ABSTRACT

Title of dissertation: GENETIC VARIATION AT THE *N-ACETYLTRANSFERASE* (*NAT*) GENES IN GLOBAL HUMAN POPULATIONS

Holly M. Mortensen, Doctor of Philosophy, 2008

Dissertation Directed by: Associate Professor Sarah A. Tishkoff
Department of Biology

Currently, studies of the possible role of natural selection in shaping the observed variation at drug metabolizing enzyme (DME) loci remain limited. Functional variability at the N-acetyltransferase (*NAT*) genes is associated with adverse drug reactions and cancer susceptibility in humans. Previous studies of small sets of ethnic groups have indicated that the *NAT* genes have high levels of amino acid variation that differ in frequency across ethnic groups. I hypothesize that this functional variation may be adaptive in different environments and is maintained due to natural selection. Presumably, change in dietary patterns has been a strong selective pressure throughout the course of human evolution. The most extreme example of a shift in the dietary patterns of modern humans is most certainly the transition from a primarily hunter-gatherer subsistence to an agriculturalist lifestyle, within the past 10,000 years. Metabolism of grain and/or dairy products likely introduced new and foreign toxins to the human body. Although we can only speculate about the selective forces acting on the *NAT* genes in the past, it is possible that the observed pattern of phenotypic variation is associated with exposure to environmental, specifically dietary, toxins.

The purpose of this study is: 1) to characterize nucleotide variation at the *NAT* drug-metabolizing genes (*NAT1, NAT2*) and the pseudo-gene (*NATP1*) in global human populations, including many previously under-represented African populations and 2) to understand the role that natural selection has played in shaping variation at *NAT1* and *NAT2* in human populations living in different environmental settings. I have resequenced ~3000 bp for each of the *NAT1*, *NAT2* and *NATP1* gene regions, in 182 African individuals and 155 individuals from a representative global panel (HGDP-CEPH), and have identified Single Nucleotide Polymorphisms (SNPs) at each locus (*NAT1* (48), *NATP1* (55) and *NAT2* (46)). I have inferred haplotype phase and characterized patterns of haplotype diversity for each *NAT* locus. I have characterized nucleotide diversity and linkage disequilibrium for this ethnically diverse population dataset, as well as performed several tests of selective neutrality. This work will contribute to our understanding of how variation at the *NAT* loci may have been adaptive for dealing with changes in diet and exposure to toxins during human evolution.

GENETIC VARIATION AT THE *N*-ACETYLTRANSFERASE (*NAT*) GENES IN GLOBAL HUMAN POPULATIONS

by

Holly M. Mortensen

Dissertation submitted to the Faculty of the Graduate School of the University of Maryland, College Park, in partial fulfillment of the requirements for the degree of Doctor of Philosophy
2008

Advisory Committee:

Associate Professor Sarah A. Tishkoff, Chair
Assistant Professor Cristian I Castillo-Davis
Associate Professor Charles F. Delwiche
Assistant Professor Eric S. Haag
Professor Charles Mitter
Associate Professor Stephen M. Mount

FOREWARD

This dissertation is divided into six chapters: Introduction, Methods, the Evolutionary characterization of each of the *NAT1*, *NAT2*, and *NATP1* gene regions, and finally, a chapter on Conclusions and Future Study. Chapter I, the Introductory chapter will present a brief overview of drug metabolizing enzyme (DME) loci in the context of human evolutionary genetic studies, discuss human genetic response to environmental toxins, and outline current knowledge of *NAT* molecular genetics, including what is currently known of *NAT* acetylator haplotypes and their effect on phenotype in human populations. Chapter II, the Methods section of this dissertation will discuss all materials used, but also gives a brief overview of the statistical and analytical methods applied to the dataset for each locus. Each subsequent data chapter (III, IV, and V) consists of Results, Discussion and Conclusion for analysis of each locus. Chapter VI, the final chapter of this dissertation, will contain all concluding remarks and suggestions for future study of the *NAT* loci.

DEDICATION

For my Grandma Erma, who always wished she could just finish that last Spanish class.

ACKNOWLEDGMENTS

I would like to give recognition to the following people and institutions for their help during my development and training, which culminated in this project and other related areas of my life.

Dr. Sarah Tishkoff, for the opportunity to devote myself to the study of African genetic diversity, both in the field and in the laboratory, for her advice, generosity and friendship, and for the inspiration to pursue my scientific interests with enthusiasm and drive.

Dr. Joanna Mountain, for being an excellent role model and mentor, for giving me my first opportunity as a Master's student to study African genetics, helping me to interact with Luca Cavalli-Sforza and Joe Greenberg (interactions that I will never forget) and for showing me her love for Tanzania, for believing in my abilities, for guiding me in my pursuit of a scientific career, and ultimately directing me to work with Sarah.

I also thank my Committee members, Cristian Castillo-Davis, Chuck Delwiche, Charlie Mitter, Eric Haag, Steve Mount, as well as past members Mary Carrington, Matt Hare and Mike Cummings, for constructive suggestions, helpful advice and encouragement with this project, and related aspects prior to and following the inception of this project.

My friends and colleagues from UC Davis: I would like to acknowledge Dr. David Glen Smith, for suggesting that I look at linguistic and genetic relationships in Native American populations, and for making sure I never forgot

where my roots were. I thank Dr. Rika Kaestle and Dr. Ripan Malhi for giving me their time and patience when I was young and impressionable.

I would like to thank my mother and father, for their emotional and logistic support throughout this process, and in helping me become the person that I am. For my Dad, specifically, I would like to thank him for always listening to what was happening and being involved in the details, being objective, and always knowing how people should act and the right thing to do. For his tireless help with computers and all things technical, for shipping the required supply of Hawaiian Kona to complete this project, and for the immense quantities of "big daddy hugs" that were sent via telephone over the years.

I would like to thank my very best friends Erika Tauriello, Jack Suter, Anne Sweeney (for helping me know I am really a rock-star), Christina and Chris McPhee, and Alessia Ranciaro, for all their help at some of the more trying times, in celebrating the small triumphs, and for helping me realize that life is still going on despite this work.

I would like to thank the Howell family for all their love and support during these last years, for reminding me who I am and who I want to be, and for helping me to feel apart of something bigger.

I would like to give my deepest gratitude to Pamela Gandy, for using her skills as a mother, a confidant and a healthcare professional extraordinaire, and for caring for my personal well-being in my years at the University of Maryland, College Park.

I am grateful to David Barker and Sotiria Boukouvala for helpful discussion of some of the more difficult aspects of *NAT* genetics, and especially for their encouragement.

The first two years of this work, in my time as a Master's student, were supported by fellowship from the Committee for African Studies and the Morrison Institute for Population and Resource Studies (special thanks to Marc Feldman), at Stanford.

As a doctoral student, my first six years were supported by a National Science Foundation Integrative Graduate Research and Training (IGERT) Fellowship in Human Evolutionary Biology, in conjunction with George Washington University. Special thanks to Bernard Wood for his undying enthusiasm and support throughout.

Support for my last year, and to attend scientific meetings, was generously provided by the Department of Biology at the University of Maryland, but mostly by my advisor, Dr. Sarah Tishkoff.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I. BACKGROUND TO DISSERTATION

It is well known that susceptibility to many complex diseases (*i.e.* cardiovascular disease, diabetes, obesity, and cancer) is influenced by both genetic and environmental factors. Characterization of genetic variation at loci known to be involved in metabolic processes is one way to address how genetic factors manifest to cause disease. The identification and subsequent characterization of genes encoding drug-metabolizing enzyme (DME) loci represent a unique opportunity to understand the larger dynamic of human disease and environmental adaptation, and to observe how environmental toxins, referred to as xenobiotics, have been a selective force in shaping the genome throughout the course of human evolution.

The evolution of our species, both within and out of Africa, can be characterized by migration events, population expansions and contractions, climatic and dietary shifts, and differential exposure to pathogens. Human migrations exposed our ancestors to varying environments and diseases. Individuals who were better adapted to specific environments had a greater probability of passing on their genes to subsequent generations. This process of natural selection has shaped the pattern of genetic variation in human populations, leaving a genetic signature at genes that may underlie variation in response to environmental toxins, and consequently pharmaceutical drugs. The application of an evolutionary perspective to the study of drug-metabolizing enzyme loci is a novel and integrative approach to identifying variants that may play a role in drug response and disease phenotypes. The greater genetic diversity within Africa reflects a larger long-term effective population size relative to other regions of the world (Bowcock et al., 1994;

Quintana-Murci et al., 1999; Reed and Tishkoff, 2006; Tishkoff et al., 1996; Tishkoff and Verrelli, 2003). This finding, in the context of earlier paleontological (Lahr and Foley, 1998; Stringer and Andrews, 1988) and archeological studies (Ambrose; Ambrose; Klien, 1992), indicates that East Africa is the most likely site of origin of anatomically modern humans. Additionally, Africans exhibit a number of genetic adaptations that have presumably evolved in response to diverse climates, diets and exposures to infectious disease (Campbell, 2008), where non-Africans represent only a subset of that diversity, making the study of the genetic relationship amongst extant African populations of particular interest to medical and drug development studies.

Drug Metabolizing Enzyme (DME) Loci

Although, there are many examples of humans showing idiosyncratic responses to pharmaceutical drugs, the roles of drug metabolizing enzyme loci (DMEs) in the metabolism of pharmaceutical substrates has been firmly established in <1% of cases (Nebert and Dieter, 2000). There appears to be a connection between metabolic reactions involving pharmaceutical substrates and metabolic reactions of substances that have more common sources (*e.g.* food sources); pharmaceutical drugs are usually plant metabolites, or derived from plant metabolites (Nebert and Dieter, 2000). Additionally, many DMEs have endogenous compounds as natural substrates and genes encoding DMEs have functioned in many critical life processes in plants and animals (Nebert, 1997) throughout evolutionary time. With this in mind, I suggest that DME loci would be more accurately referred to as DETOXIFICATION ENZYME LOCI.

Pharmaco*genetics/genomics*

Pharmaco*genetics*, coined by Vogel in 1959, refers to inherited, inter-individual variation in drug response (Weinshilboum, 2004) and focuses on the pharmacological consequences of single-gene mutations. Pharmaco*genomics* considers the genetic effects of numerous, single genes, as well as the epistatic interaction among genes, whereas pharmaco*genetics* is often used to define a narrow spectrum of inherited differences in metabolism and disposition. However, the two terms are now commonly used interchangeably, making these distinctions somewhat arbitrary.

Modern awareness and interest in pharmacogenetics did not occur until variation was observed in response to isoniazid and succinylcholine, two therapeutics commonly used to treat patients with tuberculosis and pseudocholinesterase deficiency, respectively, in the 1940's and 1950's. Isoniazid induced neuropathy, which we now know is due to variation at *N-acetyltransferase 2* (*NAT2*), was observed as a relatively common adverse reaction (Lash et al., 2003). Clinical observations of large patient differences in dose response to these drugs are now "classic" examples in pharmacogenetics, and have given way to the concept that inheritance can have an important role in individual variation in drug response.

General Discussions of Genetic Response to Environmental Toxins

In general, genetic factors may affect the toxicity of foreign compounds entering the body in two ways: (1) by influencing the individual's *response* to the compound; and (2) by affecting the *disposition* of the compound within the system (*i.e.* the individual) (Timbrell, 2000).

A well-known example of individual *response* to a compound is *G6PD* deficiency in humans, which is associated with haemolytic anemia, as well as several other clinical disorders. Variants at the *G6PD* locus, located on the long arm of the X chromosome (Tishkoff and Verrelli, 2003), contribute to increased sensitivity or response (haemolytic anemia) to drugs and food sources, such as primaquine or dapsone and fava beans (Timbrell, 2000). Heightened sensitivity to alcohol in Asian populations is another example of genetically determined increased response.

Possibly more important with regard to individual toxic response to foreign chemicals is the case when the *disposition* of the substrate is affected by genetic factors that influence the compounds toxicity. (Timbrell, 2000). In this case, variation in individual response is only partly due to genetic factors, and many other influences in the human environment may also affect drug disposition. It cannot be overlooked that aside from the individual genetic predisposition, there are several confounding factors that may influence an individual's ability to process and clear a toxin (for example, diet, presence of other drugs/xenobiotics in the system, and intracellular processes). The scale of variability is exemplified in human response to paracetamol (acetaminophen), where rates of metabolic oxidation of the drug exhibit a ten-fold range of variation (Timbrell, 2000). High inter-individual variability, such as is the case for paracetamol, suggests that a reactive metabolite, generally resulting in the addition of an endogenous functional group to the original substrate, is responsible for toxicity, rather than the parent drug itself (Timbrell, 2000).

To further describe *disposition*, the distinction between metabolic reactions involving DMEs separates these enzymes further into two groups based on the reaction

types they participate in, referred to as *phase I* and *phase II* (Figure 1.1). Xenobiotics have a variety of effects on biological systems. Many drugs must undergo biotransformation to be effective, being that a metabolite rather than the parent compound exerts the pharmacological effect. Biotransformation of a substrate can thus be beneficial, with a positive pharmacological response, or deleterious, leading to toxicity or carcinogenicity.

In most cases, endogenous biotransformation leads to inactivation (and detoxification) of the drug, lessening its toxicity and preparing it for excretion. Phase I reactions, mostly composed of CYP P450 enzymes, modify the compound by adding a functional group, typically using oxygen to create an active site (Liska, 1998). The result is a slightly more hydrophilic metabolite. Phase II reactions are conjugation reactions that generally follow Phase I reactions, "resulting in a xenobiotic that has been transformed into a water-soluble compound that can be excreted through urine or bile" (Liska, 1998). Infrequently, phase II biotransformation has been observed to be preceded by phase I, as is the case for morphine that forms its metabolite by direct conjugation (Fura, 2006). It is important to point out that just as biotransformation at phase II can increase the solubility or reactivity of a metabolite, so can the synergistic action of the DME with other DMEs or environmental agents. For this reason, the difference in activation or degradation of a substrate could be substantial (30->40 fold) (Nebert and Dieter, 2000). This can result in large inter-individual variation in drug response, and also in individual risk to environmentally caused toxicity and cancer.

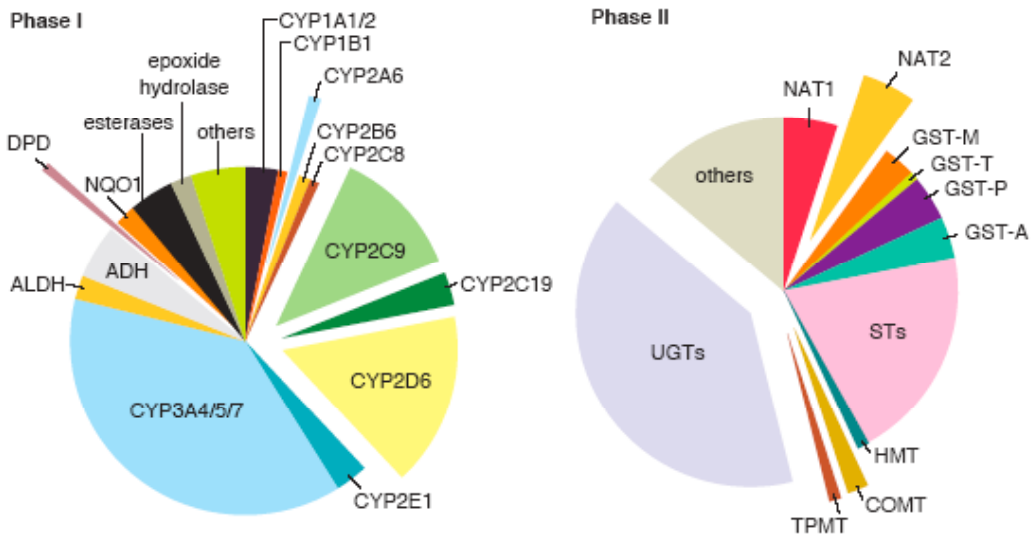


Figure 1.1

Figure 1.1: DMEs are illustrated according to the percentage of contribution to phase I and phase II metabolism. Phase I reaction specifies modification of functional groups, whereas Phase II indicates conjugation with endogenous substituent(s). Separated sections indicate that the particular enzyme polymorphism has been associated with changes in drug effects (*e.g.* NAT2 is associated with variable response to Isoniazid (INH)). (Evans and Relling, 1999).

Molecular genetics of DME loci

Structure of DMEs: Duplications-tandem orientation

The largest DME gene family described to date is the CYP P450 gene family of xenobiotic detoxifiers. The CYP gene family, in mouse and human, is characterized by dense gene clusters that arose by recurrent, local gene duplication (Thomas, 2007). According to gene duplication theory (Ohno, 1970), one cause for duplication is to provide novel protein functions. Duplications make it possible for proteins to specialize in function, contributing to the adaptive evolution of the organism. In fact, recent genetic evidence suggests that gene duplication (and loss) may have played a greater role than nucleotide substitution in the evolution of phenotypes specific to the human lineage (Demuth et al., 2006). Like the CYP P450 genes, the *N-acetyltransferase (NAT)* genes presumable arose via tandem duplication, and act as xenobiotic detoxifiers in humans. The *NAT* genes differ from most other DME genes, like the CYP P450s, in that they have relatively short, intronless coding regions and have only two functional copies in humans. In addition, unlike other DME genes in humans that have been observed to have copy number variants with as many as 13 functional gene duplications (*i.e*. P450 *CYP2D6*) (Nelson et al., 2004), there is no known copy number variation at the *NAT* loci in humans.

NAT Isozymes

In humans there are two NAT isoenzymes, NAT1 and NAT2, and their acetylation reactions are affected by genetic polymorphisms in the *NAT1* and *NAT2* genes. *NAT* isozymes function in phase II conjugation reactions. Arylamine *N*-acetyltransferase (EC

2.3.1.5) is the enzymatic activity responsible for *N*-acetylation, the major detoxification route for arylamines and arylhydrazines, including many drugs and carcinogens. This family of isoenzymes is distinct from those involved in melatonin biosynthesis (EC 2.3.1.87), such as arylalkylamine *N*-acetyltransferase, but indistinguishable from *N*-hydroxylamine *O*-acetyltransferases (EC 2.3.1.118) (Boukouvala and Fakis, 2005). NATs are also involved in bioactivation reactions with substrates derived during phase I reactions mediated by P450 CYP1A2, which form highly reactive mutagenic compounds with cellular DNA (see (Grant, 1993)for illustration of the potential pathways). It is these proposed bioactivation reactions that result in the hypothesized role of NAT in cancer predisposition.

Proposed NAT Substrates

Research attempting to elucidate the particular environmental toxins and endogenous substrates that are involved in the *N*-acetylation pathway has spanned the last five decades. *NAT* genes have been shown to be involved in the metabolic breakdown of several clinically relevant compounds (Hein, 2002), including isoniazid (INH). Other substrates include the anti-arrhythmic drug procainamide (PA), several anti-bacterial sulphonamides and anti-inflammatory 5-aminosalictic acid (5-AS), the anti-hypertensive drug hydralazine, and dapsone, used in the treatment of malaria and leprosy (Boukouvala and Fakis, 2005). Examples of toxins, other than pharmaceutical drugs, that are metabolized via N-acetylation include various heterocyclic amine and arylamine (HCA) carcinogens present in common food pyrolysis products, tobacco products, and resulting from certain industrial processes (Felton et al., 1997, Schut, 1999; Kufe et al., 2003).

Additionally, differences in NAT1 and NAT2 substrate selectivity have been observed. Both NAT1 and NAT2 are capable of *N*-acetylation of arylamines (Hein et al., 1993); however, most adverse drug reactions involving the *N*-acetylation pathway can be linked to the NAT2 isoform (Rodrigues-Lima et al., 2003) because NAT2 is expressed mainly in hepatic (liver) tissue (Kilbane et al., 1991), where metabolism of hepatotoxins occurs. Studies of the metabolism of pharmaceuticals via the NAT pathway, including INH, generally focus on the role of *NAT2* isoforms, and for this reason *NAT2* has been studied more extensively than *NAT1*.

NAT gene structure in humans and other species

The identification and cloning of genes encoding NAT isoforms has indicated the presence of NAT enzymatic activity in several eukaryotic species. Certain exceptions have been found where no gene has been identified, but NAT enzymatic activity has been reported in the same or related species (*e.g.* nematodes: *C. elegans* and *C. briggsae*; fungus: *C. albicans*; eight different plants species including *A. thaliana* and *O. sativa*) (Boukouvala and Sim, 2005). The presence of NAT protein has been predicted in all analyzed *Mammalia*, with the exception of canids that have no *NAT* genes in their genome (Trepanier et al., 1997). Figure 1.2 illustrates the phylogenetic relationship determined from amino acid alignment of NAT enzymes in both prokaryotic and eukaryotic species (Butcher et al., 2003). Amino acid sequence identity of NAT proteins is observed to be ~80% between mammals (Boukouvala and Sim, 2005).

Figure 1.3 illustrates what is currently known about the structural organization of *NAT* genes in three vertebrate species, namely human, mouse, and rat. Examination of

Figure 1.3 reveals a genomic structure where the entire coding region is contained in a single exon, forming a continuous ORF of 870 bp (Blum et al., 1990) with one or more non-coding exons (NCEs). The presence of one or more NCEs is typical of all vertebrate *NAT* genes, with the presence of an uninterrupted coding region containing a single exon and 3' UTR region (Boukouvala and Sim, 2005). Alternative splicing has also been confirmed experimentally for human *NAT1* and *NAT2* (Boukouvala and Sim, 2005; Butcher et al., 2005; Husain et al., 2004). Additionally, several studies (see Boukoulova, 2005 for references) have suggested the "differential usage of multiple, tandem poly-A signals, leading to the production of alternatively poly-Adenylated mRNAs of NAT in different vertebrate species" (Boukouvala and Fakis, 2005).

Both expressed *NAT* genes in humans, located on chromosome 8p21.3-23.1, contain 870 bp intronless, protein-coding regions. Genes with intronless coding regions make up <5% of human genes (Boukouvala et al., 2003). These genes often have upstream non-coding exons (Mummidi et al., 1997), as is the case for *NAT1* and *NAT2* (Figure 1.4). *NAT1* and *NAT2* share 87% sequence similarity (Blum et al., 1990), indicating that they likely arose via a gene duplication event. *NAT1* and *NAT2* are separated by a 146 kilobase (kb) region, which encompasses a non-transcribed pseudogene, *NATP1* that shares ~83-85% sequence homology with the coding loci (Blum et al., 1990) (Figure 1.4). Figure 1.5 illustrates the pattern of linkage disequilibrium (LD) across a >200 kb region encompassing the NAT loci, as determined by the SNP data generated by the International HapMap Consortium (Consortium, 2003). This data demonstrates the lack of LD between the individual *NAT* loci and illustrates that each locus has evolved independently from the other.

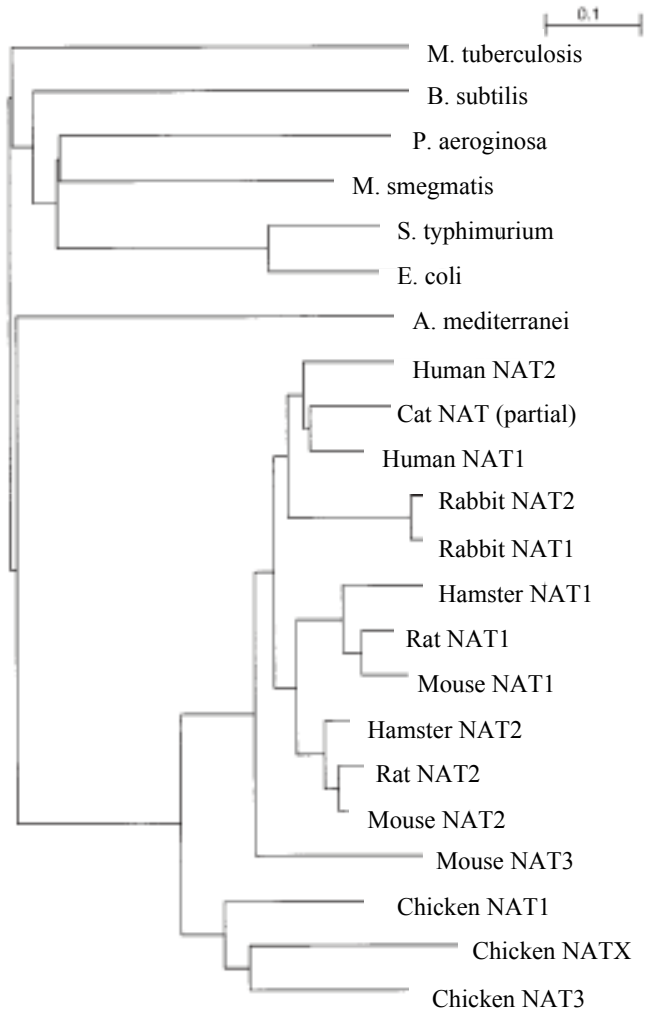


Figure 1.2

Figure 1.2: Phylogenetic tree based on amino acid alignment of Prokaryotic and Eukaryotic NATs. Chicken NATX (Genbank J03737) was uncharacterized at the time, but now corresponds to NAT3 (adapted from Butcher, 2002).

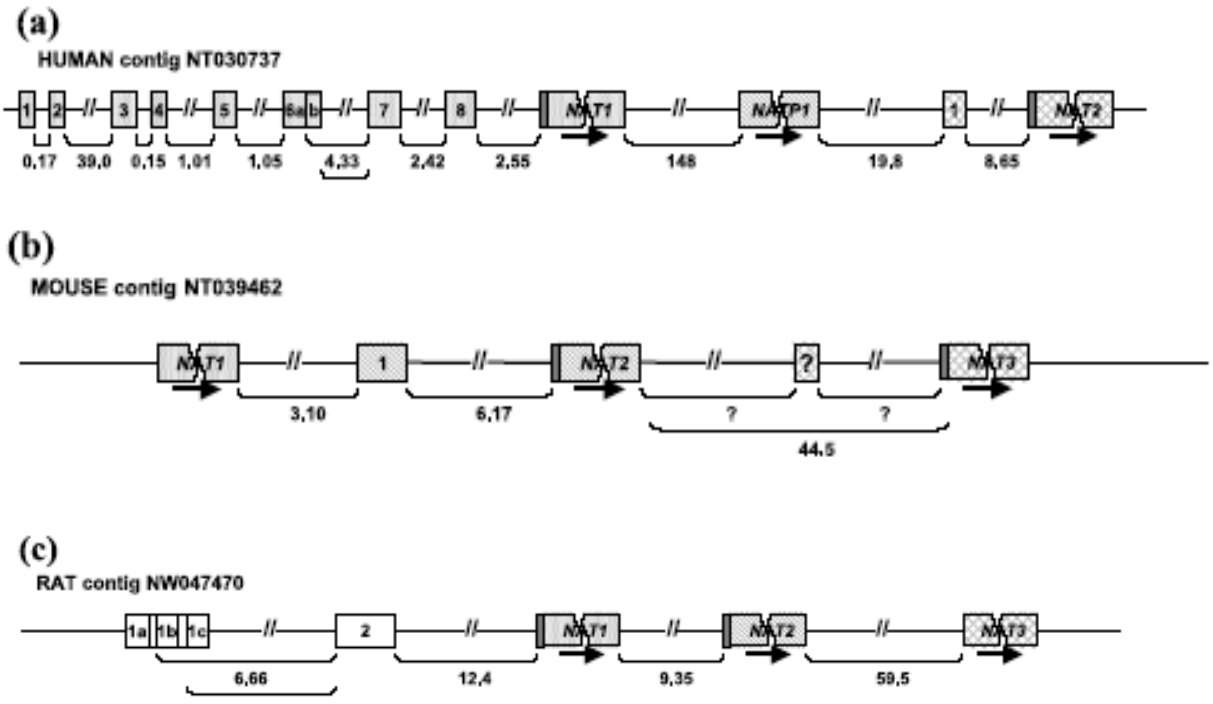


Figure 1.3

Figure 1.3: Structural organization and intron and exon structure of the *NAT1*, *NAT2*, and *NATP1* loci in the human, mouse, and rat. Distances between exon regions are indicated in kilobases (question marks indicate unknown distance). Numbered boxes indicate non-coding exons. Arrows indicate orientation of transcript (taken from (Boukouvala and Fakis, 2005). According to the Mouse Genome Sequencing Consortium (2002), the box indicated in (b) with a question mark, between *NAT2* and *NAT3* has been confirmed to be a pseudogene in mouse.

N-Acetyltransferase (*NAT*) Chromosomal Region- 8p22

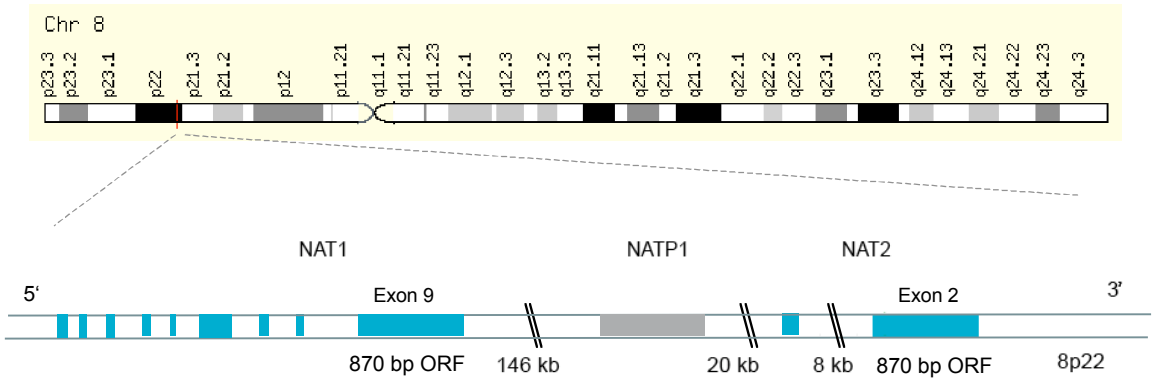


Figure 1.4

Figure 1.4: *NAT* chromosomal region 8p22, spanning >200 kb in humans. Relative positions of *NAT1* and *NAT2* ORFs, and non-coding exons (NCEs) and the single pseudogene *NATP1* are indicated.

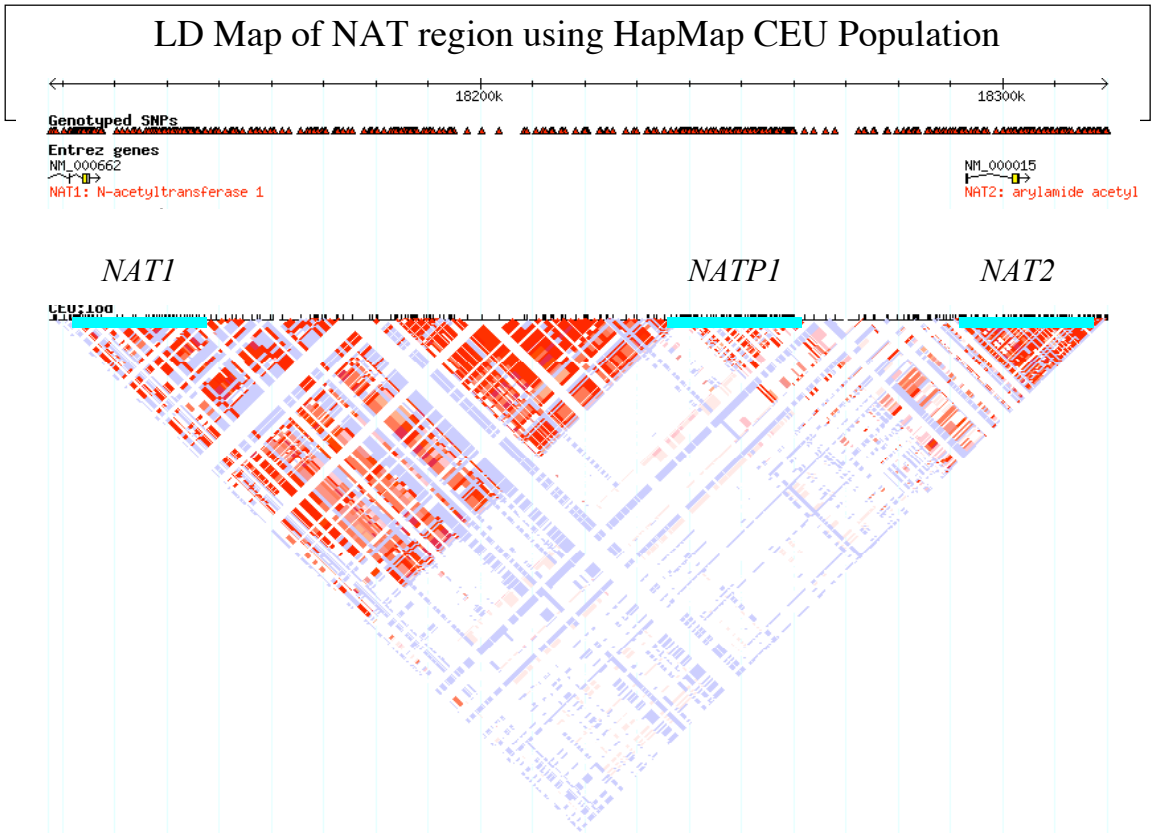


Figure 1.5

Figure 1.5: Linkage and recombination map of the *NAT* region located on chromosome 8p22. Map was generated using Haploview version 4.1 with HapMap data from Utah residents with ancestry from northern and western Europe (abbreviated as CEU). Turquoise bars indicate the locations of *NAT1*, *NATP1*, and *NAT2,* as indicated on the graph. Red signifies linkage where D'=1 with a LOD score ≥2. Blue followed by white indicates lessening of linkage parameter D', as well as confidence of the estimate (*e.g.* blue specifies D'=1 (LOD<2); pink and shades of red specify D'<1 (LOD≥2), and white specifies D'<1 (LOD<2)). Similar results were obtained from all HapMap populations (Yoruba, Nigeria (YRI); Japanese, Tokyo, Japan (JPT); Han Chinese, Beijing, China (CHB). The lack of linkage disequilibrium between the *NAT* loci at this level of heterogeneity indicates that each locus is evolving independently.

Hypothesized regulatory role for human NAT1 UTRs

Upstream NCEs are thought to be common in higher eukaryotes, especially in the presence of intronless coding regions (Mummidi et al., 1997; Sosinsky et al., 2000). It has also been shown that the introduction of intron regions between promoter and intron-less coding region enhances accumulation of the spliced message in the cytoplasm (Palmiter et al., 1991). It has been proposed by Boukouvala and Fakis (2005) that upstream NCEs may be essential in the "nucleo-cytoplasmic" transcript of NAT mRNA. Further, studies of differential utilization of upstream NCEs indicate that the process is often linked to cell-specific gene expression and regulation of both transcriptional and translational efficiency (Sonenberg, 1994). Similar to the effect of differential NCE usage, differential utilization of polyadenylation (poly-A) signals located in the 3' UTR region may increase variety of *NAT* mRNA transcripts and play a role in gene regulation (Boukouvala and Fakis, 2005). Aside from observed differences in the strength of particular poly-A signals, alternate usage of polyadenylation signals is far from clarified (Edwalds-Gilbert et al., 1997). At present, the role of 3' UTR elements in modulation of transcription and/or translation has not been verified for the *NAT* loci (Barker, pers. comm.; Boukouvala and Fakis, 2005). However, it is clear that *NAT1* has maintained 3' UTR elements for some critical purpose.

NAT1 structure and regulation

In the human genome, ~18% of all genes exhibit the use of alternate promoters (Landry et al., 2003), and it is the general consensus that splice variants that alter the 5'

UTR, but not the protein-coding region, often show altered translational efficiencies and tissue specific expression. Recent studies investigating the position of *NAT1* regulatory elements (Barker et al., 2006; Butcher et al., 2005) have found evidence for the use of alternative promoters. There are three putative promoter sites 5' of exon 9, the main *NAT1* ORF (Figure1.5). These are located upstream -245 bp, -11.8kb, and -51.5 kb from the +1 ATG start site, and are designated P1, P2, and P3, respectively (Barker et al., 2006; Butcher et al., 2005; Husain et al., 2004). Recent reports of *NAT1* transcription have also indicated the presence of at least 10 non-coding exons (NCE) (Husain et al., 2004). NCE locations have been confirmed at eight positions 51.5, 51.4, 12.3, 11.9, 10.8, 9.6, 5.2 and 2.6 kb upstream of the single coding exon 9 (Figure 1.6) (Barker et al., 2006; Boukouvala and Fakis, 2005; Butcher et al., 2005).

NAT1 is expressed ubiquitously in humans. The NAT1 isoenzyme has been found in all tissues examined thus far, including liver, intestine, bladder, breast, and placenta (see Boukouvala and Fakis, 2005 for individual studies). NAT1 activity has also been detected in fetal tissues (Pacifici et al., 1986). Most NAT1 mRNAs are found to contain the 79bp NCE 8 (Figure 1.6) (Barker et al., 2006). Less frequently, one or more other NCEs located in the 5'UTR are found in NAT1 mRNAs, resulting in altered translational efficiency (Butcher et al., 2005). Butcher (2005) has demonstrated that human cell lines derived from blood or different tissues have differing major *NAT1* transcripts that bear different combinations of NCEs. Certain NCEs (see Figure 1.6) have also been reported to be more or less tissue-restricted than others (where NCEs 4 and 8 are less restricted than 5, 6, and 7) (Boukouvala and Sim, 2005). Three polyadenylation sites have been confirmed at the 3' end of the *NAT1* exon 9 (Husain et al., 2004). These are located

downstream of the coding region at positions +1085, +1203, and +1242, relative to the +1 ATG of exon 9. One possible endogenous role for NAT1 has been elucidated in humans (Minchin, 1995), where NAT1 is capable of N-acetylation of the folate catabolite p-aminobenzoylglutamate.

NAT2 structure and regulation

In contrast to *NAT1,* human NAT2 mRNA synthesis is less complex and involves a single promoter ~8.7 kb upstream of exon 2 (Husain et al., 2007) (see Figure 1.7). The majority of *NAT2* mRNAs initiate from regions located within exon 1, a short NCE (Figure 1.7) at -8682 bp and -8752 bp 5' of the +1 ATG start site (Wakefield, pers. comm.). Exon 1 is separated from the ORF of exon 2 by an ~8kb intron. A second promoter region located just upstream from the *NAT2* exon 2 ORF has been identified (Boukoulovala, pers. comm. ; Boukouvala and Fakis, 2005) (Figure 1.7). More research needs to be done to confirm the functional effects of this putative promoter region; however, current work has indicated that mRNA expression from this putative promoter was not found in sufficient quantity in any tissue to alter the relative level of *NAT2* mRNA established by the promoter located -8.7 kb upstream of exon 2 (Husain et al., 2007). Human NAT2 isoenzyme shows a more restricted tissue-distribution profile, as compared to NAT1. According to Husain (2007), *NAT2* mRNA levels are highest in liver, with slightly lower amounts in small intestine and colon, in agreement with early findings (Hickman et al., 1998; Jenne, 1965). It has been well established that *NAT2* activity is responsible for the inactivation of the anti-tuberculosis drug isoniazid (Reif, 1953).

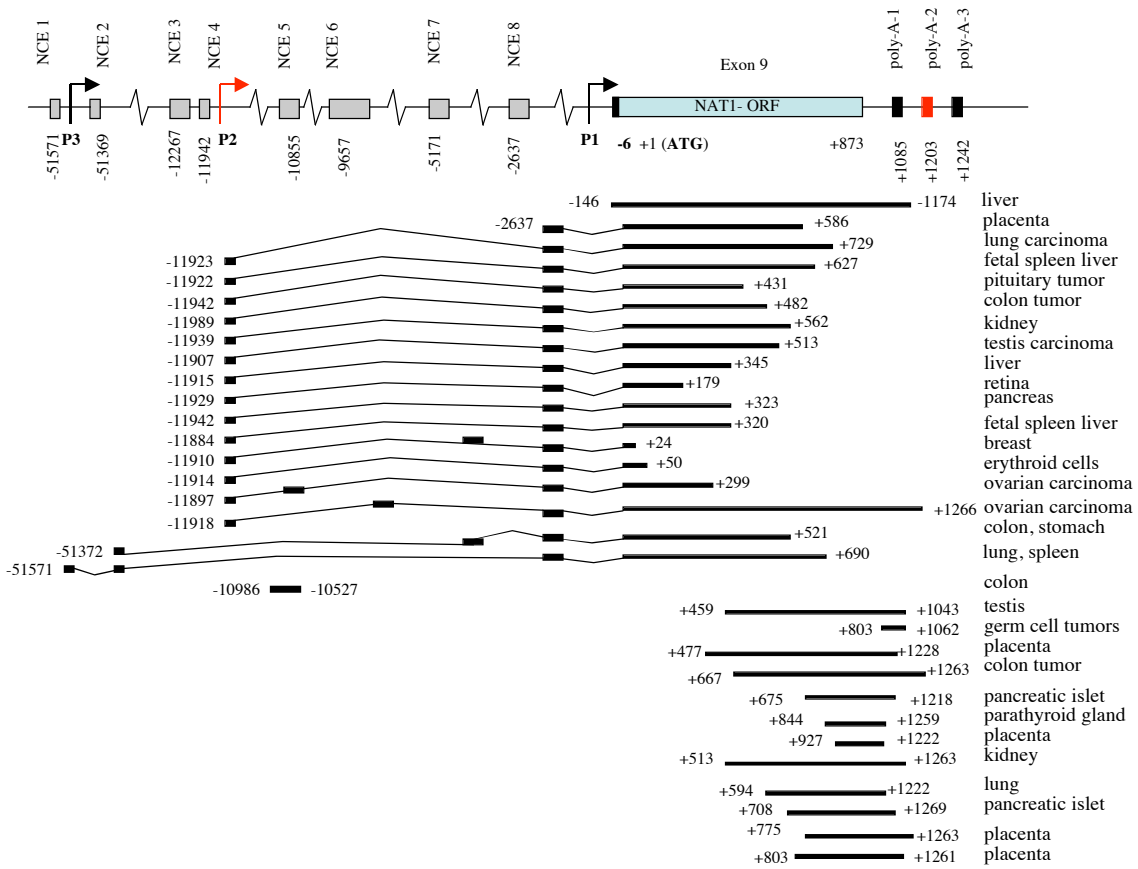


Figure 1.6

Figure 1.6: Structural organization of the *NAT1* gene region. All locations are numbered according to the +1 (ATG) site of the ORF (exon 9 for *NAT1*). Small grey boxes indicate identified non-coding exons. Black and red boxes indicate 3' poly-A sites, where the red box represents the most commonly used polyA site, poly-A-2, located at position +1203. The black box located at position -6 indicates the conserved splice site proximal to the coding region of exon 9. Black arrows indicate established promoter locations, whereas red indicates the most commonly used promoter, P2. Promoter nomenclature for *NAT1* follows that of Butcher (2005), and contrasts to that presented in both Husain (2004) and Barker (2006). Beneath the schematic of *NAT1* are selected, representative expressed sequence tags (ESTs) (adapted from (Boukouvala and Sim, 2005), where these authors recovered NAT1 sequences from dbEST database and aligned with genomic contig NT030737, which spanned 8p22-p23.1. Tissue origin is listed to the right. The top-most EST (derived from liver) was originally reported by (Ohsako and Deguchi, 1990).

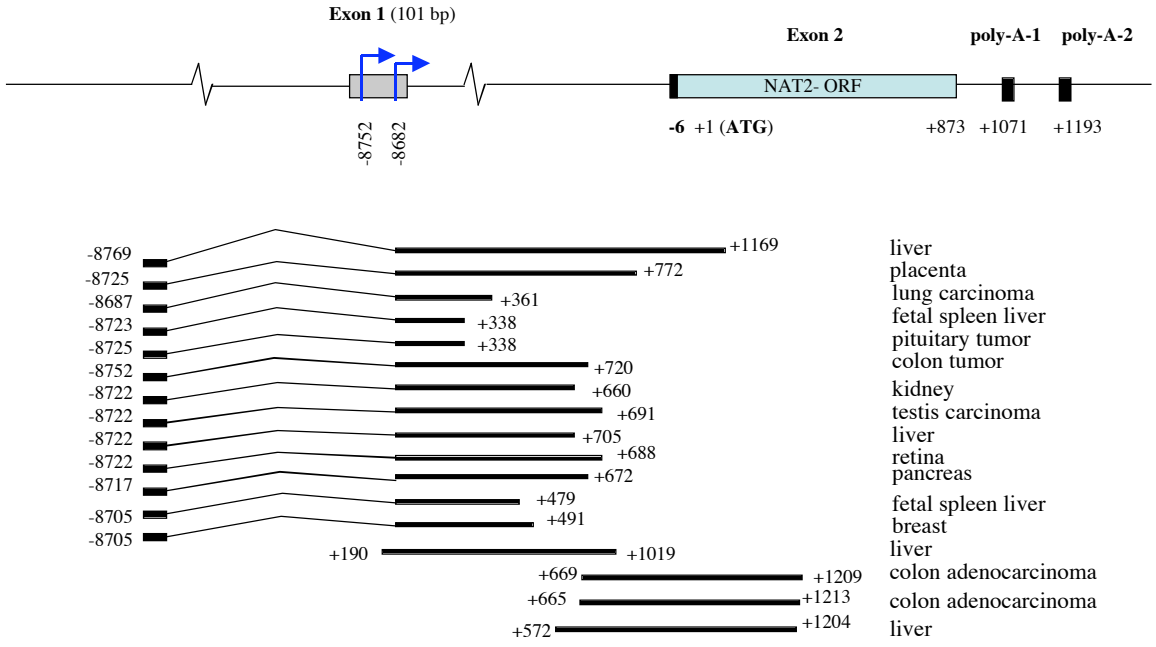


Figure 1.7

Figure 1.7: Structural organization of the *NAT2* gene region. All locations are numbered according to the +1 (ATG) site of the ORF (exon 2 for *NAT2*). The small grey box indicates exon 1, the single non-coding exon for the *NAT2* region. Blue arrows indicate identified transcription initiation locations (exact promoter locations are not currently known). The black box located at -6 indicates the conserved splice site proximal to the coding region of exon 2. The two black boxes downstream of exon 2 indicate the two polyadenylation sites identified for *NAT2*, at +1071 and +1193. Beneath the schematic of *NAT2* are selected, representative expressed sequence tags (ESTs) (adapted from (Boukouvala and Sim, 2005), where these authors recovered *NAT2* sequences from dbEST database and aligned with genomic contig NT030737, which spanned 8p22-p23.1. NAT2 specific ESTs derive from liver and colon, listed to the right. The top-most EST (derived from liver) was originally reported by (Ohsako and Deguchi, 1990).

Human NAT gene SNPs

A number of single nucleotide polymorphisms (SNPs) have been identified in *NAT1* and *NAT2* (Sim et al., 2000). *NAT1* coding region polymorphisms that *reduce* enzyme activity are found to be rare on a population basis, whereas *NAT2* polymorphisms that *reduce* enzyme activity are found to be common in human populations (Barker et al., 2006). *NAT* SNP variants have been observed to give rise to alteration in enzyme stability and protein folding, and modification in catalytic efficiency (Hein, 2003; Westwood et al., 2006). These SNPs include non-synonymous amino acid changes, as well as silent mutations in the coding region, and 5' and 3' untranslated regions (UTRs).

NAT1 polymorphisms

For many decades *NAT1* was thought to be monomorphic in human populations because of a lack of observed variation in NAT1 acetylator phenotype. With the cloning of the human *NAT* genes (Vatsis and Weber, 1993), it was demonstrated that *NAT1* is polymorphic in human populations. At present, 26 distinct *NAT1* haplotypes have been observed. The first described haplotypes were *NAT1*4* and *NAT1*10* (Vatsis and Weber, 1993). These two variant haplotypes have identical coding regions, but differ only by two mutations in the 3' untranslated region (UTR), at positions +1088 and +1095. SNP +1088 A/T, located directly following a $(TAA)_n$ repeat region (Figure 3.2), is thought to abolish the first polyadenylation signal of the gene located at +1085 (see Figure 1.6) (Boukouvala and Fakis, 2005). *NAT1*10* was originally thought to have increased acetylation activity through utilization of an alternative (stronger) poly-A signal (Hein et al., 2000). Other

studies have failed to support this hypothesis, and according to Barker (pers. comm.) increased amounts of mRNA are being produced for *10, but are not exhibiting any affect on phenotype. So, the general consensus remains that *NAT1*10* confers the same acetylator activity as does *4 (Bruhn et al., 1999; Yang et al., 2000). Interestingly, many of the known haplotypes (*e.g*. haplotypes *18A, 18*B, 26*A, 26*B, and *28) differ only by insertions or deletions in close proximity to this poly-A region of the gene. However, because of the variation in acetylation activity associated with identical *NAT1* haplotypes, accurate prediction of NAT1 phenotype is not possible (Boukouvala and Fakis, 2005).

Variation in the coding region of *NAT1* is found to be quite rare overall. NAT1 haplotypes ***4* and ***10*, which differ at 3' sites only, are observed to be most common in human populations, whereas other haplotypes are represented, typically, at frequencies of <1% (Lin et al., 1998; Upton et al., 2001). In Europeans, frequencies of haplotypes ***4* and ***10* are 75%, and 20%, whereas haplotypes ***3, *11*, and ***14* are present at <4% frequency (Boukouvala and Fakis, 2005). In Japanese and Taiwanese populations haplotypes ***4, *3*, and ***10* have been reported at frequencies of 55%, 40%, and 3%, respectively (Boukouvala and Fakis, 2005). While in Chinese populations these three haplotypes are represented in approximately equal frequencies of 30-35% (Zhao et al., 1998).

NAT2 polymorphisms

Previous reports of the impact of genetic variation at NAT2 on acetylator status have shown that the 52 previously defined *NAT2* haplotypes are based on different

combination of 16 SNPs at positions +111, +190, +191, +282, +341, +364, +411, +434, +481, +499, +590, +759, +803, +845, +857, and +859 in the coding region of the gene (http://louisville.edu/medschool/pharmacology/NAT.html). Various *NAT2* haplotype groups are designated according to specific replacement changes at only a few sites. For example *NAT2*5* and *NAT2*6* are defined by +341 (Thr114Ile) and +590 (Arg197Gln), respectively. *NAT2*4* is considered WT, and represents a haplotype associated with rapid acetylator phenotype. *NAT2* *11A, *12A, *12B, *12C, *13A, *13B, and *18 haplotypes are also considered rapid because they are defined by polymorphisms that cause silent or conservative changes to the WT haplotype. The remaining 45 *NAT2* haplotypes are defined by replacement changes in the coding region that cause reduced activity and/or stability (Boukouvala and Fakis, 2005), conferring "slow" acetylator phenotypes. The first identified and most commonly observed haplotypes in European populations are the wild-type *NAT2*4* (rapid), *NAT2*5* (slow) and *NAT2*6* (slow) haplotypes (Boukouvala and Fakis, 2005).

NAT2 Acetylator Phenotype

One of the most well-known and fully described genetic factors in human drug metabolism and disposition is the acetylator phenotype. It has been known for more than thirty years that human populations exhibit variation in their ability to acetylate certain drugs. Isoniazid, used in the treatment of tuberculosis, was in fact the prototype in the study of drug acetylation in humans. Figure 1.8, illustrates the clear tri-modal distribution of the acetylator phenotype in response to isoniazid in human populations. It has been

suggested that the acetylator phenotype results from two alleles at a single locus, where the slow acetylator type is a simple Mendelian recessive trait. The dominant-recessive relationship, however, was unclear (Timbrell, 2000), because of a lack of ability to distinguish the intermediate acetylator phenotype depending on the particular substrate administered or method used. It is now commonly thought that human *NAT2* is co-dominantly expressed. Animal studies (rabbit, mouse, and hamster) have confirmed this, indicating a co-dominant expression pattern with three possible genotypes (i.e. rapid/rapid, rapid/slow, and slow/slow) (Timbrell, 2000). Though the exact biochemical pathway has not been established, ambiguity in expression pattern of NAT2 is most likely due to the fact that NAT2 acts as a phase II enzyme, which depends on the substrate and other factors specific to the individual. Modification of the substrate by CYP1A2 is also possible and may affect NAT2 phenotype classification.

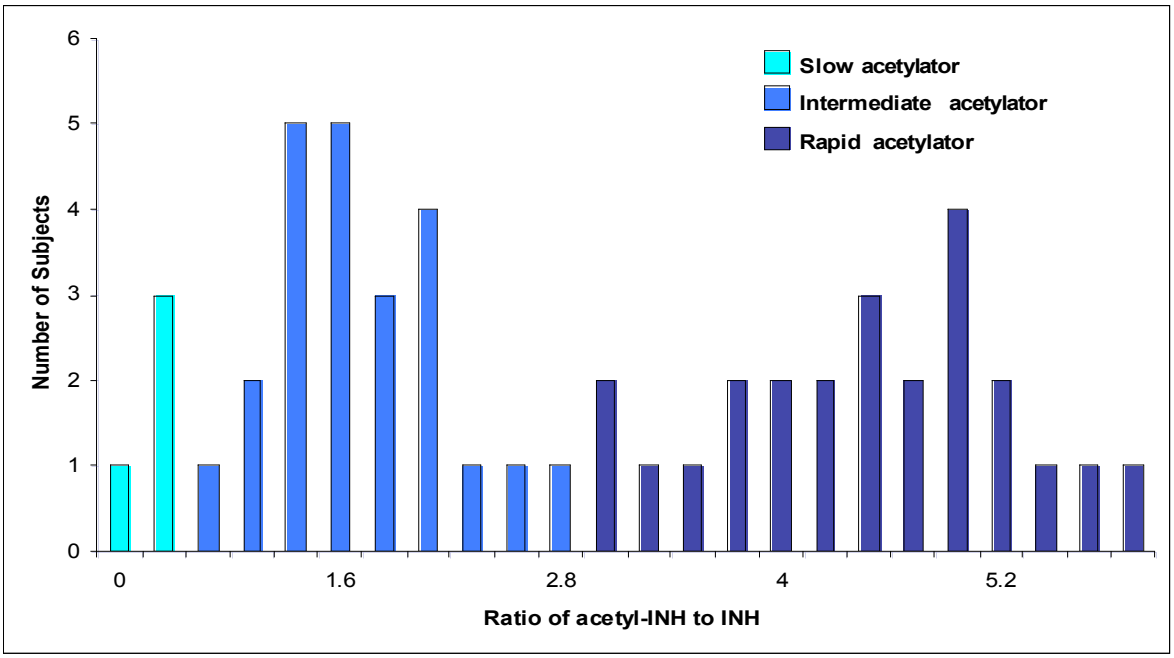


Figure 1.8

Figure 1.8: Tri-modal phenotype distribution of acetylators in response to isoniazid (INH) (adapted from (Mashimo et al., 1992)).

NAT Association Studies and Implications for Human Disease

It is currently well recognized that individual susceptibility to common disease is determined by both genetic and environmental factors. Examples of environmental influence on disease phenotypes include exposure to pollutants, industrial chemicals and harmful components of the human diet (Sim et al., 1995). The involvement of the functional *NAT* genes in human chemical carcinogenesis has been reviewed recently (Hein, 2002; Hein et al.).

Because NATs are xenobiotic metabolizing enzymes that catalyze metabolic reactions known to bioactivate or detoxify potentially harmful compounds entering the body, the association between *NAT* polymorphisms and disease has been extensively investigated (Boukouvala and Fakis, 2005). Several studies have reported association between *NAT* phenotype and cancer risk (Bell et al., 1995; Brockton et al., 2000; Roberts-Thomson et al., 1996), implicating both *NAT1* and *NAT2* genotypes. Although results have been contradictory in many studies, most investigators agree that *NAT* genotype is not a significant susceptibility factor in diseases ranging from Parkinson's disease, Alzheimer's, diabetes mellitus, lupus, and rheumatoid arthritis (Boukouvala and Fakis, 2005). However, progression of certain disorders, such as diabetes mellitus and HIV AIDS, have been indicated to interfere with individual acetylator status, causing unexpected genotype-phenotype discordances (Agundez et al., 1996; O'Neil et al., 1997; O'Neil et al., 2002).

CHAPTER II. MATERIALS AND METHODS

Samples Analyzed

I have sequenced and analyzed a total of 326, 285, and 304 globally diverse individuals for *NAT1*, *NAT2*, and *NATP1*, respectively, for a total of 652, 570, and 608 chromosomes. The description of samples analyzed for each of the three *NAT* loci is summarized in Table 2.1.

African Groups: A total of 100 individuals originating from Eastern Africa, collected by myself and other members of the Tishkoff lab, were included for analyses. These individuals were sampled in Tanzania and Sudan from the population groups listed in Table 2.1. Institutional Review Board (IRB) approval was obtained from the University of Maryland, College Park prior to sample collection. All participants >18 years of age gave informed consent before sampling. Research permits were obtained from the Tanzanian Commission for Science and Technology (COSTECH) and the Tanzanian National Institute for Medical Research (NIMR), as well as from the government of Sudan. A total of 82 individuals originating from Cameroon were collected, in collaboration with Sarah Tishkoff, by Alain Froment of the Pasteur Institute, in Paris, France after receiving permission from the government of Cameroon.

Diversity Panel: A global panel of 155 individuals was drawn from the HGDP-CEPH (Centre d'Etude du Polymorphisme Humain) Human Genome Diversity Cell Line Panel (Cann et al., 2002; Cavalli-Sforza, 2005). Population samples used are summarized in

Table 2.1, indicated using an asterisk (*), and grouped according to similar genetic ancestry as determined by STRUCTURE analysis (Pritchard et al., 2000) of genome-wide STRPs and In/del marker variants (Reed, unpublished) and of SNPs (Conrad et al., 2006; Rosenberg et al., 2005(Conrad et al., 2006; Rosenberg et al., 2005). African groups included in this sample set are Biaka Pygmy from CAR, San from Namibia, and Yoruba from Nigeria.

Laboratory Methods

DNA Extraction

During field collection 8 ml of peripheral blood was drawn from each individual involved in the study and white blood cells were isolated from the whole blood using a modified salting-out procedure (100 mM Tris-HCL pH 7.6, 40 mM EDTA pH 8.0, 50 mM NaCl, 0.2% SDS, 0.05% 8 mM Sodium Azide) (Miller et al., 1988). DNA was extracted at a later date in the Tishkoff Laboratory at the University of Maryland using the PUREGENE DNA Purification kit (Gentra Systems, Inc.). All extracted DNA was quantified using Pico Green reagent (Invitrogen) and the Wallace Victor^2 1420 Multi-Label Counter (Perkin-Elmer Life Sciences, Boston, MA) at 1.0 second per well.

Whole Genome Amplification (WGA)

DNA obtained from the CEPH panel, for the amplification of *NAT2* and *NATP1*, were degraded and/or in low concentration, prohibiting direct amplification by PCR. To alleviate this problem Whole Genome Amplification (WGA) was performed with CEPH stocks using the GenomiPhi HY DNA amplification kit (GE Heathcare). DNA replication

with WGA is extremely accurate due to the low error rate of ϕ29 DNA polymerase (1 in 10^6-10^7) compared to other enzymes. The final product of the WGA reaction results in ~40 μg of DNA per 50 μl reaction and was used directly for PCR amplification.

PCR Amplification of NAT regions

Three gene regions, *NAT1, NAT2* and *NATP1* located on chromosome 8p22 (Salem et al., 2005) (Figure 1.4), were targeted for amplification by the polymerase chain reaction (PCR) (Mullis et al., 1992). Primers were designed using the program Primer3 v 0.4.0 (http//: frodo.wi.mit.edu/) and verified by eye, using Genbank reference individuals (NAT1: X17059; NAT2: X14672; and NATP1: X17060). All primers used for this project are listed in Table 2.2. The total amplified product for each of the *NAT1, NAT2* and *NATP1* regions was obtained with a single PCR product, with the exception of *NAT2* and *NATP1* CEPH WGA products that were amplified in six overlapping segments of ~500-700bp in length (Figure 2.1). Additionally, to extend the total sequence of *NAT1* in all individuals, flanking 5' and 3' segments were included (Figure 2.1). Forward primers were used for sequencing in all cases, and reverse primers were used when necessary to allow for sequence confirmation (Figure 2.1; Table 2.2). All amplifications were performed using 1.0 unit of Platinum *HiFi* enzyme (Invitrogen) and contained 200 μM of each deoxynucleotide triphosphate (dNTP), 2 mM $MgSO_4$, and 100ng of genomic DNA in a final volume of 25 μl. Samples were denatured at 94°C for 1 minute, followed by 35 cycles of 94°C for 30 seconds, 55°C for 30 seconds, and 68°C for 1 minute per 1000 bp. The reaction was performed using a Peltier Thermal Cycler (MJ Research). PCR products

obtained were run on a 1.6% agarose gel with BenchTop pGEM DNA Marker (Promega).

NAT Resequencing

To examine nucleotide variability, direct sequencing of the *NAT* gene regions was performed using the population samples listed in Table 2.1. Because the *NAT* gene region has not been thoroughly characterized in African populations previously, this may be particularly important in the identification of novel variants.

PCR products were purified using the ExoSAP-IT process, as described by the manufacturer (US Biochemical Corp.). Sequencing reactions were subsequently prepared using this purified DNA. Nucleotide sequences for each gene region were generated in six overlapping sequence reads using the didioxy-BigDye kit (Applied Biosystems) and analyzed on the ABI 3730xl automated capillary sequencer.

Table 2.1: Populations included in the study of *NAT* nucleotide diversity				
		NAT1	***NATP1***	***NAT2***
AFRICA		**2N**	**2N**	**2N**
EAST				
Tanzania				
	Burunge (BG)	34	36	34
	Hadza (HZ)	32	28	28
	Maasai (MS)	32	26	28
	Sandawe (SW)	38	36	36
	Turu (TR)	32	30	30
Sudan	Dinka (DN)	18	30	26
CENTRAL				
CAR				
	Biaka Pygmy (BK) *	30	30	30
WEST				
Cameroon	Fulani (FU)	22	26	26
	Kanuri (KN)	26	26	24
	Lemande (LM)	26	28	28
	Mada (MD)	28	28	28
	Baka Pygmy (PB)	18	18	18
	Bakola Pygmy (PL)	14	14	14
Nigeria	Yoruba (YO)*	24	24	24
SOUTH				
Namibia	San (SA)*	14	14	14
Total Africa		**388**	**394**	**388**
EUROPE/ Middle East (ME)				
French	French (FR) *	22	22	16
Israel	Druze (DZ)*	24	22	20
Italy	Sardinian (SR)*	24	24	22
Pakistan	Brahui (PA) *	24	24	22
Russia	Russian (RU)*	24	12	12
Total Europe/ME		**118**	**104**	**92**
ASIA				
Cambodia	Cambodian (CB) *	14	14	14
China	Han (HA) *	24	24	20
Japan	Japanese (JP)*	22	20	18
New Guinea	Papuan (NG)*	18	4	4
Siberia	Yakut (SB)*	22	8	6
Total Asia		**100**	**70**	**62**
AMERICAS				
Brazil	Karitiana (BR)*	22	20	20
Mexico	Pima (PI) *	24	20	8
Total Americas		**46**	**40**	**28**
GRAND TOTAL		**652**	**608**	**570**
***HGDP-CEPH**				
2N=# of chromosomes				

Table 2.1

Table 2.1: Number and geographic distribution of samples included in the study are indicated (2N=# of chromosomes sampled). Geographic groupings of populations used in further analyses are based on results from STRUCTURE analyses using genome-wide markers (Conrad et al., 2006; Reed, unpublished; Rosenberg et al., 2005).

Table 2.2: *NAT* PCR and Sequencing Primers		
NAT1		
2856 bp	**Oligonucleotide**	**5'-3' Sequence**
	-1182_F	GCCAAGCACTGGTCCTAGAG
	-455_R	CAAAGTTTCCCGCTCAAAGA
	-574F	AAGCACTAGAACAGTAGGAA
	-194_R	CCCAGAATCCTGTGAGAAATG
	-78_F	GCCATAATTAGCCTACTCAA
	+301_R	GTGAATCATGCCAGTGCTGT
	+473_F	CTGGTATCTAGACCAAATC
	+933_F	CCTATCATGTATCTTCTGTAC
	+1387_F	TGCCATACAAGAATGAACATGA
	+1420_R	GGATTTCAACTCAGATCTAC
	+2066_R	AGCTGGAAAAGGCCAGTACA
NAT2		
3064 bp	**Oligonucleotide**	**5'-3' Sequence**
	-1134_F	TCCTACATAGTTTATGTGAC
	-545_F	ACACATCAAGAGTATTCTGT
	-443_R	GACCCACTTGATTTCCCATTT
	-77_R	GGATCTGGTGCTCAAGAATG
	-15_F	CTTGCTTAGGGGATCATGGACA
	+702_F	GTGGGCTTCATCCTCACCTAT
	+779_R	CAGCACTTCTTCAACCTCTTCC
	+1320_F	AATAGAGTCTTCCTCTCATC
	+1373_R	TGTGGCAAAGTATGGATGGA
	+1812_F	ACTGCAGATTTGTTCTTAAC
	+2412_R	ATTTCGCTTCCAGGTTGAAG
NATP1		
3000 bp	**Oligonucleotide**	**5'-3' Sequence**
	-1204_F	GGCATTACTGCCTAGAGCATTTT
	-1160_F	TCAAACTACCAATGACATTCTGC
	-577_R	GCAGATAACCTACAGAGTGGGA
	-631_F	TGGGACTCAAACTAAAGTGCTTGTA
	-60_F	CAAGGGAATCTAAGGGCAAAAG
	-14_R	TTGTTTAAGGGATCATGG
	+535_F	TGATCTCCGGGAAGAAAAGAAA
	+555_R	CCTTTACTCCTGAATCCTGAACA
	+1080_F	CCTGTGATTATCTTGGGAACCAT
	+1112_R	TGTGAGAAAATATTTAATGCGGG
	+1500_F	CAGTGCCAGACCTGGAGTAAA
	+1655_R	ATCTGCCTGCTTCTTCCTTG
	+1860_R	CCTGTGATTATCTTGGGAACCAT
*All primer numbering is with respect to the +1=**A**TG .		
NATP1 ATG start was inferred according to consensus alignment with *NAT1* and *NAT2*		
F and R denote Forward and Reverse Primers, respectively.		

Table 2.2

Table 2.2: PCR and sequencing primers for the *NAT1*, *NAT2*, and *NATP1* regions used in the present analyses. Primer numbering is according to the ATG +1 start site for all loci (consensus ATG for *NATP1*). F and R indicate forward and reverse primers in all instances.

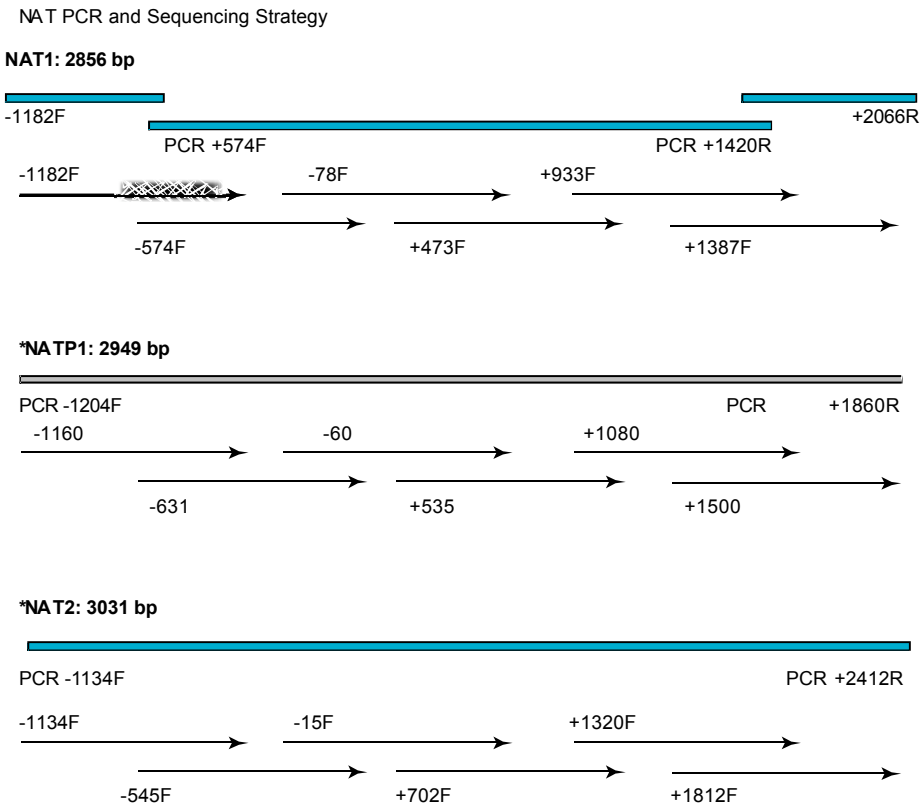


Figure 2.1

Figure 2.1: PCR and sequencing strategy used for each locus. Length of nucleotide sequence is indicated in base pairs (bp). Primer numbering corresponds to the +1ATG start site for each locus. *NATP1* numbering corresponds to the consensus +1 ATG, based on ClustalX alignment with both coding loci. *For clarity, only forward sequencing primers are indicated; see Table 3 for reverse sequencing primer numbers. For CEPH panel individuals for *NATP1* and *NAT2*, where whole genome amplification was necessary, PCR amplification was carried out in fragments of ~500-700 bp.

Data Analysis

Sequence Assembly and SNP Identification

NAT1, *NAT2* and *NATP1* sequences were edited and assembled into contigs using the programs Sequencher version 4.8 for Macintosh (Gene Codes Corp.] and the Phred, Phrap, and Consed suite for the Linux operating system (Ewing and Green, 1998a; Ewing et al., 1998b; Gordon et al., 1998). SNPs were called automatically using the Polyphred (Nickerson et al., 1997) software program, which tags SNP variants within Consed. All Polyphred SNP calls were then rechecked by eye for accuracy; SNPs identified as occurring once or twice in the dataset (singletons and doubletons, respectively) were identified at this stage. All singletons and doubletons were then confirmed by resequencing small segments of DNA containing these variants, using both forward and reverse primers.

Haplotype Reconstruction

Polyphred SNP calls were used to construct two haplotypes for each individual. Individuals with ambiguous base calls spanning greater than 200 bp in length were removed prior to phase inference (see Appendix 3). Diploid haplotypes were inferred across the *NAT1*, *NAT2*, and *NATP1* regions using the program PHASE VERSION 2.1.1 (Stephens et al., 2001), and according to the process outlined in Appendix 1, which reconstructs haplotypes from population genotype data using a coalescence model-based algorithm and is the most accurate method for inferring haplotypes to date (Kukita et al., 2005). PHASE 2.1.1 is preferable to the newer, less computationally intensive, version fastPHASE, (Scheet and Stephens, 2006) because of the greater accuracy of the

algorithm, but also because it is possible to include tri-allelic sites under a model of parent-independent mutation (PIM), which relaxes the assumption of step-wise mutation used for biallelic SNPs. PHASE 2.11 also implements a new recombination method (the –MR option), which allows the user to specify the relative physical location of each SNP. This method is slightly more accurate than the previous method that does not account for the decay of LD with distance (Stephens and Donnelly, 2003; Stephens et al., 2001).

For each locus, four PHASE runs were performed on samples grouped according to broad geographic regions (*i.e*. Africa, Europe, Asia, and the Americas) (Table 2.1). PHASE runs for each geographic region were replicated using the –x option that runs the algorithm multiple times automatically, starting from different starting points and selecting the run with the best average "goodness of fit". The "goodness of fit" measures the estimated haplotypes fit to an approximate coalescent model. SNP haplotypes are shown for each locus in Tables 3.1(a) and (c), 4.1(a) and (c), and 5.1(a) and (c), including haplotype frequencies for each population group included in the present study (Tables 3.1(b) and (d), 4.1(b) and (d), and 5.1(b) and (d). SNP haplotypes for each individual were reinserted into the original reference sequence for all analyses requiring full sequence data (see Appendix 1 for process). Sequence was unmodified, with the exception of a single problem area (repetitive sequence at the 3' end of primer -1182; see Figure 2.1) in the *NAT1* region, where a 235 bp of repetitive sequence was removed 5' of the ORF of exon 9 from all individuals. Indels and microsatellite data were not considered in the present analyses.

Nucleotide Diversity

General diversity statistics, such as *S*, HD, π, and θ_ω, were calculated, using the program DNAsp version 4.20.2 (Rozas and Rozas; Rozas et al.), at the continental, and population group levels for the *NAT1* (Table 3.2), *NAT2* (Table 4.2), and *NATP1* (Table 5.2) loci. Each of the statistics and tests of neutrality calculated using DNAsp are described, in brief, below.

Number of Segregating Sites

S, or the number of segregating sites, summarizes the diversity within a set of sequences. Under an infinite-sites model of mutation, segregating sites are the total number of mutations (Hudson, 1990).

Average Nucleotide Diversity

π, or average nucleotide diversity, is a measure of the degree of polymorphism in a sample of sequences. This concept was first introduced by Nei and Li (1979), and is defined as the average number of pairwise nucleotide differences between sequences in a sample (Tajima, 1989). π depends on both the number of polymorphic sites and their frequency. Tajima (1983) and Nei (1989) scaled this measure to account for the length of sequence being considered.

Watterson's Theta Estimator

Watterson's theta (θ_ω) is a maximum likelihood estimate of θ. θ_ω depends only on the number of segregating sites and the sample size, making this estimator independent of the frequency of polymorphic sites (Watterson, 1975). $\theta_\omega = S/(1 + 1/2 + 1/3 + ... + 1/n - 1)$, where S is the number of segregating sites and n is the sample size.

Tests of Selective Neutrality

Tests of selective neutrality at both the intra- (Tajima's D, Fu and Li's D* and F*) and inter-specific (Fu and Li's D and F, and Fay and Wu's H) levels were performed using DNAsp version 4.20.2 (Rozas and Rozas, 1995; Rozas et al.) (Table 3.2, Table 4.2, Table 5.2). Statistics estimated at the inter-species level were estimated using the published sequence for chimpanzee, *P. troglodytes* (Ptr8-WGA990). Significance was assessed for all neutrality estimates using the coalescent simulator within DNAsp (10,000 replicates), assuming no recombination. We have used the Bonferroni correction for multiple tests in order to obtain an experimentwise error rate of α, where each individual test obtains a corrected critical probability of $\alpha' = \alpha/k$, where k is equal to the number of tests carried out for the entire dataset (Sokal and Rohlf, 1995). Because we carried out six independent tests of neutrality for each NAT loci, we obtain α' values of 0.008, 0.002, and 0.0002 at the α=0.05, 0.01, and 0.001 levels, respectively.

Under a strict Wright-Fisher model of neutral evolution (Fisher, 1930; Wright, 1931) these statistics should equal zero. Negative deviation from zero can be indicative of population size expansion, positive directional selection, and/or purifying (negative)

selection operating at the loci. Conversely, positive deviation from zero can be consistent with population size reduction and/or balancing selection at the loci.

Tajima's D

Tajima's D (TD) statistic compares two theta estimators: (a) θ_π, which is based on the average number of nucleotide differences between pairs of sequences (π), and (b) θ based on heterozygosity or the number of segregating sites (S).

The mathematical relationship between these estimates of diversity is given by:

$$D = \frac{(\theta_\pi - \theta_S)}{\sqrt{\text{Var}\,(\theta_\pi - \theta_S)}}$$

It is important to note that natural selection and demographic history can cause similar departures from neutral evolution. As a result, negative values of TD can be indicative of demographic processes, such as rapid population growth, whereas positive values of TD can be indicative of a population recovering from a bottleneck. Theoretically, demographic events are expected to affect all loci similarly, whereas natural selection is expected to affect particular loci. Therefore, in order to detect adaptive variation we must have some knowledge of base-line neutral variation, or simulate expectations under particular demographic scenarios.

Sliding Window Tajima's D

A sliding window approach for TD statistic was also implemented to assess differing values of the statistic across the *NAT1, NAT2*, and *NATP1* regions at the continental, regional, and population levels (Figures 3.4, 4.4, 5.1). This method makes it possible to visualize variation in the statistic across the region at defined window lengths (by 100 sites at steps of 25 sites), where each window describes a specific topology of the genealogy for that region.

Fu and Li's D, F*, D, and F*

Estimates of Fu and Li's D and F (1993) are based on the number of derived singletons compared to the total number of singletons. This relationship is based on a coalescent framework, and the observation that "older" mutations will occur on internal branches of the genealogy, while more recent mutations will tend to occur on the external branches of a genealogy. Under a scenario of negative or purifying selection an excess of mutations on the external branches is expected because deleterious mutations will be present at low frequencies. Additionally, an excess of external mutations at low frequencies would be expected following a selective sweep, where an advantageous allele rapidly increases in frequency, together with linked variants (*i.e.* "genetic hitchhiking"). Alternatively, under a scenario of balancing selection, alleles in a given population are expected to be old and a disparity of mutation in the external branches would be observed (Fu, 1993).

The F and D statistics can be calculated at the inter- or intra-specific level (where the test statistics at the intra-specific level are designated with an asterisk, D* and F*).

Because these tests compare "mutations in the recent past with those in the relatively remote past" (Fu and Li, 1993), the use of an outgroup, if available, may be preferable. It has been noted by Fu and Li (1993) that when no outgroup is available the number of singletons may over-estimate the number of mutations in the external branches of the geneology. These authors correct for this by constructing a new statistic to represent singletons, using the method of moments. Futhermore, intra-specific estimates are included here and are useful as analogous measure to TD, which does not use an outgroup.

Fay and Wu's H

Similar to Tajima's D, the H test statistic is based on the differences between two estimators of theta: (a) θ_π , which is based on the average number of nucleotide differences between pairs of sequences (π) and (b) θ_H, an estimator based on the frequency of the derived variants (Fay and Wu, 2000). Derived, non-ancestral alleles are typically lower in frequency than ancestral alleles (Sabeti et al., 2006). As mentioned above, in a selective sweep a derived functional variant under selection may increase to intermediate or high frequency, together with linked variation, depending on the level of recombination (Fay and Wu, 2000). Fay and Wu's H statistic detects hitchhiking by measuring the excess of high compared to intermediate frequency, derived variants (Fay and Wu, 2000). In the absence of recombination, the linked variant becomes fixed; whereas if recombination is rare, all variants are present at either low or high frequencies. "A single, strong hitchhiking event will push all variation to either very low or very high frequencies, regardless of the frequency spectrum before selection" (Fay and Wu, 2000).

By focusing on specifically high frequency variants, H is less likely to be influenced by the confounding effects of background selection (Fay and Wu, 2000). However, population subdivision may be a problem for derived-allele tests like Fay and Wu's H (Sabeti et al., 2006). Background selection can be a major confounder for rare-allele tests (*e.g.* TD, D, D*, F, and F*), in addition to demographic effects such as population expansion (Sabeti et al., 2006).

Hudson, Kreitman, and Aguade (HKA) and McDonald-Kreitman (MK) tests

The Hudson, Kreitman, and Aguade (HKA) test is based on the neutral theory of molecular evolution (Kimura, 1983), and postulates that regions of the genome that evolve at high rates at the intraspecific level will also evolve at high rates at the interspecific level. The HKA test constructs a goodness of fit measure, implementing a chi-square test of the observed and expected number of differences between species for two loci (Figures 3.6 And 4.6).

"The McDonald-Kreitman test (McDonald and Kreitman, 1991) explores the fact that mutations in coding regions come in two flavors, synonymous and non-synonymous mutations" (Nielsen, 2005). For example, Figures 3.5 and 4.5 summarize counts of both silent and replacement changes within and between two species (human and *P.troglodytes*), for the *NAT1* and *NAT2* loci, with the idea that negative selection will reduce replacement substitutions and positive selection will permit replacement changes to occur in excess, relative to the number of synonymous changes.

Population Differentiation

Pairwise F_{ST} values for each *NAT* locus, by population group, were generated using Arlequin *version* 2.0 for (Excoffier, 2005). Wright's F_{ST}, sometimes referred to as the *fixation index*, measures the reduction in heterozygosity as a result of random mating, at one hierarchical level relative to another (Hartl, 1997). Wright's F_{ST} is typically used to measure the extent of population divergence among populations of the same species, relative to the net genetic diversity within the species (Charlesworth, 1998). Values between human groups, based on continental groupings (Africa, Europe and Asia) are of the range of ~0.10-0.15 for autosomal loci (reviewed in (Campbell, 2008; Tishkoff and Verrelli, 2003), depending on the particular locus. Wright suggested the following guidelines in interpretation of F_{ST}: *little* differentiation: 0-0.05; *moderate* differentiation: 0.05-0.15; *great* differentiation: 0.15-0.25; *very* great differentiation: >0.25 (Hartl, 1997).

Pairwise F_{ST} data matrices were used to generate 2-D multi-dimensional scaling (MDS) plots for the *NAT1* (Figure 3.7), *NAT2* (Figure 4.7) and *NATP1* (Figure 5.2) haplotype data, using Statistica version 8 (StatSoft, Inc., 1984-2008). MDS is a type of ordination, or clustering method, used to visually explore patterns in the data. Non-metric MDS begins with a similarity or dissimilarity matrix.

Analyses of Molecular Variance (AMOVA) calculations were performed using Arlequin *version* 2.0 (Excoffier, 2005). The STRUCTURE results of Reed and Tishkoff (unpublished) are used here to infer population groups used in the AMOVA analysis, according to the population structure listed in Table 2.3. The AMOVA, performed with a single locus, is used to determine the level of within and between population variation. Results for the AMOVA for *NAT1*, *NAT2*, and *NATP1* are shown as insets to the MDS plots (Figures 3.6, 4.6 and 5.2).

Table 2.3 Population Structure used for AMOVA	
AFRICA	
Pop 1	Dinka
Pop 2	San *
Pop 3	Biaka Pygmy *
	Baka Pygmy
	Bakola Pygmy
Pop 4	Hadza
Pop 5	Fulani
Pop 6	Burunge
	Sandawe
	Turu
Pop 7	Maasai
Pop 8	Kanuri
	Mada
Pop 9	Lemande
	Yoruba *
Total Africa	**9 populations**
EUROPE/ Middle East (ME)	
Pop 10	French *
	Druze*
	Sardinian*
	Russian *
Pop 11	Brahui*
Total Europe/ ME	**2 populations**
ASIA	
Pop 12	Papuan*
Pop 13	Cambodian *
	Han *
	Japanese *
	Yakut*
Total Asia	**2 populations**
AMERICAS	
Population 14	Karitiana*
	Pima*
Total Americas	**1 population**
GRAND TOTAL	***14 Populations***
Population groups were created with reference to STRUCTURE defined groups (Reed and Tishkoff., unpublished).	

Table 2.3

Table 2.3: Population structure used for the purpose of the AMOVA analyses of *NAT1, NAT2*, and *NATP1*. All populations represented, and corresponding 2N values, are identical to that presented in Table 2.1.

Network and Phylogenetic Analysis

Network version 4.5 was used to construct median-joining (MJ) phylogenetic networks for the *NAT1, NAT2*, and *NATP1* regions (Bandelt, 1999) (fluxus-engineering.com), shown in Figures 3.7, 4.7, and 5.3. Phylogenetic networks are preferable to simple, bifurcating trees for intra-species comparison in that differing evolutionary pathways are represented in terms of cycles or hyper-cubes. Ambiguity is represented by reticulation in the network, which can indicate the presence of homoplasy or recombination in the dataset. MJ networks illustrate how many mutational steps are required to link all haplotypes to each other, as well as indicating the frequency of observed haplotypes in the dataset. It must be noted that these methods are not intended as stand alone methods, but are useful in generating hypotheses, and may help visualize the data from an evolutionary perspective.

For analyses of the *NATP1* region (Chapter V), ClustalX version 1.83.1 (Chenna et al., 2003) was used to align all *NAT* loci with *P. troglodytes* (chimpanzee) and *Mus musculus* (mouse) (Figure 5.5). The aligned sequences were then used in the phylogenetic analyses, presented in Figure 5.6, using PAUP* 4.0 (Rogers and Swofford, 1998). In order to compute the relationship among sequences we performed a neighbor-joining analysis using Kimura two-parameter (K2P) distance, which corrects for difference in rates of transitions and transversions (Kimura, 1980).

Linkage Disequilibrium and Recombination

Linkage Disequilibrium (LD) refers to the correlation, or more specifically the non-random gametic association, of variation at two or more loci or sites within a given

population group (Mueller, 2004). The traditional measures of LD, D' and r^2, have different properties, are directly related to recombination rate, and can be applied for specific purposes. D' measures the absolute value of the deviation of observed frequencies from that of expected, for two loci. The r^2 statistic is equal to the square of the correlation coefficient between the alleles *A* at locus *x* and *B* at locus *y* (McVean et al., 2002). For evolutionary studies, knowledge of the extent of recombination becomes critical because recombination can perturb the genealogical relationship amongst haplotypes (Mueller, 2004). Model-based LD measures, such as those implemented in the LDhat (McVean et al., 2002) software package, estimate the approximate likelihood of observing LD patterns in the data under a range of recombination rates and within a coalescent framework (Hudson, 2001).

LDhat was used to generate pair-wise estimates of recombination for phased haplotype data, at each of the *NAT1, NAT2*, and *NATP1* regions, using the Composite Likelihood method of Hudson (2001). Population scaled recombination rate $\rho=4N_e r$ for diploid species was inferred using this method, where N_e is the effective population size and r is the genetic map distance across the region. The model assumes uniformity of mutation rate across the region analyzed. The diagonal matrices presented in Figures 3.8, 4.8, and 5.4 illustrate the reported value for each pair of SNPs, which is the difference between the log-composite likelihood assuming constant recombination rate over the entire region and the marginal maximum for each pair. The matrix can then be used to visualize the discordance between constant rate patterns in the data (McVean et al., 2002) Tri-allelic sites in the dataset were excluded for this portion of the analyses.

A secondary measure was taken to confirm estimates of recombination and linkage, for *NAT1* and *NAT2* using the computer program Haploview version 4.1 (Barrett et al., 2005). Haploview uses a standard expectation-maximization (EM) algorithm to "estimate the maximum-likelihood values of the four gamete frequencies" (Barrett, 2005 pp. 264) for each site, from which pairwise estimates of D' are calculated. SNPs with minor allele frequencies less that 1% were excluded for analyses of each continental group (Figures 3.9 And 4.9).

NAT2 Inferred Acetylator Status

Individual NAT2 acetylator phenotypes were predicted from phased, diploid haplotypes (Table 4.3, Figures 4.2 and 4.3) in accordance with the accepted NAT nomenclature classification of *NAT2** alleles based on functional characterization. Only SNPs with known affect on phenotype (*i.e.* in the 870bp coding region of *NAT2*) were used to infer acetylator status of individual haplotypes. Individuals with two slow activity haplotypes were classified as slow acetylators, those with two rapid acetylator types were classified as rapid acetylators, and individuals with both slow and rapid haplotypes were classified as intermediate acetylators. Haplotypes with unknown phenotypes are indicated in all cases. Further detail on acetylator inference is presented in Chapter IV.

CHAPTER III. EVOLUTIONARY CHARACTERIZATION OF THE HUMAN *NAT1* LOCUS

RESULTS

Haplotype Reconstruction and Patterns of Genetic Diversity

We sequenced 2856 bp of the *NAT1* region in 652 chromosomes, and identified 48 SNPs (Table 3.1). Seventeen *NAT1* SNPs identified are novel to the present dataset and have not been previously reported. These are located at positions -1044, -1037, -1144, -970, -920, -426, +236, +662, +758, +956, +1245, +1527, +1572, +1654, +1685, +1784, and +1969. Eight variable sites were identified in the 870 bp exon region, two of which were non-synonymous polymorphisms. The replacement changes +445G>A, +639G>T, result in amino acid substitutions Ile149Val and Ala214Ser, respectively. Both of these replacement changes have been previously reported (Genbank accession numbers rs4987076 and rs4986783). Replacement SNPs +445G and +639T are observed to be fixed in several population groups (Figure 3.1). In addition, SNP +1088T/A, following a $(TAA)_n$ repeat, is variable in humans (Figure 3.2), compared to all other NHP analyzed in this study (Figure 3.3). Variant+1088A is observed in human populations at frequencies of 13-83%. Additionally, 8 singleton mutations and two tri-allelic sites were identified in this dataset.

Summary statistics and estimates of neutrality for the *NAT1* locus are given in Table 3.2. Overall, African populations are more diverse at the *NAT1* loci than non-African populations. In particular, Pygmy groups show the most diversity when these

populations are grouped together, while Asian and Amerindian groups exhibit the least amount of diversity. Genetic diversity within individual groups shows the Fulani of Cameroon to have the highest level of nucleotide diversity (π=1.630), followed by the foraging groups such as the Hadza (π=1.490), San (π=1.450) and Baka (π=1.250) and Bakola Pygmy (π=1.250) groups.

Statistical Tests of Neutrality

Several tests of neutrality were performed for the *NAT1* locus, at both the intra and inter-species levels (Table 3.2). Overall, estimates of neutrality based on the allelic frequency distribution (TD, D*, F*, D, T) for *NAT1* exhibited consistency across estimates, with negative values in most cases. Negative values indicate a skew in the frequency spectrum of polymorphism toward rare variants, signifying that purifying or positive directional selection may have affected *NAT1*. However, we do not observe statistically significant values after Bonferroni correction, with the exception of D*,F*, and H in Asia (significant at $p<0.01$) and H values for Africa and West Africa at the $p<0.05$ level. Positive values for these estimators may be due to small sample size, and in some cases population demographic explanations may apply, especially where the group in question is believed to have experienced a recovery from a recent demographic bottleneck (*e.g.* Hadza) ((Blurton Jones et al., 1992)). The H statistic (Fay and Wu, 2000) is highly negative and significant for most population groups. H measures high frequency, derived variants and is thought to be a good measure of positive directional selection associated with a hitchhiking event or selective sweep (Sabeti et al., 2006).

Sliding window analyses of TD indicate variation for this estimator across the *NAT1* region (Figure 3.4) in all continental and population groups. Overall, TD is negative across the *NAT1* region in all populations, with the exception of a highly positive deviation from zero in the 3' UTR region of the gene corresponding to the location of SNPs at positions +1088, +1095, and +1191. Nearly every African and Asian group obtains significant TD values ($p<0.05$) at this location (Figure 3.4). In Europe, only the Brahui from Pakistan have a significant TD at this 3' location, though all have positive values. African and Asian groups that do not exhibit positive values at this position include the New Guinea population, as well as the Turu and Sandawe from Tanzania. The majority of *NAT1* coding region SNPs (Table 3.1) in human populations are present at low frequencies. Replacement SNPs +445G and 639T are present at identical frequencies that are approaching or have reached fixation in most groups. The SNPs at positions +1088, +1095, and +1191 located in the 3'UTR region, however, are present at intermediate frequency in all populations, often clustering together at frequencies of ~50%.

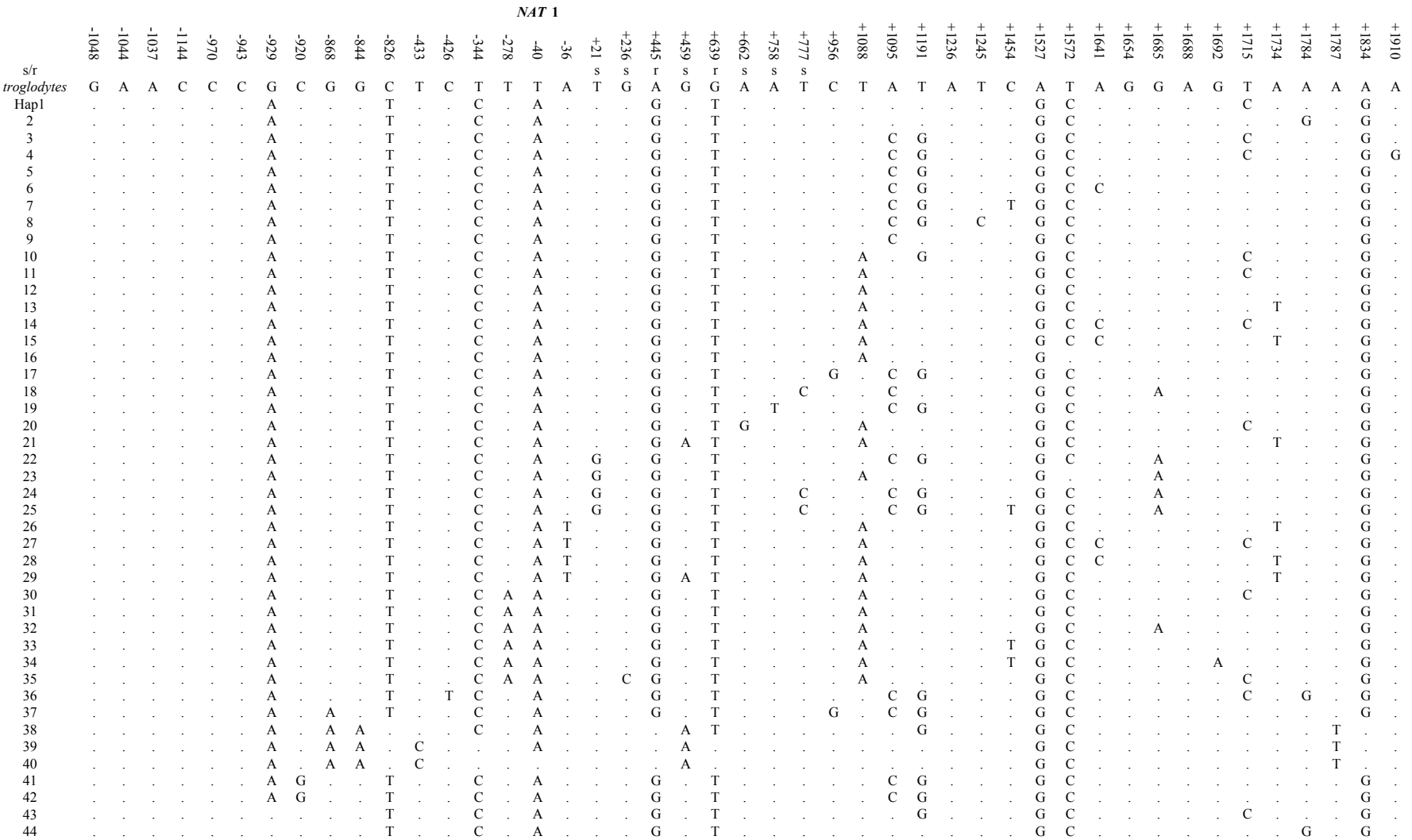

NAT 1
s/r
P. troglodytes
Hap1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
-1048 -1044 -1037 -1144 -970 -943 -929 -920 -868 -844 -826 -433 -426 -344 -278 -40 -36 +21 +236 +445 +459 +639 +662 +758 +777 +956 +1088 +1095 +1191 +1236 +1245 +1454 +1527 +1572 +1641 +1654 +1685 +1688 +1692 +1715 +1734 +1784 +1787 +1834 +1910
s r s s r s s s s r s

Table 3.1a

***NAT1* Haplotypes**

	BG	BK	DN	FU	HZ	KN	LM	MD	MS	PB	PL	SA	SW	TR	YO	Afr	DZ	FR	PA	RU	SR	CB	HA	JP	NG	SB	BR	PI	Non	ALL
Hap1	2	.	.	.	.	.	.	2	1	.	.	.	.	.	1	6	.	.	.	.	.	.	.	.	.	.	.	.	.	6
2	.	.	.	.	.	1	.	.	.	.	.	.	.	1	2	4	.	.	.	.	.	.	.	.	.	.	.	.	.	4
3	2	.	.	1	.	.	1	.	.	.	.	.	2	3	1	10	.	.	.	.	.	.	.	.	.	.	.	.	.	10
4	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
5	.	.	.	.	1	.	.	.	.	.	.	.	1	1	2	5	.	.	.	.	.	.	.	.	.	.	.	.	.	5
6	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
7	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
8	.	.	.	.	2	.	.	.	.	.	.	.	4	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	6
9	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
10	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
11	4	5	6	4	2	5	3	6	5	3	2	2	.	3	3	53	.	.	.	.	.	.	1	.	3	.	.	.	.	57
12	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
13	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
14	1	.	.	.	2	.	.	.	1	.	2	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	6
15	.	.	.	.	1	.	.	1	.	.	.	.	.	.	.	2	7	.	.	.	.	.	2	1	.	.	.	.	10	12
16	.	1	.	.	.	1	.	.	1	1	.	2	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	6
17	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
18	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
19	1	.	.	.	.	.	.	.	.	.	.	1	2	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	4
20	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
21	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
22	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
23	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
24	.	.	.	.	3	.	.	.	.	.	.	.	1	1	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	5
25	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
26	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
27	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
28	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
29	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
30	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
31	.	1	.	.	1	2	3	3	2	3	.	.	2	1	.	18	.	.	.	.	.	.	.	.	.	.	.	.	.	18
32	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
33	.	.	.	1	1	.	.	.	.	.	.	1	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
34	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	2	.	.	.	.	1	.	.	.	.	.	.	.	1	3
35	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
36	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
37	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
38	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
39	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
40	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	1	.	4	.	.	.	5	7
41	.	.	.	.	1	.	.	.	.	.	1	.	1	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
42	.	1	.	.	.	.	.	.	.	.	.	1	.	1	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
43	.	.	.	1	.	.	1	.	2	.	.	.	1	.	.	5	.	.	.	1	.	.	.	.	.	.	.	.	1	6
44	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1

Table 3.1b

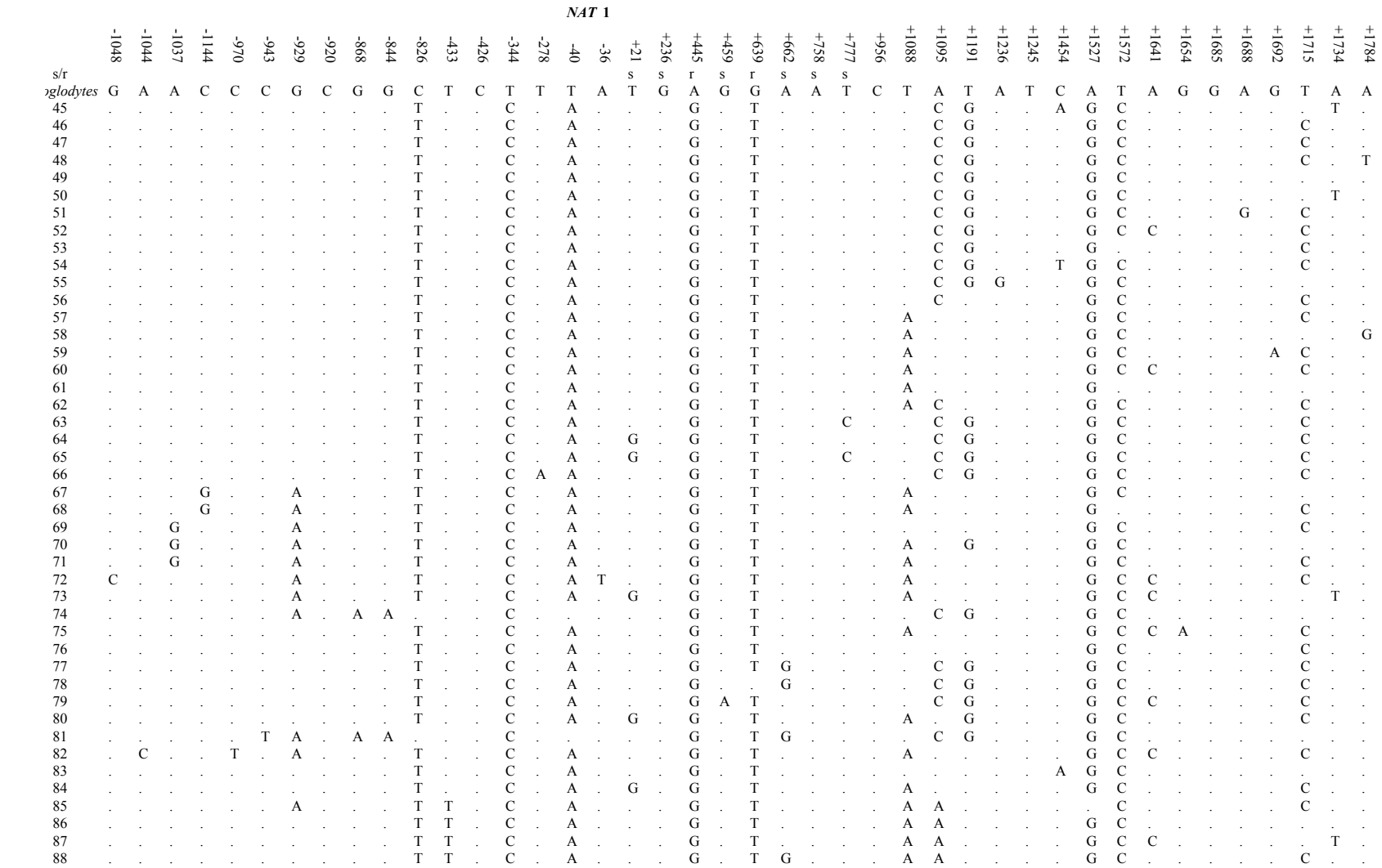

NAT 1
s/r
-1048 -1044 -1037 -1144 -970 -943 -929 -920 -868 -844 -826 -433 -426 -344 -278 -40 -36 +21 +236 +445 +459 +639 +662 +758 +777 +956 +1088 +1095 +1191 +1236 +1245 +1454 +1527 +1572 +1641 +1654 +1685 +1688 +1692 +1715 +1734 +1784
45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88

Table 3.1c

***NAT1* Haplotypes**

	BG	BK	DN	FU	HZ	KN	LM	MD	MS	PB	PL	SA	SW	TR	YO	Afr	DZ	FR	PA	RU	SR	CB	HA	JP	NG	SB	BR	PI	Non	ALL
	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
Hap45	8	13	6	11	5	11	11	8	5	2	2	3	16	17	6	124	16	12	11	17	17	1	11	13	2	13	9	12	134	258
46	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
47	3	.	.	.	.	.	.	.	.	.	.	.	.	1	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	4
48	.	.	.	.	.	.	2	1	.	1	1	1	1	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	7
49	.	.	.	.	1	.	1	.	1	.	1	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	1	1	5
50	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
51	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	6	1	.	.	.	.	.	.	.	.	7	8
52	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
53	1	.	.	1	.	1	.	2	1	.	.	.	.	1	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	7
54	.	1	.	.	.	.	.	.	.	2	1	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	4
55	.	.	.	.	1	2	.	.	5	.	.	.	.	.	.	8	.	1	.	.	.	.	2	.	.	.	.	.	3	11
56	9	.	1	1	4	2	2	1	2	1	2	.	2	.	4	31	1	3	3	3	4	4	7	7	11	8	12	4	67	98
57	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
58	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
59	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	2	.	.	.	.	.	.	.	.	.	.	.	3	3	5
60	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
61	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
62	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
63	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	1	2
64	.	.	1	.	.	.	.	.	1	.	.	.	1	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
65	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
66	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
67	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
68	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
69	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
70	.	2	.	.	.	.	.	.	.	1	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
71	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
72	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	1
73	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	1
74	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	1
75	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	3	.	.	.	1	.	.	.	.	.	5	5
76	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	7	1	.	.	.	.	.	.	.	.	.	.	8	8
77	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	1
78	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	1
79	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	1
80	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	1
81	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	1
82	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	1	.	.	3	3
83	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	1
84	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	1
85	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1
86	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	3	3
87	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	1
88	34	30	18	22	32	26	26	28	32	18	14	14	38	32	24	388	31	22	24	24	24	7	25	22	21	22	22	24	264	656

Table 3.1d

Table 3.1 (a-d are quadrants of Table 3.1): Eighty-eight *NAT1* phased haplotypes were identified, and listed in (a) and (c). The number of observed haplotypes in each population are shown in parts (b) and (d). All population abbreviations follow that presented in Table 2.1, in addition to the abbreviation 'Afr' indicating all African populations and 'Non', which denotes all non-African groups. Derived variants, as compared to *P. troglodytes*, are indicated. *s* and *r* refer to silent and replacement polymorphisms.

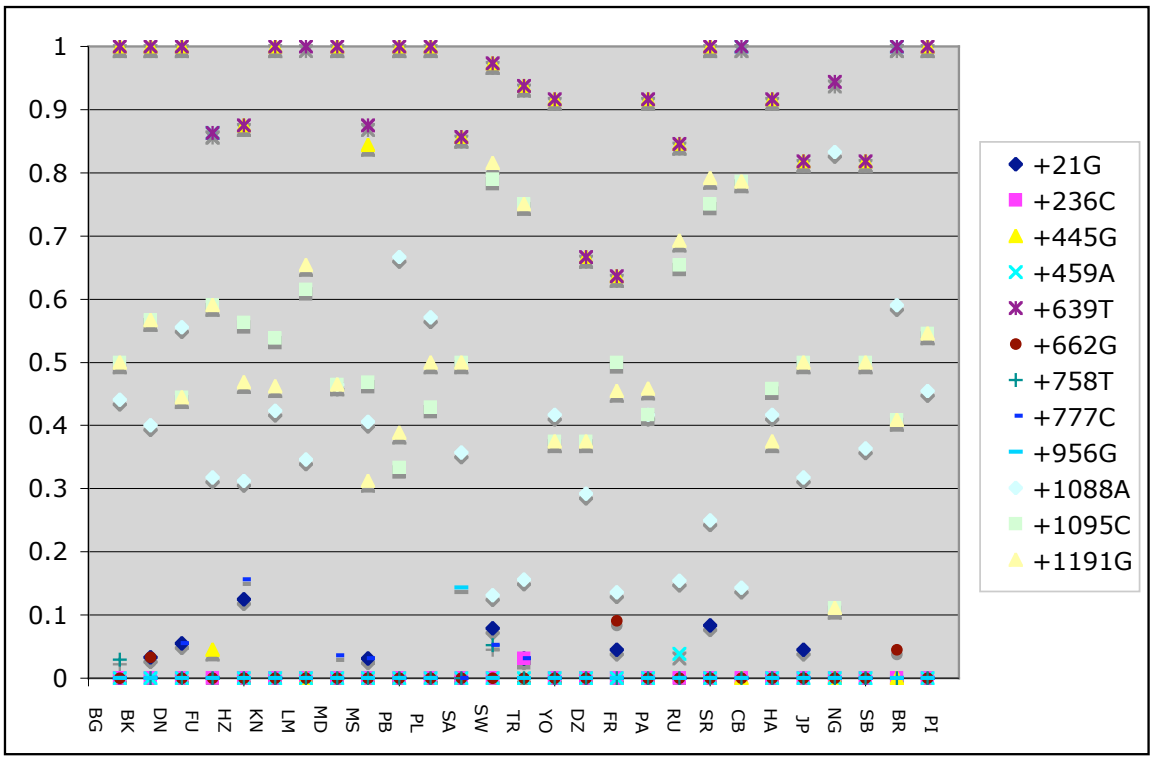


Figure 3.1

Figure 3.1: Frequency diagram of 9 *NAT1* coding region SNPs, as well as 3' SNPs at positions +1088, +1095, and +1191. Frequency of each SNP is indicated on the y-axis, and each population in the present analyses is listed on the x-axis (refer to Table 2.1 for population abbreviations).

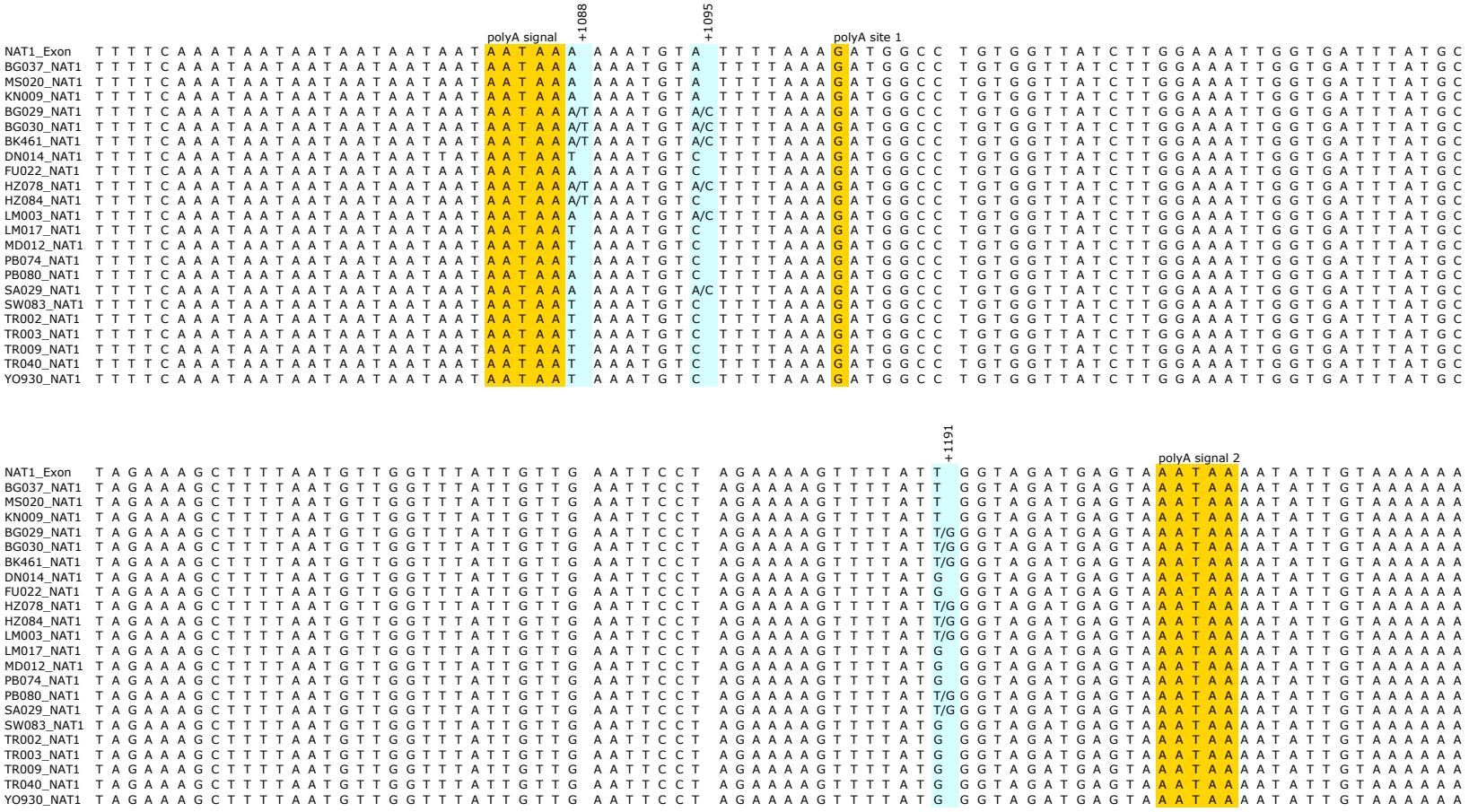


Figure 3.2

Figure 3.2: Human *NAT1* sequence alignment for a sample of individuals, indicating the location of the first and second poly-A signal and site, in relation to the three 3' mutations of interest, +1088, +1095, and +1191.

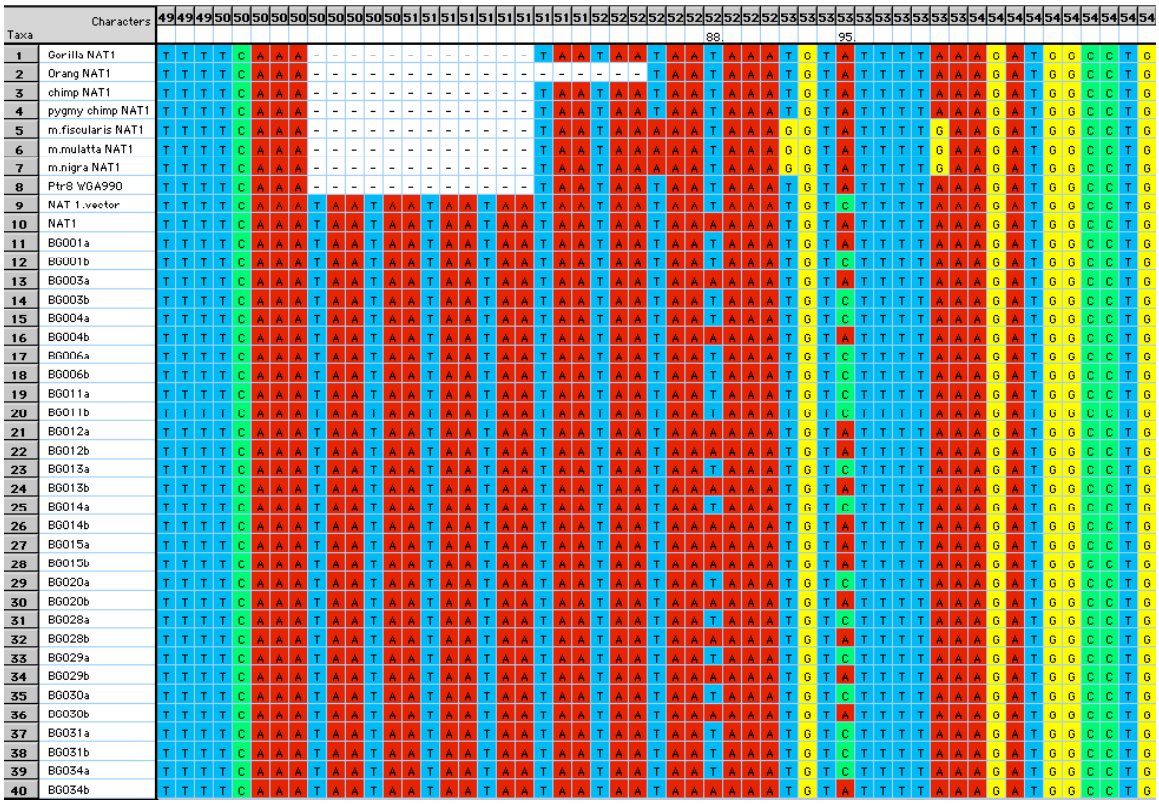


Figure 3.3

Figure 3.3: Partial alignment of the *NAT1* region containing the first poly-A signal (signal 1) in humans. The alignment illustrates variation in the $(TAA)_n$ repeat in non-human primates (NHPs), which is closely associated with the +1088 A/T SNP and +1095 A/C SNP (indicated as 88 and 95) in humans. Non-human primates included are *G. gorilla* (gorilla), *P. pygmaeus* (Sumatran orangutan), *P. paniscus* (bonobo), *P.troglodytes* (chimp), Ptr8 WGA990 (published chimp), *M. mulatta* (rhesus monkey), *M. fascicularis* (crab-eating macaque), *M. nigra* (celebes ape). The first poly-A signal is intimately associated with the +1088 variant, where an A at +1088 is known to abolish the signal in humans (Boukouvala and Fakis, 2005). SNP +1088 is not variable in NHPs, who all have a T at +1088.

	N	2N	S	s	r	H	HD	*Si*	π (∗10³)	θω(∗10³)	TD	P(TD<obs)	D*	P(D*<obs)	F*	P(F*<obs)	D	P(D<obs)	F	P(F<obs)	H	P(H<obs)
NAT1																						
Africa	194	388	43	6	2	72	0.868	4	1.220	0.002	-1.41735	0.0394	0.35031	0.6557	-0.54941	0.317	0.30558	0.646	-0.67690	0.272	-15.33997	0.007*
E Africa	93	186	36	5	2	47	0.881	11	1.200	0.002	-1.44153	0.041	-1.35749	0.085	-1.67604	0.0576	-0.83531	0.2164	-1.40566	0.100	-13.76484	0.008
W Africa	79	158	31	2	2	38	0.857	3	1.210	0.002	-1.07911	0.133	0.92079	0.801	0.12378	0.5934	0.90291	0.7875	-0.01065	0.53310	-15.07151	0.004*
Pygmy groups	31	62	21	3	0	24	0.893	7	1.280	0.002	-0.5624	0.339	-0.93306	0.138	-0.95246	0.1861	-0.74290	0.2144	-0.91957	0.2068	-3.10100	0.1093
Europe	59	118	26	3	1	21	0.612	9	0.950	0.002	-1.30037	0.073	-1.57412	0.056	-1.75807	0.0552	-0.96702	0.1668	-1.41564	0.1013	-11.11285	0.0083
Asia	51	102	23	2	2	11	0.65	15	0.780	0.001	-1.37411	0.062	-4.98532	0.0007*	-4.29272	0.0007*	-1.94898	0.0609	-2.17613	0.02910	-15.05727	0.0005*
Americas	23	46	7	1	0	7	0.675	1	0.760	0.001	0.96846	0.862	0.48112	0.8158	0.74551	0.7696	0.32361	0.5108	0.46854	0.6687	-0.95459	0.1461
Africa																						
Baka Pygmy	9	18	15	1	0	12	0.948	7	1.250	0.002	-0.2855	0.4302	-0.78067	0.177	-0.73978	0.2424	-0.38290	0.3524	-0.49437	0.3496	-2.16993	0.255
Bakola Pygmy	7	14	10	1	0	10	0.956	3	1.250	0.001	0.54222	0.7519	0.16224	0.6371	0.30185	0.608	-0.09199	0.4442	0.02549	0.5079	0.57143	0.3794
Biaka Pygmy	15	30	17	2	0	13	.795	9	1.190	0.002	-0.70869	0.2644	-1.62499	0.0903	-1.56811	0.0809	-2.10313	0.0328	-2.06871	0.0457	-2.05057	0.2591
Burunge	17	34	10	1	0	12	0.866	3	0.900	0.001	0.16842	0.633	-0.26765	0.2755	-0.15375	0.4357	0.13128	0.4869	0.03760	0.5334	-3.00891	0.1965
Dinka	9	18	12	2	0	8	0.804	4	1.180	0.001	-0.12308	0.5007	-0.12092	0.3525	-0.14046	0.4383	-0.39669	0.3368	-0.47701	0.3566	-2.06536	0.2337
Fulani	11	22	20	1	2	8	0.732	1	1.630	0.002	-0.56415	0.3196	1.31555	0.9855	0.8765	0.8448	1.44166	0.948	0.90518	0.8228	-11.42857	0.1162
Hadzabe	16	32	15	2	0	20	0.956	3	1.490	0.001	0.24574	0.6572	0.421	0.7126	0.42921	0.6728	0.32719	0.715	0.28137	0.6233	-0.52016	0.2998
Kanuri	13	26	10	0	0	9	0.791	3	1.030	0.001	0.40851	0.7178	-0.14176	0.3191	0.02762	0.4965	0.22209	0.5206	0.22368	0.58930	-4.00000	0.16470
Lemande	13	26	9	0	0	10	0.806	1	1.030	0.001	0.78181	0.8216	0.81213	0.7118	0.93585	0.8394	0.73587	0.6913	0.80718	0.785	-0.51692	0.2910
Maasai	16	32	21	3	1	17	0.933	13	1.060	0.002	-1.11609	0.1286	-2.33563	0.0366	-2.28657	0.0318	-1.46880	0.0917	-1.72305	0.0809	-8.91935	0.16170
Mada	14	28	12	1	0	12	0.873	4	1.130	0.001	0.15733	0.6221	-0.36821	0.2612	-0.24191	0.4089	-0.64204	0.2493	-0.57181	0.32410	-1.83069	0.23960
S African San	7	14	14	1	0	10	0.945	7	1.450	0.002	-0.24881	0.4417	-0.7116	0.1944	-0.6722	0.261	-0.78110	0.2478	-0.83460	0.26470	-2.63736	0.2178
Sandawe	19	38	26	4	2	16	0.815	16	1.140	0.002	-1.62799	0.0311	-2.63662	0.0204	-2.71819	0.0185	-1.62062	0.0775	-1.98874	0.0497	-10.78236	0.15370
Turu	16	32	15	3	0	12	0.714	9	0.870	0.001	-1.26692	0.095	-1.87308	0.0518	-1.97461	0.0443	-1.93044	0.029	-2.09171	0.041	-4.30242	0.19150
Yoruba	12	24	8	0	0	12	0.906	1	1.040	0.001	1.25195	0.9153	0.73702	0.8968	1.03237	0.8642	0.63214	0.6576	0.87531	0.79850	0.31159	0.34850
Europe																						
French	11	22	19	2	1	9	0.701	5	1.310	0.002	-1.03829	0.1611	0.13329	0.6014	-0.25362	0.3937	0.33499	0.5965	-0.18080	0.4592	-8.00000	0.15150
Druze	12	24	7	0	0	3	0.489	0	1.090	0.001	2.0699	0.9826	1.2873	0.861	1.75988	0.9855	1.28513	0.8598	1.76479	0.9779	-0.78261	0.26650
Sardinian	12	24	9	1	0	5	0.486	5	0.600	0.001	-0.96176	0.1842	-1.35698	0.1516	-1.4428	0.1125	-0.56674	0.2676	-0.92260	0.22210	-5.02899	0.1294
Brahui	12	24	8	0	0	8	0.797	1	1.010	0.001	1.12172	0.891	0.73702	0.8885	0.00075	0.4857	0.63214	0.6501	0.82970	0.79310	-0.44928	0.2999
Russian	12	24	10	1	0	6	0.496	5	0.640	0.001	-1.0748	0.142	-1.11112	0.1055	-1.28093	0.1238	-0.96891	0.1724	-1.27122	0.16020	-2.97101	0.2999
Asia																						
Cambodian	7	14	8	0	0	8	0.658	0	0.540	0.001	0.23911	0.6579	0.63143	0.5975	0.59556	0.7231	1.22480	0.8609	1.07118	0.8642	-1.62308	0.2999
Han	12	24	7	0	0	6	0.717	0	0.790	0.001	0.62606	0.7721	1.2873	0.8591	1.2708	0.9171	1.28513	0.8561	1.15965	0.8762	-1.12319	0.2442
Japanese	11	22	8	1	0	4	0.571	5	0.690	0.001	-0.32991	0.4136	-1.57145	0.05	-1.40583	0.1209	-1.47902	0.091	-1.50160	0.0109	-1.92208	0.20920
Papuan	9	18	19	1	2	5	0.614	15	0.980	0.002	-1.91147	0.0114	-2.44478	0.0241	-2.65381	0.0169	-0.59660	0.288	-1.26041	0.16370	-12.86275	0.0999
Yakut	11	22	4	0	0	3	0.541	1	0.540	0.000	1.147	0.8851	0.14251	0.723	0.48993	0.6889	-0.21459	0.34	0.06212	0.50170	-0.64069	0.2447
Americas																						
Karitiana	11	22	4	1	0	3	0.558	1	0.560	0.000	1.31319	0.9108	0.14251	0.723	0.54376	0.7078	-0.21459	0.3412	0.10649	0.354	0.39827	0.41020
Pima	12	24	6	0	0	6	0.717	0	0.900	0.001	1.83144	0.9714	1.23249	0.8209	1.63065	0.9748	1.20330	0.8236	1.54674	0.5964	-0.65942	0.26870

All values were calculated using DNAsp (Rosas, 1995)

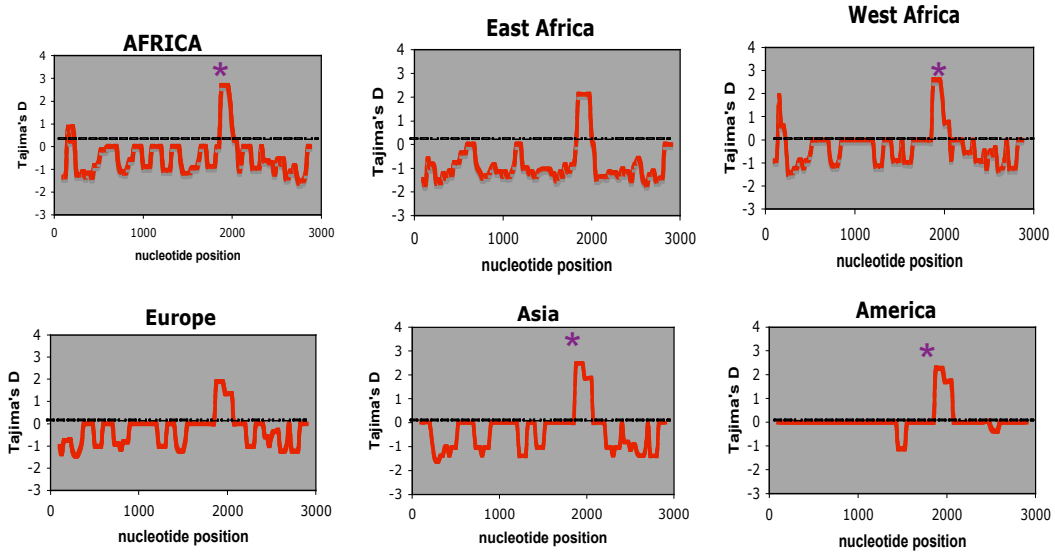
p<0.05 α' 0.008 p<0.01 α' 0.002 p<.001 α' 0.0002

Table 3.2

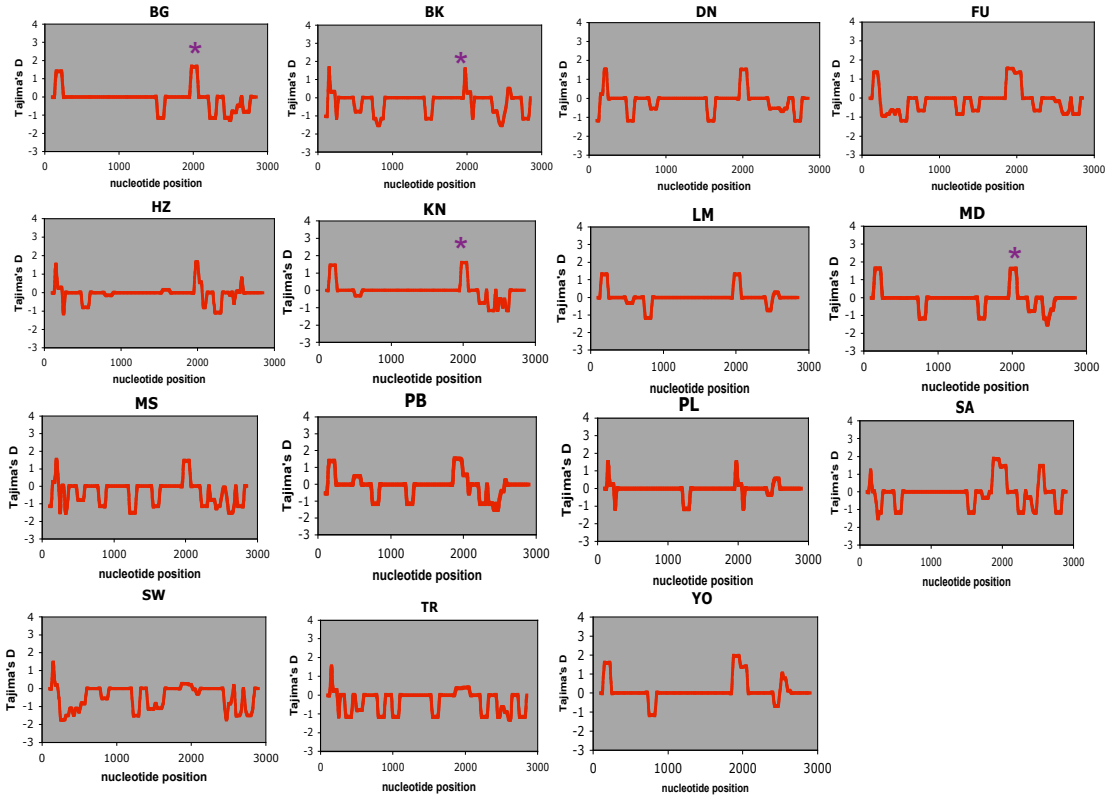
Table 3.2: Summary statistics and estimates of neutrality for the *NAT1* locus. All calculations were preformed using DNAsp version 4.20.2 (Rozas and Rozas; Rozas et al.). All abbreviations are as follows: S= number of segregating sites; 2N=number of chromosomes included for analysis; *s*=silent or synonymous variants identified; *r*=replacement or non-synonymous variants identified; H= number of haplotypes identified; HD= haplotype diversity; *Si*=number of singleton mutations; π=nucleotide diversity; θω=Waterson's theta estimator; TD=Tajima's D statistic (Tajima, 1983); D, D* and F, F* equal Fu and Li's test statistics at the inter-specific and intra-specific levels (Fu and Li, 1993), respectively; H=Fay and Wu's H statistic (Fay and Wu, 2000). P values, indicated by P (estimator<obs), indicate significance levels assessed for all neutrality estimates using the coalescent simulator within DNAsp (10,000 replicates), assuming no recombination. Following Bonferroni correction for multiple testing, we do not observe significance for any estimator, with the exception of those values specified with an asterisk.

The K_a/K_s ratio for *NAT1* of synonymous to nonsynonymous nucleotide divergence was observed to be less than one (K_a/K_s=0.242), consistent with the effects of purifying selection acting at this locus. The McDonald-Kreitman (MK) test (McDonald and Kreitman, 1991) of the observed level of intra-specific variation and levels of fixed variation between humans and chimpanzee within the coding region were not statistically significant after applying Fisher's exact test (p=0.659) (Figure 3.5). These results indicate that there is no excess or depletion of variation at the *NAT1* locus (at both coding and non-coding regions) within humans relative to chimpanzee. HKA tests comparing levels of intra- and inter-specific variation at *NAT1* to both *NATP1* and *NAT2* were not significant (Figure 3.6), with p values of 0.248 and 0.251, respectively.

NAT1 Continental Groups



NAT1 Africa



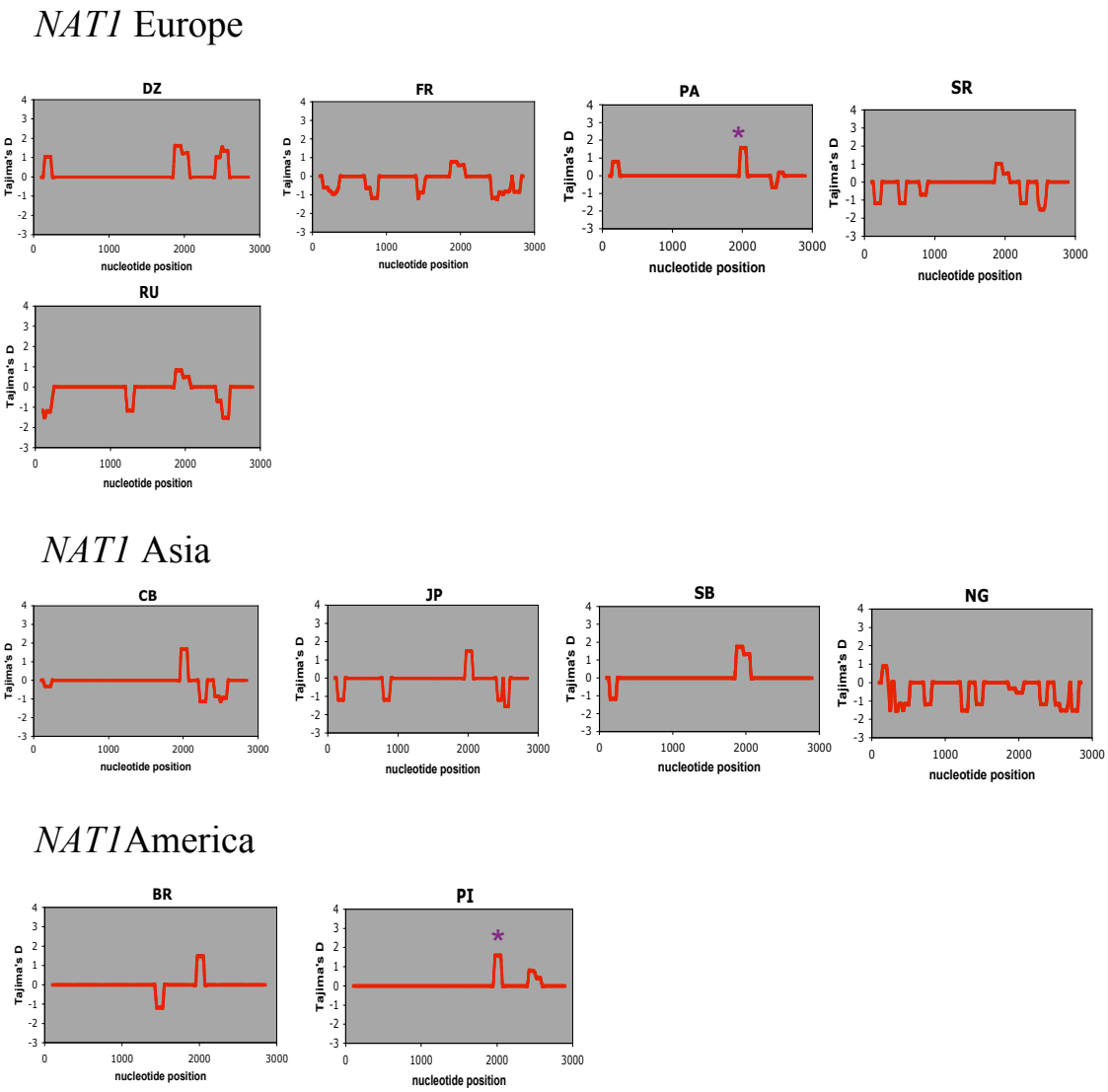


Figure 3.4

Figure 3.4: Sliding window of Tajima's D estimates the statistic across the region at defined window lengths of 100 sites at steps of 25 sites. Purple asterisks indicate significance at the $p<0.05$ level. Populations included in each continental grouping follow that presented in Table 2.1.

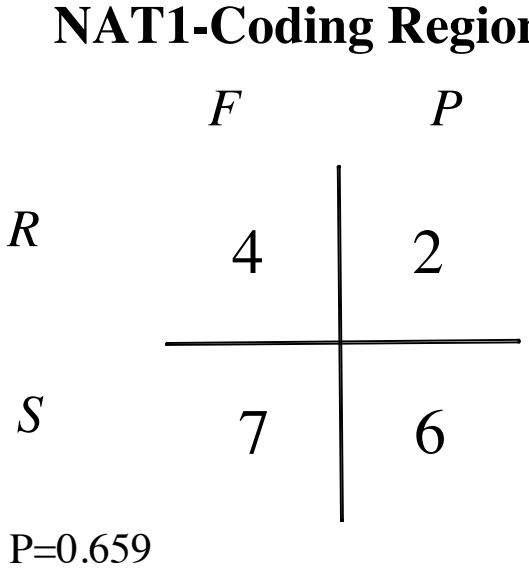


Figure 3.5

Figure 3.5: McDonald-Krietman tests for *NAT1* for the *NAT1*coding region.

Intraspecific Variability	*NAT1*	*NAT2*
Segregating Sites Obs	8	18
Segregating Sites Exp	10.36	15.64
Sample Size	529	529
Interspecies Divergence		
Number of Differences Obs	13.06	13.78
Number of Differences Exp	10.7	16.14
Total sites	870	870
chi-square value	1.318	
p-value	0.251	

Intraspecific Variability	*NAT1*	*NATP1*
Segregating Sites Obs	8	8
Segregating Sites Exp	5.94	10.06
Sample Size	556	556
Interspecies Divergence		
Number of Differences Obs	13.06	27.69
Number of Differences Exp	15.12	25.63
Total sites	870	870
chi-square value	1.331	
p-value	0.2487	

Figure 3.6

Figure 3.6: HKA tests of intra- and inter-specific variation at *NAT1*, compared to *NAT2* and *NATP1*.

Population Differentiation

Population groups do tend to cluster by broad geographic region, in general, in the MDS plots for the *NAT1* locus (Figure 3.7). However, clear distinction of clustering by geography is not observed. Given the observed pattern, the *NAT1* locus does not appear to correspond to demographic expectations, as would be expected under a neutral scenario. Asian and American populations are positioned peripherally in this analysis. Also, the Cambodian group from East Asia clusters most closely with Italian and Russian groups in the right-hand region of the plot. AMOVA results, presented in Figure 3.7, do not indicate a higher than expected percentage of observed variation within populations (99.36%) (*e.g.* populations are not more similar than expected). Additionally, variance among groups, (4.89%), accounts for a greater proportion of the total variation than does the variation among populations within groups (2.75%). *NAT1* fixation indices for F_{ST}, F_{CT}, and F_{SC} are similar to that observed for other nuclear loci in humans, and equal 0.076, 0.048, and 0.029, respectively.

Phylogenetic Haplotype Networks

The median-joining network for the *NAT1* region (Figure 3.8a) illustrates three main nodes, which correspond with variation observed at three mutations located in the 3' UTR region of the gene. *NAT1* hap46 (*NAT1*4*) and *NAT1* hap57 (*NAT1*10*) are present in high frequency (258 and 98 chromosomes, respectively) in all population groups (also see Table 3.1). These two haplotypes differ only by variants at the three 3' SNPs at positions +1088, +1095, and +1191, as indicated Figure 3.8a. Hap46 is

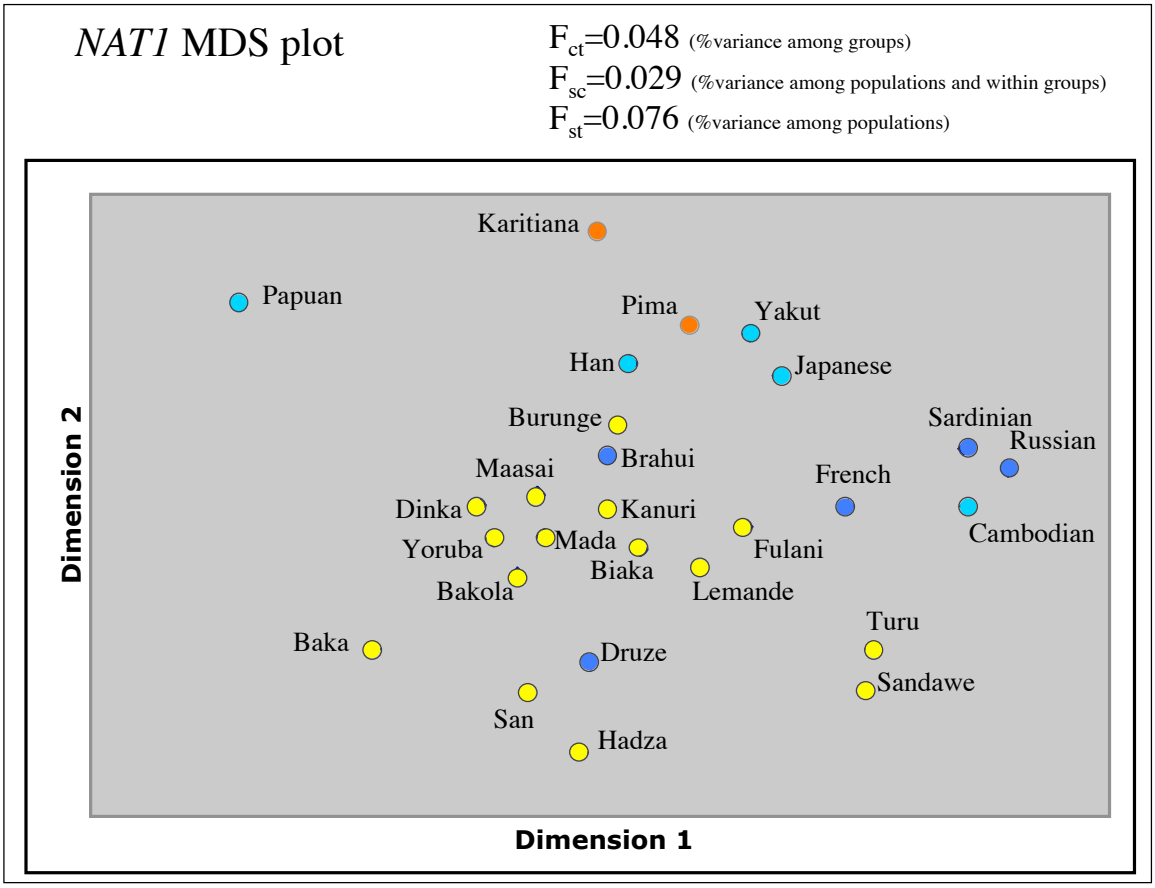


Figure 3.7

Figure 3.7: Multi-Dimensional Scaling plot for the *NAT1* loci. AMOVA results indicated in inset, where variance among groups= F_{CT}, variance among populations within groups= F_{SC}, and variance within populations= F_{ST} .Yellow=Africa; Blue=Europe; Turquoise =Asia; Orange=Americas. Population structure is specified according to the population groupings listed in Table 2.3.

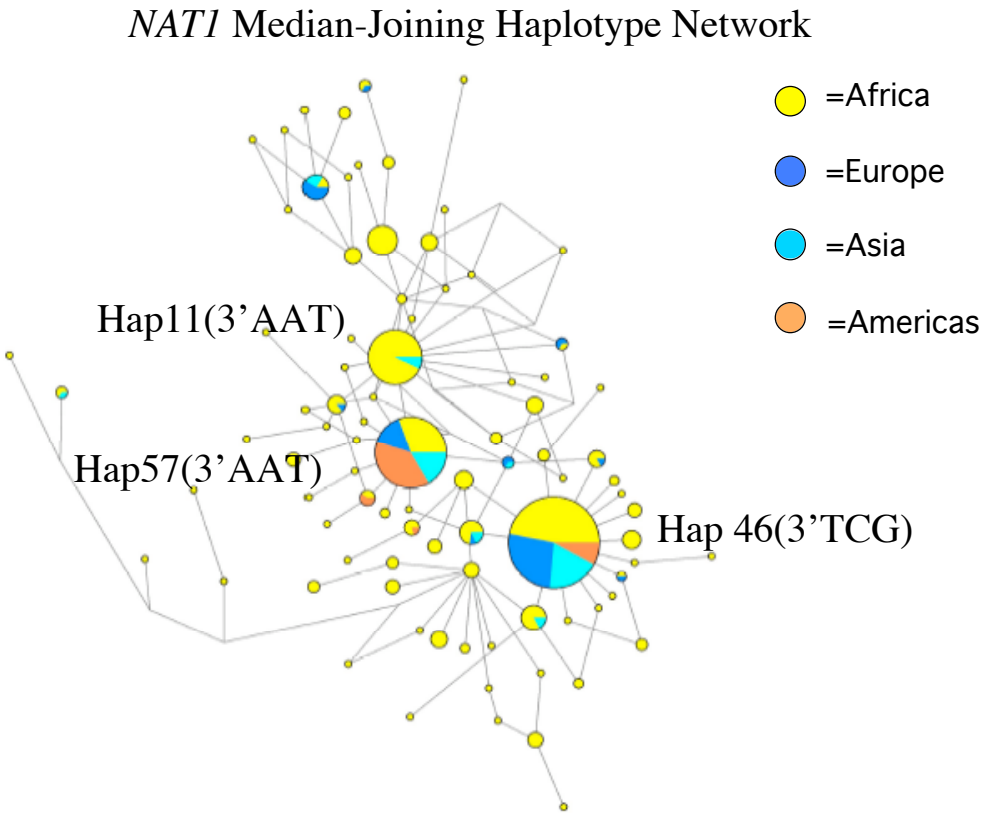
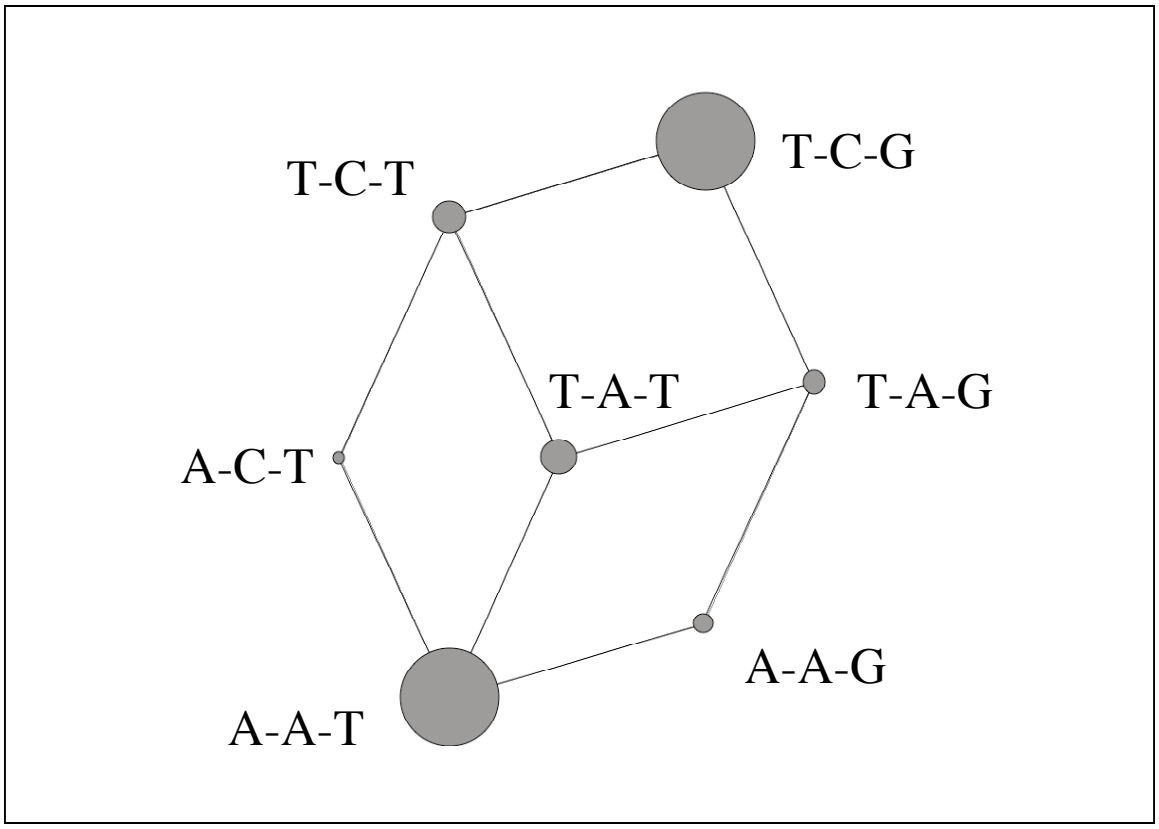


Figure 3.8a

Figure 3.8a: Median-joining network for the *NAT1* region created using Network 4.5, available at fluxus-engineering.com (Bandelt, 1999). Major geographic groupings are indicated, where Yellow=Africa; Blue=Europe; Turquoise=Asia; and Orange=Americas. The three main nodes illustrate the three phased haplotypes representing the greatest frequency of individuals (also see Table 3.1), as well as indicate the 3' SNPs at positions 1088-1095-1191.



SNP order: 1088-1095-1191

Figure 3.8b

Figure 3.8b: Median-joining network for *NAT1* 3' SNPs +1181 A/T,+1095 A/C, and +1191 T/G, created using Network 4.5, available at fluxus-engineering.com (Bandelt, 1999). Two main haplotypes are observed, in approximately equal frequencies. The **A**-A-T, specifically the +1088A, corresponds to the obliteration of the first poly-A signal and usage of the second poly-A site in all individuals (Boukouvala and Fakis, 2005).

characterized by having 3' SNPs T-C-G, where hap57 has 3' SNPs A-A-T (Table 3.1). Several low frequency haploypes are radiating from the largest node, corresponding to hap46, illustrating a star-like pattern in this region of the network. *NAT1* hap11 is the third most frequent haplotype for this *NAT1* dataset, but is confined to African (with the exception of the Sandawe) and Asian (Han and New Guinea) populations. The network illustrated in Figure 3.8b focuses only on the three 3' mutations of interest (+1088, +1095, and +1191). Two main nodes are observed representing T-C-G and A-T-T haplotypes. Both haplotypes are present at approximately equal frequencies in all population groups.

Patterns of intragenic Linkage Disequilibrium (LD)

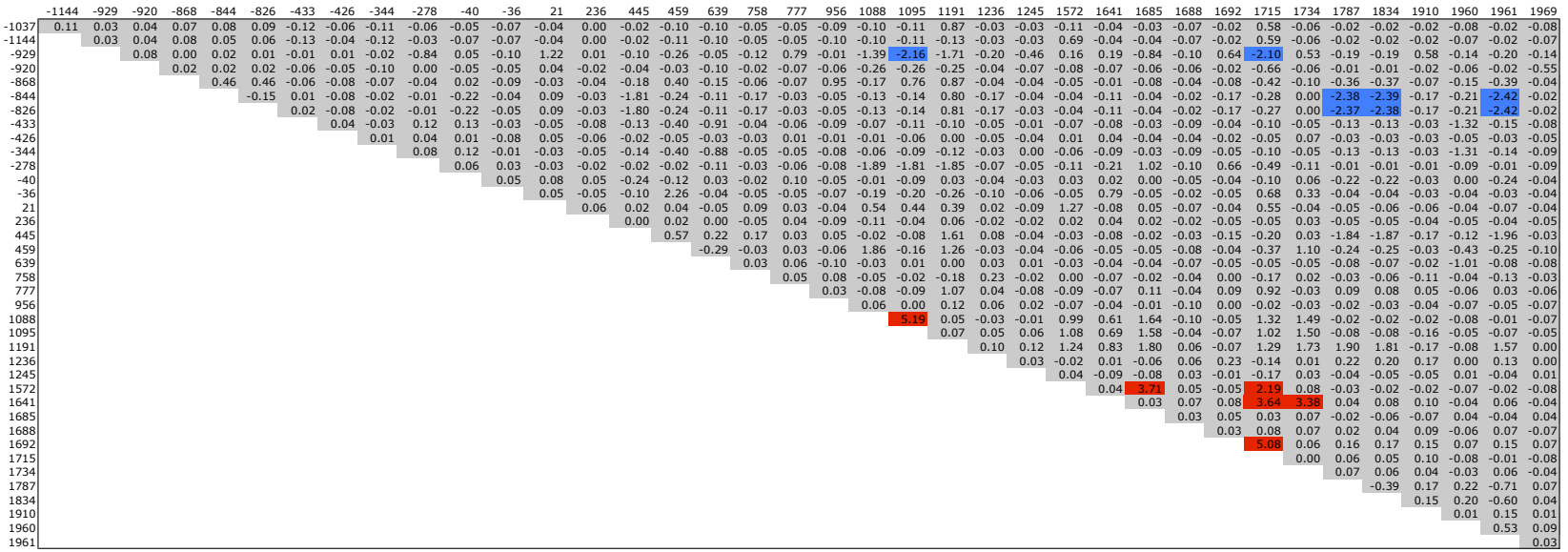
Eighty-eight distinct haplotypes were observed from the 2856 bp *NAT1* region, indicating that recombination has affected the current pattern of diversity present at this locus (Table 3.1). The results of the LDhat analyses presented in Figure 3.9, confirm higher levels of recombination for Africans, compared to non-African groups. This pattern has been observed for other nuclear loci, such as IL13 and CD4 (Tarazona-Santos and Tishkoff, 2005; Tishkoff et al., 1996). Interestingly, the significant deviation from the estimated average recombination across the *NAT1* region analyzed is outside the coding region. The majority of SNPs undergoing more recombination than expected (shown in RED) are located in the 3' UTR of *NAT1* (Figure 3.9). Results indicate that significant recombination, observed in African and non-African groups, occurs only between a few specific SNPs at *NAT1*. Specifically, the LDhat analyses for African and East African population groups indicates higher than expected recombination between six

SNP pairs, which occur outside the exon region. In close proximity to the first poly-A signal in the 3' UTR of *NAT1,* we find higher than expected recombination between SNP +1088, located in close proximity to the first poly-A signal (AATAA**A**), and 1095 (see Figure 3.2). We observe a similar pattern for West Africa, Europe and Asia for these two SNPs with the addition of a third SNP at position +1191. Higher than expected linkage between sites (shown in BLUE) is observed at the *NAT1* loci between some of the most 5' and 3' SNPs, indicative of linkage across the region. In contrast, greater than expected LD is seen in West Africa between SNPs at positions -929 and +1088, and +1095, and +1191. Finally, higher than expected linkage is observed in the Americas between the 3' SNPs at positions +1088,+1095, and +1191 (Figure 3.9).

Because of the differing patterns of linkage disequilibrium and recombination at the 3' SNPs at positions +1088, +1095, and +1191 in different population groupings, in conjunction with the increased heterozygosity of this region relative to variation across the entire *NAT1* 3kb region in general, we suspected that LDhat may be yielding spurious results. To confirm the pattern of linkage across the *NAT1* region, and in particular at these three 3' SNPs of interest, pairwise estimates of D' were calculated (Barrett et al., 2005) and are presented in Figure 3.10 . These results indicate significant linkage disequilibrium (shown in RED) between SNPs at positions +1088, +1095, and +1191, in all population groups. We observe consistency between both methods in that linkage disequilibrium is observed across the *NAT1* region, indicative of constraint preventing the accumulation of variation at this locus. Additionally, SNPs at positions +1088, +1095, and +1191 form a single haplotype block observed to be present in all population groups, with the exception of the European population (Figure 3.10). We observe strong and

significant LD across the entire *NAT1* region in all populations (with the exception of the Americas). However, low levels of LD are observed at sites flanking the SNPs at positions +1088, +1095, and +1191 (Figure 3.10), indicating that recombination may be affecting sites flanking this 3' region of interest.

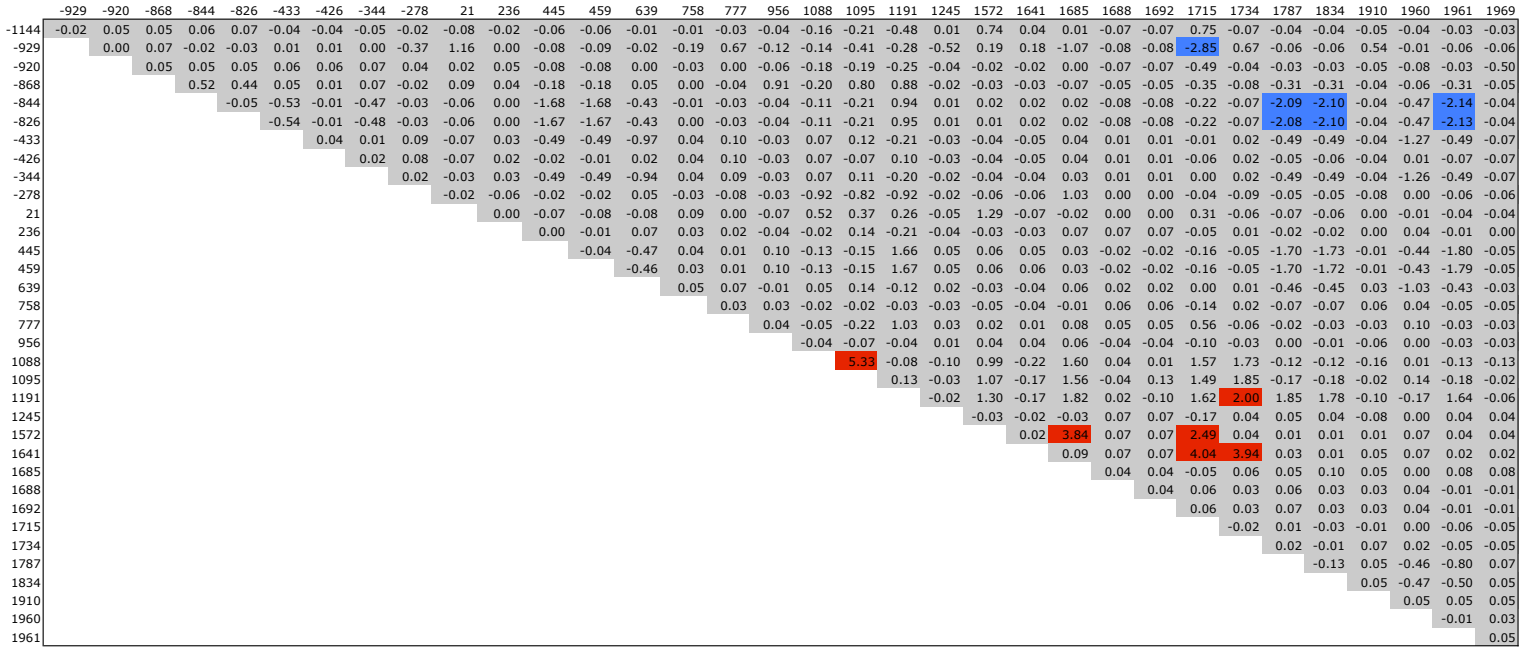
Africa
NAT1



	-1144	-929	-920	-868	-844	-826	-433	-426	-344	-278	-40	-36	21	236	445	459	639	758	777	956	1088	1095	1191	1236	1245	1572	1641	1685	1688	1692	1715	1734	1787	1834	1910	1960	1961	1969
-1037	0.11	0.03	0.04	0.07	0.08	0.09	-0.12	-0.06	-0.11	-0.06	-0.05	-0.07	-0.04	0.00	-0.02	-0.10	-0.10	-0.05	-0.05	-0.09	-0.10	-0.11	0.87	-0.03	-0.03	-0.11	-0.04	-0.03	-0.07	-0.02	0.58	-0.06	-0.02	-0.02	-0.02	-0.08	-0.02	-0.08
-1144		0.03	0.04	0.08	0.05	0.06	-0.13	-0.04	-0.12	-0.03	-0.07	-0.07	-0.04	0.00	-0.02	-0.11	-0.10	-0.05	-0.05	-0.10	-0.10	-0.11	-0.13	-0.03	-0.03	0.69	-0.04	-0.04	-0.07	-0.02	0.59	-0.06	-0.02	-0.02	-0.02	-0.07	-0.02	-0.07
-929			0.08	0.00	0.02	0.01	-0.01	0.01	-0.02	-0.84	0.05	-0.10	1.22	0.01	-0.10	-0.26	-0.05	-0.12	0.79	-0.01	-1.39	-2.16	-1.71	-0.20	-0.46	0.16	0.19	-0.84	-0.10	0.64	-2.10	0.53	-0.19	-0.19	0.58	-0.14	-0.20	-0.14
-920				0.02	0.02	0.02	-0.06	-0.05	-0.10	0.00	-0.05	-0.05	0.04	-0.02	-0.04	-0.03	-0.10	-0.02	-0.07	-0.06	-0.26	-0.26	-0.25	-0.04	-0.07	-0.08	-0.07	-0.06	-0.06	-0.02	-0.66	-0.06	-0.01	-0.01	-0.02	-0.06	-0.02	-0.55
-868					0.46	0.46	-0.06	-0.08	-0.07	-0.04	0.02	-0.09	-0.03	-0.04	-0.18	0.40	-0.15	-0.06	-0.07	0.95	-0.17	0.76	0.87	-0.04	-0.04	-0.05	-0.01	-0.08	-0.04	-0.08	-0.42	-0.10	-0.36	-0.37	-0.07	-0.15	-0.39	-0.04
-844						-0.15	0.01	-0.08	-0.02	-0.01	-0.22	-0.04	0.09	-0.03	-1.81	-0.24	-0.11	-0.17	-0.03	-0.05	-0.13	-0.14	0.80	-0.17	-0.04	-0.04	-0.11	-0.04	-0.02	-0.17	-0.28	0.00	-2.38	-2.39	-0.17	-0.21	-2.42	-0.02
-826							0.02	-0.08	-0.02	-0.01	-0.22	-0.05	0.09	-0.03	-1.80	-0.24	-0.11	-0.17	-0.03	-0.05	-0.13	-0.14	0.81	-0.17	-0.03	-0.04	-0.11	-0.04	-0.02	-0.17	-0.27	0.00	-2.37	-2.38	-0.17	-0.21	-2.42	-0.02
-433								0.04	-0.03	0.12	0.13	-0.03	-0.05	-0.08	-0.13	-0.40	-0.91	-0.04	-0.06	-0.09	-0.07	-0.11	-0.10	-0.05	-0.01	-0.07	-0.08	-0.03	-0.09	-0.04	-0.10	-0.05	-0.13	-0.13	-0.03	-1.32	-0.15	-0.08
-426									0.01	0.04	0.01	-0.08	0.05	-0.06	-0.02	-0.05	-0.03	-0.03	0.01	-0.01	-0.01	-0.06	0.00	-0.05	-0.04	0.01	0.04	-0.04	-0.04	-0.02	-0.05	0.07	-0.03	-0.03	-0.03	-0.05	-0.03	-0.05
-344										0.08	0.12	-0.01	-0.03	-0.05	-0.14	-0.40	-0.88	-0.05	-0.05	-0.08	-0.06	-0.09	-0.12	-0.03	0.00	-0.06	-0.09	-0.03	-0.09	-0.05	-0.10	-0.05	-0.13	-0.13	-0.03	-1.31	-0.14	-0.09
-278											0.06	0.03	-0.03	-0.02	-0.02	-0.02	-0.11	-0.03	-0.06	-0.08	-1.89	-1.81	-1.85	-0.07	-0.05	-0.11	-0.21	1.02	-0.10	0.66	-0.49	-0.11	-0.01	-0.01	-0.01	-0.09	-0.01	-0.09
-40												0.05	0.08	0.05	-0.24	-0.12	0.03	-0.02	0.10	-0.05	-0.01	-0.09	0.03	-0.04	-0.03	0.03	0.02	0.00	-0.05	-0.04	-0.10	0.06	-0.22	-0.22	-0.03	0.00	-0.24	-0.04
-36													0.05	-0.05	-0.10	2.26	-0.04	-0.05	-0.05	-0.07	-0.19	-0.20	-0.26	-0.10	-0.06	-0.05	0.79	-0.05	-0.02	-0.05	0.68	0.33	-0.04	-0.04	-0.03	-0.04	-0.03	-0.04
21														0.06	0.02	0.04	-0.05	0.09	0.03	-0.04	0.54	0.44	0.39	0.02	-0.09	1.27	-0.08	0.05	-0.07	-0.04	0.55	-0.04	-0.05	-0.06	-0.06	-0.04	-0.07	-0.04
236															0.00	0.02	0.00	-0.05	0.04	-0.09	-0.11	-0.04	0.06	-0.02	-0.02	0.02	0.04	0.02	-0.02	-0.05	-0.05	0.03	-0.05	-0.05	-0.04	-0.05	-0.04	-0.05
445																0.57	0.22	0.17	0.03	0.05	-0.02	-0.08	1.61	0.08	-0.04	-0.03	-0.08	-0.02	-0.03	-0.15	-0.20	0.03	-1.84	-1.87	-0.17	-0.12	-1.96	-0.03
459																	-0.29	-0.03	0.03	-0.06	1.86	-0.16	1.26	-0.03	-0.04	-0.06	-0.05	-0.05	-0.08	-0.04	-0.37	1.10	-0.24	-0.25	-0.03	-0.43	-0.25	-0.10
639																		0.03	0.06	-0.10	-0.03	0.01	0.00	0.03	0.01	-0.03	-0.04	-0.04	-0.07	-0.05	-0.05	-0.05	-0.08	-0.07	-0.02	-1.01	-0.08	-0.08
758																			0.05	0.08	-0.05	-0.02	-0.18	0.23	-0.02	0.00	-0.07	-0.02	-0.04	0.00	-0.17	0.02	-0.03	-0.06	-0.11	-0.04	-0.13	-0.03
777																				0.03	-0.08	-0.09	1.07	0.04	-0.08	-0.09	-0.07	0.11	-0.04	0.09	0.92	-0.03	0.09	0.08	0.05	-0.06	0.03	-0.06
956																					0.06	0.00	0.12	0.06	0.02	-0.07	-0.04	-0.01	-0.10	0.00	-0.02	-0.03	-0.02	-0.03	-0.04	-0.07	-0.05	-0.07
1088																						5.19	0.05	-0.03	-0.01	0.99	0.61	1.64	-0.10	-0.05	1.32	1.49	-0.02	-0.02	-0.02	-0.08	-0.01	-0.07
1095																							0.07	0.05	0.06	1.08	0.69	1.58	-0.04	-0.07	1.02	1.50	-0.08	-0.08	-0.16	-0.05	-0.07	-0.05
1191																								0.10	0.12	1.24	0.83	1.80	0.06	-0.07	1.29	1.73	1.90	1.81	-0.17	-0.08	1.57	0.00
1236																									0.03	-0.02	0.01	-0.06	0.06	0.23	-0.14	0.01	0.22	0.20	0.17	0.00	0.13	0.00
1245																										0.04	-0.09	-0.08	0.03	-0.01	-0.17	0.03	-0.04	-0.05	-0.05	0.01	-0.04	0.01
1572																											0.04	3.71	0.05	-0.05	2.19	0.08	-0.03	-0.02	-0.02	-0.07	-0.02	-0.08
1641																												0.03	0.07	0.08	3.64	3.38	0.04	0.08	0.10	-0.04	0.06	-0.04
1685																													0.03	0.05	0.03	0.07	-0.02	-0.06	-0.07	0.04	-0.04	0.04
1688																														0.03	0.08	0.07	0.02	0.04	0.09	-0.06	0.07	-0.07
1692																															5.08	0.06	0.16	0.17	0.15	0.07	0.15	0.07
1715																																0.00	0.06	0.05	0.10	-0.08	-0.01	-0.08
1734																																	0.07	0.06	0.04	-0.03	0.06	-0.04
1787																																		-0.39	0.17	0.22	-0.71	0.07
1834																																			0.15	0.20	-0.60	0.04
1910																																				0.01	0.15	0.01
1960																																					0.53	0.09
1961																																						0.03

Figure 3.9a

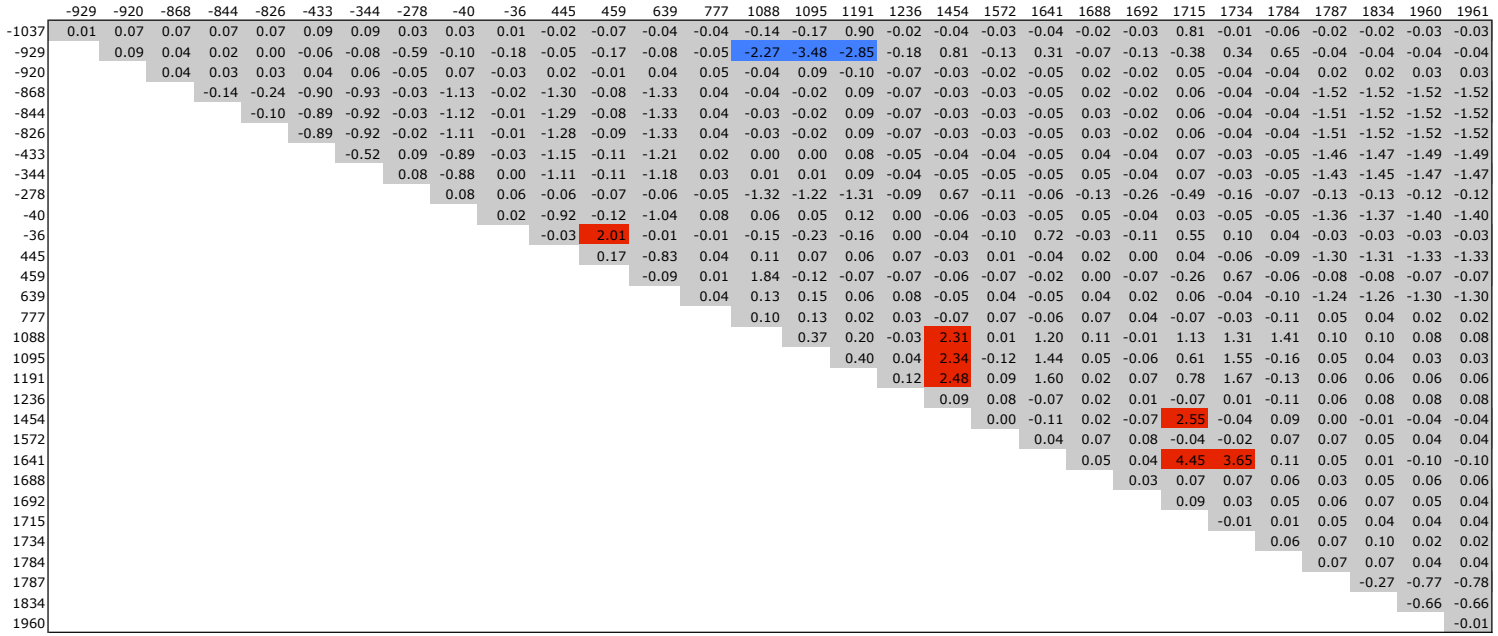
East Africa



	-929	-920	-868	-844	-826	-433	-426	-344	-278	21	236	445	459	639	758	777	956	1088	1095	1191	1245	1572	1641	1685	1688	1692	1715	1734	1787	1834	1910	1960	1961	1969
-1144	-0.02	0.05	0.05	0.06	0.07	-0.04	-0.04	-0.05	-0.02	-0.08	-0.02	-0.06	-0.06	-0.01	-0.01	-0.03	-0.04	-0.16	-0.21	-0.48	0.01	0.74	0.04	0.01	-0.07	-0.07	0.75	-0.07	-0.04	-0.04	-0.05	-0.04	-0.03	-0.03
-929		0.00	0.07	-0.02	-0.03	0.01	0.01	0.00	-0.37	1.16	0.00	-0.08	-0.09	-0.02	-0.19	0.67	-0.12	-0.14	-0.41	-0.28	-0.52	0.19	0.18	-1.07	-0.08	-0.08	-2.85	0.67	-0.06	-0.06	0.54	-0.01	-0.06	-0.06
-920			0.05	0.05	0.05	0.06	0.06	0.07	0.04	0.02	0.05	-0.08	-0.08	0.00	-0.03	0.00	-0.06	-0.18	-0.19	-0.25	-0.04	-0.02	-0.02	0.00	-0.07	-0.07	-0.49	-0.04	-0.03	-0.03	-0.05	-0.08	-0.03	-0.50
-868				0.52	0.44	0.05	0.01	0.07	-0.02	0.09	0.04	-0.18	-0.18	0.05	0.00	-0.04	0.91	-0.20	0.80	0.88	-0.02	-0.03	-0.03	-0.07	-0.05	-0.05	-0.35	-0.08	-0.31	-0.31	-0.04	-0.06	-0.31	-0.05
-844					-0.05	-0.53	-0.01	-0.47	-0.03	-0.06	0.00	-1.68	-1.68	-0.43	-0.01	-0.03	-0.04	-0.11	-0.21	0.94	0.01	0.02	0.02	0.02	-0.08	-0.08	-0.22	-0.07	-2.09	-2.10	-0.04	-0.47	-2.14	-0.04
-826						-0.54	-0.01	-0.48	-0.03	-0.06	0.00	-1.67	-1.67	-0.43	0.00	-0.03	-0.04	-0.11	-0.21	0.95	0.01	0.01	0.02	0.02	-0.08	-0.08	-0.22	-0.07	-2.08	-2.10	-0.04	-0.47	-2.13	-0.04
-433							0.04	0.01	0.09	-0.07	0.03	-0.49	-0.49	-0.97	0.04	0.10	-0.03	0.07	0.12	-0.21	-0.03	-0.04	-0.05	0.04	0.01	0.01	-0.01	0.02	-0.49	-0.49	-0.04	-1.27	-0.49	-0.07
-426								0.02	0.08	-0.07	0.02	-0.02	-0.01	0.02	0.04	0.10	-0.03	0.07	-0.07	0.10	-0.03	-0.04	-0.05	0.04	0.01	0.01	-0.06	0.02	-0.05	-0.06	-0.04	0.01	-0.07	-0.07
-344									0.02	-0.03	0.03	-0.49	-0.49	-0.94	0.04	0.09	-0.03	0.07	0.11	-0.20	-0.02	-0.04	-0.04	0.03	0.01	0.01	0.00	0.02	-0.49	-0.49	-0.04	-1.26	-0.49	-0.07
-278										-0.02	-0.06	-0.02	-0.02	0.05	-0.03	-0.08	-0.03	-0.92	-0.82	-0.92	-0.02	-0.06	-0.06	1.03	0.00	0.00	-0.04	-0.09	-0.05	-0.05	-0.08	0.00	-0.06	-0.06
21											0.00	-0.07	-0.08	-0.08	0.09	0.00	-0.07	0.52	0.37	0.26	-0.05	1.29	-0.07	-0.02	0.00	0.00	0.31	-0.06	-0.07	-0.06	0.00	-0.01	-0.04	-0.04
236												0.00	-0.01	0.07	0.03	0.02	-0.04	-0.02	0.14	-0.21	-0.04	-0.03	-0.03	0.07	0.07	0.07	-0.05	0.01	-0.02	-0.02	0.00	0.04	-0.01	0.00
445													-0.04	-0.47	0.04	0.01	0.10	-0.13	-0.15	1.66	0.05	0.06	0.05	0.03	-0.02	-0.02	-0.16	-0.05	-1.70	-1.73	-0.01	-0.44	-1.80	-0.05
459														-0.46	0.03	0.01	0.10	-0.13	-0.15	1.67	0.05	0.06	0.06	0.03	-0.02	-0.02	-0.16	-0.05	-1.70	-1.72	-0.01	-0.43	-1.79	-0.05
639															0.05	0.07	-0.01	0.05	0.14	-0.12	0.02	-0.03	-0.04	0.06	0.02	0.02	0.00	0.01	-0.46	-0.45	0.03	-1.03	-0.43	-0.03
758																0.03	0.03	-0.02	-0.02	-0.03	-0.03	-0.05	-0.04	-0.01	0.06	0.06	-0.14	0.02	-0.07	-0.07	0.06	0.04	-0.05	-0.05
777																	0.04	-0.05	-0.22	1.03	0.03	0.02	0.01	0.08	0.05	0.05	0.56	-0.06	-0.02	-0.03	-0.03	0.10	-0.03	-0.03
956																		-0.04	-0.07	-0.04	0.01	0.04	0.04	0.06	-0.04	-0.04	-0.10	-0.03	0.00	-0.01	-0.06	0.00	-0.03	-0.03
1088																			5.33	-0.08	-0.10	0.99	-0.22	1.60	0.04	0.01	1.57	1.73	-0.12	-0.12	-0.16	0.01	-0.13	-0.13
1095																				0.13	-0.03	1.07	-0.17	1.56	-0.04	0.13	1.49	1.85	-0.17	-0.18	-0.02	0.14	-0.18	-0.02
1191																					-0.02	1.30	-0.17	1.82	0.02	-0.10	1.62	2.00	1.85	1.78	-0.10	-0.17	1.64	-0.06
1245																						-0.03	-0.02	-0.03	0.07	0.07	-0.17	0.04	0.05	0.04	-0.08	0.00	0.04	0.04
1572																							0.02	3.84	0.07	0.07	2.49	0.04	0.01	0.01	0.01	0.07	0.04	0.04
1641																								0.09	0.07	0.07	4.04	3.94	0.03	0.01	0.05	0.07	0.02	0.02
1685																									0.04	0.04	-0.05	0.06	0.05	0.10	0.05	0.00	0.08	0.08
1688																										0.04	0.06	0.03	0.06	0.03	0.03	0.04	-0.01	-0.01
1692																											0.06	0.03	0.07	0.03	0.03	0.04	-0.01	-0.01
1715																												-0.02	0.01	-0.03	-0.01	0.00	-0.06	-0.05
1734																													0.02	-0.01	0.07	0.02	-0.05	-0.05
1787																														-0.13	0.05	-0.46	-0.80	0.07
1834																															0.05	-0.47	-0.50	0.05
1910																																0.05	0.05	0.05
1960																																	-0.01	0.03
1961																																		0.05

Figure 3.9b

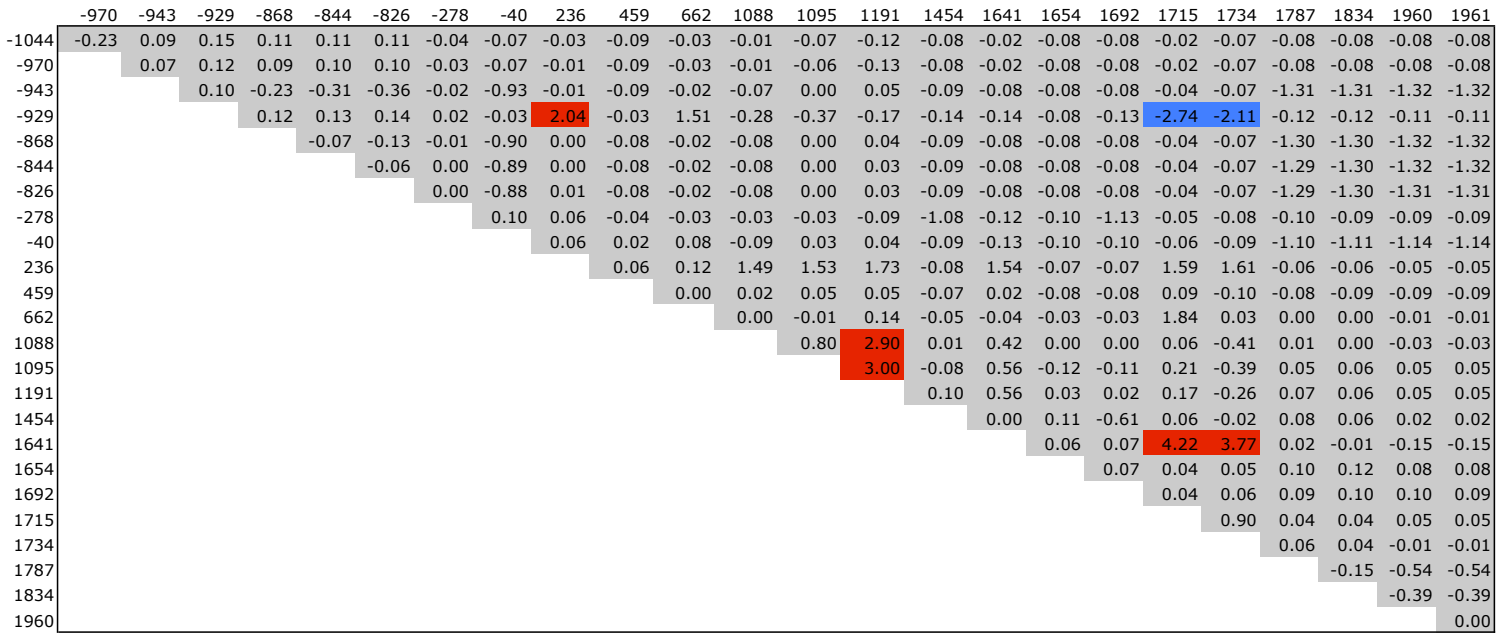
West Africa



	-929	-920	-868	-844	-826	-433	-344	-278	-40	-36	445	459	639	777	1088	1095	1191	1236	1454	1572	1641	1688	1692	1715	1734	1784	1787	1834	1960	1961
-1037	0.01	0.07	0.07	0.07	0.07	0.09	0.09	0.03	0.03	0.01	-0.02	-0.07	-0.04	-0.04	-0.14	-0.17	0.90	-0.02	-0.04	-0.03	-0.04	-0.02	-0.03	0.81	-0.01	-0.06	-0.02	-0.02	-0.03	-0.03
-929		0.09	0.04	0.02	0.00	-0.06	-0.08	-0.59	-0.10	-0.18	-0.05	-0.17	-0.08	-0.05	-2.27	-3.48	-2.85	-0.18	0.81	-0.13	0.31	-0.07	-0.13	-0.38	0.34	0.65	-0.04	-0.04	-0.04	-0.04
-920			0.04	0.03	0.03	0.04	0.06	-0.05	0.07	-0.03	0.02	-0.01	0.04	0.05	-0.04	0.09	-0.10	-0.07	-0.03	-0.02	-0.05	0.02	-0.02	0.05	-0.04	-0.04	0.02	0.02	0.03	0.03
-868				-0.14	-0.24	-0.90	-0.93	-0.03	-1.13	-0.02	-1.30	-0.08	-1.33	0.04	-0.04	-0.02	0.09	-0.07	-0.03	-0.03	-0.05	0.02	-0.02	0.06	-0.04	-0.04	-1.52	-1.52	-1.52	-1.52
-844					-0.10	-0.89	-0.92	-0.03	-1.12	-0.01	-1.29	-0.08	-1.33	0.04	-0.03	-0.02	0.09	-0.07	-0.03	-0.03	-0.05	0.03	-0.02	0.06	-0.04	-0.04	-1.51	-1.52	-1.52	-1.52
-826						-0.89	-0.92	-0.02	-1.11	-0.01	-1.28	-0.09	-1.33	0.04	-0.03	-0.02	0.09	-0.07	-0.03	-0.03	-0.05	0.03	-0.02	0.06	-0.04	-0.04	-1.51	-1.52	-1.52	-1.52
-433							-0.52	0.09	-0.89	-0.03	-1.15	-0.11	-1.21	0.02	0.00	0.00	0.08	-0.05	-0.04	-0.04	-0.05	0.04	-0.04	0.07	-0.03	-0.05	-1.46	-1.47	-1.49	-1.49
-344								0.08	-0.88	0.00	-1.11	-0.11	-1.18	0.03	0.01	0.01	0.09	-0.04	-0.05	-0.05	-0.05	0.05	-0.04	0.07	-0.03	-0.05	-1.43	-1.45	-1.47	-1.47
-278									0.08	0.06	-0.06	-0.07	-0.06	-0.05	-1.32	-1.22	-1.31	-0.09	0.67	-0.11	-0.06	-0.13	-0.26	-0.49	-0.16	-0.07	-0.13	-0.13	-0.12	-0.12
-40										0.02	-0.92	-0.12	-1.04	0.08	0.06	0.05	0.12	0.00	-0.06	-0.03	-0.05	0.05	-0.04	0.03	-0.05	-0.05	-1.36	-1.37	-1.40	-1.40
-36											-0.03	2.01	-0.01	-0.01	-0.15	-0.23	-0.16	0.00	-0.04	-0.10	0.72	-0.03	-0.11	0.55	0.10	0.04	-0.03	-0.03	-0.03	-0.03
445												0.17	-0.83	0.04	0.11	0.07	0.06	0.07	-0.03	0.01	-0.04	0.02	0.00	0.04	-0.06	-0.09	-1.30	-1.31	-1.33	-1.33
459													-0.09	0.01	1.84	-0.12	-0.07	-0.07	-0.06	-0.07	-0.02	0.00	-0.07	-0.26	0.67	-0.06	-0.08	-0.08	-0.07	-0.07
639														0.04	0.13	0.15	0.06	0.08	-0.05	0.04	-0.05	0.04	0.02	0.06	-0.04	-0.10	-1.24	-1.26	-1.30	-1.30
777															0.10	0.13	0.02	0.03	-0.07	0.07	-0.06	0.07	0.04	-0.07	-0.03	-0.11	0.05	0.04	0.02	0.02
1088																0.37	0.20	-0.03	2.31	0.01	1.20	0.11	-0.01	1.13	1.31	1.41	0.10	0.10	0.08	0.08
1095																	0.40	0.04	2.34	-0.12	1.44	0.05	-0.06	0.61	1.55	-0.16	0.05	0.04	0.03	0.03
1191																		0.12	2.48	0.09	1.60	0.02	0.07	0.78	1.67	-0.13	0.06	0.06	0.06	0.06
1236																			0.09	0.08	-0.07	0.02	0.01	-0.07	0.01	-0.11	0.06	0.08	0.08	0.08
1454																				0.00	-0.11	0.02	-0.07	2.55	-0.04	0.09	0.00	-0.01	-0.04	-0.04
1572																					0.04	0.07	0.08	-0.04	-0.02	0.07	0.07	0.05	0.04	0.04
1641																						0.05	0.04	4.45	3.65	0.11	0.05	0.01	-0.10	-0.10
1688																							0.03	0.07	0.07	0.06	0.03	0.05	0.06	0.06
1692																								0.09	0.03	0.05	0.06	0.07	0.05	0.04
1715																									-0.01	0.01	0.05	0.04	0.04	0.04
1734																										0.06	0.07	0.10	0.02	0.02
1784																											0.07	0.07	0.04	0.04
1787																												-0.27	-0.77	-0.78
1834																													-0.66	-0.66
1960																														-0.01

Figure 3.9c

Europe



	-970	-943	-929	-868	-844	-826	-278	-40	236	459	662	1088	1095	1191	1454	1641	1654	1692	1715	1734	1787	1834	1960	1961
-1044	-0.23	0.09	0.15	0.11	0.11	0.11	-0.04	-0.07	-0.03	-0.09	-0.03	-0.01	-0.07	-0.12	-0.08	-0.02	-0.08	-0.08	-0.02	-0.07	-0.08	-0.08	-0.08	-0.08
-970		0.07	0.12	0.09	0.10	0.10	-0.03	-0.07	-0.01	-0.09	-0.03	-0.01	-0.06	-0.13	-0.08	-0.02	-0.08	-0.08	-0.02	-0.07	-0.08	-0.08	-0.08	-0.08
-943			0.10	-0.23	-0.31	-0.36	-0.02	-0.93	-0.01	-0.09	-0.02	-0.07	0.00	0.05	-0.09	-0.08	-0.08	-0.08	-0.04	-0.07	-1.31	-1.31	-1.32	-1.32
-929				0.12	0.13	0.14	0.02	-0.03	2.04	-0.03	1.51	-0.28	-0.37	-0.17	-0.14	-0.14	-0.08	-0.13	-2.74	-2.11	-0.12	-0.12	-0.11	-0.11
-868					-0.07	-0.13	-0.01	-0.90	0.00	-0.08	-0.02	-0.08	0.00	0.04	-0.09	-0.08	-0.08	-0.08	-0.04	-0.07	-1.30	-1.30	-1.32	-1.32
-844						-0.06	0.00	-0.89	0.00	-0.08	-0.02	-0.08	0.00	0.03	-0.09	-0.08	-0.08	-0.08	-0.04	-0.07	-1.29	-1.30	-1.32	-1.32
-826							0.00	-0.88	0.01	-0.08	-0.02	-0.08	0.00	0.03	-0.09	-0.08	-0.08	-0.08	-0.04	-0.07	-1.29	-1.30	-1.31	-1.31
-278								0.10	0.06	-0.04	-0.03	-0.03	-0.03	-0.09	-1.08	-0.12	-0.10	-1.13	-0.05	-0.08	-0.10	-0.09	-0.09	-0.09
-40									0.06	0.02	0.08	-0.09	0.03	0.04	-0.09	-0.13	-0.10	-0.10	-0.06	-0.09	-1.10	-1.11	-1.14	-1.14
236										0.06	0.12	1.49	1.53	1.73	-0.08	1.54	-0.07	-0.07	1.59	1.61	-0.06	-0.06	-0.05	-0.05
459											0.00	0.02	0.05	0.05	-0.07	0.02	-0.08	-0.08	0.09	-0.10	-0.08	-0.09	-0.09	-0.09
662												0.00	-0.01	0.14	-0.05	-0.04	-0.03	-0.03	1.84	0.03	0.00	0.00	-0.01	-0.01
1088													0.80	2.90	0.01	0.42	0.00	0.00	0.06	-0.41	0.01	0.00	-0.03	-0.03
1095														3.00	-0.08	0.56	-0.12	-0.11	0.21	-0.39	0.05	0.06	0.05	0.05
1191															0.10	0.56	0.03	0.02	0.17	-0.26	0.07	0.06	0.05	0.05
1454																0.00	0.11	-0.61	0.06	-0.02	0.08	0.06	0.02	0.02
1641																	0.06	0.07	4.22	3.77	0.02	-0.01	-0.15	-0.15
1654																		0.07	0.04	0.05	0.10	0.12	0.08	0.08
1692																			0.04	0.06	0.09	0.10	0.10	0.09
1715																				0.90	0.04	0.04	0.05	0.05
1734																					0.06	0.04	-0.01	-0.01
1787																						-0.15	-0.54	-0.54
1834																							-0.39	-0.39
1960																								0.00

Figure 3.9d

Asia

	-868	-844	-826	-433	-344	-40	236	445	459	639	1088	1095	1191	1454	1527	1641	1715	1734	1787	1834	1960	1961
-929	0.07	0.07	0.07	0.07	0.07	0.07	0.09	0.07	0.07	0.07	6.71	6.30	6.43	0.09	0.07	0.18	5.50	0.18	0.07	0.07	0.07	0.07
-868		0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.09	0.07	0.16	0.06	0.06	0.07	0.08	0.07	0.00	0.00	0.00	0.00
-844			0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.09	0.07	0.16	0.06	0.06	0.07	0.08	0.07	0.00	0.00	0.00	0.00
-826				0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.09	0.07	0.16	0.06	0.06	0.07	0.08	0.07	0.00	0.00	0.00	0.00
-433					0.00	0.00	0.06	0.00	0.00	0.00	0.09	0.07	0.16	0.06	0.06	0.07	0.08	0.07	0.00	0.00	0.00	0.00
-344						0.00	0.06	0.00	0.00	0.00	0.09	0.07	0.16	0.06	0.06	0.07	0.08	0.07	0.00	0.00	0.00	0.00
-40							0.06	0.00	0.00	0.00	0.09	0.07	0.16	0.06	0.06	0.07	0.08	0.07	0.00	0.00	0.00	0.00
236								0.06	0.06	0.06	0.12	0.07	0.16	0.06	0.06	0.07	0.08	0.07	0.06	0.06	0.06	0.06
445									0.00	0.00	0.09	0.07	0.16	0.06	0.06	0.07	0.08	0.07	0.00	0.00	0.00	0.00
459										0.00	0.09	0.07	0.16	0.06	0.06	0.07	0.08	0.07	0.00	0.00	0.00	0.00
639											0.09	0.07	0.16	0.06	0.06	0.07	0.08	0.07	0.00	0.00	0.00	0.00
1088												0.68	0.44	0.09	0.12	0.06	6.93	0.06	0.09	0.09	0.09	0.09
1095													1.31	0.07	0.07	0.04	0.02	0.04	0.07	0.07	0.07	0.07
1191														0.16	0.16	0.04	0.01	0.04	0.16	0.16	0.16	0.16
1454															0.06	0.07	0.08	0.07	0.06	0.06	0.06	0.06
1527																0.07	0.08	0.07	0.06	0.06	0.06	0.06
1641																	0.86	0.00	0.07	0.07	0.07	0.07
1715																		0.86	0.08	0.08	0.08	0.08
1734																			0.07	0.07	0.07	0.07
1787																				0.00	0.00	0.00
1834																					0.00	0.00
1960																						0.00

America

	1088	1095	1191	1641	1715	1734
662	0.04	0.04	0.02	-0.08	-0.01	0.07
1088		-0.14	-1.96	-0.44	0.49	0.74
1095			-1.85	-0.43	0.50	0.75
1191				-0.35	0.59	0.82
1641					1.78	1.00
1715						0.50

Figure 3.9e

Figure 3.9 Intragenic linkage and recombination matrices for the *NAT1* locus, generated using Ldhat (McVean, 2002). Reported values are for each pair of SNPs, where BLUE indicates greater than expected linkage, and RED indicates greater than expected recombination. Analyses were performed for each geographic grouping following Table 2.1: (a) Africa; (b) East Africa; (c) West Africa; (d) Europe; (e) Asia and the Americas.

NAT1 Africa

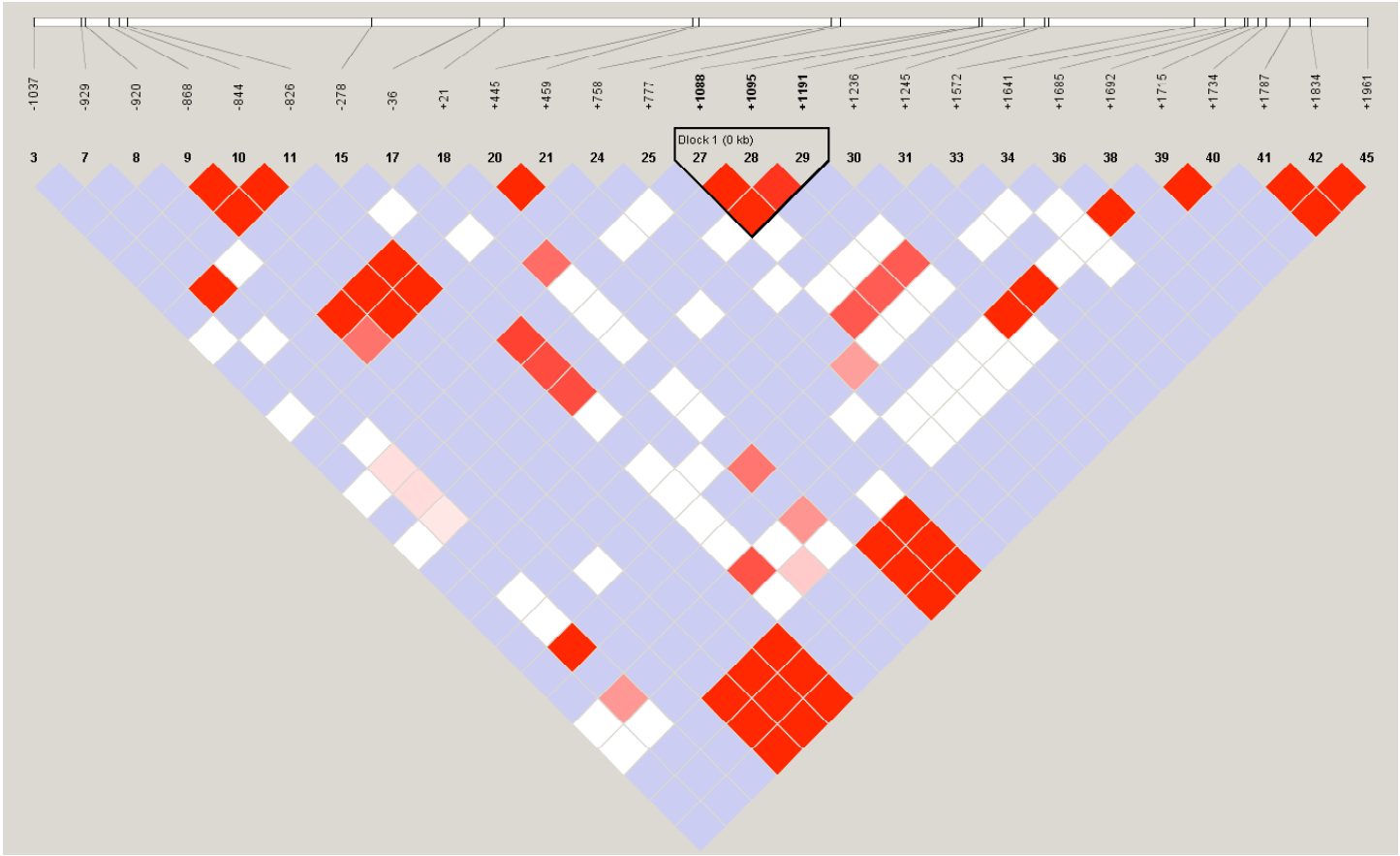


Figure 3.10a

NAT1 East Africa

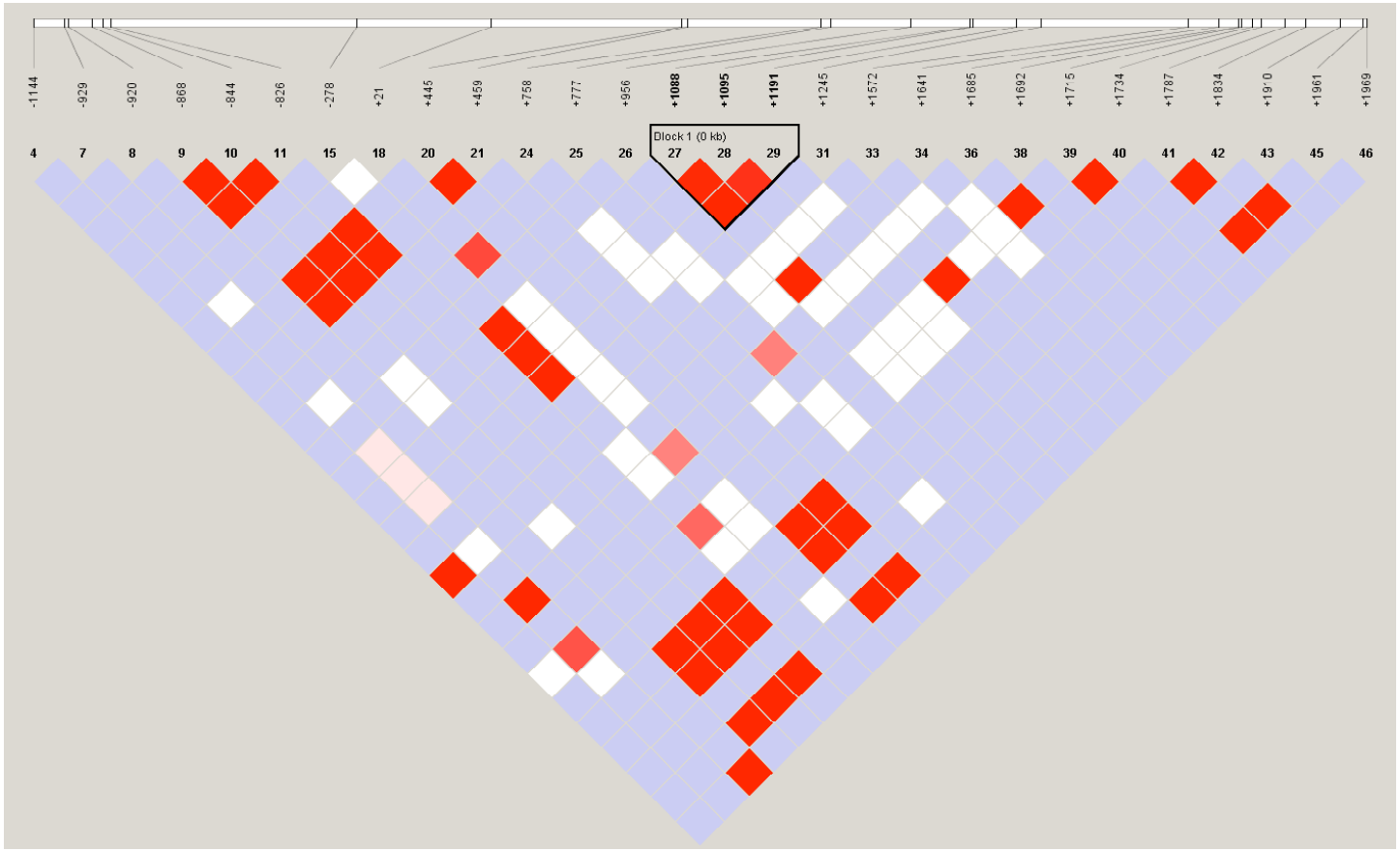


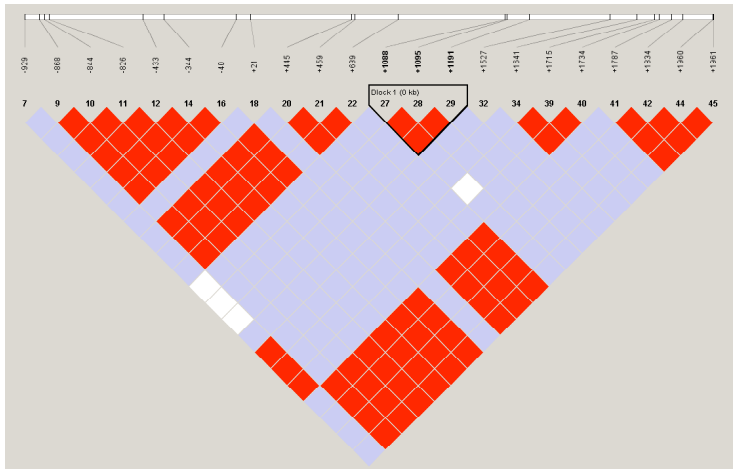
Figure 3.10b

NAT1 West Africa

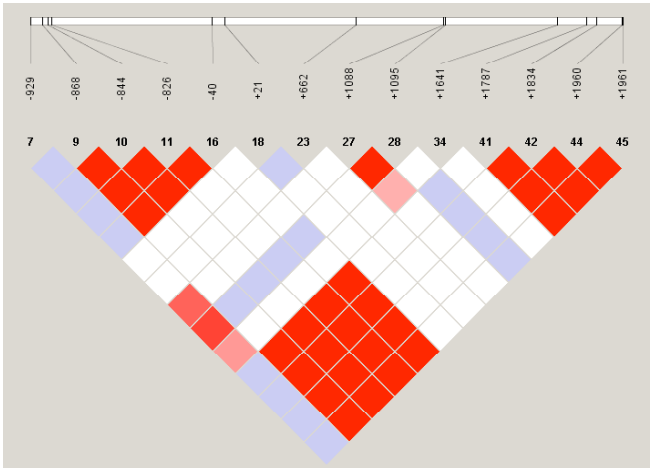
Figure 3.10c

NAT1

Asia



Europe



Americas

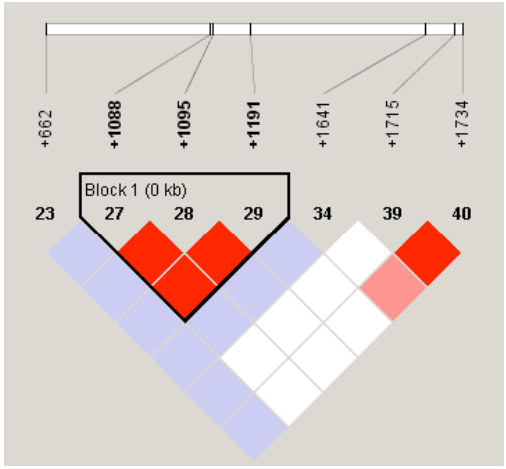


Figure 3.10d

Figure 3.10: Intragenic linkage disequilibrium and recombination graphs for the *NAT1* locus, generated using Haploview (Barrett, 2005). RED indicates D'=1 between two pairs of SNPs, indicating linkage disequilibrium is observed with high confidence (LOD≥2). Blue followed by white indicates lessening of linkage parameter D', as well as confidence of the estimate (*e.g.* blue specifies D'=1 (LOD<2); pink and shades of red specify D'<1 (LOD≥2), and white specifies D'<1 (LOD<2)). Analyses were performed for each geographic grouping as indicated in Table 2.1: (a) Africa; (b) East Africa; (c) West Africa; (d) Europe, Asia and the Americas.

DISCUSSION

We have characterized the pattern of nucleotide diversity at the *NAT1* locus, and surrounding regions, in African, European, Asian and Amerindian populations. Because of the direct role of *NAT1* in metabolism of xenobiotics and (though controversial) in cancer predisposition, *NAT1* is a potential target for natural selection. Knowledge of the pattern of genetic diversity and haplotype structure at *NAT1* in ethnically diverse population groups has important implications for understanding how *NAT1* genotypes contribute to drug-metabolism profiles and disease phenotypes at the individual level. Because of the complex expression pattern of *NAT1*, involving transcription from multiple promoters, tissue specific NCEs, as well as 3' elements, it is likely that the *NAT1* phenotype is affected to a greater extent by these adjacent elements than by polymorphism within the exon region.

Multiple Patterns of Selection

In concordance with previous studies, *NAT1* has very little observed polymorphism in the coding region, with only two amino-acid replacement causing mutations that reach near fixation levels in most population groups. Neutrality tests for the *NAT1* region are overwhelmingly negative, an observation that typically indicates the action of purifying or positive directional selection acting to constrain the region and prohibit the accumulation of variation. Other researchers (Patin et al., 2006a) have also made similar observations for the *NAT1* locus. Statistical significance was observed for tests at the intraspecific level (TD, D*, F* and H) in many cases (Table 3.2). The H

statistic was observed to be highly significant in most instances ($p<0.01$ and $p<0.001$) (Table 3.2). Because H measures high frequency variants that have swept to fixation, it is possible that the significance of these values is due to the high frequency of derived replacement changes at the two non-synonymous SNPs at positions +445G and +639T, located in the coding region of *NAT1*. However, inspection of the *NAT1* data in Table 3.1 indicate high frequency, derived variants at ten locations, including the coding region, replacement changes. High frequency, derived SNPs are located at positions -929, -826, -344, -40, +445, +639, +1088, +1527, +1572, and +1834 across the *NAT1* region. The sliding window analysis of TD confirms negative values of TD at most SNPs in the *NAT1* coding region, with one main exception. The most notable deviation at both the continental and group levels is observed for three SNPs at positions +1088, +1095, and +1191), which yield highly positive and significant TD values in nearly all populations and geographic regions. Positive values of TD could indicate the action of balancing selection acting on these 3' variants in humans.

Patin *et al*. (2006a) also indicate that there may be more than one mode of selection operating at *NAT*1 in human populations. The view of Patin *et al*. (2006a) is a conservative one, whereby balancing selection is rejected based on the lack of functional consequence of the observed replacement changes in the *NAT1* coding region. Based on the population groups studied by Patin (2006a), it is possible that they did not observe significantly positive deviation for mutations located in the 3' UTR (+1088, +1095, or +1191). The study of Patin *et al*. (2006a) included the Bakola pygmy, French, and Sardinian populations, also presented here (Figure 3.4). Each of these groups did not show significant deviation at the 3' region. However, Patin *et al*. (2006a) did study one

Bantu group from western Africa (Gabon). We find significant positive TD values at this 3' region in all Bantu groups from western Africa presented here. Patin *et al*. (2006a) focus their attention on the long basal branches of their *NAT1* phylogeny that separate haplotype *NAT1*11A,* which they deem as Eurasian specific, from the more common **4* and **10* haplotypes. With the more extensive population sampling of African groups in the present study, the **11A* haplotype was observed (Hap 40) in only two African Fulani, four New Guinea and a single Han individual (Table 3.1).

Regulatory RNA motifs

It has been argued that the relatively small number of genes in the human genome cannot account for the large variety of transcripts (Iseli et al., 2002). Three major mechanisms contribute to this complexity, alternative promoter usage, alternative splicing and polyadenylation (Iseli et al., 2002). The alignment of published *NAT1* ESTs (Boukouvala and Sim, 2005) shown in Figure 1.6, indicates differential promoter and poly-A signal usage in *NAT1* (as well as the presence of various NCEs in different tissue types). It is thought that alternative promoter usage enables the diversification of transcriptional regulation at a single locus, thereby affecting gene expression in different cell lineages, tissue types and at different developmental stages (Davuluri et al., 2008). The presence of tandem, and differentially utilized poly-A signals is a common feature of eukaryotic genes. What governs the usage of these poly-A signals has not been clarified (Boukouvala and Sim, 2005; Edwalds-Gilbert et al., 1997).

Alternative polyadenylation of NAT1 in humans and NHPs

Alternative polyadenylation has been shown to give rise to transcripts with distinct biological properties (Iseli et al., 2002). It has been observed that alternative poly-A signals are employed in *NAT* genes for rabbit, hamster and rat (Blum et al., 1989; Ebisawa et al., 1995; Land et al., 1994), where mouse has only a single polyadenylation signal (Boukouvala et al., 2003). The *NAT1* sequence alignment presented in Figure 3.2 illustrates the region following the $(TAA)_n$ repeat and surrounding the first and second poly-A signals and sites in humans. It has been observed that an A allele at SNP +1088(A/T) in humans is sufficient to completely abolish the first polyadenylation signal of *NAT1* (Boukouvala and Fakis, 2005). Figure 3.3 illustrates that the $(TAA)_n$ repeat varies amongst non-human primates (NHP), but most importantly that the SNP at +1088 is not variable in NHP where all NHP, from our closest living NHP relative the chimpanzee (represented by two species, *P.troglodytes* and *P.paniscus*) to the most distant macaque (represented by *M. mulatta*, *M. fiscularis*, and *M. nigra*) each possess at T at +1088. Additionally, the second polyadenylation signal at position +1203, appears to be highly conserved across NHP taxa. It has been indicated that this second poly-A signal may be used preferentially in humans (Boukouvala and Fakis, 2005).

A closer look at the mutations responsible for the highly positive values of TD (Figure 3.4) reveal that two main 3' haplotypes (A-T-T and T-C-A) at positions +1088, +1095, and +1191 have been maintained at high frequencies in all human groups studied (Figures 3.7a and 3.7b). As stated previously in Chapter I, the *NAT1*10* haplotype differs from *NAT1*4* (WT) by two of these three mutations (1088A and 1095A) and was

originally thought to have increased acetylation activity through utilization of the second alternative (stronger) poly-A signal (Hein et al., 2000). This finding has not been supported in subsequent studies (Barker, pers. comm.), and *NAT1*10* is thought to confer the same acetylator activity as does wild-type (Bruhn et al., 1999; Yang et al., 2000). However, it has been indicated that the presence of 1088A and 1095A (which are almost always associated with 1191T) result in increased *NAT1* enzyme production, although the mechanism has been far from clarified (Barker, pers. comm.). The persistence of these two distinct 3' haplotypes, in conjunction with the highly positive TD values for this region of *NAT1*, is consistent with the hypothesis that long-term balancing selection has maintained these two 3' haplotypes in high frequency in humans and implies these differing *NAT1* 3' haplotypes play some functional role that is currently unidentified.

Conclusions

Our analysis has highlighted some important aspects of the patterns of genetic variation at the *NAT1* locus in humans. Overall, *NAT1* variation in the coding region is low within population groups and tests of neutrality indicate purifying selection has prohibited variation to accumulate within the exon region of the gene. Being that *NAT1* is expressed ubiquitously and at different stages of development, this is not surprising. However, a different pattern of selection (balancing selection) has affected the 3' UTR region of *NAT1*. Our observations indicate that differential polyadenylation may be one way in which the NAT1 phenotype is altered and may be an important factor in currently uncharacterized modes of post-transcriptional regulation of NAT1 in humans.

CHAPTER IV. EVOLUTIONARY CHARACTERIZATION OF THE HUMAN *NAT2* LOCUS

RESULTS

Haplotype Reconstruction and Patterns of Genetic Diversity

We sequenced 3031 bp of the *NAT2* region in 570 chromosomes, and identified 46 SNPs (Table 4.1)**,** of which 22 SNPs have not been previously reported and are novel to the present dataset. These are located at positions -1014, -701, -349, -282, +308, +472, +518, +578, +632, +664, +766, +838, +1085, +1134, +1362, +1373, +1520, +1781, +1783, +1792, +1794, and +1853. Four of these SNPs (at positions +472, +518, +766 and +837) have recently been reported by Patin *et al*. (2006b), and have not yet been submitted to GenBank (Patin, pers. comm.). Eighteen variable sites were identified in the 870 bp coding region of *NAT2,* fifteen of which were replacement changes and three synonymous substitutions (Table 4.1). The replacement changes +70 T>A, +191 G>A, +308 C>T, +341 T>C, +403 C>G, +472 A>C, +518 A>G, +578 C>T, +590 G>A, +609 G>T, +665 T>G, +766 A>G, +803 A>G, +838 G>A, +857 G>A, result in amino acid substitutions Leu24Ile, Arg64Gln, Thr103Ile, Thr114Ile, Leu135Val, Thr158Leu, Lys173Arg, Thr193Met, Arg197Gln, Glu203Asp, Phe222Cys, Lys256Glu, Lys268Arg, Val280Met, Gly286Glu, respectively. Additionally, eight singleton mutations and a single tri-allelic site were identified. The remaining SNPs within the *NAT2* coding region have been previously reported with Genbank accession numbers as follows: +70 (rs45477599); +191(rs1801279); +282 (rs1041983); +341 (rs1801280); +403

NAT2

s/r	Inferred Functional Hap	-1014	-862	-859	-855	-701	-491	-413	-349	-282	-255	-234	-179	+70r	+191r	+282s	+308r	+341r	+403r	+472r	+481s	+518r	+578r	+590r	+609r	+633s	+665r	+766r	+803r	+838r	+857r	+1021	+1085	+1101	+1134	+1338	+1362	+1373	+1395	+1447	+1520	+1521	+1781	+1783	+1792	+1794
P. troglodytes		C	G	C	C	G	G	G	T	T	C	C	T	T	G	C	C	T	C	A	C	A	C	G	G	G	T	A	A	G	G	T	G	C	C	A	C	T	G	C	A	C	T	G	A	T
Hap 1	*NAT2*4*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	*NAT2*4*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	*NAT2*4*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.
4	*NAT2*12A*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	*NAT2*12A*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	*NAT2*12A*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	.
7	*Unk*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
8	*NAT2*12G*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	*NAT2*6m*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.
10	*NAT2*7B*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
11	*NAT2*7B*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	G	.	T	.	.	.	.	.	.
12	*NAT2*13A*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
13	*Unk*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
14	*NAT2*6A*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15	*NAT2*6A*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.
16	*NAT2*6A*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	.
17	*NAT2*6A*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	.
18	*NAT2*6A*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	C
19	*NAT2*6A*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	.	.	G	.
20	*NAT2*6A*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	G	.	.	A	.	.	.	.	.	G	.
21	*NAT2*6I*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	A	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	.
22	*NAT2*6C*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	G	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.
23	*NAT2*6C*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	G	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	.
24	*NAT2*6H*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	G	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	.
25	*NAT2*6G*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.
26	*NAT2*6G*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	.
27	*NAT2*6n*	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	.
28	*NAT2*14B*	.	C	.	.	.	A	.	.	.	.	.	.	.	A	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
29	*NAT2*14j*	.	C	.	.	.	A	.	.	.	.	.	.	.	A	T	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
30	*Unk*	.	C	.	.	.	A	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.
31	*NAT2*6A*	.	C	.	.	.	A	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	G	.
32	*NAT2*5B*	.	C	.	.	.	A	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
33	*NAT2*13A*	.	C	.	.	.	A	.	.	.	G	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
34	*NAT2*5A*	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
35	*NAT2*5A*	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
36	*NAT2*5B*	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
37	*NAT2*5B*	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
38	*NAT2*5B*	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.
39	*NAT2*5B*	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	G	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.
40	*NAT2*12A*	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
41	*NAT2*12A*	.	C	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
42	*NAT2*13A*	.	C	.	.	.	.	.	.	.	G	T	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
43	*NAT2*12B*	.	C	.	.	.	.	.	.	.	G	T	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
44	*NAT2*12A*	.	C	.	.	A	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
45	*NAT2*4*	.	C	.	T	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
46	*NAT2*12A*	.	C	.	T	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.
47	*NAT2*12A*	.	C	.	T	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
48	*NAT2*12A*	.	C	.	T	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.
49	*NAT2*12i*	.	C	.	T	.	A	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
50	*Unk*	.	C	.	T	.	A	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Table 4.1a

***NAT2* Haplotypes**

s/r *P. troglodytes*	*Inferred Functional Hap*	BG	BK	DN	FU	HZ	KN	LM	MD	MS	PB	PL	SA	SW	TR	YO	Afr	DZ	FR	PA	RU	SR	CB	HA	JP	NG	SB	BR	PI	Non	ALL
Hap 1	*NAT2*4*	2	2	2	.	1	5	2	2	.	.	1	5	2	3	1	28	2	.	3	.	3	3	10	9	3	2	8	8	51	79
2	*NAT2*4*	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	1	.	.	2	3
3	*NAT2*4*	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
4	*NAT2*12A*	.	2	.	.	.	3	.	.	.	1	5	.	.	.	2	13	.	.	.	.	.	.	.	.	.	.	.	.	.	13
5	*NAT2*12A*	.	.	.	.	.	.	.	.	1	.	.	.	1	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
6	*NAT2*12A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
7	*Unk*	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
8	*NAT2*12G*	.	1	.	.	.	.	.	.	.	3	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	4
9	*NAT2*6m*	.	.	.	.	.	.	.	.	.	.	.	.	2	1	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
10	*NAT2*7B*	4	.	.	.	2	.	.	.	.	.	.	1	1	1	.	9	.	.	1	1	.	5	5	1	1	.	.	.	14	23
11	*NAT2*7B*	1	.	.	.	.	.	.	1	.	.	.	.	.	1	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
12	*NAT2*13A*	.	.	.	.	1	.	1	.	.	4	.	.	.	.	.	6	.	.	.	1	.	.	.	1	.	.	.	.	2	8
13	*Unk*	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
14	*NAT2*6A*	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
15	*NAT2*6A*	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
16	*NAT2*6A*	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
17	*NAT2*6A*	10	1	7	9	10	1	2	6	4	2	.	.	13	3	2	70	5	3	6	3	8	2	4	4	.	1	.	.	36	106
18	*NAT2*6A*	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	5
19	*NAT2*6A*	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
20	*NAT2*6A*	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
21	*NAT2*6I*	1	.	.	.	1	.	2	.	1	.	.	.	.	.	1	6	.	.	.	.	.	.	.	.	.	.	.	.	.	6
22	*NAT2*6C*	.	.	.	.	.	1	.	1	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
23	*NAT2*6C*	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
24	*NAT2*6H*	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
25	*NAT2*6G*	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
26	*NAT2*6G*	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
27	*NAT2*6n*	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
28	*NAT2*14B*	1	1	1	1	.	3	4	2	2	.	.	.	.	1	2	18	.	.	.	.	.	.	.	.	.	.	.	.	.	18
29	*NAT2*14j*	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
30	*Unk*	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
31	*NAT2*6A*	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
32	*NAT2*5B*	.	.	.	1	1	.	1	3	2	1	.	.	.	1	.	10	1	1	.	.	.	.	.	.	.	.	.	.	2	12
33	*NAT2*13A*	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
34	*NAT2*5A*	.	.	.	.	.	.	1	.	.	.	.	.	.	1	2	4	1	1	1	.	3	.	.	.	.	.	1	.	7	11
35	*NAT2*5A*	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
36	*NAT2*5B*	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	4	6	10	12
37	*NAT2*5B*	9	4	9	10	3	3	7	10	6	1	.	1	6	8	.	77	5	5	1	3	4	.	.	.	.	.	2	3	23	100
38	*NAT2*5B*	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
39	*NAT2*5B*	.	.	1	.	.	.	.	.	.	.	.	1	1	1	.	4	2	1	.	1	1	.	.	.	.	.	.	.	5	9
40	*NAT2*12A*	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
41	*NAT2*12A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
42	*NAT2*13A*	.	2	.	2	.	1	1	.	.	.	.	.	1	3	5	15	.	.	.	.	.	.	.	.	.	.	.	.	.	15
43	*NAT2*12B*	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
44	*NAT2*12A*	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
45	*NAT2*4*	.	1	.	.	.	.	.	.	.	1	.	.	.	.	1	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
46	*NAT2*12A*	.	.	.	.	.	.	.	.	.	.	.	4	.	1	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	5
47	*NAT2*12A*	1	5	2	1	1	2	1	.	2	1	.	1	.	.	.	17	.	.	.	.	.	.	.	.	.	.	1	2	3	20
48	*NAT2*12A*	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
49	*NAT2*12i*	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
50	*Unk*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1

Table 4.1b

NAT2

s/r		-1014	-862	-859	-855	-701	-491	-413	-349	-282	-255	-234	-179	+70r	+191r	+282s	+308r	+341r	+403r	+472r	+481s	+518r	+578r	+590r	+609r	+633s	+665r	+766r	+803r	+838r	+857r	+1021	+1085	+1101	+1134	+1338	+1362	+1373	+1395	+1447	+1520	+1521	+1781	+1783
P. troglodytes	Inferred Functional Hap	C	G	C	C	G	G	G	T	T	C	C	T	T	G	C	C	T	C	A	C	A	C	G	G	G	T	A	A	G	G	T	G	C	C	A	C	T	G	C	A	C	T	G
Hap51	NAT2*12H	.	C	.	T	.	A	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
52	NAT2*12E	.	C	.	T	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	T	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.
53	NAT2*6A	.	C	.	T	.	A	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.
54	NAT2*4	.	C	T	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
55	NAT2*6A	.	.	.	.	.	A	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.
56	NAT2*5C	.	.	.	.	.	A	.	.	.	.	T	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
57	NAT2*5C	.	.	.	.	.	A	A	.	.	.	T	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
58	NAT2*5C	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
59	NAT2*5B	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
60	NAT2*12A	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
61	NAT2*12A	.	.	.	.	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.
62	NAT2*5C	.	.	.	.	.	.	A	.	.	.	T	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
63	NAT2*12A	.	.	.	T	.	A	.	C	.	G	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
64	NAT2*12A	.	.	.	T	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
65	Unk 12j	.	.	.	T	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	G	.	.	.	.	.	.	.	.	.	.	.	T	.	.	C
66	NAT2*4	.	.	.	T	.	.	.	C	.	G	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
67	NAT2*12A	.	.	.	T	.	.	.	C	.	G	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
68	NAT2*5B	T	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	G	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.
69	Unk 13c	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.
70	NAT2*6B	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.
71	NAT2*4	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.
72	Unk	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
73	NAT2*5A	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
74	NAT2*5A	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
75	NAT2*5A	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
78	NAT2*6A	.	C	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.
79	NAT2*6A	.	C	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
80	NAT2*4	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
81	NAT2*12A	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.
82	NAT2*5B	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
83	NAT2*5A	.	C	.	.	.	.	.	.	.	.	.	C	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
84	NAT2*12A	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.
85	NAT2*5A	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
86	NAT2*5B	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
87	NAT2*6A	.	C	.	T	.	A	.	.	C	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.
88	NAT2*5C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
89	NAT2*13A	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
90	NAT2*1AA	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
91	NAT2*6A	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
92	NAT2*7A	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.
93	NAT2*4	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.
94	NAT2*6A	.	C	.	.	.	A	.	.	.	.	T	.	.	.	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.
95	NAT2*4	.	C	.	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
96	NAT2*5A	.	C	T	.	.	A	.	.	.	.	.	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
97	NAT2*4	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	C	.
98	NAT2*12B	.	C	.	.	.	A	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
99	NAT2*5B	.	C	.	.	.	.	.	.	.	.	T	.	.	.	.	.	C	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	T	.	.	C	.
100	NAT2*12A	.	C	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Table 4.1c

***NAT2* Haplotypes**

s/r *P. troglodytes*	*Inferred Functional Hap*	BG	BK	DN	FU	HZ	KN	LM	MD	MS	PB	PL	SA	SW	TR	YO	Afr	DZ	FR	PA	RU	SR	CB	HA	JP	NG	SB	BR	PI	Non	ALL
Hap51	*NAT2*12H*	.	.	1	.	.	.	.	1	.	.	.	.	.	.	1	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
52	*NAT2*12E*	3	2	.	.	.	.	.	.	.	1	.	.	.	1	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	7
53	*NAT2*6A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
54	*NAT2*4*	.	.	.	.	.	.	.	.	1	.	.	.	2	.	.	3	2	2	2	.	2	.	.	5	.	.	.	.	13	16
55	*NAT2*6A*	.	.	1	.	.	.	.	.	3	.	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	4
56	*NAT2*5C*	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
57	*NAT2*5C*	.	.	.	.	.	.	1	1	1	.	.	.	.	1	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	4
58	*NAT2*5C*	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
59	*NAT2*5B*	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
60	*NAT2*12A*	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
61	*NAT2*12A*	.	2	.	.	.	.	.	.	.	.	3	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	5
62	*NAT2*5C*	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	2	.	.	.	.	.	.	.	.	.	2	3
63	*NAT2*12A*	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
64	*NAT2*12A*	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
65	*Unk 12j*	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
66	*NAT2*4*	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
67	*NAT2*12A*	.	1	.	.	.	.	.	.	.	.	1	.	4	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	6
68	*NAT2*5B*	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
69	*Unk 13c*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	1
70	*NAT2*6B*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	2	2
71	*NAT2*4*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	1
72	*Unk*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	1
73	*NAT2*5A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	1	.	.	1	.	.	.	.	.	4	4
74	*NAT2*5A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	1
75	*NAT2*5A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	1
78	*NAT2*6A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	2	2
79	*NAT2*6A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	1
80	*NAT2*4*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	1	3	3
81	*NAT2*12A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	1
82	*NAT2*5B*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	1
83	*NAT2*5A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	3	3
84	*NAT2*12A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	1
85	*NAT2*5A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	1
86	*NAT2*5B*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	1
87	*NAT2*6A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	1
88	*NAT2*5C*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	1
89	*NAT2*13A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	1	1	4	4
90	*NAT2*1AA*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	1
91	*NAT2*6A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	1
92	*NAT2*7A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	1
93	*NAT2*4*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	1
94	*NAT2*6A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	2	2
95	*NAT2*4*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	1
96	*NAT2*5A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	1
97	*NAT2*4*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	1
98	*NAT2*12B*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	1
99	*NAT2*5B*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	2	2
100	*NAT2*12A*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1
		34	30	26	26	28	24	28	28	28	18	14	14	36	30	24	388	22	18	24	12	24	14	24	20	4	8	22	22	216	604

Table 4.1d

Table 4.1 (a-d are quadrants of Table 4.1): One hundred *NAT2* phased haplotypes were identified, and listed in (a) and (c). The number of observed haplotypes in each population are shown in parts (b) and (d). All population abbreviations follow that presented in Table 2.1, in addition to the abbreviation 'Afr' indicating all African populations and 'Non', which denotes all non-African groups. Inferred functional haplotype refers to acetylator haplotypes inferred using SNPs +70 through +857, located within the 870 bp coding region of *NAT2*. Derived variants, based on comparison to *P. troglodytes NAT2* sequence, are indicated. *s* and *r* refer to silent and replacement polymorphisms.

(rs12720065); +481 (rs1799929); +590 (rs1799930); +609 (rs45618543); +803 (rs1208); +857 (rs1799931). Figure 4.1 illustrates the frequency of the derived variants for each SNP in the coding region of *NAT2* in each population.

NAT2 Acetylator Phenotype Inference

NAT2 acetylator status was inferred for each individual based on phased, diploid haplotypes, and focusing on the coding region SNPs with known affect on acetylator phenotype. Results are presented in Figure 4.2 for all global populations included in the present study, with Eastern and Western African groups pooled. Results for specific African groups are presented in Figure 4.3. In both Figures 4.2 and 4.3, individuals were pooled according to ethnic group affiliation, and population frequencies of inferred (diploid) phenotypes were calculated. Unknown haplotypes are indicated in grey, according to the particular level of uncertainty (*i.e.* unknown/slow, unknown/rapid, or unknown types for both inferred haplotypes). Most unknown haplotypes occur as expected in the African populations presented here, as African populations exhibit a higher level of diversity in comparison to other populations, and novel SNPs in the current dataset are in highest frequency within African groups, preventing direct inference of phenotype in some cases. The global acetylator frequencies shown in Figure 4.2 illustrates that the distribution of slow acetylators is confined to Africa and Europe. Rapid acetylator haplotypes are not fixed in any human population. However, rapid acetylators are more prevalent in Asia, and the Americas, in concordance with previous findings (Brockton et al., 2000; Fuselli et al., 2007). Interestingly, in Africa rapid

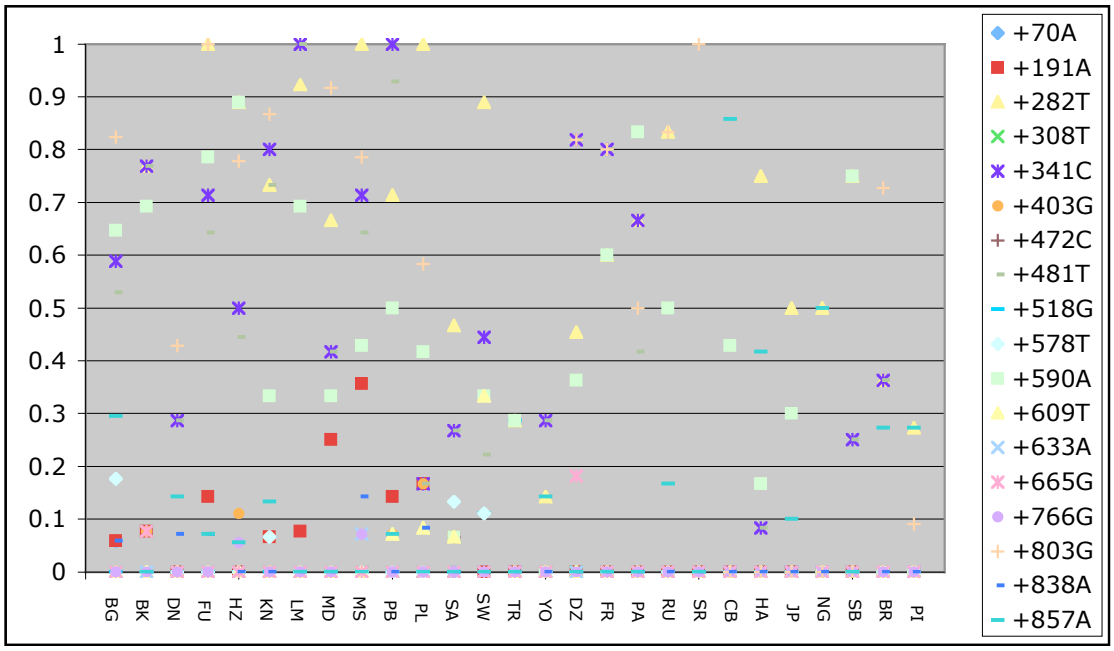


Figure 4.1

Figure 4.1: Frequency diagram of 18 *NAT2* coding region SNPs. Frequency of each SNP is indicated on the y-axis, and each population in the present analyses is listed on the x-axis (refer to Table 2.1 for population abbreviations).

acetylators are found in the highest frequencies in populations known to have some of the oldest evolutionary lineages (according to mtDNA and Y chomosome evidence), the San foragers of South Africa and the Pygmy groups, Biaka, Baka and Bakola (Campbell and Tishkoff, 2008). Population groups that have a complete absence of rapid acetylator types are the French and Russians, and in Africa the Hadza foragers of Tanzania, and the Mada and Fulani of Cameroon. From Figure 4.1, it is also clear that certain derived SNPs that are associated with "slow" and "rapid" acetylator haplotypes reach fixation in Africa. For example, in the Maasai (MS), +282T *(*6* slow or **13* rapid) and +803G (**12* slow or rapid) reach frequency of 100%. SNP +803, associated with rapid acetylator haplotypes, reaches highest frequency in all Pygmy groups. Outside of Africa, SNP +282T is notably high in Cambodians (CB), whereas SNPs +341C (**5*) and +481T reach high frequency in Russians (RU). All other groups seem to maintain multiple *NAT2* coding region SNPs at appreciable frequencies.

Statistical Tests of Neutrality

Summary statistics and estimates of neutrality for the *NAT2* locus are given in Table 4.2. Nucleotide diversity estimates (π) within continental groups for the *NAT2* locus are relatively similar between Africans and Europeans. θ_ω is more variable for all continental groups. Nucleotide diversity is slightly lower in Asian and Amerindian populations, not surprising given the smaller sample sizes represented here, and demographic history of these populations (*i.e.* population founding events). Haplotype diversity (HD) is high amongst all groups, but shows slightly lower values for non-African populations.

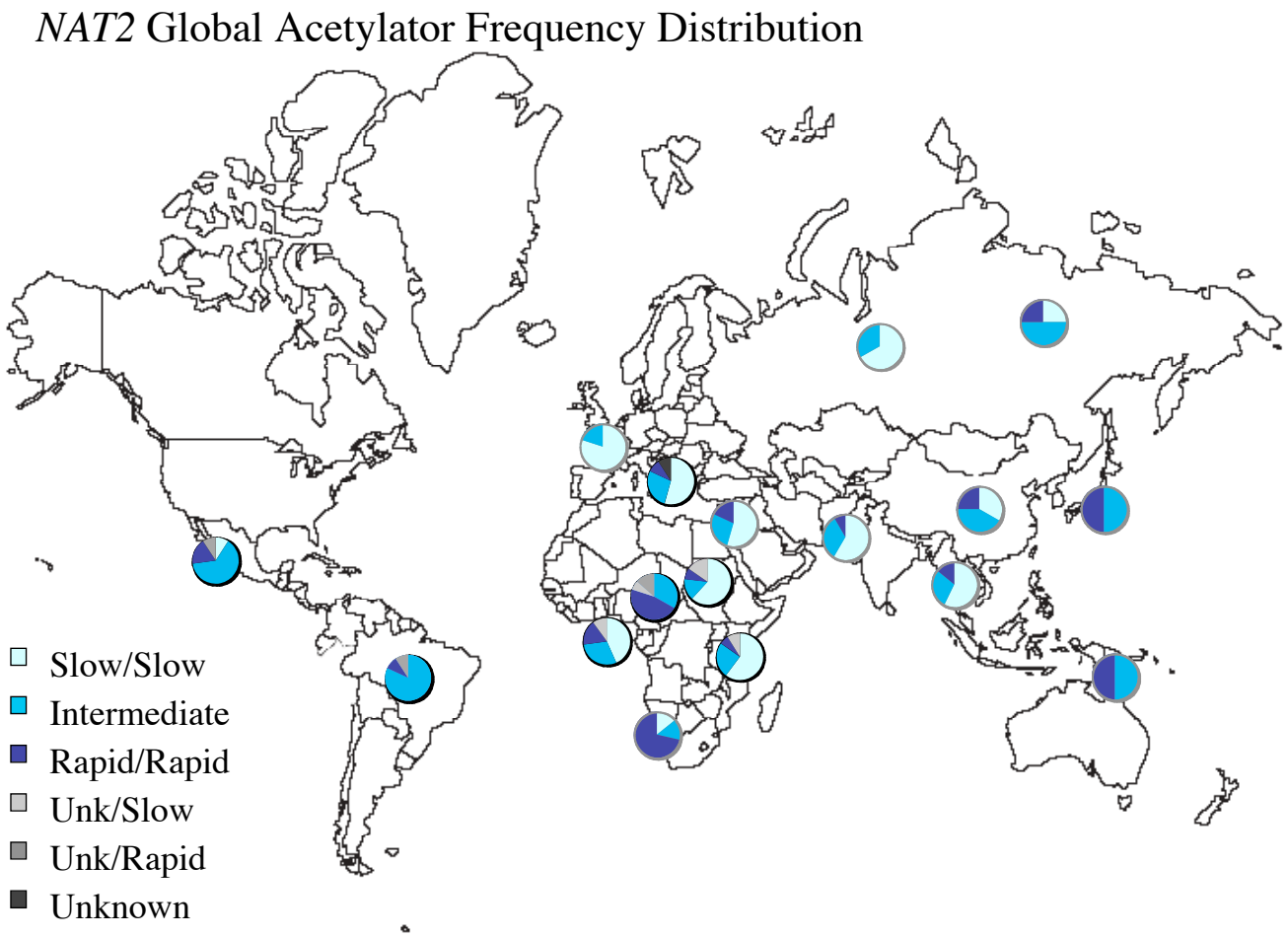


Figure 4.2

Figure 4.2: *NAT2* inferred acetylator phenotype distribution for all populations in the current study. Phenotype inference was made based on SNPs known to affect acetylator phenotype (Hein, 2003). Each individual is represented according to their inferred diploid haplotype, where light blue=slow/slow; turquoise= slow/rapid; dark blue= rapid/rapid; light grey=unknown/slow; dark grey=unknown/rapid; and black=unknown/unknown.

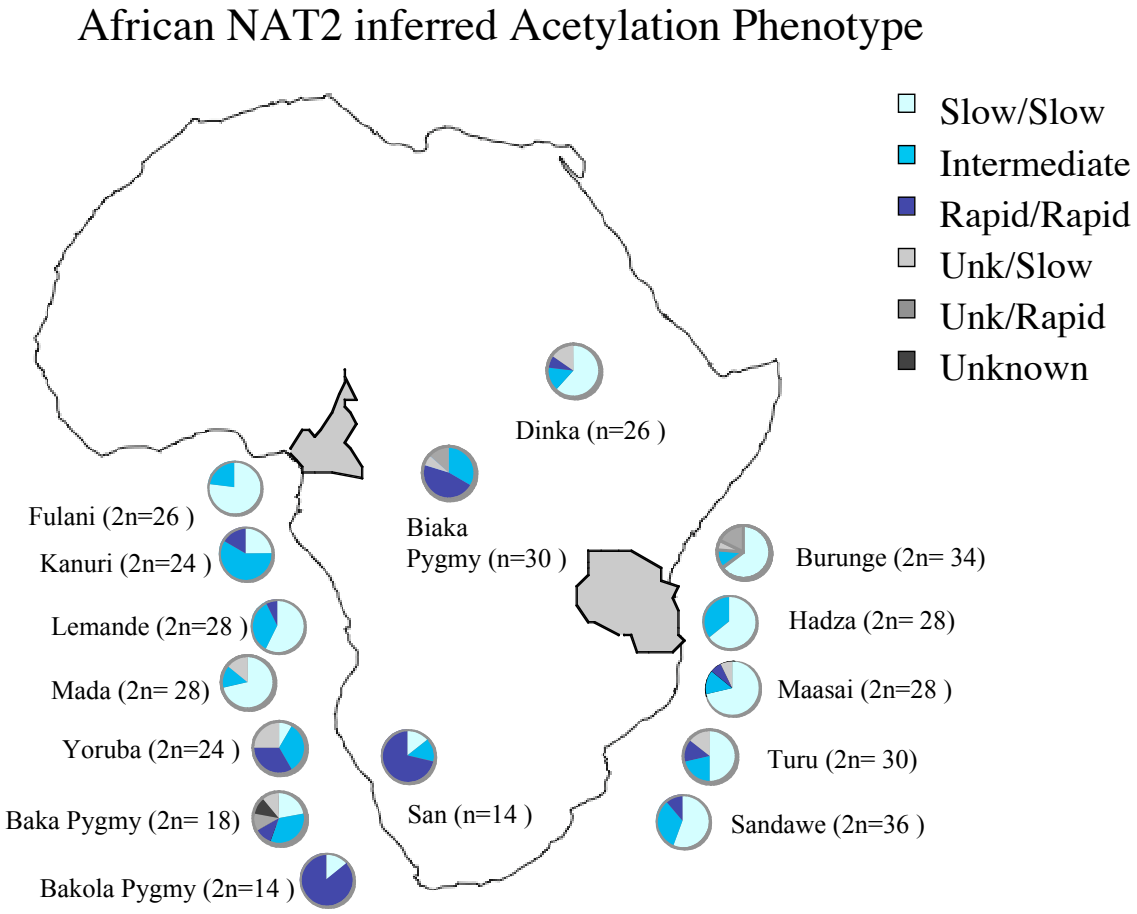


Figure 4.3

Figure 4.3: *NAT2* inferred acetylator distribution for all African populations in the current study. Phenotype inference was made based on SNPs known to affect acetylator phenotype (Hein, 2003). Each individual is represented according to their inferred diploid haplotype, where light blue=slow/slow; turquoise= slow/rapid; dark blue= rapid/rapid; light grey=unknown/slow; dark grey=unknown/rapid; and black=unknown/unknown.

	N	2N	S	s	r	H	HD	*Si*	π (∗10³)	θω(∗10³)	TD	P(TD<obs)	D*	P(D*<obs)	F*	P(F*<obs)	D	P(D<obs)	F	P(F<obs)	H	P(H<obs)
NAT2																						
Africa	194	388	45	3	15	68	.914	9	1.860	2.320	-0.556	0.571	-0.647	0.294	-0.730	0.254	-0.669	0.274	-0.745	0.245	0.822	0.394
E Africa	91	182	38	2	12	40	0.875	9	1.940	2.230	-0.382	0.4165	-0.705	0.273	-0.683	0.257	-0.751	0.242	-0.713	0.269	0.714	0.393
W Africa	53	106	27	3	9	27	.875	11	1.710	1.700	0.023	0.6698	-2.112	0.025	-1.536	0.075	-2.261	0.034	-1.622	0.068	0.013	0.322
Pygmy groups	31	62	23	2	8	24	0.952	4	1.620	1.620	0.014	0.582	0.347	0.559	0.270	0.632	0.347	0.717	0.267	0.644	0.167	0.345
Europe	46	92	21	2	5	26	0.887	3	1.730	1.360	0.797	0.830	0.493	0.611	0.724	0.794	0.502	0.777	0.740	0.813	0.323	0.356
Asia	31	62	11	2	3	15	0.798	0	0.930	0.770	0.586	0.779	1.437	1.000	1.358	0.934	1.492	0.915	1.406	0.943	0.876	0.611
Americas	14	28	10	2	3	9	0.799	0	1.040	0.850	0.717	0.802	1.406	1.000	1.398	0.946	1.503	0.920	1.498	0.949	-0.529	0.230
Africa																						
Baka Pygmy	9	18	17	2	6	12	.935	4	1.590	1.630	-0.092	0.515	0.363	0.573	0.269	0.599	0.333	0.561	0.242	0.602	-0.941	0.236
Bakola Pygmy	7	14	15	1	3	8	.857	2	1.660	1.560	0.283	0.662	0.913	0.815	0.851	0.814	0.990	0.813	0.944	0.825	0.088	0.330
Biaka Pygmy	15	30	20	2	7	16	0.945	7	1.550	1.670	-0.238	0.466	-0.554	0.225	-0.533	0.266	-0.699	0.212	-0.654	0.292	-0.248	0.301
Burunge	17	34	21	2	8	11	0.838	6	1.960	1.690	0.534	0.762	-0.222	0.321	0.029	0.513	0.048	0.733	0.263	0.613	0.299	0.347
Dinka	13	26	21	2	7	10	0.818	9	1.950	1.820	0.261	0.658	-0.930	0.151	-0.656	0.225	-1.159	0.184	-0.825	0.247	0.603	0.395
Fulani	13	26	15	2	4	8	0.757	4	1.750	1.310	1.176	0.908	0.051	0.575	0.460	0.720	0.413	0.739	0.797	0.796	-0.637	0.236
Hadzabe	14	28	20	2	5	12	0.844	6	1.610	1.700	-0.179	0.484	-0.204	0.462	-0.230	0.376	0.049	0.450	-0.034	0.506	-2.392	0.149
Kanuri	12	24	14	2	4	12	0.924	2	1.360	1.240	0.343	0.697	0.738	0.717	0.722	0.805	0.775	0.709	0.766	0.781	0.754	0.449
Lemande	14	28	20	3	6	16	0.921	6	1.840	1.700	0.310	0.693	-0.204	0.456	-0.050	0.453	0.049	0.458	0.166	0.583	0.291	0.350
Maasai	14	28	20	2	6	15	0.929	6	1.930	1.700	0.482	0.734	-0.204	0.460	0.014	0.485	-0.302	0.457	-0.045	0.496	0.275	0.354
Mada	14	28	19	2	7	10	0.828	8	1.730	1.610	0.256	0.657	-0.932	0.210	-0.656	0.238	-0.774	0.299	-0.516	0.350	-0.487	0.270
S African San	7	14	14	2	3	7	0.824	6	1.310	1.450	-0.400	0.386	-0.397	0.279	-0.455	0.298	-0.230	0.496	-0.339	0.410	-2.066	0.140
Sandawe	18	36	21	2	6	13	0.838	4	1.980	1.670	0.630	0.796	0.387	0.587	0.551	0.764	0.385	0.581	0.569	0.733	1.041	0.468
Turu	15	30	25	2	8	17	0.915	11	1.900	2.170	-0.443	0.364	-1.053	0.144	-1.008	0.168	-1.001	0.195	-0.978	0.205	0.680	0.392
Yoruba	12	24	17	2	7	15	0.946	3	1.670	1.500	0.410	0.723	0.576	0.661	0.613	0.748	0.594	0.802	0.644	0.743	1.797	0.690
Europe																						
French	8	16	12	2	3	9	0.883	1	1.760	1.190	1.814	0.978	1.082	0.856	1.478	0.963	1.177	0.961	1.647	0.964	1.600	0.735
Druze	10	20	14	2	4	11	0.916	1	1.720	1.300	1.178	0.914	1.140	0.965	1.336	0.942	1.247	0.879	1.477	0.939	1.905	0.787
Sardinian	11	22	13	2	3	8	0.831	2	1.680	1.180	1.523	0.951	0.690	0.856	1.089	0.886	0.718	0.849	1.170	0.890	2.199	0.934
Brahui	11	22	16	2	3	11	0.9	3	1.750	1.450	0.749	0.822	0.535	0.653	0.698	0.771	0.544	0.778	0.736	0.777	0.589	0.408
Russian	6	12	13	2	4	8	0.909	2	1.860	1.420	1.310	0.929	0.852	0.776	1.104	0.891	0.904	0.889	1.232	0.890	1.455	0.635
Asia																						
Cambodian	7	14	7	1	2	7	0.846	0	1.040	0.730	1.597	0.956	1.331	1.0	1.601	0.972	1.430	1.000	1.763	0.969	0.000	0.335
Han	10	20	9	2	3	7	0.753	2	0.810	0.840	-0.100	0.513	0.346	0.514	0.253	0.585	0.315	0.503	0.227	0.579	0.295	0.402
Japanese	9	18	8	1	2	5	0.732	2	0.740	0.770	-0.116	0.510	0.253	0.700	0.173	0.543	0.203	0.466	0.130	0.547	0.261	0.392
Papuan	2	4	3	1	1	2	0.5	3	0.490	0.540	-0.754	0.550	-0.754	0.535	-0.675	0.000	-1.380	0.247	-1.466	0.237	1.000	1.000
Yakut	3	6	9	2	2	5	0.933	4	1.410	1.300	0.491	0.681	0.249	0.658	0.324	0.635	-0.026	0.468	0.111	0.622	1.333	0.628
Americas																						
Karitiana	10	20	10	2	3	9	0.853	0	1.130	0.930	0.767	0.819	1.410	0.931	1.419	0.951	1.533	0.931	1.557	0.951	0.032	0.330
Pima	4	8	5	1	2	3	0.679	2	0.700	0.640	0.420	0.655	0.127	0.651	0.214	0.586	-0.053	0.630	0.072	0.569	1.071	0.810

All values were calculated using DNAsp (Rosas, 1995)

p<0.05 α' 0.008

Table 4.2

Table 4.2: Summary statistics and estimates of neutrality for the *NAT2* locus. All calculations were preformed using DNAsp version 4.20.2 (Rozas and Rozas; Rozas et al.). All abbreviations are as follows: S= number of segregating sites; 2N=number of chromosomes included for analysis; *s*=silent or synonymous variants identified; *r*=replacement or non-synonymous variants identified; H= number of haplotypes identified; HD= haplotype diversity; *Si*=number of singleton mutations; π=nucleotide diversity; $\theta\omega$=Waterson's theta estimator; TD=Tajima's D statistic (Tajima, 1989); D, D* and F, F* equal Fu and Li's test statistics at the inter-specific and intra-specific levels (Fu and Li, 1993), respectively; H=Fay and Wu's H statistic (Fay and Wu, 2000). P values, indicated by P(estimator<obs), indicate significance levels assessed for all neutrality estimates using the coalescent simulator within DNAsp (10,000 replicates), assuming no recombination. Following Bonferroni correction for multiple testing, we do not observe significance for any estimator.

NAT2 Inferred Functional Haplotypes		BG	BK	DN	FU	HZ	KN	LM	MD	MS	PB	PL	SA	SW	TR	YO	DZ	FR	PA	RU	SR	CB	HA	JP	NG	SB	BR	PI
NAT2*4	**R**	2	3	2	.	2	6	2	2	1	1	1	5	4	4	2	5	2	5	.	6	4	11	14	3	4	10	9
NAT2*5A	**S**	.	.	.	.	.	.	1	.	1	.	.	.	.	1	2	2	4	4	2	3	.	1	.	.	1	1	.
NAT2*5B	**S**	9	4	10	13	4	5	8	13	8	2	.	2	8	10	.	8	7	2	5	5	.	.	.	.	.	8	9
NAT2*5C	**S**	1	.	.	.	.	.	1	1	1	2	.	.	1	1	.	.	.	3	.	.	.	.	.	.	.	.	.
NAT2*6A	**S**	10	1	8	9	16	3	3	6	9	2	.	.	13	4	3	5	5	7	3	8	3	5	4	.	3	.	.
NAT2*6B	**S**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.
NAT2*6C	**S**	.	.	.	.	.	1	.	1	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*6G	**U**	.	.	.	.	.	.	.	.	.	1	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*6H	**U**	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*6I	**U**	1	.	.	.	1	.	2	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*6m	**U**	.	.	.	.	.	.	.	.	.	.	.	.	2	1	.	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*6n	**U**	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*7A	**S**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.
NAT2*7B	**S**	5	.	.	.	2	.	.	1	.	.	.	1	1	2	.	.	.	1	1	.	5	5	1	1	.	.	.
NAT2*12A	**R**	1	14	2	1	2	5	2	.	3	2	11	6	5	1	5	.	.	.	.	2	.	.	.	.	.	1	3
NAT2*12B	**R**	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.
NAT2*12E	**U**	3	2	.	.	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*12G	**U**	.	1	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*12H	**U**	.	.	1	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*12i	**U**	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*12j	**U**	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*13A	**R**	.	2	.	2	1	1	3	.	.	4	.	.	1	3	5	.	.	.	1	.	1	2	1	.	.	1	1
NAT2*13c	**U**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.
NAT2*14B	**S**	1	1	1	1	.	3	4	2	2	.	.	.	.	1	2	.	.	.	.	.	.	.	.	.	.	.	.
NAT2*14j	**U**	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Unknown other	**U**	1	.	.	.	.	.	.	1	1	.	.	.	.	1	2	1	.	.	.	.	.	.	.	.	.	.	.
Total Haplotpyes		34	30	26	26	28	24	28	28	28	18	14	14	36	30	24	22	18	24	12	24	14	24	20	4	8	22	22

present study
Patin [, 2006a;2006b]
[Sabbagh, 2008]

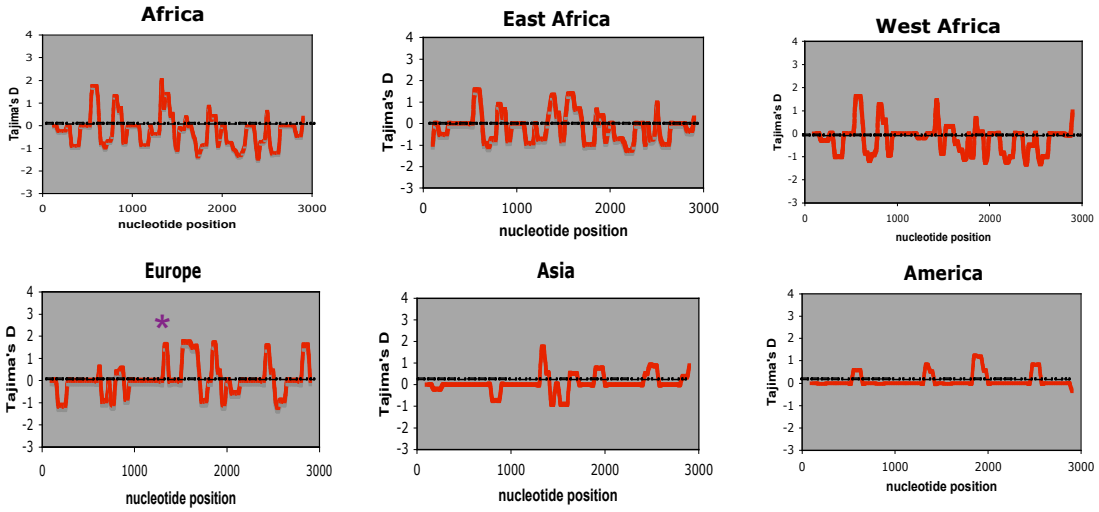
R denotes Rapid
S denotes Slow
U denotes undetermined

Table 4.3

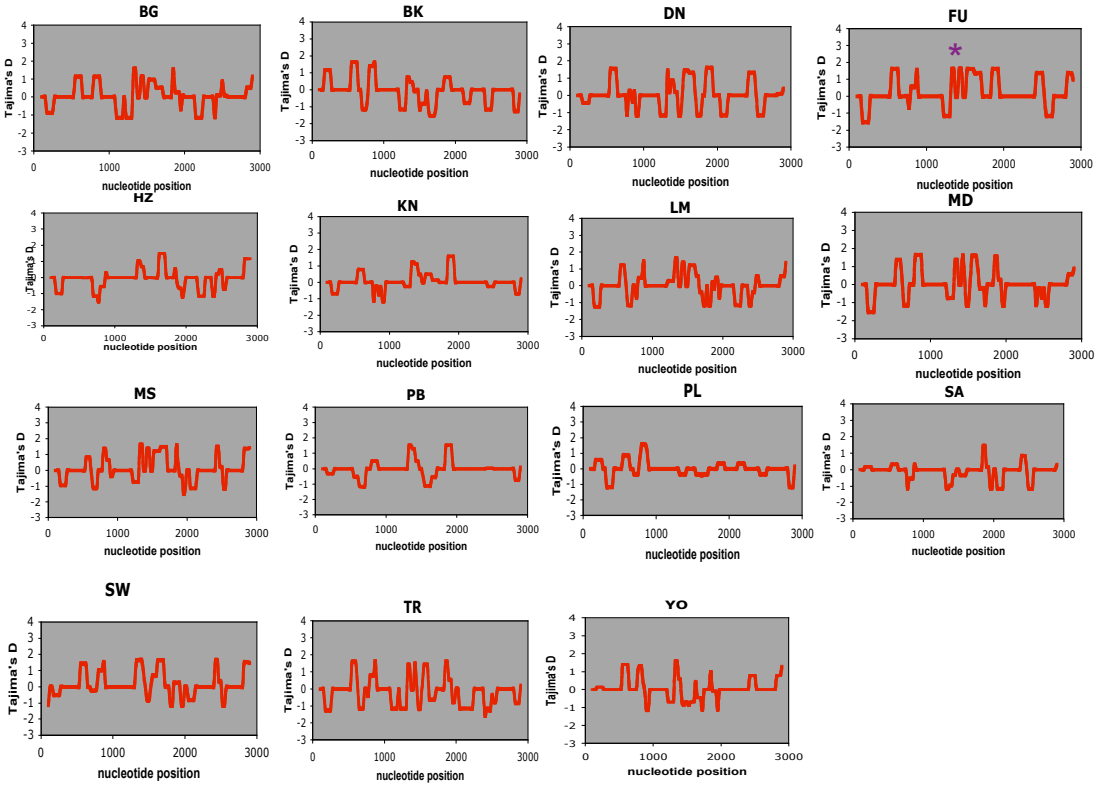
Table 4.3: Inferred *NAT2* acetylator haplotype counts for all populations in the current study. Population abbreviations follow Table 2.1. Novel haplotypes to the current study are designating in green. Those haplotypes novel to the studies of Patin *et al*. (2006a,b) are indicated in blue, and Sabbagh *et al*. (2008) indicated in yellow. R=Rapid; S=Slow, and U=Undetermined.

Several tests of neutrality were performed for the *NAT2* loci, at both the intra and inter-species levels (Table 4.2). Overall, tests of selective neutrality based on the allelic frequency spectrum showed both positive and negative values in both African and non-African groups. Significance was obtained in only two cases for D and D* in the West African grouping ($p<0.05$). However, these estimates were not significant after Bonferonni correction. Each pooled African group (Africa, East Africa, and West Africa) has negative values for TD, D, D*, F and F*, with the exception of the Pygmy grouping (Baka, Bakola and Biaka) that are positive for all estimators (Table 4.2). However, when analyzed separately, the Biaka pygmy population has negative values for these neutrality statistics. Both African and non-African groups have both positive and negative values for these estimators, though many negative values are very close to zero. Overall, European, Asian, and Amerindian groups have positive values for tests of neutrality (with the exception of the Papuan group, where 2n=4). Positive values are observed in some African populations (Baka and Bakola pygmies, Burunge, Fulani, Kanuri, Sandawe, and Yoruba), from both East and West Africa. By contrast, Biaka, Hadza, San and Turu groups have negative values for all estimators. Sliding window analyses of TD (Figure 4.4) indicate positive and negative values at several locations across the *NAT2* region. Significant values of TD ($p<0.05$) are obtained for West Africa and Europe at the continental level, and for the Fulani of Cameroon, at one location corresponding to SNP +341 that is known to be an inactivating (slow) *NAT2* mutation.

NAT2 Continental Groups



NAT2 Africa



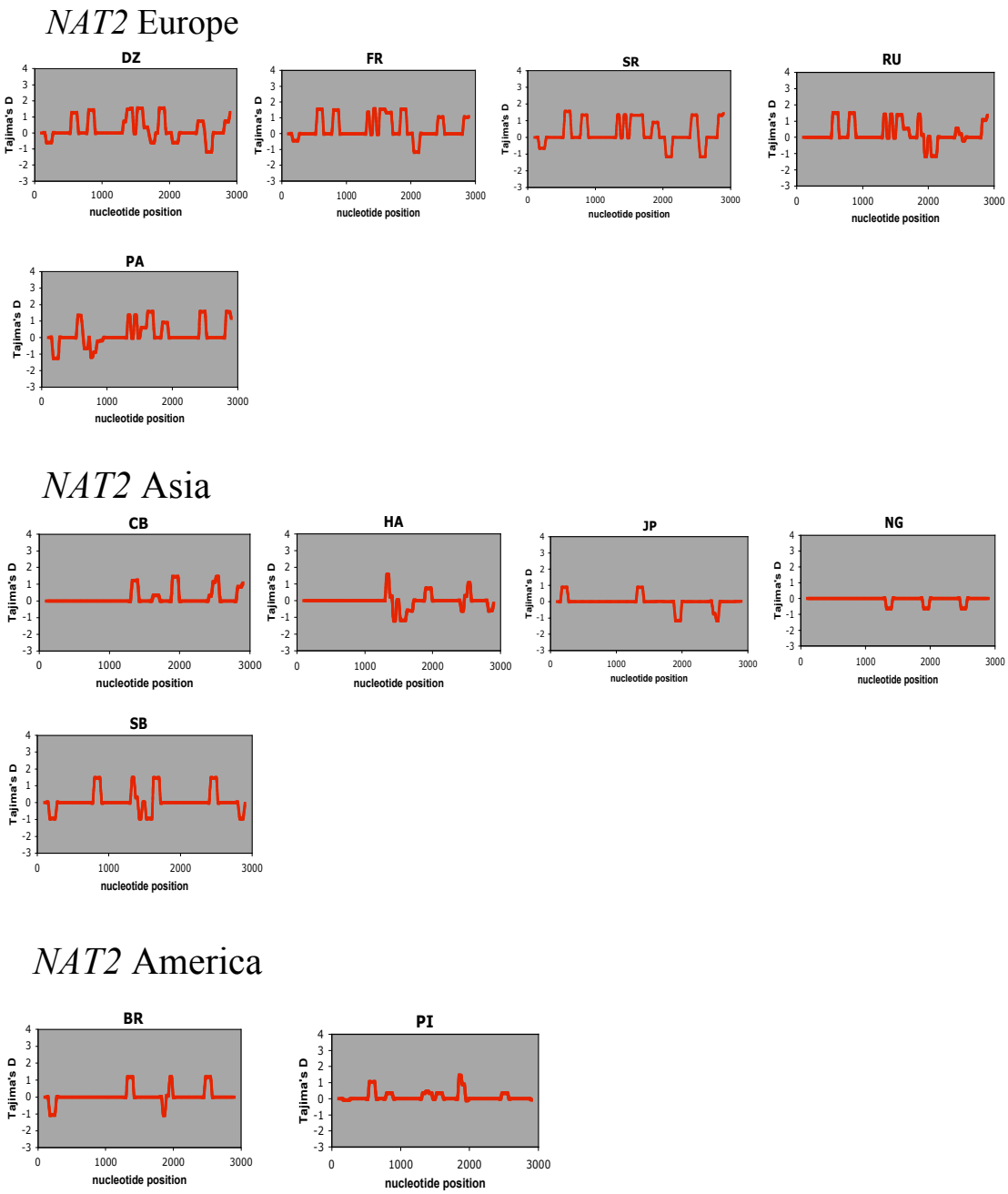


Figure 4.4

Figure 4.4: Sliding window of Tajima's D estimates across the *NAT2* region at defined window lengths of 100 sites at steps of 25 sites. Purple asterisks indicate significance at the $p<0.05$ level. Populations included in each continental grouping follow that presented in Table 2.1.

The K_a/K_s ratio for *NAT2* of synonymous to non-synonymous nucleotide divergence was observed to be higher than that observed for *NAT1*, but still less than one (K_a/K_s=0.813). The MK tests (McDonald and Kreitman, 1991) of the observed level of intra-specific variation and levels of fixed variation between humans and chimpanzee within the coding region were not statistically significant after applying Fisher's exact test (p=0.579) (Figure 4.5). These results indicate that there is no excess or depletion of variation at the *NAT2* locus between humans and chimpanzee. This finding underlines the high number of replacement changes tolerated at this loci in both humans and the chimpanzee outgroup. An HKA test comparing levels of intra- and inter-specific variation at *NAT2* relative to *NATP1* was significant with an observed p value of 0.012 (Figure 4.6).

Population Differentiation

The MDS plot for the *NAT2* locus does not reflect clustering by geographic region (Figure 4.7). Instead we observe two main clusters in the plot that correspond to rapid and slow acetylator types, where the Bakola, Biaka, San, and Hadza groups appear as outliers. The first main cluster contains African and European populations (slow); the second cluster is made up of all Asian and Amerindian groups (rapid). Particularly notable exceptions are the groups present in the center of the plot (Kanuri, Baka, Yoruba), as well as the Bakola, Biaka, and San, which all have a high proportion of intermediate and rapid inferred haplotypes. This result is consistent with the inferred frequencies of rapid and slow acetlyator haplotypes in each population, as illustrated in Figures 4.2 and 4.3. Given the observed pattern, the *NAT2* locus does not appear to

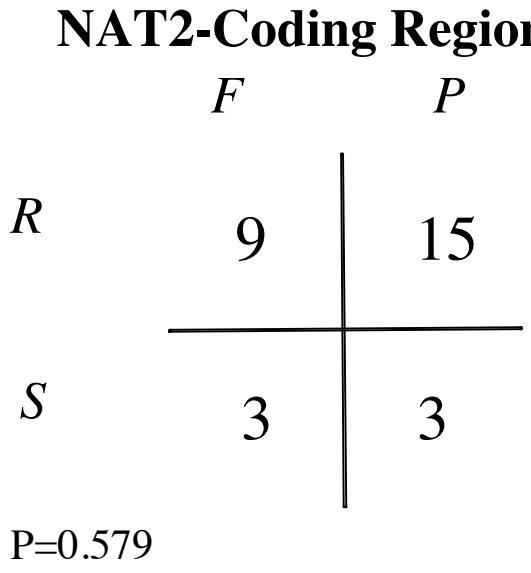


Figure 4.5

Figure 4.5: McDonald-Krietman test for the *NAT2* coding region.

	NATP1	*NAT2*
Intraspecific Variability		
Segregating Sites Obs	8	18
Segregating Sites Exp	13.76	12.24
Sample Size	564	564
Interspecies Divergence		
Number of Differences Obs	27.68	13.75
Number of Differences Exp	21.92	19.5
Total sites	870	870
chi-square value	6.287	
p-value	0.0122	

Figure 4.6

Figure 4.6: HKA tests of intra- and inter-specific variation at *NAT2*, compared to *NATP1*.

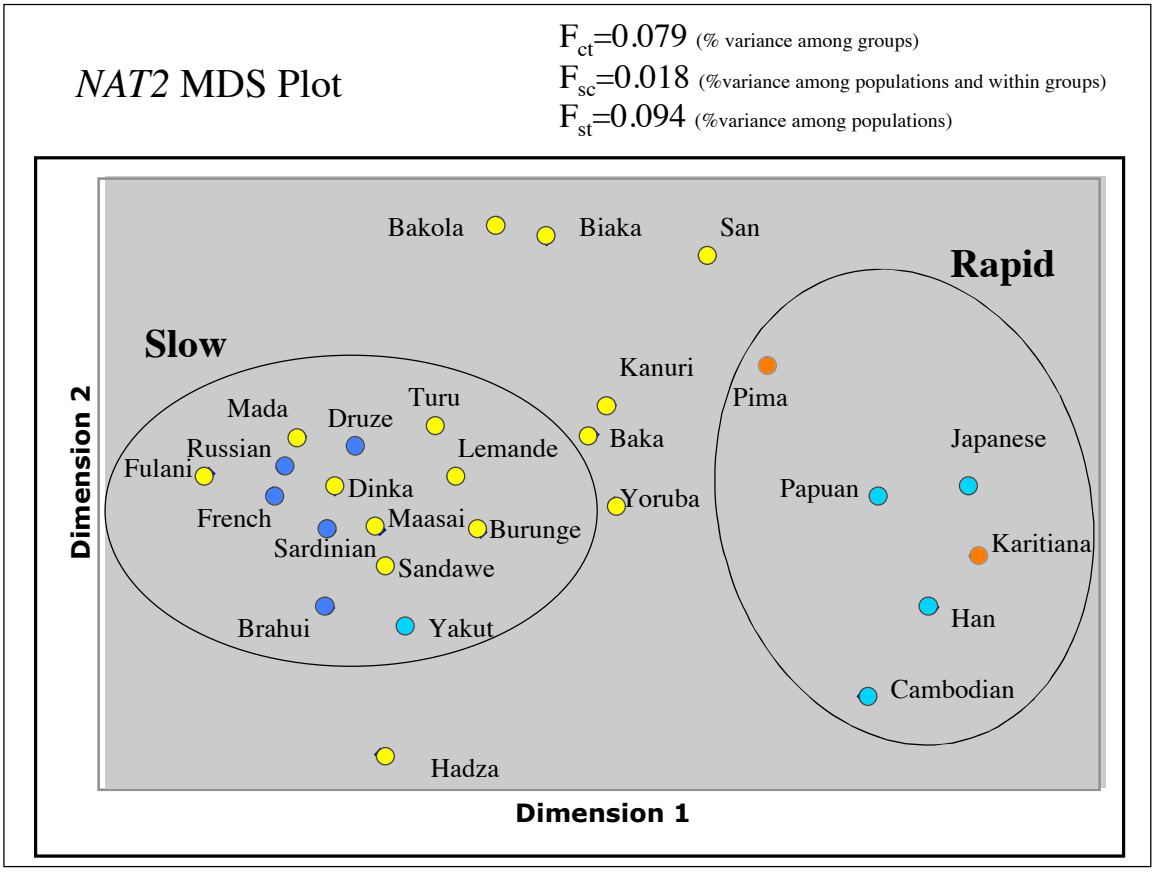


Figure 4.7

Figure 4.7: Multi-Dimensional Scaling plot for the *NAT2* locus. AMOVA results indicated in inset, where variance among groups= F_{CT}, variance among populations within groups= F_{SC}, and variance within populations= F_{ST}. Yellow=Africa; Blue=Europe; Turquoise=Asia; Orange=Americas. Population structure is specified according to the population groupings listed in Table 2.3. Populations appear to cluster based on high frequency of Rapid and Slow acetylator types, indicated with circles here. Bakola, Biaka, and San hunter-gatherers also have high frequency of rapid acetylator haplotypes (see Figure 4.3), but appear as outliers, as do the Hadza hunter-gatherers of Tanzania that have high frequencies of slow acetylator haplotypes.

correspond to demographic expectations, as would be expected under a neutral scenario. AMOVA results for *NAT2*, presented in Figure 4.7, indicate the percentage of variation among groups (7.79% of the total variance), which is slightly lower than is typically observed at other loci in humans (Campbell and Tishkoff, 2008). The percentage of variation within populations is intermediate (90.58%) when compared to other *NAT* loci, and variation among populations and within groups accounts for a small percentage of the total observed variation (1.63%). Additionally, the fixation indices for the *NAT2* locus are F_{ST}=0.094, F_{CT}=0.079, and F_{SC}=0.018.

Phylogenetic haplotype networks

The median-joining networks for the *NAT2* coding region only are illustrated in Figures 4.7a and 4.7b. Red specifies the chimpanzee outgroup, which indicates that the rapid acetylator haplotype is, in fact, ancestral in humans. In Figure 4.8a, nodes are colored according to the inferred phenotype status of haplotypes present in each node, where dark blue indicates fast acetylators, light blue indicates slow acetylator types, and grey indicates haplotypes with novel mutations that could not be classified according to the known genotype-phenotype concordance for *NAT2*. In Figure 4.8b, nodes are colored according to the proportion of haplotypes from each geographic region. Shown in both Figures 4.7a and 4.7b, lower case and italicized acetylator phenotypes designate acetylator haplotypes that could be characterized according to known acetylator mutations, but differ by SNPs that are novel to this dataset. Where classification of acetylator status was possible, rapid or slow phenotype is indicated (*e.g.* if a novel SNP was non-synonymous). Haplotypes listed as "hap" refer to the haplotype numbering

presented in Table 4.1, and differ from *4 (WT) only by previously uncharacterized SNPs, preventing addendum to previously existing nomenclature.

A third median-joining network for the entire *NAT2* region analyzed is presented in Figure 4.8c, and illustrates the frequency of haplotypes according to the major geographic divisions (Africa, Europe, Asia, and the Americas) for the entire 3031 bp resequenced region, including the coding region and 3' and 5' UTRs. Here, three nodes are observed, each representing the main haplotypes present in this *NAT2* dataset. The majority of *NAT2* haplotypes represent *NAT2*4* (WT-rapid), *NAT2*5B* (slow), and *NAT2*6A* (slow). These three nodes (4, *5B, and*6A) are indicated in Figure 4.7c and correspond to haplotypes 1, 37 and 17, respectively (refer to Table 4.1). The two derived, slow acetylator types (hap 17(*6A) and hap 37 (*5B)) are present in high frequency, and represent approximately one-third of the total dataset. Hap17 (*6A) is not observed in Amerindian individuals, Bakola Pygmies or San hunter-gatherers. This may be due to the low frequency of slow acetylators in these groups in this *NAT2* dataset. Hap37 (*5B) is not present in any Asian group, or the Bakola Pygmies and Yoruba of West Africa (see Table 4.1), however, other haplotypes that are inferred to be slow acetylator types are observed infrequently in these groups.

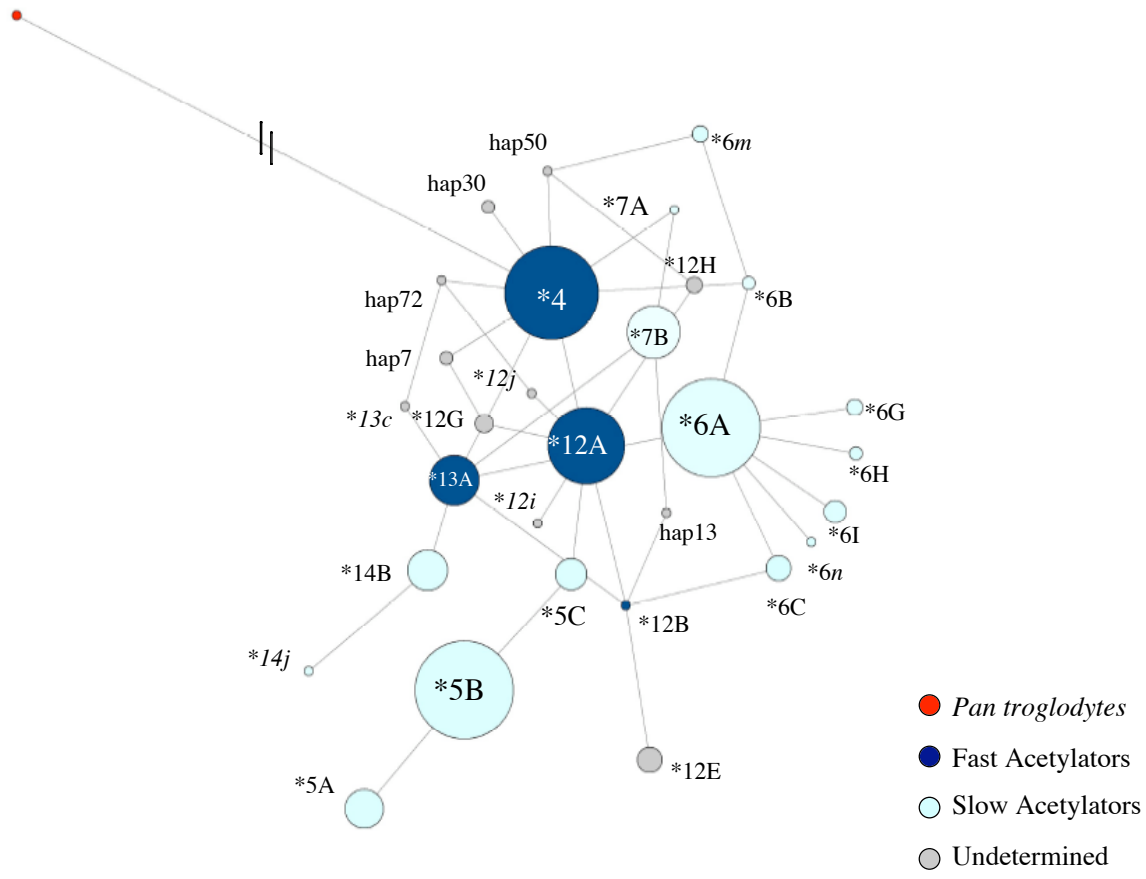


Figure 4.8a

Figure 4.8a: Median-joining network for *NAT2* inferred acetylator phenotypes, created using Network 4.5, available at fluxus-engineering.com (Bandelt, 1999). For all nodes, dark blue=Rapid acetylators, light blue=Slow acetylators, grey=haplotypes where acetylator phenotype could not be inferred, and red indicates the chimpanzee (*P. Troglodytes*) outgroup. All nomenclature follows that accepted by the NAT nomenclature committee (*e.g. NAT2**Major haplotype/sub-haplotype). Grey nodes with capitalized sub-haplotype designation represent previously described haplotypes that have not been characterized in terms of phenotype to date. Grey nodes with italicized and lower-case sub-haplotypes represent haplotypes that are novel to the current dataset, but differ by 1-2 novel SNPs, allowing for major classification based on known acetylator-defining mutations. Grey nodes bearing a "hap" label correspond to Table 4.2, and are novel haplotypes that have two SNPs that define phenotypes with conflicting acetylator status, thus prohibiting major phenotype inference.

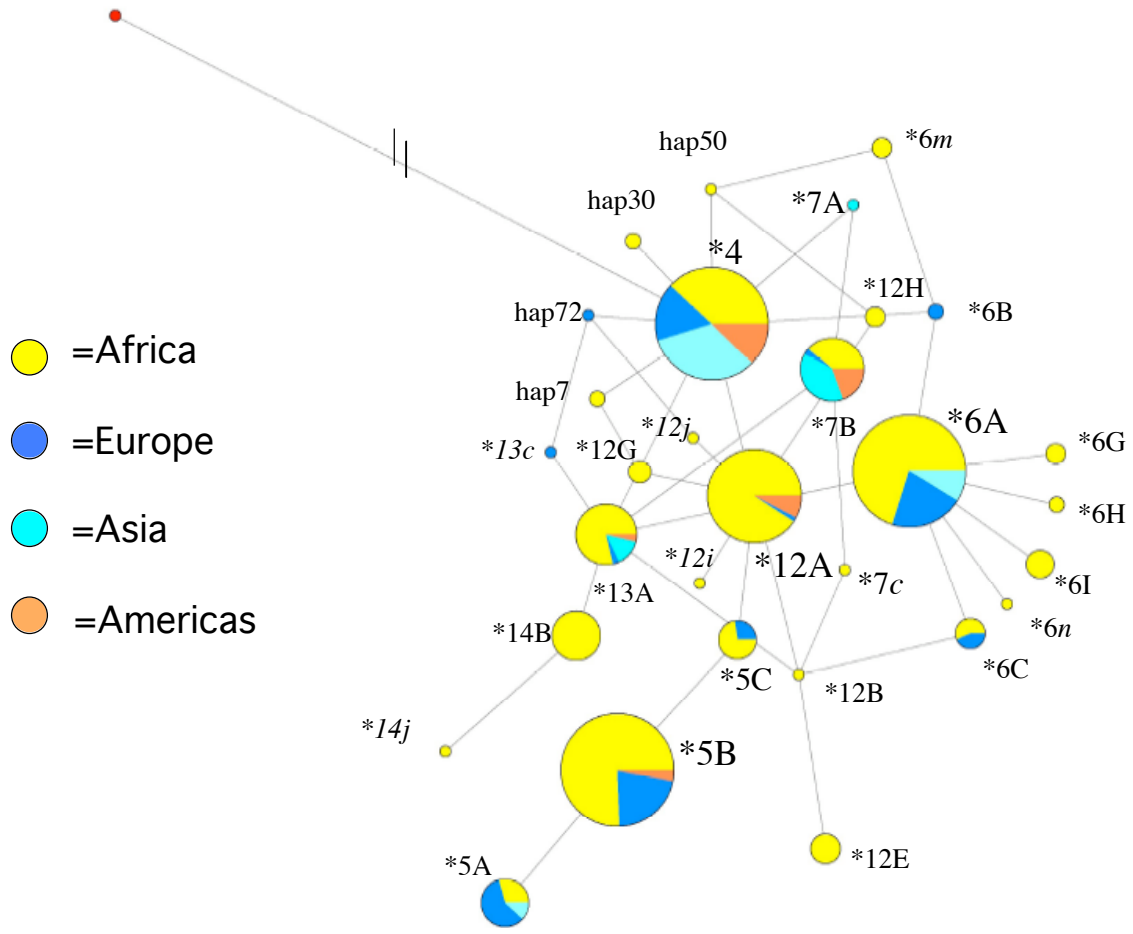


Figure 4.8b

Figure 4.8b: Median-joining haplotype network for *NAT2*. Inferred acetylator phenotypes are indicated, as in Figure 4.6a. Major geographic groups are specified according to their frequency, where Yellow=Africa; Blue=Europe; Turquoise=Asia; and Orange=Americas.

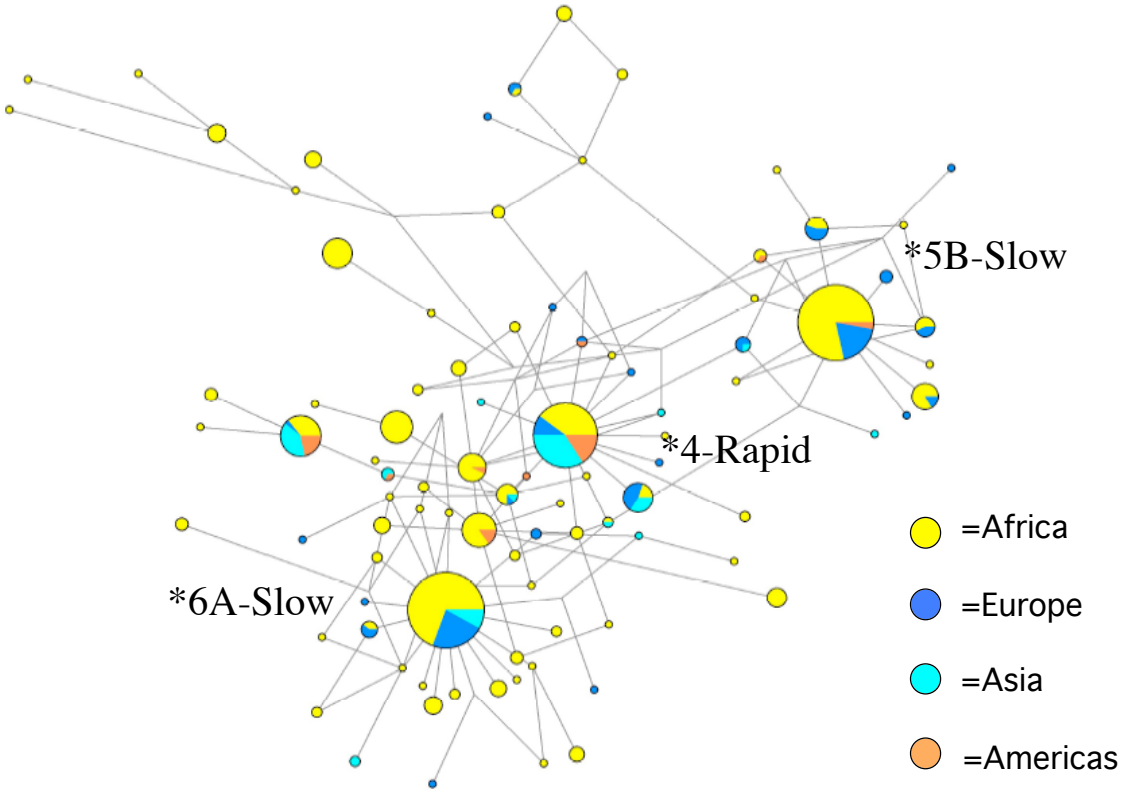


Figure 4.8c

Figure 4.8c: Median-joining network for the entire ~3kb *NAT2* region. Major geographic groups indicated, where Yellow=Africa; Blue=Europe; Turquoise=Asia; and Orange=Americas. High frequency acetylator haplotypes, Rapid *4, and Slow *5B, and *6A, in this dataset have also been indicated.

Patterns of intragenic linkage disequilibrium (LD)

One hundred distinct haplotypes were inferred from the 3031 bp *NAT2* region, indicating that recombination has affected the current pattern of diversity present at this locus (Table 4.1). The results from both the LDhat (Figure 4.9) and Haploview (Figure 4.10) analyses confirm high levels of recombination in both Africans and non-Africans. Higher than expected linkage disequilibrium exists at only a few sites within the *NAT2* region analyzed, some of which are between SNPs known to affect acetylator phenotype (*e.g.* +290, and +590). Haploview analysis indicates differing haplotype block structure across the *NAT2* region in different populations (Figure 4.10). From Figures 4.9 and 4.10 it is clear that recombination levels are high within the *NAT2* coding region in all groups.

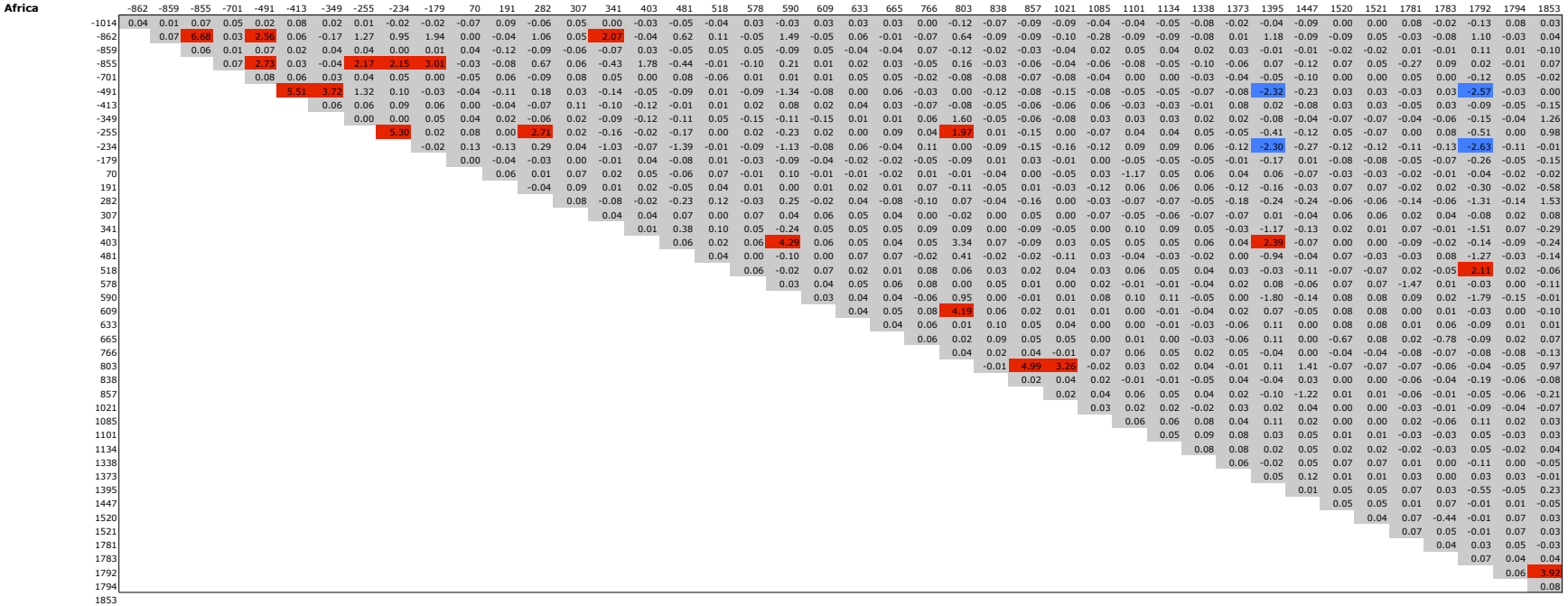
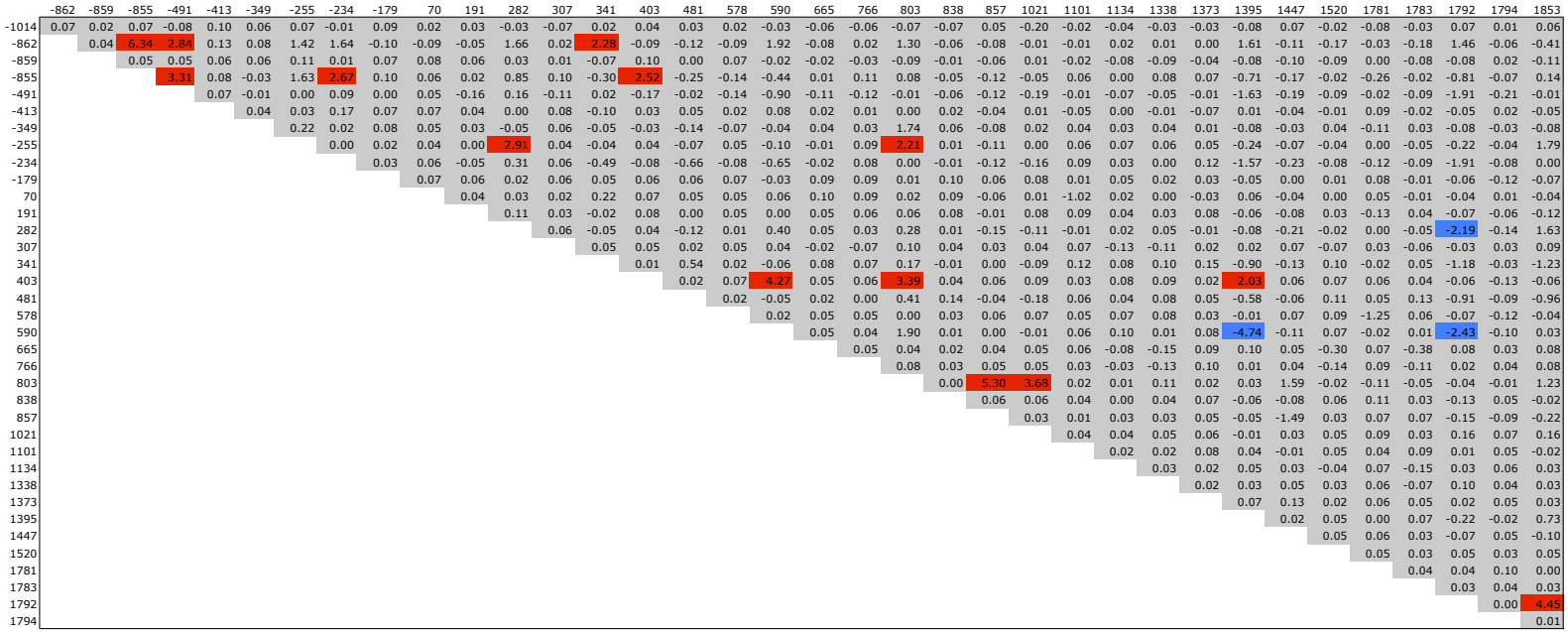


Figure 4.9a

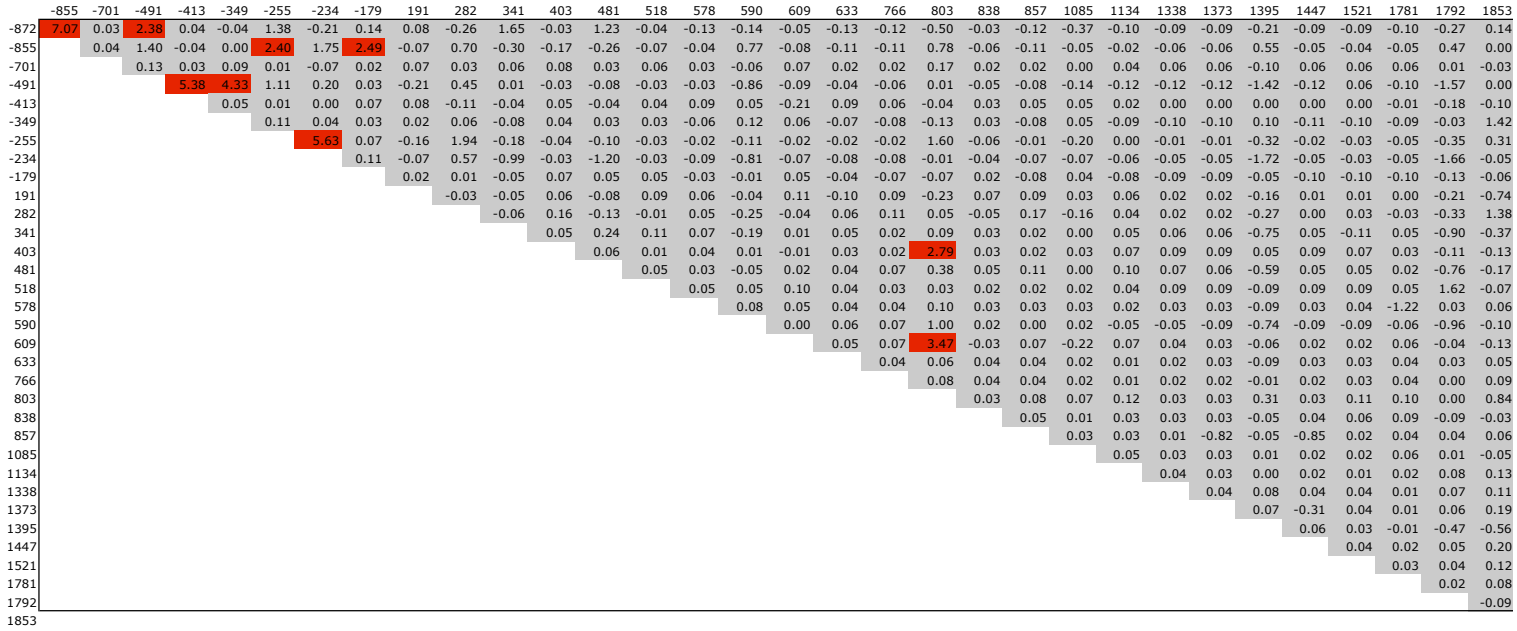
East Africa



	-862	-859	-855	-491	-413	-349	-255	-234	-179	70	191	282	307	341	403	481	578	590	665	766	803	838	857	1021	1101	1134	1338	1373	1395	1447	1520	1781	1783	1792	1794	1853
-1014	0.07	0.02	0.07	-0.08	0.10	0.06	0.07	-0.01	0.09	0.02	0.03	-0.03	-0.07	0.02	0.04	0.03	0.02	-0.03	-0.06	-0.06	-0.07	-0.07	0.05	-0.20	-0.02	-0.04	-0.03	-0.03	-0.08	0.07	-0.02	-0.08	-0.03	0.07	0.01	0.06
-862		0.04	6.34	2.84	0.13	0.08	1.42	1.64	-0.10	-0.09	-0.05	1.66	0.02	2.28	-0.09	-0.12	-0.09	1.92	-0.08	0.02	1.30	-0.06	-0.08	-0.01	-0.01	0.02	0.01	0.00	1.61	-0.11	-0.17	-0.03	-0.18	1.46	-0.06	-0.41
-859			0.05	0.05	0.06	0.06	0.11	0.01	0.07	0.08	0.06	0.03	0.01	-0.07	0.10	0.00	0.07	-0.02	-0.02	-0.03	-0.09	-0.01	-0.06	0.01	-0.02	-0.08	-0.09	-0.04	-0.08	-0.10	-0.09	0.00	-0.08	-0.08	0.02	-0.11
-855				3.31	0.08	-0.03	1.63	2.67	0.10	0.06	0.02	0.85	0.10	-0.30	2.52	-0.25	-0.14	-0.44	0.01	0.11	0.08	-0.05	-0.12	-0.05	0.06	0.00	0.08	0.07	-0.71	-0.17	-0.02	-0.26	-0.02	-0.81	-0.07	0.14
-491					0.07	-0.01	0.00	0.09	0.00	0.05	-0.16	0.16	-0.11	0.02	-0.17	-0.02	-0.14	-0.90	-0.11	-0.12	-0.01	-0.06	-0.12	-0.19	-0.01	-0.07	-0.05	-0.01	-1.63	-0.19	-0.09	-0.02	-0.09	-1.91	-0.21	-0.01
-413						0.04	0.03	0.17	0.07	0.07	0.04	0.00	0.08	-0.10	0.03	0.05	0.02	0.08	0.02	0.01	0.00	0.02	-0.04	0.01	-0.05	0.00	-0.01	-0.07	0.01	-0.04	-0.01	0.09	-0.02	-0.05	0.02	-0.05
-349							0.22	0.02	0.08	0.05	0.03	-0.05	0.06	-0.05	-0.03	-0.14	-0.07	-0.04	0.04	0.03	1.74	0.06	-0.08	0.02	0.04	0.03	0.04	0.01	-0.08	-0.03	0.04	-0.11	0.03	-0.08	-0.03	-0.08
-255								0.00	0.02	0.04	0.00	2.91	0.04	-0.04	0.04	-0.07	0.05	-0.10	-0.01	0.09	2.21	0.01	-0.11	0.00	0.06	0.07	0.06	0.05	-0.24	-0.07	-0.04	0.00	-0.05	-0.22	-0.04	1.79
-234									0.03	0.06	-0.05	0.31	0.06	-0.49	-0.08	-0.66	-0.08	-0.65	-0.02	0.08	0.00	-0.01	-0.12	-0.16	0.09	0.03	0.00	0.12	-1.57	-0.23	-0.08	-0.12	-0.09	-1.91	-0.08	0.00
-179										0.07	0.06	0.02	0.06	0.05	0.06	0.06	0.07	-0.03	0.09	0.09	0.01	0.10	0.06	0.08	0.01	0.05	0.02	0.03	-0.05	0.00	0.01	0.08	-0.01	-0.06	-0.12	-0.07
70											0.04	0.03	0.02	0.22	0.07	0.05	0.05	0.06	0.10	0.09	0.02	0.09	-0.06	0.01	-1.02	0.02	0.00	-0.03	0.06	-0.04	0.00	0.05	-0.01	-0.04	0.01	-0.04
191												0.11	0.03	-0.02	0.08	0.00	0.05	0.00	0.05	0.06	0.06	0.08	-0.01	0.08	0.09	0.04	0.03	0.08	-0.06	-0.08	0.03	-0.13	0.04	-0.07	-0.06	-0.12
282													0.06	-0.05	0.04	-0.12	0.01	0.40	0.05	0.03	0.28	0.01	-0.15	-0.11	-0.01	0.02	0.05	-0.01	-0.08	-0.21	-0.02	0.00	-0.05	-2.19	-0.14	1.63
307														0.05	0.05	0.02	0.05	0.04	-0.02	-0.07	0.10	0.04	0.03	0.04	0.07	-0.13	-0.11	0.02	0.02	0.07	-0.07	0.03	-0.06	-0.03	0.03	0.09
341															0.01	0.54	0.02	-0.06	0.08	0.07	0.17	-0.01	0.00	-0.09	0.12	0.08	0.10	0.15	-0.90	-0.13	0.10	-0.02	0.05	-1.18	-0.03	-1.23
403																0.02	0.07	4.27	0.05	0.06	3.39	0.04	0.06	0.09	0.03	0.08	0.09	0.02	2.03	0.06	0.07	0.06	0.04	-0.06	-0.13	-0.06
481																	0.02	-0.05	0.02	0.00	0.41	0.14	-0.04	-0.18	0.06	0.04	0.08	0.05	-0.58	-0.06	0.11	0.05	0.13	-0.91	-0.09	-0.96
578																		0.02	0.05	0.05	0.00	0.03	0.06	0.07	0.05	0.07	0.08	0.03	-0.01	0.07	0.09	-1.25	0.06	-0.07	-0.12	-0.04
590																			0.05	0.04	1.90	0.01	0.00	-0.01	0.06	0.10	0.01	0.08	-4.74	-0.11	0.07	-0.02	0.01	-2.43	-0.10	0.03
665																				0.05	0.04	0.02	0.04	0.05	0.06	-0.08	-0.15	0.09	0.10	0.05	-0.30	0.07	-0.38	0.08	0.03	0.08
766																					0.08	0.03	0.05	0.05	0.03	-0.03	-0.13	0.10	0.01	0.04	-0.14	0.09	-0.11	0.02	0.04	0.08
803																						0.00	5.30	3.68	0.02	0.01	0.11	0.02	0.03	1.59	-0.02	-0.11	-0.05	-0.04	-0.01	1.23
838																							0.06	0.06	0.04	0.00	0.04	0.07	-0.06	-0.08	0.06	0.11	0.03	-0.13	0.05	-0.02
857																								0.03	0.01	0.03	0.03	0.05	-0.05	-1.49	0.03	0.07	0.07	-0.15	-0.09	-0.22
1021																									0.04	0.04	0.05	0.06	-0.01	0.03	0.05	0.09	0.03	0.16	0.07	0.16
1101																										0.02	0.02	0.08	0.04	-0.01	0.05	0.04	0.09	0.01	0.05	-0.02
1134																											0.03	0.02	0.05	0.03	-0.04	0.07	-0.15	0.03	0.06	0.03
1338																												0.02	0.03	0.05	0.03	0.06	-0.07	0.10	0.04	0.03
1373																													0.07	0.13	0.02	0.06	0.05	0.02	0.05	0.03
1395																														0.02	0.05	0.00	0.07	-0.22	-0.02	0.73
1447																															0.05	0.06	0.03	-0.07	0.05	-0.10
1520																																0.05	0.03	0.05	0.03	0.05
1781																																	0.04	0.04	0.10	0.00
1783																																		0.03	0.04	0.03
1792																																			0.00	4.45
1794																																				0.01

Figure 4.9b

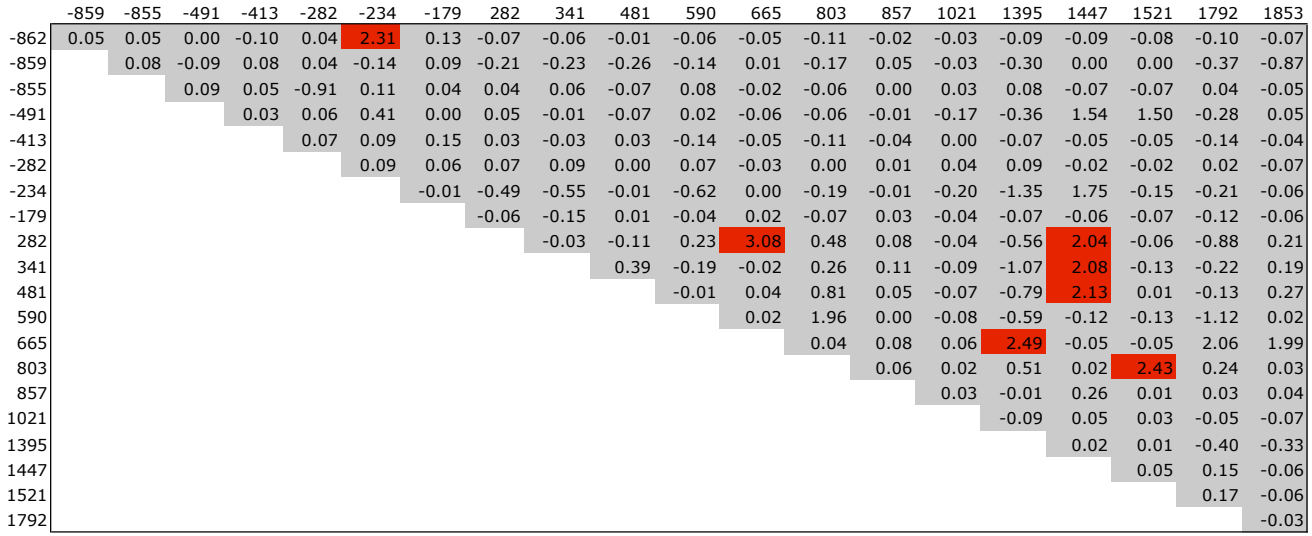
West Africa



	-855	-701	-491	-413	-349	-255	-234	-179	191	282	341	403	481	518	578	590	609	633	766	803	838	857	1085	1134	1338	1373	1395	1447	1521	1781	1792	1853
-872	7.07	0.03	2.38	0.04	-0.04	1.38	-0.21	0.14	0.08	-0.26	1.65	-0.03	1.23	-0.04	-0.13	-0.14	-0.05	-0.13	-0.12	-0.50	-0.03	-0.12	-0.37	-0.10	-0.09	-0.09	-0.21	-0.09	-0.09	-0.10	-0.27	0.14
-855		0.04	1.40	-0.04	0.00	2.40	1.75	2.49	-0.07	0.70	-0.30	-0.17	-0.26	-0.07	-0.04	0.77	-0.08	-0.11	-0.11	0.78	-0.06	-0.11	-0.05	-0.02	-0.06	-0.06	0.55	-0.05	-0.04	-0.05	0.47	0.00
-701			0.13	0.03	0.09	0.01	-0.07	0.02	0.07	0.03	0.06	0.08	0.03	0.06	0.03	-0.06	0.07	0.02	0.02	0.17	0.02	0.02	0.00	0.04	0.06	0.06	-0.10	0.06	0.06	0.06	0.01	-0.03
-491				5.38	4.33	1.11	0.20	0.03	-0.21	0.45	0.01	-0.03	-0.08	-0.03	-0.03	-0.86	-0.09	-0.04	-0.06	0.01	-0.05	-0.08	-0.14	-0.12	-0.12	-0.12	-1.42	-0.12	0.06	-0.10	-1.57	0.00
-413					0.05	0.01	0.00	0.07	0.08	-0.11	-0.04	0.05	-0.04	0.04	0.09	0.05	-0.21	0.09	0.06	-0.04	0.03	0.05	0.05	0.02	0.00	0.00	0.00	0.00	0.00	-0.01	-0.18	-0.10
-349						0.11	0.04	0.03	0.02	0.06	-0.08	0.04	0.03	0.03	-0.06	0.12	0.06	-0.07	-0.08	-0.13	0.03	-0.08	0.05	-0.09	-0.10	-0.10	0.10	-0.11	-0.10	-0.09	-0.03	1.42
-255							5.63	0.07	-0.16	1.94	-0.18	-0.04	-0.10	-0.03	-0.02	-0.11	-0.02	-0.02	-0.02	1.60	-0.06	-0.01	-0.20	0.00	-0.01	-0.01	-0.32	-0.02	-0.03	-0.05	-0.35	0.31
-234								0.11	-0.07	0.57	-0.99	-0.03	-1.20	-0.03	-0.09	-0.81	-0.07	-0.08	-0.08	-0.01	-0.04	-0.07	-0.07	-0.06	-0.05	-0.05	-1.72	-0.05	-0.03	-0.05	-1.66	-0.05
-179									0.02	0.01	-0.05	0.07	0.05	0.05	-0.03	-0.01	0.05	-0.04	-0.07	-0.07	0.02	-0.08	0.04	-0.08	-0.09	-0.09	-0.05	-0.10	-0.10	-0.10	-0.13	-0.06
191										-0.03	-0.05	0.06	-0.08	0.09	0.06	-0.04	0.11	-0.10	0.09	-0.23	0.07	0.09	0.03	0.06	0.02	0.02	-0.16	0.01	0.01	0.00	-0.21	-0.74
282											-0.06	0.16	-0.13	-0.01	0.05	-0.25	-0.04	0.06	0.11	0.05	-0.05	0.17	-0.16	0.04	0.02	0.02	-0.27	0.00	0.03	-0.03	-0.33	1.38
341												0.05	0.24	0.11	0.07	-0.19	0.01	0.05	0.02	0.09	0.03	0.02	0.00	0.05	0.06	0.06	-0.75	0.05	-0.11	0.05	-0.90	-0.37
403													0.06	0.01	0.04	0.01	-0.01	0.03	0.02	2.79	0.03	0.02	0.03	0.07	0.09	0.09	0.05	0.09	0.07	0.03	-0.11	-0.13
481														0.05	0.03	-0.05	0.02	0.04	0.07	0.38	0.05	0.11	0.00	0.10	0.07	0.06	-0.59	0.05	0.05	0.02	-0.76	-0.17
518															0.05	0.05	0.10	0.04	0.03	0.03	0.02	0.02	0.02	0.04	0.09	0.09	-0.09	0.09	0.09	0.05	1.62	-0.07
578																0.08	0.05	0.04	0.04	0.10	0.03	0.03	0.03	0.02	0.03	0.03	-0.09	0.03	0.04	-1.22	0.03	0.06
590																	0.00	0.06	0.07	1.00	0.02	0.00	0.02	-0.05	-0.05	-0.09	-0.74	-0.09	-0.09	-0.06	-0.96	-0.10
609																		0.05	0.07	3.47	-0.03	0.07	-0.22	0.07	0.04	0.03	-0.06	0.02	0.02	0.06	-0.04	-0.13
633																			0.04	0.06	0.04	0.04	0.02	0.01	0.02	0.03	-0.09	0.03	0.03	0.04	0.03	0.05
766																				0.08	0.04	0.04	0.02	0.01	0.02	0.02	-0.01	0.02	0.03	0.04	0.00	0.09
803																					0.03	0.08	0.07	0.12	0.03	0.03	0.31	0.03	0.11	0.10	0.00	0.84
838																						0.05	0.01	0.03	0.03	0.03	-0.05	0.04	0.06	0.09	-0.09	-0.03
857																							0.03	0.03	0.01	-0.82	-0.05	-0.85	0.02	0.04	0.04	0.06
1085																								0.05	0.03	0.03	0.01	0.02	0.02	0.06	0.01	-0.05
1134																									0.04	0.03	0.00	0.02	0.01	0.02	0.08	0.13
1338																										0.04	0.08	0.04	0.04	0.01	0.07	0.11
1373																											0.07	-0.31	0.04	0.01	0.06	0.19
1395																												0.06	0.03	-0.01	-0.47	-0.56
1447																													0.04	0.02	0.05	0.20
1521																														0.03	0.04	0.12
1781																															0.02	0.08
1792																																-0.09
1853																																

Figure 4.9c

Europe



	-859	-855	-491	-413	-282	-234	-179	282	341	481	590	665	803	857	1021	1395	1447	1521	1792	1853
-862	0.05	0.05	0.00	-0.10	0.04	2.31	0.13	-0.07	-0.06	-0.01	-0.06	-0.05	-0.11	-0.02	-0.03	-0.09	-0.09	-0.08	-0.10	-0.07
-859		0.08	-0.09	0.08	0.04	-0.14	0.09	-0.21	-0.23	-0.26	-0.14	0.01	-0.17	0.05	-0.03	-0.30	0.00	0.00	-0.37	-0.87
-855			0.09	0.05	-0.91	0.11	0.04	0.04	0.06	-0.07	0.08	-0.02	-0.06	0.00	0.03	0.08	-0.07	-0.07	0.04	-0.05
-491				0.03	0.06	0.41	0.00	0.05	-0.01	-0.07	0.02	-0.06	-0.06	-0.01	-0.17	-0.36	1.54	1.50	-0.28	0.05
-413					0.07	0.09	0.15	0.03	-0.03	0.03	-0.14	-0.05	-0.11	-0.04	0.00	-0.07	-0.05	-0.05	-0.14	-0.04
-282						0.09	0.06	0.07	0.09	0.00	0.07	-0.03	0.00	0.01	0.04	0.09	-0.02	-0.02	0.02	-0.07
-234							-0.01	-0.49	-0.55	-0.01	-0.62	0.00	-0.19	-0.01	-0.20	-1.35	1.75	-0.15	-0.21	-0.06
-179								-0.06	-0.15	0.01	-0.04	0.02	-0.07	0.03	-0.04	-0.07	-0.06	-0.07	-0.12	-0.06
282									-0.03	-0.11	0.23	3.08	0.48	0.08	-0.04	-0.56	2.04	-0.06	-0.88	0.21
341										0.39	-0.19	-0.02	0.26	0.11	-0.09	-1.07	2.08	-0.13	-0.22	0.19
481											-0.01	0.04	0.81	0.05	-0.07	-0.79	2.13	0.01	-0.13	0.27
590												0.02	1.96	0.00	-0.08	-0.59	-0.12	-0.13	-1.12	0.02
665													0.04	0.08	0.06	2.49	-0.05	-0.05	2.06	1.99
803														0.06	0.02	0.51	0.02	2.43	0.24	0.03
857															0.03	-0.01	0.26	0.01	0.03	0.04
1021																-0.09	0.05	0.03	-0.05	-0.07
1395																	0.02	0.01	-0.40	-0.33
1447																		0.05	0.15	-0.06
1521																			0.17	-0.06
1792																				-0.03

Figure 4.9d

Asia

	-234	282	341	481	590	857	1395	1447	1792	1853
-859	0.08	-0.20	1.50	1.44	-0.13	-0.28	-0.18	-0.19	-0.24	0.83
-234		2.19	0.03	0.02	1.15	-0.02	0.77	-0.12	-0.10	0.73
282			0.01	0.00	-0.11	0.07	0.05	-1.04	0.00	0.00
341				-0.36	-0.02	0.03	-0.07	0.00	-0.05	1.45
481					0.02	0.07	-0.07	0.01	-0.04	1.51
590						0.10	-0.13	-0.05	-0.36	-1.35
857							0.03	-0.02	-0.21	-0.27
1395								-0.01	0.13	-0.30
1447									-0.10	-0.24
1792										0.20
1853										

America

	-491	-234	282	341	481	803	857	1447	1853
-855	5.86	-0.01	0.01	-0.03	-0.03	-0.04	0.04	-0.02	-0.06
-491		0.13	-0.05	0.05	0.04	1.03	-0.01	-0.04	0.01
-234			0.04	-0.35	-0.44	0.03	0.06	-0.01	-0.26
282				0.06	0.05	-0.03	0.44	-0.78	0.13
341					-0.09	0.10	0.06	0.01	-0.09
481						0.12	0.06	0.01	-0.04
803							0.00	-0.04	0.04
857								0.43	0.07
1447									0.18

Figure 4.9e

Figure 4.8: Intragenic linkage disequilibrium and recombination matrices for the *NAT2* locus, generated using Ldhat (McVean, 2002). Reported values are for each pair of SNPs, where BLUE indicates greater than expected linkage disequilibrium, and RED indicates greater than expected recombination. Analyses were performed for each geographic grouping as indicated in Table 2.1: (a) Africa; (b) East Africa; (c) West Africa; (d) Europe; (e) Asia and the Americas.

NAT2 Africa

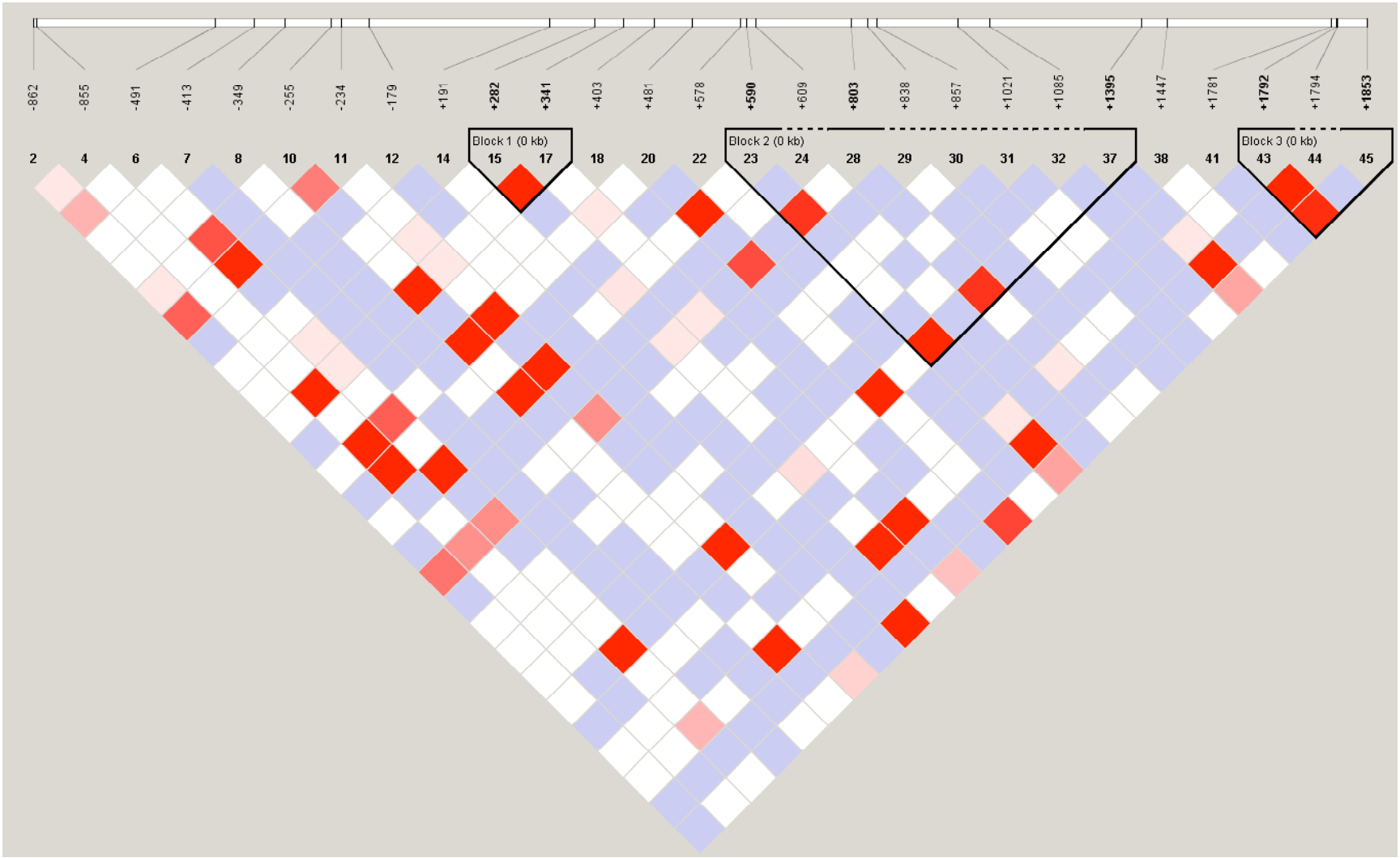


Figure 4.10a

NAT2 East Africa

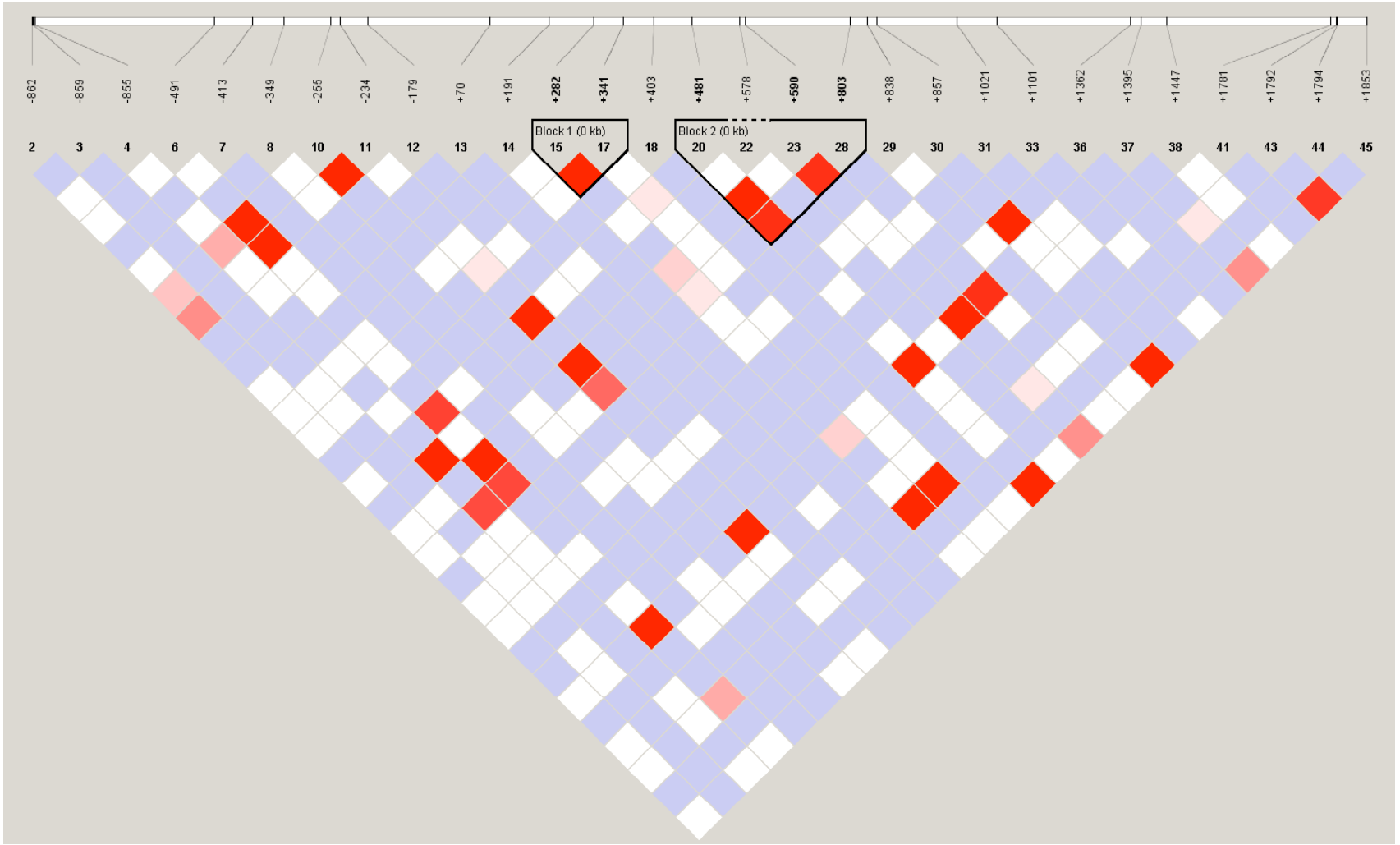


Figure 4.10b

NAT2 West Africa

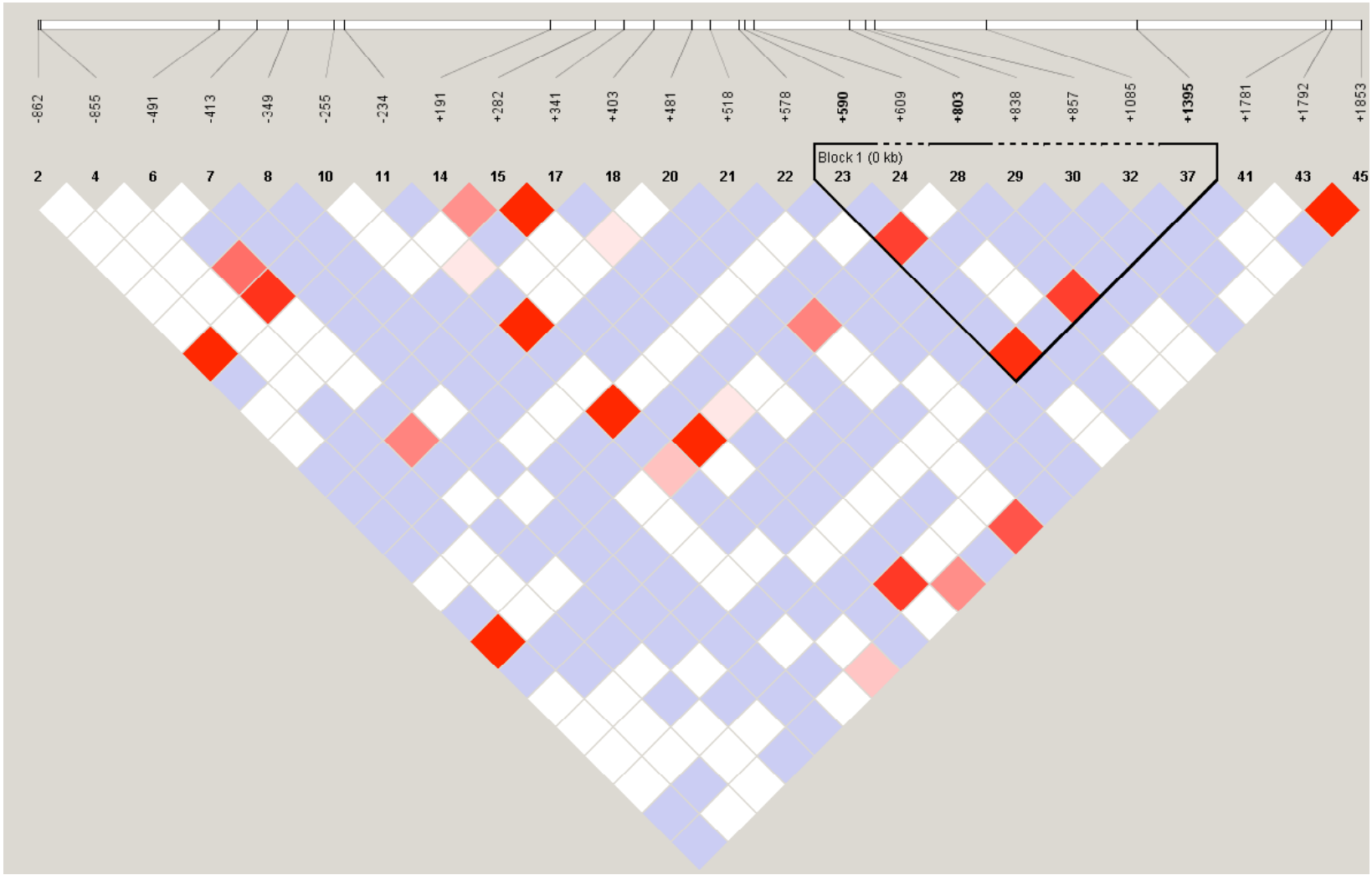
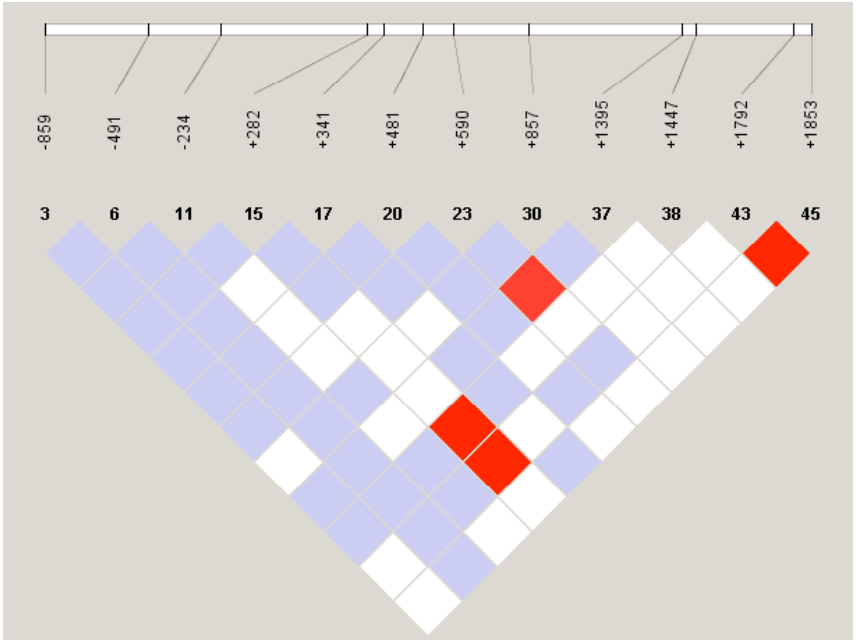


Figure 4.10c

NAT2 Europe

Figure 4.10d

NAT2 Asia



NAT2 Americas

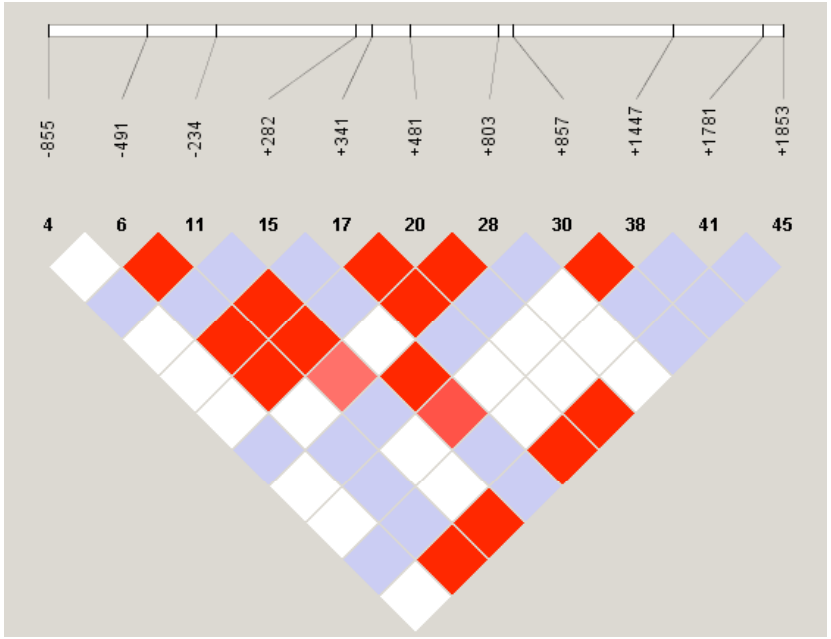


Figure 4.10e

Figure 4.10: Intragenic linkage disequilibrium and recombination graphs for the *NAT2* locus, generated using Haploview (Barrett, 2005). RED indicates D'=1 between two pairs of SNPs, indicating linkage disequilibrium is observed with high confidence (LOD≥2). Blue followed by white indicates lessening of linkage parameter D', as well as confidence of the estimate (*e.g.* blue specifies D'=1 (LOD<2); pink and shades of red specify D'<1 (LOD≥2), and white specifies D'<1 (LOD<2)). Analyses were performed for each geographic grouping as indicated in Table 2.1: (a) Africa; (b) East Africa; (c) West Africa; (d) Europe, Asia and the Americas.

DISCUSSION

We have characterized the pattern of nucleotide diversity at the *NAT2* locus in African, European, Asian and Amerindian populations. We include a large number of diverse African populations in order to identify novel variants that have not been previously characterized and gain a better understanding of the evolutionary history of the *NAT2* locus within Africa. African populations are known to show greater diversity, on average, when compared to non-African groups (European or Asian). For these reasons, we chose direct sequencing, over SNP genotyping of common variants with known effects on acetylator phenotype, as is typically done. This study does not aim to characterize the affect of novel haplotypes on phenotype, but will elucidate the frequencies of novel variants in populations previously unclassified in order to provide an unbiased description of *NAT2* sequence variation. Because of the role of *NAT2* in metabolism of drugs used in the treatment of resurgent diseases such as tuberculosis, hypertension, and HIV, understanding of the pattern of genetic variation at the *NAT2* locus in populations that experience high frequency of diseases such as these has direct implication for drug efficacy and development.

Acetylator Frequency Variation

NAT2 haplotype designation for each individual has allowed for inference of known acetylator phenotypes in all population groups. Unknown acetylator haplotypes were present at low frequencies in both African (8% of African individuals were classified as either unknown/rapid, unknown/slow or unknown/unknown) and non-African populations (3% unknown), preventing inference of acetylator phenotypes in

some cases (Table 4.3). Rapid acetylator haplotypes, the ancestral type when compared to the chimpanzee outgroup, are represented in this *NAT2* dataset by haplotypes **4* (WT), **12A* and **12B*, and **13A* (Figures 4.7a and 4.7b). Two main derived, slow-acetylator haplotypes are observed at appreciable, and approximately equal frequencies, in Africans and European groups: *NAT2*5B* and *NAT2*6A*. Overall, African populations exhibit a greater diversity of acetylator haplotypes, both rapid and slow, in comparison with non-Africans (Table 4.1). In general, we observe haplotypes *12, *13 and *14, as well as a greater diversity of *5 and *6 sub-haplotypes in Africans. Non-Africans are observed to have the highest frequencies of acetylator haplotypes *4, *5, and *6.

Patin *et al*. (2006a) observed the mutation +590 G>A, defining haplotype *NAT2*6*, to be present at intermediate frequency in all populations included in their study. *NAT2*6* is known to be common in African and European populations, but rare in Amerindian populations (Fuselli et al., 2007). *NAT2*6A* is observed in the current study at a frequency of 22.4% in African groups, 28% in Europeans, and 21.4% in Asians. *NAT2*6A* is not observed in Amerindian populations. Of the African groups studied, 33% of *NAT2*6A* haplotypes derive from East African individuals, where the highest frequencies are observed for the Hadza and Sandawe, at 10.3% and 8.3% frequency, respectively (Table 4.3). The star-like pattern of nodes that radiate from haplotype **6A* (Figure 4.8a) are composed of exclusively African haplotypes, with the exception of **6C* which is represented in the Maasai, Kanuri, Mada, and Yoruba African groups and the Druze, French, and Brahui European/Middle Eastern populations. Notably, **6G* is composed of Pygmy groups from Cameroon only (Baka and Bakola), and **6n*, novel to this dataset includes a single Sudanese (Dinka) haplotype (hap27 defined by the presence

of +308 C>T; Table 4.1). This star-like pattern of haplotypes radiating from **6A* indicates rapid expansion of the *NAT2*6* haplotypes, and implies that positive selection may be acting on a particular SNP variant of the *NAT2**6 haplotype. Further testing is necessary to confirm if +590 G>A is the exact variant responsible for the observed pattern.

Haplotype *NAT2*5B* is present in Africans at a frequency of 24.7% (28% within East Africans, specifically), and Europeans at 27% frequency. Though not represented in Asian populations, **5B* is present in Amerindian groups at 38.6% frequency (Table 4.3). This relatively high frequency of **5B* in Amerindians was also observed by Fuselli*, et al*. (2007), in their study of the Karitiana and Pima Amerindians. Patin *et al*. (Patin et al., 2006b) observed **5B* in Africa to be highest in the Chagga of Tanzania (0.438) and Somalian (0.425) populations. Patin *et al*. (Patin et al., 2006a) focus on **5B* in Eurasians, where they observe the highest frequency of **5B* in Sardinians and Ashkenazi populations (0.583 and 0.400, respectively). We observe somewhat different frequencies of **5B* in Africa; most notably, West African Fulani have the highest frequency of **5B* in our dataset (0.500), followed by the Mada of Cameroon (0.464). It has been suggested that the Fulani have certain non-African characteristics, which is interesting in light of the high frequency of **5B* in Europeans observed by Patin *et al*. (2006a). Fulani exhibit physical features, described as "non-Negroid" (Stenning, 1965), that differentiate them from other ethnic groups living in close proximity to them. Recent genetic studies investigating the low incidence of malaria in the Fulani have identified a "Caucasian" component in the Fulani when looking at diverse genetic systems implicated in immune response (HLA I A*24 (Modiano et al., 2001), and genes distinctive of T regulatory

activity (Torcia et al., 2008)). According to Torcia (2008), the Fulani are thought to have a higher immune reactivity, compared to neighboring groups. These findings, in light of the previously mentioned studies concerning the interference of metabolic and immune disorders with acetylator genotype/phenotype relationship (Chapter I), may be informative when considering the high incidence of *NAT2*5B* slow acetylators, as well as drug therapies involving *NAT2* substrates, in the Fulani.

The slow acetylator haplotype *NAT2 *14* (+191G>A), was originally described as "African-specific" because of its high frequency in African-Americans ranging from 48-55%, compared to 10% in populations of European descent. (Bell et al., 1993). The +191 G>A is thought to give rise to an "extreme" slow *NAT2* haplotype, whereby its drastic affect is due to the reduced catalytic activity associated with the polymorphism (Fretland et al., 2001). SNP +191 G>A was originally thought to show an "East-West gradient" across sub-Saharan Africa because it was found to be present in high frequencies in West Africans, low frequency in Sudanese groups (0.029) and absent in those from Somalia (Bayoumi et al., 1997). The frequency of +191A is observed to be higher in several West African groups (Guinea-Bissau (0.192); Gabonese (0.086); Dogon of Mali (0.05); and the South African Venda (0.084)), prompting some researcher to hypothesize a West African origin for the extreme slow *NAT2*14* haplotype, defined by SNP +191A (Bayoumi et al., 1997; Cavaco et al., 2003). In the present study, we do observe haplotype *NAT2*14* to be African-specific (Table 4.3). We observe an elevated frequency of *NAT2*14* in West Africa, in the Kanuri (0.125), and Lemande (0.1429) of Cameroon, as well as the Nigerian Yoruba (0.0833). *NAT2*14* is observed infrequently in East Africa (Table 4.3). Additionally, novel to this study, haplotype *NAT2*14j*, characterized by +191G>A,

+282C>T, and +633G>A, is observed in one Lemande individual from Cameroon, in heterozygous form with the defined sub-haplotype **14B* (Table 4.3).

Population Specific Selective Pressures

Levels of nucleotide and haplotype diversity are relatively consistent between African and European populations at the *NAT2* locus. Asian and Amerindian populations have a slight decrease in diversity at both the continental and population levels. At the continental level, tests of neutrality (TD, D*, F*, F, D), though insignificant, show two distinct patterns, where African groups have negative values at these estimators and non-African groups have positive values for all estimators (Table 4.2). At the population level this pattern breaks down for African groups, but not for Europeans, and we see positive values for all neutrality tests for several African groups. The similar frequencies of **5B* and **6A* in African and European populations (Figure 4.8c and Table 4.3) might be indicative of the effects of natural selection to maintain slow acetylator phenotypes at high frequency in these regions. Tests of neutrality that yield positive estimates for African groups (Bakola Pygmy, Burunge, Fulani, Kanuri, Mada, Sandawe, and Yoruba) and are, in general, lower values than observed in European groups, though significance is not observed for any population. With regard to positive values for estimates of neutrality tests in these African populations, and referring to Table 4.3, we observe either high frequency of both slow acetylator types *5 and *6 (BG, FU, MD) indicating maintenance of both types, or high frequency of slow and rapid (KN, SW, YO) phenotypes, which might be indicative of balancing selection acting to maintain both rapid and slow acetylators in these populations. However, because of the maintenance of

two main slow acetylator haplotypes, as well as the fixation of key SNPs associated with both rapid and slow acetylator phenotypes in different populations, it is possible that long-term balancing selection may be overlaid by the action of positive selection which could be acting on specific slow acetylator mutations in specific populations. These patterns, in general, support the observation originally made by Sabbagh (2008) and Magalon (2008) that multiple modes of selection are operating at *NAT2* on a population-specific basis.

Conclusion

In conclusion, our ability to differentiate demographic effects, such as the series of bottlenecks thought to have occurred during the migration of modern humans out of Africa (Rosenberg et al., 2005) from the effects of natural selection on the negative values of TD observed in Africans versus positive values in Europeans (Stajich and Hahn, 2005), requires further investigation. Furthermore, it is not completely possible to disentangle the effects of demography using the tests of neutrality presented here. To address this issue, modeling of expectations under specific demographic scenarios is necessary. However, even with more rigorous analyses, it is clear that *NAT2* has experienced complicated selection scenarios in different populations. Further, the role of selection at the *NAT2* locus may not be easily clarified because of the somewhat inexact relationship between the interaction of NAT2 enzyme with other endogenous chemicals, individual factors such as age, sex, and diet, as well as the role of Phase I CYP1A2 bioactivation in the NAT2 metabolic pathway.

CHAPTER V. EVOLUTIONARY CHARACTERIZATION OF THE HUMAN *NATP1* LOCUS

RESULTS

Patterns of genetic diversity and haplotype structure

We sequenced 2949 bp of the *NATP1* region in 608 chromosomes, and identified 55 SNPs (Table 5.1). Because the *NATP1* region is not typically characterized in studies of the *NAT* loci, the majority of SNPs identified in the *NATP1* region are novel and have not been previously reported (see Appendix 2). Of the 55 total SNPs identified as variable, only 14 SNPs have been previously reported (Patin et al., (2006a). Additionally, nine singleton mutations and two tri-allelic sites were identified in this *NATP1* dataset.

Summary statistics and estimates of neutrality for the *NATP1* locus are given in Table 5.2. Diversity levels between population groups, estimated by π and θ_{ω}, for the *NATP1* region are relatively consistent, with high levels of haplotype diversity in all populations, with the exception of the Papuan group from New Guinea (HD=0.5), due to small sample size (2n=4). Nucleotide diversity is observed to be highest in the Pygmy populations from Cameroon, where π values for the Baka and Bakola Pgymy groups are π=2.780, and π=2.610, respectively. African populations have consistently higher levels of nucleotide diversity when compared to non-African populations (Table 5.2).

NATP1

	-1039	-1034	-1003	-984	-976	-861	-805	-743	-717	-649	-646	-464	-336	-281	-272	-178	-13	+1	+56	+130	+170	+175	+293	+309	+501	+714	+883	+904	+982	+988	+989	+1072	+1096	+1123	+1129	+1131	+1153	+1183	+1205	+1211	+1226	+1362	+1387	+1417	+1461	+1483	+1490	+1507	+1587	+1607	+1617	+1665	+1720	+1761	+1795
P. troglodytes	C	T	A	T	T	T	T	G	G	A	C	T	A	T	T	A	T	A	G	T	A	A	G	G	G	G	T	G	T	A	T	A	T	G	T	T	T	G	T	A	C	T	T	C	T	T	T	A	A	A	G	A	G	T	A	G
Hap1	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
4	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.
5	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	A
6	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
7	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.
8	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.
10	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
11	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	C	.	.	.	.	.	.	G	G	A	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
12	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	C	.	C	.	.	.	.	G	G	A	.	C	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	A
13	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	T	.	C	.	C	.	.	.	.	G	G	A	.	C	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
14	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	.	T	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
16	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	C	T	.	.	C	T	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	T	.
17	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	A	C	T	.	.	C	T	.	.	.	C	.	.	.	.	G	G	.	C	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
18	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	C	T	.	.	C	T	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
19	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	C	T	.	.	C	T	.	.	.	.	.	.	.	.	.	G	.	C	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
20	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
21	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
22	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
23	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.
24	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
25	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
26	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
27	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A
28	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
29	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
30	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.
31	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
32	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
33	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	C	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
34	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	T	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.	.
35	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
36	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	.	T	.	.	T	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
37	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	G	G	.	C	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
38	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
39	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	C	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
40	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	C	T	.	.	C	T	.	.	.	C	.	.	.	.	.	G	A	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
41	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	A	C	T	.	.	C	T	.	C	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	A
42	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	.	C	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
43	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	.	C	T	.	.	C	T	.	C	C	C	.	.	.	.	G	.	A	.	C	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	C	.	A
44	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	.	C	T	.	.	C	T	.	C	C	C	.	.	.	.	.	G	G	A	.	C	.	.	.	.	C	.	.	.	.	.	.	.	.	.	C	.	A
45	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	.	C	T	.	.	C	T	.	C	C	G	C	.	.	.	.	G	.	A	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	C	.	.
46	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	.	C	T	.	.	C	T	.	C	C	G	C	.	.	.	.	G	G	A	.	C	T	.	.	.	C	.	.	.	.	.	.	.	.	.	C	.	A
47	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	.	C	T	.	.	C	T	.	C	.	C	.	.	.	.	.	G	G	A	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	C	.	A
48	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	.	C	T	.	.	C	T	.	C	.	C	.	.	.	.	.	G	G	A	.	C	.	.	.	.	C	.	.	G	.	.	.	.	.	.	C	.	A
49	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	G	.	C	T	.	.	C	T	.	C	.	C	.	.	.	.	.	G	G	A	.	C	.	.	.	.	C	.	.	.	.	.	.	.	.	.	C	.	A
50	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	C	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
51	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	G	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	.	.	.	.	T	.
52	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	G	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
53	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	G	.	G	A	.	T	.	.	C	T	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.	.
54	.	A	.	.	C	A	C	.	.	C	.	.	C	.	.	.	A	.	T	C	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
55	.	A	.	.	C	A	C	.	.	.	.	.	C	C	.	.	.	.	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
56	.	A	.	.	C	A	C	.	.	.	.	.	C	C	.	.	.	.	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
57	.	A	.	.	C	A	C	.	.	.	.	.	C	C	.	.	.	.	A	.	T	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
58	.	A	.	.	C	A	C	.	.	.	.	.	C	C	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
59	.	A	.	.	C	A	C	.	.	.	.	.	C	C	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
60	.	A	.	.	C	A	C	.	.	.	.	.	C	C	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
61	.	A	.	.	C	A	C	.	.	.	.	.	C	C	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
62	.	A	.	.	C	A	C	.	.	.	.	.	C	C	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	.	.	.	.	.	.
63	.	A	.	.	C	A	C	.	.	.	.	.	C	C	.	.	.	G	A	.	T	.	.	C	C	.	.	.	.	C	.	.	.	.	G	G	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
64	.	A	.	.	C	A	C	.	.	.	.	.	C	C	.	.	.	G	A	C	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
65	.	A	.	.	C	A	C	.	.	.	.	.	T	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	.	.	.	.	T	.
66	.	A	.	.	C	A	C	.	.	.	.	.	T	.	.	G	.	.	A	.	T	.	.	C	T	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	.	.	.	.	C	.	.
67	.	A	.	.	C	A	C	.	.	.	.	.	T	.	.	G	.	G	A	C	T	.	.	C	T	.	.	.	C	.	.	.	.	.	.	G	A	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
68	.	A	.	.	C	A	C	.	.	T	.	C	.	.	.	.	.	.	A	.	T	.	.	C	T	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.
69	.	A	.	.	C	A	C	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	C	.	.	C	.	.	.	.	G	G	A	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
70	.	A	.	.	C	A	C	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	G	.	A	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
71	.	A	.	.	C	A	C	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	G	.	A	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
72	.	A	.	.	C	A	C	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	G	.	A	.	C	T	.	C	.	.	.	A	.	C	.	.	.	.	C	.	.
73	.	A	.	.	C	A	C	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	G	.	A	.	C	T	C	.	C	.	.	A	.	C	.	.	.	.	C	.	.
74	.	A	.	.	C	A	C	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	G	G	.	C	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
75	.	A	.	.	C	A	C	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	G	G	.	C	.	.	.	C	.	.	.	.	A	.	C	.	.	.	C	.	.
76	.	A	.	.	C	A	C	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	G	G	A	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
77	.	A	.	.	C	A	C	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	G	G	A	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.
78	.	A	.	.	C	A	C	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	G	G	A	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	C	.	.

Table 5.1a

NATP1 Haplotypes

P. troglodytes	BG	BK	DN	FU	HZ	KN	LM	MD	MS	PB	PL	SA	SW	TR	YO	Afr	DZ	FR	PA	RU	SR	CB	HA	JP	NG	SB	BR	PI	Non	ALL
Hap1	1	.	.	.	.	1	.	.	.	.	.	.	1	.	.	3	3	1	1	2	4	4	7	2	.	.	6	8	38	41
2	.	3	1	.	2	1	1	1	.	.	1	4	.	3	2	19	1	1	2	.	.	.	1	4	.	2	.	.	11	30
3	.	6	1	.	5	.	2	5	.	1	.	.	1	1	2	24	3	2	1	.	.	4	7	2	.	1	.	.	20	44
4	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
5	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
6	.	.	1	.	1	.	.	.	.	1	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
7	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
8	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	2	1	1	.	.	.	.	1	.	.	.	2	1	6	8
9	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
10	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
11	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
12	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
13	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
14	.	.	1	.	.	2	1	.	.	.	1	1	1	5	2	14	1	.	1	1	2	4	5	8	3	2	9	7	43	57
15	.	.	.	.	1	.	.	.	.	.	.	.	.	2	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
16	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
17	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
18	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
19	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
20	1	1	.	.	2	.	1	.	.	.	.	.	.	1	.	6	.	1	1	.	.	.	.	.	.	.	.	.	2	8
21	5	.	2	1	.	1	.	.	2	1	.	.	1	.	2	15	.	2	.	.	.	.	.	.	.	.	.	.	2	17
22	.	2	6	3	2	5	5	1	1	1	.	.	2	4	3	35	4	3	7	.	.	.	.	.	.	.	.	.	14	49
23	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
24	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
25	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
26	.	.	1	1	.	.	.	.	.	.	.	.	1	.	.	3	.	1	2	.	2	.	.	.	.	.	.	.	5	8
27	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
28	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
29	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
30	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
31	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
32	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
33	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
34	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
35	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	5
36	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
37	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
38	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
39	1	.	.	.	1	.	2	.	1	.	.	.	.	.	2	7	.	.	.	.	.	.	.	.	.	.	.	.	.	7
40	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
41	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
42	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
43	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
44	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
45	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
46	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
47	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
49	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
50	.	.	.	.	.	3	.	1	1	.	.	.	.	1	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	6
51	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
52	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
53	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
54	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
55	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	3
56	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
57	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
58	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
59	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
60	1	1	.	.	1	.	.	.	.	1	1	1	3	.	1	10	.	.	.	.	.	.	.	.	.	.	.	.	.	10
61	5	.	3	.	6	.	2	2	1	.	.	.	4	2	.	25	.	.	.	.	.	.	.	.	.	.	.	.	.	25
62	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
63	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
64	1	.	.	.	.	1	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
65	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
66	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
67	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
68	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
69	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
70	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
71	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
72	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
73	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
74	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
75	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
76	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
77	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
78	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2

Table 5.1b

NATP1

P. troglodytes	-1039	-1034	-1003	-984	-976	-861	-805	-743	-717	-649	-646	-646	-464	-336	-281	-272	-178	-13	+1	+56	+130	+170	+175	+293	+309	+501	+714	+883	+904	+982	+988	+989	+1072	+1096	+1123	+1129	+1131	+1153	+1183	+1205	+1211	+1226	+1362	+1387	+1417	+1461	+1483	+1490	+1507	+1587	+1607	+1617	+1665	+1720	+1761	+1795
	C	T	A	T	T	T	T	G	G	A	C	T	C	A	T	T	A	T	A	G	G	T	A	A	G	T	T	G	T	A	C	T	T	T	T	G	T	T	C	T	T	T	T	A	A	T	A	A	G	A	G	T	A	G
Hap79	.	A	.	.	C	A	C	.	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	.	C	.	.	.	.	.	.	.	.	C	T	.	.	C	.	.	C	.	.	.	.	.	C	.	.
80	.	A	.	.	C	A	C	.	.	.	T	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	.	C	C	.	.	.	G	G	A	.	.	.	.	C	.	.	.	.	.	.	.	.	.	C	.	.
81	.	A	.	.	C	A	C	.	.	G	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.
82	.	A	.	.	C	A	C	.	.	G	.	.	C	.	.	.	.	.	.	A	.	T	C	.	C	T	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.
83	.	A	.	.	C	A	C	.	.	G	A	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	C	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.
84	.	A	.	.	C	A	C	.	.	G	A	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	C	.	C	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	C	.	.	.	.	.
85	.	A	.	.	C	A	C	.	.	G	A	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.
86	.	A	.	.	C	A	C	.	T	.	.	.	T	.	.	.	.	.	G	.	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	.	.	.	.
87	.	A	.	.	C	A	C	C	.	.	.	.	C	C	.	.	.	.	.	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
88	.	A	.	.	C	A	C	C	.	.	.	.	C	C	.	.	.	.	G	A	.	T	.	.	C	T	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
89	.	A	.	.	C	A	C	C	.	.	.	.	C	C	.	.	.	.	G	A	.	T	.	A	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
90	.	A	.	.	.	A	C	.	.	.	.	.	C	.	.	.	.	.	G	.	C	T	.	.	C	T	.	.	.	.	C	.	.	.	A	G	A	.	.	C	.	C	.	.	C	.	.	.	.	.	C	.	.	.
91	.	A	.	A	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
92	.	A	.	A	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
93	.	A	.	A	C	A	C	.	.	.	T	.	C	.	.	.	.	.	.	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
94	.	A	.	A	C	A	C	.	.	.	T	.	C	.	.	.	.	.	.	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	C	.	.	.	G	.	.	C	.	A
95	.	A	.	A	C	A	C	.	.	G	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
96	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
97	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.
98	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
99	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	.	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
100	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	.	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
101	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
102	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
103	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
104	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
105	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.
106	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	C	.	.	.	.	.	.	.	.	.	.	.	.	.
107	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	C	.	.	.	.	G	G	A	.	C	.	.	.	C	.	.	.	.	A	.	.	C	.	.	.
108	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
109	.	A	G	.	C	A	C	.	.	.	.	.	C	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
110	.	A	G	.	C	A	C	.	.	.	.	.	C	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
111	.	A	G	.	C	A	C	.	.	.	.	.	C	C	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
112	.	A	G	.	C	A	.	.	.	.	.	.	C	C	.	.	.	.	.	A	C	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.
113	.	.	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
114	.	.	.	.	C	A	C	.	.	.	.	.	C	.	.	G	.	G	G	A	C	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
115	.	.	.	.	C	A	C	.	.	.	.	.	T	.	.	.	.	.	.	A	C	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
116	.	.	.	.	C	A	C	.	.	.	.	.	T	.	.	.	.	.	.	A	C	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	T	.	G	.	.	.	.	.	.	.	C	.	.	.
117	.	.	.	.	C	A	C	.	.	.	.	.	T	.	.	.	.	.	G	A	C	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
118	.	.	.	.	C	A	C	.	.	.	.	.	T	.	.	G	.	.	.	A	.	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
119	.	.	.	.	C	A	C	.	.	.	.	.	T	.	.	G	.	.	.	A	C	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
120	.	.	.	.	C	A	C	.	.	.	.	.	T	.	.	G	.	G	G	A	C	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
121	.	.	.	.	C	A	C	.	.	.	.	.	T	.	.	G	.	G	G	A	C	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	C	.	.	G	C	.	.	.	.	.	.	C	.	.	.
122	.	.	.	.	C	.	C	.	.	.	.	.	C	.	.	.	.	.	.	A	C	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.
123	.	.	.	.	C	.	C	.	.	.	.	.	C	.	.	G	.	G	G	A	.	T	.	.	C	T	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.
124	.	.	.	.	C	.	C	.	.	.	.	.	T	.	.	.	.	G	G	A	.	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
125	.	.	.	.	C	.	C	.	.	.	.	.	T	.	.	G	.	.	.	A	C	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
126	.	.	.	.	C	.	C	.	.	.	.	.	T	.	.	G	.	.	.	A	C	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
127	.	.	.	.	C	.	C	.	.	.	.	.	T	.	.	G	.	G	G	A	C	T	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
128	.	.	.	.	C	.	C	.	.	.	.	.	T	.	.	G	.	G	G	A	C	T	.	.	C	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	C	.	.	.
129	T	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
130	T	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	.	.	.	.	.
131	T	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	T	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
132	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	.	.	.	C	.	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	.	.	.	.	.	.	.	T	.
133	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	.	.	.	C	.	.	C	.	.	.	.	.	.	.	T	.	.	G	.	.	.	.	.	.	.	.	.	.	.
134	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	G	.	.	.	.	.	.	.	.	.	.	.
135	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.
136	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	C	.	.	.	.	.	.	.	.
137	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.
138	.	A	.	.	C	A	C	.	.	.	.	C	C	.	.	.	.	.	.	A	.	T	.	.	C	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.
139	.	A	.	.	C	A	C	.	.	.	T	.	C	.	.	.	.	.	.	A	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.
140	.	A	.	.	C	A	C	.	.	.	T	.	C	.	.	.	.	.	.	A	C	T	.	.	C	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	G	.	.	.	.
141	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	C	.	.	.	.	.	.	.	.
142	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
143	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	.	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
144	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	G	G	.	.	.	.	.	.	.	.	T	.
145	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	C	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
146	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	C	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
147	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	C	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
148	.	A	.	.	C	A	C	.	.	.	T	.	C	.	.	.	.	.	.	A	.	T	.	.	C	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	G	.	.	.	.
149	.	A	.	.	C	A	C	.	.	.	T	.	C	.	.	.	.	.	.	A	.	T	.	.	C	.	.	.	.	C	.	.	A	.	.	.	.	.	.	.	.	.	G	C	.	.	.	.	.	G	.	.	.	.
150	.	A	.	.	C	A	C	.	.	.	T	.	C	.	.	.	.	.	.	A	.	T	.	.	C	.	.	.	.	C	.	.	A	.	.	.	.	.	.	.	C	.	.	G	.	.	.	.	.	.	.	.	.	.
151	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	T	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.
152	.	A	G	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	G	A	.	T	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	T	.
153	.	A	.	.	C	A	C	.	.	.	.	.	C	.	.	.	.	.	.	A	.	.	.	.	C	.	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.

Table 5.1c

NATP1 Haplotypes

P. troglodytes	BG	BK	DN	FU	HZ	KN	LM	MD	MS	PB	PL	SA	SW	TR	YO	Afr	DZ	FR	PA	RU	SR	CB	HA	JP	NG	SB	BR	PI	Non	ALL
Hap79	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
80	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
81	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
82	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
83	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
84	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
85	.	.	.	.	.	.	2	3	1	.	.	.	1	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	7
86	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
87	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
88	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
89	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
90	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
91	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
92	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
93	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
94	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
95	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
96	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	1	1	.	3	4
97	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
98	.	.	1	1	.	1	.	.	.	.	.	.	.	1	.	4	3	4	1	5	1	.	.	.	.	1	4	19	23
99	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
100	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
101	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
102	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
103	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
104	8	1	4	5	3	.	2	3	7	1	.	1	12	5	1	53	3	3	3	.	4	.	.	.	.	.	.	13	66
105	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
106	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
107	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
108	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
109	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
110	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
111	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
112	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
113	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
114	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
115	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
116	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
117	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
118	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
119	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
120	1	.	.	.	.	3	1	1	1	.	.	.	.	.	1	8	.	.	.	.	.	.	.	.	.	.	.	.	.	8
121	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
122	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
123	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
124	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	2	.	.	.	.	.	.	.	.	.	.	.	.	.	2
125	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
126	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
127	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
128	2	.	.	.	.	.	2	.	2	.	.	.	.	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	6
129	.	.	.	1	.	.	3	1	.	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	5
130	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
131	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1
132	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	1	.	.	.	.	.	.	.	.	.	1	1
133	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	1	.	.	.	.	.	.	1	1
134	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	1	.	.	.	.	.	.	.	.	.	1	1
135	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	1	.	.	.	.	.	.	.	.	1	1
136	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	1	.	.	.	.	.	.	.	.	.	.	1	1
137	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	1	.	.	.	1	.	.	.	.	.	.	2	2
138	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	1	.	.	.	.	.	.	.	.	1	1
139	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	1	.	.	.	.	.	.	.	.	.	.	1	1
140	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	1	.	.	.	.	.	.	.	.	1	1
141	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	1	.	.	.	.	.	.	.	.	1	1
142	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	1	.	.	.	.	.	.	.	.	.	1	1
143	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	1	.	.	.	.	.	.	.	.	1	1
144	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	1	.	.	.	.	.	1	1
145	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	1	.	.	1	1
146	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	1	.	.	.	.	1	1
147	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	1	.	1	.	2	2
148	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	1	.	.	.	1	1
149	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	2	.	.	.	2	2
150	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	1	.	1	1
151	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	1	.	.	.	.	.	1	1
152	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	1	.	.	.	.	1	1
153	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	.	.	.	.	.	.	.	.	.	.	1	1	1
	36	30	30	26	28	26	28	28	26	18	14	14	36	30	24	394	22	22	24	8	15	14	24	20	4	8	20	20	147	460

Table 5.1d

Table 5.1 (a-d are quadrants of Table 5.1): One hundred and fifty-three *NATP1* phased haplotypes were identified, and listed in (a) and (c). The number of observed haplotypes in each population are shown in parts (b) and (d). All population abbreviations follow that presented in Table 2.1, in addition to the abbreviation 'Afr' indicating all African populations and 'Non', which denotes all non-African groups. Derived variants, as compared to *P. troglodytes*, are indicated.

	N	2N	S	s	r	H	HD	*Si*	π (∗10³)	θω(∗10³)	TD	P(TD<obs)	D*	P(D*<obs)	F*	P(F*<obs)	D	P(D<obs)	F	P(F<obs)	H	P(H<obs)
NATP1																						
Africa	197	394	50	*na*	*na*	131	0.960	6	1.800	2.640	-0.896	0.193	0.249	0.644	-0.337	0.401	0.772	0.767	0.010	0.551	-8.200	0.052
E Africa	93	186	44	*na*	*na*	59	0.930	7	1.700	2.690	-1.095	0.125	-0.008	0.517	-0.574	0.304	0.755	0.754	-0.057	0.510	-10.306	0.037
W Africa	66	132	40	*na*	*na*	59	0.964	6	1.790	2.550	-0.907	0.187	0.445	0.734	-0.137	0.476	0.358	0.606	-0.223	0.448	-5.479	0.091
Pygmy groups	16	32	27	*na*	*na*	27	0.977	5	2.810	2.530	0.403	0.721	0.156	0.519	0.280	0.626	0.825	0.855	0.913	0.834	-0.702	0.268
Europe	52	104	17	*na*	*na*	23	0.915	7	0.860	1.110	-0.614	0.314	-2.365	0.031	-2.049	0.036	-1.779	0.067	-1.568	0.082	-0.879	0.201
Asia	35	70	14	*na*	*na*	15	0.824	2	0.840	0.990	-0.424	0.396	0.498	0.793	0.208	0.594	0.302	0.583	-0.034	0.501	1.103	0.625
Americas	20	40	4	*na*	*na*	7	0.776	1	0.550	0.400	0.972	0.853	0.191	0.651	0.501	0.659	1.040	1.00	1.025	0.816	-1.274	0.799
Africa																						
Baka Pygmy	9	18	27	*na*	*na*	16	.987	9	2.780	2.760	0.030	0.568	-0.261	0.382	-0.205	0.413	0.205	0.636	0.218	0.589	-1.464	0.234
Bakola Pygmy	7	14	21	*na*	*na*	12	.967	3	2.610	2.240	0.699	0.797	0.683	0.730	0.788	0.800	0.946	0.831	1.113	0.863	-0.264	0.304
Biaka Pygmy	15	30	30	*na*	*na*	19	.947	8	1.640	2.400	-1.137	0.125	-0.396	0.303	-0.748	0.220	0.455	0.723	-0.084	0.492	-3.899	0.132
Burunge	18	36	27	*na*	*na*	15	0.905	3	1.640	2.210	-0.889	0.195	0.960	0.923	0.405	0.674	1.236	0.924	0.558	0.727	-5.448	0.082
Dinka	15	30	28	*na*	*na*	20	0.954	7	1.620	2.400	-1.164	0.107	-0.631	0.209	-0.945	0.184	0.018	0.456	-0.381	0.380	0.349	0.349
Fulani	13	26	28	*na*	*na*	20	0.960	12	1.640	2.490	-1.265	0.087	-0.967	0.182	-1.244	0.131	-0.610	0.302	-1.026	0.207	-5.428	0.084
Hadzabe	14	28	26	*na*	*na*	14	0.918	10	1.570	2.350	-1.213	0.103	-0.896	0.179	-1.172	0.139	-0.464	0.327	-0.841	0.248	-2.899	0.165
Kanuri	13	26	21	*na*	*na*	16	0.945	3	1.930	1.870	0.126	0.622	0.773	0.767	0.670	0.768	0.642	0.750	0.546	0.715	-0.382	0.279
Lemande	14	28	22	*na*	*na*	16	0.950	2	1.810	2.000	-0.338	0.426	0.846	0.877	0.553	0.730	1.487	0.750	1.075	0.876	-1.783	0.188
Maasai	13	26	25	*na*	*na*	18	0.929	7	1.810	2.220	-0.678	0.274	-0.046	0.511	-0.285	0.388	0.008	0.505	-0.297	0.407	-3.126	0.140
Mada	14	28	22	*na*	*na*	19	0.955	4	1.610	1.920	-0.564	0.318	0.523	0.785	0.211	0.595	0.752	0.774	0.417	0.671	-1.328	0.216
S African San	7	14	17	*na*	*na*	11	0.934	6	1.700	1.920	-0.464	0.352	-0.228	0.383	-0.336	0.364	0.443	0.738	0.246	0.603	-1.934	0.175
Sandawe	18	36	26	*na*	*na*	19	0.879	9	1.810	2.210	-0.624	0.305	-0.857	0.198	-0.922	0.192	-0.530	0.261	-0.648	0.290	-0.654	0.271
Turu	15	30	24	*na*	*na*	15	0.929	7	1.450	2.050	-1.049	0.147	-0.196	0.452	-0.553	0.289	-0.062	0.443	-0.535	0.331	-1.241	0.229
Yoruba	12	24	32	*na*	*na*	17	0.971	13	1.920	3.000	-1.361	0.071	-0.699	0.226	-1.059	0.162	-0.400	0.348	-0.879	0.237	-3.435	0.166
Europe																						
French	11	22	7	*na*	*na*	13	0.939	1	0.810	0.650	0.775	0.819	0.650	0.861	0.795	0.784	1.366	1.000	1.414	0.923	-1.039	0.156
Druze	11	22	8	*na*	*na*	11	0.922	3	0.890	0.840	0.198	0.624	-0.220	0.436	-0.114	0.446	-0.523	0.423	-0.341	0.395	1.212	0.811
Sardinian	12	24	8	*na*	*na*	12	0.917	1	0.910	0.730	0.802	0.826	0.737	0.885	0.879	0.821	0.632	0.657	0.799	0.782	1.188	0.817
Brahui	12	24	13	*na*	*na*	14	0.906	6	0.950	1.180	-0.682	0.826	-1.380	0.130	-1.364	0.113	-1.615	0.110	-1.496	0.111	0.500	0.417
Russian	6	12	4	*na*	*na*	4	0.652	1	0.500	0.450	0.426	0.692	0.368	0.459	0.433	0.648	0.301	0.440	0.393	0.646	-0.303	0.249
Asia																						
Cambodian	7	14	5	*na*	*na*	5	0.802	2	0.750	0.640	0.613	0.757	0.020	0.537	0.202	0.546	-0.351	0.258	-0.202	0.480	0.352	0.465
Han	12	24	7	*na*	*na*	8	0.812	2	0.700	0.640	0.336	0.678	-0.029	0.606	0.090	0.534	-0.315	0.342	-0.210	0.425	0.326	0.409
Japanese	10	20	11	*na*	*na*	7	0.805	1	1.020	1.050	-0.091	0.515	0.997	0.606	0.792	0.799	0.901	0.807	0.611	0.731	1.653	0.831
Papuan	2	4	2	*na*	*na*	2	0.500	2	0.340	0.370	-0.710	0.000	-0.710	0.000	-0.604	0.000	-1.201	0.000	-1.279	0.000	0.667	1.000
Yakut	4	8	9	*na*	*na*	6	0.929	7	0.930	1.180	-1.018	0.181	-1.136	0.121	-1.227	0.179	-2.045	0.032	-2.195	0.031	1.500	0.219
Americas																						
Karitiana	10	20	4	*na*	*na*	7	0.774	1	0.560	0.480	0.528	0.738	0.387	0.725	0.492	0.699	1.133	0.718	0.963	0.777	-1.211	0.097
Pima	10	20	4	*na*	*na*	5	0.795	0	0.560	0.380	1.362	0.913	1.108	1.000	1.358	0.929	1.012	0.726	1.200	0.853	0.674	0.768

All values were calculated using DNAsp (Rosas, 1995)

p<0.05 α' ##

Table 5.2

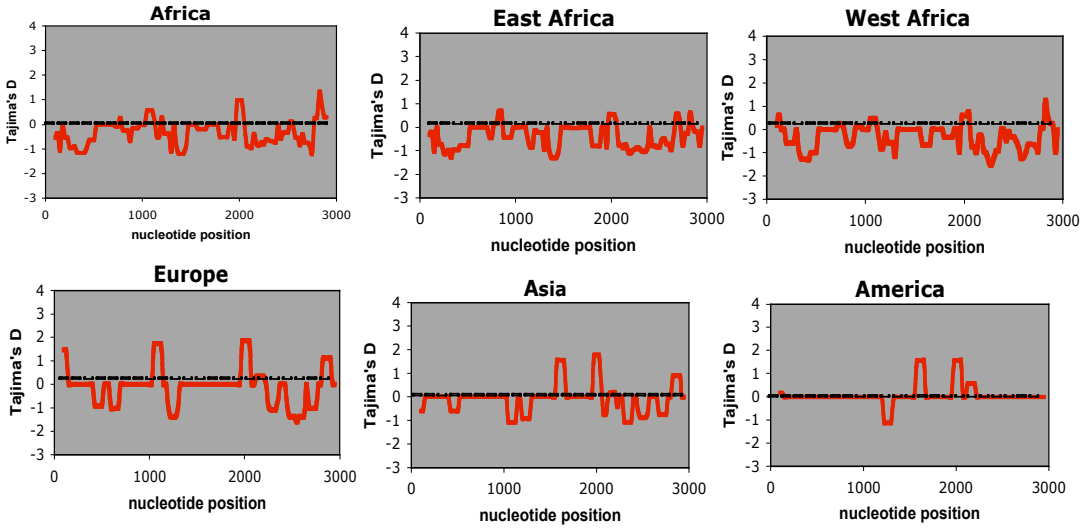
Table 5.2: Summary statistics and estimates of neutrality for the *NATP1* locus. All calculations were performed using DNAsp version 4.20.2 (Rozas and Rozas; Rozas et al.). All abbreviations are as follows: S= number of segregating sites; 2N=number of chromosomes included for analysis; *s*=silent or synonymous variants identified; *r*=replacement or non-synonymous variants identified; H= number of haplotypes identified; HD= haplotype diversity; *Si*=number of singleton mutations; π=nucleotide diversity; θω=Waterson's theta estimator; TD=Tajima's D statistic (Tajima, 1989); D, D* and F, F* equal Fu and Li's test statistics at the inter-specific and intra-specific levels (Fu and Li, 1993), respectively; H=Fay and Wu's H statistic (Fay and Wu, 2000). P values, indicated by P (estimator<obs), indicate significance levels assessed for all neutrality estimates using the coalescent simulator within DNAsp (10,000 replicates), assuming no recombination. Following Bonferroni correction for multiple testing, we do not observe significance for any estimator.

Statistical Tests of Neutrality

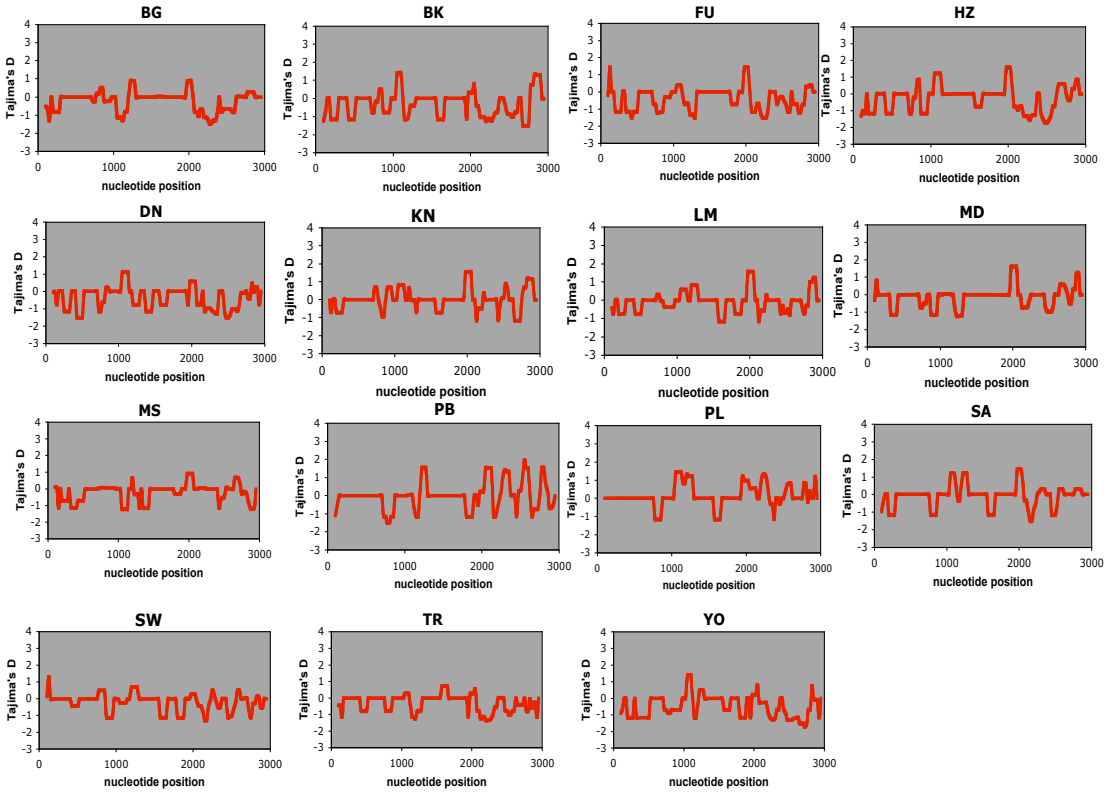
By definition, selection is not expected to act directly on a pseudogene, but non-neutral effects on neighboring loci can affect the sequence variability of a pseudogene (Martinez-Arias et al., 2001). Thus, the neutral model cannot be assumed for the *NATP1* locus.

To test whether *NATP1* evolves according to a neutral model, several tests of neutrality based on the allelic frequency spectrum were performed for the *NATP1* locus (Table 5.2). Tests of neutrality at *NATP1* are highly inconsistent across intra-specific estimators. For example, TD and D* values, which measure different (but related) aspects of the data, are highly inconsistent for *NATP1*. It would be expected that these values would be correlated, as is the case for the functional loci, but not identical (Wall and Przeworski, 2000). Values for neutrality estimators for *NATP1* do not appear to be correlated, and sliding window analysis of TD (Figure 5.1) are both highly positive and highly negative for the same population comparisons, an indication of "noise" in the dataset. Most populations have positive, though not significant, values for inter-species comparisons of D and F (Table 5.2). TD values in the sliding window analyses are not observed to be significant at the $p<0.05$ level in any population group.

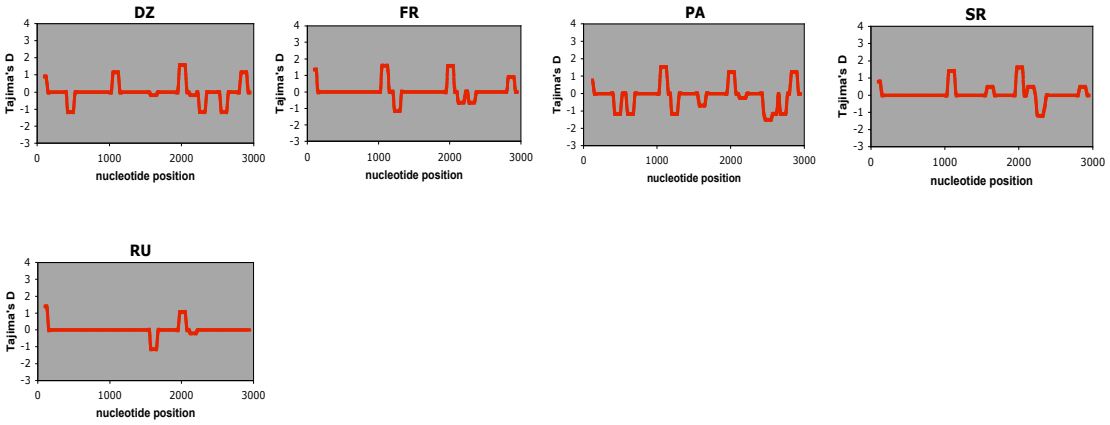
NATP1 Continental Groups



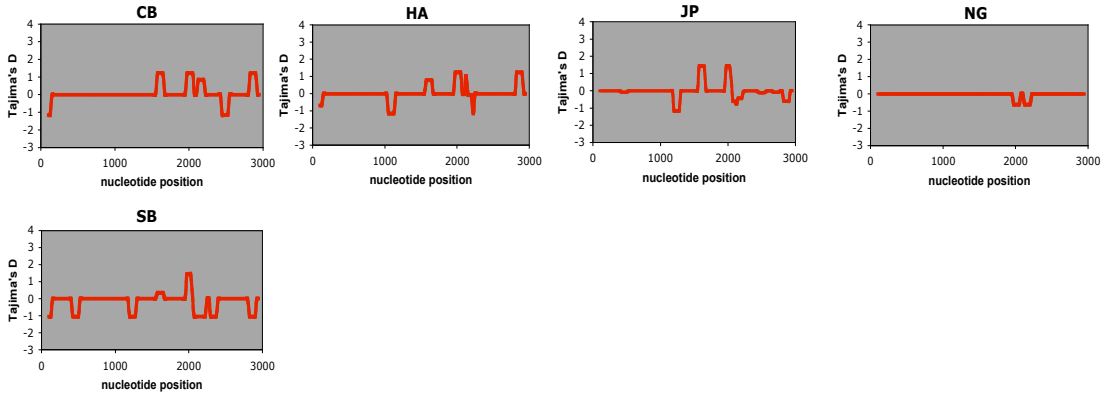
NATP1 Africa



NATP1 Europe



NATP1 Asia



NATP1 America

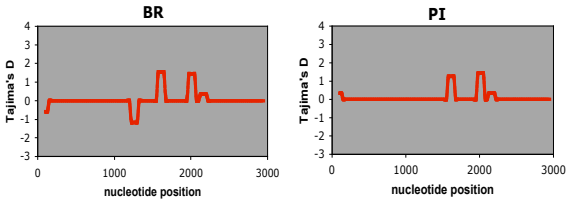


Figure 5.1

Figure 5.1: Sliding window Tajima's D estimates the statistic across the region at defined window lengths of 100 sites at steps of 25 sites. Populations included in each continental grouping follow that presented in Table 2.1.

Population Differentiation

Geographic grouping of populations as illustrated by the MDS plots for the *NATP1* locus (Figure 5.2), is consistent with demographic expectations of a neutrally evolving locus. Populations are observed to cluster into four main geographic regions: Africa, Europe, Asia, and the Americas. The percentage of variation between populations relative to within populations, as indicated by the AMOVA analysis, is greater than observed for the functional loci, reflective of elevated population differentiation. In concordance with the MDS results, the percentage of variation among groups and among populations within groups is relatively consistent for *NATP1* (6.73 and 5.09 percent, respectively). The fixation indices for the *NATP1* locus are more similar to that expected for neutrally evolving loci (F_{ST}=0.118, F_{CT}=0.067, and F_{SC}=0.054).

Phylogenetic haplotype networks

The median-joining haplotype network (Figure 5.3a) for the *NATP1* region conforms to expectation for a neutrally evolving locus. The majority of haplotypes are observed in all geographic groupings. A high number of singleton and doubleton haplotypes are seen in Africa, consistent with the longer evolutionary history allowing recombination to break-up *NATP1* haplotypes in these groups. The high level of reticulation in this network is also indicative of high levels of recombination affecting this locus. In Figure 5.3b, when only high frequency haplotypes (present in >10 individuals) are analyzed we observe three cycles or hyper-cubes, indicating that recombination or homoplasy may have affected the *NATP1* region at positions +1 and +1003; +1 and +904, and +1 and +1761.

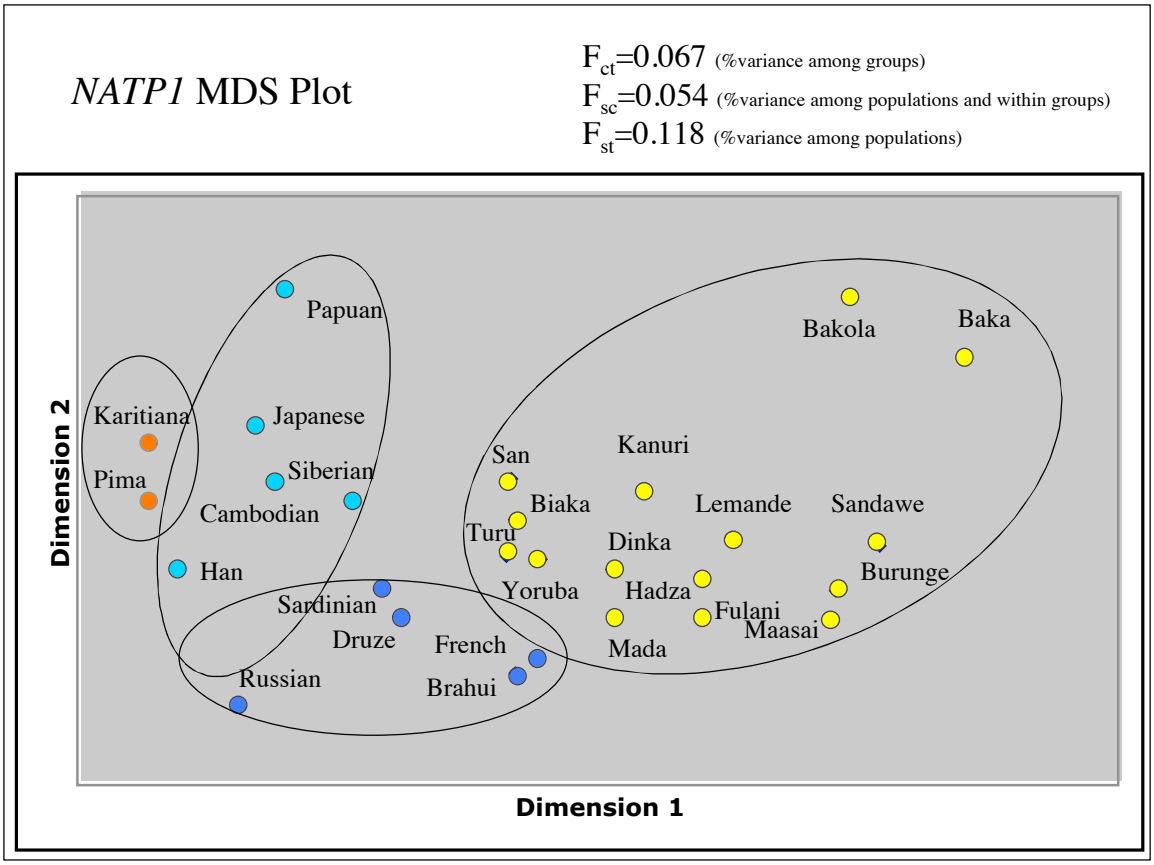


Figure 5.2

Figure 5.2: Multi-Dimensional Scaling plot for the *NATP1* loci. Continental group clusters are indicated. AMOVA results indicated in inset, where variance among groups= F_{CT}, variance among populations within groups= F_{SC}, and variance within populations= F_{ST} .Yellow=Africa; Blue=Europe; Turquoise=Asia; Orange=Americas.

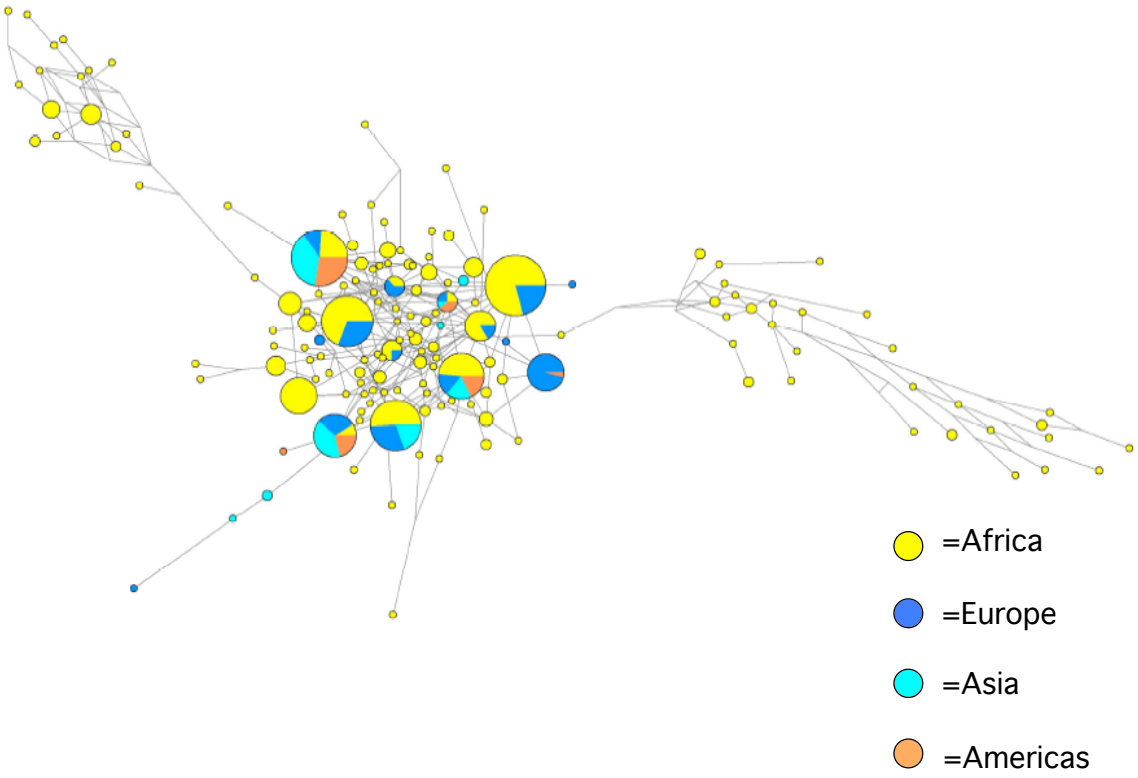


Figure 5.3a

Figure 5.3a: Median-joining network for the entire ~3kb *NATP1* region. Major geographic groups indicated, where Yellow=Africa; Blue=Europe; Turquoise=Asia; and Orange=Americas.

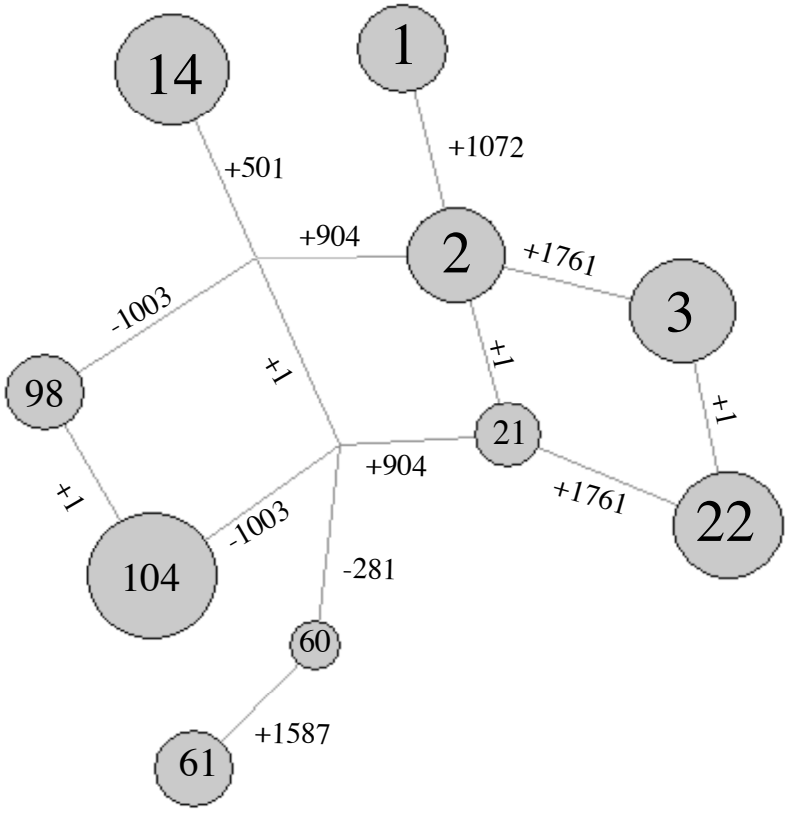


Figure 5.3b

Figure 5.3b: Median-joining network for all NATP1 haplotypes that are present in the current dataset at frequency >10 individuals. Mutations, following the consensus +1 indicates the ATG start site, according to alignment with both NAT1 and NAT2, are indicated on each branch. The sites of potential recombination or homoplasy are observable in the three hypercubes that makeup the center of the structure. All haplotype-numbering follows phased haplotypes presented in Table 5.2.

Patterns of intragenic linkage disequilibrium

One hundred and fifty-three distinct haplotypes were inferred from the 2949 bp *NATP1* region, indicating that recombination has affected the current pattern of diversity present at this locus (Table 5.1). At the continental level, recombination has affected *NATP1* in both African and European groupings (Figure 5.4). Asian and Amerindian population groups do not show evidence of greater than estimated linkage disequilibrium or hotspots of recombination at the *NATP1* locus. In general, African populations show higher levels of recombination for the *NATP1* locus relative to non-African populations, again consistent with demographic expectations. East and West Africa show very similar patterns of linkage disequilibrium and recombination, where the same general location (if not the exact SNP) shows greater than expected linkage disequilibrium or recombination in every case. One exception is the greater than expected linkage disequilibrium between SNPs at positions +904 and +1761 in East Africans. These two SNPs define a single cycle, or hyper-cube, in the haplotype network presented in Figure 5.3b.

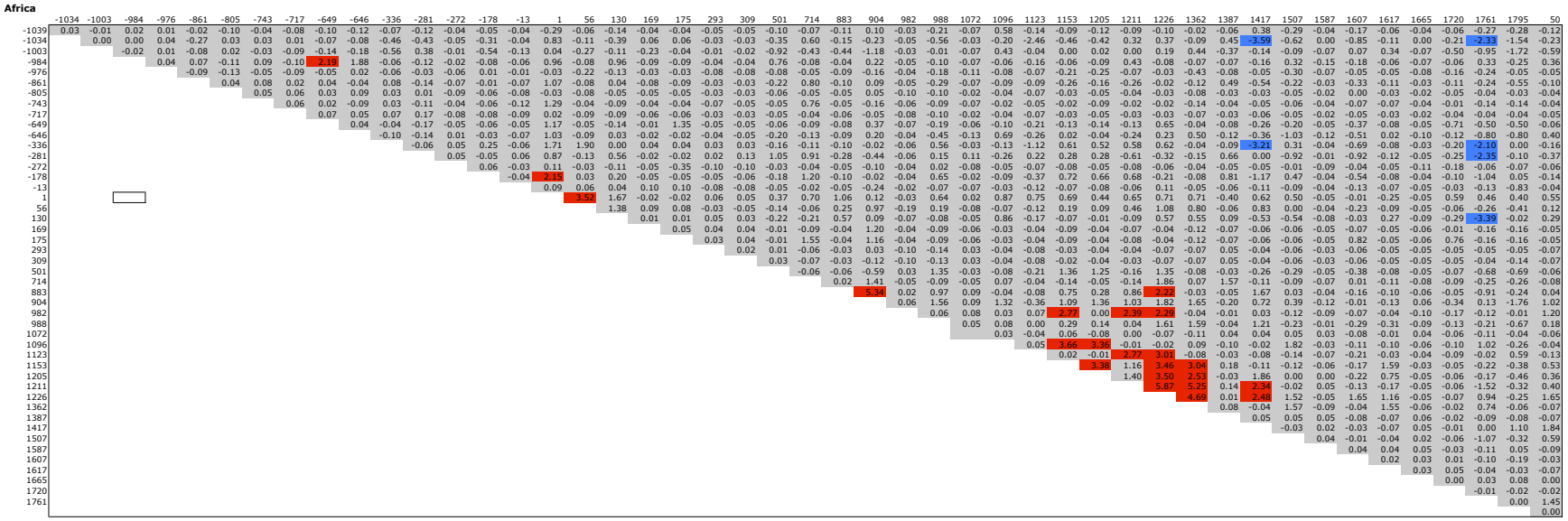


Figure 5.4a

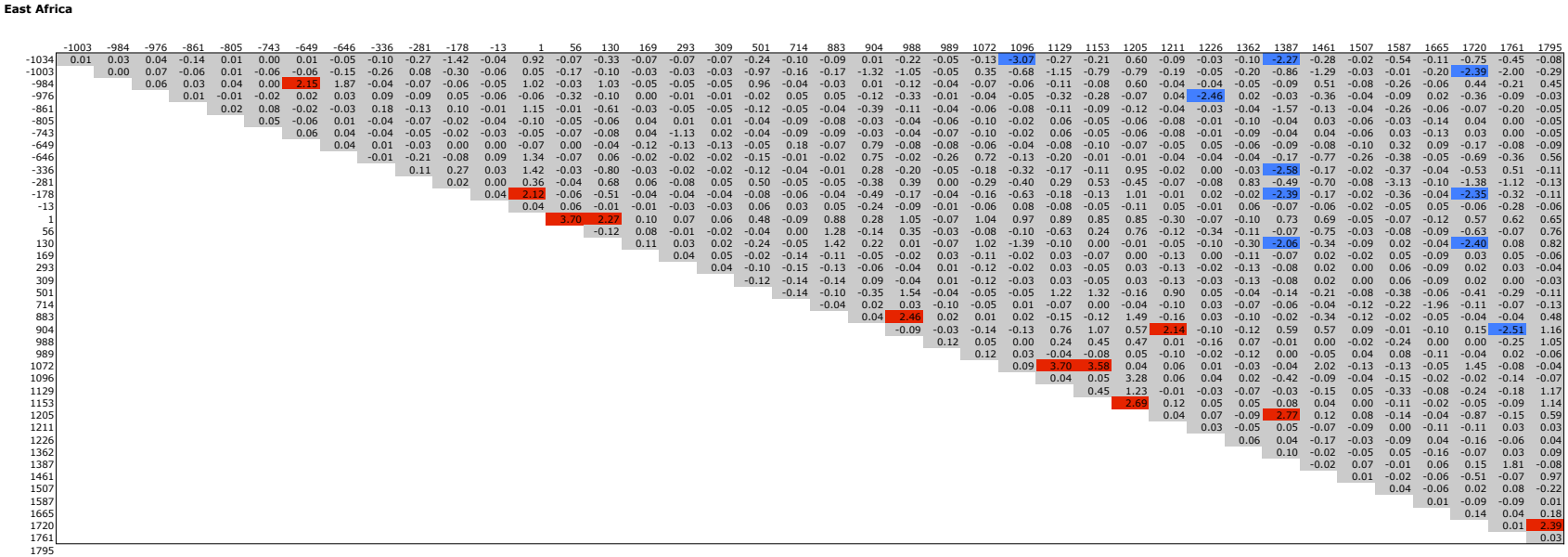
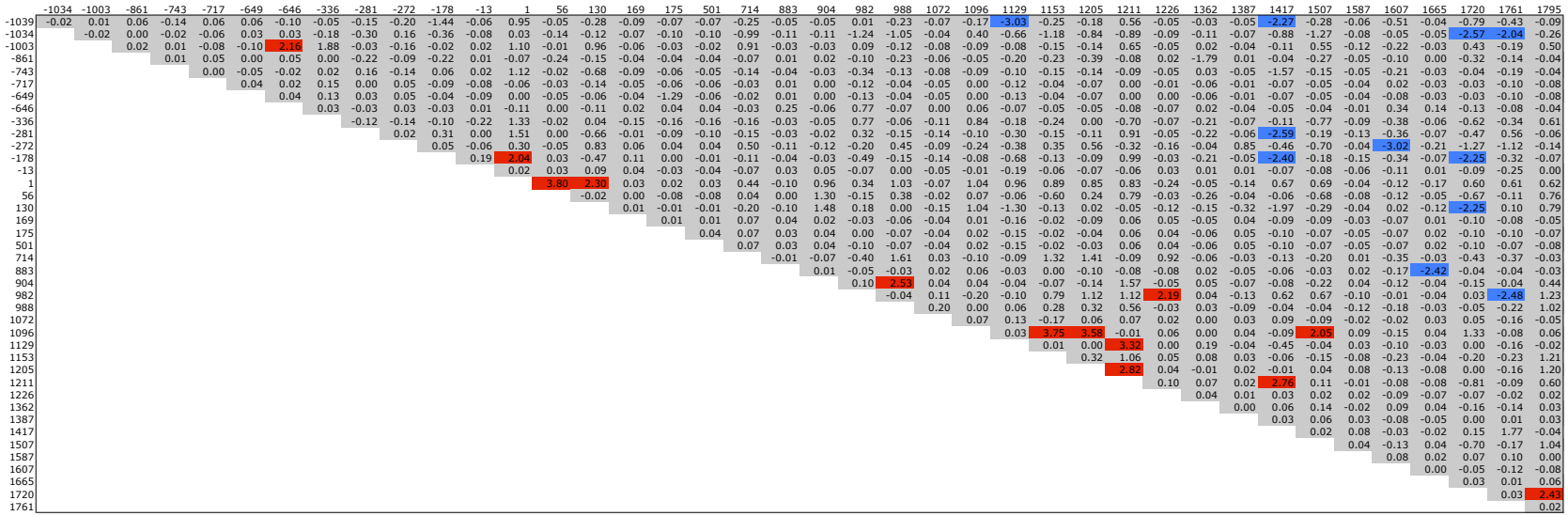

East Africa
-1003 -984 -976 -861 -805 -743 -649 -646 -336 -281 -178 -13 1 56 130 169 293 309 501 714 883 904 988 989 1072 1096 1129 1153 1205 1211 1226 1362 1387 1461 1507 1587 1665 1720 1761 1795
-1034 0.01 0.03 0.04 -0.14 0.01 0.00 0.01 -0.05 -0.10 -0.27 -1.42 -0.04 0.92 -0.07 -0.33 -0.07 -0.07 -0.07 -0.24 -0.10 -0.09 0.01 -0.22 -0.05 -0.13 -3.07 -0.27 -0.21 0.60 -0.09 -0.03 -0.10 -2.27 -0.28 -0.02 -0.54 -0.11 -0.75 -0.45 -0.08
-1003 0.00 0.07 -0.06 0.01 -0.06 -0.06 -0.15 -0.26 0.08 -0.30 -0.06 0.05 -0.17 -0.10 -0.03 -0.03 -0.03 -0.97 -0.16 -0.17 -1.32 -1.05 -0.05 0.35 -0.68 -1.15 -0.79 -0.79 -0.19 -0.05 -0.20 -0.86 -1.29 -0.03 -0.01 -0.20 -2.39 -2.00 -0.29
-984 0.06 0.03 0.04 0.00 2.15 1.87 -0.04 -0.07 -0.06 -0.05 1.02 -0.03 1.03 -0.05 -0.05 -0.05 0.96 -0.04 -0.03 0.01 -0.12 -0.04 -0.07 -0.06 -0.11 -0.08 0.60 -0.04 -0.04 -0.05 -0.09 0.51 -0.08 -0.26 -0.06 0.44 -0.21 0.45
-976 0.01 -0.01 -0.02 0.02 0.03 0.09 -0.09 0.05 -0.06 -0.06 -0.32 -0.10 0.00 -0.01 -0.01 -0.02 0.05 0.05 -0.12 -0.33 -0.01 -0.04 -0.05 -0.32 -0.28 -0.07 0.04 -2.46 0.02 -0.03 -0.36 -0.04 -0.09 0.02 -0.36 -0.09 -0.03
-861 0.02 0.08 -0.02 -0.03 0.18 -0.13 0.10 -0.01 1.15 -0.01 -0.61 -0.03 -0.05 -0.05 -0.12 -0.05 -0.04 -0.39 -0.11 -0.04 -0.06 -0.08 -0.11 -0.09 -0.12 -0.04 -0.03 -0.04 -1.57 -0.13 -0.04 -0.26 -0.06 -0.07 -0.20 -0.05
-805 0.05 -0.06 0.01 -0.04 -0.07 -0.02 -0.04 -0.10 -0.05 -0.06 0.04 0.01 0.01 -0.04 -0.09 -0.08 -0.03 -0.04 -0.06 -0.10 -0.02 0.06 -0.05 -0.06 -0.08 -0.01 -0.10 -0.04 0.03 -0.06 -0.03 -0.14 0.04 0.00 -0.05
-743 0.06 0.04 -0.04 -0.05 -0.02 -0.03 -0.05 -0.07 -0.08 0.04 -1.13 0.02 -0.04 -0.09 -0.09 -0.03 -0.04 -0.07 -0.10 -0.02 0.06 -0.05 -0.06 -0.08 -0.01 -0.09 -0.04 0.04 -0.06 0.03 -0.13 0.03 0.00 -0.05
-649 0.04 0.01 -0.03 0.00 0.00 -0.07 0.00 -0.04 -0.12 -0.13 -0.13 -0.05 0.18 -0.07 0.79 -0.08 -0.08 -0.06 -0.04 -0.08 -0.10 -0.07 -0.05 0.05 -0.06 -0.09 -0.08 -0.10 0.32 0.09 -0.17 -0.08 -0.09
-646 -0.01 -0.21 -0.08 0.09 1.34 -0.07 0.06 -0.02 -0.02 -0.02 -0.15 -0.01 -0.02 0.75 -0.02 -0.26 0.72 -0.13 -0.20 -0.01 -0.01 -0.04 -0.04 -0.04 -0.17 -0.77 -0.26 -0.38 -0.05 -0.69 -0.36 0.56
-336 0.11 0.27 0.03 1.42 -0.03 -0.80 -0.03 -0.02 -0.02 -0.12 -0.04 -0.01 0.28 -0.20 -0.05 -0.18 -0.32 -0.17 -0.11 0.95 -0.02 0.00 -0.03 -2.58 -0.17 -0.02 -0.37 -0.04 -0.53 0.51 -0.11
-281 0.02 0.00 0.36 -0.04 0.68 0.06 -0.08 0.05 0.50 -0.05 -0.05 -0.38 0.39 0.00 -0.29 -0.40 0.29 0.53 -0.45 -0.07 -0.08 0.83 -0.49 -0.70 -0.08 -3.13 -0.10 -1.38 -1.12 -0.13
-178 0.04 2.12 -0.06 -0.51 -0.04 -0.04 -0.04 -0.08 -0.06 -0.04 -0.49 -0.17 -0.04 -0.16 -0.63 -0.18 -0.13 1.01 -0.01 0.02 -0.02 -2.39 -0.17 -0.02 -0.36 -0.04 -2.35 -0.32 -0.11
-13 0.04 0.06 -0.01 -0.01 -0.03 -0.03 0.06 0.03 0.05 -0.24 -0.09 -0.01 -0.06 0.08 -0.08 -0.05 -0.11 0.05 -0.01 0.06 -0.07 -0.06 -0.02 -0.05 0.05 -0.06 -0.28 -0.06
1 3.70 2.27 0.10 0.07 0.06 0.48 -0.09 0.88 0.28 1.05 -0.07 1.04 0.97 0.89 0.85 0.85 -0.30 -0.07 -0.10 0.73 0.69 -0.05 -0.07 -0.12 0.57 0.62 0.65
56 -0.12 0.08 -0.01 -0.02 -0.04 0.00 1.28 -0.14 0.35 -0.03 -0.08 -0.10 -0.63 0.24 0.76 -0.12 -0.34 -0.11 -0.07 -0.75 -0.03 -0.08 -0.09 -0.63 -0.07 0.76
130 0.11 0.03 0.02 -0.24 -0.05 1.42 0.22 0.01 -0.07 1.02 -1.39 -0.10 0.00 -0.01 -0.05 -0.10 -0.30 -2.06 -0.34 -0.09 0.02 -0.04 -2.40 0.08 0.82
169 0.04 0.05 -0.02 -0.14 -0.11 -0.05 -0.02 0.03 -0.11 -0.02 0.03 -0.07 0.00 -0.13 0.00 -0.11 -0.07 0.02 -0.02 0.05 -0.09 0.03 0.05 -0.06
293 0.04 -0.10 -0.15 -0.13 -0.06 -0.04 0.01 -0.12 -0.02 0.03 -0.05 0.03 -0.13 -0.02 -0.13 -0.08 0.02 0.00 0.06 -0.09 0.02 0.03 -0.04
309 -0.12 -0.14 -0.14 0.09 -0.04 0.01 -0.12 -0.03 0.03 -0.05 0.03 -0.13 -0.03 -0.13 -0.08 0.02 0.00 0.06 -0.09 0.02 0.00 -0.03
501 -0.14 -0.10 -0.35 1.54 -0.04 -0.05 -0.05 1.22 1.32 -0.16 0.90 0.05 -0.04 -0.14 -0.21 -0.08 -0.38 -0.06 -0.41 -0.29 -0.11
714 -0.04 0.02 0.03 -0.10 -0.05 0.01 -0.07 0.00 -0.04 -0.10 0.03 -0.07 -0.06 -0.04 -0.12 -0.22 -1.96 -0.11 -0.07 -0.13
883 0.04 2.46 0.02 0.01 0.02 -0.15 -0.12 1.49 -0.16 0.03 -0.10 -0.02 -0.34 -0.12 -0.02 -0.05 -0.04 -0.04 0.48
904 -0.09 -0.03 -0.14 -0.13 0.76 1.07 0.57 2.14 -0.10 -0.12 0.59 0.57 0.09 -0.01 -0.10 0.15 -2.51 1.16
988 0.12 0.05 0.00 0.24 0.45 0.47 0.01 -0.16 0.07 -0.01 0.00 -0.02 -0.24 0.00 0.00 -0.25 1.05
989 0.12 0.03 -0.04 -0.08 0.05 -0.10 -0.02 -0.12 0.00 -0.05 0.04 0.08 -0.11 -0.04 0.02 -0.06
1072 0.09 3.70 3.58 0.04 0.06 0.01 -0.03 -0.04 2.02 -0.13 -0.13 -0.05 1.45 -0.08 -0.04
1096 0.04 0.05 3.28 0.06 0.04 0.02 -0.42 -0.09 -0.04 -0.15 -0.02 -0.02 -0.14 -0.07
1129 0.45 1.23 -0.01 -0.03 -0.07 -0.03 -0.15 0.05 -0.33 -0.08 -0.24 -0.18 1.17
1153 2.69 0.12 0.05 0.05 0.08 0.04 0.00 -0.11 -0.02 -0.05 -0.09 1.14
1205 0.04 0.07 -0.09 2.77 0.12 0.08 -0.14 -0.04 -0.87 -0.15 0.59
1211 0.03 -0.05 0.05 -0.07 -0.09 0.00 -0.11 -0.11 0.03 0.03
1226 0.06 0.04 -0.17 -0.03 -0.09 0.04 -0.16 -0.06 0.04
1362 0.10 -0.02 -0.05 0.05 -0.16 -0.07 0.03 0.09
1387 -0.02 0.07 -0.01 0.06 0.15 1.81 -0.08
1461 0.01 -0.02 -0.06 -0.51 -0.07 0.97
1507 0.04 -0.06 0.02 0.08 -0.22
1587 0.01 -0.09 -0.09 0.01
1665 0.14 0.04 0.18
1720 0.01 2.39
1761 0.03
1795

Figure 5.4b

West Africa



	-1034	-1003	-861	-743	-717	-649	-646	-336	-281	-272	-178	-13	1	56	130	169	175	501	714	883	904	982	988	1072	1096	1129	1153	1205	1211	1226	1362	1387	1417	1507	1587	1607	1665	1720	1761	1795
-1039	-0.02	0.01	0.06	-0.14	0.06	0.06	-0.10	-0.05	-0.15	-0.20	-1.44	-0.06	0.95	-0.05	-0.28	-0.09	-0.07	-0.07	-0.25	-0.05	-0.05	0.01	-0.23	-0.07	-0.17	-3.03	-0.25	-0.18	0.56	-0.05	-0.03	-0.05	-2.27	-0.28	-0.06	-0.51	-0.04	-0.79	-0.43	-0.09
-1034		-0.02	0.00	-0.02	-0.06	0.03	0.03	-0.18	-0.30	0.16	-0.36	-0.08	0.03	-0.14	-0.12	-0.07	-0.10	-0.10	-0.99	-0.11	-0.11	-1.24	-1.05	-0.04	0.40	-0.66	-1.18	-0.84	-0.89	-0.09	-0.11	-0.07	-0.88	-1.27	-0.08	-0.05	-0.05	-2.57	-2.04	-0.26
-1003			0.02	0.01	-0.08	-0.10	2.16	1.88	-0.03	-0.16	-0.02	0.02	1.10	-0.01	0.96	-0.06	-0.03	-0.02	0.91	-0.03	-0.03	0.09	-0.12	-0.08	-0.09	-0.08	-0.15	-0.14	0.65	-0.05	0.02	-0.04	-0.11	0.55	-0.12	-0.22	-0.03	0.43	-0.19	0.50
-861				0.01	0.05	0.00	0.05	0.00	-0.22	-0.09	-0.22	0.01	-0.07	-0.24	-0.15	-0.04	-0.04	-0.04	-0.07	0.01	0.02	-0.10	-0.23	-0.06	-0.05	-0.20	-0.23	-0.39	-0.08	0.02	-1.79	0.01	-0.04	-0.27	-0.05	-0.10	0.00	-0.32	-0.14	-0.04
-743					0.00	-0.05	-0.02	0.02	0.16	-0.14	0.06	0.02	1.12	-0.02	-0.68	-0.09	-0.06	-0.05	-0.14	-0.04	-0.03	-0.34	-0.13	-0.08	-0.09	-0.10	-0.15	-0.14	-0.09	-0.05	0.03	-0.05	-1.57	-0.15	-0.05	-0.21	-0.03	-0.04	-0.19	-0.04
-717						0.04	0.02	0.15	0.00	0.05	-0.09	-0.08	-0.06	-0.03	-0.14	-0.05	-0.06	-0.06	-0.03	0.01	0.00	-0.12	-0.04	-0.05	0.00	-0.12	-0.04	-0.07	0.00	-0.01	-0.06	-0.01	-0.07	-0.05	-0.04	0.02	-0.03	-0.03	-0.10	-0.08
-649							0.04	0.13	0.03	0.05	-0.04	-0.09	0.00	-0.05	-0.06	-0.04	-1.29	-0.06	-0.02	0.01	0.00	-0.13	-0.04	-0.05	0.00	-0.13	-0.04	-0.07	0.00	0.00	-0.06	-0.01	-0.07	-0.05	-0.04	-0.08	-0.03	-0.03	-0.10	-0.08
-646								0.03	-0.03	0.03	-0.03	0.01	-0.11	0.00	-0.11	0.02	0.04	0.04	-0.03	0.25	-0.06	0.77	-0.07	0.00	0.06	-0.07	-0.05	-0.05	-0.08	-0.07	0.02	-0.04	-0.05	-0.04	-0.01	0.34	0.14	-0.13	-0.08	-0.04
-336									-0.12	-0.14	-0.10	-0.22	1.33	-0.02	0.04	-0.15	-0.16	-0.16	-0.16	-0.03	-0.05	0.77	-0.06	-0.11	0.84	-0.18	-0.24	0.00	-0.70	-0.07	-0.21	-0.07	-0.11	-0.77	-0.09	-0.38	-0.06	-0.62	-0.34	0.61
-281										0.02	0.31	0.00	1.51	0.00	-0.66	-0.01	-0.09	-0.10	-0.15	-0.03	-0.02	0.32	-0.15	-0.14	-0.10	-0.30	-0.15	-0.11	0.91	-0.05	-0.22	-0.06	-2.59	-0.19	-0.13	-0.36	-0.07	-0.47	0.56	-0.06
-272											0.05	-0.06	0.30	-0.05	0.83	0.06	0.04	0.04	0.50	-0.11	-0.12	-0.20	0.45	-0.09	-0.24	-0.38	0.35	0.56	-0.32	-0.16	-0.04	0.85	-0.46	-0.70	-0.04	-3.02	-0.21	-1.27	-1.12	-0.14
-178												0.19	2.04	0.03	-0.47	0.11	0.00	-0.01	-0.11	-0.04	-0.03	-0.49	-0.15	-0.14	-0.08	-0.68	-0.13	-0.09	0.99	-0.03	-0.21	-0.05	-2.40	-0.18	-0.15	-0.34	-0.07	-2.25	-0.32	-0.07
-13													0.02	0.03	0.09	0.04	-0.03	-0.04	-0.07	0.03	0.05	-0.07	0.00	-0.05	-0.01	-0.19	-0.06	-0.07	-0.06	0.03	0.01	0.01	-0.07	-0.08	-0.06	-0.11	0.01	-0.09	-0.25	0.00
1														3.80	2.30	0.03	0.02	0.03	0.44	-0.10	0.96	0.34	1.03	-0.07	1.04	0.96	0.89	0.85	0.83	-0.24	-0.05	-0.14	0.67	0.69	-0.04	-0.12	-0.17	0.60	0.61	0.62
56															-0.02	0.00	-0.08	-0.08	0.04	0.00	1.30	-0.15	0.38	-0.02	0.07	-0.06	-0.60	0.24	0.79	-0.03	-0.26	-0.04	-0.06	-0.68	-0.08	-0.12	-0.05	-0.67	-0.11	0.76
130																0.01	-0.01	-0.01	-0.20	-0.10	1.48	0.18	0.00	-0.15	1.04	-1.30	-0.13	0.02	-0.05	-0.12	-0.15	-0.32	-1.97	-0.29	-0.04	0.02	-0.12	-2.25	0.10	0.79
169																	0.01	0.01	0.07	0.04	0.02	-0.03	-0.06	-0.04	0.01	-0.16	-0.02	-0.09	0.06	0.05	-0.05	0.04	-0.09	-0.09	-0.03	-0.07	0.01	-0.10	-0.08	-0.05
175																		0.04	0.07	0.03	0.04	0.00	-0.07	-0.04	0.02	-0.15	-0.02	-0.04	0.06	0.04	-0.06	0.05	-0.10	-0.07	-0.05	-0.07	0.02	-0.10	-0.10	-0.07
501																			0.07	0.03	0.04	-0.10	-0.07	-0.04	0.02	-0.15	-0.02	-0.03	0.06	0.04	-0.06	0.05	-0.10	-0.07	-0.05	-0.07	0.02	-0.10	-0.07	-0.08
714																				-0.01	-0.07	-0.40	1.61	0.03	-0.10	-0.09	1.32	1.41	-0.09	0.92	-0.06	-0.03	-0.13	-0.20	0.01	-0.35	-0.03	-0.43	-0.37	-0.03
883																					0.01	-0.05	-0.03	0.02	0.06	-0.03	0.00	-0.10	-0.08	-0.08	0.02	-0.05	-0.06	-0.03	0.02	-0.17	-2.42	-0.04	-0.04	-0.03
904																						0.10	2.53	0.04	0.04	-0.04	-0.07	-0.14	1.57	-0.05	0.05	-0.07	-0.08	-0.22	0.04	-0.12	-0.04	-0.15	-0.04	0.44
982																							-0.04	0.11	-0.20	-0.10	0.79	1.12	1.12	2.19	0.04	-0.13	0.62	0.67	-0.10	-0.01	-0.04	0.03	-2.48	1.23
988																								0.20	0.00	0.06	0.28	0.32	0.56	-0.03	0.03	-0.09	-0.04	-0.04	-0.12	-0.18	-0.03	-0.05	-0.22	1.02
1072																									0.07	0.13	-0.17	0.06	0.07	0.02	0.00	0.03	0.09	-0.09	-0.02	-0.02	0.03	0.05	-0.16	-0.05
1096																										0.03	3.75	3.58	-0.01	0.06	0.00	0.04	-0.09	2.05	0.09	-0.15	0.04	1.33	-0.08	0.06
1129																											0.01	0.00	3.32	0.00	0.19	-0.04	-0.45	-0.04	0.03	-0.10	-0.03	0.00	-0.16	-0.02
1153																												0.32	1.06	0.05	0.08	0.03	-0.06	-0.15	-0.08	-0.23	-0.04	-0.20	-0.23	1.21
1205																													2.82	0.04	-0.01	0.02	-0.01	0.04	0.08	-0.13	-0.08	0.00	-0.16	1.20
1211																														0.10	0.07	0.02	2.76	0.11	-0.01	-0.08	-0.08	-0.81	-0.09	0.60
1226																															0.04	0.01	0.03	0.02	0.02	-0.09	-0.07	-0.07	-0.02	0.02
1362																																0.00	0.06	0.14	-0.02	0.09	0.04	-0.16	-0.14	0.03
1387																																	0.03	0.06	0.03	-0.08	-0.05	0.00	0.01	0.03
1417																																		0.02	0.08	-0.03	-0.02	0.15	1.77	-0.04
1507																																			0.04	-0.13	0.04	-0.70	-0.17	1.04
1587																																				0.08	0.02	0.07	0.10	0.00
1607																																					0.00	-0.05	-0.12	-0.08
1665																																						0.03	0.01	0.06
1720																																							0.03	2.43
1761																																								0.02

Figure 5.4c

Europe

	-646	-464	1	130	169	501	904	1072	1183	1211	1387	1417	1483	1490	1617	1761
-1003	-0.06	-0.01	1.06	-0.09	0.06	-0.38	-1.87	0.04	-0.09	-0.25	-0.14	-0.14	0.03	-0.13	-0.12	-2.26
-646		0.10	-0.11	-0.09	0.02	0.00	1.41	-0.09	0.00	-0.03	-0.01	-0.01	0.06	0.06	-0.02	-0.06
-464			0.04	-0.11	-0.11	-0.09	-0.09	0.03	0.07	-0.09	-0.02	-0.02	-0.03	-0.03	-0.03	-0.03
1				0.01	0.12	-0.19	1.40	0.30	-0.05	-0.11	0.01	0.01	0.12	0.01	0.00	0.26
130					0.04	0.02	-0.03	0.07	0.05	-0.07	-0.86	-0.87	-0.02	-0.02	-0.96	-0.02
169						0.03	0.06	0.06	0.04	-0.06	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02
501							1.19	-0.29	-0.08	0.03	-0.08	-0.09	-0.09	0.01	-0.10	-0.28
904								-0.02	-0.04	2.46	0.00	-0.01	0.05	0.05	-0.03	-1.35
1072									0.01	4.04	0.04	0.04	0.05	0.05	0.06	-0.20
1183										0.06	0.09	0.08	0.06	0.06	0.04	-0.05
1211											0.17	0.16	0.14	0.14	0.08	2.25
1387												-0.10	0.02	0.02	-0.43	0.09
1417													0.03	0.03	-0.46	0.09
1483														0.05	0.01	0.09
1490															0.01	0.09
1617																0.07

Asia

	-646	1	130	501	904	988	1072	1096	1226	1387	1417	1617	1761
-1003	-0.03	-0.03	0.04	0.96	0.98	-0.04	-0.23	-0.05	-0.01	-0.06	-0.06	-0.06	0.69
-646		-0.06	0.12	-0.20	-0.33	-0.53	-0.21	-0.95	-0.17	-0.64	-0.64	-0.67	-0.15
1			0.01	-0.12	-0.11	0.04	0.13	-0.08	0.01	0.07	0.07	0.07	0.05
130				0.03	-0.03	-0.07	0.01	-0.02	-0.08	-0.05	-0.05	-0.06	0.07
501					-0.06	-0.01	0.32	1.44	-0.11	-0.09	-0.09	-0.08	-0.77
904						0.03	-0.07	-0.02	-0.01	0.02	0.01	0.00	0.00
988							0.03	0.22	-0.10	0.90	0.85	0.48	-0.10
1072								0.02	0.13	0.09	0.08	0.03	-0.19
1096									0.14	0.00	-0.03	-0.20	-0.11
1226										0.06	0.06	0.09	0.03
1387											-0.07	-0.57	-0.09
1417												-0.49	-0.08
1617													0.00

America

	169	501	904	1072
-1003	0.03	-0.32	0.47	-0.19
169		0.10	0.05	-0.14
501			-0.19	-0.19
904				0.06

Figure 5.4d

Figure 5.4: Intragenic linkage disequilibrium and recombination matrices for the *NATP1* locus, generated using LDhat (McVean, 2002). Reported values are for each pair of SNPs, where BLUE indicates greater than expected linkage disequilibrium, and RED indicates greater than expected recombination. Analyses were performed for each geographic grouping following Table 2.1: (a) Africa; (b) East Africa; (c) West Africa; (d) Europe, Asia and the Americas.

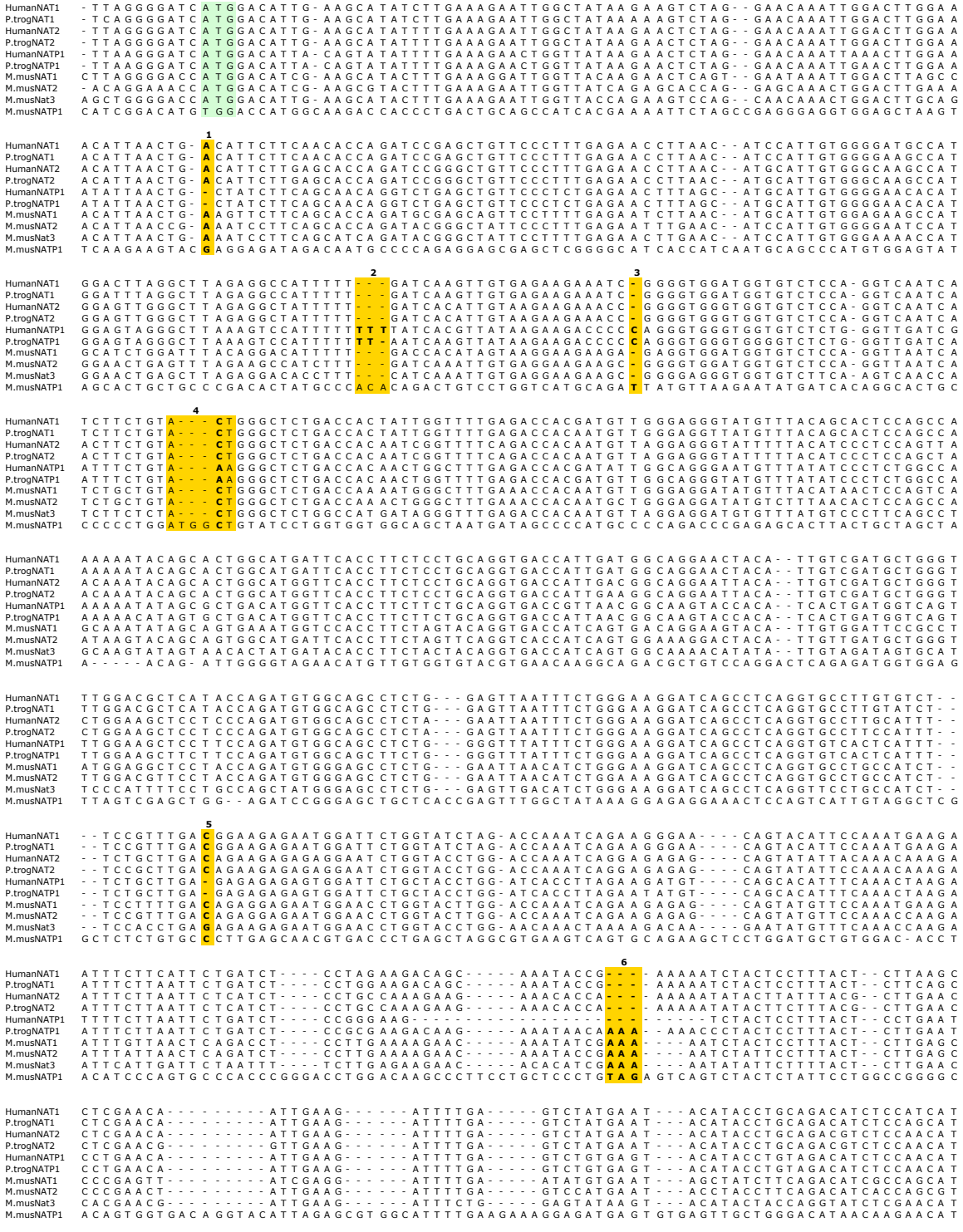
DISCUSSION

We have sequenced an ~3 kb region containing the *NAT* pseudogene, *NATP1,* in a global population dataset comprised of 608 chromosomes. We have characterized the overall haplotype structure among populations (Figure 5.3a) and observed the *NATP1* haplotypes present in humans (Figure 5.3b). Additionally, we have aligned the pseudogene in human and *Pan troglodytes* (chimpanzee) with the functional *NAT* genes, *NAT1* and *NAT2*, in both species (Figure 5.5). We have determined the phylogenetic relationship between human and chimpanzee for these loci, using the mouse outgroup, and have confirmed with this analysis that the duplication leading to *NATP1* occurred before the human-chimpanzee divergence, as indicated by the clustering of homologous genes in the phylogeny presented in Figure 5.6.

The Evolutionary Role of Pseudogenes

Understanding the forces that shape patterns of variability within and between populations and species is a major goal of evolutionary genetics (Haddrill et al., 2005). Focus in human genetics has long been in understanding the genetic basis of phenotypic variation in human populations. Gene duplication, and subsequent divergence, underlies the formation of gene families and is thought to be an important source of genetic complexity (Thomas, 2007).

Human genomes contain a large number of these gene homologies, referred to as pseudogenes, which are heritable, and presumably non-functional, neutrally evolving, and disabled at the transcriptional level (Balakirev and Ayala, 2003; Zhang and Gerstein, 2004) (see (Podlaha and Zhang, 2004) for an exception). There are two main types of



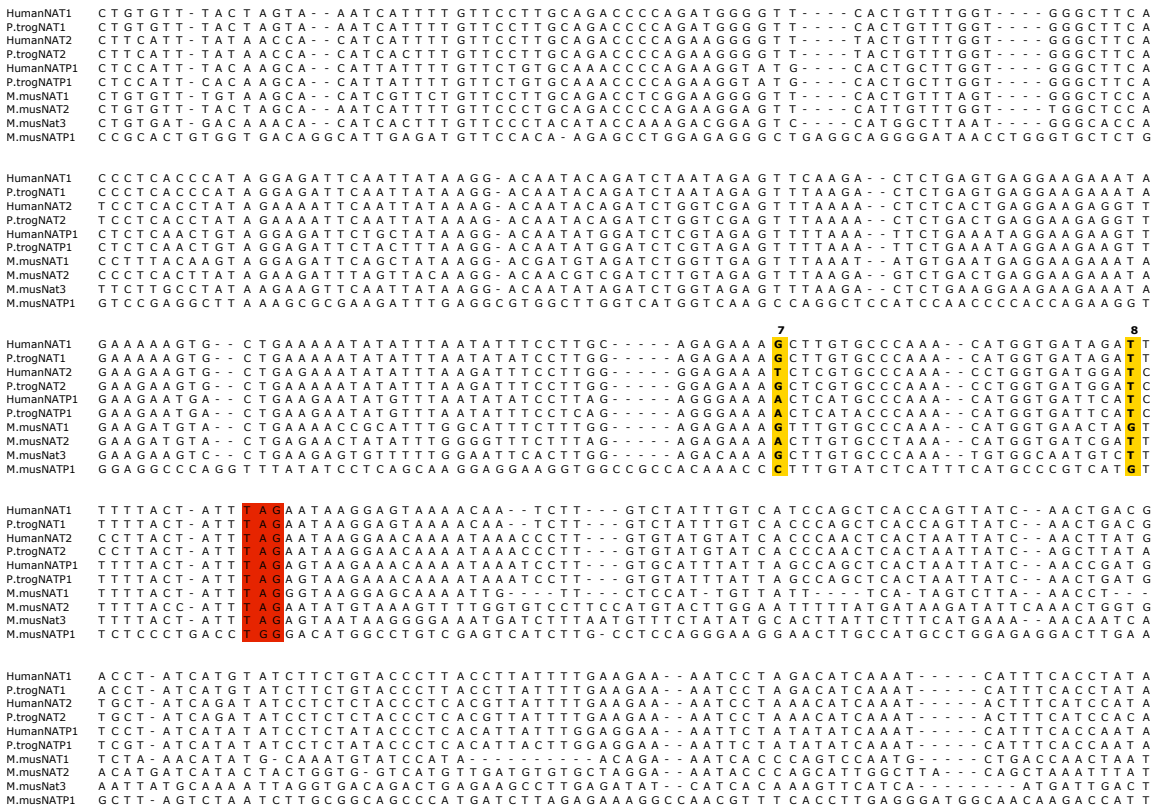


Figure 5.5: Consensus alignment of *NATP1* with *NAT1* and *NAT2* in human, *P. troglodytes, and M. musculus* performed using ClustalX version 1.83.1 (Chenna et al., 2003). Eight detrimental mutations, defining *NATP1* in humans, are highlighted in orange and numbered according to Blum et al. (1990). Consensus +1 ATG Start (shown in green) and +870 TAG Stop (shown in red) are also annotated.

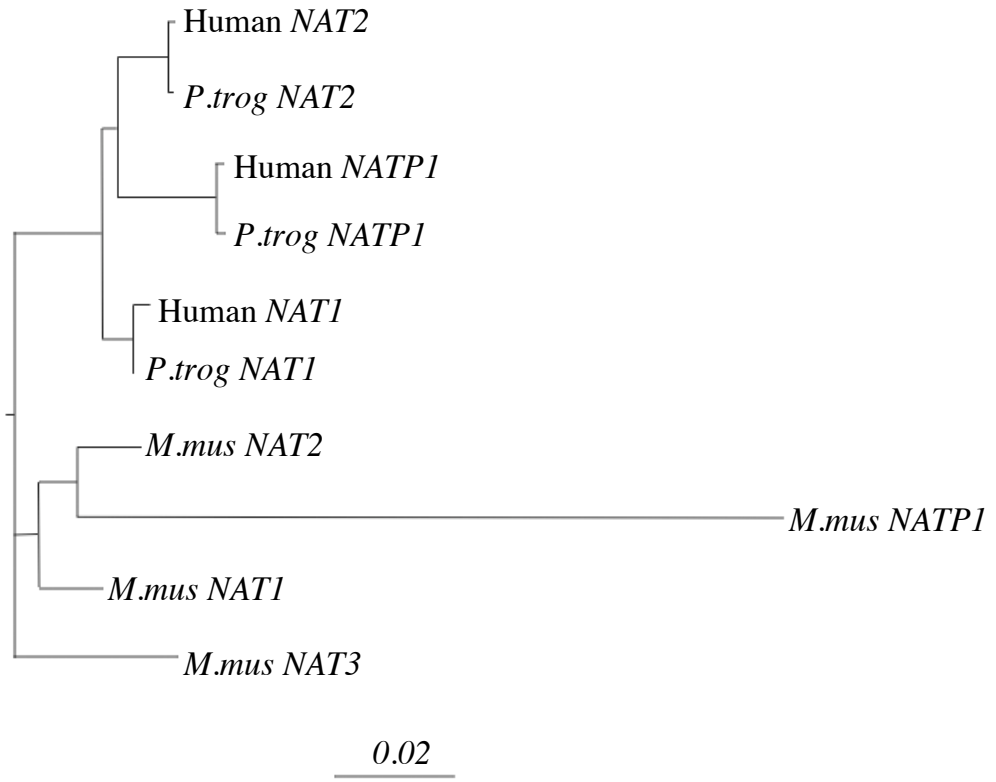


Figure 5.6

Figure 5.6: Phylogeny of the *NAT1, NAT2*, and *NATP1* gene sequences in human, chimpanzee, and mouse constructed using PAUP* 4.0 (Rogers and Swofford, 1998). The neighbor-joining algorithm was applied and set to distance, using the Kimura two-parameter distance to correct for differences in rates of transitions and transversions (Kimura, 1980).

pseudogenes typically recognized, processed (retrotransposons) and unprocessed (tandem duplications). The *NATP1* locus is considered an unprocessed pseudogene, whereby it is thought to have arisen via tandem duplication; however, whether *NATP1* arose by duplication from *NAT1* or *NAT2* has not been determined. Duplications (and losses) of functional genes are thought to occur in order to provide novel function (Ohno, 1970), perhaps in response to changes in environmental pressures.

Once a duplication is "turned-off", creating a pseudogene, why then is the structure evolutionarily conserved at all? This can be explained by temporal factors affecting the duplication and pseudogenization of a locus, however, the long-term phylogenetic stability of different genes is observed to vary greatly (Thomas, 2007). Thomas (2007) termed the rapid fluctuation in duplication and loss "birth-death evolution", and asserts that the cause for this process is not well understood. In his comparison of xenobiotic metabolizing CYP P450 genes in 11 vertebrate genomes, Thomas (2007) found that birth-death evolution of genes acting on xenobiotic substrates was much higher than in those implicated in the processing of endogenous substrates; in other words, genes involved in metabolism of external toxins were observed to be more unstable in terms of their frequent duplication and loss. Further, Thomas (2007) argues that because the xenobiotic metabolizing CYP450 genes are so unstable, they are particularly subject to the forces of natural selection, specifically positive selection for replacement changes, as a response to changing environmental conditions (*i.e.* changes in xenobiotic exposure) or pathogen response.

Variation at the *NATP1* Pseudogene

Two extensive studies of *NATP1* have been carried out to date, the first characterizing the locus as a pseudogene (Blum et al., 1990), and the second, a more recent study that uses *NATP1* for comparative purposes in a human population genetic analysis of *NAT1* and *NAT2* (Patin et al., 2006a). Blum *et al*. (1990) observe eight unique mutations (Table 5.3), which they deem are detrimental to the function of *NATP1* and define it as a pseudogene in humans. Table 5.3 illustrates the mutations observed first by Blum (1990) annotated on the consensus alignment of the *NATP1* reference (X17060) with that of *P. troglodytes* (Ptr8-WGA990) shown in Figure 5.5. From Figure 5.5, it is clear that six of these mutations are shared, and common between the *NATP1* human and chimpanzee pseudogene sequence. It is likely that there are species-specific, deleterious mutations that define *NATP1* in chimpanzee. We also observe a previously reported variable mutation (rs10088180) at +1 of the ATG start codon to be variable in this human dataset. This SNP was not identified by Blum (1990), but Patin *et al*. (2006a) observe this mutation (their NATP698 A>G) at appreciable frequencies (45-75%) in all populations included in their analyses of Africans and Eurasians.

Figure 5.6 illustrates the phylogenetic neighbor-joining tree of *NAT* genes in the human and chimpanzee reference individuals. Because sequences are clustering by gene homology, and not by species, it is clear that *NATP1* was most likely turned-off prior-to speciation. Though we cannot absolutely determine the timing of the duplication of the *NAT* loci, based on the phylogenetic analyses presented here (Figure 5.6), *NAT1* appears to be ancestral. *NAT1* duplicated to form a proto-*NAT2/NATP1* gene and the latter duplicated to give rise to *NAT2* and *NATP1* prior to the human/chimpanzee divergence.

Detrimental Mutations Defining *NATP1*			
Number	*Mutation and Position Reported*[1]	*Effect*[1]	*Mutation and Position Reported*[2]
1	77 (1bp del)	A frameshift (-1)	76 (1bp del)
2	169-171 (TTT) ins	ins 1 codon (F)	168; (TT-) in chimp
3	189 (1bp ins)	C frameshift (+1)	193 (1bp ins)
4	231 (C-->A)	Termination	230 (C-->A)
5	461 (1bp del)	Frameshift (-1)	461 (1bp del)
6	562-564 (AAA ins)	ins 1 codon (K)	562-564 (AAA ins)
7	830 (G/T-->A)	Termination	839
8	864 (1bp del)	Frameshift (-2)	864 (not observed)

[1]Blum *et al.* (1990)
[2]Mutations observed in human (Genbank:X17060) and P.troglodytes (Genbank:Ptr8-WGA990)

Table 5.3

Table 5.3: Location and effect of detrimental mutations that identify *NATP1* as a pseudogene in humans, according to Blum (1990). Positions are identified in the present study in the consensus alignment presented in Figure 5.5, in humans, and chimpanzee.

Additionally, we are unable to determine from the present analyses when the pseudogenization of the *NATP1* locus occurred. However, based on the relatively long branch length of *NATP1* in mouse, it is clear that *NATP1* in humans and chimpanzee has not been a pseudogene for as long of a duration as *NATP1* in mouse. Though unlikely, the possibility that *P. troglodytes* or other non-human primate species still possess a functional copy of *NATP1* cannot be ruled out without further analyses.

The MDS plot shown in Figure 5.2 illustrates the elevated population differentiation at the *NATP1* locus relative to *NAT1* and *NAT2*, whereby populations cluster according to major geographic regions (African, European, Asian, and Amerindian). Additionally, F_{ST} values approximate those observed for other human pseudogenes (Martinez-Arias et al., 2001), and the global average for human autosomal loci (Consortium, 2005).

Tests of neutrality for the *NATP1* locus (Table 5.2) are inconsistent across estimators. In general, inter-species comparisons (D and F) are more consistently positive, though still not significant. Negative TD values are observed for most African populations, perhaps indicative of population growth. Additionally, positive TD values outside of Africa could be interpreted as being consistent with recovery from a bottleneck following movement out of Africa (Campbell, 2008).

The haplotype phylogeny of the most common *NATP1* haplotypes, presented in Figure 5.3b, has three cycles, or areas of reticulation. Reticulation can be indicative of homoplasy or possible recombination (Bandelt, 1995). Three possible recombination events are reflected as reticulations in this network at positions +1 and +1003; +1 and +904, and +1 and +1761. Linkage disequilibrium and recombination estimates presented

in Figure 5.4, highlight similarity with the network analysis, in that East Africa has higher than expected linkage disequilibrium between sites at positions +1761 and +904, and European groups show higher than expected linkage values between sites at positions +1761 and -1003.

Conclusion

In summary, from the phylogeny presented in Figure 5.6, we have shown that *NATP1* in humans and *P. troglodytes* cluster by gene homology, relative to the other *NAT* loci. We deduce that the duplication event that led to *NAT2* and *NATP1* preceded the human-chimpanzee divergence, ~4.1 MYA (Hobolth et al., 2007). Additionally, the alignment of *NATP1* with *NAT1* and *NAT2* in both species indicates that the chimpanzee possesses six of the eight mutations that define the pseudogene in humans. Though further investigation is necessary to determine the potential for function of this region in non-human primates, it is clear that humans have experienced very different environmental pressures from chimpanzee over the course of recent evolution. Furthermore, we have observed an elevated level of population differentiation for the *NATP1* locus in humans, relative to *NAT1* and *NAT2*, which underlines the utility of this region, and pseudogenes in general, in the inference of human demographic history.

CHAPTER VI: CONCLUSIONS AND IMPLICATIONS FOR FUTURE STUDY

The *N-acetyltransferase* loci (*NAT1*, *NAT2*, and *NATP1*), spanning approximately 200 kb on chromosome 8p22, have been observed in the present study to exhibit very different patterns of intra-genetic diversity and natural selection in humans. It is critical that future efforts focus on determining the biochemical pathways involving NAT1 and NAT2, including the role of CYP1A2 may have in potential pathways and in relation to various substrates. With this information we can then begin to establish the role of NAT1 and NAT2 in detoxification and reactions leading to carcinogenicity in humans.

NAT1 exhibits low levels of variation in the coding region within human populations overall, and shows the effects of purifying selection acting to prohibit the accumulation of variation in the coding region of this gene. The role of NAT1 in many different cell types and at various stages of development, as well as its endogenous role in the metabolism of folate, are consistent with the observed constraint, preventing variation in the coding region of this locus. As a corollary to this observation, the *NAT1* locus (exon 9) is known to use differential promoters, alternative splicing, and alternative polyadenylation to increase variation of NAT1 transcript levels and patterns of expression. Most of the variation that we observe is due to SNPs located in the 3' untranslated region (UTR) of the gene that are thought to cause differential polyadenylation of *NAT1* transcripts. We observe that the pattern of variation in this region is consistent with the effects of balancing selection acting to maintain two 3' haplotypes (AAT or TCA) at three specific SNPs located at positions +1088, +1095, +1191, which are typically found in association with each other. Additionally, we observe variation in the $(TAA)_n$ repeat directly preceeding this 3' region of *NAT1* in non-

human primates, where only humans possess a polymorphism at the last $(TAA)_n$ that includes the +1088 SNP (**T**AA). This variant underlines the potential importance of alternative polyadenylation at *NAT1* in humans for generating phenotypic diversity. Though we know that the +1088A variant is sufficient to obliterate the first poly-A signal in humans, the exact effect of this 3' variants on phenotype is far from established (Barker, pers. comm.; Boukouvala and Fakis, 2005). Future work should focus on clarifying the potential role of the +1088, +1095, +1191 haplotypes on NAT1 acetylator phenotype and expression patterns. In addition, further research is needed to clarify promoter usage, the role of expression of the eight non-coding exons of *NAT1* in specific tissues, differential polyadenylation, and to distinguish other 3' motifs that may have an effect on NAT1 phenotype in humans and other species.

This study of *NAT2* variation has uncovered several previously unidentified, coding region variants, providing for an unbiased description of *NAT2* sequence variation. By doing so we have identified particular variants that may be important for other researchers attempting to determine effect of genotype on function, specifically in African populations. Interestingly, and in contrast to *NAT1*, we observe a high level of non-synonymous substitutions in the coding regions of *NAT2* (15 non-synonymous SNPs vs 3 synonymous SNPs). Additionally, we observe a significantly elevated level of diversity when compared to *NATP1*, using an HKA test (Table 4.6). We have found *NAT2* to exhibit a complex pattern of selection, whereby balancing selection, and possibly positive directional selection, are acting in a population specific manner. To disentangle the role that selection has played in shaping *NAT2* variation in different populations, and to investigate any association that particular phenotypes may have with

dietary adaptation, future research on *NAT2* variation should attempt to distinguish demographic processes that have affected the locus in different populations from the effects of natural selection. The best way to approach this would be to model natural selection under different demographic scenarios and compare expectations to the observed patterns of variation at the population level. Additionally, inference of the age of particular SNPs that influence phenotype may help to identify the timing of environmental changes that may have influenced the current pattern of diversity at *NAT*2.

The study of the *NAT* pseudogene, *NATP1*, has indicated that this locus exhibits an elevated level of population differentiation, and underlines the utility of this locus as a neutrally evolving region for the study of human demographic processes. Additionally, our study of intragenic linkage disequilibrium and recombination indicate that this region has tolerated a high level of recombination, as expected for a pseudogene. This study indicates that the duplication leading to the *NATP1* locus occurred before the chimpanzee-human divergence ~4.1 MYA(Hobolth et al., 2007). Future work with the *NATP1* locus should focus on comparative analysis of *NATP1* and other *NAT* loci across diverse species in order to clarify the timing of pseudogenization of *NATP1* in humans and of *NAT* duplication events in general.

APPENDIX 1

THE POLYPHRED TO SEQUENCE PIPELINE

WRITTEN BY HOLLY M. MORTENSEN

PURPOSE OF THE TUTORIAL:

The following tutorial will assist you in taking your Polyphred output data, following alignment and base calling using the PHRED/PHRAP/CONSED/Polyphred suite, to the necessary format for input into the program PHASE, for the inference of SNP haplotypes for your dataset. Instruction is given for PHASE 2.11(Stephens and Donnelly, 2003) but can easily be adapted to the less computationally intensive, more recent version fastPHASE (Scheet and Stephens, 2006). Following "PHASEing" you may choose to perform the last sequence of operation to take your inferred SNP haplotypes back to full sequence format. The two perl scripts used in the following exercise were written by Floyd Reed and are not included here.

INSTRUCTIONS:

1. With your edited polyphred file, create a separate folder that contains the following:

phred2phase2.pl (perl script)
polyphred.out (your edited output from CONSED)

2. from the terminal window, navigate to your folder
execute the following command:
>perl phred2phase2.pl

This command will create the following files in your folder:
phaseinput.txt
polyphredout.txt
polytable.txt

3. Open phaseinput.txt in a text editor, remove the first string of s's from the top of the document, insert number of individuals, number of SNPs, and 'P' followed by the position or relative location of each SNP included according to your reference. In the fourth line, the model applied to each site is indicated by the following string of S or M. The M character relaxes the step-wise mutation model and allows for a parent independent model (PIM) and is more appropriate for data that is not strictly bi-allelic. Additionally, tri-allelic or multi-state data need be recoded using numerical character, 1, 2, 3, 4 to represent any A,G,C, or T. Missing data is indicated by -1. Your edited haplotypes should follow the example below:

```
3
23
P 2213 2217 2224 2263 2291 2318 2332 2341 2393 2417 2435 2828 2835 2917 2983 3221 3225 3281 3497 3705 3719 3900 3923
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSMSSSSSSSSSMSSSSSS
BG001
GAACCAAGAT3GCAGGAGCA1AG
GAACCGCTAT4GCAGGAGCA1AG
BG003
```

GAACCAAAGAT3GCAGGAGCA1AG
GAACCACTAT3GCAGGAGTA1AG
BG004
GAACCGAAGAT3GCAGGAGCA1AG
GAACCGCTAT3GCAGGAGCA-1AG

4. Create a file called eg.delta. In this file translate the S and M line from the previous step to binary, 0 and 1.
So the contents of your eg.delta file, using the example above would be:

000000000000000000000000000000010000000001000000

**Note: For fastPHASE only, create a file called popinfo.txt. This file allows you to tell PHASE which populations your input derive from. A pop info file for five populations, each with six individuals would look like this:*

111111555555222222333333444444

5. Create a new folder that includes the following items:
PHASE
phaseinput.inp (edited)
eg.delta

6. From the terminal window, navigate to that folder and type the following command:
> ./PHASE -deg.delta -x5 phaseinput.inp phaseoutput.out

This will create seven output files. The phased data will be called according to what is specified in the command line. In this example the phase output is called:
phaseoutput.out

*Note: At this point it may be a good idea to get your phased haplotypes in order for double-checking. Using maclade or excel, and organize your haplotypes to check for accuracy against the polyphred polymorphism numbers (e.g. number of mutations are correct in both files)

7. Create a new file that includes the following four items:

phase2seq.pl (perl script to take phased haplotype data and reinsert into reference)
positions.txt (polyphred output positions)
reference.txt (same reference used in consed alignment! Remove all line breaks)
phased.txt (phaseoutput.out file unedited and renamed)

from the terminal window, navigate to this folder, and execute the command:
>perl phase2seq.pl

the output file will be named:
seq-output.txt

*Note: You may have to edit out spaces in this file. Other than this the sequence output is basically in phylip format, with each individual's inferred biallelic sequence (*e.g.* Sample1a, Sample1b, Sample2a, Sample2b, etc.)

APPENDIX 2
NAT OBSERVED SNP RESULTS

NAT1

SNP	NAT nomenclature	Human variant	ancestral (P. trog.)	refSNP #
1	-1048	C/G	G	rs8190843
2	-1044	A/C	A	NA
3	-1037	A/G	A	NA
4	-1144	C/G	C	NA
5	-970	C/T	C	NA
6	-943	C/T	C	rs8190844
7	-929	A/G	G	rs8190845
8	-920	C/G	C	NA
9	-868	A/G	C	rs8190846
10	-844	A/G	G	rs8190847
11	-826	C/T	C	rs8190848
	NOT OBS	A/G	T	rs8190849
	NOT OBS	A/G	A	rs8190850
	NOT OBS	C/T	C	rs28359533
	NOT OBS	A/G	G	rs8190851
	NOT OBS	C/G	G	rs8190852
	NOT OBS	C/T	C	rs8190853
	NOT OBS	A/G	A	rs8190854
12	-433	C/T	T	rs8190856
13	-426	C/T	C	NA
14	-344	C/T	T	rs4986988
15	-278	A/T	G	rs17126356
16	-40	A/T	T	rs4986989
17	-36	A/T	A	rs8190857
18	+21	G/T	T	rs4986992
19	+236	C/G	G	NA
	NOT OBS	C/T	C	rs8190858
	NOT OBS	T/C		NA
20	+445	A/G	A	rs4987076
21	+459	A/G	G	rs4986990
	NOT OBS	C/T	C	rs5030839
	NOT OBS	A/G	C	rs4986782
	NOT OBS	C/T	C	rs4987195
22	+639	G/T	G	rs4986783
23	+662	A/G	A	NA
24	+758	A/T	A	NA
25	+777	C/T	T	rs4986991
	NOT OBS	C/T	C	rs4986994
26	+956	C/G	C	NA
	NOT OBS	A/T	T	rs11985819
	NOT OBS	A/T	T	rs3197750
27	+1088	A/T	T	rs1057126/rs8190861
28	+1095	A/C	A	rs15561
29	+1191	G/T	T	rs4986993
30	+1236	A/G	A	rs4987077
31	+1245	C/T	T	NA
	NOT OBS	A/G	A	rs28359534
	NOT OBS	C/G	G	rs8190862
	NOT OBS	A/G	A	rs6982949
	NOT OBS	C/T	C	rs8190863
32	+1454	C/T	T	rs8190864
33	+1527	A/G	A	NA
34	+1572	C/T	T	NA
35	+1641	A/C	A	rs8190865
36	+1654	A/G	G	NA
	NOT OBS	A/G	G	rs8190867
37	+1685	A/G	G	NA
38	+1688	A/G	A	rs8190868
39	+1692	A/G	G	rs8190869
40	+1715	C/T	T	rs8190870
41	+1734	A/T	A	rs8190871
42	+1784	A/G	A	NA
43	+1787	A/T	A	rs8190872
44	+1834	A/G	A	rs8190873
45	+1910	A/G	A	rs9650591
46	+1960	C/G	G	rs8190874
47	+1961	A/G	G	rs8190875
48	+1969	G/T	G	NA

NAT2 SNP	NAT nomenclature1	Human variant	ancestral (P. trog.)	refSNP #
1	-1014	C/T	C	NA
2	-862	C/G	G	rs41319551
3	-859	C/T	C	rs11782802
4	-855	C/T	C	rs973874
5	-701	A/G	G	NA
6	-491	A/G	G	rs1495744
7	-413	A/G	G	rs45529434
8	-349	C/T	T	NA
9	-282	C/T	T	NA
10	-255	C/G	C	rs45605531
11	-234	C/T	C	rs7832071
12	-179	C/T	T	rs45472694
13	+70	A/T	T	rs45477599
	NOT OBS	A/T	T	rs1805158
14	+191	A/G	G	rs1801279
15	+282	C/T	C	rs1041983
16	+308	C/T	C	NA
17	+341	C/T	T	rs1801280
	NOT OBS	C/T	A	rs45532639
	NOT OBS	A/G	G	rs4986996
18	+403	C/G	C	rs12720065
	NOT OBS	A/T	A	rs4986997
19	+472	A/C	A	NA
20	+481	C/T	C	rs1799929
21	+518	A/G	A	NA
22	+578	C/T	C	NA
23	+590	A/G	G	rs1799930
24	+609	G/T	G	rs45618543
	NOT OBS	A/T	A	rs45607939
25	+632	A/G	G	NA
26	+664	G/T	T	NA
	NOT OBS	C/T	C	rs45518335
27	+766	A/G	A	NA
28	+803	A/G	A	rs1208
29	+838	A/G	G	NA
30	+857	A/G	G	rs1799931
31	+1021	C/T	T	rs2552
32	+1085	A/G	G	NA
33	+1101	C/G	C	rs45539742
34	+1134	C/T	C	NA
35	+1338	A/G	A	rs45543435
36	+1362	A/C	C	NA
37	+1373	G/T	T	NA
38	+1395	A/G	G	rs4646247
39	+1447	C/T	C	rs971473
40	+1520	A/T	A	NA
41	+1521	C/G	C	rs45547533
	NOT OBS	C/T	T	rs45438302
42	+1781	C/T	T	NA
43	+1783	C/G	G	NA
44	+1792	A/G	A	NA
45	+1794	C/T	T	NA
46	+1853	C/T	T	NA

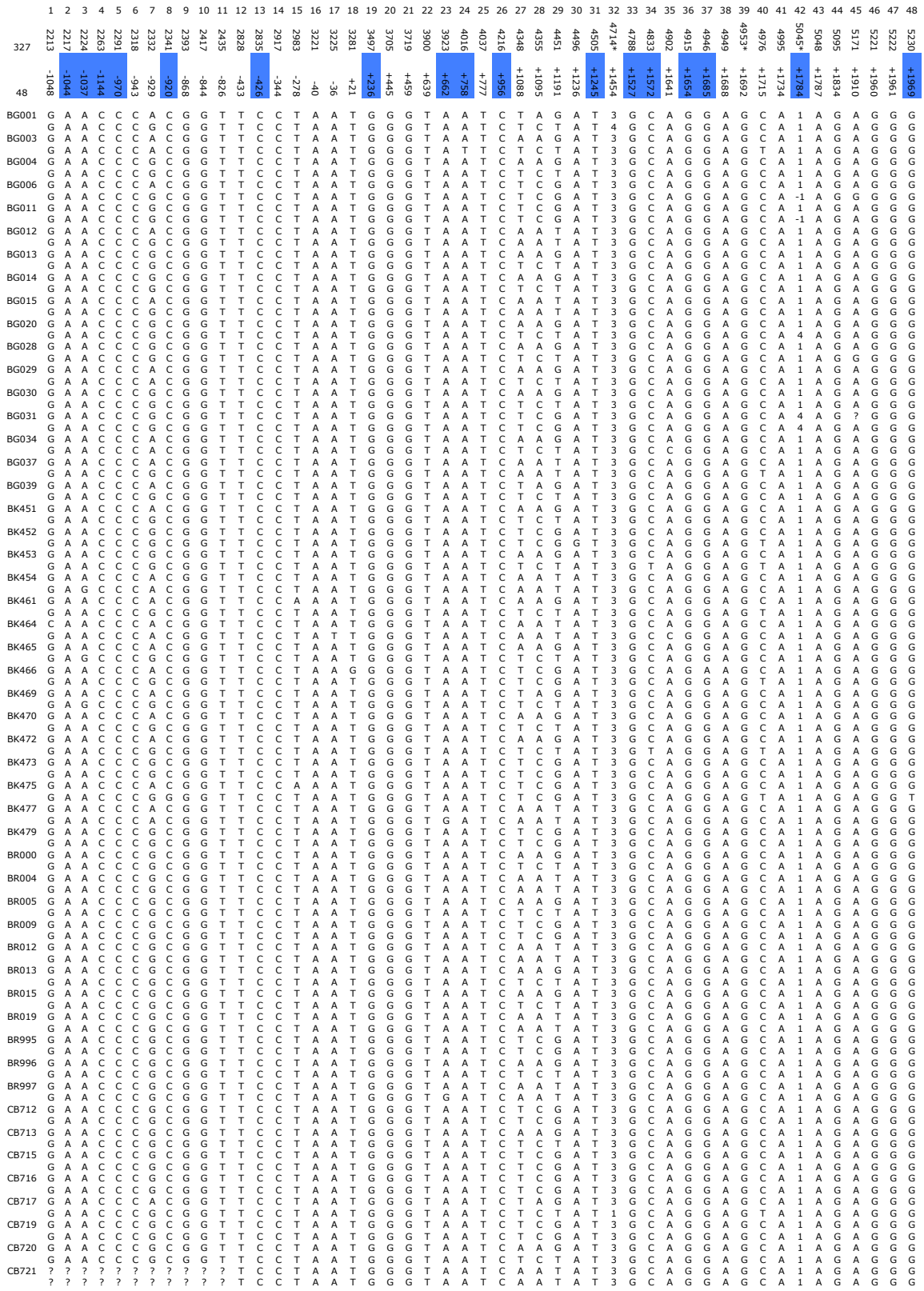
NATP1

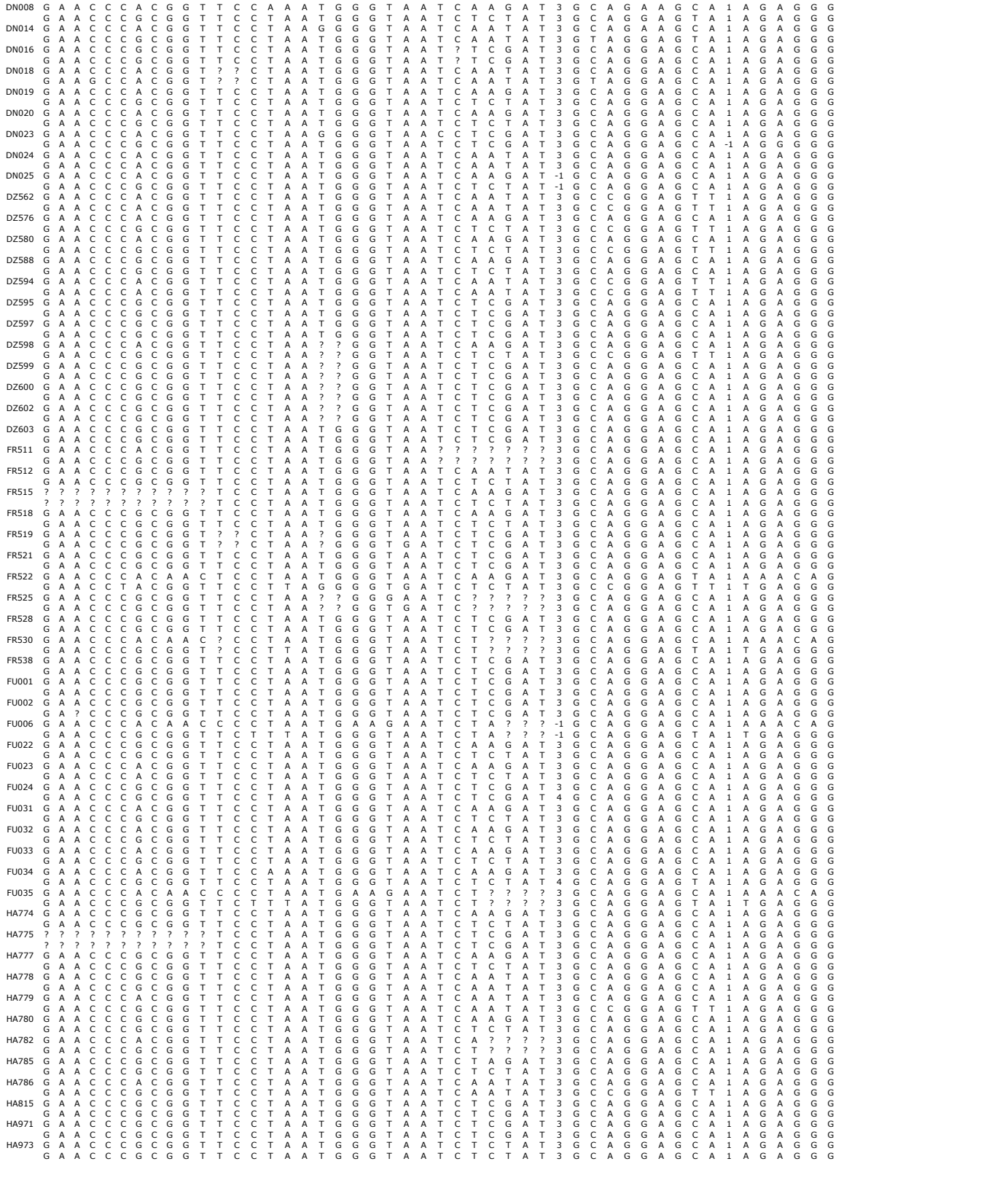
SNP	NAT nomenclature	Human variant	ancestral (P. trog.)	refSNP #
1	-1039	C/T	C	NA
2	-1034	A/T	T	NA
3	-1003	A/G	A	rs13278827
4	-984	A/T	T	NA
5	-976	C/T	T	NA
6	-861	A/T	T	NA
7	-805	A/C	T	NA
8	-743	C/G	G	NA
9	-717	G/T	G	NA
10	-649	A/G	A	NA
11	-646	C/T	C	NA
12	-464	C/T	T	NA
13	-336	C/T	A	NA
14	-281	C/T	T	rs12334336
15	-272	C/T	T	rs4487818
16	-178	A/G	A	NA
	NOT OBS	A/C	A	rs13281417
17	-13	C/T	T	NA
18	1 (ATG)	A/G	A	rs10088180
19	+56	A/G	G	NA
20	+130	C/T	T	NA
21	+169	A/T	A	NA
22	+175	A/C	A	rs2898473
23	+293	A/G	G	NA
24	+309	C/T	G	NA
25	+501	C/T	G	rs35548819
26	+714	C/T	T	NA
27	+883	C/G	G	NA
28	+904	C/T	T	rs2172426
29	+982	A/G	A	NA
30	+988	C/T	T	NA
31	+989	A/C	A	NA
32	+1072	C/T	T	rs12548816
33	+1096	A/G	G	NA
34	+1123	C/T	T	NA
35	+1129	G/T	T	NA
36	+1131	G/T	T	NA
37	+1153	A/G	G	NA
38	+1183	C/T	T	rs13254216
39	+1205	A/C	A	NA
40	+1211	C/T	C	NA
41	+1226	C/T	T	NA
42	+1362	A/T	T	NA
43	+1387	C/G	C	rs17126565
44	+1417	C/G	T	rs17126568
45	+1461	C/T	T	NA
46	+1483	C/T	T	NA
47	+1490	A/C	A	NA
48	+1507	A/G	A	NA
49	+1587	A/C	A	rs17126570
50	+1607	A/G	G	NA
51	+1617	A/G	A	rs17126572
52	+1665	C/G	G	NA
53	+1720	C/T	T	NA
54	+1761	A/T	A	rs28504072
55	+1795	A/G	G	NA

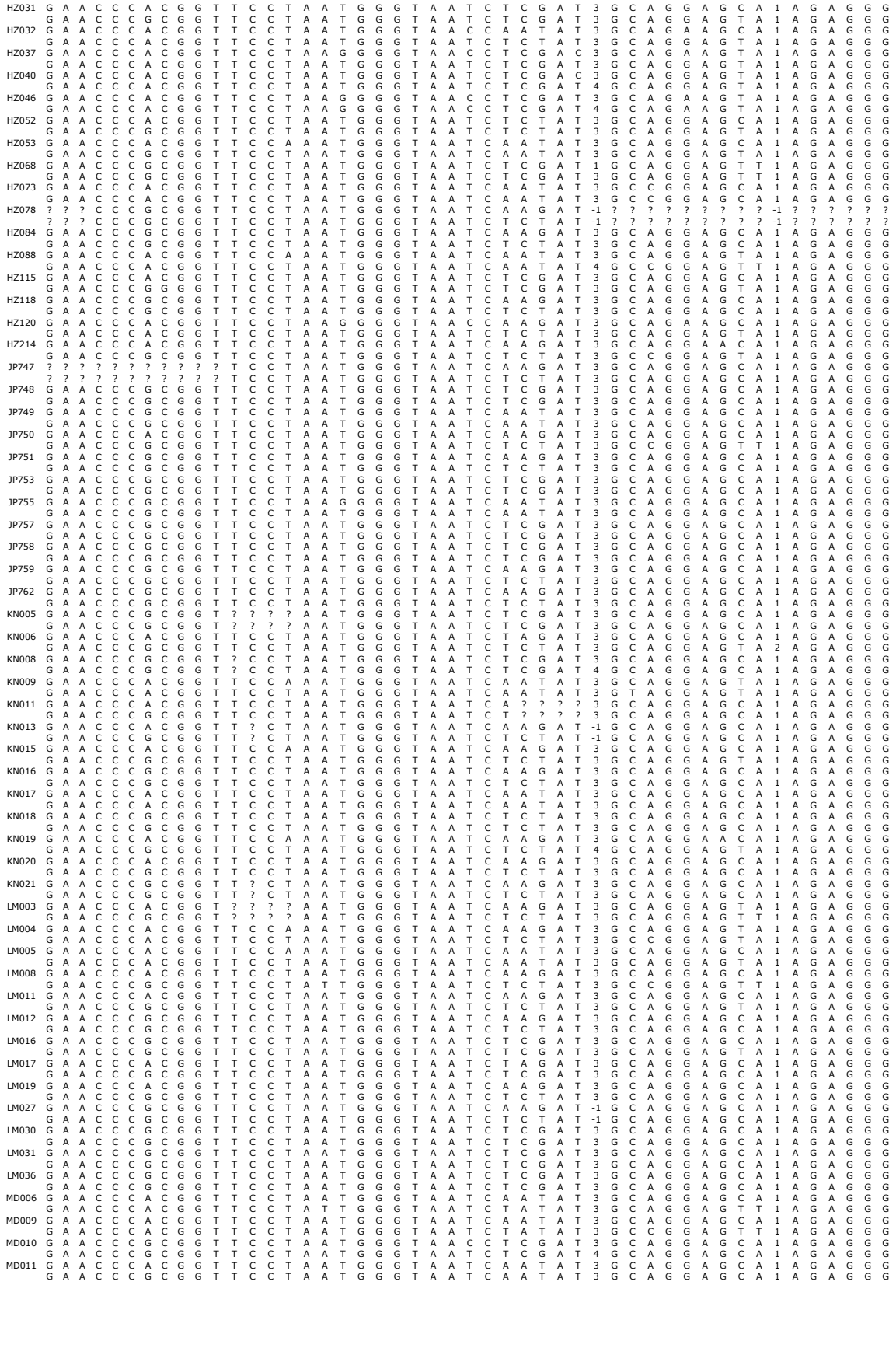
Appendix 2: Blue indicates SNPs that are novel to the present dataset. Grey indicates SNPs previously reported that were not observed in this resequencing study. All SNPs are numbered according to the +1 ATG start site and according to accepted *NAT* nomeclature. *NATP1* is numbered according to the consensus +1 ATG, determined by clustalX alignment with both *NAT1* and *NAT2*. All reference numbers were obtained from dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/).

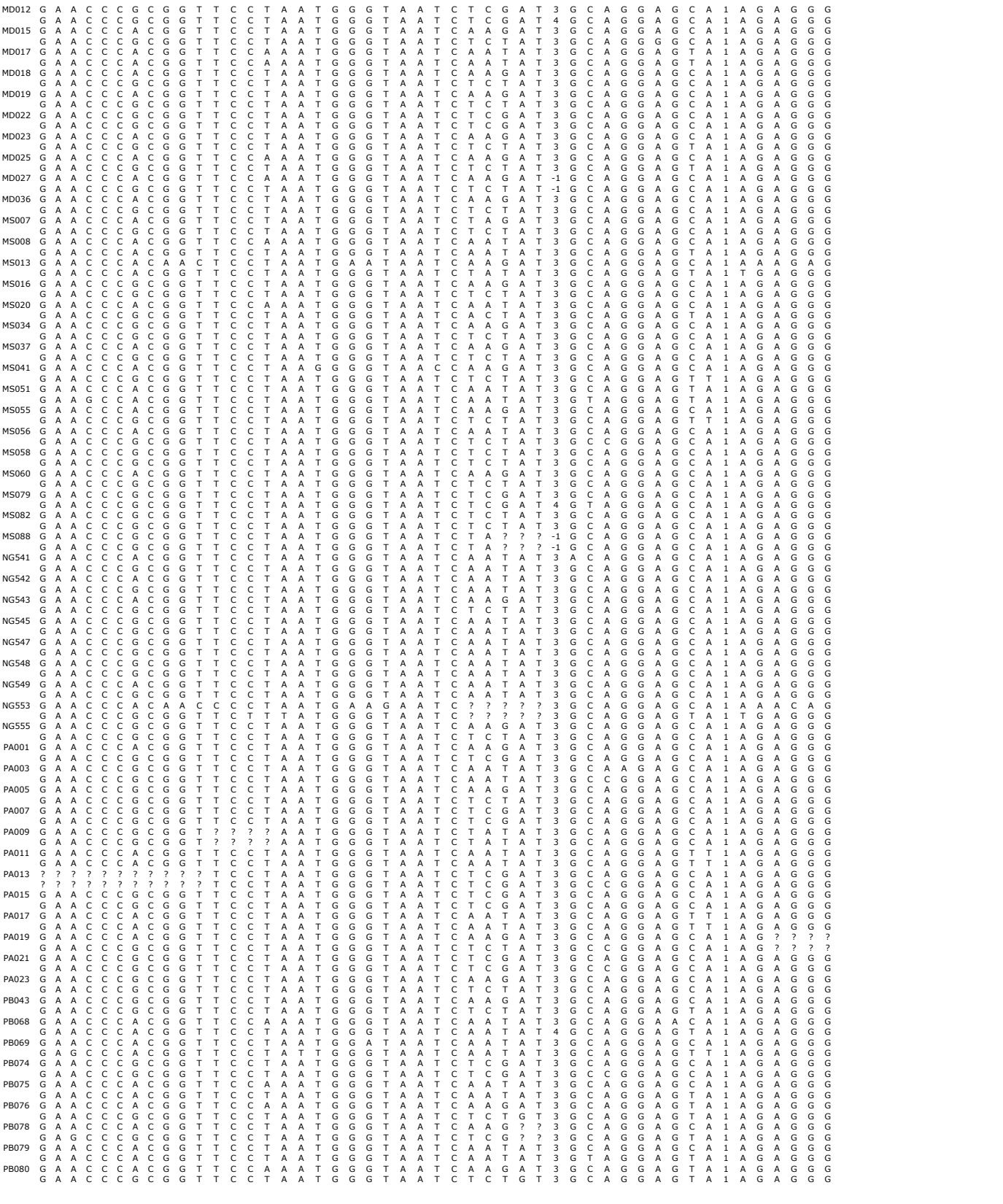
APPENDIX 3
NAT POLYPHRED OUTPUT ("UN-PHASED" NAT DATA)

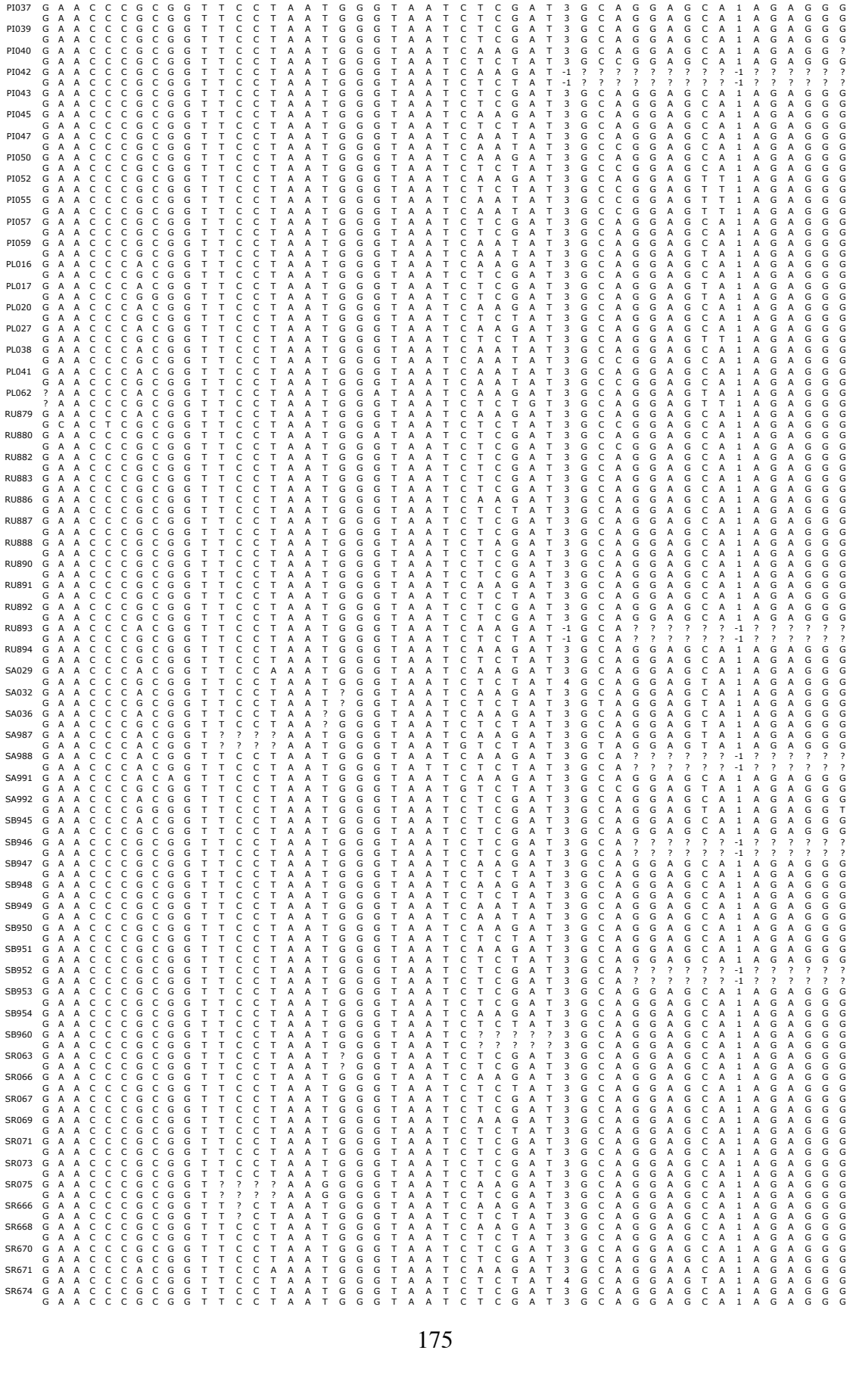
NAT1

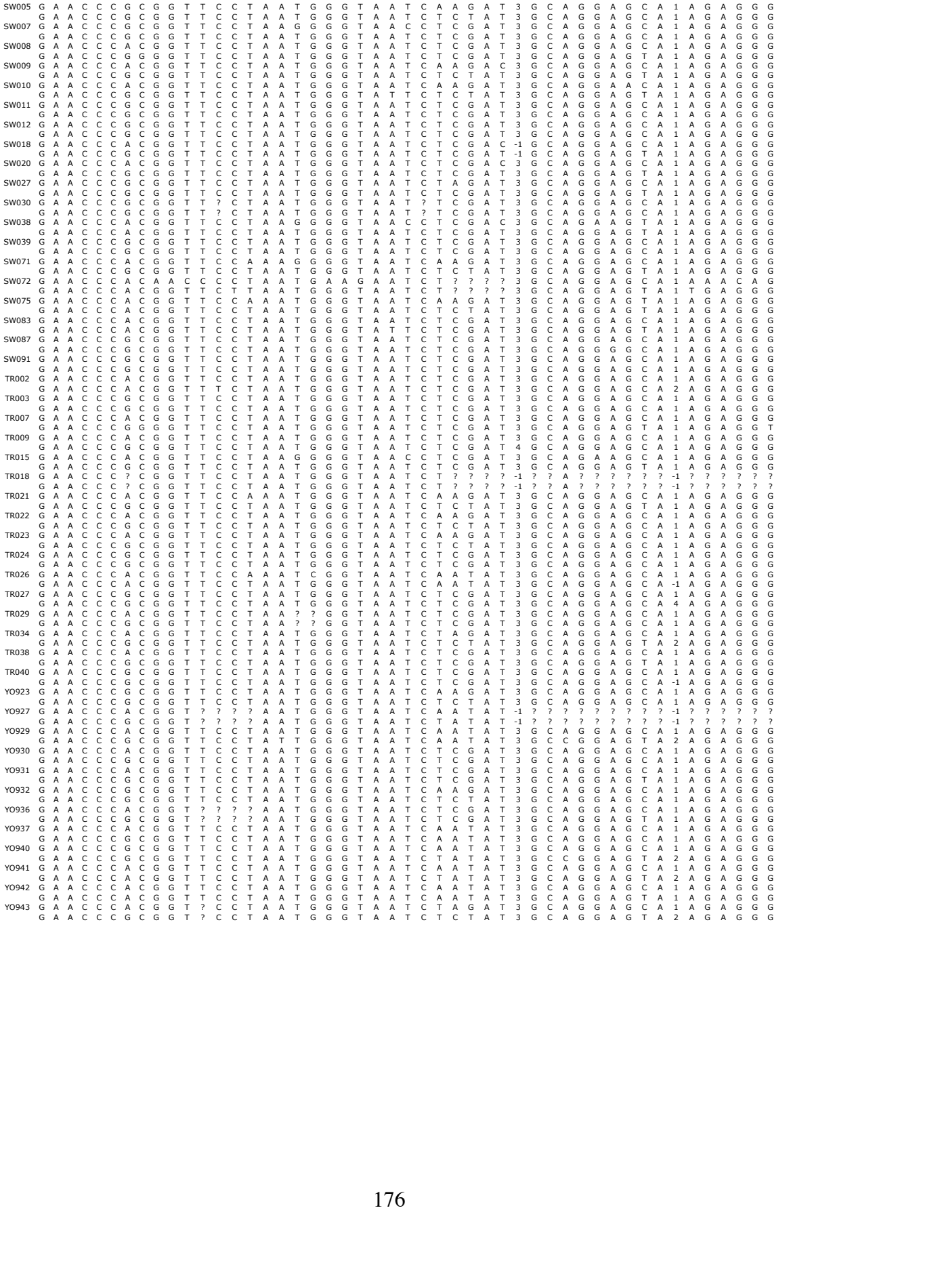




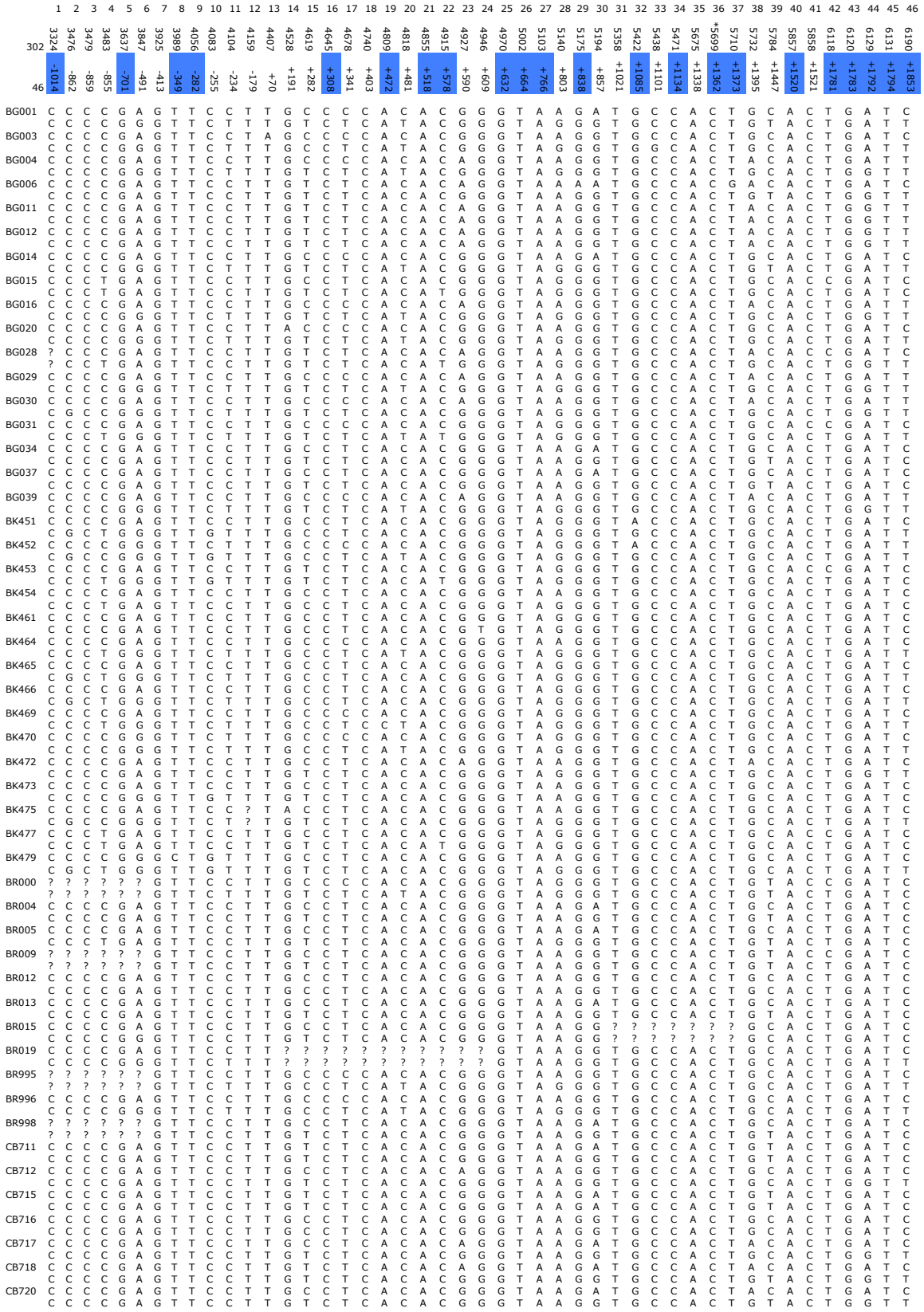


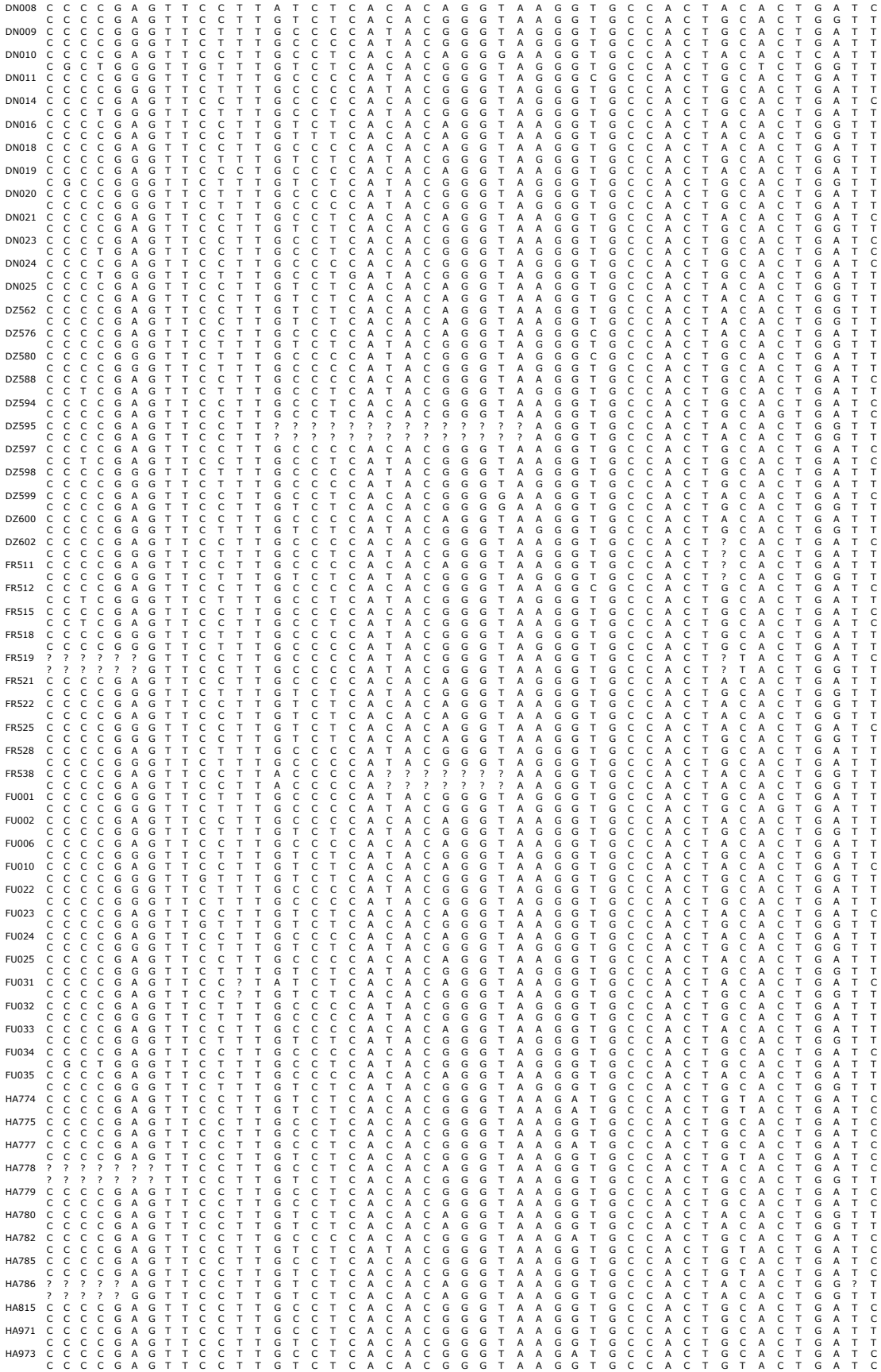


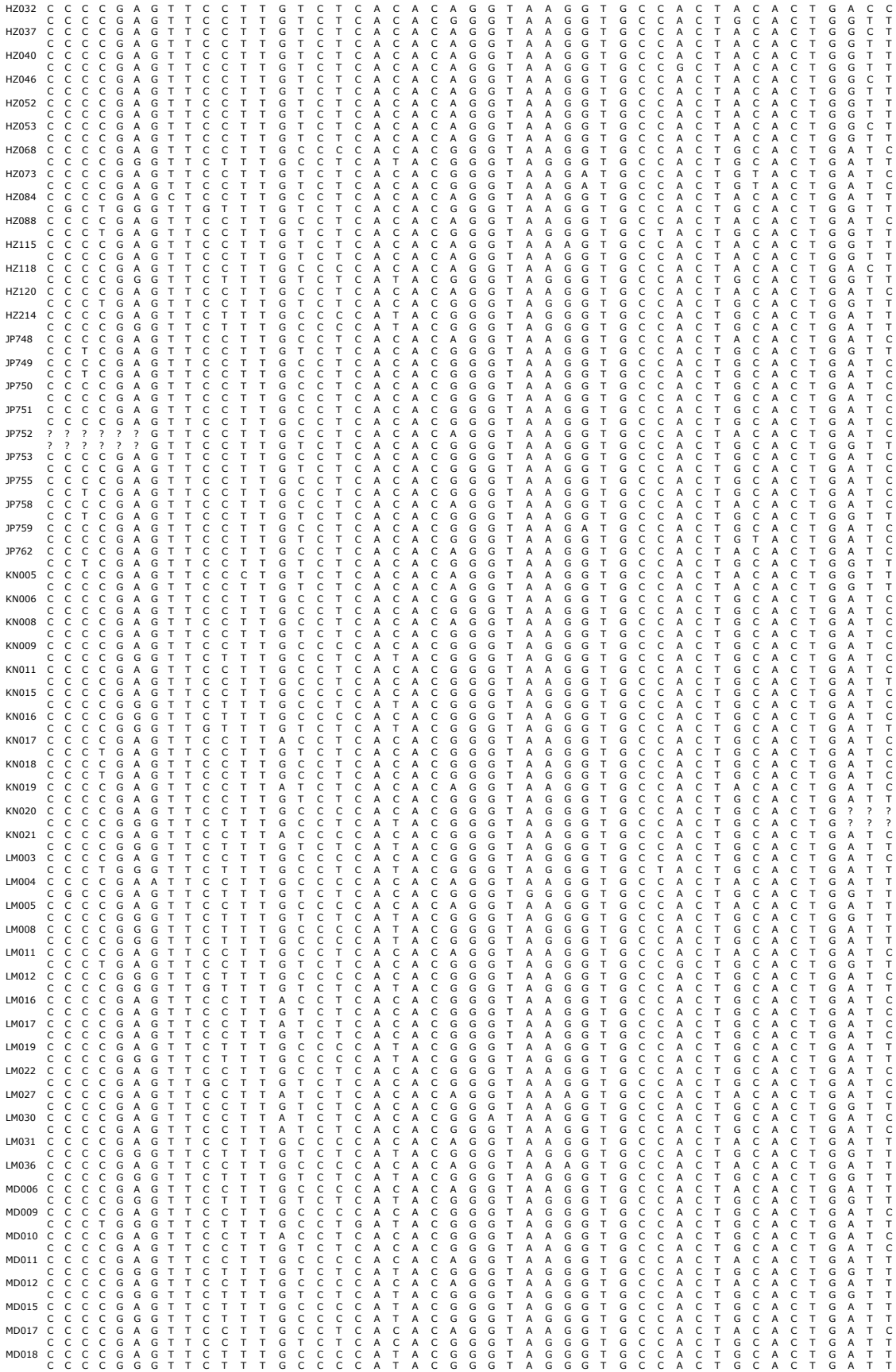


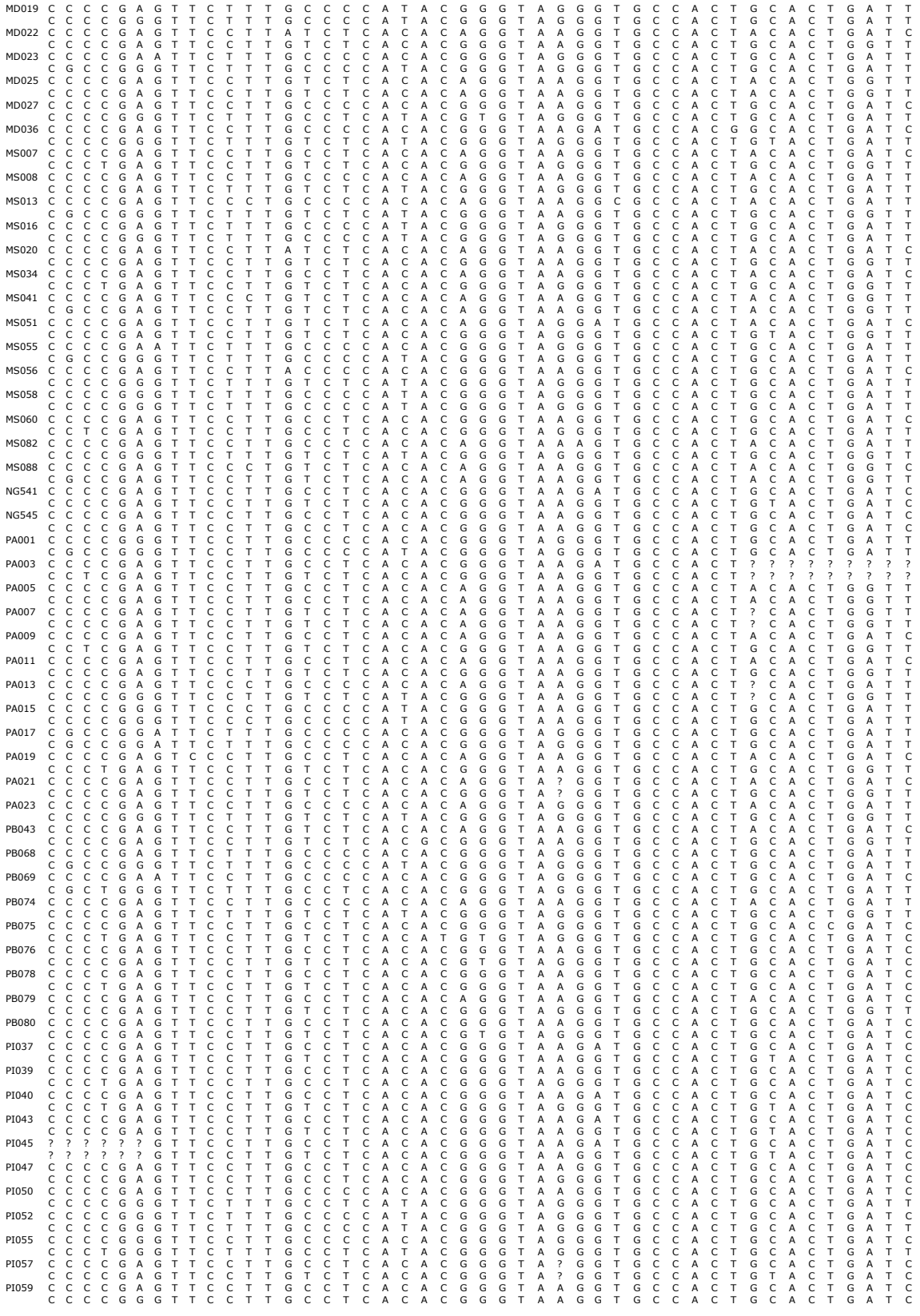


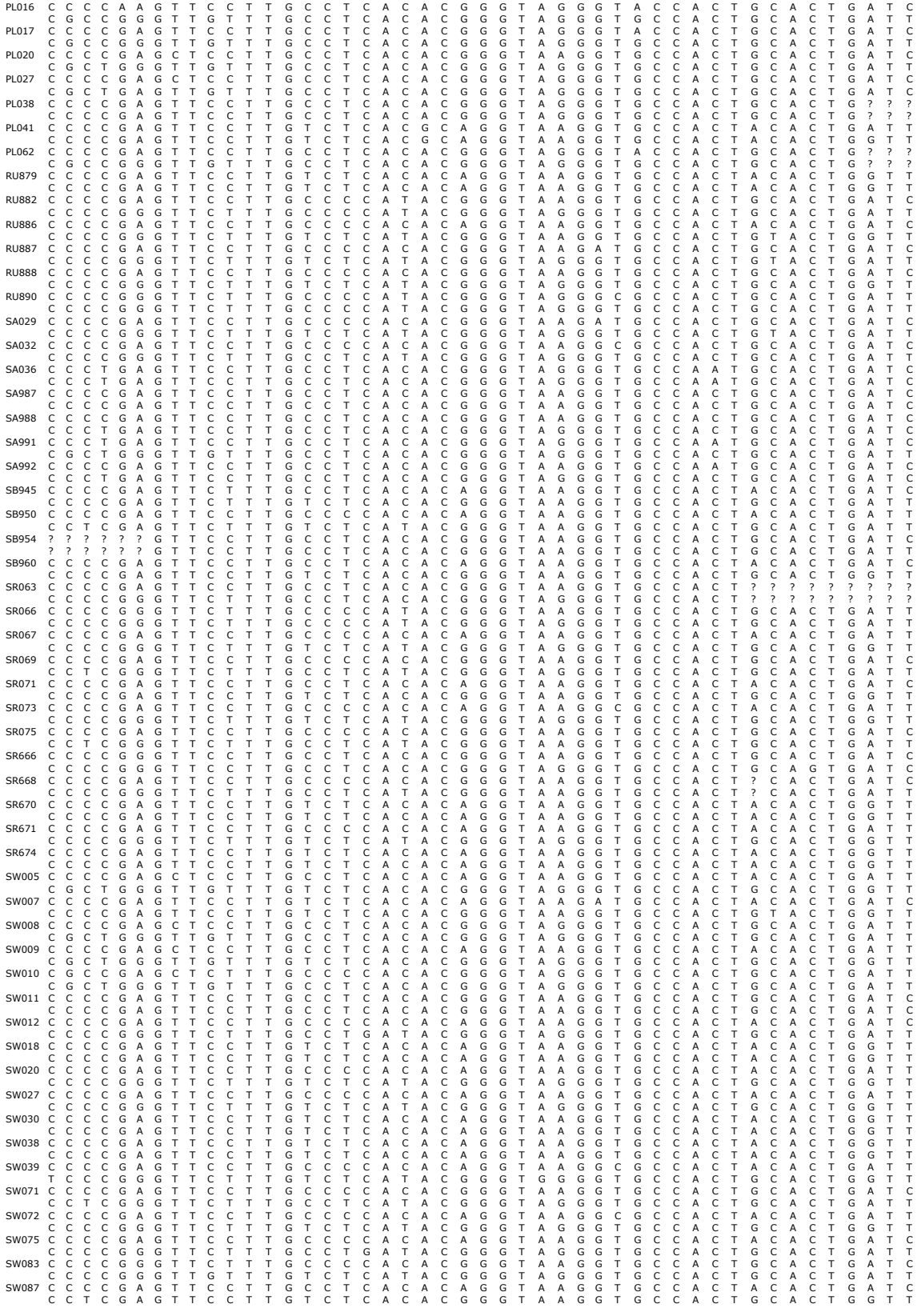
NAT2

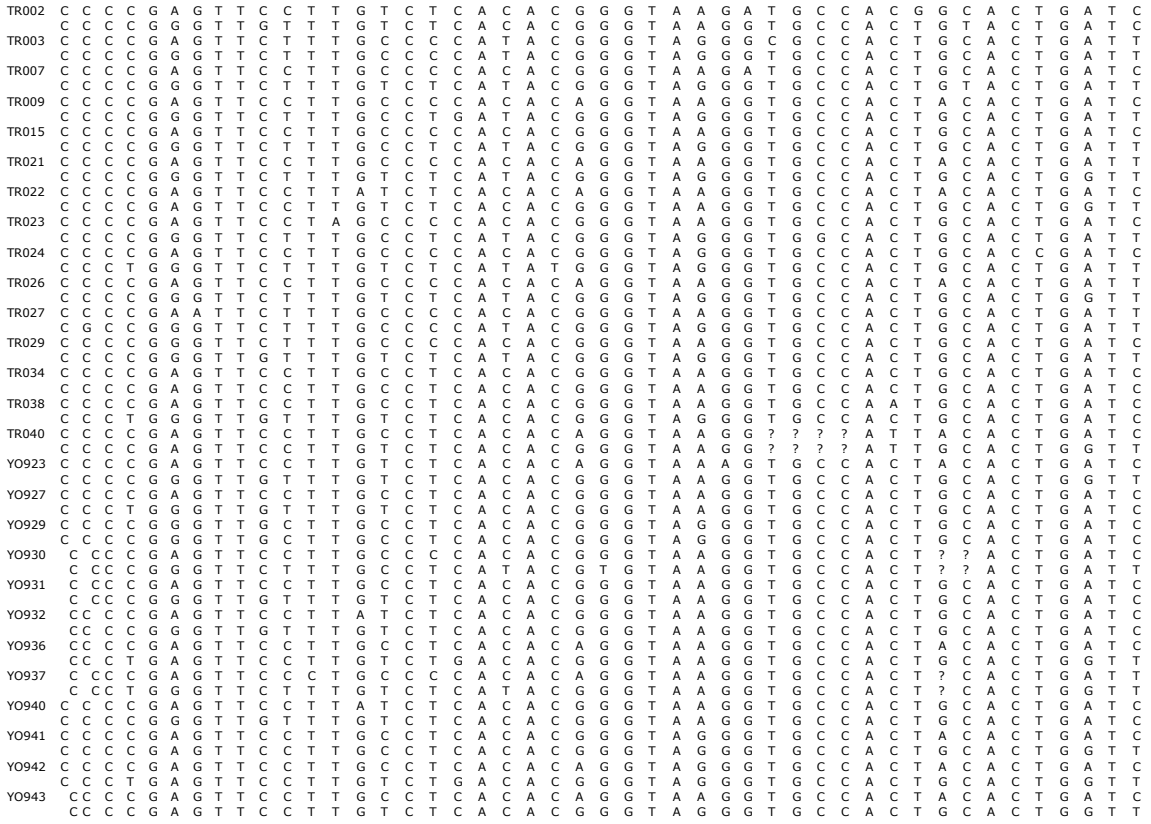




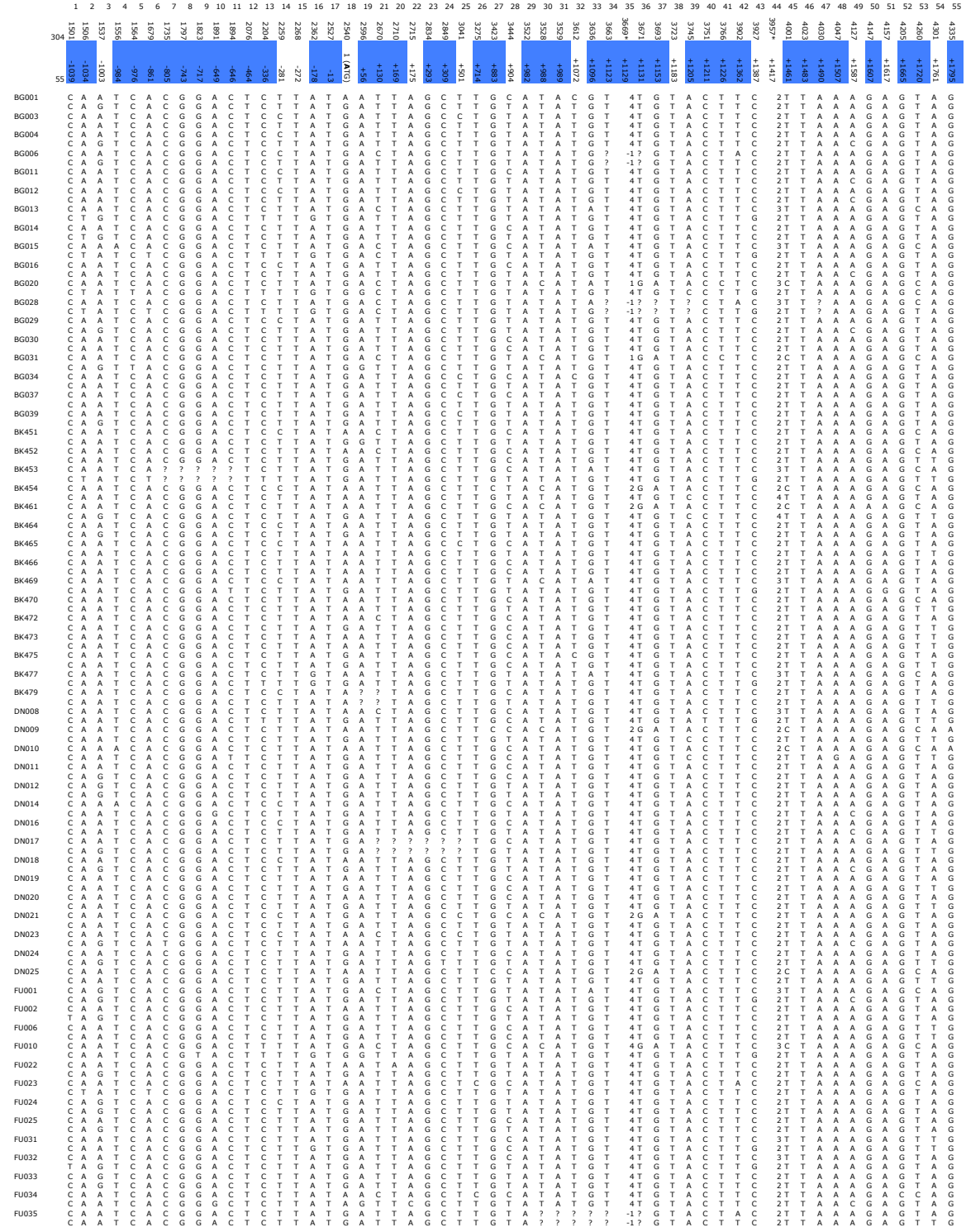

HZ032
HZ037
HZ040
HZ046
HZ052
HZ053
HZ068
HZ073
HZ084
HZ088
HZ115
HZ118
HZ120
HZ214
JP748
JP749
JP750
JP751
JP752
JP753
JP755
JP758
JP759
JP762
KN005
KN006
KN008
KN009
KN011
KN015
KN016
KN017
KN018
KN019
KN020
KN021
LM003
LM004
LM005
LM008
LM011
LM012
LM016
LM017
LM019
LM022
LM027
LM030
LM031
LM036
MD006
MD009
MD010
MD011
MD012
MD015
MD017
MD018

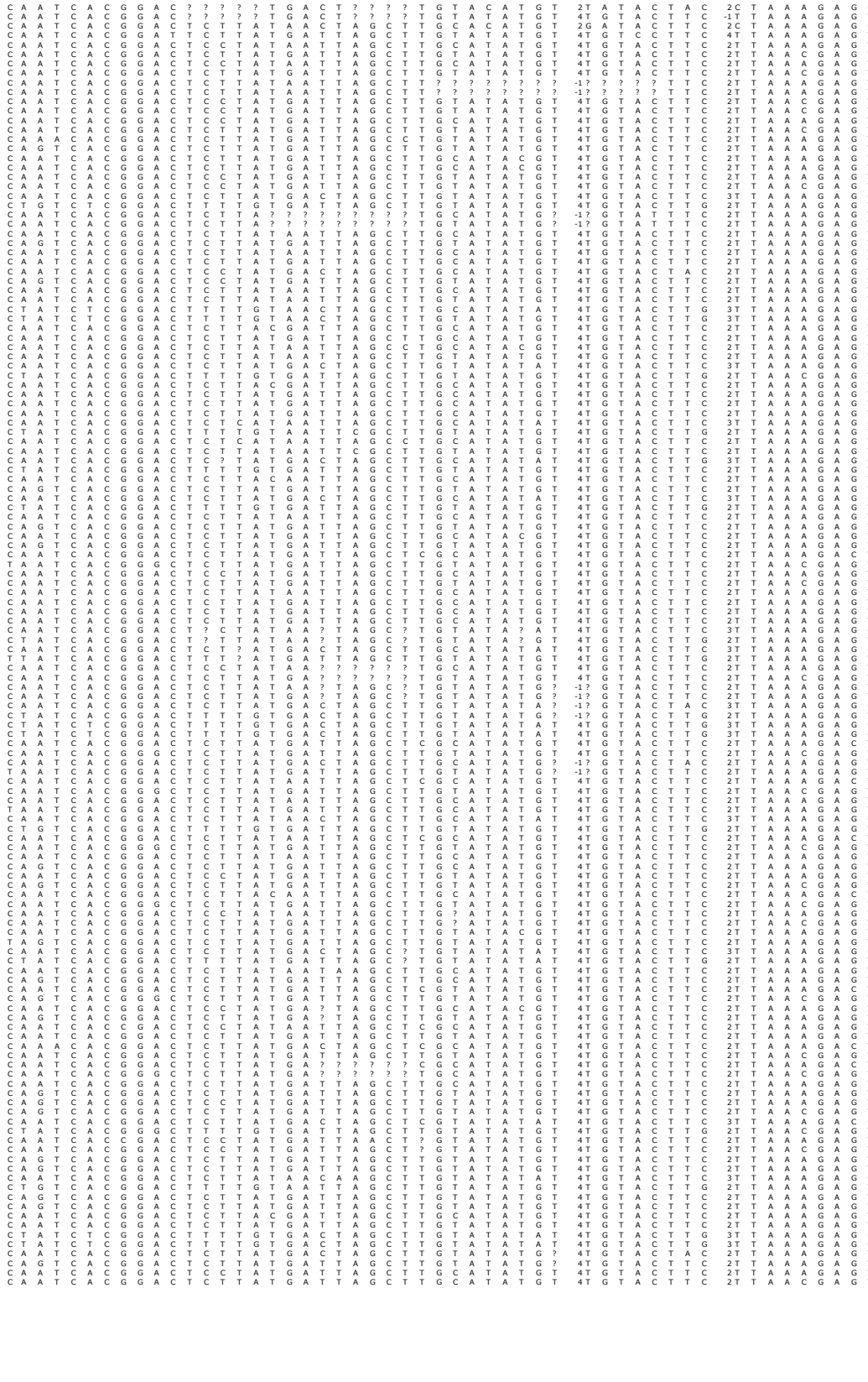

MD019
MD022
MD023
MD025
MD027
MD036
MS007
MS008
MS013
MS016
MS020
MS034
MS041
MS051
MS055
MS056
MS058
MS060
MS082
MS088
NG541
NG545
PA001
PA003
PA005
PA007
PA009
PA011
PA013
PA015
PA017
PA019
PA021
PA023
PB043
PB068
PB069
PB074
PB075
PB076
PB078
PB079
PB080
PI037
PI039
PI040
PI043
PI045
PI047
PI050
PI052
PI055
PI057
PI059

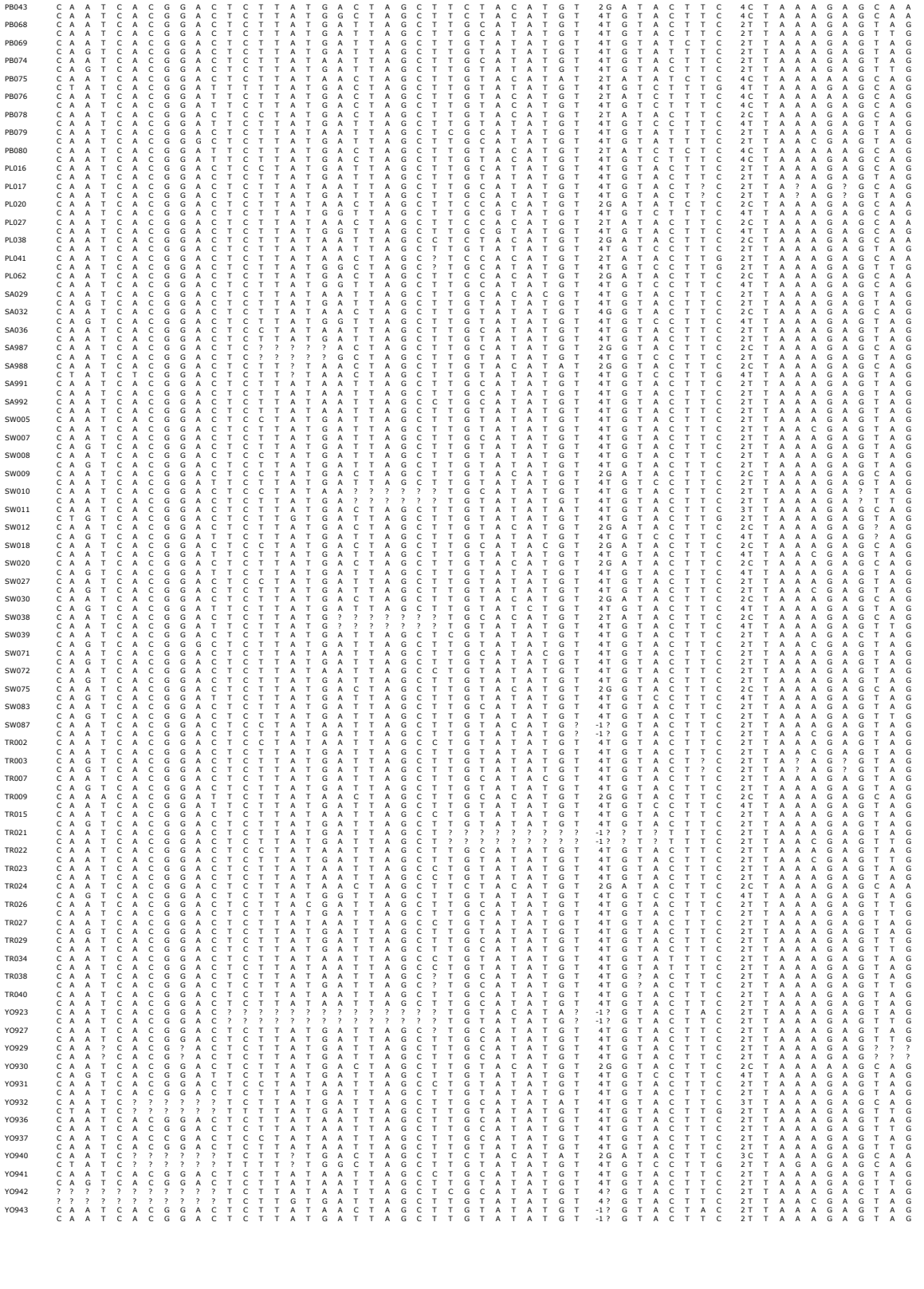


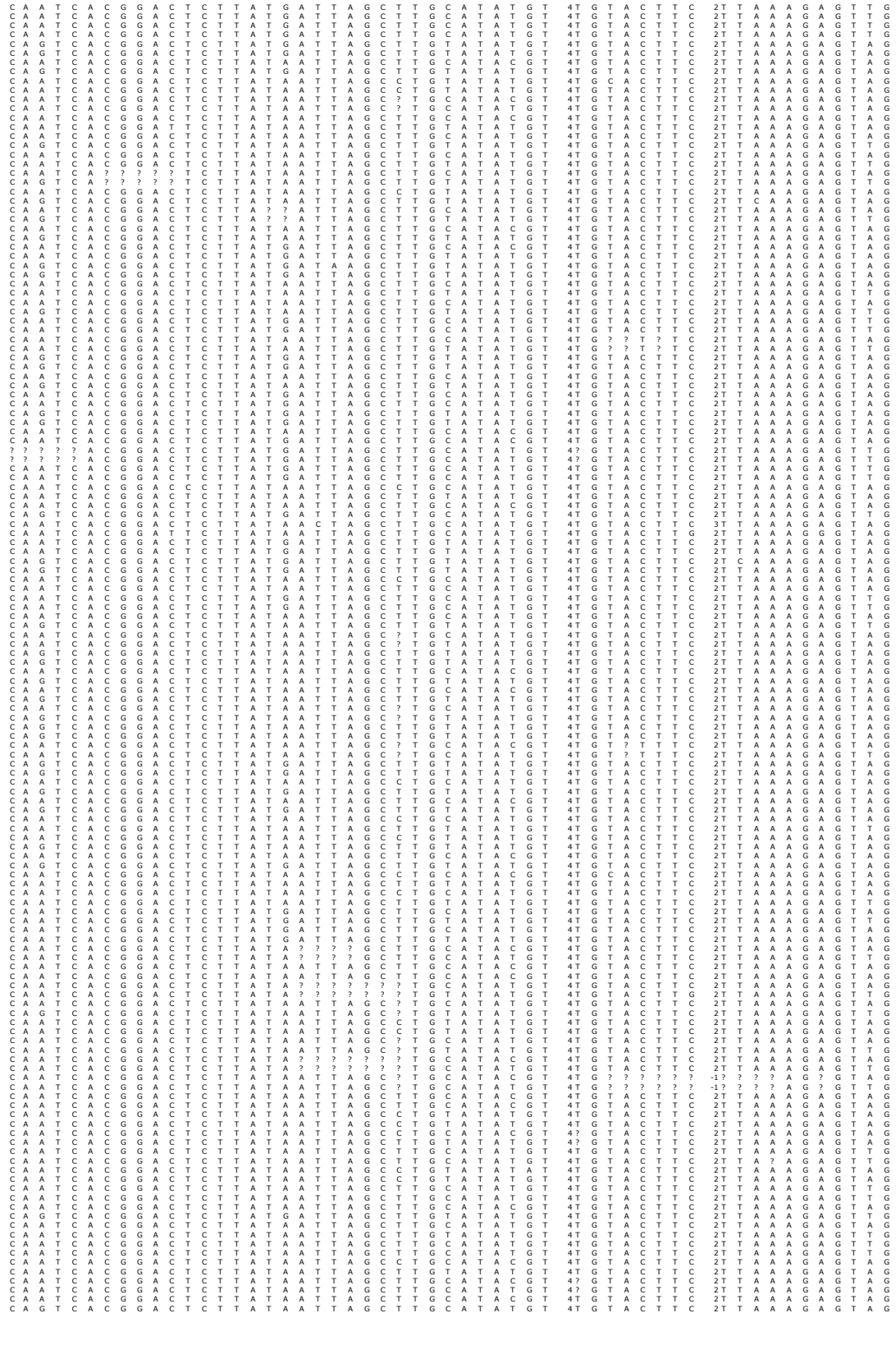

TR002 C C C C G A G T T C C T T G T C T C A A C A C G G G T A A A G A T G C C C A C G G C A C T G A T C
C C C C G G G T T T G C T T T T G G T C C T C A A A C C G G G G T T A A A G G G G T T C C C C A A C T T G G C T A C T T G A T T
TR003 C C C C G G A G T T T C C T T T T G G T C C T C A A A C C G G G G T T A A A G G G G A T C T T G G C C C C A A C T T G G C T A C T T G A T T
C C C C G G A G G T T T C C T T T T G G T C C C C A A A C C G G G G T T A A G A G G G A T T G G C C C C A A C T T G G C A A C T T G A T T
TR007 C C C C G G A G G T T T C C C T T T T G G T C C T C A A A C C G G G G T T A A A G G G G A T T G G C C C C A A C T T G G C T A C T T G A T T
C C C C G G A G G T T T C C T T T T G G T C C T C A A A C C G G G G T T A A G A G G G G T T G G C C C C A A C T T G G C T A C T T G A T T
TR009 C C C C G G A G G T T T C C C T T T T G G T C C T C A A A C C G A G G G G T T A A A G G G G G T T G G C C C C A A C T T A G C T A C T T G A T T
C C C C G G A G G T T T C C T T T T G G T C C T C G A A T A A C C G G G G G T T A A G A G G G G T T G G C C C C A A C T T G G C A A C T T G A T T
TR015 C C C C G G A G G T T T C C T T T T G G C C C C T C C A A C T A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G A T C
C C C C G G A G G T T T C C T T T T G G C C C C T C C A A T A A C C G G G G G T T A A G A G G G G T T G G C C C C A A C T T G G C A A C T T G A T T
TR021 C C C C G G A G G T T T C C T T T T G G C C C C T C C A A C T A A C C A G G G G T T A A G A G G G G T T G G C C C C A A C T T A G C A A C T T G A T T T
C C C C G G A G G T T T C C T T T T G G T C C T C C A A C T A A C C G G G G G T T A A G A G G G G T T G G C C C C A A C T T G G C A A C T T G G T T T
TR022 C C C C G G A G G T T T C C C T T T T G A T C C T C A A C T A A C C A G G G G T T A A A A G G G G T T G G C C C C A A C T T A G C A A C T T G A T T C
C C C C G G A G G T T T C C C T T T T G G T C C T C A A C C A A C C G G G G G T T A A A A G G G G T T G G C C C C A A C T T G G C A A C T T G G T T C
TR023 C C C C G G A G G T T T C C C T T A G G G C C T C T C A A T A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G A T T C
C C C C G G G G G T T T C C C T T T T G G C C C T C C A A T A A C C G G G G G T T A A G G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T
TR024 C C C C G G A G G T T T C C C T T T T G G T C C T C A A T A A C C G G G G G T T A A G G G G G G T T G G C C C C A A C T T G G C A A C T T G C T T T C
C C C T G G A G G T T T C C C T T T T G G T C C T C A A C T A A C C T G G G G G T T A A G G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T
TR026 C C C C C G G A G G T T T C C C T T T T G G T C C T C A A C T A A C C A G G G G T T A A G A G G G G T T G G C C C C A A C T T A G C A A C T T G G A T T T
C C C C C G G G G G T T T C C C T T T T G G T C C T C A A T A A C C C G G G G G T T A A G G G G G G T T G G C C C C A A C T T G G C A A C T T G G G T T T
TR027 C C C C C G G A A T T T C C T T T T G G T C C C C A A C T A A C C G G G G G T T A A G A G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T
C G C C C G G G G G T T T C C T T T T G G C C C C C A A T A A C C G G G G G T T A A G G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T
TR029 C C C C C G G G G T T T C C T T T T G G T C C C C A A C T A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G A T T C
C C C C C G G A G G T T T G C T T T T G G T C C T C A A C T A A C C G G G G G T T A A G A G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
TR034 C C C C C G A A G T T T C C C T T T T G G C C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
C C C C C G A A G T T T C C C T T T T G G C C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
TR038 C C C C C G A A G T T T C C C T T T T G G C C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
C C C C T G A G G T T T G C T T T T G G T C C T C A A C A A C C G G G G G T T A A G A G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
TR040 C C C C C G A A G T T T C C C T T T T G G T C C T C A A C A A C C A G G G G T T A A A G G G G G T T ? ? ? ? ? ? ? ? A A C T T A G C A A C T T G G A T T T C
C C C C C G A A G T T T C C C T T T T G G T C C T C A A C A A C C G G G G G T T A A A A G G G G T T ? ? ? ? ? ? ? ? A A C T T G G C A A C T T G G A G T T C
YO923 C C C C C G A A G T T T C C C T T T T G G T C C T C A A C A A C C A G G G G T T A A A G G G G G T T G G C C C C A A C T T A G C A A C T T G G A T T T C
C C C C C G A G G T T T G C T T T T G G T C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A G T T C
YO927 C C C C C G A A G T T T C C C T T T T G G T C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
C C C C T G G G G T T T G G C T T T T G G T C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
YO929 C C C C C G G G G T T T G G C C T T T T G G C C C T C A A C A A C C G G G G G T T A A G G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
C C C C C G G G G T T T G G C C T T T T G G T C C C C T C A A C A A C C G G G G G T T A A G G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
YO930 C C C C C G A G G T T T C C C T T T T G G C C C T C A A C T A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T ? ? ? A A C T T G G A T T T C
C C C C C G G G G T T T C C C T T T T G G T C C T C A A C T A A C C G G T G G T T A A A G G G G G T T G G C C C C A A C T T ? ? ? A A C T T G G A T T T C
YO931 C C C C C G A G G T T T C C C T T T T G G C C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
C C C C C G G G G T T T G C T T T T G G T C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
YO932 C C C C C G A G G T T T C C T T T T G A T C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
C C C C C G G G G T T T G C T T T T G G T C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
YO936 C C C C C G A G G T T T C C C T T T T G G T C C T C T T C A A C A A C C A G G G G T T A A A G G G G G T T G G C C C C A A C T T A G C A A C T T G G A T T T C
C C C C T G A G G T T T C C C T T T T G G T C C T C G A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A G T T T C
YO937 C C C C C G A G G T T T C C C T T C T T G G T C C T C A A C A A C C A G G G G T T A A A G G G G G T T G G C C C C A A C T T ? ? C A A C T T G G A T T T C
C C C C T G G G G T T T C C C T T T T G G T C C T C A T C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T ? ? C A A C T T G G A T T T C
YO940 C C C C C G A G T T T C C C T T T T G A T C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
C C C C C G G G T T T G C T T T T G G T C C T C A A C A A C C G G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
YO941 C C C C C G A G T T T C C C T T T T G G C C C T C A A C A A C C G G G G G T T A A G G G G G G T T G G C C C C A A C T T A G C A A C T T G G A T T T C
C C C C C G A A G T T T C C C T T T T G G C C C T C A A C A A C C G G G G G T T A A G G G G G G T T G G C C C C A A C T T G G C A A C T T G G A G T T T C
YO942 C C C C C G A A G T T T C C C T T T T G G T C C T C A A C A A C C A G G G G T T A A G G G G G G T T G G C C C C A A C T T A G C A A C T T G G A T T T C
C C C C T G A A G T T T C C C T T T T G G T C C T C G A A C A A C C G G G G G T T A A G A G G G G T T G G C C C C A A C T T G G C A A C T T G G A G T T T C
YO943 C C C C C G A A G T T T C C C T T T T G G T C C T C A A C A A C C A G G G G T T A A A G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C
C C C C G A A G T T T C C C T T T T G G T C C T C A A C A A C C G G G G G T T A A G G G G G G T T G G C C C C A A C T T G G C A A C T T G G A T T T C

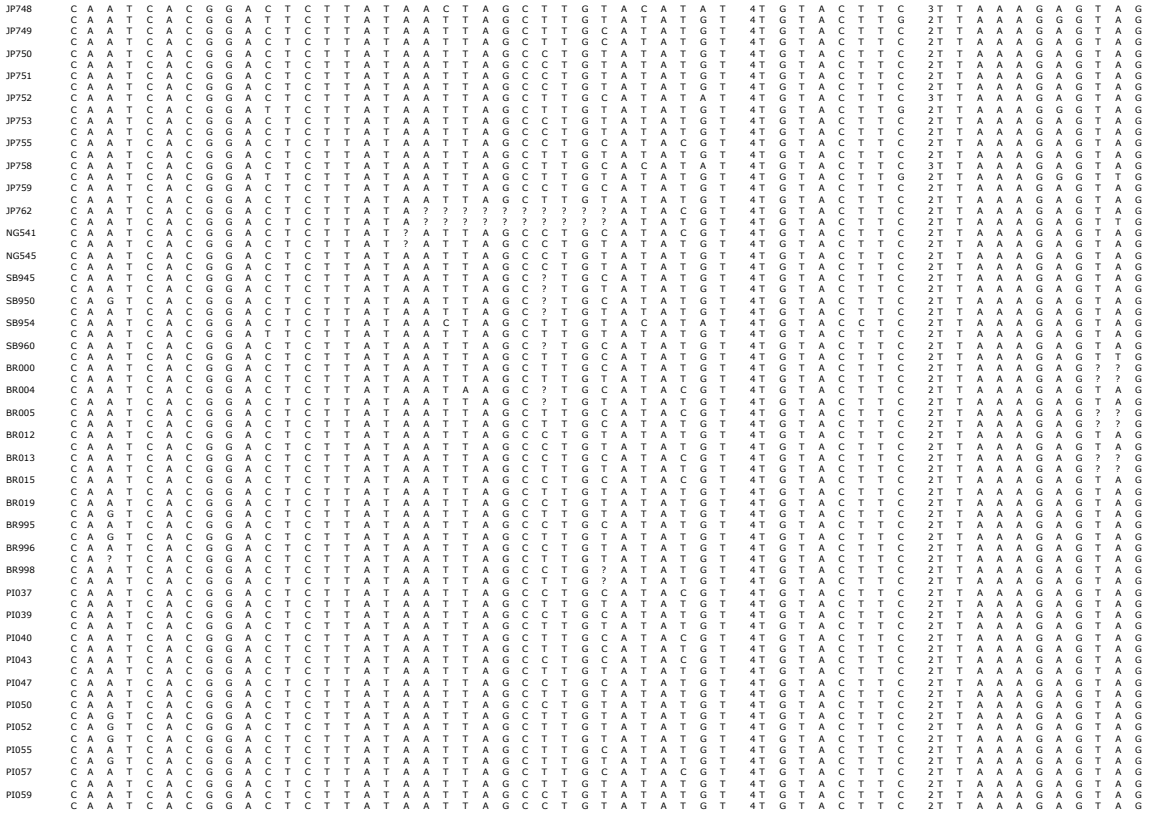
NATP1











Appendix 3: *NAT* polyphred SNP output prior to "phasing". SNP numbering is given first for polyphred, and followed in the subsequent row with corresponding position according to accepted *NAT* nomenclature, with the exception of *NATP1* that is numbered according to the consensus +1 ATG (confirmed by alignment with both *NAT1* and *NAT2*). Ambiguous calls were rechecked by eye, and all ambiguities are indicated with '?'. Tri-allelic sites in the data are coded numerically, where nucleotides A=1, G=2, C=3, T=4, and -1 indicates missing data. Individuals with ambiguous base calls spanning greater than 200 bp in length were removed prior to phase inference.

Agundez, J. A., Menaya, J. G., Tejeda, R., Lago, F., Chavez, M., and Benitez, J., 1996, Genetic analysis of the NAT2 and CYP2D6 polymorphisms in white patients with non-insulin-dependent diabetes mellitus, *Pharmacogenetics* **6**(5)**:**465-72.

Ambrose, S. H., 1982, Archaeology and linguistic reconstructions of history in East Africa., in: *The Archaeology and Linguistic Reconstructions in African History* (C. P. Ehret, M., ed.), U.C. Berkeley Press, Berkeley, pp. 104-157.

Ambrose, S. H., 1984, The introduction of pastoral adaptations to the highlands of East Africa, in: *From Hunters to Farmers: The Causes and Consequences of food Production in Africa* (J. D. B. Clark, S.A., ed.), Univ. of Cal. Press, Berkeley, Los Angeles, London, pp. 217-239.

Ambrose, S. H., 1986, Hunter-Gatherer Adaptations to Non-Marginal Environments: An Ecological and Archaeological Assessment of the Dorobo Model, *Sprache und Geschichte in Afrika* **7.2:**11-42.

Balakirev, E. S., and Ayala, F. J., 2003, Pseudogenes: are they "junk" or functional DNA?, *Annu Rev Genet* **37:**123-51.

Bandelt, H.-J. F., P; Rohl, A., 1999, Median-Joining Networks for Inferring Intraspecific Phylogenies, *Mol. Biol. Evol.* **16**(1)**:**37-48.

Bandelt, H.-J. F., P;Sykes, BC; Richards, MB., 1995, Mitochondrial portraits of human populations using median networks, *Genetics* **141:**743-753.

Barker, D. F., pers. comm., Email Correspondance (H. M. Mortensen, ed.).

Barker, D. F., Husain, A., Neale, J. R., Martini, B. D., Zhang, X., Doll, M. A., States, J. C., and Hein, D. W., 2006, Functional properties of an alternative, tissue-specific promoter for human arylamine N-acetyltransferase 1, *Pharmacogenet Genomics* **16**(7)**:**515-25.

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J., 2005, Haploview: analysis and visualization of LD and haplotype maps, *Bioinformatics* **21**(2)**:**263-5.

Bayoumi, R. A., Qureshi, M. M., al-Ameri, M. M., and Woolhouse, N. M., 1997, The N-acetyltransferase G191 A mutation among Sudanese and Somalis, *Pharmacogenetics* **7**(5)**:**397-9.

Bell, D. A., Badawi, A. F., Lang, N. P., Ilett, K. F., Kadlubar, F. F., and Hirvonen, A., 1995, Polymorphism in the N-Acetyltransferase-1 (Nat1) Polyadenylation Signal - Association of Nat1-Asterisk-10 Allele with Higher N-Acetylation Activity in Bladder and Colon Tissue, *Cancer Research* **55**(22)**:**5226-5229.

Bell, D. A., Taylor, J. A., Butler, M. A., Stephens, E. A., Wiest, J., Brubaker, L. H., Kadlubar, F. F., and Lucier, G. W., 1993, Genotype/phenotype discordance for human arylamine N-acetyltransferase (NAT2) reveals a new slow-acetylator allele common in African-Americans, *Carcinogenesis* **14**(8)**:**1689-92.

Blum, M., Grant, D. M., Demierre, A., and Meyer, U. A., 1989, Nucleotide sequence of a full-length cDNA for arylamine N-acetyltransferase from rabbit liver, *Nucleic Acids Res* **17**(9)**:**3589.

Blum, M., Grant, D. M., McBride, W., Heim, M., and Meyer, U. A., 1990, Human arylamine N-acetyltransferase genes: isolation, chromosomal localization, and functional expression, *DNA Cell Biol* **9**(3)**:**193-203.

Blurton Jones, N. G., Smith, L. C., O'Connell, J. F., Hawkes, K., and Kamuzora, C. L., 1992, Demography of the Hadza, an increasing and high density population of Savanna foragers, *Am J Phys Anthropol* **89**(2)**:**159-81.

Boukoulovala, S., pers. comm. , Email Correspondance (H. M. Mortensen, ed.).

Boukouvala, S., and Fakis, G., 2005, Arylamine N-acetyltransferases: what we learn from genes and genomes, *Drug Metab Rev* **37**(3)**:**511-64.

Boukouvala, S., Price, N., Plant, K. E., and Sim, E., 2003, Structure and transcriptional regulation of the Nat2 gene encoding for the drug-metabolizing enzyme arylamine N-acetyltransferase type 2 in mice, *Biochem J* **375**(Pt 3)**:**593-602.

Boukouvala, S., and Sim, E., 2005, Structural analysis of the genes for human arylamine N-acetyltransferases and characterisation of alternative transcripts, *Basic Clin Pharmacol Toxicol* **96**(5)**:**343-51.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., and Cavalli-Sforza, L. L., 1994, High resolution of human evolutionary trees with polymorphic microsatellites, *Nature* **368**(6470)**:**455-7.

Brockton, N., Little, J., Sharp, L., and Cotton, S. C., 2000, N-acetyltransferase polymorphisms and colorectal cancer: a HuGE review, *Am J Epidemiol* **151**(9)**:**846-61.

Bruhn, C., Brockmoller, J., Cascorbi, I., Roots, I., and Borchert, H. H., 1999, Correlation between genotype and phenotype of the human arylamine N-acetyltransferase type 1 (NAT1), *Biochem Pharmacol* **58**(11)**:**1759-64.

Butcher, N. J., Arulpragasam, A., Goh, H. L., Davey, T., and Minchin, R. F., 2005, Genomic organization of human arylamine N-acetyltransferase Type I reveals alternative promoters that generate different 5'-UTR splice variants with altered translational activities, *Biochem J* **387**(Pt 1)**:**119-27.

Butcher, N. J., Arulpragasam, A., Pope, C., and Minchin, R. F., 2003, Identification of a minimal promoter sequence for the human N-acetyltransferase Type I gene that binds AP-1 (activator protein 1) and YY-1 (Yin and Yang 1), *Biochem J* **376**(Pt 2)**:**441-8.

Campbell, M. a. T., SA, 2008, African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping, *Annual Review of Genomics and Human Genetics* **9:**in press.

Cann, H. M., de Toma, C., Cazes, L., Legrand, M. F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlaender, J. S., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R. J., Huang, X., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Mehdi, S. Q., Parham, P., Piazza, A., Pistillo, M. P., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J., and Cavalli-Sforza, L. L., 2002, A human genome diversity cell line panel, *Science* **296**(5566)**:**261-2.

Cavaco, I., Reis, R., Gil, J. P., and Ribeiro, V., 2003, CYP3A4*1B and NAT2*14 alleles in a native African population, *Clin Chem Lab Med* **41**(4)**:**606-9.

Cavalli-Sforza, L. L., 2005, The Human Genome Diversity Project: past, present and future, *Nat Rev Genet* **6**(4)**:**333-40.

Charlesworth, B., 1998, Measures of divergence between populations and the effect of forces that reduce variability, *Mol Biol Evol* **15**(5)**:**538-43.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D., 2003, Multiple sequence alignment with the Clustal series of programs, *Nucleic Acids Res* **31**(13)**:**3497-500.

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., and Pritchard, J. K., 2006, A worldwide survey of haplotype variation and linkage disequilibrium in the human genome, *Nat Genet* **38**(11)**:**1251-60.

Consortium, I. H., 2005, A haplotype map of the human genome, *Nature* **437:**1299-1320.

Consortium, T. I. H., 2003, The International HapMap Project, *Nature* **426:**789-796.

Davuluri, R. V., Suzuki, Y., Sugano, S., Plass, C., and Huang, T. H., 2008, The functional consequences of alternative promoter use in mammalian genomes, *Trends Genet* **24**(4)**:**167-177.

Demuth, J. P., De Bie, T., Stajich, J. E., Cristianini, N., and Hahn, M. W., 2006, The evolution of mammalian gene families, *PLoS ONE* **1:**e85.

Ebisawa, T., Sasaki, Y., and Deguchi, T., 1995, Complementary DNAs for two arylamine N-acetyltransferases with identical 5' non-coding regions from rat pineal gland, *Eur J Biochem* **228**(1)**:**129-37.

Edwalds-Gilbert, G., Veraldi, K. L., and Milcarek, C., 1997, Alternative poly(A) site selection in complex transcription units: means to an end?, *Nucleic Acids Res* **25**(13)**:**2547-61.

Evans, W. E., and Relling, M. V., 1999, Pharmacogenomics: translating functional genomics into rational therapeutics, *Science* **286**(5439)**:**487-91.

Ewing, B., and Green, P., 1998a, Base-calling of automated sequencer traces using phred. II. Error probabilities, *Genome Res* **8**(3)**:**186-94.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P., 1998b, Base-calling of automated sequencer traces using phred. I. Accuracy assessment, *Genome Res* **8**(3)**:**175-85.

Excoffier, L. G. L., and S. Schneider, 2005, Arlequin ver. 3.0: An integrated software package for population genetics data analysis., *Evolutionary Bioinformatics Online* **1:**47-50.

Fay, J. C., and Wu, C. I., 2000, Hitchhiking under positive Darwinian selection, *Genetics* **155**(3)**:**1405-13.

Felton, J. S., Malfatti, M. A., Knize, M. G., Salmon, C. P., Hopmans, E. C., and Wu, R. W., 1997, Health risks of heterocyclic amines, *Mutat Res* **376**(1-2)**:**37-41.

Fisher, R. A., 1930, The Genetical Theory of Natural Selection, Clarendon Press, Oxford.

Fretland, A. J., Leff, M. A., Doll, M. A., and Hein, D. W., 2001, Functional characterization of human N-acetyltransferase 2 (NAT2) single nucleotide polymorphisms, *Pharmacogenetics* **11**(3)**:**207-15.

Fu, Y. X., and Li, W. H., 1993, Statistical tests of neutrality of mutations, *Genetics* **133**(3)**:**693-709.

Fura, A., 2006, Role of pharmacologically active metabolites in drug discovery and development, *Drug Discov Today* **11**(3-4)**:**133-42.

Fuselli, S., Gilman, R. H., Chanock, S. J., Bonatto, S. L., De Stefano, G., Evans, C. A., Labuda, D., Luiselli, D., Salzano, F. M., Soto, G., Vallejo, G., Sajantila, A., Pettener, D., and Tarazona-Santos, E., 2007, Analysis of nucleotide diversity of NAT2 coding region reveals homogeneity across Native American populations and high intra-population diversity, *Pharmacogenomics J* **7**(2)**:**144-52.

Gordon, D., Abajian, C., and Green, P., 1998, Consed: a graphical tool for sequence finishing, *Genome Res* **8**(3)**:**195-202.
Grant, D. M., 1993, Molecular genetics of the N-acetyltransferases, *Pharmacogenetics* **3**(1)**:**45-50.
Haddrill, P. R., Thornton, K. R., Charlesworth, B., and Andolfatto, P., 2005, Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations, *Genome Res* **15**(6)**:**790-9.
Hartl, D. L. C., A.G., 1997, Principles of Population Genetics, Sinauer Associates, Inc. , Sunderland, Massachusetts.
Hein, D. W., 2002, Molecular genetics and function of NAT1 and NAT2: role in aromatic amine metabolism and carcinogenesis, *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* **506:**65-77.
Hein, D. W., Doll, M. A., Rustan, T. D., Gray, K., Feng, Y., Ferguson, R. J., and Grant, D. M., 1993, Metabolic activation and deactivation of arylamine carcinogens by recombinant human NAT1 and polymorphic NAT2 acetyltransferases, *Carcinogenesis* **14**(8)**:**1633-8.
Hein, D. W., Grant, Denis M., Sim, Edith., 2003, Arylamine N-Acetyltransferase (NAT) Nomenclature, *http://www.louisville.edu/medschool/pharmacology/NAT.html*.
Hein, D. W., McQueen, C. A., Grant, D. M., Goodfellow, G. H., Kadlubar, F. F., and Weber, W. W., 2000, Pharmacogenetics of the arylamine N-acetyltransferases: a symposium in honor of Wendell W. Weber, *Drug Metab Dispos* **28**(12)**:**1425-32.
Hickman, D., Pope, J., Patil, S. D., Fakis, G., Smelt, V., Stanley, L. A., Payton, M., Unadkat, J. D., and Sim, E., 1998, Expression of arylamine N-acetyltransferase in human intestine, *Gut* **42**(3)**:**402-9.
Hobolth, A., Christensen, O. F., Mailund, T., and Schierup, M. H., 2007, Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model, *PLoS Genet* **3**(2)**:**e7.
Hudson, R. R., 1990, Genetic Data Analysis. Methods for Discrete Population Genetic Data. Bruce S. Weir. Sinauer, Sunderland, MA, 1990. xiv, 377 pp., illus. $48; paper, $27, *Science* **250**(4980)**:**575.
Hudson, R. R., 2001, Two-locus sampling distributions and their application, *Genetics* **159**(4)**:**1805-17.
Husain, A., Barker, D. F., States, J. C., Doll, M. A., and Hein, D. W., 2004, Identification of the major promoter and non-coding exons of the human arylamine N-acetyltransferase 1 gene (NAT1), *Pharmacogenetics* **14**(7)**:**397-406.
Husain, A., Zhang, X., Doll, M. A., States, J. C., Barker, D. F., and Hein, D. W., 2007, Identification of N-acetyltransferase 2 (NAT2) transcription start sites and quantitation of NAT2-specific mRNA in human tissues, *Drug Metab Dispos* **35**(5)**:**721-7.
Iseli, C., Stevenson, B. J., de Souza, S. J., Samaia, H. B., Camargo, A. A., Buetow, K. H., Strausberg, R. L., Simpson, A. J., Bucher, P., and Jongeneel, C. V., 2002, Long-range heterogeneity at the 3' ends of human mRNAs, *Genome Res* **12**(7)**:**1068-74.
Jenne, J. W., 1965, Partial purification and properties of the isoniazid trans-acetylase in human
liver. Its relationship to the acetylation of p-aminosalicylic acid. , *J. Clin. Invest.* **44** **:**1992–2002.

Kilbane, A. J., Petroff, T., and Weber, W. W., 1991, Kinetics of acetyl CoA: arylamine N-acetyltransferase from rapid and slow acetylator human liver, *Drug Metab Dispos* **19**(2)**:**503-7.

Kimura, M., 1980, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *J Mol Evol* **16**(2)**:**111-20.

Kimura, M., 1983, The neutral theory of molecular evolution, Cambridge University Press, Cambridge [Cambridgeshire] ; New York, pp. xv, 367 p.

Klien, R., 1992, The archaeology of modern human origins, *Evolutionary Anthropology* **1:**5-14.

Kufe, D. M., Pollock, R. M., PhD,, Weichselbaum, R. M., Bast Jr, R. M., Gansler, T. M., MBA,, Holland, J. M., ScD (hc),, and Frei III, E. M., 2003, Holland-Frei Cancer Medicine 6 (H. C. B. D. I. c2000, ed.).

Kukita, Y., Miyatake, K., Stokowski, R., Hinds, D., Higasa, K., Wake, N., Hirakawa, T., Kato, H., Matsuda, T., Pant, K., Cox, D., Tahira, T., and Hayashi, K., 2005, Genome-wide definitive haplotypes determined using a collection of complete hydatidiform moles, *Genome Res* **15**(11)**:**1511-8.

Lahr, M. M., and Foley, R. A., 1998, Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution, *Am J Phys Anthropol* **Suppl 27:**137-76.

Land, S. J., Jones, R. F., and King, C. M., 1994, Biochemical and genetic analysis of two acetyltransferases from hamster tissues that can metabolize aromatic amine derivatives, *Carcinogenesis* **15**(8)**:**1585-95.

Landry, J. R., Mager, D. L., and Wilhelm, B. T., 2003, Complex controls: the role of alternative promoters in mammalian genomes, *Trends Genet* **19**(11)**:**640-8.

Lash, L. H., Hines, R. N., Gonzalez, F. J., Zacharewski, T. R., and Rothstein, M. A., 2003, Genetics and susceptibility to toxic chemicals: do you (or should you) know your genetic profile?, *J Pharmacol Exp Ther* **305**(2)**:**403-9.

Lin, H. J., Probst-Hensch, N. M., Hughes, N. C., Sakamoto, G. T., Louie, A. D., Kau, I. H., Lin, B. K., Lee, D. B., Lin, J., Frankl, H. D., Lee, E. R., Hardy, S., Grant, D. M., and Haile, R. W., 1998, Variants of N-acetyltransferase NAT1 and a case-control study of colorectal adenomas, *Pharmacogenetics* **8**(3)**:**269-81.

Liska, D. J., 1998, The detoxification enzyme systems, *Altern Med Rev* **3**(3)**:**187-98.

Magalon, H., Patin, E., Austerlitz, F., Hegay, T., Aldashev, A., Quintana-Murci, L., and Heyer, E., 2008, Population genetic diversity of the NAT2 gene supports a role of acetylation in human adaptation to farming in Central Asia, *Eur J Hum Genet* **16**(2)**:**243-51.

Martinez-Arias, R., Calafell, F., Mateu, E., Comas, D., Andres, A., and Bertranpetit, J., 2001, Sequence variability of a human pseudogene, *Genome Res* **11**(6)**:**1071-85.

McDonald, J. H., and Kreitman, M., 1991, Adaptive protein evolution at the Adh locus in Drosophila, *Nature* **351**(6328)**:**652-4.

McVean, G., Awadalla, P., and Fearnhead, P., 2002, A coalescent-based method for detecting and estimating recombination from gene sequences, *Genetics* **160**(3)**:**1231-41.

Miller, S. A., Dykes, D. D., and Polesky, H. F., 1988, A simple salting out procedure for extracting DNA from human nucleated cells, *Nucleic Acids Res* **16**(3)**:**1215.

Minchin, R. F., 1995, Acetylation of p-aminobenzoylglutamate, a folic acid catabolite, by recombinant human arylamine N-acetyltransferase and U937 cells, *Biochem J* **307 (Pt 1):**1-3.

Modiano, D., Luoni, G., Petrarca, V., Sodiomon Sirima, B., De Luca, M., Simpore, J., Coluzzi, M., Bodmer, J. G., and Modiano, G., 2001, HLA class I in three West African ethnic groups: genetic distances from sub-Saharan and Caucasoid populations, *Tissue Antigens* **57**(2)**:**128-37.

Mueller, J. C., 2004, Linkage disequilibrium for different scales and applications, *Brief Bioinform* **5**(4)**:**355-64.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H., 1992, Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. 1986, *Biotechnology* **24:**17-27.

Mummidi, S., Ahuja, S. S., McDaniel, B. L., and Ahuja, S. K., 1997, The human CC chemokine receptor 5 (CCR5) gene. Multiple transcripts with 5'-end heterogeneity, dual promoter usage, and evidence for polymorphisms within the regulatory regions and noncoding exons, *J Biol Chem* **272**(49)**:**30662-71.

Nebert, D. W., 1997, Polymorphisms in drug-metabolizing enzymes: what is their clinical relevance and why do they exist?, *Am J Hum Genet* **60**(2)**:**265-71.

Nebert, D. W., and Dieter, M. Z., 2000, The evolution of drug metabolism, *Pharmacology* **61**(3)**:**124-35.

Nei, M., and Jin, L., 1989, Variances of the average numbers of nucleotide substitutions within and between populations, *Mol Biol Evol* **6**(3)**:**290-300.

Nei, M., and Li, W. H., 1979, Mathematical model for studying genetic variation in terms of restriction endonucleases, *Proc Natl Acad Sci U S A* **76**(10)**:**5269-73.

Nelson, D. R., Zeldin, D. C., Hoffman, S. M., Maltais, L. J., Wain, H. M., and Nebert, D. W., 2004, Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants, *Pharmacogenetics* **14**(1)**:**1-18.

Nickerson, D. A., Tobe, V. O., and Taylor, S. L., 1997, PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing, *Nucleic Acids Res* **25**(14)**:**2745-51.

Nielsen, R., 2005, Molecular signatures of natural selection, *Annu Rev Genet* **39:**197-218.

O'Neil, W. M., Gilfix, B. M., DiGirolamo, A., Tsoukas, C. M., and Wainer, I. W., 1997, N-acetylation among HIV-positive patients and patients with AIDS: when is fast, fast and slow, slow?, *Clin Pharmacol Ther* **62**(3)**:**261-71.

O'Neil, W. M., MacArthur, R. D., Farrough, M. J., Doll, M. A., Fretland, A. J., Hein, D. W., Crane, L. R., and Svensson, C. K., 2002, Acetylator phenotype and genotype in HIV-infected patients with and without sulfonamide hypersensitivity, *J Clin Pharmacol* **42**(6)**:**613-9.

Ohno, S., 1970, Evolution by gene duplication, Springer-Verlag, Berlin, New York,, pp. xv, 160 p.

Pacifici, G. M., Bencini, C., and Rane, A., 1986, Acetyltransferase in humans: development and tissue distribution, *Pharmacology* **32**(5)**:**283-91.

Palmiter, R. D., Sandgren, E. P., Avarbock, M. R., Allen, D. D., and Brinster, R. L., 1991, Heterologous introns can enhance expression of transgenes in mice, *Proc Natl Acad Sci U S A* **88**(2)**:**478-82.

Patin, E., pers. comm., regarding Genbank submission of NAT2 sequences (H. M. Mortensen, ed.).
Patin, E., Barreiro, L. B., Sabeti, P. C., Austerlitz, F., Luca, F., Sajantila, A., Behar, D. M., Semino, O., Sakuntabhai, A., Guiso, N., Gicquel, B., McElreavey, K., Harding, R. M., Heyer, E., and Quintana-Murci, L., 2006a, Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes, *Am J Hum Genet* **78**(3)**:**423-36.
Patin, E., Harmant, C., Kidd, K. K., Kidd, J., Froment, A., Mehdi, S. Q., Sica, L., Heyer, E., and Quintana-Murci, L., 2006b, Sub-Saharan African coding sequence variation and haplotype diversity at the NAT2 gene, *Hum Mutat* **27**(7)**:**720.
Podlaha, O., and Zhang, J., 2004, Nonneutral evolution of the transcribed pseudogene Makorin1-p1 in mice, *Mol Biol Evol* **21**(12)**:**2202-9.
Pritchard, J. K., Stephens, M., and Donnelly, P., 2000, Inference of population structure using multilocus genotype data, *Genetics* **155**(2)**:**945-59.
Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, G., McElreavey, K., and Santachiara-Benerecetti, A. S., 1999, Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa, *Nature Genetics* **23**(4)**:**437-41.
Reed, F. A., and Tishkoff, S. A., 2006, African human diversity, origins and migrations, *Curr Opin Genet Dev* **16**(6)**:**597-605.
Reed, F. T., S.A., unpublished, Population Structure ascertained from dense global screening of STRPs (Marshfield Study).
Reif, B. R., 1953, Enzymatic inactivation of isonicotinic acid hydrazide in human and animal organism, *Arch. Exp. Pathol. Pharmak* **220**(4)**:**321-323.
Roberts-Thomson, I. C., Ryan, P., Khoo, K. K., Hart, W. J., McMichael, A. J., and Butler, R. N., 1996, Diet, acetylator phenotype, and risk of colorectal neoplasia, *Lancet* **347**(9012)**:**1372-4.
Rodrigues-Lima, F., Cooper, R. N., Goudeau, B., Atmane, N., Chamagne, A. M., Butler-Browne, G., Sim, E., Vicart, P., and Dupret, J. M., 2003, Skeletal muscles express the xenobiotic-metabolizing enzyme arylamine N-acetyltransferase, *Journal of Histochemistry & Cytochemistry* **51**(6)**:**789-796.
Rogers, J. S., and Swofford, D. L., 1998, A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide sequences, *Syst Biol* **47**(1)**:**77-89.
Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., and Feldman, M. W., 2005, Clines, clusters, and the effect of study design on the inference of human population structure, *PLoS Genet* **1**(6)**:**e70.
Rozas, J., and Rozas, R., 1995, DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data, *Comput Appl Biosci* **11**(6)**:**621-5.
Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X., and Rozas, R., 2003, DnaSP, DNA polymorphism analyses by the coalescent and other methods, *Bioinformatics* **19**(18)**:**2496-7.
Sabbagh, A., Langaney, A., Darlu, P., Gerard, N., Krishnamoorthy, R., and Poloni, E. S., 2008, Worldwide distribution of NAT2 diversity: implications for NAT2 evolutionary history, *BMC Genet* **9:**21.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., and Lander, E. S., 2006, Positive natural selection in the human lineage, *Science* **312**(5780)**:**1614-20.

Salem, R. M., Wessel, J., and Schork, N. J., 2005, A comprehensive literature review of haplotyping software and methods for use with unrelated individuals, *Hum Genomics* **2**(1)**:**39-66.

Scheet, P., and Stephens, M., 2006, A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase, *Am J Hum Genet* **78**(4)**:**629-44.

Sim, E., Payton, M., Noble, M., and Minchin, R., 2000, An update on genetic, structural and functional studies of arylamine N-acetyltransferases in eucaryotes and procaryotes, *Hum Mol Genet* **9**(16)**:**2435-41.

Sim, E., Stanley, L. A., Risch, A., and Thygesen, P., 1995, Xenogenetics in multifactorial disease susceptibility, *Trends Genet* **11**(12)**:**509-12.

Sokal, R. R., and Rohlf, F. J., 1995, Biometry : the principles and practice of statistics in biological research, W.H. Freeman, New York, pp. xix, 887 p.

Sonenberg, N., 1994, mRNA translation: influence of the 5' and 3' untranslated regions, *Curr Opin Genet Dev* **4**(2)**:**310-5.

Sosinsky, A., Glusman, G., and Lancet, D., 2000, The genomic structure of human olfactory receptor genes, *Genomics* **70**(1)**:**49-61.

Stajich, J. E., and Hahn, M. W., 2005, Disentangling the effects of demography and selection in human history, *Mol Biol Evol* **22**(1)**:**63-73.

Stenning, D., 1965, The Pastoral Fulani of Northern Nigeria, University of Minnesota, Minneapolis.

Stephens, M., and Donnelly, P., 2003, A comparison of bayesian methods for haplotype reconstruction from population genotype data, *Am J Hum Genet* **73**(5)**:**1162-9.

Stephens, M., Smith, N. J., and Donnelly, P., 2001, A new statistical method for haplotype reconstruction from population data, *Am J Hum Genet* **68**(4)**:**978-89.

Stringer, C. B., and Andrews, P., 1988, Genetic and fossil evidence for the origin of modern humans, *Science* **239**(4845)**:**1263-8.

Tajima, F., 1983, Evolutionary relationship of DNA sequences in finite populations, *Genetics* **105**(2)**:**437-60.

Tajima, F., 1989, Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism, *Genetics* **123**(3)**:**585-595.

Tarazona-Santos, E., and Tishkoff, S. A., 2005, Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus, *Genes Immun* **6**(1)**:**53-65.

Thomas, J. H., 2007, Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates, *PLoS Genet* **3**(5)**:**e67.

Timbrell, J., 2000, Principles of biochemical toxicology, Taylor & Francis, London ; New York, pp. ix, 394 p.

Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P., and Krings, M., 1996, Global patterns of linkage disequilibrium at the CD4 locus and modern human origins, *Science* **271**(5254)**:**1380-7.

Tishkoff, S. A., and Verrelli, B. C., 2003, Patterns of human genetic diversity: implications for human evolutionary history and disease, *Annu Rev Genomics Hum Genet* **4:**293-340.

Torcia, M. G., Santarlasci, V., Cosmi, L., Clemente, A., Maggi, L., Mangano, V. D., Verra, F., Bancone, G., Nebie, I., Sirima, B. S., Liotta, F., Frosali, F., Angeli, R., Severini, C., Sannella, A. R., Bonini, P., Lucibello, M., Maggi, E., Garaci, E., Coluzzi, M., Cozzolino, F., Annunziato, F., Romagnani, S., and Modiano, D., 2008, Functional deficit of T regulatory cells in Fulani, an ethnic group with low susceptibility to Plasmodium falciparum malaria, *Proc Natl Acad Sci U S A* **105**(2)**:**646-51.

Trepanier, L. A., Ray, K., Winand, N. J., Spielberg, S. P., and Cribb, A. E., 1997, Cytosolic arylamine N-acetyltransferase (NAT) deficiency in the dog and other canids due to an absence of NAT genes, *Biochem Pharmacol* **54**(1)**:**73-80.

Upton, A., Johnson, N., Sandy, J., and Sim, E., 2001, Arylamine N-acetyltransferases - of mice, men and microorganisms, *Trends Pharmacol Sci* **22**(3)**:**140-6.

Vatsis, K. P., and Weber, W. W., 1993, Structural heterogeneity of Caucasian N-acetyltransferase at the NAT1 gene locus, *Arch Biochem Biophys* **301**(1)**:**71-6.

Wakefield, L., pers. comm., Email Correspondance (H. M. Mortensen, ed.).

Wall, J. D., and Przeworski, M., 2000, When did the human population size start increasing?, *Genetics* **155**(4)**:**1865-74.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., et al., 2002, Initial sequencing and comparative analysis of the mouse genome, *Nature* **420**(6915)**:**520-62.

Watterson, G. A., 1975, On the number of segregating sites in genetical models without recombination, *Theor Popul Biol* **7**(2)**:**256-76.

Weinshilboum, R. W., Liewei 2004, PHARMACOGENOMICS:BENCH TO BEDSIDE , *NATURE REVIEWS DRUG DISCOVERY* **3**(September)**:**739-748.

Westwood, I. M., Kawamura, A., Fullam, E., Russell, A. J., Davies, S. G., and Sim, E., 2006, Structure and mechanism of arylamine N-acetyltransferases, *Curr Top Med Chem* **6**(15)**:**1641-54.

Wright, S., 1931, Evolution in Mendelian Populations, *Genetics* **16**(2)**:**97-159.
Yang, M., Katoh, T., Delongchamp, R., Ozawa, S., Kohshi, K., and Kawamoto, T., 2000, Relationship between NAT1 genotype and phenotype in a Japanese population, *Pharmacogenetics* **10**(3)**:**225-32.
Zhang, Z., and Gerstein, M., 2004, Large-scale analysis of pseudogenes in the human genome, *Curr Opin Genet Dev* **14**(4)**:**328-35.
Zhao, B., Lee, E. J., Yeoh, P. N., and Gong, N. H., 1998, Detection of mutations and polymorphism of N-acetyltransferase 1 gene in Indian, Malay and Chinese populations, *Pharmacogenetics* **8**(4)**:**299-304.