

Manuscript MBE-07-0187

Revised Version / 18 June 2007

Relative Rates of Evolution in the Coding and Control Regions of African mtDNAs

*Neil Howell, *† Joanna L. Elson,‡ Corinna Howell,§ and Douglass M. Turnbull‡*

*MIGENIX Corp., San Diego, California; †Department of Radiation Oncology, The University of Texas Medical Branch, Galveston; ‡Mitochondrial Research Group, School of Neurology, Neurobiology, and Psychiatry, The Medical School, The University of Newcastle upon Tyne, Newcastle upon Tyne, United Kingdom; §MitoKor Inc., San Diego, California

Address for correspondence and reprints: Dr. Neil Howell, MIGENIX Inc., Suite 210, 12780 High Bluff Drive, San Diego CA 92130

Email: NHowell@migenix.com

Telephone: 858/509-5616

FAX: 858/793-7805

Keywords: mitochondrial DNA, molecular clock, phylogenetic analysis, selection, molecular evolution

RUNNING TITLE: Evolution in African mtDNA Coding and Control Regions

Abstract

Reduced median networks of African haplogroup L mtDNA sequences were analyzed to determine the pattern of substitutions in both the non-coding control and coding regions. In particular, we attempted to determine the causes of the previously reported (Howell et al. 2004) violation of the molecular clock during the evolution of these sequences. In the coding region, there was a significantly higher rate of substitution at synonymous sites than at non-synonymous sites, as well as in the tRNA and rRNA genes. This is further evidence for the operation of purifying selection during human mtDNA evolution. For most sites in the control region, the relative rate of substitution was similar to the rate of neutral evolution (assumed to be most closely approximated by the substitution rate at 4-fold degenerate sites). However, there are a number of mutational hotspots in the control region, ~3% of the total sites, that have a rate of substitution greater than the neutral rate, at some sites by more than an order of magnitude. It is possible either that these sites are evolving under conditions of positive selection, or that the substitution rate at some sites in the control region is strongly dependent upon sequence context. Finally, we obtained preliminary evidence for “non-ideal” evolution in the control region, including haplogroup-specific substitution patterns and a decoupling between relative rates of substitution in the control and coding regions.

Introduction

The non-coding control region (also called the D-loop) of human mitochondrial DNA (mtDNA) continues to be a major source of sequence information for phylogeographic analysis and to be widely used for dating major events in human evolution and population dispersal (Helgason et al. 2006; Kayser et al. 2006; Ohashi et al. 2006; Pakendorf et al. 2006; Batini et al. 2007). However, sequence evolution in this mtDNA region is complex and the control region is unlikely to be a reliable molecular clock or to yield an unambiguous phylogeographic signal (Torrioni et al. 2006). It has been posited for two decades that the control region has a high rate of sequence divergence, both in absolute terms and relative to the rate in the coding region (Cann et al. 1987). Moreover, the control region displays marked site heterogeneity in the rates of nucleotide substitution, including the presence of hypervariable sites (Excoffier and Yang 1999; Meyer et al. 1999; Howell and Bogolin Smejkal 2000; Malyarchuk and Rogozin 2004). Furthermore, we have recently shown that the control region in a set of African mtDNAs shows violations of clock-like evolution. Models of sequence evolution that allowed substitution rates to vary among branches produced maximum likelihood trees with better support than trees where substitution was constrained to a single rate for all branches (see pages 1844-1845 of Howell et al. 2004 for further details; see also Ingman et al. 2000, Torrioni et al. 2001 and Gonder et al. 2007). Until now, there has been no specific analysis of the basis of this violation of the molecular clock.

The observed pattern of mtDNA sequence evolution will be determined by a number of factors, including population effects and selection. There is now considerable evidence that purifying selection, and possibly positive selection as well, have affected the evolution of the mtDNA *coding* region (Moilanen and Majamaa 2003; Elson et al. 2004; Loewe 2006; Broughton and Reneau 2006; Kivisild et al. 2006; Meiklejohn et al. 2007). It has been

suggested by one group that climate has been an important factor influencing human mtDNA evolution (Ruiz-Pesini et al. 2004), but that model has been challenged by other investigators (Elson et al. 2004; Kivisild et al. 2006; Sun et al. 2006).

There is little information, in contrast, on the effects of selection during evolution of the control region. It has been recognized for some time that there are segments of the control region that have extremely low mutation rates, a result suggesting the operation of purifying selection (Kocher and Wilson 1991). There is also evidence for an acceleration of the rate of substitution at evolutionarily short timescales in both the mtDNA coding and control regions (Howell et al. 2003; Ho et al. 2005; Penny 2005). While other factors might be involved, at least some of this acceleration appears to be due to purifying selection acting on the control region (Howell et al. 2003; Ho et al. 2005).

Pesole et al. (1999), on the basis of interspecific mtDNA sequence analysis, showed *lower* substitution rates in all three control region segments analyzed, relative to the synonymous substitution rate (see their Figure 1). This result, at first consideration, suggests that there is more purifying selection acting on the control region than on the least constrained positions within the coding region. On the other hand, Finnilä et al. (2001) reported that the rate of substitution at third-position sites was 22% the substitution rate in HVR1. In this report, we follow up on this type of analysis and ask the deceptively simple question: what is the rate of substitution in the control region relative to the *neutral* rate of substitution in the coding region of human mtDNA? If the former rate is significantly less than the latter, one explanation is that purifying selection has acted relatively more strongly on the control region. Conversely, if the rate of control region substitution is significantly greater than the neutral rate, then one should consider a role for positive selection (Wong and Nielsen 2004).

An important component of this study is that we specify what assumptions are made for the analysis and we then suggest what areas need to be addressed in future studies. Not

surprisingly, as is often the case for mtDNA evolution, our results are neither simple nor unambiguous. A number of questions emerge that center on the complexities of sequence evolution within the mtDNA control region.

Materials and Methods

mtDNA Sequences Analyzed and Methods of Analysis

For this initial analysis, we utilized the 96 complete African haplogroup L sequences used for the previously published molecular clock analysis (Howell et al. 2004). We recognize that this is a relatively small sequence set. However, for this first-stage analysis, we wanted to avoid the use of a larger sequence set that utilized more diverse mtDNAs (viz, those of European, Asian and African descent). One reason for this decision is that we are better able to detect the effects of selection, if any, in African mtDNAs, which are agreed to be the oldest in terms of human evolution. Furthermore, while there are now African mtDNA sequences that have been published by other groups (Kivisild et al. 2006; Gonder et al. 2007), the sequences of the control regions – a crucial part of our analysis – were not available at the time of our analysis.

As discussed in our previous publication (Howell et al. 2004), and as another reason for analysis of these sequences, we have been cognizant of the problems introduced by sequence errors and substantial effort has been put into quality control. As part of this effort, we have corrected here some minor errors in the haplogroup L networks (Appendix 1).

Assumption #1 – This set of African mtDNAs has a sufficiently low rate of sequence errors that the results and their interpretation have not been compromised.

In this study, substitutions in the control and coding regions were identified from the reduced median networks reported elsewhere (Herrnstadt et al. 2002; Howell et al. 2004). All sequence changes are expressed relative to the revised CRS of the L-strand (Andrews et al. 1999), and insertions and deletions – most of which occur in simple repeat sequences - were omitted from this analysis.

The reduced median networks were constructed from the mtDNA coding regions and control region substitutions were subsequently added in a most parsimonious (MP) fashion

(see the description on pages 1847-1849 in Howell et al. 2004). We cannot guarantee that our approach is perfectly free of bias, and a Reviewer of an earlier version of this report has noted that one can devise scenarios where such coding region-based networks are biased. Sites in the coding and in the control regions might produce, in some instances, different phylogenetic signals. In such cases, constructing networks from coding region sequences, and then adding control region sequences, could potentially introduce bias into the analysis. However, the cumulative results from many research groups provide formidable support for our approach or, at least, against treating all sites as equal when constructing human mtDNA networks.

Previous studies have found that control region networks differ in topology and contain more reticulations, using the same sequence sets, from those based on coding region sequences because of the high frequency of homoplasies within control region sequences (Finnilä et al. 2001). There were only two simple reticulations in our coding region networks (involving L3b and L3f sequences; data not shown). In contrast, control region networks with these haplogroup L sequences contain multidimensional hypercubes that cannot be analyzed or used to “add on” coding region information (data not shown).

The recent work of Non et al. (2007) is especially pertinent for this issue. They analyzed a set of 99 human mtDNA sequences and they found that there was no loss of phylogenetic signal in coding region phylogenetic trees even after deletion of 50% of the coding region. In contrast, a single maximum likelihood tree could not be generated with the HVR1 sequences and the resulting Bayesian tree contained little phylogenetic information with most of the tree in the form of an unresolved polytomy (see their Figure 3). One can “downweight” homoplastic sites (Batini et al. 2007), but this has much the same effect as our approach where control region sites were added to coding region trees under the MP condition, because the phylogenetic signal from highly variable sites – which are of little, if any, use for network construction - is nullified or dampened substantially. Also, weighting schemes

have a disadvantage because they start with assumptions about relative substitution rates that differ according to the analytical approach and sequences analyzed (Excoffier and Yang 1999; Meyer et al. 1999).

An *intraspecific* sequence set was analyzed for two reasons. In the first place, there is evidence that the substitution pathways in different species of primates are not homogeneous (Weiss and von Haeseler 2003), potentially imposing an additional – and unwanted – level of complexity to the analysis. Secondly, because of the presence of highly variable, and even hypervariable, sites in the control region, it is more difficult to accurately estimate the number of substitutions that have occurred in interspecies analyses of mtDNA control region sequences (Perna and Kocher 1995; Torroni et al. 2006).

Assumption #2 – Reduced median networks of coding region sequences, and subsequent use of MP to add control region sites, provide an acceptably accurate and less biased “readout” of the number of substitutions that have occurred in the control and coding regions of this set of African haplogroup L sequences.

For this analysis, substitutions were counted from a root, although the rooting of human mtDNA sequences has been controversial. Following the work of Salas et al. (2002; see especially their Figure 2), we assume here a haplogroup L0a root for the African sequences. There are three coding substitutions and seven control substitutions that occur in two, three or four of the five L0a sequences analyzed (see Table 1 for details). As a result, the root chosen for the analyses is a consensus sequence in which the L0a root sequence carries the non-CRS allele at these ten sites. This consensus L0a sequence was then used as the root for determining the order and number of substitutions in the L0a, L1, L2, and L3 sequences. For this analysis, we consider all substitutions at a site and include both transversions and transitions (see footnote “e” of Table 1 for examples).

Recent studies, published after our analyses were completed (Kivisild et al. 2006; Gonder et al. 2007), conclude that L0d sequences form the most basal clade within human mtDNA phylogenetic trees, but our sequence set did not include an L0d representative. A simple and complete comparison of results with different root sequences is not possible, but – on the basis of the available information – there is no evidence that our conclusions have been compromised by our choice of root sequence. In the first place, neither Kivisild et al. (2006) nor Gonder et al. (2007) provide L0d control region sequences, and a major goal of this work was a comparative analysis of sequence evolution in the control and coding regions. Secondly, there are slight differences in their resulting phylogenetic trees. Kivisild et al. (2006) use both nuclear inserts of mtDNA and a consensus *Pan* mtDNA as root sequences for their phylogenetic analyses, whereas Gonder et al. (2007) use a single *Pan troglodytes* root sequence, thus providing one possible explanation.

Based on the analyses of Gonder et al. (2007; similar results are obtained if we use the phylogenetic tree of Kivisild et al. 2006), L0a and L0d sequences share alleles at 15 sites in the coding region: 1048, 3516, 4312, 5442, 6185, 9042, 9347, 10589, 10664, 10915, 11914, 12007, 12720, 13276 and 14560. There are 12 sites (see our Table 1) that share the same allele in *all* haplogroup L mtDNA sequences (see footnote c), and there is no indication that they differ in the L0d sequences of Kivisild et al. (2006) and Gonder et al. (2007). Furthermore, none of the other 20 haplogroup L-associated substitutions are shown on the L0d networks, so we may presume that these also are shared by our L0a sequences and their L0d sequences. Thus, it appears to us that L0a and L0d sequences share key alleles at 47 sites. In contrast, our L0a consensus root sequence carries 16 L0a haplogroup specific substitutions and 26 private mutations that are not reported for the L0d sequences of Gonder et al. (2007). On the other hand, their 5 L0d TZ sequences carry a total of 30 subclade-specific substitutions that we did not observe in our L0a sequences, a net disparity of 12

substitutions. They do not enumerate the private mutations in these sequences, so we cannot assess their impact. However, given that we observed more than 570 substitutions in our coding region sequences (see Results), we conclude that our results (and, specifically, the numbers of substitutions) have not been significantly biased by our choice of root sequence.

Assumption #3 – Lacking a “pure” root sequence, a haplogroup L0a consensus sequence is used to root the networks and determine the relative order and direction in which haplogroup L substitutions have occurred.

A Qualified and Operational Definition of the Neutral Mutation Rate

It has been assumed that the rate of *synonymous* codon evolution in mammals has been neutral (Sharp et al. 1995). Furthermore, our recent analyses of human mtDNA coding region sequences indicated that the mutation pattern of synonymous substitutions was consistent with the neutral model (Elson et al. 2004; see especially Figure 1), although the fit to that model was only of borderline statistical significance for private polymorphisms. Using a powerful analytical approach, Kosakovsky Pond and Muse (2005) have shown significant site variation in the synonymous substitution rate in several genes, including the mitochondrial COX1 gene of primates (see also Hurst and Pal 2001 and Galtier et al. 2006). A number of possible explanations for this site variability are possible, and Kosakovsky Pond and Muse (2005) suggest that some type of selection might be involved.

As a result of those recent studies, it is unlikely that the synonymous substitution rate in the human mtDNA coding region is a perfect “reporter” of the neutral rate of mutation. However, lacking a complete “map” of precisely which subset of synonymous sites is non-neutral, but providing the most empirical approach, we use here the overall rate of substitution at 4-fold degenerate sites as the “benchmark” rate that is closest to the neutral rate. Thus, if the overall rate of mutation in the control region is less than the overall rate of substitution at 4-

fold degenerate sites, then one has a rationale for concluding that the *relative* effect of selection is stronger in the control region (and, of course, *vice versa*).

In the present analysis, we are limiting ourselves to the rates of mtDNA substitution that have been ascertained at the population level, rather than true mutation rates at the molecular level. Because of the complexities of human mitochondrial genetics, this rate of substitution is lower than the molecular mutation rate, but by how much is unclear because the latter is technically difficult to measure accurately (see the discussion on page 667 of Howell et al. 2003).

Assumption #4 – The overall rate of substitution at 4-fold degenerate sites, measured at the population level, is the best available estimator of the rate of neutral evolution.

With these limitations in mind, there are 3755 codons (11265 nucleotides) in the mtDNA coding region with the exclusion of termination codons and of overlapping coding frames. These codons are distributed into the following categories for measurements of relative substitution rate:

[a] 2026 nucleotide sites are 4-fold degenerate

[b] 1819 nucleotide sites are 2-fold degenerate (this number includes 90 TTA/TTG leucine codons in which both the first and third positions are 2-fold degenerate)

[c] 7420 nucleotide sites produce nonsynonymous substitutions

In addition, we also measured the aggregated substitution rates in the 22 tRNA genes (1507 nucleotide sites) and in the 2 rRNA genes (2513 nucleotide sites).

Statistical Tests

The statistical significance of differences between groups (Tables 2 and 6) was determined with χ^2 tests; 2 x 2 tests were corrected for continuity (Equation 19.10 on page

371 of Steel and Torrie 1960). The continuity correction essentially adjusts the χ^2 distribution (which is calculated from discrete data), when $df = 1$, to more nearly match the distribution based on normal deviates (which will be continuous; Yates 1934). For other comparisons, Fisher's Exact Test was used, as noted in the Results.

Results

The L0a Consensus Sequence as the Root of the Haplogroup L Network

Five L0a mtDNA sequences have been used for this analysis (Howell et al. 2004). Relative to the revised Cambridge Reference Sequence (Andrews et al. 1999), there were 89 substitutions in the coding region and 29 substitutions in the control regions of these five sequences (Table 1). These 118 substitutions can be assigned to three classes:

[a] L0a haplogroup-specific substitutions of which there are 31 in the coding region and 11 in the control region. These occur in two or more of the five sequences but do not occur, unless as private polymorphisms or as homoplasies, in any of the other haplogroup L sequences analyzed here.

[b] Haplogroup L-associated substitutions of which there are 32 in the coding region and 11 in the control region. These occur in the haplogroup L0a sequences and in at least one other haplogroup L clade. Twelve of these substitutions in the coding region and one in the control region occur in all (except in rare instances of reversion) 96 sequences analyzed here (Table 1) and do not, as a result, provide useful phylogenetic information.

[c] Private substitutions of which there are 26 in the coding region and 7 in the control region occur in only a single L0a sequence. These sequence changes are operationally defined here for this group of 5 L0a mtDNAs. This is the most “fluid” of the three classes and, as more L0a sequences are analyzed, any of these substitutions might no longer be a true private polymorphism.

Mutation Spectra of the Haplogroup L Coding and Control Regions

The consensus sequence of the haplogroup L0a mtDNAs was used to root the haplogroup L1b/L1c, L2, and L3 networks and then to determine the total number of substitutions in the control and coding regions (Table 2). There were a total of 573

substitutions at 513 sites within the coding region (Supplementary Table S1), thereby yielding an overall substitution frequency of 3.75% during the evolution of these sequences. There were also 11 substitutions at 10 sites in the non-coding regions of the coding regions, in the origin of L-strand replication, or in nucleotides where two genes overlap, but these substitutions were not included in our analyses.

It is immediately apparent that the proportions of sites mutated, and the overall substitution frequencies, are higher for both the two-fold and four-fold synonymous sites (SYN-2X and SYN-4X, respectively), than for sites involving non-synonymous changes and for sites in the rRNA and tRNA genes. This finding is compatible with our previous conclusion that sites within the human mtDNA that give rise to non-synonymous substitutions – our operational definition in that work *included* the rRNA and tRNA genes - have been subject to purifying selection (Elson et al. 2004; see also the similar results in Table 1 of Kivisild et al. 2006). The proportion of sites mutated is slightly larger for SYN-4X sites than for the SYN-2X sites (Table 2), but this difference is not statistically significant ($P > 0.25$). However, when we compare the overall substitution frequencies for these two classes of sites, which accounts for multiple changes at the same site (see further analysis below), the higher overall substitution frequency for the SYN-4X sites just reaches statistical significance, relative to that for the SYN-2X sites ($P = 0.05$) although this difference would not be meaningful if we correct for multiple tests on the whole data set.

Kivisild et al. (2006) reported that ~22% of the variable sites in the human mtDNA coding region had undergone recurrent mutations and that mutational hotspots occurred predominately within the rRNA genes. Their results differ from our observations. Among our African mtDNA sequences, recurrent mutations occurred at the following frequencies: tRNA (8.3% of the sites); rRNA (10.5%); NS (11.5%); SYN-2X (3.6%); and SYN-4X (12.4%).

Furthermore, for both the tRNA and rRNA genes, the 9 total sites of recurrent mutation all underwent only two substitutions each in our sequences.

Kivisild et al. (2006) analyzed a larger set of sequences and their sequences were more diverse and included representatives from Africans, Asians/Native Americans and Europeans. A comparative analysis of our results with theirs confirms that this is the predominant reason for the discrepancy. In their Table 6, they report that, for the 9 sites at which 5 or more recurrent mutations occur, 3 involve the rRNA genes and sites 709, 1438 and 1888. In their entire sequence set of 277 mtDNAs, they report a total of 14, 5 and 8 recurrences, respectively, although the numbers drop to 6, 3 and 2 for the 129 African mtDNAs (Figure 1 of Kivisild et al. 2006). Within our sequences, the equivalent numbers are 2, 1 and 0 (Supplementary Table 1). In their phylogenetic tree (Figure 1), Kivisild et al. (2006) report substitutions at site 709 in six subclades: L1b, L2a2, L2c, L3h, L4g and L5a. Our network confirms the changes in the L1b and L2c subclades, but our African mtDNA sequence set did not include representatives from the other four subclades. For site 1438, their phylogenetic tree shows substitutions in three subclades: L0d, L1b and L3f1. Our network also carries a substitution in the L1b sequences, but our sequence set did not include L0d sequences. One of their two L3f sequences carries this substitution but we did not detect it in any of the five L3f mtDNAs in our sequence set. Finally, Kivisild et al. (2006) report network tip changes at site 1888 in the sequences from the L0a and L3h subclades. Our sequence set did not include L3h sequences, and we did not detect this substitution in our five L0a sequences (Table 1). Overall, our results agree with theirs when sequences from the same subclades are analyzed. The exception is that they report two network tip changes (viz, within a single sequence) that we did not detect, but we do not consider this a major discrepancy given the high frequency of substitution.

In the control region, there were 261 substitutions at 104 sites (Table 2 and Supplementary Table S2). The proportion of control region sites mutated is slightly higher than that for SYN-4X sites, but the difference is not statistically significant. However, when one compares the overall substitution frequencies, that for the control region is much higher (by a factor greater than two) and the difference is highly significant. The difference in substitution frequency is 6.2-fold higher in the control region than that for the overall coding region frequency.

In Table 3, we summarize the numbers of sites with multiple changes for both the control region and for the SYN sites in the coding region, and we identify in Table 4 the control region sites that have undergone multiple substitutions in these haplogroup L sequences. There is good agreement between the hypervariable sites identified here and in other studies (Excoffier and Yang 1999; Meyer et al. 1999; Malyarchuk and Rogozin 2004). It should be noted that those various studies arrived at slightly different relative rates for sites in the control region because they analyzed different sequence sets and used different methodologies. Furthermore, and these are important points, (1) none of those studies analyzed the entire control region and – instead – focused on HVR1 and/or HVR2 and (2) none were able to compare relative substitution rates in the control region to the neutral rate of evolution in the coding region. In Table 4, nevertheless, we include relative rates from Meyer et al. (1999) for purposes of comparison.

The results in Table 3 show unequivocally that the overall increased substitution frequency in the control region, relative to both SYN-4X and SYN-2X sites, is due to the larger numbers of control region sites that have undergone three or more substitutions – which we define operationally here as hypervariable sites - in this African mtDNA sequence set. Thus, there are 5 of 2026 SYN-4X sites (0.25%) and 1 of 1819 SYN-2X sites (0.05%) that have undergone 3, 4 or 5 substitutions, whereas there are 34 of 1122 control region sites (3.0%)

that have undergone 3-11 substitutions. The mutational frequency spectrum of the control region does not fit a simple Poisson process because of the number of sites that have undergone multiple substitutions, whereas that for the SYN-2X sites does. That is, within this array of mtDNA sequences, substitutions at SYN-2X coding region sites are adequately modeled by a single rate process of 0.082 substitutions/site since the time of the last common ancestor. Interestingly, the frequency spectrum for the SYN-4X sites also does not fit a simple Poisson process, due to the presence of four hypervariable sites. Omitting the four of the five fastest sites yields a simple process that fits with a single rate of 0.091 substitutions/site, a rate that is slightly faster than that for the SYN-2X sites. Because of the known complexity of the rate process in the control region (Excoffier and Yang 1999; Meyer et al. 1999; Howell et al. 2004), including the possible presence of invariant sites, we have not tried to “force” our data into a single rate process for the control region by pruning away the most rapidly evolving sites.

Preliminary Analysis of the Substitution Process

It is important to highlight the relative rate of substitution in the control region. The three most rapidly changing sites (positions 16189, 16311 and 16519) each have a relative rate of substitution that is greater by more than 10-fold than the rate of SYN-2X and SYN-4X (neutral) substitution. One explanation for the hypervariable sites in the control region is positive selection (Wong and Nielsen 2004), but previous work has also suggested a role for sequence context-dependent mutation rates (see Discussion). As a preliminary attempt to unravel the evolutionary processes that produce hypervariable sites, we have further analyzed the control region sites in two ways. In the first approach, we used the reduced median networks to derive operational information about the direction of substitution at control region sites. In Table 4, we note for sites with five or more substitutions those for which no

substitutions in the reverse direction were detected. It was observed that seven such sites, undergoing a total of 42 substitutions, showed no reverse mutations (thus, at least one reversion event was detected at each of the other 11 sites). Given the relatively small numbers of substitutions available for analysis, however, these results are not statistically significant and we thus do not know – pending more extensive analyses – if there are sites with asymmetric substitution patterns. Thus, there were 50 control region sites with multiple substitutions: 172 forward events and 35 reversion events. Simulations with random allocation of the reversion events indicates that the distribution obtained does not deviate significantly from a random pattern (data not shown).

On the other hand, for the three “fastest” sites, it is the occurrence of multiple reversion events for each that is informative. Thus, for site 16519, we detected six forward direction substitutions (that is, to the non-CRS allele) and five reverse direction substitutions. At first glance, it is difficult to believe that such rapid substitution in *both* directions can reflect the action of positive selection as a cause. Likewise, for sites 16189 and 16311, there were two and three reversion events, respectively.

In the second approach, we analyzed this same set of hypervariable control region sites to obtain information on whether they occurred randomly with respect to haplogroup. The results (Table 5) suggest a lack of randomness for certain sites. Thus, all substitutions at sites 143, 16192, and 16309, a total of 18 changes, occurred in haplogroup L2 sequences. In actual fact, the results are even more striking in that, with the exception of one substitution at site 16192, the other 17 changes occurred in the 30 subclade L2a sequences which account for ~31% of the sequences analyzed here (data not shown). In our previous study (Howell et al. 2004), it appeared that there was a single forward mutation at site 143 and that this was subsequently followed by five reversion events in these subclade L2a sequences. This pattern, however, does not apply to sites 16192 and 16309, and multiple forward and reverse

substitutions were observed at both sites (data not shown). Conversely, none of the 18 substitutions at sites 185, 189, and 16293 occurred in the haplogroup L2 sequences. We also noted that five of the six substitutions at nucleotide position 200 occurred in the haplogroup L3 sequences. Finally, we note that none of the 10 substitutions at site 16189 occurred in the L0 and L1 sequences and it is noteworthy that these sequences carry a C:T substitution at site 16187. We know from previous studies that substitution rates in this small segment of the control region are context-dependent for some sites (Howell and Bogolin Smejkal 2000).

We caution again that these are initial results and that they are not statistically significant when analyzed on a site-by-site basis (pooling sites at this point would be *post hoc* and it cannot be justified by a testable hypothesis). Thus, we grouped the data into L2 and “Other” haplogroups and then carried out 2 x 2 Fisher’s exact tests. The P value for the test with site 16192 was 0.07, which was the lowest obtained. Nevertheless, these results are important for their heuristic value, and analysis of larger sequence sets is required for a more complete exploration of this issue.

Finally, we explored the relative rate of sequence evolution of the coding and control regions for each haplogroup (Table 6). We stratified the coding region substitutions into three classes: all, SYN-4X only and “TIP” changes (that is, only those substitutions that map to terminal positions within the networks). The SYN-4X substitutions are neutral, or closest thereto, whereas the “TIP” changes are the most recent to have occurred during evolution and should provide another way of analyzing mtDNA changes that have been subjected to less selection. By their nature, however, TIP changes will vary according to the sequence set and the resulting network or phylogenetic tree being analyzed. With all three classes of coding region substitutions, the control region substitution rate – relative to the substitution rate in the homologous coding region - was relatively higher in haplogroups L2 and L3 than in haplogroups L0 and L1. These differences do not reach statistical significance, although

significance is reached for some of the comparisons if we pool the results into L0+L1 and L2+L3 groups. Such *post hoc* analysis, however, is questionable and we prefer to await the results of further analysis with larger sequence sets.

Discussion

We analyzed the rate and pattern of substitution in the mtDNA control region relative to the coding region in a set of African haplogroup L mtDNA sequences. The significantly higher rates of synonymous substitutions (both at two-fold and four-fold degenerate sites), relative to nonsynonymous substitutions and to those in the tRNA and rRNA genes, is further support for the role of purifying selection during evolution of the human mitochondrial genome (Elson et al. 2004; Kivisild et al. 2006). In addition, a number of other findings emerged from this analysis that bear on the evolution of human mtDNA. However, this is an initial study and we are therefore cognizant of the need for further, more extensive analyses with larger and independent sequence sets.

[1] The proportion of sites mutated does not differ significantly for the synonymous 4X sites in the coding region and for the control region (Table 2). The proportions of sites that have undergone one substitution are also similar (Table 3). For the haplogroup L sequences analyzed here, therefore, there is no evidence that the rate of sequence evolution for the vast majority of sites in the control region has been the object of more – or of less - intense selection than the synonymous 4X sites, which most closely approximate the evolutionarily neutral rate of substitution. At this stage, we are unable to ascertain if there are a significant number of highly conserved or invariant sites in the control region, although this is a fertile area for subsequent analysis.

[2] As a rough estimate, ~ 3% of the control region sites have a rate of sequence evolution that “exceeds” the neutral rate of human mtDNA evolution, some by more than an order of magnitude (Table 3). The presence of hypervariable sites in the control region has been recognized previously, but what does it mean? The two most basic explanations are (1) that some control region sites are evolving under conditions of *positive* selection and/or (2) that the rate of evolution at these sites is not influenced by selection but, instead, is a reflection

of an increased rate of mutation at the molecular level. With regard to the latter possibility, the sequence context or background of some sites in the control region would presumably dictate a higher mutation rate than at other sites.

We have previously reported one instance where sequence context in the control region markedly influenced mutation rate (Howell and Bogolin Smejkal 2000). There is also evidence from other investigators. In their analysis of rate heterogeneity in the control region, Meyer *et al.* (1999) noted that sites in functionally important segments did not always have low substitutions rates, as would be expected if selection were the sole determinant of substitution rate. Malyarchuk and Rogozin (2004) obtained evidence for strand dislocation mutagenesis in the human mtDNA control region and for sequence context-dependence of mutation rates. Galtier *et al.* (2006), in an analysis of synonymous substitutions in animal mtDNA sets, detected mutation hotspots in the *coding* region. The distribution of these hotspots showed lineage-specific effects and the authors suggested a “link” between nucleotide state and mutation rate. Finally, there is the compositional segmentation analysis of Samuel *et al.* (2003) in which it was concluded that the human mtDNA mutation rate, specifically at guanine residues, at a particular site is dependent *both* upon the nucleotide and the sequence context, but *not* upon protein function. If the substitution rate is dependent upon sequence context, then the i.i.d. (independently and identically distributed) condition will not hold for the control region. It is already recognized that the i.i.d. condition does not hold for genes that encode RNA molecules with secondary structure (see especially Yu and Thorne 2006), and the mtDNA control region is predicted to have substantial secondary structure (Saccone *et al.* 1991). Taken together, the evidence indicates that sequence context is likely to determine the high substitution rates at some sites in the control region and also at a few SYN-4X sites in the coding region. As a point of speculation, it is interesting to consider whether such elevated substitution rates are “cost free” over the long-term in evolution.

[3] There was preliminary evidence that the rates of some control region substitutions may differ among different haplogroup L subclades (Table 5) and there was also a suggestion that the comparative rates of coding and control region substitution differ among haplogroups (Table 6). These results did not reach statistical significance, and their importance – therefore – lies in their heuristic value for subsequent analysis.

We are learning that the evolution of human mtDNA is very complex, but that complexity is forcing us into some new and important areas of research. This report is not meant to be an exercise in “bomb throwing” and no one here is suggesting that the use of simple models of mtDNA evolution for broader questions of human evolution and population history is inappropriate. This is not to say that choice of models might not have important effects, or that we should ignore the very large confidence intervals around the dates obtained but, thus far, the problems are most profound for deeper phylogenies (see Jones et al. 2007 for one recent example). Nevertheless, as we improve our models of mtDNA evolution, such studies of human evolution and population dispersal will benefit.

In summary, our results provide further insights into the evolution of the human mitochondrial genome. They also raise additional – and we believe important – questions on role of selection during mtDNA evolution and the role of sequence context on mutation rate. “Non-ideal” mtDNA evolution has been relatively easy to detect when analyzing evolutionary diverse sequences, but an important finding of the present study is that complex evolutionary processes can be detected even within a set of closely related human mtDNA sequences.

Appendix 1

During the analysis of the African haplogroup L networks reported here, we noted some minor errors in Figure 2 of Howell *et al.* (2004):

(a) Sequences 223 and 233 on the network are transposed; that is, the changes shown are correct if sequence 223 is relabeled 233, and *vice versa*.

(b) The branch leading to sequence 576 carries a substitution at position 16129, not 16126 as shown in the figure.

(c) The long branch descending from the node with sequences 434 and 563 carries a substitution at position 15629, not 15621.

(d) The branch leading to sequences 165 and 561 carries a substitution at position 198, not 189.

Literature Cited

- Andrews RM, Kubacka I, et al. (6 co-authors). 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genet* 23:147.
- Batini C, Coia V, Battaglia C, et al. (9 co-authors). 2007. Phylogeography of the human L1c haplogroup: genetic signatures of the prehistory of Central Africa. *Mol Phylogenet Evol* 43:635-644.
- Broughton RE, Reneau PC. 2006. Spatial covariation and nonsynonymous substitution rates in vertebrate mitochondrial genomes. *Mol Biol Evol* 23:1516-1524.
- Cann R, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. *Nature* 325:31-36.
- Elson JL, Turnbull DM, Howell N. 2004. Comparative genomics and the evolution of human mitochondrial DNA: assessing the effects of selection. *Am J Hum Genet* 74:229-238.
- Excoffier L, Yang Z. 1999. Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Mol Biol Evol* 16:1357-1368.
- Finnilä S, Lehtonen MS, Majamaa K. 2001. Phylogenetic network for European mtDNA. *Am J Hum Genet* 68:1475-1484.
- Galtier N, Enard D, et al. (5 co-authors). 2006. Mutation hot spots in mammalian mitochondrial DNA. *Genome Res* 16:215-222.
- Gonder MK, Mortensen HM, et al. (5 co-authors). 2007. Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol Biol Evol* 24:757-768.
- Helgason A, Palsson G, et al. (7 co-authors). 2006. mtDNA variation in Inuit populations of Greenland and Canada: migration history and population structure. *Am J Phys Anthropol* 130:123-134.

- Herrnstadt C, Elson JL, et al. (11 co-authors). 2002. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am J Hum Genet* 70:1152-1171.
- Ho SYW, Phillips MJ, Cooper A, Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol* 22:1561-1568.
- Howell N, Bogolin Smejkal C. 2000. Persistent heteroplasmy of a mutation in the human mtDNA control region. Hypermutation as an apparent consequence of simple repeat expansion/contraction. *Am J Hum Genet* 66:1589-1598.
- Howell N, Bogolin Smejkal C, et al. (6 co-authors). 2003. The pedigree rate of sequence divergence in the human mitochondrial genome: There is a difference between phylogenetic and pedigree rates. *Am J Hum Genet* 72:659-670.
- Howell N, Elson JL, Turnbull DM, Herrnstadt C. 2004. African haplogroup L mtDNA sequences show violations of clock-like evolution. *Mol Biol Evol* 21:1843-1854.
- Hurst LD, Pal C. 2001. Evidence for purifying selection acting on silent sites in *BRCA1*. *Trends Genet* 17:62-65.
- Ingman M, Kaesmann H, Pääbo S, Gyllensten U. 2000. Mitochondrial genetic variation and the origin of modern humans. *Nature* 408:708-712.
- Jones M, Gantenbein B, Fet V, Blaxter M. 2007. The effect of model choice on phylogenetic inference using mitochondrial sequence data: lessons from the scorpions. *Mol Phylogenet Evol* 43:583-595.
- Kayser M, Brauer S, et al. (15 co-authors). 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol Biol Evol* 23:2234-2244.

- Kivisild T, Shen P, et al. (17 co-authors). 2006. The role of selection in the evolution of human mitochondrial genomes. *Genetics* 172:373-387.
- Kocher TD, Wilson AC. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees. Pp 391-413 *in* S. Osawa and T. Honjo, eds. *Evolution of life: fossils, molecules and culture*. Springer, Tokyo.
- Kosakovsky Pond S, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol* 22:2375-2385.
- Loewe L. 2006. Quantifying the genomic decay paradox due to Muller's ratchet in human mitochondrial DNA. *Genet Res* 87:133-159.
- Malyarchuk BA, Rogozin IB. 2004. Mutagenesis by transient misalignment in the human mitochondrial DNA control region. *Ann Human Genet* 68:324-339.
- Meiklejohn CD, Montooth KL, Rand DM. 2007. Positive and negative selection on the mitochondrial genome. *Trends Ecol Evol* 23:259-263.
- Meyer S, Weiss G, von Haeseler A. 1999. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics* 152:1103-1110.
- Moilanen JS, Majamaa K. 2003. Phylogenetic network and physiochemical properties of nonsynonymous mutations in the protein-coding genes of human mitochondrial DNA. *Mol Biol Evol* 20:1195-1210.
- Non AL, Kitchen A, Mulligan CJ. 2007. Identification of the most informative regions of the mitochondrial genome for phylogenetic and coalescent analyses. *Mol Phylogenet Evol*, in press.
- Ohashi J, Naka I, et al. (8 co-authors). 2006. Mitochondrial DNA variation suggests extensive gene flow from Polynesian ancestors to indigenous Melanesians in the northwestern Bismarck Archipelago. *Am J Phys Anthropol* 130:551-556.

- Packendorf B, Novgorodov IN, et al. (6 co-authors). 2006. Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts. *Hum Genet* 120:334-353.
- Penny D. 2005. Relativity for molecular clocks. *Nature* 436:183-184.
- Perna NT, Kocher TD. 1995. Unequal base frequencies and the estimation of substitution rates. *Mol Biol Evol* 12:359-361.
- Pesole G, Gissi C, De Chirico A, Saccone C. 1999. Nucleotide substitution rate of mammalian mitochondrial genomes. *J Mol Evol* 48:427-434.
- Ruiz-Pesini E, Mishmar D, et al. (5 co-authors). 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* 303:223-226.
- Saccone C, Pesole G, Sbisa E. 1991. The main regulatory region of mammalian mitochondrial DNA: structure-function model and evolutionary pattern. *J Mol Evol* 33:83-91.
- Samuels DC, Boys RJ, Henderson DA, Chinnery PF. 2003. A compositional segmentation of the human mitochondrial genome is related to heterogeneities in the guanine mutation rate. *Nucl Acids Res* 31:6043-6052.
- Sharp PM, Averof M, et al. (5 co-authors). 1995. DNA sequence evolution: the sounds of silence. *Phil Trans Royal Soc Lond B* 349:241-247.
- Steel RGD, Torrie JH. 1960. *Principles and Procedures of Statistics*. McGraw-Hill Inc., New York.
- Sun C, Kong Q-P, Zhang Y-P. 2006. The role of climate in human mitochondrial DNA evolution: a reappraisal. *Genomics* 89:338-342.
- Torrioni A, Rengo C, et al. (12 co-authors). 2001. Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am J Hum Genet* 69:1348-1356.

- Torrioni A, Achilli A, Macaulay V, Richards M, Bandelt H-J. 2006. Harvesting the fruit of the human mtDNA tree. *Trends Genet* 22:339-345.
- Weiss G, von Haeseler A. 2003. Testing substitution models within a phylogenetic tree. *Mol Biol Evol* 20:572-578.
- Wong WSW, Nielsen R. 2004. Detecting selection in noncoding regions of nucleotide sequences. *Genetics* 167:949-958.
- Yates F. 1934. Contingency tables involving small numbers and the χ^2 test. *J Roy Stat Soc Suppl* 1:217-235.
- Yu J, Thorne JL. 2006. Dependence among sites in RNA evolution. *Mol Biol Evol* 23:1525-1537.

Table 1

Substitutions in Haplogroup L0a mtDNA Sequences

A. Coding Region

1. L0a Haplogroup-Specific Substitutions (31): C1048T^a; A2245G; C3516A; C4312T; T4586C; **T5096C**^b; G5231A; T5442C; G5460A; C5603T; **C5911T**; T6185C; C8428T; A8566G; C9042T; A9347G; G9755A; C9818T; G10589A; C10664T; T10915C; G11176A; A11641G; G11914A; G12007A; A12720G; A13276G; **A14007G**; T14308C; C15136T; G15431A
2. Haplogroup-associated Substitutions (32)^c: A750G; G769A; T825A; G1018A; A1438G; A2706G; G2758A; T2885C; C3594T; A4104G; A4769G; C7028T; A7146G; C7256T; G7521A; C8468T; C8655T; A8701G; A8860G; T9540C; A10398G; G10688A; T10810C; T10873C; G11719A; G12127A; C12705T; A13105G; C13506T; C13650T; C14766T; A15326G
3. Private Substitutions (26): T593A; T961C; G3736A; T3866C; C5012T; G5147A; G5563A; A5711G; T6221A; G6257A; G6446A; A8191G; A8460G; A8577G; A9181G; A9545G; G9554A; C11143T; A11172G; A11812G; C12432T; C13116T; T14106C; A14755G; T15099C; C15839T

B. Control Region

1. L0a Haplogroup-Specific Substitutions (11): A93G; **A95C**; T236C; G247A; C16148T; **C16168T**; T16172C; C16188G; A16230G; **A16293G**; C16320T

2. Haplogroup-associated Substitutions (11): T152C; **G185A^e**; A189G^e; A263G; **G16129A**; C16187T; T16189C; C16223T; **C16278T**; T16311C; **T16519C**

3. L0a Private Substitutions: C64T; A200G; T204C; G207A; T16093C; G16188A^d; A16215G

^a The first letter before the site number is the L-strand nucleotide in the rCRS, whereas the second letter is the nucleotide in the haplogroup L sequence(s).

^b The substitutions shown in **boldface** font occur in only two-four of the five L0a sequences.

^c Underlined substitutions occur in all of the haplogroup L mtDNA sequences analyzed here with the exception of rare reversions at sites 12705, 13506 and 15326.

^d Because the other 4 L0a sequences carry a G at position 16188, it is most likely that there was an earlier C-to-G transversion and that this particular mtDNA underwent a subsequent G-to-A transition.

^e The sequence evolution at nucleotide positions 185 and 189 in the haplogroup L sequences exemplifies the complexities and ambiguities that can occur in phylogenetic analysis of rapidly evolving sequences. The L0a root carries a G:A substitution at position 185 and an A:G substitution at position 189. Haplogroup L1b sequences carry a T at position 185, so we conclude (under the MP condition) that the ancestral sequence to this subclade underwent an A:T transversion. In contrast, L1c sequences carry an A at position 185, so we conclude that the ancestral sequence to this subclade underwent a G:A back-mutation. Thus, there have

been two substitutions at this position in the assumed evolutionary path from L0a to the L1b and L1c subclades. The situation at position 189 is even more complex. Three of the 12 L1b sequences carry a G at position 189, whereas the other nine carry an A. Analysis of the networks thus indicates an “early” G:A back-mutation at position 189 with a subsequent A:G forward-mutation in an L1b sub-branch. All L1c sequences analyzed carry a C at this position, indicating the occurrence of a transversion. However, there is insufficient phylogenetic information to allow a definitive conclusion as to whether this was a G:C or A:C transversion, which will depend on whether the G:A back-mutation occurred before the L1b/L1c “split”, or afterwards. Regardless, there are a total of three substitutions at position 189 in this portion of the network, one transversion and two transitions.

Table 2

Substitutions in the Haplogroup L Coding and Control Regions

<u>Class of Sites</u>	<u>Sites Mutated</u>	<u>Substitution Freq.</u>
Synonymous – 4X	169/2026 (8.34%)	198/2026 (9.77%)
Synonymous – 2X	138/1819 (7.59%)	144/1819 (7.92%)
Non-synonymous	113/7420 (1.52%)	129/7420 (1.74%)
rRNA Genes	57/2513 (2.27%)	63/2513 (2.51%)
tRNA Genes	36/1507 (2.39%)	39/1507 (2.59%)
Control Region	104/1122 (9.27%)	261/1122 (23.26%)

Table 3

Summary of the Number of Substitutions at Each Site

<u>Substitutions (#)</u>	<u>Control Region</u> (1122 Sites)	<u>SYN-4X</u> (2026 Sites)	<u>SYN-2X</u> (1819 Sites)
0	1018 (90.7%)	1857 (91.7%)	1681 (92.4%)
1	54 (4.8%)	148 (7.3%)	138 (7.6%)
2	16 (1.4%)	16 (0.8%)	4 (0.2%)
3	12 (1.1%)	3 (0.1%)	1 (0.05%)
4	4 (0.4%)	1 (0.05%)	0 (0%)
5	4 (0.4%)	1 (0.05%)	0 (0%)
6	5 (0.4%)	0 (0%)	0 (0%)
7	6 (0.5%)	0 (0%)	0 (0%)
8	0 (0%)	0 (0%)	0 (0%)
9	0 (0%)	0 (0%)	0 (0%)
10	2 (0.2%)	0 (0%)	0 (0%)
11	1 (0.1%)	0 (0%)	0 (0%)

Table 4

Control Region Sites with Multiple Substitutions

<u># Subs./Site</u>	<u>Sites^a</u>
11	16519 ^b (nd)
10	16189 (5), 16311 (5)
7	152 (6), 189 (6), 195 (6), <u>16093 (3)</u> , <u>16129 (5)</u> , 16192 (4)
6	143 (3), 185 (5), <u>198 (4)</u> , <u>200 (4)</u> , <u>16086 (1)</u>
5	<u>150 (6)</u> , 16293 (4), <u>16295 (1)</u> , 16309 (4)
4	16172 (4), 16187 (4), 16294 (5), 16399 (nd)
3	93 (5), 146 (6), 204 (3), 16124 (1), 16145 (2), 16213 (1), 16215 (<1), 16234 (1), 16264 (1), 16286 (1), 16292 (2), 16362 (5)
2	95 (5), 151 (4), 182 (6), 207 (3), 236 (3), 16051 (4), 16114 (1), 16185 (1), 16188 (2), 16260 (2), 16265 (2), 16270 (3), 16278 (5), 16291 (2), 16320 (2), 16355 (2)

^a For sites with five or more substitutions, underlining indicates that no reverse mutations were detected from network analysis.

^b The number in parentheses is the approximate relative substitution rate determined by Meyer et al. (1999; see their Figure 2). An (nd) means that the relative rate was not determined because the site was not in HVR1 or HVR2. Also, it should be noted that Meyer et al. (1999) used a discretized χ^2 -distribution and assigned sites to one of eight rates.

Table 5

Distribution of Substitutions by Haplogroup for Control Region Hypervariable Sites

<u>Site</u>	<u>L0 (5%)^a</u>	<u>L1 (22%)</u>	<u>L2 (45%)</u>	<u>L3 (28%)</u>
143	0	0	6	0
150	0	0	2	3
152	0	0	3	4
185	1	2	0	3
189	0	3	0	4
195	0	1	2	4
198	0	3	2	1
200	1	0	0	5
16086	0	2	3	0
16093	1	1	3	2
16129	1	1	4	1
16189	0	0	6	4
16192	0	0	7	0
16293	1	4	0	0
16295	0	0	2	3
16309	0	0	5	0
16311	0	1	4	5
16519	1	1	5	4

^a The percentage of haplogroup sequences within the total set is indicated. Thus, the five L0 sequences represent ~ 5% of the total haplogroup L sequences analyzed here.

Table 6

Ratios of Coding and Control Region Substitutions by Haplogroup

<u>Haplogroup</u>	<u>Control Region</u>	<u>Coding Region^a</u>		
		<u>All^b</u>	<u>SYN-4X</u>	<u>Tip</u>
L0	22	57 (2.59)	22 (1.00)	26 (1.18)
L1	60	166 (2.77)	52 (0.87)	95 (1.58)
L2	113	224 (1.98)	72 (0.64)	114 (1.00)
L3	70	139 (1.99)	50 (0.71)	68 (0.97)

^a“All” is the total number of substitutions in the coding region; “SYN-4X” represents only that class of substitutions; and “Tip” refers to the number of tip changes in the reduced median network of that haplogroup.

^b Simple 2 x 4 χ^2 tests were carried out to determine if the differences between control and coding regions for the four haplogroups are statistically significant. None of the three results are significant; the *P* values are approximate 0.26, 0.44 and 0.12, respectively, for the ALL, SYN-4X and TIP data.