Hum Genet (2013) 132:563–577
DOI 10.1007/s00439-013-1268-5

ORIGINAL INVESTIGATION

Human loci involved in drug biotransformation: worldwide genetic variation, population structure, and pharmacogenetic implications

Pierpaolo Maisano Delser · Silvia Fuselli

Received: 20 September 2012 / Accepted: 8 January 2013 / Published online: 26 January 2013

Abstract Understanding the role of inheritance in individual variation in drug response is the focus of pharmacogenetics (PGx). A key part of this understanding is quantifying the role of genetic ancestry in this phenotypic outcome. To provide insight into the relationship between ethnicity and drug response, this study first infers the global distribution of PGx variation and defines its structure. Second, the study evaluates if geographic population structure stems from all PGx loci in general, or if structure is caused by specific genes. Lastly, we identify the genetic variants contributing the greatest proportion of such structure. Our study describes the global genetic structure of PGx loci across the 52 populations of the Human Genome Diversity Cell-Line Panel, the most inclusive set of human populations freely available for studies on human genetic variation. By analysing genetic variation at 1,001 single nucleotide polymorphisms (SNPs) involved in biotransformation of exogenous substances, we describe the between-populations PGx variation, as well geographical groupings of diversity. In addition, with discriminant analysis of principal component (DAPC), we infer how many and which groups of populations are supported by PGx variation, and identify which SNPs actually contribute to the PGx structure between such groups. Our results show that intergenic, synonymous and non-synonymous SNPs show similar levels of genetic variation across the globe. Conversely, loci coding for Cytochrome P450s (mainly metabolizing exogenous substances) show significantly higher levels of genetic diversity between populations than the other gene categories. Overall, genetic variation at PGx loci correlates with geographic distances between populations, and the apportionment of genetic variation is similar to that observed for the rest of the genome. In other words, the pattern of PGx variation has been mainly shaped by the demographic history of our species, as in the case of most of our genes. The population structure defined by PGx loci supports the presence of six genetic clusters reflecting geographic location of samples. In particular, the results of the DAPC analyses show that 27 SNPs substantially contribute to the first three discriminant functions. Among these SNPs, some, such as the intronic rs1403527 of *NR1I2* and the non-synonymous rs699 of *AGT*, are known to be associated with specific drug responses. Their substantial variation between different groups of populations may have important implications for PGx practical applications.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-013-1268-5) contains supplementary material, which is available to authorized users.

P. Maisano Delser · S. Fuselli (✉)
Department of Life Sciences and Biotechnologies, University of Ferrara, Ferrara, Italy
e-mail: silvia.fuselli@unife.it

P. Maisano Delser
Department of Genetics, University of Leicester, Leicester, UK
e-mail: pm244@le.ac.uk

Introduction

Response and adverse reaction to drug treatment are influenced by a number of factors such as age, gender, environment, and by an individual's genetic make-up. Differences between individuals and populations are exacerbated by the polymorphic nature of the genes involved in drug biotransformation (Leabman et al. 2003; Ingelman-Sundberg et al. 2007). Investigating the population structure of pharmacogenetic loci (PGx) is important

to shed light on the evolutionary history of these genes, but at the same time, it may have important practical implications. For example, genome-wide association studies (GWAS) have recently been applied to pharmacogenomics (Cooper et al. 2008; Takeuchi et al. 2009; Teichert et al. 2009; Bailey and Cheng 2010) raising concerns about the possibility that population stratification may lead to false positive genetic associations with drug response (Nelson et al. 2008; Visscher et al. 2009).

Drug response genes: examples of spatial patterns and evolutionary inferences

Several genes determining drug response in humans show a clear geographic structure. The gene *UGT2B17*, coding for a conjugation enzyme with important roles in elimination of xenobiotics and endogenous compounds, varies in copy numbers from zero to two. This polymorphism shows very high worldwide diversity: two distinct clusters of haplotypes were identified with deleted chromosomes confined to one cluster and forming its majority. Both clusters were found in Africa, Europe, and East Asia, but at different frequencies, with gene deletion being the major allele in East Asia (Xue et al. 2008). Genes coding for Cytochrome P450 (CYP) drug metabolizing enzymes are highly polymorphic, and some genetic variants are known to affect the outcome in drug treatment to a very high extent (Ingelman-Sundberg et al. 2007). Several *CYP* genes show specific geographic patterns of genetic variation. *CYP2C9*, *CYP2C19* and *CYP2D6* defective alleles occur globally in all geographic regions, reaching extremely high frequencies in some populations. Interestingly, each of the *CYP* genes shows a distinct geographic pattern: *CYP2C9* decreased metabolic activity reaches high frequency in Europe, *CYP2C19* in East Asia, and *CYP2D6* increased metabolic activity is common in certain African and East Asian regions (Sistonen et al. 2009). The CYP3A drug metabolizing enzymes are involved in the metabolism and elimination of a wide range of xenobiotics, including about 50 % of all therapeutic drugs used in the clinics. The frequency of *CYP3A5*3*, the most common *CYP3A5* defective allele, is about 20 % in Africa, which is the major allele in non-African populations, almost fixed in some European regions (Thompson et al. 2004).

Examples of spatial structure and differences among groups of populations have also been reported for treatment responses associated with specific gene variants. A strong association between a genetic marker, the human leukocyte antigen *HLA–B*1502*, and a reactions of the skin called Stevens–Johnson syndrome induced by carbamazepine has been repeatedly shown in Han Chinese (Chung et al. 2004; Man et al. 2007). This association was absent in populations of European origin, in which another HLA variant seems to be a good marker for the same phenotype (McCormack et al. 2011). Treatment for chronic infection with hepatitis C virus shows a significantly higher efficacy and tolerability in patients of European ancestry than in patients of African ancestry. In both groups of patients, the presence of a particular allele near the *IL28B* gene, encoding interferon-l-3, is associated with an approximately twofold change in response to treatment. The difference in response rates is partially explained by the greater frequency of this allele in European than African populations (Ge et al. 2009). Given these observations and their practical pharmacogenetic consequences, it would be of great relevance to understand whether common ethnic labels may be useful to predict an individual's drug response in general, or if drug response phenotypes are more influenced by gene-specific patterns.

During the last 10 years, efforts have been made to answer similar questions considering whole-genome variation. A number of studies on human genetic diversity across the globe inferred the existence of population clusters supported by genomic data roughly corresponding to continental regions (Rosenberg et al. 2002; Li et al. 2008; Jakobsson et al. 2008), although no consensus has ever been reached on the number and definition of these clusters (Tishkoff and Kidd 2004; Barbujani and Colonna 2010). For this reason, there is still debate as to whether discrete geographical groups supported by a discontinuous distribution of human genetic variation would be useful, for example, for biomedical studies, or to understand differential response to pharmacological agents, with important consequent personal and societal implications (Weiss and Long 2009; Barbujani and Colonna 2010; Royal et al. 2010). Moreover, discontinuities between groups of populations explain only a small fraction of human genetic variance, whereas the great majority of it is explained by clinal patterns (i.e. geographic distance). The relationship between genetic and geographic distance is well known since the first half of last century. The seminal article by Haldane (1940) revealed gradients of allele frequencies of the ABO blood group in Europe, an observation that has been generalised in the 1970s to populations representative of human diversity (Lewontin 1972). Further analyses have shown that, in general, genetic differentiation between pairs of populations correlates with geographic distance separating them (Cavalli-Sforza et al. 1994), and recent genomic studies on globally diverse populations confirmed this trend at both continental and global scales (Ramachandran et al. 2005; Li et al. 2008). This pattern is expected given the recent African origin of modern humans, the founder effect associated with colonisation of new lands, and the short-range gene flow that occurred during expansions. In other words, the spatial pattern of the

polymorphism shown by our genes was largely shaped by the demographic history of our species with some exceptions in which adaptation by natural selection to specific environments may have contributed substantially (Coop et al. 2009; Novembre and Di Rienzo 2009).

As for the rest of the genome, human demographic history should be taken into account to interpret the spatial distribution of genetic variation at loci involved in biotransformation of exogenous substances. However, adaptation may have played a role in the evolution of genes coding for molecules that mediate the relationships between the organism and the environment. Indeed, signatures of natural selection have been observed at several PGx loci. Among Phase I drug metabolizing enzymes, the most convincing evidence of natural selection are provided by the pattern of evolution of *CYP3A* locus that includes four genes, *CYP3A4*, *CYP3A5*, *CYP3A7*, and *CYP3A43*. In particular, Thompson et al. (2004) found evidence of positive selection on the defective *CYP3A5*3* allele in non-African populations. Significant correlation of its allelic frequency with distance from the equator was interpreted as an adaptation underlying salt regulation. A complex pattern of natural selection was further identified for other genes of the *CYP3A* locus (Chen et al. 2009). Genetic variation in the noncoding sequence 5′ of the *CYP1A2* gene suggests that the regulatory sequences of this P450 enzyme may have evolved under positive selection, although alternative demographic explanation could not be definitively ruled out (Wooding et al. 2002). Different patterns of evolution have been observed in populations with different lifestyles at *CYP2D6*, the most variable P450 gene. In particular, the increased frequency of alleles associated with a slower rate of metabolism with transition to agriculture maybe due to the pronounced substrate-dependent activity of most of these enzymes that allowed expanding the spectrum of the metabolic response to diet compounds (Fuselli et al. 2010). Similarly, Neolithic transition would be responsible for the signature of selection showed in western/central Eurasians by NAT2, a Phase II metabolising enzyme. A possible explanation is that the slow-acetylator phenotype has been positively selected as conferring a reduced activation of environmental carcinogens (Patin et al. 2006). Finally, signatures of positive selection have been observed on both coding and regulatory regions of the *ABCBI* transporter gene in different human populations (Wang et al. 2007a).

Overview of the study

The first aim of this study was to answer the following question: is there a specific geographic structure of genetic variation involved in drug biotransformation, or do PGx loci behave like most of the rest of the genome? To answer this question, data on 1,001 single nucleotide polymorphisms (SNPs) representing 143 genes involved in drug biotransformation were extracted from a publicly available database including more than 650,000 SNPs in 52 populations from the Human Genome Diversity Panel (HGDP) (Li et al. 2008). In this study, we analyse the distribution of genetic diversity at these loci between populations and groups, and define its relationship with geography. Specifically, we use discriminant analysis of principal components (DAPC, Jombart et al. 2010) to define clusters of genetically close individuals supported by PGx variation, and describe the spatial distribution of between-clusters differentiation. The second goal of this study was to identify and describe the set of markers that actually contribute to the separation between groups of populations and, thus, give rise to specific geographic patterns. Knowing which genes involved in drug biotransformation show a detectable population structure may be of great use for practical purposes, such as drug dosage determination. Additionally, this information can help avoid population stratification in association studies on drug efficacy and tolerability.

Methods

Populations, samples and markers

In this study, we used the set of 1,043 HGDP individuals from 52 worldwide distributed populations (Rosenberg 2006) genotyped previously on Illumina's HumanHap650Y platform (Li et al. 2008). Among the SNPs genotyped, a total of 1,001 were extracted and analysed in this study.

Part of the genetic markers analysed in our study was selected starting from a previously published panel of SNPs (Visscher et al. 2009). The rest of the SNPs have been retrieved from the SNPper database (http://snpper.chip.org/, Riva and Kohane 2002) which combines information from both dbSNP and the UCSC Genome Browser.

In the study by Visscher et al. (2009), a customised SNP genotyping assay was designed to capture the genetic variation of 220 key drug biotransformation genes (i.e. phases I and II drug-metabolism enzymes, drug transporters, drug targets, drug receptors, transcription factors, ion channels and other disease-specific genes related to the physiological pathway of Adverse Drug Reactions). In total, the authors of the study could analyse 2,094 SNPs. Half of the panel consisted of tagSNPs of the candidate genes (tagSNPs selection based on the International HapMap project with four populations—CEU, CHB, JPT, YRI—LD statistic threshold $r^2 = 0.8$, minor allele

frequency >0.05). The other half included functional SNPs identified by literature review or from public databases as SNPs with known or probable effect on enzyme activity or function. Of these 2,094 SNPs selected by Visscher et al. (2009), we could extract 614 SNPs located within or near 123 genes in the dataset of HGDP populations genotyped on Illumina's HumanHap650Y platform.

Additionally, from the same dataset, we extracted 471 SNPs that represent the genetic variation at 45 Cytochrome P450 (*CYP*) loci involved in metabolism of exogenous or endogenous molecules. Among these 471, 84 were already present among the 2,094 SNPs selected by Visscher et al. (2009). Excluding these 84 overlapping SNPs between the two databases, a total of 1,001 pharmacogenetic markers (hereafter PGx loci) in 143 genes were analysed (Online Resource Table 1a).

Genetic variation in different SNP classes

Fst values were calculated for each SNP and their distribution was analysed by dividing markers on the basis of their location in non-genic or genic regions. Non-genic SNPs were further subdivided in upstream (including 5′ UTRs and promoters) or downstream (including 3′ UTRs) with respect to the closest gene, while genic SNPs were subdivided into intronic (including intron boundaries), synonymous, and non-synonymous. A second set of analyses was computed by classifying SNPs depending on the functional role of the gene in which they are located, or the functional role of the closest PGx gene in case of non-genic SNPs. Gene classes were defined according to PharmaADME core list and related gene list (pharmaADME.org), *CYP* genes metabolizing primarily exogenous or endogenous substrates were defined based on information from the literature (Nebert and Russell 2002; Guengerich 2003; Thomas 2007). Locus by locus Fst values was calculated using Arlequin suite v.3.5 (Excoffier and Lischer 2010). The independence between number of SNPs in the top 10 % Fst and gene class (or SNP role) was tested by means of a χ^2 test (SNP categories with less than 5 SNPs were excluded). To identify which SNP class (or SNP role) determined deviation from independency (i.e. $P < 0.05$), the same test was performed excluding one class (or role) of SNP at a time. In the case of SNP classes, a second analysis was performed excluding rare SNPs (minor allele frequency, MAF ≤ 0.05).

Spatial patterns of genetic variation

The role of geographic distances in shaping genetic diversity at PGx loci was tested by means of a Mantel test of matrices correlation. Genetic distances between pairs of populations were calculated as Fst values using the software Arlequin suite v.3.5 and geographic (great-circle) distances were calculated between all population pairs considering the likely routes of human migration out of Africa, following the criteria set by Ramachandran et al. (2005). The test was performed with the software Passage version 2.0 (Rosenberg and Anderson 2001) and significance was assessed with 10,000 permutations.

The amount of genetic diversity within and between the seven major geographic regions of the world (i.e. Africa, Middle East, Europe, Central/South Asia, East Asia, Oceania, and America) was evaluated by means of a hierarchical analysis of molecular variance (AMOVA) using Arlequin suite v.3.5. AMOVA allowed us to quantify genetic diversity at three levels, namely between members of the same population, between populations of the same group and between groups.

Discriminant analysis of principal components (DAPC)

DAPC is a multivariate method that may be used to identify and describe clusters of genetically related individuals (Jombart et al. 2010). First, data are converted into uncorrelated variables, which account for most of the genetic variation, using a principal component analysis (PCA). These uncorrelated components are then assessed with a discriminant analysis (DA) that aims to maximise the variation between groups relative to the diversity within group, finding linear combinations of alleles (the discriminant functions, DFs), which best separate the clusters.

Two analyses were run for the dataset including 1,001 PGx loci analysed in this study. First, DAPC was used to investigate the genetic structure of the 52 sampled populations defining a priori 52 groups. 300 principal components of PCA were retained during the preliminary variable transformation, which accounted for approximately 90 % of the total genetic variability.

The aim of the second analysis was to identify how many and which genetic clusters were supported by PGx genetic markers. The cluster algorithm *k*-means was used to find a given number of groups (k) maximising the variation between them. *k*-means was run sequentially with increasing values of k (from 1 to 100), then different cluster solutions were compared using Bayesian information criterion (BIC). Finally, the curve of BIC values as a function of k was inspected to identify the best supported number of clusters (i.e. the minimum number after which the BIC increases or decreases by a negligible amount; Jombart et al. 2010). This number was used as prior k for a further DAPC analysis.

The contribution to the first, second, and third discriminant function of this second analysis was quantified for each of the 1,001 SNPs analysed. A series of thresholds (85th, 90th, 95th and 99th percentile of the distribution of

variable contributions to DF1, DF2 and DF3) were considered to identify alleles actually contributing to the inferred structure. All the analyses were performed in the R environment 2.14.1 (R Development Core Team 2008) while the DAPC functions were implemented in the adegenet R package (Jombart 2008). A locus-specific AMOVA analysis was performed for each of the SNPs that significantly contributed to the inferred structure to quantify the amount of the total genetic variation accounted for by the between-groups component (Fct index; Arlequin suite v.3.5). For the same SNPs, the 1000 Genomes Project database (1000 genomes pilot 1 low coverage panel, http://www.1000genomes.org/) was interrogated to identify genetic variants with known or possible phenotypic effect (e.g. non-synonymous substitutions) in strong linkage disequilibrium (LD; minimum $r^2 = 0.8$, minimum $D' = 0.8$, maximum distance between variation 50 kb). LD data are available for three groups of populations: 59 Youruba from Africa (YRI), 60 Utah residents of European ancestry (CEU), and 60 Chinese + Japanese (CHB + JPT).

Results

Markers showing high degree of population differentiation

Genetic variation measured as Fst values among the 52 HGDP populations was equally distributed across markers located in different genic and non-genic regions. More specifically, the proportion of SNPs showing a high degree of population differentiation (those SNPs found within the top 10 % of all Fst values) was similar for downstream, upstream, intronic, synonymous, and non-synonymous markers. Conversely, a less homogeneous distribution of SNP-specific Fst values was observed when SNPs were subdivided according to the class of genes they represent [i.e. CYPs metabolizing exogenous substrates, CYPs with endogenous substrates, non-CYP Phase I metabolizing enzymes, Phase II metabolizing enzymes, transporters, and loci involved in pharmacodynamics (others), Table 1]. Among these classes, *CYP* genes known to principally metabolize exogenous molecules show the highest proportion of highly differentiated SNPs (proportion of SNPs in the top 10 % Fst = 15.32 % representing 8/15 genes), while Phase II DMEs are at the other extreme (6.81 % located in 6/23 genes). Because rare SNPs are more frequent in *CYP* classes than in the other gene classes, the same analysis was performed excluding rare SNPs (minor allele frequency, MAF ≤ 0.05). When rare variants are excluded (Table 1, values in parenthesis), the uneven distribution of genetic variation across gene classes is more evident and statistically significant, as shown by the result of a χ^2 test of independence between number of SNPs in the top 10 % Fst and gene categories ($P = 0.01$).

Geographic structure of PGx genetic variation

Our results indicate that the amount of genetic divergence accounted for by geographic distances between populations is 62 % (Mantel test of matrices correlation, $r = 0.79$, $P < 0.001$). An AMOVA analysis was then performed to test a population structure based on PGx loci largely corresponding to continents or subcontinents, namely Sub-Saharan Africa (7 populations), Middle East (4 populations), Europe (8 populations), Central/South Asia (9 populations), East Asia (17 populations), Oceania (2 populations), and America (5 populations). In this analysis, we tested the geographic structure proposed in previously published studies on the same set of populations (Rosenberg et al. 2002; Li et al. 2008; Biswas et al. 2009). The amount of genetic variation attributable to differences among these 7 groups is 9 % (Fct = 0.09, Fst = 0.12, $P < 0.001$ for both indices), as observed for markers representing the whole-genome variation in the same set of populations (Excoffier and Hamilton 2003; Li et al. 2008). Within geographic regions, populations seem quite homogeneous, showing Fst values from 1 to 4 %, with the exception of Oceania and America, where the within-region genetic variation equals that observed worldwide (Table 2). Within populations, the average gene diversity is highest in Middle East, followed in order of decreasing genetic variation by Central/South Asia, Europe, East Asia, Africa, Oceania and America (Online Resource Table 1b). African populations are less variable than expected; the most plausible explanation for this observation is the ascertainment bias towards populations of European origin that characterises the discovery of important PGx loci.

Genetic structure and cluster membership

When the genetic structure of the 52 HGDP populations was investigated by means of DAPC (Jombart et al. 2010), most of the genetic variation between groups was captured by the first three discriminant functions (DFs), as shown by the first three eigenvalues (i.e. between/within variance ratio of the corresponding discriminant functions) (Online Resource 2). In this analysis, the 52 populations of HGDP were used as prior clusters. The two-dimensional graphic representations of the first three discriminant functions, DF1, DF2, and DF3, look similar to what was obtained simulating genetic data according to a hierarchical island model by Jombart et al. (2010). Populations are clinally distributed along DF1 from Africa to the Middle East and Europe, to Central/South Asia, East Asia, Oceania and America. This result was expected given the observed correlation between

Table 1 Genetic variation of the 1,001 SNPs grouped by gene class and SNP role

	Mean Fst	Tot SNPs	No SNPs top 10 % Fst	Percentage of SNPs top 10 % Fst	Tot genes	No genes top 10 % Fst
Gene class						
CYP exogenous substrates	0.109 (0.137)[a]	127 (101)	21 (18)	15.32 (17.82)	15	8
CYP endogenous substrates	0.103 (0.222)	334 (296)	26 (22)	7.78 (7.43)	28	11
CYP with unknown substrates[b]	0.227	10	6	60.00	2	1
Non-CYP Phase I DMEs	0.117 (0.111)	101 (99)	11 (10)	10.89 (10.10)	20	8
Phase II DMEs[c]	0.091	88	6	6.81 (6.98)	23	6
Transporters	0.106 (0.108)	229 (221)	21	9.17 (9.50)	31	11
Others	0.104 (0.106)	111 (108)	9	8.11 (8.33)	23	7
Transcription factor	0.162	1	0	0.00	1	0
Total	0.106	1,001 (924)	100 (92)	χ^2: $P = 0.08$ (0.01)	143	52
SNP role						
Non-synonymous	0.111	67	7	10.45		
Synonymous	0.109	21	1	4.76		
Intronic	0.104	549	51	9.29		
Upstream	0.115	155	20	12.90		
Downstream	0.102	205	21	10.24		
Pseudogene (*FMO6*)	0.067	4	0	0.00		
Total	0.106	1,001	100	Yates χ^2: $P = 0.81$		

DME drug metabolizing enzymes

[a] In parenthesis are the values for SNPs with MAF > 0.05

[b] *CYP3A43* and *CYP2A7*

[c] *UGT1A* loci counted only once

Table 2 Analysis of molecular variance (AMOVA)

	Groups	Populations	Variance explained (%)		Fst	Fct	Nr loci
			Among pops within regions	Among regions			
World	1	52			0.110		1,001
World	7	52	2.67	9.56	0.123	0.096	1,001
Africa	1	7			0.048		994
Middle East	1	4			0.015		995
Europe	1	8			0.013		966
Central/South Asia	1	9			0.016		982
East Asia	1	17			0.017		954
Oceania	1	2			0.101		831
America	1	5			0.123		883

geographic and genetic distances. The divergence of African and American populations from the rest of the world is highlighted by DF2 and DF3, respectively. Although a great part of the total variance between groups is explained by DF1, DF2, and DF3, in this analysis, further DFs explain still a large portion of it (Online Resource 2). Thus, to better describe the spatial distribution of genetic variation at PGx loci, we first evaluated how many groups of populations

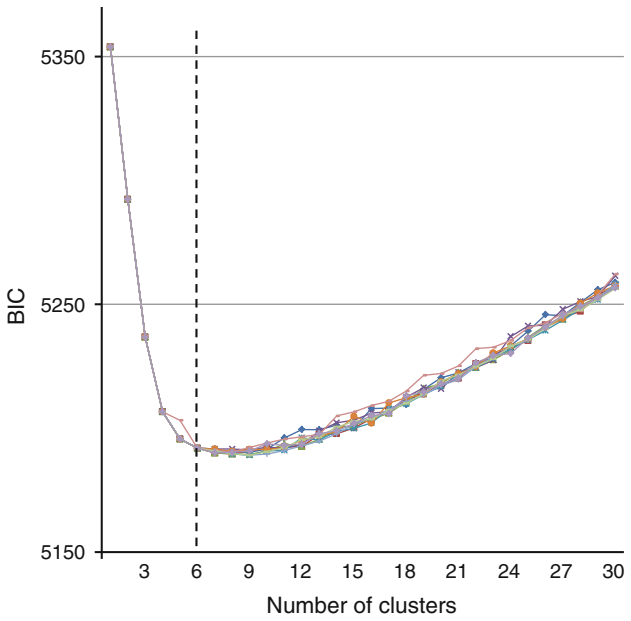

Fig. 1 Inference of the number of clusters supported by the genetic markers analysed in this study (PGx SNPs). The cluster algorithm *k*-means was used to find a given number of groups (*k*) maximising the variation between them. *k*-means was run sequentially with increasing values of *k* (from 1 to 100). The Bayesian information criterion (BIC) was used as a model selection to choose between different cluster solutions. In this figure, the curves of BIC values as a function of *k* are provided for $k = 1, 2,\ldots,30$ clusters (BIC constantly grows for $30 < k \leq 100$). *Ten overlapping curves* represent ten independent runs. The chosen number of clusters (i.e. the minimum number of clusters after which the BIC increases or decreases by a negligible amount) is indicated by the *dashed line*

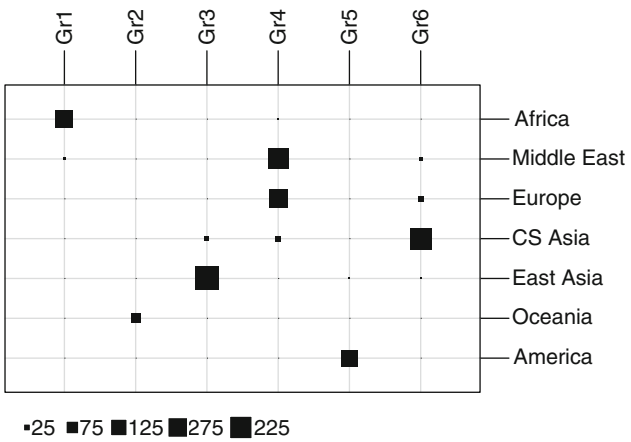

Fig. 2 Membership of individuals from each of the seven geographic regions of origin (i.e. Africa, Middle East, Europe, Central/South Asia, East Asia, Oceania, and America) to each of the six genetic clusters inferred by DAPC and *k*-means algorithm (columns, Gr1–6). The *square size* is proportional to the number of individuals (see the *legend at the bottom*)

could be useful to describe the PGx data, and then used these inferred clusters as prior clusters for DAPC analysis. The results suggest that genetic variation in our dataset supports a subdivision of the 52 populations into six clusters (Fig. 1).

We then labelled the individuals of our dataset based on their geographic regions of origin (i.e. Africa, Middle East, Europe, Central/South Asia, East Asia, Oceania, and America, see the AMOVA analysis), and estimated their membership to each of the six inferred clusters (Fig. 2). Our results show that five of the six inferred groups almost perfectly match the five geographic regions of origin, namely Africa (group 1), Central/South Asia (group 6), East Asia (group 3), Oceania (group 2) and America (group 5), while individuals from Middle East and Europe have been assigned to the same genetic group (group 4). Interestingly, when the same analysis was run considering an additional cluster (i.e. with $k = 7$), America was split into two groups, while Middle East and Europe remained in the same cluster (Online Resource 3).

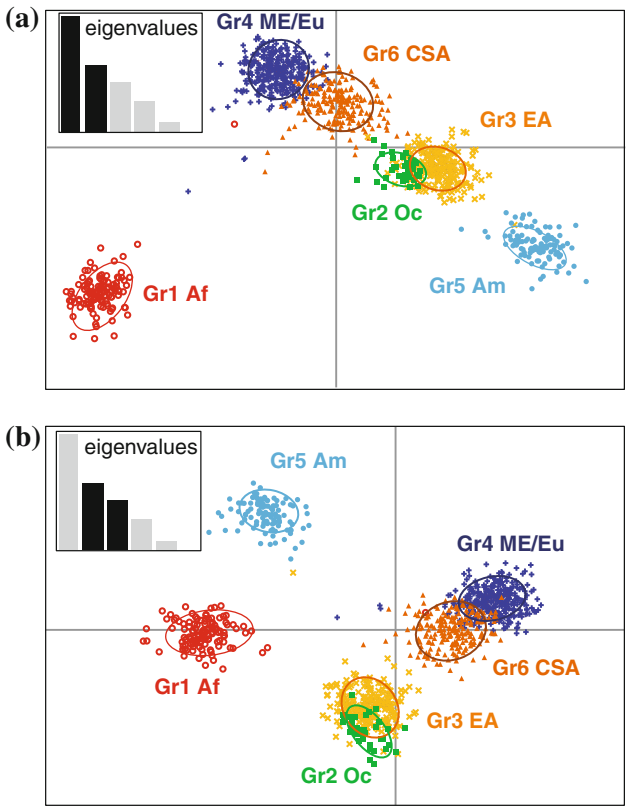

Fig. 3 Scatterplots showing **a** the first two (*x*-axis: DF1; *y*-axis: DF2); and **b** the second and the third (*x*-axis: DF2; *y*-axis: DF3) discriminant functions of 1,001 PGx SNPs. The six groups (Gr1–Gr6) obtained as the best supported structure ($k = 6$, see Fig. 1) were used as prior clusters. Groups are indicated by *different symbols* and inertia ellipses obtained by DAPC. Gr1 Af, Group1 Africa (*open circle*); Gr2 Oc, Group2 Oceania (*filled square*); Gr3 EA, Group3 East Asia (×); Gr4 ME/Eu, Group4 Middle East/Europe (+); Gr6 CSA, Group6 Central/South Asia (*filled up-pointing triangle*); Gr5 Am: Group5 America (*filled circle*)

In the scatterplot of the first two discriminant functions obtained with $k = 6$ (Fig. 3a), both the African and the American populations are clearly differentiated from the rest of the world and from each other. In the same scatterplot, DF1 displays a cline of genetic differentiation from Africa to Eurasia to Oceania and America, as observed in the previous DAPC analysis and consistent with the main expansion of modern humans from Africa. DF2 highlights the divergence of Africa and America, while Middle-Eastern and European populations are at the other extreme. Native American populations are separated from the rest of the world by DF3 (Fig. 3b).

To assess the existence of substructure within clusters, DAPC was applied separately to each of the six inferred groups. The results showed that no subclusters are supported by the data within group 3 (East Asia), group 4 (Middle East and Europe), and group 6 (Central/South Asia), as expected given the AMOVA results (Fst < 2 % in each group). Conversely, group 1 (Africa, Fst $= 5$ %) and group 2 (Oceania, Fst $= 10$ %) could be further subdivided into 2 subclusters, and group 5 (America, Fst $= 12$ %) into four subclusters (Table 3). In Africa, DAPC inferred two subgroups largely corresponding to different subsistence patterns, with subgroup 1 (SGr1.1) including farmers (Bantu, Mandenka and Yoruba), and subgroup 2 (SGr1.2) including hunter-gatherers (Pygmies and San).

Contribution of alleles to genetic clustering

The distribution of allele contributions was plotted for each of the first three DFs and four thresholds were set to identify the genetic variants that defined the six groups structure shown in Fig. 3. On average 150, 100, and 50 alleles showed a contribution above the 85th, 90th, and 95th percentile of the distribution, respectively. Given that the number of retained alleles in the case of these three thresholds was too high to be informative, we focused on the 99th percentile of the distribution. In total, 27 SNPs were found to significantly contribute the most to the six groups structure (Fig. 3). Of these, 8, 9, and 10 contributed to the first, the second and the third discriminant function, respectively (Table 4). In this paragraph, we describe SNPs with known or possible functional impact or association with specific phenotypes. Possible pharmacogenetic implications are commented in "Discussion" section. The distributions of allele frequency across populations and groups are summarised in Online Resource Table 1a and Online Resource 4.

Four of the SNPs contributing to DF1, which traces the "Out of Africa" expansion of modern humans, are located downstream or upstream the closest PGx gene, while the other four are intronic. In other words, none of these variants affects the protein structure of the enzyme encoded by the gene with which they are associated. One of these markers, rs1403527, is among the most differentiated SNPs of our dataset, showing an Fst value of 30 %. The marker is located in the intron of the gene coding for the nuclear receptor subfamily 1, group I, member 2 (*NR1I2*, OMIM 603065). The ancestral allele A has a frequency higher than 80 % in the whole world with the exception of Sub-Saharan Africa, where the average frequency is 30 %. This SNP is in strong linkage disequilibrium with several other intronic variants in all the three populations of the 1000 Genomes Project for which LD data are available. No variation affecting the coding region was found significantly associated with this SNP.

Among the 9 markers contributing to DF2, one is reported by dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) as associated with a probable pathogenic effect: rs699, located on chromosome 1, a non-synonymous substitution of the gene angiotensinogen (*AGT*, OMIM 106150) having a key role in the renin–angiotensin system. The C allele reaches a

Table 3 Number of subgroups (Sgr) inferred within each of the 6 groups, and populations or individuals in each subgroup

Inferred groups	No. of inferred subgroups	Populations or individuals in subgroups
Gr1 (Africa)	2	Sgr1.1 (Bantu NE and SEW, Mandenka, Yoruba, 1 Biaka pygmy)
		Sgr1.2 (Mbuti and Biaka pygmies, San, 1 Bantu NE)
Gr2 (Oceania)	2	Sgr2.1 (Nan Melanesian)
		Sgr2.2 (Papuan)
Gr3 (East Asia)	1	
Gr4 (Middle East and Europe)	1	
Gr5 (America)	4	Sgr5.1 (Colombians, Maya, 1 Karitiana)
		Sgr5.2 (Pima, 1 Maya)
		Sgr5.3 (Surui)
		Sgr5.4 (Karitiana)
Gr6 (Central/South Asia)	1	

Gr group, *Sgr* subgroup, *Bantu NE* Bantu North-East, *Bantu SEW* Bantu South-East + West

Table 4 Genetic features of the 27 SNPs that significantly contributed to the six-groups structure for the first three discriminant functions (DFs)

rs	Gene	Chromosome	Role	Amino acid change (position)	Relevant linked SNPs[a] (population)	Class	Locus-specific analysis of molecular variance (6 groups)
DF1							
rs1403527	*NR1I2*	chr3	Genic (intron)	–		Other	0.34
rs899729	*CYBA*	chr16	Upstream	–		Other	0.14
rs12460831	*CYP4F11*	chr19	Upstream	–		CYP endogenous substrates	0.15
rs10426628	*SULT2B1*	chr19	Genic (intron)	–		Phase II metabolizing Enzyme	0.24
rs11249454	*UGT2A1*	chr4	Genic (intron)	–		Phase II metabolizing enzyme	0.18
rs729147	*ADH7*	chr4	Downstream	–		Non-CYP Phase I metabolizing enzyme	0.16
rs6455682	*SLC22A1*	chr6	Downstream	–		Transporter	0.11
rs3850290	*SLC7A7*	chr14	Genic (intron)	–		Transporter	0.30
DF2							
rs1395	*SLC5A6*	chr2	Genic (ns)	S/F (481)		Transporter	0.22
rs406113	*GPX6*	chr6	Genic (ns)	F/L (13)		Non-CYP Phase I metabolizing enzyme	0.14
rs2472304	*CYP1A2*	chr15	Genic (intron)	–	rs1378942 (CEU) *CSK* gene, intronic	CYP exogenous substrates	0.29
rs2231164	*ABCG2*	chr4	Genic (intron)	–		Transporter	0.29
rs3813720	*ADRB1*	chr10	Downstream	–	rs1801253, CM994344[b] (YRI; CHB + JPT) ns: R/G (389)	Other	0.04
rs1339821	*CYP26C1*	chr10	Upstream	–		CYP endogenous substrates	0.18
rs2197296	*SLC22A1*	chr6	Genic (intron)	–		Transporter	0.06
rs699	*AGT*	chr7	Genic (ns)	M/T (268)		Other	0.19
rs2461817	*NR1I2*	chr3	Genic (intron)	–		Other	0.05
DF3							
rs746713	*NCF4*	chr22	Genic (Intron)	–		Other	0.12
rs2066714	*ABCA1*	chr9	Genic (ns)	I/M (883)		Transporter	0.19
rs4668115	*ABCB11*	chr2	Genic (intron)	–		Transporter	0.12
rs7011901	*CYP11B2*	chr8	Upstream	–	rs4545, CM033362[b] (CHB + JPT) ns: G/S (435)	CYP endogenous substrates	0.27
rs2762926	*CYP24A1*	chr20	Downstream	–		CYP endogenous substrates	0.20
rs338600	*CYP2S1*	chr19	Genic (intron)	–		CYP endogenous substrates	0.10
rs10743413	*SLCO1A2*	chr12	Genic (intron)	–		Transporter	0.12
rs2072671	*CDA*	chr1	Genic (ns)	K/Q (27)		Other	0.09

Table 4 continued

rs	Gene	Chromosome	Role	Amino acid change (position)	Relevant linked SNPs[a] (population)	Class	Locus-specific analysis of molecular variance (6 groups)
rs2608632	*GSTA2*	chr6	Genic (intron)	–		Phase II metabolizing enzyme	0.13
rs7151065	*SLC7A7*	chr14	Genic (intron)	–		Transporter	0.05

SNPs are listed in order of decreasing contribution

ns non-synonymous

[a] Source: 1000 Genomes Project (http://www.1000genomes.org). CEU: Utah residents (CEPH) with Northern and Western European ancestry; CHB + JPT: Han Chinese in Beijing, China + Japanese in Toyko, Japan; YRI: Yoruba in Ibadan, Nigeria

[b] Human Gene Mutation Database (HGMD) dataset (http://www.hgmd.org)

frequency higher than 50 % in 45 out of 52 populations of this study, being almost fixed in Africa, Oceania and America. Two other non-synonymous substitutions contribute substantially to DF2, namely rs1395 of the sodium-dependent multivitamin transporter (*SLC5A6*, OMIM 604024) and rs406113 of the glutathione peroxidase 6 (*GPX6*, OMIM 607913). In both cases, the aminoacid change has been predicted to be benign by Polyphen (http://genetics.bwh.harvard.edu/pph/) and no association with variation in drug metabolism has been reported. Although intronic, the tag SNP rs2472304 of *CYP1A2* (OMIM 124060) contributing to DF2, and specifically the derived A allele being the major allele in Europe, may be of pharmacogenetic interest (see "Discussion"). In individuals of European origin (CEU), rs2472304 is in strong LD with the intronic rs1378942 located in the cytoplasmic tyrosine kinase (*CSK*) gene (OMIM 124095) and significantly associated with blood pressure (Newton-Cheh et al. 2009; Wain et al. 2011). Another association between a SNP contributing to DF2 and a variant of possible phenotypic interest has been found in Yoruba from Africa (YRI) and East Asians from China and Japan (CHB + JPT). In these populations, but not in CEU, the rs3813720 located downstream of the gene *ADRB1* (OMIM 109630) is in strong LD with rs1801253, a non-synonymous SNP of the same gene with a gain of function effect (Mason et al. 1999). Rs1801253 is reported in The Human Gene Mutation Database as CM994344.

Two non-synonymous substitutions contribute to DF3, rs2066714 of the ATP-binding cassette A1 (*ABCA1*, OMIM 600046) and rs2072671 of the Cytidine deaminase (*CDA*, OMIM 123920). Polyphen predicted a benign effect of both aminoacid changes. One SNP contributing to DF3, rs7011901 located upstream of the gene *CYP11B2* (also known as aldosterone synthase, OMIM 124080), is in strong LD with the non-synonymous rs4545 of the same gene in the CHB + JPT sample of the 1000 Genomes Project. This SNP is also reported in the Human Gene Mutation Database (CM033362), and the respective aminoacid change is known to reduce the enzyme activity (Kuribayashi et al. 2003).

Discussion

One of the central problems in pharmacogenetics is to understand to what extent genetic ancestry affects drug response. To achieve this, it is first necessary to describe the distribution of variation at PGx loci across human populations and to try to predict the possible phenotypic consequences.

Overall, the results of our study showed that the genetic variation at drug metabolizing enzymes, receptors, and other molecules involved in drug biotransformation, is distributed and structured similarly to the rest of the genome. More specifically, genetic and geographic distances correlate, consistent with a model of a serial founder effect followed by expansions. Between populations, we observed a global Fst of about 10 % at PGx loci, mostly attributable to variation between geographic regions (9 %, Table 2), which are internally homogeneous. Exceptions to this pattern are Africa, America, and Oceania where populations are highly differentiated for reasons mainly due to the sampling scheme (few and scattered populations) and to the demographic history of these continents (Mulligan et al. 2004; Wang et al. 2007b; Tishkoff et al. 2009). The same geographic trend and apportionment of genetic variation have been previously observed for markers representative of the whole genome (Excoffier and Hamilton 2003; Ramachandran et al. 2005; Li et al. 2008).

Description of the distribution of human PGx variation mentioned above is based on pre-defined groups of populations (Table 2), roughly corresponding to continents, and consistent with the population structure assessed in

previous studies (Rosenberg et al. 2002; Li et al. 2008). However, because our aim was to identify the actual global structure of PGx variation, we inferred the number and the spatial distribution of population groups supported by PGx markers using *k*-means algorithm and DAPC. Our results suggest the existence of six groups of populations, three of which are highly homogeneous, while three are internally structured. As in the AMOVA analysis, the latter corresponds to Africa, Oceania and America.

The five American populations of the HGDP panel are native descendants of the first colonizers, so that variation in their genome has been affected by the initial founder effect, followed by isolation and small effective population size, especially in the case of the Brazilian Surui and Karitiana. However, for practical pharmacogenetics and epidemiology, it should be considered that in some Central and South American Countries, urban populations are quite different from native ones. In Brazil, for example, urban populations are characterised by a high level of native American, European, and African admixture, and by a large census size (Suarez-Kurtz and Pena 2006). A recent study showed that, in this country, genetic variation of *CYP2C18*, *CYP2C19*, *CYP2C9* and *CYP2C8* (together accounting for the biotransformation of 20–30 % of all drugs prescribed worldwide) correlates with the individual proportions of European, African and Amerindian biogeographical ancestry (Suarez-Kurtz et al. 2012). Given that polymorphisms in these genes have been associated with clinically relevant consequences in drug responses (see, for e.g., *CYP2C9* and the anticoagulant warfarin, or *CYP2C19* and proton pump inhibitors; Ingelman-Sundberg et al. 2007) the authors suggest that ancestry inferred by means of ancestry informative markers should be taken into account in practice. By contrast, other South American regions, in particular Andean ones, are characterised by large Native American populations (Scliar et al. 2012), and large urban centres such as Lima, have subpopulations with predominant Native American ancestry (Pereira et al. 2012). Implications of this pattern of ancestry have been shown in the case of NAT2 enzyme in Peru, where native and admixed populations show similar frequencies of different genotype coding for this protein (Fuselli et al. 2007). NAT2 metabolizes commonly prescribed antibiotics, such as the anti-tubercular drug isoniazid, and slow metabolizers have an increased risk to experience adverse reactions to normal doses of antibiotics (Roy et al. 2008). For this reason, a therapy based on individual's genotype should be considered both for native and admixed populations to avoid treatment interruption and the development of drug-resistant bacteria.

Recently, principal component analyses have been applied to genes involved in drug biotransformation (Visscher et al. 2009). However, as the authors acknowledge, the small and European-biased set of populations used in the study may have limited the ability to infer a detailed population structure. To overcome this problem, we used the HGDP panel that includes 52 populations from different parts of the world. This dataset is therefore more inclusive than those with fewer and evolutionary more distant populations analysed in previously published attempts to define discrete groups of humans using PGx genes (Wilson et al. 2001; Visscher et al. 2009). However, the HGDP dataset is not sampled densely, especially in Africa, and entire geographic regions are not represented, for example, India, most of Siberia, and North America. Such a scheme biases the estimation of groups of populations supported by genetic data (Royal et al. 2010). For example, we might have lost the fine-scale structure of the highly diversified African continent (Tishkoff et al. 2009), or the clear separation between Central/South Asia and East Asia may have been exacerbated by the absence of geographically intermediate populations. A bias that clearly affects inferences of population structure is the ascertainment bias that characterises the discovery of genetic variation involved in drug response, typically based on populations of European origin. As we have already mentioned, this may be the reason why Africa shows a lower than expected within-populations genetic variation. Finally, similar to previous studies (Wilson et al. 2001; Visscher et al. 2009), only relatively frequent SNPs can be retrieved from the freely available database used in this study (Li et al. 2008). The lack of rare genetic variants, some of which may be of great importance in drug biotransformation, can affect the identification of groups since frequent SNPs are usually old and thus shared among most human populations. Despite all these limitations, the HGDP panel and the database of markers used in this study still represent the best characterisation of human genetic variation, and allowed us to compare our results with other important studies on genome-wide distributed markers (Rosenberg et al. 2002; Zhivotovsky et al. 2003; Ramachandran et al. 2005; Li et al. 2008; Jakobsson et al. 2008).

As mentioned above, the question of the appropriate way to infer human population genetic structure important for drug metabolism was addressed in two previous studies. Wilson et al. (2001) concluded that the genetic structure of PGx loci matched that inferred using neutral markers by means of a cluster-based algorithm, but was not accurately described by commonly used ethnic labels. Notably, in their study, the proportion of membership of Bantu and Afro-Caribbean populations to the same cluster was >70 %, while the 62 % of Ethiopians clustered together with Europeans and Ashkenazi (see Table 2, Wilson et al. 2001). Conversely, applying PCA to PGx data, Visscher et al. (2009) could define groups of individuals with different biogeographic ancestry. Similarly, the six groups

inferred in our study on the basis of PGx variation roughly correspond to geographic regions, although of the 1,001 analysed SNPs, only 27 significantly contributed to this structure (Table 4). In other words, the six groups structure is mainly defined by a small proportion of the genetic variation analysed in our study. The rest appears differently or more evenly distributed across geographic regions, as supported by the fact that only 16 SNPs show a global Fst $\geq$ 30 % (Online Resource Table 1a). In this respect, however, it should be noted that the bias towards common alleles in our analysis could reduce the proportion of highly differentiated variants. Finally, as highlighted in previous studies (Sistonen et al. 2009), here we show that different SNPs can have remarkably different geographic patterns (Online Resource 4), which makes impossible to consistently define groups of individuals based on PGx variation. Altogether, these observations suggest that any use of ethnicity as proxy of genetic variation involved in drug biotransformation would overly simplify the concept of pharmacogenetics.

The main novel aspect of our study is that, in addition to characterising genetic structure between populations and groups, we also identified the SNPs contributing to the predominant discriminant functions. As already mentioned, DF1 reflects a geographic gradient tracing the main human expansion from Africa, and the distribution of allele frequency of most of the markers contributing to it follows this pattern (Online Resource 4). This gradient has been produced by a combination of events in human demographic history; it is therefore reasonable to expect that loci contributing to it have not been subjected to strong natural selection. According to this expectation, the genetic variation associated with DF1 does not include non-synonymous SNPs (Table 4). However, this view may be oversimplified given that selection does not act only on aminoacid changes, and in many cases, mutations in regulatory regions may have a strong impact on gene expression (Wray 2007). The most interesting among these markers, rs1403527, is located in the gene *NR1I2* coding for the human nuclear receptor 1I2, whose frequency pattern distinguishes Africa from the rest of the world. NR1I2 is of particular pharmacogenetic interest given its role in regulating the expression of several key proteins involved in drug metabolism and transport (Zhou et al. 2009). It functions as a xenobiotic sensor able to control (via induction) the expression of genes involved in xenobiotic metabolism, many of which have broad substrate specificity. In particular, in response to a diversity of natural and synthetic compounds, the nuclear receptor 1I2 activates the transcription of several P450 enzymes, such as CYP2B6, CYP2C8, and CYP3A4, the most important cytochrome P450 in terms of number of clinically used drugs and natural xenobiotics metabolized (Ingelman-Sundberg et al. 2007). Similarly, NR1I2 can activate the expression of Phase II drug metabolizing enzymes and transporters, the multidrug resistance protein 1 (MDR1) among the others (Synold et al. 2001). NR1I2 is also activated by several drugs such as the chemotherapeutic agent Taxol, the antibiotic Rifampicin and peptide mimetic HIV protease inhibitors such as Ritonavir (Dussault and Forman 2002). Considering its broad activity spectrum, mutated NR1I2 may be responsible for variation among individuals and populations ability to metabolize several substrates. Indeed, the molecular basis of the variable drug response for CYP3A4 substrates is still unclear (Ingelman-Sundberg et al. 2007), and it is likely that NR1I2 plays an important role in this process.

Previous studies identified several markers showing a significantly high degree of genetic differentiation in west Eurasian populations. This pattern has been attributed to a specific form of directional selection, termed by some authors *west Eurasian sweep* (Pickrell et al. 2009; Coop et al. 2009). In our study, a similar trend was shown by DF2 (Fig. 3). Among the markers contributing to DF2, the non-synonymous rs699 located in the gene *AGT* is the most interesting for pharmacogenetic and epidemiological aspects. The ancestral C allele of rs699 has been associated with hypertension (Luft 2001) and preeclampsia (Ward et al. 1993), and it is involved in metabolic pathways of drugs such as ACE inhibitors (http://www.pharmgkb.org/). Variation at this SNP correlates with distance from equator (Thompson et al. 2004), a spatial pattern that was interpreted as an adaptation underlying salt regulation, as proposed by the *sodium retention hypothesis* (Nakajima et al. 2004). More specifically, the C allele is the most frequent in 45 out of 52 populations of this study, being almost fixed in Africa, Oceania and America (Online Resource Table 1a, Online Resource 4). If the geographic pattern of rs699 has been shaped by a selective sweep as shown in previous studies, the frequency distribution of the C allele suggests a temperature-dependent selective effect (Hancock et al. 2011) rather than a phenomenon common to all non-African populations. Another interesting association between a genetic marker and a pharmacogenetic phenotype is the A allele of the tag SNP rs2472304 of *CYP1A2*, contributing to DF2, and paroxetine treatment remission in individuals from Thailand (Lin et al. 2010). This association suggests that this SNP itself, or another variant in linkage disequilibrium, is involved in drug efficacy. The genetic structure of rs2472304 clearly shows the *West Eurasian sweep* pattern (Online Resource 4), and an Fct of 0.29 (Table 4), suggesting that the remittent phenotype could be common in Europe and Central/South Asia. In Europeans, rs2472304 is in strong linkage disequilibrium with the intronic rs1378942 located in the gene *CSK*. Genome-wide analyses in populations of European origin

showed that this locus is significantly associated with systolic or diastolic blood pressure, and found substantial evidence for association with hypertension (Newton-Cheh et al. 2009; Wain et al. 2011). We searched the frequency of this SNP in all the available samples of the 1000 Genomes database and observed that its global geographic distribution was extremely similar to that of rs2472304 of *CYP1A2*, which in turn overlaps with that of the already described *AGT* rs699 (Online Resource Table 1a and Online Resource 4). Interestingly, in European populations, both *CSK* rs1378942 and *AGT* rs699 are involved in regulation of blood pressure, and both show a pattern of allele frequency that correlates with distance from equator. This observation may further support the *sodium retention hypothesis* (Nakajima et al. 2004) at least for European populations. Among the SNPs that significantly contributed to DF2, rs3813720 located in *ADRB1* is in strong LD with a gain of function mutation of the same gene (rs1801253, Mason et al. 1999) in YRI and CHB + JPT samples of the 1000 Genomes. This variant is of great pharmacogenetics and epidemiological interest (see http://www.pharmgkb.org/rsid/rs1801253?tabType=tabVip for a summary and for references).

One variant among those significantly contributing to DF3, rs7011901 located in *CYP11B2*, is in strong LD with a non-synonymous substitution (rs4545) of the same gene in Asian populations (CHB + JPT). Although some evidence suggest a phenotypic effect of this substitution (Kuribayashi et al. 2003), its biomedical interest is controversial.

The criteria used to select the genetic markers analysed in this study led our analysis to focus on genes considered of primary importance for drug metabolism. However, it should be noted that genetic variation of other loci involved in pharmacodynamics, such as the *HLA* locus (Profaizer and Eckels 2012), may contribute to the geographic structure of variable drug response.

Finally, although in this study, we did not focus on evolutionary aspects, some speculations may be attempted that could suggest further investigations. In particular, an interesting and still open question is the role of natural selection in the evolution of pharmacogenetic variation. Genetic distances between populations and groups higher or lower than expected may be due to local adaptation (i.e. local selective sweep) or to a similar pattern of polymorphism across different regions of the world (i.e. balancing selection), respectively. PGx genes are important for the interaction between humans and their environment, and for this reason can be good candidates for natural selection. Because selection acts on phenotypes, different patterns of genetic variation at non-synonymous variants and variants in regulatory regions are expected compared with synonymous, intronic and non-regulatory intergenic variants. The 1,001 SNPs analysed in this study did not show specific trends when subdivided into groups with different roles, suggesting that in general the action of natural selection, if any, did not shape the pattern of differentiation between populations at PGx loci. Conversely, highly differentiated SNPs are significantly more frequent among *CYP* preferentially involved in the metabolism of exogenous substances, as the requirement of living organisms to adapt to their environments would predict (Table 1). Given the role of molecules involved in biotransformation of xenobiotics, characterising the evolution of those genes identified in our study as showing extreme values of genetic variation and/or contributing to the first three discriminant functions, can give insights into important questions about how humans evolved in response to specific selective pressures.

Acknowledgments The authors thank Sean Hoban, Mark Jobling, Turi King, Giorgio Bertorelle, and Silvia Ghirotto for their useful suggestions.

References

Bailey KR, Cheng C (2010) Conference Scene: the great debate: genome-wide association studies in pharmacogenetics research, good or bad? Pharmacogenomics 11:305–308

Barbujani G, Colonna V (2010) Human genome diversity: frequently asked questions. Trends Genet 26:285–295

Biswas S, Scheinfeldt LB, Akey JM (2009) Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. Am J Hum Genet 84:641–650

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

Chen X, Wang H, Zhou G, Zhang X, Dong X, Zhi L, Jin L, 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

Chung WH, Hung SI, Hong HS, Hsih MS, Yang LC, Ho HC, Wu JY, Chen YT (2004) Medical genetics: a marker for Stevens–Johnson syndrome. Nature 428:486

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK (2009) The role of geography in human adaptation. PLoS Genet 5:e1000500

Cooper GM, Johnson JA, Langaee TY et al (2008) A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood 112:1022–1027

Dussault I, Forman BM (2002) The nuclear receptor PXR: a master regulator of "homeland" defense. Crit Rev Eukaryot Gene Expr 12:53–64

Excoffier L, Hamilton G (2003) Comment on "Genetic structure of human populations". Science 300:1877 (author reply 1877)

Excoffier L, Lischer H (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567

Fuselli S, Gilman RH, Chanock SJ et al (2007) Analysis of nucleotide diversity of NAT2 coding region reveals homogeneity across Native American populations and high intra-population diversity. Pharmacogenomics J 7:144–152

Fuselli S, de Filippo C, Mona S, Sistonen J, Fariselli P, Destro-Bisol G, Barbujani G, Bertorelle G, Sajantila A (2010) Evolution of detoxifying systems: the role of environment and population history in shaping genetic diversity at human CYP2D6 locus. Pharmacogenet Genomics 20:485–499

Ge D, Fellay J, Thompson AJ, Simon JS et al (2009) Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 461:399–401

Guengerich FP (2003) Cytochromes P450, drugs, and diseases. Mol Interv 3:194–204

Haldane JBS (1940) The blood-group frequencies of European peoples and racial origins. Hum Biol 12:457–480

Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A (2011) Adaptations to climate-mediated selective pressures in humans. PLoS Genet 7:e1001375

Ingelman-Sundberg M, Sim SC, Gomez A, Rodriguez-Antona C (2007) Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. Pharmacol Ther 116:496–526

Jakobsson M, Scholz SW, Scheet P et al (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11:94

Kuribayashi I, Kuge H, Santa RJ et al (2003) A missense mutation (GGC[435Gly] → AGC[Ser]) in exon 8 of the CYP11B2 gene inherited in Japanese patients with congenital hypoaldosteronism. Horm Res 60:255–260

Leabman MK, Huang CC, DeYoung J et al (2003) Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. Proc Natl Acad Sci USA 100:5896–5901

Lewontin RC (1972) The apportionment of human diversity. Evol Biol 6:381–398

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104

Lin KM, Tsou HH, Tsai IJ et al (2010) CYP1A2 genetic polymorphisms are associated with treatment response to the antidepressant paroxetine. Pharmacogenomics 11:1535–1543

Luft FC (2001) Molecular genetics of salt-sensitivity and hypertension. Drug Metab Dispos 29:500–504

Man CB, Kwan P, Baum L, Yu E, Lau KM, Cheng AS, Ng MH (2007) Association between HLA-B*1502 allele and antiepileptic drug-induced cutaneous reactions in Han Chinese. Epilepsia 48:1015–1018

Mason DA, Moore JD, Green SA, Liggett SB (1999) A gain-of-function polymorphism in a G-protein coupling domain of the human beta1-adrenergic receptor. J Biol Chem 274:12670–12674

McCormack M, Alfirevic A, Bourgeois S et al (2011) HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans. N Engl J Med 364:1134–1143

Mulligan CJ, Hunley K, Cole S, Long JC (2004) Population genetics, history, and health patterns in native Americans. Annu Rev Genomics Hum Genet 5:295–315

Nakajima T, Wooding S, Sakagami T et al (2004) Natural selection and population history in the human angiotensinogen gene (AGT): 736 complete AGT sequences in chromosomes from around the world. Am J Hum Genet 74:898–916

Nebert DW, Russell DW (2002) Clinical importance of the cytochromes P450. Lancet 360:1155–1162

Nelson MR, Bryc K, King KS et al (2008) The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83:347–358

Newton-Cheh C, Johnson T, Gateva V et al (2009) Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet 41:666–676

Novembre J, Di Rienzo A (2009) Spatial patterns of variation due to natural selection in humans. Nat Rev Genet 10:745–755

Patin E, Barreiro LB, Sabeti PC et al (2006) Deciphering the ancient and complex evolutionary history of human arylamine *N*-acetyltransferase genes. Am J Hum Genet 78:423–436

Pereira L, Zamudio R, Soares-Souza G et al (2012) Socioeconomic and nutritional factors account for the association of gastric cancer with Amerindian ancestry in a Latin American admixed population. PLoS One 7:e41200

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, Pritchard JK (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19:826–837

Profaizer T, Eckels D (2012) HLA alleles and drug hypersensitivity reactions. Int J Immunogenet 39:99–105

R Development Core Team (2008) R: a language and environment for statistical computing. www.R-project.org. Foundation for Statistical Computing, Vienna

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA 102:15942–15947

Riva A, Kohane IS (2002) SNPper: retrieval and analysis of human SNPs. Bioinformatics (Oxford, England) 18:1681–1685

Rosenberg NA (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

Rosenberg MS, Anderson CD (2001) PASSAGE: pattern analysis, spatial statistics and geographic exegesis, 1.0 edn. Department of Biology, Arizona State University, Tempe, AZ

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science (New York, NY) 298:2381–2385

Roy PD, Majumder M, Roy B (2008) Pharmacogenomics of anti-TB drugs-related hepatotoxicity. Pharmacogenomics 9:311–321

Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ, Clark AG (2010) Inferring genetic ancestry: opportunities, challenges, and implications. Am J Hum Genet 86:661–673

Scliar MO, Soares-Souza GB, Chevitarese J, Lemos L, Magalhaes WC, Fagundes NJ, Bonatto SL, Yeager M, Chanock SJ, Tarazona-Santos E (2012) The population genetics of Quechuas, the largest native South American group: autosomal sequences, SNPs, and microsatellites evidence high level of diversity. Am J Phys Anthropol 147:443–451

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

Suarez-Kurtz G, Pena SD (2006) Pharmacogenomics in the Americas: the impact of genetic admixture. Curr Drug Targets 7:1649–1658

Suarez-Kurtz G, Genro JP, de Moraes MO, Ojopi EB, Pena SD, Perini JA, Ribeiro-Dos-Santos A, Romano-Silva MA, Santana I, Struchiner CJ (2012) Global pharmacogenomics: impact of population diversity on the distribution of polymorphisms in the CYP2C cluster among Brazilians. Pharmacogenomics J 12:267–276

Synold TW, Dussault I, Forman BM (2001) The orphan nuclear receptor SXR coordinately regulates drug metabolism and efflux. Nat Med 7:584–590

Takeuchi F, McGinnis R, Bourgeois S et al (2009) A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genet 5:e1000433

Teichert M, Eijgelsheim M, Rivadeneira F, Uitterlinden AG, van Schaik RH, Hofman A, De Smet PA, van Gelder T, Visser LE, Stricker BH (2009) A genome-wide association study of acenocoumarol maintenance dosage. Hum Mol Genet 18:3758–3768

Thomas JH (2007) Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet 3:e67

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

Tishkoff SA, Kidd KK (2004) Implications of biogeography of human populations for 'race' and medicine. Nat Genet 36:S21–S27

Tishkoff SA, Reed FA, Friedlaender FR et al (2009) The genetic structure and history of Africans and African Americans. Science 324:1035–1044

Visscher H, Ross CJ, Dube MP, Brown AM, Phillips MS, Carleton BC, Hayden MR (2009) Application of principal component analysis to pharmacogenomic studies in Canada. Pharmacogenomics J 9:362–372

Wain LV, Verwoert GC, O'Reilly PF et al (2011) Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. Nat Genet 43:1005–1011

Wang H, Ding K, Zhang Y, Jin L, Kullo IJ, He F (2007a) Comparative and evolutionary pharmacogenetics of ABCB1: complex signatures of positive selection on coding and regulatory regions. Pharmacogenet Genomics 17:667–678

Wang S, Lewis CM, Jakobsson M et al (2007b) Genetic variation and population structure in native Americans. PLoS Genet 3:e185

Ward K, Hata A, Jeunemaitre X et al (1993) A molecular variant of angiotensinogen associated with preeclampsia. Nat Genet 4:59–61

Weiss KM, Long JC (2009) Non-Darwinian estimation: my ancestors, my genes' ancestors. Genome Res 19:703–710

Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, Bradman N, Goldstein DB (2001) Population genetic structure of variable drug response. Nat Genet 29:265–269

Wooding SP, Watkins WS, Bamshad MJ, Dunn DM, Weiss RB, Jorde LB (2002) DNA sequence variation in a 3.7-kb noncoding sequence 5′ of the CYP1A2 gene: implications for human population history and natural selection. Am J Hum Genet 71:528–542

Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8:206–216

Xue Y, Sun D, Daly A, Yang F, Zhou X, Zhao M, Huang N, Zerjal T, Lee C, Carter NP, Hurles ME, Tyler-Smith C (2008) Adaptive evolution of UGT2B17 copy-number variation. Am J Hum Genet 83:337–346

Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72:1171–1186

Zhou C, Verma S, Blumberg B (2009) The steroid and xenobiotic receptor (SXR), beyond xenobiotic metabolism. Nucl Recept Signal 7:e001