DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

Title Page

*The CYP2C19 intron 2 branch point SNP is the ancestral polymorphism contributing to the poor metabolizer phenotype in livers with CYP2C19*35 and CYP2C19*2 alleles*

Amarjit S. Chaudhry, Bhagwat Prasad, Yoshiyuki Shirasaka, Alison Fohner, David Finkelstein, Yiping Fan, Shuoguo Wang, Gang Wu, Eleni Aklillu, Sarah Sim, Kenneth E. Thummel and Erin G. Schuetz

Department of Pharmaceutical Sciences, St Jude Children's Research Hospital, Memphis, TN, (A.S.C., E.G.S.), Department of Pharmaceutics, School of Pharmacy, University of Washington, Seattle, Washington (B.P., Y.S., A.F., K.E.T.), Department of Computational Biology, St Jude Children's Research Hospital, Memphis, TN (D.F., Y.F., S.W., G.W.), Department of Laboratory Medicine, Division of Clinical Pharmacology, Karolinska Institutet, Stockholm, Sweden (E.A.), Section of Pharmacogenetics, Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden (S.S.).

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

Running Title Page

Running Title: *CYP2C19* nonfunctional SNP rs12769205 alters splicing

Corresponding Author: Erin G. Schuetz, Ph.D., Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105; PH: (901) 595-2205; Fax: (901) 595-3125; Email: erin.schuetz@stjude.org

Number of text pages:

Number of tables: 0

Number of figures: 9

Number of references: 39

Number of words in the *Abstract:* 249

Number of words in the *Introduction*: 380

Number of words in the *Discussion*: 1425

Abbreviations: CYP, cytochrome P450; LD, linkage disequilibrium; IGV, integrative genomic viewer; NMD, nonsense mediated decay; PM, poor metabolizer; PTC, premature termination codon; REHH, relative extended haplotype homozygosity; SNP, single nucleotide polymorphism; SV, splice variant; YRI, Yorubans.

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

ABSTRACT

CYP2C19 rs12769205 alters an intron 2 branch point adenine leading to an alternative mRNA in human liver with complete inclusion of intron 2 (exon 2B). rs12769205 changes the mRNA reading frame, introduces 87 amino acids, and leads to a premature stop codon. The 1000 genomes project indicated rs12769205 is in LD with rs4244285 on *CYP2C19*2*, but found alone on *CYP2C19*35* in Blacks. Minigenes containing rs12769205 transfected into HepG2 cells demonstrated this SNP alone leads to exon 2B and decreases *CYP2C19* canonical mRNA. A residual amount of CYP2C19 protein was detectable by quantitative proteomics with tandem mass spectrometry in *CYP2C19*2/*2* and **1/*35* liver microsomes with an exon 2 probe. However, an exon 4 probe, downstream of rs12769205, but upstream of rs4244285, failed to detect CYP2C19 protein in livers homozygous for rs12769205 demonstrating rs12769205 alone can lead to complete loss of CYP2C19 protein. *CYP2C19* genotypes and mephenytoin phenotype were compared in 104 Ethiopians. Poor metabolism of mephenytoin was seen in persons homozygous for both rs12769205 + rs4244285 (*CYP2C19*2/*2*), but with little effect on mephenytoin disposition of *CYP2C19*1/*2, CYP2C19*1/*3* or *CYP2C19*1/*35* heterozygous alleles. Extended haplotype homozygosity tests of the Hapmap Yorubans (YRI) showed both haplotypes carrying rs12769205 (*CYP2C19*35* and *CYP2C19*2*) are under significant natural selection, with *CYP2C19*35* having a higher REHH score. The phylogenetic

tree of the YRI *CYP2C19* haplotypes revealed rs12769205 arose first on *CYP2C19*35* and that rs4244285 was added later creating *CYP2C19*2.* In conclusion, rs12769205 is the ancestral polymorphism leading to aberrant splicing of *CYP2C19*35* and *CYP2C19*2* alleles in liver.

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

INTRODUCTION

CYP2C19 is an important drug-metabolizing enzyme that plays a critical role in the metabolism as well as drug-drug interactions of a variety of drugs, including proton pump inhibitors, anti-epileptics, anti-platelet drugs, and anti-depressants (Li-Wan-Po et al., 2010; Shah et al., 2012; Shirasaka et al., 2013). Similar to a variety of other CYP family members, CYP2C19 activity is polymorphic, with sub-populations of poor metabolizers, intermediate metabolizer, extensive metabolizer and ultra-rapid metabolizer (Wedlund, 2000; McGraw and Waller, 2012; Hicks et al., 2013). Multiple allelic variants of *CYP2C19* have been described. *CYP2C19*1* represents the wild-type allele. The frequent rs4244285 polymorphism, defining the *CYP2C19*2* allele, creates an exon 5 aberrant splice site, altering the reading frame of the mRNA leading to a premature stop codon and a non-functional protein (de Morais et al., 1994a).

There are numerous reports linking the *CYP2C19*2* allele to altered substrate clearance (Hirota et al., 2013; Hicks et al., 2013; Owusu et al., 2014). Over 34 *CYP2C19* variant alleles have been identified in a cytochrome P450 database (http://www.cypalleles.ki.se/cyp2c19.htm). In addition, a recent study of 2203 African Americans (Gordon et al., 2014) found that *CYP2C19* had the highest number of putative novel functional variants compared with 11 other drug metabolizing *CYP* genes. This result was interesting because it was recently reported that the *CYP2C19*2* nonfunctional allele may have been positively selected in human evolution (Janha

et al., 2014), and that inactivation of *CYP2C19* might have afforded some survival advantage. Conversely, because *CYP2C19* loss-of-function alleles confer increased risks for serious adverse cardiovascular events among clopidogrel treated patients, Clinical Pharmacogenetic Implementation Consortium (CPIC) Guidelines were issued for *CYP2C19* genotype-directed drug therapy (Scott et al., 2013).

In an effort to identify additional function-disrupting *CYP2C19* alleles, we sequenced *CYP2C19* in human livers and identified a branch point SNP (rs12769205; gene position 12662A>G) in intron 2 of *CYP2C19* that leads to intron 2 retention. Interestingly, rs12769205 is found in combination with rs4244285 (the SNP that defines all *CYP2C19*2* alleles) and 12662A>G is likely part of all *CYP2C19*2* alleles. However, rs12769205 is also found without rs4244285 on *CYP2C19*35* (allele designation assigned by the CYP allele nomenclature committee). In this report, we have investigated the functional consequence of rs12769205, whether *CYP2C19*35* is also under natural selection, and whether *CYP2C19*35* is the ancestral *CYP2C19* nonfunctional allele, arising before *CYP2C19*2*.

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

MATERIALS AND METHODS

Human liver tissue. A total of 335 human livers from 272 White and 63 Black donors were processed through the St. Jude Liver Resource at St. Jude Children's Research Hospital and were provided by the Liver Tissue Procurement and Distribution System (National Institutes of Health National Institute of Diabetes and Digestive and Kidney Diseases Contract N01-DK92310) and by the Cooperative Human Tissue Network. The St. Jude Children's Research Hospital Institutional Review Board approved the use of these human tissues for research purposes.

RNA isolation and cDNA preparation. Total RNA was extracted from human livers with rs4244285 and rs12769205 genotypes using TRIzol reagent (Invitrogen, Cat. No. 15596-026). 500 ng RNA was used to prepare cDNA using Invitrogen Thermoscript™ RT-PCR System (Cat no. 11146-024).

Genotyping of *CYP2C19* alleles in human livers. Genomic DNA from human livers was isolated using a DNeasy tissue kit (Qiagen Cat no. 69506). Genotyping of the rs4244285 and rs12769205 SNPs was performed by direct DNA sequencing. Primers used for PCR amplification and sequencing of rs4244285 were (FP) 5'-CAACCAGAGCTTGGCATATTG-3' and (RP) 5'-TGATGCTTACTGGATATTCATGC-3'; and for rs12769205 were (FP)

5'-AAAATATGAATCTAAGTCAGGCTTAGT-3' and (RP)

5'-GGAGAGCAGTCCAGAAAGGTCAGTGATA-3'. A general 25 µL PCR mixture consisted of 50 ng gDNA, 1 µM primers and Platinum PCR supermix (Invitrogen Cat No. 11306-016). For quality control, minor allele frequencies (MAFs) for rs4244285 and rs12769205 were compared to existing population genotype data from the Exome Variant Server (http://evs.gs.washington.edu/EVS/) for Whites and Blacks. The observed and reported MAFs values were in agreement.

Sequencing of *CYP2C19* exons from genomic DNA. *CYP2C19* exons were PCR amplified in a 25 µL PCR mixture consisting of 50 ng gDNA, 1 µM primers and Platinum PCR supermix mix (Invitrogen Cat No. 11306-016) and sequenced using *CYP2C19* specific primers (Supplemental Table 1).

Genotyping of African livers for *CYP2C19*2* at the cDNA level. To confirm the *CYP2C19*2* genotype assignments obtained by DNA sequencing we also genotyped the eight African livers used for subsequent analysis of the transcripts at the cDNA level using the primers and method reported by de Morais *et al* 1994a.

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

***CYP2C19* cDNA amplification.** Various portions of *CYP2C19* were PCR amplified from human liver cDNA using the following primer pairs and products were either directly sequenced or analyzed on 2% agarose gels. (1) Exon 4 (FP) (5'-ATTGAATGAAAACATCAGGATTG-3') and exon 6 (RP) (5'-GTAAGTCAGCTGCAGTGATTA-3') (de Morais et al., 1994a). *CYP2C19*1* and *CYP2C19*2* generate 284 and 244 bp products, respectively. (2) Exon 2 (FP) (5'-GAAGAGGCCATTTCCCACT-3') and exon 4 (RP) (5'-TTTCTGGAAAATAATGGAGCA-3'). Livers with and without rs12769205 generate 438 bp (retaining 169 bp intron 2) and 269 bp products, respectively. (3) Exon 2 (FP) (5'-GAAGAGGCCATTTCCCACT-3') and exon 6 (RP) (5'-GTAAGTCAGCTGCAGTGATTA-3') primers. *CYP2C19*1* generates the canonical 597 bp product. *CYP2C19*35* generates products of 766 bp (containing exon 2B) and 695 bp (exon 2B plus 70 bp deletion in exon 4). *CYP2C19*2* generates products with the 40 bp deletion of exon 5 (557 bp) and additionally containing exon 2B with or without the 70 bp deletion in exon 4 (726 bp and 655 bp, respectively).

Sequencing the full length *CYP2C19*35* cDNA and its alternative transcripts. First, cDNA from the *CYP2C19*1/*35* liver was used as template and PCR amplified with exon 2 and 6 primers (above), the PCR products were cloned into pCR 2.1 TOPO vector using the TOPO TA

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

Cloning® Kit (Invitrogen Cat No. 450641), and products transformed into One Shot® Top 10 Chemically Competent cells (Invitrogen Cat no. 1427548) and grown on LB-ampicillin plates. Forty individual colonies were picked and colony PCR carried out using M13 primers (FP) 5´-GTAAAACGACGGCCAG-3´ and (RP) 5´-CAGGAAACAGCTATGAC-3´, and the PCR products directly sequenced using the same primers.

Second, to determine whether the *CYP2C19*35* allele carried other polymorphisms in the coding region, the full length *CYP2C19*35* cDNA from the *CYP2C19*35* liver was PCR amplified using FP 5'-TTGTGGTCCTTGTGCTCTGTCTC-3'and RP 5'-GGAATGAAGCACAGCTGA-3' and the PCR product cloned into TOPO TA vector and sequenced using M13 primers mentioned above. The sequences obtained were then aligned using a software suite for sequence analysis (DNASTAR SeqMan Pro version 9.0.4. 39, 418).

***CYP2C19* genotypic data mining from 1000 genomes server (Phase 3 data).** rs4244285 and rs12769205 genotypes were downloaded for CEU, All Africans and CHB in the 1000 genomes browser (http://browser.1000genomes.org/index.html) to generate visual genotypes and study the linkage between these two alleles in different ethnic groups.

Linkage Disequilibrium Analysis. Pairwise linkage disequilibrium (LD) was calculated for 85

CEU, 246 Africans, 88 YRI and 97 CHB from the 1000 genomes browser and displayed using Haploview 4.2 software (Barrett et al., 2005). The D', along with its corresponding Logarithm of Odds (LOD) score, and r^2 LD values were determined between rs4244285, and rs3758580, rs4417205 and rs12769205 SNPs.

Generation of *CYP2C19* Minigenes. The RHCglo minigene plasmid (Singh and Cooper 2006) was generously provided by Dr. Thomas A. Cooper (Baylor College of Medicine, Houston, TX). A 264 bp fragment consisting of the last 87 nucleotides of *CYP2C19* intron 4 and full length exon 5 was amplified from the DNA of a *CYP2C19*2/*2* human liver using PCR primers (FP) 5'-ATATATGTCGACAGTTTTAAATTACAACCAGAGCTTGG-3' (having a *Sal*I site (underlined)) and (RP) 5'-ATATATCTCGAGCTTCTCCATTTTGATCAGGAAGC-3' (having *Xho*I site (underlined)), and used to replace the *Sal*I/*Xho*I fragment of the RHCglo plasmid to generate the *CYP2C19*2* minigene. The *CYP2C19*2* minigene was used as template to perform site directed mutagenesis to create the *CYP2C19*1* wild type minigene using the QuikChange II site directed mutagenesis kit (Agilent Technologies Cat no. 200523) and primers (SDM-FP) 5'-CCCACTATCATTGATTATTTCCCGGGAACCCATAACAAATTACTTAA-3' and (SDM-RP) 5'-TTAAGTAATTTGTTATGGGTTCCCGGGAAATAATCAATGATAGTGGG-3'. The *CYP2C19* exon 2/intron 2/exon 3 minigenes were generated by PCR amplifying this region

from human liver DNAs with either *CYP2C19*1* or *CYP2C19*35* genotypes using primers (FP) 5'-ATATATGGATCCCTCTCAAAAATCTATGGCCCTG-3' (having a *Bam*HI site (underlined)) and (RP) 5'-ATATATCTCGAGCCTTGGTTTTTCTCAACTCC-3' (having *Xho*I site (underlined)). The 482 bp amplified *CYP2C19* fragments were used to replace the *Bam*HI/*Xho*I fragment of the RHCglo plasmid to generate either *CYP2C19*1* or rs12769205 minigenes. The sequences of all minigenes were confirmed by DNA sequencing.

Minigene Transfection Assays. HepG2 human hepatoblastoma cells were cultured in minimal essential media supplemented with 10% fetal bovine serum, 1% penicillin, and 1% streptomycin, and maintained in a humidified incubator at 37°C in an atmosphere of 5% CO_2. For transfection studies, 150,000 cells per well were seeded in 12 well culture dishes. Twenty four hours later cells were transfected with 1000 ng of different minigene plasmids using LipoJet *in vitro* DNA and siRNA transfection kit (Ver. II, SignaGen laboratories Cat. No. SL100468). Forty-eight hours later cells were washed with phosphate-buffered saline and harvested with TRIzol reagent (Invitrogen Cat. No. 15596-026). First-strand cDNA was prepared using 500 ng RNA and oligo (dT) primers (catalog no. 11146-016, ThermoScript RT-PCR system; Invitrogen). PCR amplification was performed using (FP) RSV5U 5'-CATTCACCACATTGGTGTGC-3' and (RP) TNIE4 5'-AGGTGCTGCCGCCGGGCGGTGGCTG-3' that both anneal to the vector expressed exons in

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

order to selectively amplify only the plasmid expressed mRNA transcripts (Singh and Cooper 2006). The amplified PCR fragments were electrophoresed on 1% agarose gel.

Quantitative PCR analysis. Q-real time PCR was used to quantitate the amount of canonical *CYP2C19* transcript generated by the *CYP2C19* wild type and rs12769205 minigenes. The forward primer RSV5U 5'-CATTCACCACATTGGTGTGC-3 annealed in the vector and the reverse primer 5'-CCATTGCTGAAAACGATTCCAA-3' annealed across the *CYP2C19* exon 2 -3 junction to specifically amplify only the canonical *CYP2C19* transcript generated from the minigene. GAPDH primers were (FP) 5'-GGACCACCAGCCCCAGCAAGAG-3' and (RP) 5'-GAGGAGGGGAGATTCAGTGTGGTG-3'. Real-time PCR quantification was carried out using the SYBR GreenER quantitative PCR supermix (catalog no. 11760-100; Invitrogen) and amplifications run on an ABI PRISM 7900HT Sequence Detection System (PE Applied Biosystems, Foster City, CA). The Ct values were analyzed by the comparative Ct method to obtain relative mRNA expression levels.

CYP2C19 peptide quantification. CYP2C19 was quantified using three different surrogate peptides, (1) exon 2 specific peptide (IYGPVFTLYFGLER); (2) exon 8 specific peptide (GTTILTSLTSVLHDNK); and (3) exon 4 specific peptide (ASPCDPTFILGCAPCNVICSIIFQK).

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

The surrogate peptides for LC-MS/MS quantification were selected based on previously reported criteria (Prasad et al., 2014). Trypsin digestion and sample preparation for LC-MS/MS analysis of the genotype-defined pooled HLM (human liver microsomal) samples was performed using a previously reported protocol (Edson et al., 2013; Wang et al., 2015) with few modifications. Briefly, the pooled HLM sample (60 µL) was denatured and reduced with 40 µL of ammonium bicarbonate digestion buffer (100 mM, pH 7.8) and 10 µL of 100 mM dithiothreitol at 90 °C (5 min). The sample was then alkylated by adding 20 µL iodoacetamide (200 mM) at room temperature for 20 min. The protein was then extracted using addition of ice-cold methanol (500 µL), chloroform (200 µL) and water (400 µL). The mixture was vortexed, centrifuged at 12000x *g* for 5 min, and the upper layer was removed. The protein pellet was washed with 500 µL ice-cold methanol followed by centrifugation at 12000*g* for 5 min. The final protein pellet was dissolved in ammonium bicarbonate (40 µL) and 3% sodium deoxycholate (10 µL) before digestion by trypsin (protein:trypsin ratio of 25:1) at 37 °C for 16 h. The reaction was quenched by addition of 20 µL of heavy peptide 2 (GTTILTSLTSVLHDNK[$^{13}C_6$,$^{15}N_2$]) internal standard solution (prepared in 70% acetonitrile in water containing 0.1% formic acid) and 10 µL of the neat solvent (70% acetonitrile in water containing 0.1% formic acid). The samples were centrifuged at 4000*g* for 5 min. All of the HLM samples were digested and processed in triplicates.

The CYP2C19 surrogate peptides were then quantified using triple-quadrupole LC-MS

instruments (Xevo TQ-S coupled to ACQUITY UPLC (Waters)) in ESI positive ionization mode. Approximately 10 µg of the trypsin digest (5 µL) was injected onto the column (Acquity UPLC® HSS T3 1.8 µm, 2.1 x 100 mm, Waters) and eluted at 0.3 mL/min. A mobile phase consisting of water containing 0.1% formic acid (A) and acetonitrile containing 0.1% formic acid (B) was used. A flow rate of 0.3 mL/min was used with a gradient elution starting from 3% B and kept until 2.0 min, followed by gradient program (B concentration) of 3% to 15% (2.0-4.0 min), 15% to 25% (4.0-10 min) and 25% to 50% (10.0-14 min), and his was followed by washing with 80% mobile phase B for 0.9 min, and re-equilibration for 4.9 min. The peak retention times were confirmed by spiking either peptide standards (peptides 1 and 2) and/or trypsin digested CYP2C19 protein standard (gratis sample from Dr Nina Isoherranen). MS/MS analysis was performed by monitoring the surrogate peptides and the internal standard using instrument parameters provided (Supplemental Table 2). LC/MS/MS data were processed using the MassLynx 4.1 (Waters, Milford, MA) by integrating the peak areas generated from the ion chromatograms for the surrogate peptides and normalized by the internal standard response. Peak response for two transitions from each peptide was averaged for quantification of samples and the relative protein quantification was reported as the mean and standard deviation (SD) of peak area ratio values obtained in at least three experiments.

Ethiopian cohort (n=104). Details of the healthy unrelated Ethiopian subjects of both sexes living in Ethiopia who participated in this study were described previously (Persson et al., 1996; Aklillu et al., 2002). The study received ethics approval from the Human Ethics Committees at Huddinge University Hospital, Karolinska Institutet, Stockholm, Sweden and the National Ethics committees at the Ethiopian Science and Technology commission, Addis Ababa, Ethiopia.

CYP2C19 phenotyping of the Ethiopian cohort. Details on CYP2C19 phenotyping of the Ethiopian cohort has been previously published (Persson et al., 1996; Aklillu et al., 2002; Sim et al., 2006). Briefly, the subjects received 100 mg racemic mephenytoin (Mesantoin® Sandoz Pharmaceuticals, Basel, Switzerland) after emptying their bladder just before bedtime. Total urine was collected for 0-8 h after drug intake. Volume and pH were measured and 20 mL aliquots were stored at -20°C until analysis. The concentration ratio of *S* and *R*-mephenytoin was measured by gas chromatography as described previously (Sanz et al., 1989). Urine samples with *S/R* ratio > 0.6 were reanalyzed after acid treatment (Tybring & Bertilsson, 1992). Subjects with an *S/R* ratio greater than 0.9 and not increasing above 1.4 after acid treatment were assigned as poor metabolizers.

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

CYP2C19 genotyping of the Ethiopian cohort. Ten mL venous blood sample was taken from each subject into an EDTA-containing vacutainer tube and DNA was isolated from peripheral leukocytes using a guanidinium-isothiocynate method. Genotyping for *CYP2C19*2* and *CYP2C19*3* was done using allele specific PCR as described previously (de Morais et al., 1994a and 1994b; Persson et al., 1996). Genotyping for *CYP2C19*17* was performed as described previously (Sim et al., 2006). *CYP2C19*35* (rs12769205) was genotyped using a custom Taqman SNP genotyping assay (Item Number 4331349) from Life Technologies. Taqman genotyping assay was done using a QuantStudio 12K Flex- Real-Time PCR system (Life Technologies Holding Pte Ltd, Singapore). The final volume for each reaction was 10 µL, consisting of TaqMan fast advanced master mix (Applied Biosystems, USA), TaqMan 1X drug metabolism genotyping assays mix (Applied Biosystems, USA) and 20 ng genomic DNA. The PCR profile consisted of an initial step at 60^{o}C for 30 sec, hold stage at 95°C for 10 min. and PCR stage for 40 cycles - step 1 at 95°C for 15 sec, and step 2 at 60°C for 1 min, and post read stage at 60°C for 30 sec.

Preparation of RNA-sequencing libraries. Liver RNA was isolated at SJCRH. Liver RNA was shipped to the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) where it was analyzed for integrity and used for RNA-Seq library preparation and sequencing as described in detail in a separate report (Chhibber A et al., communicated). Raw

reads were sent to the St Jude Children's Research Hospital Computational Biology team for analysis.

RNA-Seq analysis. FASTQ sequences were mapped to the hg19 genome by STRONGARM. STRONGARM is a pipeline that employs STAR (Dobin et al., 2012), Tophat2 (Kim et al., 2013), and other mappers and was developed for the Pediatric Cancer Genome Project (Downing et al., 2012). Mapped reads were counted with HTSEQ (Theodor et al., 2014) coverage files and gene level FPKM values were then computed and data was visualized using IGV (Robinson et al., 2011). Exon junction data was extracted through the RNApeg pipeline (Edmonson et al., in preparation).

Haplotypes, relative extended haplotype homozygosity (REHH), and ancestral analysis. The Sweep program (http://www.broadinstitute.org/mpg/sweep/) was used to determine haplotypes and to generate relative extended haplotype homozygosity (REHH) plots in order to look for positive selection in the YRI population and then in other African Populations. All 216 chromosomes from the YRI in the 1000 Genomes phase 3 data (release #20130502) were used (compared to the recent Janha et al., 2014, that used 120 Yoruban chromosomes). The coordinates were lifted from hg19 to hg17 using the liftover program

(https://genome.ucsc.edu/cgi-bin/hgLiftOver) and then data in chr10:96486608-96688055 (hg17 coordinates) were used for subsequent analysis, totaling 201,447 bp. SNP positions rs12769205, rs4244285, rs4417205 and rs3758580 defined the core haplotypes. REHH plots were generated from rs77989980 (downstream of *CYP2C18*) to rs9332103 (upstream of *CYP2C9*). For ancestral analysis, the Sweep program predicts the chimp allele to be the ancestral allele for all SNPs from the HapMap Release 16. Where ancestral information was available, the ancestral core haplotype was predicted and a phylogenetic tree of haplotypes was then created.

RESULTS

A polymorphic *CYP2C19* alternative mRNA retaining intron 2. PCR amplification of the *CYP2C19* cDNA from exon 2 to exon 4 revealed the expected 269 bp product in all livers, but an additional 438 bp product polymorphically expressed in some (Fig 1). Direct sequencing of the 438 bp fragment showed complete intron 2 retention generating exon 2B. All livers expressing the *CYP2C19* alternative mRNA with exon 2B also carried a unique intron 2 SNP rs12769205, but no other SNP in intron 2 or in exons 2 or 3 was found after sequencing exon 2 through exon 3 in genomic DNA. Sequencing of the entire *CYP2C19* cDNA revealed that the rs12769205 was also found in those carrying the *CYP2C19*2* allele.

rs12769205 is in LD with rs4244285 on *CYP2C19*2* but occurs independently on the *CYP2C19*35* allele in some Blacks. To observe the frequency of rs12769205 in different populations and determine whether rs12769205 ever occurred independent of *CYP2C19*2*, rs4244285 and rs12769205 genotypes were retrieved from different populations (Fig 2) in the 1000 genome project database-phase 3 (http://browser.1000genomes.org/index.html) and plotted (Fig 2 A-D). The rs4244285 and rs12769205 SNPs appeared to be linked in *CYP2C19*2* in 99 White descendants of Northern Europeans (CEU), but rs12769205 did occur independent of rs4244285 on a new allele (*CYP2C19*35*) in 11/108 Black descendants of Yoruban West

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

Africans (YRI), and in 29/661 All Africans (that includes the YRI group). *CYP2C19*2* and *CYP2C19*35* appeared to be linked in 103 Asians (CHB). Next, we examined visual genotypes of rs4244285 and rs12769205 in each of the seven African populations representing the African super population in the 1000 Genomes project (Supplemental Figure 1). The rs12769205 SNP had the highest frequency in YRI and the lowest frequency in the Esan in Nigeria. Interestingly, there were individuals in the ESN, GWD, LWK and MSL population that were *CYP2C19*35* homozygous, but *CYP2C19*2* heterozygous, and even one YRI *CYP2C19*35* homozygous, but *CYP2C19*1* homozygous. The *CYP2C19*2/*35* diplotypes as well as *CYP2C19*2* and *CYP2C19*35* allele frequency data in different populations is provided (Supplemental Table 3). To determine if the SNPs were in LD, the D', along with its corresponding Logarithm of Odds (LOD) score, and r^2 LD values were determined between rs4244285 and rs12769205 SNPs in all populations. Haploview (Barrett et al, 2005) analysis showed that the D'/LOD = 1.0 (Supplemental Figure 2) for all pairwise comparisons of the two SNPs in all populations (note: the LD values in the Supplemental Figure 2 are scaled from 1.0 to 100 for visual clarity). The D' of 1.0 indicates that the *CYP2C19*2* rs4244285 is co-inherited with rs12769205 (on the *CYP2C19*2* haplotype) 100% of the time. The high correlation coefficients ($r^2 \geq 0.6$) between the two SNPs confirm they are significantly linked in each population. The r^2 value is lower than the D' because the allele frequencies of rs4244285 and rs12769205 are not identical -

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

rs12769205 is more frequent in YRI because rs12769205 can be found on *CYP2C19*35* that lacks rs4244285. These results prompted us to genotype human livers from White and Black donors. rs12769205 and the rs4244285 appeared to be in LD in Whites (Fig 2E) and African Americans (Fig 2F) with the *CYP2C19*2* genotypes and there was a single African American who had rs12769205 (*CYP2C19*35*) without rs4244285. This liver was further analyzed to determine if exon 2B occurred when only rs12769205 was present.

***CYP2C19*35* (rs12769205) alone leads to exon 2B.** First, the *CYP2C19*2* aberrant spliced mRNA (deletion of the first 40 bp of exon 5) was confirmed by PCR amplification of *CYP2C19* exons 4 to 6 in all samples with the rs4244285 genotype (Fig 3A). Next, *CYP2C19* exons 2 to exon 4 (Fig 3B) were amplified in the same samples yielding both the canonical 269 bp product and a 438 bp amplicon retaining intron 2 (exon 2B). *CYP2C19*2* exon 2B occurred in all livers that carried rs12769205 either alone (*CYP2C19*35*, lane 6), or in livers with rs4244285 (*CYP2C19*2* lanes 3, 4, 5, 8). Moreover, the relative abundance of canonical 269 bp product was always lower in amount in livers with the rs12769205 genotype. Lanes 3, 4, and 6 and lanes 5 and 8 indicate the smaller amount of residual wild-type mRNA amplified in samples heterozygous and homozygous for rs12769205, respectively, compared to samples 1, 2 and 7 homozygous for the *CYP2C19*1* genotype.

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

Alternative mRNAs in livers carrying the rs12769205 allele. *CYP2C19* was amplified from exons 2 to 6 and the PCR products sized on agarose gels to further determine the structure of all hepatic *CYP2C19* alternative mRNAs in livers with the rs12769205 and rs4244285 polymorphisms. Samples homozygous for *CYP2C19*1* yielded the single 597 bp wild-type product (Fig 4). *CYP2C19* PCR products from samples *heterozygous* for rs4244285 and rs12769205 yielded the 597 bp wild-type product, but other bands were difficult to resolve. To unambiguously identify the alternative mRNAs arising from rs12769205, *CYP2C19* exons 2-6 were amplified from the sample heterozygous for *CYP2C19*35*, the products were TOPO TA cloned, and individual clones sequenced. Only transcripts represented in 5% or more of clones are reported: 52/72 were wild type; 14/72 had aberrant exon 2B, and 4/14 of those lacked the first 71 bp of exon 4. These *CYP2C19*35* mRNAs corresponded to the WT (597 bp), SV1 (766 bp) and SV2 (695 bp) bands, respectively (Fig 4). The partial deletion of 71 bp of exon 4 (which alters the reading frame and truncates the protein after amino acid 165) appears to be a passenger splice variant as there were no additional SNPs associated with this spliced transcript. This strategy also allowed the unambiguous identity of the *CYP2C19* mRNAs from samples homozygous for rs12769205 and rs4244285: the alternative mRNAs shared the 40 bp deletion of exon 5 (SV1-3); SV2 and SV3 had exon 2B and SV3 had an additional 71 bp deletion of exon 4.

There was only a small amount of correctly spliced *CYP2C19* WT transcript in livers homozygous for rs12769205 and rs4244285 and this can be seen in Fig 1 (lanes 16-17) and Fig 3.

Sequencing *CYP2C19* from a *CYP2C19*1/*35* liver. To determine whether the *CYP2C19*35* allele carried other polymorphisms, the full length *CYP2C19*35* cDNA from the *CYP2C19*35* liver was amplified, cloned into the TOPO TA vector and sequenced (Supplemental Table 4). The *CYP2C19*35* mRNA contained exon 2B with the causative rs12769205, the common non-synonymous Ile331Val, and the synonymous Pro33Pro. Additional variations were discovered on the other allele but are the focus of another manuscript.

***In Silico* and *In vitro* analysis of the effect of rs12769205 on *CYP2C19* splicing.** The online bioinformatics resource Human Splicing Finder Version 2.4.1 (Desmet et al., 2009) was used to analyze intron 2 for splicing consensus sequences. The program identified a ctcctAg sequence 23 bp upstream of the end of intron 2 as the optimal branch point motif (recognized by the highest number of matrices) having a consensus value of 73.05 (on a 0-100 score range), while rs12769205 will disrupt the branch point sequence decreasing the value to 43.42 (Gooding et al., 2006).

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

To test whether the rs12769205 branch point SNP alone leads to exon 2B, and to confirm that rs4244285 leads to a 40 bp deletion of exon 5, *CYP2C19* minigenes were constructed using the RHCglo minigene vector (Singh and Cooper 2006). Two minigenes contained exons 2 and 3 and intron 2 and differed only by rs12769205 (Fig 5A), and two minigenes contained exon 5 and the last 87 bp of intron 4 and differed only by rs4244285 (*CYP2C19*2*) (Fig 5B). HepG2 cells were transfected with the four minigenes and RNA from the transfected cells was used for RT-PCR analysis using primers residing in the flanking minigene exons. The *CYP2C19*1* intron 2 minigene generated the expected wild-type transcript, whereas the rs12769205 variant generated a larger transcript containing insertion of intron 2 and one other smaller alternatively spliced fragment (Fig 5A). Q-PCR quantitation of the amount of residual wild-type transcript from the minigene revealed a 40% decrease in the amount of wild-type transcript from the rs12769205 vs. *CYP2C19*1* minigene. As expected, the *CYP2C19*1* exon 5 minigene generated the correctly spliced exon 5 transcript, whereas the rs4244285 minigene demonstrated the 40 bp deleted fragment at the start of exon 5 (Fig 5B).

Can next-generation RNA sequencing analysis of *CYP2C19* in human livers identify the 40 bp exon 5 deletion and the exon 2B splicing events? Twenty-four liver samples that had undergone next-generation RNA sequencing were analyzed for alternative *CYP2C19* mRNAs

and results visualized with the integrative genomic viewer (IGV). As is typical of RNA-Seq IGV results, there is non-uniformity in the exon peaks in part due to non-uniformity of read coverage (even when the transcripts have very similar concentrations), sequence specific read variability, and the transcriptional complexity for multigene family members, such as the CYPs. In general, the intron 2 retention (Fig 6A) and the 40 bp deletion in exon 5 (Fig 6B) was apparent by IGV in mRNA samples heterozygous for rs12769205 and rs4244285, although it was much more apparent in some samples where CYP2C19 was more highly expressed. However, it would be difficult to accurately call either of the CYP2C19 alternative mRNAs in some samples, for example the second sample from the top in Fig 6, due to low read coverage and non-uniformity of the exon architecture. Moreover, while visual inspection of each sample revealed there might be a small insertion of intron 2, when exon/intron junction analysis software was used to identify novel transcripts, the software did not call the novel exon 2B junction.

Quantitation of CYP2C19 protein in pooled human liver microsomes with different *CYP2C19* genotypes. CYP2C19 was quantified in pooled human liver microsomes with different *CYP2C19* genotypes by trypsin digestion and LC-MS/MS analysis using three different surrogate peptides specific for *CYP2C19* exons 2, 4 and 8 (Fig 7A). In theory, (a) the probes detect not just the amount of full length CYP2C19, but truncated *in-frame* CYP2C19 translated

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

from the splice variant mRNAs; and (b) because rs12769205 will frame shift the CYP2C19 protein after exon 2, and rs4244285 will frame shift the protein in the middle of exon 5, the exon 2 and 4 probes could distinguish between the functional effects of rs12769205 in intron 2 (downstream of exon 2 but *upstream* of exon 4 probe), and rs4244285 in exon 5 (downstream of both exon 2 and 4 probes) on the abundance of the *residual* CYP2C19 wild-type protein. As expected, CYP2C19 was detected with all exon probes in *CYP2C19*1/*1* livers. The amount of CYP2C19 protein detectable with the exon 2 probe in pooled livers homozygous for *CYP2C19*2/*2* or heterozygous for *CYP2C19*1/*35* was only 10% and 2%, respectively, of the amount of protein in *CYP2C19*1/*1* pooled livers (Fig 7B). The peptide quantity of CYP2C19 exon 4 and exon 8 was 95% and 81.5%, respectively, (relative to exon 2) in *CYP2C19*1/*1* samples. No wild-type CYP2C19 protein was detected in the *CYP2C19*2/*2* or *CYP2C19*1/*35* livers with either the exon 4 or exon 8 probe. The absence of detectable protein in the *CYP2C19*1/*35* liver was surprising and suggested that person carried an additional deleterious allele. That novel allele is the subject of a separate publication. Nevertheless, the absence of detectable CYP2C19 protein in the *CYP2C19*2/*2* livers with the exon 4 (but not exon 2) probe demonstrates that intron 2 rs12769205 has a functional effect on the CYP2C19 protein independent of the effect of the rs4244285. In fact, it suggests that rs12769205, because it leads to insertion of exon 2B and creates 87 altered amino acids followed by a stop codon,

confers the loss of CYP2C19 protein in *CYP2C19*2/*2* livers because of its primacy (before rs4244285) in the RNA splicing event.

Relationship of *CYP2C19* genotype to activity in Ethiopians. Data on *S*-mephenytoin hydroxylation phenotype and *CYP2C19* genotypes for *CYP2C19*2, CYP2C19*3* and *CYP2C19*17* among 104 Ethiopians was already available from earlier studies and the detailed information about the cohort has been published (Persson et al., 1996; Aklillu et al., 2002; Sim et al., 2006). We genotyped this *in vivo* cohort for the new *CYP2C19*35* allele to perform a phenotype genotype association analysis (Fig 8). Compared to *CYP2C19*1/*1* individuals, those persons heterozygous for one nonfunctional *CYP2C19* allele, either *CYP2C19*1/*35* (p=0.22) or heterozygous for *CYP2C19*1/*3* (p=0.16) did not show a significant increase in *S*/*R*-mephenytoin plasma concentrations, while those heterozygous for both rs12769205 and rs4244285 together (*CYP2C19*1/*2*) (p=0.039) did. As expected, persons homozygous for rs12769205 and rs4244285 together were poor mephenytoin metabolizers, but there were no persons homozygous for rs12769205 alone (*CYP2C19*35/*35)* to conclusively determine the independent functional effect of this SNP *in vivo*.

Extended haplotype homozygosity at the *CYP2C19* locus in Yorubans and other African

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

populations. It was recently reported that the *CYP2C19*2* nonfunctional allele may be under positive selection with *CYP2C19* inactivation conferring an evolutionary advantage in Africa (Janha et al., 2014). To investigate if the signal for selection could be attributed to rs12769205, we used Sweep, a program that uses large scale analysis of haplotype structure in the genome to detect evidence of natural selection, to reanalyze *CYP2C19*2* in 108 YRI and included rs12769205. We first used Sweep to determine *CYP2C19* haplotype structure. Sweep detected haplotype blocks 1, 2 and 3 (defining *CYP2C19*1*, *CYP2C19*2* and *CYP2C19*35*, respectively) (Fig 9A). *CYP2C19*2* rs4244285 is carried only on haplotype 2, while SNP rs12769205 is carried on haplotypes 2 (*CYP2C19*2*) and 3 (*CYP2C19*35*). Sweep then used the long range haplotype test to analyze the haplotypes for long range linkage disequilibrium. *CYP2C19* Haplotypes 2 and 3 in YRI both displayed extended homozygosity as seen by the high relative extended haplotype homozygosity (REHH) scores (Fig 9B). The REHH plots for the three haplotypes, with the core for the haplotypes centered on the genomic position of the *CYP2C19*2* variant, shows that both *CYP2C19*2* (containing both rs4244285 and rs12769205) and *CYP2C19*35* (containing only rs12769205) showed long range LD, suggesting the haplotypes rose rapidly to a high frequency before recombination could break down associations with nearby markers (Sabeti et al., 2002).

The region of longest extended homozygosity and highest REHH of 10 was seen 68-95 kb

from the core with Haplotype 3 (rs12769205 alone on *CYP2C19*35*), while across the same region for Haplotype 2 with rs12769205 + rs4244285 (*CYP2C19*2*) the REHH was 5.5-2.5. Indeed, the significant REHH for both haplotypes 2 and 3 suggests rs12769205, both alone on *CYP2C19*35* and together with rs4244285 on *CYP2C19*2*, confers an evolutionary advantage to these alleles.

The highest REHH scores were 3' of *CYP2C19*. Although the most distal 3' intergenic SNPs flanking *CYP2C19* did not extend into *CYP2C9*, we determined whether either *CYP2C9*2* or *CYP2C9*3* were on these long range *CYP2C19* extended haplotypes. However, none of the 108 YRI from the HapMap project carried either the *CYP2C9*2* (rs1799853) or *CYP2C9*3* (rs1057910) nonfunctional SNPs demonstrating that it is the two *CYP2C19* haplotypes with rs12769205 that are under natural selection.

Sweep was next used to construct a phylogenetic tree of the *CYP2C19* haplotypes (Fig 9C). The haplotypes closer to the ancestral haplotype are at the top of the figure. *CYP2C19*35* is calculated to be closer than *CYP2C19*2* to the ancestral haplotype, and hence *CYP2C19*35* is the ancestral haplotype, and rs12769205 arose before rs4244285.

Next we performed the same analysis on the 396 individuals who represented other African populations (LWK, Luhya in Webuye, Kenya; GWD, Gambian in Western divisions in the Gambia; MSL, Mende in Sierra Leone; and ESN, Esan in Nigeria) still living on the African

continent. The same three *CYP2C19* haplotypes were present in the other African populations, and at a similar frequency to the YRI (Supplemental Figure 3). Likewise, the combined other African populations displayed extended homozygosity for *CYP2C19*35*, as evidenced by the high REHH (7.6, at 95.4 kb from the core). Finally, Sweep ancestral tree analysis generated a *CYP2C19* phylogenetic tree for the other African populations that was identical to that generated for the YRI, again showing the haplotype with rs12769205 alone in the ancestral haplotype, with rs4244285 added later to generate *CYP2C19*2*.

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

DISCUSSION

We discovered that rs12769205 disrupts the branch site in *CYP2C19* intron 2 creating a novel exon 2B. This alternative *CYP2C19* mRNA will generate a nonfunctional protein since insertion of exon 2B creates an out-of-frame protein with 87 novel amino acid residues followed a premature termination codon (PTC), resulting in a truncated 197 amino acid protein. Several lines of evidence showed that rs12769205 leads to intron 2 retention (exon 2B): (a) all livers with rs12769205 generated *CYP2C19* exon 2B; (b) *in silico* splice site strength analysis predicted rs12769205 perturbed the fidelity of intron 2 branch point splice site recognition; and (c) minigenes with rs12769205 transfected into HepG2 cells showed intron 2 inclusion.

Interestingly, rs12769205, that alone defines *CYP2C19*35*, was found together with rs4244285 on the *CYP2C19*2* allele. *CYP2C19*2* was discovered in 1994 (de Morais et al., 1994a) and rs4244285, which clearly leads to altered splicing of exon 5, was the single variant thought to contribute to the *CYP2C19*2* poor metabolizer phenotype. Clearly both rs12769205 and rs4244285 are functionally important as they can individually alter the *CYP2C19* reading frame and produce a premature stop codon resulting in a truncated nonfunctional protein. This leads to the obvious question of whether rs12769205 and rs4244285 contribute equally to the PM phenotype in livers with *CYP2C19*2*. Because we did not have any individuals homozygous for *CYP2C19*35*, we cannot at this time determine, using RNA, whether the residual pool of WT

CYP2C19 transcript differed between those homozygous for the *CYP2C19*2* vs. *CYP2C19*35* alleles. To address this question we took two approaches. First, we used peptide probes to quantify the amount of remaining CYP2C19 wild-type protein in persons carrying *CYP2C19*2* and **35* alleles. While the exon 2 probe (upstream of both rs12769205 and rs4244285) detected residual CYP2C19 protein in livers homozygous for *CYP2C19*2* or with *CYP2C19*35*, the exon 4 probe failed to detect CYP2C19 protein in the same livers. Since the exon 4 probe is downstream of intron 2 rs12769205, but upstream of exon 5 rs4244285, this result suggests the intron 2 SNP can lead to complete loss of CYP2C19 protein due to its primacy in the RNA splicing event. Second, we used extended haplotype homozygosity statistics and uncovered significant evidence that the haplotypes with rs12769205 alone (*CYP2C19*35*) and with rs4244285 on *CYP2C19*2* have undergone positive selection, and that natural selection was not limited to YRI but was seen across African populations, having probably arisen in earlier human ancestors from which all other groups of Africans descended. Notably, the magnitude of evolutionary pressure for both haplotypes with rs12769205 was as great as that exerted on the human genome by infectious diseases (Janha et al., 2014), and implies that rs12769205 on both haplotypes confers an evolutionary advantage. Indeed, the signature of positive selection on the haplotype with rs12769205 alone (*CYP2C19*35*) reinforces that this polymorphisms has a significant functional effect independent of rs4244285.

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

Ancestry analysis showed that, on an evolutionary timescale, rs12769205 is the original *CYP2C19* deleterious polymorphism that arose first on *CYP2C19*35*, and then later added rs4244285 to create the new haplotype (*CYP2C19*2*). While it is possible that the gain of rs4244285 added to *CYP2C19*2* nonfunction, evidence for natural selection on the haplotype with rs12769205 alone, suggests it is sufficient to create the no function allele.

In vivo analysis was unable to conclusively demonstrate that rs12769205 alone contributes to the mephenytoin PM phenotype in Ethiopians because only persons heterozygous for *CYP2C19*35* were available, and there were no persons with *CYP2C19*35* paired with another poor metabolizer allele, and there was only a modest effect of the heterozygous genotype (intermediate phenotype) on mephenytoin disposition. Poor metabolism of mephenytoin was seen in persons homozygous for both rs12769205 + rs4244285 (*CYP2C19*2/*2*). However, there was only a moderate effect of any of the heterozygous PM genotypes (*CYP2C19*1/*2, CYP2C19*1/*35, or CYP2C19*1/*3)* on mephenytoin disposition. Indeed, persons with CYP2C19 intermediate phenotypes (e.g., *CYP2C19*1/*2*) are the most challenging populations to address for proposing clinical pharmacogenetic implementation consortium (CPIC) guidelines (drug and dose recommendations) because of the wide interindividual variability in residual *CYP2C19*1* activity (Scott et al., 2013). Consequently, the most informative subjects,

individuals homozygous for *CYP2C19*35*, or heterozygous for *CYP2C19*35* and another *CYP2C19* PM allele, need to be CYP2C19 phenotyped before comparisons can be made with *CYP2C19*2/*2* poor metabolizers and before any *CYP2C19*35* genotype directed recommendations can be made.

The discovery that rs12769205 leads to alternative *CYP2C19* splicing adds to the growing list of hepatic CYPs where we have identified SNPs leading to polymorphic splicing (*CYP3A5*3*, *CYP3A5*6* and *CYP2B6*6* (Kuehl et al., 2001; Lamba et al., 2003) and nonfunctional alleles. Sakabe and de Souza (2007) proposed that intron retention happens when introns and flanking exons are small. *CYP2C19* exons 2 and 3, 163 nt and 150 nt, respectively, are not large; and, while the average *CYP2C19* intron size is 11,092 nt (range 169 – 38,498 nt), Intron 2 is the smallest (169 nt) and is 6.8x smaller than the next smallest intron (Intron 4, 1161 nt). *CYP2C19* rs12769205 is an interesting example of a branch point SNP leading to intron retention. There are numerous hereditary disease alleles where polymorphisms in branch point motifs lead to loss of splicing activity (Taggart et al., 2012). The branch point signal, located upstream of the polypyrimidine tract, is one of three obligatory signals required for appropriate pre-mRNA splicing. Approximately 96% of branch points fall between -15 and -55 nt relative to the 3'-splice site, with the peak at position -24 nt, and the *CYP2C19* Intron 2 rs12769205 branch

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

point A is located at -23 bp relative to the 3'-splice site. This SNP would disrupt the invariant branch point adenine, a nucleotide that is absolutely required to engage in a 2'-5' phosphodiester bond with the 5′ end of the intron after the first catalytic step of the splicing reaction (Corvelo et al., 2010).

Since we have a large liver resource, it would be extremely useful to have a high-throughput RNA sequencing and analysis pipeline that could identify novel alternative splicing of ADME genes, like *CYP2C19*, that might be caused by sequence variation. This resource is the ideal tissue to look for the consequence of any sequence variant that could lead to alternative splicing because (a) many ADME genes are highly expressed in liver; (b) liver is one of the tissues that generates the highest number of alternative mRNAs per genes; and (c) alternative splicing can show tissue specificity. Hence, any polymorphism with the potential to cause alternative splicing has the best chance of being detected in liver tissue. In theory, RNA-Seq analysis can discover new mRNA transcripts and it would be incredibly valuable if the analysis tools could unambiguously call novel transcripts, such as the inclusion of exon 2B, and report those livers that had these novel transcripts. The assembly of *CYP2C19* mRNAs from short RNA-Seq reads is complicated by the high percent similarity with other neighbor CYP2C family members. For example, exon 3 in *CYP2C19*, shows 99, 95 and 97 percent identity with *2C9*, *2C8* and *2C18*, respectively; and intron 2 in *CYP2C19* shows 96% identity with *CYP2C9*

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

intron 2. Regardless, visual inspection of the IGV views of *CYP2C19* assembled exons 2-3 revealed that there might be inserted nucleotides between these exons (Fig 6). However, the IGV views still required visual analysis of the results, and that we already had PCR results to guide this analysis. The exon junction program analysis (data not shown) correctly called the *novel CYP2C19*2* exon 4/5 junction, but that was because the *CYP2C19*2* alternative mRNA was already in the reference database. The exon junction program *did not* observe a novel junction at exon 2 in rs12769205 livers, and would require that the transcript with exon 2B, first, be added to the reference database as the current programs don't have a reliable intron retention caller (D. Finkelstein, personal communication). Hence, because RNA-Seq requires a reference mRNA, and because small RNAs and long non coding RNAs can also inhabit introns, identification of polymorphic alternative mRNAs, particularly those with intron retention, may not yet be unequivocally identified by these high throughput approaches. An additional complicating factor is that, while *CYP2C19* is a highly expressed gene in human liver, the alternative *CYP2C19* transcripts generated by rs12769205 +/- rs4244285 ultimately lead to premature termination codons (PTC) and these will trigger accelerated alternative mRNA degradation through nonsense-mediated decay (NMD), decreasing the amount of alternative mRNA. Indeed, it has been suggested that the rate of intron retention in mRNA transcripts is higher than reported because NMD filters off some mRNA transcripts (Aten et al., 2013).

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

Importantly, it also makes the identification of alternative mRNAs linked to PTCs and NMD (those that are linked to functional consequences) harder to identify by an RNA-Seq approach.

Acknowledgements

We gratefully acknowledge the technical support of St Jude Children's Research Hospital: the Hartwell Center for DNA sequencing and the Computational Biology and Bioinformatics Core; the Pharmacogenetics Research Network network-wide RNA Sequencing Project (Kathy Giacomini) and the Baylor College of Medicine Human Genome Sequencing Center (Steve Scherer) for the RNA-Seq data.

Author Contributions:

Participated in research design: Schuetz, Chaudhry, Thummel

Conducted experiments: Chaudhry, Prasad, Aklillu, Sim

Contributed new reagents or analytic tools: Prasad

Performed data analysis: Schuetz, Chaudhry, Fohner, Prasad, , Finkelstein, Fan, Wu,

Wang, Aklillu,

Wrote or contributed to the writing of the manuscript: Schuetz, Chaudhry, Prasad, Finkelstein,

Wang, Aklillu, Thummel

References:

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

Anders S, Pyl PT, and Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31(2)**: 166-169.

Aten E, Sun Y, Almomani R, Santen GW, Messemaker T, Maas SM, Breuning MH, and den Dunnen JT (2013) Exome sequencing identifies a branch point variant in Aarskog-Scott syndrome. *Hum Mutat* **34**: 430-434.

Barrett JC, Fry B, Maller J, and Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**: 263-265.

Chhibber A, French CE, Yee SW, Gamazon ER, Theusch E, Qin X, Webb A, Papp AC, Wang A,

Simmons CQ, Konkashbaev A, Chaudhry AS, Mitchel K, Stryke D, Weiss ST, Kroetz DL, Sadee W, Nickerson D, Krauss RM, George AL Jr, Schuetz EG, Medina MW, Cox NJ, Scherer SE, Giacomini KM, and Brenner SE (2015) Transcriptomic variation of pharmacogenes in multiple human tissues and lymphoblastoid cell lines. *Pharmacogenomics Journal* (communicated).

Corvelo A, Hallegger M, Smith CW, and Eyras E (2010) Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol* **6(11)**: e1001016.

de Morais SM, Wilkinson GR, Blaisdell J, Nakamura K, Meyer UA, and Goldstein JA (1994a) The major genetic defect responsible for the polymorphism of S-mephenytoin metabolism in humans. *J Biol Chem* **269**: 15419-1522.

de Morais SM, Wilkinson GR, Blaisdell J, Meyer UA, Nakamura K, and Goldstein JA (1994b) Identification of a new genetic defect responsible for the polymorphism of (S)-mephenytoin metabolism in Japanese. *Mol Pharmacol* **46**: 594-598.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29(1):** 15-21.

Downing JR, Wilson RK, Zhang J, Mardis ER, Pui CH, Ding L, Ley TJ, and Evans WE (2012) The Pediatric Cancer Genome Project. *Nat Genet* **44(6):** 619-622.

Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, and Beroud C (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Research* **37**: e67.

Edson KZ, Prasad B, Unadkat JD, Suhara Y, Okano T, Guengerich FP, and Rettie AE (2013) Cytochrome P450-dependent catabolism of vitamin K: ω-hydroxylation catalyzed by human CYP4F2 and CYP4F11. Biochemistry **52(46)**: 8276-8285.

Gooding C, Clark F, Wollerton MC, Grellscheid S-N, Groom H, and Smith CWJ (2006) A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. *Genome Biology* **7**: R1.

Gordon AS, Tabor HK, Johnson AD, Snively BM, Assimes TL, Auer PL, Ioannidis JP, Peters U, Robinson JG, Sucheston LE, Wang D, Sotoodehnia N, Rotter JI, Psaty BM, Jackson RD, Herrington DM, O'Donnell CJ, Reiner AP, Rich SS, Rieder MJ, Bamshad MJ, and Nickerson

DA; On Behalf of the NHLBI GO Exome Sequencing Project (2014) Quantifying rare, deleterious variation in 12 human cytochrome P450 drug-metabolism genes in a large-scale exome dataset. *Hum Mol Genet* **23(8)**: 1957-1963.

Hicks JK, Swen JJ, Thorn CF, Sangkuhl K, Kharasch ED, Ellingrod VL, Skaar TC, Müller DJ, Gaedigk A, and Stingl JC; Clinical Pharmacogenetics Implementation Consortium (2013) Clinical Pharmacogenetics Implementation Consortium guideline for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants. *Clin Pharmacol Ther* **93**: 402-408.

Hirota T, Eguchi S, and Ieiri I (2013) Impact of genetic polymorphisms in CYP2C9 and CYP2C19 on the pharmacokinetics of clinically used drugs. *Drug Metab Pharmacokinet* **28(1)**: 28-37.

Janha RE, Worwui A, Linton KJ, Shaheen SO, Sisay-Joof F, and Walton RT (2014) Inactive alleles of cytochrome P450 2C19 may be positively selected in human evolution. *BMC Evol Biol* **14**:71.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL (2013) TopHat2: accurate

alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14(4)**: R36.

Kuehl P, Zhang J, Lin Y, Lamba J, Assem M, Schuetz J, Watkins PB, Daly A, Wrighton SA, Hall SD, Maurel P, Relling M, Brimer C, Yasuda K, Venkataramanan R, Strom S, Thummel K, Boguski MS, and Schuetz E (2001) Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nat Genet* **27**: 383-391.

Lamba V, Lamba J, Yasuda K, Strom S, Davila J, Hancock ML, Fackenthal JD, Rogan PK, Ring B, Wrighton SA, and Schuetz EG (2003) Hepatic CYP2B6 expression: gender and ethnic differences and relationship to CYP2B6 genotype and CAR (constitutive androstane receptor) expression. *J Pharmacol Exp Ther* **307**: 906-922.

Li-Wan-Po A, Girard T, Farndon P, Cooley C, and Lithgow J (2010) Pharmacogenetics of CYP2C19: functional and clinical implications of a new variant CYP2C19*17. *Br J Clin Pharmacol* **69**: 222-230.

McGraw J and Waller D (2012) Cytochrome P450 variations in different ethnic populations.

Expert Opin Drug Metab Toxicol **8**: 371-382.

Owusu Obeng A, Egelund EF, Alsultan A, Peloquin CA, and Johnson JA (2014) CYP2C19 polymorphisms and therapeutic drug monitoring of voriconazole: are we ready for clinical implementation of pharmacogenomics? *Pharmacotherapy* **34(7)**: 703-18.

Persson I, Aklillu E, Rodrigues F, Bertilsson L, and Ingelman-Sundberg M (1996) S-mephenytoin hydroxylation phenotype and CYP2C19 genotype among Ethiopians. *Pharmacogenetics* **6**: 521-526.

Prasad B and Unadkat JD (2014) Optimized approaches for quantification of drug transporters in tissues and cells by MRM proteomics. AAPS J 16(4): 634-648.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, and Mesirov JP (2011) Integrative genomics viewer. *Nat Biotechnol* **29(1)**: 24-26.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski

D, Ward R, and Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419(6909)**: 832-837.

Sakabe NJ and de Souza SJ (2007) Sequence features responsible for intron retention in human. *BMC Genomics* 8: 59.

Sanz EJ, Villen T, Alm C, and Bertilsson L (1989) S-mephenytoin hydroxylation phenotypes in a Swedish population determined after coadministration with debrisoquin. *Clin Pharmacol Ther* **45**: 495-499.

Scott SA, Sangkuhl K, Stein CM, Hulot JS, Mega JL, Roden DM, Klein TE, Sabatine MS, Johnson JA, and Shuldiner AR (2013); Clinical Pharmacogenetics Implementation Consortium. Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update. *Clin Pharmacol Ther* **94(3)**: 317-323.

Shah BS, Parmar SA, Mahajan S, and Mehta AA (2012) An insight into the interaction between clopidogrel and proton pump inhibitors. *Curr Drug Metab* **13**: 225-235.

Shirasaka Y, Chaudhry AS, Prasad B, Wong T, Calamia JC, Fohner A, Isoherranen N, Rettie AE, Schuetz EG, and Thummel KE (2015). Interindividual variability of CYP2C19-catalyzed drug metabolism due to differences in diplotypes. *Pharmacogenetics and Genomics* (Communicated).

Shirasaka Y, Sager JE, Lutz JD, Davis C, and Isoherranen N (2013) Inhibition of CYP2C19 and CYP3A4 by Omeprazole Metabolites and Their Contribution to Drug-Drug Interactions. *Drug Metab Dispos* **41**: 1414-1424.

Sim SC, Risinger C, Dahl ML, Aklillu E, Christensen M, Bertilsson L, and Ingelman-Sundberg M (2006) A common novel CYP2C19 gene variant causes ultrarapid drug metabolism relevant for the drug response to proton pump inhibitors and antidepressants. *Clin Pharmacol Ther* **79**: 103-113.

Singh G and Cooper TA (2006). Minigene reporter for identification and analysis of cis elements and trans factors affecting pre-mRNA splicing. *Biotechniques* 41(2): 177-181.

Taggart AJ, DeSimone AM, Shih JS, Filloux ME, and Fairbrother WG (2012). Large-scale

mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol* **19(7)**: 719-721.

Tybring, G and Bertilsson, L (1992) A methodological investigation on the estimation of the S-mephenytoin hydroxylation phenotype using the urinary S/R ratio. *Pharmacogenetics* **2**: 241-243.

Wang L, Prasad B, Salphati L, Chu X, Gupta A, Hop CE, Evers R, Unadkat JD (2015) Interspecies variability in expression of hepatobiliary transporters across human, dog, monkey, and rat as determined by quantitative proteomics. *Drug Metab Dispos* **43**: 367-374.

Wedlund PJ (2000) The CYP2C19 enzyme polymorphism. *Pharmacology* **61**: 174-183.

Footnotes

a) This work was supported in part by the National Institutes of Health National Institute of General Medical Sciences [Grants [GM092666] (E.G.S), [GM32165] (K.E.T) and the Pharmacogenomics Research Network (PGRN) through the network-wide RNA Sequencing Project [GM61390] (Kathy Giacomini) and [GM061388] (Steve Scherer) at the Baylor College of Medicine Human Genome Sequencing Center (Steve Scherer); the National Institutes of Health National Cancer Institute [Cancer Center Support Grant [P30 CA21765] (E.G.S); the American Lebanese Syrian Associated Charities (ALSAC) (E.G.S).

b) Person to receive reprint requests: Erin Schuetz, Ph.D. Department of Pharmaceutical Sciences, 262 Danny Thomas Place, St. Jude Children's Research Hospital, Memphis, TN-38105; Phone: (901) 595-2205; FAX: (901) 595-3125

Email: erin.schuetz@stjude.org

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

Figure Legends:

Figure 1. *CYP2C19*35* (rs12769205) leads to aberrant intron 2 retention (exon 2B). (A) *CYP2C19* cDNA was amplified from human liver cDNAs by PCR using primers in exon 2 and exon 4 and the wild-type product (269 bp) and alternatively spliced product (438 bp) analyzed on agarose gel. Homozygous wild-type (Lanes 1-7 are *CYP2C19*1/*1),* rs12769205 heterozygous (lanes 8-15) and homozygous (lanes 16-17) variant genotypes are indicated by the open, half-filled and filled boxes, respectively. Lanes marked M and –ve represent the 100 bp DNA ladder and a negative control, respectively. (B) The 438 bp fragment was excised from the gel and directly sequenced and the resulting electropherogram and nucleotide sequence are shown. The cartoon illustrates the amplification strategy, the insertion of exon 2B in samples with rs12769205, and the location of rs12769205 at the branch point adenine -23 nt upstream of the Intron 2 splice acceptor site.

Figure 2. Visual genotypes of rs12769205 and rs4244285 in different populations (1000 genomes phase 3 data) and Liver samples. Visual genotypes for rs12769205 and rs4244285 in 1000genomes samples (genotypes downloaded from http://browser.1000genomes.org/index.html): *(A)* 99 White (CEU), (B) 108 Yoruban (YRI), (C) 661 all Africans, and (D) 103 Han Chinese in Beijing, China (CHB); and in the St Jude Liver

Bank (SJLB) samples from (E) 272 White and (F) 63 African Americans. Grey, orange and red boxes indicate homozygous wild-type, heterozygous and homozygous variant genotypes, respectively. // lines indicate not all subjects are shown for that particular diplotype and population.

Figure 3. Effect of rs4244285 and rs12769205 on *CYP2C19* splicing. (A) Eight liver samples were analyzed by PCR for (A) the *CYP2C19* exon 5-40 bp deletion, caused by rs4244285, using primers in exons 4 and 6, and (B) the CYP2C19 exon 2B insertion, caused by rs12769205, using primers in exons 2 and 4 and the products analyzed on agarose gels. Arrows indicate the migration of the canonical and splice variant bands. Homozygous wild-type, heterozygous and homozygous variant genotypes are indicated by the open, half-filled and filled boxes, respectively. In Panel B, the smaller residual amount of the 269 bp wild-type mRNA is seen in rs12769205 heterozygous (Lanes 3, 4, and 6) or homozygous (lanes 5, 8) samples compared to samples 1, 2 and 7 homozygous for the *CYP2C19*1/*1* genotype.

Figure 4. Structure of alternative *CYP2C19* mRNAs in livers with rs4244285 and rs12769205 genotypes. *CYP2C19* was PCR amplified using exon 2 and 6 primers, in liver samples with indicated genotypes and the products analyzed on agarose gels. Arrows indicate

the migration of the canonical and splice variant bands. Genotypes are illustrated as depicted in Fig 3 legend. The cartoon illustrates the structures of the canonical exons (open boxes), novel exon 2B (gray box) and partial deletions of 40 bp and 71 bp at the start of exon 5 and exon 4 (black boxes), respectively in the splice variant (SV) transcripts.

Figure 5. RT-PCR analysis of minigene mRNA products. (A) Exon2-intron2-exon3 Minigenes: Two minigenes contained *CYP2C19* exon 2 + intron 2 + exon 3 (differing only by rs12769205A>G) that replaced the *BamH1/Xho1* fragment (minigene exon (Em)) of the RHCglo minigene. (B) Exon 5 minigenes: Two minigenes contained the last 87 nucleotides of *CYP2C19* intron 4 and all of exon 5 (differing only by rs4244285) that replaced the *Sal1/Xho1* fragment of the RHCglo minigene. HepG2 cells were transfected with each of the four plasmids and minigene mRNA products were analyzed by RT-PCR using the RSV5U/TNIE4 primers and the products analyzed on agarose gels. The rs12769205 minigene generated a transcript with the exon 2B insertion (A), and the rs4244285 minigene generated a transcript with the 40 bp exon 5 deletion (B).

Figure 6. Integrative genomic viewer (IGV) visualization of human liver *CYP2C19* mRNA. RNA-Seq results for human livers that were heterozygous (half filled box) or homozygous

DMD Fast Forward. Published on May 28, 2015 as DOI: 10.1124/dmd.115.064428
This article has not been copyedited and formatted. The final version may differ from this version.

wild-type (open box) for: (A) rs12769205 were visualized across *CYP2C19* exons 2-Exon3; and (B) rs4244285 were visualized across exon 5. Due to differences in *CYP2C19* read coverage, IGV was scaled between 0-50 for six of the liver samples, and between 0-250 for one sample. Panel A shows that only samples heterozygous for rs12769205 had RNA sequences with intron 2 included. Panel B shows that only samples heterozygous for rs4244285 had a decreased signal across the first 40 bp of exon 5, indicative of the heterozygous deletion of this region of the mRNA transcript.

Figure 7. CYP2C19 peptides downstream of exon 2 failed to detect residual CYP2C19 wild-type protein in *CYP2C19*2* and *CYP2C19*35* liver microsomes. (A) The location of the peptide probes used relative to *CYP2C19* exons and to rs12769205 and rs4244285. (B) Expression of CYP2C19 protein quantified in pooled human liver microsomes from *CYP2C19*1/*1*, *CYP2C19*2/*2*, and *CYP2C19*1/*35* livers using exon 2, exon 4 and exon 8 specific peptide probes. Results are graphed relative to CYP2C19 protein in *CYP2C19*1/*1* pooled liver microsomes (100%). >LLOQ, less than the lower limit of quantitation.

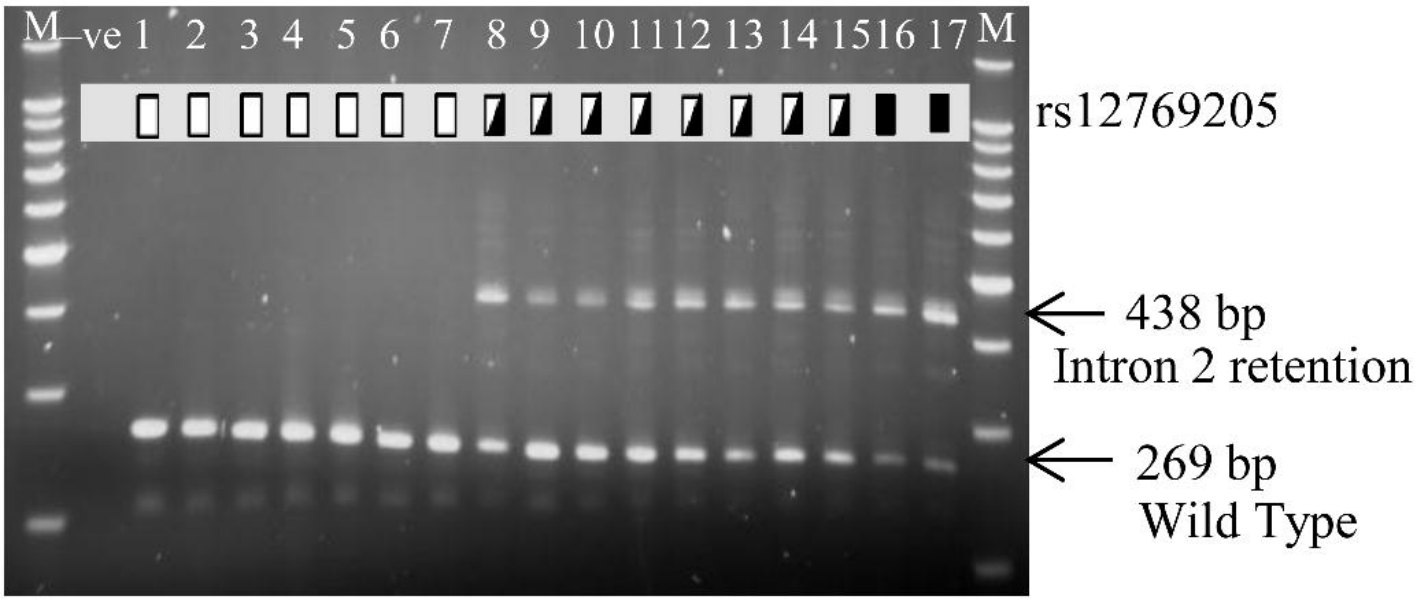
Figure 8. CYP2C19 activity in 104 Ethiopians with different *CYP2C19* genotypes.

S/*R*-mephenytoin ratio is plotted for each *CYP2C19* diplotype group. Box plots indicate the 25th

and 75th percentile, and the bold line within the box represents median and whiskers represent the range after excluding the outliers. Statistical analyses were performed using R version 3.1 (http://www.rproject.org). A general linear model was used to obtain p-values for each group compared to the *CYP2C19*1/*1* group. N, number of subjects.

Fig 9. *CYP2C19* haplotype frequencies, extended haplotype homozygosity, and ancestral tree in 216 YRI chromosomes. (A) Sweep was used to determine *CYP2C19* haplotypes (SNP positions rs12769205, rs4244285, rs4417205 and rs3758580) and their frequencies. The "." in the observed haplotypes represents nucleotides that match the ancestral allele. "GCG" below the SNP rsIDs are the Sweep predicted ancestral allele nucleotides based on the HapMap Release 16. (B) Relative extended haplotype homozygosity (REHH) for each *CYP2C19* haplotype with the core of the haplotypes centered on rs4244285. Both haplotypes containing SNP rs12769205 either alone (haplotype 3 (green, *CYP2C19*35*)), or with rs4244285 (haplotype 2 (orange, *CYP2C19*2*)) show extended haplotype homozygosity REHH. (C) Phylogenetic tree of the *CYP2C19* haplotypes. Haplotypes closer to the ancestral are at the top of the figure. The area of the squares is proportional to the frequency of the haplotype. The gray squares represent haplotypes not present in the data, but that are missing links in the phylogeny. The program determined the ancestral root of the tree was *CYP2C19*1*.

A.


M -ve 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 M
rs12769205
438 bp
Intron 2 retention
269 bp
Wild Type

B.

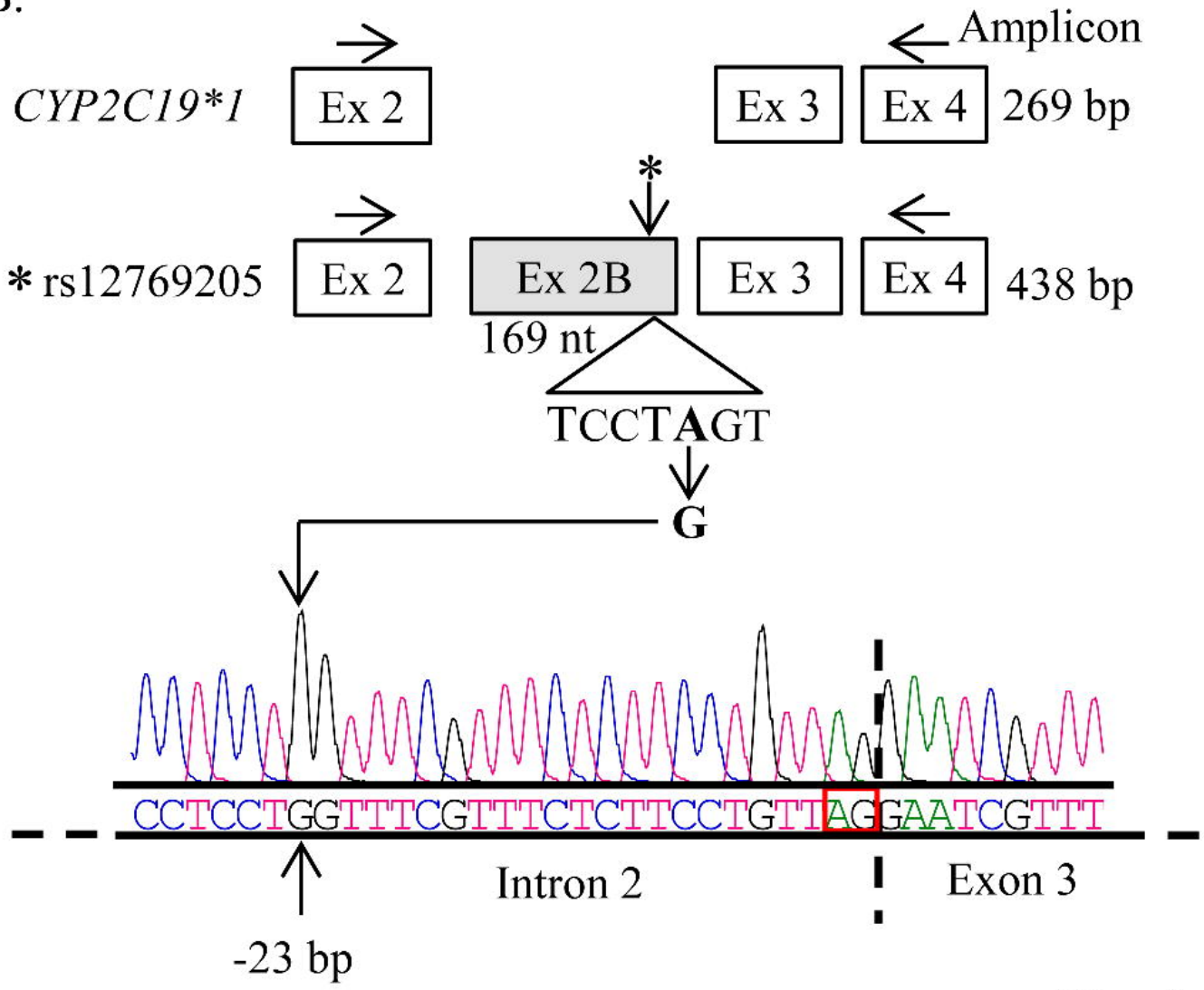

Amplicon
CYP2C19*1
Ex 2
Ex 3
Ex 4
269 bp
*
* rs12769205
Ex 2
Ex 2B
Ex 3
Ex 4
438 bp
169 nt
TCCTAGT
G
CCTCCTGGTTTCGTTTCTCTTCCTGTTAGGAATCGTTT
Intron 2
Exon 3
-23 bp

Fig. 1

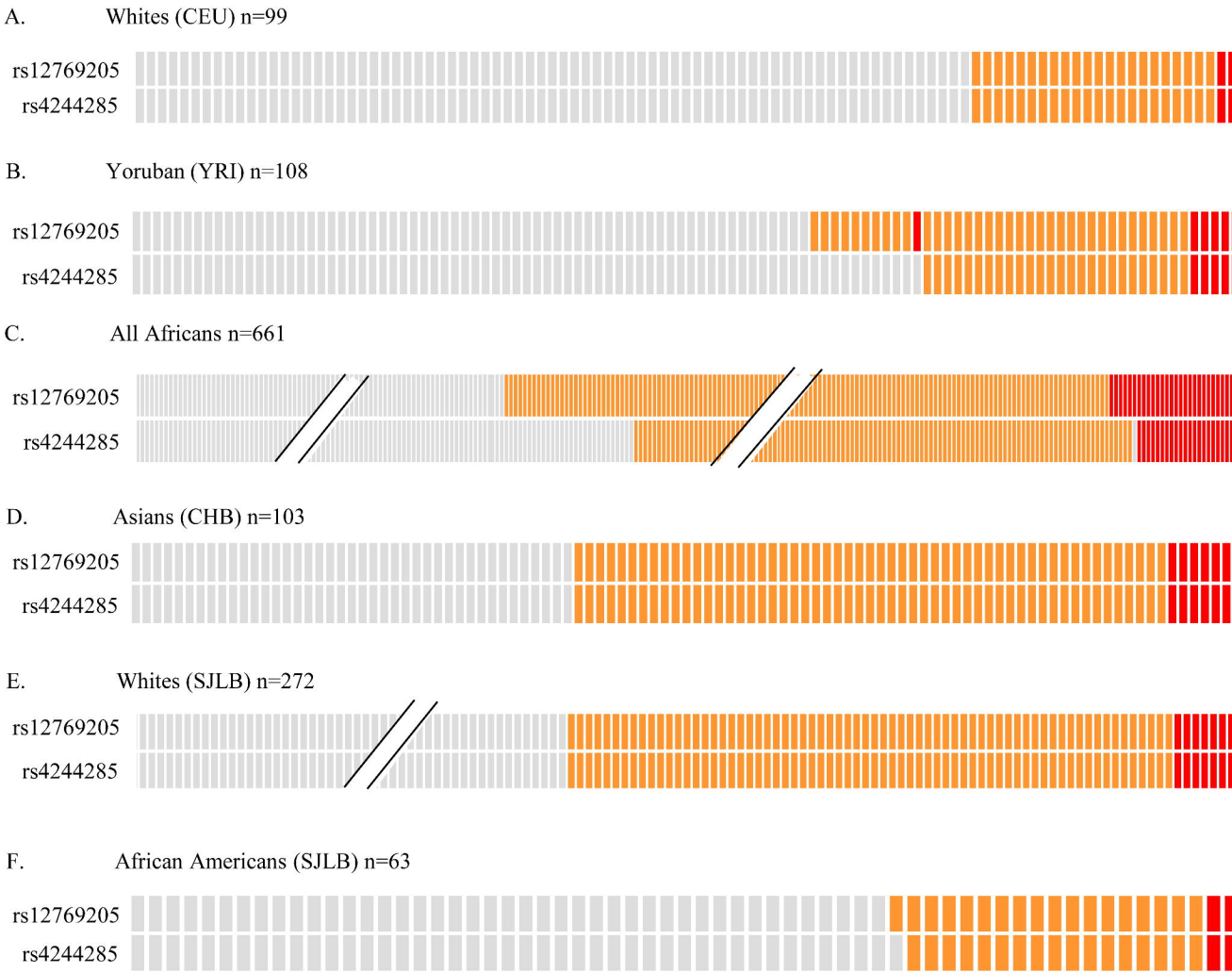
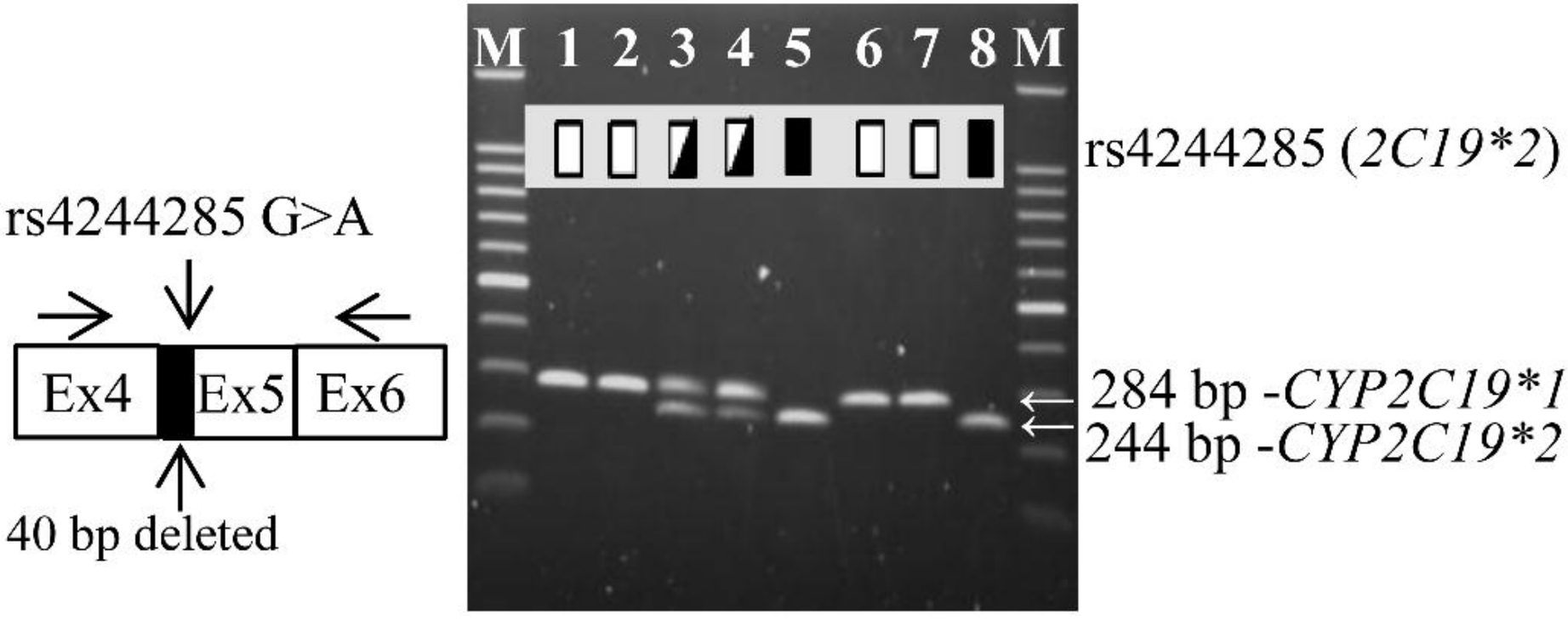

A.
Whites (CEU) n=99
rs12769205
rs4244285
B.
Yoruban (YRI) n=108
rs12769205
rs4244285
C.
All Africans n=661
rs12769205
rs4244285
D.
Asians (CHB) n=103
rs12769205
rs4244285
E.
Whites (SJLB) n=272
rs12769205
rs4244285
F.
African Americans (SJLB) n=63
rs12769205
rs4244285

Fig. 2

A.


rs4244285 G>A
Ex4
Ex5
Ex6
40 bp deleted
M 1 2 3 4 5 6 7 8 M
rs4244285 (2C19*2)
284 bp -CYP2C19*1
244 bp -CYP2C19*2

B.

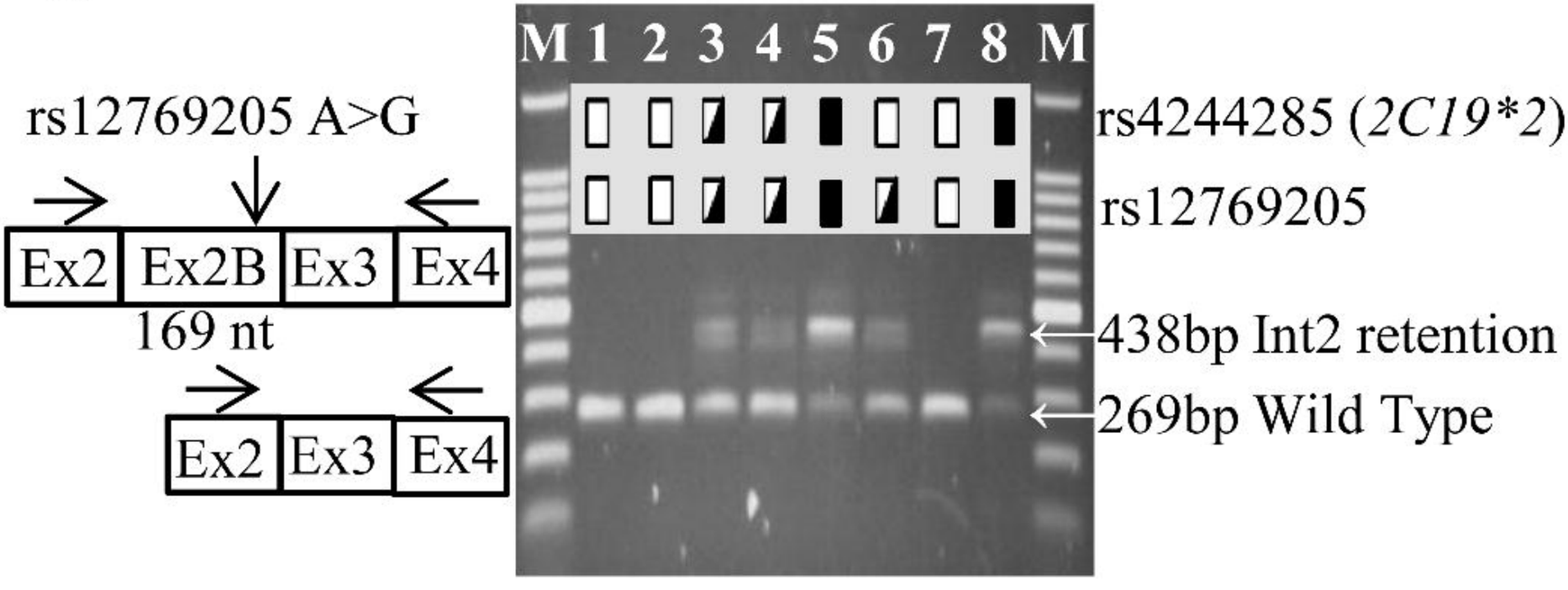

rs12769205 A>G
Ex2
Ex2B
Ex3
Ex4
169 nt
Ex2
Ex3
Ex4
M 1 2 3 4 5 6 7 8 M
rs4244285 (2C19*2)
rs12769205
438bp Int2 retention
269bp Wild Type

Fig. 3

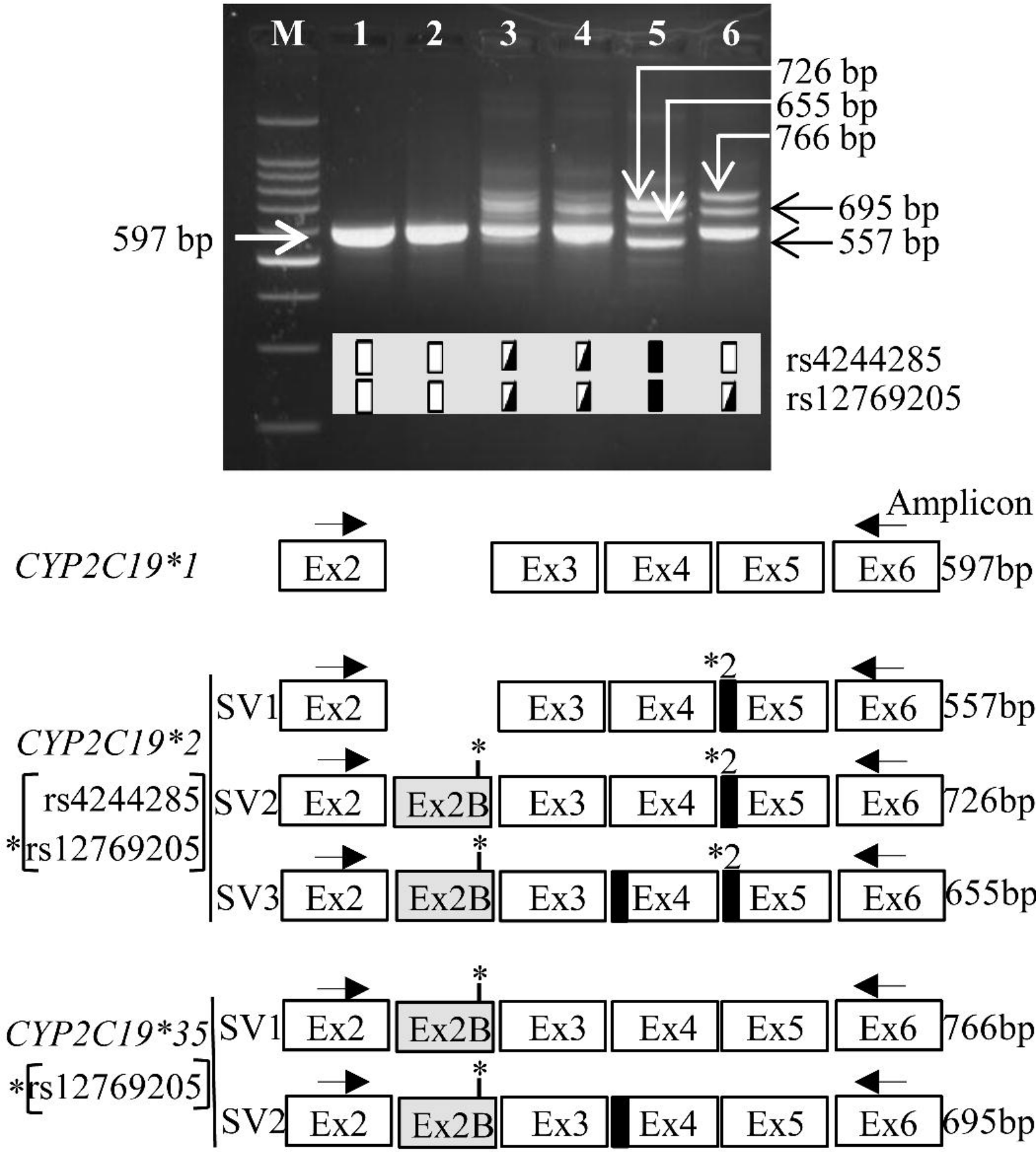

M
1
2
3
4
5
6
726 bp
655 bp
766 bp
695 bp
557 bp
597 bp
rs4244285
rs12769205
Amplicon
CYP2C19*1
Ex2
Ex3
Ex4
Ex5
Ex6
597bp
*2
SV1
557bp
CYP2C19*2
*
SV2
Ex2B
726bp
rs4244285
*rs12769205
SV3
655bp
CYP2C19*35
766bp
695bp

Fig. 4

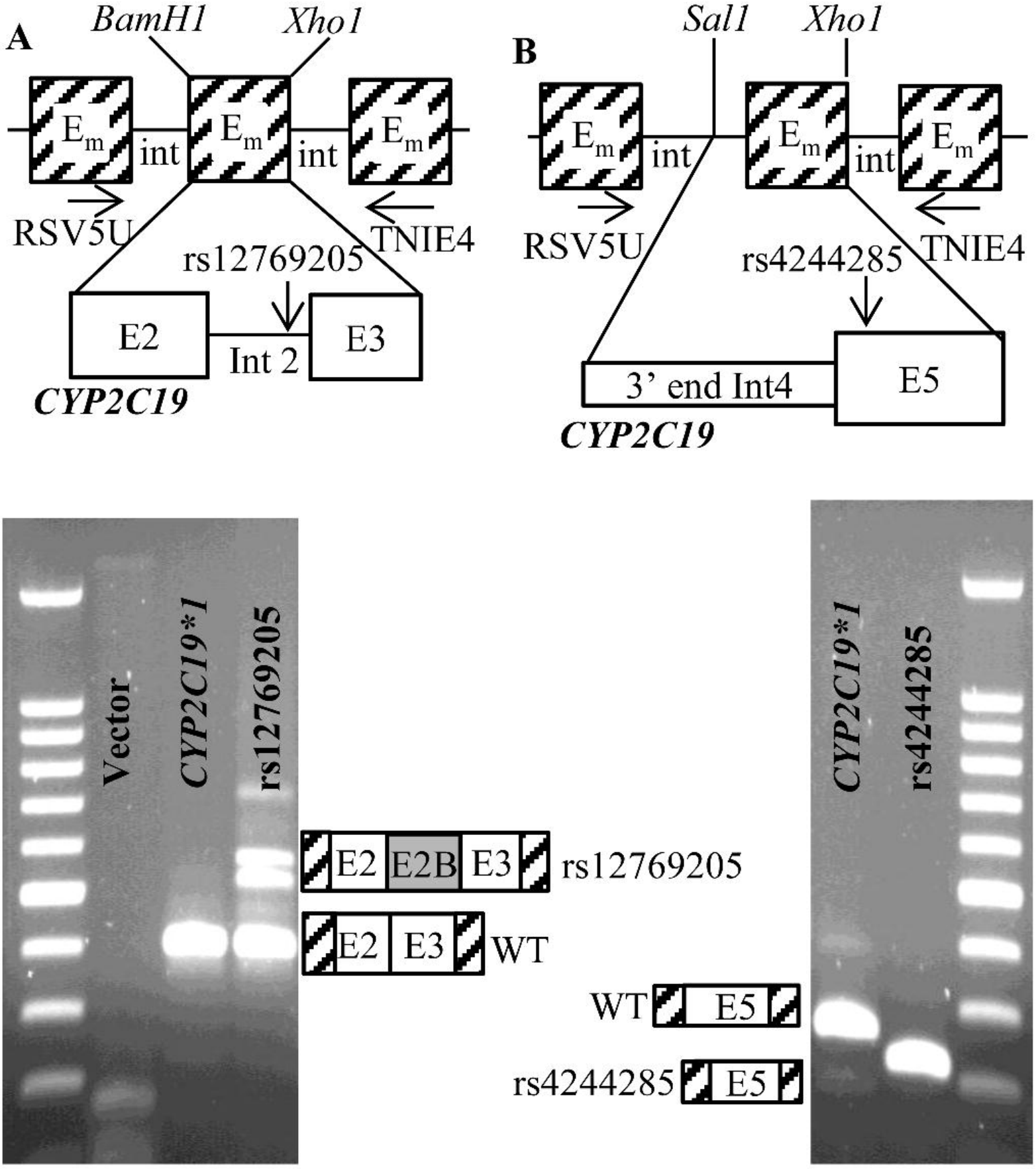
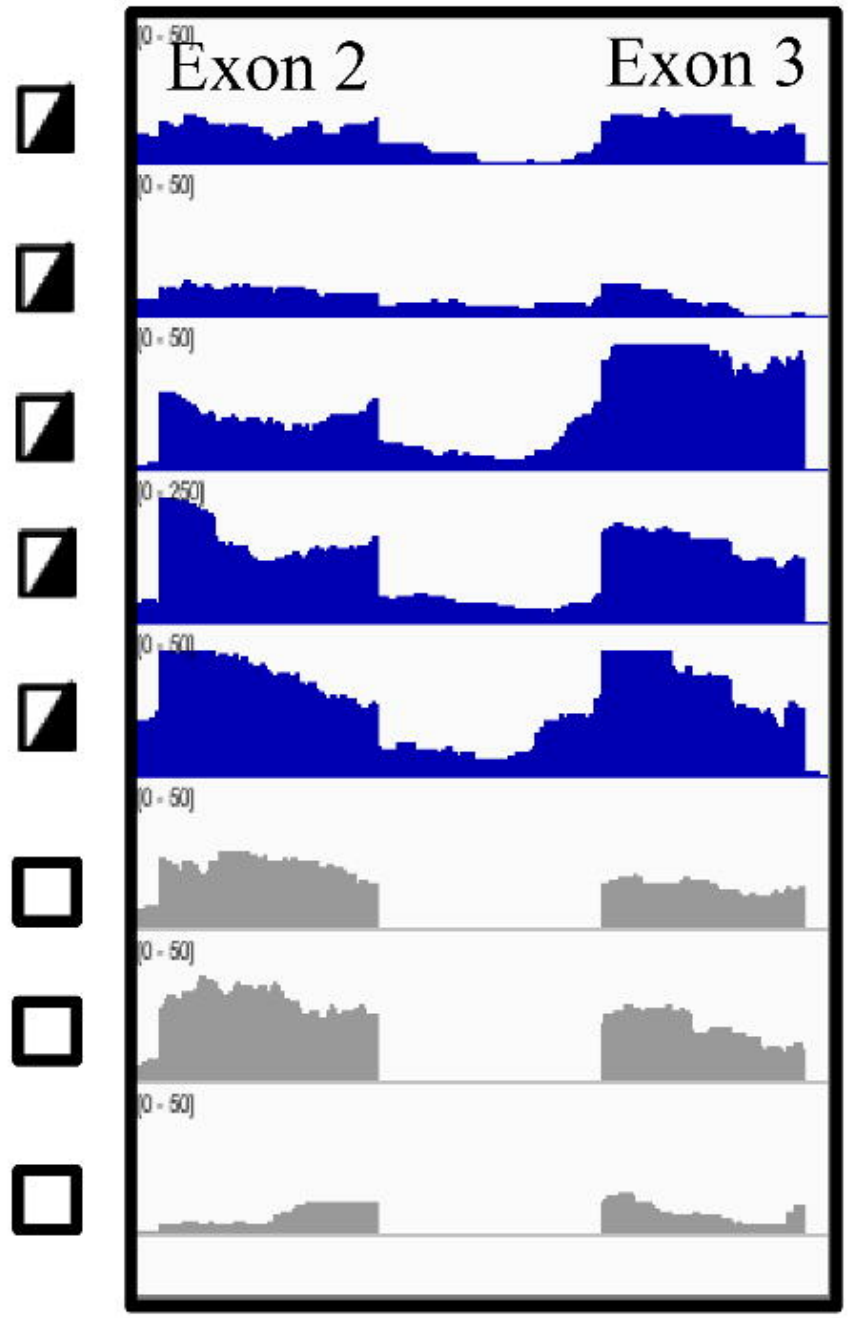

A
BamH1
Xho1
Em
int
Em
int
Em
RSV5U
TNIE4
rs12769205
E2
Int 2
E3
CYP2C19
B
Sal1
Xho1
Em
int
Em
int
Em
RSV5U
rs4244285
TNIE4
3' end Int4
E5
CYP2C19
Vector
CYP2C19*1
rs12769205
E2
E2B
E3
rs12769205
E2
E3
WT
WT
E5
rs4244285
E5
CYP2C19*1
rs4244285

Fig. 5

A.

CYP2C19 intron 2 retention

rs12769205



B.

CYP2C19 exon 5 partial deletion

rs4244285

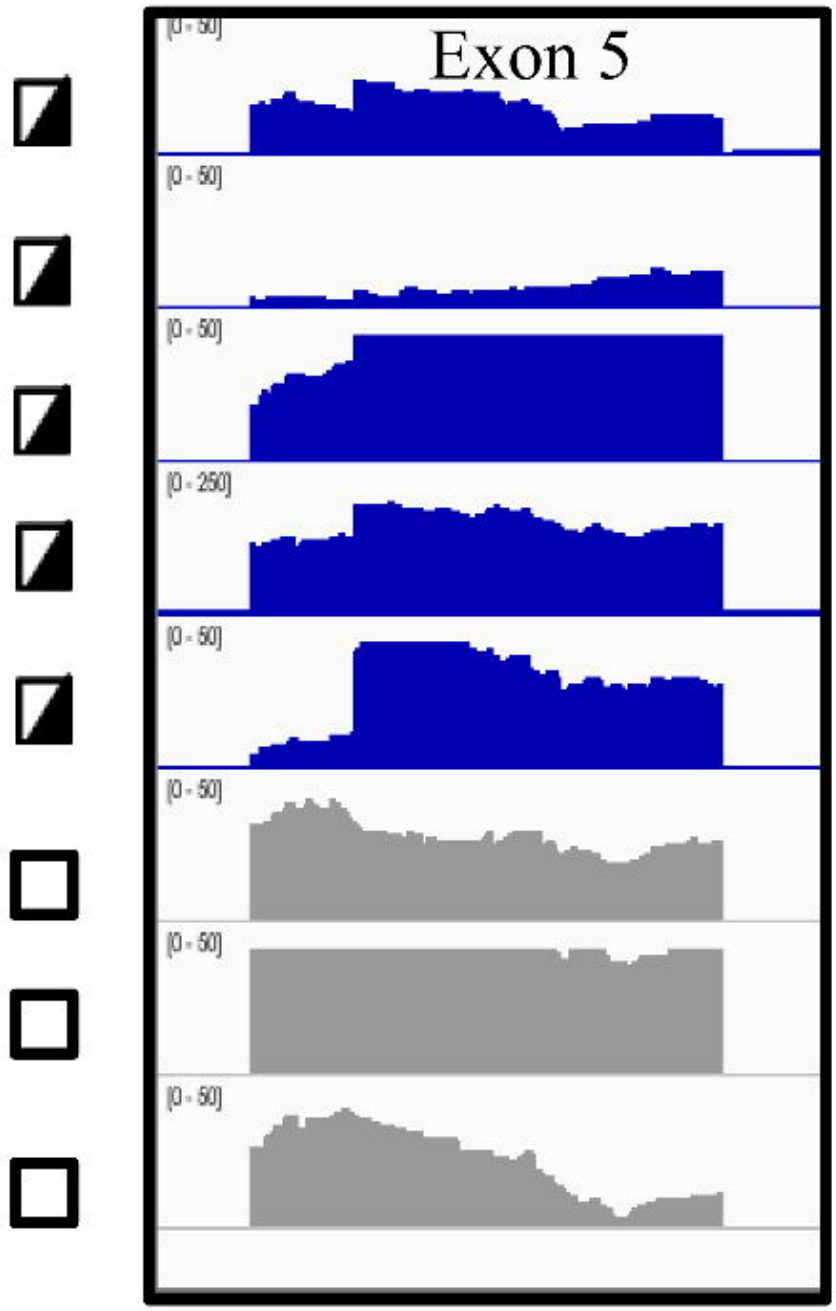
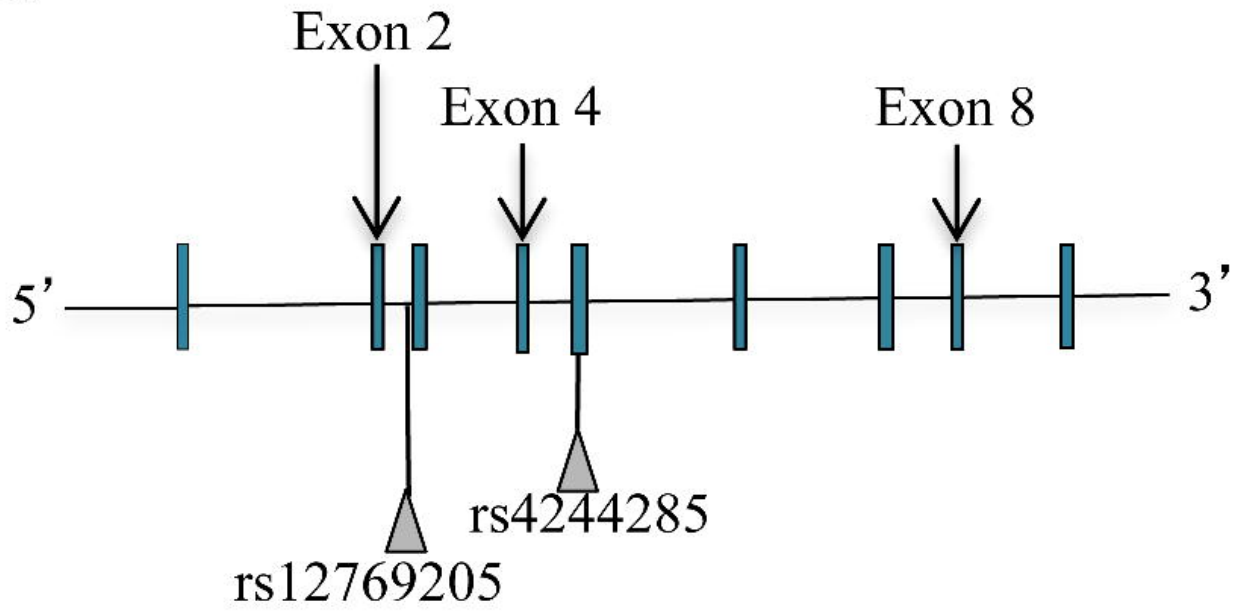


Fig. 6

A.


Exon 2
Exon 4
Exon 8
5'
3'
rs4244285
rs12769205

B.

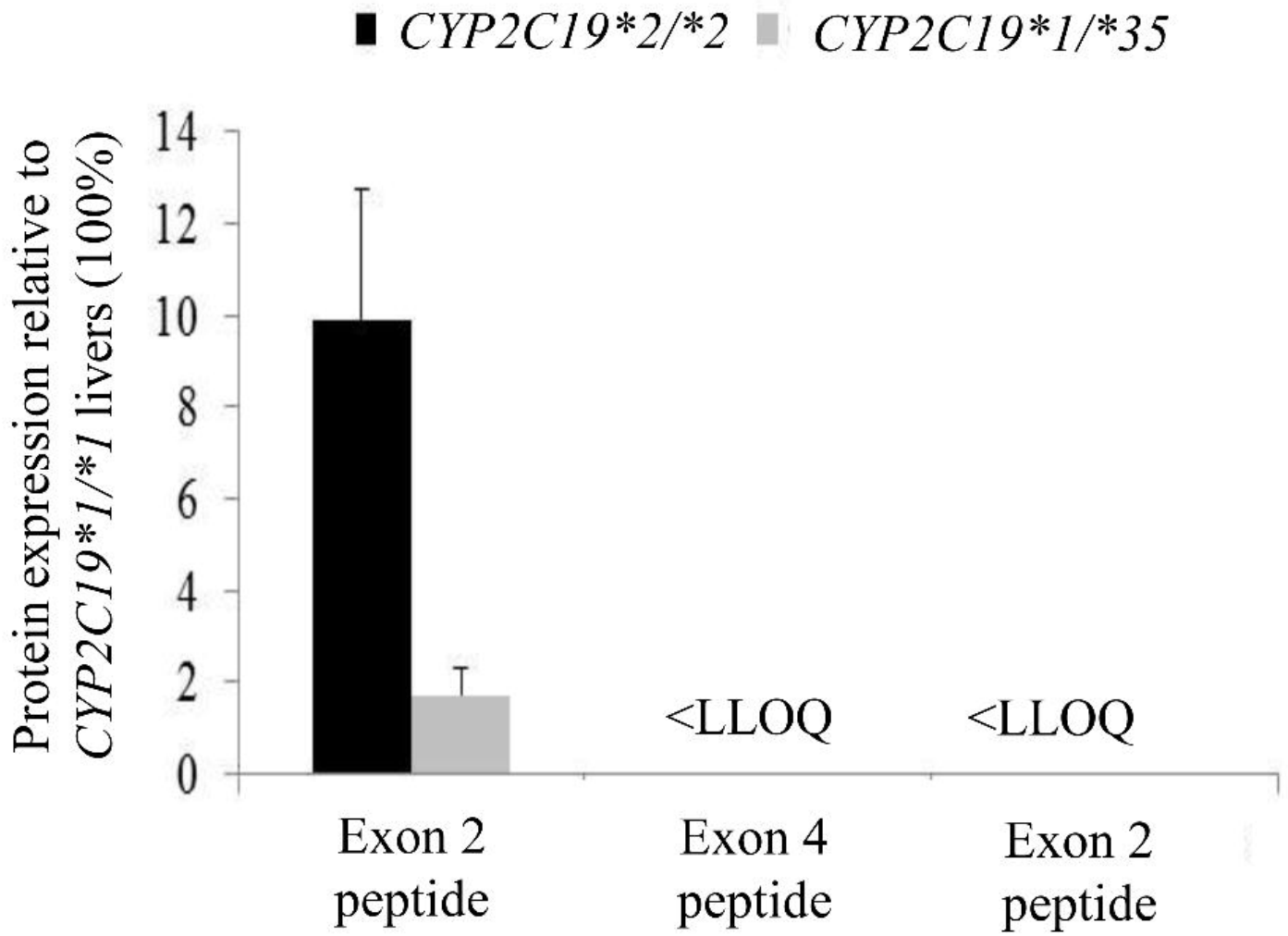

CYP2C19*2/*2
CYP2C19*1/*35
Protein expression relative to CYP2C19*1/*1 livers (100%)
14
12
10
8
6
4
2
0
<LLOQ
<LLOQ
Exon 2 peptide
Exon 4 peptide
Exon 2 peptide

Fig. 7

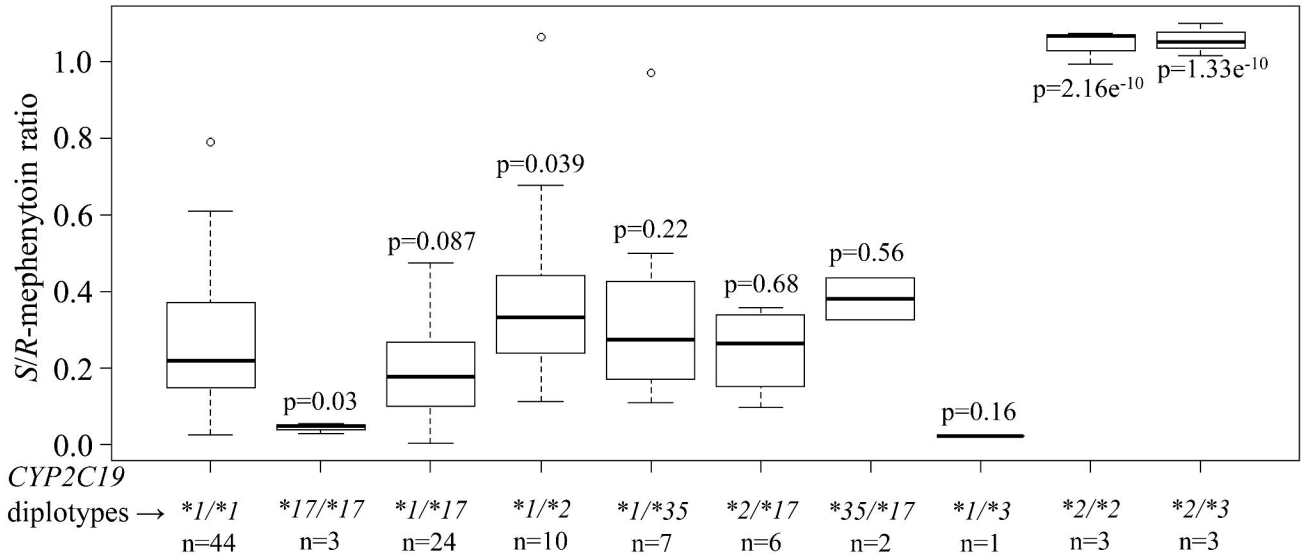

S/R-mephenytoin ratio
1.0
0.8
0.6
0.4
0.2
0.0
p=0.03
p=0.087
p=0.039
p=0.22
p=0.68
p=0.56
p=0.16
p=2.16e-10
p=1.33e-10
CYP2C19
diplotypes →
*1/*1
n=44
*17/*17
n=3
*1/*17
n=24
*1/*2
n=10
*1/*35
n=7
*2/*17
n=6
*35/*17
n=2
*1/*3
n=1
*2/*2
n=3
*2/*3
n=3

Fig. 8

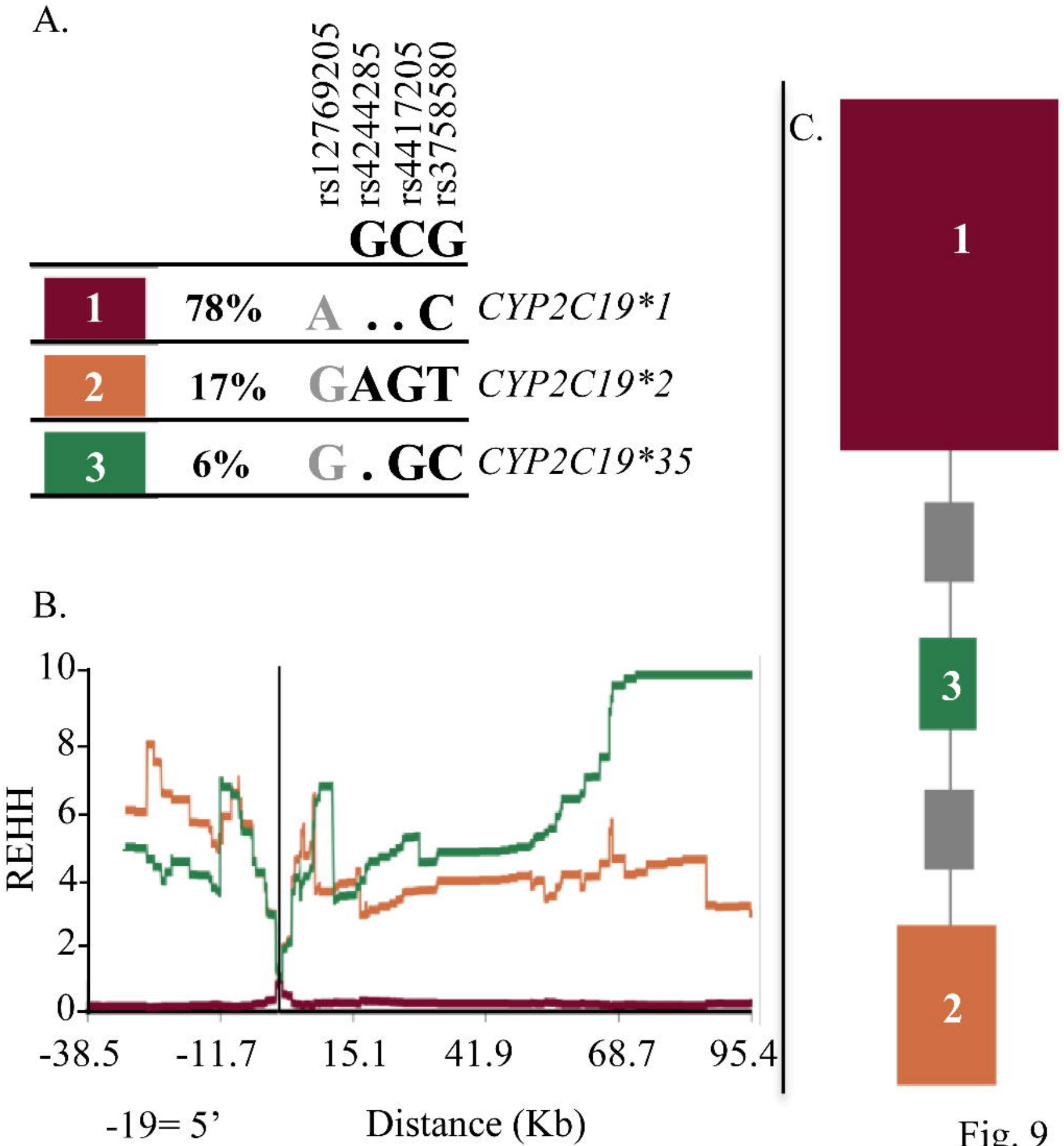

A.
rs12769205
rs4244285
rs4417205
rs3758580
GCG
1
78%
A . . C
CYP2C19*1
2
17%
GAGT
CYP2C19*2
3
6%
G . GC
CYP2C19*35
B.
REHH
10
8
6
4
2
0
-38.5
-11.7
15.1
41.9
68.7
95.4
-19= 5'
Distance (Kb)
C.
1
3
2

Fig. 9

Drug Metabolism Disposition

*The CYP2C19 intron 2 branch point SNP is the ancestral polymorphism contributing to the poor metabolizer phenotype in livers with CYP2C19*35 and CYP2C19*2 alleles*

Amarjit S. Chaudhry, Bhagwat Prasad, Yoshiyuki Shirasaka, Alison Fohner, David Finkelstein, Yiping Fan, Shuoguo Wang, Gang Wu, Eleni Aklillu, Sarah Sim, Kenneth E. Thummel and Erin G. Schuetz.

Supplemental Table 1: Primers used for PCR amplification and sequencing of *CYP2C19* exons from genomic DNA.

Exons	FP/RP	Primer sequence	PCR amplified fragment size (base pairs)
Exon 1	FP	5'-AGTGGGCCTAGGTGATTGGCCACTT-3'	410
	RP	5'-TCAAAGTATTTTACTTTACAATGATCTC-3'	
Exon 2-3	FP	5'-AAAATATGAATCTAAGTCAGGCTTAGT-3'	607
	RP	5'-GGAGAGCAGTCCAGAAAGGTCAGTGATA-3'	
Exon 4	FP	5'-TGCTTTTAAGGGAATTCATAGG-3'	383
	RP	5'-AAAATGTACTTCAGGGCTTGG-3'	
Exon 5	FP	5'-CAACCAGAGCTTGGCATATTG-3'	409
	RP	5'-TGATGCTTACTGGATATTCATGC-3'	
Exon 6	FP	5'–AAAACTGGCACAAGACAGGGATG-3'	456
	RP	5'-AAATTGGGACAGATTACAGCTGCG-3'	
Exon 7	FP	5'-AATTGCTAGAACAAATGTTCCATTTC-3'	327
	RP	5'–AGAGGGTAAGAATCATACTGTGA-3'	
Exon 8	FP	5'–CCACTGTTTCTTAAACCTTCGTGA-3'	284
	RP	5'-GAAGGCACATGTAAGTTCCAACTGA-3'	
Exon 9	FP	5'-ATCTACTCATCCCTCCTATGATTCACCG-3'	529
	RP	5'-ATGTGGCACTCAATGTAACTATTATAGA-3'	

Supplemental Table 2: Optimized MS/MS parameters of CYP2C19 surrogate peptides used for protein quantification

Peptide	Amino acid position	Parent ion	Product ion	Cone (V)	CE (eV)
IYGPVFTLYFGLER	60-73 (exon 2)	838.2	998.3	52	28
		838.2	1145.4	52	28
GTTILTSLTSVLHDNK	384-399 (exon 8)	567.3	664.4	35	17
		567.3	607.8	35	17
ASPC[160]DPTFILGC[160]APC[160]NVIC[160]SIIFQK*	161-185 (exon 4)	1434.68	535.32	60	42
		1434.68	989.44	60	42

*C[160] indicates S-carbamidomethylated cysteine after alkylation of native peptide.

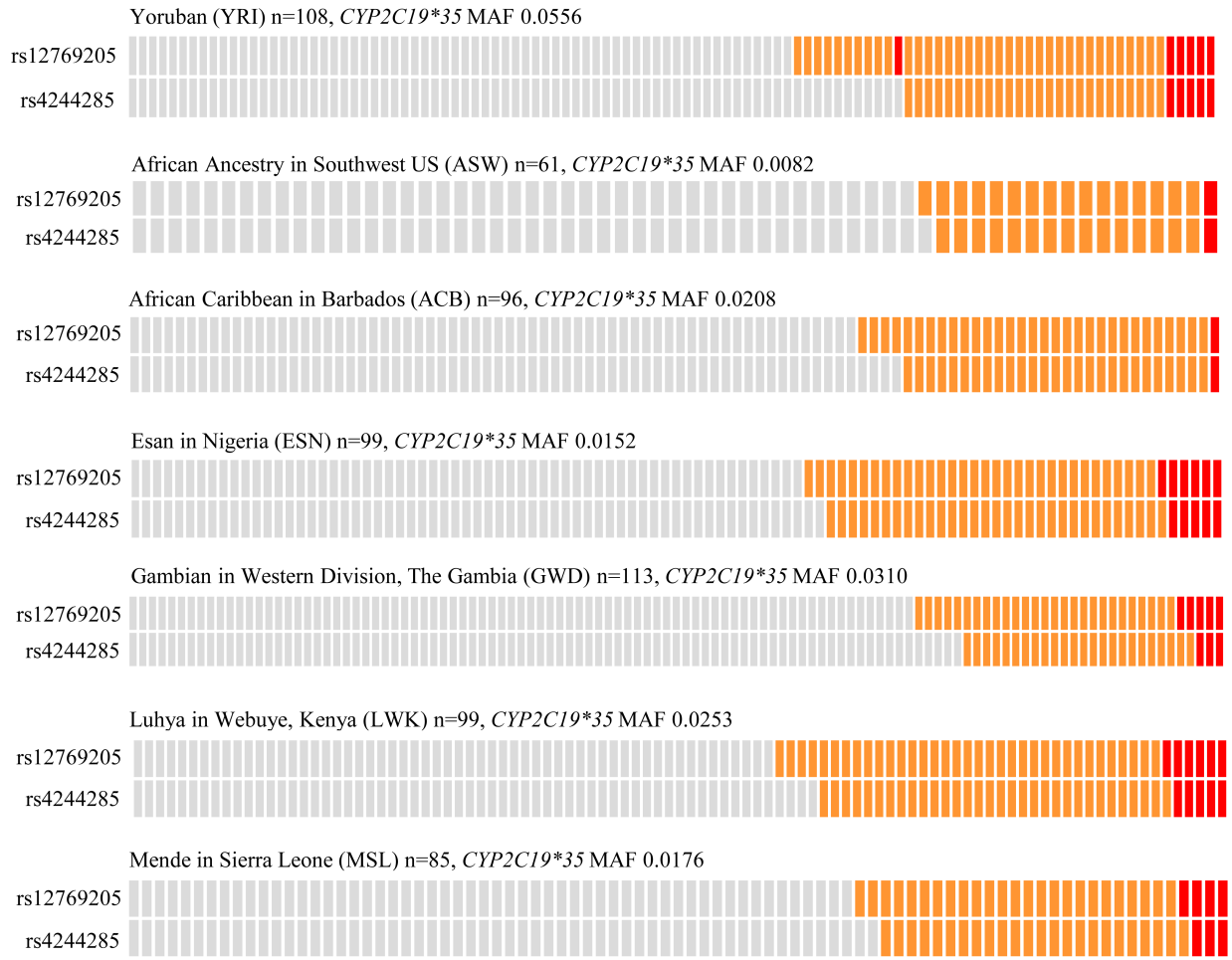
Supplemental Table 3: *CYP2C19*2/*35* diplotypes and *CYP2C19*2* and *CYP2C19*35* allele frequencies in different populations (1000 genome-Phase 3 data)

	All Africans*	Yorubans (YRI)	Caucasians (CEU)	Chinese (CHB)
Diplotype	**n =661**	**n=108**	**n=99**	**n=103**
*CYP2C19*1/*1*	430	66	75	41
*CYP2C19*1/*2*	174	26	22	55
*CYP2C19*1/*35*	28	10	0	0
*CYP2C19*2/*35*	5	0	0	0
*CYP2C19*2/*2*	23	5	2	7
*CYP2C19*35/*35*	1	1	0	0
Allele Frequency				
*CYP2C19*1*	0.8033	0.7778	0.8687	0.6650
*CYP2C19*2*	0.1702	0.1667	0.1313	0.3350
*CYP2C19*35*	0.0265	0.0556	0.0000	0.0000

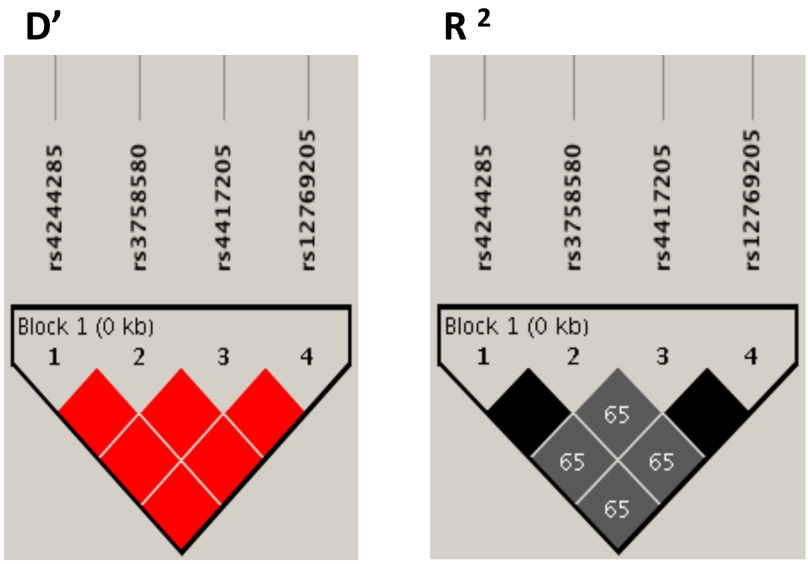
*All Africans includes all seven populations of African descent (YRI, ASW, ACB, ESN, GWD, LWK and MSL) as defined in Supplemental Figure 1.

Supplemental Table 4: Full length sequencing of the *CYP2C19*35* cDNA

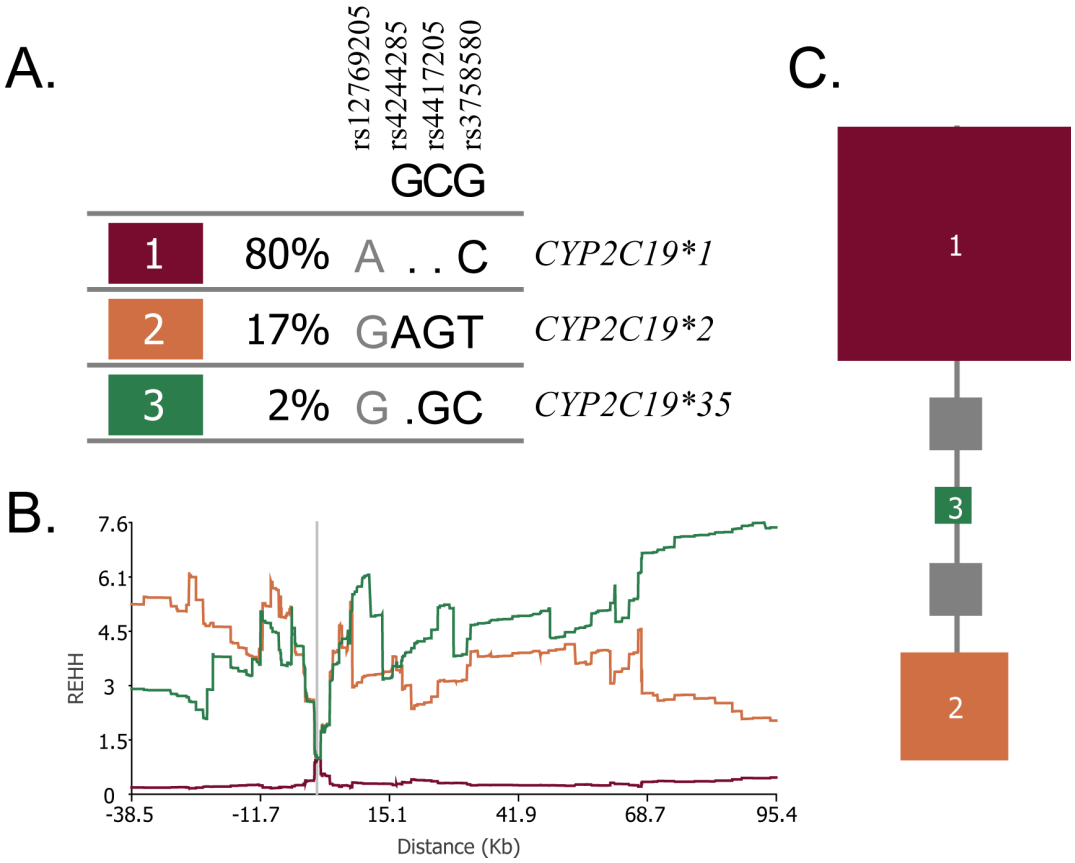
SNP rsID	Liver Genotype observed	CYP2C19*35 allele	Gene position	mRNA Position	Location in intron	Exon	Amino acid change
rs17885098	CT	T	99C>T	99C>T	-	Ex 1	Pro33Pro
rs12769205	AG	G	12662A>G	-	Ex 3-23A>G	Int 2	-
rs3758581	GG	G	80161A>G	991A>G	-	Ex 7	Ile331Val



Supplemental Figure 1: Visual genotypes of rs12769205 and rs4244285 in populations of African descent (1000 genomes phase 3 data). Visual genotypes for rs12769205 and rs4244285 in 1000 genomes samples (genotypes downloaded from http://browser.1000genomes.org/index.html). The populations of African descent are each listed with the number of individuals (n) and the *CYP2C19*35* minor allele frequency. Grey, orange and red boxes indicate homozygous wild-type, heterozygous and homozygous variant genotypes, respectively.



Supplemental Figure 2: Linkage disequilibrium (LD) map for the common *CYP2C19* allelic variants in the YRI population. (Left panel) A LD map of the *CYP2C19* SNPs rs4244285, rs3758580, rs4417205 and rs12769205 in the YRI population was created using Haploview 4.2. The red squares show complete LD with statistical significance (|D'| = 1, LOD > 2). (Right panel) R^2 LD values. The black squares show the SNPs have the highest correlation with each other, the gray boxes indicate a correlation of 0.65. The LD values in the figure are scaled from 1.0 to 100 for visual clarity.



Supplemental Figure 3: *CYP2C19* haplotype frequencies, extended haplotype homozygosity, and ancestral tree in other African populations in the 1000 genome project. The African populations (excluding YRI) living on the African continent (LWK, Luhya in Webuye, Kenya; GWD, Gambian in Western divisions in the Gambia; MSL, Mende in Sierra Leone; and ESN, Esan in Nigeria) were combined (n=396). (A) Sweep was used to determine *CYP2C19* haplotypes (SNP positions rs12769205, rs4244285, rs4417205 and rs3758580) and their frequencies were determined (see Fig. legend 9). (B) Relative extended haplotype homozygosity (REHH) for each *CYP2C19* haplotype with the core of the haplotypes centered on rs4244285. Both haplotypes containing SNP rs12769205 either alone (haplotype 3 (green, *CYP2C19*35*)), or with rs4244285 (haplotype 2 (orange, *CYP2C19*2*)) show extended haplotype homozygosity REHH. (C) Phylogenetic tree of the *CYP2C19* haplotypes. Haplotypes closer to the ancestral are at the top of the figure. The area of the squares is proportional to the frequency of the haplotype. The gray squares represent haplotypes not present in the data, but that are missing links in the phylogeny. The program determined the ancestral root of the tree was *CYP2C19*1*.