THE ROYAL SOCIETY

How clonal are human mitochondria?

Adam Eyre-Walker*, Noel H. Smith and John Maynard Smith

Centre for the Study of Evolution and School of Biological Sciences, University of Sussex, Brighton BN1 9QG, UK

Phylogenetic trees constructed using human mitochondrial sequences contain a large number of homoplasies. These are due either to repeated mutation or to recombination between mitochondrial lineages. We show that a tree constructed using synonymous variation in the protein coding sequences of 29 largely complete human mitochondrial molecules contains 22 homoplasies at 32 phylogenetically informative sites. This level of homoplasy is very unlikely if inheritance is clonal, even if we take into account base composition bias. There must either be 'hypervariable' sites or recombination between mitochondria. We present evidence which suggests that hypervariable sites do not exist in our data. It therefore seems likely that recombination has occurred between mitochondrial lineages in humans.

Keywords: mitochondria; recombination; clonal

1. INTRODUCTION

It is generally accepted that, in higher plants and animals, mitochondria are inherited from one parent, usually the mother, and that their inheritance is therefore clonal. This dogma appears to have arisen out of the belief that paternal mitochondria do not penetrate the egg (Ankel-Simons & Cummins 1996). However, it is now well-known that paternal mitochondria do enter the egg and survive for several hours in mammals (Kaneda *et al*. 1995; Ankel-Simons & Cummins 1996). Since mitochondria also contain the enzymes necessary for homologous recombination (Thyagarajan *et al*. 1996; Lunt & Hyman 1997), it seems possible that there is recombination between mitochondrial lineages and that the inheritance of mitochondria is not clonal.

Phylogenetic trees constructed using mitochondrial DNA (mtDNA) sequences contain many homoplasies, i.e. back or parallel changes at a site (e.g. see Vigilant *et al*. 1991). Homoplasies are either due to repeated mutation or to recombination. If the homoplasies are due to mutation then it is necessary to hypothesize that some sites change more often than others, i.e. that there are 'hypervariable' sites (Hasegawa *et al*. 1993; Wakeley 1993). Most of the analysis of mitochondrial DNA within species has focused on the control region, a sequence of *ca*. 1000 bp, which is involved in the control of DNA replication and transcription. Some sites are expected to change more often than others in the mitochondrial control region for two reasons: first, some sites are constrained by function (Sbisa *et al*. 1997) and, second, there is strong base composition bias within the control region. Base composition bias leads to variation in the rate of change for the following reason. Imagine that there are just two bases, C and T and that the frequency of C in a sequence is 90%. This implies that a T is nine times more likely to be replaced by a C than a C by a T. So, although over the long term, each site may have the same average rate of change, over the short term, sites occupied by T are much more likely to change than sites occupied by C. It is therefore unsurprising that Jazin *et al*. (1998) recently found evidence of variability in the rate at which new mutations appear at sites in the mitochondrial control region within human families. However, the question remains whether constraint and base composition bias are sufficient to explain the level of homoplasy within human mitochondrial sequences, or whether there must be additional sources of variability in the rate of change or recombination. This is the question we address here.

Although sites are constrained within the control region, it is not easy to identify them. It is therefore difficult to assess whether the level of homoplasy within the control region can be explained by a combination of constraint and base composition bias. We have, therefore, chosen to concentrate on variation at synonymous sites in the protein-coding regions of human mitochondrial sequences.

2. MATERIALS AND METHODS

We collected 29 largely complete mtDNA sequences from Genbank (human 'European', X93334 and human 'African', D38112) and the literature (DCM_P1–FICM (Ozawa *et al*. 1991), BrownI and BrownII (Brown *et al*. 1992*a*), BrownIII (Brown *et al*. 1992*b*), TNK203 and TNK205 (Kobayashi *et al*. 1991), Yoneda (Yoneda *et al*. 1990), CMD-1–SVR89-2 (Marzuki *et al*. 1991) and Wallace (Wallace *et al*. 1988)). The 'Cambridge' sequence was not included since it is a chimera from two different individuals (Arnason *et al*. 1996). We also used three complete chimpanzee (X93335, D38113 and D38116), two gorilla (D38114 and X93347), two orang-utan (D38115 and X97707) and one gibbon (X99256) sequences in our analyses. The non-overlapping protein-coding sequences, excluding stop codons, were extracted from each sequence. The *ND6* gene was reversed and complemented for most of the analyses since this is on the L strand, whilst all other protein-coding genes are on the H strand.

We also used the following *Pan troglodytes* sequences: 20 *ND2* sequences (Wise *et al* 1998), four complete cytochrome *b*

*Author for correspondence (a.c.eyre-walker@sussex.ac.uk).

Received 2 October 1998 *Accepted* 9 November 1998

Table 1. *Informative sites*

(The table gives the third position sites at which there are phylogenetically informative synonymous polymorphisms segregating in 29 human mitochondrial sequences. A dot indicates identity to the European sequence, a dash indicates missing information. Numbering is as in the 'Cambridge' sequence (Anderson *et al.* (1981).)

position in sequence	3	4	4	4	5	6	6	7	8	9	9	1 0	1 0	1 0	1 0	1 1	1 1	1 1	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 3	1 4	1 4	1 4	1 5	1 5	1 5
	4	8	9	9	1	4	7	0	0	0	8	3	6	8	9	2	3	4	3	3	4	5	6	7	7	3	3	6	8	0	2	3
	2	8	5	8	4	5	7	2	7	2	2	9	0	7	1	5	3	6	7	9	4	0	1	0	7	6	6	6	9	4	5	0
	3	3	8	5	7	5	6	8	1	7	4	7	7	3	5	1	5	7	2	9	1	1	2	5	1	8	5	8	3	3	6	1
European	T	C	A	A	G	C	T	T	A	C	T	A	C	T	T	A	C	G	A	C	T	G	A	C	G	G	C	C	A	G	A	G
African	.	.	.	.	A	.	.	.	.	.	.	.	.	C	C	.	.	A	G	.	.	.	.	T	.	.	.	.	.	.	.	.
DCM_P1	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	A	G	.	.	.	.	T	.	.	.	.	.	A	.	A
DCM_P2	.	.	G	.	.	T	.	.	.	.	C	.	.	C	.	.	.	A	G	.	.	.	.	T	A	.	.	.	.	A	.	A
MERFF	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	A	G	.	.	.	.	.	.	.	.	.	.	.	.	.
PD_P1	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	A	G	.	.	.	.	T	.	.	.	.	.	.	.	.
PD_P2	.	T	.	.	.	.	.	.	G	.	.	G	.	C	.	.	.	A	G	.	.	.	.	T	.	.	.	.	.	A	.	A
HCM_P1	.	.	G	.	.	T	.	.	.	.	C	.	.	C	.	.	.	A	G	.	.	.	.	T	A	.	.	.	.	A	.	A
HCM_P2	.	T	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	A	G	.	.	.	.	T	.	.	.	T	.	A	.	A
MELAS_P1	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	A	G	.	.	.	.	T	.	.	.	.	.	A	.	A
MELAS_P2	.	T	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	A	G	.	.	.	.	T	.	.	.	T	.	A	.	A
FICM	.	T	.	.	.	.	.	.	.	.	.	G	.	C	.	.	.	A	G	.	.	.	.	T	.	.	.	.	.	A	.	A
BrownI	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	G	.	.	.	G	.	.	.	.	.	.	.	.	.
BrownII	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	G	.	.	.	G	.	.	.	.	.	.	.	.	.
BrownIII	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	G	.	C	.	G	.	.	.	.	.	.	.	.	.
TNK203	.	.	.	.	A	.	.	.	.	G	.	.	T	.	.	.	.	A	G	.	.	.	.	.	.	.	G	.	G	.	.	.
TNK205	.	.	.	.	A	.	.	.	.	G	.	.	T	.	.	.	.	A	G	.	.	.	.	.	.	.	G	.	G	.	.	.
Yoneda	.	.	.	G	.	.	.	.	.	.	.	.	.	.	C	.	.	A	G	.	.	.	—	.	.	.	.	.	.	.	.	.
CMD-1	G	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	T	A	G	.	.	.	.	T	.	.	G	.	.	.	—	—
CMD-2	.	.	.	G	.	.	.	C	—	.	—	.	.	.	.	.	T	A	G	.	.	A	.	.	.	.	G	.	.	.	C	.
SVR84-3	G	—	—	—	—	.	.	.	.	.	.	.	.	.	.	.	T	A	.	.	.	.	—	—	—	—	—	—	—	—	—	—
SVR86-1	G	.	.	G	.	.	.	—	—	—	.	.	.	.	.	.	T	A	G	.	.	.	.	.	.	.	—	—	—	—	—	—
SVR87-2	G	.	.	G	.	.	C	C	.	.	.	.	.	.	.	.	T	A	G	.	.	.	.	.	.	A	—	—	—	.	—	—
SVR88-1	G	.	.	G	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	T	C	.	—	—	—	—	—	—	—	—	—	—
SVR88-3	.	.	.	G	A	.	.	.	.	.	.	.	.	.	.	.	T	A	G	T	C	A	.	T	.	.	—	—	—	A	—	—
SVR88-4	G	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	T	A	G	.	.	.	.	.	.	A	—	—	—	—	—	—
SVR89-1	G	.	.	G	.	.	C	C	.	.	.	.	.	.	.	.	.	A	G	.	.	.	—	—	—	—	—	—	—	—	—	—
SVR89-2	.	.	.	G	.	.	.	—	—	—	.	.	.	.	.	G	T	A	G	.	.	.	G	.	.	.	G	.	.	.	—	—
Wallace	.	.	.	G	.	.	.	.	—	.	.	.	.	C	.	.	.	A	G	.	.	.	.	.T	.	.	—	—	.	.	C	—

sequences (X93335, D38113, X93338 and X93339), 33 partial cytochrome *b* sequences (L-35346–35361 and L35363–L35379), four *ND4* sequences (X93335, D38113, X93344 and X93345), five *COII* partial sequences (X93335, D28113, U12697, U12705 and U12706) and four partial *ND5* sequences (X93335, D38113, X93343 and X93346).

In all analyses, only synonymous variation segregating in the third position of codons, which were otherwise invariant, were considered. All parsimony analyses were performed in PAUP, v. 3.0, using a heuristic search with 'closest' addition and 'tree bisection reconnection' branch swapping. To calculate the number of homoplasies expected if reproduction is clonal we used the amino acid frequencies and codon usage of the European sequence (X93334). To incorporate the (conservative) assumption that only transitions occur, fourfold redundant amino acids were treated as two distinct amino acids: in the case of valine, for example, one coded for by GTA or GTG and one by GTC or GTT. Sixfold redundant amino acids were treated as three distinct amino acids.

3. RESULTS

In our sample of 29 largely complete human mitochondrial sequences, there were 126 synonymous polymorphisms segregating at 3628 third sites, of which 32 were phylogenetically informative. The informative sites are shown in table 1. The most parsimonious trees constructed using the informative sites that we could find had 54 steps and, therefore, 22 homoplasies. In total, there were 1568 most parsimonious trees.

The expected number of mutational events needed to generate 126 polymorphisms in 3628 third sites is 128.2 if the inheritance of mitochondria is clonal and each site is equally likely to change. Hence, we would expect 2.2 homoplasies; the probability of getting 22 is effectively zero.

There are, however, two sources of variation in the rate of change between sites that can increase the number of homoplasies: selective constraint at a proportion of the sites and base composition bias. To test whether there is constraint at some synonymous sites in primate mitochondria we considered the degree to which human synonymous sites are saturated with substitutions relative to other primates; if there is no constraint at synonymous sites, the observed divergence should (eventually) approach the expected divergence calculated from the base composition of the sites in question (Maynard Smith & Smith 1996). We concentrated on twofold degenerate

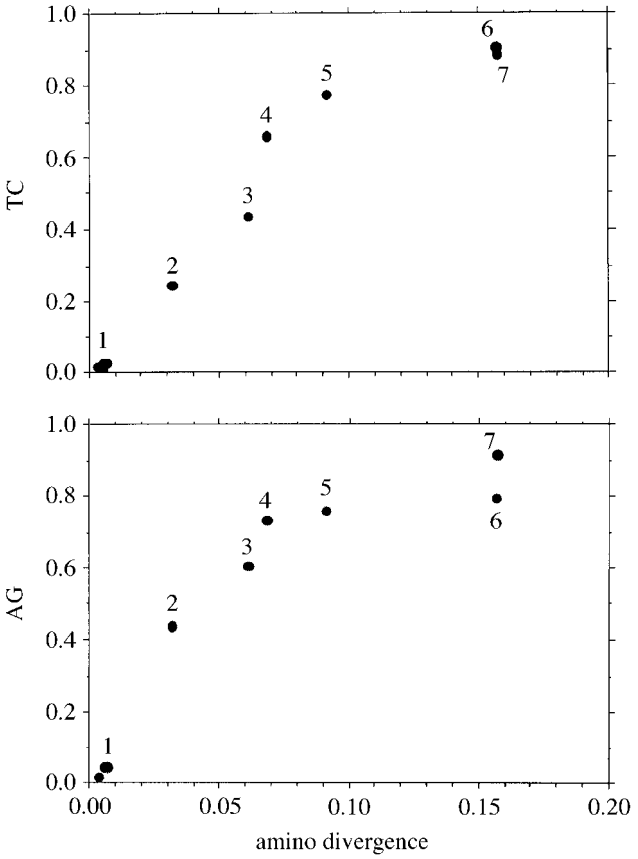

Figure 1. Observed over expected divergence (at saturation) plotted against amino acid divergence for TC- (top) and AG- (bottom) ending twofold degenerate codons. *ND6* was excluded from this analysis. 1. Within humans. 2. Common versus bonobo chimpanzee. 3. Sumatran versus Bornean orang-utan. 4. Human versus common chimpanzee. 5. Human versus gorilla. 6. Human versus orang-utan. 7. Human versus gibbon.

codons because the rate of transitional change is much greater than the rate of transversional change in mammalian mitochondria. When twofold degenerate codons become saturated, the expected divergence is

$$D_{T/C} = 1 - T_1T_2 - C_1C_2,$$
$$D_{A/G} = 1 - A_1A_2 - G_1G_2, \quad (1)$$

for T/C-ending and A/G-ending codons, respectively, where X_y is the frequency of nucleotide X in species y. If there is constraint, the observed divergence will be $(1-\delta)D$, where δ is the proportion of sites which are constrained and now X_y is the frequency in the sites which can vary.

The observed divergence, divided by the expected divergence at saturation, is plotted against the amino acid divergence in figure 1 for a variety of primates. The amino acid divergence serves as a surrogate measure of time. There is little indication that some synonymous sites are selectively constrained; for both T/C- and A/G-ending codons the trend seems to be such that the expected divergence will eventually be reached.

Synonymous codon use in mitochondria is biased (Perna & Kocher 1995); this leads to differences in the rate of change between sites because sites occupied by a rare nucleotide change more often than sites occupied by a common nucleotide. To take this into account in our calculation of the expected number of homoplasies, we calculated the effective site number, S_e: $S_e = 1/p$, where p is the probability that if two random changes occur in a sequence, they occur at the same site (Maynard Smith & Smith 1998). In the extreme case, in which each amino acid is coded for by only one codon, $S_e = 0$. More generally, if the frequency of the common codon is f, then at equilibrium $S_e = 4f(1-f)$. A second difficulty arises because transitions are more frequent than transversions. In the human data set this is a large effect: 113 out of 126 variable sites and 28 out of 32 informative sites are polymorphic for a transition. We have therefore estimated S_e on the assumption that only transitions can occur: if, as is the case, a few transversions also happen, this will lead us to overestimate p and, hence, to underestimate S_e. This is conservative, in that it leads us to overestimate the number of homoplasies expected if reproduction is clonal. Taking into account that the bias is different for different amino acids, we estimate that $S_e = 2269$ in our data set or 62% of the sites available.

Given an effective site number of 2269, the expected number of random hits needed to generate 126 polymorphisms is 129.63, implying 3.63 homoplasies, which is in sharp contrast to the observed number, $h_o = 22$. The probability of 22 homoplasies if inheritance is clonal is close to zero. To give a 1% chance of $\geqslant 22$ homoplasies, given 126 polymorphisms, requires $S_e \sim 700$ or only 19% of the sites available.

The homoplasies must therefore be due to recombination or to variation in the rate of change which cannot be attributed to base composition bias. We will refer to sites which might have a rate of change above that due to base composition bias as 'hypervariable' sites. Hypervariability can be produced by both mutation and selection; however, since there is no evidence of selection upon synonymous codon use in human mitochondria (figure l; A. Eyre-Walker, unpublished data) it seems more likely that it would be due to hypermutation. We need to differentiate between two types of hypermutation: sites at which one particular type of mutation is elevated (e.g. C→T) and sites at which several types of change are increased simultaneously (e.g. C→T and T→C). It seems most likely that hypermutation will be in one direction. However, this raises a problem: if the mutation rate in one direction is very high we would expect the site to be in the non-hypermutable state, e.g. if C→T is very frequent then we expect the site to be occupied by T most of the time. To estimate the number of sites which would have to be hypervariable to explain our data and the degree to which they would have to mutate faster than normal sites, we fitted a model, in which sites were either normal or hypermutable, to the number of sites which were polymorphic and homoplasic (see Appendix A). There are a range of solutions, examples of which are given in table 2. Two conclusions are evident. First, a substantial fraction of sites would have to be potentially hypermutable, though most would be in the non-hypermutable state. Second, sites which were in the hypermutable state would have to mutate more than 30 times faster than normal sites.

If a substantial fraction of sites are hypermutable (in one direction) then this will manifest itself in the level of

Table 2. *Number of hypermutable sites*

(The table gives the rate of change and the number of hypermutable sites required to explain the data when hypermutable sites change k times more rapidly than normal sites when they are in the hypermutable state. The total number of hypermutable sites is also given. See Appendix A for details.)

k	rate of change (u)	no. of hypermutable sites in hypermutable state	total no. of hypermutable sites
31	7.03×10^{-4}	116	3596
40	8.88×10^{-4}	54	2160
50	9.57×10^{-4}	36	1800
100	1.06×10^{-3}	17.5	1750

divergence at saturation; the divergence will be below that expected from the base composition. Let us consider twofold degenerate sites and imagine that a fraction, α, of the sites are hypermutable. Let the frequency of sites in state 1 be f_m at hypermutable sites and f_n at normal sites, e.g. the frequency of T might be f_m in the hypermutable sites of T/C-ending codons. The proportion of sites which differ between two sequences at saturation is

$$D = 2\alpha f_m(1 - f_m) + 2(1 - \alpha) f_n(1 - f_n), \quad (2)$$

but the level calculated from the base composition is

$$D^* = 2f(1 - f), \quad (3)$$

where

$$f = \alpha f_m + (1 - \alpha) f_n.$$

It is not difficult to show that $D < D^*$. The inequality also holds under more complex models where there are many different types of site, in which state 1 is favoured at some sites, while state 2 is favoured at others (A. Eyre-Walker, unpublished data). However, the divergence at twofold degenerate codons between humans and gibbons or orang-utans appears to be very close to the level you would expect at saturation given the codon bias (figure 1). There is therefore no evidence that hypermutable sites, in which the mutation rate is substantially higher in one direction, exist; a substantial fraction of the sites would have to be hypermutable to explain the level of homoplasy in our data and those sites would manifest themselves as a low level of divergence at saturation.

It is possible that there are hypermutable sites in which several directions of change are elevated simultaneously (e.g. C→T and T→C). If such sites existed, then the hypermutable sites should tend to vary in both humans and other primates (note that since the rate of non-synonymous substitution is relatively low in primate mitochondria the context of sites tends to be preserved across taxa). However, if hypermutable sites do not exist, the proportion of sites which vary in both humans and other primates (W) is expected to be the product of the proportion which vary in humans (W_h) and the proportion which vary in the other data set (W_o), if the two groups of sequences are saturated with respect to one another, e.g. $W = W_h W_o$. If the two groups of sequences are not saturated, then the overlap is expected to be greater than $W_h W_o$. Most of the variation segregating both in humans and within our primate sample are transitions. Figure 1 shows that even the divergence between humans and chimpanzees is approaching saturation with respect to transitions.

Table 3. *Homoplasic sites*

(The table gives the sites which are homoplasic in one of 20 randomly chosen most parsimonious trees, along with the average number of changes at each site and the context in which the site occurs. The colon represents the homoplasic site.)

position	average no. of hits	H strand context	L strand context
3423	2.75	CGT : GTA	TAC : ACG
4985	3.15	CCA : ACC	GGT : TGG
5147	3.00	CAC : ACC	GGT : GTG
7028	2.00	AGC : CAC	GTG : GCT
8071	2.00	CTC : TGA	TCA : GAG
10873	1.35	CCC : CTA	TAG : GGG
10915	2.00	CTG : TCC	GGA : CAG
11251	2.55	ACT : ATT	AAT : AGT
11335	2.15	CAA : AAC	GTT : TTG
11467	1.60	CTT : AAA	TTT : AAG
12372	2.55	CCT : ACT	AGT : AGG
12399	1.90	TAC : ACC	GGT : GTA
12441	2.65	TTA : GTA	TAC : TAA
12501	1.10	CAT : TGC	GCA : ATG
12612	2.00	TGT : GCA	TGC : ACA
12705	3.25	CAT : TTC	GAA : ATG
14668	1.35	ATA : ATC	GAT : TAT
15043	2.00	CGG : CGA	TCG : CCG
15256	1.65	AGT : GAC	GTC : ACT

We have performed two analyses. In the first we compared the sites which are polymorphic in humans to those that vary within chimpanzees, gorillas and orang-utans. In the second, we compared the sites which are homoplasic in the most parsimonious trees of the human data to sites which vary in another primate. The sites which are homoplasic in one out of 20 randomly chosen most parsimonious trees are shown in table 3. The results of the overlap analysis are given in table 4. There is no evidence that hypervariable sites exist; the observed number of sites which are polymorphic in humans and variable in another primate is very similar to the number expected without hypervariable sites (table 4). For example, there are 1343 codons available for analysis where we have four or more *P. troglodytes* sequences: for these there are 46 sites variable in humans and 86 sites variable in chimps and the expected overlap is 2.9 compared to the observed value of four. If we restrict the analysis to sites which are homoplasic in humans, there is one site which is polymorphic in both humans and chimpanzees; this is very close to the expected value of 0.9 sites.

Hypermutable sites often occur in particular contexts, for example CpG dinucleotides in mammals (Bulmer 1986; Bird 1987) and CCAGG sites in *Escherichia coli* (Coulondre *et al*. 1978; Halliday & Glickman 1991). Table 3 lists the contexts in which homoplasic sites occur. There

Table 4. *Overlap analysis*

(The table gives the observed number of sites which are variable in humans and another primate and the number expected without hypervariable sites.)

		all polymorphic sites		homoplasic sites	
data set	no. of sequences	$W_h W_o$	observed	$W_h W_o$	observed
P. troglodytes[a]	>4	2.9	4	0.9	1
P. troglodytes	2	0.2	0	0.0	0
P. troglodytes versus *Pan paniscus*	2	10.8	11	1.5	3
Gorilla gorilla	2	0.5	1	0.1	1
Pongo pygmaeus	2	14.9	13	1.9	2

[a]With the exception of the first common chimpanzee data set all data are from complete mitochondrial sequences.

is only one obvious pattern: four of the sites occur in a direct repeat involving the adjacent four base pairs (two either side of the site). However, this pattern is not significant. Direct repeats are relatively common since the same amino acid often occurs at adjacent sites; 405 of the 3754 third sites analysed occur in a direct repeat of the four adjacent nucleotides. The probability of observing four or more sites in a direct repeat is 0.142.

4. DISCUSSION

An analysis of 3628 synonymous third sites in the protein-coding regions of 29 human mitochondrial sequences revealed 126 polymorphisms. If all third sites were equally likely to change, the expected number of homoplasies would be 2.2, whereas the observed number is 22: the probability of such an excess is effectively zero. It follows that there has either been recombination or that all third sites are not equally likely to change.

There are three reasons why the likelihood of change may vary between sites. First, there is strong synonymous codon bias in mitochondria (Perna & Kocher 1995). Those sites at which an amino acid was coded for by the less favoured codon in the ancestral human population would be more likely to change and, hence, be homoplasic. Second, it is possible that certain synonymous sites are constrained by selection. Mutations at those sites would be less likely to be detected as polymorphisms. Third, sites may have elevated rates of mutation, either symmetrically (e.g. C→T and T→C) or asymmetrically (e.g. only C→T increased).

All these possibilities can be ruled out as causes of the excess homoplasies. First, codon bias can be allowed for by estimating an 'effective site number', S_e (Maynard Smith & Smith 1998). A conservative estimate is $S_e = 2269$ or 62% of the total number of sites. If there were only 2269 sites equally likely to change, the expected number of homoplasies would be 3.63, which is incompatible with the observed number of 22.

Selective constraints on a proportion of synonymous sites can also be rejected. If such constraints were common, the divergence at saturation between the sequences of humans and other primates would be lower than the value expected, given the observed codon bias. As shown in figure 1, the actual divergence approaches the divergence predicted from the level of codon bias at saturation.

There is no evidence of variation in the mutation rate. If some sites have an elevated mutation rate in both directions, this would cause sites that are variable in humans to also be variable in other primates. In fact, no such tendency exists. The number of polymorphic sites and of apparently homoplasic sites in humans that are variable in other primates both agree rather closely with the numbers expected if all third sites are equally likely to change.

It is in any case more likely that the rate of mutation would be elevated in one direction only (e.g. C→T). The classic examples of hypermutation, CpG dinucleotides in mammals (Bulmer 1986; Bird 1987) and CCAGG sites in bacteria (Coulondre *et al.* 1978; Halliday & Glickman 1991), conform to this pattern. However, an asymmetric increase in the mutation rate at some sites cannot account for the data. The essential difficulty is as follows: if, at some sites, C→T mutations are much more common than T→C, then at most such sites the ancestral base would be T and such sites would not contribute homoplasies. A numerical model (see the appendix and table 1) suggests that the bias in mutation rate would have to be at least 30-fold and that most third sites would have to be 'hypervariable' in this sense to account for the data. If so, the divergence at saturation would be reduced well below the level actually seen. A final reason for doubting the existence of hypervariable sites is that, in other systems, sites that are hypermutable exist in particular contexts, but there are no obvious patterns in the contexts of sites that are homoplasic in our data.

Finally, two steps which are required for recombination between mitochondrial lineages are known to exist. Paternal mitochondria enter the egg (Kaneda *et al.* 1995; Ankel-Simmons & Cummins 1996) and they contain the enzymes necessary for homologous recombination (Thyagarajan *et al.* 1996). The only barriers to recombination are the fusion of mitochondria and the time for which paternal mitochondria survive once they are in the egg. It remains unclear whether mitochondria fuse frequently (Howell 1997) and in mice there are efficient mechanisms for eliminating paternal mitochondria, so that, within several hours of fertilization, paternal mtDNA can no longer be detected (Kaneda *et al.* 1995). However, only relatively low levels of recombination are probably required to generate the patterns of homoplasy we observe.

An estimate of the relative importance of recombination and mutation is given by calculating the 'homoplasy ratio' (Maynard Smith & Smith 1998). The matrix of strains by informative sites is randomized, maintaining the number of the two alleles at each site, but allotting them randomly to strains. The most parsimonious tree is then found for the new matrix. This gives a value of h_e, the number of homoplasies expected if there is complete linkage equilibrium. For the human data, the mean value (20 trials) of h_e was 49.6 (range 47–55). The homoplasy ratio is $(h_o - h_c)/(h_e - h_c) = 0.40$. The ratio has an expected value of zero for a clonal population and one for a

population in linkage equilibrium. For the human data neither of these extreme assumptions holds.

Mitochondrial DNA has been used extensively in the study of human evolution. In many of these analyses the clonality of mitochondria has been either explicitly or implicitly assumed (Cann *et al*. 1987; Vigilant *et al*. 1991; Rogers & Harpending 1992; Hasegawa *et al*. 1993; Wakeley 1993; Rogers 1995). It is clear that many of these conclusions will have to be treated with caution or reassessed. It certainly seems dangerous to assume that mitochondria are clonal when there is evidence against and no evidence in favour of such a conjecture.

We are very grateful to Rich Kliman, Jody Hey and Lars Jermiin for helpful comments and to Cheryl Wise for making her chimpanzee *ND2* sequences available to us. N.H.S. was supported by a Wellcome Trust grant to Professor B. G. Spratt. A.E.-W. is supported by the Royal Society.

APPENDIX A

To calculate the number of hypermutable sites required to explain the level of homoplasy in our data, let us imagine there are two classes of site, normal sites and hypervariable sites. Let the proportion of sites which are hypermutable, either in the non-hypermutable or the hypermutable state, be α. At normal sites, the probability that a change occurs and is detected as a polymorphism in our sample of l sequences is u. Hypervariable sites are assumed to change at the same rate, u, in one direction and at a higher rate, uk, in the other: at equilibrium, a proportion $k/(k+1)$ of such sites will be in the non-hypermutable state and $1/(k+1)$ in the hypermutable state. We conservatively assume that the sequences form a star phylogeny; the probability of detecting x hits in l sequences when the mutation rate is m (i.e. u or uk) is therefore

$$P(l,x,m) = \frac{l!}{x!(l-x)!} m^x (1-m)^{(l-x)}, \qquad \text{(A1)}$$

The proportion of sites which are expected to be polymorphic is

$$R = \frac{\alpha}{(k+1)}(1 - P(l,0,uk)) + \left(\frac{\alpha k}{(k+1)} + (1-\alpha)\right) \times (1 - P(l,0,u)), \qquad \text{(A2)}$$

and the proportion of sites which are expected to be homoplasic is

$$H = \frac{\alpha}{(k+1)}(1 - P(l,0,uk) - P(l,1,uk)) + \left(\frac{\alpha k}{(k+1)} + (1-\alpha)\right)(1 - P(l,0,u) - P(l,1,u)). \qquad \text{(A3)}$$

The total number of sites is 3628, of which 126 are polymorphic, so $R = 126/3628$. On average, in the most parsimonious trees, there were 16.1 homoplasic sites: this is less than the observed number of 22, because some sites changed more than twice. Hence, $H = 16.1/3628$. We solved equations (A2) and (A3) for these values numerically, for $k = 31$, 40, 50 and 100. No solutions could be found for $k < 31$. The results are given in table 1.

REFERENCES

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

Ankel-Simons, F. & Cummins, J. M. 1996 Misconceptions about mitochondria and mammalian fertilization: implications for theories in human evolution. *Proc. Natl Acad. Sci. USA* **93**, 13 859–13 863.

Arnason, U., Xu, X. & Gullberg, A. 1996 Comparison between the complete mitochondrial DNA sequences of *Homo sapiens* and the common chimpanzee based on non-chimaeric sequences. *J. Mol. Evol.* **42**, 145–152.

Bird, A. P. 1987 CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.* **3**, 342–347.

Brown, M. D., Voljavec, A. S., Lott, M. L., MacDonald, I. & Wallace, D. C. 1992*a* Leber's hereditary optic neuropathy: a model for mitochondrial neurodegenerative diseases. *FASEB J.* **6**, 2791–2799.

Brown, M. D., Voljavec, A. S., Lott, M. T., Torroni, A., Yang, C.-C. & Wallace, D. C. 1992*b* Mitochondrial DNA complex I and III mutations associated with Leber's hereditary optic neuropathy. *Genetics* **130**, 163–173.

Bulmer, M. 1986 Neighbouring base effects on substitution rates in pseudogenes. *Mol. Biol. Evol.* **3**, 322–329.

Cann, R., Stoneking, M. & Wilson, A. C. 1987 Mitochondrial DNA and human evolution. *Nature* **325**, 31–36.

Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. 1978 Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**, 775–780.

Halliday, J. A. & Glickman, B. W. 1991 Mechanisms of spontaneous mutation in DNA repair-proficient *Escherichia coli*. *Mutat. Res.* **250**, 55–71.

Hasegawa, M., Di Rienzo, A., Kocher, T. D. & Wilson, A. C. 1993 Toward a more accurate time scale for the human mitochondrial DNA tree. *J. Mol. Evol.* **37**, 347–354.

Howell, N. 1997 mtDNA recombination: what do *in vitro* data mean? *Am. J. Human Genet.* **61**, 18–22.

Jazin, E., Soodyall, H., Jalonen, P., Lindholm, E., Stoneking, M. & Gyllensten, U. 1998 Mitochondrial mutation rate revisited: hot spots and polymorphism. *Nature Genet.* **18**, 109–110.

Kaneda, H., Hayashi, J. I., Takahama, S., Taya, C., Fischer Lindahl, K. & Yonekawa, H. 1995 Elimination of paternal mitochondrial DNA in intraspecific crosses during early mouse embryogenesis. *Proc. Natl Acad. Sci. USA* **92**, 4542–4546.

Kobayashi, Y., Momoi, M. Y., Tominaga, K., Shimoizumi, H., Nihei, K., Yanagisawa, M., Kagawa, Y. & Ohta, S. 1991 Respiration-deficient cells are caused by a single point mutation in the mitochondrial tRNA-Leu (UUR) gene in mitochondrial myopathy, encephalopathy, lactic acidosis, and strokelike episodes (MELAS). *Am. J. Human Genet.* **49**, 590–599.

Lunt, D. H. & Hyman, B. C. 1997 Animal mitochondrial DNA recombination. *Nature* **387**, 247.

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

Maynard Smith, J. & Smith, N. H. 1996 Synonymous nucleotide divergence: what is saturation? *Genetics* **142**, 1033–1036.

Maynard Smith, J. & Smith, N. H. 1998 Detecting recombination from gene trees. *Mol. Biol. Evol.* **15**, 590–599.

Ozawa, T. (and 11 others) 1991 Patients with idiopathic cardiomyopathy belong to the same mitochondrial DNA gene family of Parkinson's disease and mitochondrial encephalomyopathy. *Biochem. Biophys. Res. Commun.* **177**, 518–525.

Perna, N. T. & Kocher, T. D. 1995 Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J. Mol. Evol.* **41**, 353–358.

Rogers, A. R. 1995 Genetic evidence for a Pleistocene population explosion. *Evolution* **49**, 608–615.

Rogers, A. R. & Harpending, H. 1992 Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9**, 552–568.

Sbisa, E., Tanzariello, F., Reyes, A., Pesole, G. & Saccone, C. 1997 Mammalian mitochondrial D-loop structural analysis: identification of new conserved sequences and their functional and evolutionary implications. *Gene* **205**, 125–140.

Thyagarajan, B., Padua, R. A. & Campbell, C. 1996 Mammalian mitochondria possess homologous DNA recombination activity. *J. Biol. Chem.* **271**, 27 536–27 543.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. 1991 African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507.

Wakeley, J. 1993 Substitution rate heterogeneity variation among sites in hypervariable region 1 of human mitochondrial DNA. *J. Mol. Evol.* **37**, 613–623.

Wallace, D. C., Singh, G., Lott, M. T., Hodge, J. A., Schurr, T. G., Lezza, A. M. S., Elsas, L. J. & Nikoskelainen, E. K. 1988 Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science* **242**, 1427–1430.

Wise, C. A., Sraml, M. & Easteal, S. 1998 Departure from neutrality at the mitochondrial NADH dehydrogenase subunit 2 gene in humans, but not in chimpanzees. *Genetics* **148**, 409–421.

Yoneda, M., Tanno, Y., Horai, S., Ozawa, T., Miyatake, T. & Tsuji, S. 1990 A common mitochondrial DNA mutation in the tRNALys of patients with myoclonus epilepsy associated with ragged-red fibers. *Biochem. Int.* **21**, 789–796.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.