CYP2D6 worldwide genetic variation shows high frequency of altered activity variants and no continental structure
Johanna Sistonen\textsuperscript{a}, Antti Sajantila\textsuperscript{a}, Oscar Lao\textsuperscript{c}, Jukka Corander\textsuperscript{b}, Guido Barbujani\textsuperscript{d} and Silvia Fuselli\textsuperscript{a,d}

\textbf{Introduction}
Physiological responses to the same drug treatment are known to vary substantially between different individuals. In addition to external factors, these differences depend on variation at genes coding for proteins involved in the transportation of the drug to its site of action, its interaction with the target, and its metabolism. Among the genes coding for drug-metabolizing enzymes, CYP2D6 (MIM 124030), a member of the cytochrome P450 superfamily, is one of the best characterized. It is responsible for the metabolism of about 25\% of commonly prescribed drugs. Its activity ranges from complete deficiency to excessive activity, potentially causing toxicity of medication or therapeutic failure with recommended drug dosages. This study aimed to describe the CYP2D6 diversity at the global level.

\textbf{Results and conclusions} Our study shows that (i) CYP2D6 diversity is far greater within than between populations and groups thereof, (ii) null or low-activity variants occur at high frequencies in various areas of the world, (iii) linkage disequilibrium is lowest in Africa and highest in the Americas. Patterns of variation, within and among populations, are similar to those observed for other autosomal markers (e.g. microsatellites and protein polymorphisms), suggesting that the diversity observed at the CYP2D6 locus reflects the same factors affecting variation at random genome markers. \textit{Pharmacogenetics and Genomics} 17:93–101 © 2007 Lippincott Williams & Wilkins.

\textbf{Keywords:} CYP2D6, genetic variation, genotyping, polymorphism

\textbf{Methods} A total of 1060 individuals belonging to 52 worldwide-distributed populations were genotyped at 12 highly informative variable sites, as well as for gene deletion and duplications. Phenotypes were predicted on the basis of haplotype combinations.

\textbf{Background and objective} CYP2D6, a member of the cytochrome P450 superfamily, is responsible for the metabolism of about 25\% of the commonly prescribed drugs. Its activity ranges from complete deficiency to excessive activity, potentially causing toxicity of medication or therapeutic failure with recommended drug dosages. This study aimed to describe the CYP2D6 diversity at the global level.
be the possibility to develop ethnically tailored therapies [17–19]. More detailed studies including a high number of populations from different geographic origins, however, are needed to clarify to what extent the relationship between genetics and geography will be of practical use in pharmacogenetics.

This study is the first detailed description of CYP2D6 diversity at the global level, based on a mini sequencing method identifying polymorphism at 12 highly informative variable sites, as well as gene deletion and duplications. The systematic use of the same genotyping technique allowed us to generate comparable data for all populations sampled. Spatial patterns of CYP2D6 variation could be inferred from the analysis of haplotypes.

Materials and methods

DNA samples
We genotyped 1060 individuals belonging to 52 globally distributed populations. These Human Genome Diversity Panel samples were obtained from the Centre d’Etude du Polymorphisme Humain [20]. The sample set actually includes 1064 individuals, but four French individuals had to be excluded from the analyses because we could not amplify their DNAs. In some of the analyses, the population samples were grouped into eight large geographical regions, namely Sub-Saharan Africa, North Africa, the Middle East, Europe, Central/South Asia, East Asia, Oceania and the Americas. This grouping follows the original Centre d’Etude du Polymorphisme Humain documents (http://www.cephb.fr/HGDP-CEPH-Panel) with the exception of dividing Asia into two regions.

CYP2D6 genotyping
Although the terminology differs in different studies, in this paper we shall refer to the whole set of polymorphisms on a chromosome by the term haplotype. Genotyping was performed following a recently described protocol based on long PCR and single nucleotide primer extension reaction [21]. Position 1659 was added to the original 11-plex reaction described before. This genotyping protocol allowed the identification of CYP2D6 variants highly represented in different human populations (i.e. *2, *4, *10, *17, *29, *39 and *41) and variants, even if rare, known to be responsible for low or null metabolic activity (i.e. *3, *6 and *9) [2] as well as the whole gene deletion (*5) and duplications. All haplotypes not showing any of the mutations of interest were classified as *I.

Linkage disequilibrium and network of haplotypes
Haplotypes were inferred from genotypes using the software PHASE v2.1 [22,23]. Linkage disequilibrium (LD) was tested between each pair of polymorphic sites in each geographical region by calculating two statistics, namely $|D'|$ [24] and $R^2$ [25]. Only polymorphic sites with minor-allele frequencies higher than 5% in the region were considered and included in the LD analyses [26]. The significance of associations between polymorphic sites was determined by the Fisher’s exact test and Bonferroni correction, to account for multiple comparisons. Both measures of LD and Fisher’s exact test were calculated using DnaSP 3.99 [27]. The phylogenetic relationships of haplotypes were represented in a tree form using the software TCS [28].

Definition of phenotype classes
The prediction of enzyme activity corresponding to each haplotype (Fig. 1) was based on results obtained from previously published studies (for reference see http://www.cypalleles.ki.se/cyp2d6.htm). To assess the differences in CYP2D6 metabolism among regions of the world we used a conventional classification of phenotypes that is based on the assumption of dominance, in which the phenotype is determined by the most efficient haplotype in the genotype. In this way four phenotypic categories were recognized, namely poor (PM), intermediate (IM), extensive (EM) and ultrarapid metabolizers (UM) [29]; two decreased-function variants or a combination of one decreased-function variant and one nonfunctional variant were classified as IM, whereas UM was defined as a

![Fig 1](image-url)

**Fig 1** CYP2D cluster on chromosome 22 and CYP2D6 inferred haplotypes. Schematic representation of CYP2D6 gene duplication (a), gene deletion (b), normal CYP2D6 cluster (c) and CYP2D6 exons (white boxes) (d). Inferred haplotypes are named as suggested by the guidelines of Human Cytochrome P450 (CYP) Allele Nomenclature Committee. Three new haplotypes (*1661, *4180, *1661xN) were named after the carried mutation.
carrier of an active gene duplication on one chromosome in conjunction with a functional variant on the other chromosome.

**Analysis of molecular variance**

We quantified genetic diversity at three levels, namely between members of the same population, between populations of the same region and between geographical regions, by analysis of molecular variance (AMOVA [30]), using Arlequin v2.0 [31]. We typed the CYP2D6 locus in the same global sample that was analysed for 377 autosomal microsatellites short tandem repeats (STRs) by Rosenberg et al. [11], and to compare the results we chose the same grouping of populations. \( F_{ST} \) statistics, analogues of Wright’s \( F \) statistics that take the evolutionary distance between individual haplotypes into account, were estimated. These results were compared with \( F_{ST} \) values estimated from phenotypic variation.

**Geographic patterns of genetic diversity**

Matrices of geographic (great-circle) distances and genetic distances were calculated between all pairs of populations [32]. In estimating geographic distances, we considered the likely routes of human migration out of Africa, following the criteria by Ramachandran et al. [33]. Genetic distances were estimated as pairwise \( F_{ST} \) distances. Geographic and genetic distances were compared by means of nonparametric Mantel test of matrix correlation [32,34]. Geographic patterns of CYP2D6 single-haplotype diversity were summarized by a spatial autocorrelation statistic, \( I \), estimated by the software PASSAGE [32].

**Results**

**Haplotypic variation**

The inferred haplotypes of 1060 individuals genotyped for CYP2D6 are shown in Fig. 1 and their frequencies in different populations in Table 1. In addition to the already known combination of single nucleotide polymorphisms (SNPs) ([http://www.cypalleles.ki.se/cyp2d6.htm](http://www.cypalleles.ki.se/cyp2d6.htm)), we identified three new haplotypes that bear only one detected SNP, namely 4180G > C, 1661G > C or 1661G > C in a duplicated gene.

When pairs of polymorphic sites were tested for the presence of LD, the statistic \( D' \) was 1 for 78 comparisons out of 82 with four exceptions in Africa and Middle East owing to the presence of the four possible combinations of mutations 1661–2850 (Africa), 100–1661 and 1661–1846 (Middle East) and 1661–4180 (both geographical regions). The values of \( R^2 \) are shown in Fig. 2. Subsaharan Africa displayed the highest diversity, with eight frequent polymorphic positions. By contrast, only three to six variable sites reached the minor allele frequency > 5% in the other regions. Africa was the only continent where association was insignificant for some pairwise comparisons and most of the \( R^2 \) values were below 0.3, whereas all tests reached Bonferroni-corrected statistical significance in the other geographical regions. At the other extreme was Oceania for which estimating LD was impossible because only one mutation (1661G > C) was sufficiently polymorphic. The generally high values of LD and the significance of the association tests allow us to rule out a relevant role of intra-locus recombination in shaping CYP2D6 molecular variation, at least after the human migration out of Africa.

This observation is also supported by the network of haplotypes shown in Fig. 3a, in which the phylogenetic relationships between different variants are unambiguously defined with the only exception of one loop connecting haplotypes *I* and *39*. Above and beyond the clear topolgy of the tree, another important feature is that the fully functional haplotypes *I* and *2* were the most frequent variants and widely distributed in different geographical regions. The network also shows that derived variants leading to null or impaired metabolic activity such as *4* , *10* , *17* and *4I* could reach a relatively high frequency in Europe, East Asia, Africa and Western Eurasia, respectively. Haplotypes *3* and *9* were restricted to Europe, although they did not reach polymorphic frequencies (> 1%). Haplotype *6* was also subpolymorphic, but chromosomes carrying this mutation were found both in Europe and in the Middle East. The Mozabite population from North Africa had the highest frequency of gene duplications. The high values of functional-variant duplication in the Mozabites and the Near East is consistent with previous studies showing similar results in East Africa and the Middle East [35–37]. The Oceanian populations seem to be the outliers in the distribution of haplotype frequencies, showing mostly haplotype *I* and the gene duplication *AxN*, the latter associated with high metabolic activity. The only frequent mutation we detected in this region was the synonymous substitution 1661G > C in the Papuan population. Oceania and America only showed full-functional variants at high frequencies, determining a predominant high metabolic activity of CYP2D6 in these two regions of the world.

By comparing variation at the coding region, as inferred from our 12 polymorphic sites, with the chimpanzee (Pan troglodytes) sequence (GenBank accession number DQ282164), we could identify what can be tentatively considered as a candidate ancestral haplotype, namely *AI80*. This result should be taken cautiously. Indeed, the chimpanzee sequence contains several differences with respect to the human sequence available in GenBank (accession number AY545216), most of them occurring in DNA regions not assayed by the method used for this study. As a consequence, reliably rooting the human CYP2D6 tree seems to require a more extensive survey of its diversity than allowed by 12 SNPs only.
<table>
<thead>
<tr>
<th>Population</th>
<th>Functional Chr</th>
<th><em>f1</em></th>
<th><em>f2</em></th>
<th><em>f39</em></th>
<th><em>f4</em></th>
<th><em>f5</em></th>
<th><em>f6</em></th>
<th><em>f9</em></th>
<th><em>f10</em></th>
<th><em>f17</em></th>
<th><em>f29</em></th>
<th><em>f41</em></th>
<th><em>f1xN</em></th>
<th><em>2xF</em></th>
<th><em>4xN</em></th>
<th><em>10xN</em></th>
<th><em>41xN</em></th>
<th><strong>New</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Moabite (North Africa)</td>
<td>254 (44.4)</td>
<td>28.7</td>
<td>-</td>
<td>0.3</td>
<td>17.2</td>
<td>3.2</td>
<td>0.6</td>
<td>2.5</td>
<td>2.9</td>
<td>-</td>
<td>-</td>
<td>0.6</td>
<td>1.3</td>
<td>0.6</td>
<td>0.3</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Biaka</td>
<td>494 (34.4)</td>
<td>-</td>
<td>0.2</td>
<td>0.2</td>
<td>3.8</td>
<td>0.2</td>
<td>10.5</td>
<td>0.5</td>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Central/South Asia</td>
<td>320 (33.8)</td>
<td>-</td>
<td>0.2</td>
<td>0.2</td>
<td>3.8</td>
<td>0.2</td>
<td>10.5</td>
<td>0.5</td>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Han</td>
<td>90 (27.8)</td>
<td>7.8</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Naj</td>
<td>350 (35.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Hez</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Da</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Ota</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Hua</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Bui</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Daur</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Miao</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Miao</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Moabite (North Africa)</td>
<td>254 (44.4)</td>
<td>28.7</td>
<td>-</td>
<td>0.3</td>
<td>17.2</td>
<td>3.2</td>
<td>0.6</td>
<td>2.5</td>
<td>2.9</td>
<td>-</td>
<td>-</td>
<td>0.6</td>
<td>1.3</td>
<td>0.6</td>
<td>0.3</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Biaka</td>
<td>494 (34.4)</td>
<td>-</td>
<td>0.2</td>
<td>0.2</td>
<td>3.8</td>
<td>0.2</td>
<td>10.5</td>
<td>0.5</td>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Central/South Asia</td>
<td>320 (33.8)</td>
<td>-</td>
<td>0.2</td>
<td>0.2</td>
<td>3.8</td>
<td>0.2</td>
<td>10.5</td>
<td>0.5</td>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Han</td>
<td>90 (27.8)</td>
<td>7.8</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Naj</td>
<td>350 (35.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Hez</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Da</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Ota</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Hua</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Bui</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Daur</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Miao</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Miao</td>
<td>250 (25.0)</td>
<td>15.0</td>
<td>-</td>
<td>0.2</td>
<td>1.1</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td></td>
</tr>
</tbody>
</table>

NE, north-east; SE, south-east; SW, south-west. The haplotype frequencies in geographically defined groups of populations are in bold. Each group consists of populations listed above e.g. Sub-Saharan Africa includes Biaka Pygmies, Mbuti Pygmies, Mandenka, Yoruba, Bantu NE, Bantu SE, SW and San. The only exception is Mozabite which represents alone the geographical region North Africa.

*Number of chromosomes.

*Including one 4B haplotype.

*Including haplotypes carrying only 4180G > C, 1661G > C or 1661G > C in duplicated gene.
Significant values of the association:
the small sample size.

representing North Africa was excluded from this analysis because of
markers [9,38,39].

consistent with estimates based on neutral autosomal
accounted for 9.3% of the total variance, a result
regions (Table 2), the differences between regions
When the whole sample was analysed considering seven
Analysis of molecular variance

Phenotypic variation
Distribution of CYP2D6 phenotypes predicted from
genotypes is shown in Fig. 3b. Europe was characterized
by the highest frequency of PM phenotypes (8%) and it
was actually the only continent in which the distribution
is approximately bimodal [29]. In all other cases the
distribution was unimodal, but the only common feature
was the predominance of the EM class. The second most
common metabolic group in North Africa, Oceania,
Middle East and America was UM (40, 26, 12 and 8%,
respectively). Furthermore, all Oceanian and American
individuals belonged to either the UM or the EM class
which predicts high metabolic capacity, whereas PMs
were completely absent. Common decreased-function
variants, *10, *17 and *4I, led to higher number of IMs in
East Asia, Africa and Middle East than in other regions.
This characteristic has already been described in previous
studies with respect to Africa and Asia [2], but the screening of haplotype *4I allowed us to identify a
relevant number of IMs also in the Middle East.

Analysis of molecular variance
When the whole sample was analysed considering seven
regions (Table 2), the differences between regions
accounted for 9.3% of the total variance, a result
consistent with estimates based on neutral autosomal
markers [9,38,39]. CYP2D6 variances among regions were
similar to those estimated from 377 STRs by Rosenberg
et al. [11]. Europe and Central/South Asia seemed to be
more homogeneous for CYP2D6 than for STRs, so that
almost 100% of the CYP2D6 variation was accounted for
by its within-population component (ΦST = 0.00). The
high variance between populations of the Middle East
was entirely due to the presence of the highly divergent
and geographically distant sample from North Africa, the
Mozabites (28.3% of gene duplications). Oceania seemed
to harbour more variation for CYP2D6 than for STR
markers but this value was due to the presence of a silent
mutation (1661G > C) that does not influence the
protein structure; when the analysis was based on the
phenotypes, variance within Oceania was zero. The
among-population variance estimated for CYP2D6 in
America did not differ from those observed in other
regions, whereas in the study by Rosenberg et al. [11]
America showed the highest value. By and large, in the
AMOVA analysis neither CYP2D6 phenotypes nor haplo-
types showed any evident difference from neutral STRs.

Geographic patterns of genetic diversity
As a preliminary test, we compared a matrix of normalized
CYP2D6 genetic distances, FST/(1-FST), with the matrix
of geographic distances between populations by means of
Mantel test assuming an out of Africa model. The Mantel
permutation test showed that the correlation is close to
significance (P = 0.05), but explains a small fraction of
the total variation (r = 0.18), a result consistent with the
low variances previously observed between populations
and continents. To test whether the genetic diversity
observed for CYP2D6 corresponds to that inferred from
neutral markers, we compared the CYP2D6 genetic
distance matrix with a genetic distance matrix estimated
using 377 autosomal STRs [11]. Positive and statistically
significant correlation was observed between the two
matrices (r = 0.37; P < 0.01) and after controlling for the
geographic distance (r = 0.21; P < 0.05).

The analysis of spatial autocorrelation was repeated
twice: (i) considering all the populations (data not shown)
and (ii) considering only populations in Africa
and Eurasian continent (Fig. 4). Coefficients estimated at
large distances are affected by the small number of
samples in Oceania and the Americas, and by their
extreme geographical position. We placed more confi-
dence in the analysis of the samples of the old world,
whose distribution is both denser and more regular. The
full function and worldwide represented haplotypes *1
and *2 showed significant autocorrelation coefficients
only in few distance classes, and the overall pattern did
not suggest any clear interpretation (Fig. 4a). Conversely,
clear worldwide clines were apparent for haplotypes *4,
*10, *17, and, in part, *4I (Fig. 4b and c), all of them
associated with null or decreased metabolism. These four
haplotypes, each showing its maximum frequency in a
different region (respectively Europe, East Asia, Sub-
saharan Africa and Western-Central Asia), decrease in

Fig. 2

Schematic representation of pairwise linkage disequilibrium in
Subsaharan Africa, Middle East, Europe, Central/South Asia, East Asia
and America. The colour of the square represents the range of R2
values: black for R2 > 0.6; grey 0.6 ≥ R2 ≥ 0.3; white R2 < 0.3.
Significant values of the associations: *P < 0.05; **P < 0.01;
***P < 0.001 after Bonferroni correction. Mozabite population
representing North Africa was excluded from this analysis because of
the small sample size.
frequency with distance from there, suggesting that these regions were the likely centers where these haplotypes originated.

**Discussion**

Previous genetic assessments of the CYP2D6 gene variation have been performed in limited number of populations and often with varying genotyping protocols or interests [2]. To shed light on global variation at this locus, we focused on a detailed molecular study consisting of 52 widely distributed populations from all continents. Our study shows that (i) CYP2D6 diversity is far greater within than between populations and groups thereof; (ii) null or low-activity variants occur at high frequencies in various areas of the world; (iii) linkage disequilibrium is lowest in Africa and highest in the Americas; and (iv) despite the metabolic role of CYP2D6, making it susceptible to selection, the spatial patterns of diversity appear clinal, and very similar to those shown by neutral markers.

All our results suggest that the diversity observed at the CYP2D6 locus reflects the same factors affecting variation at random genome markers. High CYP2D6 genetic variances within populations are in good agreement with those estimated in studies of neutral markers (reviewed in [8]). Patterns of LD are consistent with the results of studies suggesting that through their longer evolutionary history, African populations have had a greater potential for recombination to reduce the LD generated by new
Table 2 AMOVA

<table>
<thead>
<tr>
<th>Sample</th>
<th>Number of regions</th>
<th>Number of populations</th>
<th>Haplotypes</th>
<th>Phenotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Within populations</td>
<td>Among populations within regions</td>
</tr>
<tr>
<td>World</td>
<td>1</td>
<td>52</td>
<td>89.8</td>
<td>10.2</td>
</tr>
<tr>
<td>World (Eurasia)</td>
<td>5</td>
<td>52</td>
<td>86.6</td>
<td>2.6</td>
</tr>
<tr>
<td>World</td>
<td>7</td>
<td>52</td>
<td>88.6</td>
<td>2.1</td>
</tr>
<tr>
<td>Africa</td>
<td>1</td>
<td>7</td>
<td>95.5</td>
<td>4.5</td>
</tr>
<tr>
<td>Eurasia</td>
<td>3</td>
<td>21</td>
<td>97.9</td>
<td>2.1</td>
</tr>
<tr>
<td>Europe</td>
<td>1</td>
<td>8</td>
<td>99.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Middle East</td>
<td>1</td>
<td>4</td>
<td>99.4</td>
<td>4.8</td>
</tr>
<tr>
<td>Middle East</td>
<td>1</td>
<td>3</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Central/South</td>
<td>(no Mozabites)</td>
<td>1</td>
<td>9</td>
<td>100.0</td>
</tr>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>East Asia</td>
<td>1</td>
<td>17</td>
<td>96.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Oceania</td>
<td>1</td>
<td>2</td>
<td>90.3</td>
<td>9.7</td>
</tr>
<tr>
<td>America</td>
<td>1</td>
<td>5</td>
<td>96.8</td>
<td>3.2</td>
</tr>
</tbody>
</table>

AMOVA, analysis of molecular variance.

*In Rosenberg et al. [11], number of populations = 6 (Bantu populations together).

Typically, differences in the patterns of diversity shown by different markers are attributed either to chance or to selection. Inferring selection was not the aim of the present study; however, the homogeneous geographic distribution of haplotypes *I* and *2* could be regarded as the result of a long-term selective pressure maintaining the high frequency of haplotypes coding for a full-function enzyme. Also, local high frequencies of null or reduced-activity haplotypes may indeed be due to selective pressures affecting the local populations. Selection, however, can hardly account for the global patterns of \(CYP2D6\) variation. Indeed, these patterns were very similar to those described for neutral markers, both by AMOVA and by autocorrelation analysis. This suggests that the global \(CYP2D6\) diversity was largely shaped by the same combination of gene flow and drift events that shaped the diversity of most other genome regions.

Statistics estimated from SNP data may suffer from ascertainment bias. The genotyping system used in this study allowed us to identify 12 possible mutations of \(CYP2D6\) gene, together with the whole-gene deletion and duplication. Typing of SNPs known to be polymorphic in certain populations may lead to underestimation of genetic variation in other populations. This is especially true in the case of pharmacogenetic genes, mainly characterized in European and North American individuals of European ancestry. To quantify approximately the ascertainment bias, we compared the values of \(\Phi_{ST}\) estimated from complete coding \(CYP2D6\) sequences, and from 12 SNPs, in samples coming from an analysis of \(CYP2D6\) sequence diversity (Fuselli et al., unpublished data). \(\Phi_{ST}\) values did not differ significantly over 10 populations originating from Africa, Europe and Asia (\(\Phi_{ST} = 0.09\) based on sequences, and \(\Phi_{ST} = 0.10\) based on SNPs) and in six non-African samples (\(\Phi_{ST} = 0.08\) based on sequences, and \(\Phi_{ST} = 0.09\) based on SNPs), but the 12 SNPs used for the present study underestimated variation in the four African samples (\(\Phi_{ST} = 0.02\) based on sequences, and \(\Phi_{ST} = 0.00\) based on SNPs). Therefore, we cannot rule out that a fraction, which we cannot quantify, of African diversity passed undetected in this study. This may explain why continent-specific haplotypes were observed only in Europe, and not in Africa. Africa, however, is at one extreme of the area affected by the cline, and so greater diversity there could only increase the significance of the pattern observed. Therefore, we cannot rule out that ascertainment bias has affected some of our results, but the geographic cline observed is significant despite, not because, that possible bias.

As for these spatial patterns, series of founder effects in the course of an expansion from Africa can explain the correlation between genetic and geographic distances [33,43]. The autocorrelation patterns observed in this study show that \(CYP2D6\) diversity can be described as clinal. The overall geographic gradient largely reflects the gradients shown by the four common haplotypes determining a null or reduced metabolism. Each of these haplotypes shows its maximum in a different region of the world.

Furthermore, we ascertained how many different groups of populations were supported by \(CYP2D6\) data from this study. To this aim, we used Bayesian analysis of population structure (BAPS) [44,45], a Bayesian Monte-Carlo Markov chain approach, that allowed to assign single populations to a nonpredefined number of groups.
Sampled populations were clustered using 50 parallel simulation chains over 20,000 iterations. Stability and convergence of the analysis was ensured by considering five replicates of the simulation runs. The analysis showed that 10 clusters out of 11 identified included either some but not all populations of a continent, or populations of different continents (data not shown). Therefore, it is hardly surprising that the 11 CYP2D6 clusters do not overlap with those described in any other study focused on human genetic variation at a worldwide level [9,11,14]. Contrary to what has been claimed by some authors [15], there is no guarantee that by analysing a given set of genetic markers, one can obtain information on genome diversity at large.

Although the aim of this study was not to replace genotype/phenotype correlation studies, our description of inferred phenotypes may be of significance for pharmacogenetic applications. Altered CYP2D6 metabolic activity has been associated with adverse drug reactions [1] or even fatal intoxications [46,47]. In the majority of cases, metabolism mediated by CYP2D6 contributes to inactivation of a drug. For some drugs, however, CYP2D6 catalyses the conversion of a prodrug into an active compound. Thus, adverse reactions can be caused not only by a slower than normal metabolic rate, but also by ultrarapid metabolism [48]. Our results highlight the relevance of the UM phenotype class represented in each of the eight geographical regions considered in this study, being the second most common group of individuals in North Africa, Middle East, Oceania and America. On the other hand, European populations showed the highest frequencies of the PM phenotype, and about one chromosome out of six carried the null-function haplotype *4. We, however, cannot exclude an underestimation of population/region-specific variants (either not tested or unknown) that could conceivably lead to a phenotype other than the one predicted in this study.

CYP2D6 is of great interest for clinical practice because it is responsible for the metabolism of many commonly used drugs, and its genetic polymorphism can have a strong effect on the substrate. On the basis of our study, CYP2D6 genetic variants related to altered metabolic activity are highly represented in different regions of the world. The development of ethnically tailored therapies, however, seems difficult to realize owing to the fact that there are only few rarely observed region-specific haplotypes changing the phenotype characterized to date and most of the variants seem to be geographically dispersed over all continents. Furthermore, population admixture is common or quickly increasing in many populations, which should be also taken into account when applying results obtained from pharmacogenetic studies [49]. Even if CYP2D6 polymorphism represents an excellent example of the potential clinical implications of pharmacogenetic research [50], most of the drug effects and treatment outcomes are determined by the interaction of multiple genes [51]. Naturally, more knowledge on various factors affecting the drug response has to be obtained before the pharmacogenetic approach can be extensively used in the clinical practice.

Spatial autocorrelation analysis in populations from the old world. x-axis: higher limit of geographic distance classes (in kilometers). y-axis: Autocorrelation index I. Filled symbols indicate significant values.
Acknowledgements

The authors thank Jukka Palo for comments on the manuscript. J.S. was funded by the Finnish Graduate School in Population Genetics. G.B. was supported by funds from the Italian Ministry of Universities (FISR).

References

1. Ingelman-Sundberg M. Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. Pharmacogenomics J 2005; 5:6–13.