

Identification of the most informative regions of the mitochondrial genome for phylogenetic and coalescent analyses

A.L. Non, A. Kitchen, C.J. Mulligan *

Department of Anthropology, University of Florida, Gainesville, FL, 32611, USA

Received 11 October 2006; revised 13 December 2006; accepted 21 December 2006

Available online 31 December 2006

Abstract

Analysis of complete mitochondrial genome sequences is becoming increasingly common in genetic studies. The availability of full genome datasets enables an analysis of the information content distributed throughout the mitochondrial genome in order to optimize the research design of future evolutionary studies. The goal of our study was to identify informative regions of the human mitochondrial genome using two criteria: (1) accurate reconstruction of a phylogeny and (2) consistent estimates of time to most recent common ancestor (TMRCA). We created two series of datasets by deleting individual genes of varied length and by deleting 10 equal-size fragments throughout the coding region. Phylogenies were statistically compared to the full-coding-region tree, while coalescent methods were used to estimate the TMRCA and associated credible intervals. Individual fragments important for maintaining a phylogeny similar to the full-coding-region tree encompassed bp 577–2122 and 11,399–16,023, including all or part of 12S rRNA, 16S rRNA, ND4, ND5, ND6, and cytb. The control region only tree was the most poorly resolved with the majority of the tree manifest as an unresolved polytomy. Coalescent estimates of TMRCA were less sensitive to removal of any particular fragment(s) than reconstruction of a consistent phylogeny. Overall, we discovered that half the genome, i.e., bp 3669–11,398, could be removed with no significant change in the phylogeny ($p_{AU} = 0.077$) while still maintaining overlap of TMRCA 95% credible intervals. Thus, sequencing a contiguous fragment from bp 11,399 through the control region to bp 3668 would create a dataset that optimizes the information necessary for phylogenetic and coalescent analyses and also takes advantage of the wealth of data already available on the control region.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Mitochondrial DNA; Complete mitochondrial genome sequence; Coalescence; Phylogeny

1. Introduction

Within the past few decades, the majority of genetic research on human evolutionary history has focused on the small non-coding control region of the mitochondrial genome, due to its high mutation rate and assumed selective neutrality. Within this ~1120 base pair (bp) region, most studies have been limited to the small ~400 bp hypervariable region I (HVRI, 16,024–16,383 bp, according to Anderson et al. (1981)), which comprises less than 3% of the 16,568 bp mitochondrial genome. However, because of its high frequency of homoplasies and relatively short length,

the HVRI has limited power to construct accurate topologies of networks and phylogenies (Finnila et al., 2001), to identify sub-haplogroups in a phylogeny (Palanichamy et al., 2004), and to determine founder status in a founder analysis (Bandelt et al., 2003). Even the addition of a few coding region sites, typically assayed as restriction fragment length polymorphisms (RFLPs), is unable to overcome these problems as mutation rate varies extensively between the control and coding regions (Finnila et al., 2001). The use of RFLPs is also inadequate to provide a mutation rate estimate necessary for calculating time to most recent common ancestor (TMRCA) of clades in a phylogeny (Ingman et al., 2000).

In contrast to the control region, the mitochondrial coding region is more reliable for phylogenetic analyses

* Corresponding author. Fax: +1 352 273 8284.

E-mail address: mulligan@anthro.ufl.edu (C.J. Mulligan).

and dating of evolutionary events as it has been suggested to conform to a molecular clock hypothesis with a steady and constant mutation rate (Silva et al., 2002). Due to both its longer length and slower mutation rate, the coding region has also shown a reduced occurrence of parallel mutations, or homoplasies, making phylogenetic estimates more reliable (Finnila et al., 2001). Thus, complete mitochondrial genome sequencing is becoming more frequent in evolutionary studies as sequencing costs decline. In humans, full coding region sequences have recently been used to improve phylogenetic resolution and identify new autochthonous haplogroups in India (Palanichamy et al., 2004; Sun et al., 2006), to examine the earliest human expansions throughout Eurasia (Maca-Meyer et al., 2001; Barnabas et al., 2006), and to describe global mtDNA diversity while estimating the age of modern humans (Ingman et al., 2000).

However, sequencing the entire coding region may not be necessary or desirable for all analyses. For instance, 8.8 kilobases (kb), i.e., half of the coding region, has been shown to capture the same level of diversity as the full coding region (Silva et al., 2002). Furthermore, working with large datasets carries an increased potential for sequencing errors (Sun et al., 2006). Unfortunately, little research has focused on determining which portions of the mitochondrial genome are most informative for phylogenetic analyses or estimating divergence times. Cummings et al. (1995) showed that individual mitochondrial genes are insufficient to construct interspecies topologies, but no comparable research has been performed on intra-species data. Simulation studies have found that, in the absence of recombination, extending sequence length beyond an optimal length does not increase the accuracy of coalescent estimations of a population genetic parameter like theta (Felsenstein, 2006; Pluzhnikov and Donnelly, 1996). However, no study has systematically determined the minimal amount of mtDNA sequence length sufficient for obtaining accurate estimates of divergence times.

In the present study, we investigate the relative importance of different regions of the human mtDNA genome using two criteria: (1) accurate reconstruction of a phylogeny and (2) consistent TMRCA estimates. We also determine the minimal amount of mtDNA sequence data necessary to meet these criteria. These criteria were chosen as they form the basis of most evolutionary studies and they provide two independent and statistically rigorous methods for testing the contribution of each region of the coding genome. For our study, we removed both individual genes and uniformly sized DNA fragments from the full coding region. The first criterion was assessed by statistically comparing phylogenies generated from the deleted datasets to a phylogeny estimated from the full coding region dataset (henceforth, the full-coding-region tree). The second criterion was tested using coalescent methods to estimate the TMRCA and associated 95% credible intervals of defined clades.

2. Materials and methods

2.1. Strategy for dataset selection

The dataset for this study consisted of 99 complete human mitochondrial genome sequences, 94 from India and western Eurasia (Palanichamy et al., 2004) and five from Africa (Mishmar et al., 2003). We chose this dataset because it incorporated a densely sampled population with a large number of potentially resolvable branch tips. This phylogenetically challenging dataset, which included many closely related sequences, provided a more rigorous test of our methods than would a broader sampling of older haplogroups with deeper branches that are easier to resolve. Furthermore, the sequences from India and western Eurasia belong to macrohaplogroup N and thus encompass the majority of haplotypes found outside of Africa (our results are also applicable to African haplogroups, see Section 4). Specifically, these sequences were chosen as a broad sample that represented each Western Eurasian haplogroup, including the more recent haplogroups of J, W, T, K, and V as well as the older ones such as U, R, and N (Palanichamy et al., 2004).

2.2. Sequence sub-datasets

Two series of sub-datasets were created in which the complete mitochondrial coding region was systematically reduced by deletions of major genes throughout the coding region (Dataset 1) and deletions of 10 contiguous 1545 bp fragments (Dataset 2) (Fig. 1). Genes were chosen as the unit of measure in the first dataset because a wealth of comparative datasets on individual mitochondrial genes already exists. However, since genes consist of varying lengths that might confound the analysis, we also chose to divide the full coding region into ten equal-size fragments. The control region was not included in these datasets because its mutation rate differs too greatly from that of the coding region to allow simultaneous testing in phylogenetic programs. To allow visual comparison of the various trees, a HVRI-only tree was also generated.

The 99 full coding region sequences were aligned in ClustalX 1.83 (Thompson et al., 1997), and then adjusted by hand to ensure proper alignment. This alignment was used in all subsequent phylogenetic and coalescent analyses. A general time reversible model of evolution with a proportion of invariable sites and a gamma distribution (GTR+I+G) was selected using Modeltest 3.06 (Posada and Crandall, 1998) and used in all phylogenetic and coalescent analyses.

2.3. Phylogenetic analyses

Maximum likelihood (ML) phylogenies were generated from Datasets 1 and 2 using PAUP* 4.0b10 (Swofford, 2000). The tree-bisection-and-reconnection (TBR) heuristic search method was used in all analyses. A sliding window

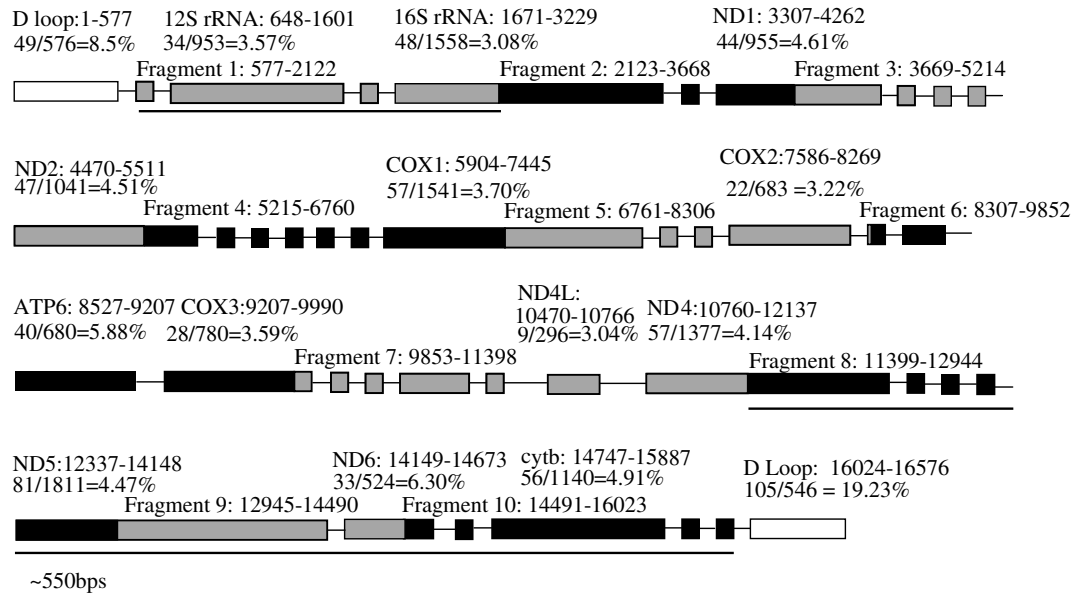


Fig. 1. Map of the mitochondrial genome. Genes deleted in Dataset 1 are represented by rectangles and the number and percentage of polymorphic sites in each gene sequence are indicated. Fragments deleted in Dataset 2 are represented by alternating segments of gray and black and include both genes and intergenic sequences. Transfer RNAs are indicated by boxes. Underlined areas represent regions identified as most informative by the phylogenetic analysis. Numbering is based on the Cambridge reference sequence [CRS] (Anderson et al., 1981) and gene positions are based on Mitomap (<http://www.mitomap.org>, 2006).

analysis was also performed on Dataset 2 to test the effects of removing larger regions of sequence (two to six contiguous fragments were removed). Trees were midpoint rooted. Seventeen sequences were removed in the phylogenetic analyses because they each differed from a second sequence by only one nucleotide, and thus, their phylogenetic location was known with near certainty. Haplogroups were identified based on definitions provided in Palanichamy et al. (2004). A phylogeny based on the HVRI was generated using the Bayesian method as implemented in MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001) because the PAUP* TBR heuristic search algorithm was unable to select a single best tree for the HVRI-only dataset.

Phylogenies generated from shortened datasets were statistically compared to the full-coding-region tree using Shimodaira's Approximately Unbiased (AU), Shimodaira–Hasegawa (SH), and Kishino–Hasegawa (KH) tests, as implemented in CONSEL (Shimodaira and Hasegawa, 2001). The AU test is the newest of the three methods and corrects for specific types of bias present in the KH and SH tests through use of a novel multiscale bootstrap technique (Shimodaira, 2002). The AU test is less conservative and considered more accurate than either KH or SH tests (Shimodaira, 2002). When only two trees are compared (as in our study), the KH and SH tests give identical results and are represented by a single line in Fig. 2.

2.4. Coalescent analyses

Using Dataset 2, a coalescent analysis in which individual fragments were deleted and a sliding window analysis in which two to six contiguous fragments were removed was

implemented in BEAST v1.3 (Drummond et al., 2005). TMRCA dates and 95% credible intervals were calculated for the root of the tree and for major haplogroups HV, R, and U in all analyses. All analyses were run under a constant population size model and a strict molecular clock with a mutation rate of 1.7×10^{-8} substitutions/site/year (Ingman et al., 2000; Pakendorf and Stoneking, 2005). Markov chains were run for 10,000,000 generations, sampled every 1000 generations, and the first one million generations were discarded as burn-in.

3. Results

3.1. Phylogenetic analyses

Thirteen ML phylogenies were generated from Dataset 1 (Fig. 1), in which individual genes were deleted. Comparisons with trees lacking COX3, ND4L, ND4, and ND6 produced p -values near 0.50, implying that the deleted-gene trees were nearly equivalent to the full-coding-region tree. Only one tree showed a significant difference when compared to the full-coding-region tree ($p_{AU} = 0.033$, deletion of 16S rRNA) although a second comparison approached significance ($p_{AU} = 0.057$, deletion of ND5) (Fig. 2a). The two trees that differed most from the full-coding-region tree were those lacking the two longest genes in the dataset (16S rRNA—1558 bp and ND5—1811 bp). In general, a weak relationship was observed between the length of the deleted gene and significance of the AU test as shown by a linear regression analysis ($r^2 = 0.52$). Removal of a single outlier in this analysis (ND4, $p_{AU} = 0.54$, 1400 bp) strengthened the relationship between gene length and significance of the

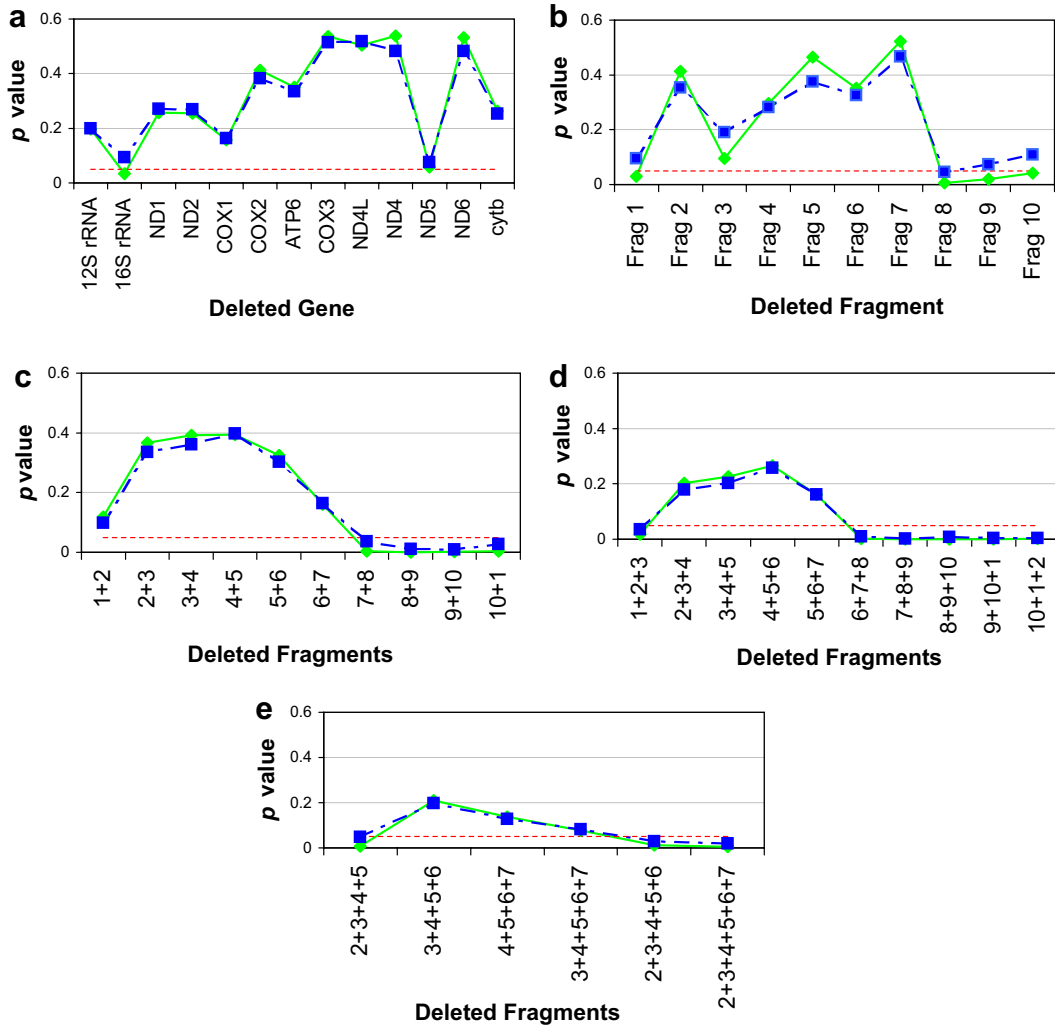


Fig. 2. Comparisons of a phylogeny based on the full-coding-region to phylogenies based on datasets with the indicated regions deleted. Any comparison that falls below $p = 0.05$ (indicated by the dotted line) means that a tree based on a deleted-region dataset differs significantly from the full-coding-region tree. Results from the AU test are represented by a solid line with diamonds while results from the SH and KH tests are represented by a single dashed line with boxes. Fragments are defined in Fig. 1. (a) Individual genes deleted (Dataset 1); (b) individual 1545 bp fragments deleted (Dataset 2); (c) segments of 2 fragments deleted; (d) segments of 3 fragments deleted; (e) segments of 4, 5, and 6 fragments deleted from the middle of the coding region (between fragments 2 through 7).

AU test ($r^2 = 0.81$). In addition to gene length, polymorphism within each gene was investigated. The genes targeted as most informative by the phylogenetic analysis did not contain the highest proportion of variable sites, i.e., ND5 and 16S rRNA contained only 4.5% and 3.1% polymorphic sites, respectively, while the most polymorphic genes were ND6 (6.3%) and ATP6 (5.9%) (Fig. 1). Average number of pairwise differences and nucleotide diversity was also calculated for each gene, but neither of these measures clearly correlated with the most informative genes as identified by the phylogenetic analysis (data not shown).

A second phylogenetic analysis was performed on a dataset in which equal-size fragments of 1545 bp were removed from the genome (Dataset 2, Fig. 1). Deletion of different regions resulted in differential loss of phylogenetic resolution despite the uniformity of fragment length, demonstrating that phylogenetic information is not uniformly distributed throughout the mitochondrial genome (Fig. 2b).

Significant loss of resolution occurred with deletion of fragments 1, 8, 9, or 10 ($p_{AU} = 0.029$, $p_{AU} = 0.0050$, $p_{AU} = 0.020$, $p_{AU} = 0.042$, respectively), which are the regions encompassing 12S rRNA and part of 16S rRNA and the majority of ND4 plus ND5, ND6, and cytb.

The results of the sliding window analysis confirmed that fragments 1, 8, 9, and 10 were most important for maintaining a phylogeny similar to the full-coding-region tree. When any of these four fragments was removed from the dataset, either alone or in combination with contiguous fragments, trees were generated that showed a significant difference from the full-coding-region tree (Fig. 2b, c, and d). In contrast, loss of any pair or trio of fragments from the middle of the coding region (between fragments 2 and 7) did not result in a tree that was significantly different than the full-coding-region tree (Fig. 2c and d). Not until four contiguous fragments were removed from the middle region did a significant change occur and then only with

loss of one group of fragments, i.e., 2+3+4+5 ($p_{AU} = 0.0070$; Fig. 2e). Removal of five fragments (3+4+5+6+7), which are equal to half of the coding region, still did not cause a significant change from the full-coding-region tree ($p_{AU} = 0.077$; Fig. 2e). It is interesting to note that deletion of fragment 2 only created a significantly different tree when combined with deletion of fragment 5 (Fig. 2e). This can be explained by the loss of two variants defining haplogroup H, i.e., bp 2706 and 7028, which are located in fragments 2 and 5, respectively. A phylogeny in which fragments 2+3+4+5 are deleted reveals a significant loss of substructure within haplogroup HV that can be attributed to loss of these two haplogroup H-defining variants (data not shown).

A visual comparison of the trees was also conducted in order to investigate finer details of the topology and branch lengths. Haplogroups U, HV, and R were highlighted in this analysis because they have a broad geographic range outside of Africa and they form monophyletic clusters in the full-coding-region tree (Fig. 3a). Phylogenies missing individual genes (Dataset 1) showed very few topological differences relative to the full-coding-tree; in fact, three trees (12S rRNA, COX1, ND2 deleted) were identical to the full-coding-region tree except for small differences in branch lengths (data not shown). A comparison of trees generated from Dataset 2 revealed the most significant change in the tree lacking fragment 8, with loss of monophyly for haplogroup R and a much shortened branch for haplogroup U (Fig. 3b). This result is expected, as fragment 8 contains the one defining variant for haplogroup R (bp 12,705) and all three haplogroup-defining variants for haplogroup U (bp 11,467, 12,308, and 12,372) as defined by (Palanichamy et al., 2004). The AU, KH, and SH tests also

showed the greatest change between a tree lacking fragment 8 and the full-coding-region tree ($p_{AU} = 0.005$ and $p_{KH/SH} = 0.045$) (Fig. 2b).

Although the HVRI dataset did not contain sufficient phylogenetic information to successfully create a single best maximum likelihood tree using PAUP*, a Bayesian tree for this region was generated for comparative purposes. The HVRI phylogeny revealed the most dramatic loss of resolution, in which the majority of the tree was reduced to an unresolved polytomy and haplogroups U, HV, and R were no longer visible (Fig. 3c).

3.2. Coalescent analyses

Estimates of TMRCA and associated credible intervals were investigated via a coalescent analysis of Dataset 2, in which ten equal-size fragments were individually removed from the dataset. Haplogroups U, HV, and R were again the focus in this analysis because they represent a broad range of coalescent events (~28,000–60,000 years before present [YBP]) and comparative TMRCA estimates have been published (Palanichamy et al., 2004; Richards et al., 2000). Means of TMRCA estimates and 95% credible intervals were plotted for Haplogroups U, HV, and R as well as the root of the entire tree for all trees generated using Dataset 2 (Fig. 4a). As a general confirmation of our methods, it is worth noting that TMRCA estimates for each haplogroup were in general agreement with previously published dates generated using non-coalescent methods (Palanichamy et al., 2004; Richards et al., 2000). Furthermore, TMRCA estimates of the root of the entire tree were within the range of published dates for the coalescence of human mtDNA, i.e., ~100,000–200,000 YBP

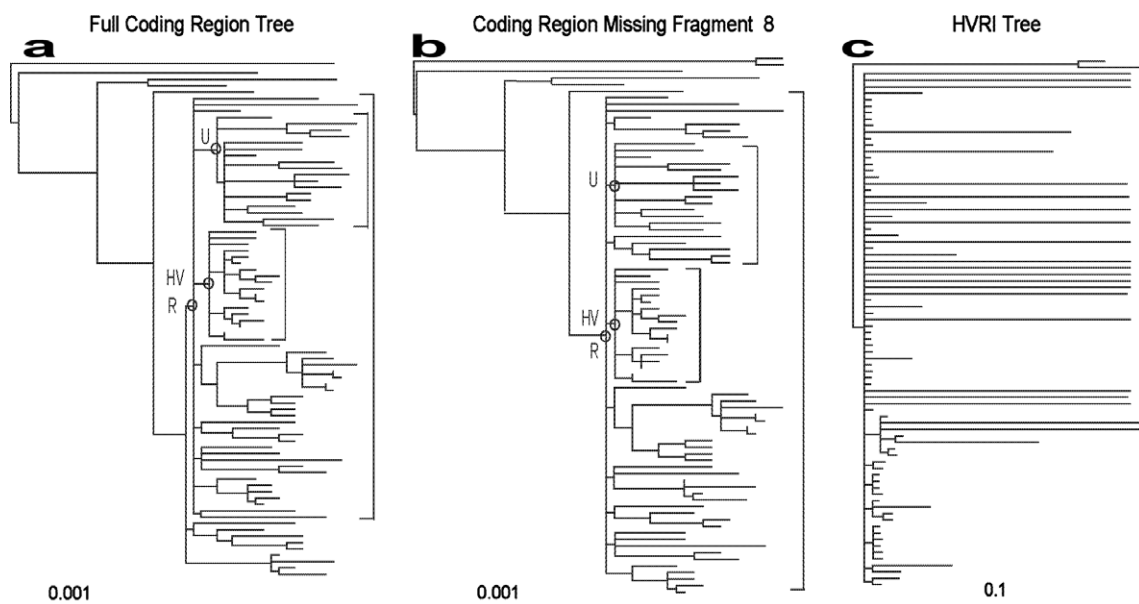


Fig. 3. Phylogenies based on: (a) full coding region dataset (ML); (b) dataset missing fragment 8 (ML) and; (c) HVRI (Bayesian). Haplogroups U, HV, and R are indicated on the phylogenies in which they are resolved (R has collapsed in 3b to include individuals of other N haplogroups, i.e., W, I1, and N1a). Scale indicates the number of substitutions.

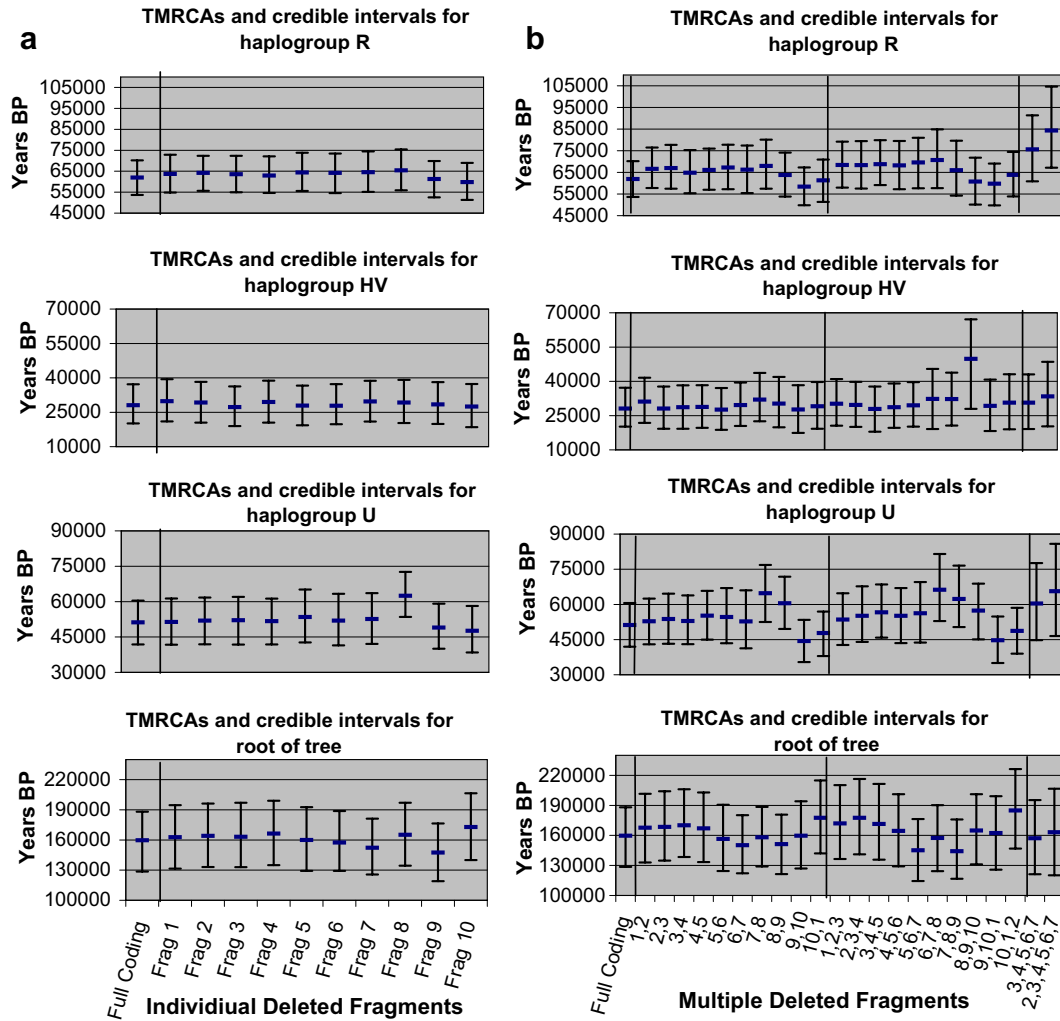


Fig. 4. Plot of TMRCA estimates and 95% credible intervals based on Dataset 2 with deletion of (a) individual fragments and (b) segments of two to six fragments.

(Cann et al., 1987; Vigilant et al., 1991; Ingman et al., 2000), as expected since our dataset includes a wide geographic range of individuals. Our analysis showed that loss of any 1545 bp fragment caused little variation in TMRCA estimates for the three haplogroups and the entire tree, as all mean dates were within the 95% credible interval of the other trees, with one exception (Fig. 4a). The exception was the tree lacking fragment 8, in which the mean TMRCA for haplogroup U was outside the 95% credible interval of the estimate from the full-coding-region tree as well as from several deleted-fragment trees (Fig. 4a). This result was consistent with that of the phylogenetic analysis, in which loss of fragment 8 also showed the most significant effect on the phylogeny (Fig. 3b), and could be at least partially explained by loss of all three U-defining variants within the deleted region.

A sliding window analysis, in which two contiguous fragments were removed, resulted in a small but consistent increase in credible intervals around all TMRCA means when compared to removal of only one fragment (Fig. 4b). Loss of two contiguous fragments caused little variation in TMRCA means, with two exceptions. When fragment pairs

7+8 and 8+9 were deleted, the TMRCA mean for haplogroup U fell outside the 95% credible interval for the full-coding-region tree. This result was also consistent with the previous analyses, in which loss of fragment 8 caused a change in the TMRCA estimate and in the phylogeny. When windows of three fragments were removed, deletion of fragments 8+9+10 had a strong effect on TMRCA estimation for haplogroup HV; the 8+9+10-deleted TMRCA mean occurred outside the credible intervals for all other TMRCA means and had a credible interval that was twice that of the full-coding-region estimate. Although individual loss of fragment 8, 9, or 10 caused little change in the TMRCA estimates for haplogroup HV, loss of all three fragments would be expected to affect the TMRCA estimate because the defining variants for haplogroup HV are located in fragments 8 (bp 11,719) and 10 (bp 14,766) (as defined by Palanichamy et al., 2004). Thus, only when both fragments 8 and 10 were removed from the dataset did the branch leading to haplogroup HV collapse, thereby incorporating many more sequences into the TMRCA calculation and pushing back the date for this haplogroup. Aside

from this exception, loss of any other three fragments caused little change in TMRCA means or credible intervals. Furthermore, loss of four contiguous fragments had minimal effect on TMRCA estimates, although the importance of combined fragments 8+10 for haplogroup HV remained evident (data not shown). Similarly, loss of five contiguous fragments from the middle of the coding region (3+4+5+6+7) maintained overlap of TMRCA credible intervals for all analyses and TMRCA means were nearly equivalent to full-coding-region means for haplogroup HV and the entire tree (Fig. 4b). Thus, half of the mitochondrial genome could be removed with no significant effect on TMRCA estimation in our dataset.

4. Discussion

The availability of full genome datasets enables an analysis of the information content throughout the mitochondrial genome in order to optimize the research design of future evolutionary studies. The goal of our study was to identify informative regions of the human mitochondrial genome using two computationally distinct measures: accurate reconstruction of a maximum likelihood phylogeny and coalescent-based estimates of TMRCA. These measures are relevant to most evolutionary studies as a well-resolved phylogeny is a prerequisite to many subsequent analyses and accurate TMRCA estimation with minimal credible intervals is crucial to allow discrimination between hypotheses involving different time frames. We find that sequencing a contiguous fragment from bp 11,399 through the control region to bp 3668 optimizes the information necessary for phylogenetic and coalescent analyses and also takes advantage of the wealth of data publicly available on the control region.

Two datasets were created so that we could study the information content of particular genes as well as the distribution of information in equal-sized fragments throughout the genome, i.e., individual genes were deleted in Dataset 1 and ten contiguous 1545 bp fragments were deleted in Dataset 2. In general, deletion of equal-size fragments had a greater effect on the phylogenetic analysis than deletion of single genes, most likely because haplogroup-defining variants often occur between genes. Specifically, phylogenetic resolution of our dataset was significantly affected by deletion of fragments 1, 8, 9, or 10 (bp 577–2122 and 11,399–16,023), which encompassed all or part of 12S rRNA, 16S rRNA, ND5, ND6, and cytb (Fig. 2). Deletion of fragments 2 and 5, in combination, also resulted in a phylogeny that was significantly different than the full-coding-region tree. In contrast, deletion of fragments 3 through 7, which is equivalent to half the mitochondrial genome, maintained a phylogeny that did not differ significantly from the full-coding-region tree. In comparison to the phylogenetic analysis, the coalescent estimations of TMRCA were less sensitive to deletion of portions of the mitochondrial genome. In fact, no deletion, even of multiple fragments, produced a TMRCA estimate that did not overlap with the full-coding-

region estimate. Deletion of fragment 8 (individually and in some combinations) was the only analysis that produced a TMRCA mean outside the full-coding-region credible interval, although both sets of credible intervals still overlapped. Deletion of fragments 3 through 7 resulted in little change in TMRCA estimation and, in fact, TMRCA means for haplogroup HV and the entire tree were virtually identical to the full-coding-region estimates (Fig. 3).

An investigation of gene characteristics revealed that neither gene length, level of polymorphism, nucleotide diversity, nor average pairwise differences was correlated with phylogenetic informativeness. One factor that did contribute to maintenance of a phylogeny was inclusion of fragments that contained defining variants for specific haplogroups. Many of the informative fragments identified by our analysis contain variants that were previously recognized as haplogroup-defining variants (Palanichamy et al., 2004). As expected, removal of fragments containing these variants caused loss of phylogenetic resolution. For example, we observed that loss of fragment 8, containing the sole R-defining variant, led to loss of monophyly of haplogroup R (Fig. 2). However, these few haplogroup-defining variants were not entirely responsible for maintaining a phylogeny as demonstrated by the same tree missing fragment 8, in which the three defining variants for haplogroup U were lost, but monophyly was maintained (although the branch leading to haplogroup U was shortened) (Fig. 3).

This analysis can be extended to additional sequences, such as those from macrohaplogroup L in Africa and macrohaplogroup M in Africa and Asia. As described above, our analysis points out the importance of including at least one unique defining site to ensure monophyly of a haplogroup. It is significant that our recommended region of analysis (bp 648–3668 and 11,399–16,033, plus the control region) includes at least one haplogroup-defining variant for haplogroups L0 (bp 3516, 13,276), L1 (3666, 13,789, 14,178), L2 (bp 13,590, 16,311, 16,390), L3 (bp 769), L4 (16,362) and M (bp 14,783, 15,043) (Kivisild et al., 2004, 2006). In fact, nearly half of the defining variants for these haplogroups (12, as listed above, out of a total of 25) are included in our recommended region. Furthermore, reconstruction of a phylogeny with deeper branches, e.g., a dataset including macrohaplogroups L and M, is theoretically less challenging than our dataset of densely sampled N haplotypes. Thus, our recommendation to sequence the half of the mitochondrial genome that encompasses fragments 1, 2, 8, 9, 10 and the control region can be extended to the study of all major human haplogroups.

Our analysis demonstrates that accurate phylogenetic reconstruction is more sensitive than coalescent TMRCA estimation to loss of mitochondrial sequence data. This result reflects the fact that small changes in tip topology may introduce significant differences between phylogenies, but would not necessarily affect TMRCA estimates that focus on interior or basal nodes. Additionally, uncertainty in tree structure is explicitly accounted for when estimating population genetic parameters in coalescent analyses,

meaning that coalescent TMRCA estimates are generally more robust to changes in topology or loss of phylogenetic resolution. For example, in our dataset, loss of fragments 2 and 5 removed two haplogroup H-defining variants, which caused loss of monophyly for haplogroup H within haplogroup HV. This resulted in a significant difference in phylogenies and haplogroup HV topology (compared to the full-coding-region tree, see Fig. 2e), but virtually no difference in TMRCA estimation for haplogroup HV (see Fig. 4b). Thus, as expected, more sequence data were needed to maintain accurate construction of a phylogeny than were necessary for consistent TMRCA estimation.

Our study suggests that the optimal dataset for mitochondrial DNA evolutionary studies includes fragments 1, 2, 8, 9, and 10 (bp 648–3668 and 11,399–16,033) based on the fact that deletion of fragments 3 through 7 caused no significant change in the resulting phylogeny and minimal variation in estimates of TMRCA. Conveniently, fragments 1+2 and 8+9+10 are adjacent to the control region. Thus, sequencing a contiguous fragment from bp 11,399 through the control region to bp 3668 would create a dataset that optimizes the information necessary for phylogenetic and coalescent analyses and also takes advantage of the wealth of data publicly available on the control region. Furthermore, our results demonstrate that only half the genome is necessary for maintaining an accurate phylogeny and consistent TMRCA estimates, suggesting that researchers could potentially cut their sequencing length in half. Given the recent report that adding loci contributes much more information to a study than increasing sequence length or adding samples (Felsenstein, 2006), our findings are consistent with a recommendation for reduced mitochondrial sequence length and analysis of additional loci.

References

- Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J., Staden, R., Young, I.G., 1981. Sequence and organization of the human mitochondrial genome. *Nature* 290, 457–465.
- Bandelt, H.J., Herrnstadt, C., Yao, Y.G., Kong, Q.P., Kivisild, T., Rengo, C., Scozzari, R., Richards, M., Villems, R., Macaulay, V., Howell, N., Torroni, A., Zhang, Y.P., 2003. Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: some caveats. *Ann. Hum. Genet.* 67, 512–524.
- Barnabas, S., Shouche, Y., Suresh, C.G., 2006. High-resolution mtDNA studies of the Indian population: implications for palaeolithic settlement of the Indian subcontinent. *Ann. Hum. Genet.* 70, 42–58.
- Cann, R.L., Stoneking, M., Wilson, A.C., 1987. Mitochondrial DNA and human evolution. *Nature* 325, 31–36.
- Cummings, M.P., Otto, S.P., Wakeley, J., 1995. Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.* 12, 814–822.
- Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–1192.
- Felsenstein, J., 2006. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Mol. Biol. Evol.* 23 (3), 691–700.
- Finnila, S., Lehtonen, M.S., Majamaa, K., 2001. Phylogenetic network for European mtDNA. *Am. J. Hum. Genet.* 68, 1475–1484.
- Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755.
- Ingman, M., Kaessmann, H., Paabo, S., Gyllenstein, U., 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* 408, 708–713.
- Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., Parik, J., Geberhiwot, T., Usanga, E., Villems, R., 2004. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *Am. J. Hum. Genet.* 75, 752–770.
- Kivisild, T., Shen, P., Wall, D.P., Do, B., Sung, R., Davis, K., Passarino, G., Underhill, P.A., Scharfe, C., Torroni, A., Scozzari, R., Modiano, D., Coppa, A., de Knijff, P., Feldman, M., Cavalli-Sforza, L.L., Oefner, P.J., 2006. The role of selection in the evolution of human mitochondrial genomes. *Genetics* 172, 373–387.
- Maca-Meyer, N., Gonzalez, A.M., Larruga, J.M., Flores, C., Cabrera, V.M., 2001. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet.* 2, 13.
- Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., Sukernik, R.I., Olckers, A., Wallace, D.C., 2003. Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. USA* 100, 171–176.
- Pakendorf, B., Stoneking, M., 2005. Mitochondrial DNA and human evolution. *Annu. Rev. Genomics Hum. Genet.* 6, 165–183.
- Palanichamy, M.G., Sun, C., Agrawal, S., Bandelt, H.J., Kong, Q.P., Khan, F., Wang, C.Y., Chaudhuri, T.K., Palla, V., Zhang, Y.P., 2004. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. *Am. J. Hum. Genet.* 75, 966–978.
- Pluzhnikov, A., Donnelly, P., 1996. Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144, 1247–1262.
- Posada, D., Crandall, K.A., 1998. MODEL TEST: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.
- Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., Villems, R., Thomas, M., Rychkov, S., Rychkov, O., Rychkov, Y., Golge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Cali, F., Vona, G., Demaine, A., Papaha, S., Triantaphyllidis, C., Stefanescu, G., Hatina, J., Belledi, M., Di Rienzo, A., Novelletto, A., Oppenheim, A., Norby, S., Al-Zaheri, N., Santachiara-Benerecetti, S., Scozzari, R., Torroni, A., Bandelt, H.J., 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* 67, 1251–1276.
- Shimodaira, H., 2002. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51, 492–508.
- Shimodaira, H., Hasegawa, M., 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17, 1246–1247.
- Silva Jr., W.A., Bonatto, S.L., Holanda, A.J., Ribeiro-Dos-Santos, A.K., Paixao, B.M., Goldman, G.H., Abe-Sandes, K., Rodriguez-Delfin, L., Barbosa, M., Paco-Larson, M.L., Petzl-Erler, M.L., Valente, V., Santos, S.E., Zago, M.A., 2002. Mitochondrial genome diversity of Native Americans supports a single early entry of founder populations into America. *Am. J. Hum. Genet.* 71, 187–192.
- Swofford, D.L., 2000. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
- Sun, C., Kong, Q.P., Palanichamy, M.G., Agrawal, S., Bandelt, H.J., Yao, Y.G., Khan, F., Zhu, C.L., Chaudhuri, T.K., Zhang, Y.P., 2006. The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. *Mol. Biol. Evol.* 23, 683–690.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876–4882.
- Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., Wilson, A.C., 1991. African populations and the evolution of human mitochondrial DNA. *Science* 253, 1503–1507.