The Pharmacogenomics Journal (2008) 8, 349–356
© 2008 Nature Publishing Group All rights reserved 1470-269X/08 $30.00
www.nature.com/tpj

ORIGINAL ARTICLE

Global patterns of variation in allele and haplotype frequencies and linkage disequilibrium across the *CYP2E1* gene

M-Y Lee[1,2], N Mukherjee[1], AJ Pakstis[1], S Khaliq[3], A Mohyuddin[3], SQ Mehdi[3], WC Speed[1], JR Kidd[1] and KK Kidd[1]

[1]*Department of Genetics, Yale University School of Medicine, New Haven, CT, USA;* [2]*Department of Preventive Medicine, School of Medicine, Keimyung University, Daegu, Republic of Korea and* [3]*Institute of Biotechnology and Genetic Engineering, University of Karachi, Karachi, Pakistan*

Correspondence:
Dr KK Kidd, Department of Genetics, Yale University School of Medicine, 333 Cedar Street, PO Box 208005, New Haven, CT 06520, USA.
E-mail: Kenneth.Kidd@yale.edu

Received 10 April 2008; revised 17 June 2008; accepted 3 July 2008; published online 29 July 2008

Cytochrome P450 2E1, gene symbol *CYP2E1*, is one of a family of enzymes with a central role in activating and detoxifying xenobiotics and endogenous compounds. Genetic variation at this gene has been reported in different human populations, and some association studies have reported increased risk for cancers and other diseases. To the best of our knowledge, multi-single-nucleotide polymorphism haplotypes and linkage disequilibrium (LD) have not been systematically studied for *CYP2E1* in multiple populations. Haplotypes can greatly increase the power both to identify patterns of genetic variation relevant for gene expression as well as to detect disease-related susceptibility mutations. We present frequency and LD data and analyses for 11 polymorphisms and their haplotypes that we have studied on over 2600 individuals from 50 human population samples representing the major geographical regions of the world. The diverse patterns of haplotype variation found in the different populations we have studied show that ethnicity may be an important variable helping to explain inconsistencies that have been reported by association studies. More studies clearly are needed of the variants we have studied, especially those in the 5′ region, such as the variable number of tandem repeats, as well as studies of additional polymorphisms known for this gene to establish evidence relating any systematic differences in gene expression that exist to the haplotypes at this gene.

The Pharmacogenomics Journal (2008) **8,** 349–356; doi:10.1038/tpj.2008.9; published online 29 July 2008

Keywords: *CYP2E1*; cytochrome P450; SNP; haplotype; linkage disequilibrium; random genetic drift

Introduction

Haplotype diversity is a key to understanding population evolution as well as disease evolution. Heterogeneity in both linkage disequilibrium (LD) and haplotype frequencies across the genome have been observed among large numbers of diverse ethnic populations in several studies.[1–4] Earlier studies from our laboratory have shown that haplotype and LD patterns at different genes associated with diseases vary widely across different populations of the world.[2,5–7] Studies on different genes associated with disease that included the Centre d'Etude du Polymorphisme Humain (CEPH) diversity panel have also shown widely varying haplotype patterns.[8,9] These earlier studies demonstrate the importance of studying the variation patterns in multiple populations representing different regions of the world for genes that have been associated with disease.

Cytochrome P450 2E1 (CYP2E1) is a member of the cytochrome P450 multifamily of enzymes that play a central role in activating and detoxifying a wide variety of xenobiotics as well as endogenous compounds. Several drug effects have been identified. The antifungal drug miconazale has been found[10] to inhibit CYP2E1 enzyme activity. Peterson *et al.*[11] have discussed the complex role CYP2E1 appears to play in the pharmacologic interaction of ciprofloxacin and pentoxifylline; genetic variation in CYP2E1 function may thus have complex secondary consequences. The review by Gonzalez and Yu[12] summarizes the evidence for the important role that genetic variation in the CYP2E1 enzyme plays in the susceptibility of patients to hepatitis induced by antituberculosis drug therapy. The PharmGKB database has links to publications showing the relationship to alcohol-related liver diseases and also reports drug response studies involving CYP2E1 for acetaminophen, alcohol, ethanol, geldanamycin and xenobiotics.

Considerable variation in allelic distributions at *CYP2E1* and of CYP2E1 enzyme activity is found among different human populations.[13–16] Several polymorphic sites in the 5′-flanking and intronic region of *CYP2E1* have been reported to be associated with increased risk factors for cancers and other diseases.[16–20] However, no consistent results were observed in studies of the effects of these single-nucleotide polymorphisms (SNPs) on the expression of the gene and activity of the enzyme, and on the susceptibility to diseases.[21–24]

The promoter region and other regulatory variation in or near the gene will function in *cis* with any amino-acid variation as one functional unit. Relevant variation can also include any variants that affect splicing or mRNA conformation. Thus, the haplotype encompassing all relevant variation is the relevant unit for association studies. LD may allow SNPs with no functional consequences to serve as surrogates for unknown and/or untyped variants with functional consequences. However, haplotype frequencies and LD patterns are expected to vary among populations.

Haplotypes and LD of the *CYP2E1* gene region have been poorly studied. The aim of the present study has been to analyze polymorphisms across most of the *CYP2E1* gene, document global ethnic variation in their allelic frequencies and study the patterns that exist in haplotypes and LD. To those ends we present data on 11 polymorphisms, their frequencies and haplotypes in over 2600 normal, healthy individuals from 50 population samples representing all major geographical regions of the world.

Results

A total of 2657 mostly unrelated individuals (by self report) were typed and analyzed for each of these polymorphisms. Allele frequencies and sample sizes for the 11 polymorphisms in all 50 populations can be found in ALFRED (http://alfred.med.yale.edu/) using the unique identifiers (UIDs) in Tables 1 and 2. Allele frequency ranges for each polymorphism are given in Figure 1, and the ancestral allele frequencies for the 10 SNPs and the most common allele frequency for the variable number of tandem repeats (VNTRs) are given in Supplementary Table S1. There were no significant deviations from Hardy–Weinberg (HW) ratios. The average heterozygosities across 50 population samples and F_{st} values for 11 markers are shown in Supplementary Figure S1. For most markers the average heterozygosities are low, ranging from 0.035 (marker 7) to 0.283 (marker 10). F_{st} values vary around the mean of 0.14 for a standard set of 369 SNPs[25] but are high at markers 10 and 11, 0.254, 0.231, respectively, at the 3′ end of the gene. Only seven of the eleven markers

Table 1 CYP2E1 polymorphisms studied

Marker	*Function*	*Polymorphism*	*dbSNP rs no.*	*Site location*	*Position*[a] *(bp)*	*Base pairs to next SNP*	*ALFRED UID*[b]	*Alleles*	*Ancestral allele*
1	None proven	VNTR		5′ upstream	135 188 828	885[c]	SI014090O	4 alleles	NA[d]
2		C_2431875_10; *Pst*I	rs3813867	5′ upstream	135 189 595	240	SI000693S	G/C	G
3		*Rsa*I	rs2031920	5′ upstream	135 189 835	703	SI000694T	C/T	C
4		C_15867697_10	rs2070672	5′ upstream	135 190 538	281	SI001473P	G/A	A
5		C_25594209_10	rs6413420	5′ upstream	135 190 819	4846	SI001475R	T/G	G
6	Val179Ile	C_30443971_10	rs6413419	Exon 4	135 195 665	1722	SI001468T	G/A	G
7	Ile321Ile	C_7468401_10	rs915909	Exon 6	135 197 387	330	SI001476S	T/C	C
8		C_30173803_10; *Msp*I	rs4646976	Intron 6	135 197 717	817	SI000692R	A/G	A
9		*Dra*I	rs6413432	Intron 6	135 198 534	2593	SI014089W	A/T	A
10		C_16026001_20	rs2070676	Intron 7	135 201 127	225	SI014088V	G/C	G
11	Phe421Phe	C_16026002_10	rs2515641	Exon 8	135 201 352		SI000174Q	T/C	C

Abbreviations: dbSNP, single-nucleotide polymorphism database; VNTR, variable number of tandem repeats.
VNTR has two common (6 and 8 repeats described by Hu *et al.*[29]) alleles; two rare alleles observed in African samples.
[a]NCBI Map Build 36.3.
[b]'UIDs' are unique identifiers in the ALFRED database for polymorphism descriptions and allele frequencies.
[c]Distance to next SNP (bp) from the proximal end of the VNTR.
[d]Not available; failed to identify ancestral allele for VNTR.

Table 2 A total of 50 populations studied: naming conventions, sample sizes and geographical regions

Name	*Abbreviations*	*Location*	N	*Population ALFRED UID*	*Sample ALFRED UID*
Africa					
Biaka	BIA	S.W. Central African Republic	70	PO000005F	SA000005F
Mbuti	MBU	Eastern Democratic Republic of Congo	39	PO000006G	SA000006G
Yoruba	YOR	Western Nigeria	78	PO000036J	SA000036J
Ibo	IBO	Southern Nigeria	48	PO000096P	SA000099S
Hausa	HAS	Northern Nigeria	39	PO000097Q	SA000100B
Chagga	CGA	Kilimanjaro area, Tanzania	45	PO000324J	SA000487T
Masai	MAS	Northern Tanzania	22	PO000456P	SA000854R
Sandawe	SND	North Central, Tanzania	40	PO000661N	SA001773S
African–Americans	AAM	United States	90	PO000098R	SA000101C
Ethiopian Jews	ETJ	Northwestern Ethiopia[a]	32	PO000015G	SA000015G
Somali	SOM	Somalia; refugees in Pakistan	20	PO000075M	SA002138O
S.W. Asia, Europe					
Yemenite Jews	YMJ	Yemen[a]	43	PO000085N	SA000016H
Druze	DRU	Israel	106	PO000008I	SA000047L
Samaritans	SAM	Israel	41	PO000095O	SA000098R
Ashkenazi	ASH	Eastern Europe[a]	83	PO000038L	SA000490N
Adygei	ADY	Krasnodar, Caucasus Mountains	54	PO000017I	SA000017I
Chuvash	CHV	Easternmost Europe near Urals	42	PO000327M	SA000491O
Hungarians	HGR	Hungary	92	PO000453M	SA002023H
Russians	RUA	Kargopol, Archangelsk region	34	PO000019K	SA001530J
Russians	RUV	Vologda, northern Russia	48	PO000019K	SA000019K
Finns	FIN	Finland	36	PO000018J	SA000018J
Danes	DAN	Denmark	51	PO000007H	SA000007H
Irish	IRI	Ireland	118	PO000057M	SA000057M
EuroAmericans	EAM	United States	92	PO000020C	SA000020C
N.W. Asia (Siberia)					
Komi Zyriane	KMZ	N.W. Asia, near Urals	47	PO000326L	SA000489V
Khanty	KTY	N.W. Asia near Urals	50	PO000325K	SA000488U
S.C. Asia					
Mohanna	MHN	Pakistan	61	PO000708P	SA002139P
Hazara	HZR	Pakistan	29	PO000575R	SA002140H
Negroid Makrani	NMK	Pakistan	28	PO000707O	SA002137N
Keralites	KER	Kerala, India[b]	30	PO000672P	SA001854S
N.E. Asia (Siberia)					
Yakut	YAK	Sakha, N.E. Siberia	51	PO000011C	SA000011C
Pacific islands					
Nasioi	NAS	Bougainville, Solomon islands	23	PO000012D	SA000012D
Micronesians	MCR	Micronesia, multiple islands	37	PO000063J	SA000063J
East Asia					
Laotians	LAO	Laos	119	PO000671O	SA001853R
Cambodians	CBD	Cambodia	25	PO000022E	SA000022E
SF Chinese	CHS	Southern Han, SF Bay Area	60	PO000009J	SA000009J
TW Chinese	CHT	Taiwan	49	PO000009J	SA000001B
Hakka	HKA	Taiwan	41	PO000003D	SA000003I
Koreans	KOR	Seoul, Korea	54	PO000030D	SA000936S
Japanese	JPN	Japan	51	PO000010B	SA000010B
Ami	AMI	Eastern mountains, Taiwan	40	PO000002C	SA000002C
Atayal	ATL	Eastern mountains, Taiwan	42	PO000021D	SA000021D
Americas					
Cheyenne	CHY	Oklahoma, USA	56	PO000023F	SA000023F

Table 2 *Continued*

Name	*Abbreviations*	*Location*	N	*Population ALFRED UID*	*Sample ALFRED UID*
Pima, Arizona	PMA	Arizona, USA	51	PO000033G	SA000025H
Pima, Mexico	PMM	Northern Mexico	53	PO000034H	SA000026I
Maya	MAY	Central Yucatan, Mexico	52	PO000013E	SA000013E
Quechua	QUE	Peru	22	PO000069P	SA000069P
Ticuna	TIC	Amazon, Brazil	65	PO000027J	SA000027J
R. Surui	SUR	Rondonia, Amazon, Brazil	47	PO000014F	SA000014F
Karitiana	KAR	Amazon, Brazil	57	PO000028K	SA000028K

Abbreviation: UID, unique identifier.
[a]Samples collected in Israel.
[b]Samples collected in USA from individuals born in Kerala.

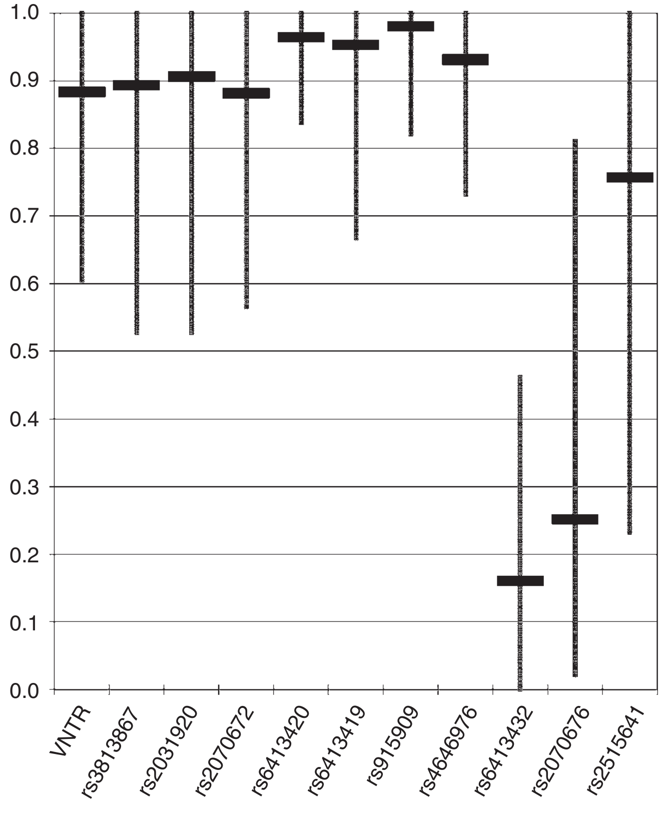

Figure 1 Graphical representation of the average and range of ancestral allele frequencies in 50 population samples for each of 11 markers at cytochrome P450 2E1 (*CYP2E1*) in 50 population samples.

segregate in all populations. The derived allele frequencies of SNPs at exon 4 (marker 6) and exon 6 (marker 7) are very low outside of Africa and these derived alleles are completely absent in the populations of East Asia and the Americas. The derived alleles of the upstream SNPs, except rs6413420 (marker 5), are observed in higher frequencies in Asia and the Americas than in Africa or Europe.

We inferred 16 common haplotypes and estimated their frequencies (Figure 2). Most of the low-frequency variation in the residual class of rare haplotypes is accounted for by a relatively small number of haplotypes in the 2–4% frequency range. The variation in haplotype frequencies among populations gives rise to a complex pattern of LD, both pairwise and as segments with high LD, that varies among populations (Supplementary Table S2 and Figure S2).

Haplotype diversity is much higher in Africa (with 6–10 common haplotypes) than outside of Africa (with about 1–6 common haplotypes). The most common 11-marker haplotype, 6GCAGGCATCC (dark green in Figure 2), is very frequent in all populations outside of Africa and in Ethiopia, but not in other populations of Africa. Two haplotypes, 6CTAGGCAACC (light yellow) and 8GCGGGCGTGT (light blue), are not seen in African populations and rarely seen ($<5\%$) in European populations, but are more frequent in most East Asian (0.0–0.273 and 0.073–0.262) and Native American (0.065–0.443 and 0.023–0.184) populations.

In order to understand the evolution of the haplotypes we estimated haplotypes with fewer SNPs across shorter segments of the gene. We identified three core regions that have evolved common haplotypes solely by accumulation of mutations from the ancestral core haplotype. These cores involve markers 1 through 5 (core A), markers 6 through 9 (core B), and markers 10 and 11 (core C; Figure 3). No recurrent mutations are required to explain all of these core haplotypes. The full 11-marker haplotypes can be explained by combinations of the haplotypes of the three cores (Figure 4; Supplementary Table S3). These combinations have arisen by accumulation of mutations (as depicted in Figure 3) and historical crossovers. It is difficult to be certain of orders of all events, mutations and crossovers, when the three cores are considered together, in part because other combinations that could have been intermediate now are either absent or exist among the rare haplotypes.

In contrast to the global frequency patterns of the whole 11-marker haplotypes, the individual core haplotypes show different global patterns (Supplementary Figures 3–5). Core A haplotypes show greater frequency similarity between African and European populations than between European and both East Asian and Native American populations. One core A haplotype, 6GCAG (no. 1 in Figure 3a), exists at frequencies of 56–95% in the Africans and Europeans.

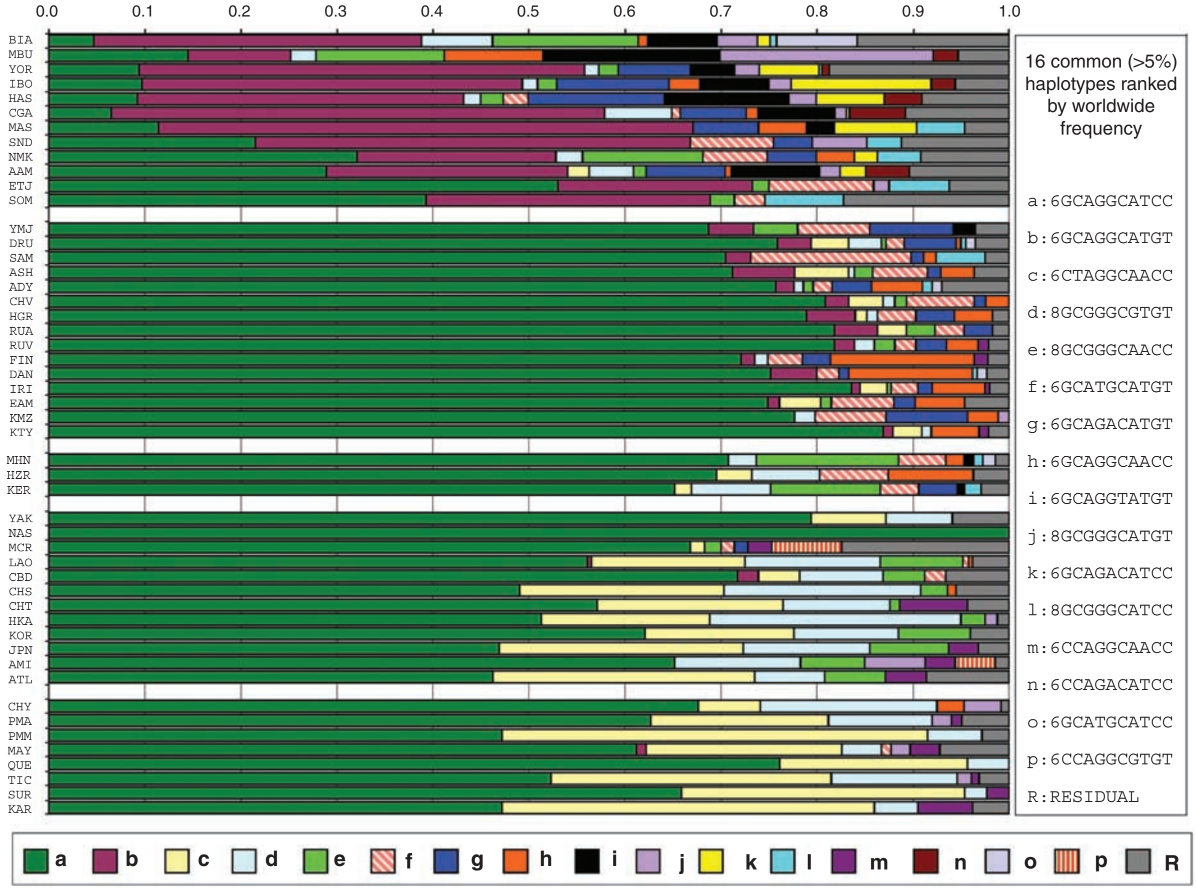

Figure 2 Frequencies for the haplotypes based on 11 markers at cytochrome P450 2E1 (*CYP2E1*) for 50 populations. For each color-coded haplotype the alleles are shown for the sites in chromosome order as numbered in Table 1. Both the full allelic description and lower case letter codes for the haplotypes are given. For each population the proportional length of each color bar represents the frequency of the respective haplotype. All haplotypes that have frequencies less than 5% in all the populations studied are grouped into the residual (gray bar) class. The ancestral haplotype, GCAGGCAAGC, for markers 2 through 11, is not found at common frequencies in any of the 50 populations studied.

Another core A haplotype, 6CTAG (no. 5 in Figure 3a), is not seen in Africans, is rare in Europeans, but is frequent in East Asian and Native American populations. To the degree these 5′ markers encompass the major regulatory regions, it is possible that East Asians and Native Americans may have a common derived variant in regulation that is very uncommon to absent in the rest of the world.

Discussion

We are unaware of any publications of the *CYP2E1* gene that have included all of the polymorphisms that we present here. Certainly, none of these markers has been studied previously on such a large and ethnically diverse set of individuals. This study is an explicit example of the type of global perspective on pharmacogenetic variation within and among populations discussed in an editorial by Marsh.[26] Even this data set does not probe the full extent of the genetic diversity of this small segment of DNA. Public databases report multiple additional polymorphisms across the gene (including 5′ and 3′-untranslated regions).

We cannot precisely relate the 16 common haplotypes (Figure 2) we have observed to the standard *CYP2E1* allelic designations in the 'cypalleles' Web site (http://www.cypalleles.ki.se/cyp2e1.htm) because, from a genetic transmission perspective, each of the haplotypes we report is an allele and the 'cypalleles' Web site does not give full haplotype specifications for the allele designations they summarize, precluding a strict comparison. Moreover, we have not included SNPs with rare or uncommon variants that have not been studied widely. To distinguish the haplotypes we have identified from those in the 'cypalleles' nomenclature, we have used letter designations rather than numbers in Figures 2 and 4 and Supplementary Table S3. As an example of the difficulty of establishing precise correspondences, the mutation G→A at marker 6 (corresponding to 179 Val→Ile) defines core B haplotype 4 (Figure 3B) and appears to represent one mutational event. That core B haplotype exists in two combinations with core A haplotypes and two combinations with core C haplotypes for a total of three

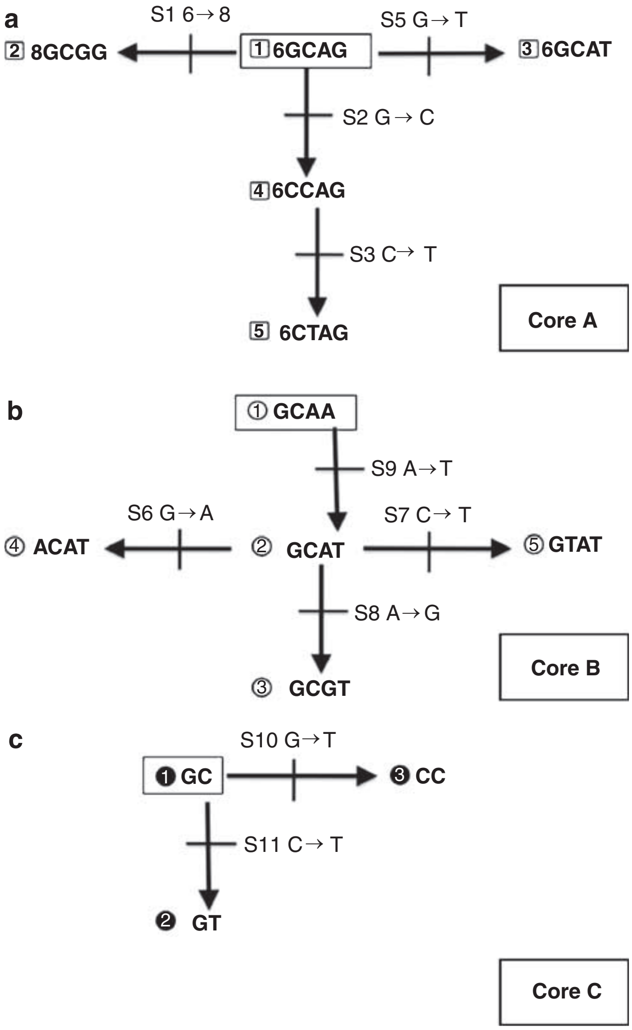

Figure 3 The evolutionary relationships among haplotypes of three core segments of cytochrome P450 2E1 (*CYP2E1*). In all cases the schema starts with the ancestral human haplotype and gives the pattern of mutational accumulation for that core. (**a**) Core A comprised of markers 1 through 5. (**b**) Core B comprised of markers 6 through 9. (**c**) Core C comprised of markers 10 and 11.

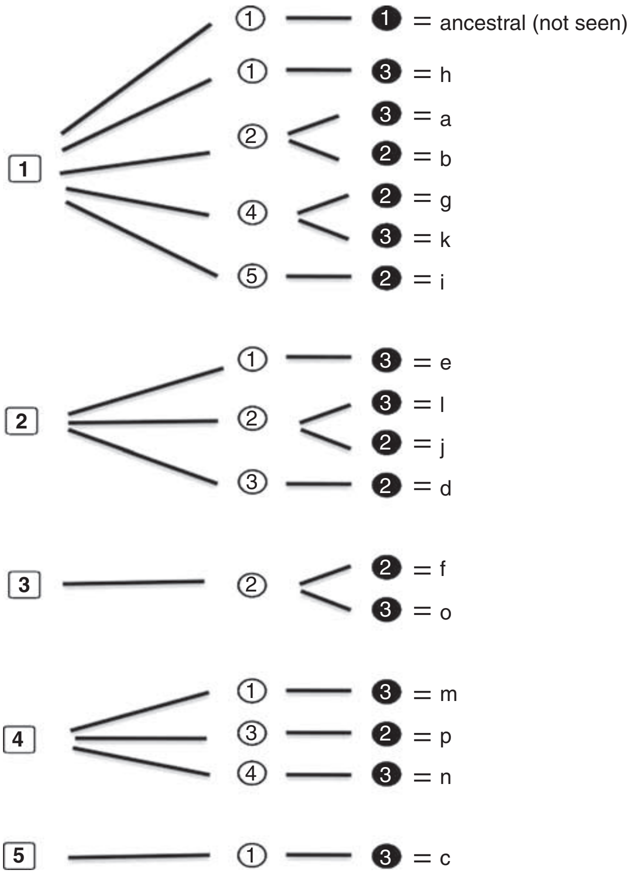

Figure 4 The composition of the 11 full haplotypes in terms of combinations of individual core haplotypes. The core haplotypes are numbered as in Figure 3 with core A on the left, core B in the center and core C on the right. The lower case letters for the full haplotypes correspond to those in Figure 2. (See also Supplementary Table S3.)

11-marker haplotypes: g, k and n (Figures 2 and 4). All three of these haplotypes correspond to allele *CYP2E1*4* in the 'cypalleles' nomenclature. We expect the haplotypes encompassing the gene to become more complex as more SNPs and rare variants are included in an even more comprehensive study of the gene.

In addition to multiple SNPs across this gene, copy number variation (CNV) encompassing *CYP2E1* has been reported.[27,28] Our typing methods are not designed to detect CNVs but we can exclude any common occurrence in our samples because there is no significant deviation from HW ratios in any of the populations.

We have studied the allele, haplotype and LD variation patterns for 11 polymorphisms in 50 populations from different geographical regions of the world across the *CYP2E1* gene and have shown that there are large differences in these patterns worldwide. The haplotypes were useful in inferring recombination events in the recent evolution of the gene. The current study focuses attention on the core haplotype lineages that appear to have involved no recombination and on the combinations that have arisen because of historical recombinations. These cores and their combinations provide the framework for future expression studies. Depending upon when in one of the evolutionary lineages a functional variant arose, we would expect it to either define a new sublineage or be inherited into the descendant haplotypes in Figure 3. Thus, the evolutionary lineages may explain multiple haplotypes (alleles) having similar functional properties, even if the causative variant has not yet been identified. Other SNPs within the molecular extent of the region spanned will likely fall within this framework; some might refine the locations of the inferred historical crossovers.

Supplementary information is available at the *The Pharmacogenomics Journal*'s Web site.

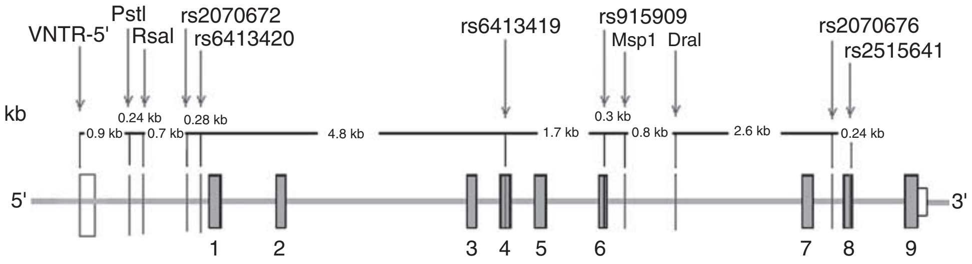

Figure 5 Map of cytochrome P450 2E1 (*CYP2E1*) on chromosome 10 and the markers typed. The filled boxes represent the exons of the gene; the number below each box is the exon number. The vertical lines represent positions of the markers studied.

Materials and methods

Samples studied

DNA was purified from lymphoblastoid cell lines from 2657 healthy adults from 50 populations from around the world (Table 2). Population membership was designated by the subjects and all blood samples were obtained with individual informed consent following protocols approved by the Institutional Review Boards at Yale University School of Medicine, at the University of Karachi, and at multiple other relevant institutions in countries where samples were collected. The average population sample size is 53 individuals.

Markers studied

We studied 10 SNPs and 1 VNTR across 13.9 kb that encompasses the 5′ region of the *CYP2E1* gene and almost the entire coding region (Figure 5). We typed the VNTR and four SNPs in the upstream region, three SNPs in the coding regions and three SNPs in the intronic regions of the gene (Table 1). The markers are referred to by their numeric position (1–11) in Table 1. The SNPs are all diallelic and the VNTR is essentially diallelic, as initially described.[29] Two other very rare VNTR alleles have been seen in some African populations in the course of this study (data not shown); they were excluded from the haplotype analyses.

Typing methods

The samples were typed by TaqMan assays (markers 2, 4–8, 10 and 11), by fragment length analysis on agarose gels (marker 1) after PCR, and by restriction fragment length after enzyme digestion of the PCR products for markers 3 (*Rsa*I) and 9 (*Dra*I).

Determining ancestral alleles

For each allele, the ancestral state in humans was determined by inference from the allele present in several other primate species. The ancestral allele of the VNTR (marker 1) could not be determined but by inference is 6.

Statistical methods

Allele frequencies of the VNTR and SNPs were calculated by gene counting assuming codominant inheritance. All the sites were also tested for Hardy-Weinberg (HW) ratios by χ^2-test and/or exact test. Expected heterozygosities were estimated as $1-\sum p_i^2$. Haplotype frequencies were estimated by the expectation–maximization (EM) algorithm using HAPLO.[30] Haplotypes with estimated frequencies of less than 5% in each of the population samples go into the residual class. The 5% threshold is a reasonable boundary for determining what are the common and rare haplotypes given the sample sizes in this study. Although some estimated haplotypes below the 5% threshold have very clear evidence of occurrence, the standard errors on estimated frequencies increase along with some erroneous inferences due to the small number of observations available in the rare zone and the fact that the LD levels between sites vary. Pairwise LD estimates were carried out as r^2 [31,32] with significance levels determined by a permutation test.[33] Comparative plots of LD for all the populations were carried out using HAPLOT.[34]

Abbreviations

CYP2E1	*cytochrome P450, family 2, subfamily E, polypeptide 1*
LD	*linkage disequilibrium*
PCR	*polymerase chain reaction*
SNP	*single-nucleotide polymorphism*
UID	*unique identifier*
VNTR	*variable number of tandem repeats*

Acknowledgments

This work was funded, in part, by National Institute of Health grants GM057672 and AA009379 to Kenneth K Kidd. None of the authors has any conflicts of interest related to the data presented here. We thank all the colleagues who helped us assemble the population samples. Special thanks are due the many hundreds of individuals from these populations who volunteered to give blood samples for studies such as this.

Duality of interest

None declared.

Electronic databases cited

ALFRED, The ALlele FREquency Databse; http://alfred.med.yale.edu; PharmGKB, The Pharmacogenetics and Pharmacogenomics Knowledge Base, http://www.PharmGKB.org

References

1 Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC *et al.* A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum Genet* 1998; **103**: 211–227.

2 Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A *et al.* Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am J Hum Genet* 2000; **66**: 1882–1899.

3 Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, Brookes AJ *et al.* Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet* 2005; **13**: 677–686.

4 Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am J Hum Genet* 2001; **69**: 831–843.

5 Mukherjee N, Kidd KK, Pakstis AJ, Speed WC, Li H, Tarnok Z *et al.* The complex global pattern of genetic variation and linkage disequilibrium at catechol-*O*-methyl transferase (*COMT*). *Mol Psychiatry* 2008, advance online publication, 24 June 2008; doi:10.1038/mp.2008.64.

6 Han Y, Gu S, Oota H, Osier MV, Pakstis AJ, Speed WC *et al.* Evidence of positive selection on a class I ADH locus. *Am J Hum Genet* 2007; **80**: 441–456.

7 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.

8 Evans DM, Cardon LR. Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *Am J Hum Genet* 2004; **75**: 687–692.

9 Gardner M, González-Neira A, Lao O, Calafell F, Bertranpetit J, Comas D. Extreme population differences across Neuregulin 1 gene, with implications for association studies. *Mol Psychiatry* 2006; **11**: 66–75.

10 Niwa T, Inoue-Yamamoto S, Shiraga T, Takagi A. Effect of antifungal drugs on Cytochrome P450 (CYP) 1A2, CYP2D6, and CYP2E1 activities in human liver microsomes. *Biol Pharm Bull* 2005; **28**: 1813–1816.

11 Peterson TC, Peterson MR, Wornell PA, Blanchard MG, Gonzalez FJ. Role of CYP1A2 and CYP2E1 in the pentoxifylline ciprofloxacin drug interaction. *Biochem Pharmacol* 2004; **68**: 395–402.

12 Gonzalez FJ, Yu AM. Cytochrome P450 and xenobiotic receptor humanized mice. *Ann Rev Pharmacol Toxicol* 2006; **46**: 41–64.

13 Hayashi S, Watanabe J, Nakachi K, Kawajiri K. Genetic linkage of lung cancer-associated MspI polymorphisms with amino acid replacement in the heme binding region of the human cytochrome P450IA1 gene. *J Biochem (Tokyo)* 1991; **110**: 407–411.

14 Kim RB, Yamazaki H, Chiba K, O'Shea D, Mimura M, Guengerich FP *et al. In vivo* and *in vitro* characterization of *CYP2E1* activity in Japanese and Caucasians. *J Pharmacol Exp Ther* 1996; **279**: 4–11.

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

16 Danko IM, Chaschin NA. Association of CYP2E1 gene polymorphism with predisposition to cancer development. *Exp Oncol* 2005; **27**: 248–256.

17 Song BJ. Ethanol-inducible cytochrome P450 (CYP2E1): biochemistry, molecular biology and clinical relevance: 1996 update. *Alcohol Clin Exp Res* 1996; **20**(Suppl): 138A–146A.

18 Iwahashi HK, Ameno S, Ameno K, Okada N, Kinoshita H, Sakae Y *et al.* Relationship between alcoholism and CYP2E1 C/D polymorphism. *Neuropsychobiology* 1998; **38**: 218–221.

19 Nguyen TT, Murphy NP, Austin CM. Amplification of multiple copies of mitochondrial Cytochrome b gene fragments in the Australian freshwater crayfish, Cherax destructor Clark (Parastacidae: Decapoda). *Anim Genet* 2002; **33**: 304–308.

20 Howard LA, Ahluwalia JS, Lin SK, Sellers EM, Tyndale RF. CYP2E1*1D regulatory polymorphism: association with alcohol and nicotine dependence. *Pharmacogenetics* 2003; **13**: 321–328. [erratum in *Pharmacogenetics* 2003; **13**: 441–442].

21 Lee HS, Yoon JH, Kamimura S, Iwata K, Watanabe H, Kim CY. Lack of association of cytochrome P4502 E1 genetic polymorphisms with the risk of human hepatocellular carcinoma. *Int J Cancer* 1997; **71**: 737–740.

22 Morita S, Yano M, Shiozaki H, Tsujinaka T, Ebisui C, Morimoto T *et al.* eCYP1A1, CYP2E1 and GSTM1 polymorphisms are not associated with susceptibility to squamous-cell carcinoma of the esophagus. *Int J Cancer* 1997; **71**: 192–195.

23 Powell H, Kitteringham NR, Pirmohamed M, Smith DA, Park BK. Expression of cytochrome P4502E1 in human liver: assessment by mRNA, genotype and phenotype. *Pharmacogenetics* 1998; **8**: 411–421.

24 Wong NA, Rae F, Simpson KJ, Murray GD, Harrison DJ. Genetic polymorphisms of cytochrome p4502E1 and susceptibility to alcoholic liver disease and hepatocellular carcinoma in a white population: a study and literature review, including meta-analysis. *Mol Pathol* 2000; **53**: 88–93.

25 Kidd KK, Pakstis AJ, Speed WC, Kidd JR. Understanding human DNA sequence variation. *J Hered* 2004; **95**: 406–420.

26 Marsh S. Pharmacogenetics: global clinical markers. *Pharmacogenomics* 2008; **9**: 371–373.

27 Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD *et al.* Global variation in copy number in the human genome. *Nature* 2006; **444**: 444–454.

28 Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007; **17**: 1665–1674.

29 Hu Y, Hakkola J, Oscarson M, Ingelman-Sundberg M. Structural and functional characterization of the 5′-flanking region of the rat and human cytochrome P450 2E1 genes: identification of a polymorphic repeat in the human gene. *Biochem Biophys Res Commun* 1999; **263**: 286–293.

30 Hawley ME, Kidd KK. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered* 1995; **86**: 409–411.

31 Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 1995; **29**: 311–322.

32 Lewontin RC. The interaction of selection and linkage. Ii. Optimum models. *Genetics* 1964; **50**: 757–782.

33 Zhao H, Pakstis AJ, Kidd JR, Kidd KK. Assessing linkage disequilibrium in a complex genetic system. I. Overall deviation from random association. *Ann Hum Genet* 1999; **63**: 167–179.

34 Gu S, Pakstis AJ, Kidd KK. HAPLOT: a graphical comparison of haplotype blocks, tagSNP sets and SNP variation for multiple populations. *Bioinformatics* 2005; **21**: 3938–3939.

Supplementary Information accompanies the paper on the The Pharmacogenomics Journal Web site (http://www.nature.com/tpj)