The Identification of Single Nucleotide Polymorphisms in the Entire Mitochondrial Genome to Increase the Forensic Discrimination of Common HV1/HV2 Types in the Caucasian Population

By

Michael DeWitt Coble

B.S. 1991, Appalachian State University, Boone, NC
M.F.S. 1997, The George Washington University, Washington, DC

A Dissertation Submitted to

The Faculty of

Columbian College of Arts and Sciences of
The George Washington University

In Partial Satisfaction of the Requirements for the
Degree of Doctor of Philosophy

January 30, 2004

Dissertation Directed By

Thomas J. Parsons

Adjunct Associate Professor of Genetics

DEDICATION

For my wife,
Karen

and

our children,
Katherine
Sophia
Matthew

Thank you for your love and encouragement

For my mother and late father,
Phyllis and Roland DeWitt Coble

Thank you for your inspiration and never ending support

ACKNOWLEDGEMENTS

I would like to thank the many people in my life that helped me achieve this goal. First, I wish to acknowledge my advisor, Dr. Thomas Parsons, for support and guidance leading to the completion of this dissertation. I am grateful that he provided me the opportunity to join the team at AFDIL, and to contribute to the mission of the laboratory. I would also like to thank current and previous members of the Research section for their support and friendship, especially Jodi Irwin for being such a great friend and colleague over the years.

This work was supported by a National Institutes of Justice (NIJ), Office of Justice Programs, Department of Justice grant 2000-1J-CX-K010. The opinions and assertions contained herein are solely those of the author and are not to be construed as official or as views of the U.S. Department of Defense, U.S. Department of the Army, or the U.S. Department of Justice.

I want to thank Dr. Lois Tully at the NIJ for her support and her recognition of the significance of this project from the beginning. To current and former members of the NIJ team: Jennifer O'Callaghan, Jessica Saunier, Christine Harvie, and Ilona Letmanyi: thanks for all of the help and assistance you have given to me over the years. A very special thanks goes to Rebecca Just, Supervisory Research Technologist at AFDIL. You have taken the reins of the SNP typing portion of the NIJ grant, and have made it successful. Your constant, hard work has been crucial for the success of this project, and you have been so patient with me as I have heavily relied on your laboratory and organizational skills. I would also like to thank the many student interns that have helped

me over the years: Rachel Barry, Trina Bersola, Serena Filosa, Victoria Glynn, Carrie Guyan, William Ivory, and Devon Pierce. I appreciate all of your help, and it has been a pleasure getting to know each of you. I am also grateful for the new friendships made through collaborations associated with this project: Dr. Walther Parson and Harold Niederstaetter at the Institute of Legal Medicine in Innsbruck, Austria; and Dr. John Butler and Dr. Pete Vallone at the National Institutes of Standards and Technology. And, to the rest of the team at AFDIL, words cannot tell you how much I have enjoyed working with so many experts in the mtDNA world. To Rob Fisher, Chad Ernst, and Chris Los, thanks for being there. I would be remiss without thanking Col. Brion Smith, the Chief Deputy Medical Examiner of the DOD DNA Registry. I can never repay you for your kind words and support. Your leadership is second to none.

I also extend my deepest gratitude to my dissertation advisory committee members Dr. Marc Allard, Dr. Carl Merril, and Dr. Dudley Strickland. Thank you for your time and useful comments. I would also like to thank Dr. Diana Lipscomb and Dr. Moses Schanfield for their willingness to serve on my dissertation defense committee.

I must also acknowledge members of The George Washington University Institute of Biomedical Sciences and the program in Genetics. Thank you for accepting me into the program; it has been a great experience. For all of my "classmates" that started with me in the IBS program, especially Henry Nguyen, Lori Brandes, and Jennifer Mariner, I am forever indebted to you all. I am extremely grateful for Dr. Diana Johnson, the Director of the Genetics program. You have always given your time for my "quick questions" as a doctoral candidate.

I would like to friends and family members for their support and encouragement over the years. I would like to thank my wife, Karen, and my children, Katherine, Sophia, and Matthew. They have tolerated my absence at the dinner table while being stuck on the beltway, and have given me nothing but encouragement and love in return. I would also like to thank my mother, Phyllis, and my brother, Douglas for their continuous love, support, and understanding. I would also like to thank my mother and father-in-law, Gerda and Arthur Shipe for their help and assistance. For my good friend, Jose Ruiz, your sense of humor has kept me going all this time. It's been great having our laboratories near to one another, and I look forward to continuing our friendship beyond the dissertations.

Finally, I want to give a special thanks to my late father, Roland DeWitt Coble. Thanks, Dad. You were always there with love, support, and encouragement. You are my inspiration. I wish you could be here to share this momentous occasion with me.

ABSTRACT

Mitochondrial DNA (mtDNA) typing has found an important niche in the forensic testing of degraded samples and shed hairs. Currently, most forensic mtDNA laboratories focus on sequence information within the two hypervariable regions (HV1 and HV2) of about 600+ bases within the control region. The distribution of mtDNA types is highly skewed toward rare types, making the significance of a match for a mtDNA type previously unseen in a database quite substantial. There are also a number of common types observed in various populations. One limitation of mtDNA testing is the low power of discrimination associated with common HV1/HV2 types. For example, in the European Caucasian forensic database, there are approximately twenty common HV1/HV2 types that occur at a population frequency of 0.5% or greater, for an aggregate frequency of about twenty-one percent of the population.

We have sequenced the entire mtDNA genome (mtGenome) of 241 Caucasian individuals who match one of eighteen common HV1/HV2 types in order to identify single nucleotide polymorphisms (SNPs) in the coding region useful for additional discrimination. Focusing on SNPs that were shared, neutral and non-redundant, we have developed a set of eight multiplex panels containing 59 informative sites suitable for SNP typing assays. Each panel contains seven to eleven SNPs, and is specific to discriminating one or more of the common HV1/HV2 types in the Caucasian population. The discrimination provided by the multiplex panels provides maximal discrimination while preserving limited DNA extract from forensic casework. Applying all eight multiplex panels to the 241 sequences resolved the individuals into 106 haplotypes, 56 of

which were unique, a nearly 6-fold improvement over the initial 18 common HV1/HV2 types.

We have also investigated evolutionary properties of the 59 discriminating SNPs by characterizing the mutation rates in the mtDNA coding region using phylogenetic trees constructed by parsimony. Most of the SNPs that discriminated among the 18 common HV1/HV2 types in Caucasians (51/58. 88%) can be classified as having relatively slow mutation rates, indicating that these sites are narrowly useful for resolving within these specific common HV1/HV2 types. The remaining SNPs in the multiplex assay could be classified as having relatively fast rates. The decision to use brute-force whole mtGenome sequencing was necessary to discover sites specific for resolving common HV1/HV2 types in Caucasians. The strategy of mtGenome sequencing for identifying discriminatory SNPs will be required to resolve common HV1/HV2 types in other forensically important groups (African Americans, Hispanics).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I. Introduction

I. 1. Mitochondrial DNA as a Genetic Marker

The human mitochondrial DNA (mtDNA) genome was sequenced and enumerated over twenty years ago (Anderson *et al.*, 1981). The approximately 16,569 base pair genome encodes 13 polypeptides, 22 tRNAs, and 2 rRNA subunits (Figure 1). This closed, double-stranded, circular genome can be classified according to function: the coding region (about 15.5 kb of the genome) and the non-coding control region (about 1.1 kb of the genome). All of the genes in the coding region are highly concatenated and only occasionally have non-coding "spacer" sequences separating the genes. The control region has an important regulatory function for the mitochondria and contains sequences to initiate both transcription and DNA replication of the heavy strand. Many laboratories have focused on sequences within the non-coding control region of the mtDNA genome (mtGenome), specifically hypervariable regions I and II (HV1 and HV2), they span roughly positions 16024-16383 and 57-372, respectively (numbered correspondingly to the reference sequence, Anderson *et al.*, 1981). Although the boundaries of HV1 and HV2 are not rigidly defined, the dense array of polymorphisms within HV1/HV2 has made this region an attractive target for sequencing studies of mtDNA variation.

One role of this cellular organelle is to provide energy in the form of Adenosine Tri-Phosphate (ATP) through oxidative phosphorylation (OXPHOS) reactions inside the mitochondria. The small number of polypeptides encoded with the mtDNA genome represents only a small fraction of the total proteins necessary for mitochondrial function. Most of these proteins are encoded in the nuclear DNA genome and are subsequently exported to the mitochondria.

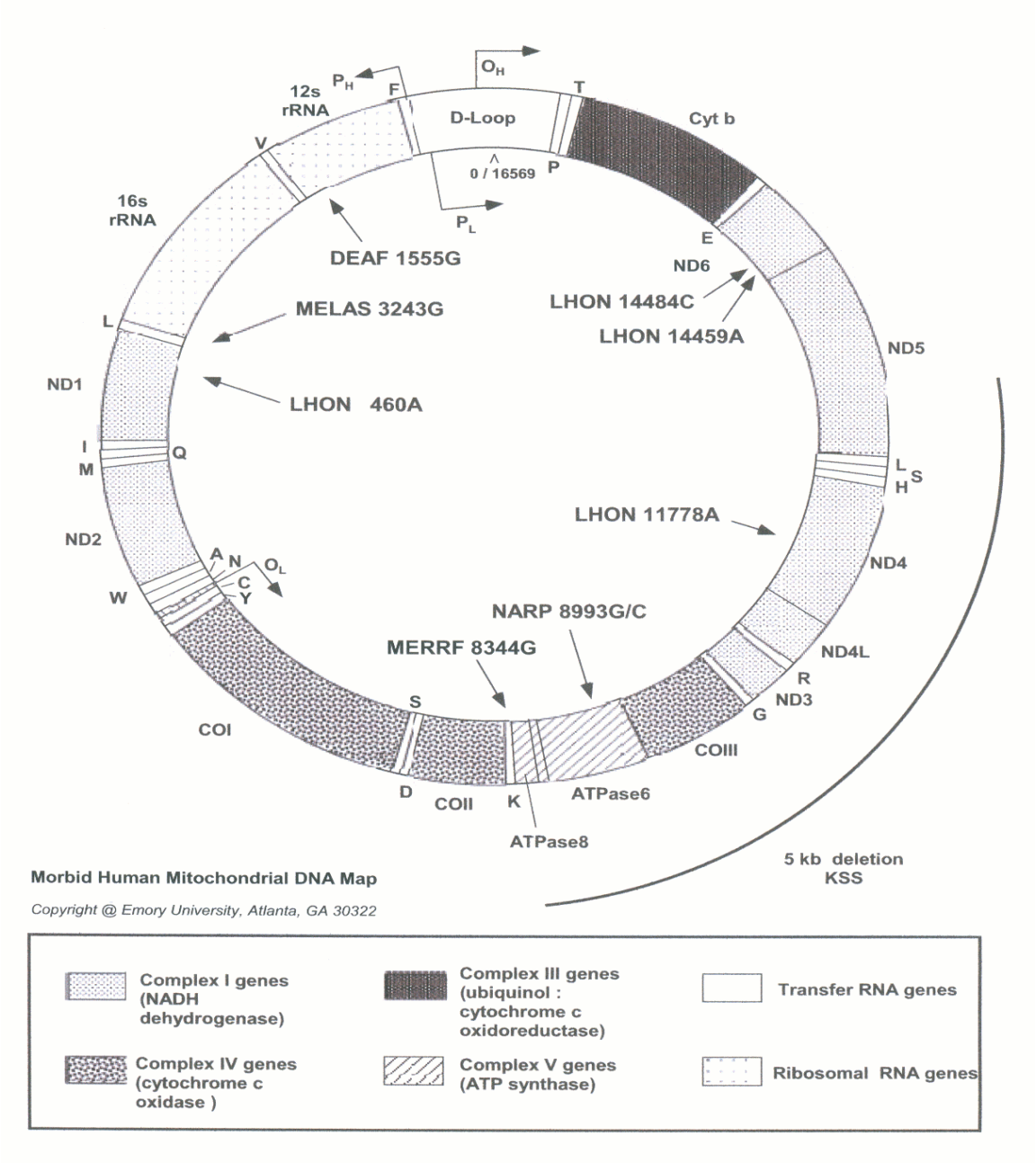


Figure 1. The Human Mitochondrial DNA Genome. The genes encoded by the mitochondrial DNA (mtDNA) genome are noted. Point mutations associated with mitochondrial diseases are noted in the center of the genome. Diagram provided by MitoMap (http://www.mitomap.org/).

Several characteristics of mtDNA make it an ideal choice for studying human evolution and genetic variation. First, human mtDNA is wholly maternally inherited and passed to the offspring from the expansion of mitochondria in the oocyte (Giles *et al.* 1980). Second, since the mtDNA has been shown not to recombine in humans (Ingman *et al.* 2000; Elson *et al.* 2001), accumulated differences between any two molecules reflect changes that have occurred since they shared a common ancestor. Finally, mtDNA has been shown to have a mutation rate up to ten times the mutation rate of single copy nuclear DNA genes (Brown *et al.* 1979; Horai *et al.* 1995).

I. 1. 1. RFLP and Control Region Analyses

Early studies of mtDNA identified mitochondrial mutations based on Restriction Fragment Length Polymorphism (RFLP) data, pioneered by Wesley Brown and Doug Wallace (Brown 1980; Denaro *et al.* 1981). The use of "low resolution" restriction analysis (using 5-6 restriction enzymes) revealed a phylogeny that was "star-like", with a central mitochondrial type shared among diverse populations throughout the world (Johnson et al 1983). Using "high resolution" restriction analysis (12 restriction enzymes), an analysis by Cann *et al.* (1987) revealed a more highly resolved phylogenetic tree of mtDNA. In this phylogeny, a set of African sequences were observed to be basal to other African and non-African sequences. This "Out of Africa" (OOA) hypothesis and the African "mitochondrial Eve" ancestry of modern humans proposed by Cann *et al.* (1987) was fundamental in establishing the potential of the mtDNA molecule as a genetic marker for studies of human evolution.

In the late 1980s and early 1990s, as DNA sequencing became cheaper, easier, and faster to perform, mtDNA sequencing of the control region became an additional method utilized for genetic studies of human mtDNA evolution. Vigilant *et al.* (1991) sequenced a portion of the control region and confirmed the OOA hypothesis (Cann *et al.*, 1987). However, a subsequent reanalysis of this data by Templeton (1992), found phylogenetic trees that were shorter in length than the most parsimonious tree (MPT) of Vigilant *et al.* (1991). Some of the phylogenetic trees produced by Templeton (1992) did not support an African base of the human evolution. This gave rise to a period of intense debate, but additional data and better methods of constructing phylogenetic trees have since solidified support for the OAA hypothesis using control region sequences (e.g. Penny *et al.*, 1995; Watson *et al.*, 1997).

The analysis of RFLP variation in global populations permitted the identification of a number of monophyletic clades in which all mtDNAs could be classified. These clades, or haplogroups, were distinguished by ancient mutations that occurred thousands of years ago. Continent-specific markers have been identified for Africans (Chen *et al.,* 1995), European Caucasians (Torroni *et al.,* 1994; 1996) and Asians (Schurr *et al.,* 1990). Haplogroups are designated in the literature by a capital letter followed by a number representing subclusters of the haplogroup (Schurr *et al.,* 1990; Torroni *et al.,* 1996; Richards *et al.,* 1998). Nearly all African mtDNAs can be classified into one of three main haplogroups (L1, L2 or L3). Haplogroup L1 is considered to be ancestral since it is found at the root of the human mtDNA phylogeny. The Eurasian super-haplogroups M and N are believed to have left Africa approximately 50,000 – 60,000 years ago to give

rise to the current haplogroups in Asia (A, B, C, D, F) and Europe (H, I, J, K, T, U, V, W, X).

Since the mtGenome is one linked molecule, sequence motifs in the control region also correlate strongly with haplogroups defined by RFLPs (Torroni *et al.*, 1996; Macaulay *et al.*, 1999a). However, due to the high mutation rate and propensity for reversion mutations (discussed further below), haplogroup assignment based on control region data is tentative. For example, the control region polymorphism 73A is associated with haplogroup H (Torroni *et al.*, 1996; Macaulay *et al.*, 1999a; Allard *et al.*, 2002). This is a relatively fast site in the control region (Meyer *et al.*, 1999; Allard *et al.*, 2002), making the site an often-observed homoplasious character in phylogenetic analyses. There are several examples in the literature where individuals belonging to haplogroup H have undergone reversions to 73G (Torroni *et al.*, 1996; Macaulay *et al.*, 1999a; Allard *et al.*, 2002). Since the control region sequence motif 73G is associated with haplogroup U, it is impossible to unambiguously classify individuals having the 73G polymorphism to a haplogroup without additional information from the coding region.

I. 1. 2. Molecular Evolution Studies Using mtDNA

Additional molecular evolution studies using mtDNA control region sequences have demonstrated that mtDNA evolves in a complex manner. For example, Aquadro and Greenberg (1983) determined that base composition is unequal in human mtDNA, and found that transitions occur much more frequently than transversions (32:1 ratio). Additionally, Aquadro and Greenberg (1983) recognized several instances of reversions,

or homoplasies, in their sequences, and that the distribution of substitutions in their data was non-random.

The site-to-site variability in mutation rates in the control region is one of the more peculiar features of mtDNA. Most sites in the control region have very low mutation rates, or are invariant, while other sites mutate extremely quickly. This extreme rate heterogeneity has been established by many studies, using a variety of methods (Wakeley, 1993; Hasegawa *et al.*, 1993; Meyer *et al.*, 1999; Excoffier and Yang, 1999; Pesole and Saccone, 2001; Malyarchuck *et al.*, 2002). Nonetheless, the complexities of mtDNA evolution have yet to be fully characterized or explained. For example, recent studies using pedigree analyses have determined an unexpectedly high empirical mutation rate in the coding region (Howell *et al.*, 1996; Parsons *et al.*, 1997; Sigurdardottir *et al.*, 2000; Heyer *et al.*, 2001). The initial estimates of mutation rates determined by pedigree studies were ~10-fold higher than those estimated by phylogenetic studies (Howell *et al.*, 1996; Parsons *et al.*, 1997). The question of whether mutational fast sites were the source of this discrepancy between pedigree and phylogenetic rate estimations was a source of much debate (Paabo, 1996; Jazin *et al.*, 1998; Parsons and Holland, 1998). Howell *et al.* (2003) has recently published the results of a large-scale pedigree study and has again confirmed the discrepancy of the mutation rate in pedigree versus phylogenetic estimations. It may be that there is no single factor that accounts for this discrepancy. Mutational hotspots, genetic drift, selection, and the inability of phylogenetic methods to properly account for mutation rate heterogeneity may all contribute to the disparity between the phylogenetic derived

mutation rate and the empirical pedigree mutation rate (Parsons *et al.*, 1997; Howell *et al.*, 2003).

I. 1. 3. Entire mtGenome Sequencing

Most of the population genetic analyses of mtDNA for nearly twenty years have come from the few hundreds of nucleotides within the two hypervariable regions and the RFLP fragments, discussed above. The emergence of "mitogenomics" began with the landmark publication of 53 complete human mtGenomes from diverse global populations in late 2000 by Ingman and colleagues. The robust phylogenetic tree from this global sample confirmed previous research (Cann *et al.*, 1987; Vigilant *et al.*, 1991) that humans migrated out of Africa to populate the world. Soon afterward, Finnila *et al.* (2000) used Conformation-Sensitive Gel Electrophoresis (CSGE) for analysis of coding region sequences of 22 Finnish individuals belonging to haplogroup U. Finnila *et al.* (2001) followed up this effort with a phylogenetic analysis of 192 mtGenomes from Finnish samples spanning all European Caucasian haplogroups. The maximal information afforded by complete mtGenomes has allowed a "coming of age" for mitochondrial gene trees (Richards and Macaulay, 2001).

As more laboratories generate entire mtGenomes, the fruits of mitogenomics are coming to bear in a number of areas. In addition to the confirmation of the out of African hypothesis (Ingman *et al.*, 2000 and Maca-Meyer *et al.*, 2001), the lack of recombination in mtDNA has been rigorously confirmed (Ingman *et al.* 2000; Elson *et al.* 2001) – despite a flurry of debate to the contrary (Awadalla *et al.*, 1999; Eyre-Walker *et al.*, 1999;

Macaulay *et al.*, 1999b; Jorde and Bamshad, 2000; Kivisild and Villems, 2000; Kumar *et al.*, 2000; Parsons and Irwin, 2000; Wiuf, 2001; Eyre-Walker and Awadalla, 2001; Innan and Nordborg, 2002; Hagelberg, 2003). Entire mtGenomes have been used to discover novel polymorphisms associated with haplogroups (Finnila *et al.*, 2001 and Herrnstadt *et al.*, 2002). Torroni *et al.* (2001) observed differences in the mutation rate of African haplogroup L2 subtypes. Mishmar *et al.* (2003) proposed that selection plays a role in the mtGenome variation among different haplogroups, and Meyer and von Haeseler (2003) have studied site-specific substitution rates in the entire mtGenome.

I. 2. MtDNA as a Tool for Forensics

MtDNA analyses have also found an important role in the forensic DNA community (reviewed in Holland and Parsons 1999). One advantage of using mtDNA in forensic casework is the high copy number of mtDNA molecules within the cell. Nuclear autosomal loci used in forensics are present in only two copies per cell whereas mtDNA is present in approximately 500 to 2,000 copies per mammalian cell (Piko *et al.*, 1976; Michaels *et al.*, 1982). It is correspondingly more likely that some amplifiable copies of the mtDNA will survive in highly degraded samples, permitting the analysis of evidence samples that otherwise would not give a nuclear DNA profile. Samples that are typically submitted for mtDNA analysis include degraded bloodstains, bones, saliva, fingernails, and hair shafts. MtDNA typing of hair shafts is a particularly important application as shed hairs are commonly found as sources of evidence. Currently, forensic hair comparisons of evidentiary and reference specimens are based upon a set of

morphological characteristics. These analyses tend to be subjective, relying on the experience and judgment of the examiner (Bisbing, 1982). In an attempt to allay the subjectivity associated with hair analysis, an increasing number of crime labs are moving toward performing mtDNA testing. Moreover, mtDNA can further discriminate among hairs that cannot be otherwise excluded as coming from the same individual.

I. 2. 1. Forensic mtDNA Testing

Currently, most forensic laboratories using mtDNA typing focus on the sequence information within HV1 and HV2 (Holland *et al.* 1993; Wilson *et al.* 1993). As is often the case in identifying skeletal remains of missing soldiers (one of the principal missions at the Armed Forces DNA Identification Laboratory – AFDIL), the analyst compares the sequences of the skeletal remains to a maternal family reference sequence. If these sequences differ at multiple positions, an unequivocal exclusion can be made. If the sequences match, this is consistent with the hypothesis of maternal relatedness. As a consequence of the strict maternal inheritance of mtDNA, an individual is expected to match all of his maternal relatives (barring mutation), of which there may be a large number in the population. It is also possible that individuals not related maternally could, by chance, exhibit the same HV1/HV2 type due to homoplasmic mutations. Finally, unlike STR loci, the mtGenome is one linked molecule not subject to recombination. Multiplication of the population frequencies of polymorphisms in HV1 and HV2 cannot be used to determine the frequency of a mtDNA profile. For the reasons listed above, an mtDNA match by itself cannot be considered a conclusive identifier. In order to

determine the significance of an mtDNA match, one must make reference to the observed frequency of that polymorphism in a relevant database (Holland and Parsons, 1999).

I. 2. 2. The Frequency Distribution of Caucasian Haplotypes

An international effort has resulted in the formation of a forensic mtDNA database containing over 4800 forensic mtDNA profiles. This database has recently been made available on the internet (via *Forensic Science Communications* at http://www.fbi.gov - see Monson *et al.* 2002). The HV1/HV2 sequence range for this database spans from: HV1, positions 16024-16365 and HV2, positions 73-340. This sequence range is also used by AFDIL, and will heretofore be considered the default sequencing regions involving HV1 and HV2. Pairwise comparisons of the Caucasian sequences in the database (N=1655) provide an informative picture of the distribution of mtDNA HV1/HV2 types (Figure 2). Insertions in the highly variable C-stretch region of HV2 were ignored in this comparison, given the unstable hypermutability in this region (Stewart *et al.*, 2001). The frequency distribution of haplotypes in the Caucasian population is highly skewed in a continuous manner. At the one end of this "L-shaped" curve are the number of HV1/HV2 sequences that represent "unique" types seen only once in the database (839/1665 [50.4%] of the HV1/HV2 types). At the other end of the distribution are a small number of common, shared HV1/HV2 types in the Caucasian database. The most common HV1/HV2 type occurs at a frequency of about 7% in the population (Figure 2 - see also Lutz-Bonengel *et al.*, 2003). Although this distribution

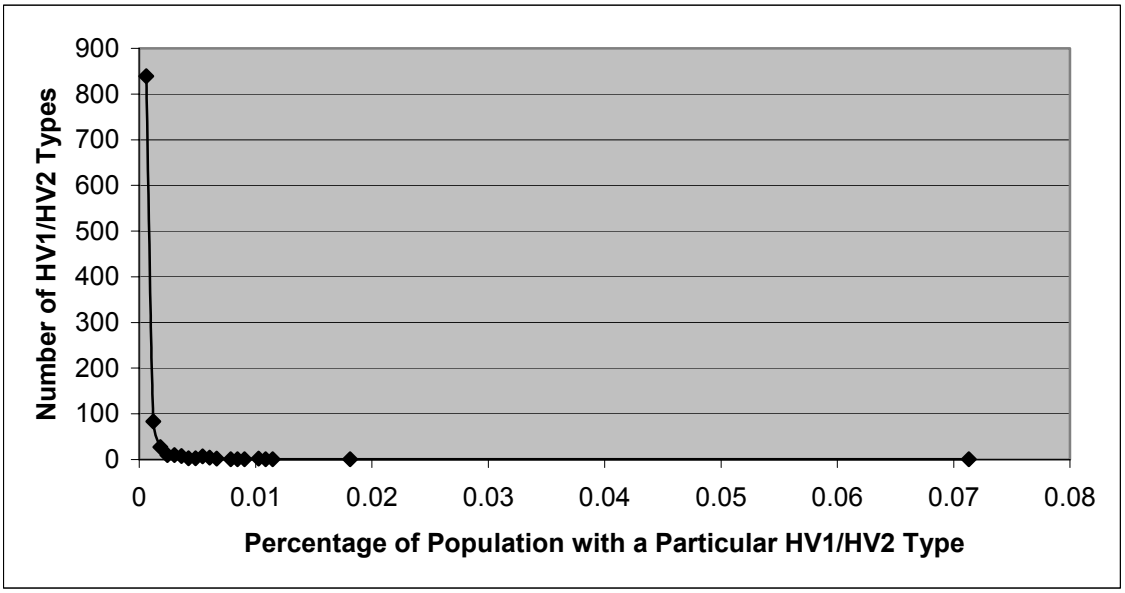


Figure 2. Graph of the Number of HV1/HV2 Types Versus the Percentage of the Caucasian Population Having a Particular HV1/HV2 Type. The data represents the pair wise comparison of 1665 European Caucasians. These percentages do not consider length variation in the HV2 C-stretch region (303-309).

is specific for Caucasians, the general features of the distribution are common in all populations studied to date: the presence of a large number of sequences that are unique in the database, and a small number of common types, the latter representing an appreciable proportion of the individuals in the population. The frequency distribution of HV1/HV2 types illustrates the greatest limitation of mtDNA testing: the small number of common types in the population for which the power of discrimination is low.

I. 2. 3. Challenges Associated with Forensic mtDNA Testing

For the forensic scientist, the challenges associated with generating a profile from highly degraded skeletal remains are demanding. Often times, the analyst must expend a large quantity of the limited DNA extract for generating an HV1 and HV2 sequence using overlapping "mini-primer sets" that target short amplicons (150 bp or less; Gabriel *et al.*, 2001). About half the time, the resulting HV1/HV2 profile will be observed as unique in the forensic population database. About twenty percent of the time, the resulting HV1/HV2 profile will belong to one of the common types (those that occur at a prevalence of 0.5% or greater in the frequency distribution – see Figure 2). Thus, about one-fifth of the time, the forensic scientist is confronted with results for which the power of discrimination is particularly low. For example, in a number of cases processed at AFDIL, the task of identifying skeletal remains occurs within a "closed population" of individuals (e.g. ten soldiers in a flight crew that perished in an aircraft accident during World War II). Most of the time, the HV1/HV2 mtDNA profiles generated from the

plane crash can be matched with maternal relatives of the missing. However, if multiple reference sequences for different service members share one of the common HV1/HV2 types among them, the forensic scientist would be unable to differentiate among them, and the identification of the remains could not be made.

I.3. The Central Effort of the Dissertation

I. 3. 1. Sequence Information in the mtGenome

The central effort of this dissertation was to mitigate one of the current limitations of forensic mtDNA typing: the inability to distinguish between unrelated individuals that share common HV1/HV2 types. We hypothesized that accessing sequence information present in the coding region of the mtGenome would be useful for further discrimination of common HV1/HV2 types. To this end, we proposed to sequence the entire mtGenome of non-related individuals sharing the more frequently occurring common types within the Caucasian population, to discover sites useful for discrimination.

The decision to use brute-force sequencing of the entire mtGenome of individuals sharing common HV1/HV2 types, in order to identify sites useful for forensic discrimination, was based on two hypotheses that have, in the final analysis, proven to be correct. First, we hypothesized that the greatest amount of total variation would be found in the coding region. The coding region has an evolutionary rate about four-fold lower than the non-coding control region (Aquadro and Greenberg, 1983; Horai and Hayasaka, 1990). However, the coding region is about 15-fold larger than the control region, making it a potentially rich source of additional sequence variation for the discrimination

of common HV1/HV2 types. Second, we hypothesized that sequence data in the published literature would not be greatly useful for identifying sites that would resolve common HV1/HV2 types. At the onset of our project, only a few complete human mtGenomes were published. A number of partially sequenced mtGenomes available at the time (Awadalla *et al.*, 1999) were shown to be filled with errors (Macaulay *et al.*, 1999b; Kivisild and Villems, 2000). Much of the variation known about the coding region came from the extensive number of studies using RFLP data, which only surveys about one-fifth of the mtGenome (Wallace, 1994). However, much of the RFLP data was of dubious value, as these sites tend to occur at slow evolving sites in the mtGenome, often defining global clades. Because the mtGenome is one linked molecule, the quickly evolving control region types would in effect map upon the more stable, slow-evolving RFLP variants without the latter adding much, if any, resolution for common HV1/HV2 types.

Recently, some researchers have recommended increasing forensic discrimination by sequencing one or two highly variable genes or short fragments in the regions of the mtGenome where multiple variable sites have previously been identified (Andreasson *et al.*, 2002; Lee *et al.*, 2002; Lutz-Bonengel *et al.*, 2003). We evaluated such an approach, but determined that there were several important flaws. First, given the linkage of mtDNA polymorphisms, even if a gene is observed to be highly variable, that variation is not guaranteed (or likely) to resolve individuals sharing common HV1/HV2 types in the quickly evolving control region. Next, by focusing only on a few fragments of the mtGenome, one would lose access to sites that may be both rare and widely scattered over the mtGenome. Lastly, sequencing portions of the coding region of forensic

casework or reference samples could have the unintended effect of revealing information of medical genetic significance.

I. 3. 2. Ethical Considerations Associated with mtGenome Sequencing

The risk of unintentionally determining information of medical significance is an important factor to consider as forensic scientists begin to venture into the coding region. Given the vital role of the mitochondria for the cellular energy production, many genetic diseases are associated with mutations in the mtDNA (reviewed by Wallace 1999). Phenotypic changes in the shape of proteins, ribosomes, or tRNAs encoded in the mtGenome can have the potential of altering cellular energy production. It is believed that such point mutations can lead to a "leaky" phenotype, where the altered protein or tRNA still partially functions (Scheffler, 1999). However, reduction in OXPHOS capacity can have significant medical consequences that may be differentially manifested over time (some point mutation diseases are late onset), or among tissues (Scheffler, 1999). Changes in the metabolic state may in turn induce apoptosis of the cell, as the mitochondrial permeability transition plays a central role in the apoptotic cell death pathway (reviewed in Liu *et al.*, 1997).

There are more than one hundred characterized mtDNA genetic diseases either associated or implicated with mtDNA mutations (Kogelnik *et al.*, 1998). The MitoMap web site contains an expanding list of disease-associated/implicated mtDNA mutations (http://www.mitomap.org/). Given the pivotal role of mtDNA function in basic physiology, the sequencing of genes or gene fragments to increase forensic

discrimination cannot confidently avoid encountering a variant that now, or may in the future, be associated with a disease or medically relevant condition.

For example, a single transition at nucleotide position 14459 in the NADH dehydrogenase 6 (ND6) gene (resulting in an amino acid change from alanine to valine) can cause sudden-onset blindness called Leber's Hereditary Optic Neuropathy (LHON), as a consequence of the death of the optic nerve (Wallace 1999). The mean age of onset is 27.6 years (ranging from 8 to 60 years) in patients having LHON (Wallace *et al.*, 1999). Additional mtDNA mutations associated with late-onset genetic diseases associated with point mutations that affect mtDNA genes include Parkinson's disease (Brown *et al.* 1996), Alzheimer's disease (Lin *et al.* 2002) and diabetes (Cavelier *et al.* 2002).

Therefore, unless one focuses on non-phenotypic changes in proteins, ribosomes, or tRNAs of the mtGenome, the forensic scientist could be faced with moral and ethical issues generally associated with clinical genetics and genetic counseling. For example, the group of forensic scientists in Europe focused on sequencing short fragments (50+ nucleotides) of the coding region is using the Pyrosequencing system (Pyrosequencing - Uppsala, Sweden) to increase forensic discrimination (Andreasson *et al.* 2002). One fragment in their assay occurs from the nucleotides 3216-3403. However, within this fragment resides the 3243A-G mutation, a site correlated to three diseases: diabetes mellitus (DM), diabetes mellitus with deafness (DMDF), and mitochondrial encephalomyopathy, lactic acidosis and stoke-like episodes (MELAS). If, having discovered this mutation as part of a family reference sequence to identify a set of skeletal remains, is the forensic scientist now obligated to tell the individual who gave the

reference sample and/or other maternal relatives? Conversely, is there a practical means to ensure that the donor or other parties would not discover this information and its medical significance?

I. 3. 3. Focus on Neutral SNPs for Discrimination

Based on these considerations, we propose to limit our search for forensically-informative single nucleotide polymorphism (SNP) sites to 1) non-coding sequences in the control region outside of HV1/HV2, 2) short "spacer" regions throughout the coding region that are non-coding (ranging in length from one base to thirty bases), or 3) silent (synonymous) mutations in the protein coding genes (with one specific exception, see below). In checking discriminatory sites for neutrality, it is necessary to bear in mind that not all third codon positions in the mtGenome are neutral. Mitochondria have their own genetic code (slightly different from the eukaryotic genetic code for the nuclear genome) and there is not necessarily a four-fold degenerate third codon position. It is possible to have a non-synonymous change in the amino acid sequence resulting from transversions in third codon positions. Conversely, it is also possible to have mutations at the first codon position and still code for the same amino acid (e.g. there are two tRNAs that code for the amino acids leucine and serine).

Although we have focused on sites that are either non-coding, or at third codon position changes in the coding region, it is surely the case that most of the non-synonymous variation in the mtGenome is neutral as well (e.g. Kimura, 1983). For example, two of the three SNPs that are diagnostic polymorphisms for the Eurasian

superhaplogroup N (8701 and 10398) represent first position codon changes that result in a change from threonine to alanine. Given the millions of individuals in the world that carry these mutations, we can be confident that typing these SNPs would not reveal any unanticipated information of medical significance. If a non-synonymous codon SNP, or a tRNA or rRNA SNP is found to be of exceptional forensic utility, we might then evaluate, on a case-by-case basis, grounds for making an exception to our highly conservative rules for SNP selection. Evaluation of such sites could be based on their characterization in the literature and how widely these sites vary in the population.

I. 3. 4. Practical Applications of Informative SNPs

We envision that molecular assays specific for selected SNP sites are ideal for accessing information outside the CR in order to increase forensic resolution. This avoids the medical genetic problems of general sequencing, by assaying only "neutral" variants of interest. Further, many SNP assays utilize very small amplicons, making them ideal for ancient DNA casework. A number of SNP assays, such as the SNaPShot™ (Applied Biosystems, Foster City, CA) kit, can be readily multiplexed. Multiplexing informative SNP sites has several advantages for analyzing forensic mtDNA casework. In difficult cases involving degraded DNA, there is often a limited amount of extract remaining for further testing. By accessing multiple sites in a single amplification, casework extract can be preserved. In addition to saving extract, multiplex assays can often be conducted and analyzed much faster than sequencing. One practical limitation of multiplex assays, however, is that they can often be difficult to optimize.

Additionally, we have found in our experience with the SNaPShot™ assay, that we are limited as to the number of sites that can be placed in the multiplex. Each SNP extension primer requires the addition of a 5' four nucleotide base poly-thymidine tail to aid in electrophoretic separation (Vallone *et al.*, 2003). The maximum number of sites that can be multiplexed in the SNaPShot™ assay is about ten.

I. 3. 5. The Strategy of Identifying Discriminatory SNPs

Our approach to identifying SNPs useful for resolving common HV1/HV2 types was based on sequencing the entire mtGenome of multiple Caucasian individuals sharing common HV1/HV2 types. Pairwise comparisons of the forensic database reveal 22 common HV1/HV2 types that occur at a frequency of 0.5% or greater among the Caucasian population (Figure 2). A significant number of the sequences in the forensic database were generated at AFDIL, so we had access to appropriate samples for mtGenome sequencing. However, due to limitations of sample availability, we focused on a subset of 18 of the 22 common Caucasian types (Table1). We have designated these 18 common HV1/HV2 types according to their putative haplogroup association based on HV1 sequence data (Table 1). This was done as a convenient method for naming the 18 HV1/HV2 types, and is not to be confused with the nomenclature in the current literature, since haplogroups cannot unambiguously be determined by sequence data in HV1/HV2 (discussed above). The 18 common HV1/HV2 types were classified into five Caucasian haplogroups: H, J, K, T and V (Table 1). The most common

N	HV1/HV2 Type	Frequency	HV1/HV2 Sequence (+ 263A-G, 315.1C)
31	H1	7.1%	CRS
25	H2	1.8%	152 C
11	H3	0.8%	16129 A
8	H4	0.5%	16263 C
12	H5	0.9%	16304 C
11	H6	0.9%	73 G
7	H7	0.7%	16162 G 16209 C 73 G
25	V1	1.0%	**16298 C**
15	J1	1.1%	**16069 T 16126 C** 73 G 185 A 228 A 295 T
8	J2	0.6%	**16069 T 16126 C** 73 G 228 A 295 T
13	J3	0.7%	**16069 T 16126 C** 73 G 185 A 188G 228 A 295 T
8	J4	0.5%	**16069 T 16126 C** 16145 A 16172 C 16222 T 16261 T 73 G 242 T 295 T
21	T1	1.1%	**16126 C 16294 T** 16296 T 16304 C 73 G
10	T2	0.6%	**16126 C** 16163 G 16186 T 16189 C **16294 T** 73 G 152 C 195 C
8	T3	0.5%	**16126 C 16294 T** 16296 T 73 G
14	K1	1.0%	**16224 C 16311 C** 73 G 146 C 152 C
7	K2	0.5%	16093 C **16224 C 16311 C** 73 G
7	K3	0.5%	**16224 C 16311 C** 73 G
241		20.8%	

Table 1 The Common HV1/HV2 Types and Their Frequencies Found in the Caucasian Population. The sample size column (N) refers to the number of individuals sequenced over the entire mtDNA genome for this study. The sequences listed are polymorphisms compared to the Cambridge Reference Sequence (Anderson *et al.* 1981), with the exception of the mutations 263A-G and 315.1 C, which are shared by all. Length variants in the 303-309 C-stretch region have been ignored. Haplogroup-associated polymorphisms in HV1 are indicated in bold. The frequency of the common types were determined from HV1/HV2 data for 1655 Caucasians (Monson *et al.*, 2002).

HV1/HV2 type, designated as H1, is found in about 7% of the population (Figure 2 - see also Lutz-Bonengel *et al.*, 2003).

I. 3. 6. Criteria for SNP Selection

We devised the following specific criteria for the selection of SNP assay sites: 1) the SNP sites must be neutral with respect to amino acid or phenotypic changes, 2) the sites should vary within (or among) multiple individuals, 3) the sites should not be redundant to other selected sites, and should be selected in combinations that maximize discrimination. Once SNP sites were discovered to resolve specific common HV1/HV2 types, our goal was to make these results practical for the forensic laboratory. We chose to arrange the SNP sites useful for resolving particular common type(s) into multiplex panels. Thus, when the forensic analyst encounters one of the common Caucasian HV1/HV2 types in a casework sample, she can choose the appropriate SNP panel to assay and provide maximum discrimination. It should be noted that common HV1/HV2 types are not unique to Caucasians. Other forensically important populations studied to date (African-American, Hispanics, etc…) show the same type of frequency distribution observed in Caucasians (Figure 2). It is uncertain if the SNP sites useful for one resolving common types in one haplogroup will also resolve common types in a distantly related haplogroup.

At the outset of this project, we were unsure how likely a SNP site that resolves one common HV1/HV2 type would resolve: a) other common HV1/HV2 types from the same haplogroup, b) other common HV1/HV2 types from different, but related

haplogroups, or c) other common HV1/HV2 types from the distantly related haplogroups (e.g. Caucasian versus African). There was also substantial uncertainty as to the monophyly of sequences that match the common HV1/HV2 types. For example, five of the seven common HV1/HV2 types we tentatively identified as belonging to haplogroup H (types H2-H6 – see Table 1) differ from the most common HV1/HV2 type (H1) by one mutation. Some of these mutations (e.g. 73, 152, 16129, 16304) have been classified as relatively fast sites in the literature (Meyer *et al.*, 1999; Allard *et al.*, 2002; Malyarchuck *et al.*, 2002). Additionally, we were uncertain if the common HV1/HV2 type we classified as H6 (73G) would turn out to be a mixture of individuals belonging to haplogroup H or haplogroup U (see above). It was unclear if the common HV1/HV2 types were widely paraphyletic (or polyphyletic) due to reversions and homoplasy within the control region, and what effect this would have on developing SNP panels.

Despite the uncertainties, we did expect some sites to retain utility among closely related types. Much of the utility of a SNP site depends on the nature of the site itself. Some polymorphisms have arisen once during human evolution and are associated with all the individuals within a clade, or haplogroup. Within the haplogroup, there are additional polymorphisms in the coding region that have arisen within two or more subclusters of sequences (Herrnstadt et al., 2002). We might then expect a predominance of slow, rare sites that discriminate within the common HV1/HV2 types. These sites may be considered "needles in the haystack" that are specific to particular common type(s). We also expected that we might observe a number of sites in the coding region that may be universal "hot spots". Such SNPs would be useful not only for resolving closely related common types, but for resolving common types of any group or race. It was also

possible that the sites discovered would prove to be a continuous combination of the above: sites that are specific to a common type and sites that are universal fast sites, depending on the extent of rate heterogeneity in the coding region (discussed below).

I. 4. Mutation Rate Variation in the Coding Region

I. 4. 1. Characterization of the Relative Mutation Rates in HV1/HV2

One salient characteristic of human mtDNA evolution is extreme substitution rate heterogeneity from site to site (Wakeley, 1993). Several studies have sought to characterize the relative mutation rates of sites in the two HV regions (Wakeley, 1993; Hasegawa *et al.*, 1993; Excoffier and Yang, 1999; Meyer *et al.*, 1999; Pessole and Saccone, 2001; Malyarchuck *et al.*, 2002). The mutation rate spectrum in HV1 and HV2 is best described by a gamma distribution, with an "L-shaped" curve (Figure 3). In such a distribution, a few sites exhibit very high mutation rates (so called, “hotspots”) while most sites have a range of low substitution rates (or are practically invariant). The shape parameter (α) of the gamma distribution is used as a measure of the amount of rate variation (reviewed in Yang 1996). The α parameter is inversely related to the extent of rate variation: low values of α (≤ 1) produce an “L-shaped” distribution while increased values of α (> 1) are indicative of intermediate rates (i.e. few sites are either invariant or fast). As α approaches infinity, all sites have the same rate (Figure 3).

Differential substitution rates have been shown to create artifacts in phylogenetic reconstruction (Tateno *et al.* 1994; Yang 1995) if not properly accounted for. Mutation

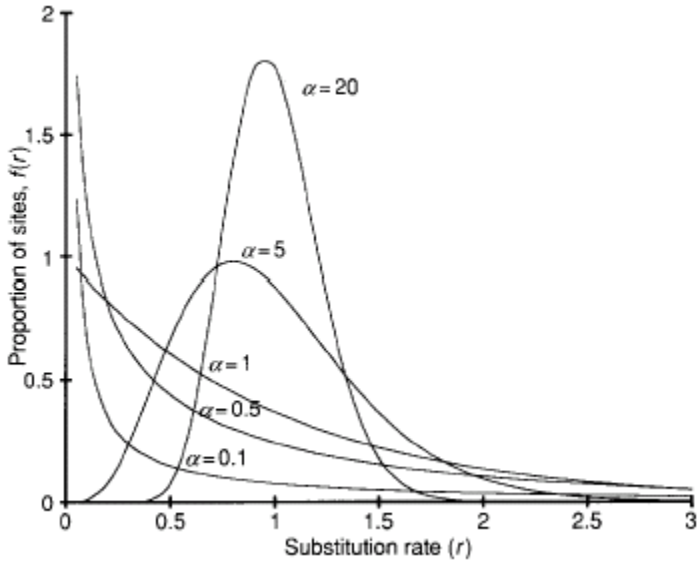


Figure 3. The Gamma Distribution of Substitution Rates. The density function (proportion of sites) is graphed against the substitution rates at sites. The gamma distribution has a shape parameter, α, which is inversely related to the degree of rate variation. Low values of α (≤ 1) the gamma curve is "L-shaped", meaning that most sites are invariable, while a few sites have a fast mutation rate. As α becomes larger (>1), the curve becomes "bell-shaped", meaning that nearly all sites have an intermediate rate with a few sites invariant and a few sites fast. Figure adapted from Yang, 1996.

rate heterogeneity can mimic population expansion (Lundstrom *et al.* 1992; Aris-Brosou and Excoffier 1996) and bias estimations of population genetic parameters such as estimation of effective population size (Fu, 1994) and calibration of the molecular clock (Adachi and Hasegawa, 1995; Yang, 1996). Wakeley (1994) demonstrated that ignoring the mutation rate variation could greatly underestimate the transition:transversion ratio.

I. 4. 2. Methods to Determine Mutation Rate Variation

Several methods have been used to determine the mutation rate variation within the mtDNA control region. These techniques include accessing the relative mutation rates from phylogenetic trees evaluated using parsimony (Wakeley 1993; Hasegawa *et al.* 1993), maximum likelihood (Excoffier and Yang 1999; Meyer *et al.* 1999), and pair wise distance methods (Pesole and Saccone 2001). The determination of relative mutation rates using phylogenetic trees inferred by using parsimony has been shown to underestimate the true mutation rate by overestimating α (Yang, 1996). Maximum likelihood (ML) methods utilize an evolutionary model to estimate DNA substitutions and have been shown to perform better than parsimony-based methods to determine unbiased measures of α (Yang 1994). One disadvantage of the ML method is the intense computational requirements for analysis, limiting the number of sequences that can be evaluated at one time. A pair wise distance method has also been used to determine relative mutation rates (e.g. in the small-ribosomal-subunit RNA; see Van de Peer *et al.* 1993). This method is based on the idea that in comparing two sequences, the probability of observing a difference at a particular site depends on the site-specific mutation rate and

the evolutionary distance between the two sequences. Thus, when comparing two very closely related sequences, there is a much higher probability that any observed nucleotide differences between the two sequences are a result of changes at fast evolving sites (Van de Peer *et al.*, 1993; Pesole and Saccone, 2001; Meyer and von Haeseler, 2003). Since this method simply compares sequences and calculates the "disparity index" (Van de Peer *et al.*, 1993), a much larger number of sequences can be analyzed without having to rely on phylogenetic tree building methods (Van de Peer and De Wachter 1993; Pesole and Saccone 2001).

I. 4. 3. Assumptions of the Rate Variation in the Coding Region

Characterization of the site-by-site spectrum of relative mutation rates in the coding region is now feasible with the recent, growing number of published mtGenome sequences. Until very recently, the assumption that the slower evolving coding region has a relatively homogeneous mutation rate appears to have been the rule. For example, Eyre-Walker *et al.* (1999) observed 22 homoplasic sites in third position codons from a partial dataset of mtGenome sequences. By assuming, "there is no evidence of variation in the mutation rate", this group concluded that the appearance of excessive homoplasy in the mtDNA coding region was due to recombination. Criticisms of the data used in that study (Macaulay *et al.* 1999b; Kivisild and Villems 2000), the methods used (Jorde and Bamshad 2000; Kumar *et al.* 2000), and recent analyses of entire mtGenomes (Ingman *et al.* 2000; Elson *et al.* 2001; and Herrnstadt *et al.* 2002) have abrogated the recent controversy of recombination in mtGenome (see also, Innan and Nordborg 2002).

There appears to be inconsistency in the literature as to the relative amount of homoplasy in the coding region. For example, Herrnstadt *et al.* (2002) notes the "relatively large number of sites" showing the "extensive" amount of homoplasy in their data set of 560 coding region mtGenomes. On the other hand, another group (e.g. Yao *et al.* 2003) examines the same data and notes that, "Homoplasy in the coding region is much less than in the control region and may have only a few hot spots (see, e.g., table 2 of Herrnstadt *et al.* [2002])."

Recently, Meyer and von Haeseler (2003), using the pair wise distance comparison method between pairs of sequences (based upon Van de Peer *et al.* 1993), identified site-specific mutation rates and estimated the α parameter for the mtGenome. In the pair wise method, Van de Peer *et al.* (1993) first estimated genetic distances between pairs of sequences to identify sites that have undergone multiple mutations. Then, Van de Peer *et al.* (1993) used probabilistic analysis to determine the site-specific rates at each position. Meyer and von Haeseler (2003) determined that the statistical basis used by Van de Peer *et al.* (1993) was not well established. Meyer and von Haeseler (2003) used the pair wise distance method as in Van de Peer *et al.* (1993) followed by a maximum likelihood model to estimate the site-specific rates in the mtGenome.

I. 4. 4. Coding Region Rate Variation and SNPs that Increase Forensic Discrimination of Common HV1/HV2 Types

In this dissertation, we have characterized the relative mutation rates of all nucleotide positions in the coding region of the mtGenome, using phylogenetic trees generated from both parsimony and neighbor joining methods. We generated phylogenies using these two methods and mapped character (site) changes upon the inferred topology in order to determine relative mutation rates in the human mtDNA coding region. Previous simulation studies to estimate the mutation rate variation using parsimony found that the number of DNA sequences under study was the most powerful factor in determining parameter estimations (Deng and Fu 2000; see also previous research in Yang and Wang 1995).

We have constructed a phylogenetic tree using the coding region of 646 human mtDNA coding regions (Ingman *et al.* 2000; Maca-Meyer *et al.* 2001; and Herrnstadt *et al.* 2002), representing the largest such analysis of mutation rate variation to our knowledge. The 241 mtGenomes selected for sequencing (Table 1) to resolve individuals matching in HV1/HV2 were not included in the analysis for mutation rate variation. Mutation rates gleaned from the analysis of these data would reflect the attribution bias within this data set. For example, the mutation rates in the control region would be zero for nearly all sites except for the few sites that were used to define the common HV1/HV2 types. A similar distortion of mutation rates would also be observed for the coding region of the 241 sequences from common types. Therefore, the published sequences represent a better source of data to determine the "global" rate spectrum.

It is our desire to understand the coding region mutation rates, both from the standpoint of basic characterization of molecular evolution and in relation to the SNP sites we have discovered that discriminate among common HV1/HV2 types. We are interested in knowing the proportion of SNP sites identified as useful for resolving common HV1/HV2 types in the Caucasian population that are characterized as fast sites. If nearly all of the sites useful for resolving common types in the Caucasian population are comprised of universal fast sites, then the strategy of sequencing the entire mtGenomes of common types in African-Americans or Hispanics might be unnecessary. However, if fast sites comprise only a small proportion of the discriminatory sites in our SNP panels, then those rare, slow sites within common types will need to be discovered by mtGenome sequencing, as we extend this effort to additional populations.

Chapter II. Identification of Polymorphic Sites Useful for the Forensic Discrimination of Common HV1/HV2 types.

This section will address Specific Aim #1: The identification of coding region sites to increase the power of discrimination for common Caucasian HV1/HV2 types. We will describe the results of sequencing 241 individuals among 18 common HV1/HV2 types in the Caucasian population. The objective in the evaluation of these results will be to develop panels of SNP assay sites useful in forensic casework.

II. 1. Materials and Methods

II. 1. 1. Sample Selection for Whole Genome Sequencing

In order to determine the population frequencies of common HV1/HV2 types, a pair wise comparison of 1665 Caucasian HV1/HV2 sequences was performed using the mtDNA database from Monson *et al.*, 2002 (Figure 2). The regions of HV1 and HV2 used for this study spans from: HV1=16024-16365 and HV2=73-340. A population frequency of 0.5% was chosen as an operational lower cut-off since an appreciable proportion of the common HV1/HV2 types are represented at or above this value. We identified 22 common HV1/HV2 types that occur at 0.5% or greater. However, due to limitations of sample availability, we focused on a subset of 18 of the 22 common Caucasian types (Table 1.1). The C-stretch length heteroplasmy variants in the HV1 and HV2 regions were ignored for sample selection since these variants are not used for

the purpose of exclusion in forensic cases (Holland and Parsons 1999; Stewart *et al.* 2001).

Many of the Caucasian database samples identified as belonging to the 18 common HV1/HV2 types (Table 1) were originally sequenced and subsequently stored at AFDIL. Therefore, we were able to retrieve 241 bloodstain cards from individuals belonging to the 18 common HV1/HV2 types that were sequenced in this study. Assurances were made to guarantee the anonymity of the contributor's name with respect to the sequence generated, and the Armed Forces Institute of Pathology Internal Review Board approved the project for the Use of Human Subjects. Five 3mm punches from the bloodstain cards were placed in sterile 1.7mL eppendorf tubes. To protect the genetic privacy of the individuals sequenced in this study, no information relating to the sample number or name was recorded on the eppendorf tube. The eppendorf tubes were then randomized and numbered according to the common HV1/HV2 type. The samples were named according to the following example: sample J3-07 refers to the seventh individual belonging to the third common HV1/HV2 type we identified from haplogroup J. The nomenclature used here is not to be confused with the mtDNA sub-haplogroups in the current literature (e.g. H1, T1, J2, etc… see Herrnstadt *et al.* 2002 for a review).

II. 1. 2. DNA Extraction

Total genomic DNA was extracted from bloodstains dried on paper cards (Fitzco, Minneapolis, MN). For the 31 samples belonging to the HV1/HV2 type H1, a single 3mm punch from the card was extracted with a 5% (w/v) final concentration of Chelex

100 beads (BioRad Laboratories, Hercules, CA). The extraction procedure followed the published protocol of Walsh *et al.* 1991. The remaining 209 samples were extracted via a high throughput procedure developed at AFDIL. Bloodstain paper cards were punched using a Wallac DBS Puncher (Perkin-Elmer Life Sciences, La Jolla, CA) to obtain a 3mm punch in a 96 well tray for extraction. The punches were then extracted using the Qiagen BioRobot^{TM} 9604 (Qiagen, Gaithersburg, MD) instrument using the QIAmp 96-well DNA Swab extraction kit. The protocol for extraction followed the manufacture's suggestion for liquid blood, with an inclusion of a 45-minute lysis incubation rather than the suggested 15-minute incubation.

II. 1. 3. PCR Amplification and Sequencing Overview

The strategies for amplification and sequencing the mtGenome vary widely among the different laboratories that have published entire sequence data. Levin *et al.* (1999) amplified the mtGenome with 58 amplicons (average size of 551 base pairs) and used these same amplification primers for sequencing (116 sequencing reactions). The shorter fragment sizes for sequencing using the Levin *et al.* (1999) strategy provides sequence information from both strands of DNA. However, this strategy was excessively redundant (nearly 32,000 base pairs are amplified for one mtGenome) and inefficient for our purposes. Ingman *et al.* (2000), using the strategy of Rieder *et al.* (1998), amplified the mtGenome in 24 amplicons (average size of 893 base pairs), and sequenced the fragments with the same amplification primers (48 sequencing reactions). Torroni *et al.* (2001) sequenced the mtGenomes of Africans by amplifying 11 amplicons (average size

of 1810 base pairs) and sequencing each fragment with 3 forward primers (36 sequencing reactions).

We developed a high throughput method (Levin *et al.*, 2003) to amplify and sequence the mtGenome using selected primers from Levin *et al.* (1999). We amplified the mtGenome into 12 overlapping fragments (average size of 1556 base pairs). For sequencing of each amplicon, we used a set of overlapping forward and reverse sequencing primers to sequence both strands. Each sequence primer was chosen to produce, on average, ~450 base pairs of information. In a few of the amplicons, there was a lack of primer pairs to provide coverage from both strands. In these sections, two primers were used to sequence the same stand. This prevented us from having to rely on one strand of sequence information in significant portions of the mtGenome, a potential source of "phantom mutations" (Bandelt *et al.*, 2002).

II. 1. 4. PCR Amplification of the mtDNA Genome

The entire mtDNA genome was amplified in 12 overlapping fragments that produced PCR products ranging from 825 to 1856 base pairs (bp). The primers used to amplify amplicons 01-11 (Table 2) were based upon primer information from Levin *et al.* (1999). It should be noted that Levin *et al.* (1999) was published before the reanalysis of the original Cambridge Reference Sequence (Andrews *et al.* 1999). One of the errors in the original Anderson *et al.* (1981) sequence was a Guanine at position 14368. This affected the penultimate 3' base of our amp09 reverse primer. The corrected base at

PCR Amp Number	Fragment Length (base pairs)	Name	Sequence (5' -->3')
1	1856	F361	ACAAAGAACCCTAACACCAGC
		R2216	TGTTGAGCTTGAACGCTTTC
2	1565	F1993	AAACCTACCGAGCCTGGTG
		R3557	AGAAGAGCGATGGTGAGAGC
3	1543	F3441	ACTACAACCCTTCGCTGACG
		R4983	GGTTTAATCCACCTCAACTGCC
4	1730	F4797	CCCTTTCACTTCTGAGTCCCAG
		R6526	ATAGTGATGCCAGCAGCTAGG
5	1790	F6426	GCCATAACCCAATACCAAACG
		R8311	GACGATGGGCATGAAACTG
6	1685	F8164	CGGTCAATGCTCTGAAATCTGTG
		R9848	GAAAGTTGAGCCAATAATGACG
7	1856	F9754	AGTCTCCCTTCACCATTTCCG
		R11600	CTGTTTGTCGTAGGCAGATGG
8	1721	F11403	GACTCCCTAAAGCCCATGTCG
		R13123	AGCGGATGAGTAAGAAGATTCC
9	1596	F12793	TTGCTCATCAGTTGATGATACG
		R14388	TTAGCGATGGAGGTAGGATT**G**G
10	1208	F14189	ACAAACAATGGTCAACCAGTAAC
		R15396	TTATCGGAATGGGAGGTGATTC
11	825	F15260	AGTCCCACCCTCACACGATTC
		R16084	CGGTTGTTGATGGGTGAGTC
12 (CR)	1340	F15878	TTAACTCCACCATTAGCACC
		R649	TTTGTTTATGGGGTGATGTGA

Table 2. Oligonucleotide Sequences for PCR Primers Used to Amplify the Entire mtDNA Genome into 12 Overlapping Fragments. The amplicon number, fragment size in base pairs, primer name and sequence are noted. The Guanine base underlined and in bold for the reverse amplicon 09 primer was corrected from an error (Cytosine) in the original rCRS sequence (Anderson *et al.*, 1981).

position 14368 (C) is noted in Table 2. The control region amplicon (amp12) was amplified using primer sequences developed at AFDIL. A summary of the amplicons and primer sequences can be found in Table 2.

The PCR mixture for each amplicon contained extracted DNA (1 μL – 3 ng/μL), AmpliTaqGold® DNA polymerase (1 μL – 5U/μL) (Applied Biosystems, Foster City, CA), 10X PCR buffer (5 μL) containing 100 mmol/L Tris-HCl, 50 mmol/L KCl, 15 mmol/L $MgCl_2$ and 0.01% (w/v) gelatin, pH 8.3 (Applied Biosystems, Foster City, CA), dNTP's (0.2 mmol/L) (Invitrogen, Carlsbad, CA), 2 μL of forward and reverse amplicon primers (10 μmol/L) (MWG Biotech, High Point, NC) plus sterile dH_2O (Invitrogen, Carlsbad, CA) to a final volume of 50 μL. Thermal cycling was conducted in either a PerkinElmer Applied Biosystems 9700 thermocycler (Applied Biosystems, Foster City, CA) or an MWG Primus thermocycler integrated on a robotic liquid handling platform (MWG Biotech, High Point, NC) with the following conditions: 10 min at 96 ^{o}C (activation of AmpliTaq Gold®), plus 40 cycles of 94 ^{o}C for 15 sec, 56 ^{o}C for 30 sec, and 72 ^{o}C for 1 min.

An aliquot of 5μl of PCR product was fractionated by gel electrophoresis in a 0.7% agarose gel containing 0.3 μg/mL of ethidium bromide to assess the purity and size of the DNA fragments. The PCR products were purified with Shrimp Alkaline Phosphatase (SAP)/Exonuclease I (ExoI) (Amersham Pharmacia, Piscataway, NJ). 5 μL of exonuclease I (10 U/μL) and 10 μL of Shrimp Alkaline Phosphatase (1 U/μL) were added to each tube containing PCR product. The ExoI enzyme digests unused amplification primers by degrading single-stranded DNA. The SAP enzyme inactivates unused dNTP's by removing the 5' phosphate group. The tubes were incubated at 37 ^{o}C

for 15 min followed by 94 ^{o}C for 15 min in a PerkinElmer Applied Biosystems 9700 thermocycler (Appled Biosystems, Foster City, CA).

II. 1. 5. Sequencing of the mtDNA Genome

Cycle sequencing was performed with the ABI PRISM® BigDye® Terminator (Version 1.0) cycle sequencing kit (Applied Biosystems, Foster City, CA). The sequencing mixture consisted of: 9 μl dH_2O, 6 μl of BigDye® dilution buffer (400mM TRIS, 10mM $MgCl_2$, pH 9.0), 1.6 μl BigDye® Terminator (Applied Biosystems, Foster City, CA), 0.4 μl ABI PRISM® dGTP BigDye® Terminator (Applied Biosystems, Foster City, CA), 1μl of forward or reverse primer (10 μM each), and 2 μl of PCR product for a total volume of 20 μl. Thermal cycling was conducted in either a Perkin-Elmer Applied Biosystems 9700 thermocycler (Applied Biosystems, Foster City, CA) or an MWG Primus thermocycler (MWG Biotech, High Point, NC) with the following conditions: an initial 1 minute denaturation at 96^{o}C; followed by 25 cycles of 15 seconds at 94^{o}C (denaturation), 5 seconds at 50^{o}C (annealing), and 2 minutes at 60^{o}C (extension). The DNA product was purified by filtration though a 96 well spin plate matrix (Edge BioSystems, Gaithersburg, MD).

Electrophoresis and sequencing were performed with either an ABI 377 (for the 31 H1 individuals) or 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA). For the ABI 377, products were analyzed on 5% acrlyamide, 6M Urea gels (Long Ranger® Singel®, Cambrex Corporation, East Rutherford, NJ). Samples analyzed on the ABI 3100 used POP-6^{TM} polymer (Applied Biosystems, Foster City, CA) on a 36 cm

capillary. Sequences were aligned with the revised Cambridge Reference Sequence (rCRS) (Andrews *et al.* 1999) and edited using Sequencher Plus 4.0.5b11 (GeneCodes, Ann Arbor, MI). A list of differences compared to the rCRS were recorded for each individual.

In most cases, sequence information was acquired for both the forward and reverse directions. However, in some regions the reverse primer (typically) was ineffective at providing acceptable sequence data, perhaps due to secondary structure. Therefore, two separate reactions generating clear sequence information using the same primer was required to provide complete coverage. A total of 95 sequencing reactions plus 1 pGEM reaction (as a positive control) were conducted in a 96 well format.

The following sequencing primers from Levin *et al.*(1999) (for amps 01-11) and from AFDIL (amp 12) were used to sequence the 12 PCR amplicons:

Amp01 (F361/R2216): F361, R921, F1234, R1425, F873, R2216, F1657, R1769

Amp02 (F1993/R3557): F1993, R2660, R2834, R3557, F2417, R3006, F3234

Amp03 (F3441/R4983): F3441, R3940, F3931 (2X), R4982, F4392, R4162

Amp04 (F4797/R6526): F4797 (2X), R6526, F5700 (2X), F5318, R5882, F6242

Amp05 (F6426/R8311): F6426 (2X), R7255, F7645 (2X), R8311, F7075, R7792

Amp06 (F8164/R9848): F8164 (2X), R9059, F8903, R9848, F8539, R9403, F9309

Amp07 (F9754/R11600): F9754 (2X), R10556, F11001 (2X), R11600, F10386, R11267

Amp08 (F11403/R13123): F11403 (2X), F12357, R13123, F11901 (2X), F12601, R12876

Amp09 (F12793/R14388): F12793, R13611, F13518 (2X), R14388, F13188, R13343, F13899, R13935

Amp10 (F14189/R15396): F14189 (2X), R15396, F14909, R14996, F14470

Amp11 (F15260/R16084): F15260, R16084, F15574, R15774

Amp12 (F15878/R649): F15971, R16175 (2X), F16450 (2X), R274, F314 (2X), R649 (2X), F16190, R16400

The (2X) above refers to primers that were standardly used twice in one amplicon in order to provide complete sequence coverage. The primer sequence information can be found in the Appendix 1.

II. 1. 6. Robotic Sequencing and Data Quality Control

High-throughput sequencing of the entire mtDNA genome was performed on an MWG RoboAmp® 4200 robotic platform. The features of this particular robot include: a single tip liquid handler, a roaming arm capable of moving plates, an integrated MWG Primus thermal cycler with high pressure lid, refrigerated racks for keeping reagents cool, and a lid handler arm capable of removing the plastic lid covers of the 96 well plastic ware. Seven different individuals plus one negative control were amplified for each 96-well plate (rows 1-7 for individual mtDNAs; row 8 for negative controls). The robot was programmed to pipette the amplification master-mix of all reagents to each column (49 μl). One microliter of the template DNA was then added to each row (using disposable

tips to prevent cross-contamination). The plate was then moved to the Primus thermal cycler for PCR amplification. In the meantime, additional plates were prepared for amplification in the ABI 9700 thermal cycler.

SAP/ExoI post-PCR cleanup was also performed on the MWG RoboAmp® 4200. 15 μl of the SAP/ExoI mixture was added to each tube, and the incubation was performed in either the MWG Primus thermal cycler or the ABI 9700 thermal cycler. The sequencing reactions were prepared on the MWG RoboAmp® 4200 in the following manner: 1) 17 μl of dH_2O, BigDye® dilution buffer, BigDye® terminator kit, and dGTP BigDye® terminator kit – in a single mastermix – was added to each well of the 96-well plate (without changing the pipette tip); 2) 2 μl of PCR from each amplicon was added to the specific target wells (without changing tips); 3) 1 μl of sequencing primer was added to the target well (tip changed each time). The robotic arm then moved the sequencing plates to the Primus thermal cycler. Additional sequencing plates were prepared by the liquid handling portion of the robot (during the sequencing step in the Primus thermal cycler) and were placed by hand on the ABI 9700 thermal cycler for sequencing.

The post-sequencing cleanup was performed by transfer of the sequencing reaction from the sequencing plate to the Edge BioSystem (Gaithersburg, MD) 96-well block via a multi-channel pipette. Samples were dried in a HetoVac heated vacuum drier (Appropriate Technical Resources, Laurel, MD), and re-suspended in 10 μl of HiDi Formamide (Applied Biosystems, Foster City, CA) for loading onto the ABI 3100.

A careful attention to detail was required in order to avoid errors in sequencing and the transcription of mutations onto paper (e.g. "phantom mutations" see Bandelt *et al.* 2002). We have taken several steps to avoid some of the common mistakes that lead to

sequencing errors (Yao *et al.* 2003). We decided from the beginning to sequence both strands of mtDNA, or to sequence one strand at least twice. In one quality control experiment, a scan of 139 "windows" of data within the Sequencher Plus program (120 bp of data per window) from 20 random genomes was performed. We found that on average, about 37% of the mtGenome contig was determined with two strands of sequence information. Therefore, about 63% of the time three or more strands of sequence data were used for calling bases.

After sequences have been determined, another major potential source of error is the process of data transcription and database compilation (Yao *et al.*, 2003). In our approach, each mtGenome sequence was inspected and reviewed by two individuals who independently compiled a list of the polymorphisms compared to the rCRS (Andrews *et al.*, 1999). The two reviews were compared and, in the rare instance of a discrepancy, were resolved. The consensus list of differences to the rCRS were then entered into a master database program using a graphical user interface (developed by Jodi Irwin, AFDIL). After entering the list of mutations into the master database, another independent reviewer checked the data before it was "locked" and no longer subjected to changes. Each site was then evaluated for its position within genes and proteins (codon position), and whether the site was synonymous or non-synonymous as determined by the web program MitoAnalyzer (Lee and Levin 2002). The entire sequence information for all mtDNA genomes sequenced in this study can be found in Appendix 2. The characterization of each polymorphic site (gene region, synonymous/non-synonymous changes, amino acid changes, disease association, etc…) can be found in Appendix 3.

II. 1. 7. Phylogenetic Presentation of Data

Phylogenetic trees, built using maximum parsimony, were used to present and visualize the data. Parsimony trees for all haplogroup analyses were generated using the software program Winclada version 1.00.08 (Nixon 2002; http://www.cladistics.com/) running the program NONA version 2 (Goloboff 1999; http://www.cladistics.com/aboutNona.htm). Text files of each mtGenome were compiled into Sequencher (Gene Codes, Ann Arbor, MI) and aligned with the African haplogroup L3 sequence, "Mandenka (Genbank ID – AF346995)" from Ingman et al (2000). This African sequence served as an outgroup for phylogenetic analysis in all Caucasian groups. Unweighted parsimony analyses were performed using the heuristic search option using 10 replicates with 2 starting trees per replication. The multiple Tree Bisection Reconnection (TBR+TBR) search strategy was used with all insertions removed from the data matrix.

II. 2. Results

II. 2. 1. Superhaplogroup H/V Analysis

The Caucasian haplogroups H and V are sister haplogroups found at a frequency of ~45% and ~5%, respectively, within European and United States Caucasian populations (Torroni *et al.* 1996; Macaulay et al 1999a; Allard *et al.* 2002). The H/V cluster is grouped by the coding region diagnostic polymorphisms 11719G and 14766C (Macaulay *et al.* 1999a; Saillard *et al.* 2000). Haplogroup H is grouped by the coding region mutations at positions 2706 and 7028. Haplogroup H individuals are usually characterized by the presence of the control region polymorphism 73A and the lack of defining polymorphisms associated with other haplogroups (Torroni *et al.*, 1996). However, a number of reversals at position 73 have been observed in the literature (Torroni *et al.*, 1996; Macaulay *et al.*, 1999a; Allard *et al.*, 2002). In addition to the superhaplogroup HV-specific mutations, haplogroup V is characterized by having coding region-specific mutations 4580G-A and 15904C-T, and control region-specific mutations 72T-C and 16298T-C.

II. 2. 2. Phylogenetic Analysis of the H/V Cluster

A total of 105 unrelated individuals from the seven common haplogroup H HV1/HV2 types (H1-H7) and 25 individuals belonging to the common haplogroup V HV1/VH2 type (designated as "V1") were selected for whole genome sequencing. Phylogenetic analysis using parsimony was performed on these 130 sequences. Multiple

most parsimonious trees (MPTs) having a tree length of 335 (consistency index 0.86, retention index 0.91) were obtained. One representative MPT is shown in Figure 4. Nearly all haplogroup H and V sequences differed from one another, with 130 sequences resolved into 115 haplotypes. To condense the size of the tree for ease of viewing, several monophyletic branches were collapsed according to the common HV1/HV2 type. A small number of characters (e.g. diagnostic sites to distinguish haplogroups H and V, neutral sites that distinguish the monophyletic clades of H4, H5 and H7, position 3010, and additional branches of noted interest) have been mapped upon this tree.

The tree of the eight common H and V HV1/HV2 types (Figure 4) bifurcates basally, separating haplogroups H and V, consistent with the diagnostic coding region mutations that distinguish these haplogroups. It is interesting, however, that some of the HV1/HV2 types we designated as "H", based on their control region sequence, actually prove to belong to haplogroup V (H1-20 and H1-29 – indicated by the joined arrows) when sequence information in the coding region is included in the analysis (Figure 4). It is apparent that these sequences underwent a C to T reversion at position 16298 in HV1. Conversely, we also observe an instance where a V1 individual (V1-11) is actually a haplogroup H sequence that independently gained the 16298C polymorphism, thus appearing to belong to haplogroup V (see arrow, Figure 4). These results are not surprising since 16298 has been demonstrated to be a relatively "fast" site for mutations to occur (Wakeley 1993; Hasegawa *et al.* 1993; see also Allard *et al.* 2002 – table 2). Position 16298 is the only HV1/HV2 site that distinguishes haplogroup V's from H's (position 72, specific to haplogroup V, is outside of the range for HV2), so homoplasy at that site results in inaccurate haplogroup designation when only HV1/HV2 are

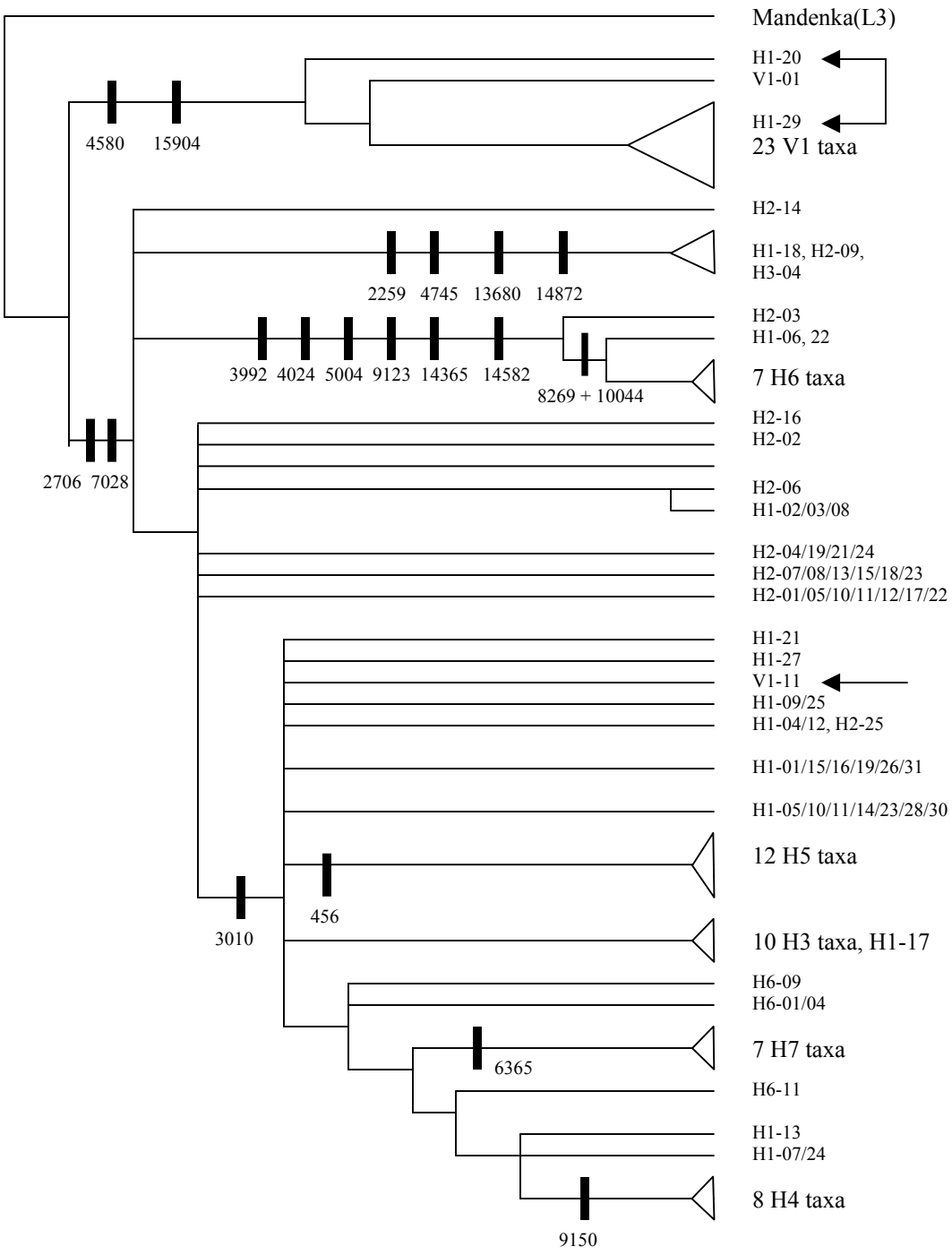

Mandenka(L3)
H1-20
V1-01
H1-29
23 V1 taxa
4580
15904
H2-14
H1-18, H2-09,
H3-04
2259
4745
13680
14872
H2-03
H1-06, 22
3992
4024
5004
9123
14365
14582
8269 + 10044
7 H6 taxa
H2-16
H2-02
2706
7028
H2-06
H1-02/03/08
H2-04/19/21/24
H2-07/08/13/15/18/23
H2-01/05/10/11/12/17/22
H1-21
H1-27
V1-11
H1-09/25
H1-04/12, H2-25
H1-01/15/16/19/26/31
H1-05/10/11/14/23/28/30
12 H5 taxa
456
3010
10 H3 taxa, H1-17
H6-09
H6-01/04
7 H7 taxa
6365
H6-11
H1-13
H1-07/24
8 H4 taxa
9150

Figure 4. Phylogenetic Analysis of the Superhaplogroup HV Cluster. The mtGenomes of 105 haplogroup H individuals having common HV1/HV2 types and 25 haplogroup V individuals were sequenced. The African haplogroup L3 sequence "Mandenka (Genbank ID – AF346995)" from Ingman et al (2000) was used as an outgroup sequence. One most parsimonious tree is shown (Tree length of 335, consistency index 86, retention index 91). Trees were generated using the software program Winclada version 1.00.08 (Nixon 2002) running NONA (Goloboff 1999). Vertical bars on branches indicate selected character state changes of interest. The arrows indicate instances of reversion of the haplogroup V diagnostic polymorphism 16298.

considered.

Three of the seven HV1/HV2 types (H4, H5, and H7) form monophyletic clusters in the HV tree (Figure 4). Individuals in the common HV1/HV2 type H4 are grouped by the coding region polymorphism 9150G, and members of the common HV1/HV2 type H7 are grouped by the 6365C polymorphism. The individuals of the common HV1/HV2 type H5 share the control region polymorphism 456C. The remaining sequences examined (H1, H2, H3, and H6) form paraphyletic groups on the tree. This is not surprising since the HV1/HV2 sites that define these HV1/HV2 types (H2: 152T-C; H3: 16129G-A; H6: 73A-G) have been shown to have fast mutation rates (Hasegawa *et al.* 1993; Excoffier and Yang 1999; Meyer *et al.*, 1999; Pesole and Saccone 2001; Allard *et al.* 2002; Malyarchuck *et al.* 2002). For example, individuals H1-18, H2-09, and H3-04 form a monophyletic clade based on the shared mutations at positions 2259T, 4745G, 13680T, and 14872T (Figure 4). Homoplasic mutations in HV1/HV2 have separated these sequences into three common HV1/HV2 types.

One of the interesting common HV1/HV2 types is H6 (73A-G). Of the eleven sequences, seven form a monophyletic cluster distinguished by the eight polymorphisms 3992T, 4024G, 5004C, 8269A, 9123A, 10044G, 14365T, and 14582G (Figure 4). The longer branch length for this cluster suggests this HV1/HV2 type is older than other individuals of haplogroup H, an observation also made by Richards *et al.*, 2000 and Herrnstadt *et al.*, 2002. In fact there are an average of 9.3 mutations relative to the rCRS in the coding region among these seven H6 sequences compared to an average of 3.1 mutations in all other subhaplogroup H clusters combined (excluding the rare rCRS alleles 750A, 1438A, 4769A, 8860G, and 15326A). Two of the H1 individuals (H1-06

and H1-22) fall in this cluster – indicative of a probable reversion at position 73 to the rCRS allele. The other four H6 mtDNAs fall elsewhere on the tree, apparently the result of non-"H6" sequences gaining the 73A-G mutation in HV2.

II. 2. 3. Identification of Sites to Distinguish Individuals Matching HV1/HV2 Type H1

A total of 31 randomly selected, unrelated individuals belonging to HV1/HV2 type H1 were sequenced for the entire mtGenome. Parsimony analysis of the data resulted in a single MPT having a tree length of 128 (consistency index 0.99, retention index 0.97). Further random addition repetitions (100 total) failed to find additional trees of shorter length.

We observed 75 variable sites outside of HV1/HV2 that resolved the 31 H1 individuals into 28 haplotypes (Figure 5). The most common haplotype among the 31 H1 individuals (using the entire sequence information) was shared by 3 individuals (H1-02, H1-03, and H1-08). These individuals have only one mutation in their entire coding region (position 10211 C-T) compared to the rCRS sequence, itself a member of haplogroup H. Two additional individuals (H1-06 and H1-22) matched completely over the mtGenome with 8 mutations at positions 3992T, 4024G, 5004C, 8269A, 9123A, 10044G, 14365T, and 14582G (Figure 5). The only other HV1/HV2 types that share these 8 mutations are individuals belonging to the common HV1/HV2 type H6 (Figure 4), which is defined by the presence of 73G. It is probable that the H1-06 and H1-22 individuals were originally of the H6 type and have undergone a reversion at position 73

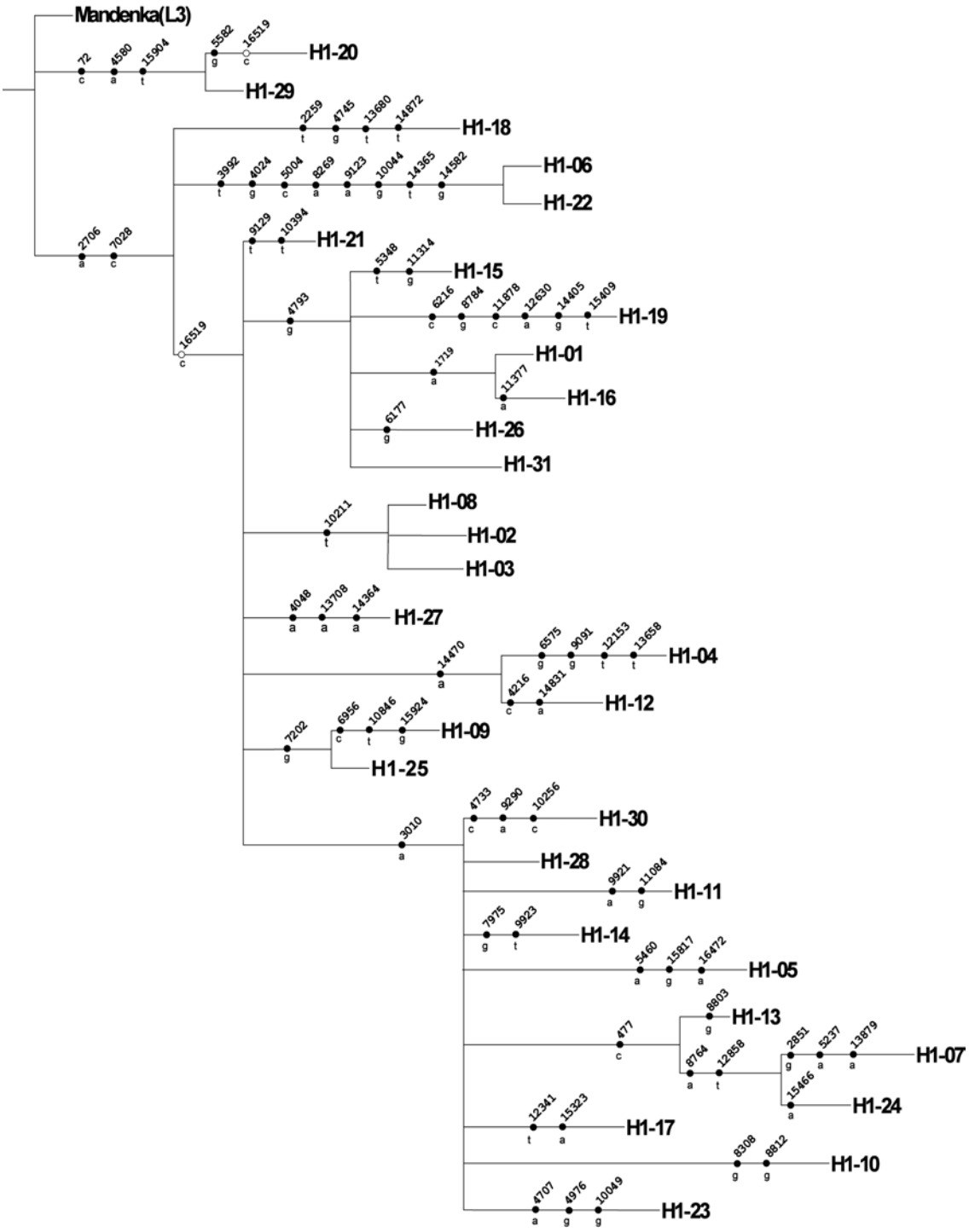

Mandenka(L3)
72 c
4580 a
15904 t
5582 g
16519 c
H1-20
H1-29
2259 t
4745 g
13680 t
14872 t
H1-18
3992 t
4024 g
5004 c
8269 a
9123 a
10044 g
14365 t
14582 g
H1-06
H1-22
2706 a
7028 c
9129 t
10394 t
H1-21
5348 t
11314 g
H1-15
4793 g
6216 c
8784 g
11878 c
12630 a
14405 g
15409 t
H1-19
16519 c
1719 a
H1-01
11377 a
H1-16
6177 g
H1-26
H1-31
10211 t
H1-08
H1-02
H1-03
4048 a
13708 a
14364 a
H1-27
14470 a
6575 g
9091 g
12153 t
13658 t
H1-04
4216 c
14831 a
H1-12
7202 g
6956 c
10846 t
15924 g
H1-09
H1-25
3010 a
4733 c
9290 a
10256 c
H1-30
H1-28
9921 a
11084 g
H1-11
7975 g
9923 t
H1-14
5460 a
15817 g
16472 a
H1-05
8803 g
H1-13
477 c
8764 a
12858 t
2851 g
5237 a
13879 a
H1-07
15466 a
H1-24
12341 t
15323 a
H1-17
8308 g
8812 g
H1-10
4707 a
4976 g
10049 g
H1-23

Figure 5. Phylogenetic Analysis of the H1 Individuals Using Data from the Entire mtGenome. The mtDNA genome of 31 non-related individuals belonging to the common HV1/HV2 type, H1, were sequenced and analyzed. The African haplogroup L3 sequence "Mandenka (Genbank ID – AF346995)" from Ingman et al (2000) was used as an outgroup sequence. Trees were generated using the software program Winclada version 1.00.08 (Nixon 2002) running NONA (Goloboff 1999). A single MPT having a tree length of 128 (consistency index 0.99, retention index 0.97) is shown. Open circles that mark character changes on a branch are instances of homoplasy. The 31 H1 sequences could be resolved into 28 haplotypes.

to appear as HV1/HV2 type H1 (Figure 4).

II. 2. 4. Application of the Criteria for Site Selection Using H1 as an Example

The primary goal of the genome sequencing strategy was to identify SNPs for practical forensic assays to resolve Caucasian individuals sharing common HV1/HV2 types. We adopted criteria that the sites should be a) neutral, b) shared, and c) non-redundant (see Chapter 1). Restricting the data to only neutral sites (i.e. removing all sites that potentially create changes in the phenotype), there are 44 variable sites that resolve the 31 H1 individuals into 23 haplotypes (Figure 6). It would be an easy task to simply assay one neutral SNP from each branch (Figure 6) to resolve these particular H1 individuals. However, polymorphisms that occur in only a single individual may be "private mutations" that are quite rare or unique in the population. Private, or very rare, mutations are poor choices for generally useful SNP assays to increase forensic discrimination. Considering only neutral SNPs that are shared in multiple H1 individuals, we identified 13 variable sites that resolve the 31 individuals into 14 haplotypes (Figure 7). However, SNP assay site selection was not finalized until all mtGenomes were sequenced in order to include sites that were shared among different common HV1/HV2 types. For example, individual H1-18 has three neutral mutations (4745, 13680, and 14872 – Figure 6) not shared among other H1 individuals. None of these sites would meet the second criterion for consideration as a potential SNP site. However, after sequencing the H2 and H3 common types, we observed that these sites are shared among different types (Figure 4) and would be included as useful SNP assay

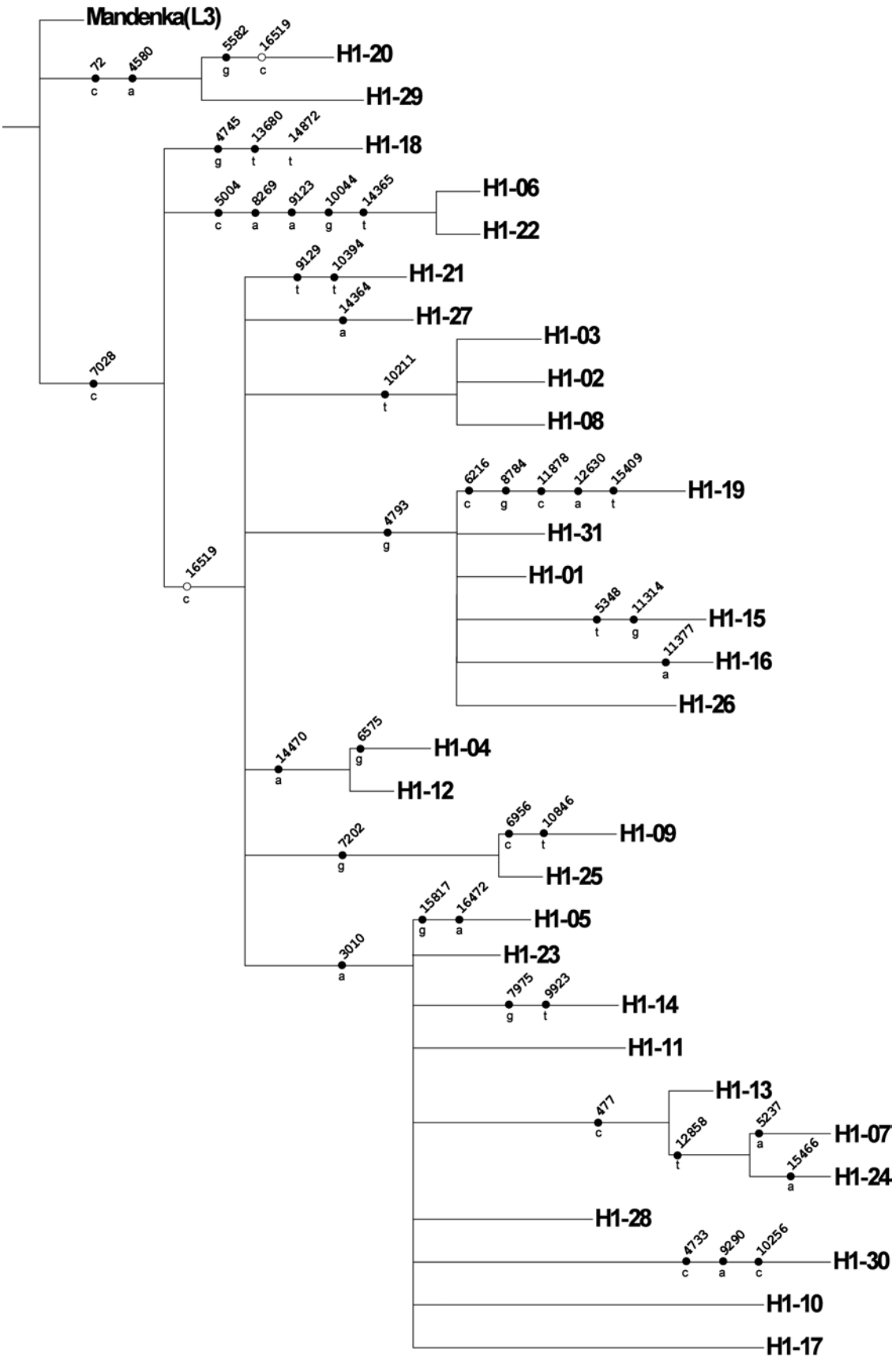

Mandenka(L3)
72 c
4580 a
5582 g
16519 c
H1-20
H1-29
4745 g
13680 t
14872 t
H1-18
5004 c
8269 a
9123 a
10044 g
14365 t
H1-06
H1-22
9129 t
10394 t
H1-21
14364 a
H1-27
7028 c
10211 t
H1-03
H1-02
H1-08
16519 c
6216 c
8784 g
11878 c
12630 a
15409 t
H1-19
4793 g
H1-31
H1-01
5348 t
11314 g
H1-15
11377 a
H1-16
H1-26
14470 a
6575 g
H1-04
H1-12
7202 g
6956 c
10846 t
H1-09
H1-25
3010 a
15817 g
16472 a
H1-05
H1-23
7975 g
9923 t
H1-14
H1-11
477 c
H1-13
12858 t
5237 a
H1-07
15466 a
H1-24
H1-28
4733 c
9290 a
10256 c
H1-30
H1-10
H1-17

Figure 6. Phylogenetic Analysis of the H1 Individuals Using Only Neutral SNPs from the Entire mtGenome. The mtDNA genome of 31 non-related individuals belonging to the common HV1/HV2 type, H1, were sequenced and analyzed. The African haplogroup L3 sequence "Mandenka (Genbank ID – AF346995)" from Ingman et al (2000) was used as an outgroup sequence. Trees were generated using the software program Winclada version 1.00.08 (Nixon 2002) running NONA (Goloboff 1999). Open circles that mark character changes on a branch are instances of homoplasy. A single MPT having a tree length of 97 (consistency index 0.98, retention index 0.97) is shown. The 31 H1 sequences could be resolved into 23 haplotypes using only neutral SNPs.

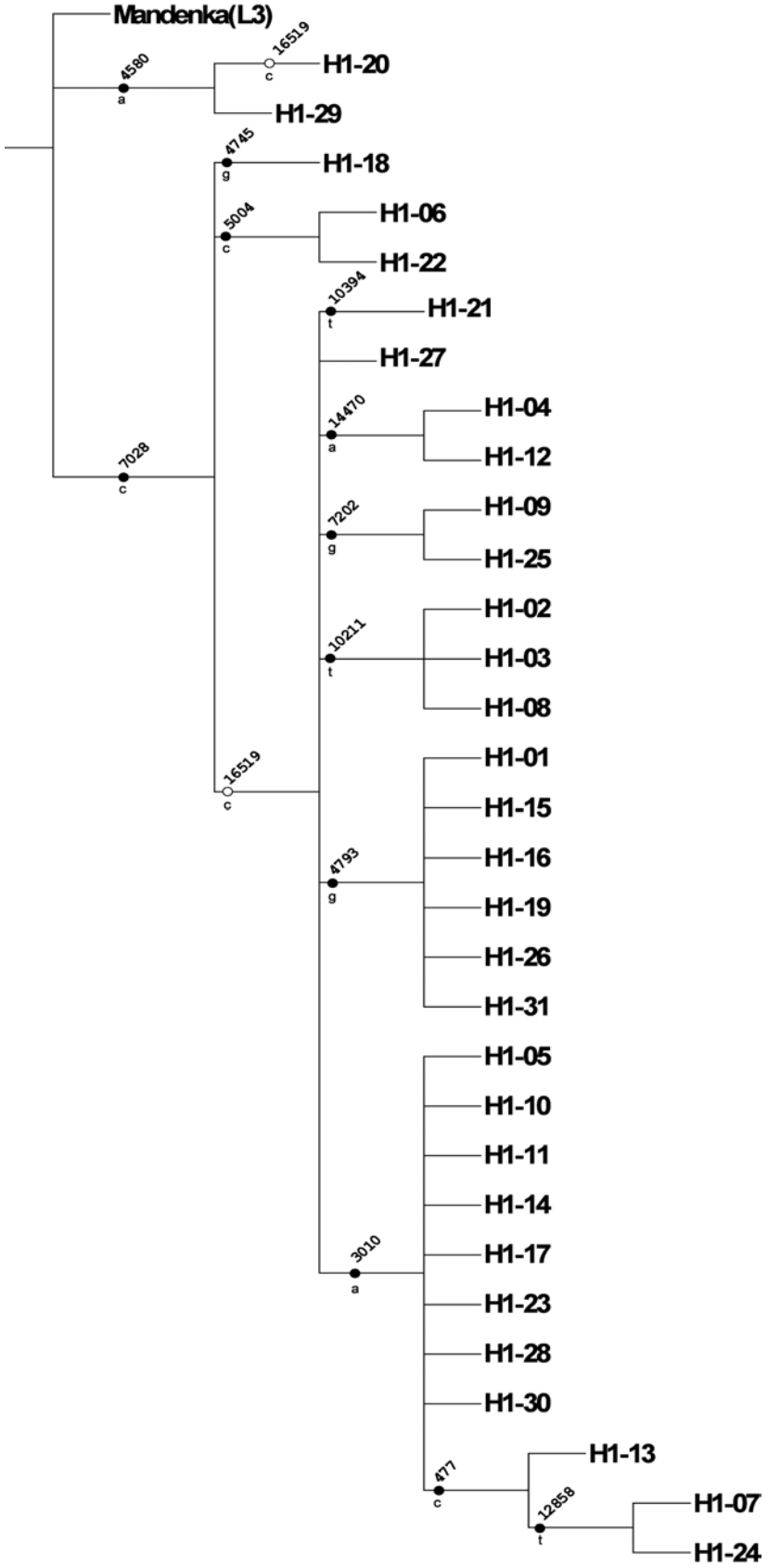

Mandenka(L3)
16519
c
4580
a
H1-20
H1-29
4745
g
H1-18
5004
c
H1-06
H1-22
10394
t
H1-21
H1-27
14470
a
H1-04
H1-12
7028
c
7202
g
H1-09
H1-25
10211
t
H1-02
H1-03
H1-08
16519
c
H1-01
H1-15
4793
g
H1-16
H1-19
H1-26
H1-31
H1-05
H1-10
H1-11
H1-14
3010
a
H1-17
H1-23
H1-28
H1-30
H1-13
477
c
12858
t
H1-07
H1-24

Figure 7. Phylogenetic Analysis of the H1 Individuals Using only Shared, Neutral Sites from the Entire mtGenome. The mtDNA genome of 31 non-related individuals belonging to the common HV1/HV2 type, H1, were sequenced and analyzed using the phylogenetic program Winclada version 1.00.08 (Nixon 2002) running NONA (Goloboff 1999). The African haplogroup L3 sequence "Mandenka (Genbank ID – AF346995)" from Ingman et al (2000) was used as an outgroup sequence. Open circles that mark character changes on a branch are instances of homoplasy. A single MPT having a tree length of 14 (consistency index 0.92, retention index 0.96) is shown. The 31 H1 sequences could be resolved into 14 haplotypes using these neutral, shared SNPs.

sites.

The third criterion, non-redundancy, was often the most difficult to implement. Since the mtGenome is one linked molecule, several sites would often separate out the same set of individuals. The decision to choose SNPs among these multiple potential assay sites was made by comparisons to other HV1/HV2 types. If one site in particular varied widely among other common HV1/HV2 types, then this SNP was more attractive for selection. For this reason, efficient SNP selection was conducted after the completion of mtGenome sequencing (see above).

Included among the 13 variable neutral sites are three polymorphisms within the control region: 16519T-C, 477T-C and the 523/524 AC deletion. Position 16519 C is the only homoplasic character mapped upon this particular tree, resolving individual H1-20 from individual H1-29 (the two haplogroup V sequences that appear as HV1/HV2 type H1), and is present in all other H1 individuals with the exception of H1-18, H1-06, and H1-22 (Figure 6). Previous research (see Chen *et al.* 1995; Richards *et al.* 1998; Finnila *et al.* 2001; Ingman and Gyllensten, 2001) has revealed position 16519 to be one of the most quickly evolving sites in the mtDNA genome.

A schematic diagram of how the selected neutral SNP sites resolve H1 individuals into component haplotypes is represented in Figure 8. The most common type after the application of the neutral SNPs occurs in 7 of the 31 (23%) of the samples (Figure 8). The process of SNP identification for all other common HV1/HV2 types followed the same strategy as outlined above.

This section has presented in detail our strategy for SNP selection for the common HV1/HV2 type H1. This strategy was also used for identifying SNPs to discriminate

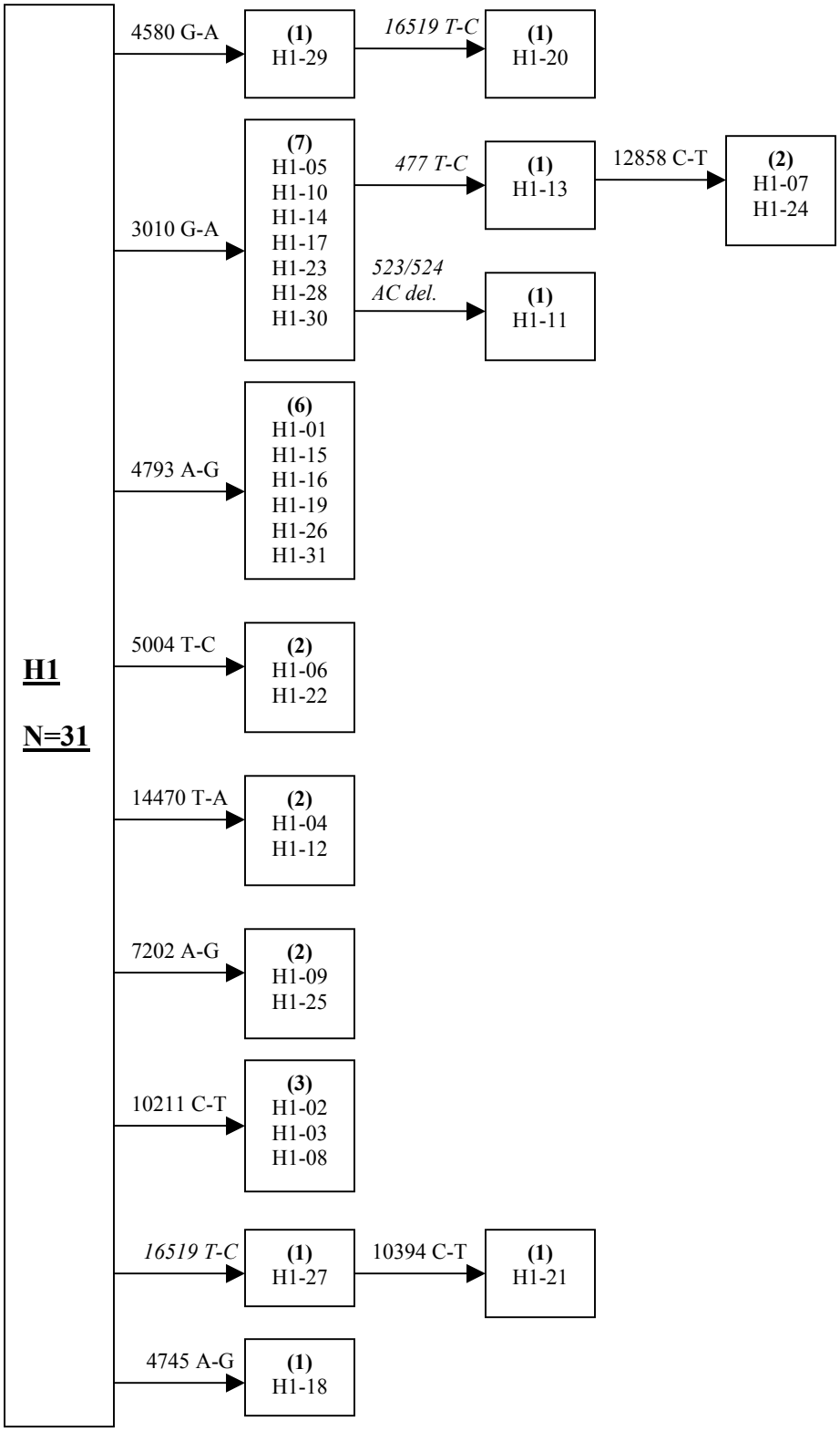

H1
N=31
4580 G-A
(1)
H1-29
16519 T-C
(1)
H1-20
3010 G-A
(7)
H1-05
H1-10
H1-14
H1-17
H1-23
H1-28
H1-30
477 T-C
(1)
H1-13
12858 C-T
(2)
H1-07
H1-24
523/524
AC del.
(1)
H1-11
4793 A-G
(6)
H1-01
H1-15
H1-16
H1-19
H1-26
H1-31
5004 T-C
(2)
H1-06
H1-22
14470 T-A
(2)
H1-04
H1-12
7202 A-G
(2)
H1-09
H1-25
10211 C-T
(3)
H1-02
H1-03
H1-08
16519 T-C
(1)
H1-27
10394 C-T
(1)
H1-21
4745 A-G
(1)
H1-18

Figure 8. Schematic diagram of the resolution of the HV1/HV2 common Caucasian type H1 after applying thirteen neutral SNPs over the entire mtDNA genome.
Nucleotide positions within the control region (outside of HV1/HV2) are in italics. Thirteen shared, neutral polymorphisms resolved the 31 sequences into 14 haplotypes.

HV1/HV2 types, with the results summarized below.

II. 2. 5. Neutral Sites to Resolve H2 Individuals

The HV1/HV2 type H2 (152T-C) is the second most common type in Caucasians, found at a frequency of 2.1% (Table 2). A total of 25 randomly selected, non-related individuals belonging to HV1/HV2 type H2 were selected for whole genome sequencing. We identified 14 neutral, shared SNP sites in the H2 individuals (Figure 9). The most common type after the application of the neutral, shared SNPs occurred in 4 of the 25 (16%) of the H2 common HV1/HV2 types (Figure 9).

II. 2. 6. Neutral Sites to Resolve all Other Haplogroup H Common HV1/HV2 Types

Five neutral, shared SNPs were identified in the common Caucasian HV1/HV2 type H3, resolving the eleven individuals into six haplotypes (Figure 10). Only one neutral SNP site (9380G-A) was observed among the HV1/HV2 common type H4 (Figure 11). This site resolved the H4 individuals into two haplotypes (Figure 11). Four of the eight H4 individuals we sequenced (H4-01, H4-05, H4-06 and H4-08) matched completely over the mtGenome.

Three neutral, shared SNPs were able to resolve the twelve HV1/HV2 type H5 individuals into four haplotypes (Figure 12). Two of these SNPs were found in the control region outside of HV1/HV2: position 16519T-C and the deletion of the AC repeat unit at positions 523 and 524. The 523/524 AC-del was found in eight of the twelve

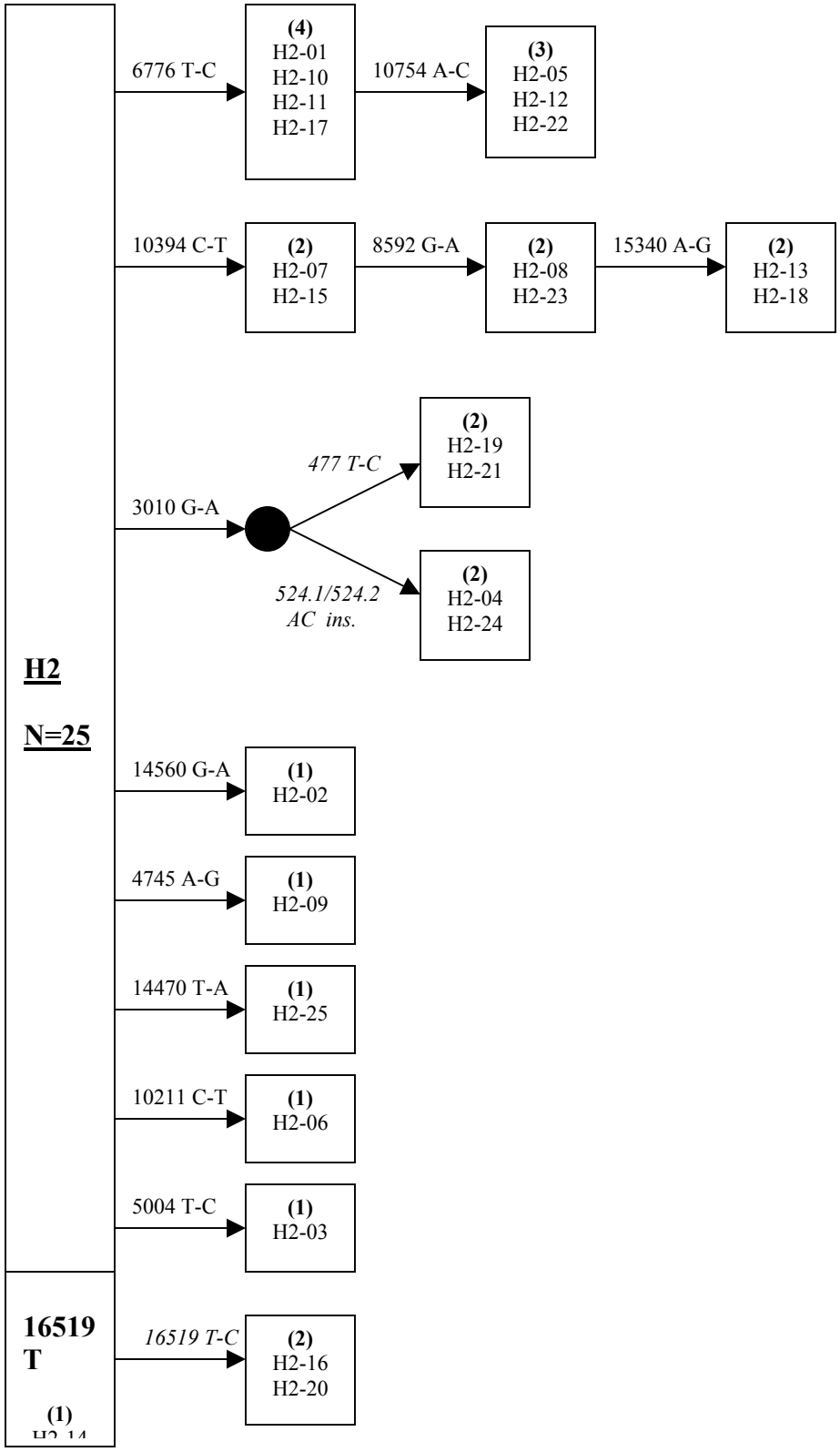

H2
N=25
16519 T
(1)
6776 T-C
(4)
H2-01
H2-10
H2-11
H2-17
10754 A-C
(3)
H2-05
H2-12
H2-22
10394 C-T
(2)
H2-07
H2-15
8592 G-A
(2)
H2-08
H2-23
15340 A-G
(2)
H2-13
H2-18
3010 G-A
477 T-C
(2)
H2-19
H2-21
524.1/524.2 AC ins.
(2)
H2-04
H2-24
14560 G-A
(1)
H2-02
4745 A-G
(1)
H2-09
14470 T-A
(1)
H2-25
10211 C-T
(1)
H2-06
5004 T-C
(1)
H2-03
16519 T-C
(2)
H2-16
H2-20

Figure 9. Schematic diagram of the resolution of the HV1/HV2 Common Caucasian type H2 after applying eight neutral SNPs over the entire mtDNA genome.
Nucleotide positions within the control region (outside of HV1/HV2) are in italics. Shared, neutral SNPs were useful for resolving the 25 sequences into 14 haplotypes.

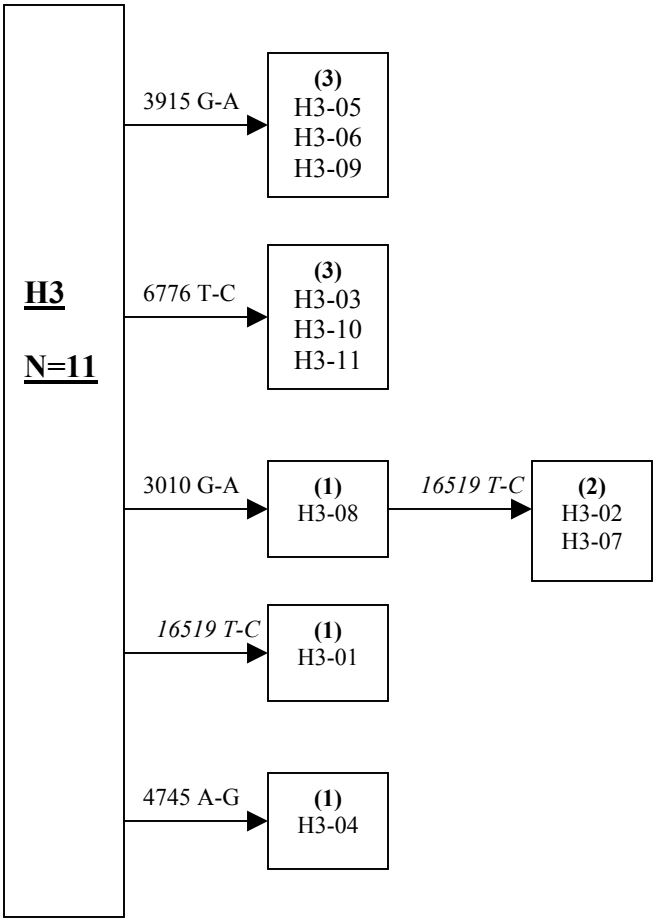


Figure 10. Schematic diagram of the resolution of the HV1/HV2 Common Caucasian type H3 after applying five neutral SNPs over the entire mtGenome. Nucleotide positions within the control region (outside of HV1/HV2) are in italics. Neutral, shared SNPs resolved the 11 sequences into six haplotypes.

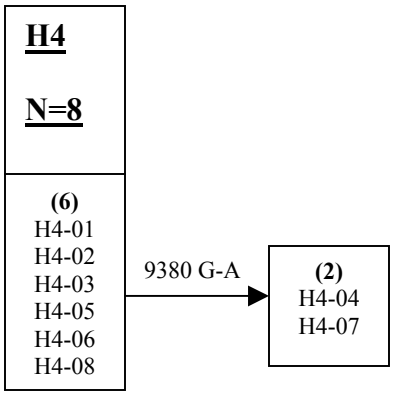


Figure 11. Schematic diagram of the resolution of the HV1/HV2 Common Caucasian type H4 after applying one neutral SNPs over the entire mtGenome. The neutral SNP observed in the data set of eight non-related individuals was useful for resolving the sequences into two haplotypes.

individuals.

The only neutral coding region polymorphism was 11719G-A, found in two individuals (H5-03 and H5-12 – Figure 12). It is interesting to note this polymorphism since this SNP is a marker of the haplogroup HV cluster. Haplogroup H individuals typically have the 11719G variant, and these two HV1/HV2 type H5 individuals have the non-haplogroup H allele at this position. Herrnstadt *et al.* (2002) observed the same mutational event in 6/226 Caucasian coding region mtGenomes belonging to haplogroup H. It would be interesting to know if the six samples in the Herrnstadt *et al.* (2002) data set belong to our HV1/HV2 type H5. However, the data published thus far from the Herrnstadt *et al.* (2002) study lacks control region information.

The common HV1/HV2 type H6 can be divided into two categories: seven of eleven individuals belonging to the "long branch" cluster (discussed above, see Figure 4) and four individuals that lack this set of mutations. All eleven individuals within H6 can be resolved into five haplotypes using four shared, neutral SNPs (Figure 13).

We found no neutral, shared SNPs in the common HV1/HV2 type H7 individuals. Three of the seven H7 individuals matched exactly in the coding region (H7-01, H7-05, and H7-07). Given the relatively small sample size for H7 (n=7), increasing the number of sequences may be useful for identifying SNPs.

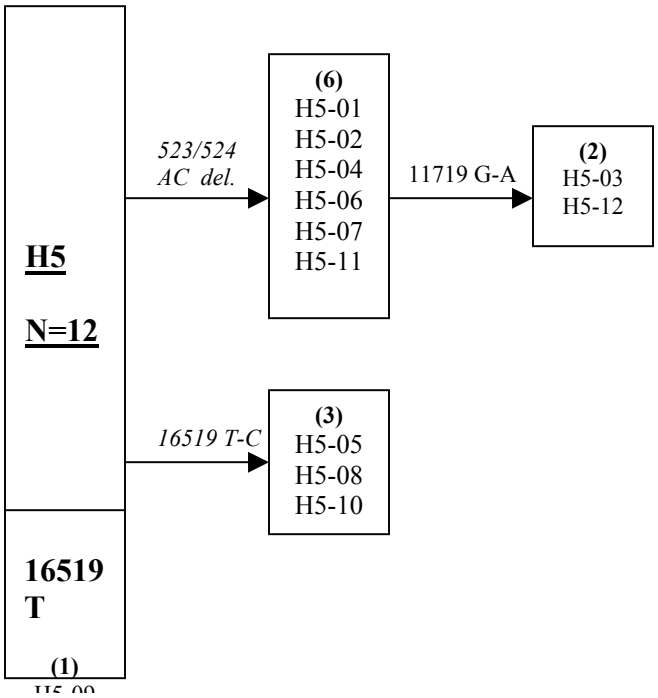


Figure 12. Schematic diagram of the resolution of the HV1/HV2 Common Caucasian type H5 after applying three neutral SNPs over the entire mtGenome. Nucleotide positions within the control region (outside of HV1/HV2) are in italics. Shared, neutral SNPs identified were useful for resolving the 12 sequences into four haplotypes.

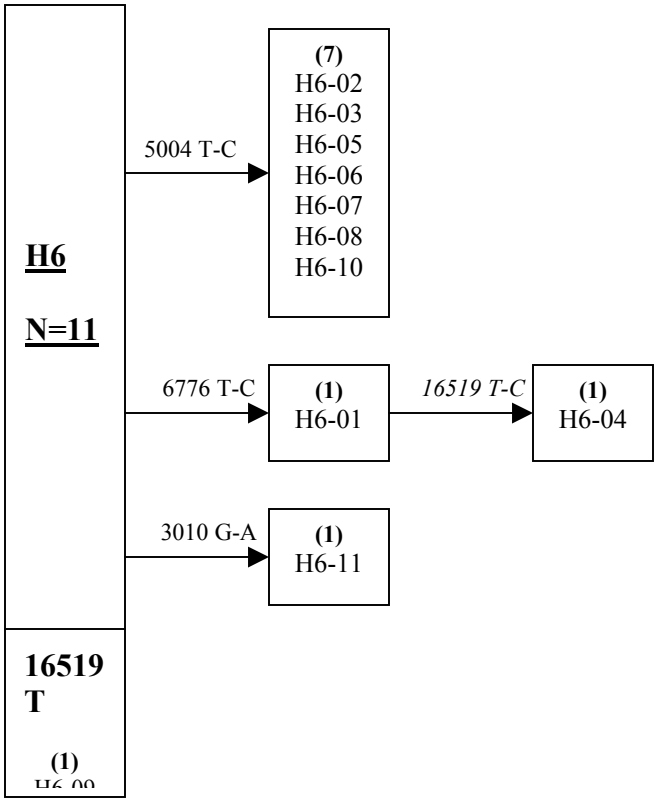


Figure 13. Schematic diagram of the resolution of the Common Caucasian HV1/HV2 type H6 after applying four neutral SNPs over the entire mtGenome. Nucleotide positions within the control region (outside of HV1/HV2) are in italics. Neutral, shared SNPs were useful for resolving the 11 sequences into five haplotypes.

II. 2. 7. Haplogroup V Analysis

Twenty-five unrelated individuals grouped by the haplogroup V polymorphism at position 16298T-C were sequenced in this study. These individuals were assigned the common HV1/HV2 type nomenclature, V1. Parsimony analysis of the entire mtDNA genome for V1 generated eight MPTs having a tree length of 71 (consistency index 0.91; retention index 0.83). One representative MPT is shown for these 25 individuals in Figure 14.

As noted previously, one individual (V1-11) classified as being a member of haplogroup V, based on HV1/HV2 sequence information, was actually a member of haplogroup H (Figure 4). To distinguish V1 individuals that were misclassified according to their HV1/HV2 haplotype, the haplogroup V diagnostic polymorphism 4580 was utilized as a SNP site. Entire mtGenome sequence information was used to resolve the 25 V1 sequences into 23 haplotypes (Figure 14). Two sequences (V1-01 and V1-19) have undergone a reversion at the haplogroup V-associated polymorphism 72T-C (Figure 14 – V1-01 is basal to the clade defined by 72T-C and the V1-19 branch is noted by the reversion to 72T). Previous studies have not characterized position 72 as a "fast" site (Meyer *et al.* 1999; Allard *et al.* 2002). Three individuals (V1-06, V1-10 and V1-15) matched exactly in the coding region, and could not be distinguished from one another. We identified eight neutral, shared SNPs to resolve the 25 V1 individuals into nine haplotypes (Figure 15).

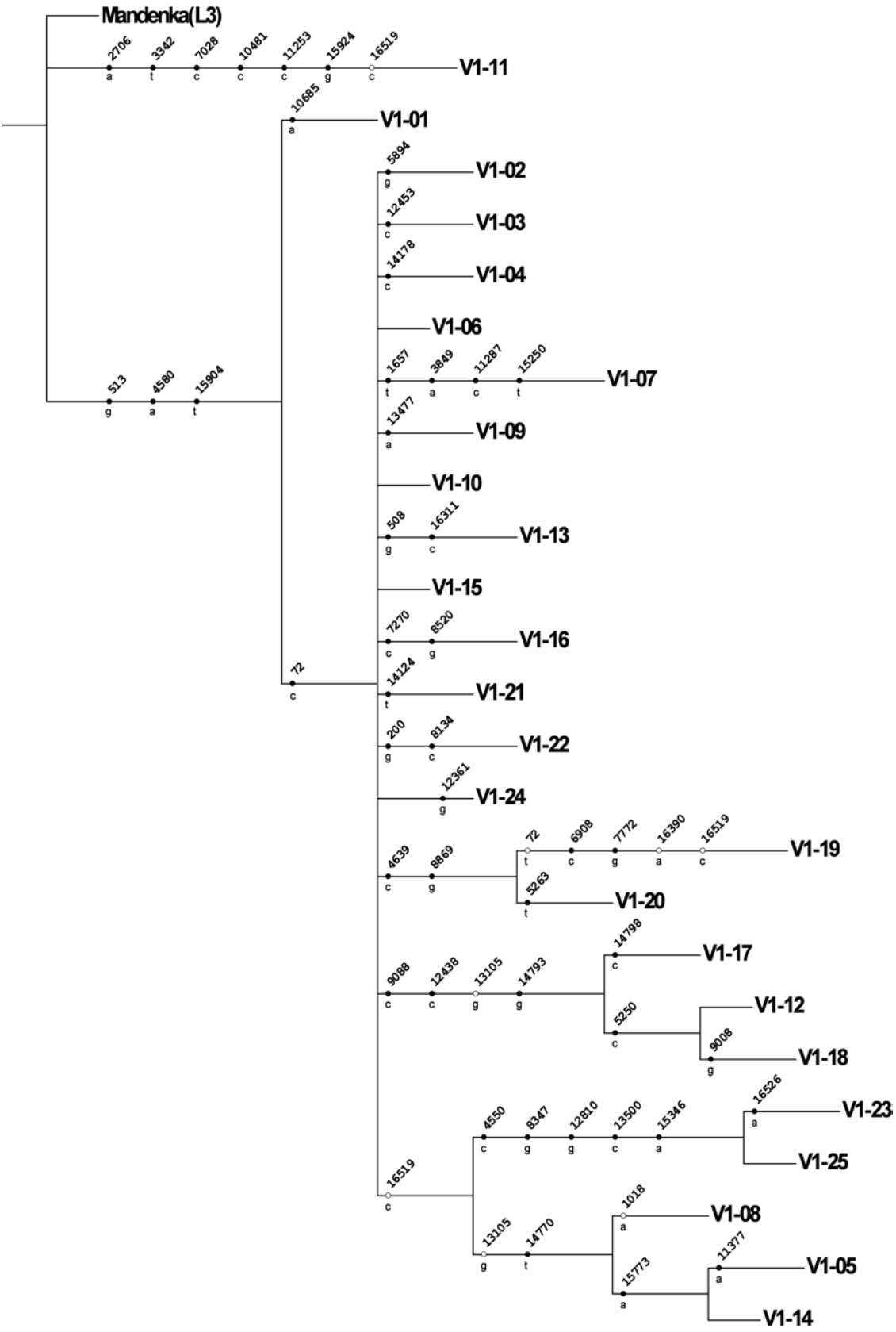

Mandenka(L3)
2706 a
3342 t
7028 c
10481 c
11253 c
15924 g
16519 c
V1-11
10685 a
V1-01
5894 g
V1-02
12453 c
V1-03
14178 c
V1-04
V1-06
1657 t
3849 a
11287 c
15250 t
V1-07
513 g
4580 a
15904 t
13477 a
V1-09
V1-10
508 g
16311 c
V1-13
V1-15
7270 c
8520 g
V1-16
72 c
14124 t
V1-21
200 g
8134 c
V1-22
12361 g
V1-24
72 t
6908 c
7772 g
16390 a
16519 c
V1-19
4639 c
8869 g
5263 t
V1-20
14798 c
V1-17
9088 c
12438 c
13105 g
14793 g
V1-12
5250 c
9008 g
V1-18
16526 a
V1-23
4550 c
8347 g
12810 g
13500 c
15346 a
V1-25
16519 c
1018 a
V1-08
13105 g
14770 t
11377 a
V1-05
15773 a
V1-14

Figure 14. Phylogenetic Analysis of the V1 Individuals Using Data from the mtGenome. The mtGenome of 25 unrelated individuals belonging to the common HV1/HV2 type V1 were sequenced and analyzed. The African haplogroup L3 sequence "Mandenka (Genbank ID – AF346995)" from Ingman et al (2000) was used as an outgroup sequence. Trees were generated using the software program Winclada version 1.00.08 (Nixon 2002) running NONA (Goloboff 1999). A single MPT having a tree length of 71 (consistency index 0.91, retention index 0.83) is shown. Open circles that mark character changes on a branch are instances of homoplasy. The 25 sequences were resolved into 23 haplotypes.

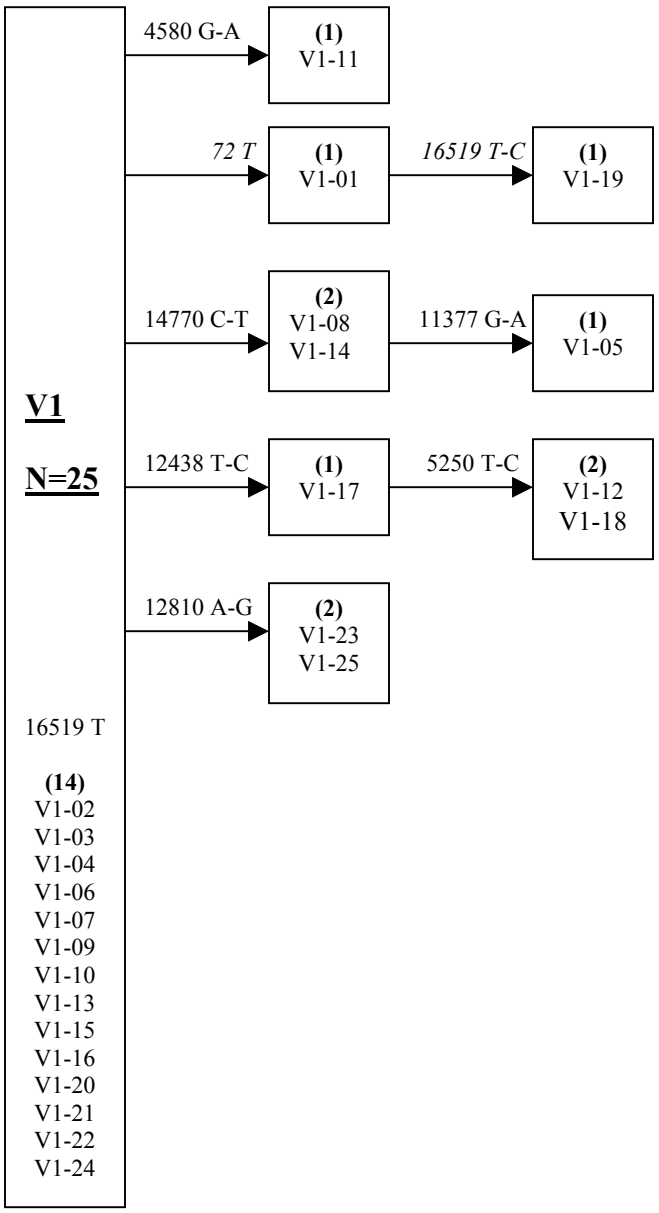


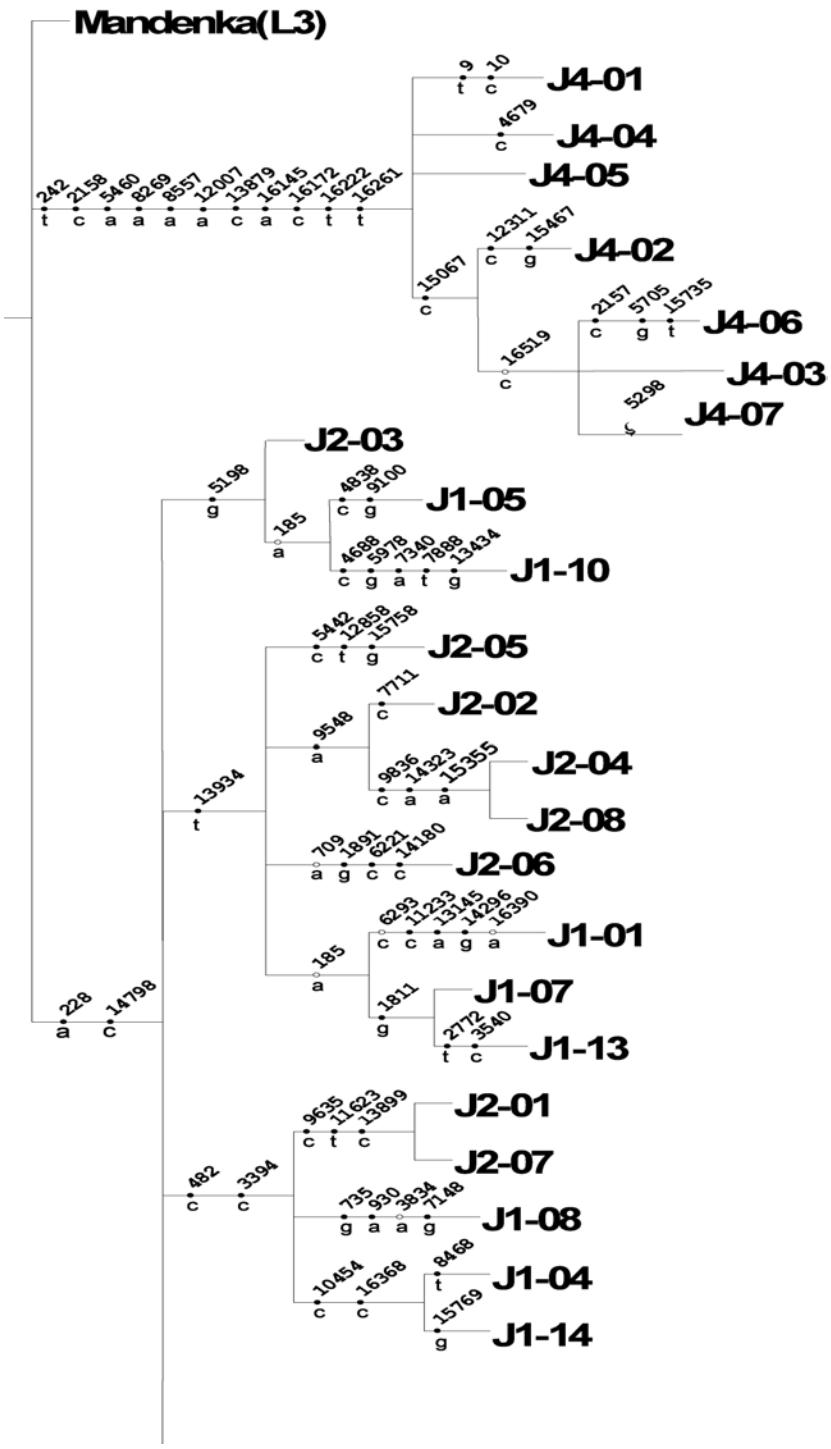
Figure 15. Schematic diagram of the resolution of the HV1/HV2 common Caucasian HV1/HV2 type V1 after applying eight neutral SNPs over the entire mtGenome.
Nucleotide positions within the control region (outside of HV1/HV2) are in italics. The eight shared, neutral sites resolved the 25 sequences into nine haplotypes.

II. 2. 8. Superhaplogroup J/T Analysis

The Caucasian haplogroups J and T are sister haplogroups each found at a frequency of ~10% within European and United States Caucasian populations (Torroni *et al.* 1996; Macaulay et al 1999a; Allard *et al.* 2002). The J/T cluster is grouped by the coding region diagnostic polymorphisms 4216T-C, 11251A-G and 15452C-A and the HV1 diagnostic polymorphism at 16126T-C (Torroni *et al.* 1994; Torroni *et al.* 1996; Macaulay *et al.* 1999a; Finnila *et al.* 2001). Haplogroup J is grouped by the coding region mutations 10398A-G, 12612A-G and 13708G-A and the HV1/HV2 diagnostic mutations 16069C-T and 295C-T (Finnila *et al.* 2001; Herrnstadt *et al.* 2002; Allard *et al.* 2002). Haplogroup T is grouped by coding region-specific mutations 709G-A, 1888G-A, 4917A-G, 8697G-A, 10463T-C, 13368G-A, 14905G-A, 15607A-G and 15928G-A; and the HV1-specific polymorphism 16294C-T (Finnila *et al.* 2001; Herrnstadt *et al.* 2002). We have identified four common HV1/HV2 J types (J1-J4) and three common HV1/HV2 T types (T1-T3) occurring at a frequency of 0.5% or greater in the Caucasian mtDNA database (Table 1).

II. 2. 9. Phylogenetic Analysis of the J Cluster

We sequenced 43 entire mtGenomes for the four common HV1/HV2 type J individuals. Multiple MPTs were constructed having a tree length of 168 (consistency index 0.94, retention index 0.93). One representative is shown for these 43 individuals in Figure 16. The J4 HV1/HV2 individuals form a “long branch” monophyletic clade


Mandenka(L3)
9 10
t c
J4-01
4679
c
J4-04
J4-05
242 2158 5460 8269 8557 12007 13879 16145 16172 16222 16261
t c a a a a c a c t t
12311 15467
c g
J4-02
15067
c
2157 5705 15735
c g t
J4-06
16519
c
J4-03
5298
J4-07
J2-03
5198
g
4838 9100
c g
J1-05
185
a
4688 5978 7340 7888 13434
c g a t g
J1-10
5442 12858 15758
c t g
J2-05
7711
c
J2-02
9548
a
9836 14323 15355
c a a
J2-04
J2-08
13934
t
709 1891 6221 14180
a g c c
J2-06
6293 11233 13145 14296 16390
c c a g a
J1-01
185
a
1811
g
J1-07
2772 3540
t c
J1-13
228 14798
a c
9635 11623 13899
c t c
J2-01
J2-07
482 3394
c c
735 930 3834 7148
g a a g
J1-08
10454 16368
c c
8468
t
J1-04
15769
g
J1-14

A

A

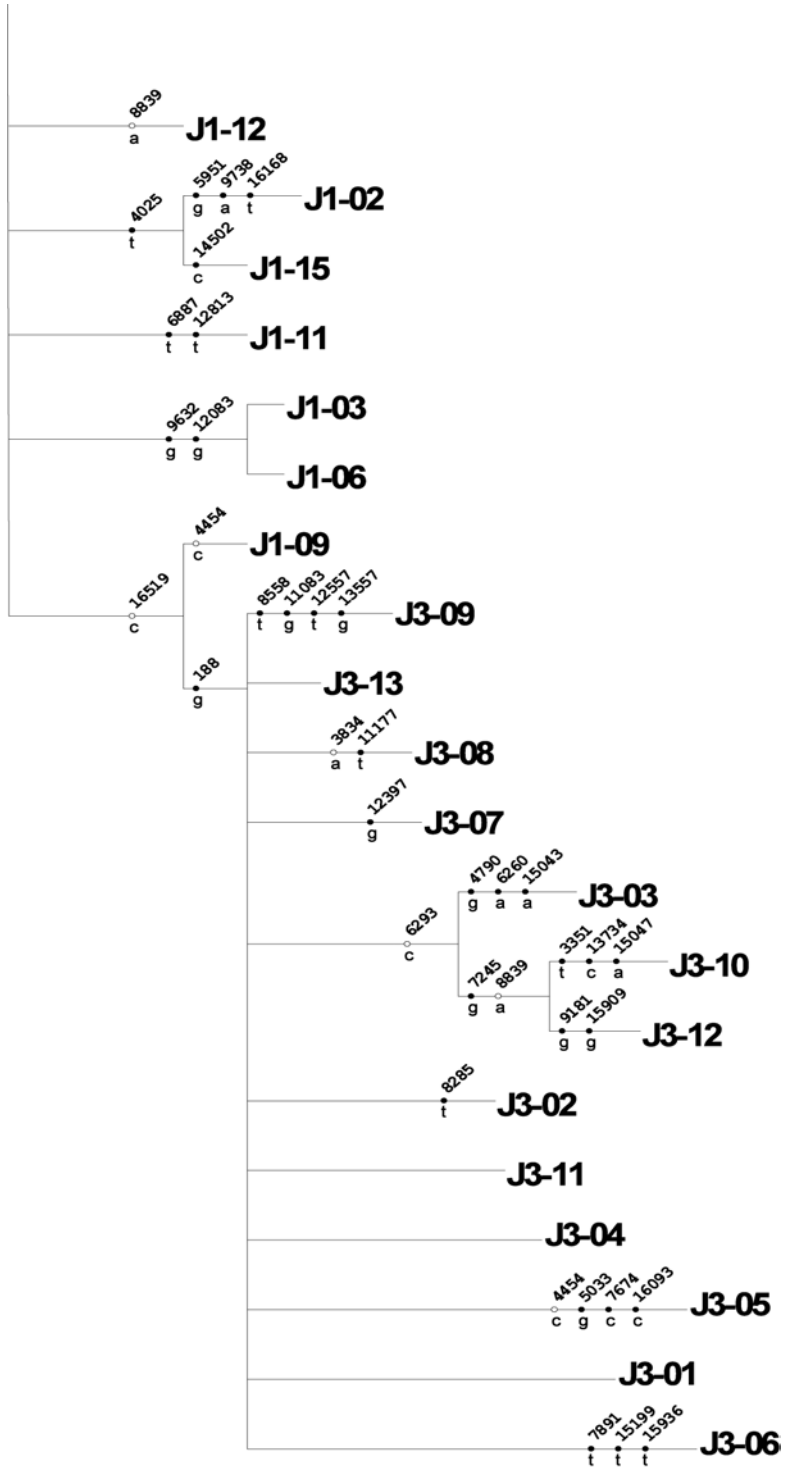

8839 a
J1-12
5951 g
9738 a
16168 t
J1-02
4025 t
14502 c
J1-15
6887 t
12813 t
J1-11
9632 g
12083 g
J1-03
J1-06
4454 c
J1-09
16519 c
8558 t
11083 g
12557 t
13557 g
J3-09
188 g
J3-13
3934 a
11177 t
J3-08
12397 g
J3-07
4790 g
6260 a
15043 a
J3-03
6293 c
3351 t
13734 c
15047 a
J3-10
7245 g
8839 a
9181 g
15909 g
J3-12
8285 t
J3-02
J3-11
J3-04
4454 c
5033 g
7674 c
16093 c
J3-05
J3-01
7891 t
15199 t
15936 t
J3-06

Figure 16. Phylogenetic Analysis of the Four Common HV1/HV2 type J Individuals Using Data from the mtGenome. The mtGenome of 43 unrelated individuals belonging to four common HV1/HV2 types (J1, J2, J3 and J4) of haplogroup J was sequenced and analyzed. The African haplogroup L3 sequence "Mandenka (Genbank ID – AF346995)" from Ingman et al (2000) was used as an outgroup sequence. Trees were generated using the software program Winclada version 1.00.08 (Nixon 2002) running NONA (Goloboff 1999). One representative MPT having a tree length of 168 (consistency index 0.94, retention index 0.93) is shown over two pages, connected by the capital letter A. Open circles that mark character changes on a branch are instances of homoplasy. The J4 sub-group forms a monophyletic cluster basal to the other three sub-groups.

grouped by six coding region mutations (2158T-C, 5460G-A, 8269G-A, 8557G-A, 12007G-A and 13879T-C). The J3 HV1/HV2 individuals also form a monophyletic group grouped by the 16519T-C mutation. One individual (J1-09) appears within the J3 clade and is likely grouped there due to a reversion at position 188A-G (the one mutation differing between J1s and J3s – Table 1).

The remaining common HV1/HV2 types J1 and J2 form paraphyletic groups in the tree (Figure 16). This is most likely because these two HV1/HV2 types differ from each another at position 185G-A, a site shown to be fast by both phylogenetic (Meyer *et al.* 1999; Allard *et al.* 2002) and empirical pedigree studies (Parsons *et al.* 1997). Individuals matching exactly over the mtDNA coding region include: two individuals from J1 (J1-03 and J1-06), two individuals from J2 (J2-04 and J2-08), four individuals from J3 (J3-01, J3-04, J3-11 and J3-13), and two J4 individuals (J4-03 and J4-05).

2.2.10 Neutral Sites to Resolve Haplogroup J Common HV1/HV2 Types

We have identified seven sites to resolve the fifteen common HV1/HV2 type J1 sequences into eight haplotypes (Figure 17). Three of the seven sites were found within the control region (outside of HV1/HV2): 16368T-C, 16519T-C and 482T-C. Four neutral, shared coding region sites (5198A-G, 9548G-A, 9836T-C and 12858C-T) in addition to the control region site 482T-C were useful for resolving the eight HV1/HV2 type J2 sequences into six haplotypes (Figure 18a). Five neutral sites were identified to resolve the thirteen J3 HV1/HV2 individuals into six haplotypes (Figure 18b). And finally, the coding region site 15067T-C and the control region polymorphism 16519T-C

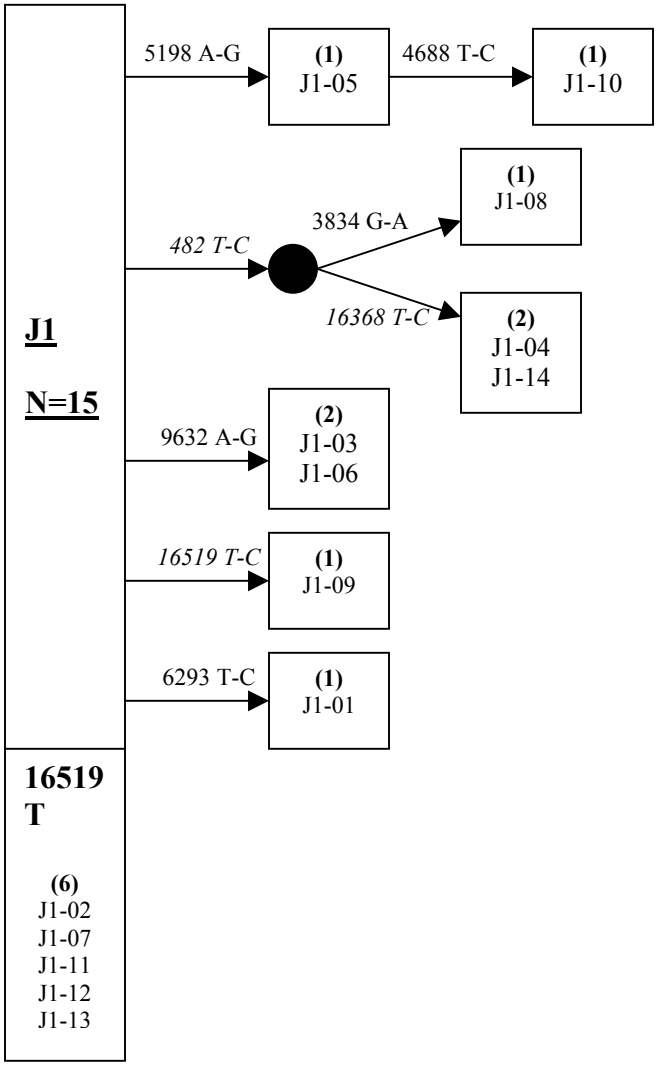
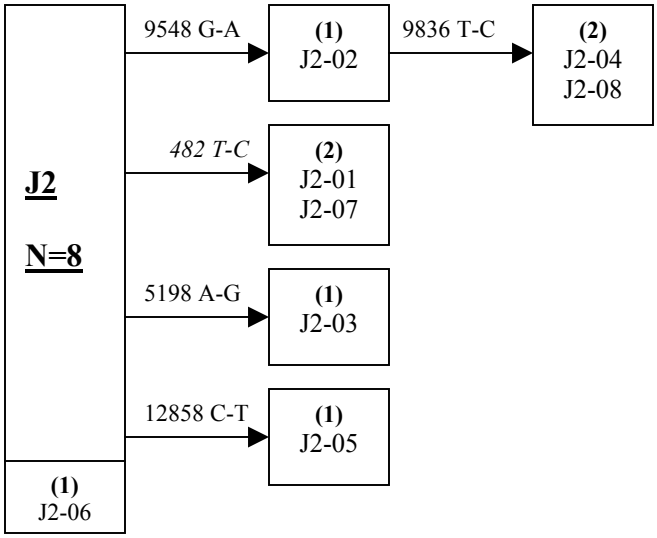
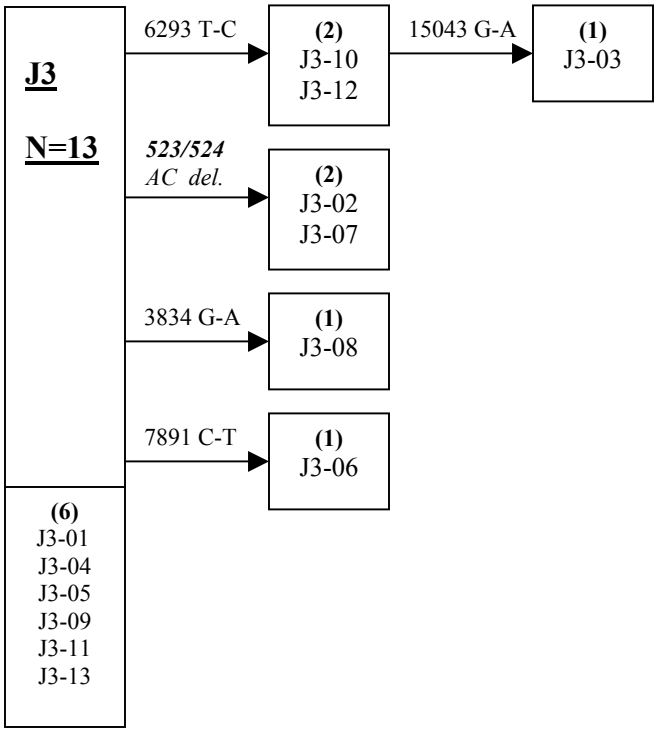


Figure 17. Schematic diagram of the resolution of the common Caucasian HV1/HV2 type J1 after applying eight neutral SNPs over the entire mtGenome. Nucleotide positions within the control region (outside of HV1/HV2) are in italics. The eight shared, neutral sites resolved the 15 sequences into eight haplotypes.

A.


J2
N=8
9548 G-A
(1)
J2-02
9836 T-C
(2)
J2-04
J2-08
482 T-C
(2)
J2-01
J2-07
5198 A-G
(1)
J2-03
12858 C-T
(1)
J2-05
(1)
J2-06

B.


J3
N=13
6293 T-C
(2)
J3-10
J3-12
15043 G-A
(1)
J3-03
523/524
AC del.
(2)
J3-02
J3-07
3834 G-A
(1)
J3-08
7891 C-T
(1)
J3-06
(6)
J3-01
J3-04
J3-05
J3-09
J3-11
J3-13

C.

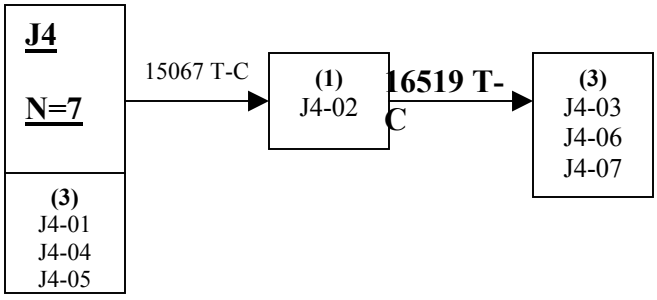

J4
N=7
15067 T-C
(1)
J4-02
16519 T-C
(3)
J4-03
J4-06
J4-07
(3)
J4-01
J4-04
J4-05

Figure 18. Schematic diagram of the resolution of three common Caucasian HV1/HV2 types belonging to haplogroup J after identifying shared, neutral SNPs over the mtGenome. Nucleotide positions within the control region (outside of HV1/HV2) are in italics. A.) Resolution of the eight J2 common type individuals into six haplotypes. B.) Resolution of the thirteen J3 common type individuals into six haplotypes. C.) Resolution of the seven J4 common type individuals into three haplotypes.

resolved the seven HV1/HV2 type J4 individuals into three haplotypes (Figure 18c). It is notable that position 16519T-C was found to be useful for resolving the J4 HV1/HV2 common type. We observed among the 15 common HV1/HV2 type J1 individuals and the 8 common HV1/HV2 type J2 individuals, position 16519 was essentially fixed for the rCRS base (T - with one exception in individual J1-09). All 12 common HV1/HV2 type J3 individuals were fixed for 16519C.

II. 2. 11. Phylogenetic Analysis of the Haplogroup T Cluster

We sequenced the mtGenomes of 39 individuals belonging to three common HV1/HV2 haplogroup T types (Table 1). Parsimony analysis was performed as described previously. A single MPT was generated having a tree length of 136 (consistency index 1.00, retention index 1.00). The MPT is shown in Figure 19. The T2 type individuals form a monophyletic clade defined by one coding region mutation (12633G-A) and the three HV1 polymorphisms (Figure 19). The T1 and T3 common HV1/HV2 types form a bifurcating cluster defined by the coding region mutations 11812A-G and 14233A-G. The HV1/HV2 type T3 individuals are grouped by four coding region mutations (2850T-C, 7022T-C, 13965T-C and 14687A-G). The HV1/HV2 type T1 cluster is grouped by coding region mutations at positions 930G-A and 5147G-A. Two of the T3 mtDNAs (T3-02 and T3-08) appear in the branch with T1 individuals (Figure 19). This is not surprising since the only difference between the common types T1 and T3 in HV1/HV2 is the polymorphism at position 16304T-C, a site previously characterized as being relatively fast (Wakeley 1993; Hasegawa *et al.* 1993; Excoffier

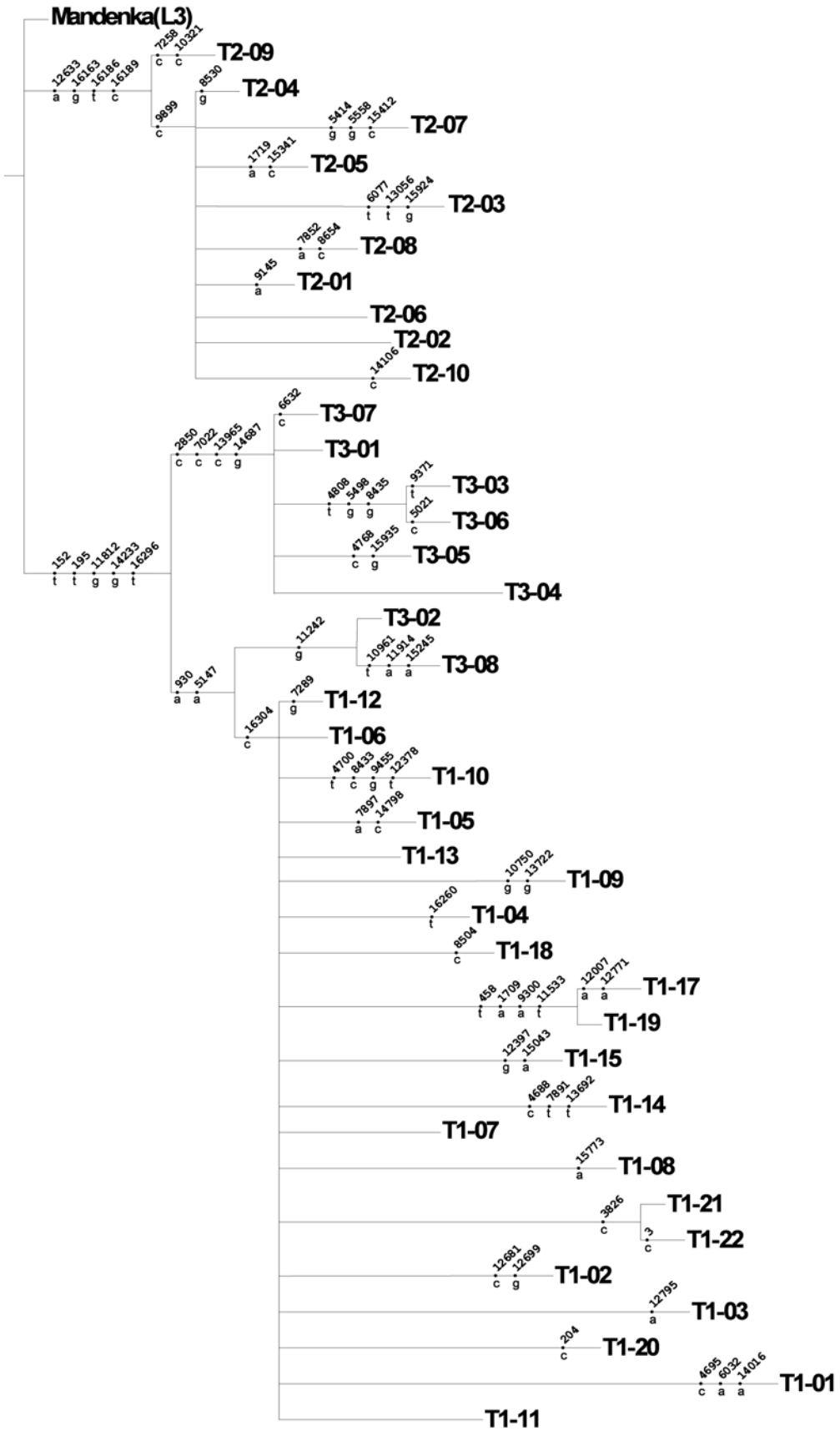

Mandenka(L3)
12633 a
16163 g
16186 t
16189 c
7258 c
10321 c
T2-09
8530 g
T2-04
9899 c
5414 g
5558 g
15412 c
T2-07
1719 a
15341 c
T2-05
6077 t
13056 t
15924 g
T2-03
7852 a
8654 c
T2-08
9145 a
T2-01
T2-06
T2-02
14106 c
T2-10
6632 c
T3-07
2850 c
7022 c
13965 c
14687 g
T3-01
4808 t
5498 g
8435 g
9371 t
T3-03
5021 c
T3-06
4768 c
15935 g
T3-05
T3-04
152 t
195 t
11812 g
14233 g
16296 t
T3-02
11242 g
10961 t
11914 a
15245 a
T3-08
930 a
5147 a
7289 g
T1-12
16304 c
T1-06
4700 t
8433 c
9455 g
12378 t
T1-10
7897 a
14798 c
T1-05
T1-13
10750 g
13722 g
T1-09
16260 t
T1-04
8504 c
T1-18
458 t
1709 a
9300 a
11533 t
12007 a
12771 a
T1-17
T1-19
12397 g
15043 a
T1-15
4688 c
7891 t
13692 t
T1-14
T1-07
15773 a
T1-08
3826 c
T1-21
3 c
T1-22
12681 c
12699 g
T1-02
12795 a
T1-03
204 c
T1-20
4695 c
6032 a
14016 a
T1-01
T1-11

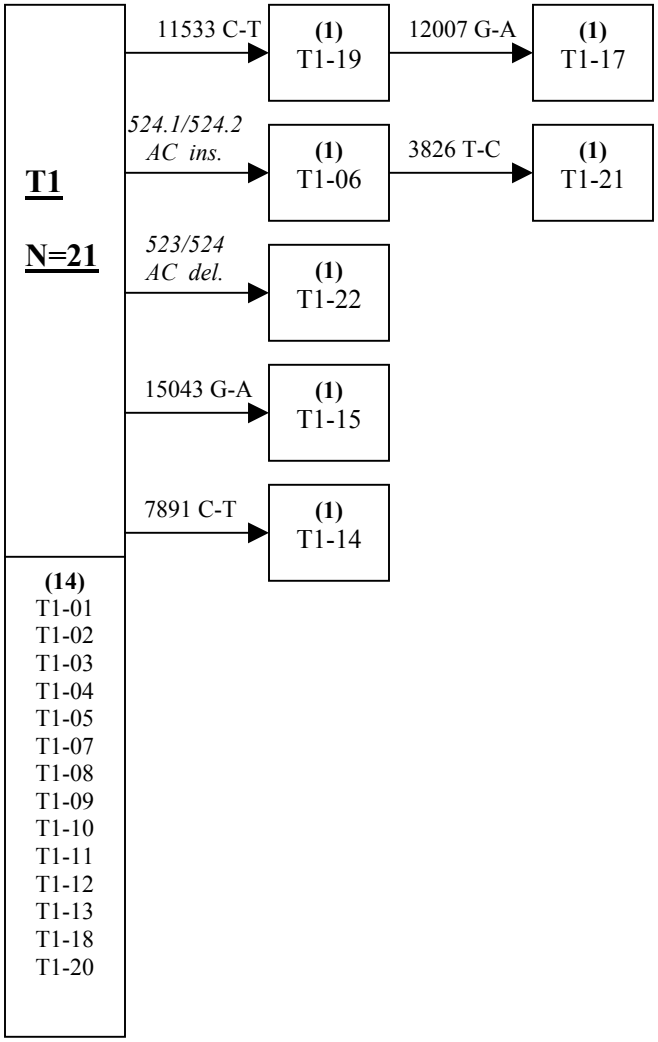
Figure 19. Phylogenetic Analysis of the Three Common HV1/HV2 type T Individuals Using Data from the mtGenome. The mtGenome of 39 non-related haplogroup T individuals belonging to three common HV1/HV2 types (T1, T2 and T3) were sequenced and analyzed using the phylogenetic program Winclada version 1.00.08 (Nixon 2002) running NONA (Goloboff 1999). The African haplogroup L3 sequence "Mandenka (Genbank ID – AF346995)" from Ingman et al (2000) was used as an outgroup sequence. One MPT was generated with a tree length of 136 (consistency index 1.00, retention index 1.00). Open circles that mark character changes on a branch are instances of homoplasy. The T2 individuals form a monophyletic clade basal to the other two common HV1/HV2 types.

and Yang 1999; Meyer *et al.* 1999; Malyarchuck *et al.* 2002; and Allard *et al.* 2002). Individuals matching exactly over the mtDNA coding region include: a set of six individuals from type T1 (T1-04, T1-06, T1-07, T1-11, T1-13 and T1-20), a set of two individuals from type T1 (T1-21 and T2-22), two individuals from type T2 (T2-02 and T2-06) and two individuals from type T3 (T3-01 and T3-04).

II. 2. 12. Neutral Sites to Resolve Haplogroup T Common HV1/HV2 Types

We identified seven shared, neutral sites that resolved the HV1/HV2 type T1 individuals (Figure 20a) into eight haplotypes. The 524.1/524.2 AC insertion outside of the HV1/HV2 region was useful for resolving one individual (T1-06) while the 523/524 AC deletion was useful for resolving another individual (T1-22). Sixteen of the twenty-one T1 individuals (76%) could not be resolved with the SNPs we have identified (Figure 20a). Only one shared, neutral site (9899T-C) was identified as variable for the HV1/HV2 type T2 individuals (shared among nine of the ten sequences - Figure 19). The common HV1/HV2 type T3 individuals could be resolved into two types (Figure 20b): sequences that were defined by the neutral site 5147G-A and sequences that were defined by the neutral site 7022T-C. Two additional neutral sites were identified for resolving T3 individuals in each branch. Overall, we identified four shared, neutral sites to resolve the eight T3 individuals into four haplotypes (Figure 20b). The often-useful control region polymorphism at position 16519 was found to be fixed for the non-rCRS variant in all 39 HV1/HV2 type T individuals.

A.


T1
N=21
11533 C-T
(1)
T1-19
12007 G-A
(1)
T1-17
524.1/524.2
AC ins.
(1)
T1-06
3826 T-C
(1)
T1-21
523/524
AC del.
(1)
T1-22
15043 G-A
(1)
T1-15
7891 C-T
(1)
T1-14
(14)
T1-01
T1-02
T1-03
T1-04
T1-05
T1-07
T1-08
T1-09
T1-10
T1-11
T1-12
T1-13
T1-18
T1-20

B.

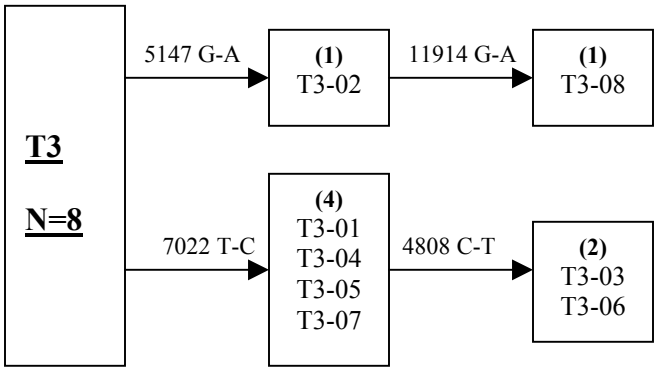

T3
N=8
5147 G-A
(1)
T3-02
11914 G-A
(1)
T3-08
7022 T-C
(4)
T3-01
T3-04
T3-05
T3-07
4808 C-T
(2)
T3-03
T3-06

Figure 20. Schematic diagram of the resolution of two HV1/HV2 common Caucasian types belonging to haplogroup T after the identification of neutral SNPs over the mtGenome. Nucleotide positions within the control region (outside of HV1/HV2) are in italics. A.) Resolution of the 21 common HV1/HV2 type T1 individuals into eight haplotypes using seven shared, neutral SNPs. B.) Resolution of the eight common HV1/HV2 type T3 individuals into four haplotypes using four shared, neutral SNPs. The T2 common HV1/HV2 type shared only one neutral SNP for resolving the sequences into two haplotypes.

II. 2. 13. Phylogenetic Analysis of the Haplogroup K Cluster

The Caucasian haplogroup K is a subgroup of the Haplogroup U cluster (Richards *et al.* 1998; Finnila *et al.* 2001). Haplogroup K occurs at a frequency of ~ 9% in North American and European populations (Torroni *et al.* 1996; Allard *et al.* 2002). Haplogroup K is defined by coding region polymorphisms at positions 1811A-G, 9055G-A, 12308A-G and 12372G-A and coding region polymorphisms at positions 16224T-C, 16311T-C, and 73G (Torroni *et al.* 1996; Macaulay *et al.* 1999a; Finnila *et al.* 2001; Allard *et al.*, 2002). We have identified three common HV1/HV2 types of haplogroup K (K1, K2 and K3) occurring at a frequency of 0.5% or greater (Table 1).

We sequenced 28 mtGenomes for the three common HV1/HV2 haplogroup K types (Table 1). Parsimony analysis was performed as described previously. Nineteen MPTs were generated with a tree length of 133 (consistency index 0.96, retention index 0.96). One representative MPT is shown in Figure 21. The common HV1/HV2 type K1 forms a paraphyletic group with 10 of the 14 individuals being grouped by the coding region mutations 4561T-C, 9716T-C and 10398G-A (Figure 21). The remaining four K1 individuals (K1-05, K1-07, K1-12 and K1-13) are on a bifurcated branch defined by 709A-G and 1189T-C polymorphisms (Figure 21). These four K1 individuals form a clade distinct from the K2 and K3 individuals by mutations at 9093A-G and 11377G-A (Figure 21). Since this group of four K1 individuals are closely related to the K2/K3 cluster, these individuals "appear" as the HV1/HV2 type K1 due to a T-C mutation at 152, a previously described "fast" site in HV2. Four K1 individuals (K1-01, K1-02, K1-06 and K1-10) match exactly in their coding region, while two additional sets of two K1

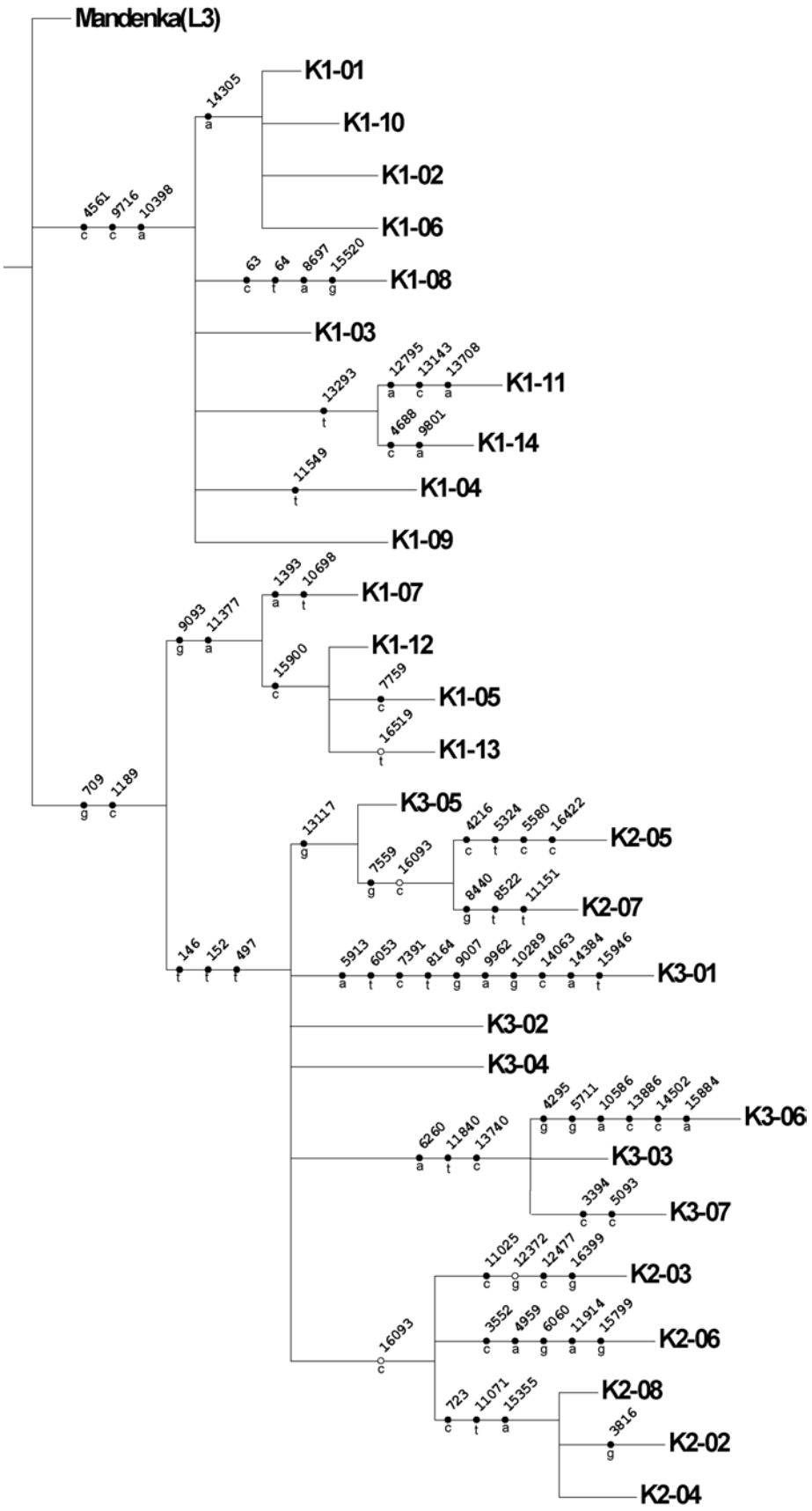

Mandenka(L3)
K1-01
K1-10
K1-02
K1-06
K1-08
K1-03
K1-11
K1-14
K1-04
K1-09
K1-07
K1-12
K1-05
K1-13
K3-05
K2-05
K2-07
K3-01
K3-02
K3-04
K3-06
K3-03
K3-07
K2-03
K2-06
K2-08
K2-02
K2-04

Figure 21. Phylogenetic Analysis of the Three Haplogroup K common HV1/HV2 type Individuals Using Data from the mtGenome. The mtGenome of 28 unrelated haplogroup K individuals belonging to three common HV1/HV2 types (K1, K2 and K3) were sequenced and analyzed using the phylogenetic program Winclada version 1.00.08 (Nixon 2002) running NONA (Goloboff 1999). The African haplogroup L3 sequence "Mandenka (Genbank ID – AF346995)" from Ingman et al (2000) was used as an outgroup sequence. One representative MPT with a tree length of 133 (consistency index 0.96, retention index 0.96) is shown. Open circles that mark character changes on a branch are instances of homoplasy.

mtDNAs also match exactly (K1-03 and K1-09; and K1-12 and K1-13).

The remaining K2 and K3 mtDNAs form a clade defined by the control region mutations 146T, 152T and 497C-T (Figure 21). Five of the seven K2 sequences form a cluster defined by the transition at 16093T-C (Figure 21). The two additional K2 sequences (K2-05 and K2-07) form a cluster with a K3 individual (K3-05) defined by 13117A-G (Figure 21). The nucleotide position 16093, a polymorphism associated with our common HV1/HV2 type K2, has been demonstrated to be a relatively fast site (Parsons *et al.* 1997; Meyer *et al.* 1999, Allard *et al.* 2002; and Malyarchuk *et al.* 2002). So, it is not surprising to observe two apparent K2 individuals within the K3-defined branch. Two K2 sequences (K2-04 and K2-08) and two K3 sequences (K3-02 and K3-04) match exactly over the entire mtGenome.

II. 2. 14. Neutral Sites to Resolve Haplogroup K Common HV1/HV2 Types

Two shared, neutral coding region SNPs were identified for grouping the K1 individuals into two clusters: one cluster grouped by 9716T-C (ten K1 individuals), and the other cluster grouped by 11377G-A (four sequences) (Figure 22a). The 9716 cluster could further be resolved by five additional shared, neutral SNPs (Figure 22a). The 11377 cluster was resolved by the control region polymorphism at 16519 and the 523/524 AC deletion (Figure 22a). In all, the fourteen K1 sequences were resolved into eight haplotypes using nine shared, neutral SNPs.

The common HV1/HV2 type K2 individuals were resolved into four haplotypes by using three neutral sites. Two SNPs (11914G-A and 15355G-A) were identified in the

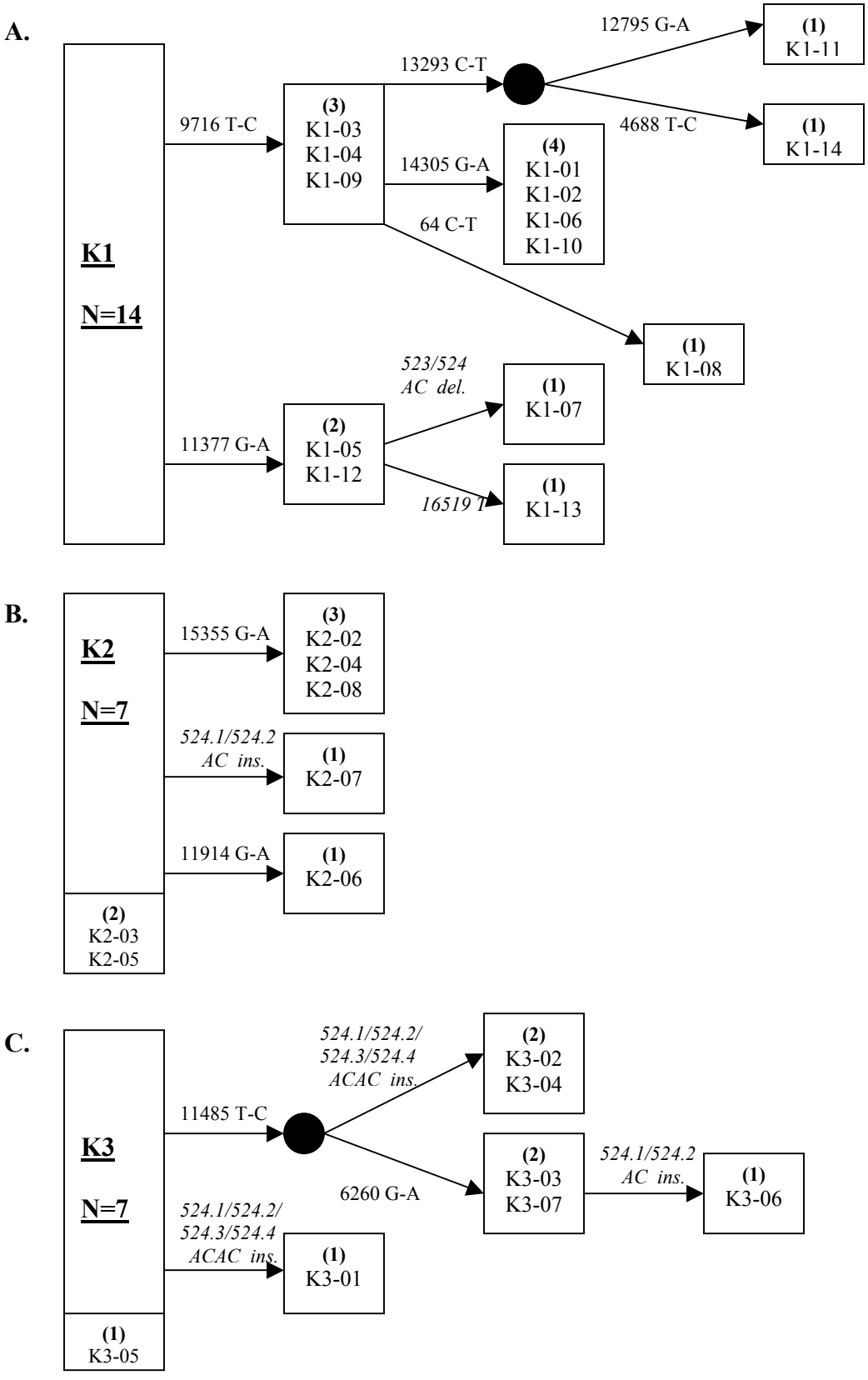

A.
K1
N=14
9716 T-C
11377 G-A
(3)
K1-03
K1-04
K1-09
(2)
K1-05
K1-12
13293 C-T
14305 G-A
64 C-T
12795 G-A
4688 T-C
(4)
K1-01
K1-02
K1-06
K1-10
(1)
K1-11
(1)
K1-14
(1)
K1-08
523/524
AC del.
(1)
K1-07
16519 T
(1)
K1-13
B.
K2
N=7
15355 G-A
524.1/524.2
AC ins.
11914 G-A
(3)
K2-02
K2-04
K2-08
(1)
K2-07
(1)
K2-06
(2)
K2-03
K2-05
C.
K3
N=7
11485 T-C
524.1/524.2/
524.3/524.4
ACAC ins.
6260 G-A
524.1/524.2/
524.3/524.4
ACAC ins.
(2)
K3-02
K3-04
(2)
K3-03
K3-07
524.1/524.2
AC ins.
(1)
K3-06
(1)
K3-01
(1)
K3-05

Figure 22. Schematic diagram of the resolution of three HV1/HV2 Common Caucasian types belonging to Haplogroup K after applying neutral SNPs over the mtGenome. Nucleotide positions within the control region (outside of HV1/HV2) are in italics. A.) Resolution of the fourteen common HV1/HV2 type K1 individuals into eight haplotypes using nine shared, neutral SNPs. B.) Resolution of the seven common HV1/HV2 type K2 individuals into four haplotypes using three shared, neutral SNPs. C.) Resolution of the seven common HV1/HV2 type K3 individuals into five haplotypes using four neutral SNPs.

coding region, while the 524.1/524.2 AC insertion was seen in one K2 individual (K2-07 - Figure 22b). The seven K3 individuals were resolved into five haplotypes by four neutral sites (Figure 22c). One individual (K3-06) was resolved by the 524.1/524.2 AC insertion (Figure 22c). Three of the remaining sequences (K3-01, K3-02 and K3-04) share the double insertion of the AC repeating dinucleotide (524.1/524.2/524.3/524.4 ACAC insertion). The observation of this double insertion at such a high frequency of K3 individuals (43%) was interesting since the frequency of this particular insertion in a random sample of 68 unrelated individuals was found to be 1% (Bodenteich *et al.* 1992). Finnila *et al.* (2001) observed the same double AC repeat insertion in 1 of 192 Finnish samples (0.5% - in one haplogroup H individual). Although K3-02 and K3-04 match exactly over their entire mtGenome, the K3-01 individual has ten additional coding region polymorphisms (not counting the diagnostic mutations associated with haplogroup K) suggesting that these sequences do not all share a common ancestor.

II. 3. Discussion

II. 3. 1. Summary of the Variation in the 241 Individuals

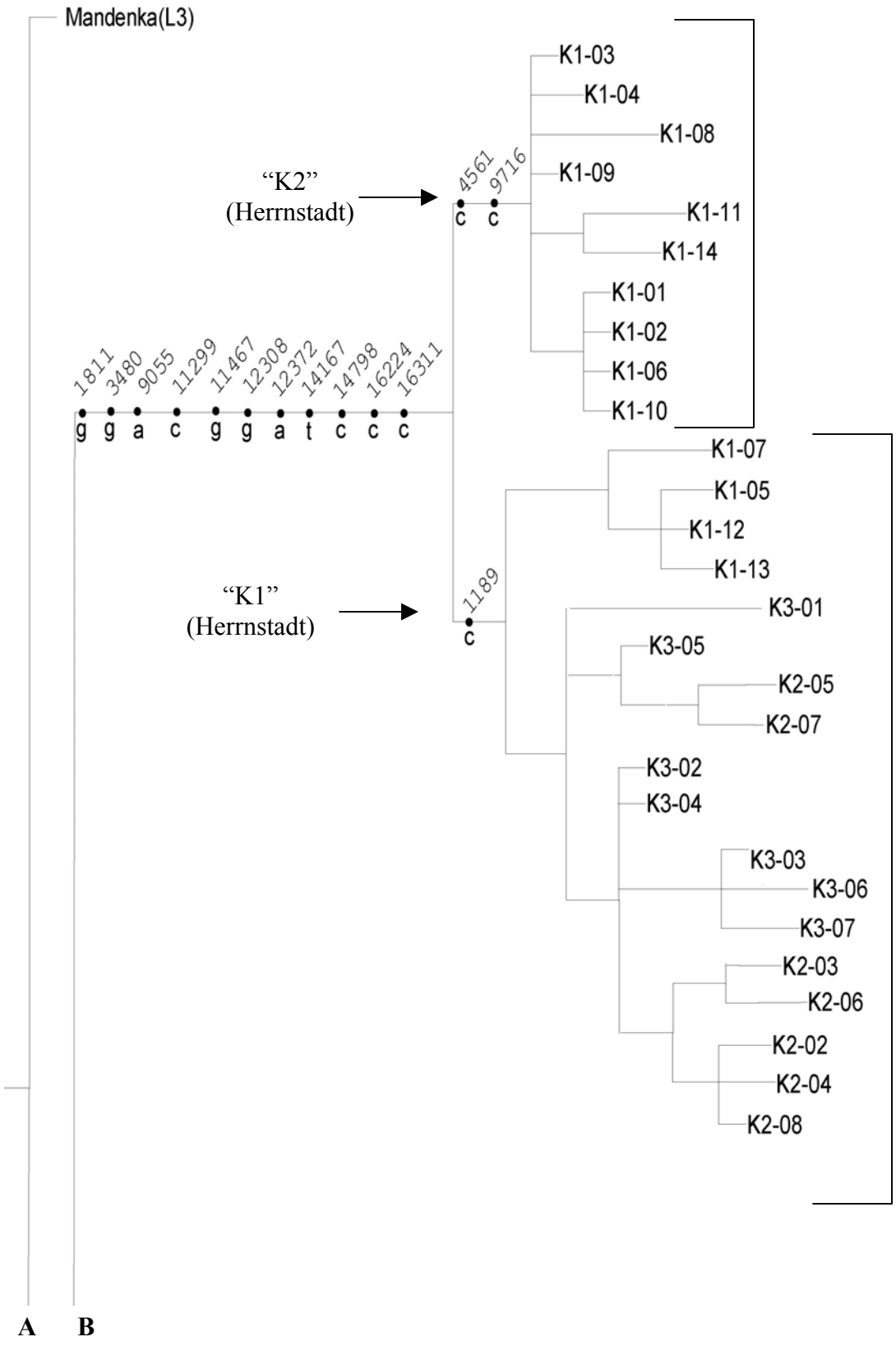
The 241 Caucasian individuals selected in this study, representing 18 common HV1/HV2 types, could be resolved into 209 haplotypes when the entire mtGenomes were sequenced. Only 32 of the 241 (13%) matched one or more individuals over the entire mtGenome. The average number of differences between individuals that match in HV1/HV2 was 6.2, with the greatest number of differences occurring between two K1 common type individuals: 22. The common HV1/HV2 types showed variable degrees of diversity at the entire mtGenome level compared to one another. For example, none of the 12 common HV1/HV2 type H5 individuals matched one another, with an average number of 6 differences and a maximum number of 12 differences between any two H5 individuals. At the other end of the spectrum, one-half of the common HV1/HV2 type H4 individuals (n=8) matched one another in the mtGenome, with an average number of ~1 differences and a maximum number of 2 differences between any two H4 individuals. It is possible that these disparities were due to sampling effects in some cases, but they more likely reflect real differences in the time since the sequences, grouped according to HV1/HV2, have diverged.

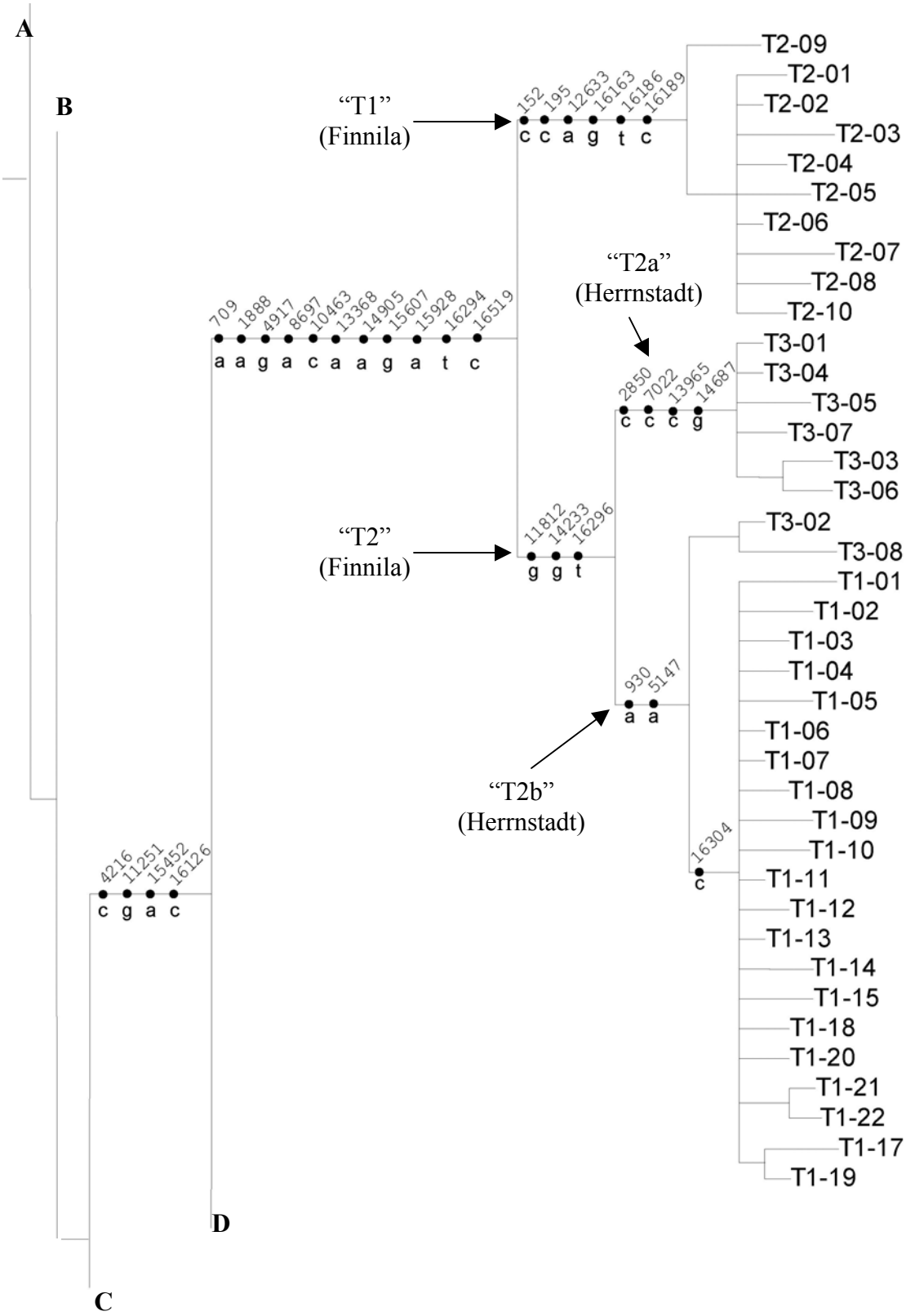
Homoplasies were found to be quite common in the HV1/HV2 region. The high amount of homoplasy in the HV1/HV2 region sometimes caused individuals to be grouped to one common HV1/HV2 type, when in fact, they are not closely related to other individuals within the group. For example, two individuals were misidentified as belonging to the common HV1/HV2 type H1, until the mtGenomes of these two

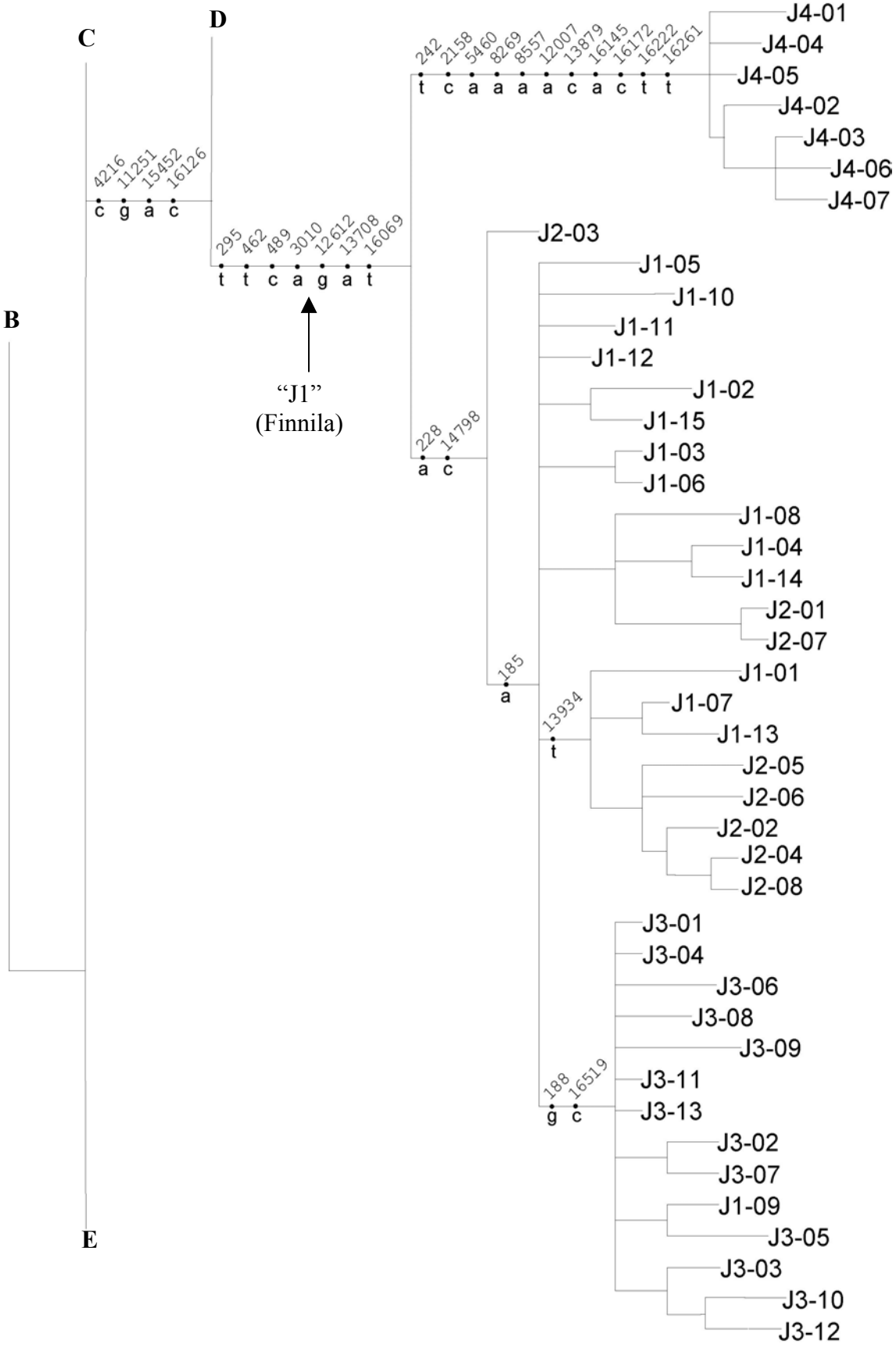
individuals were sequenced and seen to be consistent with haplogroup V. These individuals appeared to be the H1 common type because of a reversion at the 16298 polymorphism (Figure 4). Some of the common HV1/HV2 types form monophyletic clades (e.g. the common HV1/HV2 type, J4 – Figure 16) while other common HV1/HV2 types were identified to be polyphyletic or paraphyletic (for example, H1's and H2's – Figure 4).

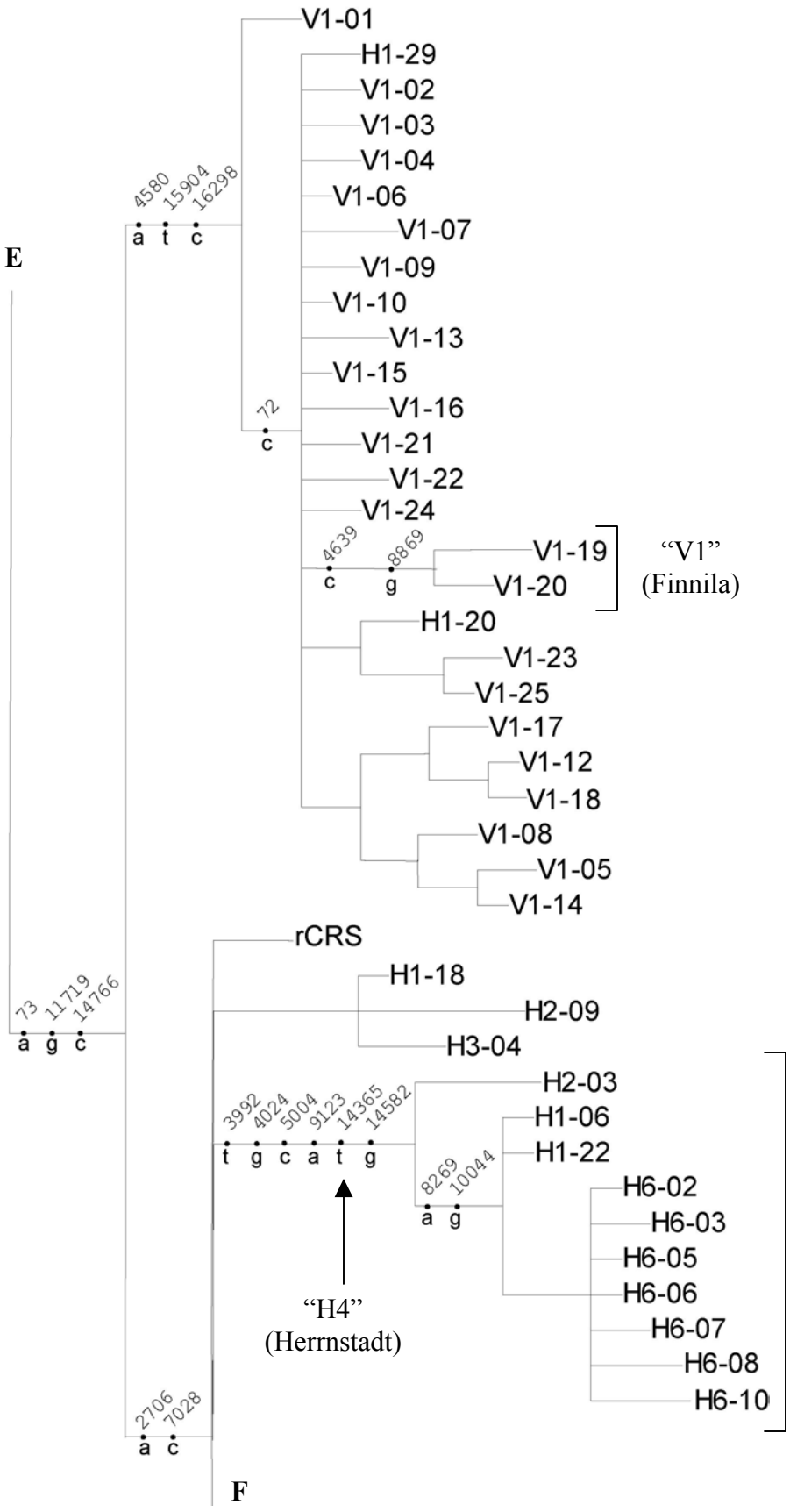
A total evidence phylogenetic tree (Figure 23) was constructed using parsimony with diagnostic polymorphism from recognized mtDNA haplogroups and sub-haplogroups mapped onto the tree. The total evidence tree (Figure 23) substantiates our rather arbitrary method of nomenclature for the common HV1/HV2 types. For example, we observed the established haplogroup H subgroup, H1 (defined by the 3010 A polymorphism; Finnila *et al*., 2001), among six of the seven common HV1/HV2 haplogroup H types (Figure 23). The polyphyletic and paraphyletic grouping of common HV1/HV2 types resulting from homoplasies in HV1 and HV2 may be responsible for varying amount of mtGenome diversity observed among the common HV1/HV2 types.

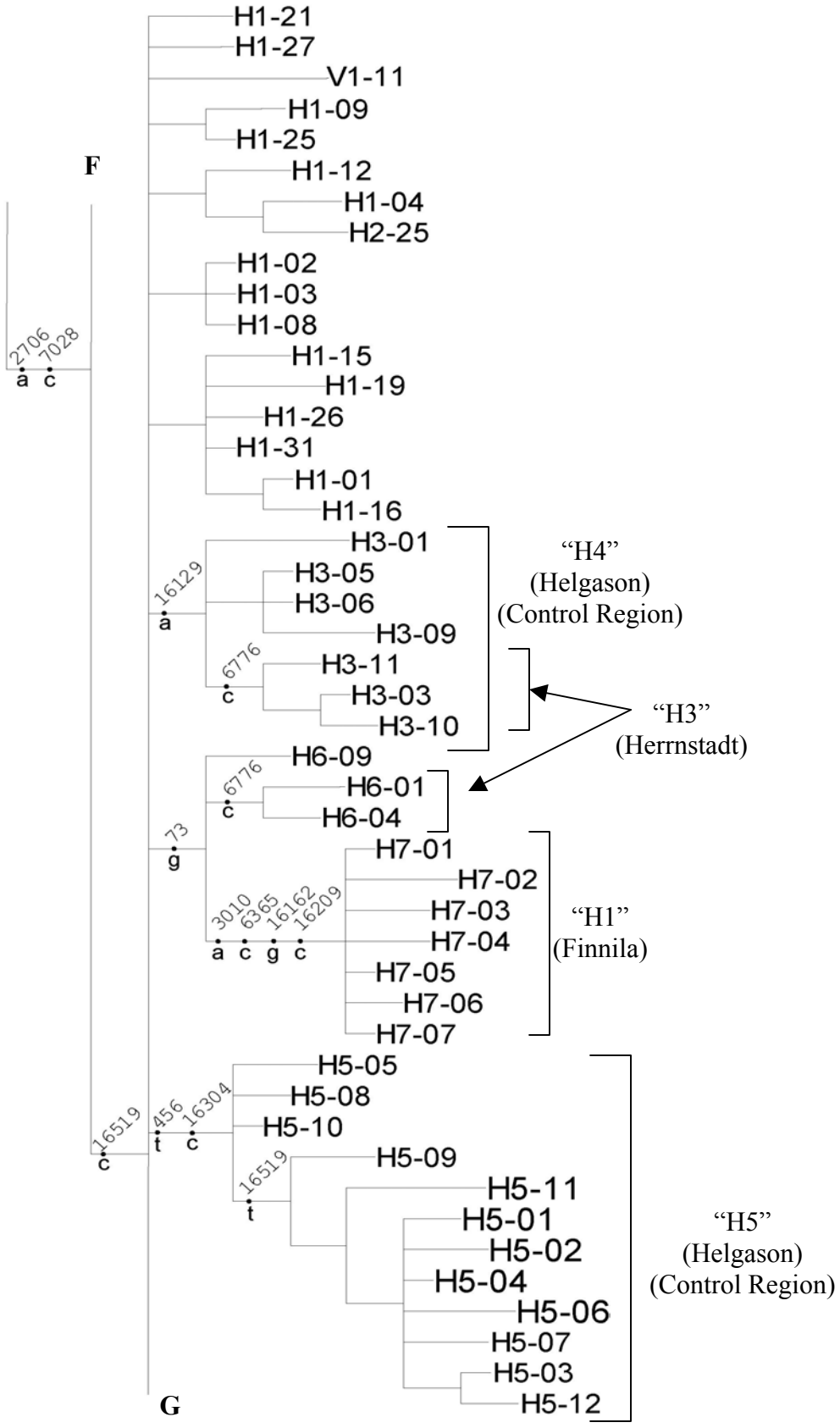
The average number of sequence differences over the entire mtGenome among individuals belonging to the common HV1/HV2 types did not vary considerably among the haplogroups: haplogroup H (6.9); haplogroup J (5.6), haplogroup K (7.2), haplogroup T (3.7), and haplogroup V (6.2). However, the variation within a particular common type was sometimes observed to vary widely (e.g. the common types H4 and H5, discussed above).


Mandenka(L3)
K1-03
K1-04
K1-08
"K2"
(Herrnstadt)
4561
9716
K1-09
c c
K1-11
K1-14
K1-01
K1-02
K1-06
K1-10
1811 3480 9055 11299 11467 12308 12372 14167 14798 16224 16311
g g a c g g a t c c c
K1-07
K1-05
K1-12
K1-13
"K1"
(Herrnstadt)
1189
c
K3-01
K3-05
K2-05
K2-07
K3-02
K3-04
K3-03
K3-06
K3-07
K2-03
K2-06
K2-02
K2-04
K2-08
A
B


A
B
C
D
"T1"
(Finnila)
152 195 12633 16163 16186 16189
c c a g t c
"T2"
(Finnila)
11812 14233 16296
g g t
"T2a"
(Herrnstadt)
2850 7022 13965 14687
c c c g
"T2b"
(Herrnstadt)
930 5147
a a
709 1888 4917 8697 10463 13368 14905 15607 15928 16294 16519
a a g a c a a g a t c
4216 11251 15452 16126
c g a c
16304
c
T2-09
T2-01
T2-02
T2-03
T2-04
T2-05
T2-06
T2-07
T2-08
T2-10
T3-01
T3-04
T3-05
T3-07
T3-03
T3-06
T3-02
T3-08
T1-01
T1-02
T1-03
T1-04
T1-05
T1-06
T1-07
T1-08
T1-09
T1-10
T1-11
T1-12
T1-13
T1-14
T1-15
T1-18
T1-20
T1-21
T1-22
T1-17
T1-19


B
C
D
E
4216 11251 15452 16126
c g a c
295 462 489 3010 12612 13708 16069
t t c a g a t
242 2158 5460 8269 8557 12007 13879 16145 16172 16222 16261
t c a a a a c a c t t
228 14798
a c
185
a
13934
t
188 16519
g c
"J1"
(Finnila)
J4-01
J4-04
J4-05
J4-02
J4-03
J4-06
J4-07
J2-03
J1-05
J1-10
J1-11
J1-12
J1-02
J1-15
J1-03
J1-06
J1-08
J1-04
J1-14
J2-01
J2-07
J1-01
J1-07
J1-13
J2-05
J2-06
J2-02
J2-04
J2-08
J3-01
J3-04
J3-06
J3-08
J3-09
J3-11
J3-13
J3-02
J3-07
J1-09
J3-05
J3-03
J3-10
J3-12


V1-01
H1-29
V1-02
V1-03
V1-04
V1-06
V1-07
V1-09
V1-10
V1-13
V1-15
V1-16
V1-21
V1-22
V1-24
V1-19
V1-20
"V1"
(Finnila)
H1-20
V1-23
V1-25
V1-17
V1-12
V1-18
V1-08
V1-05
V1-14
rCRS
H1-18
H2-09
H3-04
H2-03
H1-06
H1-22
H6-02
H6-03
H6-05
H6-06
H6-07
H6-08
H6-10
"H4"
(Herrnstadt)
E
F
4580 15904 16298
a t c
72
c
4639 8869
c g
73 11719 14766
a g c
3992 4024 5004 9123 14365 14582
t g c a t g
8269 10044
a g
2706 7028
a c


F
H1-21
H1-27
V1-11
H1-09
H1-25
H1-12
H1-04
H2-25
H1-02
H1-03
H1-08
H1-15
H1-19
H1-26
H1-31
H1-01
H1-16
H3-01
H3-05
H3-06
H3-09
H3-11
H3-03
H3-10
H6-09
H6-01
H6-04
H7-01
H7-02
H7-03
H7-04
H7-05
H7-06
H7-07
H5-05
H5-08
H5-10
H5-09
H5-11
H5-01
H5-02
H5-04
H5-06
H5-07
H5-03
H5-12
2706 a
7028 c
16129 a
6776 c
73 g
3010 a
6365 c
16162 g
16209 c
16519 c
456 t
16304 c
16519 t
"H4" (Helgason) (Control Region)
"H3" (Herrnstadt)
"H1" (Finnila)
"H5" (Helgason) (Control Region)
G

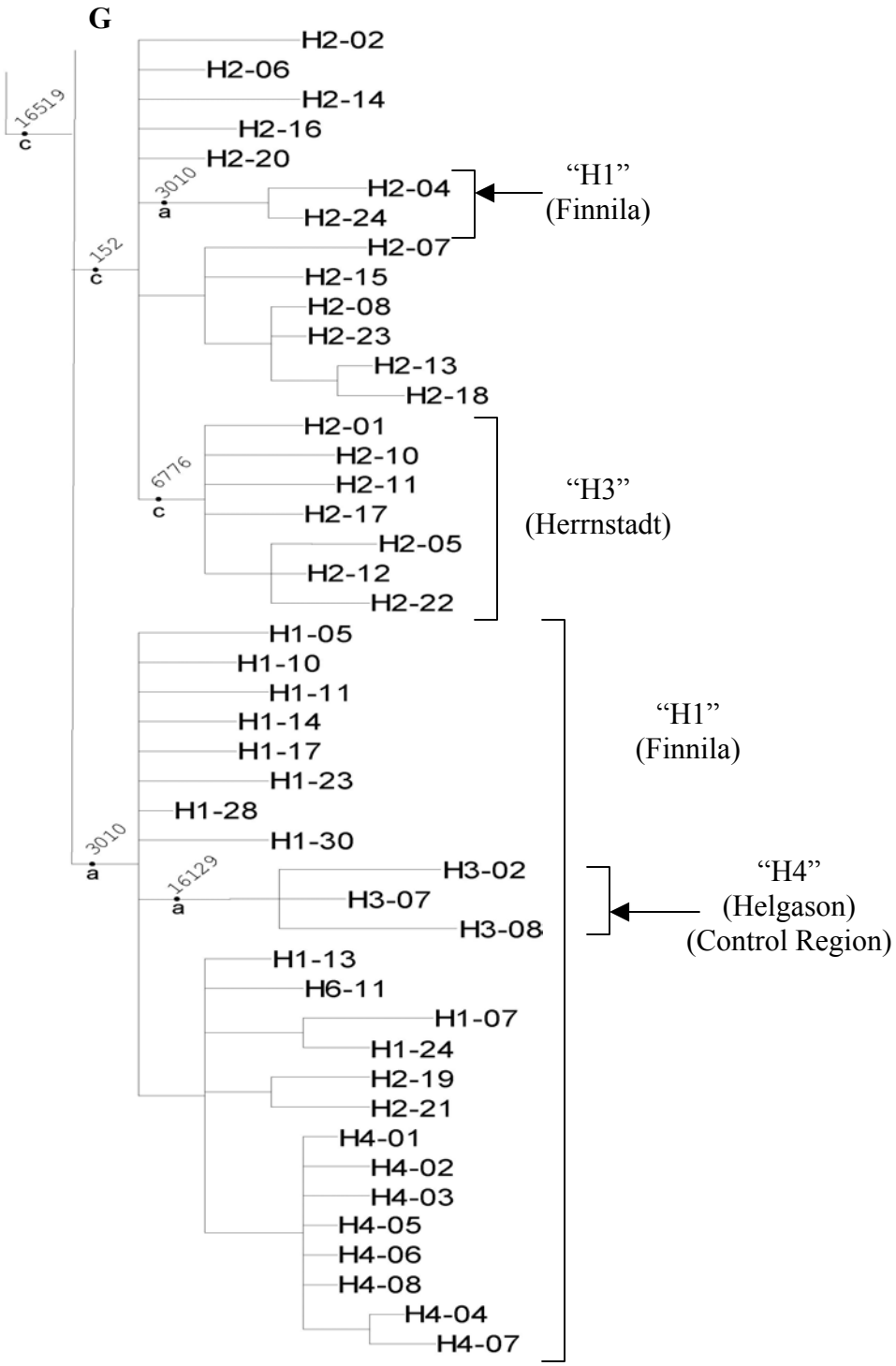

G
H2-02
H2-06
H2-14
H2-16
H2-20
H2-04
H2-24
H2-07
H2-15
H2-08
H2-23
H2-13
H2-18
H2-01
H2-10
H2-11
H2-17
H2-05
H2-12
H2-22
H1-05
H1-10
H1-11
H1-14
H1-17
H1-23
H1-28
H1-30
H3-02
H3-07
H3-08
H1-13
H6-11
H1-07
H1-24
H2-19
H2-21
H4-01
H4-02
H4-03
H4-05
H4-06
H4-08
H4-04
H4-07
16519
c
3010
a
152
c
6776
c
3010
a
16129
a
"H1"
(Finnila)
"H3"
(Herrnstadt)
"H1"
(Finnila)
"H4"
(Helgason)
(Control Region)

Figure 23. Total Evidence Phylogenetic Tree of all 241 mtGenomes. The mtGenomes of 241 individuals having common HV1/HV2 types were sequenced. The African haplogroup L3 sequence "Mandenka (Genbank ID – AF346995)" from Ingman *et al*., (2000) was used as an outgroup sequence. One most parsimonious tree (of 126 total) is shown (tree length of 646, consistency index of 0.86, retention index of 0.96). Trees were generated using the software program Winclada version 1.00.88 (Nixon 2002) running NONA (Goloboff 1999). Diagnostic polymorphisms associated with established mtDNA haplogroups (or sub-haplogroups) are mapped onto the tree with the corresponding published reference. The tree is connected from one page to another by capital letters.

An evaluation of the amount of variation observed within the various genes/regions of the mtGenome was performed (Table 3). The least variable regions of the mtGenome were found to be the two ribosomal RNA genes (Table 3). Outside of the control region, the most highly variable regions were the short, interspersed non-coding spacer regions of the mtGenome (Table 3), with variation observed at nearly 7% of these sites. The most variable protein-coding genes were the ATP Synthase genes 8 and 6, with variation observed at nearly 5% of the sites in the ATP Synthase 8 gene.

That the ATP Synthase genes had the highest sequence variation among the coding region genes agrees with other findings (Ingman and Gyllensten, 2001; Mishmar *et al.*, 2003). These results are somewhat surprising since the ATP6 gene has been demonstrated to be highly conserved among distantly related species (Wallace *et al.* 1987; Saccone *et al.* 2000).

The type of variation within the protein-coding genes was evaluated by tabulating the number of synonymous and non-synonymous variants within each gene (Table 4). A high proportion of the changes that occur in the ATP6 and ATP8 genes were observed to be nonsynonymous, resulting in amino acid substitutions (Table 4). The high proportion of polymorphic variation in the ATP6 gene in humans (compared to related species) and the increased amounts of nonsynonymous changes in the ATP6 and ATP8 genes implies natural selection may have acted upon the variation in these regions at some point in human evolution.

The non-coding sites outside of HV1/HV2 within the control region did not show a greatly higher proportion of variable sites than the coding region. This proportion, 4.1%, was greater than almost all of the protein-coding genes, but less than twice the

Gene/Region	Length of Region	Number of Discriminatory Sites	%[c]
CR Outside of HV1/HV2	513	21	4.1%
All tRNA's Combined	1507	35	2.3%
NC Region Outside of CR	91	6	6.6%
12s rRNA	954	13	1.4%
16s rRNA	1559	16	1.0%
NADH Dehydrogenase 1	957	22	2.3%
NADH Dehydrogenase 2	1044	37	3.5%
Cytochrome C Oxidase I	1542	38	2.5%
Cytochrome C Oxidase II	684	18	2.6%
ATP Synthase 8	207	10	4.8%
ATP Synthase 6	681	25	3.7%
Cytochrome C Oxidase III	783	18	2.3%
NADH Dehydrogenase 3	348	8	2.3%
NADH Dehydrogenase 4L	297	6	2.0%
NADH Dehydrogenase 4	1380	33	2.5%
NADH Dehydrogenase 5	1812	56	3.1%
NADH Dehydrogenase 6	525	15	2.9%
Cytochrome b	1137	35	3.1%
TOTAL	**15959[a]**	**410[b]**	**2.6%[d]**

Table 3. Summary of the Number of Discriminatory Sites Among Common HV1/HV2 Types in 241 Caucasian mtGenome, by Gene/Region. Discriminatory sites refers to sites that vary within one or more common HV1/HV2 types. a) The numbers for length of region do not sum to the value listed in the total because of overlap in some of the genes. b) The numbers for discriminatory sites do not sum to the value in the total since two of the discriminatory sites occur in overlapping regions. c) Percent (%) of discriminatory sites refers to the number of sites that show discriminatory variation divided by the length of the gene/region x 100. d) The total percent refers to the total number of sites that show variation outside the HV1/HV2 regions divided by the total number of bases in the mtGenome outside of HV1/HV2.

Gene	Length	Synonymous	Nonsynonymous	Total	% NonSyn.
ND1	956	14	8	22	36.4%
ND2	1,042	25	11	36	30.6%
CO1	1,542	29	9	38	23.7%
CO2	684	14	4	18	22.2%
ATP8	207	3	5	8	62.5%
ATP6	681	7	20	27	74.1%
CO3	784	14	4	18	22.2%
ND3	346	5	2	7	28.6%
ND4L	297	5	1	6	16.7%
ND4	1,378	30	7	37	18.9%
ND5	1,812	39	15	54	27.8%
ND6	525	8	7	15	46.7%
CYB	1,141	23	15	38	39.5%
Total	11,341	216	108	324	33.1%

Table 4. Synonymous and Nonsynonymous Mutations Among Common HV1/HV2 Types in 241 Caucasian mtDNA sequences, by Genes. The gene region refers to the 13 protein coding genes (ND – NADH dehydrogenase; CO – Cytochrome c oxidase; ATP – ATP synthase; CYB – Cytochrome b). The length of the gene is given in nucleotides. Both synonymous and nonsynonymous mutations are noted as differences in from the rCRS. The numbers do not include diagnostic polymorphisms associated with haplogroup designation. The percentage of nonsynonymous mutations was calculated by dividing the number of nonsynonymous mutations by the total number of mutations in the gene x 100.

average of 2.5% for the entire mtGenome (outside of HV1/HV2). This may be a consequence of selective constraints on variation in the control region due to regulatory functions (e.g. Saccone *et al.*, 1991).

The control region outside of the HV1/HV2 region did, however, contain sites that were highly variable among our common HV1/HV2 types. For example, the most variable SNP observed among our 18 common types was 16519, varying in nine common HV1/HV2 types. An additional variable element that proved useful for discrimination was the AC indel occurring between nucleotides 515-524 (Bodenteich *et al.*, 1992). Indels within this region were seen in 11 of the 18 common HV1/HV2 types. The AC indel, although not a true SNP, and 16519 are examples of mutational hotspots that were observed to vary widely across a number of haplotypic backgrounds. Both of these sites will be extremely informative to assay for increased forensic discrimination.

II. 3. 2. Forensic Discrimination and SNP Analysis

The specific aim of this chapter was to identify shared, neutral SNPs to increase the power of discrimination for common Caucasian HV1/HV2 types. The 18 common HV1/HV2 types represents approximately 21% of the Caucasian population, with the most common HV1/HV2 type observed in about 7% of the population (Table 1). The greatest limitation of current forensic mtDNA testing lies with the small number of common types where the power of discrimination is low. Cases involving the re-association of skeletal remains, such as those at AFDIL, typically require mtDNA testing for forensic identification. The forensic scientist is often challenged with generating an HV1/HV2 mtDNA profile from limited quantities of highly degraded DNA only to be stymied by one of the handful of common types that occur in a cumulative 21% of the Caucasian population. In cases of mass disasters, such as the World Trade Center disaster of September 11, 2001, mtDNA is being used to assist in the identification of highly degraded remains. However, potentially a large number of the individuals recovered from this disaster will share one of the few common Caucasian HV1/HV2 types, making the identification of these remains inconclusive.

This dissertation identified a significant amount of variation in the coding and non-coding regions outside of the HV1/HV2 region of the 241 individuals that were sequenced. Sequencing the entire mtGenome of 241 individuals among the 18 common HV1/HV2 types revealed 59 SNP sites that met our criteria as useful forensic markers (Table 5). Eight of these SNPs were found outside of HV1/HV2 in the control region; one SNP was found in an intergenic, coding region spacer; forty-eight SNPs were silent,

synonymous changes in the protein-coding genes; and one SNP was within the 16S rRNA. The SNP site in the 16S rRNA, 3010, is an exception to our conservative criteria for SNP selection, where we avoided sites that reside in rRNA or tRNA genes. This SNP was determined to be a valuable site for inclusion in our assay based upon: a) the 3010G/A polymorphism is a well characterized mutation, published widely in the literature, and not suspected to be associated with genetic diseases (Mehta *et al.*, 1989; Finnila *et al.*, 2001); and b) this site was very informative for forensic discrimination, especially among the common type of haplogroup H. The wide variation of 3010 in the population assured us that this polymorphism is likely governed by Kimura's Neutral Theory of Molecular Evolution (Kimura, 1983).

The 59 forensic SNP sites selected (Table 5) have been arranged into eight multiplex "panels" (A-H) that target common HV1/HV2 types. The arrangement of the SNPs was made to provide general utility for the forensic scientist. Multiplex assays can be challenging to optimize, require months to develop, and use platforms not necessarily available to many laboratories. However, for forensic purposes, multiplex assays can target specific sites while at the same time preserving a limited volume of extract. As a "middle ground" for practical SNP multiplex panel development, we batched SNPs that were useful for resolving particular common types and would be readily available for use in the laboratory. Each panel contains seven to eleven sites, an ideal number of markers to combine for multiplexing, based upon our experience with an allele-specific primer extension (ASPE) assay using SNaPShot™ (Vallone *et al.*, in press). We have found this system to be advantageous for use in the forensic laboratory since the assay is sensitive, able to detect mixtures/heteroplasmy, gives robust results using degraded materials, and

A	B	C	D	E	F	G	H
477	477	72	482	4808	64	3826	64
3010	3010	513	5198	5147	4745	3834	4688
4580	3915	4580	6260	9380	10211	4688	11377
4793	5004	5250	9548	9899	10394	6293	12795
5004	6776	11719	9635	11914	10685	7891	13293
7028	8592	12438	11485	15067	11377	11533	14305
7202	10394	12810	11914	16519	14470	12007	16519
10211	10754	14770	15355		14560	12795	
12858	11864	15833	15884		16390	15043	
14470	15340	15884	16368		14869	16390	
16519	16519	16519				16519	
H1	H2 H3 H6	V1 H5	J1 J2 K2 K3	J4 T2 T3 H4	V1 H1 H2 H3	J1 J3 T1	K1

Table 5. The Eight Multiplex Panels of SNPs to Resolve Common Caucasian HV1/HV2 Types. Sites within each panel were selected to provide the maximal resolution for the common HV1/HV2 types (listed at the bottom of the panel)

can be operated on "standard" instrumentation (e.g. the ABI 377 or ABI 3100).

The multiplex panels were developed to complement the current mtDNA typing of HV1/HV2. Once a common HV1/HV2 common type has been identified by sequencing, for example, H1, the forensic scientist can then prepare a single amplification using Multiplex A (Table 5) for the greatest opportunity for discrimination. However, not all the SNPs identified for the resolution of H1 could be placed into multiplex A. Therefore, if multiplex A failed to provide resolution, the forensic scientist can conduct a second amplification using multiplex F to target additional SNP sites for H1 discrimination (Table 5). Some sites were included in multiple panels to provide redundancy for discrimination. For example, the highly discriminatory polymorphism 16519 was included in six of the eight panels (Table 5). This way, for any given common HV1/HV2 type, a single multiplex can be utilized for maximum resolution if only enough extract remains for a single amplification.

A summary of the utility of the eight multiplex panels is found in Table 6. Originally, the 241 individuals sequenced belonged to 18 common types that, together, comprise about 21% of the Caucasian population. Applying all eight multiplex panels (Table 5) to these 241 sequences, were able to resolve the individuals into 106 haplotypes, 56 of which were unique. This represents a nearly 6-fold improvement over the initial 18 common HV1/HV2 types (Table 1).

The eight multiplex panels (Table 5) developed did not include the extremely informative 515-524 AC indel. It is possible to type the number of repeats in this region using any number of platforms, including sequencing a short fragment around this element in the control region. Including the AC indel polymorphism with the eight

multiplex panels would resolve the 241 individuals into 113 different types, 66 of which were unique (Table 6).

Due to limitations in sample availability, we were unable to perform mtGenome sequencing on four of the common HV1/HV2 types which occur at 0.5% or greater in the Caucasian population. However, one could predict that at least some of the SNPs identified among the 18 common types we tested would be useful for discrimination of closely related common types. Such was the case in the sites that resolved common HV1/HV2 types H1 and H2. Multiplexes A and B show a significant amount of overlap among the included SNPs (Table 5). The common type H2 differs from common type H1 at the extremely fast site, 152 (Meyer *et al.*, 1999; Allard *et al.*, 2002; Malyarchuck *et al.*, 2002). The most frequently occurring common Caucasian HV1/HV2 type that was not tested was 146C, 263G, and 315.1C, present at 0.6% of the population. This common type again differs from the H1 common type only at mutational fast site, 146 (Meyer *et al.*, 1999; Allard *et al.*, 2002; Malyarchuck *et al.*, 2002). Thus, one would expect that multiplex A and B would be useful for discrimination of this particular common type.

We have demonstrated that entire mtGenome sequencing was successful for identifying shared, neutral SNPs to increase the forensic discrimination of common HV1/HV2 types in the Caucasian population. Some sites in the SNP assay have an apparent general utility (e.g. 16519 and the AC indel). However, most sites are specific to single types, or to a few closely related common types. A similar approach of sequencing mtGenomes of common HV1/HV2 types for SNP discovery to increase forensic discrimination will apparently be needed for other forensically important groups (e.g. African Americans, Hispanics, etc…).

8 Multiplexes		8 Multiplexes + AC indel polymorphism	
# of types	*# individuals/type*	*# of types*	*# individuals/type*
2	14	1	14
1	9	1	13
3	8	1	9
2	7	1	8
1	6	3	7
1	5	2	6
6	4	2	5
7	3	5	4
27	2	6	3
56	1	25	2
		66	1

Table 6. Discrimination of the Common HV1/HV2 types with the Application of Multiplex Panels. The discrimination of 241 common HV1/HV2 type individuals resolved by the eight multiplex panels (left) or the eight panels including the 515-525 AC indel polymorphism (right).

Chapter III. Characterization of the Relative Mutation Rates in the Coding Region of the mtDNA Genome

This section will address Specific Aim #2: To characterize the spectrum of relative mutation rates in the coding region of the mtDNA genome using phylogenetic trees generated from both parsimony and neighbor joining methods. An understanding of the relative mutation rates in the coding region will facilitate future efforts to resolve common HV1/HV2 types in other forensically important groups. If the SNPs that resolve common HV1/HV2 types in Caucasians were mostly fast evolving sites, then the need to continue to sequence entire mtGenomes would be ameliorated. However, if discriminating SNPs are characterized as slow, rare sites specific to a common type (or related common types) then continued mtGenome sequencing will be necessary for SNP discovery.

III. 1. Materials and Methods

A total of 646 coding region mtGenomes were analyzed in this study. Entire mtGenome sequences from the published literature of Ingman *et al.*, 2000 and Maca-Meyer *et al.*, 2001 (representing 53 and 33 mtGenomes, respectively) were downloaded as Sequencher files from Max Ingman's mtDB website at Uppsala University (http://www.genpat.uu.se/mtDB/). An additional 560 coding region sequences from Herrnstadt *et al.*, 2002 were obtained from the MitoKor website

(http://www.mitokor.com/science/560mtdnasrevision.php). These 560 coding region sequences were corrected of the errors that were discovered in the original Herrnstadt *et al.* (2002) data set (Herrnstadt *et al.*, 2003). All sequences were combined in Sequencher, and after proper alignment, all insertions and deletions were removed. A final file of only the coding region sequences (nucleotide positions 577-16023) was exported from Sequencher in the Nexus format (Maddison *et al.* 1997). Additional Nexus files were constructed consisting of the control region, coding region, and full mtGenome sequences using only the 53 human genomes from Ingman *et al.*, 2000.

Phylogenetic trees were generated using PAUP*4.0b10 (Swofford 2003; http://paup.csit.fsu.edu/). Parsimony analyses of unweighted sequences were performed using the heuristic search option using the "closest" addition of sequences and the "tree bisection reconnection" options. Further random addition repetitions (100 total) were performed to try to find trees of shorter lengths.

The mutation rate spectrum was determined by counting the number of character changes as mapped onto one of the MPTs. The distribution of character changes on all trees and the average number of character changes per site were determined within the tree score function of PAUP*. The individual character output of all nonconstant characters were saved to a text file and analyzed with a spreadsheet program. To test the sensitivity of this approach to minor variation in tree topology, we tabulated the distribution of characters from multiple MPTs and compared the rate spectrum from alternative trees. We also compared the rate spectrum from a tree of length 2362, ten steps greater than the MPT. Parsimony-based approximations of the gamma shape

parameter, α, were determined using the Yang and Kumar (1996) method and was calculated within PAUP* using the tree scores function.

A single Neighbor Joining (NJ) tree (using the uncorrected "p" model as a distance option) was obtained using PAUP*. The NJ tree had a tree length of 2353 (consistency index 0.69, retention index 0.91). Character changes were mapped onto the NJ tree, and a rate spectrum produced as described above. In addition to the combined coding region of 646 mtGenomes (above), we analyzed separately the 53 human mtGenomes of Ingman *et al.*, 2000. After alignment using Sequencher, and removal of indels, we exported Nexus files containing sequences of: the entire mtGenome (16569 bases), the coding region only (positions 577-16023; 15447 bases), the control region (positions 16024-16569 and 1-576; 1122 bases), and HV1 (positions 16024-16383; 360 bases). Multiple parsimony trees were evaluated for each region as described above.

Mutations were denoted by the nucleotide base of the revised Cambridge Reference Sequence (Anderson *et al.*, (1981); Andrews *et al.*, (1999)) followed by the nucleotide position, followed by the variant nucleotide base (e.g. C10400T refers to a nucleotide change from the rCRS nucleotide base of Cytosine to Thymine at position 10400). Amino acid changes in coding region genes were determined using the web-based program MitoAnalyzer (Lee and Levin, 2002 - http://www.cstl.nist.gov/biotech/strbase/mitoanalyzer.html).

III. 2. Results

III. 2. 1. Rate Variation in the Coding Region

We first analyzed the data set of 53 human mtGenomes of Ingman *et al.*, 2000. Parsimony trees were evaluated for the estimation of rate heterogeneity in four regions: HV1, control region, coding region, and the entire mtGenome. In addition to using maximum parsimony, we also estimated the rate heterogeneity using trees constructed by the NJ method. Both parsimony and NJ trees had very similar structure, reconstructing the major haplogroup divisions consistent with previous phylogenetic analyses (e.g. see figure 4 of Macaulay *et al.* 1999a). However, in two instances, the parsimony method found trees that were shorter in length than the NJ method (for the control region and entire mtGenome data sets (8 steps and 3 steps, respectively)). Both methods found the same length tree for HV1 and coding region datasets. Despite the minor differences among each method, the estimation of the α shape parameter gave very similar results (Table 7).

All regions of the genome (in part and in total) show high amounts of rate heterogeneity as measured by estimations of the α shape parameter (Table 7). We determined the amount of rate heterogeneity in the control region to be approximately one-half of the rate variation in the coding region ($\alpha = 0.0038$ vs. 0.0075, respectively). The rate variation for the entire Ingman *et al.*, 2000 dataset ($\alpha = 0.005$) was also estimated (Table 7). The rate variation in the HV1 region ($\alpha = 0.209$) was also observed to be high ($\alpha < 1$) (Table 7).

Table 7. Mutation Rate Estimation Based on Parsimony and Neighbor Joining Constructed Phylogenetic Trees. Tree lengths and the α estimation of the gamma distribution were determined using the Yang and Kumar (1996) method for various regions of the Ingman et al (2000) data and 646 coding region human mtDNA sequences (Ingman et al 2000; Maca-Meyer et al 2001; and Herrnstadt et al 2003).

	Parsimony		Neighbor Joining	
Data Set (# genomes)	**Tree Length**	**α estimation**	**Tree Length**	**α estimation**
Ingman HV1 (53)	144	0.2091	144	0.2081
Ingman Control Region (53)	273	0.0038	281	0.0036
Ingman Coding Region (53)	588	0.0075	588	0.0074
Ingman Full Data (53)	873	0.0050	876	0.0067
Total Coding Data (646)	2352	0.0086	2353	0.0083

To increase the sample size for this study an additional 593 coding region sequences from the published data set of Maca-Meyer *et al.*, 2001 and the revised data of Herrnstadt *et al.*, 2003 were added to the Ingman data (total sample size of 646). These data were analyzed for the mutation rate heterogeneity in the coding region only since the Herrnstadt *et al.*, 2003 data lacks sequence information from the control region. A heuristic search using parsimony found multiple trees having one step shorter than the NJ method (MPT length of 2352 - consistency index 0.69, retention index 0.91). Although the NJ tree was longer than the MPTs, we did not observe any major differences in the mutation rate spectrum of fast sites between the NJ and parsimony trees (data not shown) or in the estimation of the α parameter (0.0083 as estimated for the NJ tree, and 0.0086 as estimated by parsimony). The α parameter of the combined data set was very similar to the coding region estimation using the Ingman *et al.*, 2000 data alone ($\alpha = 0.0086$ vs. 0.0075 for the analysis of 646 coding region sequences and 53 sequences, respectively).

III. 2. 2. Relative Mutation Rates in the Coding Region

We characterized the relative mutation rate spectrum in the coding region by counting the number of substitutions per site upon the MPT. The majority of variable sites within the entire dataset changed only once on a tree, while the fastest site changed 15 times within any given MPT.

Even though the analysis yielded several thousand MPTs, there was no significant difference in the mutation rate spectrum among the multiple MPTs. For example, in a survey of 107 MPTs, 72 trees showed exactly the same rate spectrum. The remaining

trees showed only minor differences in the rate spectrum compared to each other, typically affecting only one pair of sites between two trees. For example, for the majority of the parsimony trees (72/107 surveyed) site 3760 changed once while site 15257 changed twice. In two alternative MPTs, site 3760 changed twice while site 15257 changed only once. Between various trees, the difference in the number of changes per site never exceeded one, and always involved a single pair of sites.

We also characterized the site distribution of a "less than optimal" phylogenetic tree. A parsimony-evaluated tree 10 steps greater than the MPT (tree length of 2362) was generated. Only very minor differences were observed in the distribution of character changes between these two trees (Table 8). The estimation of the α shape parameter was also similar between the two trees ($\alpha = 0.0086$ for the MPT and $\alpha = 0.0083$ for the MPT+10 steps tree). These results indicate that our strategy for accessing the mutation rate spectrum is not in any significant way dependent on the choice of MPT, or even on having found, for certain, the shortest possible tree(s) for the data.

Figure 24 shows a graph of the relative mutation rates over the mtDNA coding region of 646 mtDNAs. Figure 25 shows a graph of HV1 and HV2 of relative mutation rates using the 53 human mtDNA sequences of Ingman et al. (2000). The relative rates in HV1 (Figure 25) were observed to have more "intermediate" values than what is observed in coding region genes (Figure 26), or HV2 where the rates tend to be either relatively fast or nearly invariant. This could explain the observation that an α value of HV1 nearly 25-fold higher than the α value from the coding region, or the entire control region (Table 7).

Table 8. Comparison of the Mutation Rate Spectrum in 646 mtDNA Coding Region Genomes Using Parsimony Analysis. The spectrum of sites in the left column was determined by parsimony analysis of the most parsimonious tree (MPT; Tree Length = 2352). The spectrum of sites in the right column was determined by a less than optimal parsimony tree (MPT+10 steps; Length = 2362). Individual branch lengths (Length) represent the number of times the character changed on the tree.

MPT (L=2352)		MPT+10 (L=2362)	
Length	Character	Length	Character
15	709	15	709
13	11914	13	11914
13	5460		
12	13708	12	13708
		12	5460
10	15924	10	15924
10	1719		
9	3010	9	1719
9	10398	9	10398
8	8251	8	8251
8	14470	8	14470
8	15784	8	15784
		8	3010
7	961	7	961
7	3316	7	3316
7	12007		
6	5237	6	5237
6	10915	6	10915
6	11719	6	11719
6	12346	6	12346
6	13105	6	13105
6	13928	6	13928
6	14569	6	14569
6	14766	6	14766
6	15301	6	15301
6	15670	6	15670
6	15884	6	15884
		6	12007

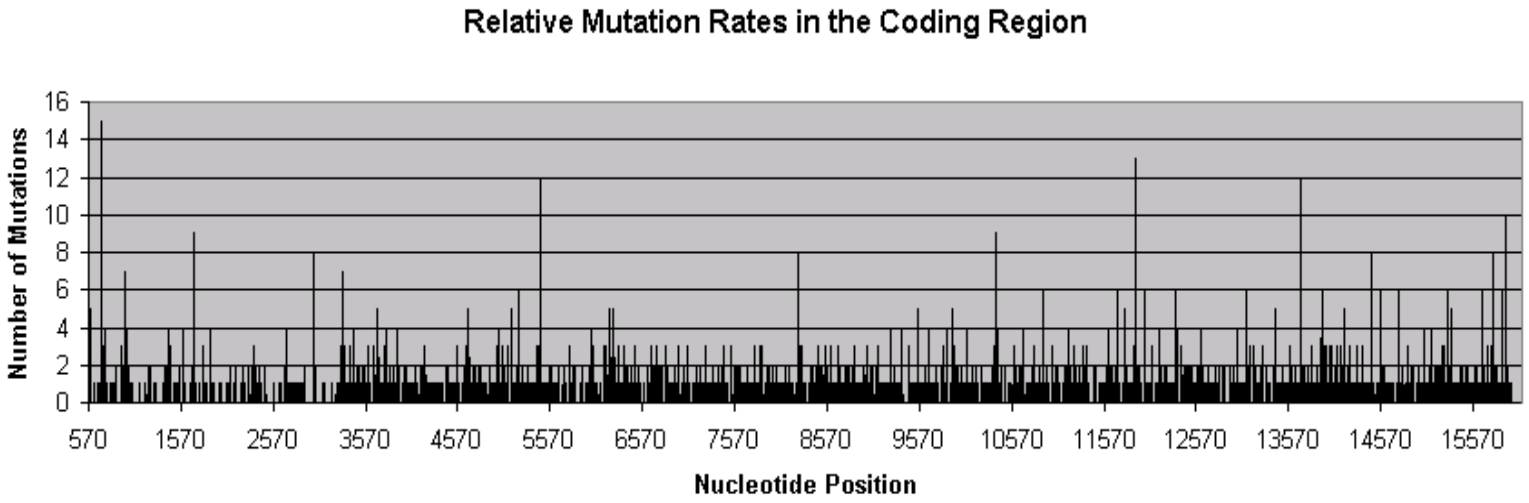
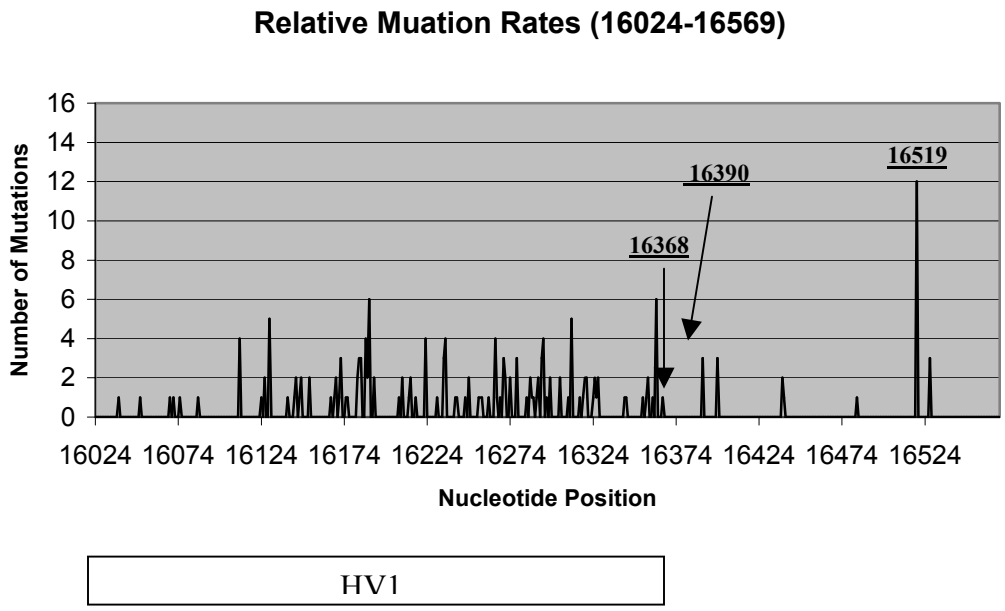
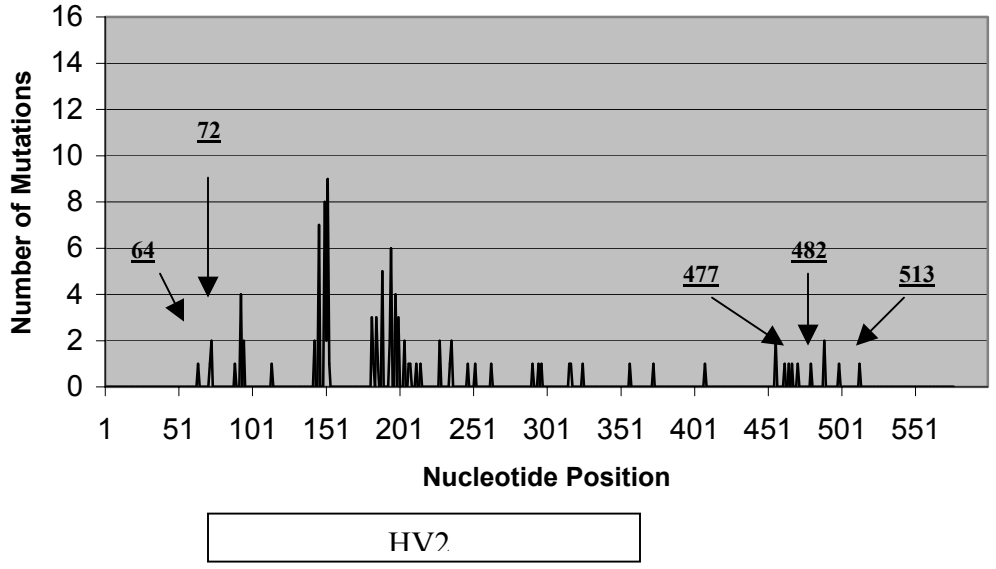


Figure 24. Relative Mutation Rates Over the mtDNA Coding Region. The relative rates were determined by counting the number of character changes occurring on one representative MPT, constructed by parsimony, using 646 coding region mtDNAs (Ingman et al., 2000; Maca-Meyer et al., 2001; Herrnstadt et al., 2002). The nucleotide positions are numbered according to the rCRS.


Relative Muation Rates (16024-16569)
Number of Mutations
16
14
12
10
8
6
4
2
0
16390
16519
16368
16024 16074 16124 16174 16224 16274 16324 16374 16424 16474 16524
Nucleotide Position

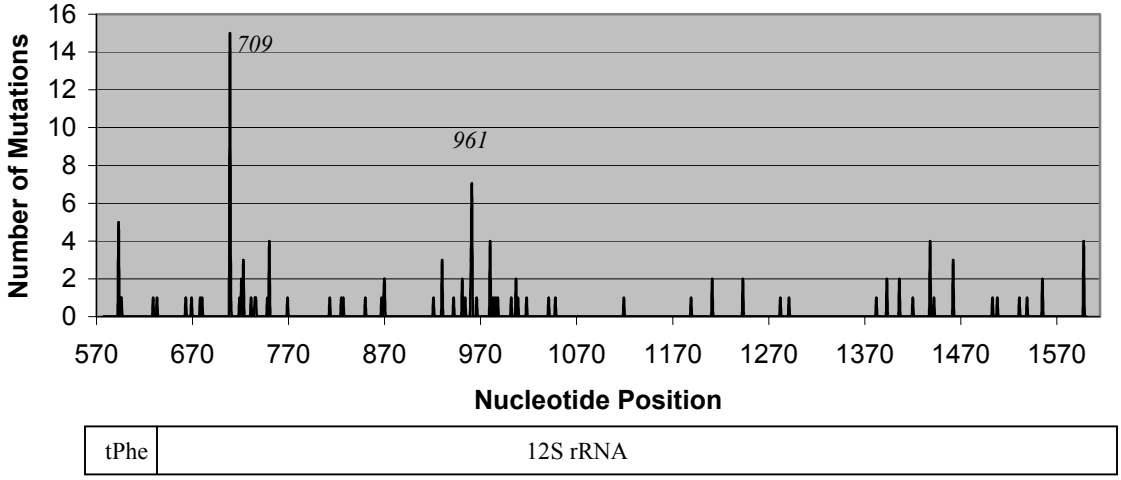
HV1


Relative Muation Rates (1-577)
Number of Mutations
16
14
12
10
8
6
4
2
0
72
64
477
482
513
1 51 101 151 201 251 301 351 401 451 501 551
Nucleotide Position

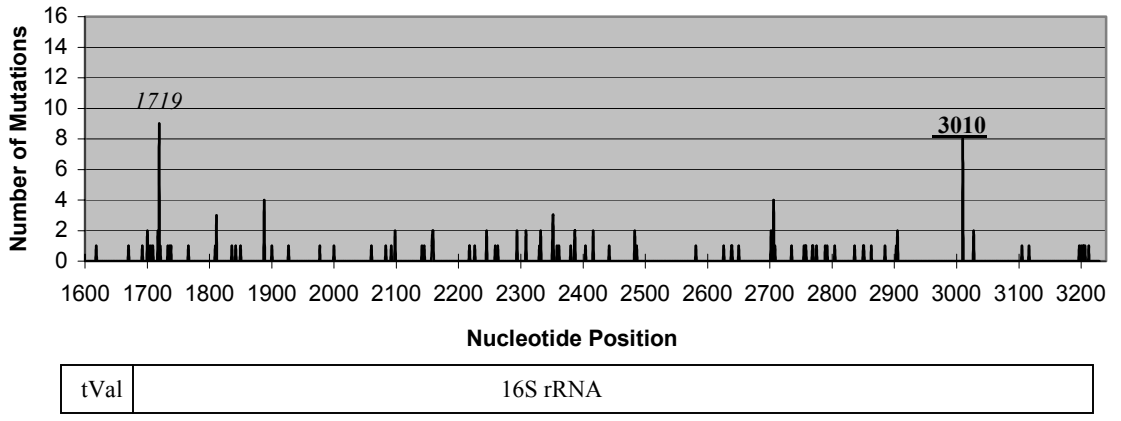
HV2

Figure 25. Relative Mutation Rates in the Control Region. The relative rates were determined by counting the number of character changes occurring on one representative MPT, constructed by parsimony, using 53 human mtDNAs (Ingman et al., 2000). The nucleotide positions are numbered according to the rCRS. The top panel shows the relative mutation rates from nucleotide positions 16023-16569. The bottom panel shows the relative mutation rates from nucleotide positions 1-577. The regions containing HV1 (top) and HV2 (bottom) are noted below the nucleotide positions. SNPs present in the 8 multiplex panels are mapped, in bold and underlined, onto the figure.

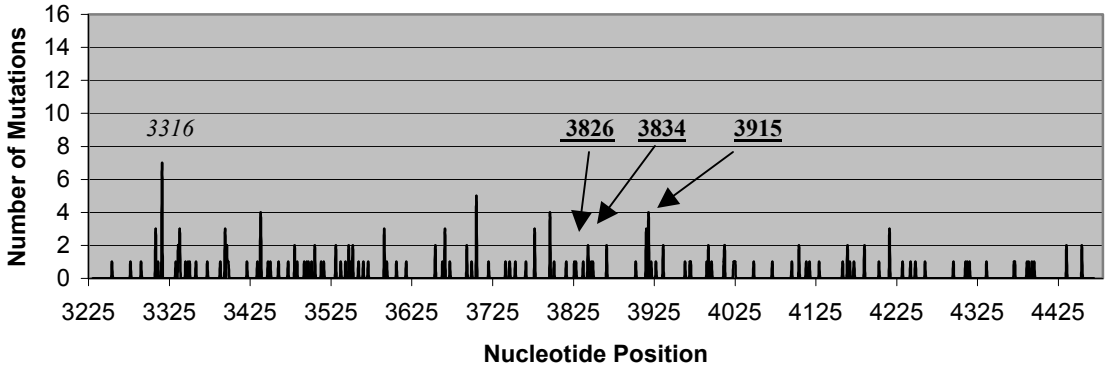
$tRNA^{Phe}$ and 12S rRNA



$tRNA^{Val}$ and 16S rRNA

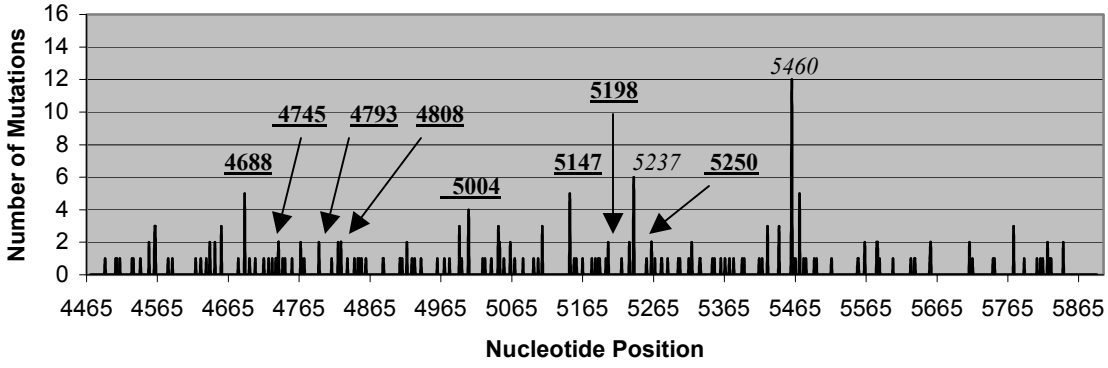


tRNA^{Leu}, ND1, tRNA^{Ile}, tRNA^{Gln}, tRNA^{Met}



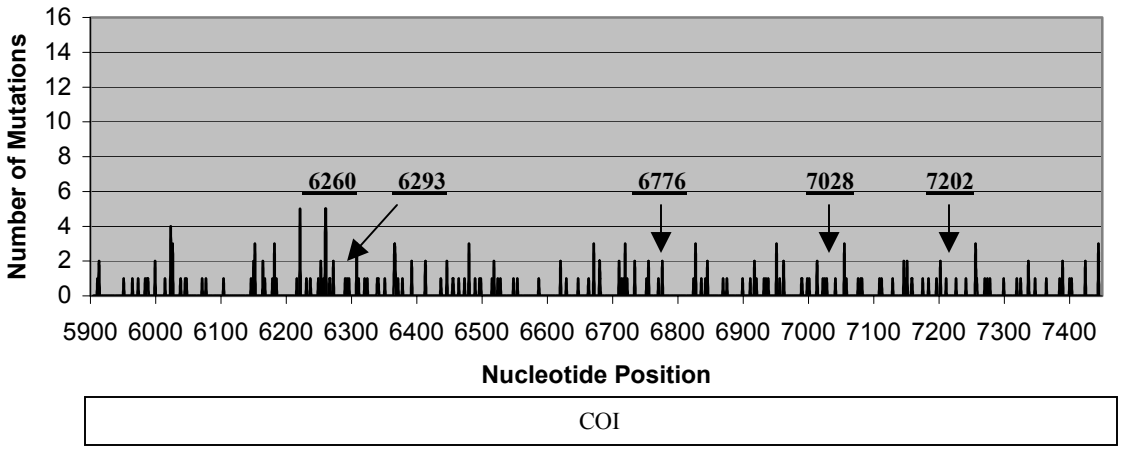
tVal	ND1	tIle tGln tMet

ND2, tRNA^{Trp}, tRNA^{Ala}, tRNA^{Asn}, tRNA^{Cys}, tTNA^{Tyr}

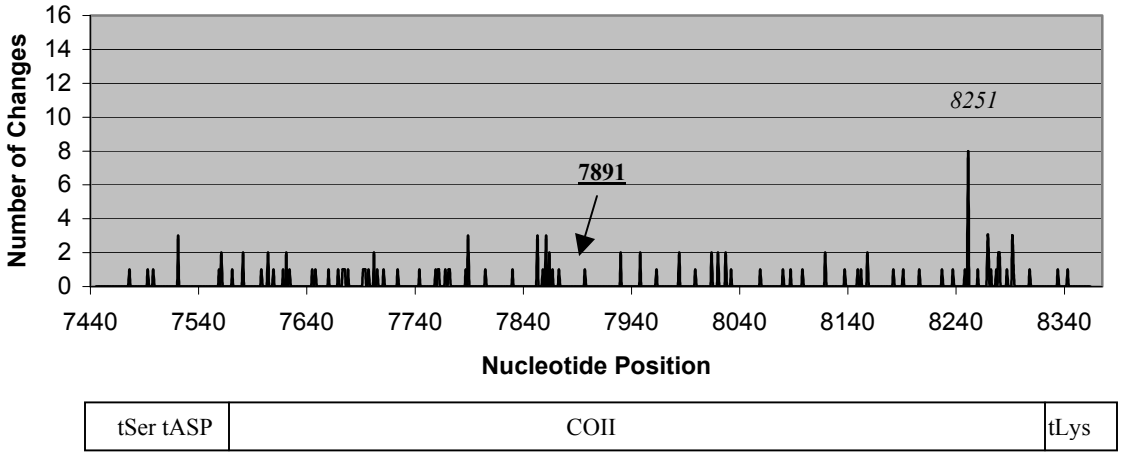


ND2	tTrp tAla tAsn tCys tTyr

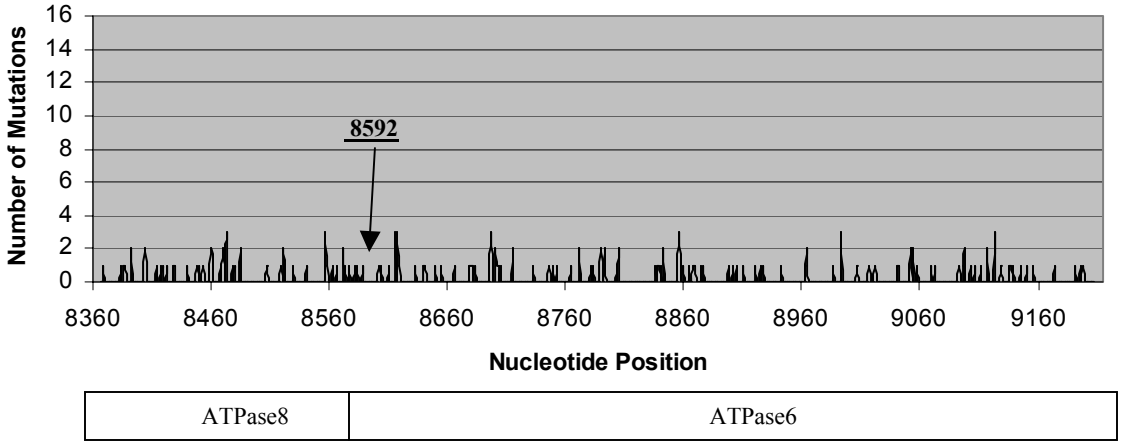
COI



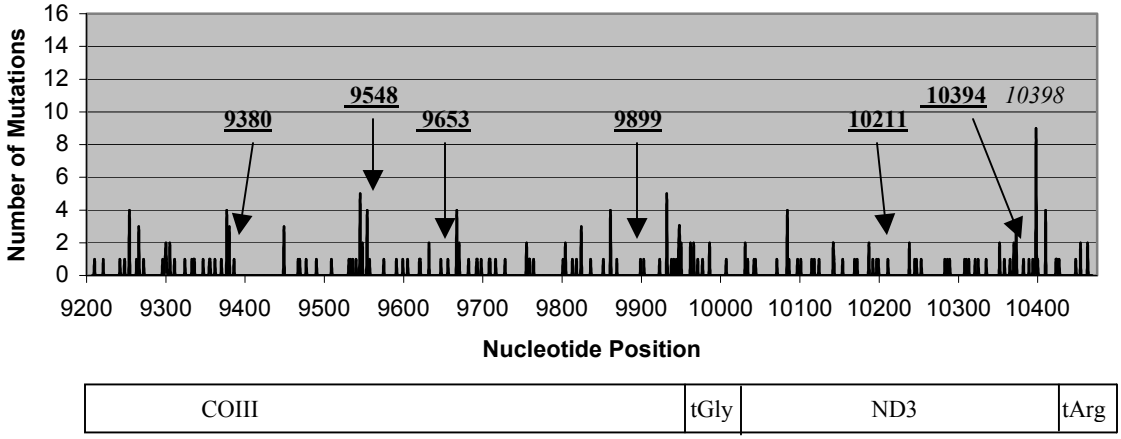
tRNA^{Ser}, tRNA^{Asp}, COII, tRNA^{Lys}



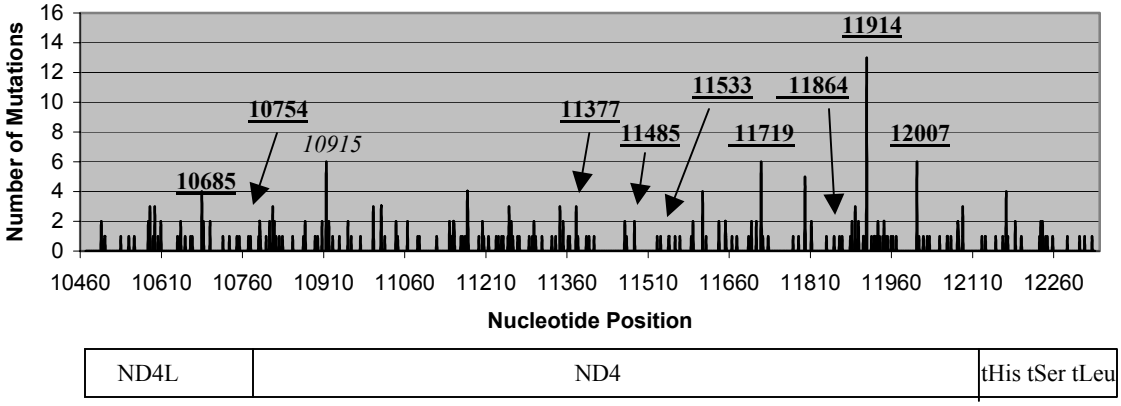
ATPase8/ATPase6



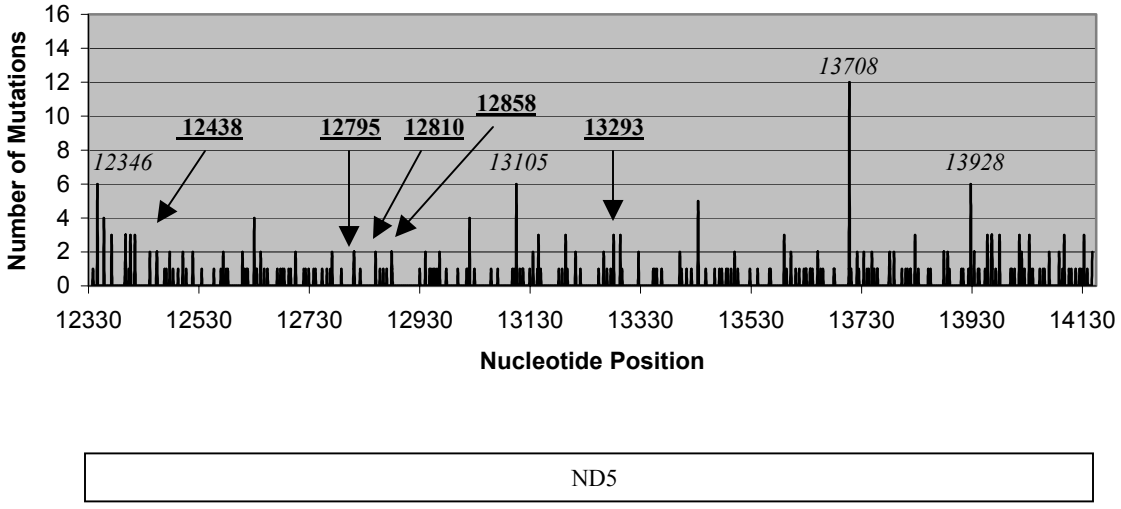
COIII, tRNA^{Gly}, ND3, tRNA^{Arg}



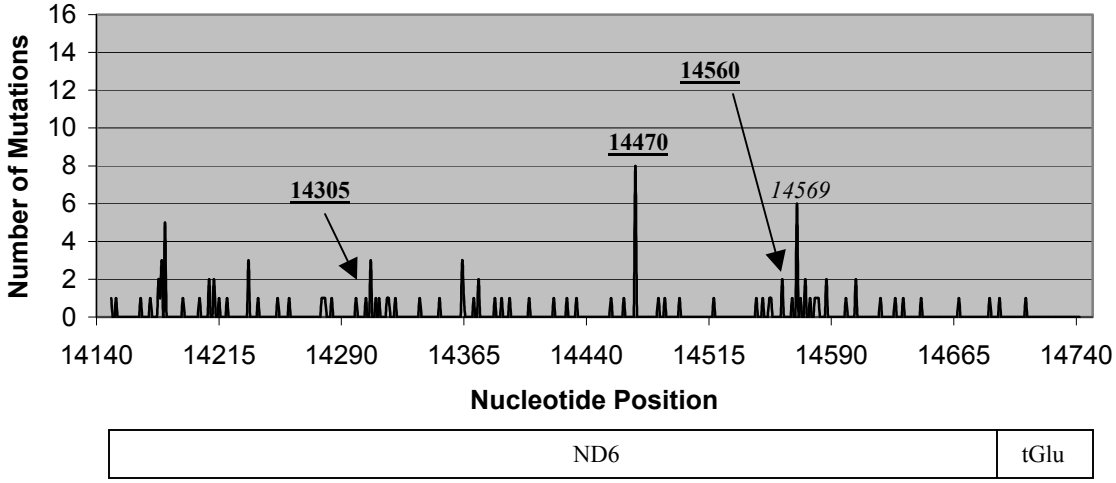
ND4, ND4L, tRNA^{His}, tRNA^{Ser2}, tRNA^{Leu2}



ND5



ND6, tRNA^{Glu}



CytB, tRNA^{Thr}, tRNA^{Pro}

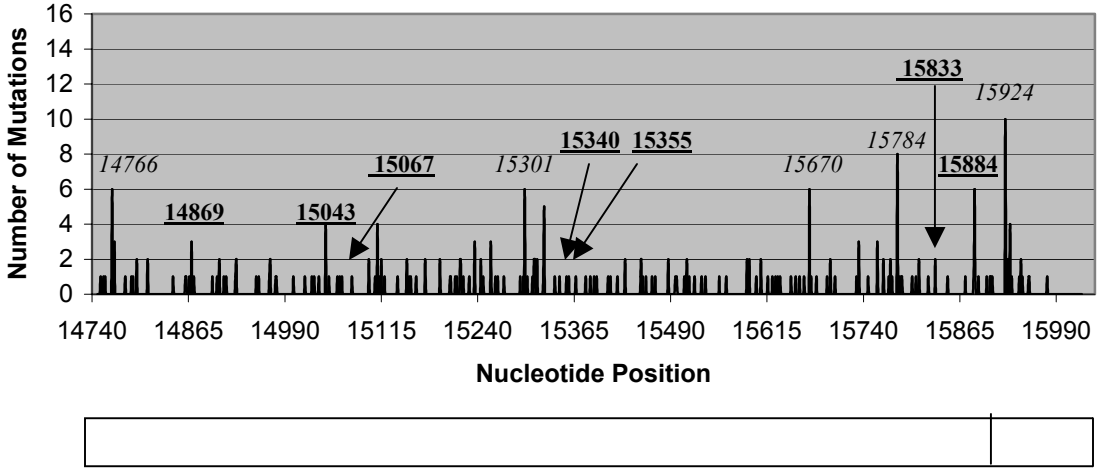


Figure 26. Relative Mutation Rates in the Coding Region, by Genes. The relative rates were determined by counting the number of character changes occurring on one representative MPT, constructed by parsimony, using 646 human mtDNAs (Ingman et al., 2000; Maca-Meyer et al., 2001; Herrnstadt et al., 2002). The nucleotide positions are numbered according to the rCRS. The 12 panels show the relative mutation rates delineated by genes (indicated by the boxes below the nucleotide positions), in the coding region. Mutations showing six or more character changes on the MPT, and not part of the SNP multiplex assay panels are noted in italics in each panel. SNPs within the eight multiplex panels are indicated by bold, underlined numbers.

III. 3. Discussion

III. 3. 1. Rate Variation in the Coding Region

We have determined the overall amount of mutation rate variation and the specific relative mutation rates of all sites in the human mtGenome coding region by mapping character changes onto a phylogenetic trees produced by parsimony and NJ analysis. The amount of rate heterogeneity was measured by estimating the α shape parameter of the gamma distribution as determined by the Yang and Kumar, (1996) method. Previous critiques of using parsimony analysis to estimate the mutation rate have suggested the parsimony method underestimates the true substitution rate by overestimating the α shape parameter (reviewed in Yang, 1996). However, simulations by Deng and Fu, (2000) showed that α is not always underestimated, especially when the number of sequences is larger and the phylogeny is accurately reconstructed.

Previous studies (Meyer *et al.*, 1999; Excoffier and Yang, 1999; Pesole and Saccone, 2001) have established that there is high mutation rate heterogeneity within the control region (specifically within regions HV1 and HV2). Using only the HV1 region from the 53 human mtGenomes of Ingman *et al.*, 2000, our parsimony analysis gave an α value of 0.21 that is very similar to a value of 0.26 as estimated using a maximum likelihood approach by Meyer *et al.*, 1999 (see their Table 1). Excoffier and Yang (1999), also using a maximum likelihood approach, calculated an α value for HV1 ranging from 0.25 – 0.28 (depending on the model of evolution used). The estimation of α using other methods (e.g. "method of moments" – Johnson and Kotz, 1969; and the

ML method of Sullivan *et al.*, 1995) gave slightly higher estimations ($\alpha = 0.41$ and 0.34, respectively), indicating that these methods would, in this instance, overestimate the mutation rate.

Previous studies have focused on assigning various CR sites into a small number of rather arbitrary rate categories, with particular interest in identifying fast sites (Wakeley, 1993; Hasegawa *et al.*, 1993). A result of this is a perspective that emphasizes "hot spots" and their effect on properties of mtDNA evolution (Howell *et al.*, 1996; Paabo, 1996; Parsons *et al.*, 1997; Jazin *et al.*, 1998; Parsons and Holland, 1998). We don't feel that the categorical approach is particularly helpful for illuminating what is essentially a continuous rate variation. Table 9 presents a summary of the mutational spectrum of sites having 6 or more substitutions. A complete list of the site-specific mutation rate spectrum for the coding region can be found in Appendix 4. The majority of the fastest sites were found within protein coding genes (20/25). One might expect that most of the changes in the protein coding would occur at 3rd position silent (synonymous) polymorphisms. In fact, we observed nearly as many non-synonymous mutations within the fastest sites as synonymous mutations (8 vs. 11, respectively – Table 9). One striking observation is that all four of the fastest sites represented in the ND5 gene are non-synonymous changes, while all four of the fastest ND4 sites are synonymous changes (Table 9).

Evidence that natural selection has contributed to the pattern of sequence variation in the human mtDNA genome has been accumulating (Excoffier, 1990; Nachman *et al.*, 1996; Templeton, 1996; Wise *et al.*, 1998; Torroni *et al.*, 2001). Recently, Mishmar *et al.*, (2003) analyzed the nonsynonymous versus synonymous ratio

Table 9. Mutation Rate Spectra of the Coding Region of 646 Human Mitochondrial DNA Genomes using Parsimony Analysis. The number of branch changes per character (Length) of the fastest sites are represented. The nucleotide position number (character), gene affected, and codon position are enumerated according to the rCRS sequence (Anderson *et al.*, 1981; Andrews *et al.*, 1999).

Length	Character	Gene	Codon
15	709	12S	*
13	11914	ND4	3
12	5460	ND2	1
12	13708	ND5	1
10	15924	tRNA(thr)	*
9	1719	16S	*
9	10398	ND3	1
8	3010	16S	*
8	8251	COII	3
8	14470	ND6	3
8	15784	CYTB	3
7	961	12S	*
7	3316	ND1	1
6	5237	ND2	3
6	10915	ND4	3
6	11719	ND4	3
6	12007	ND4	3
6	12346	ND5	1
6	13105	ND5	1
6	13928	ND5	2
6	14569	ND6	3
6	14766	CYTB	2
6	15301	CYTB	3
6	15670	CYTB	3
6	15884	NC	*

(Ka/Ks) for all 13 mtDNA protein coding genes and concluded that regional mtDNA variation may have been influenced by positive selection. In particular, non-synonymous variation in the ATP6 gene was significantly higher in Asian-derived populations within the arctic zone compared to populations of Caucasians and Africans. We have also observed a considerable amount of nonsynonymous variation occurs in the ATP6 and ATP8 genes of the 241 Caucasian mtGenomes sequenced in this study (Table 4). Mishmar *et al.* (2003) have hypothesized that rapid changes in mtDNA protein coding regions, tweaking the ability of the mitochondria to produce energy via oxidative phosphorylation and generating heat, would be selectively beneficial to migrating populations adapting to new climates and diets.

III. 3. 2. Comparison of Rate Spectrum Analyses

To our knowledge, only one other study (Meyer and von Haeseler, 2003) has evaluated the relative mutation rates over the mtDNA coding region. Using the 53 entire mtGenome sequences of Ingman *et al.* (2000), Meyer and von Haeseler (2003) utilized a pair wise comparison method to determine genetic distances, and then evaluated this matrix with maximum likelihood to estimate site-specific substitution rates in the mtGenome. They estimated the α shape parameter in the mtGenome to be 0.002. Our estimation of 0.005, using the same data, was quite similar. Both of these values have confirmed the existence of extreme mutation rate heterogeneity in the mtDNA genome.

There is, however, a significant discrepancy between the site-specific rates that we determined and those identified by Meyer and von Haeseler (2003) (Table 10). We

constructed a skeleton tree (Figure 27) based on the well-established phylogeny of human mtDNA (using the NJ tree [Figure 2, Ingman *et al.*, 2000] and the 53 complete human mtGenomes of Ingman *et al.*, 2000). In this tree, African sequences of haplogroup L1 and L2 are basal to the clades of the African haplogroup L3 and Eurasian macro-haplogroups M and N (Figure 27). We mapped many of the fastest sites as determined by Meyer and von Haeseler (2003) onto the skeleton tree in Figure 27. Many of the "fastest" sites determined by the pair wise method estimated by ML from Meyer and von Haeseler (2003) were actually diagnostic mutations for major mtDNA haplogroups. For example, the macro-haplogroup M-specific transition at position C10400T carries a rate score of 100 (as calculated by the method of Meyer and von Haeseler, 2003). However, as Yao *et al.* (2003) recently noted, not one homoplasmic change at this position has been observed in over 900 coding region sequences or fragments.

It appears that the pair wise method mistakes high frequency polymorphisms for sites with high mutation rates. For example, three of the macro-haplogroup N polymorphisms (A8701, T9540, and T10873) are found in 20 of the 53 individuals in the Ingman *et al.* (2000) data set (Figure 27). These three sites have a mutation rate score of 155, suggesting that these variants are extremely fast sites, according to Meyer and von Haeseler (2003). In contrast, the mapping of these sites onto a phylogenetic tree evaluated by parsimony would require a single change on the tree to map each site. The phylogenetic analysis would not suggest that these sites have high mutation rates. In fact these sites are stable, ancient mutations.

Although the parsimony method has been criticized for underestimating the substitution rate, we found our estimation of the α shape parameter to be quite similar to

Table 10 Comparison of the Mutation Rate Spectrum Determined by a Pair Wise Distance Method (with ML) and Parsimony. Rate Scores were determined using a pair wise distance method (evaluated using ML) tabulated from Meyer and von Haeseler, 2003. Character refers to the nucleotide position according to the rCRS (Anderson *et al.*, 1981; Andrews *et al.*, 1999). The length column values are the number of mutations (changes) that occurred on the MPT analyzed using parsimony (this study). Boxes surrounding the rate score, character, and length are sites that both studies determined to be fast evolving.

Rate Score	Character	Length
175.21	15301	6
162.82	10398	9
155.20	8701	2
155.20	9540	1
155.20	10873	1
129.16	12705	2
119.30	7521	3
112.03	769	1
112.03	1018	1
112.03	3594	1
112.03	4104	2
112.03	7256	3
112.03	13650	1
105.84	11914	13
100.77	10400	1
100.77	14783	1
100.77	15043	4
96.96	10688	2
96.89	13105	7
89.38	825	1
89.38	2758	1
89.38	2885	1
89.38	8468	1
89.38	8655	1
89.38	10810	2
89.38	13506	1

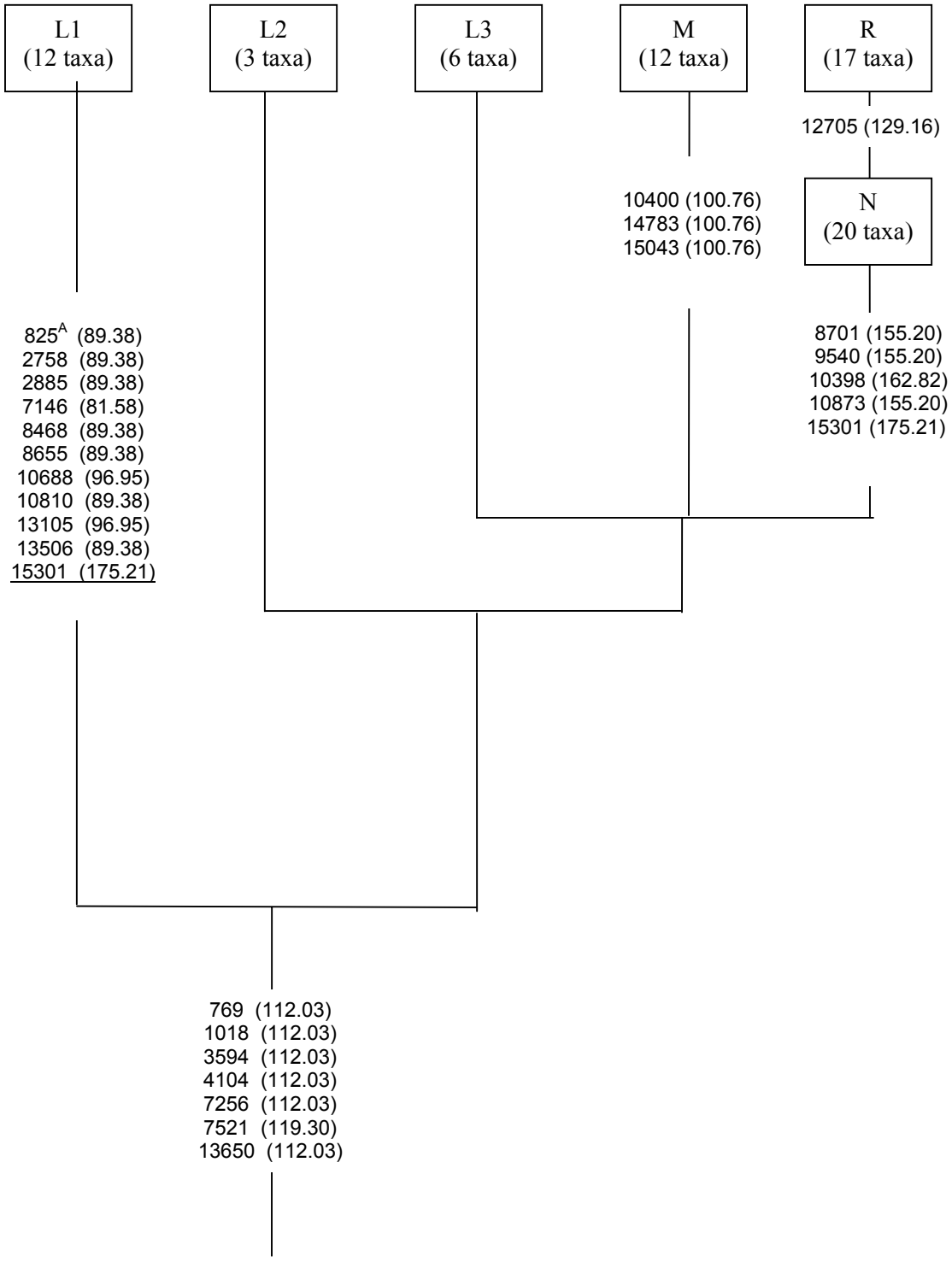

L1
(12 taxa)
L2
(3 taxa)
L3
(6 taxa)
M
(12 taxa)
R
(17 taxa)
12705 (129.16)
N
(20 taxa)
10400 (100.76)
14783 (100.76)
15043 (100.76)
825A (89.38)
2758 (89.38)
2885 (89.38)
7146 (81.58)
8468 (89.38)
8655 (89.38)
10688 (96.95)
10810 (89.38)
13105 (96.95)
13506 (89.38)
15301 (175.21)
8701 (155.20)
9540 (155.20)
10398 (162.82)
10873 (155.20)
15301 (175.21)
769 (112.03)
1018 (112.03)
3594 (112.03)
4104 (112.03)
7256 (112.03)
7521 (119.30)
13650 (112.03)

Figure 27. Skeleton Tree of the 53 Human Genomes and Mutation Rate Scores.
The skeleton tree was built using the neighbor joining method from the 53 complete human mtDNA genomes of Ingman et al (2000). Haplogroup associated polymorphisms with mutation "rate scores" of the fastest sites according to Meyer and von Haeseler (2003) are mapped on the branches. The numbers of taxa represented in the Ingman data set are noted for some of the haplogroups. Underlined sites show parallel mutations among different haplogroups.

other methods of estimating the substitution rate. One advantage of using parsimony over maximum likelihood or pair wise methods is the ability to analyze a greater number of sequences. Tourasse and Gouy (1997) have shown that increasing the number of sequences in a data set of rRNA weakens the bias inherent in using parsimony. They contend that increasing the sample size will increase the genetic diversity present in smaller data sets, leading to better estimations of population genetic parameters. For example, in small data sets, some sites that are truly variable may appear as invariant. Increasing the number of samples, and the genetic diversity of the data, can decrease the α shape parameter (Wakeley, 1993; Tourasse and Gouy, 1997). We also contend that the insight gained from using tree-based methods can be valuable to distinguish ancient, haplogroup-associated sites from sites that are truly fast.

III. 3. 3. Mutation Rate Variation and Forensic SNPs for Discrimination

We were interested in characterizing the mutation rate heterogeneity in the coding region in order to understand the nature of the sites we discovered for resolving common types (Table 5). We were curious if the sites that resolved the common HV1/HV2 types in Caucasians were a) slow, rare sites that were specific to one common type; b) universal fast sites that would resolve any common HV1/HV2 type; or c) a combination of both.

Perhaps one of the more interesting findings of the site-specific spectrum of mutation rates in the coding region was the observation that many of the fastest sites were non-synonymous changes in the mtGenome (Table 9). Recently, it has been suggested that highly variable SNPs in the mtDNA coding region should be targeted for increased forensic discrimination (Andreasson *et al.*, 2002; Lutz-Bonengel *et al.*, 2003). The sites

identified here as fast, non-synonymous polymorphisms would need to be evaluated as being truly neutral and not associated with mtDNA diseases before being implemented into a SNP assay (as outlined in our criteria). Simply identifying universal fast sites for SNP assays would not be prudent for increased forensic discrimination without such consideration.

Within our 241 Caucasian mtGenome sequences, we observed variation at many of the fastest sites (Table 11). Fifteen of the twenty-five fastest sites in our rate spectrum have been observed to vary in one or more individuals in our database. Six of these sites were neutral and varied in two or more individuals (Table 11) and have been incorporated into our eight multiplex panels (Table 5). These include: four protein-coding sites (11914, 14470, 11719, and 12007); one rRNA site (3010); and one site in the non-coding intergenic spacer regions (15884).

The strategy to use brute-force sequencing to discover neutral sites to resolve common HV1/HV2 types proved to be a sound approach. Most of the SNPs utilized in the eight multiplex panels were the slow, rare sites that were specific to a common type ("needles" that couldn't be pulled out the "haystack of mtGenomes" in the literature). The nature of the discriminating SNPs in the eight multiplex panels is best illustrated in Figure 26, where all 51 coding region SNP sites have been mapped onto the relative rates for each coding region gene. The eight remaining multiplex SNPs in the control region (excluding the AC indel) have been mapped onto the control region relative rates (Figure 25). A small number (7/58, 12%) of the shared, neutral SNPs incorporated in the eight multiplex panels were identified as having relatively fast rates in the mtGenome (Table 11; Figure 25; and Figure 26). It appears that forensic discrimination of common

Table 11. Mutation Rate Spectrum Observed in 241 Common HV1/HV2 Caucasian Types. Fast sites were determined by the number of changes on a MPT using 646 coding region sequences as described in Materials and Methods. Length is the number of changes observed on the MPT, Character is the nucleotide position numbered according to the rCRS (Anderson *et al.*, 1981; Andrews *et al.*, 1999), the Gene and Codon refers to the location in the mtGenome where the site occurs. A "yes" in the 241 Caucasians column indicates that at least one individual was observed to have a mutation at the particular site. Sites indicated by, "Yes-SNP", were included in the multiplex SNP panel (Table 5).

Length	Character	Gene	codon	241 Caucasians
15	709	12S	*	Yes
13	11914	ND4	3	Yes-SNP
12	5460	ND2	1	Yes
12	13708	ND5	1	Yes
10	15924	tRNA(thr)	*	Yes
9	1719	16S	*	Yes
9	10398	ND3	1	Yes
8	3010	16S	*	Yes-SNP
8	8251	COII	3	
8	14470	ND6	3	Yes-SNP
8	15784	CYTB	3	
7	961	12S	*	
7	3316	ND1	1	
6	5237	ND2	3	Yes
6	10915	ND4	3	Yes
6	11719	ND4	3	Yes-SNP
6	12007	ND4	3	Yes-SNP
6	12346	ND5	1	
6	13105	ND5	1	Yes
6	13928	ND5	2	
6	14569	ND6	3	
6	14766	CYTB	2	
6	15301	CYTB	3	
6	15670	CYTB	3	
6	15884	CYTB	nc	Yes-SNP

HV1/HV2 types in other populations of interest will require a similar approach of targeted mtGenome sequencing for SNP discovery, rather than relying on universal fast sites for resolving common types.

Chapter IV. Summary

In this dissertation, we have sought to ameliorate one of the greatest limitations of current forensic mtDNA testing for HV1/HV2 in Caucasians: the lack of resolution that is obtained when a small number of common types are encountered in a case. We have sequenced the entire mtGenome of 241 Caucasian individuals belonging to 18 common HV1/HV2 types. We observed a significant amount of variation in the mtGenome that is useful for distinguishing individuals who would otherwise match for HV1/HV2. We chose to focus on SNPs in the coding region that meet the criteria of being neutral, shared, and non-redundant. Based upon these criteria, we have developed a set of eight multiplex panels containing 59 discriminatory sites useful for resolving individuals sharing 18 common Caucasian HV1/HV2 types. Applying the SNP panels (including the control region AC indel repeat information) to the 241 individuals produced 113 haplotypes, a 6-fold improvement in resolution. The SNP multiplex panels are meant to complement current mtDNA testing of HV1/HV2, and can be developed to work on instrumentation already in the laboratory (via the SNaPShot^{TM} kit). When a common type is encountered in casework, the forensic scientist can then utilize the appropriate SNP panel for further resolution.

We have also characterized site-specific mutation rates in the coding region. We have observed that nearly half of the fastest sites in the coding region produce non-synonymous changes. Most of the sites we discovered to be useful for forensic discrimination are sites that are specific to a particular common HV1/HV2 type. A few of the SNP sites in our multiplex panels contain sites that may be useful in resolving

common HV1/HV2 types from various backgrounds. However, the resolution of common types in other forensically important populations will likely require a similar strategy of entire mtGenome sequencing to discover discriminatory SNPs.

Appendix 1. Oligonucleotide Sequences for PCR Primers Used to Sequence the Entire mtDNA Genome from 12 Overlapping Fragments. The PCR amplicon (with primers used to amplify), primer name, and primer sequence are listed below. Primers that were used twice within the same amplicon are noted as (2X).

PCR Amplicon	Name	Sequence (5' --> 3')
Amp01 (F361/R2216)	F361	ACAAAGAACCCTAACACCAGC
	F873	GGTTGGTCAATTTCGTGCCAG
	R921	ACTTGGGTTAATCGTGTGACC
	F1234	CTCACCACCTCTTGCTCAGC
	R1425	AATCCACCTTCGACCCTTAAG
	F1657	CTTGACCGCTCTGAGCTAAAC
	R1769	GCCAGGTTTCAATTTCTATCG
	R2216	TGTTGAGCTTGAACGCTTTC
Amp02 (F1993/R3557)	F1993	AAACCTACCGAGCCTGGTG
	F2417	CACTGTCAACCCAACACAGG
	R2660	AGAGACAGCTGAACCCTCGTG
	R2834	CCCAACCTCCGAGCAGTACATG
	R3006	ATGTCCTGATCCAACATCGAG
	F3234	AGATGGCAGAGCCCGGTAATC
	R3557	AGAAGAGCGATGGTGAGAGC
Amp03 (F3441/R4983)	F3441	ACTACAACCCTTCGCTGACG
	F3931 (2X)	TCAGGCTTCAACATCGAATACG
	R3940	TGAAGCCTGAGACTAGTTCGG
	R4162	TGAGTTGGTCGTAGCGGAATC
	F4392	CCCATCCTAAAGTAAGGTCAGC
	R4983	GTTTAATCCACCTCAACTGCC
Amp04 (F4797/R6526)	F4797 (2X)	CCCTTTCACTTCTGAGTCCCAG
	F5318	CACCATCACCCTCCTTAACC
	F5700 (2X)	TAAGCACCCTAATCAACTGGC
	R5882	GCTGAGTGAAGCATTGGACTG
	F6242	CGCATCTGCTATAGTGGAGG
	R6526	ATAGTGATGCCAGCAGCTAGG
Amp05 (F6426/R8311)	F6426 (2X)	GCCATAACCCAATACCAAACG
	F7075	GAGGCTTCATTCACTGATTTCC
	R7255	GAGGCTTCATTCACTGATTTCC
	F7645 (2X)	TATCACCTTTCATGATCACGC
	R7792	GGGCAGGATAGTTCAGACGG
	R8311	AAGTTAGCTTTACAGTGGGCTCTAG
Amp06 (F8164/R9848)	F8164 (2X)	CGGTCAATGCTCTGAAATCTGTG
	F8539	CTGTTCGCTTCATTCATTGCC
	F8903	CCCACTTCTTACCACAAGGC
	R9059	GTGGCGCTTCCAATTAGGTG
	F9309	TTTCACTTCCACTCCATAACGC
	R9403	GTGCTTTCTCGTGTTACATCG
	R9848	GAAAGTTGAGCCAATAATGACG

Appendix 1. (Continued).

PCR Amplicon	Name	Sequence (5' --> 3')
Amp07 (F9754/R11600)	F9754 (2X)	AGTCTCCCTTCACCATTTCCG
	F10386	GGATTAGACTGAACCGAATTGG
	R10556	GGAGGATATGAGGTGTGAGCG
	F11001 (2X)	AACGCCACTTATCCAGTGAACC
	R11267	TGTTGTGAGTGTAAATTAGTGCG
	R11600	CTGTTTGTCGTAGGCAGATGG
Amp08 (F11403/R13123)	F11403 (2X)	GACTCCCTAAAGCCCATGTCG
	F11901 (2X)	TGCTAGTAACCACGTTCTGGTG
	F12357	AACCACCCTAACCCTGACTTCC
	F12601	TTCATCCCTGTAGCATTGTTCG
	R12876	GATATCGCCGATACGGTTG
	R13123	AGCGGATGAGTAAGAAGATTCC
Amp09 (F12793/R14388)	F12793	TTGCTCATCAGTTGATGATACG
	F13188	CACTCTGTTCGCAGCAGTATG
	R13343	TTGAAGAAGGCGTGGGTACAG
	F13518 (2X)	CATCATCGAAACCGCAAAC
	R13611	TCGAGTGCTATAGGCGCTTGTC
	F13899	TTTCTCCAACATACTCGGATTC
	R13935	TGTGATGCTAGGGTAGAATCCG
	R14388	TTAGCGATGGAGGTAGGATTGG
Amp10 (F14189/R15396)	F14189 (2X)	ACAAACAATGGTCAACCAGTAAC
	F14470	TCCAAAGACAACCATCATTCC
	F14909	TACTCACCAGACGCCTCAACCG
	R14996	CGTGAAGGTAGCGGATGATTC
	R15396	TTATCGGAATGGGAGGTGATTC
Amp11 (F15260/R16084)	F15260	AGTCCCACCCTCACACGATTC
	F15574	CGCCTACACAATTCTCCGATC
	R15774	ACTGGTTGTCCTCCGATTCAGG
	R16084	CGGTTGTTGATGGGTGAGTC
Amp12 (F15878/R649)	F15971	TTAACTCCACCATTAGCACC
	R16175 (2X)	TGGATTGGGTTTTTATGTA
	F16190	CCCCATGCTTACAAGCAAGT
	R16400	GTCAAGGGACCCCTATCTGA
	F16450 (2X)	GCTCCGGGCCCATAACACTTG
	R274	TGTGTGGAAAGTGGCTGTGC
	F314 (2X)	CCGCTTCTGGCCACAGCACT
	R649 (2X)	TTTGTTTATGGGGTGATGTGA

Appendix 2. Sequence Polymorphisms of 241 Individuals over the Entire mtGenome, by Common HV1/HV2 Type. Polymorphisms are compared to the rCRS (Anderson *et al.,* 1981; Andrews *et al.,* 1999). All individuals had the following polymorphisms compared to the rCRS sequence: 263G, 315.1C, 750G, 1438G, 3107del, 4769G, 8860G, and 15326G. Polymorphisms shared among a particular haplogroup are noted.

H1s

H1-01	1719A	4793G	16519C					
H1-02/03/08	10211T	16519C						
H1-04	6575G	9091G	12153T	13658T	14470A	16519C		
H1-05	3010A	5460A	15817G	16472A	16519C			
H1-06/22	3992T	4024G	5004C	8269A	9123A	10044G	14365T	14582G
H1-07	477C	2851G	3010A	5237A	8764A	12858T	13879A	16519C
H1-09	6956C	7202G	10846T	15924R	16519C			
H1-10	3010A	8308G	8812G	16519C				
H1-11	523A del	524C del	3010A	9921A	11084G	16519C		
H1-12	4216C	14470A	14831A	16519C				
H1-13	477C	3010A	8803G	16519C				
H1-14	3010A	7975G	9923T	16519C				
H1-15	4793G	5348T	11314G	16519C				
H1-16	1719A	4793G	11377A	16519C				
H1-17	3010A	12341T	15323A	16519C				
H1-18	2259T	4745G	13680T	14872T				
H1-19	4793G	6216C	8784G	11878C	12630A	14405G	15409T	16519C
H1-20	72C	2706G	4580A	5582G	7028T	15904T	16519C	
H1-21	9129T	10394T	16519C					
H1-23	3010A	4707A	4976G	10049G	16519C			
H1-24	477C	3010A	8764A	12858T	15466A	16519C		
H1-25	7202G	16519C						
H1-26	4793G	6177G	16519C					
H1-27	4048A	13708A	14364A	16519C				
H1-28	3010A	16519C						
H1-29	72C	2706G	4580A	7028T	15904T			
H1-30	3010A	4733C	9290A	10256C	16519C			
H1-31	4793G	16519C						

Appendix 2. (Continued)

H2s 152C

H2-01	6776C	11182G	15758G	16519C			
H2-02	10685A	11914A	13104G	14560A	16519C		
H2-03	523A del	524C del	3992T	4024G	5004C	7741C	8433C
	8975Y	9123A	14365T	14582G			
H2-04	64T	524.1A	524.2C	3010A	11824R	11864C	15497A
H2-05	3992T	4418C	5772A	6776C	10754C	16519C	
H2-06	10211T	16519C					
H2-07	10394T	13168T	14155T	14577C	14869C	16519C	
H2-08/23	8592A	10394T	16519C				
H2-09	2259T	4745G	7337A	13326C	13680T		
	13851T	14831A	14872T	16519C			
H2-10	6776C	11590G	12217G	14687G	16519C		
H2-11	3640A	5213T	5893T	6776C	16519C		
H2-12	6776C	10754C	16519C				
H2-13	8592A	10394T	15340G	16519C			
H2-14	3511G	5460A	12616C				
H2-15	10394T	11020G	15110A	16519C			
H2-16	4080C	4646C	16519C				
H2-17	3970T	6776C	12236A	16519C			
H2-18	739T	8592A	10394T	15340G	16519C		
H2-19	477C	3010A	8645G	12354C	16519C		
H2-20	2772T	16519C					
H2-21	477C	3010A	5654C	9025A	16519C		
H2-22	4823C	6776C	10754C	15244G	16519C		
H2-24	524.1A	524.2C	3010A	11864C	15497A	16519C	
H2-25	3918A	9091G	14470A	16519C			

H3s - 16129A

H3-01	55C	57C	6253C	11050C	11410C	14959T	16519C
H3-02	3010A	3221G	8705C	9581C	11204C	16519C	
H3-03	2581G	6776C	16519C				
H3-04	2259T	4745G	13680T	14560A	14872T		
H3-05/06	3915A	16519C					
H3-07	3010A	15323A	16519C				
H3-08	460C	3010A	3621C	7951G	8701G	12493G	
H3-09	3915A	593C	6296T	14869A	16519C		
H3-10	2581G	6638C	6776C	16519C			
H3-11	6776C	8470G	16519C				

Appendix 2. (Continued)

H4s - 16263

H4-01/05/06/08	477C	3010A	9150G	16519C		
H4-02	477C	3010A	5945T	9150G	16519C	
H4-03	477C	3010A	9150G	15055C	16519C	
H4-04	477C	3010A	9150G	9380A	16519C	
H4-07	477C	3010A	9150G	9380A	15734A	16519C

H5s - 16304C

H5-01	456T	523A del	524C del	4336C	7691C	15833T		
H5-02	456T	523A del	524C del	4336C	11151T	13023T	15833T	
H5-03	456T	523A del	524C del	4336C	11719A	15833T		
H5-04	456T	523A del	524C del	4336C	15833T			
H5-05	456T	3200C	3438A	16519C				
H5-06	456T	523A del	524C del	4336C	6779G	10084C	12864C	15833T
H5-07	456T	523A del	524C del	4336C	8027A	8803T	15833T	
H5-08	456T	5899.1C	15930A	16519C				
H5-09	456T	2626C	8020A					
H5-10	456T	16519C						
H5-11	456T	513A	523A del	524C del	4336C	5817T		
	9725C	10915C	12384C	15884A	16527T			
H5-12	456T	523A del	524C del	4336C	11719A	11731G	15833T	

H6s - 73G

H6-01	3200C	6776C	7148C	10310A	13896C	16519C	
H6-02/05/06	523A del	524C del	3992T	4024G	5004C		
	8269A	9123A	10044G	14365T	14582G		
H6-03	523A del	524C del	3992T	4024G	5004C	8269A	
	9123A	9997C	10044G	14365T	14582G		
H6-04	761G	6776C	16519C				
H6-07	523A del	524C del	3992T	4024G	5004C	8269A	
	9123A	10044G	13161C	14365T	14582G		
H6-08	523A del	524C del	3992T	4024G	5004C	7058C	8269A
	9123A	10044G	12687T	14365T	14582G	15172A	
H6-09	1842G	11641G	16519C				
H6-10	523A del	524C del	3992T	4024G	5004C	8269A	9123A
	10007C	10034C	10044G	11440A	14365T	14582G	1656G
H6-11	477C	3010A	15325C	16519C			

H7s - 16126G	**16209C**	**73G**				
H7-01/05/07	3010A	6365C	16519C			
H7-02	1555G	3010A	4639C	6365C	10993A	16519C
H7-03	1116G	3010A	6365C	7961C	16519C	
H7-04	3010A	5773A	5922T	6365C	16519C	
H7-06	3010A	4843T	6365C	16519C		

Appendix 2. (Continued)

V1s -	**16298C**							
V1-01	2706G	4580A	7028T	10685A	15904T			
V1-02	72C	2706G	4580A	5894G	7028T	15904T		
V1-03	72C	2706G	4580A	7028T	12453C	15904T		
V1-04	72C	2706G	4580A	7028T	14178C	15904T		
V1-05	72C	2706G	4580A	7028T	11377A	13105G	14770T	15773A
	15904T	16519C						
V1-06/10/15	72C	2706G	4580A	7028T	15904T			
V1-07	72C	290.1A	2706G	3849A	4580A	7028T	11287C	15250T
	15904T	1657T						
V1-08	72C	1018A	2706G	4580A	7028T	13105G	14770T	15904T
	16519C							
V1-09	72C	2706G	4580A	7028T	8289 - 9 bp ins (CCCCCTCTA)			
	13477A	15904T						
V1-11	513A	3342T	10481C	11253C	15924G	16519C		
V1-12	72C	2706G	4580A	5250C	7028T	9088C	12438C	13105G
	14793G	15904T						
V1-13	72C	2706G	4580A	508G	7028T	15904T	16311Y	
V1-14	72C	2706G	4580A	7028T	13105G	14770T	15773A	15904T
	16519C							
V1-16	72C	2706G	4580A	7028T	7270C	8520G	15904T	
V1-17	72C	2706G	4580A	7028T	9088C	12438C	13105G	14793G
	14798C	15904T						
V1-18	72C	2706G	4580A	5250C	7028T	9008G	9088C	12438C
	13105G	14793G	15904T					
V1-19	498.1C	2706G	4580A	4639C	6908C	7028T	7772G	8869G
	15904T	16390A	16519C					
V1-20	72C	2706G	4580A	4639C	5263T	7028T	8869G	15904T
V1-21	72C	2706G	4580A	7028T	14124T	15904T		
V1-22	72C	200R	2706G	4580A	7028T	8134C	15904T	
V1-23	72C	2706G	4550C	4580A	7028T	8347G	12810G	13500C
	15346A	15904T	16519C	16526A				
V1-24	72C	2706G	4580A	7028T	12361G	15904T		
V1-25	72C	2706G	4550C	4580A	7028T	8347G	12810G	13500C
	15346A	15904T	16519C					

Appendix 2. (Continued)

Haplogroup J polymorphisms -								
	462T	**489C**	**2706G**	**3010A**	**4216C**	**7028T**		
	10398G	**11251G**	**11719A**	**12612G**	**13708A**	**14766T**	**15452A**	
J1s -	**73G**	**185A**	**228A**	**295T**	**16069T**	**16126C**		
J1-01	6293C	11233C	13145A	13934T	14296G	14798C	16390A	
J1-02	4025T	5951G	9738A	14798C	16168Y			
J1-03/06	9632G	12083G	14798C					
J1-04	482C	3394C	8468T	10454C	14798C	16368C		
J1-05	4838C	5198G	9100G	14798C				
J1-07	1811G	13934T	14798C					
J1-08	482C	735G	930A	3394C	3834A	7184G	14798C	
J1-09	4454C	14798C	16519C					
J1-10	522D	523D	4688C	5198G	5978G			
	7340A	7888T	13434G	14798C				
J1-11	6887T	12813T	14798C					
J1-12	8839A	14798C						
J1-13	1811G	2772Y	3540C	13934T	14798C			
J1-14	482C	3394C	10454C	14798C	15769G	16368C		
J1-15	4025T	14502C	14798C					
J2s -	**73G**	**228A**	**295T**	**16069T**	**16126C**			
J2-01	482C	3394C	9635C	11623T	13899C	14798C		
J2-02	7711C	9548A	13934T	14798C				
J2-03	5198G	14798C						
J2-04/08	9548A	9836C	13934T	14323A	14798C	15355A		
J2-05	5442C	12858T	13934T	14798C	15758G			
J2-06	709A	1891G	6221C	13934T	14180C	14798C		
J2-07	482C	3394C	9635C	11623T	13899C	14798C		
J3s -	**73G**	**185A**	**188G**	**228A**	**295T**	**16069T**	**16126C**	
J3-01/04/11/13	14798C	16519C						
J3-02	522D	523D	8285T	14798C	16519C			
J3-03	4790G	6260A	6293C	14798C	15043A	16519C		
J3-05	4454C	5033G	7674C	14798C	16093Y	16519C		
J3-06	7891T	14798C	15199T	15936T	16519C			
J3-07	522D	523D	12397G	14798C	16519C			
J3-08	3834R	11177T	14798C	16519C				
J3-09	8558T	11083G	12557T	13557G	14798C	16519C		
J3-10	3351T	6293C	7245G	8839A	13734C	14798C	15047A	16519C
J3-12	6293C	7245G	8839A	9181G	14798C	15909G	16519C	

Appendix 2. (Continued)

J4s -	**73G**	**242T**	**295T**	**16069T**	**16126C**			
	16145A	**16172C**	**16222T**	**16261T**				
J4-01	9T	10C	2158C	5460A	8269A			
	8557A	12007A	12612G	13879C				
J4-02	2158C	5460A	8269A	8557A	12007A			
	12311C	12612G	13879C	15067C	15467G			
J4-03/05	2158C	5460A	8269A	8557A	12007A			
	12612G	13879C	15067C	16519C				
J4-04	2158C	4679C	5460A	8269A	8557A	12007A	12612G	13879C
J4-06	2157C	2158C	5460A	5705G	8269A	8557A		
	12007A	12612G	13879C	15067C	15735T	16519C		
J4-07	2158C	5298G	5460A	8269A	8557A			
	12007A	12612G	13879C	15067C	16519C			
J4-08	185A	1462A	2158C	5460A	6345C	7299G		
	8269A	8557A	12007A	12612G	13879C			

Haplogroup T polymorphisms -

	709A	**1888A**	**2706G**	**4216C**	**4917G**	**7028T**	**8697A**	**10463C**
	11251G	**11719A**	**13368A**	**14766T**	**14905A**	**15452A**	**15607G**	**15928A**
T1s	**16126C**	**16294T**	**16296T**	**16304C**	**73G**			
T1-01	930A	4695C	5147A	6032A	11812G	14016A	14233G	16519C
T1-02	930A	5147A	11812G	12681Y	12699G	14233G	16519C	
T1-03	930A	5147A	11812G	12795A	14233G	16519C		
T1-04/06/07 11/13/20	930A	5147A	11812G	14233G	16260Y	16519C		
T1-05	930A	5147A	7897A	11812G	14233G	14798C	16519C	
T1-08	930A	5147A	11812G	14233G	15773A	16519C		
T1-09	930A	5147A	11812G	13722G	14233G	16519C		
T1-10	930A	4700T	5147A	8433C	9455G			
	11812G	12378T	14233G	16519C				
T1-12	930A	5147A	7289G	11812G	14233G	16519C		
T1-14	930A	4688C	5147A	7891T	11812G	13692T	14233G	16519C
T1-15	930A	5147A	11812G	12397G	14233G	15043A	16519C	
T1-17	458T	930A	1709A	5147A	9300A	11533T		
	11812G	12007A	12771A	14233G	16519C			
T1-18	930A	8504C	5147A	11812G	14233G	16519C		
T1-19	458T	930A	1709A	5147A	9300A			
	11533T	11812G	14233G	16519C				
T1-21/22	524.1A	524.2C	930A	3826C	5147A	11812G	14233G	16519C

Appendix 2. (Continued)

T2s	**16126C**	**16163G**	**16186T**	**16189C**	**16294T**	**73G**	**152C**	**195C**
T2-01	9145A	9899C	12633A	16519C				
T2-02/06	9899C	12633A	16519C					
T2-03	6077T	9899C	12633A	13056T	15924G	16519C		
T2-04	8530G	9899C	12633A	16519C				
T2-05	1719R	9899C	12633A	15341C	16519C			
T2-07	5414G	5558G	9899C	12633A	15412C	16519C		
T2-08	7852A	8654C	9899C	12633A	16519C			
T2-09	7258C	10321C	12633A	16519C				
T2-10	9899C	12633A	14106C	16519C				

T3s	**16126C**	**16294T**	**16296T**	**73G**				
T3-01/04	2850C	7022C	11812G	13965C	14233G	14687G	16519C	
T3-02	930A	5147A	11242G	11812G	14233G	16519C		
T3-03	2850C	4808T	5498G	7022C	8435G	9371T		
	11251G	11812G	13965C	14233G	14687G	16519C		
T3-05	2850C	4768Y	7022C	11812G	13965C			
	14233G	14687G	15935G	16519C				
T3-06	2850C	4808T	5021Y	5498G	7022C	8435G		
	11812G	13965C	14233G	14687G	16519C			
T3-07	2850C	6632C	7022C	11812G	13965C	14233G	14687G	16519C
T3-08	930A	5147A	10961T	11242G	11812G			
	11914A	14233G	15452A	16519C				

Haplogroup K polymorphisms -	**1811G**	**2706G**	**3480G**	**7028T**	**9055A**
9698C	**10550G**	**11299C**	**11467G**	**11719A**	**12308G**	**12372A**
14167T	**14766T**	**14798C**	**16519C**			

K1s-	**16224C**	**16311C**	**73G**	**146C**	**152C**			
K1-01/02/06/10	709A	4561C	9716C	14305A				
K1-03/09	709A	4561C	9716C					
K1-04	709A	4561C	9716C	11549T				
K1-05	498D	1189C	7759C	9093G	10398G	11377A	15900C	
K1-07	498D	523A del	523C del	1189C	1393R			
	9093G	10398G	10698T	11377A				
K1-08	63C	64T	709A	4561C	8697A	9716C	15520G	
K1-11	709A	4561C	9716C	12795A	13143C	13293T	13708A	
K1-12/13	498D	1189C	9093G	10398G	11377A	15900C		
K1-14	709A	4561C	4688C	9716C	9801A	13293T		

Appendix 2. (Continued)

K2s-	**16093C**	**16224C**	**16311C**	**73G**				
K2-02	497T	524.1A	524.2C	723C	10398G			
	11071T	11485C	1189C	15355A				
K2-03	497T	3816G	10398G	11025C	1189C	12477C	16399G	
K2-04/08	497T	524.1A	524.2C	723C	1189C			
	10398G	11071T	11485C	15355A				
K2-05	497T	5324T	1189C	4216Y	5580C			
	7559G	10398G	13117G	16422C				
K2-06	497T	1189C	3552C	4959A	6060G			
	8276.1C	10398G	11914A	15799G				
K2-07	497T	524.1A	524.2C	1189C	7559G			
	8440G	8522T	10398G	11151T	13117G			

K3s-	**16224C**	**16311C**	**73G**					
K3-01	524.1A	524.2C	523.3A	523.4C	1189C	5913A	6053T	7391C
	8164T	9007G	9962A	10289G	10398G	14063C	14384A	15946T
K3-02/04	497T	524.1A	524.2C	523.3A	523.4C	1189C	10398G	11485C
K3-03	497T	524.1A	524.2C	1189C	6260A			
	10398G	11485C	11840T	13740C				
K3-05	497T	1189C	10398G	13117G				
K3-06	497T	524.1A	524.2C	1189C	4295G	5711G	6260A	10398G
	10586A	11485C	11840T	13740C	13886C	14502C	15884A	
K3-07	497T	1189C	3394C	5093C	6260A			
	10398G	11485C	11840T	13740C				

Appendix 3. Characterization of the Polymorphic Sites Identified in the 214 common HV1/HV2 mtGenomes. Each polymorphism identified among the 241 common HV1/HV2 mtGenomes are listed according to the nucleotide position (site), the gene or region within the mtGenome, the amino acid codon affected (and change, according to the one letter amino acid abbreviation), if the change represents a synonymous or non-synonymous amino acid change, if the polymorphism been previously observed in the literature, if the polymorphism associated with a disease, if the polymorphism was selected for multiplex SNP assay development (see Table 5) or if not, the reason, and the common types the polymorphism was useful for resolving. The 59 Polymorphisms selected for multiplex SNP development are in bold text. Abbreviations for disease-associated polymorphisms include: DM – Diabetes Mellitus; DEAF – Maternally inherited Deafness or Aminoglycoside-induced Deafness; MHCM – Maternally inherited Hypertonic CardioMyopathy; ADPD – Alzheimer's Disease and Parkinson's Disease; LHON – Leber's Hereditary Optic Neuropathy; AD - Alzheimer's Disease; GER/SIDS – GastroEsophogeal Reflux disease/Sudden Infant Death Syndrome; MELAS – Mitochondrial Encephalomyopathy, Lactic Acidosis, and Stroke-like episodes; CPEO – Chronic Progressive External Ophthalmoplegia; LIMM – Lethal Infantile Mitochondrial Myopathy. Polymorphisms not selected as discriminatory SNPs among the 241 common HV1/HV2 mtGenomes are noted if the polymorphism created a potential phenotypic change in the amino acid, tRNA, or rRNA (Pheno. Δ); were observed only once among the 241 mtGenomes (Singleton); were shared among all members of a particular haplogroup/haplogroups, making them uninformative (Uninform.); or were redundant with other polymorphic sites (Redundant). Polymorphisms that were used to determine the common HV1/HV2 types are listed as "Diagnostic."

Site	Gene/ Region	(Codon) AA Δ	Syn/ Non-Syn.	Prev. Obs.	Disease	SNP (reason)	Common Types
3	CR	-	-	No	-	No (Singleton)	T1
9	CR	-	-	Yes	-	No (Singleton)	J4
10	CR	-	-	No	-	No (Singleton)	J4
55	CR	-	-	No	-	No (Singleton)	H3
57	CR	-	-	No	-	No (Singleton)	H3
63	CR	-	-	Yes	-	No (Singleton)	K1
64	**CR**	**-**	**-**	Yes	**-**	**YES**	**H2/K1**
72	**CR**	**-**	**-**	Yes	**-**	**YES**	**H1/V1**
73	CR	-	-	Yes	-	No (Diagnostic)	-
146	CR	-	-	Yes	-	No (Diagnostic)	-
152	CR	-	-	Yes	-	No (Diagnostic)	-
185	CR	-	-	Yes	-	No (Diagnostic)	-
188	CR	-	-	Yes	-	No (Diagnostic)	-
195	CR	-	-	Yes	-	No (Diagnostic)	-
200	CR	-	-	Yes	-	No (Singleton)	V1
204	CR	-	-	Yes	-	No (Singleton)	T1
228	CR	-	-	Yes	-	No (Diagnostic)	-
242	CR	-	-	Yes	-	No (Diagnostic)	-
263	CR	-	-	Yes	-	No (Diagnostic)	All - rCRS
291.1 A	CR	-	-	No	-	No (Singleton)	V1
295	CR	-	-	Yes	-	No (Diagnostic)	-
456	CR	-	-	Yes	-	No (Uninform.)	H5
458	CR	-	-	Yes	-	No (Uninform.)	T1
460	CR	-	-	No	-	No (Singleton)	H3
462	CR	-	-	Yes	-	No (Uninform.)	All J's
477	**CR**	**-**	**-**	**Yes**	**-**	**YES**	**H1/H2/H6**
482	**CR**	**-**	**-**	**Yes**	**-**	**YES**	**J1/J2**
489	CR	-	-	Yes	-	No (Uninform.)	All J's
497	CR	-	-	Yes	-	No (Uninform.)	All K2s K3s
498	CR	-	-	Yes	-	No (Redundant)	K1
498.1 C	CR	-	-	No	-	No (Singleton)	V1
508	CR	-	-	Yes	-	No (Singleton)	V1
513	**CR**	**-**	**-**	**Yes**	**-**	**YES**	**H5/V1**
523 del	**CR**	**-**	**-**	**Yes**	**-**	**YES**	**H's J's K's T's**
524 del	**CR**	**-**	**-**	**Yes**	**-**	**YES**	**"**
524.1 A	**CR**	**-**	**-**	**Yes**	**-**	**YES**	**H's K's T's**
524.2 C	**CR**	**-**	**-**	**Yes**	**-**	**YES**	**"**
524.3 A	**CR**	**-**	**-**	**Yes**	**-**	**YES**	**K3**
524.4 C	**CR**	**-**	**-**	**Yes**	**-**	**YES**	**"**
593	tPhe	-	-	Yes	No	No (Pheno. Δ)	H3
709	12S	-	-	Yes	No	No (Pheno. Δ)	K1/J2
723	12S	-	-	Yes	No	No (Pheno. Δ)	K2
735	12S	-	-	Yes	No	No (Pheno. Δ)	J1
739	12S	-	-	No	No	No (Pheno. Δ)	H2
750	12S	-	-	Yes	No	No (Pheno. Δ)	All - rCRS
761	12S	-	-	No	No	No (Pheno. Δ)	H6
930	12S	-	-	Yes	No	No (Pheno. Δ)	J1
1018	12S	-	-	Yes	No	No (Pheno. Δ)	V1

Appendix 3. (Continued)

Site	Gene/ Region	(Codon) AA Δ	Syn/ Non-Syn.	Prev. Obs.	Disease	SNP (reason)	Common Types
1116	12S	-	-	No	No	No (Pheno. Δ)	H7
1189	12S	-	-	Yes	No	No (Pheno. Δ)	K2/K3
1393	12S	-	-	Yes	No	No (Pheno. Δ)	K1
1438	12S	-	-	Yes	DM	No (Pheno. Δ)	All - rare rCRS
1462	12S	-	-	Yes	No	No (Pheno. Δ)	J4
1555	12S	-	-	Yes	DEAF	No (Pheno. Δ)	H7
1656	12S	-	-	No	No	No (Pheno. Δ)	H6
1657	tVal	-	-	No	No	No (Pheno. Δ)	V1
1709	16S	-	-	Yes	No	No (Pheno. Δ)	T1
1719	16S	-	-	Yes	No	No (Pheno. Δ)	H1/T2
1811	16S	-	-	Yes	No	No (Pheno. Δ)	J1
1842	16S	-	-	Yes	No	No (Pheno. Δ)	H6
1888	16S	-	-	Yes	No	No (Pheno. Δ)	All T's
1891	16S	-	-	No	No	No (Pheno. Δ)	J2
2157	16S	-	-	No	No	No (Pheno. Δ)	J4
2158	16S	-	-	Yes	No	No (Pheno. Δ)	J4
2232.1 A	16S	-	-	No	No	No (Pheno. Δ)	T1
2259	16S	-	-	Yes	No	No (Pheno. Δ)	H1/H2/H3
2325.1 T	16S	-	-	No	No	No (Pheno. Δ)	T1
2581	16S	-	-	Yes	No	No (Pheno. Δ)	H3
2626	16S	-	-	Yes	No	No (Pheno. Δ)	H5
2706	16S	-	-	Yes	No	No (Pheno. Δ)	H1/V1
2772	16S	-	-	Yes	No	No (Pheno. Δ)	H2/J1
2850	16S	-	-	Yes	No	No (Pheno. Δ)	T3
2851	16S	-	-	Yes	No	No (Pheno. Δ)	H1
3010	**16S**	**-**	**-**	**Yes**	**No**	**YES**	**H1/H2/H3/H6**
3200	16S	-	-	Yes	No	No (Pheno. Δ)	H5/H6
3221	16S	-	-	Yes	No	No (Pheno. Δ)	H3
3342	ND1	(3) none	Syn.	No	No	No (Singleton)	V1
3351	ND1	(3) none	Syn.	Yes	No	No (Singleton)	J3
3394	ND1	(1) Y-H	Non-Syn.	Yes	No	No (Redundant)	J1/J2/K3
3438	ND1	(3) none	Syn.	Yes	No	No (Singleton)	H5
3480	ND1	(3) none	Syn.	Yes	No	No (Uninform.)	All K's
3511	ND1	(1) T-A	Non-Syn.	No	No	No (Pheno. Δ)	H2
3540	ND1	(3) none	Syn.	No	No	No (Singleton)	J1
3552	ND1	(3) none	Syn.	Yes	No	No (Singleton)	K2
3621	ND1	(3) none	Syn.	No	No	No (Singleton)	H3
3640	ND1	(1) A-T	Non-Syn.	No	No	No (Pheno. Δ)	H2
3816	ND1	(3) none	Syn.	Yes	No	No (Singleton)	K2
3826	**ND1**	**(1) none**	**Syn.**	**Yes**	**No**	**YES**	**T1**
3834	**ND1**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**J1/J3**
3849	ND1	(3) none	Syn.	Yes	No	No (Singleton)	V1
3915	**ND1**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H3**
3918	ND1	(3) none	Syn.	Yes	No	No (Singleton)	H2
3970	ND1	(1) none	Syn.	Yes	No	No (Singleton)	H2
3992	ND1	(2) T-M	Non-Syn.	Yes	No	No (Redundant)	H1/H2/H6

Appendix 3. (Continued)

Site	Gene/ Region	(Codon) AA Δ	Syn/ Non-Syn.	Prev. Obs.	Disease	SNP (reason)	Common Types
4024	ND1	(1) T-A	Non-Syn.	Yes	No	No (Redundant)	H1/H2/H6
4025	ND1	(2) T-M	Non-Syn.	Yes	No	No (Pheno. Δ)	J1
4048	ND1	(1) D-N	Non-Syn.	Yes	No	No (Pheno. Δ)	H1
4080	ND1	(3) none	Syn.	No	No	No (Singleton)	H2
4216	ND1	(1) T-H	Non-Syn.	Yes	No	No (Pheno. Δ)	H1/K2
4295	tIle	-	-	Yes	MHCM	No (Pheno. Δ)	K3
4336	tGln	-	-	Yes	ADPD	No (Pheno. Δ)	H5
4418	tMet	-	-	Yes	No	No (Pheno. Δ)	H2
4454	tMet	-	-	Yes	No	No (Pheno. Δ)	J1/J3
4550	ND2	(3) none	Syn.	No	No	No (Redundant)	V1
4561	ND2	(2) V-A	Non-Syn.	Yes	No	No (Pheno. Δ)	K1
4580	**ND2**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H1/V1**
4639	ND2	(2) I-T	Non-Syn.	Yes	No	No (Pheno. Δ)	H7/V1
4646	ND2	(3) none	Syn.	Yes	No	No (Singleton)	H2
4679	ND2	(3) none	Syn.	Yes	No	No (Singleton)	J4
4688	**ND2**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**J1/K1/T1**
4695	ND2	(1) F-L	Non-Syn.	Yes	No	No (Pheno. Δ)	T1
4700	ND2	(3) none	Syn.	Yes	No	No (Singleton)	T1
4707	ND2	(1) L-I	Non-Syn.	Yes	No	No (Pheno. Δ)	H1
4733	ND2	(3) none	Syn.	Yes	No	No (Singleton)	H1
4745	**ND2**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H1/H2/H3**
4768	ND2	(2) M-T	Non-Syn.	No	No	No (Pheno. Δ)	T3
4769	ND2	(3) none	Syn.	Yes	No	No (Uninform.)	All - rCRS
4790	ND2	(3) none	Syn.	No	No	No (Singleton)	J3
4793	**ND2**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H1**
4808	**ND2**	**(3) none**	**Syn.**	**No**	**No**	**YES**	**T3**
4823	ND2	(3) none	Syn.	Yes	No	No (Singleton)	H2
4838	ND2	(3) none	Syn.	No	No	No (Singleton)	J1
4843	ND2	(2) T-M	Non-Syn.	Yes	No	No (Pheno. Δ)	H7
4917	ND2	(1) N-D	Non-Syn.	Yes	LHON	No (Pheno. Δ)	All T's
4959	ND2	(1) A-T	Non-Syn.	Yes	No	No (Pheno. Δ)	K2
4976	ND2	(3) none	Syn.	Yes	No	No (Singleton)	H1
5004	**ND2**	**(1) none**	**Syn.**	**Yes**	**No**	**YES**	**H1/H2/H6**
5021	ND2	(3) none	Syn.	No	No	No (Singleton)	T3
5033	ND2	(3) none	Syn.	No	No	No (Singleton)	J3
5093	ND2	(3) none	Syn.	No	No	No (Singleton)	K3
5147	**ND2**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**T3**
5198	**ND2**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**J1/J2**
5213	ND2	(3) none	Syn.	No	No	No (Singleton)	H2
5237	ND2	(3) none	Syn.	Yes	No	No (Singleton)	H1
5250	**ND2**	**(1) none**	**Syn.**	**No**	**No**	**YES**	**V1**
5263	ND2	(2) A-V	Non-Syn.	Yes	No	No (Pheno. Δ)	V1
5298	ND2	(1) I-V	Non-Syn.	No	No	No (Pheno. Δ)	J4
5324	ND2	(3) none	Syn.	Yes	No	No (Singleton)	K2
5348	ND2	(3) none	Syn.	Yes	No	No (Singleton)	H1
5414	ND2	(3) none	Syn.	Yes	No	No (Singleton)	T2
5442	ND2	(1) F-L	Non-Syn.	Yes	No	No (Pheno. Δ)	J2
5460	ND2	(1) A-T	Non-Syn.	Yes	AD	No (Pheno. Δ)	H1/H2

Appendix 3. (Continued)

Site	Gene/ Region	(Codon) AA Δ	Syn/ Non-Syn.	Prev. Obs.	Disease	SNP (reason)	Common Types
5498	ND2	(3) none	Syn.	No	No	No (Redundant)	T3
5558	tTrp	-	-	No	No	No (Pheno. Δ)	T2
5580	NC	none	-	Yes	No	No (Singleton)	K2
5582	NC	none	-	Yes	No	No (Singleton)	H1
5654	tAla	-	-	Yes	No	No (Pheno. Δ)	H2
5705	tAsp	-	-	No	No	No (Pheno. Δ)	J4
5711	tAsn	-	-	Yes	No	No (Pheno. Δ)	K3
5772	tCys	-	-	No	No	No (Pheno. Δ)	H2
5773	tCys	-	-	Yes	No	No (Pheno. Δ)	H7
5817	tCys	-	-	No	No	No (Pheno. Δ)	H5
5893	NC	none	-	Yes	No	No (Singleton)	H2
5894	NC	none	-	Yes	No	No (Singleton)	V1
5899.1 C	NC	none	-	No	No	No (Singleton)	H5
5913	COI	(1) D-N	Non-Syn.	Yes	No	No (Pheno. Δ)	K3
5922	COI	(1) none	Syn.	No	No	No (Singleton)	H7
5945	COI	(3) none	Syn.	No	No	No (Singleton)	H4
5951	COI	(3) none	Syn.	Yes	No	No (Singleton)	J1
5978	COI	(3) none	Syn.	No	No	No (Singleton)	J1
6032	COI	(3) none	Syn.	No	No	No (Singleton)	T1
6053	COI	(3) none	Syn.	No	No	No (Singleton)	K3
6060	COI	(1) I-V	Non-Syn.	No	No	No (Pheno. Δ)	K2
6077	COI	(3) none	Syn.	Yes	No	No (Singleton)	T2
6177	COI	(1) M-V	Non-Syn.	No	No	No (Pheno. Δ)	H1
6216	COI	(1) none	Syn.	Yes	No	No (Singleton)	H1
6221	COI	(3) none	Syn.	Yes	No	No (Singleton)	J2
6253	COI	(2) M-T	Non-Syn.	Yes	No	No (Pheno. Δ)	H3
6260	**COI**	**(3) none**	**Syn.**	Yes	**No**	**YES**	**J3/K3**
6293	**COI**	**(3) none**	**Syn.**	Yes	**No**	**YES**	**J1/J3**
6296	COI	(3) none	Syn.	Yes	No	No (Singleton)	H3
6345	COI	(1) F-L	Non-Syn.	No	No	No (Pheno. Δ)	J4
6365	COI	(3) none	Syn.	Yes	No	No (Redundant)	H7
6575	COI	(3) none	Syn.	No	No	No (Singleton)	H1
6632	COI	(3) none	Syn.	Yes	No	No (Singleton)	T3
6638	COI	(3) none	Syn.	No	No	No (Singleton)	H3
6776	**COI**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H2/H3/H6**
6779	COI	(3) none	Syn.	No	No	No (Singleton)	H5
6887	COI	(3) none	Syn.	No	No	No (Singleton)	J1
6908	COI	(3) none	Syn.	No	No	No (Singleton)	V1
6956	COI	(3) none	Syn.	Yes	No	No (Singleton)	H1
7022	COI	(3) none	Syn.	Yes	No	No (Redundant)	T3
7028	**COI**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H1/V1**
7058	COI	(3) none	Syn.	Yes	No	No (Singleton)	H6
7148	COI	(3) none	Syn.	No	No	No (Singleton)	H6
7184	COI	(3) none	Syn.	Yes	No	No (Singleton)	J1
7202	**COI**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H1**
7245	COI	(1) T-A	Non-Syn.	No	No	No (Pheno. Δ)	J3
7258	COI	(2) I-T	Non-Syn.	Yes	No	No (Pheno. Δ)	T2
7270	COI	(2) V-A	Non-Syn.	Yes	No	No (Pheno. Δ)	V1

Appendix 3. (Continued)

Site	Gene/ Region	(Codon) AA Δ	Syn/ Non-Syn.	Prev. Obs.	Disease	SNP (reason)	Common Types
7289	COI	(3) none	Syn.	No	No	No (Singleton)	T1
7299	COI	(1) M-V	Non-Syn.	Yes	No	No (Pheno. Δ)	J4
7337	COI	(3) none	Syn.	Yes	No	No (Singleton)	H2
7340	COI	(3) none	Syn.	Yes	No	No (Singleton)	J1
7391	COI	(3) none	Syn.	No	No	No (Singleton)	K3
7559	tAsp	-	-	Yes	No	No (Pheno. Δ)	K2
7674	COII	(2) I-T	Non-Syn.	Yes	No	No (Pheno. Δ)	J3
7691	COII	(1) F-L	Non-Syn.	No	No	No (Pheno. Δ)	H5
7711	COII	(3) none	Syn.	Yes	No	No (Singleton)	J2
7741	COII	(3) none	Syn.	No	No	No (Singleton)	H2
7759	COII	(3) none	Syn.	Yes	No	No (Singleton)	K1
7772	COII	(1) T-A	Non-Syn.	Yes	No	No (Pheno. Δ)	V1
7852	COII	(3) none	Syn.	No	No	No (Singleton)	T2
7888	COII	(3) none	Syn.	No	No	No (Singleton)	J1
7891	**COII**	**(3) none**	**Syn.**	**No**	**No**	**YES**	**J3/T1**
7897	COII	(3) none	Syn.	Yes	No	No (Singleton)	T1
7951	COII	(3) none	Syn.	No	No	No (Singleton)	H3
7961	COII	(1) none	Syn.	No	No	No (Singleton)	H7
7975	COII	(3) none	Syn.	No	No	No (Singleton)	H1
8020	COII	(3) none	Syn.	Yes	No	No (Singleton)	H5
8027	COII	(1) A-T	Non-Syn.	Yes	No	No (Pheno. Δ)	H5
8134	COII	(3) none	Syn.	No	No	No (Singleton)	V1
8164	COII	(3) none	Syn.	No	No	No (Singleton)	K3
8269	COII	(3) none	Syn.	Yes	No	No (Redundant)	H1/H6
8276.1	NC	none	-	No	No	No (Singleton)	K2
8285	NC	none	-	No	No	No (Singleton)	J3
8288.1-9	NC	none	-	No	No	No (Singleton)	V1
8308	tLys	-	-	Yes	No	No (Pheno. Δ)	H1
8347	tLys	-	-	No	No	No (Pheno. Δ)	V1
8433	ATP8	(2) I-T	Non-Syn.	No	No	No (Pheno. Δ)	H2/T1
8435	ATP8	(1) T-A	Non-Syn.	Yes	No	No (Pheno. Δ)	T3
8440	ATP8	(3) none	Syn.	Yes	No	No (Singleton)	K2
8468	ATP8	(1) none	Syn.	Yes	No	No (Pheno. Δ)	J1
8470	ATP8	(3) none	Syn.	Yes	No	No (Singleton)	H3
8504	ATP8	(1) Y-H	Non-Syn.	No	No	No (Pheno. Δ)	T1
8520	ATP8	(2) E-G	Non-Syn.	No	No	No (Pheno. Δ)	V1
8522	ATP8	(1) P-S	Non-Syn.	Yes	No	No (Pheno. Δ)	K2
8530	ATP8	(3) none	**	Yes	No	No (Pheno. Δ)	T2
	ATP6	(1) N-D	**				
8557	ATP8	(3) none	**	Yes	No	No (Pheno. Δ)	J4
	ATP6	(1) A-T	**				
8558	ATP8	(1) P-S	Non-Syn.	Yes	No	No (Pheno. Δ)	J3
	ATP6	(2) A-V	Non-Syn.				
8592	**ATP6**	**(3) none**	**Syn.**	**No**	**No**	**YES**	**H2**
8645	ATP6	(2) N-S	Non-Syn.	No	No	No (Pheno. Δ)	H2
8654	ATP6	(2) I-T	Non-Syn.	No	No	No (Pheno. Δ)	T2
8697	ATP6	(3) none	Syn.	Yes	No	No (Redundant)	K1's/All T's
8701	ATP6	(1) T-A	Non-Syn.	Yes	No	No (Pheno. Δ)	H3

Appendix 3. (Continued)

Site	Gene/ Region	(Codon) AA Δ	Syn/ Non-Syn.	Prev. Obs.	Disease	SNP (reason)	Common Types
8705	ATP6	(2) M-T	Non-Syn.	Yes	No	No (Pheno. Δ)	H3
8764	ATP6	(1) A-T	Non-Syn.	Yes	No	No (Pheno. Δ)	H1
8784	ATP6	(3) none	Syn.	Yes	No	No (Singleton)	H1
8803	ATP6	(1) T-A	Non-Syn.	Yes	No	No (Pheno. Δ)	H1
8803	ATP6	(1) T-S	Non-Syn.	No	No	No (Pheno. Δ)	H5
8812	ATP6	(1) T-A	Non-Syn.	No	No	No (Pheno. Δ)	H1
8839	ATP6	(1) A-T	Non-Syn.	Yes	No	No (Pheno. Δ)	J1/J3
8860	ATP6	(1) T-A	Non-Syn.	Yes	No	No (Pheno. Δ)	All - rCRS
8869	ATP6	(1) M-V	Non-Syn.	Yes	No	No (Pheno. Δ)	V1
8975	ATP6	(2) L-P	Non-Syn.	No	No	No (Pheno. Δ)	H2
9007	ATP6	(1) T-A	Non-Syn.	Yes	No	No (Pheno. Δ)	K3
9008	ATP6	(2) T-S	Non-Syn.	No	No	No (Pheno. Δ)	V1
9025	ATP6	(1) G-S	Non-Syn.	No	No	No (Pheno. Δ)	H2
9055	ATP6	(1) A-T	Non-Syn.	Yes	No	No (Pheno. Δ)	All K's
9088	ATP6	(1) S-P	Non-Syn.	No	No	No (Pheno. Δ)	V1
9091	ATP6	(1) T-A	Non-Syn.	No	No	No (Pheno. Δ)	H1/H2
9093	ATP6	(3) none	Syn.	Yes	No	No (Redundant)	K1
9100	ATP6	(1) I-V	Non-Syn.	No	No	No (Pheno. Δ)	J1
9123	ATP6	(3) none	Syn.	Yes	No	No (Redundant)	H1/H2/H6
9129	ATP6	(3) none	Syn.	Yes	No	No (Singleton)	H1
9145	ATP6	(1) A-T	Non-Syn.	Yes	No	No (Pheno. Δ)	T2
9150	ATP6	(3) none	Syn.	Yes	No	No (Uninform.)	All H4's
9181	ATP6	(1) none	Syn.	Yes	No	No (Singleton)	J3
9290	COIII	(3) none	Syn.	No	No	No (Singleton)	H1
9300	COIII	(1) A-T	Non-Syn.	Yes	No	No (Pheno. Δ)	T1
9371	COIII	(3) none	Syn.	No	No	No (Singleton)	T3
9380	**COIII**	**(3) none**	**Syn.**	**Yes**	No	**YES**	**H4**
9455	COIII	(3) none	Syn.	No	No	No (Singleton)	T1
9548	**COIII**	**(3) none**	**Syn.**	**Yes**	No	**YES**	**J2**
9581	COIII	(3) none	Syn.	No	No	No (Singleton)	H3
9632	COIII	(3) none	Syn.	Yes	No	No (Redundant)	J1
9635	**COIII**	**(3) none**	**Syn.**	**No**	No	**YES**	**J2**
9698	COIII	(3) none	Syn.	Yes	No	No (Uninform.)	All K's
9716	COIII	(3) none	Syn.	Yes	No	No (Redundant)	K1
9725	COIII	(3) none	Syn.	No	No	No (Singleton)	H5
9738	COIII	(1) A-T	Non-Syn.	No	No	No (Pheno. Δ)	J1
9801	COIII	(1) V-M	Non-Syn.	Yes	No	No (Pheno. Δ)	K1
9836	COIII	(3) none	Syn.	Yes	No	No (Singleton)	J2
9899	**COIII**	**(3) none**	**Syn.**	**Yes**	No	**YES**	**T2**
9921	COIII	(1) A-T	Non-Syn.	No	No	No (Pheno. Δ)	H1
9923	COIII	(3) none	Syn.	Yes	No	No (Pheno. Δ)	H1
9962	COIII	(3) none	Syn.	Yes	No	No (Pheno. Δ)	K3
9997	tGly	-	-	No	MHCM	No (Pheno. Δ)	H6
10007	tGly	-	-	Yes	No	No (Pheno. Δ)	H6
10034	tGly	-	-	Yes	No	No (Pheno. Δ)	H6
10044	tGly	-	-	Yes	GER/SIDS	No (Pheno. Δ)	H1
10049	tGly	-	-	No	No	No (Pheno. Δ)	H1
10084	ND3	(2) I-T	Non-Syn.	Yes	No	No (Pheno. Δ)	H5

Appendix 3. (Continued)

Site	Gene/ Region	(Codon) AA Δ	Syn/ Non-Syn.	Prev. Obs.	Disease	SNP (reason)	Common Types
10211	**ND3**	**(3) none**	**Syn.**	**Yes**	No	**YES**	**H1/H2**
10256	ND3	(3) none	Syn.	Yes	No	No (Singleton)	H1
10289	ND3	(3) none	Syn.	Yes	No	No (Singleton)	K3
10310	ND3	(3) none	Syn.	Yes	No	No (Singleton)	H6
10321	ND3	(2) V-A	Non-Syn.	Yes	No	No (Pheno. Δ)	T2
10394	**ND3**	**(3) none**	**Syn.**	**Yes**	No	**YES**	**H1/H2**
10398	ND3	(1) T-A	Non-Syn.	Yes	No	No (Pheno. Δ)	K1
10454	tArg	-	-	Yes	No	No (Pheno. Δ)	J1
10463	tArg	-	-	Yes	No	No (Pheno. Δ)	All T's
10481	ND4L	(3) none	Syn.	No	No	No (Singleton)	V1
10550	ND4L	(3) none	Syn.	Yes	No	No (Uninform.)	All K's
10586	ND4L	(3) none	Syn.	Yes	No	No (Singleton)	K3
10685	**ND4L**	**(3) none**	**Syn.**	**Yes**	No	**YES**	**V1/H2**
10698	ND4L	(1) none	Syn.	No	No	No (Singleton)	K1
10750	ND4L	(2) N-S	Non-Syn.	Yes	No	No (Pheno. Δ)	T1
10754	**ND4L**	**(3) none**	**Syn.**	**Yes**	No	**YES**	**H2**
10846	ND4	(3) none	Syn.	No	No	No (Singleton)	H1
10915	ND4	(3) none	Syn.	Yes	No	No (Singleton)	H5
10961	ND4	(1) none	Syn.	Yes	No	No (Singleton)	T3
10993	ND4	(3) none	Syn.	Yes	No	No (Singleton)	H7
11020	ND4	(3) none	Syn.	No	No	No (Singleton)	H2
11025	ND4	(2) L-P	Non-Syn.	No	No	No (Singleton)	K2
11050	ND4	(3) none	Syn.	No	No	No (Singleton)	H3
11071	ND4	(3) none	Syn.	No	No	No (Redundant)	K2
11083	ND4	(3) none	Syn.	No	No	No (Singleton)	J3
11084	ND4	(1) T-A	Non-Syn.	Yes	MELAS	No (Pheno. Δ)	H1
11151	ND4	(2) A-V	Non-Syn.	No	No	No (Pheno. Δ)	H5/K2
11177	ND4	(1) P-S	Non-Syn.	Yes	No	No (Pheno. Δ)	J3
11182	ND4	(3) none	Syn.	No	No	No (Singleton)	H2
11204	ND4	(1) F-L	Non-Syn.	Yes	No	No (Pheno. Δ)	H3
11233	ND4	(3) none	Syn.	Yes	No	No (Singleton)	J1
11242	ND4	(3) none	Syn.	Yes	No	No (Redundant)	T3
11251	ND4	(3) none	Syn.	Yes	No	No (Uninform.)	All J's/T's
11253	ND4	(2) I-T	Non-Syn.	Yes	No	No (Pheno. Δ)	V1
11287	ND4	(3) none	Syn.	No	No	No (Singleton)	V1
11299	ND4	(3) none	Syn.	Yes	No	No (Uninform.)	All K's
11314	ND4	(3) none	Syn.	Yes	No	No (Singleton)	H1
11377	**ND4**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H1/K1/V1**
11410	ND4	(3) none	Syn.	Yes	No	No (Singleton)	H3
11440	ND4	(3) none	Syn.	No	No	No (Singleton)	H6
11467	ND4	(3) none	Syn.	Yes	No	No (Uninform.)	All K's
11485	**ND4**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**K2/K3**
11533	**ND4**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**T1**
11549	ND4	(1) none	Syn.	Yes	No	No (Singleton)	K1
11590	ND4	(3) none	Syn.	Yes	No	No (Singleton)	H2
11623	ND4	(3) none	Syn.	Yes	No	No (Singleton)	J2
11641	ND4	(3) none	Syn.	Yes	No	No (Singleton)	H6
11719	**ND4**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H5**

Appendix 3. (Continued)

Site	Gene/ Region	(Codon) AA Δ	Syn/ Non-Syn.	Prev. Obs.	Disease	SNP (reason)	Common Types
11731	ND4	(3) none	Syn.	No	No	No (Singleton)	H5
11812	ND4	(3) none	Syn.	Yes	No	No (Uninform.)	All T1's/T3's
11824	ND4	(3) none	Syn.	No	No	No (Singleton)	H2
11840	ND4	(1) none	Syn.	Yes	No	No (Singleton)	K3
11864	**ND4**	**(1) none**	**Syn.**	**Yes**	**No**	**YES**	**H2**
11878	ND4	(3) none	Syn.	No	No	No (Singleton)	H1
11914	**ND4**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H2/K2/T3**
12007	**ND4**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**T1**
12083	ND4	(1) S-A	Non-Syn.	No	No	No (Pheno. Δ)	J1
12153	tHis	-	-	Yes	No	No (Pheno. Δ)	H1
12217	tSer	-	-	No	No	No (Pheno. Δ)	H2
12236	tSer	-	-	Yes	No	No (Pheno. Δ)	H2
12308	tLeu	-	-	Yes	CPEO	No (Pheno. Δ)	All K's
12311	tLeu	-	-	Yes	CPEO	No (Pheno. Δ)	J4
12341	ND5	(2) T-I	Non-Syn.	No	No	No (Pheno. Δ)	H1
12354	ND5	(3) none	Syn.	No	No	No (Singleton)	H2
12361	ND5	(1) T-A	Non-Syn.	Yes	No	No (Pheno. Δ)	V1
12372	ND5	(3) none	Syn.	Yes	No	No (Uninform.)	All K's
12378	ND5	(3) none	Syn.	No	No	No (Singleton)	T1
12384	ND5	(3) none	Syn.	No	No	No (Singleton)	H5
12397	ND5	(1) T-A	Non-Syn.	Yes	No	No (Pheno. Δ)	J3/T1
12438	**ND5**	**(3) none**	**Syn.**	**No**	**No**	**YES**	**V1**
12453	ND5	(3) none	Syn.	Yes	No	No (Singleton)	V1
12477	ND5	(3) none	Syn.	Yes	No	No (Singleton)	K2
12493	ND5	(1) M-V	Non-Syn.	No	No	No (Pheno. Δ)	H3
12557	ND5	(2) T-I	Non-Syn.	Yes	No	No (Pheno. Δ)	J3
12612	ND5	(3) none	Syn.	Yes	No	No (Uninform.)	All J's
12616	ND5	(1) none	Syn.	Yes	No	No (Singleton)	H2
12630	ND5	(3) none	Syn.	Yes	No	No (Singleton)	H1
12633	ND5	(3) none	Syn.	Yes	No	No (Uninform.)	All T2'S
12681	ND5	(3) none	Syn.	Yes	No	No (Singleton)	T1
12687	ND5	(3) none	Syn.	No	No	No (Singleton)	H6
12699	ND5	(3) none	Syn.	No	No	No (Singleton)	T1
12771	ND5	(3) none	Syn.	Yes	No	No (Singleton)	T1
12795	**ND5**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**K1/T1**
12810	**ND5**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**V1**
12813	ND5	(3) none	Syn.	No	No	No (Singleton)	J1
12858	**ND5**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H1/J2**
12864	ND5	(3) none	Syn.	Yes	No	No (Singleton)	H5
13023	ND5	(3) none	Syn.	No	No	No (Singleton)	H5
13056	ND5	(3) none	Syn.	No	No	No (Singleton)	T2
13104	ND5	(3) none	Syn.	Yes	No	No (Singleton)	H2
13105	ND5	(1) I-V	Non-Syn.	Yes	No	No (Pheno. Δ)	V1
13117	ND5	(2) S-W	Non-Syn.	Yes	No	No (Pheno. Δ)	K2/K3
13143	ND5	(3) none	Syn.	Yes	No	No (Singleton)	K1
13145	ND5	(2) S-N	Non-Syn.	Yes	No	No (Pheno. Δ)	J1
13161	ND5	(3) none	Syn.	Yes	No	No (Singleton)	H6
13168	ND5	(1) none	Syn.	No	No	No (Singleton)	H2

Appendix 3. (Continued)

Site	Gene/ Region	(Codon) AA Δ	Syn/ Non-Syn.	Prev. Obs.	Disease	SNP (reason)	Common Types
13293	**ND5**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**K1**
13326	ND5	(3) none	Syn.	Yes	No	No (Singleton)	H2
13368	ND5	(3) none	Syn.	Yes	No	No (Uninform.)	All T's
13434	ND5	(3) none	Syn.	Yes	No	No (Singleton)	J1
13477	ND5	(1) A-T	Non-Syn.	Yes	No	No (Pheno. Δ)	V1
13500	ND5	(3) none	Syn.	Yes	No	No (Redundant)	V1
13557	ND5	(3) none	Syn.	No	No	No (Singleton)	J3
13658	ND5	(2) T-I	Non-Syn.	No	No	No (Pheno. Δ)	H1
13680	ND5	(3) none	Syn.	Yes	No	No (Redundant)	H1/H2/H3
13692	ND5	(3) none	Syn.	Yes	No	No (Singleton)	T1
13708	ND5	(1) A-T	Non-Syn.	Yes	LHON	No (Pheno. Δ)	H1/K1
13722	ND5	(3) none	Syn.	Yes	No	No (Singleton)	T1
13734	ND5	(3) none	Syn.	Yes	No	No (Singleton)	J3
13740	ND5	(3) none	Syn.	Yes	No	No (Singleton)	K3
13851	ND5	(3) none	Syn.	Yes	No	No (Singleton)	H2
13879	ND5	(1) S-P	Non-Syn.	Yes	No	No (Pheno. Δ)	All J4'S
13879	ND5	(1) S-T	Non-Syn.	Yes	No	No (Pheno. Δ)	H1
13886	ND5	(2) L-P	Non-Syn.	Yes	No	No (Pheno. Δ)	K3
13896	ND5	(3) none	Syn.	No	No	No (Singleton)	H6
13899	ND5	(3) none	Syn.	Yes	No	No (Singleton)	J2
13934	ND5	(2) T-M	Non-Syn.	Yes	No	No (Pheno. Δ)	J2
13965	ND5	(3) none	Syn.	Yes	No	No (Redundant)	T3
14016	ND5	(3) none	Syn.	Yes	No	No (Singleton)	T1
14063	ND5	(2) I-T	Non-Syn.	Yes	No	No (Pheno. Δ)	K3
14106	ND5	(3) none	Syn.	Yes	No	No (Singleton)	T2
14124	ND5	(3) none	Syn.	No	No	No (Singleton)	V1
14155	ND6	(3) none	Syn.	No	No	No (Singleton)	H2
14167	ND6	(3) none	Syn.	Yes	No	No (Uninform.)	All K's
14178	ND6	(1) I-V	Non-Syn.	Yes	No	No (Pheno. D)	V1
14180	ND6	(2) Y-C	Non-Syn.	Yes	No	No (Pheno. D)	J2
14233	ND6	(3) none	Syn.	Yes	No	No (Uninform.)	All T1's/T3's
14296	ND6	(3) none	Syn.	Yes	No	No (Singleton)	J1
14305	**ND6**	**(3) none**	**Syn.**	**No**	**No**	**YES**	**K1**
14323	ND6	(3) none	Syn.	Yes	No	No (Singleton)	J2
14364	ND6	(1) none	Syn.	Yes	No	No (Singleton)	H1
14365	ND6	(3) none	Syn.	Yes	No	No (Redundant)	H1/H2/H6
14384	ND6	(2) A-V	Non-Syn.	Yes	No	No (Pheno. Δ)	K3
14405	ND6	(2) V-A	Non-Syn.	Yes	No	No (Pheno. Δ)	H1
14470	**ND6**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H1/H2**
14502	ND6	(1) I-V	Non-Syn.	Yes	No	No (Pheno. Δ)	J1/K3
14560	**ND6**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H2/H3**
14577	ND6	(1) I-V	Non-Syn.	Yes	No	No (Pheno. Δ)	H2
14582	ND6	(2) V-A	Non-Syn.	Yes	No	No (Pheno. Δ)	H1/H2/H6
14687	tGlu	-	-	Yes	No	No (Pheno. Δ)	H2/T3
14766	CytB	(2) T-I	Non-Syn.	Yes	No	No (Pheno. Δ)	All J, K, Ts
14770	**CytB**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**V1**
14793	CytB	(2) H-R	Non-Syn.	Yes	No	No (Pheno. Δ)	V1
14798	CytB	(1) F-C	Non-Syn.	Yes	No	No (Pheno. Δ)	T1/V1

Appendix 3. (Continued)

Site	Gene/ Region	(Codon) AA Δ	Syn/ Non-Syn.	Prev. Obs.	Disease	SNP (reason)	Common Types
14831	CytB	(1) A-T	Non-Syn.	No	No	No (Pheno. Δ)	H1/H2
14869	**CytB**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H2/H3**
14872	CytB	(3) none	Syn.	Yes	No	No (Redundant)	H1/H2/H3
14905	CytB	(3) none	Syn.	Yes	No	No (Uninform.)	All T's
14959	CytB	(3) none	Syn.	No	No	No (Singleton)	H3
15043	**CytB**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**J3/T1**
15047	CytB	(1) G-S	Non-Syn.	Yes	No	No (Pheno. Δ)	J3
15055	CytB	(3) none	Syn.	No	No	No (Singleton)	H4
15067	**CytB**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**J4**
15110	CytB	(1) A-T	Non-Syn.	Yes	No	No (Pheno. Δ)	H2
15172	CytB	(3) none	Syn.	Yes	No	No (Singleton)	H6
15199	CytB	(3) none	Syn.	Yes	No	No (Singleton)	J3
15244	CytB	(3) none	Syn.	Yes	No	No (Singleton)	H2
15245	CytB	(1) G-S	Non-Syn.	No	No	No (Pheno. Δ)	T3
15250	CytB	(3) none	Syn.	No	No	No (Singleton)	V1
15323	CytB	(1) A-T	Non-Syn.	Yes	No	No (Pheno. Δ)	H1/H3
15325	CytB	(3) none	Syn.	Yes	No	No (Singleton)	H6
15326	CytB	(1) T-A	Non-Syn.	Yes	No	No (Pheno. Δ)	All - rCRS
15340	**CytB**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**H2**
15341	CytB	(1) F-L	Non-Syn.	No	No	No (Pheno. Δ)	T2
15346	CytB	(3) none	Syn.	Yes	No	No (Redundant)	V1
15355	**CytB**	**(3) none**	**Syn.**	**Yes**	**No**	**YES**	**J2/K2**
15409	CytB	(3) none	Syn.	Yes	No	No (Singleton)	H1
15412	CytB	(3) none	Syn.	Yes	No	No (Singleton)	T2
15452	CytB	(1) L-I	Non-Syn.	Yes	No	No (Pheno. Δ)	All J's/T's
15466	CytB	(3) none	Syn.	Yes	No	No (Singleton)	H1
15467	CytB	(1) T-A	Non-Syn.	Yes	No	No (Pheno. Δ)	J4
15497	CytB	(1) G-S	Non-Syn.	Yes	No	No (Pheno. Δ)	H2
15520	CytB	(3) none	Syn.	No	No	No (Singleton)	K1
15607	CytB	(3) none	Syn.	Yes	No	No (Uninform.)	All T's
15734	CytB	(1) A-T	Non-Syn.	Yes	No	No (Pheno. Δ)	H4
15735	CytB	(2) A-V	Non-Syn.	No	No	No (Pheno. Δ)	J4
15758	CytB	(1) I-V	Non-Syn.	Yes	No	No (Pheno. Δ)	H2/J2
15769	CytB	(3) none	Syn.	No	No	No (Singleton)	J1
15773	CytB	(1) V-M	Non-Syn.	Yes	No	No (Pheno. Δ)	V1/T1
15799	CytB	(3) none	Syn.	No	No	No (Singleton)	T1/V1
15817	CytB	(3) none	Syn.	No	No	No (Singleton)	H1
15833	**CytB**	**(1) none**	**Syn.**	**Yes**	**No**	**YES**	**H5**
15884	**NC**	**-**	**-**	**Yes**	**No**	**YES**	**H5/K3**
15900	tThr	-	-	Yes	No	No (Pheno. Δ)	K1
15904	tThr	-	-	Yes	No	No (Pheno. Δ)	H1/V1
15909	tThr	-	-	No	No	No (Pheno. Δ)	J3
15924	tThr	-	-	Yes	LIMM	No (Pheno. Δ)	H1/T2/V1
15928	tThr	-	-	Yes	No	No (Pheno. Δ)	All T's
15930	tThr	-	-	Yes	No	No (Pheno. Δ)	H5
15935	tThr	-	-	No	No	No (Pheno. Δ)	T3
15936	tThr	-	-	No	No	No (Pheno. Δ)	J3
15946	tThr	-	-	Yes	No	No (Pheno. Δ)	K3

Appendix 3. (Continued)

Site	Gene/ Region	(Codon) AA Δ	Syn/ Non-Syn.	Prev. Obs.	Disease	SNP (reason)	Common Types
16069	CR	-	-	Yes	No	No (Diagnostic)	-
16093	CR	-	-	Yes	No	No (Diagnostic)	-
16126	CR	-	-	Yes	No	No (Diagnostic)	-
16129	CR	-	-	Yes	No	No (Diagnostic)	-
16145	CR	-	-	Yes	No	No (Diagnostic)	-
16162	CR	-	-	Yes	No	No (Diagnostic)	-
16163	CR	-	-	Yes	No	No (Diagnostic)	-
16168	CR	-	-	Yes	No	No (Singleton)	J1
16172	CR	-	-	Yes	No	No (Diagnostic)	-
16186	CR	-	-	Yes	No	No (Diagnostic)	-
16189	CR	-	-	Yes	No	No (Diagnostic)	-
16209	CR	-	-	Yes	No	No (Diagnostic)	-
16222	CR	-	-	Yes	No	No (Diagnostic)	-
16224	CR	-	-	Yes	No	No (Diagnostic)	-
16260	CR	-	-	Yes	No	No (Singleton)	T1
16261	CR	-	-	Yes	No	No (Diagnostic)	-
16263	CR	-	-	Yes	No	No (Diagnostic)	-
16294	CR	-	-	Yes	No	No (Diagnostic)	-
16296	CR	-	-	Yes	No	No (Diagnostic)	-
16298	CR	-	-	Yes	No	No (Diagnostic)	-
16304	CR	-	-	Yes	No	No (Diagnostic)	-
16311	CR	-	-	Yes	No	No (Diagnostic)	-
16368	**CR**	**-**	**-**	Yes	**No**	**YES**	**J1**
16390	**CR**	**-**	**-**	Yes	**No**	**YES**	**J1/V1**
16399	CR	-	-	Yes	No	No (Singleton)	K2
16422	CR	-	-	Yes	No	No (Singleton)	K2
16472	CR	-	-	No	No	No (Singleton)	H1
16519	**CR**	**-**	**-**	**Yes**	**No**	**YES**	**H's J's K's V's**
16526	CR	-	-	Yes	No	No (Singleton)	V1
16527	CR	-	-	Yes	No	No (Singleton)	H5

Appendix 4. The Entire Mutation Rate Spectrum of the Coding Region of 646 Human Mitochondrial DNA Genomes using Parsimony Analysis. The site-by-site rate spectrum was determined from parsimony analysis of 646 coding region genomes (Ingman et al., 1999; Maca-Meyer et al., 2001; and Herrnstadt et al., 2002). The numbers refer to the revised Cambridge Reference Sequence rCRS (Anderson et al., 1981; Andrews et al., 1999). Length (L) of characters were determined by counting the number of changes on a most parsimonious tree. Characters are ordered from the fastest changing sites to the slowest. Invariant sites (L=0) are not shown.

L	rCRS	L	rCRS	L	rCRS	L	rCRS	L	rCRS	L	rCRS
15	709	4	3796	3	6182	3	13194	2	3027	2	6253
13	11914	4	3918	3	6308	3	13281	2	3336	2	6272
12	5460	4	5004	3	6366	3	13293	2	3396	2	6365
12	13708	4	6023	3	6480	3	13590	2	3480	2	6367
10	15924	4	9254	3	6671	3	13827	2	3505	2	6392
9	1719	4	9377	3	6719	3	13958	2	3531	2	6413
9	10398	4	9554	3	6827	3	13966	2	3547	2	6446
8	3010	4	9667	3	6951	3	13980	2	3552	2	6518
8	8251	4	9861	3	7055	3	14016	2	3654	2	6620
8	14470	4	10084	3	7256	3	14034	2	3693	2	6680
8	15784	4	10410	3	7444	3	14097	2	3843	2	6710
7	961	4	10685	3	7521	3	14133	2	3866	2	6734
7	3316	4	11176	3	7789	3	14180	2	3936	2	6755
6	5237	4	11611	3	7853	3	14233	2	3992	2	6776
6	10915	4	12172	3	7861	3	14308	2	4012	2	6845
6	11719	4	12358	3	8269	3	14364	2	4104	2	6917
6	12007	4	12630	3	8292	3	14769	2	4164	2	6962
6	12346	4	13020	3	8473	3	14869	2	4185	2	7013
6	13105	4	15043	3	8557	3	15236	2	4435	2	7146
6	13928	4	15110	3	8616	3	15257	2	4454	2	7151
6	14569	4	15930	3	8618	3	15734	2	4553	2	7202
6	14766	3	723	3	8697	3	15758	2	4639	2	7337
6	15301	3	930	3	8856	2	721	2	4646	2	7389
6	15670	3	960	3	8994	2	870	2	4736	2	7424
6	15884	3	1462	3	9123	2	951	2	4767	2	7561
5	593	3	1811	3	9266	2	1007	2	4793	2	7581
5	3705	3	2352	3	9380	2	1211	2	4820	2	7604
5	4688	3	3308	3	9449	2	1243	2	4824	2	7621
5	5147	3	3338	3	9824	2	1393	2	4917	2	7702
5	5471	3	3394	3	9948	2	1406	2	5048	2	7864
5	6221	3	3591	3	10373	2	1555	2	5063	2	7930
5	6260	3	3666	3	10589	2	1700	2	5201	2	7948
5	6261	3	3777	3	10598	2	1717	2	5231	2	7984
5	9545	3	3915	3	10816	2	2098	2	5262	2	8014
5	9932	3	4216	3	11002	2	2158	2	5319	2	8020
5	11800	3	4561	3	11017	2	2159	2	5563	2	8027
5	13434	3	4562	3	11253	2	2245	2	5580	2	8119
5	14182	3	4655	3	11347	2	2294	2	5581	2	8158
5	15326	3	4991	3	11377	2	2308	2	5656	2	8270
4	750	3	5046	3	11893	2	2332	2	5711	2	8279
4	980	3	5108	3	12092	2	2351	2	5821	2	8280
4	1438	3	5426	3	12372	2	2387	2	5843	2	8393
4	1598	3	5442	3	12397	2	2416	2	5913	2	8404
4	1888	3	5773	3	12406	2	2483	2	5999	2	8460
4	2706	3	6026	3	12414	2	2702	2	6150	2	8470
4	3438	3	6152	3	13145	2	2905	2	6164	2	8485

Appendix 4. (Continued)

L	rCRS	L	rCRS	L	rCRS	L	rCRS	L	rCRS	L	rCRS
2	8521	2	11065	2	13749	2	15944	1	1706	1	2836
2	8572	2	11143	2	13781	1	595	1	1709	1	2850
2	8701	2	11150	2	13789	1	596	1	1721	1	2851
2	8715	2	11151	2	13879	1	629	1	1733	1	2863
2	8772	2	11177	2	13886	1	633	1	1736	1	2885
2	8790	2	11204	2	13934	1	663	1	1738	1	2903
2	8793	2	11257	2	13965	1	669	1	1766	1	3105
2	8805	2	11299	2	14020	1	678	1	1809	1	3116
2	8843	2	11353	2	14070	1	680	1	1836	1	3197
2	8964	2	11467	2	14088	1	710	1	1842	1	3200
2	9052	2	11485	2	14148	1	719	1	1850	1	3203
2	9053	2	11593	2	14178	1	731	1	1900	1	3204
2	9055	2	11641	2	14209	1	735	1	1927	1	3206
2	9098	2	11653	2	14212	1	736	1	1977	1	3212
2	9117	2	11701	2	14374	1	748	1	2000	1	3254
2	9300	2	11710	2	14560	1	769	1	2060	1	3277
2	9305	2	11812	2	14574	1	813	1	2083	1	3290
2	9548	2	11887	2	14587	1	825	1	2092	1	3311
2	9632	2	11899	2	14605	1	827	1	2141	1	3333
2	9670	2	11935	2	14798	1	850	1	2145	1	3337
2	9755	2	11946	2	14812	1	867	1	2157	1	3345
2	9804	2	12083	2	14905	1	869	1	2217	1	3348
2	9947	2	12189	2	14927	1	921	1	2218	1	3350
2	9950	2	12236	2	14971	1	942	1	2226	1	3358
2	9962	2	12239	2	15099	1	954	1	2259	1	3372
2	9966	2	12441	2	15115	1	966	1	2263	1	3388
2	9986	2	12454	2	15148	1	982	1	2330	1	3397
2	10031	2	12477	2	15172	1	984	1	2358	1	3398
2	10142	2	12501	2	15191	1	986	1	2361	1	3421
2	10187	2	12519	2	15217	1	988	1	2380	1	3434
2	10238	2	12574	2	15218	1	1002	1	2386	1	3447
2	10352	2	12609	2	15244	1	1009	1	2404	1	3450
2	10370	2	12642	2	15313	1	1018	1	2442	1	3460
2	10454	2	12705	2	15314	1	1041	1	2486	1	3472
2	10463	2	12771	2	15317	1	1048	1	2581	1	3483
2	10499	2	12811	2	15431	1	1119	1	2626	1	3492
2	10586	2	12850	2	15452	1	1189	1	2638	1	3495
2	10609	2	12879	2	15487	1	1282	1	2639	1	3497
2	10646	2	12940	2	15511	1	1291	1	2650	1	3501
2	10688	2	12966	2	15589	1	1382	1	2708	1	3513
2	10700	2	13135	2	15592	1	1420	1	2735	1	3516
2	10792	2	13212	2	15607	1	1442	1	2755	1	3537
2	10810	2	13263	2	15697	1	1503	1	2757	1	3543
2	10819	2	13326	2	15766	1	1508	1	2758	1	3549
2	10876	2	13401	2	15775	1	1531	1	2768	1	3559
2	10907	2	13500	2	15812	1	1539	1	2769	1	3565
2	10920	2	13602	2	15833	1	1618	1	2775	1	3571
2	10955	2	13651	2	15883	1	1670	1	2789	1	3594
2	11016	2	13722	2	15927	1	1692	1	2792	1	3606
2	11044	2	13734	2	15928	1	1703	1	2804	1	3618

Appendix 4. (Continued)

L	rCRS	L	rCRS	L	rCRS	L	rCRS	L	rCRS	L	rCRS
1	3663	1	4395	1	5069	1	5744	1	6489	1	7211
1	3672	1	4491	1	5081	1	5746	1	6496	1	7226
1	3699	1	4506	1	5096	1	5788	1	6498	1	7241
1	3720	1	4508	1	5102	1	5806	1	6515	1	7257
1	3741	1	4512	1	5153	1	5811	1	6524	1	7258
1	3746	1	4529	1	5156	1	5814	1	6528	1	7270
1	3753	1	4531	1	5165	1	5824	1	6548	1	7274
1	3766	1	4541	1	5178	1	5826	1	6554	1	7278
1	3801	1	4580	1	5183	1	5894	1	6587	1	7299
1	3816	1	4586	1	5187	1	5911	1	6629	1	7319
1	3826	1	4619	1	5189	1	5951	1	6647	1	7325
1	3828	1	4626	1	5198	1	5964	1	6663	1	7347
1	3837	1	4634	1	5220	1	5973	1	6681	1	7364
1	3847	1	4679	1	5255	1	5984	1	6713	1	7385
1	3849	1	4695	1	5263	1	5987	1	6716	1	7400
1	3902	1	4703	1	5267	1	5988	1	6722	1	7403
1	3921	1	4715	1	5276	1	6014	1	6752	1	7476
1	3927	1	4722	1	5277	1	6038	1	6770	1	7493
1	3963	1	4727	1	5285	1	6045	1	6824	1	7498
1	3969	1	4732	1	5300	1	6047	1	6836	1	7559
1	3970	1	4733	1	5302	1	6071	1	6842	1	7571
1	3990	1	4735	1	5315	1	6077	1	6869	1	7598
1	3996	1	4742	1	5330	1	6104	1	6875	1	7609
1	4011	1	4745	1	5331	1	6146	1	6899	1	7618
1	4023	1	4755	1	5348	1	6167	1	6911	1	7624
1	4024	1	4769	1	5351	1	6179	1	6920	1	7645
1	4025	1	4772	1	5360	1	6185	1	6932	1	7648
1	4048	1	4811	1	5366	1	6216	1	6935	1	7660
1	4071	1	4823	1	5372	1	6231	1	6938	1	7669
1	4095	1	4833	1	5378	1	6237	1	6956	1	7673
1	4113	1	4843	1	5390	1	6249	1	6989	1	7674
1	4117	1	4848	1	5393	1	6257	1	6990	1	7675
1	4129	1	4850	1	5414	1	6266	1	6998	1	7678
1	4158	1	4853	1	5418	1	6267	1	7001	1	7692
1	4167	1	4859	1	5459	1	6290	1	7022	1	7693
1	4172	1	4883	1	5463	1	6293	1	7025	1	7694
1	4203	1	4884	1	5465	1	6296	1	7028	1	7697
1	4232	1	4907	1	5477	1	6311	1	7041	1	7705
1	4242	1	4908	1	5480	1	6320	1	7058	1	7711
1	4248	1	4910	1	5492	1	6324	1	7076	1	7724
1	4260	1	4924	1	5495	1	6339	1	7080	1	7744
1	4295	1	4928	1	5516	1	6340	1	7082	1	7759
1	4310	1	4937	1	5553	1	6341	1	7109	1	7761
1	4312	1	4960	1	5554	1	6351	1	7112	1	7762
1	4315	1	4970	1	5584	1	6371	1	7129	1	7768
1	4336	1	4977	1	5603	1	6378	1	7149	1	7771
1	4370	1	4994	1	5628	1	6437	1	7158	1	7772
1	4371	1	5024	1	5633	1	6455	1	7159	1	7787
1	4386	1	5027	1	5634	1	6456	1	7175	1	7805
1	4388	1	5036	1	5655	1	6464	1	7184	1	7830
1	4392	1	5054	1	5715	1	6473	1	7196	1	7858

Appendix 4. (Continued)

L	rCRS	L	rCRS	L	rCRS	L	rCRS	L	rCRS	L	rCRS
1	7867	1	8558	1	8943	1	9490	1	10154	1	10828
1	7873	1	8563	1	8965	1	9509	1	10169	1	10834
1	7897	1	8566	1	8987	1	9531	1	10172	1	10853
1	7963	1	8574	1	9006	1	9533	1	10192	1	10873
1	7999	1	8577	1	9007	1	9536	1	10197	1	10894
1	8032	1	8581	1	9017	1	9540	1	10199	1	10895
1	8059	1	8584	1	9022	1	9557	1	10211	1	10899
1	8080	1	8588	1	9041	1	9575	1	10245	1	10927
1	8087	1	8602	1	9042	1	9591	1	10247	1	10939
1	8098	1	8604	1	9058	1	9599	1	10253	1	10961
1	8137	1	8610	1	9070	1	9605	1	10283	1	10978
1	8149	1	8619	1	9072	1	9620	1	10286	1	11023
1	8152	1	8633	1	9093	1	9621	1	10289	1	11047
1	8182	1	8640	1	9097	1	9647	1	10308	1	11084
1	8191	1	8641	1	9103	1	9656	1	10310	1	11087
1	8206	1	8650	1	9106	1	9682	1	10313	1	11119
1	8227	1	8655	1	9111	1	9692	1	10321	1	11147
1	8237	1	8666	1	9128	1	9693	1	10325	1	11152
1	8248	1	8679	1	9129	1	9698	1	10335	1	11164
1	8260	1	8680	1	9136	1	9708	1	10358	1	11167
1	8272	1	8681	1	9137	1	9709	1	10365	1	11172
1	8277	1	8684	1	9139	1	9716	1	10382	1	11197
1	8287	1	8702	1	9145	1	9728	1	10389	1	11206
1	8293	1	8703	1	9150	1	9758	1	10394	1	11215
1	8308	1	8705	1	9156	1	9759	1	10397	1	11229
1	8334	1	8733	1	9174	1	9764	1	10400	1	11233
1	8343	1	8746	1	9192	1	9801	1	10423	1	11239
1	8369	1	8750	1	9196	1	9813	1	10424	1	11242
1	8383	1	8752	1	9198	1	9818	1	10427	1	11251
1	8386	1	8764	1	9210	1	9836	1	10448	1	11260
1	8387	1	8781	1	9221	1	9852	1	10505	1	11269
1	8414	1	8784	1	9242	1	9869	1	10506	1	11272
1	8417	1	8794	1	9248	1	9899	1	10535	1	11290
1	8419	1	8803	1	9263	1	9903	1	10550	1	11296
1	8422	1	8836	1	9272	1	9923	1	10560	1	11302
1	8428	1	8838	1	9296	1	9938	1	10595	1	11314
1	8429	1	8839	1	9299	1	9941	1	10604	1	11332
1	8440	1	8841	1	9301	1	9944	1	10640	1	11339
1	8448	1	8860	1	9311	1	9971	1	10644	1	11348
1	8450	1	8865	1	9324	1	9977	1	10654	1	11362
1	8453	1	8868	1	9333	1	10007	1	10664	1	11365
1	8461	1	8869	1	9335	1	10034	1	10667	1	11380
1	8468	1	8870	1	9336	1	10042	1	10724	1	11396
1	8472	1	8875	1	9347	1	10044	1	10736	1	11401
1	8478	1	8877	1	9355	1	10071	1	10750	1	11410
1	8480	1	8898	1	9356	1	10086	1	10754	1	11470
1	8507	1	8901	1	9362	1	10097	1	10771	1	11527
1	8508	1	8905	1	9370	1	10101	1	10775	1	11533
1	8519	1	8911	1	9386	1	10115	1	10790	1	11548
1	8522	1	8921	1	9467	1	10118	1	10793	1	11549
1	8530	1	8925	1	9469	1	10124	1	10804	1	11560
1	8541	1	8928	1	9477	1	10143	1	10822	1	11569

Appendix 4. (Continued)

L	rCRS	L	rCRS	L	rCRS	L	rCRS	L	rCRS	L	rCRS
1	11590	1	12258	1	12954	1	13630	1	14061	1	14552
1	11617	1	12285	1	12957	1	13637	1	14094	1	14553
1	11654	1	12308	1	12961	1	13638	1	14106	1	14566
1	11665	1	12317	1	12978	1	13641	1	14110	1	14571
1	11674	1	12331	1	12999	1	13650	1	14118	1	14577
1	11696	1	12338	1	13015	1	13656	1	14125	1	14580
1	11722	1	12396	1	13017	1	13660	1	14127	1	14581
1	11732	1	12403	1	13029	1	13680	1	14131	1	14582
1	11778	1	12453	1	13059	1	13681	1	14139	1	14599
1	11788	1	12468	1	13071	1	13707	1	14149	1	14620
1	11840	1	12471	1	13098	1	13710	1	14152	1	14629
1	11854	1	12483	1	13101	1	13711	1	14167	1	14634
1	11864	1	12492	1	13104	1	13725	1	14173	1	14645
1	11869	1	12507	1	13111	1	13740	1	14179	1	14668
1	11884	1	12535	1	13116	1	13743	1	14193	1	14687
1	11908	1	12557	1	13117	1	13752	1	14203	1	14693
1	11909	1	12570	1	13129	1	13753	1	14215	1	14709
1	11923	1	12579	1	13143	1	13758	1	14220	1	14743
1	11928	1	12582	1	13149	1	13759	1	14239	1	14751
1	11929	1	12612	1	13182	1	13780	1	14251	1	14755
1	11932	1	12616	1	13188	1	13782	1	14258	1	14757
1	11938	1	12618	1	13197	1	13803	1	14278	1	14783
1	11944	1	12633	1	13221	1	13810	1	14279	1	14791
1	11947	1	12634	1	13254	1	13813	1	14280	1	14793
1	11950	1	12635	1	13269	1	13818	1	14284	1	14845
1	11953	1	12648	1	13276	1	13819	1	14299	1	14861
1	11959	1	12654	1	13296	1	13833	1	14305	1	14862
1	11963	1	12672	1	13353	1	13851	1	14311	1	14866
1	11969	1	12678	1	13356	1	13855	1	14313	1	14872
1	12011	1	12681	1	13359	1	13880	1	14318	1	14896
1	12019	1	12684	1	13368	1	13889	1	14319	1	14902
1	12026	1	12693	1	13404	1	13911	1	14323	1	14911
1	12030	1	12696	1	13413	1	13914	1	14338	1	14914
1	12049	1	12720	1	13422	1	13924	1	14350	1	14926
1	12063	1	12723	1	13448	1	13926	1	14365	1	14953
1	12064	1	12729	1	13464	1	13927	1	14371	1	14956
1	12070	1	12738	1	13473	1	13933	1	14384	1	14970
1	12082	1	12741	1	13477	1	13941	1	14388	1	14978
1	12084	1	12753	1	13485	1	13943	1	14393	1	14979
1	12091	1	12762	1	13488	1	13954	1	14405	1	15001
1	12127	1	12768	1	13494	1	13967	1	14420	1	15016
1	12133	1	12788	1	13506	1	13973	1	14428	1	15025
1	12134	1	12810	1	13528	1	14000	1	14434	1	15028
1	12153	1	12822	1	13542	1	14002	1	14455	1	15034
1	12166	1	12858	1	13563	1	14007	1	14463	1	15047
1	12171	1	12864	1	13565	1	14022	1	14484	1	15058
1	12175	1	12870	1	13594	1	14025	1	14488	1	15061
1	12200	1	12880	1	13611	1	14037	1	14497	1	15064
1	12234	1	12930	1	13617	1	14040	1	14518	1	15077
1	12245	1	12948	1	13626	1	14053	1	14544	1	15107
1	12248	1	12950	1	13629	1	14059	1	14548	1	15109

Appendix 4. (Continued)

L	rCRS	L	rCRS
1	15113	1	15521
1	15119	1	15530
1	15136	1	15535
1	15151	1	15553
1	15153	1	15562
1	15160	1	15601
1	15204	1	15616
1	15205	1	15622
1	15211	1	15626
1	15213	1	15629
1	15221	1	15632
1	15229	1	15646
1	15247	1	15652
1	15263	1	15657
1	15266	1	15663
1	15274	1	15672
1	15295	1	15679
1	15299	1	15693
1	15304	1	15703
1	15311	1	15731
1	15325	1	15746
1	15340	1	15773
1	15346	1	15777
1	15355	1	15787
1	15358	1	15789
1	15367	1	15790
1	15380	1	15803
1	15386	1	15808
1	15391	1	15824
1	15394	1	15849
1	15409	1	15872
1	15412	1	15885
1	15421	1	15889
1	15451	1	15900
1	15454	1	15904
1	15458	1	15905
1	15466	1	15907
1	15470	1	15932
1	15496	1	15942
1	15497	1	15946
1	15498	1	15954
1	15508	1	15955
1	15514	1	15978

REFERENCES

Adachi, J., and Hasegawa, M. (1995) Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J. Mol. Evol. 40:622-628.

Allard, M.W., Miller, K., Wilson, M., Monson, K., and Budowle, B. (2002) Characterization of the Caucasian haplogroups present in the SWGDAM forensic mtDNA dataset for 1771 human control region sequences. Scientific Working Group on DNA Analysis Methods. J. Forensic Sci. 47: 1215-1223.

Anderson, S., Bankier, A.T., Barrell, B.G., deBrujin, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe,B.A., Sanger, F., Schreier, P.H., Smith, A.J.H., Staden, R., and Young, I.G. (1981) Sequence and organization of the human mitochondrial genome. Nature 290: 457-465.

Andreasson, H., Asp, A., Alderborn, A., Gyllensten, U., and Allen, M. (2002) Mitochondrial sequence analysis for forensic identification using pyrosequencing technology. Biotechniques 32: 124-133.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999) Reanalysis and revision of the Cambridge Reference Sequence for human mitochondrial DNA. Nature Genetics 23: 147.

Aquadro, C.F., and Greenberg, B.D. (1983) Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics 103: 287-312.

Aris-Brosou, S., and Excoffier, L. (1996) The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol. Biol. Evol. 13: 494-504.

Awadalla, P., Eyre-Walker, A., and Smith, J.M. (1999) Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286: 2524-2525.

Bandelt, H.J., Quintana-Murci, L., Salas, A., and Macaulay, V. (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am. J. Hum. Genet. 71: 1150-1160.

Bisbing, R. E. (1982). The forensic identification and association of human hair. in Forensic Science Handbook, R. Saferstein, ed. (New Jersey: Prentice Hall Regents), pp. 184-221.

Bodenteich, A., Mitchell, L.G., Polymeropoulos, M.H., and Merril, C.R. (1992) Dinucleotide repeat in the human mitochondrial D-loop. Hum. Mol. Genet. 1: 140.

Brown, M.D., Shoffner, J.M., Kim, Y.L., Jun, A.S., Graham, B.H., Cabell, M.F., Gurley, D.S., and Wallace, D.C. (1996) Mitochondrial DNA sequence analysis of four Alzheimer's and Parkinson's disease patients. Am. J. Med. Genet. 61: 283-289.

Brown, W.M., George, M. Jr., and Wilson, A.C. (1979) Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA 76: 1967-1971.

Brown, W.M. (1980) Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc. Natl. Acad. Sci. USA 77: 3605-3609.

Cann, R.L., Stoneking, M., and Wilson, A.C. (1987) Mitochondrial DNA and human evolution. Nature 325: 31-36.

Cavelier, L., Erikson, I., Tammi, M., Jalonen, P., Lindholm, E., Jazin, E., Smith, P., Luthman, H., and Gyllensten, U. (2002) MtDNA mutations in maternally inherited diabetes: presence of the 3397 ND1 mutation previously associated with Alzheimer's and Parkinson's disease. J. Neuropathol. Exp. Neurol. 61: 634-639.

Chen, Y.S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A.S., and Wallace, D.C. (1995) Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am. J. Hum. Genet. 57: 133-149.

Denaro, M., Blanc, H., Johnson, M.J., Chen, K.H., Wilmsen, E., Cavalli-Sforza, L.L., and Wallace, D.C. (1981) Ethnic variation in Hpa 1 endonuclease cleavage patterns of human mitochondrial DNA. Proc. Natl. Acad. Sci. USA 78: 5768-5772.

Deng, H.W., and Fu, Y.X. (2000) Counting mutations by parsimony and estimation of mutation rate variation across nucleotide sites – a simulation study. Mathematical and Computer Modelling 32: 83-95.

Elson, J.L., Andrews, R.M., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (2001) Analysis of European mtDNAs for recombination. Am. J. Hum. Genet. 68:145-153.

Excoffier, L. (1990) Evolution of human mitochondrial DNA: evidence for departures from a pure neutral model of populations at equilibrium. J. Mol. Evol. 30: 125-139.

Excoffier, L., and Yang, Z. (1999) Substitution rate variation among sites in the mitochondrial hypervariable region I of humans and chimpanzees. Mol. Biol. Evol. 16: 1357-1368.

Eyre-Walker, A., Smith, N.H., and Smith, J.M. (1999) How clonal are human mitochondria? Proc. R. Soc. Lond. B Biol. Sci. 266: 477–483.

Eyre-Walker, A., and Awadalla, P. (2001) Does human mtDNA recombine? J. Mol. Evol. 53: 430-435.

Finnila, S., Hassinen, I.E., Ala-Kokko, L., and Majamaa, K. (2000) Phylogenetic network of the mtDNA haplogroup U in Northern Finland based on sequence analysis of the complete coding region by conformation-sensitive gel electrophoresis. Am. J. Hum. Genet. 66: 1017-1026.

Finnila, S., Lehtonen, M.S., and Majamaa, K. (2001) Phylogenetic network for European mtDNA. Am. J. Hum. Genet. 68: 1475-1484.

Fu, Y.X. (1994) A Phylogenetic Estimator of Effective Population Size or Mutation Rate. Genetics 136: 685-692.

Gabriel, M.N., Huffine, E.F., Ryan, J.H., Holland, M.M., and Parsons, T.J. (2001) Improved MtDNA sequence analysis of forensic remains using a "mini-primer set" amplification strategy. J. Forensic Sci. 46: 247-253.

Giles, R.E., Blanc, H., Cann, H.M., and Wallace, D.C. (1980) Maternal inheritance of human mitochondrial DNA. Proc. Natl. Acad. Sci. USA 77: 6715-6719.

Goloboff, P. (1999) NONA (NO NAME) ver. 2.0. Published by the author, Tucumán, Argentina. (http://www.cladistics.com/aboutNona.htm).

Hagelberg, E. (2003) Recombination or mutation rate heterogeneity? Implications for Mitochondrial Eve. Trends Genet. 19: 84-90.

Hasegawa, M., Di Rienzo, A., Kocher, T.D., and Wilson, A.C. (1993) Toward a more accurate time scale for the human mitochondrial DNA tree. J. Mol. Evol. 37: 347-354.

Helgason, A., Hickey, E., Goodacre, S., Bosnes, V., Stefansson, K., Ward, R., and Sykes, B. (2001) mtDna and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. Am. J. Hum. Genet. 68: 723-737.

Herrnstadt, C., Elson, J.L., Fahy, E., Preston, G., Turnbull, D.M., Anderson, C., Ghosh, S.S., Olefsky, J.M., Beal, M.F., Davis, R.E., and Howell, N. (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.

Herrnstadt, C., Preston, G., and Howell, N. (2003) Errors, phantoms and otherwise, in human mtDNA sequences. Am. J. Hum. Genet. 72: 1585-1586.

Heyer, E., Zietkiewicz, E., Rochowski, A., Yotova, V., Puymirat, J., and Labuda, D. (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. Am. J. Hum. Genet. 69: 1113-1126.

Holland, M.M., Fisher, D.L., Mitchell, L.G., Rodriquez, W.C., Canik, J.J., Merril, C.R., and Weedn, V.W. (1993) Mitochondrial DNA sequence analysis of human skeletal remains: identification of remains from the Vietnam War. J. Forensic Sci. 38: 542-553.

Holland, M.M., and Parsons, T.J. (1999) Mitochondrial DNA Sequence Analysis - Validation and Use for Forensic Casework. Forensic Science Review 11: 21-50.

Horai, S., and Hayasaka, K. (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am. J. Hum. Genet. 46: 828-842.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K., and Takahata, N. (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc. Natl. Acad. Sci. USA 92: 532-536.

Howell, N., Kubacka, I., and Mackey, D.A. (1996) How rapidly does the human mitochondrial genome evolve? Am. J. Hum. Genet. 59: 501-509.

Howell, N., Smejkal, C.B., Mackey, D.A., Chinnery, P.F., Turnbull, D.M., and Herrnstadt, C. (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.

Ingman, M., Kaessmann, H., Pääbo, S., and Gyllensten, U. (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408: 708-713.

Ingman, M., and Gyllensten, U. (2001) Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. J. Hered. 92: 454-461.

Ingman, M., and Gyllensten, U. (2003) Mitochondrial genome variation and evolutionary history of Australian and New Guinean Aborigines. Gen. Res. 13: 1600-1606.

Innan, H., and Nordborg, M. (2002) Recombination or mutational hot spots in human mtDNA? Mol. Biol. Evol. 19: 1122-1127.

Jazin, E., Soodyall, H., Jalonen, P., Lindholm, E., Stoneking, M., and Gyllensten, U. (1998) Mitochondrial mutation rate revisited: hot spots and polymorphism. Nat. Genet. 18: 109-110.

Johnson, M.J., Wallace, D.C., Ferris, S.D., Rattazzi, M.C., and Cavalli-Sforza, L.L. (1983) Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. J. Mol. Evol. 19: 255-271.

Johnson, N.L., and Kotz, S. (1969) Discrete Distributions Volume One, Boston: Houghton Mifflin.

Jorde, L.B., and Bamshad, M. (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288: 1931.

Kimura, M. (1983) The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press.

Kivisild, T., and Villems, R. (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288: 1931.

Kogelnik, A.M., Lott, M.T., Brown, M.D., Navathe, S.B., and Wallace, D.C. (1998) MITOMAP: a human mitochondrial genome database--1998 update Nucleic Acids Res. 26: 112-115. (http://www.mitomap.org/)

Kumar, S., Hedrick, P., Dowling, T., and Stoneking, M. (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288: 1931.

Lee, M.S., and Levin, B.C. (2002) MitoAnalyzer, a computer program and interactive web site to determine the effects of single nucleotide polymorphisms and mutations in human mitochondrial DNA. Mitochondrion 1: 321-326. (http://www.cstl.nist.gov/biotech/strbase/mitoanalyzer.html)

Lee, S.D., Lee, Y.S., and Lee, J.B. (2002) Polymorphism in the mitochondrial cytochrome B gene in Koreans. An additional marker for individual identification. Int. J. Legal Med. 116: 74-78.

Levin, B.C., Cheng, H., and Reeder, D.J. (1999) A human mitochondrial DNA standard reference material for quality control in forensic identification, medical diagnosis, and mutation detection. Genomics 55: 135-146.

Levin, B.C., Holland, K.A., Hancock, D.K., Coble, M., Parsons, T.J., Kienker, L.J., Williams, D.W., Jones, M., and Richie, K.L. (2003) Comparison of the Complete mtDNA Genome Sequences of Human Cell Lines - HL-60 and GM10742A - From Individuals With Pro-Myelocytic Leukemia and Leber Heredity Optic Neuropathy, Respectively, and the Inclusion of HL-60 in the NIST Human Mitochondrial DNA Standard Reference Material - SRM 2392-I. Mitochondrion 2: 387-400.

Lin, M.T., Simon, D.K., Ahn, C.H., Kim, L.M., and Beal, M.F. (2002) High aggregate burden of somatic mtDNA point mutations in aging and Alzheimer's disease brain. Hum. Mol. Genet. 11: 133-145.

Liu,Y., Cortopassi, G., Steingrimsdottir, H., Waugh, A.P., Beare, D.M., Green, M.H., Robinson, D.R., and Cole, J. (1997) Correlated mutagenesis of bcl2 and hprt loci in blood lymphocytes. Environ. Mol. Mutagen. 29: 36-45.

Lundstrom, R., Tavare, S., and Ward, R.H. (1992) Modeling the evolution of the human mitochondrial genome. Math. Biosci. 112: 319-335.

Lutz-Bonengel, S., Schmidt, U., Schmitt, T., and Pollak, S. (2003) Sequence polymorphisms within the human mitochondrial genes MTATP6, MTATP8, and MTND4. Int. J. Leg. Med. 117: 133-142.

Maca-Meyer, N., Gonzalez, A.M., Larruga, J.M., Flores, C., and Cabrera, V.M. (2001) Major genomic mitochondrial lineages delineate early human expansions. BMC Genet. 2: 13-20.

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonne-Tamir, B., Sykes, B., and Torroni, A. (1999a) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am. J. Hum. Genet. 64: 232-249.

Macaulay, V., Richards, M., and Sykes, B. (1999b) Mitochondrial DNA recombination-no need to panic. Proc. R. Soc. Lond. B Biol. Sci. 266: 2037-2039.

Maddison, D.R., Swofford, D.L., and Maddison, W.P. (1997) NEXUS: An extensible file format for systematic information. Systematic Biology 46: 590-621.

Malyarchuk, B.A., Rogozin, I.B., Berikov, V.B., and Derenko, M.V. (2002) Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region. Hum. Genet. 111: 46-53.

Mehta, A.B., Vulliamy, T., Gordon-Smith, E.C., and Luzzatto, L. (1989) A new genetic polymorphism in the 16S ribosomal RNA gene of human mitochondrial DNA. Ann. Hum. Genet. 53: 303-310.

Meyer, S., Weiss, G., and 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.

Meyer, S., and Von Haeseler, A. (2003) Identifying site-specific substitution rates. Mol. Biol. Evol. 20: 182-189.

Michaels, G.S., Hauswirth, W.W., and Laipis, P.J. (1982) Mitochondrial DNA copy number in bovine oocytes and somatic cells. Developmental Biology (Orlando) 94: 246-251.

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., and Wallace, D.C. (2003) Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. USA 100: 171-176.

Monson, K.L., Miller, K.W.P., Wilson, M.R., DiZinno, J.A., and Budowle, B. (2002) The mtDNA population database: an integrated software and database resource for forensic comparison. Forensic Sci. Comm., Vol 4, #2. (http://www.fbi.gov/hq/lab/fsc/backissu/april2002/miller1.htm)

Nachman, M.W., Brown, W.M., Stoneking, M., and Aquadro, C.F. (1996) Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics 142: 953-963.

Nixon, K.C. (2002) WinClada version 1.00.08. Published by the Author, Ithaca, NY. (http://www.cladistics.com/).

Paabo, S. (1996). Mutational hot spots in the mitochondrial microcosm. Am. J. Hum. Genet. 59: 493-496.

Parsons, T.J., Muniec, D.S., Sullivan, K., Woodyatt, N., Alliston-Greiner, R., Wilson, M.R., Berry, D.L., Holland, K.A., Weedn, V.W., Gill, P., and Holland, M.M. (1997). A high observed substitution rate in the human mitochondrial DNA control region. Nature Genetics 15: 363-368.

Parsons, T.J., and Holland, M.M. (1998) Mitochondrial mutation rate revisited: hot spots and polymorphism. Nat. Genet. 18: 110.

Parsons, T.J., and Irwin, J.A. (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288: 1931.

Penny, D., Steel, M., Waddell, P.J., and Hendy, M.D. (1995) Improved analyses of human mtDNA sequences support a recent African origin for Homo sapiens. Mol. Biol. Evol. 12: 863-882.

Pesole, G., and Saccone, C. (2001) A novel method for estimating substitution rate variation among sites in a large dataset of homologous DNA sequences. Genetics 157: 859-865.

Piko, L., and Matsumoto, L. (1976) Number of mitochondria and some properties of mitochondrial DNA in the mouse egg. Developmental Biology (Orlando) 49: 1-10.

Richards, M.B., Macaulay, V.A., Bandelt, H.J., and Sykes, B.C. (1998) Phylogeography of mitochondrial DNA in western Europe. Ann. Hum. Genet. 62: 241-260.

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., Papiha, S., Triantaphyllidis, C., Stefanescu, G., Hatina, J., Belledi, M., Di Rienzo, A., Novelletto, A., Oppenheim, A., Norby, S., Al-Zaheri, N., Santachiara-Benerecetti, S., Scozari, R., Torroni, A., and Bandelt, H.J. (2000) Tracing

European founder lineages in the Near Eastern mtDNA pool. Am. J. Hum. Genet. 67: 1251-1276.

Richards, M., and Macaulay, V. (2001) The mitochondrial gene tree comes of age. Am. J. Hum. Genet. 68: 1315-1320.

Rieder, M.J., Taylor, S.L., Tobe, V.O., and Nickerson, D.A. (1998) Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. Nucleic Acids Res. 26: 967-973.

Saccone, C., Pesole, G., and Sbisa, E. (1991) The main regulatory region of mammalian mitochondrial DNA: structure-function model and evolutionary pattern. J. Mol. Evol. 33: 83-91.

Saccone, C., Gissi, C., Lanave, C., Larizza, A., Pesole, G., and Reyes, A. (2000) Evolution of the mitochondrial genetic system: an overview. Gene 261: 153-159.

Saillard, J., Magalhaes, P.J., Schwartz, M., Rosenberg, T., and Norby, S. (2000) Mitochondrial DNA variant 11719G is a marker for the mtDNA haplogroup cluster HV. Hum. Biol. 72: 1065-1068.

Scheffler, I.E. (1999) Mitochondria. New York: Wiley-Liss, Inc.

Schurr, T.G., Ballinger, S.W., Gan, Y.Y., Hodge, J.A., Merriwether, D.A., Lawrence, D.N., Knowler, W.C., Weiss, K.M., and Wallace, D.C. (1990) Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages. Am. J. Hum. Genet. 46: 613-623.

Sigurdardottir, S., Helgason, A., Gulcher, J.R., Stefansson, K., and Donnelly, P. (2000) The mutation rate in the human mtDNA control region. Am. J. Hum. Genet. 66: 1599-1609.

Stewart, J.E., Fisher, C.L., Aagaard, P.J., Wilson, M.R., Isenberg, A.R., Polanskey, D., Pokorak, E., DiZinno, J.A., and Budowle, B. (2001) Length variation in HV2 of the human mitochondrial DNA control region. J. Forensic Sci. 46: 862-870.

Sullivan, J., Holsinger, K.E., and Simon, C. (1995) Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmodontine rodents. Mol. Biol. Evol. 12: 988-1001.

Swofford, D.L. (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Tateno, Y., Takezaki, N., and Nei, M. (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate

varies with site. Mol. Biol. Evol. 11: 261-277.

Templeton, A.R. (1992) Human origins and analysis of mitochondrial DNA sequences. Science 255: 737.

Templeton, A.R. (1996) Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. Genetics 144: 1263-1270.

Torroni, A., Lott, M.T., Cabell, M.F., Chen, Y.S., Lavergne, L., and Wallace, D.C. (1994) MtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. Am. J. Hum. Genet. 55: 760-776.

Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., Obinu, D., Savontaus, M.L., and Wallace, D.C. (1996) Classification of European mtDNAs from an analysis of three European populations. Genetics 144: 1835-1850.

Torroni, A., Rengo, R., Guida, V., Cruciani, F., Sellitto, D., Coppa, A., Calderon, F.L., Simionati, B., Valle, G., Richards, M., Macaulay, V., and Scozzari, R. (2001) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? Am. J. Hum. Genet. 69: 1348-1356.

Tourasse, N.J., and Gouy, M. (1997) Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. Mol. Biol. Evol. 14: 287-298.

Vallone, P.M., Hamm, R.S., Coble, M.D., Butler, J.M., and Parsons, T.J. (2003) A multiplex allele specific primer extension assay for 11 forensically informative SNPs distributed throughout the mitochondrial genome. Int. J. Legal Med. (*Manuscript in press*)

Van de Peer, Y., and De Wachter, R. (1993) TREECON: a software package for the construction and drawing of evolutionary trees. Comput. Appl. Biosci. 9: 177-182.

Van de Peer, Y., Neefs, J.M., De Rijk, P., and De Wachter, R. (1993) Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: calibration of the molecular clock. J. Mol. Evol. 37: 221-232.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., and Wilson, A.C. (1991) African populations and the evolution of human mitochondrial DNA. Science 253: 1503-1507.

Wakeley, J. (1993) Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. 37: 613-623.

Wakeley, J. (1994) Substitution-rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. 11: 436-442.

Wallace, D.C., Ye, J.H., Neckelmann, S.N., Singh, G., Webster, K.A., and Greenberg, B.D. (1987) Sequence analysis of cDNAs for the human and bovine ATP synthase beta subunit: mitochondrial DNA genes sustain seventeen times more mutations. Curr. Genet. 12: 81-90.

Wallace, D.C. (1994) Mitochondrial DNA sequence variation in human evolution and disease. Proc. Natl. Acad. Sci. USA 91: 8739-8746.

Wallace, D.C. (1999) Mitochondrial diseases in man and mouse. Science 283: 1482-1488.

Wallace, D.C., Brown, M.D., and Lott, M.T. (1999) Mitochondrial DNA variation in human evolution and disease. Gene 238: 211-230.

Watson, E., Forster, P., Richards, M., and Bandelt, H.J. (1997) Mitochondrial footprints of human expansions in Africa. Am. J. Hum. Genet. 61: 691-704.

Wilson, M.R., Holland, M.M., Stoneking, M., DiZinno, J.A., and Budowle, B. (1993) Guidelines for the use of mitochondrial DNA sequencing in forensic science. Crime Laboratory Digest 20: 68-77.

Wise, C.A., Sraml, M., and Easteal, S. (1998) Departure from neutrality at the mitochondrial NADH dehydrogenase subunit 2 gene in humans, but not in chimpanzees. Genetics 148: 409-421.

Wiuf, C. (2001) Recombination in human mitochondrial DNA? Genetics 159: 749-756.

Yang, Z. (1994) Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39: 105-111.

Yang, Z. (1995) Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. J. Mol. Evol. 40: 689-697.

Yang, Z., and Wang, T. (1995) Mixed model analysis of DNA sequence evolution. Biometrics 51: 552-561.

Yang, Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. TREE 11: 367-372.

Yang, Z., and Kumar, S. (1996) Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. 13: 650-659.

Yao, Y.G., Macauley, V., Kivisild, T., Zhang, Y.P., and Bandelt, H.J. (2003) To trust or not to trust an idiosyncratic mitochondrial data set. Am. J. Hum. Genet. 72: 1346-1349.