Group-based pharmacogenetic prediction: is it feasible and do current NHS England ethnic classifications provide appropriate data?

Catherine J. E. Ingram[1*], Rosemary Ekong[1*$], Naser Ansari-Pour[2* $], Neil Bradman[3$] and Dallas M Swallow[1**].

[1] Research Department of Genetics, Evolution and Environment
University College London, Darwin Building, Gower Street, London WC1E 6BT, UK

[2] Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7LF, UK

[3] Henry Stewart Group, 40-41 Museum Street, London WC1A 1LT

*   contributed equally

**corresponding author d.swallow@ucl.ac.uk; 02076797320; Orcid ID: 0000-0001-7310-2735

$ R. Ekong Orcid ID: 0000-0002-5984-7893

N. Ansari-Pour Orcid ID:0000-0003-0908-0484

N. Bradman Orcid ID: 0000-0003-1408-5853

ABSTRACT

Inter-individual variation of drug metabolising enzymes (DMEs) leads to variable efficacy of many drugs and even adverse drug responses. Consequently, it would be desirable to test variants of many DMEs before drug treatment. Inter-ethnic differences in frequency mean that the choice of SNPs to test may vary across population groups. Here we examine the utility of testing representative groups as a way of assessing what variants might be tested. We show that publicly available population information is potentially useful for determining loci for pre-treatment genetic testing, and for determining the most prevalent risk haplotypes in defined groups. However, we also show that the NHS England classifications have limitations for grouping for these purposes, in particular for people of African descent. We conclude: (a) genotyping of hospital patients and people from the hospital catchment area confers no advantage over using samples from appropriate existing ethnic group collections or publicly available data, (b) given the current NHS England Black African grouping, a decision as to whether to test, would have to apply to all patients of recent Black African ancestry to cover reported risk alleles and (c) the current scarcity of available genome and drug effect data from Africans is a problem for both testing and treatment decisions.

INTRODUCTION

Over the past 50 years, pharmacogenetic studies have identified a large number of genetic variants that influence drug response [1, 2]. Some of these variants may be associated with adverse drug reactions (ADR) [3-5] while others may alter drug efficacy [6]. Thus genotyping of such single nucleotide polymorphisms (SNPs) can be useful for determining dose and drug suitability [7, 8]. ADRs have been estimated to cost the UK National Health Service (NHS) in excess of £600m per year [9], and it is hoped that applying pharmacogenetics more widely can reduce the economic burden by reducing the number of ADRs, as well as avoiding ineffective or unnecessary therapy [10].

Advice regarding pharmacogenetic testing, guidance on specific conditions and directives on drug labelling is already offered in many countries. In the US, the Food and Drug Administration (FDA) continuously updates an extensive document on drug labelling (https://www.fda.gov/downloads/Drugs/ScienceResearch/UCM578588.pdf). The labelling of some medications (e.g. Maraviroc, Cetuximab, Trastuzumab, Dasatinib) makes clear that genetic testing is required or in some cases simply recommended prior to their administration (e.g. Warfarin, Irinotecan). In the UK, the Summary of Product Characteristics (SmPC) for a drug are approved by the European Medicines Agency in association with MHRA, but do not necessarily include pharmacogenetic advice.

There is well documented interethnic variation in allelic distribution of genes affecting drug metabolism, with metabolising enzymes, receptors and transporters all implicated [11-13]. Genetic variation in drug metabolising enzymes (DMEs) that occurs at significantly different frequencies [14] or is private to definable populations [12] may result in varying risk for a given ADR. Ethnic identity has therefore been proposed as a source of information in medicinal intervention [15], as a potentially useful indicator for assigning patients into groups at high and low risk of ADR, and has already been used to select patients for genetic testing to prevent ADRs prior to the administration of carbamazepine [16]. Carbamazepine (and various other drugs) can trigger severe dermatological reactions that are associated with carrying the *HLA-B*1502* allele, in the Han Chinese population, where it is found at high frequency. FDA-approved labelling and also British National Formulary (BNF) (https://bnf.nice.org.uk/drug/carbamazepine.html#preTreatmentScreeningInformation) recommend testing for this variant in high risk (i.e. Han Chinese and Thai origin) individuals.

Hospitals in England collect ethnic or country of origin information from all their patients, as is also the practise in many other countries. In anticipation of the possibility of a decision by the NHS to undertake widespread genetic testing, (https://www.england.nhs.uk/wp-content/uploads/2018/08/national-genomic-test-directory-faqs.pdf) we have addressed the following questions: (a) Can the ethnic/country of origin labels currently in use be applied as a way of deciding whether it is appropriate to genotype only patients of groups known to have specific therapeutically relevant variants, but not patients belonging to other ethnic groups, or should the genotype of all patients who will require the therapeutic intervention be ascertained? (b) Is it necessary to genotype samples from representative patients attending a hospital or living within a hospital's catchment area itself to establish what SNPs should be tested; or is it satisfactory to genotype samples from existing 'ethnic group' collections, or use genotypes reported in existing databases? Using the NHS England classification system, the three patient groups most frequently treated at University College Hospital (UCH) in London are White British, Black Africans (notably a very broad label), and Bangladeshi, so we focused on these.

As a proof of concept, we examined 39 therapeutically relevant SNPs in six key genes of phase I and phase II of drug metabolism in samples that represented three frequently encountered UCH patient groups. We compared three data sets: (a) those from samples collected in UCH and UCH catchment area, (b) a set of samples that we have 'in-house', collected in the countries of origin, and (c) publicly available data from the 1000 Genomes Project [17]. In view of the fact that the NHS England Asian groups are separated by country rather than by continent, as is done for the Black Africans, we also examined groups which, in the NHS England classification, would be expected to self-classify as Indian (class H) and compared those with the Bangladeshi (class K) groups to examine variation among groups from the same continent. In addition, we have examined population-specific haplotype configurations of the therapeutically relevant variants, since phase affects predicted activity levels.

RESULTS

Selection of groups to study

Prior to volunteer recruitment we obtained from the Department of Information and Communication Technology at UCH (UCH ICT) patient ethnicity and country of origin data for all patients that had been recorded during the two-year period (2006-2007) and also outpatients, for six months (March to October 2009) at the start of the project. This analysis showed that White British, Black African and Bangladeshi categories were the top self-declared ethnicities of patients and was in agreement with a 2009 survey for Camden, now superseded by the 2011 census data (Table 1). On the basis of this we opted to focus the study on these. In addition to the 236 samples collected for the project, we tested 295 that had been collected outside the UK in locations as near as possible to the immediate ancestral origin of the local London groups that we were testing (see Methods section). Also for comparison we extracted publicly available data (n=886 from the 1000G study) (Table 1).

Selection of therapeutically relevant SNPs to test

We identified from the literature those Phase I and Phase II drug metabolising enzymes most frequently cited in pharmacogenetic studies [18], and combined this information with the FDA Table of Pharmacogenomic Biomarkers in Drug Labelling (https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm) to short list priority genes to type in this study. From this the phase I enzymes CYP2C9, CYP2C19 and CYP3A5, and the Phase II enzymes UGT1A and NAT2 were selected (Table 2). Despite its importance in drug toxicity and efficacy CYP2D6 was not included because of the complexity of genetic variation observed in the gene, including the existence of extensive copy number variation and multiple functional alleles with unknown combinatory effects [19]. In addition, the phase I DME, FMO2, which carries a common loss of function allele that has reached fixation in Europeans, but is expressed in many Africans, was included, because of its potential importance in relation to TB therapy in Africans [20, 21]. A total of 53 SNPs from 6 genes

(Supplementary Table S1) was selected and typed in all samples: 39 of these have been judged to be therapeutically relevant (see Table 2 for clinically important drugs metabolised by enzymes encoded by the genes analysed here). The remaining 14 SNPs (located within and around the DMEs and with frequencies greater than approximately 0.1 worldwide (see Supplementary Table S2) were of no described function but were typed to assist with haplotype analysis. Since *UGT1A1*28* and the alleles *UGT1A1*36* and **37*, form a small 'microsatellite'(5 to 8 copies of a TA in the promoter), and thus are not available in the public data sets, this was replaced by rs887829 (*UGT1A1*80*) which is in high LD with it in all the populations that we have tested (TA7/8 vs TA6/5 r^2 >> 0.9; unpublished data and [22, 23]).

Prevalence of therapeutically relevant variation in the different groups

In all, genotypes of 1417 individuals were obtained for the 39 therapeutically relevant SNPs and 14 others. Analysis of the therapeutically relevant SNPs showed that 11 out of 39 were monomorphic in the groups we studied, of which 7 and 4 were in *CYP2C19* and *CYP2C9* respectively. These were all SNPs reported in the literature at very low frequencies, but were tested here because of the patchy non-European coverage of the published data. The frequencies of the 28 polymorphic therapeutically relevant variations are given in Table 3 (further details of these SNPs are given in Supplementary Table S3). Of these, rs72552267 (*CYP2C19*6*), was observed once (in a London white British individual) in the entire dataset. This variant was therefore removed from further analysis. Frequencies of the 14 other non-therapeutically relevant SNPs are given in Supplementary Table S2.

Table 3 shows that both the occurrence and frequency of many of the alleles differ between the Africans, White British and Bangladeshi. As previously reported in the review of Bains [37] there was more variation in the African than the other groups (i.e. more derived alleles were detected). The non-truncated form of *FMO2* (ancestral C allele at rs6661174, *FMO2*1*) was present in many of the Africans and did not occur in the White British, but was detected in the Bangladeshi group, an observation reported here for the first time.

The allele frequencies in the three London groups (White British, Black African and Bangladeshi) were, in most cases, broadly similar to the equivalent groups collected in their countries of origin and/or the 1000G samples (Table 3 and Supplementary Figures 1-7, which show confidence intervals). There were, however, a few marked differences between the African groups in the two genes *NAT*2 and *CYP3A5* (Table 3, and Supplementary Figure 1a and 2a). There were non-overlapping confidence intervals between the London collected Africans and the 1000G Africans, with the 'in-house' (country of origin) collection lying somewhere in between. For this reason, the 1000G samples and the 'in house' groups of Africans were subdivided and plotted by country of origin /ethnicity (Supplementary Figure 1b and 2b). This revealed considerable heterogeneity: for example, the Ethiopian Somali group from our 'in-house' collection was a clear outlier for the *NAT2* SNP rs1801280 (Supplementary Figure 1b) and also the *CYP3A5* SNP rs776746 (Supplementary Figure 2b).

In order to understand better what was driving the frequency differences between the combined African sets of the 1000 genomes samples and the very heterogeneous London collection (Supplementary Table S3), we grouped the London Africans from the Horn of Africa together (Ethiopia, Eritrea and Somalia), and compared the allele frequencies for these two SNPs with those of the Nigerians (the largest west African group). We observed highly significant differences between them (*NAT2*, rs1801280 T>C Horn of Africa 0.68; Nigeria 0.28, χ^2 P-value = 0.00007 for allele counts; *CYP3A5*, rs776746 A>G Horn of Africa 0.65; Nigeria 0.19, χ^2 P-value $\leq$ 0.00001). For both SNPs, the frequencies of the Horn of Africa set collected in London (0.68 and 0.65) were even greater than those of the Somalis collected in Ethiopia (frequencies below 0.6 and shown in Supplementary Figures 1b and2b).

Haplotypes

Because in many cases multiple low activity therapeutically relevant variants occur within a single gene, it was of interest to know whether the alleles occur on different haplotypes – i.e. whether they are in *cis* or *trans*. Where two low activity alleles occur in *trans* it is most likely that the compound heterozygotes will have activities in between those of the respective low activity homozygotes. In contrast, carrying two low activity alleles *in cis* means that the effect is much less; there will be just one low activity chromosome and one with the normal activity.

To construct haplotypes and to assist with accuracy of assignment, we made use of the full set of SNPs for each gene region and we obtained haplotypes for each gene separately, and also for *CYP2C9* and *CYP2C19* together, since they are adjacent to each other on the same chromosome (not shown): Table 4 shows the most informative haplotypes and their frequencies in all the main groups. The haplotypes that carry low activity therapeutically relevant variants are indicated in bold in Table 4 and it can be seen that for the CYP and NAT genes, the low activity alleles are mostly in *trans*-configuration and recombinant haplotypes are extremely rare. The inclusion of non-therapeutically relevant SNPs shows that the low activity *CYP3A5*3*, that varies in frequency so dramatically across populations, nevertheless mainly occurs on the same haplotype (T-CC**C**A) in all groups, and in *trans* with the rare low activity alleles *6 and *7.

In the case of *UGT1A*, there is much more recombination, and the addition of a non-functional SNP (rs2602381) located just upstream of the therapeutically relevant variants simply increased the number of haplotypes, without providing further information (not shown). Thus, in Table 4, rs2602381 is omitted and the shorter haplotypes are presented. The most frequent haplotypes are quite different in the different groups. This is due to allele frequency differences as well as differences in recombination in the Africans. It can be seen that the combination of the low activity *28/*27 (as indicated by the surrogate SNP rs887829 (*80) always occurs together with the low activity *60 allele (G) at rs4124874, but many people (in particular, Africans) carry *60 in the absence of *28/*27 (e.g. haplotype TGGC).

Analysis of the Indian groups

Although we had only collected 14 samples from Indians in our London collection (NHS England ethnicity class H), we generated data for this group since this is also a major ethnicity group in London (Table 1). We also tested other Indian groups available the 'in-house' collection of Gujarati Indians and two 1000G Indian groups (see Figures S3, S4c-7c). When compared with the Bangladeshi groups, the Indian groups showed similar frequencies and had the same set of therapeutically relevant variants with one very minor exception, (rs28371685 in *CYP2C9),* which was only observed once, in the large 1000 genomes ITU (Indian Telugu in the UK) group (Supplementary Table S3, Supplementary Figures S3-7). It is noteworthy that *FMO2*1* was, as in the Bangladeshi, also detected in the three large Indian groups, though not in the small London group probably because of sample size.

Discussion

This study confirms that there are some large differences in both allele occurrence and allele frequency between the different self-identified NHS England ethnic/ancestry groups. Consequently, it should, in principle, not be necessary to test everyone for some therapeutically relevant SNPs. However, the results of tests to compare allele frequencies between the London collected data and data obtained from other sources, highlighted the problem of non-representative sampling. Here we show that while both the White British and the Bangladeshi are very similar in terms of allele frequency, regardless of collection location or source of the data, this is not the case for the Black Africans, differing geographically as well as across ethnic groups as we also show here in the Supplementary Figures S1b, S2b, S4b – S7b. This is not surprising since it is well known that the African continent is genetically very diverse [38-40]. Also, in our experience, even collecting a representative cohort of samples from London Africans from UCH and its surrounding areas is difficult. Our own recruitment, in which we recorded more detail than the standard questions asked of NHS England patients was somewhat different in composition from that recorded in publicly available demographic statistics used to construct our 'in-house' data set https://data.london.gov.uk/dataset?q=nationality. These databases provide more information on the country of origin of London residents than does the NHS England categorisation, but other sources which provide data at borough level (https://data.london.gov.uk/dataset/detailed-country-birth-2011-census-borough) show that there is considerable heterogeneity across London.

In the case of two of the genes tested here (*CYP3A5* and *NAT2*), the evidence that we present shows major differences between East and West Africans, but fewer differences between peoples of the Indian subcontinent (which are subdivided into the different classes, Bangladeshi and Indians). This means that the current NHS England classifications are likely to be of limited utility for genetic testing purposes; it would be important to have more precise information about the ethnicity and country of origin of patients of African ancestry before using group-based testing.

Our study also highlights the lack of publicly available genome data to match the known African communities of London. It is thus important for the scientific community to obtain more African genome data of suitable quality from a wider range of geographic locations/ethnic groups that can be made publicly available.

In order to avoid adverse drug reactions or lack of efficacy it is more useful to know whether or not particular variants occur in a given population, rather than consider subtle differences in frequencies. However, a problem is that aside from the frequent occurrence of recent mixed ancestry, totally excluding the presence of such therapeutically relevant variants is not possible, and rare alleles may easily be missed even if data exists for hundreds of individuals (see the upper limit CIs for the 0 allele frequencies in the Supplementary Figures).

Since the risk of homozygosity for the deficiency alleles (sometimes called human knockouts, https://www.sciencemag.org/news/2017/04/human-knockouts-may-reveal-why-some-drugs-fail) is usually the most important issue it might in principle be possible to decide a threshold allele frequency below which testing is not needed. However, a careful evaluation is required for any gene/drug combination. Also, it has been reported that heterozygosity should be taken into account in drug dosing in many, if not most, cases. Supplementary Table S3 shows the list of SNPs together with the literature references for recommended dosing or suitability of the drug, where homozygotes and heterozygotes may or may not be recommended intermediate doses, or alternatives, according to the drug in question (see the Dean *et al* on-line reference links and other literature references in Supplementary Table S3). There is less detailed information for the functional impact of variants that occur only in individuals of African ancestry than there is for those present in people of European ancestry, but they seem to appear mainly in *trans* like the European variants, and it is probable that all low activity variants will behave much the same and their effects will be additive in compound heterozygotes.

The situation with *UGT1A1* is more complicated. Although allele frequencies vary for rs887829 (now known as *UGT1A1*80*) (Table 3), used as surrogate for the well-known **28/*27* alleles (rs8175347), it is common in most populations. Homozygosity or compound heterozygosity of **28/ (TA)7* and **27/ (TA)8,* seems, according to the literature, to be the most relevant therapeutically, leading to irinotecan toxicity, as well as high bilirubin levels, in response to a number of xenobiotics. A recent review which includes African Americans [41] suggests that *UGT1A1*60* and **93* in the phenobarbital response enhancer, which are in high LD with **28/*27* (in *cis*), have little independent effect on irinotecan toxicity, and results on bilirubin levels are conflicting, even though functional data shows evidence of altered transcriptional regulation for **60* [42]. It may be that **60* and **93* are important in the context of other drugs or other combinations of alleles, or different combinations of *CYP3A4* and *CYP3A5* alleles, since these enzymes are also involved in the metabolism of irinotecan. We show here that *UGT1A1*60* (rs4124874) is particularly frequent both in East and West Africa (Supplementary Figure S5), and while most people who have **28/*27* have **60,* even more people have**60* but not **28/*27*. So, it will be important for a future study on Africans to determine the extent to which **60* is associated with an effect on its own. In addition, there is a lack of good drug toxicity and bilirubin data from Africa. Without this additional information it seems, at present, that testing **28/*27* or a surrogate for it, is the best course of action in all the major groups that we studied here. *UGT1A7*3* is also in high LD with the *UGT1A1* loci in Europeans and Africans, making the effects hard to disentangle. There is, in contrast to *UGT1A1*60,* better evidence that *UGT1A7*3*, and also *UGT1A1*6* (that we did not study here), have an effect in East Asian patients, where they are common and where **28/*27* is rare [43, 44].

For each gene the risks are potentially different, both the importance and chances of missing therapeutically relevant variants are different and consequences of doing so vary with the administered drug. We have nevertheless attempted to suggest whether or not particular SNPs could be tested in restricted NHS England groups in Supplementary Table S3. For example, our analysis suggests that it would only be necessary to pre-test for *FMO2* in Africans and individuals from the Indian subcontinent and not white British, before ethionamide therapy. This however is speculative, since proper pharmacokinetic studies are needed. In fact, metabolic studies relevant to other genes are also lacking for many populations, in particular for Africans.

In conclusion, to answer the questions originally posed:
(*a) Can the ethnic/country of origin labels currently in use be applied as a way of deciding whether it is appropriate to genotype only patients of groups known to have specific therapeutically relevant variants, but not patients belonging to other ethnic groups, or should the genotype of all patients who will require the therapeutic intervention be ascertained?*

The ethnic/country of origin labels currently in use by NHS England appear useful for giving guidance as to whether or not to test for particular therapeutically relevant variants in white British and Bangladeshi/Indian patients. However, such guidance is more problematic for people classified as Black Africans. Given the current NHS England Black African grouping, a decision as to whether to test, would have to apply to all patients of recent Black African ancestry. Recording the country of origin and recent ancestry of patients currently recorded as Black African would be beneficial.

(b) Is it necessary to genotype samples from patients attending a hospital or living within a hospital's catchment area itself to establish what SNPs should be tested, or is it satisfactory to genotype samples in existing 'ethnic group' collections, or use genotypes reported in existing databases?

Genotyping of representative hospital patients and people from the hospital catchment area confers no advantage over using samples from appropriate existing ethnic group collections or publicly available data.

In summary, this study shows that information from public data can be valuable in assessing which loci might be useful for genetic testing for the purpose of drug administration, in order to maximise coverage, though more publicly available genome/DME data is needed, particularly for the African continent. For such information to be fully exploited, more appropriate ethnicity information should be collected by NHS England. Although, in the long term, whole genome sequencing may become cheap enough for all patients to be tested, much more knowledge is needed about the subtleties of the functional consequences of specific alleles and haplotypes in relation to particular drugs among peoples of different ancestries.

MATERIALS AND METHODS

Subjects

Volunteers were recruited primarily from the three frequently represented groups (self-declared White British, Black Africans and Bangladeshi) at clinics at UCH Outpatients Department, and from local community groups in the London area, particularly the London Borough of Camden in which UCH is located. Ethical approval was obtained by the National Research Ethics Service (Bromley Local Research Ethics Committee 09/H0805/33 and UCL Hospitals NHS Foundation Trust (09/0146). The volunteers provided fully informed consent and completed questionnaires, to provide more detail of their self-declared ethnicity. The paperwork and samples (mainly buccal swabs and a few blood samples) were coded, and both were anonymised. In total, 78 White British, 77 Black African, 67 Bangladeshi samples and 14 Indian samples were collected. The country of origin/ethnic breakdown of these groups is shown in Supplementary Table S4. Our prior power calculations had shown that we required a sample size of 40 individuals to have 95% confidence of detecting alleles occurring at an allele frequency of 0.1.

The samples labelled 'country of origin', were prepared from buccal swabs donated during the years 1998 to 2005 with informed consent from individuals in the countries listed in Supplementary Table S4. DNA extraction for the broad purpose of studying human genetic variation was conducted under ethics approvals of the joint UCL/UCLH Committees on the Ethics of Human Research Committee A (references 99/0196) and locally where such procedures existed at the time collections were made. Examples of earlier publication of data from these collections are given in Supplementary Table S4. The African samples were selected to most closely match, geographically, the three most frequent countries of origin listed in https://data.london.gov.uk/dataset?q=nationality (Nigeria, Ghana and Somalia).

Genomic DNA was extracted using standard protocols.

Publicly available data from the 1000 genomes (1000G) project (https://www.internationalgenome.org/) were used to act as further 'proxy' populations. We used the White British from Great Britain (GBR; n=91), the Bangladeshi (BEB; n=86) and all Black African groups (n=504) who would have self-identified as Black African, had they been seen in UCH. In addition, we retrieved data from two Indian datasets, the Gujarati Indians in Houston (GIH; N=103) and Indian Telugu in the UK (ITU; N=102). Further details are shown in Supplementary Table S4.

Genotyping

The selected SNPs were genotyped (blind of ethnicity, because coded) using KASP™ (Kompetitive Allele Specific PCR) (LGC Biosearch technologies) where possible. Sanger sequencing and RFLP were conducted 'in house' where KASP™ was problematic and protocols were optimised using samples with known genotypes. Details are given in Supplementary Table 1.

Data Handling

All the data files were converted to PLINK [45] format for further analysis. The dataset generated (N=531) was modified to create a PED file readable by PLINK. For this, the two indel variants (rs9332131 and rs41303343) were coded into an acceptable diallelic format identical to that reported in the 1000G dataset. A corresponding MAP file (containing locus information on the variants) was also generated. Since there was some missing data in the 'in-house' samples typed (7%) both the samples and the SNPs were filtered for more than 50% missing data by using the *mind 0.5* and *geno 0.5* commands respectively and seven samples were removed from the dataset leaving 524 individuals, but all therapeutically relevant SNPs were retained. The seven samples removed were two Somali, two Black African, one Asante, one Bangladeshi and one White British. For ease of handling, all three data sets were merged using the *merge* command. To enable this, five SNPs reported on the opposite strand in the 1000G data were reversed (to the chromosomal strand) in the 'in-house' data using the *flip* function.

Statistical analysis

Allele and genotype frequency, and Hardy-Weinberg equilibrium calculations were undertaken in PLINK 1.9 [45]. To compare the allele frequencies across different groups (both within and between NHS England ethnic categories), 95% CI of the frequencies were calculated with the standard error defined as $\sqrt{{pq}/{2N}}$ where p and q are the major and minor allele frequencies, and N is the sample size of the given group. The upper limit CI for zero observations (i.e. MAF=0) was calculated as the maximum frequency at which an allele might be present using the formula q=$1-(1-P)^{\frac{1}{2N}}$, where P is probability of being observed and is set as 95% and N is the sample size of the given group[46]. These calculations were implemented in the R environment for statistical computing [47].

Haplotypes

Haplotypes within each locus were inferred using PHASE version 2.1.1 under default conditions [48] using all SNPs within and around each gene, for each major population group separately (all African, all Bangladeshi and all British).

Code Availability

Data pertaining to the 1000G SNPs were extracted using VCFtools version 0.1.13 (https://vcftools.github.io/index.html), from the 1000G Phase 3 VCF files for the relevant chromosomes (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/).

For systematic and automated haplotype analysis of large SNP data, we developed an R-based tool to convert PLINK PED/MAP files to PHASE input, and to summarise haplotype inference results in multiple groups from PHASE output. This tool is now publicly available on Github (https://github.com/nansari-pour/PLINKtoPHASE).

Acknowledgements

We thank all the sample donors who participated in this study and the UCLH clinicians, Aroon Hingorani, Alastair Forbes, Simon Woldman, Steve Hurel, Clare Dollery and others who helped us with access to patient volunteers in their clinics.

We also thank Mark Thomas for access to some of the samples and Pieta Nosanea, Esther Williams, Sarah Steward and Ayele Tarekegn for help with sample collections from African and Indian volunteers in their native countries; Ranji Areseratnam for technical assistance.

This research was funded by the University College London Hospitals Comprehensive Biomedical Research Centre.

Declaration of interest

During this study NB had a controlling interest in a company interested in developing diagnostic technology to identify variation in drug metabolising enzymes to improve healthcare. Neither NB nor the company now have that objective. None of the other authors have any potential conflicts of interest to declare.

References

1. Weinshilboum RM, Wang L. Pharmacogenetics and pharmacogenomics: development, science, and translation. *Annu Rev Genomics Hum Genet* 2006; **7:** 223-245.

2. Daly AK. Pharmacogenetics: a general review on progress to date. *Br Med Bull* 2017; **124**(1)**:** 65-79.

3. Browning LA, Kruse JA. Hemolysis and methemoglobinemia secondary to rasburicase administration. *Ann Pharmacother* 2005; **39**(11)**:** 1932-1935.

4. Khan S, Mandal RK, Elasbali AM, Dar SA, Jawed A, Wahid M*, et al*. Pharmacogenetic association between NAT2 gene polymorphisms and isoniazid induced hepatotoxicity: trial sequence meta-analysis as evidence. *Biosci Rep* 2019; **39**(1).

5. Lonjou C, Borot N, Sekula P, Ledger N, Thomas L, Halevy S*, et al*. A European study of HLA-B in Stevens-Johnson syndrome and toxic epidermal necrolysis related to five high-risk drugs. *Pharmacogenet Genomics* 2008; **18**(2)**:** 99-107.

6. Perry CM. Maraviroc: a review of its use in the management of CCR5-tropic HIV-1 infection. *Drugs* 2010; **70**(9)**:** 1189-1213.

7. Stehle S, Kirchheiner J, Lazar A, Fuhr U. Pharmacogenetics of oral anticoagulants: a basis for dose individualization. *Clin Pharmacokinet* 2008; **47**(9)**:** 565-594.

8. Johnson JA, Cavallari LH. Warfarin pharmacogenetics. *Trends Cardiovasc Med* 2015; **25**(1)**:** 33-41.

9. Davies EC, Green CF, Taylor S, Williamson PR, Mottram DR, Pirmohamed M. Adverse

drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. *PLoS One* 2009; **4**(2)**:** e4439.

10. Dressler LG. Integrating personalized genomic medicine into routine clinical care: addressing the social and policy issues of pharmacogenomic testing. *N C Med J* 2013; **74**(6)**:** 509-513.

11. Hovelson DH, Xue Z, Zawistowski M, Ehm MG, Harris EC, Stocker SL*, et al*. Characterization of ADME gene variation in 21 populations by exome sequencing. *Pharmacogenetics and genomics* 2017; **27**(3)**:** 89.

12. Creemer OJ, Ansari-Pour N, Ekong R, Tarekegn A, Plaster C, Bains RK*, et al*. Contrasting exome constancy and regulatory region variation in the gene encoding CYP3A4: an examination of the extent and potential implications. *Pharmacogenetics and genomics* 2016; **26**(6)**:** 255-270.

13. Gurwitz D, Motulsky AG. 'Drug reactions, enzymes, and biochemical genetics': 50 years later. *Pharmacogenomics* 2007; **8**(11)**:** 1479-1484.

14. Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG*, et al*. Population genetic structure of variable drug response. *Nat Genet* 2001; **29**(3)**:** 265-269.

15. Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ*, et al*. The importance of race and ethnic background in biomedical research and clinical practice. *N Engl J Med* 2003; **348**(12)**:** 1170-1175.

16. Ferrell PB, Jr., McLeod HL. Carbamazepine, HLA-B*1502 and risk of Stevens-Johnson syndrome and toxic epidermal necrolysis: US FDA recommendations. *Pharmacogenomics* 2008; **9**(10)**:** 1543-1546.

17. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015; **526**(7571)**:** 68-74.

18. Holmes MV, Shah T, Vickery C, Smeeth L, Hingorani AD, Casas JP. Fulfilling the promise of personalized medicine? Systematic review and field synopsis of pharmacogenetic studies. *PLoS One* 2009; **4**(12)**:** e7960.

19. Zanger UM, Raimundo S, Eichelbaum M. Cytochrome P450 2D6: overview and update on pharmacology, genetics, biochemistry. *Naunyn Schmiedebergs Arch Pharmacol* 2004; **369**(1)**:** 23-37.

20. Francois AA, Nishida CR, de Montellano PRO, Phillips IR, Shephard EA. Human flavin-containing monooxygenase 2.1 catalyzes oxygenation of the antitubercular drugs thiacetazone and ethionamide. *Drug Metabolism and Disposition* 2009; **37**(1)**:** 178-186.

21. Veeramah KR, Thomas MG, Weale ME, Zeitlyn D, Tarekegn A, Bekele E*, et al*. The

potentially deleterious functional variant flavin-containing monooxygenase 2*1 is at high frequency throughout sub-Saharan Africa. *Pharmacogenet Genomics* 2008; **18**(10)**:** 877-886.

22. Horsfall LJ, Zeitlyn D, Tarekegn A, Bekele E, Thomas MG, Bradman N*, et al*. Prevalence of clinically relevant UGT1A alleles and haplotypes in African populations. *Ann Hum Genet* 2011; **75**(2)**:** 236-246.

23. Gammal RS, Court MH, Haidar CE, Iwuchukwu OF, Gaur AH, Alvarellos M*, et al*. Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for UGT1A1 and Atazanavir Prescribing. *Clin Pharmacol Ther* 2016; **99**(4)**:** 363-369.

24. Lee CR, Pieper JA, Frye RF, Hinderliter AL, Blaisdell JA, Goldstein JA. Differences in flurbiprofen pharmacokinetics between CYP2C9*1/*1, *1/*2, and *1/*3 genotypes. *Eur J Clin Pharmacol* 2003; **58**(12)**:** 791-794.

25. Perini JA, Vianna-Jorge R, Brogliato AR, Suarez-Kurtz G. Influence of CYP2C9 genotypes on the pharmacokinetics and pharmacodynamics of piroxicam. *Clin Pharmacol Ther* 2005; **78**(4)**:** 362-369.

26. Rettie AE, Haining RL, Bajpai M, Levy RH. A common genetic basis for idiosyncratic toxicity of warfarin and phenytoin. *Epilepsy Res* 1999; **35**(3)**:** 253-255.

27. Tang C, Shou M, Mei Q, Rushmore TH, Rodrigues AD. Major role of human liver microsomal cytochrome P450 2C9 (CYP2C9) in the oxidative metabolism of celecoxib, a novel cyclooxygenase-II inhibitor. *J Pharmacol Exp Ther* 2000; **293**(2)**:** 453-459.

28. Furuta T, Ohashi K, Kosuge K, Zhao XJ, Takashima M, Kimura M*, et al*. CYP2C19 genotype status and effect of omeprazole on intragastric pH in humans. *Clin Pharmacol Ther* 1999; **65**(5)**:** 552-561.

29. Goldstein JA, Faletto MB, Romkes-Sparks M, Sullivan T, Kitareewan S, Raucy JL*, et al*. Evidence that CYP2C19 is the major (S)-mephenytoin 4'-hydroxylase in humans. *Biochemistry* 1994; **33**(7)**:** 1743-1752.

30. Hirani VN, Raucy JL, Lasker JM. Conversion of the HIV protease inhibitor nelfinavir to a bioactive metabolite by human liver CYP2C19. *Drug Metab Dispos* 2004; **32**(12)**:** 1462-1467.

31. Inomata S, Nagashima A, Itagaki F, Homma M, Nishimura M, Osaka Y*, et al*. CYP2C19 genotype affects diazepam pharmacokinetics and emergence from general anesthesia. *Clin Pharmacol Ther* 2005; **78**(6)**:** 647-655.

32. Scott SA, Sangkuhl K, Stein CM, Hulot JS, Mega JL, Roden DM*, et al*. Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update. *Clin Pharmacol Ther* 2013; **94**(3)**:** 317-323.

33. Tseng E, Walsky RL, Luzietti RA, Jr., Harris JJ, Kosa RE, Goosen TC*, et al*. Relative contributions of cytochrome CYP3A4 versus CYP3A5 for CYP3A-cleared drugs assessed in vitro using a CYP3A4-selective inactivator (CYP3cide). *Drug Metab Dispos* 2014; **42**(7)**:** 1163-1173.

34. Phillips IR, Shephard EA. Drug metabolism by flavin-containing monooxygenases of human and mouse. *Expert Opin Drug Metab Toxicol* 2017; **13**(2)**:** 167-181.

35. McDonagh EM, Boukouvala S, Aklillu E, Hein DW, Altman RB, Klein TE. PharmGKB summary: very important pharmacogene information for N-acetyltransferase 2. *Pharmacogenet Genomics* 2014; **24**(8)**:** 409-425.

36. Innocenti F, Ratain MJ. Pharmacogenetics of irinotecan: clinical perspectives on the utility of genotyping. *Pharmacogenomics* 2006; **7**(8)**:** 1211-1221.

37. Bains RK. African variation at Cytochrome P450 genes: Evolutionary aspects and the implications for the treatment of infectious diseases. *Evol Med Public Health* 2013; **2013**(1)**:** 118-134.

38. Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet* 2008; **9:** 403-433.

39. Choudhury A, Aron S, Sengupta D, Hazelhurst S, Ramsay M. African genetic diversity provides novel insights into evolutionary history and local adaptations. *Hum Mol Genet* 2018; **27**(R2)**:** R209-R218.

40. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K*, et al*. The African Genome Variation Project shapes medical genetics in Africa. *Nature* 2015; **517**(7534)**:** 327-332.

41. de Man FM, Goey AKL, van Schaik RHN, Mathijssen RHJ, Bins S. Individualization of Irinotecan Treatment: A Review of Pharmacokinetics, Pharmacodynamics, and Pharmacogenetics. *Clin Pharmacokinet* 2018; **57**(10)**:** 1229-1254.

42. Sugatani J, Mizushima K, Osabe M, Yamakawa K, Kakizaki S, Takagi H*, et al*. Transcriptional regulation of human UGT1A1 gene expression through distal and proximal promoter motifs: implication of defects in the UGT1A1 gene promoter. *Naunyn Schmiedebergs Arch Pharmacol* 2008; **377**(4-6)**:** 597-605.

43. Han FF, Guo CL, Yu D, Zhu J, Gong LL, Li GR*, et al*. Associations between UGT1A1*6 or UGT1A1*6/*28 polymorphisms and irinotecan-induced neutropenia in Asian cancer patients. *Cancer Chemother Pharmacol* 2014; **73**(4)**:** 779-788.

44. Cui C, Shu C, Cao D, Yang Y, Liu J, Shi S*, et al*. UGT1A1*6, UGT1A7*3 and UGT1A9*1b polymorphisms are predictive markers for severe toxicity in patients with metastatic gastrointestinal cancer treated with irinotecan-based regimens. *Oncol Lett* 2016;

12(5)**:** 4231-4237.

45. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**(1)**:** 7.

46. Ansari Pour N, Plaster CA, Bradman N. Evidence from Y-chromosome analysis for a late exclusively eastern expansion of the Bantu-speaking people. *Eur J Hum Genet* 2013; **21**(4)**:** 423-429.

47. R-Core Team (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.

48. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003; **73**(5)**:** 1162-1169.

Table legends

Table 1. NHS England ethnic categories, population data and samples tested. Data from UCH records for all patients for a two-year period (2006-2007) before the proposal was funded; outpatients only for a six-month period at the start of the project (2009); and 2011 census commissioned data for Camden. Sample collections studied are also shown.

Table 2. List of clinically important drugs metabolised by the enzymes encoded by the six genes included in this study.

Table 3. Allele frequencies of therapeutically relevant variations tested in groups representing the three major NHS England classification of ethnicities in London. Frequencies of the rarer allele worldwide shown in all cases. Alleles shown on the chromosomal 'top' strand.

Table 4. Inferred gene-wide haplotypes and their respective frequencies in the groups representing the three major NHS England classification of ethnicities in London. SNP order shown in side table.

Table 1. Table of NHS England* ethnic categories, population data and samples tested.

NHS England ethnic categories, data from UCH records for all patients for a two-year period (2006-2007) before the proposal was funded; outpatients only for a six-month period at the start of the project (2009); and 2011 census commissioned data for Camden.
Sample collections tested are also shown. Data for the groups tested are shown in bold

NHS England Categories		Population data			Samples Tested		
		UCH 2006/7 all	UCH 2009 clinics	Camden 2011**			
White							
A	British	**164,218**	**9,337**	**97,400**	**78**	**50**	**91**
B	Irish	6,773	460	7,132			
C	Any other White background	35,496	1,846	23,672			
Mixed							
D	White and Black Caribbean	1,660	56	2,502			
E	White and Black African	1,272	39	1,809			
F	White and Asian	1,622	57	3,912			
G	Any other mixed background	3,544	97	1,032			
Asian or Asian British							
H	Indian	**8,128**	**252**	**6,184**	**14**	**44**	**205**
J	Pakistani	2,629	104	1,502			
K	Bangladeshi	**9,124**	**295**	**12,517**	**67**		**86**
L	Any other Asian background	5,326	172	620			
Black or Black British							
M	Caribbean	8,339	368	3,608			
N	African	**15,455**	**485**	**11,800**	**77**	**201**	**505**
P	Any other Black background	2,531	81	2,550			
Other							
S	Other Ethnic Background	14,954	615	10,054			
R	Chinese	3,639	135	6,592			
Z	Not stated	50,575	954	N/A			

*Scotland uses different classifications:
https://www.ndc.scot.nhs.uk/Dictionary-A-Z/Definitions/index.asp?Search=E&ID=243&Title=Ethnicity%20Code
**https://data.london.gov.uk/dataset/detailed-ethnicity-by-age---sex-ward-tools---2011-census--

Table 2. List of clinically important drugs metabolised by the enzymes encoded by the six genes included in this study.

Gene	Drug	Reference
CYP2C9	Warfarin, phenytoin, flurbiprofen, celecoxib, piroxicam	[24-27]
CYP2C19	Omeprazole, mephenytoin, diazepam, clopidogrel, nelfinavir	[28-32]
CYP3A5	Tacrolimus, verapamil, vardenafil, midazolam	[33]
FMO2	Ethionamide, methimazole, thiacetazone	[34]
NAT2	Isoniazid, hydralazine, sulfamethoxazole	[35]
UGT1A1	Irinotecan	[36]

Table 3. Allele frequencies of therapeutically relevant variations tested in groups representing the three major NHS England classification of ethnicities in London. Frequencies of the rarer allele worldwide shown in all cases. Alleles shown on the chromosomal 'top' strand

Chromosome	Gene	SNP rsID	DME locus Nomenclature	Allele of interest	**Allele Frequency**							
					White British			**Bangladeshi**		**Black African**		
					London (N=77)	**UK in-house (N=50)**	**GBR_1000G (N=91)**	**London (N=66)**	**BEB_1000G (N=86)**	**London (N=75)**	**African_in-house (N=198)**	**African_1000G (N=504)**
1	*FMO2*	rs6661174	*FMO2*2*	C (*FMO2*1*)#	0.000	0.000	0.000	0.008	0.012	0.207	0.169	0.155
2	*UGT1A*	rs11692021	*UGT1A7*3*	C	0.377	0.330	0.341	0.406	0.390	0.281	0.288	0.234
2	*UGT1A*	rs4124874	*UGT1A1*60* ##	G	0.460	0.380	0.379	0.528	0.611	0.854	0.907	0.922
2	*UGT1A*	rs10929302	*UGT1A1*93*	A	0.279	0.240	0.231	0.385	0.407	0.304	0.340	0.345
2	*UGT1A*	rs887829	*UGT1A1*28* ^	T	0.315	0.261	0.264	0.417	0.419	0.477	0.507	0.499
7	*CYP3A5*	rs41303343	*CYP3A5*7*	A	0.000	0.000	0.000	0.000	0.000	0.093	0.062	0.119
7	*CYP3A5*	rs10264272	*CYP3A5*6*	T	0.000	0.000	0.000	0.000	0.000	0.127	0.147	0.175
7	*CYP3A5*	rs776746	*CYP3A5*3*	C#	0.912	0.906	0.945	0.637	0.634	0.333	0.199	0.151
8	*NAT2*	rs1801279	*NAT2*14*	A	0.000	0.000	0.000	0.000	0.000	0.113	0.105	0.111
8	*NAT2*	rs1801280	*NAT2*5*	C	0.422	0.440	0.467	0.364	0.361	0.433	0.334	0.292
8	*NAT2*	rs1799930	*NAT2*6*	A	0.295	0.320	0.275	0.341	0.267	0.192	0.231	0.227
8	*NAT2*	rs1799931	*NAT2*7*	A	0.032	0.030	0.022	0.091	0.099	0.033	0.036	0.025
10	*CYP2C19*	rs12248560	*CYP2C19*17*	T$^{\$}$	0.183	0.153	0.242	0.159	0.111	0.220	0.207	0.233
10	*CYP2C19*	rs72552267	*CYP2C19*6*	A	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	*CYP2C19*	rs17884712	*CYP2C19*9*	A	0.000	0.000	0.000	0.000	0.000	0.008	0.003	0.011
10	*CYP2C19*	rs4986893	*CYP2C19*3*	A	0.000	0.000	0.000	0.000	0.023	0.013	0.000	0.002
10	*CYP2C19*	rs6413438	*CYP2C19*10*	T	0.000	0.000	0.000	0.000	0.000	0.007	0.003	0.002
10	*CYP2C19*	rs4244285	*CYP2C19*2*	A	0.099	0.170	0.143	0.414	0.326	0.182	0.130	0.178
10	*CYP2C19*	rs192154563	*CYP2C19*16*	T	0.000	0.000	0.000	0.000	0.000	0.000	0.013	0.001
10	*CYP2C19*	rs55640102	*CYP2C19*12*	T	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000
10	*CYP2C9*	rs72558189	*CYP2C9*14*	A	0.000	0.000	0.000	0.011	0.029	0.000	0.000	0.000
10	*CYP2C9*	rs1799853	*CYP2C9*2*	T	0.115	0.088	0.088	0.011	0.017	0.008	0.024	0.001
10	*CYP2C9*	rs7900194	*CYP2C9*8*	A	0.000	0.014	0.000	0.000	0.000	0.067	0.012	0.050
10	*CYP2C9*	rs9332131	*CYP2C9*6*	Δ	0.000	0.000	0.000	0.000	0.000	0.014	0.008	0.010
10	*CYP2C9*	rs57505750	*CYP2C9*31*	C	0.000	0.000	0.000	0.000	0.000	0.007	0.000	0.002
10	*CYP2C9*	rs28371685	*CYP2C9*11*	T	0.000	0.000	0.000	0.000	0.000	0.021	0.026	0.028
10	*CYP2C9*	rs1057910	*CYP2C9*3*	C	0.125	0.040	0.071	0.102	0.116	0.000	0.008	0.000
10	*CYP2C9*	rs28371686	*CYP2C9*5*	G	0.000	0.000	0.000	0.000	0.000	0.007	0.010	0.016

$^{\$}$ Associated with an increased enzymatic activity.

^ proxy for *UGT1A1*28*

# for *FMO2* and *CYP3A5* the most common allele world-wide is not the ancestral allele
## for *UGT1A1*60* the low activity allele is more frequent than the high activity allele

Table 4. Inferred gene-wide haplotypes and their respective frequencies in the groups representing the three major NHS England classification of ethnicities in London. SNP order shown in side table.

Gene	Haplotype	W_British	UK_British	GBR	Bangladeshi	BEB	B_African	African in-house_all	African 1000G_all
UGT1A	**CGAT**	0.273	0.240	0.231	0.333	0.326	0.233	0.225	0.195
	CGGC	0.065	0.040	0.066	0	0.006	0.033	0.038	0.013
	CGG**T**	0.026	0.020	0.033	0.008	0	0	0.013	0.026
	T**GAT**	0.006	0	0	0.053	0.081	0.073	0.111	0.150
	T**G**GC	0.058	0.070	0.049	0.091	0.186	0.387	0.422	0.410
	T**G**G**T**	0	0	0	0	0.012	0.147	0.124	0.128
	TTGC	0.558	0.600	0.610	0.439	0.331	0.120	0.056	0.078
	CGAC	0	0	0	0	0	0	0.003	0
	CTGC	0.013	0.030	0.011	0.076	0.058	0.007	0.010	0
CYP3A5	G-CCTG	0	0	0	0	0.006	0.207	0.280	0.250
	G-CTTG	0.026	0.030	0.011	0.318	0.320	0.053	0.053	0.077
	G-**T**CTG	0	0	0	0	0	0.127	0.141	0.175
	T**A**CCTG	0	0	0	0	0	0.093	0.061	0.119
	T-CC**C**A	0.909	0.890	0.929	0.583	0.605	0.320	0.207	0.148
	T-CC**C**G	0.006	0.020	0.016	0.015	0.023	0.013	0.003	0.003
	T-CCTA	0	0	0.005	0.008	0	0.087	0.144	0.137
	T-CCTG	0.058	0.060	0.038	0.068	0.041	0.100	0.106	0.091
	G-CCTA	0	0	0	0	0	0	0.003	0
	T-CTTG	0	0	0	0	0	0	0.003	0
	G-CC**C**G	0	0	0	0.008	0	0	0	0
	G-CC**C**A	0	0	0	0	0.006	0	0	0
NAT2	**A**TGGG	0	0	0	0	0	0.113	0.104	0.111
	G**C**GGA	0	0	0	0	0	0	0	0.001
	G**C**GGG	0.422	0.440	0.467	0.364	0.360	0.433	0.343	0.291
	GT**A**GA	0.292	0.320	0.275	0.341	0.262	0.193	0.212	0.227
	GTG**A**G	0.032	0.030	0.022	0.091	0.099	0.033	0.035	0.025
	GTGGG	0.253	0.210	0.236	0.205	0.273	0.227	0.290	0.345
	G**CA**GG	0	0	0	0	0	0	0.003	0
	GT**A**GG	0	0	0	0	0.006	0	0.010	0
	GTGGA	0	0	0	0	0	0	0.003	0
CYP2C19	CA**A**GCGCGATCT	0	0	0	0	0	0.007	0.003	0.011
	CAG**A**CGCGATCC	0	0	0	0	0.023	0.013	0	0.002
	CAGGCGCAACCT	0.247	0.230	0.198	0.136	0.221	0.020	0.013	0.003
	CAGGCGCGACCT	0.175	0.170	0.225	0.129	0.169	0.180	0.215	0.177
	CAGGCGCGATCC	0.130	0.030	0.066	0.098	0.116	0.193	0.184	0.123
	CAGGCGCGATCT	0.039	0.030	0.038	0.030	0.017	0.147	0.184	0.230
	CAGGCGCGCTCT	0	0	0	0	0	0	0.005	0.003
	CAGGCGTGCTCT	0	0	0	0	0	0	0	0.009
	CAGG**T**GCGATCT	0	0	0	0	0	0	0	0.002
	CGGGC**A**CGATCC	0	0.160	0.143	0.371	0.320	0.173	0.124	0.176
	CGGGC**A**CGATCT	0	0	0	0.008	0.006	0	0	0.002
	CGGGCGCGATCC	0	0	0	0.008	0	0.027	0.023	0.029
	CGGGCGCGAT**T**T	0	0	0	0	0	0	0	0.001
	TAGGCGTGCTCC	0	0	0	0	0	0.013	0.025	0.028
	TAGGCGTGCTCT	0.201	0.160	0.242	0.152	0.110	0.193	0.184	0.205
	CAGGCGCAACCC	0.110	0.210	0.088	0.008	0.017	0.013	0.020	0
	CAGGCGCAATCT	0	0	0	0	0	0	0.003	0
	CGGGCGCGATCT	0	0	0	0	0	0	0.003	0
	CGGGCGCGAT**T**C	0	0	0	0	0	0	0.013	0
	CGGG**TA**CGATCC	0	0	0	0	0	0	0.003	0
	CAGGC**A**CGATCC	0.097	0	0	0.038	0	0.007	0	0
	TAGGCGCGCTCT	0	0	0	0	0	0.007	0	0
	TAGG**T**GTGCTCT	0	0	0	0	0	0.007	0	0
	CAGGCGCGACCC	0	0	0	0.008	0	0	0	0
	CGGGC**A**CGACCT	0	0	0	0.008	0	0	0	0
	TAGGCGCGATCC	0	0	0	0.008	0	0	0	0
	CAGGC**A**CAACCT	0	0.010	0	0	0	0	0	0
CYP2C9	AGCGAG**C**CACA	0	0	0	0	0	0.007	0	0.002
	AGCGAGTCACA	0.325	0.390	0.341	0.227	0.314	0.080	0.126	0.125
	AGCGAGTCACC	0.065	0.020	0.066	0.038	0.070	0.080	0.078	0.054
	GGC**A**AATCACA	0	0	0	0	0	0	0	0.001
	GGC**A**AATCACC	0	0	0	0	0	0.053	0.010	0.049
	GGCGAATCACA	0.019	0.070	0.060	0.023	0.012	0.007	0.023	0.004
	GGCGAATCACC	0.104	0.240	0.082	0.402	0.331	0.307	0.253	0.217
	GGCGAATCA**G**C	0	0	0	0	0	0	0.010	0.016
	GGCGAGTCACA	0.013	0.010	0.038	0.053	0.017	0.053	0.058	0.059
	GGCGAGTCACC	0.208	0.140	0.231	0.144	0.110	0.367	0.381	0.436
	GGCGAGT**T**ACC	0	0	0	0	0	0.020	0.025	0.028
	GGCG**-**ATCACC	0	0	0	0	0	0.013	0.008	0.01
	GG**T**GAGTCACA	0	0	0	0	0	0	0.003	0.001
	GGCGAATC**C**CC	0.123	0.040	0.066	0.083	0.099	0	0.008	0
	GG**T**GAATCACC	0.117	0.070	0.088	0.008	0.017	0.007	0.015	0
	GG**T**GAGTCACC	0	0	0	0	0	0	0.003	0
	GGCGAATCA**G**A	0	0	0	0	0	0.007	0	0
	G**A**CGAATCACC	0	0	0	0.008	0	0	0	0
	GGCGAATC**C**CA	0	0	0	0.015	0	0	0	0
	A**A**CGAGTCACC	0	0	0	0	0.012	0	0	0
	G**A**CGAATC**C**CC	0	0	0	0	0.017	0	0	0
	AGCGAATCACA	0.019	0	0.022	0	0	0	0	0
	AGCGAATC**C**CC	0	0	0.005	0	0	0	0	0
	AGC**A**AATCACA	0	0.020	0	0	0	0	0	0
	AGCGAGTC**C**CA	0.006	0	0	0	0	0	0	0

SNP order

Gene	SNP*	*Allele Nomenclature
UGT1A	**rs11692021**	UGT1A7*3
	rs4124874	UGT1A1*60
	rs10929302	UGT1A1*93
	rs887829	UGT1A1*80 !!
CYP3A5	rs4646458	
	rs41303343	CYP3A5*7
	rs10264272	CYP3A5*6
	rs4646449	
	rs776746	CYP3A5*3
	rs2687087	
NAT2	**rs1801279**	NAT2*14
	rs1801280	NAT2*5
	rs1799930	NAT2*6
	rs1799931	NAT2*7
	rs4646247	
CYP2C19	**rs12248560**	CYP2C19*17
	rs12769205	CYP2C19*35
	rs17884712	CYP2C19*9
	rs4986893	CYP2C19*3
	rs6413438	CYP2C19*10
	rs4244285	CYP2C19*2
	rs12247175	
	rs4494250	
	rs12253253	
	rs4917623	
	rs192154563	CYP2C19*16
	rs2104162	
CYP2C9	rs7899661	
	rs72558189	CYP2C9*14
	rs1799853	CYP2C9*2
	rs7900194	CYP2C9*8
	rs9332131	CYP2C9*6
	rs4918766	
	rs57505750	CYP2C9*31
	rs28371685	CYP2C9*11
	rs1057910	CYP2C9*3
	rs28371686	CYP2C9*5
	rs2860975	

BOLD text =Low activity alleles

!! -now known as *80 but surrogate for *27/*28