HMG Advance Access published November 16, 2010

Global Patterns of Genetic Diversity and Signals of Natural Selection for Human ADME Genes

Jing Li[1, 2, 3], Luyong Zhang[2,*], Hang Zhou[1], Mark Stoneking[3], Kun Tang[1,*]

1 CAS-MPG Partner Institute and Key Laboratory for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

2 National Drug Screening Laboratory, China Pharmaceutical University, Nanjing, China

3 Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[*] corresponding authors:

Kun Tang Telephone: 86-21-54920277 Fax: 86-21-54920451 Email: tangkun@picb.ac.cn

Luyong Zhang Telephone: 86-25-85391036 Fax: 86-25-85303260 Email: lyzhang@cpu.edu.cn

Abstract

Genetic polymorphisms in many genes related to drug absorption, distribution, metabolism and excretion (ADME genes) contribute to the high heterogeneity of drug responses in humans. However, the extent to which genetic variation in ADME genes may contribute to differences among human populations in drug responses has not been studied. In this work, we investigate the global distribution of genetic diversity for 31 core and 252 extended ADME genes. We find that many important ADME genes are highly differentiated across continental regions. Additionally, we analyze the genetic differentiation associated with clinically relevant, functional polymorphism alleles, which is important for evaluating potential among-population heterogeneity in drug treatment effects. We find that ADME genes show significantly greater variation in levels of population differentiation, and we find numerous signals of recent positive selection on ADME genes. These results suggest that genetic differentiation at ADME genes could contribute to population heterogeneity in drug responses.

Introduction

An individual's response to drug treatment is an essential aspect of the therapeutic outcome, as aberrant drug responses can lead to a lack of therapeutic effect or adverse drug response (ADR). Moreover, drug response is highly variable both at the intra-population and inter-population levels (1). ADR incidents due to poor understanding of variability in drug response can have severe medical and economic consequences (2).

Many factors can influence one's response to drug therapy, including age, gender and genotype. Genes involved in absorption, distribution, metabolism and excretion (ADME) of drugs, and various drug target genes comprise the most important genetic determinants (3, 4). Based on metabolic properties *in vivo*, the ADME proteins can be generally classified into 3 groups (5, 6): Phase I metabolizing enzymes, which primarily consist of the cytochrome P450 enzymes and carry out enzymatic oxidation, reduction, and hydrolysis reactions that expose or add functional groups to produce polar molecules; Phase II metabolizing enzymes, which add endogenous compounds to the molecules after Phase I metabolism and further increase solubility, and include arylamine N-acetyltransferase (NAT), cytosolic glutathione S-transferases (GST), etc; and drug transporters, which include efflux transporters (e.g., the ATP binding cassette (ABC) proteins) and uptake transporters (e.g., the solute carrier proteins (SLC)), and play key roles in drug absorption, distribution, and excretion in the human body(7).

Genetic polymorphisms in the ADME genes are known to contribute significantly to variation in drug response (8). For example, genetic variants can alter

the drug response phenotypes by changing the expression level (e.g., CYP2D6), the protein coding sequence (e.g., CYP2C9), or the mRNA splicing form (e.g., CYP3A5) of ADME genes (9-13). Functional genetic polymorphisms have also been discovered and analyzed in many other common ADME genes, including CYP1A1, CYP1A2, CYP2E1, CYP3A4, ABCB1, SLCO1B3, UGT2B7, UGT2B15 (11, 12, 14-17). As a result, genetic tests for relevant ADME polymorphisms are becoming increasingly common; according to a recent study by the Federal Drug Administration, approximately one-quarter of the prescriptions written in the United States in 2006 contained pharmacogenetic labeling (18).

Substantial genetic differences and phenotypic variation may also occur for the ADME genes across different populations. Recent genome-wide genetic diversity studies have demonstrated significant population structure at both global and local levels (19-21). Ample evidence exists for strong inter-population differences in drug responses for specific genes, such as CYP2C9, CYP2D6 and NAT2, and for drugs such as Clopidogrel, Mercaptopurine, Omeprazole, Warfarin, etc. (13, 22-24). However, development and clinical trials of drugs in use today were predominantly carried out in populations of European ancestry from the US or Europe. Many countries, especially in the developing world, rely on US FDA/European Medicines Agency (EMEA) guidelines for safety levels and optimal therapeutic dosages. A comprehensive understanding of the inter-ethnic genetic differences in the ADME genes and their impact on drug response is therefore crucial to guide the effective global prescription of drugs.

Nonetheless, most pharmacogenetic studies to date have focused on a limited number of ADME genes in a limited number of populations; only a few studies investigated population differentiation in ADME genes at the global or continental scale. One such study found substantial allele frequency differences among statistically inferred clusters of ancestry as well as ethnically labeled groups(25). Another investigated the genetic diversity pattern of CYP2D6 across 52 world-wide populations and found that the CYP2D6 genetic diversity distribution did not differ from that of random markers (26). This analysis was subsequently extended to three P450 enzymes - CYP2C9, CYP2C19, and CYP2D6 - in more populations, and revealed that the three genes have distinctive global diversity patterns (27). Analysis of the SLCO1B1 gene revealed a similar pattern of SLCO1B1 global genetic diversity, compared to other genomic markers(28). Finally, the most comprehensive study to date of ADME genes genotyped 167 polymorphisms in 27 ADME genes in East Asian, European and African groups (29). However, the general population differentiation pattern of ADME genes remains unclear as only a limited numbers of SNPs were analyzed for each gene. A systematic and comprehensive worldwide survey of the general patterns of genetic diversity is needed to reveal the population differentiation patterns for each gene and for the ADME genes as a group.

Moreover, many ADME genes play important roles in defense against xenobiotics. These genes might have undergone substantial local selective pressures during early human migrations throughout different geographical regions. Natural selection leaves specific signals in the genetic diversity patterns around the selected

loci, such as high/low population differentiation levels, skewed allele frequency spectra, and/or extended haplotype homozygosity (30). Indeed, evidence of recent positive selection has been reported for numerous ADME genes, such as ABCB1(31), CYP3A5(32, 33), NAT2(34, 35), and several ABC transporters (25). Therefore, a systematic survey of signals natural selection in ADME genes may help explain the inter-population differentiation patterns in the ADME genes. It may also help identify additional functional polymorphisms(36).

In this study, we systematically analyze the patterns of genetic diversity and signals of natural selection in 283 ADME genes (31 core and 252 extended ADME genes) in 62 global populations. We further discuss how the inter-population genetic diversity profiles and potential functional polymorphisms might impact the response to various drug compounds. This study is the first to systematically investigate worldwide population differentiation and potential signals of natural selection in a comprehensive list of ADME genes. Follow-up research based on this framework may help evaluate the portability of specific drugs across different populations.

Results

We obtained genome-wide SNP data for different populations from two data sources. The first is Human Genome Diversity Panel (CEPA-HGDP)(19), and the second is HapMap 3 dataset(37). A total of 580,612 SNPs were used, analyzed in 1951 individuals from 62 populations, and which were further assigned into 7 continental regions, namely Africa, Middle East, Europe, South/Central Asia, East

Asia, Oceania, and America (see Supplementary Table S1). We defined the ADME gene lists according to the PharmaADME database (4, 38). The proposed ADME consensus genes are divided into two lists – the core gene list and the extended gene list. The core genes are deemed to be the most important as they are directly involved in drug metabolism and/or have a significant impact on drugs pharmacokinetic profile. The extended genes include other genes thought to be associated with drug metabolism. The PharmaADME consensus has been used in previous studies (4), and the core gene list has been adopted by several commercial genotyping products (e.g. Affymetrix DMETTM, Illumina VeraCode® ADME and TaqMan® DME).

Given the significance of the core ADME genes in both research and industry, we investigated whether this group of genes has a different global population differentiation pattern when compared to the extended genes and to control genic and non-genic regions. In total, we studied 31 core ADME genes and 252 extended ADME genes (see Supplementary Table S2).

Global population differentiation in ADME genes

To investigate global differentiation among ADME genes, we calculated the weighted average F_{st} of multiple sites from haplotypes of each gene across all 62 populations (hereafter referred to as GA- F_{st}, see Methods) for the core ADME genes and extended genes. Genes in high LD were concatenated into single loci (see Materials and Methods). GA- F_{st} value measures population differentiation based on the full polymorphism composition of a genomic region. High GA- F_{st} values indicate

that the haplotype composition is substantially different among different populations, suggesting that the functional variants associated with specific haplotypes are also distributed heterogeneously across populations. On the other hand, genes with low GA- F_{st} may have limited genetic as well as functional heterogeneity across populations. The GA- F_{st} values vary widely among the ADME genes, with the maximum GA- F_{st} found in the core gene CYP3A5 at 0.310, and the minimum of 0.0526 in the extended gene SLC04C1 (Supplementary Table S3).

Under neutral evolution, population differentiation is influenced solely by random genetic drift (which increases differentiation) vs. migration (which decreases differentiation), and these are expected to have the same average effect across the genome. On the other hand, natural selection impacts population differentiation in a locus-specific manner: local positive selection tends to increase population differentiation, whereas negative or balancing selection tends to decrease population differentiation. GA- F_{st} is therefore often used to detect departures from evolutionary neutrality (39-41). In this study, we sampled 500 random non-genic regions which are expected to be mainly influenced by neutral demographic processes (see Materials and Methods). Figure 1 shows the comparison of the GA- F_{st} distribution of the non-genic regions, and those of the two ADME gene groups (the core and extended ADME genes). It is apparent from Figure 1 that compared to the GA- F_{st} distribution of the non-genic regions, those of the two ADME gene groups (the core and extended ADME genes) have a much bigger range. The F test of equal variance is highly significant between the core genes (0.130±6.2%, mean ± sd) and non-genic regions

(0.109±3.0%, mean± sd, P_value = 5.882×10^{-11}), as well as between the extended genes (0.119±4.1%, mean ± sd) and non-genic regions (P_value = 5.267×10^{-9}, Table 1). The average GA- F_{st} value is significantly higher in the extended ADME genes (Mann Whitney U test P_value = 0.005, Table 1) than in the non-genic regions. Although the average GA- F_{st} value for core ADME genes does not differ significantly from that for non-genic regions (Table 1) this is likely to reflect limited power due to the small number of core ADME genes, as the average GA- F_{st} value for core ADME genes is actually larger than that for extended ADME genes (Table 1). These results strongly suggest an excess of positive selection on the ADME genes.

To determine whether the ADME genes are under a stronger influence of selection than other genes in the genome, we compared the two ADME gene categories against a group of randomly sampled genes (see Materials and Methods) and between each other. The mean GA- F_{st} values do not differ significantly among the three gene categories (Table 1). However, the core ADME genes have a significantly higher variance in the GA- F_{st} distribution than both the extended ADME genes (F test P_value = 0.001) and the random genes (F test P_value = 0.012). Moreover, the difference in the variance in the GA- F_{st} distribution for the extended ADME genes vs. the random genes approaches, but does not reach, statistical significance (F test P_value = 0.067, Table 1). These results together suggest that while most of the ADME related genes were not more influenced by natural selection than the other genes in the genome, some core genes do show more evidence of selection, which led to greater variance in the distribution of GA- F_{st} values.

We use the lowest and highest 1% of the non-genic GA- F_{st} values as the cutoffs for evidence of departure from neutrality, which are 0.0503 and 0.2032 respectively. No ADME genes have GA- F_{st} values lower than the lower cutoff, whereas three core genes (CYP3A4, CYP1A2 and CYP3A5) and 12 extended genes (ALDH2, GSTCD, CYP3A7, CYP3A43, ABCC12, SLC28A2, ABCC11, PPARA, CYP26A1, CYP26C1, CES2 and CYP1B1) are above the upper cutoff (Supplementary Table 3). These genes are thus candidates for functional heterogeneity across ethnic groups.

Pairwise population differentiation in ADME genes

The GA- F_{st} values indicate the overall degree of population differentiation in the ADME genes. It is also desirable to know how exactly are the inter-population differences distributed, e.g. whether the global differentiation is mainly explained by a few outlier populations, or by differences among major geographical regions. To answer this question, we used the pairwise weighted average F_{st} for the multiple sites from haplotypes of each gene (hereafter referred as PA- F_{st}, see Methods) to measure the differentiation between each pair of populations. The PA- F_{st} values are plotted as contour maps (Figure 2, Supplementary Figure S1). In these contour maps, PA- F_{st} values are represented as boxes of black/white gradients, where darker gradients represent higher PA- F_{st} values. Each population was also assigned to one of the seven continental regions as mentioned above (Supplementary Table S1). We observed various kinds of pairwise population differentiation patterns among different ADME genes. It seems most of these patterns can be described as differentiation among the 7

continental regions. To formally evaluate whether the variances among continental regions are higher than those among populations within a region, we carried out an analysis of molecular variance (AMOVA) and compared the variances between the two levels with an *F*-test (see Materials and Methods). Indeed, most ADME genes are more differentiated among regions than within regions (see Supplementary Table S3 for details).

Some typical population differentiation patterns in the ADME genes are presented in Figure 2. CYP3A4 (Figure 2a) shows distinct differences in the genetic profiles between African and non-African populations (*F* test P_value =0.006). CYP1A2 (Figure 2b) exhibits strong differences between European and non-European populations (*F* test P_value =0.001). East Asians are clearly differentiated from other populations at NAT2 (*F* test P_value =0.002, Figure 2c). On the other hand, significant differentiation does not always occur only between continental regions. For example, UGT2B7 (Figure 2d) seems to be as highly differentiated among populations within continental regions as among continental regions (*F* test P_value = 0.106). The inter-population differentiation for UGT2B7 is especially high in East Asia and America (Figure 2d). This suggests that for drugs processed by such genes, there may be significant differences in drug response even among closely related ethnic groups.

In some cases, no strong differences can be observed among populations or regions. For example, NAT1 shows low PA- F_{st} values across all pairwise comparisons (P_value=0.031), which implies that its sequence is highly conserved

among human populations (Figure 2e). Still, many of the PA- F_{st} patterns for ADME genes are strongly different from neutral patterns (Figure 2f), suggesting that natural selection has played a major role in shaping the population differentiation of ADME genes. Genes with high population differentiation may indicate substantial diversity in drug response, and care has to be taken when a drug metabolized or transported by these genes is to be applied to groups with very different genetic profiles.

Global population differentiation of functional ADME SNPs

All the above analyses are based on general patterns of genetic variation in each gene. However, the genetic variants most relevant to clinical application are the functional variants that directly affect enzyme efficacy or expression. We first looked for high frequency non-synonymous SNPs, which are assumed to have higher potential functional impact on the drug response profile. Eighteen candidate SNPs (SNP1-18, Table 2) were identified and many have been previously reported to have functional significance (References see Table 2). We further included two reported functional SNPs in non-coding regions (SNP19, 20, Table2). Finally, six additional candidate SNPs were added from the pharmaADME core ADME SNP list (http://www.pharmaadme.org/). For any candidate SNPs that are not represented in our merged genotyped data, we found tagging SNPs (see Materials and Methods). We then studied the global population differentiation for these candidate functional SNPs and tested whether the GA- F_{st} of the specific genes predicts the population differentiation of these functional SNPs.

The 26 candidate functional SNPs are listed in Table 2. The GS- F_{st} values (0.109±6.4%) of these SNPs fall in a similar range as the GA- F_{st} values of the ADME genes. Two SNPs, rs776746 (in CYP3A5) and rs4149117 (in SLCO1B3) are of particular interest for their high GS- F_{st} values (0.332 and 0.208). They therefore may contribute significantly to inter-ethnic variation in response to drugs metabolized by these genes.

More interestingly, we found that the GS- F_{st} values of these functional SNPs strongly correlate with the GA- F_{st} of their respective genes ($r^2 = 0.80$; Figure 3). This suggests that the population differentiation level of the functional SNPs can be well interrogated by the GA- F_{st} of the corresponding gene. In the absence of full characterization of potential functional variants, the GA- F_{st} patterns of ADME genes may therefore provide an indirect assessment for potential inter-population variation in drug responses.

Detecting local natural selection in the ADME genes

The PA-F_{st} analyses revealed highly variable patterns of ADME genetic differentiation among worldwide populations, and many of them are incompatible with neutral expectations. To formally analyze the signatures of selection in the ADME genes, we applied two commonly used statistics: lnRsb and CLR (see Materials and Methods). The lnRsb method has been shown to possess good power to detect recent positive selection, with higher power toward fixed or nearly fixed sweeps (42). The CLR test is a model based approach, which preferentially detects

fixed sweeps(43). Therefore lnRsb is used here to describe the distribution of putative signals of positive selection, and CLR is used for validating fixed sweeps.

In general, we found that a substantial fraction of ADME genes exhibit evidence of recent positive selection. There is good agreement between the results of the lnRsb and CLR tests, which supports the reliability of the analysis (Figure 4 and Supplementary Figure S2). Figure 4 shows the signals of positive selection in the core ADME genes in different populations. In the African populations, signals of selection identified by lnRsb seem to occur sporadically in different genes and populations. Nonetheless, the CLR test supports the indication of selection on DPYD in the Mandenka and Luhya populations, and on ABCB1 in the African American and Yoruba populations. Less clear patterns in the African populations may result from the overall high level of population differentiation in Africans. On the other hand, signals of selection in the non-African populations seem to depend highly on the geographic region. Specifically, the Phase I drug metabolizers CYP2E1, CYP3A4 and CYP3A5, Phase II drug metabolizers GSTP1 and NAT1, as well as the drug transporters ABCB1 and SLCO1B1 show rather consistent signals of positive selection across the Middle East, Europe and Central South Asia. In East Asia, the signals of selection are less consistently shared among all the populations. But signals are found in numerous populations for the Phase I drug metabolizers CYP2C8, CYP2C9, CYP2D6, CYP2E1, CYP3A4 and CYP3A5, the Phase II drug metabolizers GSTM1, GSTP1 and UGT1A1, as well as the drug transporter ABCB1. The two drug transporters SLC22A1 and SLC22A2 appear to be selected in multiple American populations

(Figure 4).

Several genes also exhibit significant signals in the CLR test, including CYP2E1 (in Russian and Brahui) and SLCO1B1 (in Sindhi and Pathan). CYP3A4 and CYP3A5 exhibit significant signals in the CLR test in numerous Eurasian populations (Russian, Sardinian, Tuscan, Oroqen and Uygur from CEPA-HGDP; Central European and Toscani from HapMap). We analyzed the allele frequencies of the SNPs in the CYP3A cluster, where CYP3A4 and CYP3A5 are located in close proximity (Figure 5). It can be seen from Figure 5a that the frequencies of the derived alleles are either near 0% or near 100% in the Eurasian populations. This extreme lack of variation is a sign of a strong sweep (44). Allele frequencies are also found highly polarized in the other Eurasian populations with significant lnRsb signatures (data not shown). By contrast, the African populations exhibit highly variable allele frequency distribution at this locus, and few derived alleles approach 100% fixation (Figure 5b).

Discussion

Genetic polymorphisms account for a substantial proportion of inter-individual and inter-ethnic heterogeneity for drug responses (45). Nonetheless, the development and clinical tests of the majority of drugs currently in use were conducted in cohorts of limited ethnic diversity, with a strong bias towards European ancestry. A comprehensive understanding of population differentiation in the genetic determinants of drug response is vital for guiding the extrapolation of drug usage to diverse ethnicities. We systematically addressed this issue by examining the global and local

population differentiation profiles in 283 ADME genes across 62 worldwide ethnic groups. We further conducted a full scan for evidence of natural selection on ADME genes. Our results provide insights into the evolution of genetic diversity patterns in the ADME genes and provide new candidates for important functional variants.

We found that ADME genes have very different population differentiation patterns both globally and regionally. Some genes have high global population differentiation levels (e.g. CYP3A5), while others are more conserved (e.g. SLC04C1 and NAT1). For many genes the high differentiation across populations mainly reflects differentiation among continental regions. However, for some genes global diversity mostly reflects differentiation within a single continental region, such as Africa, Europe or East Asia (Figure 2). It is worth emphasizing that the population differentiation of functional variants strongly correlates with the average differentiation (GA- F_{st}) based on all SNPs within the gene (Figure 4). This suggests that the allele frequency differences of functional variants can be estimated by examining their surrounding common polymorphisms. High GA- F_{st} values for a gene therefore may imply the existence of functional variants contributing to inter-ethnic heterogeneity in drug responses.

Various lines of evidence suggest that natural selection has influenced many ADME genes. The GA- F_{st} has a much wider distribution in the ADME genes than in the random neutral regions (Figure 1), suggesting enrichment for signals of both positive selection and negative/balancing selection. PA- F_{st} maps also revealed clear departures from neutrality in many genes. Moreover, two formal tests of positive

selection, the lnRsb and CLR tests, revealed widespread signals of recent positive selection. These signals tend to be continental-specific for a given gene (Figure 5, Figure S6). A potential explanation for the enrichment of signals of selection is that from an evolutionary viewpoint, the ADME genes are mainly involved in defense against xenobiotics. The varying environments and diets encountered during the recent migrations of humans out-of-Africa might have exerted strong selection pressures on various ADME genes. Local adaption consequently resulted in high differentiation of the ADME genes across different ethnic groups.

A potential caveat to these results is that many of the groups examined in this study have small sample sizes; 10 of the 62 populations have less than 10 individuals. While larger sample sizes would improve the overall power of our analyses, small sample sizes alone do not account for our results. The global GA- F_{st} value is calculated as the weighted average of the population specific F_{st} values, therefore populations with small sample size contribute less to the GA- F_{st} values. Indeed, after removing all the populations with sample sizes less than 10, the main GA- F_{st} pattern remains unchanged (see Supplementary Table. S4). The PA-F_{st} values could, in principle, be more affected by small population sizes. However, the pairwise F_{st} patterns were examined in the context of continental regions, such that individual PA-F_{st} values are unlikely to change the general patterns. Finally, for the tests of positive selection, small sample size does not dramatically reduce the power of the tests used here (21).

Many ADME genes examined in this study play pivotal roles in response to daily

drug therapies, and some of them also exhibit high population differentiation and/or strong evidence of selection. This could contribute to variation in therapeutic outcomes and to elevated risks of adverse reactions across populations. To examine this more closely, we summarized various factors concerning drug response for each core ADME gene, including its substrates, GA- F_{st} value, PA- F_{st} summary pattern, and signals of positive selection. Table 3 gives this summary for 6 genes of interest: CYP1A2, CYP2C19, CYP2C9, CYP3A4, NAT1, and ABCB1. The data for the remaining genes are listed in supplementary Table S3.

CYP3A4 and CYP3A5 are located near each other in chromosomal region 7q21.1. CYP3A4 metabolizes an estimated 50% of the currently used drugs (46, 47). Both CYP3A genes show strong population differentiation globally (GA- F_{st} 0.223 and 0.310 respectively; CYP3A5 has the highest GA- F_{st} of all ADME genes). In agreement with a previous study(48), we find strong signals of selection on CYP3A4 and CYP3A5 in many Eurasian populations. For example, in CYP3A4, there are 27 populations with significant lnRsb signals, and 7 populations with significant CLR signals. The allele frequency spectrum for CYP3A5 shows a clear selection sweep signal in multiple Eurasian populations. Previously, an intron SNP (rs776746) was found to completely interrupt the CYP3A5 expression by alternative splicing, which might explain the strong signal of selection in this region (49). Other candidate causal SNPs in CYP3A4 have been proposed but no clear conclusions were obtained (48). Details of the PA- F_{st} contour maps show that the genetic diversity of CYP3A genes mainly comes from high differentiation between Africans and non-Africans (Figure 2).

Given the wide substrate spectrum of CYP3A4, Africans may have some general differences in drug responses from non-Africans. Therefore, special care should be taken when CYP3A substrates are given to individuals of African ancestry.

CYP2C9 is another important drug metabolizer. It carries out the metabolism of anti-inflammatory, antidiabetic drugs, such as Celecoxib, lornoxicam, diclofenac, ibuprofen, etc. Warfarin, one of the major substrates used as an anticoagulant, is also metabolized by CYP2C9, and at a significantly higher rate in Europeans than in African-Americans or Asians (50). The GA- F_{st} value for CYP2C9 is low (0.067) and there is no clear indication of population differentiation either globally or specifically between Europeans and African/Asians (see Supplementary Figure S1). However, several African and East Asian populations (Cambodian, Miaozu, Mongola, Oroqen, Pygmies, Tujia, Yizu, Yoruba, YRI) exhibit evidence of selection on this gene (Figure 5), which suggests that positive selection may have played some role in the differentiation among continental regions. There is more population differentiation between the Americas and the other continental regions, which suggests that populations from the Americas may also differ in warfarin metabolism rate.

CYP1A2 mainly metabolizes antidepressants such as imipramine, clomipramine, etc. It has a high GA- F_{st} value (0.277), which is mainly accounted for by the highly distinctive genetic pattern in Europeans (Figure 2c). The dosage safety and efficacy should therefore be carefully examined for these drugs outside Europe.

ABCB1 is the most studied drug transporter and plays a pivotal role in drug

resistance (51). The global population differentiation of ABCB1 is not unusual (GA-F_{st} = 0.115); the major differences occur between Africans and non-Africans, and many non-African populations show signals of recent positive selection (Figure 5). Previous studies already revealed substantial population differentiation and large allele frequency differences in candidate functional SNPs between Africans and non-Africans(52, 53). The very broad substrate spectrum of ABCB1 suggests that general drug resistance phenotypes may exhibit non-negligible differences between individuals of African vs. non-African ancestry.

NAT2 is one of the N-acetyltransferase genes in humans involved in detoxification of a large number of chemicals. Many genetic polymorphisms influence NAT2 activity, resulting in the phenotype of either slow or rapid metabolizer. The GA-F_{st} of NAT2 is not extreme (0.128), although East Asians do tend to differ from the other populations in pairwise comparisons (Figure 2c). However, signals of selection come from various other continental regions but not East Asia (Figure 5). It is therefore of interest to study the detailed distribution of the genetic variants of NAT2 in Asian populations and test how they influence general detoxification in East Asians.

In conclusion, we have identified a significant tendency for ADME genes to exhibit more variation than other regions of the genome in the amount of differentiation among human populations. There is a strong correlation between the population differentiation of functionally significant SNPs and the population differentiation of other SNPs in such genes, suggesting that GA- F_{st} values for genes

can help identify functional SNPs. We have also identified numerous signals of recent positive selection in ADME genes. Further investigation of these signals should help in avoiding adverse drug reactions in particular populations, and illuminate important selective events in the history of human populations.

To fully appreciate how ethnicity would influence ADME gene diversity and function, one promising strategy is to carry out deep sequencing in all ADME genes in various ethnic groups. This becomes feasible with the development of next-generation sequencing technologies (54, 55). In the future, cohort studies of comprehensive drug-response phenotypes supported with full genetic information should enable a systematic understanding of the ethnicity / drug-response interplay.

Materials and methods

Genetic data

The genome-wide SNP data analyzed here are from two primary data sources, namely, the Human Genome Diversity Panel (CEPH-HGDP)(19), and the Phase III Hapmap (http://hapmap.ncbi.nlm.nih.gov/). The CEPH-HGDP panel includes 940 individuals from 51 populations, while the Phase III Hapmap data are from 1,011 individuals from 11 populations. These 62 globally-distributed populations were further grouped into seven continental regions: Africa, Middle East, Europe, South/Central Asia, East Asia, Oceania, and America. The population ID, location, and sample size are listed in Supplementary Table S1.

Phasing and merging Datasets

The HGDP-CEPH individuals were previously genotyped with Illumina HumanHap 650K Beadchips, resulting in genotypes for approximately 650,000 SNPs(56). After downloading the data we phased the data with fastPHASE (57), using the parameters -Ku100 -Kl5 -Ki10 to determine the number of haplotype clusters (all other parameter values used the default settings). The phased data were then merged in PLINK (58) with haplotype data from Phase III Hapmap, and non-overlapping SNPs were removed. The merged dataset thus consists of haplotypes for 580,612 SNPs in 1952 individuals from 62 populations.

Inferring Ancestral States

In some analyses, such as identifying recent selective sweeps, it is desirable to specify the ancestral and derived alleles of each SNP. To do so, SNP annotations were created using the TAMAL database, based chiefly on data from UCSC genome browser files (59). Using the physical position of the SNP in the human genome, the aligned chimpanzee and macaque allele information were obtained, and the ancestral allele of the human SNP was defined if one of the human alleles was identical to both the chimpanzee and the macaque allele at the corresponding position. In the merged dataset, there are 461,461 SNPs for which the ancestral allele could be identified by this procedure.

Obtaining gene information for ADME genes

The ADME gene lists were obtained from the PharmaADME database (4, 38). The PharmaADME consortium consists of individuals from academia and the pharmaceutical and genomic technology industries (www.PharmaADME.org). Our analyses were based on two sets of ADME genes: the core set, containing the most important genes in drug metabolism; and the extended set, containing other drug metabolism-related genes (4, 38). ADME genes located on sex chromosomes were excluded, resulting in 31 core ADME genes and 252 extended ADME genes (Supplementary Table 2). Gene coordinate information was obtained from the Refseq database to infer the start and end position for each gene (60).

Two additional groups of genes/regions were defined to compare to the ADME genes. The first consists of 500 randomly selected non-genic regions, selected to be at least 200kb away from any RefSeq genes (http://www.ncbi.nlm.nih.gov/RefSeq/) and matching the physical size distribution of the ADME genes. For the set of random genes, 500 random genes were directly sampled from the RefSeq database without replication.

Population differentiation and F_{st} estimation

Three F_{st} statistics, namely F_{st} averaged across all sites within a gene over all populations (GA- F_{st}), pairwise F_{st} averaged across all sites within a gene between each pair of populations (PA- F_{st}), and F_{st} calculated per site (GS- F_{st}), were all obtained with Arlequin v.3.11(61). The F_{st} for a single site is equivalent to the

unbiased F_{st} described by Weir and Cockerham, and the F_{st} for a haplotype is simply a weighted average F-statistic over the corresponding loci (62). A non-parametric test (e.g. Mann Whitney U test) was used to evaluate the significance of differences between F_{st} distributions among ADME gene sets, randomly selected genes, and nogenic regions., Similarly, an F-test was used to evaluate the significance of differences between the variance of the various F_{st} distributions. All significance tests were performed using the R package (http://www.r-project.org/).

Based on the HapMap LD database, we concatenated two neighboring genes into a single locus if any pair of SNPs, one from either gene, had an LD value (r^2) higher than 0.5 in any of the HapMap LD populations. Thus, in GA- F_{st} distributions, there are 28 ADME core genes (or groups), and 221 ADME extended genes (or groups), see Fig. 1.

The PA-F_{st} values were also calculated with the Arlequin package. In particular, a 62 *62 PA- F_{st} matrix was calculated among the 62 populations. Each PA- F_{st} value was assigned into one of the following bins: <0.01, 0.01~0.05, …, 0.25~0.30, >0.30. Each bin was then represented by a gray scale gradient in the contour maps, where the darker gradients represent higher PA- F_{st} values.

To evaluate whether the genetic variance among the 7 continental regions is significantly different from the genetic variance among populations within each region, analysis of molecular variance (AMOVA) was carried out using Arlequin, and an *F*-test was used to assess the statistical significance of the variance components.

Selection of candidate functional SNPs

We first identified all the non-synonymous SNPs from the HapMap phase II data (http://hapmap.ncbi.nlm.nih.gov/), which are located within the ADME genes and have minor allele frequencies higher than 0.05 in each of the three groups – Yoruba (YRI), Central Europeans (CEU) and Chinese-Japanese combined (CHB, JBT). Based on these criteria, we identified 18 non-synonymous SNPs. We further included one intronic SNP (SNP19 in Table 2, rs776746 in CYP3A5, which disrupts transcript splicing (63), and one 5' UTR SNP (SNP20 in Table 2, rs7662029 in the proximal promoter region of UGT2B7, which modulates the transcription of UGT2B7 (64). The pharmaADME consortium also proposed a list of core candidate functional SNPs (http://www.pharmaadme.org/), within which 22 SNPs have minor allele frequencies higher than 0.05 in all three HapMap II groups (YRI, CEU and CHB-JBT). Sixteen of these overlap with the 20 functional SNPs identified above (SNP1–14, SNP19 and SNP20 in Table 2); the other six SNPs were also added to the list of candidate functional SNPs (Table 2). For any candidate SNPs that are not represented in our merged genotyped data, we identified corresponding tagging SNPs with high r^2 value (at least 0.8) with the SNP of interest.

Detecting signals of selection

Two approaches were used to detect signals of recent positive selection. The first is based on the idea that unusually long haplotype homozygosity, associated with a high allele frequency, is unexpected under neutrality but is expected with strong

positive selection (42). Here, we used a modified InRsb approach, which is based on comparing the EHH of the same allele in different populations, to identify candidate regions that have experienced recent local selection (42). We used the same procedure described previously to identify candidate regions(21), except that we used a cutoff of 1% to identify the top SNPs, and candidate regions were defined as having at least 5 top SNPs within a 100kb region.

The second approach used to identify signals of recent positive selection is based on the maximum composite likelihood ratio (CLR) statistic, which uses the spatial distribution of allele frequencies along the genome and compares the hypothesis of a complete selective sweep against the null hypothesis of no sweep. We used the Sweepfinder program (65) to carry out these calculations. Given that the alternative hypothesis is a complete sweep, this test has the most power to detect complete or near-complete sweeps, but may lack power to detect partial sweeps. We used a strict cutoff for the empirical p-value of 0.01, to consider a signal as a significant indication of positive selection.

Conflict of interest

The authors declare no conflict of interest.

References:

1 Grossman, I. (2007) Routine pharmacogenetic testing in clinical practice: dream or reality? *Pharmacogenomics*, **8**, 1449-1459.

2 Lazarou, J., Pomeranz, B.H. and Corey, P.N. (1998) Incidence of adverse drug reactions in hospitalized patients - A meta-analysis of prospective studies. *Jama-J Am Med Assoc*, **279**, 1200-1205.

3 Fuhr, U., Jetter, A. and Kirchheiner, J. (2007) Appropriate phenotyping procedures for drug metabolizing enzymes and transporters in humans and their simultaneous use in the "cocktail" approach. *Clinical Pharmacology & Therapeutics*, **81**, 270-283.

4 Daly, T.M., Dumaual, C.M., Miao, X., Farmen, M.W., Njau, R.K., Fu, D.J., Bauer, N.L., Close, S., Watanabe, N., Bruckner, C. *et al.* (2007) Multiplex assay for comprehensive genotyping of genes involved in drug metabolism, excretion, and transport. *Clin Chem*, **53**, 1222-1230.

5 Tetko, I.V., Bruneau, P., Mewes, H.W., Rohrer, D.C. and Poda, G.I. (2006) Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov Today*, **11**, 700-707.

6 Balani, S.K., Miwa, G.T., Gan, L.S., Wu, J.T. and Lee, F.W. (2005) Strategy of utilizing in vitro and in vivo ADME tools for lead optimization and drug candidate selection. *Curr Top Med Chem*, **5**, 1033-1038.

7 Benet, L.Z. (2009) The Drug Transporter-Metabolism Alliance: Uncovering and Defining the Interplay. *Mol Pharmaceut*, **6**, 1631-1643.

8 Ma, M.K., Woo, M.H. and McLeod, H.L. (2002) Genetic basis of drug metabolism. *Am J Health Syst Pharm*, **59**, 2061-2069.

9 Karazniewicz-Lada, M., Luczak, M. and Glowka, F. (2009) Pharmacokinetic studies of enantiomers of ibuprofen and its chiral metabolites in humans with different variants of genes coding CYP2C8 and CYP2C9 isoenzymes. *Xenobiotica*, **39**, 476-485.

10 Stevens, J.C., Marsh, S.A., Zaya, M.J., Regina, K.J., Divakaran, K., Le, M. and Hines, R.N. (2008) Developmental changes in human liver CYP2D6 expression. *Drug Metab Dispos*, **36**, 1587-1593.

11 Nebert, D.W. (1997) Polymorphisms in drug-metabolizing enzymes: what is their clinical relevance and why do they exist? *Am J Hum Genet*, **60**, 265-271.

12 Coutts, R.T. (1994) Polymorphism in the metabolism of drugs, including antidepressant drugs: comments on phenotyping. *J Psychiatry Neurosci*, **19**, 30-44.

13 Mega, J.L., Close, S.L., Wiviott, S.D., Shen, L., Hockett, R.D., Brandt, J.T., Walker, J.R., Antman, E.M., Macias, W., Braunwald, E. *et al.* (2009) Cytochrome p-450 polymorphisms and response to clopidogrel. *N Engl J Med*, **360**, 354-362.

14 Ingelman-Sundberg, M. (2002) Polymorphism of cytochrome P450 and xenobiotic toxicity. *Toxicology*, **181-182**, 447-452.

15 Zhou, S.F., Liu, J.P. and Chowbay, B. (2009) Polymorphism of human cytochrome P450 enzymes and its clinical impact. *Drug Metab Rev*, **41**, 89-295.

16 Ada, A.O., Suzen, S.H. and Iscan, M. (2004) Polymorphisms of cytochrome P450 1A1, glutathione S-transferases M1 and T1 in a Turkish population. *Toxicol*

Lett, **151**, 311-315.

17 Magalon, H., Patin, E., Austerlitz, F., Hegay, T., Aldashev, A., Quintana-Murci, L. and Heyer, E. (2008) Population genetic diversity of the NAT2 gene supports a role of acetylation in human adaptation to farming in Central Asia. *Eur J Hum Genet*, **16**, 243-251.

18 Gardiner, S.J. and Begg, E.J. (2006) Pharmacogenetics, drug-metabolizing enzymes, and clinical practice. *Pharmacol Rev*, **58**, 521-590.

19 Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L. *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100-1104.

20 Hunley, K.L., Healy, M.E. and Long, J.C. (2009) The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: implications for biological race. *Am J Phys Anthropol*, **139**, 35-46.

21 Lopez Herraez, D., Bauchet, M., Tang, K., Theunert, C., Pugach, I., Li, J., Nandineni, M.R., Gross, A., Scholz, M. and Stoneking, M. (2009) Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One*, **4**, e7888.

22 Cadamuro, J., Dieplinger, B., Felder, T., Kedenko, I., Mueller, T., Haltmayer, M., Patsch, W. and Oberkofler, H. (2009) Genetic determinants of acenocoumarol and phenprocoumon maintenance dose requirements. *Eur J Clin Pharmacol*, **66**, 236-260.

23 Lee, S.J., Kim, W.Y., Kim, H., Shon, J.H., Lee, S.S. and Shin, J.G. (2009) Identification of new CYP2C19 variants exhibiting decreased enzyme activity in the metabolism of S-mephenytoin and omeprazole. *Drug Metab Dispos*, **37**, 2262-2269.

24 Marsh, S. and Van Booven, D.J. (2009) The increasing complexity of mercaptopurine pharmacogenomics. *Clin Pharmacol Ther*, **85**, 139-141.

25 Wilson, J.F., Weale, M.E., Smith, A.C., Gratrix, F., Fletcher, B., Thomas, M.G., Bradman, N. and Goldstein, D.B. (2001) Population genetic structure of variable drug response. *Nat Genet*, **29**, 265-269.

26 Sistonen, J., Sajantila, A., Lao, O., Corander, J., Barbujani, G. and Fuselli, S. (2007) CYP2D6 worldwide genetic variation shows high frequency of altered activity variants and no continental structure. *Pharmacogenet Genomics*, **17**, 93-101.

27 Sistonen, J., Fuselli, S., Palo, J.U., Chauhan, N., Padh, H. and Sajantila, A. (2009) Pharmacogenetic variation at CYP2C9, CYP2C19, and CYP2D6 at global and microgeographic scales. *Pharmacogenet Genomics*, **19**, 170-179.

28 Pasanen, M.K., Neuvonen, P.J. and Niemi, M. (2008) Global analysis of genetic variation in SLCO1B1. *Pharmacogenomics*, **9**, 19-33.

29 Man, M., Farmen, M., Dumaual, C., Teng, C.H., Moser, B., Irie, S., Noh, G.J., Njau, R., Close, S., Wise, S. *et al.* (2010) Genetic Variation in Metabolizing Enzyme and Transporter Genes: Comprehensive Assessment in 3 Major East Asian Subpopulations With Comparison to Caucasians and Africans. *J Clin Pharmacol*.

30 Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. and Clark, A.G. (2007) Recent and ongoing selection in the human genome. *Nat Rev Genet*, **8**, 857-868.

31 Hesselson, S.E., Matsson, P., Shima, J.E., Fukushima, H., Yee, S.W., Kobayashi, Y., Gow, J.M., Ha, C., Ma, B., Poon, A. *et al.* (2009) Genetic variation in the proximal promoter of ABC and SLC superfamilies: liver and kidney specific expression and promoter activity predict variation. *PLoS One*, **4**, e6942.

32 Thompson, E.E., Kuttab-Boulos, H., Witonsky, D., Yang, L., Roe, B.A. and Di Rienzo, A. (2004) CYP3A variation and the evolution of salt-sensitivity variants. *Am J Hum Genet*, **75**, 1059-1069.

33 Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol*, **4**, e72.

34 Patin, E., Harmant, C., Kidd, K.K., Kidd, J., Froment, A., Mehdi, S.Q., Sica, L., Heyer, E. and Quintana-Murci, L. (2006) Sub-Saharan African coding sequence variation and haplotype diversity at the NAT2 gene. *Hum Mutat*, **27**, 720.

35 Luca, F., Bubba, G., Basile, M., Brdicka, R., Michalodimitrakis, E., Rickards, O., Vershubsky, G., Quintana-Murci, L., Kozlov, A.I. and Novelletto, A. (2008) Multiple advantageous amino acid variants in the NAT2 gene in human populations. *PLoS One*, **3**, e3136.

36 Sun, C., Southard, C., Witonsky, D.B., Kittler, R. and Rienzo, A.D. (2010) Allele-Specific Down-Regulation of RPTOR Expression Induced by Retinoids Contributes to Climate Adaptations. *PLoS Genet*, **6**, e1001178.

37 Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu,

F., Bonnen, P.E., de Bakker, P.I., Deloukas, P., Gabriel, S.B. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52-58.

38 Dumaual, C., Miao, X., Daly, T.M., Bruckner, C., Njau, R., Fu, D.J., Close-Kirkwood, S., Bauer, N., Watanabe, N., Hardenbol, P. *et al.* (2007) Comprehensive assessment of metabolic enzyme and transporter genes using the Affymetrix (R) Targeted Genotyping System. *Pharmacogenomics*, **8**, 293-305.

39 Coulombe-Huntington, J., Lam, K.C., Dias, C. and Majewski, J. (2009) Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet*, **5**, e1000766.

40 Barendse, W., Harrison, B.E., Bunch, R.J., Thomas, M.B. and Turner, L.B. (2009) Genome wide signatures of positive selection: the comparison of independent samples and the identification of regions associated to traits. *BMC Genomics*, **10**, 178.

41 Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M., Cavalli-Sforza, L.L., Feldman, M.W. and Pritchard, J.K. (2009) The role of geography in human adaptation. *PLoS Genet*, **5**, e1000500.

42 Tang, K., Thornton, K.R. and Stoneking, M. (2007) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol*, **5**, e171.

43 Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W. *et al.* (2009) Signals of

recent positive selection in a worldwide sample of human populations. *Genome Res*, **19**, 826-837.

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

45 van der Weide, J. and Steijns, L.S. (1999) Cytochrome P450 enzyme system: genetic polymorphisms and impact on clinical pharmacology. *Ann Clin Biochem*, **36 (Pt 6)**, 722-729.

46 Lai, J., Vesprini, D., Chu, W., Jernstrom, H. and Narod, S.A. (2001) CYP gene polymorphisms and early menarche. *Mol Genet Metab*, **74**, 449-457.

47 Thier, R., Bruning, T., Roos, P.H., Rihs, H.P., Golka, K., Ko, Y. and Bolt, H.M. (2003) Markers of genetic susceptibility in human environmental hygiene and toxicology: the role of selected CYP, NAT and GST genes. *Int J Hyg Environ Health*, **206**, 149-171.

48 Qiu, H., Taudien, S., Herlyn, H., Schmitz, J., Zhou, Y., Chen, G., Roberto, R., Rocchi, M., Platzer, M. and Wojnowski, L. (2008) CYP3 phylogenomics: evidence for positive selection of CYP3A4 and CYP3A7. *Pharmacogenet Genomics*, **18**, 53-66.

49 Chen, X., Wang, H., Zhou, G., Zhang, X., Dong, X., Zhi, L., Jin, L. and He, F. (2009) Molecular population genetics of human CYP3A locus: signatures of positive selection and implications for evolutionary environmental medicine. *Environ Health Perspect*, **117**, 1541-1548.

50 Momary, K.M., Shapiro, N.L., Viana, M.A., Nutescu, E.A., Helgason, C.M. and Cavallari, L.H. (2007) Factors influencing warfarin dose requirements in African-

Americans. *Pharmacogenomics*, **8**, 1535-1544.

51 Leschziner, G.D., Andrew, T., Pirmohamed, M. and Johnson, M.R. (2007) ABCB1 genotype and PGP expression, function and therapeutic drug response: a critical review and recommendations for future research. *Pharmacogenomics J*, **7**, 154-179.

52 Tang, K., Ngoi, S.M., Gwee, P.C., Chua, J.M., Lee, E.J., Chong, S.S. and Lee, C.G. (2002) Distinct haplotype profiles and strong linkage disequilibrium at the MDR1 multidrug transporter gene locus in three ethnic Asian populations. *Pharmacogenetics*, **12**, 437-450.

53 Tang, K., Wong, L.P., Lee, E.J., Chong, S.S. and Lee, C.G. (2004) Genomic evidence for recent positive selection at the human MDR1 gene locus. *Hum Mol Genet*, **13**, 783-797.

54 Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J. and Turner, D.J. (2010) Target-enrichment strategies for next-generation sequencing. *Nat Methods*, **7**, 111-118.

55 Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat Rev Genet*, **11**, 31-46.

56 Rosenberg, N.A. (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet*, **70**, 841-847.

57 Scheet, P. and Stephens, M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes

and haplotypic phase. *Am J Hum Genet*, **78**, 629-644.

58 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **81**, 559-575.

59 Hemminger, B.M., Saelim, B. and Sullivan, P.F. (2006) TAMAL: an integrated approach to choosing SNPs for genetic studies of human complex traits. *Bioinformatics*, **22**, 626-627.

60 Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**, D61-65.

61 Excoffier, L., Laval, G. and Schneider, S. (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol Bioinform Online*, **1**, 47-50.

62 Weir, B.S. and Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358-1370.

63 Coto, E., Tavira, B., Marin, R., Ortega, F., Lopez-Larrea, C., Ruiz-Ortega, M., Ortiz, A., Diaz, M., Corao, A.I., Alonso, B. *et al.* (2010) Functional polymorphisms in the CYP3A4, CYP3A5, and CYP21A2 genes in the risk for hypertension in pregnancy. *Biochem Biophys Res Commun*, **397**, 576-579.

64 Tojcic, J., Benoit-Biancamano, M.O., Court, M.H., Straka, R.J., Caron, P. and Guillemette, C. (2009) In vitro glucuronidation of fenofibric acid by human UDP-

glucuronosyltransferases and liver microsomes. *Drug Metab Dispos*, **37**, 2236-2243.

65 Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G. and Bustamante, C. (2005) Genomic scans for selective sweeps using SNP data. *Genome Res*, **15**, 1566-1575.

66 Kim, W.J., Lee, J.H., Yi, J., Cho, Y.J., Heo, K., Lee, S.H., Kim, S.W., Kim, M.K., Kim, K.H., In Lee, B. *et al.* (2010) A nonsynonymous variation in MRP2/ABCC2 is associated with neurological adverse drug reactions of carbamazepine in patients with epilepsy. *Pharmacogenet Genomics*, **20**, 249-256.

67 Tomlinson, B., Hu, M., Lee, V.W., Lui, S.S., Chu, T.T., Poon, E.W., Ko, G.T., Baum, L., Tam, L.S. and Li, E.K. (2010) ABCG2 polymorphism is associated with the low-density lipoprotein cholesterol response to rosuvastatin. *Clin Pharmacol Ther*, **87**, 558-562.

68 Agundez, J.A., Garcia-Martin, E. and Martinez, C. (2009) Genetically based impairment in CYP2C8- and CYP2C9-dependent NSAID metabolism as a risk factor for gastrointestinal bleeding: is a combination of pharmacogenomics and metabolomics required to improve personalized medicine? *Expert Opin Drug Metab Toxicol*, **5**, 607-620.

69 Ross, K.A., Bigham, A.W., Edwards, M., Gozdzik, A., Suarez-Kurtz, G. and Parra, E.J. (2010) Worldwide allele frequency distribution of four polymorphisms associated with warfarin dose requirements. *J Hum Genet*, **55**, 582-589.

70 Kelemen, L.E., Goodman, M.T., McGuire, V., Rossing, M.A., Webb, P.M., Kobel,

M., Anton-Culver, H., Beesley, J., Berchuck, A., Brar, S. *et al.* (2010) Genetic variation in TYMS in the one-carbon transfer pathway is associated with ovarian carcinoma types in the Ovarian Cancer Association Consortium. *Cancer Epidemiol Biomarkers Prev*, **19**, 1822-1830.

71 Kim, H.N., Kim, N.Y., Yu, L., Kim, Y.K., Lee, I.K., Yang, D.H., Lee, J.J., Shin, M.H., Park, K.S., Choi, J.S. *et al.* (2009) Polymorphisms of drug-metabolizing genes and risk of non-Hodgkin lymphoma. *Am J Hematol*, **84**, 821-825.

72 Szental, J.A., Baird, P.N., Richardson, A.J., Islam, F.M., Scholl, H.P., Charbel Issa, P., Holz, F.G., Gillies, M. and Guymer, R.H. (2010) Analysis of glutathione S-transferase Pi isoform (GSTP1) single-nucleotide polymorphisms and macular telangiectasia type 2. *Int Ophthalmol*.

73 Agundez, J.A., Golka, K., Martinez, C., Selinski, S., Blaszkewicz, M. and Garcia-Martin, E. (2008) Unraveling ambiguous NAT2 genotyping data. *Clin Chem*, **54**, 1390-1394.

74 Pinsonneault, J., Nielsen, C.U. and Sadee, W. (2004) Genetic variants of the human H+/dipeptide transporter PEPT2: analysis of haplotype functions. *J Pharmacol Exp Ther*, **311**, 1088-1096.

75 Kim, D.H., Sriharsha, L., Xu, W., Kamel-Reid, S., Liu, X., Siminovitch, K., Messner, H.A. and Lipton, J.H. (2009) Clinical relevance of a pharmacogenetic approach using multiple candidate genes to predict response and resistance to imatinib therapy in chronic myeloid leukemia. *Clin Cancer Res*, **15**, 4750-4758.

76 Sallinen, R., Kaunisto, M.A., Forsblom, C., Thomas, M., Fagerudd, J.,

Pettersson-Fernholm, K., Groop, P.H. and Wessman, M. (2010) Association of the SLC22A1, SLC22A2, and SLC22A3 genes encoding organic cation transporters with diabetic nephropathy and hypertension. *Ann Med*, **42**, 296-304.

77 Vladutiu, G.D. and Isackson, P.J. (2009) SLCO1B1 variants and statin-induced myopathy. *N Engl J Med*, **360**, 304.

78 Kiyotani, K., Mushiroda, T., Kubo, M., Zembutsu, H., Sugiyama, Y. and Nakamura, Y. (2008) Association of genetic polymorphisms in SLCO1B3 and ABCC2 with docetaxel-induced leukopenia. *Cancer Sci*, **99**, 967-972.

79 Jeannesson, E., Siest, G., Bastien, B., Albertini, L., Aslanidis, C., Schmitz, G. and Visvikis-Siest, S. (2009) Association of ABCB1 gene polymorphisms with plasma lipid and apolipoprotein concentrations in the STANISLAS cohort. *Clin Chim Acta*, **403**, 198-202.

80 Sookoian, S., Castano, G., Gianotti, T.F., Gemma, C. and Pirola, C.J. (2009) Polymorphisms of MRP2 (ABCC2) are associated with susceptibility to nonalcoholic fatty liver disease. *J Nutr Biochem*, **20**, 765-770.

81 Kwara, A., Lartey, M., Sagoe, K.W., Kenu, E. and Court, M.H. (2009) CYP2B6, CYP2A6 and UGT2B7 genetic polymorphisms are predictors of efavirenz mid-dose concentration in HIV-infected patients. *AIDS*, **23**, 2101-2106.

82 Carr, D.F., la Porte, C.J., Pirmohamed, M., Owen, A. and Cortes, C.P. (2010) Haplotype structure of CYP2B6 and association with plasma efavirenz concentrations in a Chilean HIV cohort. *J Antimicrob Chemother*, **65**, 1889-1893.

83 Shuldiner, A.R., O'Connell, J.R., Bliden, K.P., Gandhi, A., Ryan, K., Horenstein,

R.B., Damcott, C.M., Pakyz, R., Tantry, U.S., Gibson, Q. *et al.* (2009) Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *JAMA*, **302**, 849-857.

Legends to Figures:

Figure 1. Histograms of GA- F_{st} distributions of (a) 28 ADME core genes (or groups), (b) 221 ADME extended genes (or groups), (c) 500 randomly selected non-genic regions, and (d) 500 randomly selected genes.

Figure 2. The contour maps constructed from PA- F_{st} matrices for 62 populations. In these maps, (a) to (f) illustrate the pairwise genetic differences for CYP3A4, CYP1A2, NAT2, UGT2B7, NAT1 and non-genes, respectively. Lines separate the continental regions, with the names of regions marked at the corresponding position to the right of each map. The number 1 to 62 labeled in the verge of maps means: **Africa:** 1.Bantu, 2.Biaka_Pygmies, 3.Mandenka, 4.Mbuti_Pygmies, 5.San, 6.Yoruba, 7.Mozabite, 8.asw, 9.lwk, 10.mkk, 11.yri; **Middle East:** 12.Bedouin, 13.Druze, 14.Palestinian; **Europe:** 15.Adygei, 16.Basque, 17.French, 18.Italian, 19.Orcadian, 20.Russian, 21.Sardinian, 22.Tuscan, 23.ceu, 24.tsi, **South Central Asia:** 25.Brahui, 26.Balochi, 27.Hazara, 28.Makrani, 29.Sindhi, 30.Pathan, 31.Kalash, 32.Burusho, 33.gih; **East Asia:** 34.Cambodian, 35.Dai, 36.Daur, 37.Han, 38.Hezhen, 39.Japanese, 40.Lahu, 41.Miaozu, 42.Mongola, 43.Naxi, 44.Oroqen, 45.She, 46.Tu, 47.Tujia, 48.Uygur, 49.Xibo, 50.Yakut, 51.Yizu, 52.chb, 53.chd, 54.jpt; **Oceania:** 55.Melanesian, 56.Papuan**, America:** 57.Colombian, 58.Karitiana, 59.Maya, 60.Pima, 61.Surui, 62.mex (Supplementary table S1. gives more detailed information)

Figure 3. The scatter plot of GA- F_{st} vs. GS- F_{st} for functionally important SNPs.

Figure 4. Signals of positive selection signal in the 31 core ADME genes among the 62 worldwide populations. The blue blocks indicate the significant signals found by

the lnRsb method; red indicates the significant signals found by Sweepfinder; and black indicates the overlapping signals for both methods.

Figure 5. The derived allele frequency distributions of the SNPs in CYP3A4 in (a) European populations, (b) African populations.

Table 1. The results of statistical tests of the significance of differences in the distribution of GA- F_{st} values among ADME core genes, extended genes, randomly selected genes, and non-genic regions.

	Mann-Whitney u test			*F*-test	
	Mean 1	Mean 2	P value	Ratio of variance	P value
Core genes vs. Extended genes	0.130	0.119	0.720	2.281	0.001
Core genes vs. Non-genes	0.130	0.109	0.192	4.244	5.882×10^{-11}
Core genes vs. Rand-genes	0.130	0.120	0.712	1.861	0.012
Extended genes vs. Non-genes	0.119	0.109	0.005	1.861	5.267×10^{-9}
Extended genes vs. Rand-genes	0.119	0.120	0.959	0.816	0.067

Table 2. The 26 high frequency candidate functional SNPs and their GS- F_{st} values, and the corresponding genes and their GA- F_{st} values. The GS- F_{st} values for the tagging SNPs are given when the candidate SNPs are absent in our dataset.

No.	Gene_Name	GA-F_{st}	rsSNP ID	Position	Tagging SNP	Effect	GS-F_{st}
1	ABCC2	0.064	rs2273697	chr10:101553805	-	Nonsynonymous(66)	0.044
2	ABCG2	0.141	rs2231142	chr4:89271347	-	Nonsynonymous(67)	0.160
3	CYP2C8	0.101	rs10509681	chr10:96788739	-	Nonsynonymous(68)	0.076
4	CYP2C9	0.067	rs1057910	chr10: 96731043	-	Nonsynonymous(69)	0.035
5	DPYD	0.098	rs1801265	chr1:98121473	-	Nonsynonymous(70)	0.136
6	GSTP1	0.083	rs1695	chr11:67109265	-	Nonsynonymous(71)	0.088
7	GSTP1	0.083	rs1138272	chr11:67110155	-	Nonsynonymous(72)	0.038
8	NAT2	0.128	rs1208	chr8:18302596	-	nonsynonymous(73)	0.159
9	SLC15A2	0.127	rs1143671	chr3:123129976	-	Nonsynonymous(74)	0.129
10	SLC15A2	0.127	rs2257212	chr3:123126494	-	nonsynonymous(74)	0.129
11	SLC22A1	0.079	rs628031	chr6:160480835	-	nonsynonymous(75)	0.046
12	SLC22A2	0.098	rs316019	chr6:160590272	-	nonsynonymous(76)	0.028
13	SLCO1B1	0.102	rs2306283	chr12:21221005	-	nonsynonymous(77)	0.116
14	SLCO1B3	0.172	rs4149117	chr12:20902747	-	nonsynonymous(78)	0.208
15	ABCB1	0.115	rs9282564	chr7: 87067376	-	nonsynonymous(79)	0.075
16	ABCC2	0.064	rs8187710	chr10:101601284	-	nonsynonymous(80)	0.110
17	UGT2B7	0.093	rs7439366	chr4:69998927	rs7375178	nonsynonymous(81)	0.088
18	ABCG2	0.141	rs2231137	chr4:89280138	rs4148150	nonsynonymous(75)	0.127
19	CYP3A5	0.310	rs776746	chr7:99108475	-	Intron, nonfunctional CYP3A5(63)	0.332
20	UGT2B7	0.093	rs7662029	chr4:69996501	-	proximal promoter region(64)	0.065
21	CYP2B6	0.054	rs3745274	chr19:46204681	rs11673270	nonsynonymous(82)	0.083
22	CYP2C19	0.085	rs4244285	chr10:96531606	rs12767583	synonymous(83)	0.059
23	NAT2	0.128	rs1801280	chr8:18302134	rs1208	nonsynonymous(73)	0.159
24	SLC15A2	0.127	rs1143672	chr3:123130858	rs874742	nonsynonymous(74)	0.131
25	SLC15A2	0.127	rs2293616	chr3:123124383	rs874742	synonymous(74)	0.131
26	UGT2B7	0.093	rs7668258	chr4:69996667	rs12513195	-(81)	0.092

Table 3. Summary information for several core ADME genes of interest. Similar information for the other core genes is provided in supplementary table S2.

Gene name	Example substrates	GA-F_{st}	Global diversity pattern	Populations with InRsb signals	Populations with SweepFinder signals
CYP1A2	Amitriptyline, Imipramine, Clomipramine, Clozapine, Caffeine, etc.	0.277	EU-others	LWK, MKK	NA *
CYP2C19	Amitriptyline, antiepileptics, nordazepam, diazepam, phenytoin, etc.	0.085	AM-others	NA	NA
CYP2C9	Celecoxib, lornoxicam, lornoxicam, diclofenac, ibuprofen, etc.	0.067	AM-others	Cambodian, Miaozu, Mongola, Oroqen, Pygmies, Tujia, Yizu, Yoruba, YRI	NA
CYP3A4	Diltiazem, nifedipine, felodipine, verapamil, ciclosporin, etc. (approximately 50% of all current clinical used drugs)	0.223	AF-others	Adygei, Bedouin, Brahui, Burusho, Daur, Druze, Hazara, Kalash, Makrani, Mozabite, Naxi, Oraqen, Palestinian, Pathan, Russian, Sardinian, Sindhi, Tu, Tuscan, Uygur, Yakut, CEU, GIH, LWK, MEX, MKK, TSI	Makrani, Oroqen, Russian, Sardinian, Uygur, CEU, TSI
NAT2	sulfamethazine, retigabine, hydralazine, sulfasalazine, ribavirin, etc.	0.128	EA - others	Adygei, Bedouin, Colombian, Italian, LWK, MEX, MKK	NA
ABCB1	Colchicine, tacrolimus, quinidine, etoposide, etc.	0.115	AF-others	Adygei, Balochi, Bedouin, Brahui, Burusho, Cambodian, Colombian, Druze, Hazara, Italian, Japanese, Makrani, Mozabite, Naxi, Oroqen, Papuan, Pathan, Russian, Sardinian, Sindhi, Tu, Tujia, Xibo, Yakut, ASW, CEU, GIH, YRI	Pathan, Sindhi, MEX

* : 'AF','AM' ,'EA' , 'EU' in column 4 indicate Africa, America, East Asia, Europe, respectively. 'NA' in column 5 and 6 means no signal of selection was found.

Figure 1.

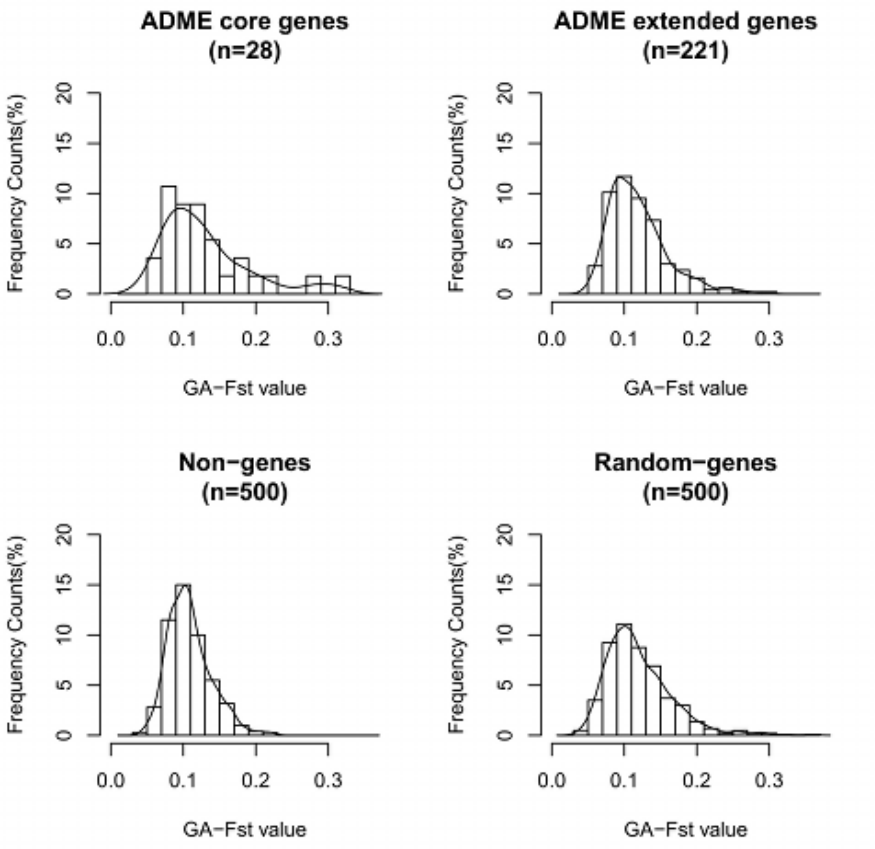

Figure 2.

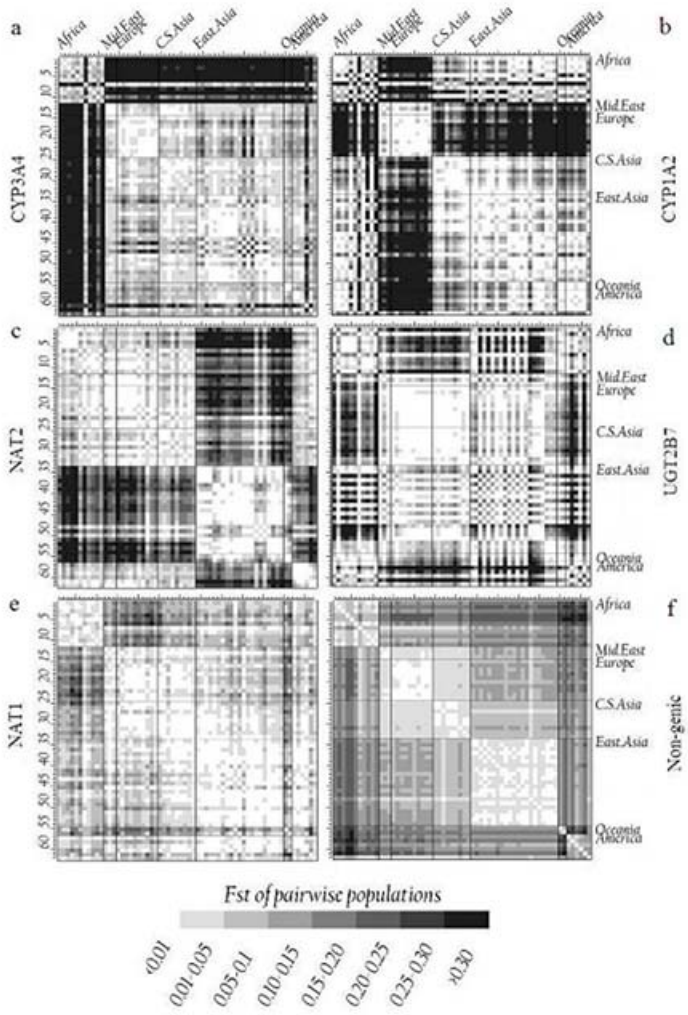

Figure 3.

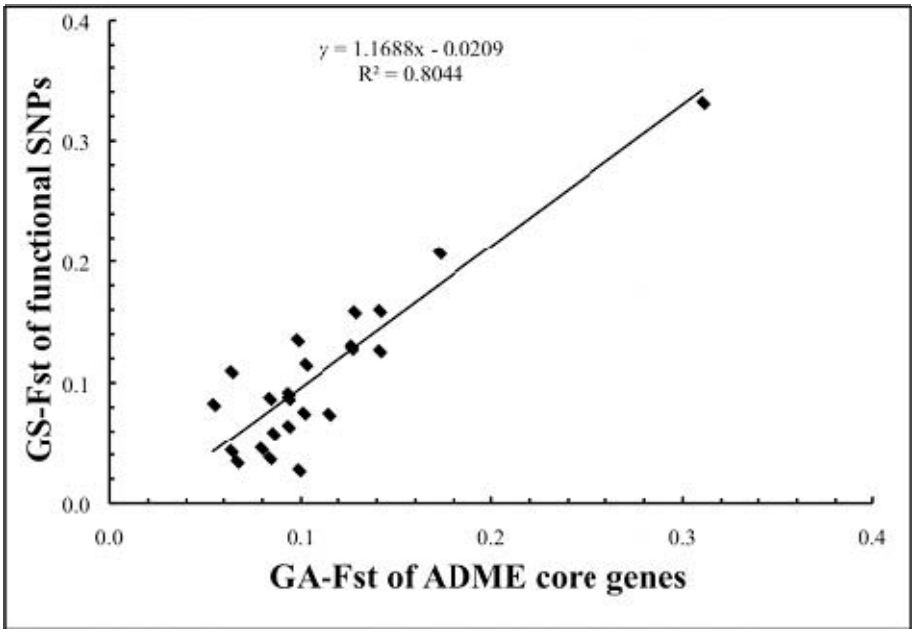

Figure 4.

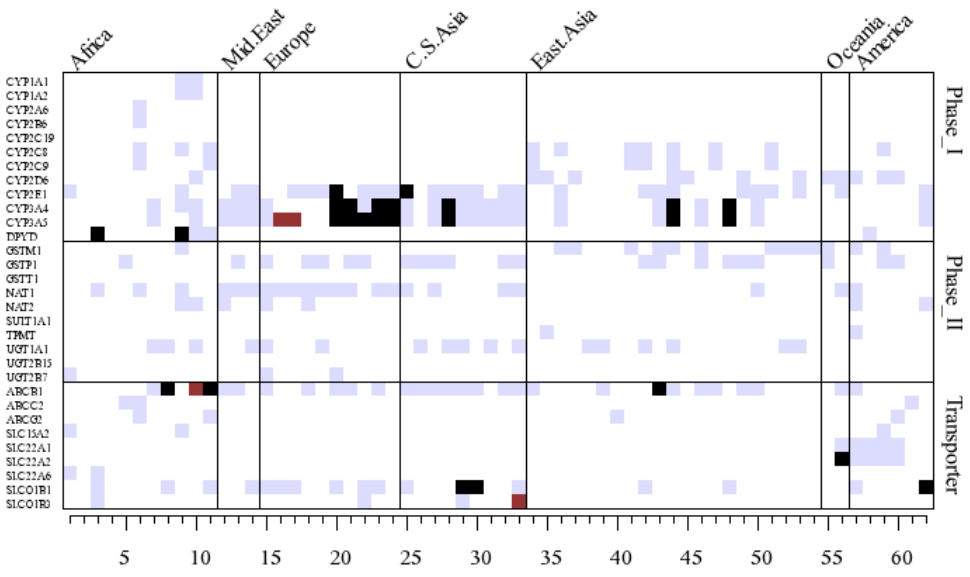

Figure 5.