Letters to the Editor

Am. J. Hum. Genet. 72:1585–1586, 2003

Errors, Phantom and Otherwise, in Human mtDNA Sequences

To the Editor:

The good news is that a very large number of human mtDNA sequences from diverse populations and ethnic groups are becoming available for analysis. The bad news is that many of these sequences contain errors (Dennis 2003; Forster 2003). In at least one instance, that of the Icelandic population, it appears that mtDNA sequence errors were a contributing factor (although not the only one) to an erroneous conclusion about the genetic diversity of these people (Arnason 2003). Forster (2003) cites other examples where mtDNA sequence errors have compromised analyses of population genetics and human evolution. In a reanalysis of mtDNA sequences in the Ladin population of the Alps, the original conclusions on population diversity were not overturned after the use of more accurate sequences (Vernesi et al. 2002). At this point, we do not know the extent of the damage, so to speak, caused by mtDNA sequence errors. Nevertheless, it is clear that correcting such errors must be undertaken as quickly as possible.

As a result of our reduced median network analyses (Herrnstadt et al. 2002), we released a database of 560 human mtDNA coding region sequences. A small number of errors in these sequences were detected by Dr. Hans-Jürgen Bandelt, and we were able to correct these, as noted in an erratum that was published soon after our original report (Herrnstadt et al. 2002). Subsequently, a systematic approach to the detection of phantom sequence errors was published in this *Journal* (Bandelt et al. 2002). As defined by these investigators, phantom errors are those that arise during the sequencing process itself. Dr. Bandelt contacted us again and suggested that there were phantom mutations in our mtDNA database. Specifically, the likely errors involved G→C transversions at nt 7927 and nt 7985. Such a result was surprising to us, because we believed that our sequencing approach and quality control measures had avoided such errors. Therefore, we used Dr. Bandelt's information as a starting point for a comprehensive reanalysis of our database.

After reanalysis, which included inspection of the electropherograms for all G→C and C→G transversions, we found that 41 of these mtDNA sequences contained at least one such phantom error. In fact, there were more such phantom errors than those suggested by Dr. Bandelt. In addition to the phantom transversions at positions 7927 and 7985, we detected instances of other such errors that included ones at nucleotide positions 500, 14160, 14460, 14974, and 16239. However, these errors did not occur randomly throughout the database. Instead, we could "isolate" the errors to a short time period that was relatively early during our large-scale mtDNA sequencing program. With the benefit of hindsight, it appears that the frequency of these errors was caused by two technical factors (see also Bandelt et al. 2002). The first was that one particular capillary array of the ABI 3700 DNA Analyzer produced suboptimal base separations, whereas the second was that the sequencing chemistry at that time utilized an early version of reagents that was optimized subsequently.

In addition to these 41 sequences, we also found that an additional 26 mtDNA sequences contained errors that arose during data entry or editing. As a result of this reanalysis, we have corrected the database of 560 sequences, which is available through the MitoKor Web site (the URL address is given below).

Have these errors invalidated our network analyses? Not to a substantial degree. Many of the sequence errors generated private polymorphisms, which were not included in our analyses. Furthermore, a substantial proportion of the branches in these networks were established by multiple substitutions (see figs. 1–4 in Herrnstadt et al. 2002), and, so far, we have no evidence from additional network analysis that the original results need major revision. Can we now guarantee that our mtDNA database is error free? No. Although such is our goal, it is not practical, and it is probably not technically feasible.

It is now clear that many mtDNA databases or sequence sets contain errors (Forster 2003). The solution to this problem is further effort, both at the front end (the sequencing process itself) and at the back end (increased quality control) of mtDNA database construction.

Acknowledgments

We thank Dr. Hans-Jürgen Bandelt (University of Hamburg) for bringing the issue of mtDNA sequence errors to our at-

tention. The expert assistance of Brian Hulihan (MitoKor) with the Web site and with the files of mtDNA sequences is gratefully acknowledged.

CORINNA HERRNSTADT,[1] GWEN PRESTON,[1] AND NEIL HOWELL[1,2]

[1]*MitoKor, San Diego; and* [2]*Department of Radiation Oncology, The University of Texas Medical Branch, Galveston*

Electronic-Database Information

The URL for data presented herein is as follows:

MitoKor, http://www.mitokor.com/science/560mtdnasrevision.php (for the revised 560 mtDNA coding-region sequences; "zip" and "sit" files also available)

References

Arnason E (2003) Genetic heterogeneity of Icelanders. Ann Hum Genet 67:5–16

Bandelt H-J, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71:1150–1160

Dennis C (2003) Error reports threaten to unravel databases of mitochondrial DNA. Nature 421:773–774

Forster P (2003) To err is human. Ann Hum Genet 67:2–4

Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, Anderson C, Ghosh SS, Olefsky JM, Beal MF, Davis RE, Howell N (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am J Hum Genet 70:1152–1170 (erratum 71:448–449)

Vernesi C. Fuselli S, Castri L, Bertorelle G, Barbujani G (2002) Mitochondrial diversity in linguistic isolates of the Alps: a reappraisal. Hum Biol 74:725–730

Address for correspondence and reprints: Dr. Neil Howell, MitoKor, Inc., 11494 Sorrento Valley Road, San Diego, CA 92121. E-mail: howelln@mitokor.com

0002-9297/2003/7206-0024$15.00

Am. J. Hum. Genet. 72:1586–1590, 2003

South Asia, the Andamanese, and the Genetic Evidence for an "Early" Human Dispersal out of Africa

To the Editor:

The out-of-Africa model of anatomically modern human evolution posits an African origin 100,000–200,000 years ago, followed by subsequent dispersal(s) to Eurasia and other continents within the last 100,000 years (Stringer and Andrews 1988). Although alternative models have been proposed, the out-of-Africa scenario receives the most support both from archeological and genetic evidence (Lahr and Foley 1994). However, the route(s) followed by the African migrants remain poorly understood. One proposed route was through northern Africa toward the Levant, which finds support in the archeological and fossil records (Lahr and Foley 1994). This exit of modern humans out of Africa would have taken place during the Upper Paleolithic era (~45,000 years ago), which considerably postdates the earliest evidence of modern human presence in the Sahul. Indeed, luminescence dating, paleovegetation changes, and skeletal remains suggest that Australia was inhabited by modern humans by 60,000 years ago (Roberts and Jones 1994; Johnson et al. 1999; Miller et al. 1999; Thorne et al. 1999), implying a substantially earlier migration from Africa to Australia. To take this evidence into account, as well as morphological and archeological features of many Australian fossils, a second migration of modern humans, known as the "southern route" hypothesis, was suggested to have occurred during Middle Paleolithic times (60,000–100,000 years ago) from eastern Africa to Sahul via South Asia (Cavalli-Sforza et al. 1994; Lahr and Foley 1994).

In the January 2003 issue of the *Journal,* Endicott et al. (2003) investigated the genetic affinities of 11 Andaman islanders, a group of people in the Indian Ocean with phenotypic similarities to some African populations (i.e., "Negrito" features) and reputed to be possible descendants of early migrants out of Africa to Sahul, following the southern route. The authors claim that the results of their investigation "support the growing evidence of an early movement of humans through southern Asia." In our opinion, Endicott and colleagues' results do not support any relationship between the present Andamanese population and the hypothesized early southern migration. The authors identified three different mtDNA haplotypes in 11 Andaman islanders, two belonging to haplogroup M2 and one belonging to M4. These haplogroups had previously been reported only in the Indian subcontinent (Kivisild et al. 1999*b*; Bamshad et al. 2001). The Andaman M4 haplotype has been found previously in mainland India (Kivisild et al. 1999*b*), whereas the two Andaman M2 haplotypes are (so far) unique to the Andamanese. Given that (1) the latter two types occupy a basal position in the M2 network, which has an estimated coalescence time of 63,000 ± 6,000 years (Kivisild et al. 1999*b*), and (2) they are not found in mainland India, Endicott et al. (2003) conclude they represent an "early" settlement of the Andaman Islands. These two points need discussion.

Regarding point 1, the age of a haplogroup cannot be automatically equated to the age of subsets of this haplogroup. The founding type of haplogroup M2, characterized by 16223T and 16319A relative to the Cambridge reference sequence (CRS) (Anderson et al. 1981) (fig. 1),