). Moreover, it seems that this branch additionally underwent further mutations at positions 146 and 16129; missing variants G769A (non-L3), C10664T (L0), C14620T (L0f), C14766T (non-HV), *C16278T*; the ambiguity C12705Y might point to a reading problem at this position since the variant C12705T has frequently been overlooked in this data set

EF184600 Haplogroup L0f
A73G T152C G185A A189G G207A G247A A263G 315insC T391C 523-524delAC A750G G769A T825A G1018A C1048T A1438G A2245G A2706G G2758A T2885C C3417T C3510A C3516A C3594T A4104G C4312T T4586C T4695C A4769G C4964T A5276G T5442C C5603T A5605G T6152C C6164T T6185C A6923G C7028T A7146G T7148C C7256T G7521A C8468T C8655T A8701G A8860G C9042T G9055A A9347G T9540C T9581C C9620T C9818T T9978G A10055G A10398G G10589A C10664T G10688A T10810C T10873C T10909C T10915C A11641G G11719A C12705T A12720G T12903C A13105G A13276G A13470G C13506T C13650T C14620T C14766T C15136T A15326G G15431A T15852C C16169T T16172C C16187T T16189C C16223T A16230G A16265C C16270T C16278T T16311C C16327T T16368C C16465T T16519C
Comments: This sequence is closely related to the preceding EF184598 (see comments there); missing variants *G11914A* (non-L1'5), G12007A (non-L1'5), C14109T (L0f); phantom mutation T9978G

EF184603 Haplogroup L0a2a1
C64T A93G T152C A189G T236C G247A A263G 315insC 523-524delAC A750G G769A T825A G1018A C1048T A1438G A2245G A2706G G2758A T2885C C3516A C3594T A4104G C4312T T4586C T4598C A4769G G5147A G5231A T5442C G5460A C5603T A5711G T6185C G6257A C7028T A7146G C7256T

A7424G G7521A 8281-8289del C8428T A8460G C8468T A8566G C8655T A8701G A8860G C9042T A9347G T9540C G9755A C9818T T9978G A10398G G10589A C10664T G10688A T10810C T10873C T10915C C11143T A11172G G11176A A11641G G11719A G11914A G12007A C12705T A12720G A13105G A13276G C13506T C13650T T14182C T14308C A14755G C14766T C15136T A15326G G15431A T16093C C16148T T16172C C16187T C16188G T16189C C16223T A16230G T16311C C16320T T16519C
Comments: missing variant *T146C*; phantom mutation T9978G

EF184607 Haplogroup L0a2
A73G A93G T146C T152C A189G T236C G247A A263G 523-524delAC 526insG A538C A750G G769A T825A G1018A C1048T A1438G A2245G A2706G G2758A T2885C C3516A C3594T G3882A A4104G C4312T T4452C T4586C A4769G A4917G G5147A G5231A T5442C G5460A C5603T A5711G T6185C G6257A C7028T A7146G C7256T G7521A 8281-8289del C8428T A8460G C8468T A8566G C8655T A8701G A8860G C9042T A9347G T9540C G9755A C9818T T9963G A10398G G10589A C10664T G10688A T10873C A11172G G11176A A11641G G11719A G11914A G12007A C12705T A12720G A13105G A13276G C13506T A13582G C13650T T14308C C14766T C15136T A15326G G15431A C16148T T16172C C16187T C16188A T16189C A16212G C16223T A16230G T16311C C16320T
Comments: missing variants *C64T*, 315insC, T10810C (non-L2'6), T10915C (L0), *T16519C*; phantom mutation T9963G

EF184609 Haplogroup L0k1
A73G T146C T152C A189G T195C C198T G207A G247A 315insC 523-524delAC A750G G769A T825A T850C G1018A C1048T T1243C A1438G A2706G G2758A C2836A T2885C T3338A C3516A C3594A T3653A A4104G C4312T G4541A A4769G T4907C T5442C T6185C C6938T C7028T G7457C G7521A C8468T C8655T A8701G A8860G T8911C G8994A C9042T A9136G A9347G T9540C C9818T A10398G A10499G G10589A C10664T G10688A A10765T A11653G G11719A G11914A G12007A T13020C A13105G A13276G C13420T T14371C T14374C A16166C C16187T T16189C T16209C C16214T C16223T A16230G C16278T C16291G T16311C T16519C
Comments: This sequence shares with EF184610 four mutations (underlined) beyond the common L0k1 motif; missing variants T4586C (L0abfk), A5811G (L0k), A7146G (non-L2'5), C7256T (non-L3'4), A7257G (L0k), T10810C (non-L2'6), T10873C (non-N), A10876G (L0k), T10915C (L0), C10920T (L0k), C10939T (L0k1), C11296T (L0k), T11299C (L0k), C12705T (non-R), A12720G (L0), C13506T (non-L2'6), G13590A (L0k),C13650T (non-L3'4), T13819C (L0k), G13928C (L0k), T14020C (L0k), T14182C (L0k), C14766T (non-HV), A15326G (non-H2a2a), *T16172C*; phantom mutation A10765T (eliminating a stop codon and leading to a reading-frame shift) [8]; documentation error: C3594A instead of the expected C3594T

EF184610 Haplogroup L0k1
A73G T146C T152C A189G T195C C198T G207A G247A 309insC 315insC 523-524delAC A750G T825A T850C G1018A C1048T T1243C A1438G A2363G A2706G G2758A C2836A T2885C C3516A C3594T A4104G C4312T G4541A T4586C A4769G T4907C A5811G T6185C C6938T C7028T A7146G C7256T

A7257G G7521A C8468T C8655T A8701G A8860G T8911C G8994A C9042T A9136G A9347G T9540C C9818T A10398G A10499G G10589A C10664T G10688A C10727G T10810C T10873C A10876G T10915C C10920T C10939T C11296T T11299C A11653G G11719A G11914A T11988C G12007A G12070A T13020C A13105G A13276G C13506T G13590A C13650T T13819C G13928C T14182C T14371C T14374C G14569A C14766T A15326G 15788-15792del A16166C T16172C C16187T T16189C T16209C C16223T A16230G C16278T C16291G T16311C T16519C
Comments: This sequence is related to the preceding sequence (see comments above); missing variants G769A (non-L3), T5442C (L0), C10939T (L0k1), C12705T (non-R), T14020C (L0k), *C16214T*; phantom mutation C10727G

EF184625 Haplogroup L3a
A73G T152C A263G 309insC 315insC T721C A750G A1438G 2156insA G2702A A2706G A3796G C4088T T4733C A4769G C6500T C7028T A8701G A8860G T9540C T9951K T9963K T9967K C10314T A10398G T10873C G11719A C12705T C12816T T14461C C14766T A14851G A14927G T15109C G15301A A15326G G15553A C16223T A16254G A16293G A16316G
Comments: phantom mutations T9951K, T9963K, T9967K; this sequence was misclassified as L3f by Pereira et al. [8]

EF184626 Haplogroup M1a1
A73G A183G T195C A263G 315insC T489C A750G A813G A1438G A2706G G3705A A4769G G6446A T6671C T6680C T6944C C7028T A8701G A8860G T9540C A10398G C10400T G11719A C12346T C12705T A12950C T14110C C14766T T14783C G15043A G15301A A15326G G16129A A16182C A16183C T16189C C16223T T16249C T16311C T16359C T16519C
Comment: missing variants T10873C (non-N), C12403T (M1)

EF184635 Haplogroup M1a1
A73G T195C A263G 309insCC A750G A813G A1438G A2706G G3705A A4769G G6446A T6671C T6680C C7028T A8701G A8860G G9055A T9540C A10398G C10400T A10841G T10873C T10881G G11719A C12346T C12403T G12561A C12705T A12950C C14766T T14783C G15043A G15301A A15326G T16189C C16223T T16249C T16311C T16359C T16519C
Comment: missing variants 315insC, T489C (M), T14110C (M1), *G16129A*

EF184637 Haplogroup M1a5
A73G T195C A263G 315insC A750G A813G A1438G A2706G A2963G A4769G G6446A T6671C T6680C C7028T A8701G A8860G G9379C T9540C A10398G C10400T T10873C G11719A C12403T C12705T A12950C T13215C A13722G G14323A T14515C C14766T T14783C G15043A G15301A A15326G C15770A A15799G T16189C C16223T T16249C T16311C T16519C
Comment: missing variants T489C (M), T14110C (M1)

2. Sequences DQ826448 and DQ834253-DQ834261
Reference: Phan,V.C., Nong,V.H., Nguyen,B.N., Tran,T.M.N., Le,T.B.T., Do,Q.H., Nguyen,N.L., Bui,T.H., Pham,D.M., Tran,T.T., Tong,Q.M., Nguyen,T.T.,

Nguyen,D.T., Le,T.T.H., Nguyen,D.C., Le,Q.H., Dang,D.H., Quyen,D.T., Van,D.H., Trinh,V.B., Le,B.Q. and Nguyen,D.B.
General comments: All sequences lack A263G, 315insC, and C14766T, besides those listed in addition for the specific mtDNAs, which was likely incurred by the use of a wrong reference sequence. A salient feature of this small data set is the tremendous amount of transversions, nearly all of which are phantom mutations; one has to reckon with a certain number of phantom transitions as well, which are however difficult to pinpoint

DQ826448 Haplogroup M7b1
Submitted to GenBank: 28-JUN-2006
Released in GenBank: 10-JUL-2006
A73G C150T T199C T489C A1438G A2706G A3133T A3243G C3311T T3644G G4048A C4071T A4164G T4492G A4769G G5460A C6455T T6680C T6937C G6954A C7028T T7572C A7724T A8701G A8860G T9187A T9824C A10574G T10657A T10873C G11719A C12705T T12811C G15043A G15301A A15326G G16130A C16192T C16223T T16297C
Comments: missing variants A750G (non-H2a2), A5351G (M7b'd), T7684C (M7b'd), G7853A (M7b'd), T9540C (non-N), *A10398G* (non-N), C10400T (M), C12405T (M7b'd), T14783C (M); phantom mutations A3133T, T4492G, T9187A, T10657A, and perhaps T3644G (which was also reported in sequence AY339545 in GenBank); base shift from 16129 to 16130; this sequence was misclassified as haplogroup HV by Pereira et al.[8]

Sequences DQ834253-DQ834261
Submitted to GenBank: 30-JUN-2006
Released in GenBank: 11-JUL-2006
The nine individuals belong to the Kinh, Tay, Muong ethnic group; the corresponding amino acid changes in the cyt*b* gene were published in Vietnam (http://dieuduong.com.vn/images/file/Hoi%20nghi%20khoa%20hoc%20tuoi%20tre/3/313_cyto.pdf). The variation in the mitochondrial *ND2* gene of the sequences DQ834253-DQ834258 were discussed in another article published in Vietnam (http://vst.vista.gov.vn/home/database/Folder.2004-04-19.4917/MagazineName.2004-12-29.3445/2007/2007_00001/MArticle.2008-06-18.0820/view)

DQ834253 Haplogroup B4a1c
A73G T146C C269G 523-524delAC G709A A750G A1438G A2706G A4769G T5465C T6815C C7028T T8093C 8281-8289del A8860G G9123A T10238C G11719A A12904G A15326G T16093C A16182C T16217C A16237G C16261T T16519C
Comments: missing variant *T16189C* (B); phantom mutation C269G

DQ834254 Haplogroup B5a1
A73G A210G T216G 523-524delAC A750G A1438G A1789G A2706G A4769G A6012G A6575G A6659G T6706G T6707A C6805T C6960T C7028T T7546C G8584A A8860G T9950C A10398G T11204C G11719A A15235G A15326G G16129A A16183C T16189C T16249C T16519C
Comments: missing variants *G709A* (B5), A3537G (B5a), *8281-8289del* (B), T16140C (B5), C16266A (B5a); phantom mutations T216G, T6706G, T6707A

DQ834255 Haplogroup N9a
A73G C150T C269G T279G C330G A750G A1438G A2706G A4769G T4856C C7028T G8020A A8860G T11204C G11719A A12358G G12372A C12705T T12811C A15326G C16223T C16261T T16519C
Comments: missing variant G5231A (N9a), G5417A (N9), C16257A (N9a); phantom mutations C269G, T279G, C330G (see Brandstätter et al. [9]); this mtDNA was misclassified as HV by Pereira et al.[8]

DQ834256 Haplogroup M7b1'2
A73G C150T A210G C269G T489C A750G A1438G A2706G G4048A C4071T A4164G A4769G A5351G G5460A C6186T C6187G C6455T T6680C A6914G C7028T G7075A A7403G A7515G T7684C A7844G G7853A C8076A A8192T A8389C A8701G A8860G T9164G A9180T T9540C T9824C A10398G C10400T T10873C G11719A C12405T C12705T T12811C T14783C G15301A A15326G A15644T G16129A T16189C C16223T C16257T T16297C
Comments: missing variants *T199C*, G15043A (M); phantom mutations C269G, C6187G, C8076A, A8192T, A8389C, T9164G, A9180T, A15644T, and the two transitions G7075A and A7403G that were also inflicted upon the next sequence

DQ834257 Haplogroup M21d
A73G T216G T489C G709A A750G A1438G G1598A A1763G A2706G C3819T G3915A A4769G T5108C C6231T C6455T C7028T G7075A A7187G A7403G A7528G T7538G T7861C A7920G G7929A T7999C C8509T A8701G A8860G T9540C A10398G C10400T T10873C T11482C G11719A C12705T T14783C G15043A A15236G G15301A A15326G A16181G C16192T C16218T C16223T C16242T C16291T T16304C T16519C
Comment: Phantom mutations T216G, G7075A, A7403G and T7538G (see preceding sequence)

DQ834258 Haplogroup N9a
A73G C150T A750G A1438G A2706G A4769G T4856C G5231A G5417A C7028T C7945T T7954C A8701G A8860G T9540C G10586A G11719A A12358G G12372A C12705T A15326G T15593C C16223T C16257A C16261T T16519C
Comments: this mtDNA may be a recombinant since it bears A8701G and T9540C characteristic of non-N status; this sequence was misclassified as haplogroup HV by Pereira et al.[8]

DQ834259 Haplogroup R
A73G C150T C269G A750G T1119C A1438G T2442C A2706G T3394C C3435T C3497T C3571T T3944C A4343G A4769G C7028T C7990T A8379G A8425G A8860G T9128C T9179A A9180G A9214G A9531G G11440A G11719A C12882T C13035T C14803T A14818G A15326G G15346A A15408G T15447C A15720G G16060A T16061G G16129A C16179T A16183C T16189C G16274A C16279T
Comments: phantom mutations C269G, T9179A, T16061G; this sequence was misclassified as haplogroup HV by Pereira et al.[8]

DQ834260 Haplogroup M
A73G T152C T199C T204C C269G T489C 523-524delAC A750G A1438G A2706G A3676C A4021G A4769G A4901G G6026T A6494G T6497C C7028T A7289G

T7854C A8379G A8701G A8860G C8970T G9380A T9540C T10256C A10398G C10400T T10873C G11719A C12705T G13708A A13893G G14040A T14783C G15043A A15122G C15147T G15216A G15301A A15326G T16140C T16172C A16183C T16189C C16223T C16279T
Comments: phantom mutations C269G, A3676C, G6026T; note that A3676C was also reported by Sudoyo et al.[10], but the latter sequence data contained many errors[11]

DQ834261 Haplogroup F1a1
A73G G94A C269G 523-524delAC A750G A1438G A2706G A3676C C3970T C4086T T4646C A4769G T6392C T6961C G6962A C7028T C7990T T8069C 8281-8289del A8379G C8409G A8860G G9053A A9217G G9548A G10310A T10609C G11719A G12406A A12715G C12882T G13759A G13928C C14791G A14879G C14953T A15189C A15326G G16129A A16162G T16172C T16304C T16519C
Comments: missing variant 249del (F); phantom mutations C269G, A3676C (this mutation also occurs in the unrelated sequence DQ834260; see above), C8409G, C14791G, A15189C

3. Sequences EF446784 and EF488201
Reference: Noer,A.S., Syukriani,Y.F., Moeis,M.R., Ariwahjoedi,B., Atiemah,E.S. and Siti,H.
General comment: The excess of heteroplasmic-like patterns are likely due to contamination or poor sequence quality

EF446784 Haplogroup Q1b
Human mitochondrial genome of a Javanese origin
Submitted to GenBank: 07-FEB-2007
Released in GenBank: 11-MAR-2007
A73G T89C T146C A263G 309insC 315insC T489C 523-524delAC A750G A1438R A2706R T3398Y G3531K T4117Y A4769R A4985R G5460A A5843G C7028Y T7268G A8701R G8790R A8860R T9540Y A10398R C10400Y T10873Y C11335Y G11719R C12705Y G12940A T13500C C13764Y T14025Y C14766T T14783C G15043A T15067C G15257A G15301A A15326G A15924R G16129A T16144C C16148T T16172C C16223T A16241G A16265C T16311C A16343G C16355T
Comments: G4985A and T11335C are among the listed errors in the original Cambridge reference sequence (CRS)[1]; missing variant C8964T (Q1)

EF488201 Haplogroup N9a6a
Human Mitochondrial Genome of a Sundanese origin
Submitted to GenBank: 09-MAR-2007
Released in GenBank 20-MAR-2007
A73G T146C C150T A263G 309insCC 315insC A750G A1438R A2706R 3106insC A4769R T4856Y A4985R A5186R G5231R G5417A C7028Y A8701R G8790R A8860G C11335Y G11719R A12358R G12372R C12705T C13503M T13674K C13742Y C14766T A15080G A15326G C16223Y C16257M C16261Y C16292Y C16294Y T16362Y T16519C
Comments: 3106insC, G4985A, and T11335C are among the listed errors in the CRS[1]; in region 3106-3107, there is only a single cytosine residue, not the CC doublet in the revised CRS.[1] This common ‘3106’ error occurs in further instances listed

below

4. Sequences DQ418488, DQ437577, DQ462232-DQ462234, DQ519035

Sequences DQ418488, DQ462232-DQ462234, and DQ519035 were generated by the State Key Laboratory of Forensic Sciences, College of Medicine, Xi'an Jiaotong University, 76# Yanta West Road, Xi'an, Shaanxi 710061, P.R.China

General comments: All six sequences miss various variants and show mosaic features pointing to artefactual recombination; five of the unrelated sequences bear T5093C and the remaining one has T5095C

DQ418488 Haplogroup D4c1b1
Homo sapiens isolate Tibet03 mitochondrion
Submitted to GenBank: 27-FEB-2006
Released in GenBank: 01-APR-2006
A73G 249delA A263G 309insCC 315insC T489C A750G A856G A1438G A2706G 4769G T5093C C7028T C8414T A8860G T9540C T10410C T10873C G11719A A14692G C14766T T14783C G15043A G15301A A15326G C16223T T16224C C16245T C16292T T16519C
Comments: missing variants *T146C*, *T195C*, C2766T (D4c1), *G3010A* (D4), C4883T (D), C5178A (D), A8701G (non-N), *A10398G* (non-N), C10400T (M), C12705T (non-R), C14668T (D4), *T16362C*; variant 249delA may have been introduced by an artefactual recombination; phantom mutation T5093C; this sequence was misclassified as pre-HV by Pereira et al.[8]

DQ462232 Haplogroup A5
Homo sapiens isolate HUI MITG 3M mitochondrion
Submitted to GenBank: 27-FEB-2006
Released in GenBank: 01-APR-2006
A73G T146C A263G 309insC 315insC T489C 523-524delAC A750G A1438G G1709A A1736G A2706G T4248C A4769G T5093C C7028T A8563G C8794T A8860G A11081T G11719A C12705T C14766T A15326G T16126C C16223T C16290T T16311C G16319A T16519C T16555G
Comments: missing variants *T152C*, A235G (A), A663G (A), A4824G (A), C11536T (A5); variant T489C may have been introduced by an artefactual recombination; phantom mutation T5093C; this sequence was misclassified as A1 by Pereira et al.[8]

DQ462233 Haplogroup A
Homo sapiens isolate HUI MITG 5F mitochondrion
Submitted to GenBank: 27-FEB-2006
Released in GenBank: 04-APR-2006
A73G T152C A263G 315insC 523-524delAC A750G A1438G A1736G A2706G A3708G T4248C A4769G T5093C C7028T C8794T A8860G G11719A C12888T T13469A C14766T A15326G A15758G C16223T T16519C
Comments: missing variants A235G (A), A663G (A), A4824G (A), C12705T (non-R), C16290T (A), *G16319A* (A); the entire control region points to a non-A sequence; phantom mutations T5093C, T13469A.

DQ462234 Haplogroup C4b

Homo sapiens isolate UIG MITG 1M mitochondrion
Submitted to GenBank: 27-FEB-2006
Released in GenBank: 04-APR-2006
A73G T195C 249delA A263G 309insC 315insC T489C A750G A1438G 2232insA A2706G T3552A A3816G A4769G T5095C G6026A C7028T G7702A G8027A G8584A A8701G A8860G T9540C A9545G T10410C T10873C C11215T G11719A C12705T A13263G T14318C C14766T T14783C T15204C G15301A A15326G A15487T C16072T C16223T C16292T T16298C C16327T
Comments: missing variants A4715G (M8), C7196A (M8), *A10398G* (non-N), C10400T (M), *G11914A* (C), G11969A (C4), G15043A (M); this sequence bears the mutation T10410C characteristic of D4a1

DQ519035 Haplogroup M
Homo sapiens from China mitochondrion
Submitted to GenBank: 28-MAR-2006
Released in GenBank: 28-MAY-2006
A73G T152C A263G 309insC 315insC T489C A540T A750G A1438G A2706G A4769G T5093C C5228T C7028T T7751C C8794T A8860G A9531G T9540C C9696T T9758C T10410C T10873C G11719A C14668T C14766T T14783C G15043A G15301A A15326G C15400T A16183C T16189C C16223T T16362C
Comments: missing variants A8701G (non-N), *A10398G* (non-N), C10400T (M), C12705T (non-R), this sequence bears *T16362C* (D), C14668T (D4), and T10410C (D4a1) but otherwise lacks all other expected variants, so that this is likely a recombinant sequence; phantom mutations A540T, T5093C. This sequence was misclassified as pre-HV by Pereira et al.[8]

DQ437577 Haplogroup T2d
Homo sapiens isolate Mongol mitochondrion
Tuo,Y., Hou,Q.F. and Li,S.B.
Submitted to GenBank: 07-MAR-2006
Released in GenBank: 18-MAR-2006
A73G T152C C194T A263G 309insC 315insC G709A A1438G G1888A A2706G T4216C A4769G T5093C A5747G C7028T C8684T A8774G A8860G T10410C A11251G G11719A A11812G T13260C G13368A T13469A G13708A A14233G C14766T G14905A A15326G C15452A A15607G G15928A T16126C A16175G C16294T C16296T A16497G T16519C
Comments: missing variants A750G (non-H2a2), A4917G (T), G8697A (T), T10463C (T); the sequence bears the mutation T10410C characteristic of D4a1; phantom mutation T13469A; this sequence was misclassified as JT by Pereira et al.[8]

5. Sequences DQ358973-DQ358977
Detjen,K.A., Tinschert,S., Kaufmann,D., Algermissen,B., Nurnberg,P., and Schuelke,M.
Submitted to GenBank: 11-JAN-2006
Released in GenBank: 06-MAR-2006
General comment: All sequences lack 315insC and A750G

DQ358976 Haplogroup K1b2

A73G T146C T195C A263G T1189C A1438G A1811G A2706G A3480G A4769G G5913A C7028T G7211A A8860G G9055A T9698C A10398G A10550G G10646A T11289C A11467G G11719A A12308G G12372A T12738G G13194A C14167T C14766T T14798C A15326G T16224C T16311C T16519C
Comment: additional missing variant T11299C (K)

6. Sequences DQ523619-DQ523681
Fraumene,C., Belle,E.M., Castri,L., Sanna,S., Mancosu,G., Cosso,M., Marras,F., Barbujani,G., Pirastu,M. and Angius,A.
Submitted to GenBank: 02-MAY-2006
Released in GenBank: 03-OCT-2006
Reported in reference 12
General comments: All sequences except for DQ523681 are exactly 16569 bases long, with no insertions or deletions when aligned; in particular, the nearly omnipresent 315insC and the frequent indels at 524 are absent in the entire data set. Evidently, a partially revised CRS has been used as a default sequence, which then still retained the '3106' error [1].

7. Sequences EF660912-EF661013
Gasparre,G., Porcelli,A.M., Bonora,E., Pennisi,L.F., Toller,M., Iommarini,L., Ghelli,A., Moretti,M., Betts,C.M., Martinelli,G.N., Ceroni,A.R., Curcio,F., Carelli,V., Rugolo,M., Tallini,G. and Romeo,G.
Submitted to GenBank: 11-JUN-2007
Released in GenBank: 04-JUL-2007
Reported in reference 13
General comments: 24 sequences in this data set are incomplete in different parts and have a remarkable high frequency of phantom mutations; e.g. mutations G16023T occur in five, C12562G in three, and A10658T and T15813G each in two unrelated sequences; the otherwise rather infrequent A574C occurs four times in this data set. Pereira et al. [8] highlighted five mutations (G4720A, G5185A, G11403A, T10657G, and A15606G) that would not be expected in natural mtDNA sequences; except for A15606G was reported in Bayat et al.[14], all these mutations appear to be lab-specific and which we will regard as phantom mutations. In addition, there are several indels constituting framshift mutations. All these features taken together clearly point to phantom mutations and reading difficulties rather than cancer-specific mutations

EF660912 Haplogroup K1a
A73G T146C C150T T195C A263G 315insC C497T A750G T1189C A1438G A1811G A2706G A3480G A4769G C7028T A8860G G9055A T9698C A10398G A10550G T11299C A11467G G11719A A12308G G12372A C14167T C14766T T14798C A15326G G16023T T16086C T16224C T16311C G16319A T16519C
Comment: phantom mutation G16023T

EF660917 Haplogroup I5a1
A73G T199C T250C A263G 315insC 573insC A750G A1438G G1719A A2706G A4529T A4769G T5074C C7028T A8188G G8251A 8281-8289del A8860G T10034C T10238C A10398G G11719A G12501A C12705T A12961G T13602C

A13780G C14766T G15043A A15326G A15924G T15968C G16129A C16148T C16223T G16391A T16519C
Comment: missing variant A14233G (I5)

EF660918 Haplogroup H13a1a1
A263G 309insC A750G C1352A A1438G C2259T A4745G A4769G G7337A A8860G T11204C T12235C C13680T C14872T A15326G T15900C
Comments: missing variant 315insC; phantom mutation C1352A

EF660923 Haplogroup H1
A263G 315insC T684W A750G A1438G G3010A A4769G A8860G T11736C A15326G C16270T T16276C T16519C G16566A
Comment: phantom mutation T684W

EF660925 Haplogroup H29
A93G A263G 309insC 315insC 523-524delAC 573insCCC A750G A1438G A4769G A5582G A8860G T9077C A12397G T13635A A15326G T16189C T16519C
Comment: phantom mutation T13635A

EF660928 Haplogroup T2b
A73G T195C A263G 315insC A750G G930A A1438G G1888A A2706G T4216C G4580R A4769G A4917G G5147A G5821A C7028T G8697A A8860G T10463C A11251G G11719A A11812G G13368A A14233G C14766T G14905A A15326G C15452A A15607G G15928A T16126C C16294T C16296T T16304C T16519C
Comment: missing variant *G709A* (T)

EF660929 Haplogroup J1c1c1
(1-21)missing T23A 31insC 71insG A73G 114insC T146C G185A A188G C222T G228A A263G C295T 315insC C462T A750G A1438G G2702A A2706G G3010A T4216C A4769G C7028T C7441A A8860G A10398G G10685A A11251G G11719A A12612G T13281C G13708A A13933G G13980A C14766T T14798C A15326G C15452A C16069T T16126C G16213A C16294T T16519C
Comments: sequence is incomplete; missing variant T489C (J); phantom mutations T23A, 31insC, 71insG, 114insC, C7441A

EF660930 Haplogroup U5a2b
(1-38)missing A73G A263G 315insC A750G 960insC A1438G A2706G T3197C A4769G C7028T G7337A A8860G G9477A G9548A 10115-10116delTA C11177A A11467G G11719A A12308G G12372A T13617C T13753C C14766T A14793G A15326G C16192T C16256T C16270T T16311C G16526A
Comments: sequence is incomplete; phantom mutations C11177A, 10115-10116delTA (frame shift mutation in the *MT-ND3* gene)

EF660934 Haplogroup H1h
T152C A263G 309insC 315insC 523-524delAC A750G A1438G G3010A A4769G G7013A T7818R A8860G A11167G G11914A T11935G A14887G A15326G T16189C A16194C T16243C C16261T A16343G T16519C
Comment: phantom mutations T7818R, T11935G, A16194C

EF660935 Haplogroup HV1a2
A263G 315insC A374G A750G A1438G G1664C A2706G G4596A A4769G C7028T A8014T T8277C T8279C A8280C A8860G G8994A A12361G T12601C A15218G A15326G T15796C C16067T
Comment: phantom mutations G1664C, A8280C

EF660937 Haplogroup N1c
A73G T152C A189G T195C G207A A210G A263G 315insC A750G A1438G G1719A A2706G A4769G A5319G C7028T T8222C A8308G A8860G C8943T T9119R T10238C T11025C T11437C G11719A G12501A C12562G A12612G C12705T A13637G A13780G G14560A C14766T A15326G T16086C C16201T C16223T A16265G C16355T T16519C
Comments: missing variant *T204C*; phantom mutations T9119R, C12562G

EF660949 Haplogroup H1e1
T146C A263G 315insC A750G A1438G A1555G G3010A A4769G G5460A A8512G A8860G A15326G T15813G C16174T T16519C
Comment: phantom mutation T15813G

EF660950 Haplogroup U5a
A73G A263G 315insC A750G A1438G A2706G T3197C A3714G A4769G C7028T A8860G G9387S G9477A C10619T A10768G A11467G G11719A A12308G G12372A T13617C T14034C C14766T A14793G A15326G A15791G T16172C C16192T C16256T C16270T
Comment: phantom mutation G9387S

EF660951 Haplogroup H
A263G 315insC A750G A1438G A4769G T5442C A8860G G9932A C12562G C14149T A15326G T16519C
Comment: phantom mutation C12562G

EF660958 Haplogroup HV
A263G 309insC 315insC A750G A1438G A2706G A4769G C7028T A8860G C12562G C14830T A15326G T16311C
Comment: phantom mutation C12562G

EF660966 Haplogroup W4
A73G G143A A189G C194T T195C T196C T204C G207A A263G 309insC 315insC G709A A750G T1243C A1438G A2706G A3505G A4769G G5046A G5147A G5460A C7028T G8251A A8860G G8994A C11674T G11719A T12414C C12705T C14766T A15326G G15884C C16187T C16223T C16234T T16249C C16266T C16287T C16292T T16519C
Comment: missing variant A11947G (W)

EF660967 Haplogroup J2a2a
A73G C150T T195C A235G A263G C295T 309insC 315insC T489C G564C A565C A567C A574C A750G A1438G 2149insAG A2706G T4216C A4769G T6671C C7028T C7476T C8386T A8860G A10398G A10499G C10684G A11002G T11204C A11251G G11377A G11719A A12570G A12612G A14656G C14766T G15257A

A15326G C15452A A15679G C16069T T16126C
Comments: missing variant *G13708A* (J); phantom mutations G564C, A565C, A567C, C10684G

EF660968 Haplogroup HV
T195C A263G 315insC A357C A750G A1438G A2706G A4769G G6962A C7028T A8706G A8860G A15326G T16298C
Comment: phantom mutation A357C

EF660971 Haplogroup R0a1a
T58C C64T T146C A263G 309insC 315insC 524insAC A750G A827G A1438G T2442C A2706G T3847C A4769G A5605G G6026R G6480A C7028T G8292A A8860G C11761T C13188T C14766T A15326G T16126C T16362C T16519C
Comment: missing variant C16355T (R0a1a)

EF660972 Haplogroup T2b
A73G T195C A263G 315insC A750G G930A A1438G A2706G T4216C A4769G T4823C A4917G G5147A G5185A (5204-5274)missing C6935T C7028T (8183-8532)missing G8697A A8860G T10463C A11251G G11719A A11812G G13368A A14233G C14766T G14905A A15326G A15607G G15928A T16126C C16294T C16296T T16304C T16519C
Comments: sequence is incomplete; missing variants *G709A* (T), G1888A (T), C15452A (JT); phantom mutation G5185A

EF660976 Haplogroup H
(1-55)missing A263G 309insC 315insC A750G A1438G A4769G (5434-5716)missing C6215A T8618C A8860G A15326G T16519C
Comments: sequence is incomplete; phantom mutation C6215A

EF660978 Haplogroup T2b2
A73G A263G 315insC A750G G930A A1438G G1888A A2706G T4216C A4769G A4917G G5147A C7028T G8697A A8860G T10463C C11242G A11251G G11719A A11812G G13368A A14233G C14766T G14905A A15326G C15452A A15607G G15928A T16126C C16192T C16294T T16519C
Comment: missing variants *G709A* (T), *T16304C* (T2b)

EF660984 Haplogroup J2a2a
A73G C150T T195C A235G C262T A263G C295T 315insC T489C A750G A1438G 2463insA A2706G T4216C G4309A A4769G T6671C C7028T C7476T C8386T A8860G (10281-10436)missing A10499G (10575-10652)missing T10657G A10658T A11002G A11251G G11377A G11719A A11797G A12570G A12612G C13056T A13419G G13708A C14766T G15257A A15326G C15452A A15679G C16069T T16126C C16169T (16549-16569)missing
Comments: sequence is incomplete; phantom mutations T10657G, A10658T

EF660987 Haplogroup I4
A73G T199C T204C T250C A263G 309insC A750G A1438G G1719A A2706G A4769G C7028T G8251A G8519A A8860G T10034C T10238C A10398G A10819G G11719A A13780G C14766T A15326G C16223T T16304C G16391A T16519C

Comments: missing variants 315insC, A4529T (N1e'I), G12501A (N1), C12705T (non-R), G15043A (N1a'e'I), *A15924G* (N1e'I), *G16129A*

EF660988 Haplogroup H13a2a
(1-38)missing A263G 309insC 315insC A750G A1008G C2259T 3571insC A4134G (4168-4219)missing (4254-4388)missing A4769G A5656G A8860G G9575A A12662G A13105G C14872T A15326G T16172C A16183C (16186-16228)missing C16282A A16285C T16519C
Comments: sequence is incomplete; missing variants *G709A* (H13a2), A1438G (non-H2); phantom mutations 3571insC, C16282A, A16285C

EF660989 Haplogroup U2e
A73G T152C A263G 315insC A750G G988A A1438G A1811G (2442-2610)missing A2706G A3720G A4769G A5390G C7028T A8860G A10876G A11467G G11719A A12308G G12372R T13020R T13734R G14560R C14766T G14859T A15326G C15661T A15907G A16051G G16129C A16183R T16189R A16258C T16362C
Comments: this sample was probably contaminated, displaying an excess of seeming heteroplasmy; phantom mutations T13020R, T13734R, T16189R; missing variants A508G (U2e), T5426C (U2e), C6045T (U2e), T6152C (U2e), T10320C (U2e), T13734C (U2e), *T16189C*

EF660990 Haplogroup H14a
T152C A263G 309insC 315insC 523-524delAC A574C A750G A1438G A4769G A6035G T7645C A8860G A10217G G10573A A15326G T15456G C15457G C15460A (15462-15603)missing A15606G A15607T C16256T T16311C T16352C
Comment: phantom mutations T15456G, C15457G, C15460A, A15607T

EF660991 Haplogroup H4a1a1a
A73G A263G 309insC 315insC 523-524delAC A750G A1438G A3305G C3992T A4024G A4769G T5004C (5484-5724)missing G8269A A8860G G9123A T10007C T10034C A10044G G12406A (13344-13533)missing A14582G A15326G A16317G T16362C
Comments: sequence is incomplete; missing variant C14365T (H4a)

EF660992 Haplogroup T2b
1-5del A73G A263G 315insC T334C G709A A750G G930A A1438G A2706G T4216C (4295-4388)missing A4769G A4917G G5147A C7028T G8697A A8860G A9180G G9966A T10463C A11251G G11440A G11719A A11812G G12056A G13368A A14233G C14766T G14905A A15326G C15452A A15607G G15928A T16126C C16292T C16294T C16296T T16304C T16519C
Comments: sequence is incomplete; missing variant G1888A (T)

EF660993 Haplogroup N1b1d
A73G T152C G185A A188G A263G 315insC A750G A1438G G1598A C1703T G1719A C2639T A2706G 3571insC C3921A A4769G C4960T G5471A C7028T G8251A A8261G C8410T C8472T T8763C A8836G A8860G C9335T T10238C A11362G G11719A G12501A C12705T A12822G A14053G C14766T A15326G T15813G G16129A G16145A C16176G C16223T C16291T T16297C G16390A
Comment: phantom mutations 3571insC, T15813G

EF660994 Haplogroup H
A73G A263G 309insC 315insC A750G A1438G T1452C T1716C A4769G A8860G A9377G G10586A 11085-11086delCA A15326G C16186T T16519C
Comment: phantom mutation 11085-11086delCA (frame shift mutation in the *MT-ND4* gene)

EF660995 Haplogroup H
T195C A263G 315insC A335G A750G A1438G A4769G G4831A A8860G 11038delA T13602C G15047A A15326G A16265C T16519C
Comment: phantom mutation 11038delA (frame shift mutation in the *MT-ND4* gene)

EF660996 Haplogroup H1q
A263G 309insC 315insC A750G A1438G G3010A (3528-3628)missing A4769G T4859C A8860G C10043T (10644-10684)missing 13235insT A15326G T16189C T16519C
Comments: sequence is incomplete; phantom mutation 13235insT (frame shift mutation in the *MT-ND5* gene)

EF660998 Haplogroup H1q
A263G 315insC A750G A1438G G3010A G3244A (4168-4388)missing T4688C A4769G T4859C G8839A A8860G T10656G A10658T G14569A A15326G G16023T T16172C C16173T C16188G T16189C T16519C
Comments: sequence is incomplete; phantom mutations T10656G, A10658T, G16023T

EF660999 Haplogroup H1
T152C A263G T310C 315insC A750G A1438G G3010A (4168-4388)missing A4769G C8346T A8860G (10633-10681)missing (14374-14488)missing A15326G G16023T A16207G T16519C
Comments: sequence is incomplete; phantom mutation G16023T

EF661000 Haplogroup T1a1
A73G T152C T195C A263G 315insC G709A A750G A1438G G1888A A2706G G3244A T4216C G4720A A4769G A4917G G4996A C7028T G8697A A8860G T9899C T10463C A11251G G11719A C12633A G13368A A14274G C14766T G14905A A15326G C15452A G15928A G16023T T16126C A16163G C16186T T16189C C16294T T16519C
Comments: missing variant A15607G (T); phantom mutations G4720A, G16023T

EF661001 Haplogroup T2e1
C41T A73G C150T A263G 315insC G709A A750G A1438G G1888A A2706G T4216C A4769G A4917G C5396T C7028T (8315-8532)missing G8697A A8860G G9932A T10463C A11251G G11719A A11812G G13368A A14233G C14766T G14905A A15326G C15452A A15607G G15928A G16023T T16126C G16153A C16294T T16519C
Comments: sequence is incomplete; phantom mutation G16023T

EF661005 Haplogroup V1a2

T72C A263G 309insC 315insC A750G A1438G A2706G G4580A T4639C A4769G C5263T C7028T A7055G A8860G A8869G G11403A A12490G A15326G C15904T T16298C
Comment: phantom mutation G11403A

EF661006 Haplogroup U2e1
A73G T152C T217C A263G 315insC C340T A508G A750G G988A A1438G A1811G A2706G 3229insA G3392A A3720G A4769G A5390G T5426C A5900C C6045T T6152C T6620C C7028T A8860G G9182A A10876G A11467G G11719A A12308G G12372A T13020C T13572C T13734C C14766T A15326G C15661T A15907G A16051G G16129C A16183C T16189C C16256T T16362C
Comment: phantom mutation A5900C

EF661010 Haplogroup H3
A263G 315insC C411A A750G A1438G T4222C A4769G C5960T T6776C A8860G T12811C T13474C A15326G T15804C C15823T A15860G T16093C C16266T T16311C T16519C
Comment: C411A is restricted to this data set.

8. Sequences EU370391-EU370397
Abu-Amero,K.K., Larruga,J.M., Cabrera,V.M. and Gonzalez,A.M.
Submitted to GenBank: 26-DEC-2007
Released in GenBank: 25-MAR-2008
Reported in reference 15
General comment: All sequences are incomplete in several parts

9. Sequences FJ236978-FJ236983
Ennafaa,H., Cabrera,V.M., Abu-Amero,K.K., Gonzalez,A.M., Amor,M.B., Bouhaha,R., Dzimiri,N., Elgaaied,A.B. and Larruga,J.M.
Submitted to GenBank: 26-SEP-2008
Released in GenBank: 22-MAR-2009
Reported in reference 16
General Comment: All sequences have the '3106' error

10. Sequences AM260596-AM260597
Annunen-Rasila,J., Finnila,S., Mykkanen,K., Moilanen,J.S., Veijola,J., Poyhonen,M., Viitanen,M., Kalimo,H. and Majamaa,K.
Submitted to GenBank: 25-APR-2006
Released in GenBank: 01-AUG-2006
Reported in reference 17
General comment: The control region sequence information is not reported

AM260596 Haplogroup U2e1a1
A750G A1438G A2706G C3116T T3197C G4113A A4769G A5390G T5426C C6045T T6152C C7028T G8857A A8860G A10876G C11197T A11467G G11719A T11732C A12308G G12372A T13020C T13734C C14766T A15326G A15907G

Comment: missing variants A1811G (U2'3'4'7'8'9), A3720G (U2e)

AM260597 Haplogroup U3b1
A750G A1438G A1811G A2706G C3546A A4769G A6359G C7028T A8860G T9656C A11467G G11719A A12308G G12372A T13743C A14139G C14766T A15326G T15454C
Comment: missing variants A4188G (U3b), C4640A (U3b)

11. Sequence AY289073
Ingman,M. and Gyllensten,U.
Submitted to GenBank: 04-OCT-2006
Released in GenBank: 02-MAY-2003
Reported in reference 18

AY289073 Haplogroup U1a
A73G T199C A263G C285T 309insC 315insC A385G 523-524delAC A750G T1005C A1438G C2218T A2706G 3158insT G3591A A4769G G4991A G6026A C7028T T7581C A8701G A8860G A9288G A11467G G11719A C12403T T12957C A13104G A13422G A14070G G14364A C14766T G15148A A15326G A15954C A16183C T16189C 16193insC T16249C C16400T
Comments: missing variants A12308G (U), G12372A (U), T12879C (U1); this sequence may be a recombinant with a haplogroup M1 type, whereby A8701G (non-N), C12403T (M1) may have been introduced

12. Sequences AY195745, AY195756, AY195767, and AY195775
Mishmar,D., Ruiz-Pesini,E., Golik,P., Macaulay,V., Clark,A.G., Hosseini,S., Brandon,M., Easley,K., Chen,E., Brown,M.D., Sukernik,R.I., Olckers,A. and Wallace,D.C.
Submitted to GenBank: 11-DEC-2002; updated on 10-JUN-2004
Released in GenBank: 10-JUN-2004
Reported in reference 19

AY195745 Haplogroup T2b
A73G T152C T195C A263G G709A A750G G930A A1438G G1888A A2706G T4216C A4769G A4917G G5147A C7028T A8860G T10463C A11251G G11719A A11812G G13368A A14233G C14766T A14836G G14905A C14974G A15326G C15452A A15607G G15928A T16126C C16294T T16304C T16519C
Comments: missing variants 315insC, G8697A (T); phantom mutation C14974G

AY195756 Haplogroup N1b1c
A73G T146C C150T T152C A263G 309insCC 315insC C320T 523-524delAC A750G T961C 965insCCCC A1438G G1598A C1703T G1719A C2639T A2706G T3083C C3921A A4769G C4960T G5471A C7028T G8251A C8472T A8836G A8860G A8962G A9093C C9335T T9957C T10238C A11362G G11719A G12501A C12705T A12822G C14766T A15326G G16145A C16176G C16223T G16390A T16519C
Comments: phantom mutations C320T, A9093C; the former is caused by the length

heteroplasmy of the C stretch 303–309 (see Brandstätter et al. [9])

AY195775 Haplogroup H1b
A263G 309insC 315insC C320T 523-524delAC A750G A1438G G3010A A3796G A4769G T8298C A8860G A15326G T16189C T16356C T16362C T16519C
Comments: phantom mutation C320T, caused by the length heteroplasmy of the C stretch 303–309 (see Brandstätter et al. [9])

AY195767 Haplogroup T2b1
A73G A263G 309insC 315insC C317G G709A A750G G930A A1438G G1888A A2012R A2706G T4216C A4769G A4917G G5147A C7028T G8697A A8860G T10463C A11251G G11719A A11812G G13368A G14016A A14233G C14766T G14905A A15326G C15452A A15607G G15928A T16126C C16294T C16296T T16304C T16519C
Comments: phantom mutation C317G, caused by the length heteroplasmy of the C stretch 303–309 (see Brandstätter et al. [9])

13. Sequences EU095205, EU095208, EU095250
Fagundes,N.J.R., Kanitz,R., Eckert,R., Valls,A.C.S., Bogo,M.R., Salzano,F.M., Glenn Smith,D., Silva,W.A. Jr., Zago,M.A., Ribeiro-dos-Santos,A.K., Santos,A.E.B., Petzl-Erler,M.L. and Bonatto,S.L.
Submitted to GenBank: 13-AUG-2007
Released in GenBank: 07-MAR-2008
Reported in reference 20

EU095205 Haplogroup A2
C64T A73G T146C A153G G207A A235G A263G 310insT C317G 523-524delAC A663G A750G A1438G A1736G A2246G A2706G T4248C A4769G A4824G T6216C C7028T G8027A C8794T A8860G G11719A G12007A C12705T C14766T A15326G C16111T C16223T C16290T C16291T G16319A T16362C
Comment: phantom mutations 310insT, C317G, caused by the length heteroplasmy of the C stretch 303–309 (see Brandstätter et al. [9])

EU095208 Haplogroup B2b
A73G A263G 310insT C317G G499A 523-524delAC A750G A827G A1438G A2706G A3547G A4385G A4769G G4820A T4977C C6473T G6755A C7028T T7278C C7810T 8281-8289del A8860G T9950C C11177T G11719A G13590A T14110C C14766T A15326G C15535T A16183C T16189C T16217C T16519C
Comments: phantom mutations 310insT, C317G, caused by the length heteroplasmy of the C stretch 303–309 (see Brandstätter et al. [9])

EU095250 Haplogroup X2a1a1
A73G G143A T195C A200G T204C G207A A263G 315insC C317G A750G A1438G G1719A C2393T A2706G T3552C A4769G A6113G T6221C C6371T C7028T A8860G A8913G G11719A A12397G C12705T A13966G T14470C T14502C C14766T A15326G A16183C T16189C G16213A C16223T C16278T C16291T G16319A T16357C T16519C
Comments: phantom mutation C317G, caused by the length heteroplasmy of the C

stretch 303–309 (see Brandstätter et al. [9])

14. Sequences AY339402- AY339593

Moilanen,J.S., Finnila,S., Lehtonen,M.S. and Majamaa,K.
Submitted to GenBank: 11-JUL-2003
Released in GenBank: 10-OCT-2007
Reported in reference 21
General comment: The data set reported in reference 21 was not entirely produced by direct sequencing. Some sequences may have an omission of C340T. Six sequences are annotated in detail as follows:

AY339437 Haplogroup V1a1
T72C A227G A263G 309insC 310insT C317G T485C A750G A1438G A2706G G4580A T4639C A4769G C5263T C7028T A8860G A8869G A15326G C15904T A16183G T16298C
Comment: phantom mutation 310insT, C317G

AY339463 Haplogroup W1b
A73G A189G T195C T204C G207A A227G A263G C330G G709A A750G T1243C A1438G A2706G A3505G A4769G T4928C G5046A G5460A C7028T C7864T G8251A A8860G G8994A G9612A C11674T G11719A A11947G T12414C C12705T C14766T A15326G G15884C C16223T C16292T T16519C
Comment: missing variant 315insC; phantom mutation C330G (see Brandstätter et al. [9])

AY339546 Haplogroup U2e1a1
A73G T152C A263G 309insC 315insC A508G 524insAC A750G A1438G A1811G A2706G C3116T T3197C A3720G A4769G A5390G T5426C C6045T T6152C C7028T A8860G A10876G C11197T A11467G G11719A T11732C A12172G A12308G G12372A T13020C T13734C G13928A C14766T A15326G A15907G A16051G G16129C A16183C T16189C T16362C T16519C
Comment: missing variants T217C (U2e1), C340T (U2e1)

AY339549 Haplogroup U4d1
A73G T195C A263G 309insCC 315insC G499A 524insAC A750G A1438G A1811G 2405insC A2706G C2772T T4646C A4769G A5984G T5999C A6047G C6653A C6938T C7028T T8260C A8860G A11467G G11719A A12308G G12372A C14620T C14766T A15326G T15693C T16356C T16519C
Comment: missing variants T629C (U4d), C11332T (U4)

AY339581 Haplogroup J1b1a1a
A73G C242T A263G C264T C295T 315insC C462T T489C A750G A1438G T2158C A2706G G3010A T4216C A4769G G5460A C7028T G8269A T8286C 8287insC G8557A A8860G A10398G A11251G G11719A A12612G G13708A T13879C C14766T T15067C A15326G C15452A C16069T T16126C G16145A T16172C C16222T A16247G C16261T G16274A T16519C
Comment: missing variant G12007A (J1b1a)

AY339582 Haplogroup J1b1a1a
A73G C242T A263G C264T C295T C462T T489C A750G A1438G T2158C A2706G G3010A T4216C A4769G G5460A C7028T G8269A T8286C 8287insC G8557A A8860G A10398G A11251G G11719A A12612G G13708A T13879C C14766T T15067C A15326G C15452A C16069T T16126C G16145A T16172C C16186T C16222T C16261T G16274A T16519C
Comment: missing variants 315insC, G12007A (J1b1a)

15. Sequence AF46968, AF346973, and AF347006
Ingman,M., Kaessmann,H., Paabo,S. and Gyllensten,U.
Submitted to GenBank: 09-FEB-2001
Released in GenBank: 22-AUG-2003
Reported in reference 22

AF346968 Haplogroup L1c1a1a1a
44insC A73G C151T T152C C186A A189C T195C T204C G247A A263G A297G 315insC G316A C467T 523-524delAC A750G G769A T825A G1018A A1438G A2308G 2395delA A2706G G2758A T2885C C3594T G3666A A3796T A4104G A4769G A5951G A5984G T6071C G6182A C7028T A7055G A7146G C7256T T7389C G7521A G8027A T8087C C8468T C8655T A8701G A8860G T8928C A9072G T9311C T9540C T10321C G10586A G10688A T10810C T10873C A11167G C11257T G11719A T11899C C12705T A12810G A12930T A13105G A13485G C13506T C13650T T13789C T14000A T14034C T14088C A14148G T14178C G14560A C14766T C14911T A15326G T15663C G16129A T16189C C16214T C16234T T16249C T16271C G16274A C16278T C16294T T16311C C16360T T16519C
Comments: missing variants A3843G (L1c1), T4454A (L1c1a), and *C16187T* (non-L2'3'4'6); possibly a recombinant [23]

Sequence AF346973 Haplogroup F4a
A73G T146C A263G 309insC 315insC C317A A750G A1438G A2706G T3290C G3316A C3970T G4512A A4769G C5263T T6392C C7028T T7561C A8860G G10310A T10915C G11719A T12134C C12153T G12630A T13602C G13928C C14766T G15110A A15326G T15670C G15803A T16126C A16207G T16304C T16362C A16399G
Comments: missing variant 249delA (F); this sequence was ambiguously classified as F4/M9a by Pereira et al. [8]

AF347006 Haplogroup V7a
T72C A93G A95C A263G 309insC 315insC 523-524delAC A750G A1438G A2706G C3549T G4580A A4769G C7028T G7444A A8860G T11899C C14097T C14766T A15326G C15904T G16153A T16298C
Comment: this mtDNA might be a recombinant having gained C14097T and C14766T (non-HV) [23]

16. Editing errors in sequences from GenBank

The following sequences in GenBank have editing errors and are included Pereira et al.'s study. [8]
1) DQ523681 reported by Fraumene et al. [12]: site 16569 missing G
2) EF556162 reported by Behar et al. [24]: 16569insG
3) EU742151 reported by Feder et al. [25]: 16569insGATC
4) AP008336 reported by Tanaka et al. [26]: 16569insG
5) AP008866 reported by Tanaka et al. [26]: 16569insGATCACAG
6) EF060364 reported by La Morgia et al. [27]:
16569insGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCAT
GCATT; this sequence was corrected on 20-AUG-2009

17. Inconsistence of mtDNA sequence variation in the presumably same individuals reported by Kivisild et al. [28] and Hartmann et al. [29]

General comments: Several samples analyzed by Hartmann et al. [29] seem to be identical to samples analyzed previously by Kivisild et al. [28] The latter study only reported the coding region (which is located in region 436-16021) sequence variation [28]. However, there are some inconsistencies regarding the sequence variation in the overlapping region of the presumably same samples between these two reports

(1) EU597516 (HGDP00167) & DQ112952 (As 18) Haplogroup M2a1
Comment: DQ112952 lacked the phylogenetically expected M2 variant A11083G compared to EU597516

DQ112952
Submitted to GenBank: 30-JUN-2005
Updated in GenBank: 18-OCT-2006
Deposited in GenBank by Shen,P. and Oefner,P.
Reported in reference 28
(1-435)missing C447G T489C 523-524delAC A750G A1438G T1780C A2706G A4769G G4924A A4965G G5252A C7028T T7961C A8396G A8502G A8701G A8860G T9540C T9758C T9965C A10398G C10400T T10873C G11719A C12705T A12810G C14766T T14783C G15043A G15301A A15326G T15670C (16022-16569)missing

EU597516
Submitted to GenBank: 06-APR-2008
Released in GenBank: 26-MAR-2008
Deposited in GenBank by Hartmann,A., Thieme,M., Nanduri,L.K., Stempfl,T., Moehle,C., Kivisild,T. and Oefner,P.J.
Reported in reference 29
A73G T195C T204C A263G 309insC 315insC C447G T489C 523-524delAC A750G A1438G T1780C A2706G A4769G G4924A A4965G G5252A C7028T T7961C A8396G A8502G A8701G A8860G T9540C T9758C T9965C A10398G C10400T T10873C A11083G G11719A C12705T A12810G C14766T T14783C G15043A G15301A A15326G T15670C C16223T C16270T G16319A T16352C T16519C

(2) EU597569 (HGDP00709) & DQ112790 (Am 15) Haplogroup B2e
Comment: EU597569 lacked C15265T compared to DQ112790

DQ112790
Submitted to GenBank: 30-JUN-2005
Updated in GenBank: 18-OCT-2006
Deposited in GenBank by Shen,P. and Oefner,P.
Reported in reference 28
(1-435)missing G499A A750G A827G A1438G A2706G A3547G A4769G G4820A T4977C C6119T C6473T C7028T 8281-8289del A8860G T9950C C11177T G11719A G13590A C14049T C14766T C15265T A15326G C15535T (16022-16569)missing

EU597569
Submitted to GenBank: 06-APR-2008
Released in GenBank: 26-MAR-2008
Deposited in GenBank by Hartmann,A., Thieme,M., Nanduri,L.K., Stempfl,T., Moehle,C., Kivisild,T. and Oefner,P.J.
Reported in reference 29
A73G C194T T199C A263G 309insCC 315insC G499A A750G A827G A1438G A2706G A3547G A4769G G4820A T4977C C6119T C6473T C7028T 8281-8289del A8860G T9950C C11177T G11719A G13590A C14049T C14766T A15326G C15535T A16183C T16189C T16217C T16519C

(3) EU597580 (HGDP00710) & DQ112791 (Am 16) Haplogroup B2
Comments: EU597580 missed the expected B4b variant G499A; the presence of the rare variant C15265T in EU597580 and DQ112790 that may belong to different subhaplogroups of B2 is suspicious. However, we were informed by Peter Oefner who checked the original sequencing traces that the genotypes were as reported

DQ112791
Submitted to GenBank: 30-JUN-2005
Updated in GenBank: 18-OCT-2006
Deposited in GenBank by Shen,P. and Oefner,P.
Reported in reference 28
(1-435)missing G499A A750G A827G A1438G A2706G A3547G T3760G A4769G G4820A T4977C T5095C C6473T C7028T C7813T 8281-8289del A8860G T9950C C11177T G11719A G13590A C14766T A15326G C15535T (16022-16569)missing

EU597580 Submitted to GenBank: 06-APR-2008
Released in GenBank: 26-MAR-2008
Deposited in GenBank by Hartmann,A., Thieme,M., Nanduri,L.K., Stempfl,T., Moehle,C., Kivisild,T. and Oefner,P.J.
Reported in reference 29
A73G A263G 309insCC 315insC A750G A827G A1438G A2706G A3547G T3760G A4769G G4820A T4977C T5095C C6473T C7028T C7813T 8281-8289del A8860G T9950C C11177T G11719A G13590A C14766T C15265T A15326G C15535T C16150T A16183C T16189C T16217C T16519C

18. Sequences EU443443 - EU443514

Rao,V.R., Kumar,S., Padmanabham,P.B.S.V., Ravuri,R.R., Uttaravalli,K., Koneru,P., Mukherjee,P.A., Das,B., Kotal,M., Xaviour,D. and Saheb,S.Y.
Submitted to GenBank: 31-JAN-2008
Released in GenBank: 13-AUG-2008
Reported in reference 30
General comments: Ten sequences in this data set had rare frame shift mutations in the coding regions; some sequences missed expected variants

EU443443 Haplogroup M2a1
A73G T204C A263G 315insC A335G C447G T489C G526A A750G A1438G T1780C A2706G G3483A A4769G G5252A A5990G C7028T T7961C A8396G A8502G A8701G A8860G T9540C T9758C A10398G C10400T T10873C A11083G G11719A C12705T A12810G A14596G C14766T T14783C G15043A G15301A A15326G T15670C 15719insT C16176T C16223T C16270T G16274A G16319A T16352C T16519C
Comment: phantom mutation 15719insT (frame shift mutation in the *MT-CYB* gene)

EU443444 Haplogroup M2a1
A73G T204C A263G 309insC 315insC A335G C447G T489C G526A A750G A1438G T1780C A2706G G3483A A4769G G5252A A5990G C7028T T7961C A8396G A8502G A8701G A8860G T9540C T9758C A10398G C10400T T10873C A11083G G11719A C12705T A12810G A14596G C14766T T14783C G15043A G15301A A15326G T15670C 15719insT C16176T C16223T C16270T G16274A G16319A T16352C T16519C
Comment: phantom mutation 15719insT (frame shift mutation in the *MT-CYB* gene)

EU443455 Haplogroup M2a1
A73G A200G T204C A263G 315insC C447G T489C A750G A1438G G1462A T1780C A2706G T4216C A4769G G5252A C7028T T7961C A8396G A8502G A8701G A8860G T9540C T9758C A10398G C10400T T10873C A11083G 11543insA G11719A C12705T A12810G C14766T T14783C G15043A G15301A A15326G T15670C C15700T A15924G T16017C T16075C C16223T C16270T G16274A C16278T G16319A T16352C T16519C
Comment: phantom mutation 11543insA (frame shift mutation in the *MT-ND4* gene)

EU443456 Haplogroup M2a1
A73G T204C A263G 315insC C447G T489C A750G A1438G G1462A T1780C A2706G T4216C A4769G G5252A C7028T T7961C A8396G A8502G A8701G A8860G T9540C T9758C A10398G C10400T T10873C 11543insA G11719A C12705T A12810G C14766T T14783C G15043A G15301A A15326G T15670C C15700T A15924G T16017C T16075C C16223T C16270T G16274A C16278T G16319A T16352C T16519C
Comments: missing variant A11083G (M2); phantom mutation 11543insA (frame shift mutation in the *MT-ND4* gene)

EU443468 Haplogroup M2a1
A73G T195C C198T T204C G207A A263G 315insC C447G T482C T489C A750G A1438G T1780C A2706G A4769G G5252A C7028T C7151T T7961C A8396G A8502G A8701G A8860G T9540C T9758C A10398G C10400T T10873C A11083G

G11719A T11864C C12705T A12810G C14751T C14766T G15043A G15301A A15326G T15670C T16093C C16223T C16270T G16319A T16352C T16519C
Comment: missing variant T14783C (M)

EU443476 Haplogroup M2a1
A73G T195C T199C T204C A263G 309insC 315insC C447G T489C A750G T961C 965insC A1438G T1780C A2706G A4769G G5252A C7028T T7961C A8396G A8502G A8701G A8860G T9540C T9758C 9959delT A10398G C10400T A10819G T10873C A11083G G11719A G12501A C12705T A12810G C14766T T14783C G15043A G15301A A15326G T15670C C16223T A16230G T16243C C16270T G16319A T16352C T16519C
Comment: phantom mutation 9959delT (frame shift mutation in the *MT-CO3* gene)

EU443477 Haplogroup M2a1
A73G T195C T199C T204C A263G 309insC 315insC C447G T489C A750G A1438G T1780C A2706G A4769G G5252A C7028T T7961C A8396G A8502G 8527insA A8701G A8860G T9540C T9758C A10398G C10400T A10819G T10873C A11083G G11719A G12501A C12705T A12810G C14766T T14783C G15043A G15301A A15326G T15670C C16223T A16230G T16243C C16270T G16319A T16352C T16519C
Comment: phantom mutation 8527insA (frame shift mutation in the *MT-ATP8* and *MT-ATP6* genes)

EU443478 Haplogroup M2a1
A73G T195C T199C T204C A263G 309insC 315insC C447G T489C A750G A1438G T1780C A2706G A4769G G5252A 6322delG C7028T T7961C A8396G A8502G A8701G A8860G T9540C T9758C A10398G C10400T A10819G T10873C A11083G G11719A G12501A C12705T A12810G C14766T T14783C G15043A G15301A A15326G T15670C C16223T A16230G C16242T T16243C C16270T G16319A T16352C T16519C
Comment: phantom mutation 6322delG (frame shift mutation in the *MT-CO1* gene)

EU443479 Haplogroup M2a1
A73G T195C T199C T204C A263G 309insC 315insC C447G T489C A750G A1438G T1780C A2706G A4769G G5252A C7028T T7961C A8396G A8502G A8701G A8860G 9537delC T9540C T9758C A10398G C10400T A10819G T10873C A11083G G11719A G12501A C12705T A12810G C14766T T14783C G15043A G15301A A15326G T15670C 15943-15944delTT C16223T A16230G T16243C C16270T G16319A T16352C T16519C
Comment: phantom mutation 9537delC (frame shift mutation in the *MT-CO3* gene)

EU443491 Haplogroup M2a2
A73G A263G 315insC C447G T489C A750G A1438G T1780C A2706G A4769G C7028T G7702A T7961C A8396G A8502G A8701G A8860G T9935C A10398G C10400T T10873C C11041A A11083G G11719A T12657C C12705T A12810G G13708A C14766T T14783C G15043A G15301A A15326G T15670C C16223T A16240C G16274A T16311C G16319A T16519C
Comment: missing variant T9540C (non-N)

EU443497 Haplogroup M2b1
A73G T152C C182T T195C A263G 309insC 315insC C447G T471C T489C 523-524delAC C549T A750G A1438G A1453G T1780C A2706G G2831T C3630T A4769G C5263T T5420C A5480G G5744A A5747G G6260A A6647G C7028T G7337A A8502G T8632C A8701G A8860G T9233C T9540C T9899C A10398G C10400T T10873C A11083G G11719A 12617insT C12705T T13254C T14783C G15043A G15301A A15326G T15670C G15777C 16169insC T16189C 16193insC C16223T G16274A G16319A C16320T T16519C
Comment: phantom mutation 12617insT (frame shift mutation in the *MT-ND5* gene)

EU443512 Haplogroup M2b
A73G T146C T152C C182T T195C A263G 315insC C447G T489C 523-524delAC A567G A750G A1438G A1453G T1780C A2706G G2831T T3398C C3630T 4511insT A4769G G5744A A6647G C7028T A8502G A8701G A8860G T9540C T9899C A10398G C10400T T10873C A11083G 11376delA G11719A G12630A C12705T T13254C T14783C G15043A G15301A A15326G T15670C T15862G 16169insC A16183C T16189C 16193insC C16223T G16274A G16319A C16320T T16519C
Comment: phantom mutations 4511insT (frame shift mutation in the *MT-ND2* gene) and 11376delA (frame shift mutation in the *MT-ND4* gene)

Acknowledgement

We thank Miss Rui Bi for double-checking the data.

Supplemental References

1. Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet *23*, 147.
2. Behar, D.M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., Comas, D., et al. (2008). The dawn of human matrilineal diversity. Am J Hum Genet *82*, 1130-1140.
3. Kong, Q.-P., Bandelt, H.-J., Sun, C., Yao, Y.-G., Salas, A., Achilli, A., Wang, C.-Y., Zhong, L., Zhu, C.-L., Wu, S.-F., et al. (2006). Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet *15*, 2076-2086.
4. Palanichamy, M.g., Sun, C., Agrawal, S., Bandelt, H.-J., Kong, Q.-P., Khan, F., Wang, C.-Y., Chaudhuri, T.K., Palla, V., and Zhang, Y.-P. (2004). Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet *75*, 966-978.
5. Torroni, A., Achilli, A., Macaulay, V., Richards, M., and Bandelt, H.-J. (2006). Harvesting the fruit of the human mtDNA tree. Trends Genet *22*, 339-345.
6. van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat *30*, E386-394.
7. Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., Salas, A., Oppenheimer, S., Macaulay, V., and Richards, M.B. (2009). Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet *84*, 740-759.
8. Pereira, L., Freitas, F., Fernandes, V., Pereira, J.B., Costa, M.D., Costa, S., Máximo, V., Macaulay, V., Rocha, R., and Samuels, D.C. (2009). The diversity present in 5140 human mitochondrial genomes. Am J Hum Genet *84*, 628-640.
9. Brandstätter, A., Sänger, T., Lutz-Bonengel, S., Parson, W., Béraud-Colomb, E., Wen, B., Kong, Q.-P., Bravi, C.M., and Bandelt, H.-J. (2005). Phantom mutation hotspots in human mitochondrial DNA. Electrophoresis *26*, 3414-3429.
10. Sudoyo, H., Suryadi, H., Lertrit, P., Pramoonjago, P., Lyrawati, D., and Marzuki, S. (2002). Asian-specific mtDNA backgrounds associated with the primary G11778A mutation of Leber's hereditary optic neuropathy. J Hum Genet *47*, 594-604.
11. Ji, Y., Jia, X., Zhang, Q., and Yao, Y.-G. (2007). mtDNA haplogroup distribution in Chinese patients with Leber's hereditary optic neuropathy and G11778A mutation. Biochem Biophys Res Commun *364*, 238-242.
12. Fraumene, C., Belle E.M., Castrì L., Sanna S., Mancosu G., Cosso M., Marras F., Barbujani G., Pirastu M., and Angius A. (2006). High resolution analysis and phylogenetic network construction using complete mtDNA sequences in Sardinian genetic isolates. Mol Biol Evol *23*, 2101-2111.
13. Gasparre, G., Porcelli, A.M., Bonora, E., Pennisi, L.F., Toller, M., Iommarini, L., Ghelli, A., Moretti, M., Betts, C.M., Martinelli, G.N., et al. (2007). Disruptive mitochondrial DNA mutations in complex I subunits are markers of oncocytic phenotype in thyroid tumors. Proc Natl Acad Sci U S A *104*, 9001-9006.
14. Bayat, A., Walter, J., Lamb, H., Marino, M., Ferguson, M.W., and Ollier, W.E. (2005). Mitochondrial mutation detection using enhanced multiplex denaturing high-performance liquid chromatography. Int J Immunogenet *32*, 199-205.
15. Abu-Amero, K.K., Larruga, J.M., Cabrera, V.M., and González, A.M. (2008).

Mitochondrial DNA structure in the Arabian Peninsula. BMC Evol Biol 8, 45.
16. Ennafaa, H., Cabrera, V.M., Abu-Amero, K.K., González, A.M., Amor, M.B., Bouhaha, R., Dzimiri, N., Elgaaïed, A.B., and Larruga, J.M. (2009). Mitochondrial DNA haplogroup H structure in North Africa. BMC Genet *10*, 8.
17. Annunen-Rasila, J., Finnilä , S., Mykkänen, K., Pöyhönen, J.S., Veijola, J., Poyhonen, M., Viitanen, M., Kalimo, H., and Majamaa, K. (2006). Mitochondrial DNA sequence variation and mutation rate in patients with CADASIL. Neurogenetics *7*, 185-194.
18. Ingman, M., and Gyllensten, U. (2003). Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. Genome Res *13*, 1600-1606.
19. Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., et al. (2003). Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A *100*, 171-176.
20. Fagundes, N.J., Kanitz, R., Eckert, R., Valls, A.C., Bogo, M.R., Salzano, F.M., Smith, D.G., Silva, W.A., Jr., Zago, M.A., Ribeiro-dos-Santos, A.K., et al. (2008). Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. Am J Hum Genet *82*, 583-592.
21. Finnilä, S., Lehtonen, M.S., and Majamaa, K. (2001). Phylogenetic network for European mtDNA. Am J Hum Genet *68*, 1475-1484.
22. Ingman, M., Kaessmann, H., Pääbo, S., and Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. Nature *408*, 708-713.
23. Kong, Q.-P., Salas, A., Sun, C., Fuku, N., Tanaka, M., Zhong, L., Wang, C.-Y., Yao, Y.-G., and Bandelt, H.-J. (2008). Distilling artificial recombinants from large sets of complete mtDNA genomes. PLoS ONE *3*, e3016.
24. Behar, D.M., Metspalu, E., Kivisild, T., Rosset, S., Tzur, S., Hadid, Y., Yudkovsky, G., Rosengarten, D., Pereira, L., Amorim, A., et al. (2008). Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora. PLoS One *3*, e2062.
25. Feder, J., Blech, I., Ovadia, O., Amar, S., Wainstein, J., Raz, I., Dadon, S., Arking, D.E., Glaser, B., and Mishmar, D. (2008). Differences in mtDNA haplogroup distribution among 3 Jewish populations alter susceptibility to T2DM complications. BMC Genomics *9*, 198
26. Tanaka, M., Cabrera, V.M., González, A.M., Larruga, J.M., Takeyasu, T., Fuku, N., Guo, L.J., Hirose, R., Fujita, Y., Kurata, M., et al. (2004). Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res *14*, 1832-1850.
27. La Morgia, C., Achilli, A., Iommarini, L., Barboni, P., Pala, M., Olivieri, A., Zanna, C., Vidoni, S., Tonon, C., Lodi, R., et al. (2008). Rare mtDNA variants in Leber hereditary optic neuropathy families with recurrence of myoclonus. Neurology *70*, 762-770.
28. Kivisild, T., Shen, P., Wall, D.P., Do, B., Sung, R., Davis, K., Passarino, G., Underhill, P.A., Scharfe, C., Torroni, A., et al. (2006). The role of selection in the evolution of human mitochondrial genomes. Genetics *172*, 373-387.
29. Hartmann, A., Thieme, M., Nanduri, L.K., Stempfl, T., Moehle, C., Kivisild, T., and Oefner, P.J. (2009). Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. Hum Mutat *30*, 115-122.
30. Kumar, S., Padmanabham, P.B., Ravuri, R.R., Uttaravalli, K., Koneru, P., Mukherjee, P.A., Das, B., Kotal, M., Xaviour, D., Saheb, S.Y., et al. (2008). The

earliest settlers' antiquity and evolutionary history of Indian populations: evidence from M2 mtDNA lineage. BMC Evol Biol *8*, 230.