European Journal of Human Genetics (2004) 12, 916–924
© 2004 Nature Publishing Group All rights reserved 1018-4813/04 $30.00
www.nature.com/ejhg

ARTICLE

Molecular diversity at the CYP2D6 locus in the Mediterranean region

Silvia Fuselli[1], Isabelle Dupanloup[1,5], Elena Frigato[1], Fulvio Cruciani[2], Rosaria Scozzari[2], Pedro Moral[3], Johanna Sistonen[4], Antti Sajantila[4] and Guido Barbujani*[,1]

[1]*Department of Biology, University of Ferrara, via Borsari 46, 44100 Ferrara, Italy;*[2]*Department of Genetics and Molecular Biology, University of Rome ''La Sapienza'', P.le Aldo Moro, 5, 00185 Roma, Italy;*[3]*Department of Animal Biology, Faculty of Biology, University of Barcelona, Avinguda Diagonal, Barcelona, Spain;*[4]*Department of Forensic Medicine, University of Helsinki, P.O. Box 40, 00014 Helsinki, Finland;*[5]*Centre for Integrative Genomics, Université de Lausanne, CH-1015 Lausanne, Switzerland*

Despite the importance of cytochrome *P*450 in the metabolism of many drugs, several aspects of molecular variation at one of the main loci coding for it, CYP2D6, have never been analysed so far. Here we show that it is possible to rapidly and efficiently genotype the main European allelic variants at this locus by a SNaPshot method identifying chromosomal rearrangements and nine single-nucleotide polymorphisms. Haplotypes could be reconstructed from data on 494 chromosomes in six populations of the Mediterranean region. High levels of linkage disequilibrium were found within the chromosome region screened, suggesting that CYP2D6 may be part of a genomic recombination block, and hence that, aside from unequal crossingover that led to large chromosomal rearrangements, its haplotype diversity essentially originated through the accumulation of mutations. With the only, albeit statistically insignificant, exception of Syria, haplotype frequencies do not differ among the populations studied, despite the presence among them of three well-known genetic outliers, which could be the result of common selective pressures playing a role in shaping CYP2D6 variation over the area of Europe that we surveyed.

European Journal of Human Genetics (2004) **12,** 916–924. doi:10.1038/sj.ejhg.5201243
Published online 1 September 2004

Keywords: genetic diversity; single-nucleotide polymorphisms; SNaPshot genotyping; linkage disequilibrium; mutation; selection

Introduction

The human genome includes at least 57 genes coding for cytochrome *P*450 proteins, and 29 pseudo-genes.[1] Among them, CYP2D6 (OMIM 124030) has a crucial role in the metabolism of over 40 drugs, including *β*-adrenergic blocking agents, antiarrhythmics, antipsychotics, antidepressants, and narcotic analgesics. Genetic variation at *P*450-CYP2D6 causes differences in the catalytic activity of the enzyme. In studies based on the metabolism of probe drugs, several phenotypic classes are defined. Individuals defined as poor metabolizers (PM) generally carry two nonfunctional CYP2D6 alleles, while the presence of one functional allele is generally sufficient to determine the extensive metabolizer (EM) phenotype. In addition, individuals carrying chromosomal rearrangements with two or more active copies of the CYP2D6 gene tend to be classified as ultrarapid metabolizers (UM). However, the limits among phenotypic classes are uncertain (e.g. several authors recognize an additional class of intermediate metabolizers, IM,[2] and the continuous distribution of phenotypes suggests interaction with other genes and/or

*Correspondence: Dr G Barbujani, Department of Biology, University of Ferrara, via Borsari 46, 44100 Ferrara, Italy. Tel: +39 0532 291312; Fax: +39 0532 29761; E-mail: g.barbujani@unife.it

Received 30 January 2004; revised 13 April 2004; accepted 12 May 2004

more complex interactions among alleles than simple dominance (see Figure 1 of Bertilsson *et al*[3]). Therefore, the relationship between metabolic ratios (ie the phenotypes) and the genotypes of CYP2D6 is far from clarified.

The CYP2D6 gene is located on 22q13.1, at the 3′-end of the CYP2D cluster, downstream of the CYP2D8P and CYP2D7P pseudogenes.[4,5] To date, more than 70 different variants have been described (www.imm.ki.se/CYPalleles/cyp2d6), differing for single-base changes, short insertions and deletions, or for major rearrangements such as deletion[6] and duplications[7] of the whole gene. For the sake of clarity, in what follows we shall refer to these variants as *haplotypes*, whereas we shall use the term *alleles* only to refer to the alternative nucleotides at polymorphic DNA sites. Therefore, in this study we shall call *haplotypes* the genetic variants of CYP2D6 that are called *alleles* in most previous studies. In other words, here haplotype means a set of alleles transmitted together. The information about their association on a chromosome is called their *phase*.

Genetic variation at CYP2D6 is high, both among populations and among individuals of the same population.[8] Two common haplotypes, *1 and *2, represent each between 7 and 40% of the gene pool in most world populations, and haplotype *5, the whole gene deletion, between 1 and 5%. Other defective haplotypes seem to show continental specificities, although no continent so far has been studied in sufficient detail to allow robust conclusions. Haplotype *4 represents 15–20% of the European gene pool,[2,9] haplotype *10 has a high frequency, up to 65% in Asia,[10] haplotype *17 is present at frequencies around 20–30% in Africa,[11] and all are rare or absent elsewhere in the world. Haplotype *4 is a loss-of-function haplotype, whereas haplotypes *10 and *17 confer a decreased enzyme activity.[12,13] The high frequency of these haplotypes in certain populations is hard to account for under a simple mutation-selection model. Haplotypes containing multiple gene copies were found at high frequencies only in specific regions like Ethiopia (29%)[14] and Saudi Arabia (21%),[15] while their frequency in the rest of the world ranges between 1 and 3%.

In this study, we genotyped CYP2D6 in six populations of the Mediterranean region, defining the haplotype phases from genotype data by computational methods. We further analysed these data, looking for levels of linkage disequilibrium (LD), trying to establish the evolutionary relationships among haplotypes, and testing for population differentiation among six European and Near-Eastern population samples. The shape of the evolutionary tree, and the similarities observed among populations living at large geographic distances, suggest that selective pressures have been important in determining variation at this locus.

Materials and methods

Sampling

We genotyped nine SNPs in the coding region of the CYP2D6 locus (Figure 1a) in 247 individuals from six populations sampled in the course of a project aimed at describing genetic diversity around the Mediterranean sea. We had available DNA from 51 Syrians,[16] 28 Ladin speakers from the Italian Alps,[17] 51 Southern Spaniards, 38 Basques, 48 Sardinians, and 31 Central Italians. All of them gave their informed consensus to the authors of the original studies for which the blood samples were collected. Basques, Ladin speakers, and Sardinians are known to be among the most genetically divergent European populations for several loci.[18,19] For each individual, we also tested for the presence of gene duplication or deletion of the whole coding region.

Genotyping

We used a combination of long polymerase chain reactions (PCRs) and a multiplex single-base extension system to detect gene deletion and duplication, and to characterize

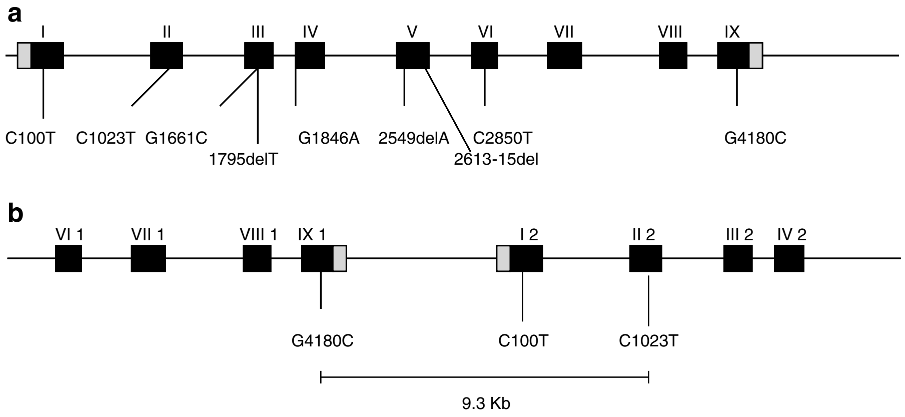

Figure 1 Structure of the CYP2D6 gene (**a**), and schematic representation of a region of chromosome 22 with a duplicated CYP2D6 (**b**). Black boxes are exons.

nine variable positions (Figure 1a) in the coding region of CYP2D6. In this way, we could identify the mutations which are generally supposed to define[20,21] more than 90% of the total known variants among Europeans. Mutation 1023 C>T is rare in Europe, but it was also considered in the screening because it is known to be common in African populations.

The genotyping method we summarize here will be described in greater detail in a forthcoming paper (Sistonen *et al.*, submitted). Three parallel long PCRs were run for each sample. We looked for CYP2D6 gene duplication using the primer pair pCYP-207-F and pCYP-32-R.[22] A control band was obtained adding the pCYP-13-F forward primer[6] to the reaction mix. To detect the presence of the entire gene deletion, having at the same time an internal control band, we used again a set of three different primers: pCYP-13-F/pCYP-207-F/ pCYP-24-R[6,22] (Figure 2).

We obtained a 5.1 kb fragment containing all nine CYP2D6 exons with primers CYP2D6-F and CYP2D6-R.[23] This product was used as a template to type 9 positions in one reaction, based on single-base primer extension with fluorescent labelled ddNTPs (ABI Prism SNaPshot Multiplex Kit, Applied Biosystems). The SNaPshot results were double-checked in the case of one (Basque) individual carrying the new haplotype *4Bas, using a combination of nested PCR and restriction fragment length polymorphism (RFLP) as described in Sachse *et al.*[20]

An additional PCR-SNaPshot assay was used to distinguish between different types of allele duplications. We amplified a 9.3 kb fragment spanning the genomic region between exon 9 and intron 2 of two subsequent CYP2D6 genes (Figure 1b), using primers P2x2F and P2x2R.[24] The 9.3 kb fragment was used as a template in a specific SNaPshot reaction for exon 9, 1 and 2 polymorphic positions.

Haplotype determination

As a preliminary step, we inferred the haplotypes from SNP data using a PHASE reconstruction method.[25] PHASE is a Bayesian statistical method for haplotype reconstruction, which incorporates the prior assumption that unresolved haplotypes will tend to be similar to known haplotypes, observed in unambiguous genotypes. PHASE does not rely on other assumptions, such as Hardy–Weinberg equilibrium. Computer simulations studies[25,26] clearly suggest that PHASE performs better than other algorithms, such as EM, in a wide range of conditions.

Hardy–Weinberg equilibrium

Departures from Hardy–Weinberg equilibrium were assessed in each sample, using the exact Fisher test implemented in the Arlequin package.[27] The test was performed for each SNP separately, as well as for the haplotypes defined by PHASE.

Linkage disequilibrium

Once the haplotypes had been defined by PHASE, we wanted to know whether recombination has played a significant role in shaping CYP2D6 variation in the sampled populations. For that purpose, we tested for LD between each pair of SNPs in each sample. The significance of allelic association was evaluated by means of the extension of Fisher's exact test on contingency tables implemented in the Arlequin software,[27] and using Bonferroni correction for multiple comparisons.[28] Further, the D' statistic,[29] measuring the degree of association between SNP loci, was computed in each sample.

An evolutionary tree of haplotypes

The relationships among the haplotypes defined by the PHASE algorithm were summarized by a statistical

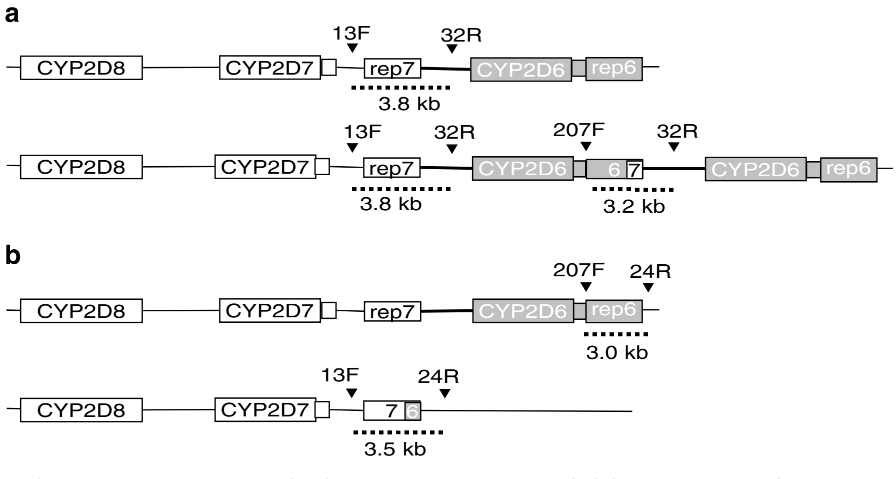

Figure 2 A schematic description of CYP2D6 duplication-specific (**a**) and deletion-specific (**b**) genotyping reactions. Arrows correspond to the positions of primers, which are described in the text.

parsimony tree or cladogram[30] using the software TCS.[31] With this method, a tree is constructed by connecting haplotypes when the probability that there is no intermediate haplotype in the sample is greater than 0.95. The tree obtained in this way does not have a root, that is, it is actually a network. One of its advantages, over strictly bifurcating trees, is that it allows one to identify, through the presence of loops in the reconstructed phylogeny, possible episodes of recombination and homoplasies.

Intra-population diversity indices

We calculated in our samples three summary indices of diversity, by means of the Arlequin software[27]: (1) gene diversity, that is, the probability that two randomly chosen haplotypes are different in the sample; (2) average gene diversity over the polymorphic sites, that is, the probability that two randomly chosen nucleotides in the sample at the same site are different, and (3) the mean number of pairwise differences between all pairs of haplotypes, π.

For comparison, we estimated these statistics in two additional, African, samples, one from Ghana (sample size: 193)[11] and another from Tanzania (sample size: 106).[32] The haplotypes were reconstructed in these samples from SNP data, according to the Human Cytochrome *P*450 (CYP) Allele Nomenclature (www.imm.ki.se/CYPalleles/cyp2d6).

Population differentiation

An analysis of molecular variance (AMOVA)[33] was used to estimate how genetic variation is partitioned among population and individuals. Genetic distances (Φst) were also estimated between populations using the Arlequin software.[27] The statistical significance of the variance components and of the genetic distances was evaluated by permuting haplotypes among populations.

Results

Haplotype determination

We carried out our survey using a SNaPshot test designed to determine the individual genotype at the nine most common polymorphic positions. Like all other SNP-based approaches, this genotyping procedure does not provide the haplotype phase, which can be directly determined only by using other labour-intensive approaches. Therefore, we inferred the haplotype phases from genotype data by a computational method. Table 1 shows the association of the SNPs into haplotypes. The haplotypes inferred in this way correspond exactly to those empirically proposed by previous studies (www.imm.ki.se/CYPalleles/cyp2d6). In addition, among the Basques, we identified for the first time a haplotype that could be considered a new variant of *4 (1661G>C, 1846G>A, and 4180G>C), bearing the splice site mutation responsible for loss of activity, but without the substitution 100 C>T, usually associated with a transition at 1846. We further confirmed this result by nested PCR followed by RFLP analysis for the four aforementioned SNPs.[20]

Haplotype and genotype frequencies, and Hardy–Weinberg equilibrium

The frequencies of 12 haplotypes are given in Table 2. Genotype frequencies are available upon request. Four genotypes, *1/*1, *1/*2, *2/*2, and *1/*4, account for more than half of the total in each population; none of the populations of this study showed significant departures from the Hardy–Weinberg equilibrium expectations. Also present were all the haplotypes reported to be associated with lower or null metabolic activity that could be identified by our SNaPshot test (*3, *6, *9, *10, and *17). Their frequencies (Table 2) are in good agreement with previous studies on Europe.[2,9,20] One individual from Andalusia was found to be a carrier of the typically African haplotype *17. The high frequency of haplotypes with duplications in Syria (*1 × N + *2 × N = 7.8%) agrees with previous observations of high frequencies of duplications in the Near East and in Africa.[14,15]

Table 1 SNP positions and frequencies of haplotypes defined by the PHASE algorithm in the whole set of samples

	100[a]	*1023*	*1661*	*1707*	*1846*	*2549*	*2613-15*	*2850*	*4180*	
Haplotype	*C>T*	*C>T*	*G>C*	*T>G*	*G>A*	*delA*	*delAGA*	*C>T*	*G>C*	*Estimated frequency*
*1[b]	0	0	0	0	0	0	0	0	0	40.8
*2[b]	0	0	1	0	0	0	0	1	1	37.1
*3	0	0	0	0	0	1	0	0	0	1.5
*4[b]	1	0	1	0	1	0	0	0	1	15.8
*4Bas	0	0	1	0	1	0	0	0	1	0.2
*6	0	0	0	1	0	0	0	0	0	0.8
*9	0	0	0	0	0	0	1	0	0	0.6
*10	1	0	1	0	0	0	0	0	1	2.9
*17	0	1	1	0	0	0	0	1	1	0.2

[a] Nucleotide numbers are given considering +1 the A of ATG-translation initiation codon.
[b] Including *1 × N, *2 × N, and *4 × N (duplication or multiduplication of haplotype *1, *2, *4 respectively).

Linkage disequilibrium

We measured LD between pairs of SNPs distributed across the 4.1 kb genomic region we screened. However, sites with minor allele frequency <0.1 were excluded, because for them the power of the methods to detect LD is severely reduced.[34] Gene duplications and deletions were excluded from this analysis. Both in the analysis of each single population, and in their joint analysis, the measure of LD was the highest observable ($|D'|=1$) and remained significant after the Bonferroni correction ($P<0.001$) between all pairs of alleles.

Simple calculations on the number of haplotypes (Table 3) confirm a strong association among mutations. If the only evolutionary force generating new variants from one founder haplotype were the process of mutation, one would expect to observe up to $n+1$ haplotypes, where n is the number of SNPs considered. On the contrary, if mutations were reassorted in all possible manners by recombination, one would expect to observe 2^n different haplotypes. Table 3 shows that the total number of haplotypes is less than $n+1$ in each of the six sampled populations. Therefore, both tests of LD, and counts of the numbers of different haplotypes, suggest that recurrent recombination has not played a significant role in shaping CYP2D6 variation at SNP sites. This result is in agreement with the identification of an LD block across a 390 kb region spanning CYP2D6,[35] and allows us to treat the DNA region of interest as a single non-recombining genomic block[36] in the subsequent tests.

Evolutionary trees of haplotypes

The network illustrating the relationships among the haplotypes found in our six populations is shown in Figure 3. As there is no information yet on CYP2D6 in the closest evolutionary relatives of humans, the great apes, it was impossible to root this network, and hence we have no objective basis to define the historical sequence of mutations that led to the current diversity. Three clusters are apparent in the network. The first cluster corresponds to haplotype *1 and its presumably derived (because rare) variants, namely the two nonfunctional haplotypes *3 and *6 and the reduced-function haplotype *9. The second cluster is mainly represented by haplotype *2. The 1023 C>T mutation-step leads to *17, known to be responsible

Table 2 Frequencies of observed CYP2D6 haplotypes in six European populations

Haplotype	*Southern Spaniards* *% (n = 102)*	*Basques* *% (n = 76)*	*Sardinians* *% (n = 96)*	*Central Italians* *% (n = 62)*	*Alps* *% (n = 56)*	*Syrians* *% (n = 102)*	*Enzyme activity*
*1	41.18	40.79	36.46	33.87	33.93	47.06	Normal
*2	35.29	32.89	40.63	35.48	32.14	30.39	Normal
*3	0.00	0.00	3.13	0.00	7.14	0.00	No
*4	17.65	21.05[a]	12.50	12.90	19.64	9.80	No
*5	1.96	2.63	1.04	0.00	1.79	0.98	No
*6	0.00	1.32	0.00	3.23	0.00	0.98	No
*9	0.98	0.00	0.00	1.61	1.79	0.00	Decr
*10	0.98	1.32	4.17	8.06	0.00	2.94	Decr
*17	0.98	0.00	0.00	0.00	0.00	0.00	Decr
*1 × N[b]	0.00	0.00	0.00	0.00	1.79	3.92	Incr
*2 × N[c]	0.98	0.00	2.08	3.23	0.00	3.92	Incr
*4 × N[d]	0.00	0.00	0.00	1.61	1.79	0.00	No
Total	100.00	100.00	100.00	100.00	100.00	100.00	

[a] Including a new variant of haplotype *4 (1661G>C, 1846G>A and 4180G>C).
[b,c,d] Duplication or multiduplication of haplotype *1, *2, *4, respectively.

Table 3 Observed and expected numbers of different haplotypes by population

	Chromosomes[a]	*Pol sites*[b]	*Observed SNP haplotypes*	*n+1*[c]	*Theor max*[d]
Southern Spaniards	100	7	6	8	128
Basques	74	6	6	7	64
Sardinians	95	6	5	7	64
Central Italians	62	7	6	8	128
Alps	55	7	5	8	128
Syrians	101	6	5	7	64

[a] The chromosomes with the whole gene deletion were excluded from this table, and so the number of chromosomes is not necessarily twice the number of individuals in the population sample.
[b] Referred to the nine SNPs of Table 1.
[c] n is the total number of SNPs considered.
[d] Theoretical maximum number of haplotypes = 2^n, n is the total number of SNPs considered.

for the expression of an enzyme with altered substrate affinity.[13] The branch topology of the *10–*4 cluster is not totally resolved, as shown by the presence of a closed loop.

Population diversity

Measures of genetic diversity within the six populations of this study were compared with those estimated from two African populations (Table 4). No Mann–Whitney non-parametric test on the six statistics estimated reached significance. Therefore, contrary to what is observed for the generality of nuclear markers,[37] we found no evidence of excess diversity in Africa, regardless of whether or not major gene rearrangements (deletions and duplications) were considered. Actually, when deletions and duplications were excluded from the analyses, the African samples tended to show lower internal diversity than the samples of this study. In principle, this result might mean that CYP2D6 is exceptionally variable in Europe with respect to most other genome markers, but probably it reflects an inadequate characterization of the African samples. In other words, variation in Africa may have been underestimated, because the sites typed in these populations, with the exception of the 1023 C>T transition, are sites that were found to be polymorphic in European subjects.[9] This interpretation is also supported by the poor correlation observed among Africans between CYP2D6 haplotypes and phenotypes, suggesting that the nucleotide substitutions affecting gene expression have not been thoroughly described in those samples.[11,38]

The AMOVA analysis in the six samples of this study revealed that within-population variation, that is, genetic differences between individuals within populations, account for 100% of total genetic variation (data not given). Therefore, the Φ_{ST} distances between the sampled populations are essentially 0. Nevertheless, the Syrian sample shows some increased genetic divergence from the other populations. Aklillu *et al*[39] proposed that selective pressures associated with diet may have favoured the spread of duplicated or anyway very active alleles in East Africa. The high frequency of duplicated haplotypes in Syria (as well as in Arabia[15]) may be due to the same phenomenon.

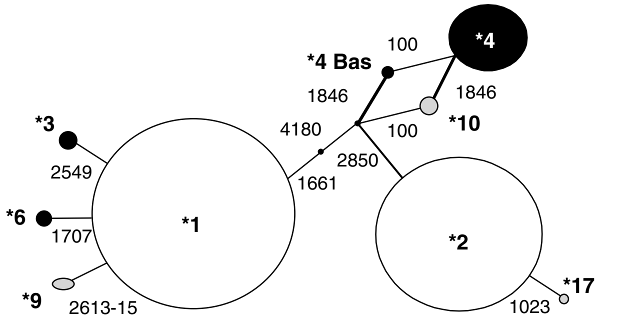

Figure 3 Tree of haplotypes inferred from the analysis of nine SNPs. Figures on each branch indicate the mutated site for which two haplotypes differ. The size of the circles is proportional to the haplotype frequency in the samples of this study. Full-function, decreased-function and null-function haplotypes are, respectively, in white, grey and black.

Discussion

In this study, we typed nine SNPs at the CYP2D6 locus in nearly 500 chromosomes from six populations of the Mediterranean region. The haplotypes correspond to those commonly found using more complex genotyping techniques (http://www.imm.ki.se/CYPalleles/cyp2d6.htm), confirming the association on the chromosomes of alleles in haplotypes that can be recognized on the basis of a small number of key mutations. In addition, we described among the Basques a previously unknown haplotype, which we provisionally label *4Bas, because of the presence of the substitution in position 1846, which characterises all *4 haplotypes.

LD statistics reached the highest possible values among all suitable mutations, an indication, although not proof, that CYP2D6 may be transmitted as a single block of DNA,

Table 4 Measures of internal genetic diversity in two African populations, and in the six populations of this study

	Ghana[11]	*Tanzania*[32]	*South Spain*	*Basques*	*Sardinians*	*Central Italians*	*Alps*	*Syrians*
Individuals	193	106	51	38	48	31	28	51
Gene copies	386	212	102	76	96	62	56	102
NH	7	8	8	7	7	7	8	8
NS	14	11	11	11	11	11	11	11
S (Pol sites)	8	9	9	7	8	7	9	8
Gene diversity	0.71±0.01	0.81±0.01	0.68±0.02	0.69±0.03	0.69± 0.03	0.74±0.03	0.75±0.03	0.68±0.03
AGD	0.19±0.12	0.19±0.12	0.20±0.12	0.20±0.12	0.19±0.12	0.20±0.13	0.22±0.13	0.19±0.12
π	2.31±1.27	2.13±1.19	2.16±1.21	2.20±1.23	2.09±1.18	2.24±1.25	2.42±1.33	2.06±1.16

Deletions and duplications were considered along with nucleotide substitutions.
NH: number of different haplotypes; NS: total number of sites being characterized; AGD: average gene diversity over the polymorphic sites; π: average pairwise sequence difference (average mismatch).

not disrupted by recombination. That finding has significant practical implications. In general, the human genome is believed to be largely composed of blocks (around 22 kb in size in European populations), within which recombination is scarce or absent, separated by hotspots of meiotic recombination[36] (for a partly different view, see Anderson and Slatkin[40]). In agreement with previous work suggesting the existence of a large LD block covering the CYP2D cluster,[35] our results show that, for most practical purposes, the CYP2D6 gene may be treated as a nonrecombining unit in the genome. That finding, if confirmed, would provide what seems a simple way to rapidly characterize a rather large genome region, thus increasing the power of association tests.[41]

Absence of inferred recombinant gametes also simplified the task to establish the evolutionary relationships among alleles, which we represented in tree form. The obvious next question to ask is whether this variation reflects the chance effects of mutation and drift, or whether sequence changes are reflecting some sort of selective pressures. To address this question, it will be necessary to sequence the entire gene in different populations, looking for the excess of intermediate-frequency haplotypes which typically reflect balancing selection.[42] That balancing selection may be contributing to shaping diversity at CYP2D6 is suggested by the fact that subfunctional, and even non-functional, haplotypes such as *4 occur at high frequencies, apparently inconsistent with a high selection coefficient against them. At this stage, one can only speculate that protein variants that might interact with a broader range of xenobiotics, or with classes of xenobiotics that are not commonly metabolized by the common haplotypes, would have an obvious evolutionary advantage.

The six populations of the Mediterranean shores we studied do not appear genetically differentiated, although Syria, a geographic outlier, also shows a higher (if insignificant) frequency of gene duplications. Garte *et al*[43] observed little genetic heterogeneity among Europeans for many genes involved in drug metabolism, including NAT2, CYP1A1, and GSTM1. The GSTT1 locus seemed to be the only exception, whereas CYP2D6 variation was not analysed because of the inconsistency of data sets available in the literature, mainly due to the different laboratory typing methods. In this paper, we used consistently the same method for genotyping all populations, so that we could examine the hypothesis of a European genetic homogeneity at the CYP2D6 locus too. An unexpected finding was that genetic differences are minimal at CYP2D6, despite the inclusion in our study of three of the main genetic outliers of Europe, namely the Basques, the Sardinians, and the populations of the Eastern Italian Alps. These populations are known to differ substantially from their neighbours, both at autosomal and at uniparentally transmitted loci,[17,44–46] which is largely regarded as an effect of reproductive isolation, caused by both geographic and linguistic barriers to gene flow.[47,48] However, a thorough study of autosomal SNPs in these populations has not been carried out yet. Therefore, the possibility should be considered, at least in principle, that most autosomal SNP loci are poorly differentiated in Europe, and that CYP2D6 be no exception. Alternatively, should greater differences emerge among Europeans for other autosomal SNPs (possibly reflecting population divergence through independent random genetic drift[18]), one could speculate that European populations underwent the same selective pressures for loci of pharmacogenetic interest, so that their CYP2D6 haplotype frequencies, as well as those of NAT2, CYP1A1, and GSTM1, are similar.

To further progress in our understanding of the origins and maintenance of diversity at CYP2D6, a crucial step will be the sequencing of the gene in some primate species. Only in this way shall we be able to identify the direction of the evolutionary change, thus estimating the age of the observed alleles, and hence potentially correlating those ages with the dates at which the main human demographic transitions took place. In parallel, identifying the ancestral alleles is crucial to understanding whether current diversity evolved from a fully functional haplotype, or if more complicated phenomena occurred. However, the results of this study already indicate that: (1) substitutions altering the amino-acid sequence (100 C>T, 1707 T>G, 1846 G>A, 2549 delA, 2613-15 delAGA, 2850 C>T, 4180 G>C) have been maintained within coding regions of the gene; (2) that these mutations occurred early enough in evolution to be shared by most or all Mediterranean populations; and (3) that, despite the evolutionary disadvantage associated with nonfunctional (*3, *4, *5, *6) and subfunctional (*9, *10, *17) haplotypes, these haplotypes are maintained at subpolymorphic, and even polymorphic (haplotype *4, reaching a frequency >0.20 among the Basques) frequencies in the populations.

Only a study of the DNA sequences, and a sound understanding of the relationships between genotypes and phenotypes, will tell us what sort of evolutionary pressures underlies this pattern of diversity. Traditionally, defective or abnormal genotypes are identified in individuals showing an abnormal ability to metabolize one molecule, often sparteine or debrisoquine. However, it is known that variants of the *P*450 protein with reduced ability to metabolize a certain class of molecules may have an increased ability to metabolize other molecules,[49] which may easily pass undetected if the starting point of the analysis is the phenotype. The only way to address this problem is to reverse the approach, namely to start from a genetic characterization of the subjects, and to associate to each genotype the distribution of phenotypic values, represented by data on the metabolism of several drugs. This approach, if successful, may pave the ground for an individual-specific pharmaceutical treatment.

Acknowledgements

We are grateful to Giorgio Bertorelle, Francesca Visentini, and Eduardo Tarazona-Santos for several comments and suggestions. This study was supported by the FISR fund of the Italian Ministry of the Universities, and by funds provided by the Universities of Ferrara, Helsinki, and Rome 'La Sapienza'.

References

1 Danielson PB: The cytochrome *P*450 superfamily: biochemistry, evolution and drug metabolism in humans. *Curr Drug Metab* 2002; **3**: 561–597.

2 Griese EU, Zanger UM, Brudermanns U *et al*: Assessment of the predictive power of genotypes for the *in-vivo* catalytic function of CYP2D6 in a German population. *Pharmacogenetics* 1998; **8**: 15–26.

3 Bertilsson L, Dahl ML, Dalen P, Al-Shurbaji A: Molecular genetics of CYP2D6: clinical relevance with focus on psychotropic drugs. *Br J Clin Pharmacol* 2002; **53**: 111–122.

4 Kimura S, Umeno M, Skoda RC *et al*: The human debrisoquine 4-hydroxylase (CYP2D) locus: sequence and identification of the polymorphic CYP2D6 gene, a related gene, and a pseudogene. *Am J Hum Genet* 1989; **45**: 889–904.

5 Heim MH, Meyer UA: Evolution of a highly polymorphic human cytochrome *P*450 gene cluster: CYP2D6. *Genomics* 1992; **14**: 49–58.

6 Steen VM, Andreassen OA, Daly AK *et al*: Detection of the poor metabolizer-associated CYP2D6(D) gene deletion allele by long-PCR technology. *Pharmacogenetics* 1995; **5**: 215–223.

7 Johansson I, Lundqvist E, Bertilsson L, Dahl ML, Sjoqvist F, Ingelman-Sundberg M: Inherited amplification of an active gene in the cytochrome *P*450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. *Proc Natl Acad Sci USA* 1993; **90**: 11825–11829.

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

9 Marez D, Legrand M, Sabbagh N *et al*: Polymorphism of the cytochrome *P*450 CYP2D6 gene in a European population: characterization of 48 mutations and 53 alleles, their frequencies and evolution. *Pharmacogenetics* 1997; **7**: 193–202.

10 Garcia-Barcelo M, Chow LY, Chiu HF *et al*: Genetic analysis of the CYP2D6 locus in a Hong Kong Chinese population. *Clin Chem* 2000; **46**: 18–23.

11 Griese EU, Asante-Poku S, Ofori-Adjei D, Mikus G, Eichelbaum M: Analysis of the CYP2D6 gene mutations and their consequences for enzyme function in a West African population. *Pharmacogenetics* 1999; **9**: 715–723.

12 Johansson I, Oscarson M, Yue QY, Bertilsson L, Sjoqvist F, Ingelman-Sundberg M: Genetic analysis of the Chinese cytochrome *P*4502D locus: characterization of variant CYP2D6 genes present in subjects with diminished capacity for debrisoquine hydroxylation. *Mol Pharmacol* 1994; **46**: 452–459.

13 Bapiro TE, Hasler JA, Ridderstrom M, Masimirembwa CM: The molecular and enzyme kinetic basis for the diminished activity of the cytochrome *P*450 2D6.17 (CYP2D6.17) variant. Potential implications for CYP2D6 phenotyping studies and the clinical use of CYP2D6 substrate drugs in some African populations. *Biochem Pharmacol* 2002; **64**: 1387–1398.

14 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.

15 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.

16 Vernesi C, Di Benedetto G, Caramelli D *et al*: Genetic characterization of the body attributed to the evangelist Luke. *Proc Natl Acad Sci USA* 2001; **98**: 13460–13463.

17 Vernesi C, Fuselli S, Castri L, Bertorelle G, Barbujani G: Mitochondrial diversity in linguistic isolates of the Alps: a reappraisal. *Hum Biol* 2002; **74**: 725–730.

18 Cavalli-Sforza LL, Menozzi P, Piazza A: *The History and Geography of Human Genes*. Princeton: Princeton University Press, 1994.

19 Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G: Geographic patterns of mtDNA diversity in Europe. *Am J Hum Genet* 2000; **66**: 262–278.

20 Sachse C, Brockmoller J, Bauer S, Roots I: Cytochrome *P*450 2D6 variants in a Caucasian population: allele frequencies and phenotypic consequences. *Am J Hum Genet* 1997; **60**: 284–295.

21 Sachse C, Brockmoller J, Hildebrand M, Muller K, Roots I: Correctness of prediction of the CYP2D6 phenotype confirmed by genotyping 47 intermediate and poor metabolizers of debrisoquine. *Pharmacogenetics* 1998; **8**: 181–185.

22 Lovlie R, Daly AK, Molven A, Idle JR, Steen VM: Ultrarapid metabolizers of debrisoquine: characterization and PCR-based detection of alleles with duplication of the CYP2D6 gene. *FEBS Lett* 1996; **392**: 30–34.

23 Lundqvist E, Johansson I, Ingelman-Sundberg M: Genetic mechanisms for duplication and multiduplication of the human CYP2D6 gene and methods for detection of duplicated CYP2D6 genes. *Gene* 1999; **226**: 327–338.

24 Johansson I, Lundqvist E, Dahl ML, Ingelman-Sundberg M: PCR-based genotyping for duplicated and deleted CYP2D6 genes. *Pharmacogenetics* 1996; **6**: 351–355.

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

26 Xu CF, Lewis K, Cantone KL *et al*: Effectiveness of computational methods in haplotype prediction. *Hum Genet* 2002; **110**: 148–156.

27 Schneider S, Roessli D, Excoffier L: *Arlequin ver. 2000: A Software for Population Genetics Data Analysis*. Switzerland: Genetics and Biometry Laboratory, University of Geneva, 2000.

28 Sokal RR, Rohlf FJ: *Biometry*, 3rd edn. San Francisco: Freeman, 1995.

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

30 Templeton AR, Crandall KA, Sing CF: A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* 1992; **132**: 619–633.

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

32 Wennerholm A, Johansson I, Hidestrand M, Bertilsson L, Gustafsson LL, Ingelman-Sundberg M: Characterization of the CYP2D6*29 allele commonly present in a black Tanzanian population causing reduced catalytic activity. *Pharmacogenetics* 2001; **11**: 417–427.

33 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.

34 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.

35 Hosking LK, Boyd PR, Xu CF *et al*: Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity. *Pharmacogenomics J* 2002; **2**: 165–175.

36 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.

37 Tishkoff SA, Verrelli BC: Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet* 2003; **4**: 293–340.

38 Gaedigk A, Bradford LD, Marcucci KA, Leeder JS: Unique CYP2D6 activity distribution and genotype-phenotype discordance in black Americans. *Clin Pharmacol Ther* 2002; **72**: 76–89.

39 Aklillu E, Herrlin K, Gustafsson LL, Bertilsson L, Ingelman-Sundberg M: Evidence for environmental influence on CYP2D6-catalysed debrisoquine hydroxylation as demonstrated by phenotyping and genotyping of Ethiopians living in Ethiopia or in Sweden. *Pharmacogenetics* 2002; **12**: 375–383.

40 Anderson EC, Slatkin M: Population-genetic basis of haplotype blocks in the 5q31 region. *Am J Hum Genet* 2004; **74**: 40–49.

41 Nielsen DM, Weir BS: Association studies under general disease models. *Theor Popul Biol* 2001; **60**: 253–263.

42 Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989; **123**: 585–595.

43 Garte S, Gaspari L, Alexandrie AK *et al*: Metabolic gene polymorphism frequencies in control populations. *Cancer Epidemiol Biomarkers Prev* 2001; **10**: 1239–1248.

44 Piazza A: Who are the Europeans? *Science* 1993; **260**: 1767–1769.

45 Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P, Comas D: Human mitochondrial DNA variation and the origin of Basques. *Ann Hum Genet* 1995; **59**: 63–81.

46 Scozzari R, Cruciani F, Pangrazio A *et al*: Human Y-chromosome variation in the Western Mediterranean area: implications for the peopling of the region. *Hum Immunol* 2001; **62**: 871–884.

47 Barbujani G, Sokal RR: Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc Natl Acad Sci USA* 1990; **87**: 1816–1819.

48 Rosser ZH, Zerjal T, Hurles ME *et al*: Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet* 2000; **67**: 1526–1543.

49 Ramamoorthy Y, Tyndale RF, Sellers EM: Cytochrome *P*450 2D6.1 and cytochrome *P*450 2D6.10 differ in catalytic activity for multiple substrates. *Pharmacogenetics* 2001; **11**: 477–487.

50 Sistonen J, Fuselli S, Levo A, Sajantila A: CYP2D6 genotyping by a multiplex primer extension reaction; Submitted for publication.