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. ©2009 by The American Society of Human Genetics. All rights reserved.

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. ©2009 by The American Society of Human Genetics. All rights reserved.