CYP2D6 worldwide genetic variation shows high frequency of altered activity variants and no continental structure

Johanna Sistonen[a], Antti Sajantila[a], Oscar Lao[c], Jukka Corander[b], Guido Barbujani[d] and Silvia Fuselli[a,d]

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.

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.

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. *Pharmacogenetics and Genomics* 17:93–101 © 2007 Lippincott Williams & Wilkins.

Pharmacogenetics and Genomics 2007, **17**:93–101

Keywords: CYP2D6, genetic variation, genotyping, polymorphism

Departments of [a]Forensic Medicine, [b]Mathematics and Statistics, University of Helsinki, Finland, [c]Department of Forensic Molecular Biology, Erasmus University Medical Center, DR Rotterdam, The Netherlands and [d]Department of Biology, University of Ferrara, Ferrara, Italy

Correspondence and requests for reprints to Johanna Sistonen, Department of Forensic Medicine, P.O. Box 40 00014 University of Helsinki, Finland
Tel/fax: +358 9 191 27450/27518;
e-mail: johanna.sistonen@helsinki.fi

Received 21 April 2006 **Accepted** 6 August 2006

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 used drugs belonging to classes such as antidepressants, neuroleptics, β-blockers and antiarrhythmics [1]. The *CYP2D6* gene is highly polymorphic with more than 50 variants described to date (*http://www.cypalleles.ki.se/cyp2d6.htm*). The phenotypic consequences of this variation are considerable. The CYP2D6 enzyme activity ranges from complete deficiency to ultrarapid metabolism, possibly giving rise to profound toxicity of medication or therapeutic failure with recommended drug dosages.

Previous genetic studies showed high levels of *CYP2D6* polymorphism, both within and between populations [2], and a surprisingly high frequency of null and reduced-function variants. These findings raise several questions of evolutionary and applied relevance. First, such a high diversity can hardly be maintained by a simple mechanism of directional selection common to all populations, or by genetic drift alone. As a consequence, more complex processes must be envisaged, and any explanation of the observed diversity must account for the local occurrence at substantial frequencies of null or reduced-activity variants. Second, *CYP2D6* sequence diversity is clearly associated with phenotypic variation in the gene's expression and activity, which in turn is part of a complex network of interactions of extreme pharmacogenetic and pharmacogenomic interest. A third question bears on the interpretation of human diversity in general, which has recently been and still is the subject of intense debate (for reviews, see e.g. [3–8]). Some studies of neutral markers described a gradation of genetic diversity in the geographical space, with allele frequencies forming clines over much of the planet [9,10]. Geographic structuring, however, is also evident [11,12], which is interpreted by some authors as evidence that a concordant clustering of genotypes in major continental or subcontinental clusters is both possible, and useful for medical practice [13]. In particular, the main focus of the challenging debate about individual's ancestry and drug response [14–16] seems to

1744-6872 © 2007 Lippincott Williams & Wilkins

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 Subsaharan 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 **1*.

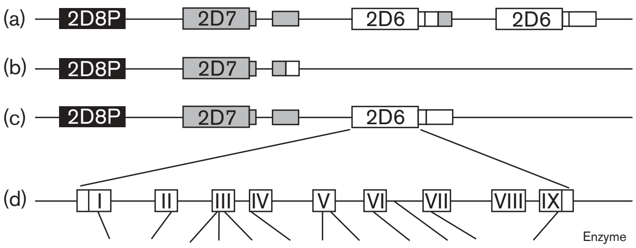
Linkage disequilibrium and network of haplotypes

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



Haplotype	100	1023	1659	1661	1707	1846	2549	2613-15	2850	2988	3183	4180	Dup	Enzyme activity
*1	C	C	G	G	T	G	A	AGA	C	G	G	G	–	norm
*2	C	C	G	C	T	G	A	AGA	T	G	G	C	–	norm
*3	C	C	G	G	T	G	–	AGA	C	G	G	G	–	none
*4	T	C	G	C	T	A	A	AGA	C	G	G	C	–	none
*4B	T	C	G	G	T	A	A	AGA	C	G	G	C	–	none
*5	–	–	–	–	–	–	–	–	–	–	–	–	–	none
*6	C	C	G	G	–	G	A	AGA	C	G	G	G	–	none
*9	C	C	G	G	T	G	A	–	C	G	G	G	–	decr
*10	T	C	G	C	T	G	A	AGA	C	G	G	C	–	decr
*17	C	T	G	C	T	G	A	AGA	T	G	G	C	–	decr
*29	C	C	A	C	T	G	A	AGA	T	G	A	C	–	decr
*39	C	C	G	C	T	G	A	AGA	C	G	G	C	–	norm
*41	C	C	G	C	T	G	A	AGA	T	A	G	C	–	decr
*1661	C	C	G	C	T	G	A	AGA	C	G	G	G	–	norm
*4180	C	C	G	G	T	G	A	AGA	C	G	G	C	–	norm
*1xN	C	C	G	G	T	G	A	AGA	C	G	G	G	+	incr
*2xN	C	C	G	C	T	G	A	AGA	T	G	G	C	+	incr
*4xN	T	C	G	C	T	A	A	AGA	C	G	G	C	+	none
*10xN	T	C	G	C	T	G	A	AGA	C	G	G	C	+	decr
*41xN	C	C	G	C	T	G	A	AGA	T	A	G	C	+	decr
*1661xN	C	C	G	C	T	G	A	AGA	C	G	G	G	+	incr

CYP2D cluster on chromosome 22 and *CYP2D6* inferred haplotypes. Schematic representation of *CYP2D6* gene duplication (a), gene deletion (b), normal *CYP2D* 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. Φ 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*), 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 **1* and **39*. Above and beyond the clear topology of the tree, another important feature is that the fully functional haplotypes **1* 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 **41* 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 **1* and the gene duplication **1xN*, 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 **4180*. 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 1 ***CYP2D6* haplotype frequencies in single populations and geographically defined groups of populations**

		Functional			Nonfunctional				Reduced					Duplications					
Population	Chr[a]	**1*	**2*	**39*	**3*	**4*[b]	**5*	**6*	**9*	**10*	**17*	**29*	**41*	**1xN*	**2xN*	**4xN*	**10xN*	**41xN*	New[c]
Biaka Pygmies	72	18.1	50.0	–	–	1.4	1.4	–	–	8.3	12.5	2.8	1.4	–	–	1.4	–	–	2.8
Mbuti Pygmies	30	10.0	60.0	–	–	–	10.0	–	–	–	3.3	3.3	–	13.3	–	–	–	–	–
Mandenka	48	25.0	12.5	–	–	12.5	6.3	–	–	6.3	18.8	6.3	10.4	–	–	–	–	–	2.1
Yoruba	50	42.0	12.0	–	–	–	4.0	–	–	4.0	6.0	12.0	2.0	–	4.0	14.0	–	–	–
Bantu NE	24	29.2	25.0	–	–	–	4.2	–	–	–	16.7	16.7	–	–	–	4.2	–	–	4.2
Bantu SE,SW	16	37.5	12.5	–	–	–	18.8	–	–	–	25.0	6.3	–	–	–	–	–	–	–
San	14	0.0	64.3	–	–	–	14.3	–	–	–	7.1	–	–	14.3	–	–	–	–	–
Subsaharan Africa	**254**	**24.4**	**32.7**	–	–	**2.8**	**5.9**	–	–	**4.3**	**12.2**	**6.7**	**2.8**	**2.4**	**0.8**	**3.5**	–	–	**1.6**
Mozabite (North Africa)	**60**	**11.7**	**28.3**	–	–	**11.7**	**3.3**	–	–	–	**8.3**	–	**8.3**	–	**28.3**	–	–	–	–
Bedouin	98	35.7	20.4	–	–	5.1	4.1	2.0	–	–	3.1	–	22.4	1.0	2.0	–	–	–	4.1
Druze	96	29.2	27.1	–	–	7.3	6.3	–	–	–	1.0	–	15.6	10.4	3.1	–	–	–	–
Palestinian	102	40.2	27.5	–	–	7.8	1.0	2.0	–	2.0	2.0	–	12.7	–	4.9	–	–	–	–
Middle East	**296**	**35.1**	**25.0**	–	–	**6.8**	**3.7**	**1.4**	–	**0.7**	**2.0**	–	**16.9**	**3.7**	**3.4**	–	–	–	**1.4**
French	50	28.0	32.0	–	–	16.0	4.0	–	4.0	2.0	–	–	10.0	2.0	–	2.0	–	–	–
French Basque	48	29.2	20.8	–	–	20.8	10.4	–	6.3	6.3	–	–	4.2	–	2.1	–	–	–	–
Sardinian	56	28.6	35.7	–	1.8	21.4	1.8	–	–	5.4	–	–	3.6	–	1.8	–	–	–	–
North Italian	28	39.3	28.6	–	–	14.3	3.6	3.6	–	–	–	–	3.6	–	–	3.6	3.6	–	–
Tuscan	16	50.0	18.8	–	–	18.8	–	–	6.3	–	–	–	6.3	–	–	–	–	–	–
Orcadian	32	46.9	18.8	–	–	12.5	–	3.1	6.3	3.1	–	–	6.3	3.1	–	–	–	–	–
Adygei	34	41.2	29.4	–	–	8.8	2.9	–	–	2.9	–	–	14.7	–	–	–	–	–	–
Russian	50	32.0	34.0	–	–	20.0	–	–	–	–	–	–	8.0	–	4.0	–	–	–	2.0
Europe	**314**	**34.4**	**28.7**	–	**0.3**	**17.2**	**3.2**	**0.6**	**2.5**	**2.9**	–	–	**7.0**	**0.6**	**1.3**	**0.6**	**0.3**	–	**0.3**
Brahui	50	54.0	20.0	–	–	4.0	6.0	–	–	2.0	–	–	14.0	–	–	–	–	–	–
Balochi	50	36.0	32.0	–	–	8.0	2.0	–	–	8.0	–	2.0	10.0	–	2.0	–	–	–	–
Hazara	50	40.0	26.0	–	–	12.0	6.0	–	–	4.0	–	–	8.0	2.0	2.0	–	–	–	–
Makrani	50	42.0	30.0	–	–	6.0	8.0	–	–	4.0	–	–	10.0	–	–	–	–	–	–
Sindhi	50	44.0	20.0	–	–	12.0	4.0	–	–	6.0	–	–	14.0	–	–	–	–	–	–
Pathan	50	42.0	30.0	–	–	10.0	–	–	–	2.0	–	–	14.0	2.0	–	–	–	–	–
Kalash	50	50.0	36.0	–	–	8.0	–	–	–	–	–	–	6.0	–	–	–	–	–	–
Burusho	50	38.0	40.0	2.0	–	6.0	4.0	–	–	4.0	–	–	6.0	–	–	–	–	–	–
Uygur	20	45.0	25.0	–	–	5.0	5.0	–	–	5.0	–	–	15.0	–	–	–	–	–	–
Central/South Asia	**420**	**43.3**	**29.0**	**0.2**	–	**8.1**	**3.8**	–	–	**3.8**	–	**0.2**	**10.5**	**0.5**	**0.5**	–	–	–	–
Han	90	27.8	7.8	–	–	2.2	1.1	–	–	56.7	–	–	–	2.2	–	–	1.1	1.1	–
Tujia	20	35.0	15.0	–	–	–	10.0	–	–	40.0	–	–	–	–	–	–	–	–	–
Yizu	20	45.0	15.0	–	–	5.0	5.0	–	–	15.0	–	–	15.0	–	–	–	–	–	–
Miaozu	20	45.0	10.0	–	–	–	10.0	–	–	35.0	–	–	–	–	–	–	–	–	–
Oroqen	20	50.0	10.0	–	–	5.0	5.0	–	–	25.0	–	–	–	–	–	–	5.0	–	–
Daur	20	25.0	15.0	–	–	10.0	–	–	–	50.0	–	–	–	–	–	–	–	–	–
Mongola	20	25.0	25.0	–	–	10.0	–	–	–	35.0	–	–	–	–	–	–	5.0	–	–
Hezhen	20	40.0	30.0	–	–	–	10.0	–	–	20.0	–	–	–	–	–	–	–	–	–
Xibo	18	16.7	38.9	–	–	–	5.6	–	–	38.9	–	–	–	–	–	–	–	–	–
Dai	20	10.0	5.0	–	–	5.0	15.0	–	–	55.0	–	–	10.0	–	–	–	–	–	–
Lahu	20	20.0	15.0	–	–	15.0	5.0	–	–	25.0	–	–	20.0	–	–	–	–	–	–
She	20	25.0	10.0	–	–	–	15.0	–	–	50.0	–	–	–	–	–	–	–	–	–
Naxi	20	40.0	5.0	–	–	–	5.0	–	–	50.0	–	–	–	–	–	–	–	–	–
Tu	20	15.0	20.0	–	–	–	10.0	–	–	45.0	–	–	5.0	–	5.0	–	–	–	–
Yakut	50	32.0	38.0	–	–	2.0	10.0	–	–	12.0	–	–	2.0	–	4.0	–	–	–	–
Japanese	62	37.1	16.1	–	–	–	3.2	–	–	40.3	–	–	–	–	–	–	3.2	–	–
Cambodian	22	31.8	4.5	4.5	–	–	4.5	–	–	54.5	–	–	–	–	–	–	–	–	–
East Asia	**482**	**30.9**	**16.4**	**0.2**	–	**2.7**	**5.8**	–	–	**39.4**	–	–	**2.3**	**0.4**	**0.6**	–	**1.0**	**0.2**	–
Papuan	34	55.9	–	–	–	–	2.9	–	–	–	–	–	–	11.8	–	–	–	–	29.4
NAN Melanesian	44	84.1	–	–	–	–	–	–	–	4.5	–	–	–	11.4	–	–	–	–	–
Oceania	**78**	**71.8**	–	–	–	–	**1.3**	–	–	**2.6**	–	–	–	**11.5**	–	–	–	–	**12.8**
Pima	50	74.0	10.0	–	–	8.0	–	–	–	–	–	–	–	4.0	4.0	–	–	–	–
Maya	50	48.0	38.0	–	–	6.0	–	–	–	–	2.0	–	–	6.0	–	–	–	–	–
Colombian	26	50.0	42.3	–	–	–	7.7	–	–	–	–	–	–	–	–	–	–	–	–
Karitiana	48	62.5	29.2	–	–	–	–	–	–	–	–	–	–	–	8.3	–	–	–	–
Surui	42	61.9	38.1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
America	**216**	**60.2**	**30.1**	–	–	**3.2**	**0.9**	–	–	–	**0.5**	–	–	**2.3**	**2.8**	–	–	–	–
Total	**2120**	**0.38**	**0.25**	**<0.01**	**<0.01**	**0.07**	**0.04**	**<0.01**	**<0.01**	**0.11**	**0.02**	**0.01**	**0.07**	**0.02**	**0.02**	**0.01**	**<0.01**	**<0.01**	**0.01**

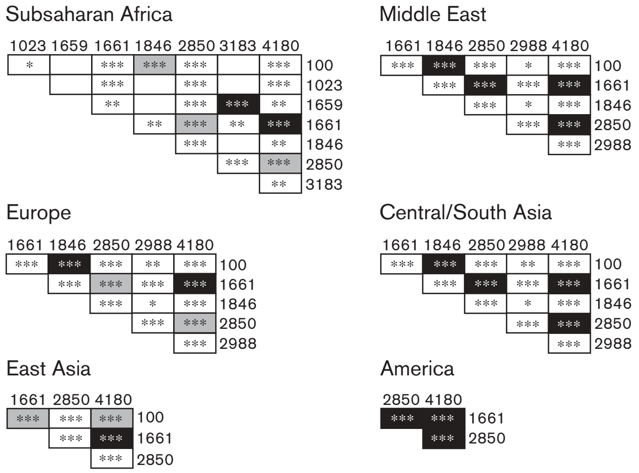
NE, north-east; SE, south-east; SW, south-west. The haplotype frequencies in geographically defined groups of populations are in bold. Each group consists of populations listed above e.g. SubSaharan Africa includes Biaka Pygmies, Mbuti Pygmies, Mandenka, Yoruba, Bantu NE, Bantu SE, SW and San. The only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

[a] Number of chromosomes.

[b] Including one **4B* haplotype.

[c] Including haplotypes carrying only 4180G>C, 1661G>C or 1661G>C in duplicated gene.

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 R^2 values: black for $R^2 > 0.6$; grey $0.6 \geq R^2 \geq 0.3$; white $R^2 < 0.3$. Significant values of the association: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ after Bonferroni correction. Mozabite population representing North Africa was excluded from this analysis because of the small sample size.

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 **41*, 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 **41* allowed us to identify a relevant number of IMs also in the Middle East.

Analysis of molecular variance

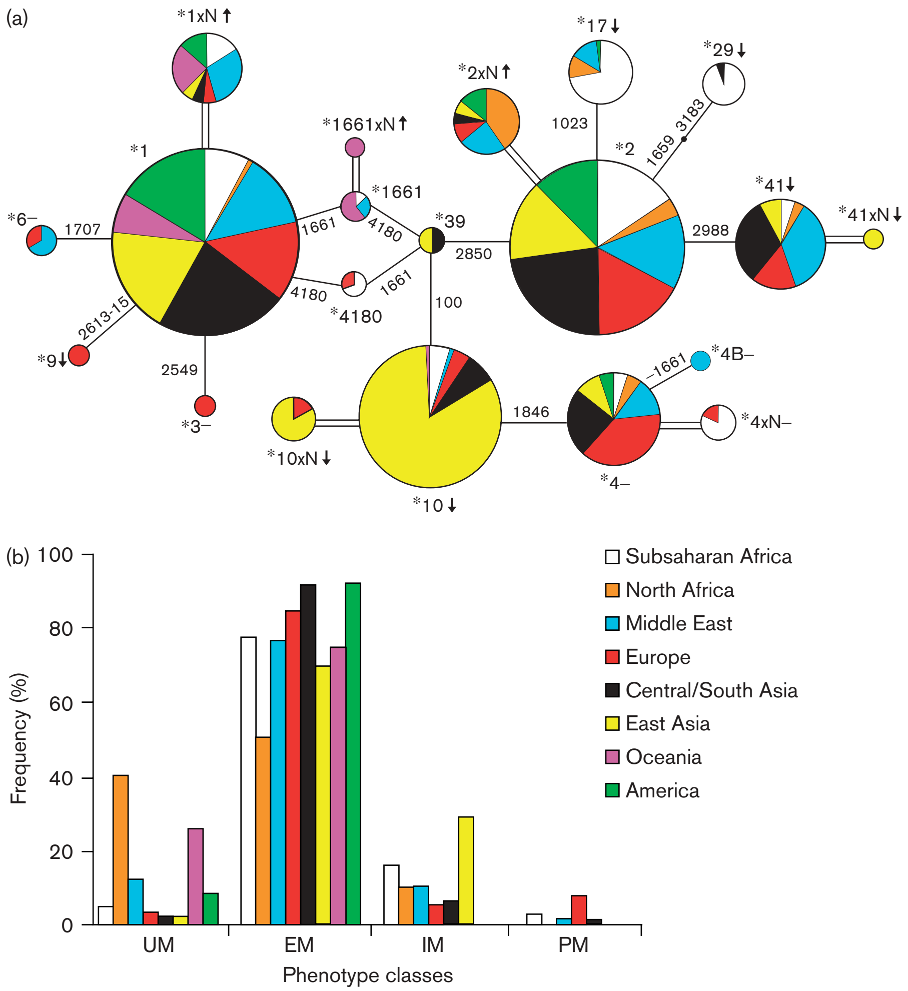
When the whole sample was analysed considering seven regions (Table 2), the differences between regions accounted for 9.3% of the total variance, a result consistent with estimates based on neutral autosomal markers [9,38,39]. *CYP2D6* variances among regions were similar to those estimated from 377 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 ($\Phi_{ST} = 0.00$). 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 haplotypes 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, $F_{ST}/(1\text{-}F_{ST})$, 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 confidence 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, **41* (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, Subsaharan Africa and Western-Central Asia), decrease in

Fig. 3



CYP2D6 haplotype and phenotype diversity in different geographical regions. (a) *CYP2D6* haplotypes are represented in a network. The size of the circle is proportional to the haplotype frequency in the whole dataset. Mutations separating haplotypes are marked in the figure. Double lines correspond to gene duplication. The altered enzymatic activity related to a haplotype is represented as follows: increased (↑), decreased (↓), null (–). (b) Frequency of *CYP2D6* phenotype classes is shown in different geographical regions. Phenotypes are predicted from genotypes following the model described in Material and methods. UM: ultrarapid metabolizers; EM: extensive metabolizers; IM: intermediate metabolizers; PM: poor metabolizers.

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

Sample	Number of regions	Number of populations	Haplotypes			Phenotypes		
			Within populations	Among populations within regions	Among regions	Within populations	Among populations within regions	Among regions
World	1	52	89.8	10.2		90.5	9.5	
World (Eurasia)	5	52	86.6	2.6	10.8	88.8	5.5	5.8
World	7	52	88.6	2.1	9.3	89.6	3.9	6.5
Africa	1	7[a]	95.5	4.5		94.5	5.5	
Eurasia	1	21	97.9	2.1		94.5	5.5	
Eurasia	3	21	97.5	1.0	1.5	93.0	1.7	5.2
Europe	1	8	99.8	0.2		100.0	0.0	
Middle East	1	4	95.4	4.6		93.4	6.6	
Middle East (no Mozabites)	1	3	100.0	0.0		98.0	2.0	
Central/South Asia	1	9	100.0	0.0		100.0	0.0	
East Asia	1	17	96.0	4.0		93.1	6.9	
Oceania	1	2	90.3	9.7		100.0	0.0	
America	1	5	96.8	3.2		98.0	2.0	

AMOVA, analysis of molecular variance.

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

mutations or founder effects [40,41]. The broad geographic cline of *CYP2D6* diversity parallels those observed by Serre and Paabo [10], Ramachandran *et al.* [33] and, with protein markers, by Cavalli-Sforza *et al.* [42].

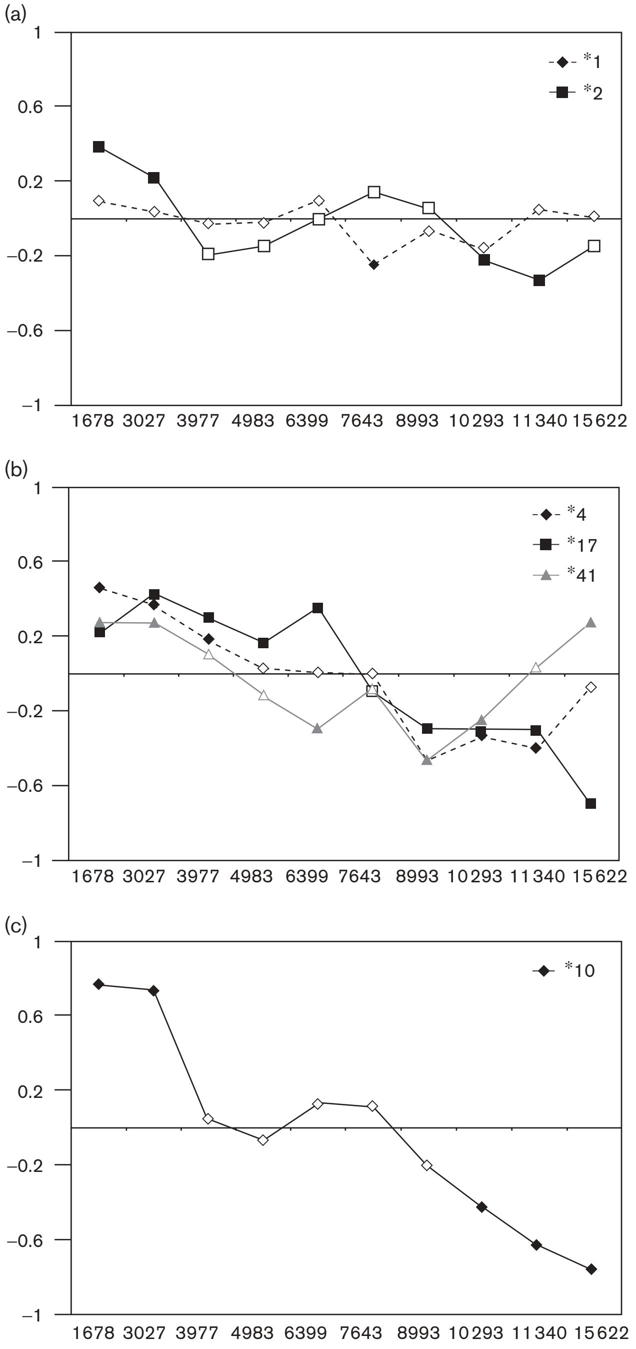
Typically, differences in the patterns of diversity shown by different markers are attributed either to chance or to selection. Inferring selection was not the aim of the present study; however, the homogeneous geographic distribution of haplotypes **1* and **2* could be regarded as the result of a long-term 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 Φ_{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). Φ_{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 *becaus*e, 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.

Fig. 4



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.

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 ultarapid 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.

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.

2 Bradford LD. CYP2D6 allele frequency in European Caucasians, Asians, Africans and their descendants. *Pharmacogenomics* 2002; **3**:229–243.

3 Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, *et al.* The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med* 2003; **348**:1170–1175.

4 Cooper RS, Kaufman JS, Ward R. Race and genomics. *N Engl J Med* 2003; **348**:1166–1170.

5 Kittles RA, Weiss KM. Race, ancestry, and genes: implications for defining disease risk. *Annu Rev Genomics Hum Genet* 2003; **4**:33–67.

6 Jorde LB, Wooding SP. Genetic variation, classification and 'race'. *Nat Genet* 2004; **36**:S28–S33.

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

8 Barbujani G. Human races: classifying people vs. understanding diversity. *Curr Genom* 2005; **6**:215–226.

9 Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, *et al.* Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res* 2002; **12**:602–612.

10 Serre D, Paabo S. Evidence for gradients of human genetic diversity within and among continents. *Genome Res* 2004; **14**:1679–1685.

11 Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, *et al.* Genetic structure of human populations. *Science* 2002; **298**:2381–2385.

12 Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB. Human population genetic structure and inference of group membership. *Am J Hum Genet* 2003; **72**:578–589.

13 Tang H, Quertermous T, Rodriguez B, Kardia SL, Zhu X, Brown A, *et al.* Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am J Hum Genet* 2005; **76**:268–275.

14 Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, *et al.* Population genetic structure of variable drug response. *Nat Genet* 2001; **29**:265–269.

15 Risch N, Burchard E, Ziv E, Tang H. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol* 2002; **3**:comment 2007.

16 Tate SK, Goldstein DB. Will tomorrow's medicines work for everyone? *Nat Genet* 2004; **36**:S34–S42.

17 Bloche MG. Race-based therapeutics. *N Engl J Med* 2004; **351**:2035–2037.

18 Taylor AL, Ziesche S, Yancy C, Carson P, D'Agostino R Jr, Ferdinand K, *et al.* Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *N Engl J Med* 2004; **351**:2049–2057.

19 Rahemtulla T, Bhopal R. Pharmacogenetics and ethnically targeted therapies. *BMJ* 2005; **330**:1036–1037.

20 Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, *et al.* A human genome diversity cell line panel. *Science* 2002; **296**:261–262.

21 Sistonen J, Fuselli S, Levo A, Sajantila A. CYP2D6 genotyping by a multiplex primer extension reaction. *Clin Chem* 2005; **51**:1291–1295.

22 Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**:978–989.

23 Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003; **73**:1162–1169.

24 Lewontin RC. The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics* 1964; **49**:49–67.

25 Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet* 1968; **38**:226–231.

26 Goddard KA, Hopkins PJ, Hall JM, Witte JS. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am J Hum Genet* 2000; **66**:216–234.

27 Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 2003; **19**:2496–2497.

28 Clement M, Posada D, Crandall KA. TCS: a computer program to estimate gene genealogies. *Mol Ecol* 2000; **9**:1657–1659.

29 Zanger UM, Raimundo S, Eichelbaum M. Cytochrome P450 2D6: overview and update on pharmacology, genetics, biochemistry. *Naunyn Schmiedebergs Arch Pharmacol* 2004; **369**:23–37.

30 Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 1992; **131**:479–491.

31 Schneider S, Roessli D, Excoffier L. *Arlequin, v 2.000: a software for population genetics data analysis*. Geneva: Genetics and Biometry Laboratory, University of Geneva; 2000.

32 Rosenberg MS. *PASSAGE: Pattern analysis, spatial statistics and geographic exegesis. Version 1.0*. Tempe, AZ: Department of Biology, Arizona State University; 2001.

33 Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 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* 2005; **102**:15942–15947.

34 Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Res* 1967; **27**:209–220.

35 Aklillu E, Persson I, Bertilsson L, Johansson I, Rodrigues F, Ingelman-Sundberg M. Frequent distribution of ultrarapid metabolizers of debrisoquine in an Ethiopian population carrying duplicated and multiduplicated functional CYP2D6 alleles. *J Pharmacol Exp Ther* 1996; **278**:441–446.

36 McLellan RA, Oscarson M, Seidegard J, Evans DA, Ingelman-Sundberg M. Frequent occurrence of CYP2D6 gene duplication in Saudi Arabians. *Pharmacogenetics* 1997; **7**:187–191.

37 Fuselli S, Dupanloup I, Frigato E, Cruciani F, Scozzari R, Moral P, *et al.* Molecular diversity at the CYP2D6 locus in the Mediterranean region. *Eur J Hum Genet* 2004; **12**:916–924.

38 Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. An apportionment of human DNA diversity. *Proc Natl Acad Sci USA* 1997; **94**:4516–4519.

39 Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, *et al.* The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am J Hum Genet* 2000; **66**:979–988.

40 Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, *et al.* The structure of haplotype blocks in the human genome. *Science* 2002; **296**:2225–2229.

41 Tishkoff SA, Verrelli BC. Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. *Curr Opin Genet Dev* 2003; **13**:569–575.

42 Cavalli-Sforza LL, Menozzi P, Piazza A. *The history and geography of human genes*. Princeton, New Jersey: Princeton University Press; 1994.

43 Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, *et al.* A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am J Hum Genet* 1998; **62**:1389–1402.

44 Corander J, Waldmann P, Sillanpaa MJ. Bayesian analysis of genetic differentiation between populations. *Genetics* 2003; **163**:367–374.

45 Corander J, Waldmann P, Marttinen P, Sillanpaa MJ. BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* 2004; **20**:2363–2369.

46 Sallee FR, DeVane CL, Ferrell RE. Fluoxetine-related death in a child with cytochrome P-450 2D6 genetic deficiency. *J Child Adolesc Psychopharmacol* 2000; **10**:27–34.

47 Koski A, Ojanperä I, Vuori E, Sistonen J, Sajantila A. A fatal doxepin poisoning associated with a defective CYP2D6 genotype. *Am J Foren Med Path* in press.

48 Gasche Y, Daali Y, Fathi M, Chiappe A, Cottini S, Dayer P, *et al.* Codeine intoxication associated with ultrarapid CYP2D6 metabolism. *N Engl J Med* 2004; **351**:2827–2831.

49 Suarez-Kurtz G. Pharmacogenomics in admixed populations. *Trends Pharmacol Sci* 2005; **26**:196–201.

50 Weinshilboum R. Inheritance and drug response. *N Engl J Med* 2003; **348**:529–537.

51 Evans WE, Relling MV. Moving towards individualized medicine with pharmacogenomics. *Nature* 2004; **429**:464–468.