

Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity

LK Hosking¹
PR Boyd¹
CF Xu¹
M Nissum³
K Cantone¹
IJ Purvis¹
R Khakhar¹
MR Barnes¹
U Liberwirth³
K Hagen-Mann³
MG Ehm²
JH Riley¹

¹GlaxoSmithKline Medicines Research Centre, Stevenage, Herts, UK and New Frontiers Science Park North, Harlow, Essex, UK; ²GlaxoSmithKline, Research Triangle Park, NC, USA; ³MWG-Biotech, Ebersberg, Germany

Correspondence:

L Hosking, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Rd, Stevenage, Herts SG1 2NY, UK
Tel: +44 1438 764950
Fax: +44 1438 768097
E-mail: lkh47482@gsk.com

ABSTRACT

The cytochrome p450 enzyme, CYP2D6, metabolises approximately 20% of marketed drugs. CYP2D6 multiple variants are associated with altered enzyme activities. Genotyping 1018 Caucasians for CYP2D6 polymorphisms (G1846A, delT1707, delA2549 and A2935C), known to result in the recessive CYP2D6 poor drug metaboliser (PM) phenotype, identified 41 individuals with predicted PM phenotype. These 41 individuals were classified as 'cases'. Single nucleotide polymorphisms (SNPs) mapping within an 880 kb region flanking CYP2D6, were identified to evaluate potential association between genetic variation and the CYP2D6 PM phenotype. The 41 PM cases and 977 controls were genotyped and analysed for 27 SNPs. Associations were observed across a 390 kb region between 14 SNPs and the PM phenotype (P values from 6.20×10^{-4} to 4.54×10^{-35}). Haplotype analysis revealed more significant levels of association ($P = 3.54 \times 10^{-56}$). Strong ($D' > 0.7$) linkage disequilibrium (LD) between SNPs was observed across the same 390 kb region associated with the CYP2D6 phenotype. The observed phenotype:genotype association reached genome-wide levels of significance, and supports the strategy for potential application of LD mapping and whole genome association scans to pharmacogenetic studies.

The Pharmacogenomics Journal (2002) 2, 165–175. doi: 10.1038/sj.tpj.6500096

Keywords: CYP2D6; linkage disequilibrium (LD); single nucleotide polymorphism (SNP); haplotype

INTRODUCTION

Association between a genetic marker and a phenotype may exist as a result of linkage disequilibrium (LD) between the disease causing allele and the marker allele. LD estimations have already contributed to the localisation of causal genes in several monogenic disorders,^{1,2} and are currently being applied to complex disease gene identification.^{3,4} A proof of concept study designed to determine the usefulness of association studies for mapping complex disease susceptibility genes proved successful.⁵ Single nucleotide polymorphisms (SNPs) surrounding APOE, a previously confirmed late onset Alzheimer's susceptibility gene on chromosome 19, were identified and associated strongly with late onset Alzheimer's disease.⁵ Calpain 10 was associated with susceptibility to type II diabetes mellitus⁶ following analysis of genetic variation within the gene, in Mexican Americans. LD mapping has successfully identified a region containing IBD5, a susceptibility gene for Crohn disease, which contained a common haplotype strongly associated with the disease.⁷ Recently identified SNPs in the insulin receptor gene were associated with migraine by LD mapping.⁸ LD therefore plays an important role in mapping complex disease genes⁶ and in proposed genome-wide scans.⁹ With the release of the first draft sequence of the entire human genome, and the concomitant output from The SNP Consortium¹⁰ culminating in the release

Received: 12 November 2001

Revised: 15 January 2002

Accepted: 18 January 2002

of 2 400 000 SNPs (October 2001), it has become possible to rapidly identify and map genetic variation to specific fully sequenced chromosomal regions. There has been ongoing debate regarding the study design for LD mapping, such as the density of the SNP map, the allele frequencies of the SNPs and the number of samples required to detect a phenotype:genotype association.^{11,12}

It is known that there are inter-individual differences in response to medicine, such as efficacy and adverse events. Altered response may be a result of underlying genetic variation between different individuals. For example, Kuivenhoven *et al* observed that different alleles of the cholesteryl ester transfer protein gene, were associated with altered progression of coronary atherosclerosis in individuals taking Pravastatin.¹³ Genetic variation within genes encoding drug metabolising enzymes is widely reported to affect medicine response.¹⁴ *CYP2D6* metabolises around 20% of marketed drugs.¹⁵ More than 68 variations have been identified in the *CYP2D6* gene (<http://www.imm.ki.se/CYPalleles>), some of which may lead to different drug metabolising activities. Normal drug metabolism is classified as 'extensive', drug metabolism requiring multiple copies of *CYP2D6* as 'ultra-rapid', and impaired drug metabolism is referred to as either 'intermediate' or, in the case of complete lack of enzyme activity, 'poor'. This can result in altered medicine safety, toxicity and efficacy in individuals harbouring the relevant *CYP2D6* polymorphisms. Several studies identified specific *CYP2D6* alleles used to differentiate various drug metabolising phenotypes.^{16–18} *CYP2D6* poor metaboliser (PM) phenotype is generally reported to exist at a frequency of around 5–10% in the Caucasian population.¹⁹ Many other genes contributing to drug absorption, distribution, metabolism and elimination may harbour genetic variation resulting in altered drug efficacy and safety. It is therefore advantageous to identify which additional chromosomal regions contribute to drug response.

This study attempts to evaluate the feasibility of LD mapping and genome-wide association scans in identifying chromosomal regions harbouring genetic variation leading to altered drug response. Twenty-seven SNPs mapping to an 880 kb region flanking *CYP2D6*, were genotyped in *CYP2D6* PMs and controls. LD across the region was estimated, and association between SNPs and the PM phenotype was determined.

RESULTS

Identification of 'Cases' for Phenotype:Genotype Association Study

More than 18 *CYP2D6* mutations contribute to the recessive PM phenotype, five of which (G1934A, delA2637, delT1795, A3023C and a whole gene deletion, *CYP2D6D*) would detect almost all (>99%) PMs in Caucasians.¹⁹ Conforming to standard gene polymorphism nomenclature (<http://www.imm.ki.se/CYPalleles>), they are now referred to as G1846A, delA2549, delT1707, A2935C and *CYP2D6D*, respectively. Four of these polymorphisms (G1846A, delA2549, delT1707, A2935C) were directly typed using Taqman in 1018 Caucasians to identify individuals with predicted PM phenotype.

The minor allele frequencies were 19%, 1.75%, 1.0% and 0.05% respectively, similar to that reported for Caucasians.¹⁹ Allele frequencies have been calculated without taking into account the frequency of the whole gene deletion. The frequency of the second most common polymorphism, the whole gene deletion (*CYP2D6D*), is reported to be ~2% in Caucasians¹⁹, predicting <1 individual homozygous for this deletion in a sample of 1018 Caucasians. If an individual is homozygous for the deletion, no PCR amplification would be expected when assaying for any of the other polymorphisms within the *CYP2D6* gene. Every individual in this study generated PCR products when genotyped for the four *CYP2D6* polymorphisms listed, suggesting that none of the individuals was homozygous for *CYP2D6D*. Compound heterozygotes for *CYP2D6D* and an additional *CYP2D6* polymorphism would be detected as homozygotes for the additional *CYP2D6* polymorphism, and correctly identified as *CYP2D6* PMs. Direct genotyping for the whole gene deletion, *CYP2D6D*, was therefore not performed. Forty-one individuals were predicted to express PM phenotype: 32 were either homozygous for G1846A or G1846A/*CYP2D6D* heterozygotes, and nine were compound heterozygotes (G1846A/delA2549; G1846A/delT1707 or delA2549/delT1707). These 41 individuals were subsequently referred to as the cases or PMs. The remaining 977 individuals were classified as the controls, which consisted of extensive, intermediate and rapid metabolisers, as well as ~0.3% PMs.

Selection of SNPs Flanking the *CYP2D6* Gene

Seventy-five SNPs within an 880 kb region on human chromosome 22q13.1–13.2 were identified and verified, from the public databases or by *de novo* sequencing. Forty of the 75 SNPs were selected on the basis of allele frequency and physical location, and genotyped in cases and controls. Two SNPs were non-polymorphic within this sample and were therefore removed from further analysis. Genotype distributions for four SNPs deviated from Hardy–Weinberg equilibrium (HWE), due to either non-specific assays (primers also mapped to pseudogenes, or other members of a gene family) or genotype errors, and these four SNPs were removed from further analysis. Two additional assays proved to be non-specific (mapping to other chromosomal regions, in addition to chromosome 22), and were removed from the study. Five SNPs had allele frequencies <5%, and were also removed from further analysis. Minor allele frequencies for the remaining 27 SNPs ranged from 0.10 to 0.49 (Table 1).

The relative physical locations of the 27 SNPs are represented in Figure 1 (<http://www.ensembl.org>) and map to an 880 kb contig flanking *CYP2D6*. SNP 1 mapped to base 230 and SNP 27 mapped to base 879049 in the contig shown. The gene encoding the drug metabolising enzyme *CYP2D6* maps at approximately 540 kb in the contig, between SNP 17 and SNP 18. The distances between the 27 SNPs ranged from 21 bp to 100 kb, with an average inter-SNP distance of 34 kb (Table 1). In addition to *CYP2D6*, the 880 kb region contains 11 transcripts including *NHP2L1*, *Q9NSP4*, *NAGA*, *NDUFA6*, and *TCF20*. There is a relative

Table 1 Association between 27 SNPs surrounding the *CYP2D6* gene and the *CYP2D6* PM phenotype

SNP ID	Database ID	Inter-SNP distance (bp)	Frequency in cases (41) ^a	Frequency in controls (977) ^a	<i>P</i> ^b
1	TSC313423	88746	0.11	0.22	7.6E-02
2	TSC117740	88818	0.08	0.21	3.40E-02
3	TSC0110320	27634	0.16	0.23	3.40E-01
4	dbSNP5272	33264	0.09	0.19	1.20E-01
5	TSC008552	14379	0.07	0.11	6.50E-01
6	dbSNP4021020	36	0.06	0.10	5.00E-01
7	dbSNP4021021	30497	0.08	0.11	1.90E-01
8	dbSNP4021022	10990	0.00	0.15	8.5E-05
9	TSC0004117	40371	0.62	0.21	1.82E-11
10	dbSNP4021023	33236	0.01	0.21	3.44E-06
11	TSC0118628	21091	0.16	0.49	6.70E-09
12	TSC118637	33697	0.82	0.28	7.16E-25
13	dbSNP4021024	57804	0.11	0.49	2.04E-11
14	TSC0110363	19297	0.12	0.47	6.60E-11
15	dbSNP4021015	10805	0.00	0.34	2.60E-14
16	dbSNP4021016	21	0.00	0.22	5.80E-08
17	dbSNP4021017	28550	0.03	0.35	5.80E-11
18	TSC0110364	18194	0.86	0.25	8.00E-27
19	TSC0110368	99901	0.13	0.44	3.10E-07
20	TSC0035446	17792	0.82	0.17	4.54E-35
21	TSC0089659	39219	0.08	0.27	6.20E-04
22	TSC0117810	25623	0.49	0.48	5.10E-01
23	TSC0091223	59114	0.46	0.49	4.70E-01
24	dbSNP4021011	23686	0.26	0.19	2.10E-01
25	TSC0098529	30887	0.11	0.14	9.10E-01
26	dbSNP4021012	25719	0.19	0.28	2.3E-01
27	TSC0004225		0.13	0.18	6.93E-01

^aAllele frequencies were calculated using gene counting. The number of genotypes for each assay differed from 732 to 1014 due to assay failure rate variation.

^b*P* values were obtained using Fisher's exact test.

increase in overall GC content around SNPs 8 and 21, in comparison with the rest of the 880 kb region. None of the 27 *CYP2D6* surrounding SNPs mapped to *CYP2D6* itself.

Association Between SNP Markers and the *CYP2D6* PM Phenotype

Table 1 and Figure 2a present the association between SNP markers and the *CYP2D6* PM phenotype. Fourteen SNPs showed significant association with the PM trait ($P < 0.01$). All 14 mapped within a 390 kb region flanking *CYP2D6* (SNP 8 to SNP 21). Ten of the 14 SNPs associated with the PM phenotype had *P* values $< 10^{-7}$, demonstrating a level of significance sufficient to detect a phenotype:genotype relationship within an entire genome-wide scan, calculated using the Bonferroni adjustment for multiple testing, assuming genotyping with 100 000 SNPs. SNP 20 had the highest level of significance ($P = 4.54 \times 10^{-35}$) with the PM phenotype and yet resides ~120 kb away from *CYP2D6* itself. The minor allele frequency for SNP 20 was 0.17 and 0.82 in the controls and cases, respectively. Likewise, the minor allele frequency for SNP 12 was 0.28 and 0.82 in the controls and cases, respectively. Conversely for SNP 11, minor allele frequencies decreased from 0.49 in the controls to 0.16 in the cases.

To evaluate whether genome-wide significance levels

could be reached using smaller numbers of samples, the analysis was repeated using either fewer controls, or fewer total samples with a constant 4% proportion of PMs. Generally, the level of significance decreased concomitantly with numbers of controls when numbers of cases remained the same (Figure 2b). Genome-wide significance (P values = 1.6×10^{-13} , 9.6×10^{-9} , 5.7×10^{-11} and 1.3×10^{-13}) for four markers (SNPs 12, 15, 18 and 20 respectively) was observed even when only 41 controls were used in the analysis. As expected, the level of significance observed also decreased concomitantly with the total number of samples analysed (data not shown). When analysing 200 samples (8 cases and 192 controls), five SNPs showed significant association ($P < 0.01$), and one of those (SNP 20) reached a level of significance sufficient for a genome-wide association scan ($P < 10^{-7}$).

Haplotype and Linkage Disequilibrium Analysis

Haplotype frequencies derived from windows of five adjacent SNPs were compared in cases and controls. The highest levels of significance were observed for haplotypes containing SNPs with the greatest genotypic association with PM (SNPs 12, 18 and 20). The region of observed association was the same as that detected using single marker analysis, but the level of significance was considerably higher (Figure 3).

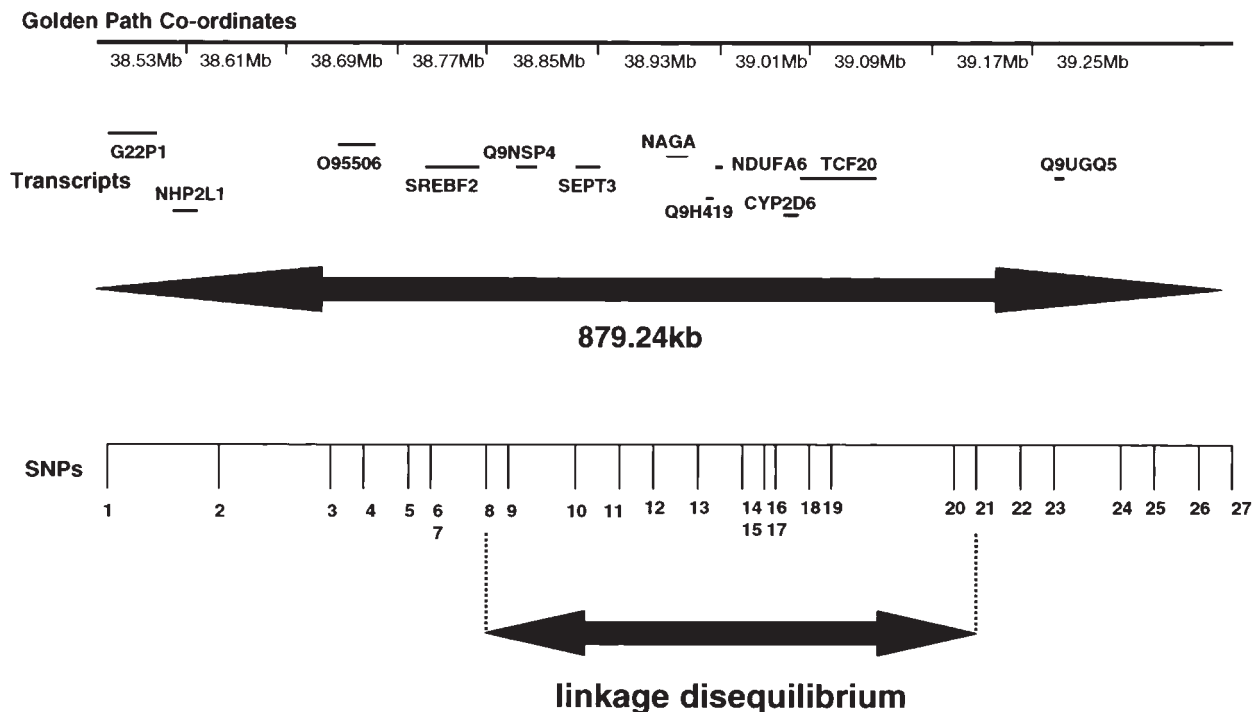


Figure 1 Relative locations of the 27 SNPs flanking *CYP2D6*. The 880 kb region flanking *CYP2D6* containing the relative locations of all 27 SNPs is visually represented in this ensembl output (<http://www.ensembl.org>). Golden path co-ordinates are recorded below the bold black line, and transcripts are represented by bars. Numbers 1–27 represent the SNPs in locus order. *CYP2D6* is located between SNPs 17 and 18. The chromosomal region in linkage disequilibrium with the *CYP2D6* PM phenotype is indicated.

The haplotypes derived from SNP marker set 12–16, spanning ~120 kb, showed association with the PM phenotype with $P = 3.54 \times 10^{-56}$. Haplotypes derived from markers 19–23 spanning ~180 kb, were also significantly associated with the *CYP2D6* PM phenotype ($P = 1.82 \times 10^{-45}$). The estimated frequencies of the individual haplotypes varied enormously between case and control samples (Table 2). One particular haplotype (2_1_1_1_1), derived from marker set 12–16 was present within cases at a frequency of 0.83 and controls with a frequency of 0.17. In the control samples 10 haplotypes derived from marker set 12–16, were estimated to represent >99% of the chromosomes, whereas in the cases >99% of chromosomes were represented by only three haplotypes: 2_1_1_1_1; 1_2_2_1_1 and 1_1_1_1_1. Statistical evaluation of multiple markers physically close to each other appears to be more powerful than markers analysed independently for detection of association with the PM phenotype in this sample.

LD was examined between 27 SNPs flanking *CYP2D6* in the entire sample set of 1018 Caucasians, and the three subgroups separately (230 Corriell panel individuals, 459 North American and 329 UK samples). Similar patterns of three distinct LD blocks were observed in the entire sample set and the three subgroups (Figure 4, data not shown for subgroup analysis). The broadest block of strong LD ($D' > 0.7$) was identified spanning a 390 kb region between SNP 8 and SNP 21 (Figure 4). This region of strong LD correlated with the

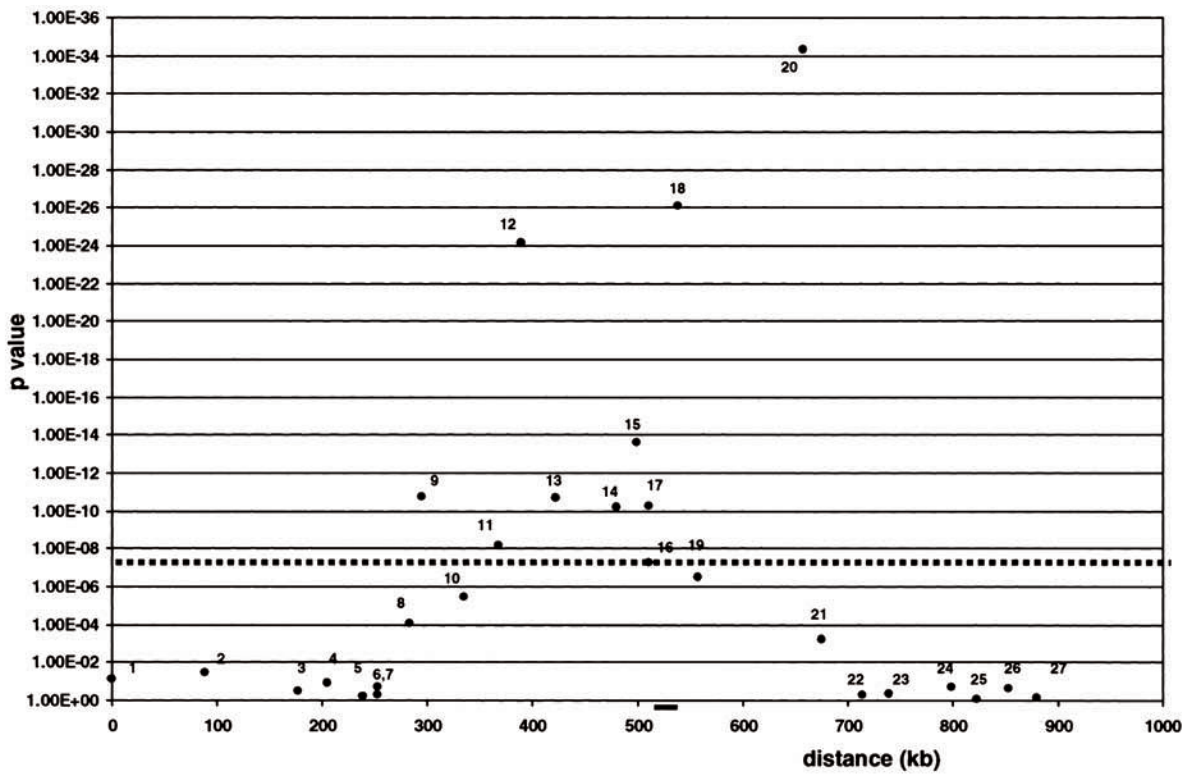
390 kb region harbouring genetic variation associated with the *CYP2D6* PM phenotype.

DISCUSSION

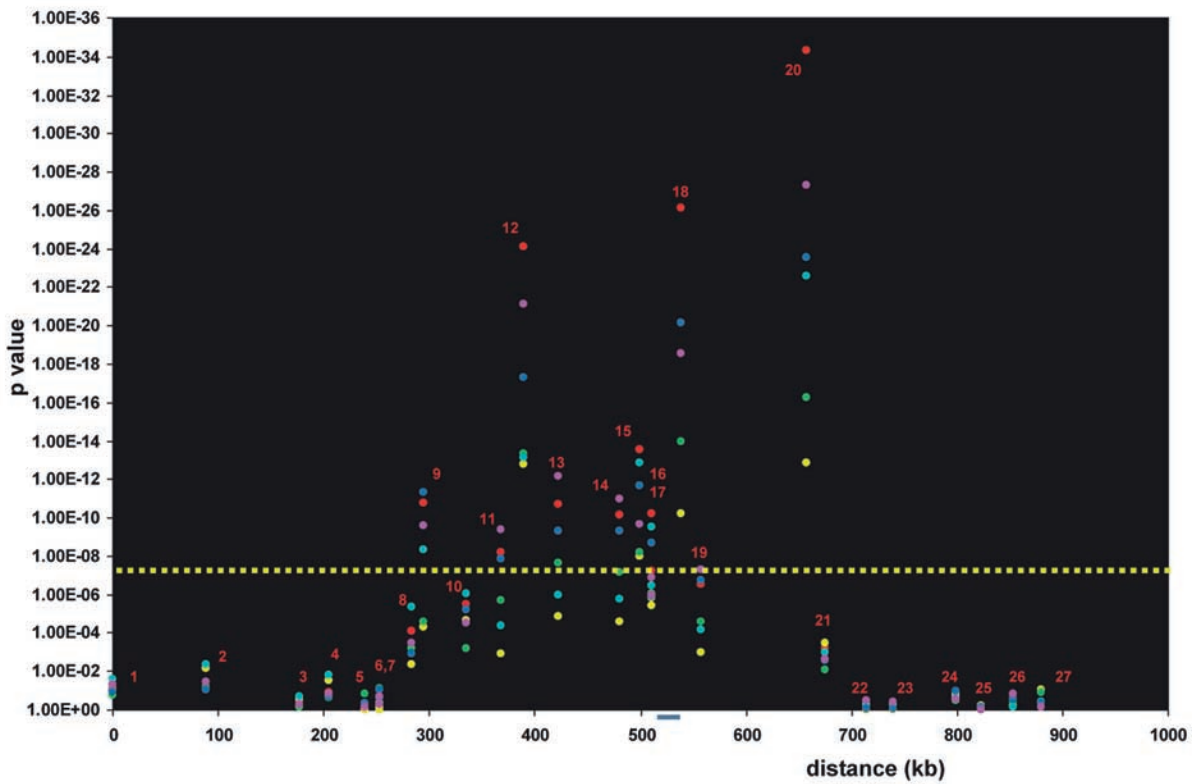
The potential use of LD mapping for the identification of common disease genes is currently the subject of much discussion.^{20,21} This particular set of experiments was perfor-

Figure 2 Association between SNPs and the *CYP2D6* PM phenotype. (a) Fisher's exact test was performed on all cases and all controls. Due to the close physical proximity of some SNPs, the resolution is not sufficient to clearly discriminate all markers. Marker pair 6/7 appears superimposed in this figure at an approximate map location of 253 kb. *CYP2D6* is indicated by the short horizontal bar underneath the x axis, and the horizontal dashed line represents the significance level required to detect association in whole genome scans ($P < 10^{-7}$). Numbers represent the SNPs referred to in the text. (b) Fisher's exact test was performed on all cases and: all controls (red); 205 controls (pink); 164 controls (blue); 123 controls (turquoise); 82 controls (green) and 41 controls (yellow). Due to the close physical proximity of some SNPs, the resolution is not sufficient to clearly discriminate all markers. Marker pairs 6/7 and 16/17 appear superimposed in this figure at approximate map locations of 253 and 510 kb, respectively. *CYP2D6* is indicated by the short horizontal blue bar underneath the x axis, and the horizontal dashed yellow line represents the significance level required to detect association in whole genome scans ($P < 10^{-7}$). Numbers represent the SNPs referred to in the text.

a



b



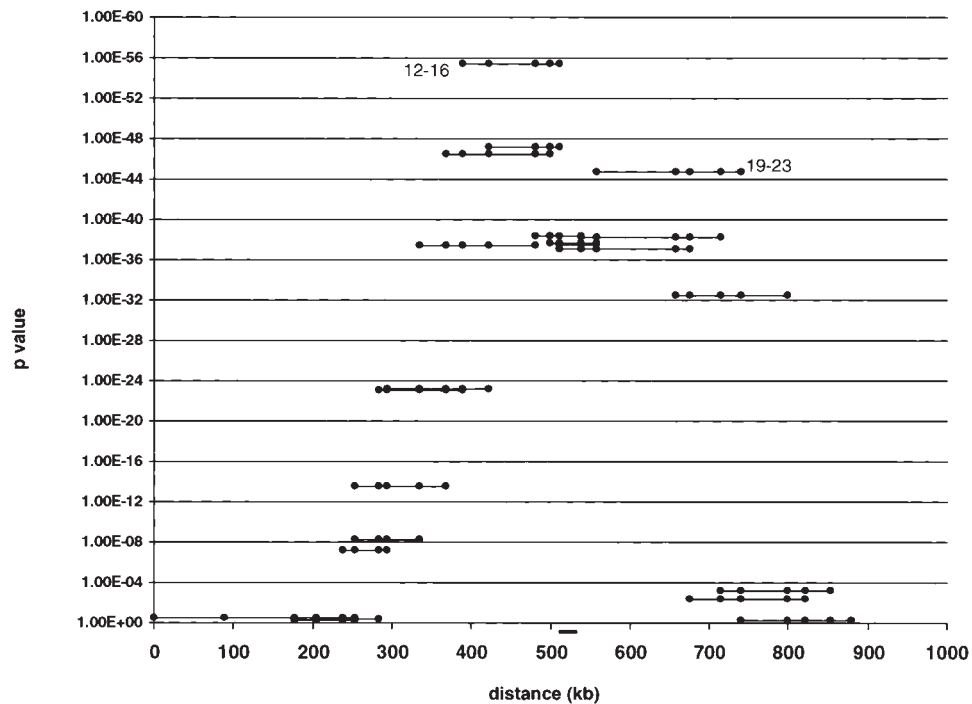


Figure 3 Association between haplotypes and the *CYP2D6* PM phenotype. Haplotypes were derived by EM algorithm from windows of five adjacent SNPs. Each window of five markers is represented by a horizontal line containing five data points, and marker sets 12–16 and 19–23 are indicated. Due to the close physical proximity of some SNPs, the resolution is not sufficient to clearly discriminate all markers. Marker pairs 6/7 and 16/17 appear superimposed in this figure at approximate map locations of 253 and 510 kb, respectively. Associations between estimated haplotype frequencies and the *CYP2D6* PM phenotype are described by *P* values. *CYP2D6* is indicated by the short horizontal bar underneath the *x* axis.

Table 2 Examples of different haplotype frequencies in cases and controls

SNP set	Haplotypes	Frequency (controls) ^a	Frequency (cases) ^b
12–16	1_2_2_1_1	0.46	0.10
	1_1_1_2_2	0.22	0.00
	2_1_1_1_1	0.17	0.83
	2_1_1_2_1	0.09	0.00
19–23	2_1_1_1_1	0.24	0.06
	1_1_2_2_2	0.17	0.02
	2_1_1_2_2	0.15	0.00
	1_1_1_1_1	0.11	0.03
	1_2_1_2_2	0.09	0.39
	1_2_1_1_1	0.07	0.27
	1_1_2_1_1	0.06	0.00

^aFrequencies were estimated for haplotypes derived from two of the most highly significant marker sets (12–16 and 19–23, $P = 3.54 \times 10^{-56}$ and $P = 1.82 \times 10^{-45}$, respectively) and all haplotypes with estimated frequencies >5% in all samples are listed.

med to determine whether LD mapping could be applied to a genome-wide search for genetic variation underlying altered drug response. In brief, genetic variation flanking *CYP2D6* was identified and correlated to the recessive *CYP2D6* PM phenotype.

There has been much debate regarding the density and frequency of SNPs, and the LD between them that is required to detect association. A simulation study³ estimated that useful LD extended for only 3 kb, and therefore one SNP every 3 kb would be required to detect association. However, empirical data have shown that LD varies throughout the genome,^{20,22–24} and experimental data suggest that fewer SNPs are required to detect association.^{5,8} In this study 27 SNPs were identified over an 880 kb region flanking *CYP2D6*, with an average spacing of one every 34 kb. Highly significant association ($P < 10^{-7}$) between the recessive *CYP2D6* PM phenotype and 10 SNPs was detected within 390 kb surrounding *CYP2D6*. Consistent with other SNP association studies,²⁵ SNPs furthest from the causal polymorphisms showed least association with the phenotype and mapped outside the 390 kb region bounded by LD. The strongest level of single marker SNP association ($P = 4.54 \times 10^{-35}$) with the *CYP2D6* PM trait was observed 120 kb away from *CYP2D6* (SNP 20).

At present, the number of samples required to detect genetic association in drug trials has not been fully determined.¹² As expected, power is reduced when the number of controls or total number of samples are reduced. The significance of the association between SNP 12 and the *CYP2D6* PM phenotype, decreased from $P = 7.16 \times 10^{-25}$ to $P = 1.59 \times 10^{-13}$, when analysing 977 controls or 41 con-

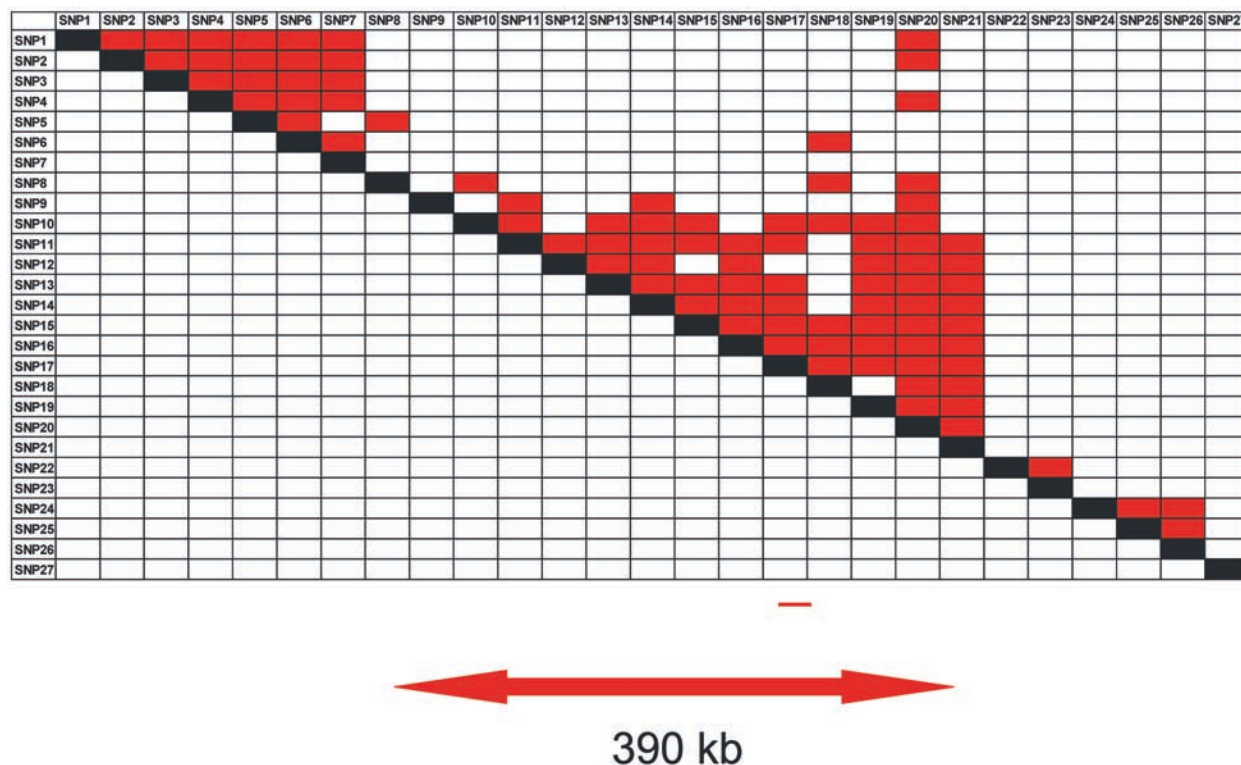


Figure 4 Linkage disequilibrium between 27 SNPs flanking *CYP2D6*. Strong LD ($D' > 0.7$) was detected across a 390 kb region between SNPs 8 and 21, indicated by the red arrow. Absolute D' values greater than 0.7 are shaded in red. The relative location of *CYP2D6* is indicated by the short horizontal red bar.

trols, respectively. Haplotype analysis has previously shown increased power over single marker analysis.^{7,25,26} A simulation study demonstrated that four haplotypes constructed from two neighbouring SNPs gave increased power to detect association over SNPs analysed singly.²⁶ Martin *et al* confirmed this experimentally, observing that 2–3 SNP haplotypes surrounding *APOE*, were more significantly associated with Alzheimer's disease than single SNPs.²⁵ Recently, a common haplotype of the cytokine region on chromosome 5q31 flanking *IBD5*, has been shown to confer susceptibility to Crohn disease,⁷ even when the causal polymorphism has not been identified. It is important, therefore to ensure the SNPs selected capture the common underlying haplotypes. In this study, haplotype frequencies from five markers were estimated using the EM algorithm, and compared between cases and controls. Haplotype frequencies differed between cases and controls (Table 2). Two haplotypes derived from marker set 12–16, (1_1_1_2_2 and 2_1_1_2_1), were present in control chromosomes with frequencies of 0.22 and 0.09, respectively, but were completely absent in case chromosomes. These differences were reflected in association analysis. Highly significant association ($P < 10^{-7}$) of the haplotypes (Figure 3) and the PM phenotype was detected over the 390-kb region demonstrating significant LD flanking *CYP2D6*. The level of significance increased dramatically ($P = 3.54 \times 10^{-56}$) from levels achieved with single marker analysis ($P = 4.54 \times 10^{-35}$),²⁵ consistent with Martin *et al*²⁵

underlining the importance of haplotype determination and analysis in association studies.

LD mapping identified three discrete blocks of LD across an 880 kb region on chromosome 22, the major one being a 390 kb region from SNP 8 to SNP 21 spanning *CYP2D6*. It is known that the extent of LD across the human genome varies enormously. Other recent studies^{21,24} reported LD blocks ranging from 10 kb to 300 kb. The reason for the broad range of LD identified in this study across the *CYP2D6* gene is unclear. LD is also influenced by many factors such as selection, local mutation rates, changes in population size, admixture and local rates of recombination, and generally extends further in regions of low recombination.²⁷ The sex-averaged recombination rate for three STS markers mapping to this region ranged between 0.83–0.85 cM Mb⁻¹,²⁷ which was lower than the recombination rates across the rest of the chromosome. The molecular basis for recombination remains unknown, but contributory factors may include high GC content,²⁷ which in this locus appeared to increase at the approximate location of SNPs 8 and 21. All SNPs within the 390 kb showed significant association with the PM phenotype, suggesting that the density of SNPs in such regions can be reduced. Identification of chromosomal regions harbouring genetic variation associated with drug metabolising phenotypes may, therefore, require fewer SNPs in areas of low recombination, and more SNPs in recombination hot-spots.

LD mapping has been informative in this study, correctly determining a large (390 kb) chromosomal region associated with *CYP2D6* poor drug metabolism. Ten SNPs demonstrating genome-wide significant ($P < 10^{-7}$) association mapped within the 390 kb region. Therefore, if the SNP density had been reduced so that only one of these 10 SNPs had been selected for a genome-wide association scan, the chromosomal region underlying the *CYP2D6* PM phenotype would still have been detected. There are, however, disadvantages in analysing broad regions of strong LD. Six other transcripts also map within the 390 kb region harbouring *CYP2D6*, and as in other studies⁷ the mapping performed in this study would not have been able to identify the causal gene variants. In addition, although haplotype analysis confirmed and, in fact, proved more powerful than single marker analysis for detecting association between SNPs and the *CYP2D6* PM phenotype, it was not able to further refine the broad 390 kb region identified, as markers in the region are in LD with each other (Figure 4). Knowing which chromosomal regions are associated with altered drug response could serve two purposes. Firstly, to aid the identification of novel genetic variation underlying altered drug response, and secondly to aid the identification of individuals potentially exhibiting altered drug response even if the causal polymorphisms are unknown. Ultimately, medicine treatment may be modified to potentially avoid certain adverse drug events in particular patients.²⁸

The pattern of LD observed may alter if several common polymorphisms all contributed equally to the PM trait. In this instance, the major *CYP2D6* causal polymorphism (G1846A) is present at a much higher frequency (~19%) than four other *CYP2D6* PM determining polymorphisms (0.05–2%). Therefore, in contrast to some drug metabolising enzymes where several common alleles give rise to a similar drug metabolising phenotype,²⁹ the *CYP2D6* PM phenotype should be considered as an essentially monogenic trait with additional small genetic contributions from three low allele frequency polymorphisms. Data from multivariant pharmacogenetic loci, which may exemplify the value of LD mapping in pharmacogenetic studies, have yet to be generated.

In summary, the right SNPs at the right physical locations with an appropriate range of allele frequencies must be genotyped in sufficient case and control samples in order to detect association. These data suggest that LD estimations and genome-wide association scans may be feasible for mapping essentially monogenic pharmacogenetic loci, necessary for the subsequent identification of genetic variation underlying medicine response.

MATERIALS AND METHODS

Samples

One thousand and eighteen Caucasians were included in the study. Two hundred and thirty samples originated from CEPH or Coriell cell repositories (Camden, NJ, USA). The remaining samples were taken from patients in Glaxo-SmithKline (GSK) Clinical Pharmacology studies with consent for non-identified genotyping: 459 from North America and 329 from the United Kingdom. DNA was isolated from

GSK samples from 10 ml of blood by Cambridge Molecular (Cambridge, UK).

Polymorphism Identification

One hundred and twenty-three SNPs were identified within an 880 kb region surrounding *CYP2D6* from The SNP Consortium (TSC) (<http://snp.cshl.org/>) release 5 (Sept 2000) and dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/>) public databases: 88 from TSC and 35 from dbSNP. Seventy-five SNPs were selected for even distribution across the region, and were verified in a panel of 16 Caucasians. Twenty-three of the 75 SNP primer sets did not produce PCR products from the panel of 16 Caucasian individuals. A further nine primer sets did not produce any sequence information. However, 39 PCR products from the remaining 43 putative SNPs were sequenced and confirmed as real polymorphisms. In regions of low SNP density, 72 genomic fragments were sequenced in the panel of 16 Caucasians and 36 SNPs were verified, giving a total of 75 SNPs in the region. Forty verified SNPs were then selected based on their physical location and allele frequency and genotyped on the entire sample.

Genotyping

Genotyping was performed in Germany using the mass spectrometry MALDI-TOF technology, applying an allele specific primer extension reaction to amplified genomic DNA.^{30–32} PCR was performed in 10 μ l volumes in 384-well plates, using 35 ng of genomic DNA, 5 pmol of each primer (Table 3), 1 \times PCR buffer (ABGene, Epsom, UK), 1.5 mM MgCl₂, 200 μ M dNTPs (Amersham Pharmacia Biotech, Piscataway, NJ, USA), 0.4 U Thermo-Start DNA Polymerase (ABGene). Amplifications were carried out in a Primus 384 thermocycler (MWG-Biotech), thermal cycle: 95°C for 15 min, followed by 95°C for 1 min, annealing temperature (depending on SNP) for 1 min, 72°C for 1 min, repeated 42 times, ending with 72°C for 10 min and rapid cooling to 8°C. Allele specific primer extension (Table 3) was performed in 384-well plates by combining 10 μ l purified PCR product with 20 pmol of specific extension primer (MWG-Biotech), 1 μ l Thermo Sequenase DNA polymerase, 2 μ l Thermo Sequenase reaction buffer, 2 mM of dNTPs and ddNTPs, respectively (all Amerham Pharmacia Biotech), to a total volume of 20 μ l. The resulting mixture was subjected to 40 rounds of thermal cycling (94°C for 10 s, annealing temperature for 20 s, 72°C for 30 s). dNTPs and ddNTPs were selected to give an extension of a single base for one allele and two bases for the other allele. PCR and extension products were purified using MultiScreen384 PCR and SNP filter plates (Millipore, Bedford, MA, USA), respectively, on a dedicated automated robotics platform (RoboSNiP 1600, MWG-Biotech). Mass spectra were acquired on a linear MALDI-TOF MS (Bruker Daltonics, Billerica, MA, USA) in positive mode at 19 kV total extraction voltage and 17.4 kV delayed extraction. Twenty shots were accumulated per spectrum. Samples were prepared on AnchorChip targets (Bruker) using a 3-hydroxypicolinic acid matrix. Genotypes were identified from the acquired spectra using the Genotools software (Bruker) for automated spectra evaluation.

Table 3 PCR and extension primer sequences used for MALDI-TOF assays

SNP ID	PCR forward primer	PCR reverse primer	Extension primer	Fragment length
1	GTCTCATCATGAGGACAGAACAGG	TCTCCTTCAGTTTTCTCTGGC	CCCGTGAATAATGGAAAAA	316
2	GTGCGAAACTGTCTCAAGC	CTCTTGACACATGTCCAAGGC	GAGGAGCCTCCAGGAATCA	322
3	CGACCTTAAAGTGACAGCATG	GTGTGCCCATCACTAGATTG	TGGATAAGTCCACTGCTCT	272
4	GCAGTACAAGATCCTTGATG	CTTCAAGTCTTACGTGTGGC	CCTGTCCCTGAACACCAACC	170
5	CTTCACCTGCAGCATACCAC	GTTTCGTAGTCACCTCTAC	ACCTGGAGCCCAGTACAAAA	270
6	GCTAAGGAAGGAGACTTGAG	CATCCTTCAACTGAGAAGCC	ACCTCCTCCAGACTCTTCGAC	678
7	GCTAAGGAAGGAGACTTGAG	CATCCTTCAACTGAGAAGCC	ACCCCTTGAAGGACATGATAC	678
8	GGACTATCTGTCGTCTATTG	CAGCATCAAGTACCTCTGTAG	CTATGTTGCTGCTAGACG	640
9	CACTGGTGGTTCACAATCAG	CCATCACTAGTGCCGAATGC	TGCCAGGGTTGATGGTGGGG	251
10	GTAGCTTCCAAGTGCCTCAG	CAGCATCGTGTCTCCAGC	GATAGGAAGGAGGGAGGG	399
11	GGTTCAGTTGTGCCAGTAA	GCATTAGTGTGTCCAGGGC	GGGAAGACCACACTGGGCA	143
12	GTGCTTGCCAAATGATTTGC	CACGACATCCTTCTAGACGC	TCTCTGTGTTACTTCCAACC	264
13	GGAAGAATACCATTGTGGGG	GCATCTTTGTGCACACATGC	ACAGGTGTATAATACTCCA	688
14	TTACAAGTAGCGGAGCCAAGCCTTTG	CTCAGAGCCTTGCATGTGAC	GATAATCTGTGCCCTGACACT	277
15	GTGCTGCTTGTGTCAATCAC	GTTGTGTACACCATGCAAC	GAAGGGAAATGTGAGTTACA	577
16	CTTATGAGAGGTGACCAGAG	CTTGGTTACCCACTCTTAC	GAAGGGTGTGCTTGTGTCA	400
17	CTTATGAGAGGTGACCAGAG	CTTGGTTACCCACTCTTAC	TCACGATTGCAAGAGCACAC	400
18	GTTCTGTCTACATGGCAGC	CCTCATGCTGCCATATCTTG	ACTCTTGGCCTTTTGAGGTG	282
19	CTGGTAAGAATGGTCCC	CTGAGCTTTGCTGCTTGAC	CATTAGGGTGTGCTGACAGGC	179
20	GCTCAGAGACTTGACAAAAGC	GTCTTCTCACTTGCCAGCTC	TGTGACTGAGTTAGCAGGC	381
21	GCTCCATGATCCTCCCTGAC	GCAGAGAGCCCAGCATGAC	CCCCATTAGTGCTCACGGG	208
22	CTTCATTAGGGTAATCACGGGGCTC	CCACAAGGCTATGTGGCTTTGGT	TCCCGGCAGAGACCAGTCC	330
23	GTGCTACCTTGATTTGAGC	GCTGCTACCATCATCTTGC	CAGGTTTCTGTCTGCATGAG	173
24	GTACCAGTTAGGTCACAGGC	GACCTGTGTACACCAAGC	GCCCCGGGGTCTAGAGG	265
25	GCAAGATGCTTTTTAGGACC	GCAGCTCAGAGATGGCCTGGC	CCAGTCTGCCCTCCATCGT	356
26	GCAAGAGAAAGCACAGGCTTG	GTGCTCTTGGCAGTAACAGCC	CAAGGAACTGAACACAGCT	371
27	GGTGTGATAGTCTCAATGTATC	GCATATACACATCCAGAGAGAAC	CTCACTTTGAGACCGAGTTAT	572

Four polymorphisms within *CYP2D6* (G1846A, delT1707, A2935C and delA2549) were genotyped using Taqman technology to identify individuals with predicted *CYP2D6* PM phenotype within the 1018 Caucasian samples.³³

Statistical Analysis

Genotype frequencies of each marker were examined for significant ($P < 0.01$) deviation from Hardy–Weinberg equilibrium. LD was measured using the standardised D' first described by Lewontin.³⁴ D' could be described as a normalised value of LD. It is LD relative to the maximum value for a given set of allele frequencies at a pair of sites. It is calculated by dividing the raw D value by the absolute maximal value possible.

Fisher's exact test was employed to analyse association between genotypes and phenotype using the PROC FREQ procedure implemented in SAS statistical analysis software. PROC FREQ computes exact P -values (significant $P < 0.01$) for general $R \times C$ tables using the network algorithm developed by Mehta and Patel.³⁵ This algorithm provides a substantial advantage over direct enumeration, which can be very time-consuming and feasible only for small sample numbers. To determine the effect of reducing control numbers on levels of significant association, analysis was performed using randomly selected decreasing numbers of controls ($n = 977, 205, 164, 123, 82$ and 41) and 41 cases. This resulted in case:control ratios of 1:24, 1:5, 1:4, 1:3, 1:2 and

1:1, respectively. Analysis was also performed on decreasing total sample numbers, each group of which contains a constant 4% proportion of PMs. Multiple testing was considered and corrected for using the Bonferroni adjustment, while recognising that a methodology adjusting for correlated tests would be more appropriate.³⁶

Haplotype frequencies were estimated through the expectation maximization (EM) algorithm from unrelated individuals.³⁷ The multilocus genotypes from each individual were used to enumerate all possible haplotypes, which were assigned multiple random starting frequencies sampled from the multivariate uniform distribution. A sample that yielded the maximum EM likelihood was taken. The frequencies were updated with frequencies calculated from all the possible haplotypes from each individual in the sample, continuing until the frequencies were constant from iteration to iteration. The following regression-based model (Zaykin *et al*, personal communication) was used to relate inferred haplotype probabilities to the PM phenotype for each individual:

$$Y = \mu + Xb + e$$

where Y is a column vector ($1 \times k$) containing the PM status for each person, μ is the overall population mean, X is a matrix ($k \times h$) containing k row vectors each consisting of the EM-inferred haplotype probabilities conditional on each individual's genotype, b is a vector ($h \times 1$) of haplotype

parameters, and e is a column vector ($1 \times k$) of random error terms for each individual. The hypothesis of no overall association of haplotypes with the trait is tested as $H_0: b_k = 0$, and the individual haplotype k can be tested as $H_0: b = 0$. D'Agostino³⁸ and Weir³⁹ discuss the relation and asymptotic equivalence between ANOVA performed on the binary phenotype and chi-square contingency table tests. Extrapolation of this (Zaykin *et al*, personal communication) has demonstrated equivalence between ANOVA and the regression test used to generate P values in this regression model. Therefore, an F statistic was calculated to estimate P values for these tests. Haplotypes were constructed in this manner from 27 SNPs spanning the entire 880 kb region and consisted of marker sets of five neighbouring SNPs. A series of 27 sequential overlapping haplotypes was derived. Markers within *CYP2D6* were not included in haplotype analysis. Association between derived haplotypes and the PM phenotype was determined, allowing incorporation of information from multiple markers which are physically close on a chromosome.

ACKNOWLEDGMENTS

We would like to thank Shela Varsani and Ros Cutts for data handling, Mike Stubbins for helpful discussions about the *CYP2D6* variants, and Dmitri Zaykin for help with the haplotyping methodology. We thank Linda McCarthy for her critical reading of the manuscript. We are grateful for the GSK sequencing group for the assistance in identification and verification of the SNPs, and Yingkun Brunner, Dajana Preuss and Elvira Deravanessian for assistance with the MALDI-TOF genotyping.

DUALITY OF INTEREST

None declared.

CONSENT

Informed consent was obtained from all genotyped individuals included in the study.

REFERENCES

- Snarey A, Thomas S, Schneider M, Pound S, Barton N, Wright A *et al*. Linkage disequilibrium in the region of the autosomal dominant polycystic kidney disease gene (PDK1). *Am J Hum Gen* 1994; **55**: 365–371.
- Kerem B, Rommens J, Buchanan J, Markiewicz D, Cox T, Chakravarti A. Identification of the cystic fibrosis gene: genetic analysis. *Science* 1989; **245**: 1073–1080.
- Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Gen* 1999; **22**: 139–144.
- Rannala B, Reeve J. High resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *Am J Hum Gen* 2001; **69**: 159–178.
- Martin E, Gilbert J, Lai, Riley J, Rogala A, Slotterback B *et al*. Analysis of association at single nucleotide polymorphisms in the *APOE* region. *Genomics* 2000; **75**: 7–12.
- Horikawa Y, Oda N, Cox N, Li X, Orho-Melander M, Hara M *et al*. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet* 2000; **26**: 163–175.
- Rioux J, Daly M, Silverberg M, Lindblad K, Steinhart H, Cohen Z *et al*. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* 2001; **29**: 223–228.
- McCarthy L, Hosford D, Riley J, Bird M, White N, Hewett D *et al*. Single nucleotide polymorphism alleles in the insulin receptor are associated with typical migraine. *Genomics* 2001; **78**: 135–149.
- Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **27**: 1516–1517.
- Marshall E. Drug firms to create public database of genetic mutations. *Science* 1999; **284**: 406–407.
- Cardon L, Bell J. Association study designs for complex diseases. *Nat Rev Genetics* 2001; **2**: 91–99.
- Elston R, Idury R, Cardon L, Lichter J. The study of candidate genes in drug trials: sample size considerations. *Statistics in Medicine* 1999; **18**: 741–751.
- Kuivenhoven J, Jukema J, Zwinderman A, de Knijff P, McPherson R, Brusckhe A *et al*. The role of a common variant of the cholesteryl ester transfer protein gene in the progression of coronary atherosclerosis. The Regression Growth Evaluation Statin Study Group. *New Eng J Med* 1998; **338**: 86–93.
- Nebert D. Polymorphisms in drug-metabolising enzymes: what is their clinical relevance and why do they exist? *Am J Hum Gen* 1997; **60**: 265–271.
- Evans W, Relling M. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* 2000; **286**: 487–491.
- Marez D, Legrand M, Sabbagh N, Guidice J, Spire C, Lafitte J *et al*. Polymorphism of the cytochrome P450 CYP2D6 gene in a European population: characterisation of 48 mutations and 53 alleles, their frequencies and evolution. *Pharmacogenetics* 1997; **7**: 193–202.
- Griese E, Zanger U, Brudermanns U, Gaedigk A, Mikus G, Morike K *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.
- Stuven T, Griese E, Kroemer H, Eichelbaum M, Zanger U. Rapid detection of CYP2D6 null alleles by long-distance and multiplex-polymerase chain reaction. *Pharmacogenetics* 1996; **6**: 417–421.
- Sachse C, Brockmoller J, Bauer S, Roots I. Cytochrome P450 2D6 variants in a Caucasian population: allele frequencies and phenotypic consequences. *Am J Hum Gen* 1997; **60**: 284–295.
- Reich D, Cargill M, Bolk S, Ireland J, Sabeti P, Richter D *et al*. Linkage disequilibrium in the human genome. *Nature* 2001; **411**: 199–204.
- Daly M, Rioux J, Schaffner S, Hudson T, Lander E. High-resolution haplotype structure in the human genome. *Nat Genet* 2001; **29**: 229–232.
- Taillon-Miller P, Bauer-Sardina I, Saccone N, Putzel J, Laitinen T, Cao A *et al*. Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat Genet* 2000; **25**: 324–328.
- Moffatt M, Traherne J, Abecasis G, Cookson W. Single nucleotide polymorphism and linkage disequilibrium within the TCR alpha/delta locus. *Hum Mol Gen* 2000; **9**: 1011–1019.
- Abecasis G, Noguchi E, Heinzmann A, Traherne J, Bhattacharyya S, Leaves N *et al*. Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Gen* 2001; **68**: 191–197.
- Martin E, Lai E, Gilbert J, Rogala A, Afshari A, Riley J *et al*. SNPing away at complex diseases: analysis of single nucleotide polymorphisms around *APOE* in Alzheimer's disease. *Am J Hum Gen* 2000; **67**: 383–394.
- Zollner S, von Haeseler A. A coalescent approach to study linkage disequilibrium between single nucleotide polymorphisms. *Am J Hum Gen* 2000; **66**: 615–628.
- Yu A, Zhao C, Fan Y, Jang W, Mungall A, Deloukas P. Comparison of human genetic and sequence-based physical maps. *Nature* 2001; **409**: 951–953.
- Roses A. Pharmacogenetics and the practice of medicine. *Nature* 2000; **405**: 857–865.
- Lee F, Zhao B, Seow-Choen F. Relationship between polymorphism of N-acetyltransferase gene and susceptibility to colorectal carcinoma in a Chinese population. *Pharmacogenetics* 1998; **8**: 513–517.
- Braun A, Little D, Koster H. Detecting CFTR gene mutations by using primer oligo base extension and mass spectrometry. *Clin Chem* 1997; **43**: 1151–1158.
- Little D, Cornish T, O'Donnell M, Braun A, Cotter R, Koester H. MALDI on a chip: analysis of arrays of low-femtomole to subfemtomole quantities of synthetic oligonucleotides and DNA diagnostic products dispensed by a piezoelectric pipet. *Anal Chem* 1997; **69**: 4540–4546.
- Ross P, Hall L, Smirnov I, Haff L. High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nat Biotechnol* 1998; **16**: 1347–1351.
- Livak K, Marmaro J, Todd J. Towards fully automated genome-wide polymorphisms screening. *Nat Genet* 1995; **9**: 341–342.

- 34 Lewontin R. The interaction of selection and linkage. *Genetics* 1964; **49**: 49–67.
- 35 Mehta C, Patel N. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J Am Stat Assoc* 1983; **78**: 427–434.
- 36 Westfall P, Zaykin D, Young S. *Biostatistical Methods*. Humana Press: New Jersey, 2001.
- 37 Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc* 1977; **39**: 1–38.
- 38 D'Agostino R. Relation between the chi-squared and ANOVA tests for testing equality of k independent dichotomous populations. *The American Stat* 1972; **26**: 30–32.
- 39 Weir BS. *Genetic Data Analysis II*. Sinauer Associates: Massachusetts, 1996.