Published in final edited form as:

Pharmacogenomics. 2011 November ; 12(11): 1545–1558. doi:10.2217/pgs.11.88.

Characterization of genetic variation and natural selection at the arylamine *N*-acetyltransferase genes in global human populations

Holly M Mortensen[1,2], **Alain Froment**[3], **Godfrey Lema**[4], **Jean-Marie Bodo**[5], **Muntaser Ibrahim**[6], **Thomas B Nyambo**[4], **Sabah A Omar**[7], and **Sarah A Tishkoff**[1,8,†]

[1]Department of Biology, University of Maryland, College Park, MD, USA

[2]National Health & Environmental Effects Laboratory, Office of Research & Development, US Environmental Protection Agency, Research Triangle Park, NC, USA

[3]UMR 208, IRD-MNHN, Musée de l'Homme, 75116 Paris, France

[4]Department of Biochemistry, Muhimbili University of Health & Allied Sciences, Dar es Salaam, Tanzania

[5]Ministère de la Recherche Scientifique et de l'Innovation, BP 1457, Yaoundé, Cameroon

[6]Department of Molecular Biology, Institute of Endemic Diseases, University of Khartoum, 15-13 Khartoum, Sudan

[7]Kenya Medical Research Institute, Center for Biotechnology Research & Development, 54840-00200 Nairobi, Kenya

[8]Department of Genetics, University of Pennsylvania, PA, USA

Abstract

Functional variability at the arylamine *N*-acetyltransferase genes is associated with drug response in humans and may have been adaptive in the past owing to selection pressure from diet and exposure to toxins during human evolution.

Aims—We have characterized nucleotide variation at the *NAT1* and *NAT2* genes, and at the *NATP1* pseudogene in global human populations, including many previously under-represented African populations, in order to identify potential functional variants and to understand the role that natural selection has played in shaping variation at these loci in globally diverse populations.

†Author for correspondence: Tel.: +1 215 746 2670, Fax: +1 215 573 2326, tishkoff@mail.med.upenn.edu.

Financial & competing interests disclosure

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Materials & methods—We have resequenced approximately 2800 bp for each of the *NAT1* and *NAT2* gene regions, as well as the pseudogene *NATP1*, in 197 African and 132 non-African individuals.

Results & conclusion—We observe a signature of balancing selection maintaining variation in the 3'-UTR of *NAT1*, suggesting that these variants may play a functional role that is currently undefined. In addition, we observed high levels of nonsynonymous functional variation at the *NAT2* locus that differs amongst ethnically diverse populations.

Keywords

Africa; arylamine *N*-acetyltransferase; drug-metabolizing enzyme loci; natural selection; xenobiotic-metabolizing loci

Characterization of nucleotide variation at genes encoding xenobiotic-metabolizing enzymes in ethnically diverse populations is critical for identifying variants that play a role in xenobiotic response and for understanding the role of these genes in adaptation to diverse environments and diets during human evolution. *N*-acetylation is a major detoxification route for many drugs and carcinogens [1]. *NAT1* and *NAT2* are enzymatically distinct, with certain aromatic amine or hydrazine drugs being preferentially acetylated by *NAT1* (e.g., p-aminosalicylic acid and p-aminobenzoic acid) and others by *NAT2* (e.g., isoniazid and sulfamethazine). NAT isoenzymes are also involved in bioactivation reactions with substrates derived during Phase I metabolism mediated by CYP450 genes. These reactions generate reactive mutagenic compounds that bind to DNA [2]. It is these proposed bioactivation reactions that may result in the hypothesized role of *N*-acetylation in cancer predisposition [2]. Examples of toxins that are metabolized via *N*-acetylation include various heterocyclic amine carcinogens present in food, pyrolysis products, tobacco products and products resulting from certain industrial processes (e.g., preparation of various textiles, hair dyes, rubber and plastics) [3–5].

Human NAT1 is expressed ubiquitously and at various stages in development [6,7]. Human NAT2 isozyme shows a more restricted tissue-distribution profile, and is expressed mainly in hepatic tissue [8,9]. Most adverse drug reactions involving the *N*-acetylation pathway can be linked to the NAT2 isoform, including response to isoniazid commonly used to treat TB. Supplementary Figure 1 (see www.futuremedicine.com/doi/suppl/10.2217/pgs.11.88) illustrates the structure of the human *NAT* region, located on chromosome 8p21.3–23.1 [10]. *NAT1* and *NAT2* are separated by a 168-kb region [11,12] that encompasses a nontranscribed pseudogene *NATP1* [13]. Both functional *NAT* genes contain 870-bp intronless, protein-coding exons. *NAT1* and *NAT2* share 87% sequence similarity with each other and approximately 83–85% sequence homology with *NATP1* [13], indicating that these loci arose via gene duplication. Recent studies investigating the exon–intron structure of the *NAT1* locus have identified at least eight noncoding exons and have predicted adjacent alternative promoters potentially regulating the tissue-specific expression of the gene [14–17] (Supplementary Figure 1). In addition, three sites have been identified in the 3 ´-UTR of *NAT1*, located downstream of the coding region at positions +1085, +1203 and +1242 relative to the +1 ATG of exon 9, which are predicted to influence polyadenylation of the mRNA product (Supplementary Figure 1). By contrast, the majority of *NAT2* mRNAs

initiate from positions −8682 and −8752 relative to the +1 ATG start site of exon 1 [17]. The *NAT2* exon 1 is a short noncoding exon that is separated from the ORF coded by exon 2 by an approximately 8-kb intron (Supplementary Figure 1).

We have surveyed nucleotide sequence variation in both coding and flanking regions of the *NAT1* and *NAT2* genes, as well as the *NAT* pseudogene *NATP1*, in ethnically diverse global populations that include many previously under-represented African populations. Knowledge of the pattern of genetic diversity and haplotype structure at the *NAT* loci in ethnically diverse population groups has important implications for understanding how *NAT* genotypes contribute to xenobiotic-metabolism profiles and disease phenotypes. We have identified a number of polymorphisms that show signatures of natural selection and may influence NAT1 and NAT2 function. Our results contribute to understanding of how variation at the *NAT* loci may have been adaptive for dealing with exposure to toxins during human evolution, and have important implications for understanding individual variation in drug response.

Materials & methods

Population samples

A total of 170 individuals originating from Tanzania, Sudan and Cameroon were included in the study (Table 1). Institutional Review Board approval was obtained from the University of Maryland, College Park (MD, USA) prior to sample collection. Written informed consent was obtained from all participants and research/ethics approval and permits were obtained from the following institutions prior to sample collection: Tanzanian Commission for Science and Technology and Tanzanian National Institute for Medical Research in Dar es Salaam, Tanzania; the University of Khartoum in Sudan; the Nigerian Institute for Research and Pharmacological Development, Abuja, Nigeria; the Ministry of Health and National Committee of Ethics, Cameroon. During field collection 8 ml of peripheral blood was drawn from each individual involved 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% sodium dodecyl sulfate, 0.05% 8 mM sodium azide) [18]. DNA was extracted at a later date using the Puregene® DNA purification kit (Gentra Systems, Inc., MN, USA). All extracted DNA was quantified using Pico Green reagent (Invitrogen, CA, USA) and the Wallac Victor^{2}™ 1420 MultiLabel Counter (PerkinElmer Life Sciences, MA, USA) at 1.0 s per well.

In addition, a global panel of 166 individuals was included from the Human Genome Diversity Cell Line Panel-Centre d'Etude du Polymorphisme Humain (CEPH) [19,20]. African groups included from the CEPH diversity panel included Biaka pygmy from the Central African Republic, San from Namibia and Yoruba from Nigeria. Population samples used are summarized in Table 1, and grouped according to similar genetic ancestry as determined by structure analysis [21] of genome-wide short tandem repeat polymorphisms, indel marker variants [22], and SNPs [23,24].

PCR amplification & resequencing

PCR and sequencing primers were designed using the program Primer3 version 0.4.0 [101]. All primers used for this project are listed in Supplementary Table 1. All nucleotide positions referenced in this article are numbered according to consensus *NAT* nomenclature [25–27]. DNA samples obtained from the CEPH diversity panel were subject to whole-genome amplification (WGA) using the GenomiPhi® HY DNA amplification kit (GE Heathcare, Buckinghamshire, UK) prior to PCR amplification. DNA replication with WGA is extremely accurate due to the low error rate of φ29 DNA polymerase (1 in 10^6–10^7) compared with other enzymes. 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 approximately 500–700 bp in length. Forward primers were used for sequencing in all cases, and reverse primers were used when necessary to allow for sequence confirmation. All amplifications were performed using 1.0 unit of Platinum HiFi enzyme (Invitrogen, CA, USA) and contained 200 μM of each deoxynucleotide triphosphate, 2 mM $MgSO^4$, and 100 ng of genomic DNA in a final volume of 25 μl. Samples were denatured at 94°C for 1 min, followed by 35 cycles of 94°C for 30 s, 55°C for 30 s, and 68°C for 1 min per 1000 bp. The reaction was performed using a Peltier Thermal Cycler (MJ Research, MA, USA). PCR products obtained were run on a 1.6% agarose gel with BenchTop pGEM DNA Marker (Promega, WI, USA).

To examine nucleotide variability and identify novel variants, direct sequencing of the *NAT* gene regions was performed using the population samples listed in Table 1. PCR products were purified using the ExoSAP-IT process, as described by the manufacturer (US Biochemical Corp., OH, USA). 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, CA, USA) and analyzed on the ABI 3730×l automated capillary sequencer.

Data analysis

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 [28–30]. SNPs were called automatically using the Polyphred [31] 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 then confirmed by resequencing, using both forward and reverse primers. All regions were included for further analysis with the exception of a single problem area of *NAT1*, which contained high numbers of repetitive elements, where 235 bps of sequence at the 3′ end of primer -1182 was removed from all individuals (Supplementary Table 1). Indels and microsatellite data were not considered in the present analyses.

Diploid haplotypes were inferred across the *NAT1*, *NAT2*, and *NATP1* regions using the program PHASE version 2.1.1 [32], which reconstructs haplotypes from population genotype data using a coalescence model-based algorithm [33]. PHASE 2.1.1 enables

inclusion of triallelic sites under a model of parent-independent mutation. PHASE 2.11 also implements a recombination method (the –MR option), which allows the user to specify the relative physical location of each SNP and accounts for the decay of linkage disequilibrium (LD) with distance [34,35]. No individual was included in phase inference with missing data spanning greater than 200 bp in length. The total number of individuals included for analyses of the coding and noncoding regions is listed for each NAT locus in Table 1. The number of individuals included for analyses of the coding region only for NAT2 are listed in Table 2. The total number of individuals included in phase inference is listed for each *NAT* locus in Table 1. 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 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', which measures the estimated haplotypes fit to an approximate coalescent model.

General diversity statistics and tests of selective neutrality were calculated using the program DNAsp version 4.20.2 [36,37] at the continental and population group levels for all loci (Table 3 & Supplementary Tables 2–4). Populations with less than ten chromosomes were included in tests of selective neutrality only with pooled geographic groupings, and were not included in individual analyses (refer to Table 1 & Supplementary Tables 2–4). The published sequence for chimpanzee, *Pan troglodytes* (Ptr8-WGA990) was used for all interspecific analyses. Significance was assessed for all neutrality estimates using the coalescent simulator within DNAsp (10,000 replicates), assuming no recombination. We used the Bonferroni correction for multiple tests in order to obtain an experiment-wise 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 [38]. A sliding window approach for inferring Tajima's D (TD) statistic was also implemented to assess differing values of the statistic across the gene regions (Supplementary Figure 2). 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.

Pairwise population genetic distance (F_{ST}) values between populations using phased haplotypes for each *NAT* locus were generated using Arlequin version 2.0 [39]. Pairwise F_{ST} data matrices were used to generate 2D multidimensional scaling (MDS) plots for the *NAT1*, *NAT2* and *NATP1* haplotype data (Figure 1), using Statistica version 8 (StatSoft, Inc., 1984–2008). Analyses of molecular variance (AMOVA) calculations were performed using Arlequin version 2.0 [39] to determine the level of within and between population variation.

Network version 4.5 was used to construct median-joining (MJ) phylogenetic networks [40] for the *NAT2* region (Figure 2). Phylogenetic networks are preferable to simple, bifurcating trees for intraspecies comparison in that differing evolutionary pathways are represented in terms of cycles or hypercubes.

Pairwise LD between SNPs across the *NAT1* and *NAT2* regions was inferred using Haploview version 4.1 (Supplementary Figure 3) [41]. Haploview uses a standard

expectation-maximization algorithm to estimate the maximum-likelihood values of gamete frequencies for each SNP, from which pairwise estimates of D´ are calculated. SNPs with minor allele frequencies less that 1% were excluded for analyses of each continental group. LD between *NAT* loci was analyzed using the Haploview 4.1 map builder and HapMap populations. The LD map of the entire *NAT* region using HapMap Utah residents with northern and western European ancestry is illustrated in Supplementary Figure 4.

Results

Resequencing & SNP identification

We have characterized nucleotide variation in a total of 326, 301 and 304 globally diverse individuals for *NAT1*, *NAT2*, and *NATP1*, respectively (Table 1). Because Africa is under-represented in prior studies, we have included individuals from 15 geographically and ethnically diverse African populations practicing different subsistence modes (e.g., hunting and gathering, agro-pastoralism, agriculture and pastoralism) and representing all four major linguistic families of Africa.

We resequenced 2723 bp of the *NAT1* region, encompassing the 870-bp intronless coding region (exon 9) and 1853 bp of flanking sequence (1048 bp 5´ and 1040 bp 3´) (Supplementary Figure 1) and identified 48 SNPs, 17 of which have not been previously reported (Supplementary Tables 5 & 6). A total of eight SNPs were identified in the 870 bp exon region, two of which were nonsynonymous polymorphisms (+445G>A, +639G>T) that had been previously reported (Supplementary Tables 5 & 6). A SNP located in the 3´ region of *NAT1* following a $(TAA)_n$ repeat at position +1088, a site thought to play a role in polyadenylation of the NAT1 mRNA, was found to be highly variable, with frequencies of the +1088A variant ranging from 13–83% (Supplementary Figure 5). This SNP was in strong LD with two additional SNPs also at high frequency at positions +1095 and +1191.

We resequenced 2808 bp of the *NAT2* region, encompassing the 870-bp intronless coding region (exon 2) and 1938 bp of the flanking regions (1014 bp 5´ and 924 bp 3´) (Supplementary Figure 1) and identified 46 SNPs (including one triallelic SNP at +1362 [Supplementary Table 7]), 16 of which have not been previously reported (Supplementary Table 6). A total of 18 SNPs were identified in the 870 bp coding region of *NAT2*, three of which have not been previously reported and fifteen nonsynonymous substitutions (Supplementary Tables 6 & 7).

In order to distinguish effects of natural selection and demography on patterns of genetic variation at *NAT1* and *NAT2* loci, we analyzed nucleotide variation at the *NATP1* pseudogene. We resequenced 2834 bp of the *NATP1* region, including the 870-bp region homologous to the *NAT1/NAT2* coding exons, 1039 bp 5´ and 925 bp 3´ of this region. This region overlaps in part with that analyzed by Patin *et al*. [42]. We identified 55 SNPs, 30 of which have not been previously reported in dbSNP [102] or by the NAT nomenclature committee (Supplementary Table 7) [101]. We confirmed the presence of eight mutations that result in a nonfunctional protein product as described by Blum [13].

Haplotype variation & population differentiation

A total of 88 distinct haplotypes were observed for the 2723 bp *NAT1* region (Supplementary Table 5), indicating that recombination has affected the pattern of diversity at this locus. The pattern of LD for the entire *NAT* region is presented in Supplementary Figure 4. The LD results for *NAT1* (Supplementary Figure 3) confirm lower levels of LD in Africans compared with non-African groups at this locus, a pattern observed for other nuclear loci, [43–45] and consistent with the longer evolutionary history of African groups. Pairwise estimates of D´ indicate statistically significant LD between three 3´ SNPs at positions +1088, +1095 and +1191, which form a single haplotype block in all population groups, with the exception of Europeans.

A total of 100 *NAT2* haplotypes were inferred from the 2808 bp region, indicating that recombination has affected the pattern of diversity at the *NAT2* locus (Supplementary Table 6 & Supplementary Figure 3). Higher than expected LD 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). LD analysis indicates differing haplotype block structure across the *NAT2* region in different populations (Supplementary Figure 3).

Multidimensional scaling plots estimated from pairwise F_{ST} values were used to determine how populations cluster based on genetic differentiation (Figure 1). At *NATP1*, populations clustered based on geographic location as expected for a neutral locus (Figure 1). By contrast, distinct clustering by geography is not observed for the *NAT1* and *NAT2* loci (Figure 1). The MDS plot for the *NAT2* locus reflects clustering that corresponds to rapid and slow acetylator types, where the Bakola, Biaka, San and Hadza groups appear as outliers, likely due to high levels of genetic drift in these small hunter-gatherer populations (Figure 1). Groups clustering in the center of the plot (Kanuri, Baka and Yoruba), as well as the Bakola, Biaka and San hunter–gatherers, have a high proportion of inferred intermediate acetylator phenotypes. AMOVA results, indicating hierarchical levels of variation within and between groups, are given in Figure 1. The level of between relative to within population variation at *NATP1* is ~12%, similar to what is observed at other neutral loci in humans [46]. By contrast, the level of between relative to within population variation for *NAT2* is 9.4% and for *NAT1* is 7.6%.

Nucleotide diversity & tests of selective neutrality

Summary statistics of nucleotide diversity and statistical tests of neutrality based on allele frequency distributions at the *NAT1* and *NAT2* loci for populations pooled by major geographic region are shown in Table 3, and for individual populations in Supplementary Tables 2–4. Overall, African populations have higher levels of genetic diversity at the *NAT* loci than non-African populations. TD estimates of neutrality for *NAT1*, based on the allele frequency distribution, are negative in most cases, signifying an excess of rare variation, although none of these values were significantly different from expectation under a neutral model. However, Fay and Wu's H statistic [47] is highly negative and significant for *NAT1*, following Bonferroni correction, for most population groups (Table 3 & Supplementary Table 2). Using a sliding window analysis of TD at *NAT1* with 100-bp window lengths, we observed negative, but nonsignificant, values across most regions of the *NAT1* gene in all

populations and continental groups (Supplementary Figure 2). Notably, we observed a positive peak in TD values in most populations in the 3´-UTR region of the gene, corresponding to the location of SNPs at positions +1088, +1095 and +1191, which was statistically significant in the pooled West African, Asian and Native American populations and in several individual populations (Supplementary Figure 2). At *NAT2*, we observed negative, but nonsignificant, values of TD in Africans, both pooled and in individual populations, and positive, but nonsignificant, values in non-Africans for pooled and individual populations (Table 3 & Supplementary Table 3).

The K_a:K_s ratio of nonsynonymous to synonymous substitutions was observed to be less than one at *NAT1* (K_a:K_s = 0.242), consistent with the effects of purifying selection acting at this locus. The K_a:K_s ratio for *NAT2* was observed to be higher than that observed for *NAT1*, but still less than one (K_a:K_s = 0.813). In addition, we tested for adaptive evolution at *NAT1* and *NAT2* coding regions using a McDonald–Kreitman test [48], which compares the ratio of polymorphic (among humans) and fixed (between humans and chimpanzee) nonsynonymous and synonymous changes [48]. The McDonald–Kreitman tests were not significant for either the *NAT1* (Fisher's exact test p = 0.659) or *NAT2* (Fisher's exact test p = 0.579) loci. The McDonald–Kreitman test results for *NAT2* indicate that high numbers of replacement changes are tolerated at this locus.

We also tested for significant differences in levels of genetic diversity among the *NAT1*, *NAT2*, and *NATP1* loci using Hudson–Kreitman–Aguade tests [49] to compare ratios of intra- and inter-specific variation between each pair of loci. The Hudson–Kreitman–Aguade tests comparing variation at *NAT1* to both *NATP1* and *NAT2* were not significant (p = 0.248 and 0.251, respectively). By contrast, an Hudson–Kreitman–Aguade test comparing levels of intra- and inter-specific variation at *NAT2* relative to *NATP1* was significant (p = 0.012).

NAT2 acetylator phenotype inference

NAT2 acetylator status was inferred for each individual based on phased, diploid haplotypes, and considering only the coding region SNPs with known affect on acetylator phenotype [26]. However, in some cases individuals had novel variants with unknown phenotype (e.g., *NAT2**22 and *NAT2*23*) (SI4). A median-joining network for *NAT2* haplotypes is illustrated in Figure 2. The chimpanzee out-group is observed to branch from the human *NAT2*4* rapid node, indicating that the rapid acetylator haplotype is ancestral in humans. Inferred frequencies of rapid, intermediate and slow acetylators for global populations are shown in Table 2 & Figure 3. No acetylator phenotype is fixed in any human population. However, rapid acetylators are more prevalent in Asia and the Americas, in concordance with previous findings [50,51]. Interestingly, in Africa rapid acetylators are found at highest frequencies in foraging populations known to have some of the oldest evolutionary lineages according to mtDNA and Y chomosome evidence [52–54], the San of South Africa and the Pygmies from Cameroon (Biaka, Baka and Bakola), an observation consistent with prior studies [42,55–57]. Populations that have a complete absence of rapid acetylator types are the French and Russian, and in Africa the Hadza foragers of Tanzania, and the Mada and Fulani of Cameroon.

Discussion

Here, we present a comprehensive study of global human variation across the *NAT1*, *NAT2*, and *NATP1* loci, encompassing the 870 bp coding regions, as well as 5′- and 3′-UTRs likely to harbor regulatory elements that may influence the variable, tissue specific expression observed for NAT isozymes. We include a large number of diverse African populations that have been under-represented in prior studies in order to identify novel functional variants and gain a better understanding of the evolutionary history of the *NAT* loci within Africa. Because African populations are known to show greater genetic diversity when compared with non-African populations [22], we chose direct sequencing, over SNP genotyping of common variants with known effects on acetylator phenotype, in order to provide an unbiased description of *NAT* sequence variation.

Multiple patterns of selection at the *NAT1* locus

Because of the role of the *NAT1/NAT2* loci in the metabolism of xenobiotics present in dietary and other environmental sources, and their probable role in epigenetic regulation [58], the *NAT* loci are potential targets for natural selection. Our study confirms, in concordance with previous studies [42], that *NAT1* has only two observed nonsynonymous mutations that reach near fixation levels in most population groups. This is evident in the lack of clustering observed in the MDS plots plots generated from the pairwise population genetic distances (Figure 1). Also, consistent with this observation, tests of neutrality at *NAT1* indicate that purifying selection has prohibited nonsynonymous variation from accumulating within the coding region of the gene. In addition, analysis of the allele frequency distribution indicates an excess of rare variation (Table 3 & Supplementary Table 2), as expected under a model of purifying/background selection, consistent with observations from prior study of smaller population sets [42].

In contrast to the pattern observed in the coding region of *NAT1*, the sliding window analysis of TD indicates highly positive and significant values of TD statistic at three *NAT1* SNPs located in the 3′-UTR at positions +1088, +1095 and +1191 in nearly all populations and geographic regions (Supplementary Figure 2). In addition, these sites are in nearly complete LD and form two common haplotypes, A-T-T and T-C-A, which are maintained at high and approximately equal frequencies in nearly every population (Supplementary Figures 3 & 5). Position +1088 is known to be a polyadenylation site [59,60]. However, the functional effects of SNPs in the 3′UTR region of *NAT1* (e.g., +1088,+1095 and +1191) are poorly understood [6,61,62]. The two most common *NAT1* haplotypes, *NAT1*4* (reference sequence) and *NAT1*10*, differ by two of these three mutations (1088A and 1095A). Whether *NAT1*4* (reference sequence) and *NAT1*10* confer the same or different acetylator activity has not been confirmed [59,63,64]. These results suggest the possibility that balancing selection may be maintaining haplotype variation at these *NAT1* 3′-UTR sites, and suggests that they may play some functional role, possibly related to mRNA stability.

NAT2 acetylator frequency variation

Knowledge of patterns of genetic variation at *NAT2*, which plays a role in metabolism of drugs used to treat several common diseases such as tuberculosis and hypertension, in

ethnically diverse populations will be important for developing more efficient treatment approaches for use in ethnically diverse populations. Indeed, we show that the frequency of the *NAT2* rapid and slow acetylator genotypes differ across ethnically diverse African populations, even among those from similar geographic regions.

The NAT2 acetylator phenotype is one of the best characterized variable xenobiotic-metabolizing enzymes traits in humans. The acetylator alleles are thought to act in a codominant manner, with the rapid/slow heterozygotes showing intermediate activity. However, depending on the particular substrate administered, the intermediate (rapid/slow) acetylator phenotype may show substantial variation in activity. In addition, modification of the NAT2 substrates may also be influenced by other loci (e.g., *CYP1A2, CYP2A6, CYP2A13* and *GSTM1*) [65].

Based on results of prior studies, we were able to infer acetylator phenotypes for most of the *NAT2* haplotypes identified in the current study. *NAT2* haplotypes with undetermined phenotype effect were present at low frequencies in both African (8% frequency) and non-African populations (3% frequency), preventing inference of acetylator phenotypes in some cases (Table 2). The most common rapid acetylator haplotypes in our dataset were **4* (reference sequence), **12A* and **12B* and **13A* (Figure 3 & Table 2). The most common slow acetylator haplotypes were 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 (Figure 3 & Table 2).

The slow acetylator haplotype *NAT2 *14* (+191G>A), which is thought to give rise to an 'extreme' slow *NAT2* phenotype [66,67], was originally described as 'African-specific' because of its high frequency in African–Americans ranging from 48–55%, compared with 10% in populations of European descent [67,68]. We also observe haplotype *NAT2*14* to be African-specific, with a high frequency of *NAT2*14* in several West African groups (Kanuri [0.125], Lemande [0.1429], Yoruba [0.0833]) and a low frequency in East Africa (Table 2). This observation is consistent with other studies which show a west to east gradient of decreasing frequency of *NAT2*14* across subsaharan Africa, consistent with a possible West African origin [68,69].

Population-specific selective pressure at the *NAT2* locus

Levels of nucleotide and haplotype diversity are relatively similar between African and European populations at the *NAT2* locus (Table 3 & Supplementary Table 3). Asian and Amerindian populations have a slight decrease in diversity at both the continental and population levels (Supplementary Table 3). Tests of neutrality based on the allele frequency distribution at *NAT2* show distinct patterns for African and non-African populations, where consistently negative and positive values for estimators of neutrality are observed for Africans and non-Africans, respectively. The high frequency of *NAT2*5B* and **6A* slow acetylator haplotypes in African and European populations, but not in Asian populations (Table 2), may be indicative of the effects of natural selection acting to maintain slow acetylator phenotypes at high frequency in those geographic regions. Given our observation that two main slow acetylator haplotypes are maintained at high frequency in most

populations, and that fixation of either rapid or slow acetylator haplotypes has occurred in several ethnically diverse populations, it is possible that long-term balancing selection may be overlaid by the action of positive selection acting on specific acetylator variants in specific populations (Table 2 & Supplementary Figure 6). These patterns, in general, support observations of previous studies [56,57,70,71] which suggest that multiple modes of selection are operating at *NAT2* on a population-specific basis.

Future studies to interpret the effects of natural selection on the pattern of variation at the *NAT1* locus should focus on understanding the potential functional consequences of the three common 3′ variable sites at positions +1088, +1095 and +1191, particularly in regard to mRNA stability and tissue specific expression. In addition, the elucidation of functional effect on NAT2 enzyme activity of the novel *NAT2* haplotypes described in the current study will be informative for understanding differential response to pharmaceutical drugs and other substrates metabolized by NAT2 in ethnically diverse African populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would first like to thank the Africans who contributed their samples to this study. Special thanks to Floyd Reed for his assistance in writing scripts and creating the program analysis pipeline used in the present study, Alessia Ranciaro for her contribution to the whole-genome re-amplification of samples, as well as Wen-Ya Ko and Michael Campbell for helpful discussion.

This work was supported by the US National Science Foundation (NSF) IGERT grant 9987590 to Holly M Mortensen and Sarah A Tishkoff, NSF grants BCS 0196183, and BCS-0827436, NIH grants R01GM076637 and DP1-OD-006445-01 to Sarah A Tishkoff.

References

Bibliography

Papers of special note have been highlighted as:

- of interest

1. Hein DW. Molecular genetics and function of *NAT1* and *NAT2*: role in aromatic amine metabolism and carcinogenesis. Mutat. Res. 2002; 506:65–77. [PubMed: 12351146]
2. Grant DM. Molecular genetics of the *N*-acetyltransferases. Pharmacogenetics. 1993; 3(1):45–50. [PubMed: 8097948]
3. Kufe, DM.; Pollock, RM.; Weichselbaum, RM., et al. Holland-Frei Cancer Medicine 6 (5th Edition). BC Decker, PA, USA: 2003.
4. Schut HAJ, Snyderwine EG. DNA adducts of heterocyclic amine food mutagens: implications for mutagenesis and carcinogenesis. Carcinogenesis. 1999; 20(3):353–368. [PubMed: 10190547]
5. Felton JS, Malfatti MA, Knize MG, et al. Health risks of heterocyclic amines. Mutat. Res. 1997; 376(1–2):37–41. [PubMed: 9202736]
6. Boukouvala S, Fakis G. Arylamine *N*-acetyltransferases: what we learn from genes and genomes. Drug Metab. Rev. 2005; 37(3):511–564. [PubMed: 16257833]
7. Pacifici GM, Bencini C, Rane A. Acetyltransferase in humans: development and tissue distribution. Pharmacology. 1986; 32(5):283–291. [PubMed: 3487092]

8. Kilbane AJ, Petroff T, Weber WW. Kinetics of acetyl CoA: arylamine *N*-acetyltransferase from rapid and slow acetylator human liver. Drug Metab. Dispos. 1991; 19(2):503–507. [PubMed: 1676662]
9. Husain A, Zhang X, Doll MA, States JC, Barker DF, Hein DW. Identification of *N*-acetyltransferase 2 (NAT2) transcription start sites and quantitation of NAT2-specific mRNA in human tissues. Drug Metab. Dispos. 2007; 35(5):721–727. [PubMed: 17287389]
10. Hickman D, Risch A, Buckle V, et al. Chromosomal localization of human genes for arylamine *N*-acetyltransferase. Biochem. J. 1994; 297(Pt 3):441–445. [PubMed: 8110178]
11. Matas N, Thygesen P, Stacey M, Risch A, Sim E. Mapping *AAC1*, *AAC2* and *AACP*, the genes for arylamine *N*-acetyltransferases, carcinogen metabolising enzymes on human chromosome 8p22, a region frequently deleted in tumours. Cytogenet. Cell Genet. 1997; 77(3–4):290–295. [PubMed: 9284941]
12. Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2011; 39(Database issue):D38–D51. [PubMed: 21097890]
13. Blum M, Grant DM, McBride W, Heim M, Meyer UA. Human arylamine *N*-acetyltransferase genes: isolation, chromosomal localization, and functional expression. DNA Cell Biol. 1990; 9(3): 193–203. [PubMed: 2340091]
14. Barker DF, Husain A, Neale JR, et al. Functional properties of an alternative, tissue-specific promoter for human arylamine *N*-acetyltransferase 1. Pharmacogenet. Genomics. 2006; 16(7): 515–525. [PubMed: 16788383]
15. Butcher NJ, Arulpragasam A, Goh HL, Davey T, Minchin RF. 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. 2005; 387(Pt 1):119–127. [PubMed: 15487985]
16. Husain A, Barker DF, States JC, Doll MA, Hein DW. Identification of the major promoter and non-coding exons of the human arylamine *N*-acetyltransferase 1 gene (*NAT1*). Pharmacogenetics. 2004; 14(7):397–406. [PubMed: 15226672]
17. Boukouvala S, Sim E. Structural analysis of the genes for human arylamine *N*-acetyltransferases and characterisation of alternative transcripts. Basic Clin. Pharmacol. Toxicol. 2005; 96(5):343–351. [PubMed: 15853926]
18. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988; 16(3):1215. [PubMed: 3344216]
19. Cann HM, de Toma C, Cazes L, et al. A human genome diversity cell line panel. Science. 2002; 296(5566):261–262. [PubMed: 11954565]
20. Cavalli-Sforza LL. The Human Genome Diversity Project: past, present and future. Nat. Rev. Genet. 2005; 6(4):333–340. [PubMed: 15803201]
21. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155(2):945–959. [PubMed: 10835412]
22. Tishkoff SA, Reed FA, Friedlaender FR, et al. The genetic structure and history of Africans and African Americans. Science. 2009; 324(5930):1035–1044. [PubMed: 19407144] ▪ Uses a very large multilocus dataset to understand the current population substructure and previous migration patterns of African and African–American populations.
23. Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 2005; 1(6):e70. [PubMed: 16355252]
24. Conrad DF, Jakobsson M, Coop G, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. 2006; 38(11):1251–1260. [PubMed: 17057719]
25. Vatsis KP, Weber WW, Bell DA, et al. Nomenclature for *N*-acetyltransferases. Pharmacogenetics. 1995; 5(1):1–17. [PubMed: 7773298]
26. Hein DW, Grant DM, Sim E. Update on consensus arylamine *N*-acetyltransferase gene nomenclature. Pharmacogenetics. 2000; 10(4):291–292. [PubMed: 10862519]

27. Hein DW, Boukouvala S, Grant DM, Minchin RF, Sim E. Changes in consensus arylamine *N*-acetyltransferase gene nomenclature. Pharmacogenet. Genomics. 2008; 18(4):367–368. [PubMed: 18334921]
28. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998; 8(3):186–194. [PubMed: 9521922]
29. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998; 8(3):175–185. [PubMed: 9521921]
30. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Res. 1998; 8(3):195–202. [PubMed: 9521923]
31. Nickerson DA, Tobe VO, Taylor SL. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. 1997; 25(14):2745–2751. [PubMed: 9207020]
32. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 2001; 68(4):978–989. [PubMed: 11254454]
33. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am. J. Hum. Genet. 2005; 76(3):449–462. [PubMed: 15700229]
34. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 2006; 78(4):629–644. [PubMed: 16532393]
35. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. 2003; 73(5):1162–1169. [PubMed: 14574645]
36. Rozas J, Rozas R. DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. Comput. Appl. Biosci. 1995; 11(6):621–625. [PubMed: 8808578]
37. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 2003; 19(18):2496–2497. [PubMed: 14668244]
38. Sokal, RR.; Rohlf, FJ. Biometry (3rd Edition). NY, USA: WH Freeman and Co.; 1995. The principles and practice of statistics in biological research; p. 887
39. Excoffier LGL, Laval G, Schneider S. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evol. Bioinform. Online. 2005; 1:47–50. [PubMed: 19325852]
40. Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 1999; 16(1):37–48. [PubMed: 10331250]
41. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005; 21(2):263–265. [PubMed: 15297300]
42. Patin E, Barreiro LB, Sabeti PC, et al. Deciphering the ancient and complex evolutionary history of human arylamine *N*-acetyltransferase genes. Am. J. Hum. Genet. 2006; 78(3):423–436. [PubMed: 16416399] ▪ First comprehensive population genetic analysis of the *NAT* loci in humans, including the pseudogene *NATP1*.
43. Tarazona-Santos E, Tishkoff SA. Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (*IL13*) locus. Genes Immun. 2005; 6(1): 53–65. [PubMed: 15602587]
44. Tishkoff SA, Dietzsch E, Speed W, et al. Global patterns of linkage disequilibrium at the *CD4* locus and modern human origins. Science. 1996; 271(5254):1380–1387. [PubMed: 8596909]
45. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science. 2002; 296(5576):2225–2229. [PubMed: 12029063]
46. Tishkoff SA, Verrelli BC. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu. Rev. Genomics Hum. Genet. 2003; 4:293–340. [PubMed: 14527305]
47. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000; 155(3):1405–1413. [PubMed: 10880498]
48. McDonald JH, Kreitman M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature. 1991; 351(6328):652–654. [PubMed: 1904993]

49. Hudson RR. Estimating the recombination parameter of a finite population model without selection. Genet. Res. 1987; 50(3):245–250. [PubMed: 3443297]
50. Brockton N, Little J, Sharp L, Cotton SC. *N*-acetyltransferase polymorphisms and colorectal cancer: a HuGE review. Am. J. Epidemiol. 2000; 151(9):846–861. [PubMed: 10791558]
51. Fuselli S, Gilman RH, Chanock SJ, et al. Analysis of nucleotide diversity of *NAT2* coding region reveals homogeneity across Native American populations and high intra-population diversity. Pharmacogenomics J. 2007; 7(2):144–152. [PubMed: 16847467]
52. Knight A, Underhill PA, Mortensen HM, et al. African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr. Biol. 2003; 13(6):464–473. [PubMed: 12646128]
53. Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA. Whole-mtDNA genome sequence analysis of ancient African lineages. Mol. Biol. Evol. 2007; 24(3):757–768. [PubMed: 17194802]
54. Tishkoff SA, Gonder MK, Henn BM, et al. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol. Biol. Evol. 2007; 24(10):2180–2195. [PubMed: 17656633]
55. Patin E, Harmant C, Kidd KK, et al. Sub-Saharan African coding sequence variation and haplotype diversity at the *NAT2* gene. Hum. Mutat. 2006; 27(7):720. [PubMed: 16786516]
56. Sabbagh A, Langaney A, Darlu P, Gerard N, Krishnamoorthy R, Poloni ES. Worldwide distribution of NAT2 diversity: implications for NAT2 evolutionary history. BMC Genet. 2008; 9:21. [PubMed: 18304320]
57. Luca F, Bubba G, Basile M, et al. Multiple advantageous amino acid variants in the *NAT2* gene in human populations. PloS One. 2008; 3(9):e3136. [PubMed: 18773084] ▪ A model is proposed by these authors to account for increase in *NAT2* slow acetylator types, which incorporates dietary shift from foraging to a primarily agricultural existence, and relates to the abundance of folate in the diet and rate of folate catabolism.
58. Wakefield L, Boukouvala S, Sim E. Characterisation of CpG methylation in the upstream control region of mouse NAT2: evidence for a gene-environment interaction in a polymorphic gene implicated in folate metabolism. Gene. 2010; 452(1):16–21. [PubMed: 20026257] ▪ Suggests epigenetic regulation of the murine NAT2 expression (orthologous to human NAT1), and provides evidence that this modification varies between tissue types, and is altered by NAT2 depletion or supplementation with folate.
59. Bell DA, Badawi AF, Lang NP, Ilett KF, Kadlubar FF, Hirvonen A. 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 Res. 1995; 55(22):5226–5229. [PubMed: 7585580]
60. Bell DA, Stephens EA, Castranio T, et al. Polyadenylation polymorphism in the acetyltransferase 1 gene (*NAT1*) increases risk of colorectal cancer. Cancer Res. 1995; 55(16):3537–3542. [PubMed: 7627961]
61. Sim E, Westwood I, Fullam E. Arylamine N-acetyltransferases. Expert Opin. Drug Metab. Toxicol. 2007; 3(2):169–184. [PubMed: 17428149]
62. Zhu Y, States JC, Wang Y, Hein DW. Functional effects of genetic polymorphisms in the *N*-acetyltransferase 1 coding and 3′ untranslated regions. Birth Defects Res. A Clin. Mol. Teratol. 2011; 91(2):77–84. [PubMed: 21290563]
63. Bruhn C, Brockmoller J, Cascorbi I, Roots I, Borchert HH. Correlation between genotype and phenotype of the human arylamine *N*-acetyltransferase type 1 (NAT1). Biochem. Pharmacol. 1999; 58(11):1759–1764. [PubMed: 10571250]
64. Yang M, Katoh T, Delongchamp R, Ozawa S, Kohshi K, Kawamoto T. Relationship between *NAT1* genotype and phenotype in a Japanese population. Pharmacogenetics. 2000; 10(3):225–232. [PubMed: 10803678]
65. Jensen LJ, Kuhn M, Stark M, et al. STRING 8 – a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 2009; 37(Database issue):D412–D416. [PubMed: 18940858]

66. Fretland AJ, Leff MA, Doll MA, Hein DW. Functional characterization of human N-acetyltransferase 2 (*NAT2*) single nucleotide polymorphisms. Pharmacogenetics. 2001; 11(3):207–215. [PubMed: 11337936]

67. Bell DA, Taylor JA, Butler MA, et al. Genotype/phenotype discordance for human arylamine *N*-acetyltransferase (*NAT2*) reveals a new slow-acetylator allele common in African–Americans. Carcinogenesis. 1993; 14(8):1689–1692. [PubMed: 8102597]

68. Bayoumi RA, Qureshi MM, al-Ameri MM, Woolhouse NM. The N-acetyltransferase G191 A mutation among Sudanese and Somalis. Pharmacogenetics. 1997; 7(5):397–399. [PubMed: 9352576]

69. Cavaco I, Reis R, Gil JP, Ribeiro V. *CYP3A4*1B* and *NAT2*14* alleles in a native African population. Clin. Chem. Lab. Med. 2003; 41(4):606–609. [PubMed: 12747609]

70. Magalon H, Patin E, Austerlitz F, et al. Population genetic diversity of the *NAT2* gene supports a role of acetylation in human adaptation to farming in Central Asia. Eur. J. Hum. Genet. 2008; 16(2):243–251. [PubMed: 18043717]

71. Sabbagh A, Darlu P, Crouau-Roy B, Poloni ES. Arylamine *N*-acetyltransferase 2 (NAT2) genetic diversity and traditional subsistence: a worldwide population survey. PloS One. 2011; 6(4):e18507. (2011). [PubMed: 21494681]

72. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 1975; 7(2):256–276. [PubMed: 1145509]

73. Tajima F. Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123(3):585–595. [PubMed: 2513255]

Websites

101. Primer3 version 0.4.0. http://frodo.wi.mit.edu

102. NCBI dbSNP Short Genetic Variations. www.ncbi.nlm.nih.gov/projects/SNP

103. Hein, DW.; Grant, DM.; Sim, E.; Minchin, RF.; Boukouvala, S. Arylamine *N*-Acetyltransferase (NAT) Nomenclature. http://louisville.edu/medschool/pharmacology/consensus-human-arylamine-n-acetyltransferase-gene-nomenclature

Executive summary

- Knowledge of the pattern of genetic diversity and haplotype structure at the *NAT* loci in ethnically diverse population groups has important implications for identifying variants that play a role in xenobiotic response and for understanding the role of these genes in adaptation to diverse environments and diets during human evolution.
- In the present *NAT* analysis, we survey nucleotide sequence variation in both coding and flanking regions (~2800 bp) of the *NAT1* and *NAT2* genes, as well as the *NAT* pseudogene *NATP1*, in ethnically diverse global populations that include many previously under-represented African populations.

Materials & methods

- To examine nucleotide variability, direct sequencing of the *NAT* gene regions was performed using a global panel of 326 individuals, with an emphasis on under-represented African populations.
- Diploid haplotypes were inferred across the *NAT1*, *NAT2* and *NATP1* regions using the program phase version 2.1.1, which reconstructs haplotypes from population genotype data using a coalescence model-based algorithm.
- General diversity statistics and tests of selective neutrality were calculated. Pairwise population genetic distance values between populations using phased haplotypes for each *NAT* locus were generated. Median-joining phylogenetic networks for the *NAT2* region were constructed. Pairwise linkage disequilibrium (LD) between SNPs within the *NAT1* and *NAT2* regions was inferred.

Results

- We have characterized nucleotide variation in a total of 326, 301 and 304 globally diverse individuals for *NAT1*, *NAT2* and *NATP1*, respectively.
- We have included individuals from 15 geographically and ethnically diverse African populations practicing different subsistence modes (e.g., hunting and gathering, agro-pastoralism, agriculture and pastoralism) and representing all four major linguistic families of Africa.
- We resequenced 2723 bp of the *NAT1* region, and identified 48 SNPs, 17 of which have not been previously reported. We resequenced 2808 bp of the *NAT2* region and identified 46 SNPs (including one triallelic SNP at +1362), eight of which have not been previously reported.
- A SNP located in the 3' region of *NAT1* following a $(TAA)_n$ repeat at position +1088, a site thought to play a role in polyadenylation of the NAT1 mRNA, was found to be highly variable, with frequencies of the +1088A variant ranging from 13–83%.

- A total of 88 distinct haplotypes were observed for the 2723 bp *NAT1* region, indicating that recombination has affected the pattern of diversity at this locus. Pairwise estimates of D' indicate statistically significant LD between three 3' SNPs at positions +1088, +1095 and +1191, which form a single haplotype block in all population groups, with the exception of Europeans.
- A total of 100 distinct *NAT2* haplotypes were inferred from the 2808 bp region, indicating that recombination has affected the pattern of diversity at the *NAT2* locus. Higher than expected LD 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). LD analysis indicates differing haplotype block structure across the *NAT2* region in different populations.
- We inferred frequencies of *NAT2* rapid, intermediate and slow acetylators for this global population dataset, and report ten novel *NAT2* haplotypes.
- Using a sliding window analysis of Tajima's D at *NAT1* we observed a positive peak in Tajima's D values in most populations in the 3'-UTR region of the gene, corresponding to the location of SNPs at positions +1088, +1095 and +1191, which was statistically significant in the pooled West African, Asian and Native American populations and in several individual populations. This result indicates that SNPs at these sites are being maintained at high frequency.
- In Africa, the inferred acetylator phenotypes are found at highest frequencies in foraging populations, the San of South Africa and the Pygmies from Cameroon (Biaka, Baka and Bakola).

Conclusion

- Our *NAT1* results suggest the possibility that balancing selection may be maintaining haplotype variation at three *NAT1* 3'-UTR positions (+1088, +1095 and +1191), and suggests that these SNP variants may play some functional role, possibly related to mRNA stability.
- These *NAT1* 3' sites are in nearly complete LD and form two common haplotypes, A-T-T and T-C-A, which are maintained at high and nearly equal frequencies in most populations.
- We observe high levels of nonsynonymous functional variation at the *NAT2* locus that differs amongst ethnically diverse populations.
- We show that the frequency of the *NAT2* rapid and slow acetylator genotypes differ across ethnically diverse African populations, even among those from similar geographic regions.
- The most common rapid acetylator haplotypes in our dataset were **4*, **12A* and **12B* and **13A*. The most common slow acetylator haplotypes were observed at appreciable and approximately equal frequencies in Africans and European groups: *NAT2*5B* and *NAT2*6A*.

- Given our observation that two main *NAT2* slow acetylator haplotypes are maintained at high frequency in most populations, and that fixation of either rapid or slow acetylator haplotypes has occurred in several ethnically diverse populations, it is possible that long-term balancing selection may be overlaid by the action of positive selection acting on specific acetylator variants in specific populations.

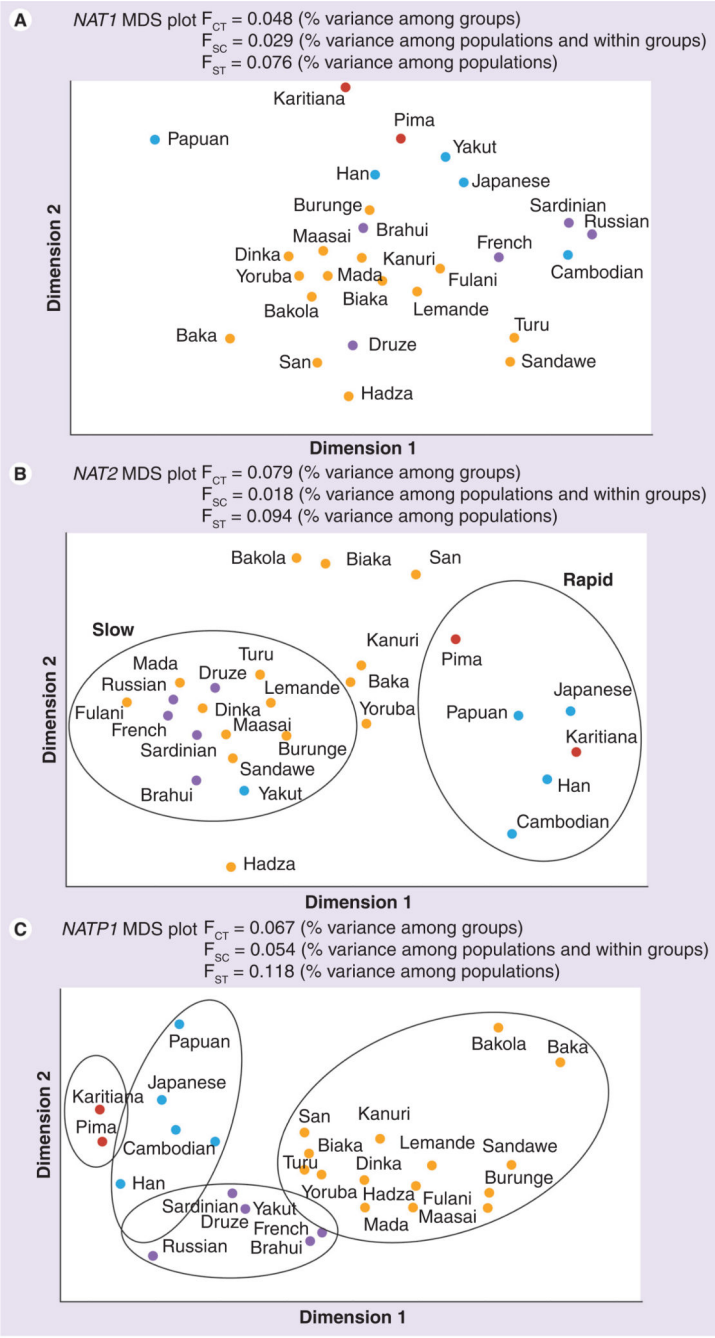

Figure 1. Multidimensional scaling plots of population pairwise F_{ST} values for the (A) *NAT1*, (B) *NAT2* and (C) *NATP1* gene regions

Analysis of molecular variance results are indicated in the inset, where variance among groups = F_{CT}, variance among populations within groups = F_{SC}, and variance among populations = F_{ST}. Yellow = Africa; Purple = Europe; Blue = Asia; Red = Americas. Population structure is specified according to the population groupings listed in Table 1. MDS: Multidimensional scaling.

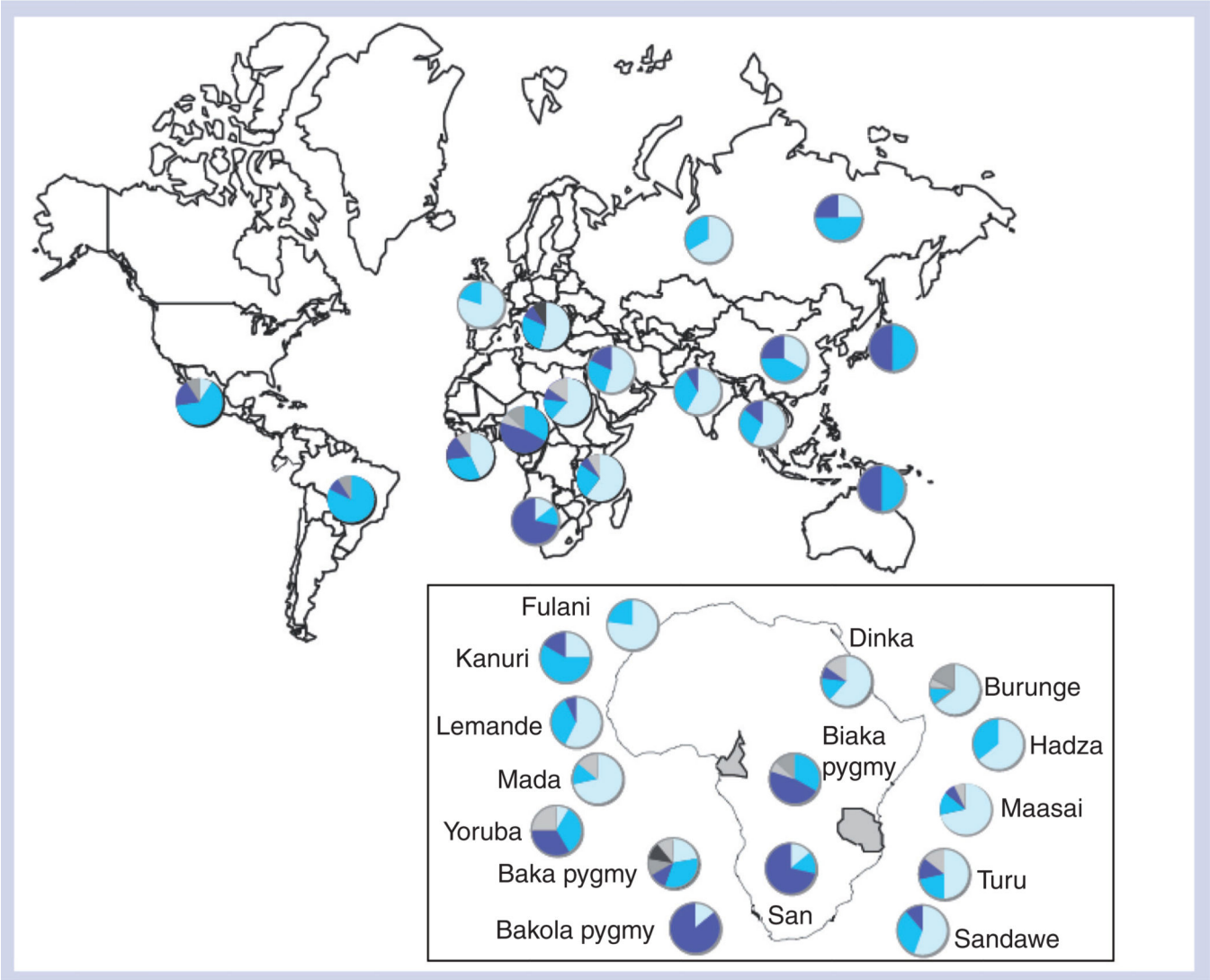

Figure 2. NAT2-inferred acetylator phenotype distribution for all populations in the current study

Phenotype inference was made based on SNPs known to affect acetylator phenotype [103]. Each individual is represented according to their inferred diploid haplotype, where light blue = slow/slow; turquoise = slow/rapid; dark blue = rapid/rapid; light gray = unknown/slow; dark gray = unknown/rapid; black = unknown/unknown. Inset shows NAT2-inferred acetylator phenotype distribution within Africa for each population group. Populations shown flanking Africa are those sampled from Cameroon and Tanzania (highlighted in gray).

Refer to Table 1 for population counts included in the present study.

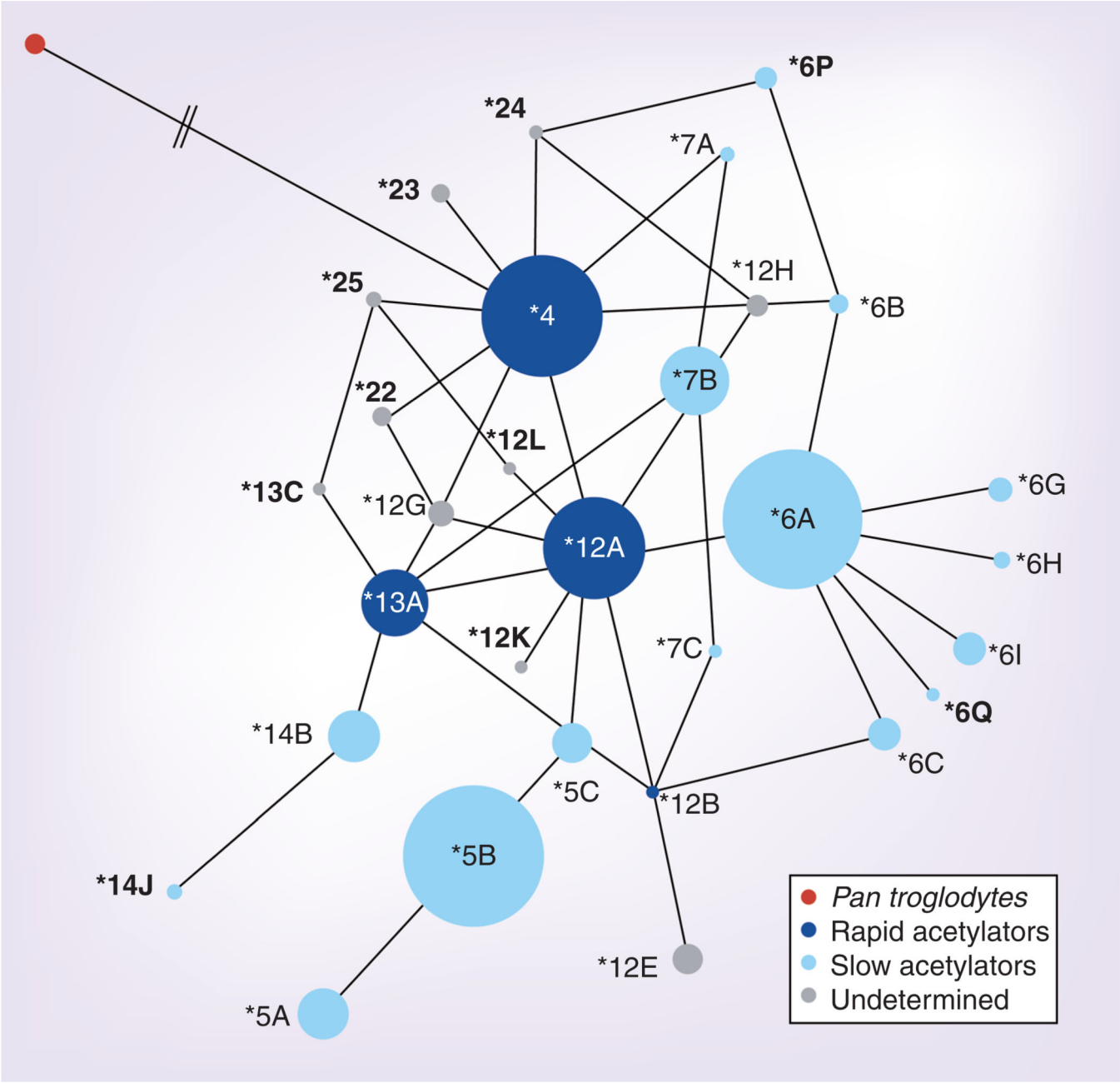

Figure 3. Median-joining network for NAT2-inferred acetylator phenotypes

For all nodes, dark blue = rapid acetylators, light blue = slow acetylators, gray = haplotypes where acetylator phenotype could not be inferred, and red indicates the chimpanzee (*Pan troglodytes*) outgroup. All nomenclature follows that recommended by the NAT nomenclature committee (e.g., NAT2*Major haplotype/sub-haplotype). Bold sub-haplotypes represent haplotypes that are novel to the current dataset.

Table 1

Populations included in the study of *NAT* nucleotide diversity.

	Population	*NAT1* 2N	*NATP1* 2N	*NAT2* 2N
East Africa				
Tanzania	Burunge[6]	34	36	34
	Hadza[4]	32	28	28
	Maasai[7]	32	26	28
	Sandawe[6]	38	36	36
	Turu[6]	32	30	30
Sudan	Dinka[1]	18	30	26
Central Africa				
Central African Republic	Biaka pygmy[3†]	30	30	30
West Africa				
Cameroon	Fulani[5]	22	26	26
	Kanuri[8]	26	26	24
	Lemande[9]	26	28	28
	Mada[8]	28	28	28
	Baka pygmy[3]	18	18	18
	Bakola pygmy[3]	14	14	14
Nigeria	Yoruba[9†]	24	24	24
South Africa				
Namibia	San[2†]	14	14	14
Total Africa		388	394	388
Europe/Middle East				
France	French[10†]	22	22	16
Israel	Druze[10†]	24	22	20
Italy	Sardinian[10†]	24	24	22
Pakistan	Brahui[11†]	24	24	22
Russia	Russian[10†]	24	12	12
Total Europe/Middle East		118	104	92
Asia				

	Population	*NAT1* 2N	*NATP1* 2N	*NAT2* 2N
Cambodia	Cambodian[13†]	14	14	14
China	Han[13†]	24	24	20
Japan	Japanese[13†]	22	20	18
New Guinea	Papuan[12†]	18	4	4
Siberia	Yakut[13†]	22	8	6
Total Asia		100	70	62
Americas				
Brazil	Karitiana[14†]	22	20	22
Mexico	Pima[14†]	24	20	20
Total Americas		46	40	40
Grand total		652	608	582

Superscript numbers indicate defined population groups based on the whole-genome structure results [35].

†Human Genome Diversity Cell Line Panel–Centre d'Etude du Polymorphisme Humain. 2N: Number of chromosomes.

Table 2

NAT2-inferred functional haplotypes by population.

	2N	*NAT2* haplotype designation																												
		*4	*5A	*5B	*5C	*6A	*6B	*6C	*6G	*6H	*6I	*6P†	*6Q†	*7A	*7B	*7C	*12A	*12B	*12E	*12G	*12H	*12K†	*12L†	*13A	*13C†	*14B	*14j†	*22†,‡	*23†,‡	*24†,‡	*25†,‡
Africa																														
Burunge	34	2	–	9	1	10	–	–	–	–	1	–	–	–	5	–	1	–	3	–	–	–	–	–	–	1	–	–	1	–	–
Biaka pygmy	30	3	–	4	–	1	–	–	–	–	–	–	–	–	–	–	14	1	2	1	–	1	–	2	–	1	–	–	–	–	–
Dinka	26	2	–	10	–	8	–	–	–	–	–	–	1	–	–	–	2	–	–	–	1	–	1	–	–	1	–	–	–	–	–
Fulani	26	.	–	13	–	9	–	–	–	–	–	–	–	–	–	–	1	–	–	–	–	–	–	2	–	1	–	–	–	–	–
Hadza	28	2	–	4	–	16	–	–	–	–	1	–	–	–	2	–	2	–	–	–	–	–	–	1	–	–	–	–	–	–	–
Kanuri	24	6	–	5	–	3	–	1	–	–	–	–	–	–	–	–	5	–	–	–	–	–	–	1	–	3	–	–	–	–	–
Lemande	28	2	1	8	1	3	–	–	–	1	2	–	–	–	–	–	2	–	–	–	–	–	–	3	–	4	1	–	–	–	–
Mada	28	2	–	13	1	6	–	1	–	–	–	–	–	–	1	–	–	–	–	–	1	–	–	–	–	2	–	1	–	–	–
Maasai	28	1	1	8	1	9	–	1	–	–	1	–	–	–	–	1	3	–	–	–	–	–	–	–	–	2	–	–	–	–	–
Baka pygmy	18	1	–	2	2	2	–	–	1	–	–	–	–	–	–	–	2	–	1	3	–	–	–	4	–	–	–	–	–	–	–
Bakola pygmy	14	1	–	–	–	–	–	–	2	–	–	–	–	–	–	–	11	–	–	–	–	–	–	–	–	–	–	–	–	–	–
San	14	5	–	2	–	–	–	–	–	–	–	–	–	–	1	–	6	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Sandawe	36	4	–	8	1	13	–	–	–	1	–	2	–	–	1	–	5	–	–	–	–	–	–	1	–	–	–	–	–	–	–
Turu	30	4	1	10	1	4	–	–	–	–	–	1	–	–	2	–	1	–	1	–	–	–	–	3	–	1	–	–	1	–	–
Yoruba	24	2	2	–	–	3	–	1	–	–	1	–	–	–	–	–	5	–	–	–	1	–	–	5	–	2	–	1	–	1	–
Europe																														
Druze	22	5	2	8	–	5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1	–	–	–	–	–	1
French	18	2	4	7	–	5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Brahui	24	5	4	2	3	7	2	–	–	–	–	–	–	–	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Russian	12	–	2	5	–	3	–	–	–	–	–	–	–	–	1	–	–	–	–	–	–	–	–	1	–	–	–	–	–	–	–
Sardinian	24	6	3	5	–	8	–	–	–	–	–	–	–	–	–	–	2	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Asia																														
Cambodian	14	4	–	–	–	3	–	–	–	–	–	–	–	1	5	–	–	–	–	–	–	–	–	1	–	–	–	–	–	–	–
Han	24	11	1	–	–	5	–	–	–	–	–	–	–	–	5	–	–	–	–	–	–	–	–	2	–	–	–	–	–	–	–

	2N	*NAT2* haplotype designation																												
		***4**	***5A**	***5B**	***5C**	***6A**	***6B**	***6C**	***6G**	***6H**	***6I**	***6P**†	***6Q**†	***7A**	***7B**	***7C**	***12A**	***12B**	***12E**	***12G**	***12H**	***12K**†	***12L**†	***13A**	***13C**†	***14B**	***14j**†	***22**†,‡	***23**†,‡	***24**†,‡	***25**†,‡
Japanese	20	14	–	–	–	4	–	–	–	–	–	–	–	–	1	–	–	–	–	–	–	–	–	1	–	–	–	–	–	–	–
Papuan	4	3	–	–	–	–	–	–	–	–	–	–	–	–	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Yakut	8	4	1	–	–	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Americas																															
Karitiana	22	10	1	8	–	–	–	–	–	–	–	–	–	–	–	–	1	1	–	–	–	–	–	1	–	–	–	–	–	–	–
Pima	22	9	–	9	–	–	–	–	–	–	–	–	–	–	–	–	3	–	–	–	–	–	–	1	–	–	–	–	–	–	–

†Haplotypes novel to the present study.

‡Novel haplotypes that bear two SNPs that define phenotypes with conflicting acetylator status.

2N: Number of chromosomes.

Pharmacogenomics. Author manuscript; available in PMC 2015 November 20.

Table 3

Summary statistics of polymorphism data by population group.

	2N	Segregating sites (n)	Synonymous substitutions	Nonsynonymous substitutions	Haplotypes (n)	Haplotype diversity	Singletons (n)	Nucleotide diversity per bp ($\times 10^{-3}$)	Watterson's theta $\theta\omega(\times 10^{-3})^{\dagger}$	Tajima's $D^{\ddagger}$	Fay and Wu's $H^{\S}$
NAT1											
Africa	388	43	6	2	72	0.868	4	1.220	0.002	-1.417^{*}	-15.340^{**}
East Africa	186	36	5	2	47	0.881	11	1.200	0.002	-1.441^{*}	-13.765^{**}
West Africa	158	31	2	2	38	0.857	3	1.210	0.002	−1.079	-15.071^{**}
Pygmy groups	62	21	3	0	24	0.893	7	1.280	0.002	−0.562	−3.101
Europe	118	26	3	1	21	0.612	9	0.950	0.002	−1.300	-11.113^{**}
Asia	102	23	2	2	11	0.65	15	0.780	0.001	−1.374	-15.057^{***}
Americas	46	7	1	0	7	0.675	1	0.760	0.001	0.968	−0.955
NAT2											
Africa	388	45	3	15	68	0.914	9	1.860	2.320	−0.556	0.822
East Africa	182	38	2	12	40	0.875	9	1.940	2.230	−0.382	0.714
West Africa	106	27	3	9	27	0.875	11	1.710	1.700	0.023	0.013
Pygmy groups	62	23	2	8	24	0.952	4	1.620	1.620	0.014	0.167
Europe	92	21	2	5	26	0.887	3	1.730	1.360	0.797	0.323
Asia	62	11	2	3	15	0.798	0	0.930	0.770	0.586	0.876
Americas	28	10	2	3	9	0.799	0	1.04	0.85	0.717	−0.529
NATP1											
Africa	394	50	NA	NA	131	0.960	6	1.800	2.640	−0.896	−8.200
East Africa	186	44	NA	NA	59	0.930	7	1.700	2.690	−1.095	-10.306^{*}
West Africa	132	40	NA	NA	59	0.964	6	1.790	2.550	−0.907	−5.479
Pygmy groups	32	27	NA	NA	27	0.977	5	2.810	2.530	0.403	−0.702
Europe	104	17	NA	NA	23	0.915	7	0.860	1.110	−0.614	−0.879
Asia	70	14	NA	NA	15	0.824	2	0.840	0.990	−0.424	1.103
Americas	40	4	NA	NA	7	0.776	1	0.550	0.400	0.972	−1.274

† Watterson (1975) [72].

‡ Tajima (1989) [73].

§ Fay and Wu (2000) [47].

* $p < 0.05$; α ’ 0.008.

** $p < 0.01$; α ’ 0.002.

*** $p < 0.001$; α ’ 0.0002.

2N: Number of chromosomes; NA: Not applicable.