Published in final edited form as:
Pharmacogenet Genomics. 2008 October ; 18(10): 877–886. doi:10.1097/FPC.0b013e3283097311.

The potentially deleterious functional variant flavin-containing monooxygenase *2*1* is at high frequency throughout sub-Saharan Africa

Krishna R. Veeramah[a,b], **Mark G. Thomas**[b], **Michael E. Weale**[e], **David Zeitlyn**[f], **Ayele Tarekegn**[g], **Endashaw Bekele**[g], **Nancy R. Mendell**[h], **Elizabeth A. Shephard**[c], **Neil Bradman**[a], and **Ian R. Phillips**[a]

a*The Centre for Genetic Anthropology, University College London*

b*Research Department of Genetics, Evolution and Environment, University College London*

c*Research Department of Structural and Molecular Biology, University College London*

d*School of Biological and Chemical Sciences, Queen Mary, University of London*

e*Department of Medical and Molecular Genetics, King's College London, Guys Tower, Guy's Hospital, London*

f*Department of Anthropology, University of Kent, Canterbury, UK*

g*Addis Ababa University, Addis Ababa, Ethiopia, State University of New York at Stony Brook, Stony Brook, New York, USA*

h*Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, New York, USA*

Abstract

Background—The drug-metabolizing enzyme flavin-containing monooxygenase 2 (*FMO2*) is the predominant *FMO* isoform present in the lung of most mammals, including non-human primates. All Europeans and Asians tested have been shown to be homozygous for a nonfunctional variant, *FMO2*2A*, which contains a premature stop codon due to a single-nucleotide change in exon 9 (g. 23238C > T). The ancestral allele, *FMO2*1*, encodes a functionally active protein and has been found in African–Americans (26%) and Hispanics (2% to 7%). Possessing this variant increases the risk of pulmonary toxicity when exposed to thioureas, a widely used class of industrial compounds. FMO2 may also be involved in the metabolism of drugs that are used to treat diseases that are prevalent in Africa.

Results and Conclusion—We conducted a survey of g.23238C > T variation across Africa that revealed that the distribution of this SNP is relatively homogeneous across sub-Saharan Africa, with approximately one third of individuals possessing at least one *FMO2*1* allele, though in some populations the incidence of these individuals approached 50%. Thus many sub-Saharan Africans may be at substantially increased health risk when encountering thiourea-containing substrates of *FMO2*. Analysis of HapMap data with the Long-Range Haplotype test found no evidence for positive selection of either 23238C > T allele and maximum-likelihood coalescent analysis indicated that this

Correspondence to Dr Krishna R. Veeramah, The Centre for Genetic Anthropology, Department of Biology, University College London, Room 108, Wolfson House, 4 Stephenson Way, London, NW1 2HE, UK, Tel: + 44 0 207 679 7418; fax: + 44 0 207 679 5052; e-mail: krishna.veeramah@ucl.ac.uk.

Conflicts of interest: none declared.

mutation occurred some 500,000 years before present. This study demonstrates the value of performing genetic surveys in Africa, a continent in which human genetic diversity is thought to be greatest, but where studies of the distribution of this diversity are few.

Keywords

drug metabolism; flavin-containing monooxygenase; flavin-containing monooxygenase 2; long-range haplotype test; pharmacogenetics; sub-Saharan Africa

Introduction

Flavin-containing monooxygenases

Flavin-containing monooxygenases (FMOs, EC1.14.13.8) catalyze the nicotinamide adenine dinucleotide phosphate-dependent oxidative metabolism of a variety of foreign chemicals that contain, as their site of oxidation, a soft nucleophilic heteroatom, such as nitrogen, phosphorous, sulphur or selenium [1–4]. Substrates include therapeutic drugs, dietary-derived compounds and environmental pollutants.

Humans possess five functional *FMO* genes, designated *FMO1–FMO5* [5–7]. All but the *FMO5* gene is present within a 220-kb cluster on chromosome 1q24.3 [7]. *FMO5* is located about 26 Mb closer to the centromere at 1q21.1 [7]. A sixth gene, *FMO6*, present within the cluster, does not produce a correctly spliced mRNA and thus seems to be a pseudogene [8]. A second *FMO* gene cluster, containing five pseudogenes, *FMO7P–FMO11P*, is located about 4 Mb centromeric of the *FMO* gene cluster [7].

Earlier studies on flavin-containing monooxygenase 2

In most mammals, including nonhuman primates, *FMO2* is the major isoform expressed in the lung [6,9–12]. A single nucleotide polymorphism (SNP) (g.23238C > T, dbSNP #rs6661174) in exon 9 that converts a glutamine codon at position 472 to a stop codon (Q472X), resulting in the production of a truncated polypeptide that is functionally inactive [10], has been identified in humans. In populations of European (n = 79) and Asian (n = 118) origin, all individuals tested have been found to be homozygous for this allele (*FMO2*2A*) [10,13]. An allele, *FMO2*1*, that has, however, been shown earlier to encode a full-length, functionally active protein [10,14] has been found in African–Americans (26%, n = 180) [10,13,15] and Hispanics (2–7%, n = 280 and 327) [16].

Substrates of human *FMO2* include thioether-containing organophosphate pesticides, such as phorate and disulfoton [17]. In this case, products of the *FMO2*-catalyzed reaction are substantially less toxic than the parent compounds [18] and as a consequence the enzyme has a protective role. *FMO2* has also, however, been shown to catalyze S-oxygenation of thiourea and some of its derivatives, such as phenylthiourea, α-naphthylthiourea and ethylenethiourea [19], producing sulfenic and/or sulfinic acid metabolites, which are more toxic than the parent compound [18]. Sulfenic acid derivatives of thioureas can deplete glutathione, leading to oxidative stress [20]; they can also bind to sulphydryl groups on proteins and thus may directly perturb cell function [21]. Thus, if exposed to thiourea or its derivatives, individuals who possess an *FMO2*1* allele are predicted to be at increased risk of pulmonary toxicity. With an estimated global production of 10 000 ton [22], thioureas are present in a wide range of industrial, household and medical products and, consequently, exposure to these chemicals is widespread.

FMOs are also involved in the metabolism of therapeutic drugs, including several that are used to treat multidrug-resistant tuberculosis [23–25], which is a major health problem in Africa,

with an estimated 544 000 deaths in 2005 (http://www.who.int/mediacentre/factsheets/fs104/en/). Evidence that at least one of these drugs, ethionamide, is a substrate for human *FMO2* is observed [4], but it is not known whether metabolism of the drug by *FMO2* will increase or decrease its efficacy or toxicity.

The rationale for studying flavin-containing monooxygenase 2 in Africans

It has been shown earlier that most African–Americans have a significant European contribution to their ancestry (about 4–30% [26–28]) so it is likely that functional *FMO2* will be found at an even higher incidence in sub-Saharan Africans than in African–Americans. As this may be important in regard to drug efficacy and public safety we assessed the distribution of the *FMO2*1* and *FMO2*2A* alleles in multiple populations across Africa. Samples from the Middle East (Turkey and Yemen) were also characterized to determine whether the *FMO2*1* allele was present at appreciable frequencies in populations outside but close to Africa.

In addition, we used the Long-Range Haplotype test [29], which examines the level of allele-specific haplotype linkage disequilibrium, to analyze data from the International HapMap project for evidence of positive selection at the g.23238C > T SNP and used sequence data from the National Institute of Environmental Health Sciences (NIEHS) SNP program to estimate the time of origin of the *FMO2*2A* allele. This will help to provide preliminary insights into the evolutionary history and future of the *FMO2* enzyme.

Materials and methods

Sample collection

DNA samples were prepared from buccal swabs from a sample of 18-year-old males unrelated at the paternal grandfather level from the following locations in and around Africa: Algeria–Mostaganem (*n* = 43), Algeria–Port Say (*n* = 118), Cameroon–Mayo Darle (*n* = 119), Cameroon–Lake Chad (*n* = 76), Ethiopia–Gambella (*n* = 106), Ethiopia–Addis Ababa (*n* = 24), Ethiopia–Borena (and surrounding area) Wollo (*n* = 36), Ethiopia–Dessie (and surrounding area) Wollo (*n* = 26), Ghana–Sandema (*n* = 90), Ghana–Navrongo (*n* = 45), Malawi–Lilongwe (*n* = 144), Malawi–Mangochi (*n* = 60), Malawi–Mzuzu (*n* = 56), Morocco–Ifrane (*n* = 70), Mozambique–Sena (*n* = 84), Nigeria–Calabar (*n* = 88), Senegal southern region (*n* = 94), Senegal–Dakar (*n* = 95), South Africa–Pretoria (*n* = 41), Sudan northern region (*n* = 136), Sudan southern region (*n* = 126), Tanzania–Kilimanjaro (*n* = 50), Turkey–East Anatolia (*n* = 31), Turkey–West Anatolia (*n* = 28), Uganda–Ssese Islands (*n* = 39), Yemen–Sena (*n* = 34), Yemen–Hadramaut region (*n* = 83), Zimbabwe–Mposi (*n* = 34). All samples were collected anonymously with informed consent. Sociological data, including age, current residence, birthplace, self-declared cultural identity and religion of the individual and of the individual's father, mother, paternal grandfather and maternal grandmother were also collected. In addition, the African populations sampled were grouped into four geographic regions (north Africa-NA, west Africa-WA, central east Africa-CEA, south-east Africa-SEA), as delineated in Table 1. The two Anatolian–Turkish samples were considered to be from a single region (TU), as were the two Yemeni samples (YE).

g.23238C > T genotyping

A 68-base pair (bp) region containing the g.23238C > T SNP was amplified by PCR using the primers *FMO2*-1414-UM (5′-TGG CTG TGA GAC TCT ATT TCG GAC CCT GCA ACT CCG A-3′) and *FMO2*-1414-LM (5′-CCA TTG CCC AGG CCC AAC CAG GCG ATA TT-3′). Each primer contained a single mismatch to its target sequence at the 3′-end penultimate nucleotide (underlined). The design of the primers was such that the amplification product would contain recognition sites for the restriction endonucleases (REs) *Mbo*I (GATC), if the

target sequence contained a C at position 23238, and *Mse*I (TTAA), if the target sequence contained a T at position 23238.

DNA was amplified in 10–l reaction volumes containing 0.4 μmol/l of each primer, 0.13 units *Taq* DNA polymerase (HT Biotech, Cambridge, UK), 9.3 nmol/l TaqStart monoclonal antibody (BD Biosciences Clontech, Oxford, UK), 200 μmol/l dNTPs and reaction buffer supplied with the *Taq* polymerase. The cycling parameters were: 5 min of preincubation at 93° C, after 37 cycles of 93°C for 1 min, 55°C for 1 min and 72°C for 1 min.

The resultant PCR product was used for two independent, complementary RE digestions that each targeted one of the two introduced RE sites. RE digestions were carried out in 10 μl volumes containing 4 μl of PCR product, 0.7 units RE (*MboI* or *MseI*), BSA and reaction buffer according to the supplier's recommendations (New England Biolabs, Hitchin, UK). All reactions were incubated overnight at 37°C. After RE digestion DNA fragments were resolved by electrophoresis through a 3.5% agarose gel. When full-length PCR product is digested with *Mbo*I, *FMO2*1* alleles are cleaved, resulting in two fragments of length 35 and 33 bp, respectively. When full-length PCR product is digested with *Mse*I, *FMO2*2A* alleles are cleaved, resulting in two fragments of length 38 and 30 bp, respectively.

Data analysis

Tests for departure of observed genotype frequencies from those expected under Hardy–Weinberg equilibrium [30] were performed using Arlequin software (Laurent Excoffier, Guillaume Laval and Stephan Schneider, Zoological Institute, University of Bern, Bern, Switzerland) [31]. Pairwise F_{ST} values were estimated from analysis of molecular variance F_{ST} values [32].

Logistic regression analysis was carried out to evaluate the differences in the *FMO2*1* allele frequency among subgroups within regions and among regions in which the subgroups had similar allele frequencies. This was undertaken by first testing for fit of the subgroup frequencies to a model, which allowed only for regional differences in the *FMO2* allele frequencies. Pearson χ^2 tests were subsequently carried out to test for overall heterogeneity within individual regions. If significant heterogeneity was found in a region, further pairwise comparisons of the subgroups within the region were made by Fisher's exact tests. For logistic regression analysis and post-hoc region and subgroup comparisons, individuals were categorized into two groups on the basis of whether or not they possessed at least one *FMO2*1* allele. (In this way the sample size equaled the number of individuals studied, *n*, rather than the number of chromosomes, 2n, thus ensuring that the observations were truly independent).

Principal coordinates analysis was performed, using GENSTAT5 software (VSN International Ltd, Hemel Hempstead, United Kingdom), on pairwise similarity matrices. Here similarity was quantified as being equal to the value of the genetic distance subtracted from 1.0 ($1\text{-}F_{ST}$). Values along the main diagonal, representing the similarity of each population sample to itself, were calculated from the estimated genetic distance between two copies of the same sample. For analysis of molecular variance-based F_{ST} distances, the resulting similarity of a sample to itself simplifies to $n/(n-1)$.

A Mantel test for the correlation between a matrix of pairwise F_{ST} values and a corresponding matrix of pairwise geographic distances was performed within the R-programming environment, using routines found in the APE package (Emmanuel Paradis, Université Montpellier II, Montpellier, France).

The Long-Range Haplotype test

The Long-Range Haplotype test was carried out using the Phase II International HapMap Project data release (http://www.hapmap.org) from four different populations. HapMap Phase II encompasses the following: the rel#21 YRI build, consisting of 3 241 616 SNPs genotyped in 30 parent–offspring trios from the Yoruba in Ibadan, Nigeria, the rel#21 CEU build, consisting of 1 105 072 SNPs genotyped in 30 parent–offspring trios from the Centre d'Etude du Polymorphisme Humain (Centre d'Etude du Polymorphisme Humain-Utah residents with ancestry from northern and western Europe) panel and the rel#21 CHB and JPT build, consisting of 3 305 784 SNPs genotyped in 45 unrelated Han Chinese from Beijing, China and 45 unrelated Japanese from Tokyo, Japan. On account of their high genetic similarity, it is accepted practice to pool the CHB and JPT datasets. In each dataset approximately one SNP is genotyped every 2 kb across the human genome.

The iHS method of Voight *et al.* [33] was applied to the g.23238C > T SNP in the HapMap Phase II YRI dataset and to the *FMO2* gene in the HapMap Phase II YRI, CEU and pooled JPT and CHB datasets, using the web-based tool Haplotter (http://pritch.bsd.uchicago.edu/data.html).

Estimating the age of the g.23238C > T mutation

Individuals of various ethnicities, including a subset of HapMap samples, have been sequenced for all exons of selected environmental response genes, including *FMO2*, as part of the NIEHS SNP program (NIEHS SNPs. NIEHS Environmental Genome Project, University of Washington, Seattle, Washington [http://egp.gs.washington.edu] [(June, 2007)]). We utilized *FMO2* sequencing data from the NIEHS SNP program to estimate the time when the g.23238C > T mutation occurred (see Supplementary Section 1 for a full explanation).

Results

The distribution of 23238C > T in Africa

The g.23238C > T allele frequencies and geographic locations for populations typed in this study are shown in Table 1 and Fig. 1. The overall *FMO2*1* allele frequency for all samples from Africa (n = 1800) was 0.153, with 28.3% of individuals having at least one *FMO2*1* allele. Across all 24 populations in Africa the observed percentage of individuals who have at least one *FMO2*1* allele ranged from 4.3 to 49.1. For samples from sub-Saharan Africa (n = 1569) the overall *FMO2*1* allele frequency was 0.170, with 31.4% of individuals having at least one *FMO2*1* allele and across these 21 populations the observed range of frequencies of *FMO2*1*-carrying individuals was 17.8–49.1%. The YE sample (n = 117) had an overall *FMO2*1* allele frequency of 0.047, with 8.5% of individuals having at least one *FMO2*1* allele. The *FMO2*1* allele was not observed in the Anatolian–Turkish sample (n = 59). No population deviated significantly from Hardy–Weinberg equilibrium (P > 0.12).

Using logistic regression on the proportion of individuals with at least one *FMO2*1* allele, significant differences were found both among regions [P < 0.0001, degrees of freedom (df) = 5] and among populations within regions (P < 0.04, df = 23). The major factor contributing to among-region differences is likely to be the noticeably lower *FMO2*1* frequencies observed in nonsub-Saharan African populations in comparison with sub-Saharan African populations. Pearson's χ^2 tests were carried out to explore within-region differences (see Table 2). The only statistically heterogeneous region was CEA (P < 0.003, df = 5). Exclusion of CEA from the logistic regression analysis resulted in no significant differences (P = 0.25, df 15) among populations within the remaining regions.

To make pairwise comparisons of regions using a Fisher's exact test, populations within each region were pooled, except in the case of CEA, which we had identified earlier as having statistically significant heterogeneity and therefore was excluded from this analysis. From these pairwise comparisons (Table 3) the following arrangement of regions based on frequencies of individuals with at least one *FMO2*1* allele could be discerned:

$$TU<(YE=NA)<(SEA=WA)$$

Further examination of populations in CEA, with pairwise Fisher's exact tests (Table 4), showed the populations in this region to be roughly split into two main groups, one consisting of north Sudan and the four Amharic populations and the other consisting of the Anuak and south Sudan. A principal coordinates analysis plot of pairwise F_{ST} values for all populations in Africa (Fig. 2) showed the Anuak of Gambella and south Sudan to be genetically close to each other with respect to the g.23238C > T SNP, probably because both of them possessing slightly elevated *FMO2*1* frequencies in comparison with the other African populations surveyed here. Addis Ababa seems to be somewhat separated from all populations, but this may be a stochastic effect because of its low sample size. A Pearson's χ^2 test comparing the frequencies of individuals with at least one *FMO2*1* allele in all populations in sub-Saharan Africa was significant ($P < 0.003$), but removing only the Anuak and south Sudan populations resulted in nonsignificance ($P = 0.526$), emphasizing that these two populations are outliers from the overall allele distribution observed across sub-Saharan Africa.

A significant correlation between matrices of pairwise genetic distances (F_{ST}) and geographic distances (kilometre) was found using the Mantel test when all populations typed in this study ($P < 0.001$) and only African populations ($P < 0.003$) were considered, but not when only sub-Saharan African populations were analyzed ($P = 0.741$), confirming the generally similar distribution of g.23238C > T alleles across sub-Saharan Africa.

When samples were grouped by self-declared ethnic identity [they were included as a separate group if there were 15 samples or more with the same self-declared ethnic identity (see Supplementary Table 1)], no significant differences were found between the same ethnic group living in multiple locations (Fisher's exact, $P > 0.24$) (see Supplementary Table 2), for example, the Amharic speakers who were sampled in three locations (Pearson's χ^2, $P = 0.47$ df = 2), or among different ethnic groups collected at the same location (Fisher's exact, $P > 0.09$).

Examining flavin-containing monooxygenase 2 for evidence of natural selection

Genotyping of the g.23238C > T SNP and many neighbouring SNPs by the International HapMap project allowed us to investigate, using the Long-Range Haplotype test [29], whether a signal suggestive of positive selection of either allele at this locus could be detected. The *FMO2*1* allele frequency in the YRI dataset is 0.175, which is similar to that observed in sub-Saharan Africa. In contrast, *FMO2*1* was absent in the CEU and CHB + JPT datasets, consistent with previous studies [10,13].

The method of Voight *et al.* [33], which uses the statistic, standardized iHS, allows direct comparisons of SNPs of different frequencies and provides a measure of haplotype conservation around the target SNP in comparison with the rest of the genome. We used the web-based tool Haplotter, which applies the method of Voight *et al.* [33] on HapMap Phase II data, to look for evidence of recent positive selection at the g.23238C > T locus in the YRI dataset.

The standardized iHS for this locus is 1.653, a value which lies in the 95th percentile on a standard normal curve. This indicates that the increased level of haplotype homozygosity on the derived T allele (as iHS is positive) is not significantly different (*P* value > 0.05, two-tailed test) from that we would expect from the genome as a whole [an iHS ≥ 2 (*P* value ≤ 0.05), would be considered statistically significant) and therefore provides no evidence of recent positive selection for either allele.

This particular analysis was not possible using the CEU or JPT and CHB datasets because the g.23238C > T SNP is monomorphic in these populations. Examination of the whole *FMO2* gene, which, however, involves examining the proportion of SNPs in the gene that have extreme iHS values in comparison with other genes (see Voight *et al.* [33]), again using Haplotter, showed no evidence of selection in any population (*P* values: CEU = 0.084736, YRI = 0.406732, JPT and CHB = 0.608040).

Analysis of National Institute of Environmental Health Sciences flavin-containing monooxygenase 2 resequencing data

The NIEHS SNP program identified, from whole gene sequencing, 19 *FMO2* coding-region variants among the Panel 2 samples (see Table 5), 14 of which were reported earlier [15] in African–Americans. Four mutations were synonymous, nine were nonsynonymous, one was found in the 3′ untranslated region, two were insertions (one of which was found in the 3′ untranslated region), one was a deletion and two were premature stop codons (including 23238C > T).

After haplotype inference of the 12 NIEHS Yoruba individuals (24 chromosomes), four chromosomes were shown to possess the 23238C allele (see Table 5). Three of these chromosomes had identical haplotypes (haplotype 1), with two synonymous changes (g. 13733G > A, g.22027G > A) in comparison with an ancestral reference sequence (elucidated from chimpanzee and macaque data), one of which was found only on a 23238C background (g.22027G > A). The fourth chromosome had an additional, nonsynonymous, mutation [g. 18237G > A (R238Q)] that was only found on a 23238C background (haplotype 2).

Addition of the 15 phased NIEHS African–American samples (30 chromosomes) showed that a further seven chromosomes possessed the 23238C SNP. Six of the seven had the two synonymous mutations whereas the other lacked the g.13733G > A variant (haplotype 3). The R238Q variant was also found in two 23238C African–American individuals whereas an additional nonsynonymous mutation [g.19910G > C (R391T)] was found in a further two 23238C chromosomes (haplotype 4).

The 23238T-possessing chromosomes found in the Yoruba, African–American, European, Hispanic and Asian NEIHS samples possessed a number of variants including nonsynonymous and synonymous mutations as well as insertions and deletions, often in combination. For example, the g.7702_7703insGAC insertion is found on the same background as a deletion (g. 10951delG), a stop codon (g.23238T) and two nonsynonymous mutations [g.7731T > C (F81S) and g.13732C > T (S195L)] (*n* = 15, haplotypes 21, 22 and 23).

Utilizing phased *FMO2* genomic data for the Yoruba NEIHS samples produced an estimate of the time of occurrence of the 23238C > T mutation of 502 404 years before (lower boundary: 2*4889*0.816*19.4 = 154790 years before, upper boundary: 2*8751*1.648*36.1 = 1 041 243 years before), using the coalescent-based method described by Griffiths and Majoram [34] (see Supplementary Section 1 for a full explanation).

Discussion

Functional flavin-containing monooxygenase 2 is found at high frequency throughout sub-Saharan Africa

The g.23238C > T SNP allele distribution reported in this study is consistent with our expectation based on the proportion of *FMO2*1* in African–American and Hispanic individuals. The ancestral allele of g.23238C > T is present at even higher frequencies in most sub-Saharan populations than in the admixed populations of the Americas, with approximately one-third of individuals possessing this variant.

Our results suggest that frequencies of g.23238C > T alleles are fairly similar throughout most of sub-Saharan Africa. Two groupings, the Anuak and south Sudan, which, however, display significantly higher frequencies of the ancestral allele (including geographically neighbouring populations) are present. The Anuak in Ethiopia are thought to be an immigrant population associated with a larger group of Anuak, who reside in southeastern Sudan (personal correspondence Ambaye–Ogato). This may go some way to explaining the similar high 23238T allele frequencies observed in these two populations.

The substantial difference in *FMO2*1* allele frequencies between northern African and sub-Saharan African populations is consistent with other genetic studies [35,36] which show the Saharan desert acting as a major barrier to gene flow. The presence of the *FMO2*1* allele at a low frequency in the Maghreb as well as in the Yemen could be a consequence of the Arab slave trade during the 8–19th centuries [37,38]. The absence of *FMO2*1* from the Turkish datasets is in agreement with earlier work, which has shown that the *FMO2*1* allele is not present in populations that are not of recent African descent [10,13].

The possible consequences of flavin-containing monooxygenase 2 functionality in Africans

Given the observed similarity in the distribution of the g.23238C > T polymorphism across sub-Saharan Africa, [population approximately 726 million World Bank estimate (http://www.worldbank.org)] and assuming the *FMO2*1* allele in Africans results in a fully functional *FMO2* enzyme [14]) we estimate that some 220 million individuals may express a functional enzyme. A considerable number are therefore potentially at risk of thiourea toxicity. This need to urgently establish the scale of potential risk is reinforced by the current widespread use of ethionamide in the treatment of tuberculosis. Drugs that are primarily metabolized by FMOs may, in general, have certain advantages over those metabolized by cytochrome P450 enzymes, because FMOs are not as readily inhibited or induced, thus reducing the risk of drug–drug interactions [39]. If, however, *FMO2* is involved in the metabolic pathway of drugs used to treat common diseases in Africa and if products of enzymatic activity have a toxic effect then great caution should be applied in the distribution and use of such drugs.

The evolution of flavin-containing monooxygenase 2 in humans

Sequence data from the NIEHS SNP programme for Yoruba and African–American samples support the suggestion that the *FMO2*1* allele results in functionally active *FMO2*. Although, we offer no statistical support, because of uncertainty in regard to certain aspects of the NIEHS data (i.e. there is possible error in haplotype inference because of the presence of very rare variants and there are large regions where successful sequencing coverage in all samples has not been achieved), it would seem that a large majority of variants that may affect the functional activity of the enzyme lie on an *FMO2*2A* background. This suggests that chromosomes possessing this allele are in mutational free fall because of the loss of function caused by the g.23238C > T mutation, whereas chromosomes with *FMO2*1* may have been evolutionarily conserved as they still retain enzymatic activity. Given the small number of g.23238C-

possessing individuals (n = 4/12) in the NIEHS Yoruba dataset it is, however, necessary to be cautious in drawing conclusions about *FMO2* activity in Africa from these data alone.

The Long-Range Haplotype test revealed no evidence for positive selection on either allele at the g.23238C > T SNP in any of three HapMap populations (YRI, CEU, CHB and JPT). Therefore, on the basis of current knowledge it is not possible to conclude that the high frequency of the derived *FMO2*2A* allele is a consequence of selective advantage. Sabeti *et al.* [29] have, however, indicated that the extended haplotype homozygosity statistic is unable to detect positive selection that has occurred more than 30 000 years ago, so we cannot dismiss the possibility that a strong selective pressure existed before this date which resulted in the increase in *FMO2*2A* allele frequency and the complete loss of the *FMO2*1* allele outside of Africa. Though not statistically significant the relatively high iHS value for the g.23238C > T SNP in the Yoruba, the almost significant P value for the FMO2 gene in the Europeans (P = 0.085) and the marked difference in European/Asian and Africa 23238C frequency suggests that this scenario may be a possibility. Additional analysis and resequencing is required (using an approach similar to that of Xue *et al.* [40]) to establish whether or not there is evidence of selection acting on either allele, either within sub-Saharan Africa or outside this region.

If selection is not the cause of the elevated frequency of the derived allele then, given the presence of *FMO2*1* throughout sub-Saharan Africa at roughly similar frequencies, the most likely explanation for why the *FMO2*1* allele is not present outside Africa is because it was lost in a bottleneck when anatomically modern humans migrated out of Africa sometime after 65 000 years ago [41,42] and that therefore the g.23238C > T SNP must have a sub-Saharan African origin before this event. Dating when the g.23238C > T SNP arose, through the use of NIEHS sequencing data, seem to be, notwithstanding the need to apply somewhat crude assumptions, to support the ancient origin of this SNP with a time of 502 404 years before, well before any estimates of the first exodus of modern humans from Africa.

Conclusion

Sub-Saharan Africa is thought to possess more human genetic diversity than the rest of the world combined. It is, however, not yet clear how this diversity is distributed and indeed what part of that diversity is not present outside the continent. Paucity of such knowledge can lead to inappropriate therapeutic, prophylactic and diagnostic intervention and increase the risk of an adverse drug reaction. Surveys such as the one performed here are not only of benefit to the indigenous populations of Africa, but are also of increasing importance in the planning of healthcare in the developed world, where the number of individuals of recent African descent is growing and, in some areas, such as the Americas and Europe, is already substantial. A need for more studies on human genetic diversity in Africa is observed; research from which all people of recent African descent, wherever living, should benefit.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the Biotechnology and Biological Sciences Research Council, the Melford Charitable Trust and NIMH R01 MH071523. The authors thank all DNA sample donors and Professor Robert C. Griffiths for his help in using recomb58.

References

1. Cashman JR. Human flavin-containing monooxygenase: substrate specificity and role in drug metabolism. Curr Drug Metab 2000;1:181–191. [PubMed: 11465082]

2. Phillips IR, Francois AA, Shephard EA. The flavin-containing monooxygenases (FMOs): genetic variation and its consequences for the metabolism of therapeutic drugs. Curr Pharmacogenomics 2007;5:292–313.
3. Phillips IR, Shephard EA. Flavin-containing monooxygenases (FMOs): mutations, disease and drug response. Trends Pharmacol Sci 2008;29:294–301. [PubMed: 18423897]
4. Krueger SK, Williams DE. Mammalian flavin-containing monooxygenases: structure/function, genetic polymorphisms and role in drug metabolism. Pharmacol Ther 2005;106:357–387. [PubMed: 15922018]
5. Lawton MP, Cashman JR, Cresteil T, Dolphin CT, Elfarra AA, Hines RN, et al. A nomenclature for the mammalian flavin-containing monooxygenase gene family based on amino acid sequence identities. Arch Biochem Biophys 1994;308:254–257. [PubMed: 8311461]
6. Phillips IR, Dolphin CT, Clair P, Hadley MR, Hutt AJ, McCombie RR, et al. The molecular biology of the flavin-containing monooxygenases of man. Chem Biol Interact 1995;96:17–32. [PubMed: 7720101]
7. Hernandez D, Janmohamed A, Chandan P, Phillips IR, Shephard EA. Organization and evolution of the flavin-containing monooxygenase genes of human and mouse: identification of novel gene and pseudogene clusters. Pharmacogenetics 2004;14:117–130. [PubMed: 15077013]
8. Hines RN, Hopp KA, Franco J, Saeian K, Begun FP. Alternative processing of the human *FMO6* gene renders transcripts incapable of encoding a functional flavin-containing monooxygenase. Mol Pharmacol 2002;62:320–325. [PubMed: 12130684]
9. Yueh MF, Krueger SK, Williams DE. Pulmonary flavin-containing monooxygenase (FMO) in rhesus macaque: expression of *FMO2* protein, mRNA and analysis of the cDNA. Biochim Biophys Acta 1997;1350:267–271. [PubMed: 9061021]
10. Dolphin CT, Beckett DJ, Janmohamed A, Cullingford TE, Smith RL, Shephard EA, Phillips IR. The flavin-containing monooxygenase 2 gene (FMO2) of humans, but not of other primates, encodes a truncated, nonfunctional protein. J Biol Chem 1998;273:30599–30607. [PubMed: 9804831]
11. Krueger SK, Yueh MF, Martin SR, Pereira CB, Williams DE. Characterization of expressed full-length and truncated FMO2 from rhesus monkey. Drug Metab Dispos 2001;29:693–700. [PubMed: 11302936]
12. Janmohamed A, Hernandez D, Phillips IR, Shephard EA. Cell-specific, tissue-specific, sex-specific and developmental stage-specific expression of mouse flavin-containing monooxygenases (Fmos). Biochem Pharmacol 2004;68:73–83. [PubMed: 15183119]
13. Whetstine JR, Yueh MF, McCarver DG, Williams DE, Park CS, Kang JH, et al. Ethnic differences in human flavin-containing monooxygenase 2 (FMO2) polymorphisms: detection of expressed protein in African–Americans. Toxicol Appl Pharmacol 2000;168:216–224. [PubMed: 11042094]
14. Krueger SK, Martin SR, Yueh MF, Pereira CB, Williams DE. Identification of active flavin-containing monooxygenase isoform 2 in human lung and characterization of expressed protein. Drug Metab Dispos 2002;30:34–41. [PubMed: 11744609]
15. Furnes B, Feng J, Sommer SS, Schlenk D. Identification of novel variants of the flavin-containing monooxygenase gene family in African–Americans. Drug Metab Dispos 2003;31:187–193. [PubMed: 12527699]
16. Krueger SK, Siddens LK, Martin SR, Yu Z, Pereira CB, Cabacungan ET, et al. Differences in *FMO2*1* allelic frequency between Hispanics of Puerto Rican and Mexican descent. Drug Metab Dispos 2004;32:1337–1340. [PubMed: 15355885]
17. Henderson MC, Krueger SK, Siddens LK, Stevens JF, Williams DE. S-oxygenation of the thioether organophosphate insecticides phorate and disulfoton by human lung flavin-containing monooxygenase 2. Biochem Pharmacol 2004;68:959–967. [PubMed: 15294458]
18. Neal RA, Halpert J. Toxicology of thiono–sulfur compounds. Annu Rev Pharmacol Toxicol 1982;22:321–339. [PubMed: 7044288]
19. Henderson MC, Krueger SK, Stevens JF, Williams DE. Human flavin-containing monooxygenase form 2 S-oxygenation: sulfenic acid formation from thioureas and oxidation of glutathione. Chem Res Toxicol 2004;17:633–640. [PubMed: 15144220]

20. Krieter PA, Ziegler DM, Hill KE, Burk RF. Increased biliary GSSG efflux from rat livers perfused with thiocarbamide substrates for the flavin-containing monooxygenase. Mol Pharmacol 1984;26:122–127. [PubMed: 6431260]
21. Onderwater RC, Commandeur JN, Menge WM, Vermeulen NP. Activation of microsomal glutathione S-transferase and inhibition of cytochrome P450 1A1 activity as a model system for detecting protein alkylation by thiourea-containing compounds in rat liver microsomes. Chem Res Toxicol 1999;12:396–402. [PubMed: 10328749]
22. UN Environment Programme. Geneva, Switzerland: The International Labour Organization and the World Health Organization; 2003. The Concise International Chemical Assessment Document 49 (CICADA 49).
23. Vannelli TA, Dykman A, Ortiz de Montellano PR. The antituberculosis drug ethionamide is activated by a flavoprotein monooxygenase. J Biol Chem 2002;277:12824–12829. [PubMed: 11823459]
24. Fraaije MW, Kamerbeek NM, Heidekamp AJ, Fortin R, Janssen DB. The prodrug activator EtaA from *Mycobacterium tuberculosis* is a Baeyer–Villiger monooxygenase. J Biol Chem 2004;279:3354–3360. [PubMed: 14610090]
25. Qian L, Ortiz de Montellano PR. Oxidative activation of thiacetazone by the *Mycobacterium tuberculosis* flavin monooxygenase EtaA and human FMO1 and FMO3. Chem Res Toxicol 2006;19:443–449. [PubMed: 16544950]
26. Reed TE. Caucasian genes in American Negroes. Science 1969;165:762–768. [PubMed: 4894336]
27. Destro-Bisol G, Maviglia R, Caglia A, Boschi I, Spedini G, Pascali V, et al. Estimating European admixture in African–Americans by using microsatellites and a microsatellite haplotype (CD4/Alu). Hum Genet 1999;104:149–157. [PubMed: 10190326]
28. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, et al. Estimating African–American admixture proportions by use of population-specific alleles. Am J Hum Genet 1998;63:1839–1851. [PubMed: 9837836]
29. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature 2002;419:832–837. [PubMed: 12397357]
30. Guo SW, Thompson EA. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics 1992;48:361–372. [PubMed: 1637966]
31. Schneider, S.; Roessli, D.; Excoffier, L. Arlequin: a software for population genetics data analysis. User manual ver 2.000. Genetics and Biometry Lab. Department of Anthropology, University of Geneva; 2000.
32. Reynolds J, Weir BS, Cockerham CC. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 1983;105:767–779. [PubMed: 17246175]
33. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol 2006;4:e72. [PubMed: 16494531]
34. Griffiths RC, Marjoram P. Ancestral inference from samples of DNA sequences with recombination. J Comput Biol 1996;3:479–502. [PubMed: 9018600]
35. Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, et al. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet 2002;70:1197–1214. [PubMed: 11910562]
36. Salas A, Richards M, De la FT, Lareu MV, Sobrino B, Sanchez-Diz P, et al. The making of the African mtDNA landscape. Am J Hum Genet 2002;71:1082–1111. [PubMed: 12395296]
37. Fisher, HJ. Slavery in the history of muslim black Africa. Vol. 1st. London: C. Hurst & Co. Ltd; 2001.
38. Richards M, Rengo C, Cruciani F, Gratrix F, Wilson JF, Scozzari R, et al. Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. Am J Hum Genet 2003;72:1058–1064. [PubMed: 12629598]
39. Cashman JR. Some distinctions between flavin-containing and cytochrome P450 monooxygenases. Biochem Biophys Res Commun 2005;338:599–604. [PubMed: 16112078]
40. Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, Kim Y, et al. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am J Hum Genet 2006;78:659–670. [PubMed: 16532395]

41. Mellars P. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. Proc Natl Acad Sci USA 2006;103:9381–9386. [PubMed: 16772383]

42. Reed FA, Tishkoff SA. African human diversity, origins and migrations. Curr Opin Genet Dev 2006;16:597–605. [PubMed: 17056248]

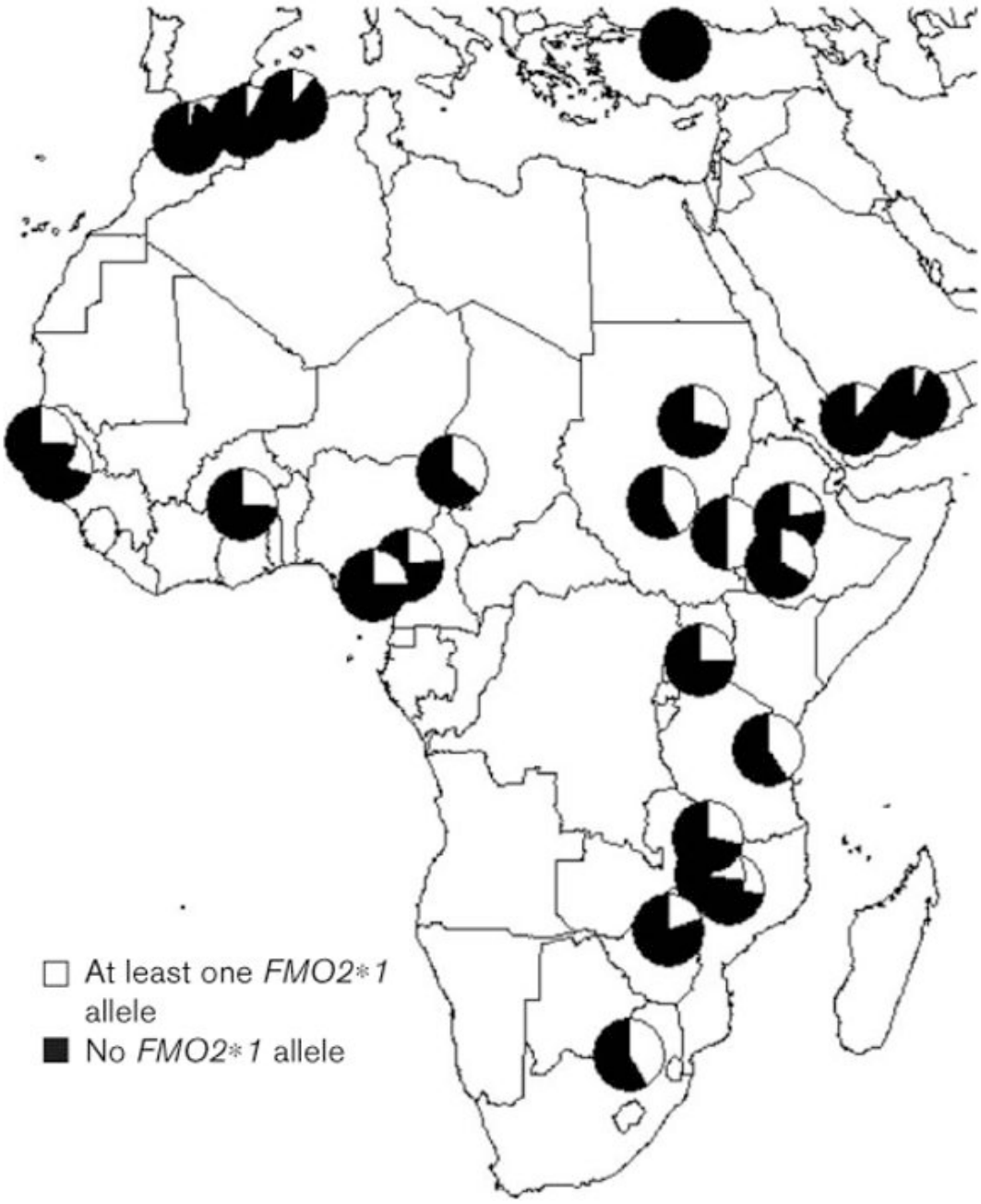

Fig. 1.
Map showing the percentage of individuals with at least one *FMO2*1* allele in Africa and two nearby countries.

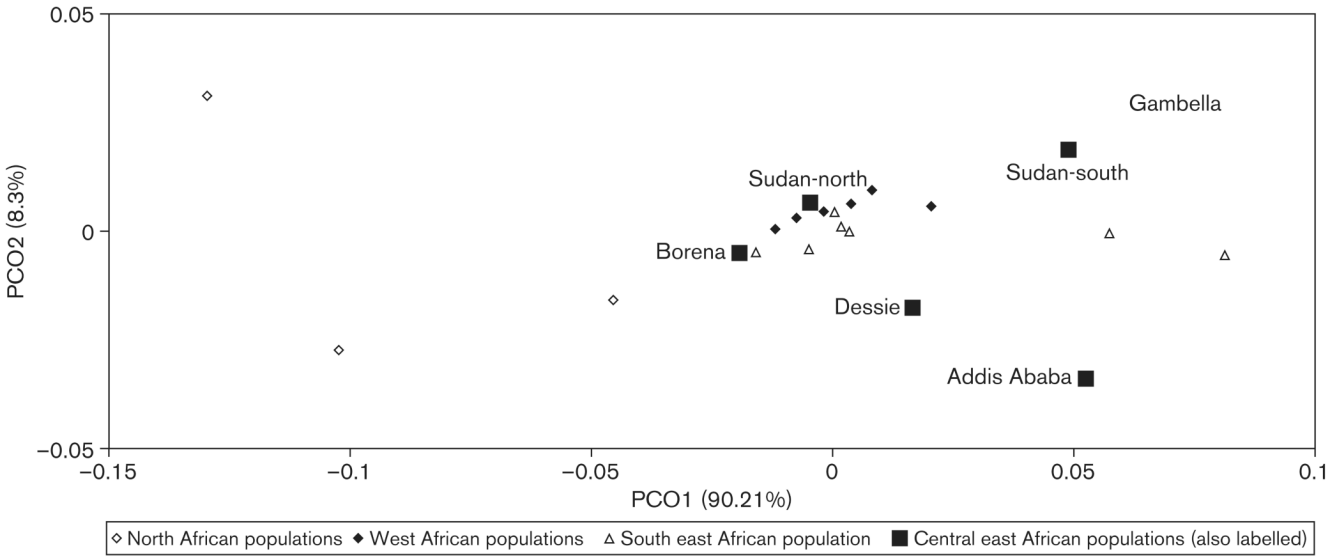

Fig. 2.
Principal coordinates analysis (PCO) plot of 23238 C > T-based population F_{ST} values.

Table 1

otype and allele frequencies

Country										
ion	Cultural identity	*FMO2*1/FMO2*1*	*FMO2*1/FMO2*2A*	*FMO2*2A/FMO2*2A*	*n*	*FMO2*1* frequency	*FMO2*2A* frequency	At least one *FMO2*1* allele	Latitude	Longitude
e	Various	2	39	78	119	0.181	0.819	34.45%	6.542	11.454
	Various	2	25	49	76	0.191	0.809	35.53%	12.28	14.752
	Bulsa	3	21	66	90	0.15	0.85	26.67%	10.726	− 1.279
	Kasena	1	7	37	45	0.1	0.9	17.78%	10.884	− 1.085
	Igbo	0	22	66	88	0.125	0.875	25.00%	4.957	8.314
	Manj	3	25	66	94	0.165	0.835	29.79%	12.986	− 15.88
	Wolof	1	24	70	95	0.137	0.863	26.32%	14.687	− 17.452
	Anuak	5	47	54	106	0.269	0.731	49.06%	8.25	34.583
a	Amhara	1	7	16	24	0.188	0.813	33.33%	9.006	38.852
lo	Amhara	0	7	29	36	0.097	0.903	19.44%	10.75	38.767
o	Amhara	1	6	19	26	0.154	0.846	26.92%	11.231	39.526
	Various	2	37	97	136	0.151	0.849	28.68%	15.213	33.036
	Various	10	43	73	126	0.25	0.75	42.06%	10.854	29.772
	Various	4	35	105	144	0.149	0.851	27.08%	− 13.983	33.774
	Various	1	16	43	60	0.15	0.85	28.33%	− 14.467	35.267
	Various	0	17	39	56	0.152	0.848	30.36%	− 11.465	34.023
e	Sena	2	22	60	84	0.155	0.845	28.57%	− 17.442	35.027

Country										
ion	Cultural identity	*FMO2*1/FMO2*1*	*FMO2*1/FMO2*2A*	*FMO2*2A/FMO2*2A*	*n*	*FMO2*1* frequency	*FMO2*2A* frequency	At least one *FMO2*1* allele	Latitude	Longitude
a										
	Bantu	2	15	24	41	0.232	0.768	41.46%	− 25.753	28.297
o	Chagga	2	18	30	50	0.22	0.78	40.00%	− 5.383	38.05
	Bantu	0	10	29	39	0.128	0.872	25.64%	− 0.452	32.564
	Shona	0	7	27	34	0.103	0.309	20.59%	− 17.309	31.328
n	Unspecified	0	5	38	43	0.058	0.942	11.63%	35.94	0.09
	Unspecified	0	10	108	118	0.042	0.958	8.47%	35.083	− 2.183
	Berbers	0	3	67	70	0.021	0.979	4.29%	33.588	− 5.165
a	Anatolian Turks	0	0	31	31	0	1	0.00%	40.277	33.254
ia	Anatolian Turks	0	0	28	28	0	1	0.00%	39.684	31.212
	Unspecified	0	4	30	34	0.059	0.941	11.76%	15.409	44.242
	Unspecified	1	5	77	83	0.042	0.958	7.23%	16.811	49.942
	Total	43	477	1456	1976	0.142	0.858	26.32%		

Pharmacogenet Genomics. Author manuscript; available in PMC 2009 October 1.

Table 2

Pearson's χ^2 test on individual regions

Region	df	χ^2	*P* value
CEA	5	18.12	0.003*
NA	2	2.15	0.34
SEA	7	7.51	0.37
WA	6	7.34	0.29
Yemen	1	0.64	0.43

df, degrees of freedom.

** P* value is less than 0.05.

CEA, central east Africa; NA, north Africa; SEA, south-east Africa; WA, west Africa.

Table 3

Fisher's exact tests between regions

	WA	SEA	NA	TU
SEA	0.7915	\	\	}
NA	0.0001*	0.0001*	\	}
TU	0.0001*	0.0001*	0.0294*	}
YE	0.0001*	0.0001*	0.8361	0.0321

*P value is less than 0.05.

NA, north Africa; SEA, south-east Africa; TU, Turkey; WA, west Africa; YE, Yemen.

Table 4

Fisher's exact tests between CEA populations

	Gambella	Addis Ababa	Borena, Wollo	Dessie, Wollo	Sudan north
Addis Ababa	0.1812				
Borena, Wollo	0.0018*	0.2418			
Dessie, Wollo	0.0492*	0.7598	0.5475		
Sudan north	0.0013*	0.6340	0.2982	1.0000	
Sudan south	0.2933	0.5008	0.0180*	0.1883	0.0277*

*P value is less than 0.05.

CEA, central east Africa.

g. 19679A > G (E314G)	g. 19839A > G (Synon) *	g. 19910G > C (R391T) *	g. 22027G > A (Synon) *	g. 22060T > G (N413K) *	g. 23087A > G (Synon)	g. 23238C > T (Q472X) *	g. 23300A > G (3′ UTR) *	g.23412_23413insT (3′UTR) *	Ethnic idenity of NIEHS samples					
A	**A**	**G**	**G**	**T**	**A**	**C**	**A**	**−**	**AA**	**YR**	**AS**	**EU**	**HI**	**Total**
			A						2	3	0	0	0	5
			A						2	1	0	0	0	3
			A						1	0	0	0	0	1
		C	A						2	0	0	0	0	2
	G					**T**	G		1	0	0	0	0	1
	G					**T**	G		1	4	1	2	3	11
G	G					**T**		+	0	0	0	1	0	1
	G			G		**T**			0	1	0	0	0	1
	G					**T**			9	0	22	23	13	67
	G					**T**	G		0	0	0	2	2	4
	G				G	**T**			0	0	0	0	1	1
G	G					**T**		+	0	0	0	1	0	1
	G			G		**T**			0	1	0	0	0	1
	G					**T**	G		1	0	0	0	0	1
G	G					**T**		+	0	0	0	1	0	1
	G			G		**T**			0	0	0	1	0	1
	G					**T**			2	3	0	7	4	16
	G					**T**		+	0	4	0	0	0	4
	G					**T**	G		0	0	5	3	6	14
G	G					**T**		+	1	0	12	0	9	22
	G			G		**T**			6	5	0	0	1	12
	G					**T**			2	0	0	0	0	2
	G					**T**	G		0	0	0	1	0	1
	G					**T**			0	1	0	0	2	3
	G					**T**			0	1	8	1	1	11
G	G					**T**		+	0	0	0	1	2	3

g. 19679A > G (E314G)	g. 19839A > G (Synon) *	g. 19910G > C (R391T) *	g. 22027G > A (Synon) *	g. 22060T > G (N413K) *	g. 23087A > G (Synon)	g. 23238C > T (Q472X) *	g. 23300A > G (3′ UTR) *	g.23412_23413insT (3′UTR) *	Ethnic idenity of NIEHS samples					
A	A	G	G	T	A	C	A	–	AA	YR	AS	EU	HI	Total
									30	24	48	44	44	190

Pharmacogenet Genomics. Author manuscript; available in PMC 2009 October 1.