European Journal of Human Genetics (2001) 9, 969–972

www.nature.com/ejhg

LETTER

Mutation and linkage disequilibrium in human mtDNA

Molecular evolutionary biologists found a treasure in mitochondrial DNA (mtDNA) for reconstruction of the historical relationships between groups.[1] Because it appears that mtDNA does not undergo recombination and is maternally inherited, examination of different mtDNA sequences appears to faithfully record the history of sequential mutation over time in maternal lineages. In addition, the lack of recombination is fundamental in the forensic application of mtDNA and is critical in understanding the transmission of diseases due to mtDNA mutations.[2]

In recent years, there have been reports of exceptions to this dogma, including a variety of inheritance patterns in plants and invertebrates, but it appeared that the conventional assumptions about mtDNA were still true in vertebrates. However, a recent report[3] suggested that a negative relation between a measure of linkage disequilibrium (r^2) and physical distance could mainly be due to recombination. Because this is potentially an extremely significant finding, it is important to examine in detail the basis of these conclusions.

The recombinational explanation appears to be accepted by some researchers[4–7] although others have pointed various problems with this interpretation.[8–11] On the other hand, mutation is known to occur at a high rate in mammalian mtDNA and transitional variants are produced at a greater than 10-fold higher rate than transversional variants. Therefore, if these observations[3] can be easily be accounted for by mutation, as we will show below, then there is no reason to resort to a recombinational explanation. We will first present a background discussion of measures of linkage disequilibrium and then concentrate on the mutational explanation for the patterns observed.

Linkage disequilibrium is the amount of statistical association of alleles (nucleotides) at different loci (sites). A number of evolutionary factors – selection, genetic drift, gene flow, mutation, and genetic hitchhiking – can generate linkage disequilibrium.[12] However, for tightly linked, neutral loci, a balance between the generation of linkage disequilibrium, primarily by genetic drift (and mutation), and its reduction by recombination, is generally assumed. Mutation may play a dual role and can both generate disequilibrium by new mutations and reduce disequilibrium through recurrent mutation, but the former is of greater significance in short-term evolutionary history.

That the association of alleles at different loci, linkage disequilibrium, approaches random proportions asymptotically by recombination was first noted in 1909.[13] The simplest measure of linkage disequilibrium for two biallelic loci, D, is the difference between the observed and expected frequencies of a gamete (haplotype). Let us assume that there are two alleles at each of two loci and that the most frequent allele at locus A is designated A_1 (with frequency p_1) and the most frequent allele at locus B is designated B_1 (with frequency q_1). There are four different gametes, which have the observed frequencies, x_i, as given below. The expected frequencies of the gametes are the product of the frequencies of the constituent alleles. For example, the expected frequency of gamete A_1B_1 is p_1q_1. The deviation between the observed and expected frequencies of this gamete is

$$D = x_1 - p_1q_1$$

and is the basic linkage disequilibrium measure[14] (this can also be calculated as

$$D = x_1x_4 - x_2x_3.$$

	A_1	A_2	
B_1	$x_1 = p_1q_1 + D$	$x_3 = p_2q_1 - D$	q_1
B_2	$x_2 = p_1q_2 - D$	$x_4 = p_2q_2 + D$	q_2
	p_1	p_2	

However, the largest value that D can take is a function of the allele frequencies at the two loci and its maximum (minimum) of 0.25 (-0.25) is only possible when all alleles have frequencies of 0.5. As a result, a standardized modification of D

$$D' = D/D_{max}$$

was proposed that has the same maximum, 1 or -1, for all combinations of allele frequencies, where D_{max} is the maximum disequilibrium possible for the given allele frequencies.[15] When $D>0$, then D_{max} is the lesser of p_1q_2 and p_2q_1 and when $D<0$, then D_{max} is the lesser of p_1q_1 and p_2q_2 (we will use the absolute value of D' below).

Theoretical developments examining the impact of genetic drift on linkage disequilibrium[16] used another 'standardized' measure

$$r^2 = \frac{D^2}{p_1p_2q_1q_2}.$$

Like D, r^2 does not have the same maximum for all combination of allele frequencies and only has a maximum of 1 if the two loci have the same allele frequencies[17] and if there is positive association in gametes (haplotypes) between the most common alleles at the two loci (see below for situations when there is a negative association). Interestingly,

if we define $r'=r/r_{max}$, where r_{max} is the maximum possible value of r for the given allele frequencies, then

$$r' = \frac{D/(p_1p_2q_1q_2)^{1/2}}{D_{max}/(p_1p_2q_1q_2)^{1/2}} = D'.$$

Therefore, if r^2 is standardized by the maximum r^2 value for the given allelic frequencies, then it is equal to $(D')^2$,[9].

Because there has been no direct demonstration of recombination for vertebrate mtDNA[9] and there are now several large data sets[18,19] that show no evidence of recombination for any of the measures of disequilibrium, we assume that unique (single) mutations generate different gametes and then that genetic drift may increase their frequencies. Let us assume that gamete A_1B_1 is the ancestral gamete and gametes A_1B_2 and A_2B_1 are produced by independent mutations from B_1 to B_2 and A_1 to A_2, respectively.

	A_1	A_2	
B_1	x_1	$x_3 = q_1 - x_1$	q_1
B_2	$x_2 = p_1 - x_1$	0	q_2
	p_1	p_2	

The sum of the gametic frequencies is $x_1+x_2+x_3=1$ and after substitution, $x_1=p_1+q_1-1$. Assuming that the ancestral alleles A_1 and B_1 are the most frequent and their frequencies are equal to p_1 and q_1, then

$$D' = -(p_1 - x_1)(q_1 - x_1)/p_2q_2 = -1.$$

In other words, D' gives the maximum absolute disequilibrium possible for these allele frequencies, suggestive of the lack of recombination.

On the other hand, for the above gametic array

$$r^2 = \frac{[-(p_1 - x_1)(q_1 - x_1)]^2}{p_1p_2q_1q_2} = \frac{p_2q_2}{p_1q_1}.$$

For example, when x_1=0.6, x_2=0.24, x_3=0.16, and x_4=0, then r^2=0.06. In this case, only mutation has been responsible for the gametic array and the low r^2 value does not reflect past recombination.

It has been argued that r^2 has more 'power' to detect lack of recombination than D'.[6] However, in cases like this, low r^2 values may falsely suggest that recombination has occurred when mutation is a simpler explanation. A better approach to determine how likely it would be to observe the maximum D' for given allele frequencies is to calculate the probability of observing an equal or more extreme two-locus association by chance alone (P), Fisher's exact test.[20] The exact probability depends upon sample size and if we use the same gametic frequencies as above, then for total sample sizes of 25, 50, and 100, P=0.54, 0.17, and 0.01, respectively. In other words, the statistical significance of this low (0.06) value of r^2, suggested to indicate recombination, is highly dependent upon the sample size.

Now let us assume again that the various observed sequences[3] are generated by mutation and construct a neighbour-joining, phylogenetic tree for the 14 variable sites analysed from the 45 (22 unique sequences for these 14 sites) sequences analysed[3] as given in Figure 1.[10] As examples of linkage disequilibrium, let us examine some instructive site pairs on this tree. There are three pairs of closely linked sites in the cytochrome b gene, positions 14783, 15043, and 15301, with very high r^2 values.[21] For example, for the site pair 14783–15043, there are 35 TG, 0 TA, 0 CG, and 9 CA sequences (Table 1a). All of the measures of linkage disequilibrium for these gametes reflect high disequilibrium. However, all 9 CA sequences can be the result of two unique transition mutations (T to C at 14783 and G to A at 15043) that are both present in the adjacent groups of six and three sequences at the bottom right of the phylogenetic tree (the other two sites pairs, 14783–15301 and 15043–15301, have nearly the same pattern). In other words, the high disequilibrium values can be both related to the phylogenetic history of the sequences and a function of the high number of identical sequences.

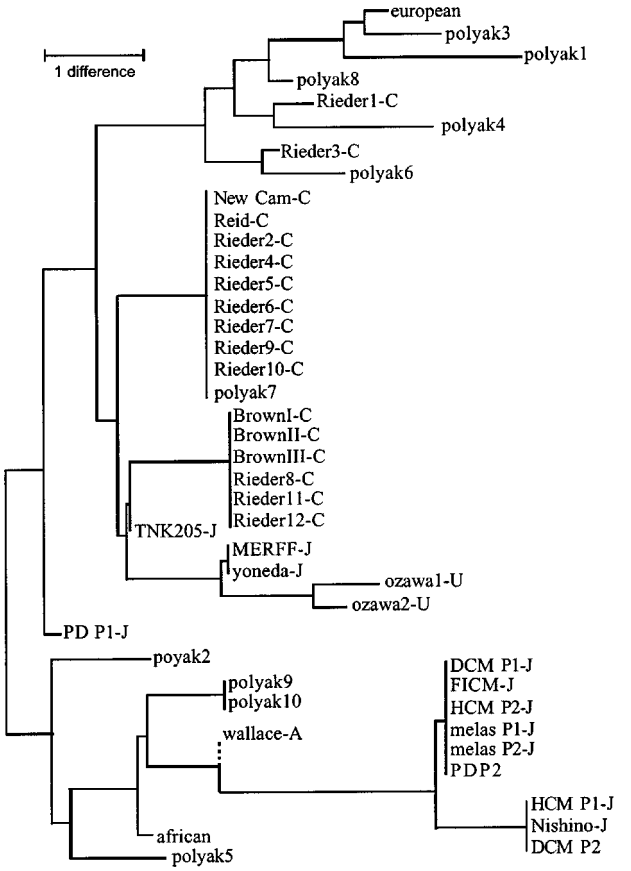

Figure 1 The neighbour-joining tree based on the 14 variable sites in the 45 human mtDNA sequences.[3] Note that the scale indicates one nucleotide difference and that many of the sequences are represented multiple times. The wallace-A sequence is connected to the tree with a dotted line because it had missing data at three sites.[10]

A group of site pairs that differs between the linkage disequilibrium measures is a set of six site pairs approximately 8000 bp apart. These site pairs again involve the three sites in close proximity above and two other sites, 6455 and 7028, on the opposite side of the molecule. For example, for the site pair 7028 – 14783, there are 24 TT, 9 TC, 11 CT, and 0 CC sequences (Table 1b). In this case, $|D'|$ indicates high disequilibrium while r^2 indicates low disequilibrium. The exact probability is nearly significant, suggesting that there is relatively high disequilibrium. Further, from examination of the phylogenetic tree, the low disequilibrium indicated by r^2 is consistent with a mutational explanation. In this case, a unique mutation at 14783 (T to C as before in the two adjacent groups of six and three sequences) and two independent transition mutations at 7028 (T to C in the group of 10 sequences and the poyak2 sequence) produce the gametic array in Table 1b.

Another informative example is for the distant site pair 6455 – 14783 in which there are 33 CT, 6CC, 2 TT, and 3 TC sequences (Table 1c). In this case, $|D'|$ is half the maximum possible and r^2 is again low but the exact probability is significant. If we examine the phylogenetic tree, we find that all the 6 CC sequences are found in the lower branch connecting to a clade of six haplotypes, the 2 TT sequences are in the adjacent ozawa1-U and ozawa2-U sequences, and the 3 TC sequences are in the lower branch of three. In other words, a single mutation at 14783 and two independent transitional mutations at 6455, one leading to the ozawa sequences and one leading to the branch of three, provide a simple explanation for this gametic array. Here again the exact probability is a better indicator of the association between sites than r^2.

If recombination were significant, then one would predict that the two sites, 6455 and 7028 (Table 1d), only 574 nucleotides apart, would have higher linkage disequilibrium than site 14783, on the other side of the molecule, and either of this sites. Contrary to this prediction, r^2 between these close sites is only 0.04, falsely suggesting recombination, while $|D'|=1$. In fact, the two closest of all pairs of sites, 11251 and 11299, only 49 nucleotides apart, have $r^2=0.02$ while again $|D'|=1$. If one were to standardize r^2 by the maximum possible value for the given allele frequencies, as shown above, then for both these situations $(r')^2=(D')^2=1$.

The other four distant site pairs with the three sites in close proximity and either 6455 or 7028 also produce low r^2 values and either intermediate to maximal $|D'|$ values. In other words, the low r^2 values between these distance sites pairs suggest that mutation, either single or double, can generate low r^2 values. Overall, mapping nucleotide substitutions on the phylogenetic tree reveals single transitional changes at four sites, two independent transitional changes at eight sites, and single, followed by backward, transitional changes at two sites.

In addition to the argument of a negative association of linkage disequilibrium and distance, an excess of homoplasmic sites has been cited as support for mtDNA recombination.[7] In the new data set in which this was examined,[18] a significant excess of homoplasmic sites was not found. Further, in this study,[18] it was suggested that in the earlier data[3] ''there are substantial uncertainties and complexities that influence the estimation of both the expected and the observed number of homoplasies''.

Overall, the pattern of linkage disequilibrium observed[3] is entirely consistent with mutation as an explanation. Before recombination is accepted as important in human (vertebrate) evolution, either direct evidence from family studies, observation of mtDNA sequences that could only be considered recombinant, or definitive evidence from population studies, in which mutation can be excluded as an explanation, are necessary. Of course, if recombination does occur but at a low rate relative to mutation, then any signal in the amount of linkage disequilibrium from recombination would be obscured by mutation. However, recent evidence is already mounting against mtDNA recombination from population studies.[18,19]

Table 1 Four examples of site pairs from the 45 human mtDNA sequence[3] that illustrate (*a*) similarity between the linkage disequilibrium measures and difference, (*b*) – (*d*), between the measures. Only 44 sequences are given for the first three site pairs because wallace-A did not have information for site 14783.

(a)		Site 14783			
		T	C		$\|D'\|=1.0$
Site 15043	G	35	0	0.80	$r^2=1.00$
	A	0	9	0.20	$P<0.0001$
		0.80	0.20		
(b)		Site 7028			
		T	C		$\|D'\|=1.0$
Site 14783	T	24	11	0.80	$r^2=0.09$
	C	9	0	0.20	$P=0.08$
		0.75	0.25		
(c)		Site 6455			
		C	T		$\|D'\|=0.5$
Site 14783	T	33	2	0.80	$r^2=0.12$
	C	6	3	0.20	$P=0.05$
		0.89	0.11		
(d)		Site 6455			
		C	T		$\|D'\|=1.0$
Site 7028	T	28	5	0.73	$r^2=0.04$
	C	12	0	0.27	$P=0.31$
		0.67	0.33		

Philip Hedrick and Sudhir Kumar
Department of Biology, Arizona State University, Tempe, AZ 85287, USA

Acknowledgments

We would like to thank T Dowling, R Lewontin and M Stoneking for discussions on this topic. P Hedrick has been supported by the Ullman Professorship and S Kumar by NIH, NSF, and Burroughs-Wellcome Fund grants for this research.

References

1 Avise J: Phylogeography: the history and formation of species. Harvard University Press, Cambridge, MA, 2000.
2 Howell N: Human mitochondrial disease: answering questions and questioning answers. *Int Rev Cytol* 1997: **186**: 49 – 116.
3 Awadalla P, Eyre-Walker A, Maynard Smith J: Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science* 1999: **286**: 2524 – 2525.
4 Hey J: Human mitochondrial DNA recombination: can it be true? *Trends Ecol Evol* 2000: **15**: 181 – 182.
5 Strauss E: mtDNA shows signs of paternal influence. *Science* 1999: **286**: 2436.
6 Awadalla P, Eyre-Walker A, Maynard Smith J: Questioning evidence for recombination in human mitochondrial DNA. *Science* 2000: **288**: 1931a.
7 Eyre-Walker A: Do mitochondria recombine in humans? *Phil Trans Roy Soc Lond B* 2000: **355**: 1573 – 1580.
8 Kivisild T, Villems R: Questioning evidence for recombination in human mitochondrial DNA. *Science* 2000: **288**: 1931a.
9 Jorde LB, Bamshad M: Questioning evidence for recombination in human mitochondrial DNA. *Science* 2000: **288**: 1931a.
10 Kumar S, Hedrick PW, Dowling T, Stoneking M: Questioning evidence for recombination in human mitochondrial DNA. *Science* 2000: **288**: 1931a.
11 Parsons TJ, Irwin JA: Questioning evidence for recombination in human mitochondrial DNA. *Science* 2000: **288**: 1931a.
12 Hedrick PW: Genetics of populations (2nd ed) Jones and Bartlett, Boston, 2000.
13 Weinberg W: Uber Vererbungsgesetze beim Menschen. *Z Abst V Vererb* 1909: **1**: 277 – 330.
14 Lewontin RC, Kojima K: The evolutionary dynamics of complex polymorphisms. *Evolution* 1960: **14**: 450 – 472.
15 Lewontin RC: The interaction of selection and linkage. 1. General considerations; heterotic models. *Genetics* 1964: **49**: 49 – 67.
16 Hill WG, Robertson A: Linkage disequilibrium in finite populations. *Theoret Appl Genet* 1968: **38**: 226 – 231.
17 Hedrick PW: Inference of recombinational hotspots using gametic disequilibrium values. *Heredity* 1988: **60**: 435 – 438.
18 Elson, JL, Andrews, RM, Chinnery PF, Lightowlers, RN, Turnbull DM, Howell N: Analysis of European mtDNAs for recombination. *Am J Hum Genet* 2001: **68**: 145 – 153.
19 Ingman M, Kaessmann H, Paabo S, Gyllensted U: Mitochondrial genome variation and the origin of modern humans. *Nature* 2000: **408**: 708 – 713.
20 Lewontin RC: The detection of linkage disequilibrium in molecular sequence data. *Genetics* 1995: **140**: 377 – 388.
21 Kumar S: LDA: Linkage Disequilibrium Analysis, Version 1.0, Arizona State Univ., Tempe, AZ, 2000.