THE ROYAL SOCIETY

Mitochondrial DNA recombination—no need to panic

Vincent Macaulay*, Martin Richards and Bryan Sykes

Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK

Recombination has recently been invoked as an explanation for the large amount of homoplasy observed in a collection of complete or nearly complete human mitochondrial sequences. Here we show that some of the data on which this conclusion was based are likely to be unreliable and that if these data are excluded, the results are no longer significant.

Keywords: mitochondria; recombination; errors in data

In a recent article in this journal (Eyre-Walker *et al.* 1999), a large amount of homoplasy at the third positions of the protein-coding segments of human mitochondrial DNA (mtDNA) was inferred. Having put forward arguments dispensing with a number of other possible mechanisms that might account for their observation, the authors came to the conclusion that such high levels of homoplasy could only be the result of recombination affecting what had been thought of as a molecule with clonal inheritance. The implications of large amounts of recombination for human evolutionary studies based on mtDNA phylogenies would be considerable. However, we think the explanation of the signal detected lies elsewhere.

It appears that one of the data sets used by Eyre-Walker *et al.* (1999), which accounts for ten out of their 29 sequences, contains a number of errors. This is the data reported by Marzuki *et al.* (1991). First, an error has been introduced in the transcription of this data set. At nucleotide position 4985 an adenine appears in the nine sequences in which this position was sequenced in figure 2 of Marzuki *et al.* (1991), but it is reported as a guanine in table 1 of Eyre-Walker *et al.* (1999). This site is one which appears as a rare variant or, possibly, as a mistake in the Cambridge reference sequence (CRS) of human mtDNA (Anderson *et al.* 1981), as carefully discussed by Howell *et al.* (1992). Second, it seems that table 1 of Marzuki *et al.* (1991) itself displays examples of confusion caused by scoring other sites of this kind, which directly affect the Eyre-Walker *et al.* (1999) study. Specifically, nucleotide positions 3423 and 11 335 are thought to fall in this category and in other data sets have almost always been found with thymine and cytosine, respectively, contra the CRS (Howell *et al.* 1992). In the Marzuki *et al.* (1991) data they are reported as polymorphic, with guanine and thymine, respectively, in the majority. Contrast this with a recent study by Hofmann *et al.* (1997), where these positions were found fixed contra the CRS in 67 Germans. This latter study encompasses much of West Eurasian mtDNA variation (Macaulay *et al.* 1999) and represents a good comparison with the Marzuki *et al.* (1991) data since seven out of those ten sequences can be assigned to West Eurasian clusters on the basis of characteristic sequence motifs (CMD-2, SVR87-2 and SVR89-1 to cluster H, SVR88-1 to cluster U, SVR84-3 to cluster K, SVR88-3 to cluster I and SVR89-2 to cluster J). Finally, in a reply to Howell *et al.* (1992), Marzuki *et al.* (1992) alluded to positions not in fact sequenced in their study which were not indicated in the original data table. This leads to problems, for example, at nucleotide position 14 365, another site which is a rare variant in the CRS.

Given the problems that seem likely to affect the ten Marzuki *et al.* (1991) sequences, what picture emerges if we focus on the remaining 19? Only 26 of the phylogenetically informative characters from table 1 of Eyre-Walker *et al.* (1999) survived in this subset. These admit a single most parsimonious tree (figure 1) of length 30 steps, hence with only four recurrent mutations (at nucleotide positions 4985, 5147, 8071 and 10 915) compared to 22 in the full data set. In order to confirm that this dramatic reduction can be attributed to discarding these particular ten sequences, we left out a randomly chosen set of ten sequences from the full data and evaluated the number of homoplasies in a putative most parsimonious tree constructed from the remainder, using the heuristic option of PAUP (Swofford 1993); we repeated this operation 24 times. The number of recurrent mutations ranged from seven to 17 with a mean of 11.9 and a standard deviation of 2.4, fully demonstrating the anomaly of the Marzuki *et al.* (1991) data.

Is it necessary to invoke recombination in order to explain this much lower level of homoplasy? In order to test this, we asked, as Eyre-Walker *et al.* (1999) did, how much parallel mutation would be expected (given the number of sites assayed and the number observed to be polymorphic) under the simplest model of evolution, in which mtDNA evolves clonally and the synonymous mutations at third positions occur at the same rate at every site. In the process of discarding the ten Marzuki

*Author for correspondence (vincent.macaulay@cellsci.ox.ac.uk).

Received 14 May 1999 *Accepted* 9 June 1999

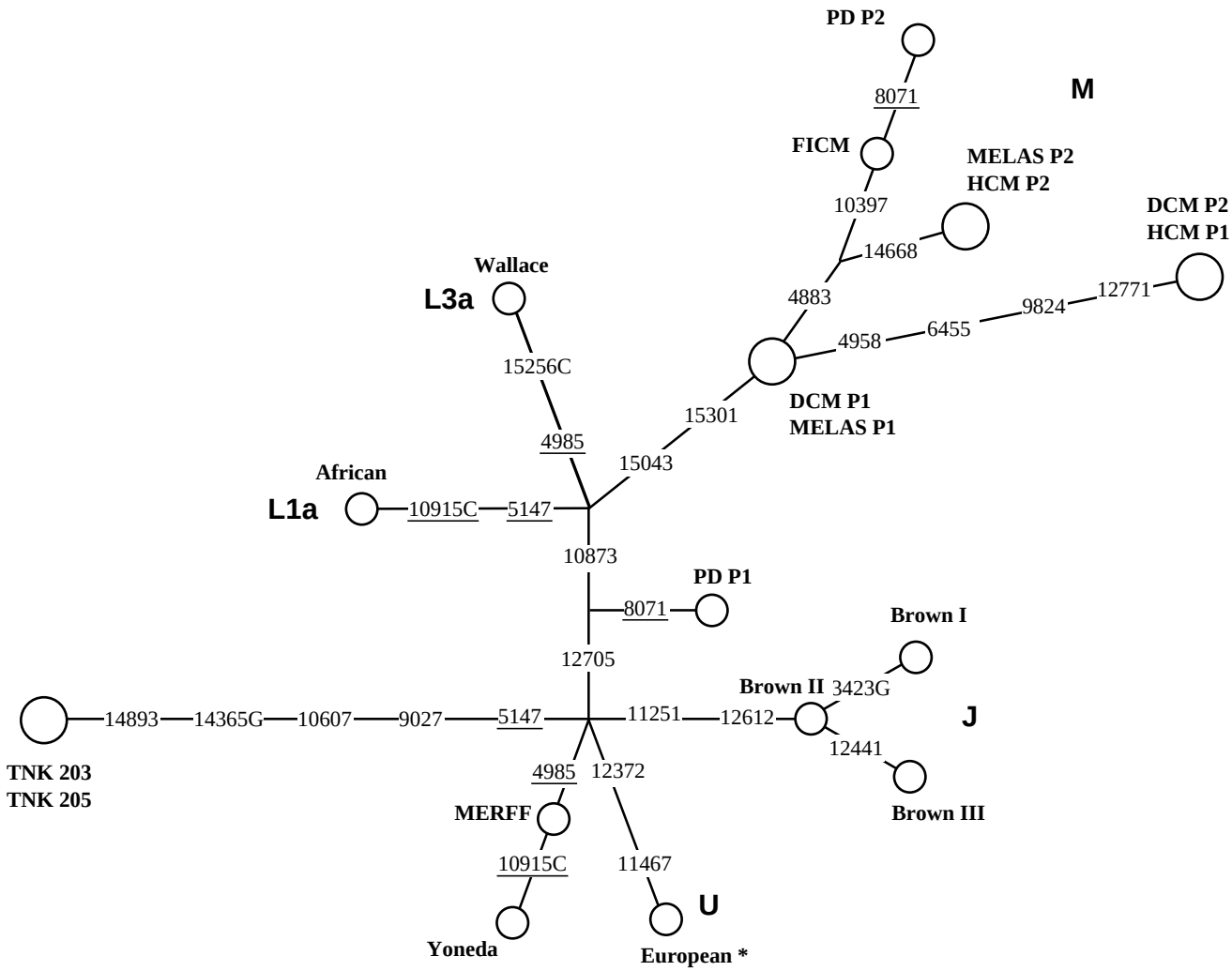

Figure 1. Reduced median network (Bandelt *et al*. 1995) for the Eyre-Walker *et al*. (1999) data, excluding the ten Marzuki *et al*. (1991) sequences. The network was constructed using the program Network (Röhl 1997). The circles represent sequences, with the area proportional to the frequency and the links represent mutations from the sequence labelled 'European' (not the CRS), which is marked with an asterisk. Transversions are specified. Cluster designations are included where motif positions can be identified. The reduced median network is also the unique most parsimonious tree, of length 30 steps and includes four homoplasies (indicated by underlining of the positions involved).

et al. (1991) sequences, 22 sites cease to be polymorphic, so the 126 polymorphisms in the 3628 third positions reported in Eyre-Walker *et al*. (1999) are reduced to 104. Under the simple clonal model, 1.5 homoplasies (with a standard deviation of 1.3) are expected. There is a probability of 7% that four or more homoplasies would be observed: the model cannot be rejected at the 5% level. It seems that, for the time being at least, we need look for no more exotic explanation of the data than this.

Recombination encompassing sufficiently large segments of DNA should be manifest as grid-like topologies in phylogenetic network diagrams, in particular in the median network of the data set (Bandelt *et al*. 1995). We constructed both full and reduced median networks of the Marzuki *et al*. (1991) data set alone, excluding three positions where missing data could not be unambiguously accommodated. The more easily visualized reduced median network is shown in figure 2. Interestingly, the networks indeed show the grid-like pattern characteristic of recombination, involving nucleotide positions 12 399 and 12 441. Nevertheless, in view of the strong probability of errors in the Marzuki *et al*. (1991) data, coupled with the fact that variants at these positions are rare, having yet to be recorded in other data sets (e.g. Hofmann *et al*. 1997) and with the lack of such grid-like patterns in more extensive West Eurasian coding-region data sets (e.g. Torroni *et al*. 1994, 1996; Hofmann *et al*. 1997; Macaulay *et al*. 1999), we should be wary of jumping to conclusions: indeed it is suspicious that the two samples appear to be adjacent in the Marzuki *et al*. (1991) labelling scheme. Recombination on the page or in the database would seem more likely than recombination in the cell.

We thank Adam Eyre-Walker, John Maynard Smith and Hans-Jürgen Bandelt for critical readings of the manuscript and Alan Cooper, Andrew Rambaut, Paul Harvey and members of the University of Oxford Human Population Genetics Journal Club for helpful discussions.

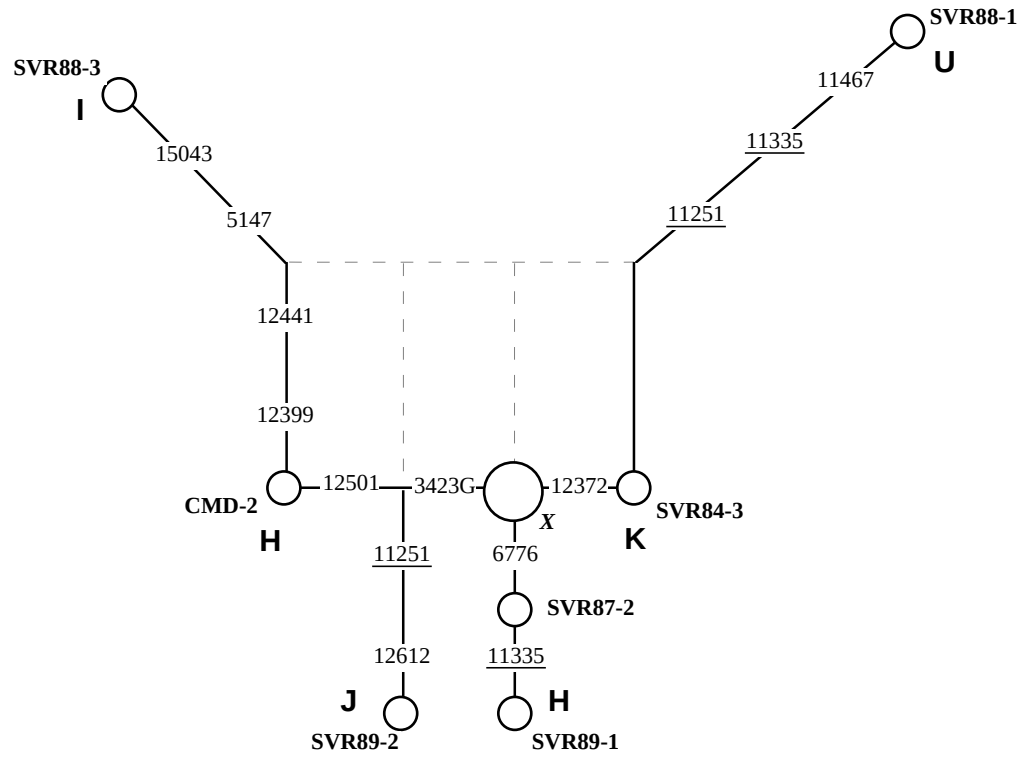

Figure 2. Reduced median network for the Marzuki *et al*. (1991) data alone. Characters with missing data were included if all the most parsimonious trees recovered by a PAUP heuristic search were unanimous in their choice of state, leaving only three characters with missing data (nucleotide positions 7028, 12 705 and 13 368) which were omitted. The node marked *X* contains sequences CMD-1, SVR86-1 and SVR88-4 and departs from the 'European' sequence at nucleotide positions 3423, 11 335, 11 467, 12 372 and 15 256. Within the network the unique most parsimonious tree is indicated by the solid lines. The network clearly displays an apparent recombination event involving nucleotide positions 12 399 and 12 441, occurring between the adjacently labelled Marzuki *et al*. (1991) sequences SVR88-1 and SVR88-3.

REFERENCES

Anderson, S. (and 13 others) 1981 Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465.

Bandelt, H.-J., Forster, P., Sykes, B. C. & Richards, M. B. 1995 Mitochondrial portraits of human populations using median networks. *Genetics* **141**, 743–753.

Eyre-Walker, A., Smith, N. H. & Maynard Smith, J. 1999 How clonal are human mitochondria? *Proc. R. Soc. Lond.* B **266**, 477–483.

Hofmann, S., Jaksch, M., Bezold, R., Mertens, S., Aholt, S., Paprotta, A. & Gerbitz, K. D. 1997 Population genetics and disease susceptibility: characterization of central European haplogroups by mtDNA gene mutations, correlations with D loop variants and association with disease. *Hum. Mol. Genet.* **6**, 1835–1846.

Howell, N., McCullough, D. A., Kubacka, I., Halvorson, S. & Mackey, D. 1992 The sequence of human mtDNA: the question of errors versus polymorphisms. *Am. J. Hum. Genet.* **50**, 1333–1337.

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonné-Tamir, B., Sykes, B. & Torroni, A. 1999 The emerging tree for West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am. J. Hum. Genet.* **64**, 232–249.

Marzuki, S., Noer, A. S., Lertrit, P., Thyagarajan, D., Kapsa, R., Utthanaphol, P. & Byrne, E. 1991 Normal variants of human mitochondrial DNA and translation products: the building of a reference data base. *Hum. Genet.* **88**, 139–145.

Marzuki, S., Lertrit, P., Noer, A. S., Kapsa, R. M. I., Sudoyo, H., Byrne, E. & Thyagarajan, D. 1992 Reply to Howell *et al.*: the need for a joint effort in the construction of a reference data base for normal sequence variants of human mtDNA. *Am. J. Hum. Genet.* **50**, 1337–1340.

Röhl, A. 1997 *Network: a program package for calculating phylogenetic networks*. Hamburg: Mathematisches Seminar, University of Hamburg.

Swofford, D. L. 1993 *PAUP: phylogenetic analysis using parsimony*. Champaign, IL: Illinois Natural History Survey.

Torroni, A., Lott, M. T., Cabell, M. F., Chen, Y. S., Lavergne, L. & Wallace, D. C. 1994 mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. *Am. J. Hum. Genet.* **55**, 760–776.

Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., Obinu, D., Savontaus, M.-L. & Wallace, D. C. 1996 Classification of European mtDNAs from an analysis of three European populations. *Genetics* **144**, 1835–1850.