Recombination in Human Mitochondrial DNA?

Carsten Wiuf

Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom

Manuscript received April 10, 2001
Accepted for publication July 17, 2001

ABSTRACT

The possibility of recombination in human mitochondrial DNA (mtDNA) has been hotly debated over the last few years. In this study, a general model of recombination in circular molecules is developed and applied to a recently published African sample ($n = 21$) of complete mtDNA sequences. It is shown that the power of correlation measures to detect recombination in circular molecules can be vanishingly small and that the data are consistent with the given model and no recombination only if the overall heterogeneity in mutation rate is <0.09.

RESEARCH based on the phylogenetic analyses of complete human mtDNA sequences has suggested that mtDNA undergoes recombination (Awadalla *et al.* 1999; Eyre-Walker *et al.* 1999). Others challenged this conclusion, arguing that the data applied were flawed (Macaulay *et al.* 1999) or that the methodology was inappropriate (Jorde and Bamshad 2000; Kivisild and Villems 2000; Kumar *et al.* 2000; Parsons and Irwin 2000). Recently, it was reported that two samples of complete mtDNA sequences, not published previously, show no evidence for recombination (Ingman *et al.* 2000; Elson *et al.* 2001). In all these analyses, it is assumed that recombination occurs in a simple, easily detectable, fashion. In this article, recombination in circular molecules is discussed from a model perspective and it is shown that effects of recombination can be much more intricate and less intuitive than assumed in previous articles. It is shown that the power to detect recombination using correlation measures can be vanishingly small. Further, as an illustration, the mtDNA sample of African origin ($n = 21$) in Ingman *et al.* (2000) is reanalyzed for the presence of recombination. It is found that data are inconsistent with standard models of evolution without recombination, unless a high heterogeneity in mutation rate is assumed.

A decay of linkage disequilibrium (LD) with physical distance is expected in recombining sequences. Putative evidence of decay with physical distance has been the core argument in favor of recombination in human mtDNA. In all published analyses of LD in mtDNA known to the author, it is assumed that genetic distance increases with increasing physical distance, or that genetic distance is (roughly) proportional to physical distance. However, this is not necessarily the case. Depending on the nature of the recombination process, the relation between the two measures of distance can be more complex than that. The aim of this article is to demonstrate how modeling of mtDNA evolution can play a role in understanding the potential effects of recombination. First, I describe a general model of recombination in circular molecules and, second, this model is applied to the data of African origin.

Imagine that a molecule, M_1, in a generation recombines with probability r. In that case, a second molecule, M_2, is chosen and a new molecule, M, is formed consisting of a part, P_1, copied from M_1 and a part, P_2, copied from M_2. This process is called the mixing process. In general, the possible forms of P_1 and P_2 are restricted only by $P_1 \cap P_2 = \varnothing$ and $P_1 \cup P_2 = M$; that is, the parts are disjoint and together they constitute a whole molecule. The interpretation of r would vary with the type of genetic system in question. For example, in the context of mtDNA, r could be the probability that two mtDNA molecules meet and recombine, or r could be the probability that a part of a nuclear pseudogene is copied onto a mtDNA molecule. For now the exact definition of "generation" is left out, as well as any attempt to model aspects of the reproductive structure other than recombination, such as the number of offspring, possible paternal leakage (*i.e.*, paternal mitochondria that enter the egg), etc. These aspects are of course of considerable importance in describing the genealogical structure of a sample of mtDNA sequences and are required in the analysis of data. First, I am concerned with the effect of mixing M_1 and M_2.

Several copies of the mitochondrial genome are present in each mitochondrion (generally 5–10 copies; King and Stansfield 1990). Further, mitochondria are known to possess at least some of the enzymes required for recombination (Thyagarajan *et al.* 1996), and rearrangements of mitochondria genomes by within-lineage recombination occur at low frequencies in human cells (Kajander *et al.* 2000). However, the exact mechanism by which this occurs is not known nor is it known

Address for correspondence: Department of Statistics, 1 S. Parks Rd., Oxford OX1 3TG, England. E-mail: wiuf@stats.ox.ac.uk

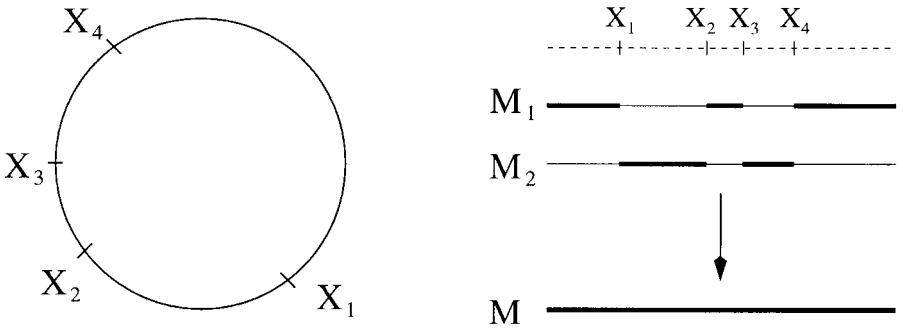


FIGURE 1.—Recombination in a circular molecule. Two molecules, M_1 and M_2, are in this example broken in four places, X_1, X_2, X_3, and X_4, and combined into M. The left shows a molecule with the four breakpoints. The right shows how the fragments of M_1 and M_2 are put together; here the circular molecules are represented as straight lines. M_1 contributes $P_1 = (X_1, X_2) \cup (X_3, X_4)$ and M_2 contributes $P_2 = (X_2, X_3) \cup (X_4, X_1)$, where (u, v) denotes the arc counted clockwise from u to v. Viewed from the present toward the past, the effect of arecombination is to split the ancestral nucleotides of M into two groups: those that belong to M_1 and those that belong to M_2. In contrast to recombination in linear genomes, the two "ends" of M on the right share the same ancestor M_1. The molecule M provides no information about the state of the nucleotides in M_1 and M_2 marked by thin lines.

whether rearrangements could occur between paternal leaked genomes and maternal genomes. In certain fungi, the mitochondria are inherited biparentally and recombination occurs between both parental genomes (SAVILLE *et al.* 1998). Other genetic systems exhibit recombination in various other ways. Bacterial genomes undergo recombination by processes involving direct contact between two cells (conjugation), by picking up DNA pieces from the environment (transformation), or by virus-mediated DNA transfer (transduction; GRIFFITHS *et al.* 1996). Viral genomes likewise undergo recombination after formation of a circular genome (GRIFFITHS *et al.* 1996). The mtDNA genome might recombine in similar ways. In contrast to linear genomes, however, an even number of breakpoints (crosses) is required to divide a molecule into two disjoint parts (Figure 1). Each recombination most likely involves only two crosses, though recombination in bacteria by conjugation occasionally results in four or more crosses (GRIFFITHS *et al.* 1996). In this article two simple cases are considered: (A) a model with two breakpoints and (B) one with four breakpoints. In the analysis of data only examples of A are applied, as A is likely to be more realistic biologically than B.

Consider the molecule M. Viewed from the present toward the past, the nucleotides in M share different ancestors. Those in P_1 have ancestor M_1 and those in P_2 have ancestor M_2 (Figure 1). Let y and z be two nucleotides in M separated by a physical distance $d = d(y, z) \leq \frac{1}{2}$ (the smaller of the two arcs from y to z), such that the length of the whole molecule is 1. The genetic distance, $R(d)$, between y and z is here defined as the probability that y and z have different ancestors, given that M is created by recombination in the present generation. This definition is very convenient and relies only on the way nucleotides copied from M_1 and M_2 are joined into M; in fact $R(d)$ is independent of the probability r of a recombination. It is then easy to compare the effect of mixing in different models without reference to how frequently recombination events occur. In the standard model of recombination for nuclear sequences (*i.e.*, no heterogeneity in recombination rate), a linear relation is obtained between the two distance measures.

MATERIALS AND METHODS

Genetic distance, $R(d)$: Consider two nucleotides y and z at distance $d = d(y, z)$. In the two cases A and B, respectively, $R(d)$ is given by

$$R(d) = P(X_1 < d \leq X_2) \leq 1 \quad (1)$$

and

$$R(d) = P(X_1 < d \leq X_2) + P(X_3 < d \leq X_4) \leq 1, \quad (2)$$

respectively, where X_1, X_2, X_3, and X_4 denote the breakpoints ordered clockwise from y (Figure 2). In Equations 1 and 2, $R(d)$ is not necessarily increasing in d and $R(d)$ might depend on y due to rate heterogeneity or hotspots. To proceed farther, some restrictions are put on the distribution of the breakpoints. Choose one breakpoint, X, randomly among all nucleotides and let Z_i, $i = 1, 2$, or $i = 1, 2, 3, 4$, be the arc lengths numbered consecutively clockwise from X. Thus, rate homogeneity is assumed and no region (*e.g.*, the control region) evolves in any different ways from the rest of the molecule (with respect to recombination). Without loss of generality it can be assumed that in model A, $Z_1 \leq \frac{1}{2}$. Equation 1 becomes

$$R(d) = 2d - 2\int_0^d P(Z_1 \leq x)\,dx \leq 2d. \quad (3)$$

It can be shown that $R(d)$ is increasing in d with a gradually decreasing slope (APPENDIX). Equation 2 does not take a similar simple form, but $R(d)$ can be bounded upward,

$$R(d) \leq \min\{4d, 1\} \quad (4)$$

(APPENDIX).

The examples of model A shown in Figure 3 are as follows: A1, $Z_1 = L \leq 0.5$, $Z_2 = 1 - L$, $R(d) = \min\{2d, 2L\}$; and A2, $Z_1 \sim U(0, 0.5)$ (Z_1 is uniform on the interval from 0 to 0.5), $Z_2 = 1 - Z_1$, $R(d) = 2d(1 - d)$. Example A2 is equivalent to choosing two points at random on the circle. This provides two extreme models: one with high interference (the relative positions of the two breakpoints are correlated) and no variation in the length of the exchanged segment (A1); and one with no interference (the two breakpoints are uncorrelated) and relatively high variance, $\mathrm{Var}(Z_1) = \frac{1}{48}$ (A2). The maximal

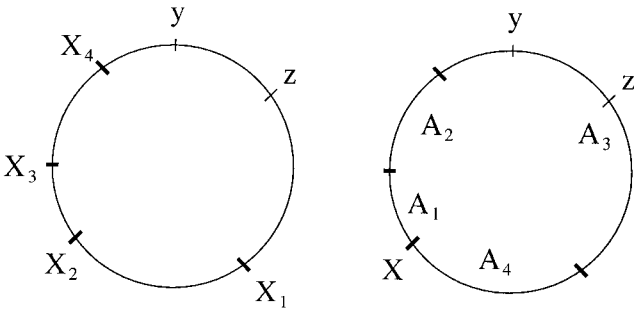


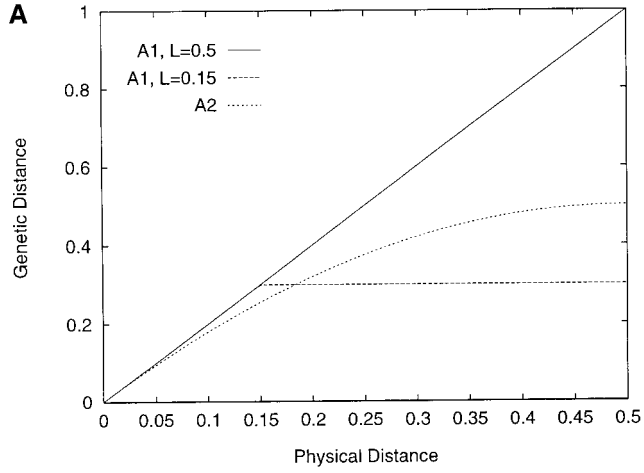
FIGURE 2.—Notation. Left, the nucleotides y and z are fixed and their distance is given by the smaller of the two arcs from y to z. The breakpoints X_1, X_2, X_3, and X_4 are numbered consecutively clockwise from y. In this example no breakpoints fall between y and z. Right, a breakpoint, X, is chosen at random among all nucleotides and the fragments between breakpoints are numbered clockwise from X. The fragment Z_3 comprises both y and z.

variance is $\frac{1}{16}$, which is obtained only if $P(Z_1 = 0) + P(Z_1 = \frac{1}{2}) = 1$ (HOFFMANN-JØRGENSEN 1994).

Examples of model B shown in Figure 3 are as follows: B1, $Z_1 = Z_2 = Z_3 = L \leq \frac{1}{4}$, $Z_4 = 1 - 3L$. The form of $R(d)$ depends in a complicated way on L. For $0 \leq L \leq \frac{1}{6}$, $R(d) = 4d$, if $0 \leq d < L$; $R(d) = -2d + 6L$, if $L \leq d < 2L$; $R(d) = 2d - 2L$, if $2L \leq d < 3L$; and $R(d) = 4L$, if $3L \leq d < 0.5$. For $\frac{1}{6} \leq L \leq \frac{1}{5}$, $R(d) = 4d$, if $0 \leq d < L$; $R(d) = -2d + 6L$, if $L \leq d < 2L$; $R(d) = 2d - 2L$, if $2L \leq d < 1 - 3L$; and $R(d) = 2(1 - 4L)$, if $1 - 3L \leq d < 0.5$. For $\frac{1}{5} \leq L \leq \frac{1}{4}$, $R(d) = 4d$, if $0 \leq d < L$; $R(d) = -2d + 6L$, if $L \leq d < 1 - 3L$; $R(d) = -4d + 2$, if $1 - 3L \leq d < 2L$; and $R(d) = 2(1 - 4L)$, if $2L \leq d < 0.5$. And in the second example, B2, $(Z_1, Z_2, Z_3) \sim D_3(1, 1, 1)$ (Dirichlet with parameters 1, 1, and 1), $Z_4 = 1 - (Z_1 + Z_2 + Z_3)$, $R(d) = 4d(1 - d)\{d^2 + (1 - d)^2\}$. Example B2 is equivalent to choosing four points at random on the circle. Similar to A1 and A2, example B1 shows high interference whereas B2 does not.

The genealogy of a sample: It is assumed that recombination happens relatively rarely such that a single molecule is not likely to experience more than one recombination when passed on from parent to child. One human generation is counted as one generation. During one generation the molecule might be copied several times. The next generation of females and males is chosen from the previous generation by choosing randomly N females and N males from the reproductive N females. Assume a molecule M_1 is drawn at random from the female pool. With probability r a recombination event occurs and a second parent, M_2, from the male pool is chosen, and M_1 and M_2 are joined according to the mixing process. Thus, heteroplasmy is introduced through recombination of paternal leaked mtDNA, and r encompasses the probability of leakage, that two molecules meet (one female and one male) and that the two molecules mix. The inbreeding effective population size, N_e, is $N_e = N$ (EWENS 1979) and standard techniques lead to a coalescent approximation of the distribution of a sample's history (HUDSON 1983). Going backward in history, the time between events, coalescence, or recombination is exponentially distributed with rate $k(k - 1)/2 + kR/2$ while there are k lineages ancestral to the sample. The event is a coalescence with probability $(k - 1)/(k - 1 + R)$ and a recombination with probability $R/(k - 1 + R)$. Time is measured in units of N_e generations, and $R = 2N_e r$. This provides a very efficient way of simulating sample histories.

Model specifications: The mutation process is assumed to



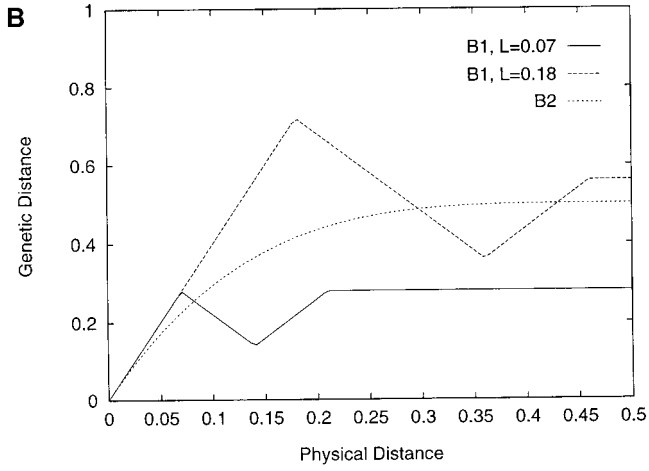


FIGURE 3.—Examples of models A and B. (A) Examples of model A. A1, one segment of fixed length, L percentage of an entire molecule, is exchanged. If $L = 0.5$ ~8300 nucleotides are exchanged, and if $L = 0.15$ ~2500 nucleotides are exchanged. A2, breakpoints are chosen at random. (B) Examples of model B. B1, two segments of fixed length are exchanged. The two exchanged segments and the segment between them each comprise L percentage of an entire molecule. If $L = 0.18$ ~$2 \times 3000 = 6000$ nucleotides are exchanged, and if $L = 0.07$ ~$2 \times 1150 = 2300$ nucleotides are exchanged. B2, four breakpoints are randomly chosen.

be a two-state Jukes-Cantor model with rate heterogeneity (GU and LI 1998). The observed pairwise sequence divergence in the sample is $\hat{\pi} = 4.63 \times 10^{-3}$ (INGMAN *et al.* 2000) and the scaled mutation rate, $\theta = 2N_e u$ (where u is the probability of a mutation per molecule per generation), is estimated from $\hat{\pi}$ and the coalescent model for two values of the rate heterogeneity parameter, α: $\alpha = 0.2$ and $\alpha = \infty$ (all sites evolve with the same rate; APPENDIX). For $\alpha = 0.2$, $\hat{\theta} = 4.88 \times 10^{-3}$, and for $\alpha = \infty$, $\hat{\theta} = 4.63 \times 10^{-3}$. This estimate of θ does not presuppose that recombination does not occur (APPENDIX). If the sequence divergence, π, is changed then θ is changed accordingly, but not the time depth in the genealogy of the sequences. The recombination rate is varied over $R = 0.1$, 1, and 10, which on average gives $\approx 3.6R$ recombination events in the sample's history (HUDSON 1983). From the equation $R = 2N_e r$, the probability of a recombination (incorporating

the probability that the two recombining molecules meet) is found to vary over $r = 10^{-5}$, 10^{-4}, and 10^{-3}, assuming $N_e = 5000$. Simulation results are given for four models, explained in MATERIALS AND METHODS: A1, $L = 0.015$; A1, $L = 0.15$; A1, $L = 0.5$; and A2, $Z_1 \sim U(0, \frac{1}{2})$.

RESULTS

The mixing process: Figure 3 displays a number of examples of the cases A and B with a general mathematical treatment of the cases A and B put in MATERIALS AND METHODS. A linear relation between physical and genetic distance is achieved only if one-half of a molecule is exchanged. This has been proposed as a model of recombination in bacteria (HUDSON 1994). In example A2 in Figure 3A, both breakpoints are chosen at random and all pairs of nucleotides with a physical distance >0.25 have more or less the same genetic distance, $0.375 \leq R(d) \leq 0.5$. Such pairs provide information about variation in LD for distant pairs of nucleotides, rather than information about the decay of LD. Examples A1 and A2 differ in the amount of interference (in A1 there is no interference), but the shape of $R(d)$ is very similar in the two examples.

If there are many breakpoints, there is no straightforward relation between genetic and physical distance. In the examples displayed in Figure 3B, two show an increase followed by a decrease in genetic distance, whereas the last example shows a rapid increase toward the maximal genetic distance (here 0.5). In the latter, most pairs of nucleotides are expected to show the same level of LD. These examples are by no means exhaustive, but they are sufficient to establish that the relation between physical and genetic distance is not necessarily simple. In general, there are two things in play:

Interference: If the end points of the two exchanged segments are correlated (as in Figure 3B, example 1, but not in example 2), the genetic distance does not increase overall with physical distance. In Figure 3B, the "valley" is caused by interference.

Circularity: The genetic distance between two points is affected by the molecule bending back onto itself. In general, the maximal genetic distance between two sites is limited to $\frac{1}{2}$, but this limit might be lower if two (or more) segments are exchanged. As an example, consider Figure 3B, example 1, where the distance levels off at $d = 0.46$. The two exchanged segments plus the shortest segment between them take up $3L \leq \frac{3}{4}$ of the molecule. If the distance, d, between two points exceeds $1 - 3L$, an effect similar to interference appears and the genetic distance levels off. This should not be confused with the phenomenon that appears in Figure 3, A and B, example 1, $L = 0.07$, which is due to a small segment being exchanged.

African mtDNA data: How do different models affect the conclusions drawn in recent studies (INGMAN *et al.* 2000; ELSON *et al.* 2001) that mtDNA data are consistent with the hypothesis of no recombination? To answer this question, a specific model of the genealogy is required, in addition to a specification of the mixing process. One such model is derived in MATERIALS AND METHODS and assumes that paternal leakage is the principal source generating heteroplasmy within a cell.

A brief summary of the African sample (INGMAN *et al.* 2000) is useful. It comprises $n = 21$ complete sequences of ~16,500 nucleotides each. There are 367 polymorphic sites and of these 198 are informative (from Figure 4, INGMAN *et al.* 2000). The data were found to be consistent (using Tajima's D) with a population of constant size and thus conform to the model described in MATERIALS AND METHODS. Tajima's D might not be adequate in the presence of recombination, and a further test is proposed below. To assess the extent of recombination, Pearson-correlation coefficients ρ_{Δ^2} and $\rho_{D'}$, respectively, were calculated between physical distance and level of LD measured by the quantities Δ^2 and D' (for a definition see DEVLIN and RISCH 1995; Δ^2 is also known at r^2), respectively. The correlation coefficients were found to be consistent with a hypothesis of no recombination ($\rho_{D'} = 1 \times 10^{-3}$ and $\rho_{\Delta^2} = 2.23 \times 10^{-6}$). If only three out of four possible haplotypes are present in two sites, the quantity $|D'|$ is 1. In this sample the proportion of pairs where $|D'| < 1$, $f(D')$, is 0.093 (obtained from Figures 1 and 4 in INGMAN *et al.* 2000).

Because of the possibly complicated (and unknown) relation between physical and genetic distance, the correlation coefficient of physical distance with D' (or Δ^2) is not necessarily a good indicator of the presence of recombination. The fraction $f(D')$ might prove better; if recombination is frequent, many pairs of sites will show four distinct haplotypes (irrespectively of the mixing process) and in each of these cases, $D' < 1$. Thus, $f(D')$ relates to the number of homoplasies in the sample and therefore to the amount of recombination. Unfortunately (in this context), homoplasies might also appear as a product of the mutation process.

Table 1 shows the power of ρ_{Δ^2}, $\rho_{D'}$, and $f(D')$ for various choices of mixing process, amount of recombination, and mutation process (further details of parameters and the simulation method are given in MATERIALS AND METHODS). The fraction $f(D')$ seems superior to the two correlation measures when the exchanged segment is small, and $\rho_{D'}$ has in most cases higher power than ρ_{Δ^2}. (That D' is a better measure of LD than Δ^2 has repeatedly been stressed in the literature; DEVLIN and RISCH 1995; GUO 1997). In the first example, ~250 nucleotides (A1, $L = 0.015$) are exchanged at a recombination event, while in the second (A1, $L = 0.15$), 2500 nucleotides are exchanged. In example A2 ~4000 nucleotides are exchanged on average. If more than one segment is exchanged (*e.g.*, as in example B) the power of the correlation measures can be vanishingly small

TABLE 1

The power of ρ_{Δ^2}, $\rho_{D'}$, and $f(D')$

		A1, $L = 0.015$			A1, $L = 0.15$			A1, $L = 0.05$			A2		
R	α	ρ_{Δ^2}	$\rho_{D'}$	$f(D')$	ρ_{Δ^2}	$\rho_{D'}$	$f(D')$	ρ_{Δ^2}	$\rho_{D'}$	$f(D')$	ρ_{Δ^2}	$\rho_{D'}$	$f(D')$
0.1	∞	5.2	5.7	6.2	5.8	8.8	7.2	11	13	8.4	7.4	10	7.5
	0.2	5.3	5.4	5.0	7.1	6.8	4.8	11	11	7.5	7.6	8.2	7.6
1	∞	5.0	6.5	6.5	17	33	32	51	59	46	32	42	35
	0.2	5.4	5.8	5.5	15	18	20	51	51	32	33	36	25
10	∞	4.9	18	41	77	85	97	99	99	100	95	95	99
	0.2	6.0	6.0	19	76	72	91	99	97	98	95	92	94

Shown is the number of times in percentage ρ_{Δ^2} and $\rho_{D'}$ were below the 5% fractile and the number of times in percentage $f(D')$ was outside the 2.5–97.5% confidence interval (CI). The null distribution was obtained simulating under $R = 0$. See MATERIALS AND METHODS for a description of the models. If the exchanged segment is small (A1, $L = 0.015$; ~250 nucleotides), the power of the correlation measures are low, even for high rates of recombination, R. A total of 10,000 samples were simulated for the null distribution ($R = 0$) and 1000 for $R > 0$. Standard errors are of the order $\sqrt{p/m}$ where p is the power and m is the number of simulations performed.

whereas the power of $f(D')$ is relatively high (results not shown). This is expected because $\rho_{D'}$ (or ρ_{Δ^2}) is designed to detect recombination only if genetic and physical distance is monotonically related, whereas $f(D')$ is designed to detect recombination from an excess of homoplasies.

Other mutation processes, *e.g.*, processes that allow for different rates in different regions or for transition/transversion bias, have reduced power compared to the mutation process applied in the simulations in Table 1, simply because more variable sequence patterns are expected.

The probability of no recombination events in a sample's history is

$$P(\text{no recombination}) = \prod_{k=2}^{n} \frac{k-1}{k-1+R}$$

(MATERIALS AND METHODS). If $R = 0.1$, P (no recombination) ≈ 0.70, and most samples have not experienced recombination in its history. Consequently, recombination is hard to detect, irrespective of the mixing process. If $R = 1$, P(no recombination) ≈ 0.04, and most samples have experienced at least one recombination event. With an effective population size of 5000 and $R = 1$ the probability of a recombination (*i.e.*, the probability of leakage, that two molecules meet, one sperm-derived and one egg-derived, and that they mix) is 10^{-4} per molecule per generation. This could easily turn out to be orders of magnitude too high, but seems to provide a lower bound to which recombination could be inferred from phylogenetic analyses (provided the exchanged segment in a recombination event is fairly large; Table 1). In general, the power increases with increasing sample size, but only slowly. Table 2 shows the power for three values of sample sizes $n = 21$, 50, and 100 assuming model A2.

But are the data consistent with any model of recombination? In Tables 3–5, this is investigated using the statistics $\rho_{D'}$, $f(D')$, and I = no. inf/no. polym, the ratio of informative to polymorphic sites (informative sites are polymorphic sites where two different nucleotides are present in at least two sequences each). In general, the expectation of I does not vary much with R, but the variance decreases with increasing R (results not shown). Therefore, I indicates whether the data fit a constant population size model and can be thought of as a complementary test to Tajima's D. The observed value of I (0.53) is consistent with all the investigated models. In contrast, the observed value of $f(D')$ (0.093) does not conform to a model with no recombination ($P < 0.002$ if $\alpha = 0.2$ and $P < 0.0002$ if $\alpha = \infty$). If $\alpha \approx 0.09$, $P \approx 0.05$ (results not shown). That is, more homoplasies are observed than can be explained by a model without recombination, unless $\alpha < 0.09$. The observed value of $\rho_{D'}$ (0.001) is consistent, in some cases, with high levels of recombination. Thus, even though the correlation coefficients can be explained by a model without recombination, other aspects of the data cannot. Similar results were obtained using a sequence divergence higher and lower than the one applied here (results not shown).

What is the true value of the rate heterogeneity parameter? Estimates of α vary and depend on the region(s) under scrutiny (MEYER *et al.* 1999) and are confounded with other kinds of variations in the mutation rate, such as region-specific variation or biases in the rate of transversions to transitions. Further, clonal reproduction is assumed; that is, $R = 0$. Thus, it can be difficult to get a firm idea of the true value of α from data. Using tree-puzzle (STRIMMER and VON HAESELER 1996) and assuming clonality, I found $\hat{\alpha} = 0.10 \pm 0.05$ for the African data set. This is not surprising: If $R = 0$, the data are consistent with the given model only if $\alpha < 0.09$, and $\hat{\alpha} = 0.10 \pm 0.05$ allows for this to be true.

TABLE 2

The effect of sample size on power

	$n = 21$			$n = 50$			$n = 100$		
R	ρ_{Δ^2}	$\rho_{D'}$	$f(D')$	ρ_{Δ^2}	$\rho_{D'}$	$f(D')$	ρ_{Δ^2}	$\rho_{D'}$	$f(D')$
0.1	7.6	8.2	7.6	7.8	11	8.6	9.1	10	9.7
1	33	36	25	36	42	36	37	44	37
10	95	92	94	97	97	99	98	97	99

Shown is the power (in percentage) of ρ_{Δ^2}, $\rho_{D'}$, and $f(D')$ for $\alpha = 0.2$, $\pi = 4.63 \times 10^{-3}$, and three sample sizes, n, in example A2. The null distribution was obtained simulating under $R = 0$. See MATERIALS AND METHODS for a description of the models. A total of 10,000 samples were simulated for the null distribution ($R = 0$) and 1000 for $R > 0$. Standard errors are of the order $\sqrt{p/m}$ where p is the power and m is the number of simulations performed.

If $R > 0$, any estimate of α is likely to be downward biased, predicting more rate heterogeneity than is actually there. In that case, the true value is thus expected to be higher than the estimated.

DISCUSSION

These findings suggest that human mtDNA might be recombining. A number of comments should be made at this stage.

The mutation process: The excess of homoplasies observed in the African sample could be generated by a complicated mutation process. Strong rate heterogeneity ($\alpha > 0.09$) in itself is not sufficient. Rate heterogeneity does not lead to a decay of LD with physical or genetic distance but to higher variance in the number of mutations in the sample. Linked mutations (one mutation happens as a result of another mutation) could possibly explain an excess of homoplasies. Estimates of the rate heterogeneity in the hypervariable regions vary; one study reports 0.26 in hypervariable region (HVR)I and 0.13 in HVRII (MEYER *et al.* 1999). Outside the control region, α is estimated to be 0.15 in the African data (V. MACAULAY, personal communication). At the present stage it is difficult to judge whether skewness in rate heterogeneity could explain the data.

The recombination (mixing) process: The models applied to analyze data are mathematically simple and assume that all sites potentially are sites for crossing over. In reality, this might be a very crude assumption (*e.g.*, some bacteria have just a few crossing-over sites) and might complicate detection of recombination from DNA sequences data even further.

Demography: In the analysis a population of constant size is assumed. If the population has been expanding, recombination is in general more difficult to detect, because the genealogy of the sample becomes star-like. Also the expected number of homoplasies in the data is fewer relative to a population of constant size (assuming the same number of mutations), because each mutation is more likely to affect only a single sequence.

mtDNA replication: It is not known how mtDNA repli-

TABLE 3

Data's consistency with models, *I*

	A1, $L = 0.015$		A1, $L = 0.15$		A1, $L = 0.5$		A2	
R, α	∞	0.2	∞	0.2	∞	0.2	∞	0.2
0	+	+	+	+	+	+	+	+
0.1	+	+	+	+	+	+	+	+
1	+	+	+	+	+	+	+	+
10	+	+	+	+	+	+	+	+

Using the *I*-statistics, data are found to be consistent with all the investigated models. Even high amounts of rate heterogeneity ($\alpha = 0.2$) do not change this. + indicates that the observed value of I (0.53) was within the 2.5–97.5% CI obtained from simulations under $R = 0$. See MATERIALS AND METHODS for a description of the models. A total of 10,000 samples were simulated for $R = 0$ and 1000 for $R > 0$. Standard errors are of the order $\sqrt{p/m}$, where p is the power and m is the number of simulations performed.

TABLE 4

Data's consistency with models, $f(D')$

	A1, $L = 0.015$		A1, $L = 0.15$		A1, $L = 0.5$		A2	
R, α	∞	0.2	∞	0.2	∞	0.2	∞	0.2
0	−	−	−	−	−	−	−	−
0.1	−	−	−	−	−	−	−	−
1	−	−	−	+	+	+	+	+
10	+	+	+	+	+	+	+	+

Using the $f(D')$-statistics, data are found to be consistent with models predicting a low amount of recombination but not with models without recombination. + indicates that the observed value of $f(D')$ (0.093) was within the 2.5–97.5% CI obtained from simulations under $R = 0$. See MATERIALS AND METHODS for a description of the models. A total of 10,000 samples were simulated for $R = 0$ and 1000 for $R > 0$. Standard errors are of the order $\sqrt{p/m}$, where p is the power and m is the number of simulations performed.

TABLE 5

Data's consistency with models, $\rho_{D'}$

	A1, $L = 0.015$		A1, $L = 0.15$		A1, $L = 0.5$		A2	
R, α	∞	0.2	∞	0.2	∞	0.2	∞	0.2
0	+	+	+	+	+	+	+	+
0.1	+	+	+	+	+	+	+	+
1	+	+	+	+	+	+	+	+
10	+	+	+	+	−	−	+	+

Using the $\rho_{D'}$-statistics, data are found to be consistent even with some models predicting high amounts of recombination. + indicates that the observed value of $\rho_{D'}$ (0.001) was within the 2.5–97.5% CI obtained from simulations under $R = 0$. See MATERIALS AND METHODS for a description of the models. A total of 10,000 samples were simulated for $R = 0$ and 1000 for $R > 0$. Standard errors are of the order $\sqrt{p/m}$, where p is the power and m is the number of simulations performed.

cates (HOWELL 1997). The power to detect recombination depends strongly on the replication process of mtDNA. Only if the two recombining molecules have different ancestral molecules many human generations back in time can the recombination event be detected. Otherwise there will not be sufficient time for mutations to accumulate in the lineages of the recombining molecules. Recombination between maternal mtDNA molecules can be detected only if the recombining lineages coexist without coalescing over many generations.

Paternal leakage: Paternal leakage has been reported in inbred lines of mice (GYLLENSTEN *et al.* 1991). But sperm-derived mtDNA has also been found to disappear rapidly and completely in an early stage of embryogenesis (SHITARA *et al.* 1998). This process in mice might well be similar in humans. If paternal leakage is rare, the chances to detect recombination from mtDNA sequence data will be further undermined.

Concluding remarks: Even small levels of recombination that may not be immediately detectable in the data can have pronounced effects if recombination is ignored in an analysis of the data (SCHIERUP and HEIN 2000). A reconstructed tree will tend to be star-like and an excess of homoplasies is expected. As an example, evidence for population growth in human data is mainly based on analysis of mtDNA data. These conclusions would be challenged if human mtDNA is recombining.

In conclusion, it seems vital and important that assessment of recombination in the mtDNA is based on proper modeling. Significant correlation of LD with physical distance might be a sign of recombination, but recombination cannot be ruled out as a result of a nonsignificant correlation. Phylogenetic and population genetic analyses might prove insufficient to judge whether human mtDNA is recombining partly because different candidate models vary considerably in what they predict and partly because the power to detect recombination from decay in LD might be vanishingly small. It is therefore not surprising that the original results in AWADALLA *et al.* (1999) and EYRE-WALKER *et al.* (1999) have proven difficult to reproduce.

P. Donnelly, V. Macaulay, G. McVean, M. Przeworski, and K. Strimmer are thanked for commenting on the manuscript. The author was supported by Biotechnology and Biological Sciences Research Council grant 43/MMI9788 and by the Carlsberg Foundation, Denmark.

LITERATURE CITED

AWADALLA, P., A. EYRE-WALKER and J. MAYNARD SMITH, 1999 Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science **286:** 2524–2525.

DEVLIN, B., and N. RISCH, 1995 A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics **29:** 311–322.

ELSON, J. L., R. M. ANDREWS, P. F. CHINNERY, R. N. LIGHTOWLERS, D. M. TURNBUL *et al.*, 2001 Analysis of European mtDNAs for recombination. Am. J. Hum. Genet. **68:** 145–153.

EWENS, W., 1979 *Mathematical Population Genetics.* Springer-Verlag, New York.

EYRE-WALKER, A., N. H. SMITH and J. MAYNARD SMITH, 1999 How clonal are human mitochondria? Proc. R. Soc. Lond. Ser. B **266:** 477–483.

GRIFFITHS, A. J. F., J. H. MILLER, D. T. SUZUKI, R. C. LEWONTIN and W. M. GELBART, 1996 *An Introduction to Genetic Analysis.* W. H. Freeman and Company, New York.

GU, X., and W.-H. LI, 1998 Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proc. Natl. Acad. Sci. USA **95:** 5899–5905.

GUO, S.-W., 1997 Linkage disequilibrium measures for fine-scale mapping: a comparison. Hum. Hered. **47:** 310–314.

GYLLENSTEN, U., D. WHARTON, A. JOSEFSSON and A. C. WILSON, 1991 Paternal inheritance of mitochondrial DNA in mice. Nature **352:** 255–257.

HOFFMANN-JØRGENSEN, J., 1994 *Probability with a View Towards Statistics.* Chapman & Hall, New York.

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

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

HUDSON, R. R., 1994 Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation. J. Evol. Biol. **7:** 535–548.

INGMAN, M., H. KAESSMANN, S. PÄÄBO and U. GYLLENSTEN, 2000 Mitochondrial genome variation and the origin of modern humans. Nature **408:** 708–713.

JORDE, L. B., and M. BAMSHAD, 2000 Questioning evidence for recombination in human mitochondrial DNA. Science **288:** 1931.

KAJANDER, O. A., A. T. ROVIO, K. MAJAMAA, J. POULTON, J. N. SPELBRINK *et al.*, 2000 Human mtDNA sublimons resemble rearranged mitochondrial genomes found in pathological states. Hum. Mol. Genet. **9:** 2821–2835.

KING, R. C., and W. D. STANSFIELD, 1990 *A Dictionary of Genetics.* Oxford University Press, Oxford.

KIVISILD, T., and R. VILLEMS, 2000 Questioning evidence for recombination in human mitochondrial DNA. Science **288:** 1931.

KUMAR, S., P. HEDRICK, T. DOWLING and M. STONEKING, 2000 Questioning evidence for recombination in human mitochondrial DNA. Science **288:** 1931.

MACAULAY, V., M. RICHARDS and B. SYKES, 1999 Mitochondrial DNA recombination—no need to panic. Proc. R. Soc. Lond. Ser. B **266:** 2037–2039.

MEYER, S., G. WEISS and A. VON HAESELER, 1999 Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics **152:** 1103–1110.

PARSONS, T. J., and J. A. IRWIN, 2000 Questioning evidence for recombination in human mitochondrial DNA. Science **288:** 1931.

SAVILLE, B. J., Y. KOHLI and J. B. ANDERSON, 1998 mtDNA recombination in a natural population. Proc. Natl. Acad. Sci. USA **95:** 1331–1335.

Schierup, M., and J. Hein, 2000 Consequences of recombination on traditional phylogenetic analysis. Genetics **156:** 879–891.

Shitara, H., J.-I. Hayashi, S. Takahama, H. Kaneda and H. Yonekawa, 1998 Maternal inheritance of mouse mtDNA in interspecific hybrids: segregation of the leaked paternal mtDNA followed by the prevention of subsequent paternal leakage. Genetics **148:** 851–857.

Strimmer, K., and A. von Haeseler, 1996 Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13:** 964–969.

Thyagarajan, B., R. A. Padua and C. Campbell, 1996 Mammalian mitochondria possess homologous DNA recombination activity. J. Biol. Chem. **271:** 27536–27543.

Communicating editor: J. Hein

APPENDIX

Genetic distance, $R(d)$: Equation 3 follows from

$$R(d) = P(X_1 < d \leq X_2) = 2\int_0^d P(x < d < x + Z_1)\,dx$$
$$= 2\int_0^d P(x \leq Z_1)\,dx = 2d - 2\int_0^d P(Z_1 \leq x)\,dx \leq 2d,$$

using that X_1 is uniformly distributed. By differentiation with respect to d,

$$R'(d) = 2 - 2P(Z_1 \leq d),$$

and it follows that $R(d)$ is increasing with a gradually decreasing slope. Equation 4 follows from reasoning similar to that above and $P(X_3 < d \leq X_4) \leq P(X_1 < d \leq X_4) \leq 2d$.

Model specifications: Assume the standard model of heterogeneity in mutation rates; that is, the mutation rate θ_i of column i in the alignment is given by $\theta_i = \theta u_i$, where u_i is gamma distributed, $u_i \sim \Gamma(\alpha, \alpha)$ (Gu and Li 1998). The probability, π_t, that two nucleotides sharing an ancestor t generations ago are different becomes

$$\pi_t = \frac{1}{2} - \frac{1}{2}\left(1 + \frac{2\theta t}{\alpha}\right)^{-\alpha}$$

(Gu and Li 1998). In the standard coalescent model t is exponentially distributed with parameter 1 and the probability, π, that two nucleotides are different is

$$\pi = \int_0^\infty \pi_t e^{-t} dt = \frac{1}{2} - \frac{1}{2}\int_0^\infty e^{-t}\left(1 + \frac{2\theta t}{\alpha}\right)^{-\alpha} dt. \quad \text{(A1)}$$

If $\alpha = \infty$ then $\pi = \theta/(1 + 2\theta)$, and if $\alpha = 0$ then $\pi = 0$. In (A1), π is also the expected pairwise sequence divergence in a random sample, irrespective of whether $R = 0$ or $R > 0$. If π is estimated by $\hat{\pi}$, the observed average pairwise sequence divergence in the sample and α is assumed known; then an estimate of θ can be produced from (A1). If $R > 0$, the average $\hat{\pi}$ has less variance than if $R = 0$.