Can Indirect Tests Detect a Known Recombination Event in Human mtDNA?

Daniel James White*1 and Neil John Gemmell†

*Department of Biological Sciences, University of Canterbury, Christchurch, New Zealand; and †Centre for Reproduction and Genomics, Department of Anatomy and Structural Biology, University of Otago, Dunedin, New Zealand

Whether human mitochondrial DNA (mtDNA) recombines sufficiently to influence its evolution, evolutionary analysis, and disease etiology, remains equivocal. Overall, evidence from indirect studies of population genetic data suggests that recombination is not occurring at detectable levels. This may be explained by no, or low, recombination or, alternatively, current indirect tests may be incapable of detecting recombination in human mtDNA. To investigate the latter, we have tested whether six well-established indirect tests of recombination could detect recombination in a human mtDNA data set, in which its occurrence had been empirically confirmed. Three showed statistical evidence for recombination ($r^2$ vs. distance, the Homoplasy test, Neighborhood Similarity Score), and three did not ($D^2$ vs. distance, Max Chi Squared, Pairwise Homoplasy Index). Possible reasons for detection failure are discussed. Further, evidence from earlier studies suggesting a lack of recombination in mtDNA in humans is reconsidered, taking into account the appropriateness of the tests used, based on our new findings.

Whether recombination occurs in human mitochondrial DNA (mtDNA) at a level significant enough to alter the evolution of the molecule, and evolutionary analysis based on mtDNA, remains contentious (Slate and Gemmell 2004; reviewed by White et al. 2008). At present, only one definitive case of mtDNA recombination has been detected using direct means (Kraysberg et al. 2004), in a patient suffering from a mitochondrial myopathy, although the occurrence of tetraplasmy in 10 individuals who carried mutations in both coding and control regions of mtDNA (Zsurka et al. 2005), suggests recombination may be more widespread in somatic tissue. Further, an indirect observation suggests that mtDNA recombinants can reach the germ line in humans, thereby potentially becoming inheritable (Zsurka et al. 2007).

Currently, evidence for mtDNA recombination in the germ and soma lines of humans, using direct means, remains absent or scarce, respectively. An alternative approach is the use of indirect tests, and numerous authors have sought to detect recombination in this way (Awadalla et al. 1999; Eyre-Walker et al. 1999; Wiuf 2001; Piganeau and Eyre-Walker 2004). Indirect tests analyze the distribution of polymorphisms in DNA sequence, and assess the probability that the patterns observed can be explained by crossover events. Despite a wide range of tests utilizing a variety of statistical frameworks (see http://www.bioinf.-manchester.ac.uk/recombination/programs.shtml for an up to date list), indirect evidence for widespread recombination in humans remains equivocal (Jorde and Bamshad 2000; Kivisild and Villems 2000; Elson et al. 2001; Wiuf 2001; Innan and Nordborg 2002; Piganeau and Eyre-Walker 2004). However, mtDNA recombination has been detected in animals in similar studies (Piganeau et al. 2004; Tsousis et al. 2005), raising the possibility that the lack of evidence for recombination in human mtDNA is not due to an absence of recombination per se, but rather the ineffectiveness of current tests to detect it.

The power of many of these tests has been assessed previously using sequence data, simulated under various parameters chosen for the purpose of the study (Posada and Crandall 2001). However, mtDNA evolves under a unique set of parameters, and it remains unclear how effective the available indirect tests are at detecting recombination in mitochondrial sequence. For example, mtDNA is circular and its mutation rate heterogeneously distributed across the genome (Yang 1996), both of which may strongly influence correlation measures of detecting recombination in mtDNA (Wiuf et al. 2001). In addition, when compared with nuclear DNA, mtDNA mutation rates are higher (Ingman et al. 2000; Mishmar et al. 2003) and can be correlated (Nielsen 1997). Further, relative to animals, human mtDNA sequence diversity is low (Jorde et al. 1998). Thus, it is entirely possible that even if recombination in mtDNA occurs naturally in humans, indirect tests currently available will be unable to detect it.

In the only empirically confirmed occurrence of mtDNA recombination in humans, Kraysberg et al. (2004) recovered 33 recombinant maternal molecules from 450 single-molecule polymerase chain reaction clones, derived from the muscle tissue of an individual known to possess a 10:1 excess of paternal-to-maternal mtDNA. The availability of these sequence data has enabled us, for the first time, to assess empirically the effectiveness of indirect tests at detecting recombination in human mtDNA. The recombination tests were selected for evaluation based on previous inclusion in both successful, and unsuccessful, studies of recombination detection. They include three population based tests; the correlation between physical distance and linkage disequilibrium (LD) using both the $D^2$ and $r^2$ measures of LD (Awadalla et al. 1999) and the Homoplasy test (Smith and Smith 1998). Three general methods were also assessed: Max Chi Squared (Smith 1992), Neighborhood Similarity Score (NSS, Jakobsen and Easteal 1996), and the more recently developed Pairwise Homoplasy Index (PHI, Bruen et al. 2006).

Of these six tests of recombination, three were not able to detect recombination when we know it happened: the Max Chi Squared, PHI, and LD using $D^2$ tests. Results from our correlations between LD and physical distance were mixed however as LD versus $r^2$ did show a significant negative correlation, indicative of recombination. Recombination was also detected by the NSS and Homoplasy tests. As

---

1 Present address: Centre for Reproduction and Genomics, Department of Anatomy and Structural Biology, University of Otago, P.O. Box 913, Dunedin, New Zealand.

Key words: recombination, indirect, detection, mtDNA, human.

E-mail: daniel.white@otago.ac.nz.

doi:10.1093/molbev/msp073

Advance Access publication April 15, 2009

© The Author 2009. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.

For permissions, please e-mail: journals.permissions@oxfordjournals.org
these results come from one sample, it is possible they are not the most statistically probable, assuming $P$ values for each test are normally distributed for this type of data set but rather represent less likely outcomes due to sampling (Innan and Nordborg 2002). To control for this, all six data sets were reevaluated using 100 samples of simulated sequence, generated using parameters that re-created the level sets were reevaluated using 100 samples of simulated sequence used in these studies. The Kraytsberg nucleotide diversity of the Kraytsberg data set, $\pi = 1.25 \times 10^{-3}$, in comparison with those used in earlier simulation studies, may preclude efficient detection of recombination with these methods. Both methods have been shown to perform less well with low sequence divergence (Posada and Crandall 2001; Wiuf et al. 2001; Bruen et al. 2006), and the Kraytsberg $\pi$ is lower again than the least divergent simulated sequences used in these studies. The Kraytsberg nucleotide divergence at $1.25 \times 10^{-3}$ substitutions per site per generation is, however, within the range of values predicted for human populations ($6.0 \times 10^{-5}$ to $6.8 \times 10^{-3}$, Ingman et al. 2000). Taken together, these findings suggest the diversity of the Kraytsberg data set, and perhaps human mtDNA in general, may be too low for PHI and Max Chi Squared to work effectively.

The failure of the $D'$ method to detect recombination is less surprising, as it has been shown to be less powerful than $r^2$ at detecting recombination under the finite sites model (McVean et al. 2002), a model that describes well the distribution of polymorphisms at synonymous sites in human mtDNA (Piganeau and Eyre-Walker 2004). As such, the $r^2$ measure has previously been recommended for recombination detection in human mtDNA, using synonymous sites (Piganeau and Eyre-Walker 2004). Our result strongly supports this recommendation.

Recombination in human mtDNA could have substantial impact on mitochondrial disease etiology, transmission, and inheritance. Further, if unaccounted for, mitochondrial recombination may lead to inaccurate inferences drawn from evolutionary analysis (Schierup and Hein 2000a, 2000b; Posada and Crandall 2002). Although there now exists compelling indirect evidence of widespread recombination in animals (Piganeau et al. 2004; Tsaouisis et al. 2005), perhaps as a result of greater diversity in animal mtDNA and older mtDNA lineages leading to more opportunities for recombination to occur (Jorde et al. 1998), for humans results thus far are equivocal (table 2).

Results provided here suggest one reason for the apparent lack of indirect evidence of recombination in human mtDNA may derive from an inability of many of the contemporary tests being used to detect naturally occurring recombination in human mtDNA. If we were to revise results of previous indirect studies, based on our findings, the landscape of evidence for recombination in human mtDNA alters. In table 2, we summarize the majority of studies thus far of human mtDNA recombination, using indirect statistical approaches. If we ignore the negative findings from tests that we have shown to be unable to detect recombination in the Kraytsberg mtDNA sequence data, in cases where recombination was detected by another method, the proportion of positive results changes from 20.3% to 23.5%. If we remove these test results from all studies, the proportion of data sets that give a positive result increases from 20.3% to 31.6%. Further development of these

---

**Table 1**

*P* Values Obtained from Various Indirect Tests Using a Real Data Set

<table>
<thead>
<tr>
<th>Test</th>
<th>$P$ Value</th>
<th>% Simulations $^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D'$</td>
<td>0.959 (0.936)</td>
<td>44</td>
</tr>
<tr>
<td>$r^2$</td>
<td>&lt;0.001 (0.002)</td>
<td>94</td>
</tr>
<tr>
<td>Homoplasy</td>
<td>&lt;0.001 (&lt;0.001)</td>
<td>99</td>
</tr>
<tr>
<td>Max Chi Sqd</td>
<td>0.061</td>
<td>82</td>
</tr>
<tr>
<td>NSS</td>
<td>0</td>
<td>95</td>
</tr>
<tr>
<td>PHI $p^{b}$</td>
<td>0.524</td>
<td>94</td>
</tr>
<tr>
<td>PHI $n$</td>
<td>0.392</td>
<td>95</td>
</tr>
</tbody>
</table>

$^a$ Numbers in brackets refer to $P$ values obtained using synonymous sites only.

$^b$ For PHI, $P$ values were obtained after randomly permuting positions of sites ($p$), as well as from estimations based on the assumption of a normal distribution of PHI values ($n$).

---

Properties, for example, high auto-correlated mutation rates (Yang 1995; Nielsen 1997), among-site heterogeneity in mutation rates (Yang 1996) and a complicated relationship between physical and genetic distance due to its circular nature (Wiuf 2001), the findings from our study are likely to be different to those previously published. In particular, the low, presumably more realistic (Ingman et al. 2000) nucleotide diversity of the Kraytsberg data set, $\pi = 1.25 \times 10^{-3}$, in comparison with those used in earlier simulation studies, may preclude efficient detection of recombination with these methods. Both methods have been shown to perform less well with low sequence divergence (Posada and Crandall 2001; Wiuf et al. 2001; Bruen et al. 2006), and the Kraytsberg $\pi$ is lower again than the least divergent simulated sequences used in these studies. The Kraytsberg nucleotide divergence at $1.25 \times 10^{-3}$ substitutions per site per generation is, however, within the range of values predicted for human populations ($6.0 \times 10^{-5}$ to $6.8 \times 10^{-3}$, Ingman et al. 2000). Taken together, these findings suggest the diversity of the Kraytsberg data set, and perhaps human mtDNA in general, may be too low for PHI and Max Chi Squared to work effectively.
Table 2
A Summary of Indirect Studies of Recombination Detection in Human mtDNA

<table>
<thead>
<tr>
<th>Study</th>
<th>Data Set</th>
<th>Test/s Used</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Awadalla et al. (1999)</td>
<td>45 mtDNA sequences, global</td>
<td>$r^2$</td>
<td>0.012</td>
</tr>
<tr>
<td>Eye-Walker et al. (1999)</td>
<td>29 mtDNA seqs, global</td>
<td>H</td>
<td>0 (effectively)</td>
</tr>
<tr>
<td>Ingman et al. (2000)</td>
<td>53 mtDNA seqs from 14 of the major</td>
<td>$D'$ $r^2$</td>
<td>ns ns</td>
</tr>
<tr>
<td></td>
<td>linguistic phyla</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jorde and Bamshad (2000)</td>
<td>Awadalla data sets plus four</td>
<td>$D'$ $r^2$</td>
<td>ns 2/8 data sets significant (&lt;0.03 and &lt;0.02)*</td>
</tr>
<tr>
<td></td>
<td>European data sets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kumar et al. (2000)</td>
<td>Awadalla data sets</td>
<td>$D'$ FET</td>
<td>ns ns</td>
</tr>
<tr>
<td>Elson et al. (2001)</td>
<td>64 British mtDNAs and two Africans</td>
<td>$D'$ $r^2$ $d$</td>
<td>ns ns ns</td>
</tr>
<tr>
<td>Herrnstadt et al. (2002)</td>
<td>56 African</td>
<td>$r^2$ $d$</td>
<td>ns ns ns</td>
</tr>
<tr>
<td></td>
<td>69 Asian</td>
<td>$r^2$ $d$</td>
<td>sig sig</td>
</tr>
<tr>
<td></td>
<td>435 European</td>
<td>$r^2$ $d$</td>
<td>ns ns ns</td>
</tr>
<tr>
<td></td>
<td>226 Haplogroup H</td>
<td>$r^2$ $d$</td>
<td>ns ns ns</td>
</tr>
<tr>
<td>Piganeau and Eyre-Walker</td>
<td>Awadalla data set</td>
<td>$r^2$ $D'$ H GC MS</td>
<td>ns ns 0.0001 ns ns</td>
</tr>
<tr>
<td>(2004)</td>
<td>Ingman data set</td>
<td>$r^2$ $D'$ H GC MS</td>
<td>ns ns 0.0052 ns ns</td>
</tr>
<tr>
<td></td>
<td>192 Finnish seqs</td>
<td>$r^2$ $D'$ H GC MS</td>
<td>&lt;0.05 ns ns ns ns</td>
</tr>
<tr>
<td></td>
<td>Herrnstadt African</td>
<td>$r^2$ $D'$ H GC MS</td>
<td>&lt;0.05 ns 0.04 ns ns</td>
</tr>
<tr>
<td></td>
<td>Herrnstadt Asian</td>
<td>$r^2$ $D'$ H GC MS</td>
<td>ns ns 0.01 ns ns ns</td>
</tr>
<tr>
<td></td>
<td>Herrnstadt European</td>
<td>$r^2$ $D'$ H GC MS</td>
<td>na na 0.0001 ns ns</td>
</tr>
</tbody>
</table>

Global refers to globally distributed.

ns: P values > 0.05; sig: P values ≤ 0.05; na: not available; $r^2$: LD versus distance using $r^2$ with a random permutation step to estimate statistical significance; $D'$: LD versus distance using $D'$ with a random permutation step to estimate statistical significance; d: LD versus distance using d with a random permutation step to estimate statistical significance; H: homoplasy test; GC: Geneconv; and MS: Max Chi Squared.

* Both data sets from Awadalla et al. (1999).
* These results are for polymorphisms in their data set with no minimum minor allele frequency cutoff.
* From Finnskjaer et al. (2001).

Tests may be required for optimization in mtDNA, but for now, we have highlighted some tests of choice. Although NSS detected recombination here, it has been suggested that precaution be taken when using this test to detect recombination in mtDNA, especially in divergent sequences, due to an increase in false positive generation as a consequence of substitution rate correlation (Bruen et al. 2006), a trait known to mtDNA (Yang 1995). Therefore, the best performing of the available options tested on the Kraytsberg data set, and presumably other human mtDNA data sets too, are LD versus distance using $r^2$ and the Homoplasy test.

Materials and Methods

For the 33 recombinant molecules recombinants, plus representative parental molecules, Kraytsberg et al. (2004) obtained 6,854 bp of mtDNA sequence (fig. 1), which was kindly made available to us by Konstantin Khrapko (Yale University) for use in our recombination detection tests. Codon positions of the polymorphic sites are summarized in figure 1. One hundred samples of simulated sequence (length 6,854 bp) were evolved in TREEVOLVE (Grassly et al. 1999), with the level of recombination set to the recombination parameter of the Kraytsberg data set (rho = 5.492), as estimated in LAMARC (Kuhner 2006), and a mutation rate per site per generation of $5 \times 10^{-7}$.

Indirect Tests of Recombination

Theory predicts that the strength of LD between two alleles should decrease with physical distance in the presence of recombination (Hill and Robertson 1968). Thus, a significant negative correlation between LD and distance could indicate recombination. There are several measures of LD, and two have been used for assessing recombination in mtDNA: $r^2$ (Hill and Robertson 1968) and $D'$ (Lewontin 1964). Both are assessed here.

The homoplasy test investigates whether more homoplasies (co-occurrence of a polymorphism on separate branches of a phylogenetic tree) occur in the most parsimonious tree, than expected by recurrent mutation under a model of clonal inheritance (Smith and Smith 1998). If an excess of homoplasies is detected, then recombination may be the most likely explanation. The incidence of homoplasy was investigated in the data using the Homoplasy Test, implemented in Linux using a C translation of the original QBASIC version, kindly provided by David Posada (University of Vigo).

The Max Chi Squared method compares the arrangement of segregating sites between two sequences either side of a putative recombination break point, with all other sequences in the alignment (Smith 1992). A sliding window of sequence to analyze was set to the number of polymorphic sites divided by 1.5. Significance was estimated as the proportion of 1,000 permuted test scores, calculated after randomly positioning the informative sites, below the observed test score.

NSS describes the extent of clustering of compatibilities (either compatible or incompatible) of adjacent informative sites in a sequence alignment (Jakobsen and Eastal 1996). Higher NSS values than expected by chance can be explained by recurrent mutation, gene conversion, or recombination. Significance of the observed NSS is achieved by randomly permuting the order of informative sites 1,000 times and determining the fraction of random scores that are at least as high as the observed data.
The PHI measures the mean refined incompatibility score between sites within a window of sequence of preset length and reflects the minimum number of homoplasies on any tree required to describe the genealogical history of a pair of sites (Bruen et al. 2006). If recombination is responsible for homoplasies, PHI scores should be lower than if recurrent mutation is responsible, as recurrent mutation is not correlated with physical distance. An estimate of the statistical significance of the PHI was achieved by randomly permuting site positions in the alignment (simulating no recombination) 1,000 times, and calculating the proportion of times the permuted PHI score is less than or equal to the observed score.

Acknowledgments

We thank Konstantin Khapko for the data set and David Posada for the C version of the Homoplasy Test. We are also indebted to Dr Vladimir Mencl and Iris Vargas Jentzsch for their patient, computer support. We thank Hideki Innan and two anonymous reviewers for their comments, which improved this manuscript. This study was funded by the Royal Society of New Zealand, Marsden Fund Grant UOC402.

Literature Cited


