Characterization of ADME gene variation in 21 populations by exome sequencing

Daniel H. Hovelson[c], Zhengyu Xue[a], Matthew Zawistowski[c], Margaret G. Ehm[b], Elizabeth C. Harris[a], Sophie L. Stocker[d], Annette S. Gross[d], In-Jin Jang[e], Ichiro Ieiri[g], Jong-Eun Lee[f], Lon R. Cardon[b], Stephanie L. Chissoe[b], Gonçalo Abecasis[c] and Matthew R. Nelson[b]

Objective **Proteins involving absorption, distribution, metabolism, and excretion (ADME) play a critical role in drug pharmacokinetics. The type and frequency of genetic variation in the ADME genes differ among populations. The aim of this study was to systematically investigate common and rare ADME coding variation in diverse ethnic populations by exome sequencing.**

Materials and methods **Data derived from commercial exome capture arrays and next-generation sequencing were used to characterize coding variation in 298 ADME genes in 251 Northeast Asians and 1181 individuals from the 1000 Genomes Project.**

Results **Approximately 75% of the ADME coding sequence was captured at high quality across the joint samples harboring more than 8000 variants, with 49% of individuals carrying at least one 'knockout' allele. ADME genes carried 50% more nonsynonymous variation than non-ADME genes ($P=8.2\times10^{-13}$) and showed significantly greater levels of population differentiation ($P=7.6\times10^{-11}$). Out of the 2135 variants identified that were predicted to be deleterious, 633 were not on commercially available ADME or general-purpose genotyping arrays. Forty deleterious variants within important ADME genes, with frequencies of at least 2% in at least one population, were identified as candidates for future pharmacogenetic studies.**

Conclusion **Exome sequencing was effective in accurately genotyping most ADME variants important for pharmacogenetic research, in addition to identifying rare or potentially de novo coding variants that may be clinically meaningful. Furthermore, as a class, ADME genes are more variable and less sensitive to purifying selection than non-ADME genes.** ***Pharmacogenetics and Genomics*** **27:89–100 Copyright © 2017 The Author(s). Published by Wolters Kluwer Health, Inc.**

Pharmacogenetics and Genomics 2017, **27**:89–100

Keywords: absorption, distribution, genetic variants, excretion, exome sequencing, metabolism, multiethnic populations, pharmacogenetics

[a]GlaxoSmithKline Research and Development, Durham, North Carolina, [b]GlaxoSmithKline Research and Development, King of Prussia, Pennsylvania, [c]Department of Biostatistics and/or Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA, [d]GlaxoSmithKline Research and Development, Sydney, New South Wales, Australia, [e]Seoul National University, [f]DNA Link Inc., Seoul, Korea and [g]Kyushu University, Fukuoka, Japan

Correspondence to Matthew R. Nelson, PhD, Department of Genetics, GlaxoSmithKline Research and Development, 709 Swedeland Road, King of Prussia, PA 19406, USA
Tel: + 1 610 270 6167; e-mail: matthew.r.nelson@gsk.com

Present address: Zhengyu Xue and Elizabeth C. Harris: PAREXEL International Corp., Durham, North Carolina, USA

Present address: Sophie L. Stocker: St Vincent's Hospital Sydney, Darlinghurst, New South Wales, Australia

Received 11 May 2016 **Accepted** 15 November 2016

Introduction

The selection of a therapeutically effective and safe dose is crucial for drug development and requires assessment of intrinsic and extrinsic factors that influence drug pharmacokinetics (concentration and duration of drug exposure). Proteins involving absorption, distribution, metabolism, and excretion (ADME) play an important role in determining the pharmacokinetic profile of a drug. There is considerable genetic variation in genes encoding ADME proteins, both within and between populations, which can affect drug pharmacokinetics [1,2]. For example, the product labels for over 20 psychotropic medicines have been updated with ADME genetic variation that contributes toward interindividual pharmacokinetic variability [3]. Interpopulation differences can also be very important. For example, at the same dose, two-fold higher rosuvastatin concentrations are observed in individuals of East Asian than European ancestry and as a consequence the recommended initial rosuvastatin daily dose in East Asians (5 mg) is half that in other populations (10 mg) [4]. This interethnic difference in pharmacokinetics has been related, in part, to interethnic differences in the frequencies of genetic variants of

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (*www.pharmacogeneticsandgenomics.com*).

1744-6872 Copyright © 2017 The Author(s). Published by Wolters Kluwer Health, Inc. DOI: 10.1097/FPC.0000000000000260

functional consequence in the genes encoding drug transporters, namely, BCRP (*ABCG2*) and OATP1B1 (*SLCO1B1*), which are important determinants of rosuvastatin pharmacokinetics [5–7]. Given the importance of developing medicines for patients worldwide, and the increasing globalization of clinical drug development, identifying and quantifying all ADME genetic variations that contribute to interethnic differences in drug pharmacokinetics, efficacy, and safety is of extreme interest [8,9].

Ultimately, an understanding of variations in ADME genes across global populations may direct strategies for preclinical and clinical development, and aid in interpreting pharmacokinetic data from multiregional clinical studies. Despite this, there have been few systematic efforts to comprehensively identify and characterize all common and low-frequency variants in ADME genes across diverse populations. Some studies have explored the impact of population-specific variation of known ADME gene variants for specific genes and/or populations [10–13]. Li *et al.* [14] used publicly available Human Genome Diversity Project [15] and phase III HapMap data [16] to explore patterns of selection in ADME genes across global populations, reporting greater than expected diversity in ADME genes. Ramos *et al.* [17] observed wide population differentiation in variants assayed on the Affymetrix DMET Plus Array (Affymetrix, Santa Clara, California, USA) in 1478 samples from 19 populations. However, the variants investigated in most studies suffer from ascertainment bias, in that most single nucleotide variants (SNVs) were discovered in populations of European ancestry or in small numbers of patients with aberrant drug pharmacokinetics. Rare variation is abundant in protein-coding regions [18,19], and there is a need to identify the low frequency and rare ADME coding variants that may be important for studies of drug pharmacokinetics and provide insights into the selective forces that shaped the patterns of variation in these genes. Despite the recent availability of samples with exome and whole-genome sequence data, no studies to date have leveraged the power of next-generation sequencing (NGS) to more comprehensively investigate the ADME coding variation in samples collected from diverse ethnic populations.

We therefore sought to use data derived from commercial exome-sequencing capture arrays to systematically investigate SNVs, insertions, and deletions (indels) and copy number variants (CNVs) in samples from diverse ethnic populations with the purpose of identifying coding variants that may influence ADME protein function and could be incorporated into drug disposition studies. Using an industry-standard list of ADME genes as a starting point [14,20], we evaluated the capability of exome sequencing to capture ADME coding variation across 298 ADME genes (38 core and 260 extended), highlighting the discovery of novel and uncharacterized ADME coding variation. We compiled a list of 1062 SNVs from the literature known to be important for drug ADME, and report allele frequencies from the 21 different populations in this study for polymorphic coding variants on this list. For this analysis and interpretation, we paid special attention to Northeast Asian populations as this region is now home to more than 22% of the global population and is contributing an increasing proportion of patients to global clinical trials. Furthermore, Northeast Asian regulatory authorities consider the relevance of foreign clinical data to their local populations. Consequently, it is important to understand the profile of ADME gene variants of functional significance in Northeast Asian populations.

In addition, population genetic parameters (e.g. variation per kilobase, F_{ST} values) providing insight into the population history of these genes as well as evolutionary forces are summarized, and the rates of variation in ADME genes are compared with rates of variation found elsewhere in the genome. Finally, the strengths and limitations of using exome sequencing as an approach to carrying out ADME genetic studies are explored.

Materials and methods

Samples and data

The Northeast Asian Variation Analysis (NEAVA) samples included 251 healthy individuals recruited from Kyushu, Japan ($N=125$), and Seoul, South Korea ($N=126$). All participants provided written informed consent for this genetic research, which was reviewed and approved by institutional review boards and independent ethics committees according to local guidelines. Genomic DNA extracted from blood was used for exome sequencing using the Agilent SureSelect Human All Exon 50 Mb Kit (Agilent, Santa Clara, California, USA) as exome captures and the HiSeq 2000 (Illumina, San Diego, California, USA) as pair-end 150 bp reads. Targeted exome sequencing was carried out to an average depth of 60 × exome-wide on all participants.

Exome sequence data of 1181 samples from 19 populations from the 1000 Genomes Project (1000G) [21] were combined with the NEAVA data for this analysis (Table S1, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). Three different exome capture and sequencing protocols (NimbleGen SeqCap EX Exome v2, NimbleGen v1 2.1M Human Exome, and Agilent SureSelect All Exon v2) were used for 1000G samples [21,22], whereas a fourth protocol (the Agilent SureSelect Human All Exon kit) was used for the NEAVA samples. Therefore, a set of regions captured at high depth across all 1432 samples were identified for most downstream analyses and capture depth was calculated across all regions in the intersection of 1000G capture protocol targets as well as across the full set of Agilent SureSelect Human All Exon capture regions (Fig. S1, Supplemental digital content 1, *http://links.lww.com/FPC/B140*).

Bioinformatics analyses

Approximately 526 Kbp of coding sequence for 38 core and 260 extended ADME genes as defined in a study by Pharm-ADME [20], with minor modifications, were identified on the basis of the UCSC knownGene table [23] (Supplemental digital content 2, *http://links.lww.com/FPC/B141*). Sequencing reads were mapped and aligned to the reference human genome (GRCh37) and decoy sequences used in the 1000G [24] by BWA software v6.1 [25]. Base qualities were calibrated using GATK v1.6-13 TableRecalibration tool [26] after removing duplicate reads using Picard's MarkDuplicates tool v1.73 (*https://broadinstitute.github.io/picard/*). Unique reads with Phred-scaled quality of at least 20 were retained for variant calling.

SNV detections and calls were carried out across ± 50 bp of the Agilent's targeted regions using the gotCloud variant calling pipeline [27]. The quality of the variant calls was evaluated by comparing the sequencing-based genotypes with publicly available array-based genotypes for six 1000G samples (Table S4, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). Indel sites were discovered from the NEAVA samples ($N=251$) using both Dindel [28] and the GATK v2.1-12 UnifiedGenotyper tool [26], and then, genotypes were called for the 1432 joint samples using the UnifiedGenotyper tool at the indel sites identified from the NEAVA samples. CNVs were detected using XHMM [29] and compared with getDeletions (Adrian Tan, unpublished data) as a sanity check.

Variants were annotated using TabAnno software [30]. Predicted deleteriousness for SNVs was based on the Condel score [22] by Ensembl's Variant Effect Predictor tool [31]. The variants were defined as deleterious if the Condel score of 0.50 or more; novel if not reported in both dbSNP137 and the 1000G phase I; or uncharacterized if not reported in dbSNP129, but in dbSNP137 and/or the 1000G phase I. A list of characterized ADME variants was compiled on the basis of publically accessible ADME variation databases and publications by September 2013 (Supplemental digital content 3, *http://links.lww.com/FPC/B142*).

Statistical analyses

Principal components analysis (PCA) was carried out on the SNVs with minor allele frequency (MAF) of more than 5% by PLINK v1.07 [32]. Plots (Figs S3A and S3B, Supplemental digital content 1, *http://links.lww.com/FPC/B140*) show samples cluster well at both the continental and the population level.

All Weir and Cockerham F_{ST} statistics reported were calculated using VCFtools [33,34]. Per-gene F_{ST} statistics were calculated across the full joint sample for all ADME and non-ADME genes. Between-population F_{ST} statistics were calculated for all two-way population comparisons across the 21 joint sample populations (Tables S5A–S5D, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). The Mann–Whitney test was used to determine differences in variation between ADME and non-ADME genes. Variant discovery curves were used to compare the rate at which variations were discovered in population samples of different sizes, which were plotted by number of variants per kilobase (VPK) coding sequence versus number of sampled haplotypes.

Data validation, variant call quality control, and ADME variation imputability

To evaluate NEAVA data quality, validation using different genotyping platforms, Sanger sequencing and Axiom BioBank Genotyping Array, was performed by DNA Link Inc. (Seoul, South Korea). A total 155 variants were selected for validation by Sanger sequencing, including 55 novel nonsynonymous sites (52 SNV and three Indel) in core ADME genes, 50 random singletons, and 50 random nonsingletons (Supplemental digital content 4, *http://links.lww.com/FPC/B143*). Overall, 96 NEAVA patients (48 Japanese and 48 Korean) were randomly selected and genotyped by Axiom Biobank Genotyping Array according to the manufacturer's instruction (Affymetrix).

The quality of the sequencing-based variant calls was first evaluated by comparing called genotypes to publicly available array-based genotypes. Using Affymetrix DMET Plus Array data (Affymetrix) for six of the integrated 1000G samples (Table S4, Supplemental digital content 1, *http://links.lww.com/FPC/B140*), genotype concordance rates for the sequencing-based variant calls (using genotypes from 422 sites on the DMET chip called variants in the present sample) were ~ 99.5%, after applying appropriate genotype depth and quality filters (minGD = 10). When comparing variant calls in the present study with published Illumina OMNI array genotypes at 47 067 overlapping sites, genotype concordance rates for these six 1000G patients were ~ 99.7%. Genotype concordance rates for sites in both DMET and OMNI were more than 99.9%.

A subset of the NEAVA samples ($N=96$; 48 Japanese, 48 Korean) were also randomly selected for genotyping by the Axiom Biobank Genotyping Array (Affymetrix). Nonreference genotype concordance between Exome-sequencing and Axiom array genotypes was high, with more than 99% nonreference genotype concordance at the 89 958 variant sites in the consensus coding regions also present and called polymorphic on the Axiom array. Sanger sequencing of a subset of ADME variation yielded a false-positive rate estimate of 0.7%.

The existing data from European ancestry ($N=5399$) were used for the assessment of imputability for the ADME variation (8161 SNVs identified in this study).

The imputation was performed using a cosmopolitan haplotype reference panel from the 1000 Genomes Project phase 3 on the basis of genotyping by Axiom Biobank Genotyping Array (Affymetrix) and using Hidden Markov Model methods as implemented in MaCH and minimac [35,36].

Results

Capture of ADME variation with exome-sequencing

Two sets of exome sequence data, NEAVA ($N=251$) and 1000G ($N=1181$), were integrated to jointly identify and call SNVs, short indels, and CNVs for 1432 individuals (Table S1, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). The 298 ADME genes selected for analysis were divided into 38 core genes known to influence drug biotransformation and/or disposition, and 260 extended genes with less direct evidence (Supplemental digital content 2, *http://links.lww.com/FPC/B141*; Supplemental digital content 5, *http://links.lww.com/FPC/B144*). Different exome-sequencing protocols were used across the NEAVA and 1000G studies; therefore, most downstream analyses focused primarily on a set of 'consensus' coding regions (ADME: 386 Kbp; non-ADME: 20.5 Mbp) captured at high average depth ($\geq 20\times$) in both studies (Fig. S1, Supplemental digital content 1, *http://links.lww.com/FPC/B140*).

This exome-sequencing approach captured the majority of core and extended ADME gene coding sequences at a depth of at least 20× across the joint sample (Fig. S2, Supplemental digital content 1, *http://links.lww.com/FPC/B140*; Supplemental digital content 2, *http://links.lww.com/FPC/B141*). Coverage was considerably higher for ADME than non-ADME genes (Table S2, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). Of 38 core ADME genes, three (8%; *GSTP1*, *GSTT1*, and *NAT2*) had less than 50% of the coding bases captured in the consensus regions. There were 53 of 260 extended ADME genes (20%) with less than 50% of the coding sequence captured (Supplemental digital content 2, *http://links.lww.com/FPC/B141*). In total, 87% of the core and 72% of the extended ADME, and 59.3% of non-ADME coding bases were captured in the consensus regions. Thus, despite poor coverage for some genes, most ADME coding bases were captured well across all exome-sequencing protocols used, particularly for the core genes.

High-quality SNVs and indels were generated, and CNVs detection was explored in this study (Table 1, Supplemental digital content 6, *http://links.lww.com/FPC/B145*). The quality of variant identification and calls was assessed with both Sanger sequencing and genome-wide SNP arrays. In all, 155 variants, containing singletons, nonsingletons, and novel nonsynonymous (NS) SNVs and indels, were selected randomly to confirm heterozygous calls by Sanger sequencing on NEAVA samples. Of 144 successfully sequenced sites (11 failed probe design, Supplemental digital content 4, *http://links.lww.com/FPC/B143*), 143 variants were validated, yielding a heterozygous false call rate of 0.7% (exact 95% confidence interval = 0.02–3.8%). Of the 155 variants attempted to validate by Sanger, 27 variants locate at *CYP2A*s, *CYP2C*s, *CYP2D*s, *CYP3A*s, and *UGT1A*s, which are known as highly homologous regions. Of these, 21 variants (including 11 novel SNVs) were validated by Sanger sequencing. Six variants did not fulfill the probe design criteria; for these, no wet-lab was performed. Genotype concordance for 96 NEAVA patients with Axiom BioBank array genotypes and six 1000G patients with the publically available array data was high, with at least 99.3% at the variant sites in the consensus genome-wide coding regions (Table 4S, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). The expected clustering of samples at the continental and individual population level from PCA on genotypes in the consensus genome-wide regions also indicated the high overall data quality (Fig. S3A and S3B, Supplemental digital content 1, *http://links.lww.com/FPC/B140*).

A majority of coding variants identified in this study were nonsynonymous and much of this variation is currently unreported in the ADME databases [38,46,47]. Of the 8161 ADME coding variants identified in the consensus regions, 62% were nonsynonymous and 43% of these NS variants were predicted to be functionally deleterious by Condel [24]; 21% were novel (not in dbSNP137), of which 70% were nonsynonymous and 48% were predicted to be deleterious, and 55% of all ADME NS variants were added to dbSNP since 2008 (in dbSNP137, but not in dbSNP129), which were predominantly low frequency compared with ADME coding variants present in dbSNP129, median MAF = 0.0045 (dbSNP129) versus 0.0003 (dbSNP137, but not dbSNP129), $P=2.2\times10^{-16}$.

Furthermore, a minority of the ADME variants observed in this study are currently captured by standard genotyping arrays (Table S3, Supplemental digital content 1, *http://links.lww.com/FPC/B140*), including less than 20% of the NS variants in core ADME genes included on the Affymetrix DMET Plus Array or Illumina VeraCode ADME Core Panel. This study successfully rediscovered 97.3% of the variants in the 'consensus' ADME regions, with MAF of more than 0.2%, from the NHLBI Exome Sequencing Project [48] that have been included in the commercialized genotyping arrays; however, more than 55% of the ADME NS variants identified in this study are absent from existing arrays, of which, 633 variants were predicted to be deleterious. These results highlight a key advantage of exome sequencing over genotyping arrays in characterizing putatively functional variation across ADME genes in populations of interest.

In addition, imputability for ADME variation was assessed using the existing data from European ancestry

Table 1 Summary of single nucleotide variation, insertion/deletion, and copy number calls in ADME coding sequence

	All coding	ADME coding	Core genes	Extended genes
Single nucleotide variant calls[a]				
Variants				
Nonsynonymous	219 400	5043	896	4147
Nonsynonymous (deleterious)[b]	35 281	2156	404	1752
Nonsense	5287	169	35	134
Synonymous	147 596	2949	450	2499
All SNVs	372 500	8161	1381	6780
Variants per individual				
Nonsynonymous	5613	145	26.6	116.4
Nonsynonymous (deleterious)[b]	304	24.7	4.6	20.0
Nonsense	61	< 1	< 1	< 1
Synonymous	6706	150	26.8	121.3
All SNVs	12 380	295	53.5	237.8
Short insertions/deletions[c]				
Variants				
Insertions	481	8	2	6
Deletions	209	11	1	10
All insertions/deletions	690	19	3	16
Variants per individual				
Insertions	52	1.97	0.11	1.04
Deletions	55	0.51	0.94	0.40
All insertions/deletions	107	2.48	1.05	1.44

	Common name	Deletion/ duplication	Called[e]	OnAffyDMET	inDGV	References
Copy number variation[d]						
Genes						
CES1	–	Deletion	Y	N	Y	Ulloa *et al.* [37]
	–	Duplication	Y	N	Y	
CYP2A6	CYP2A6*4	Deletion	Y	Y	Y	The Human Cytochrome P450 (CYP) Allele Nomenclature Database [38] Martis *et al.* [39]
	CYP2A6*1 × 2	Duplication	Y	Y	Y	
CYP2B6	CYP2B6*29 or *30	Deletion	Y	N	Y	
CYP2D6	CYP2D6*5_gene deletion	Deletion	Y	Y	Y	The Human Cytochrome P450 (CYP) Allele Nomenclature Database [38] Gaedigk *et al.* [40]
	CYP2D6*X × 2	Duplication	Y	Y	Y	
CYP2E1	CYP2E1*1C × 2	Duplication	Y	N	Y	Martis *et al.* [39]
CYP21A2	CYP21A2*7	Duplication	Y	N	Y	The Human Cytochrome P450 (CYP) Allele Nomenclature Database [38]
CYP4A11	–	Deletion	Y	N	Y	
	–	Duplication	Y	N	Y	
DHRSX	–	Deletion	Y	N	Y	
	–	Duplication	Y	N	Y	
GSTM1	GSTM1*0 Null_gene deletion	Deletion	Y	Y	Y	Marenne *et al.* [41] Xu *et al.* [42] McLellan *et al.* [43]
	GSTM1*X2_gene duplication	Duplication	Y	N	Y	
SULT1A1	SULT1A1 CNV	Duplication	Y	N		Gaedigk *et al.* [40] Hebbring *et al.* [44]
UGT2B17	UGT2B17*2_deletion of gene	Deletion	Y	Y	Y	Gaedigk *et al.* [40] Ménard *et al.* [45]

CNV, copy number variant; N, no, Y, yes.

[a] Totals include all high-quality variants falling in consensus capture regions (see Materials and Methods section).

[b] Deleterious defined as Condel score ≥ 0.47.

[c] Includes all high-quality coding insertion/deletion variants (< 50 bp) in consensus capture regions (see Materials and Methods section).

[d] GATK UnifiedGenotyper 'Best Practices' filters applied (see Materials and Methods section).

[e] All ADME CNVs present in the relaxed-threshold CNV callset.

($N = 5399$). Out of the 8161 ADME SNVs identified from this study, 4454 SNVs (54.5%) were imputed, of which, 2214 SNVs (27.1%) attained the imputation score of 0.3 or more as the conventional cut-off for analyses. The imputability distribution on the basis of the MAF is summarized (Table S7, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). As expected, at least 94% of the ADME SNVs with MAF of at least 5% can be imputed at the imputation score of 0.3 or more threshold.

In consensus ADME regions, this analysis rediscovered 99% of the variants described in the original 1000G analysis. For the rest, 96% of variants were detected, for example, the *CYP2D6*4* (rs3892097) and *CYP3A5*3* (rs776746) variants were observed in many patients with less than 20× average depth, but most fell outside of consensus regions.

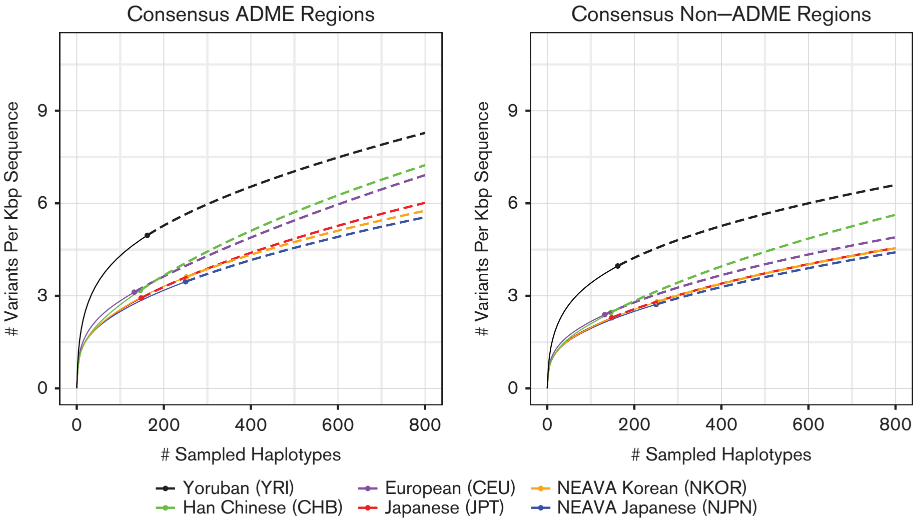
Patterns of ADME variation

We observed considerable heterogeneity in the total abundance of variation across the studied populations, and in ADME genes, these differences primarily appear to be driven by NS variation. To control for different sample sizes between populations, we compared the rate of variants among populations with variant discovery curves (Fig. 1; Fig. S4, Supplemental digital content 1, *http://links.lww.com/FPC/B140*) [49]. With respect to the abundance of variants for ADME and non-ADME genes, the order for populations was broadly in line with expectations on the basis of their demographic histories. The African populations showed the highest levels of variation, followed by Americans, Asians, and Europeans. For six selected populations (YRI, CHB, CEU, JPT, NKOR, and NJPN), the nonsynonymous and synonymous variation in ADME and non-ADME genes showed similar trends (Fig. S5, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). Among Northeast Asian populations, the patterns of predicted variation were consistent with ancestral bottleneck events in migration to these isolated territories. Across the global populations, the rates of variation in the ADME genes were predicted to be 25% higher than that in non-ADME genes, after adjusting for differences in sequencing coverage. These differences were driven by the rates of NS variants, where we observed 50% more VPK in ADME compared with non-ADME genes, Mann–Whitney $P=8.2\times10^{-13}$, (Fig. 2; Fig. S5, Supplemental digital content 1, *http://links.lww.com/FPC/B140*).

The per individual carriage rates of ADME NS variants varied considerably among populations (Fig. S6, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). Overall, Northeast Asian individuals carried fewer ADME NS variants (~139 NS SNVs per person) than other populations (European = ~ 144; American = ~ 145; South Asian = ~ 145; African = ~ 176). On average, each Northeast Asian individual carried ~ 10 NS ADME variants currently unreported in the ADME variation databases (Supplemental digital content 3, *http://links.lww.com/FPC/B142*), of which 25% are predicted to be deleterious. Quantities of uncharacterized (not in dbSNP129) deleterious ADME variants were similar in Northeast Asian, American, and European populations (2.7, 2.8, and 2.3 per individual, respectively), but higher for the African patients (6.0/person). Overall, Northeast Asian patients carried on average two frameshift indels in ADME coding regions, slightly greater than the European or African patients (1.6 and 1.7, respectively).

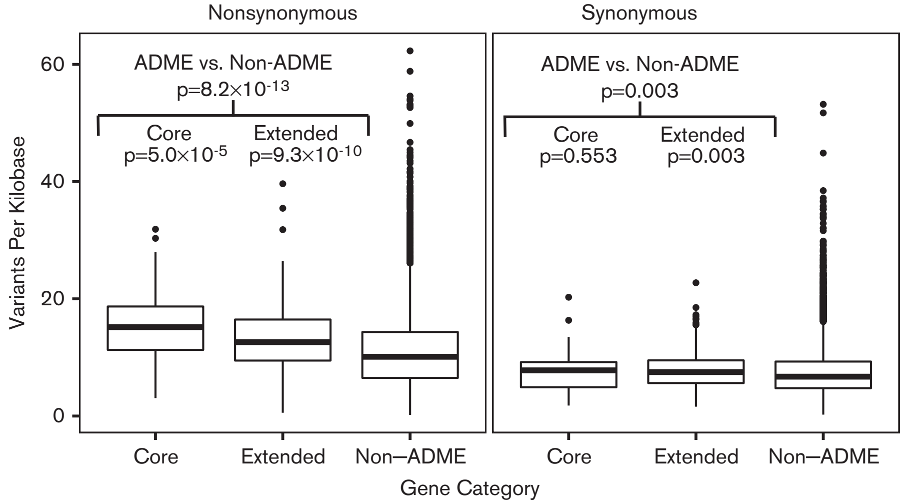
Knockout (nonsense or essential splice site variant) alleles leading to truncated or nonfunctional proteins are particularly important for drug ADME. In this study, ~ 49% of patients carried at least one putative 'knockout' variant, with 3% of all samples carrying at least one uncharacterized knockout allele. The carriage of one or more knockout alleles was the highest in the African population at 73% and the lowest in Northeast Asians at

Fig. 1



Coding variation in ADME versus non-ADME genes. Variant discovery curves for four continental ancestral populations (African, American, Northeast/East Asian , and European; see Table S1, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). Variant discovery curves were used to compare the rate at which variation was discovered in population samples of different sizes. Predicted numbers of variants per kilobase coding sequence are plotted on the *y*-axis; numbers of sampled haplotypes are plotted on the *x*-axis. The solid, dotted, and dashed lines represent the hypergeometric expectation, observed number of variants, and jackknife projections, respectively, for each continental population.

Fig. 2



Differences in variation between ADME (core + extended) and non-ADME genes driven by nonsynonymous variation. Box and whisker plots showing the distribution of gene-level variants per kilobase (VPK) coding sequence across the core ($N = 38$) and the extended ($N = 260$) ADME genes, and non-ADME ($N = 15\,124$) genes in the consensus capture sequence (see Fig. S1, Supplemental digital content 1, *http://links.lww.com/FPC/B140* and Supplemental digital content 2, *http://links.lww.com/FPC/B141*). *P*-values correspond to two-sample Wilcoxon (Mann–Whitney) test results. Upper and lower box hinges correspond to 25th and 75th percentile; whiskers extend from hinges to the highest and lowest values within 1.5 times the interquartile range (IQR; distance between the 25th and the 75th percentile). Points beyond the whiskers represent outliers (e.g. VPK values greater/less than 1.5 × IQR).

40% (Fig. S6C, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). Although a greater proportion of European samples (56%) carried one or more knockout alleles than Northeast Asians, 3.6% of Northeast Asians carry at least one uncharacterized 'knockout' variant, compared with less than 1% of Europeans, reflecting the population bias in the databases of known variation.

A majority of the ADME NS variants identified appeared to be private to a single population or continent. In this analysis, 54% of the ADME NS variants were singletons. In total, over 60% of the NS variants observed were private to a single population, with 33% absent in dbSNP137 (Fig. S7, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). Of the private ADME NS variants, 79% were rare (< 1% MAF) and 46% were predicted to be deleterious.

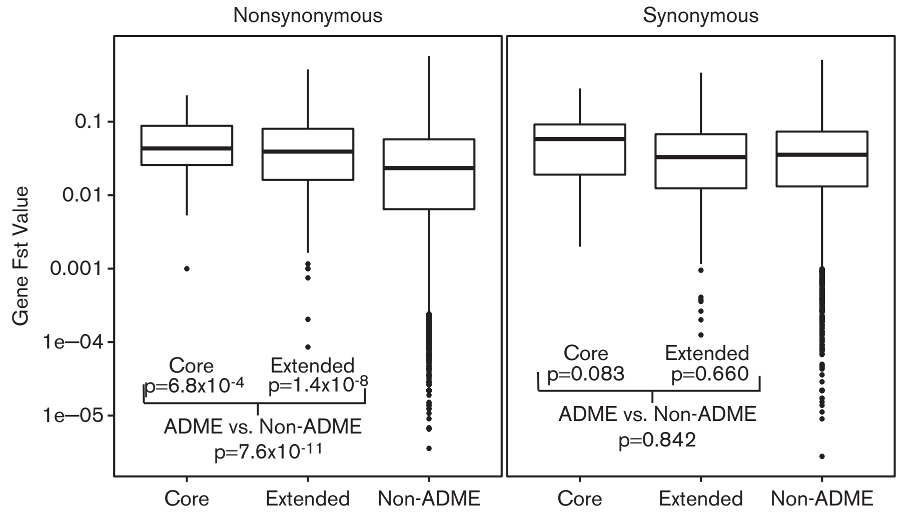
As expected, ADME variation across populations was shared on the basis of geographic location. Using joint site frequency spectrum analyses that included all two-way population combinations [47], it appeared that more variants were shared between the geographically closer populations than distant ones, especially for variants with low MAF (Fig. S8, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). The Northeast Asian populations have low allele sharing with individuals of African (1000G YRI) and European (1000G CEU) ancestry, but comparatively little differentiation among Northeast Asian populations.

Although previous array-based ADME research has suggested that ADME genes may show nominally higher amounts of genetic differentiation across populations than non-ADME genes [14], this sequence-based approach enabled us to examine this hypothesis more completely. Gene-level F_{st} analyses were carried out to quantify genetic differentiation across the 21 populations in this study (Table S5, Supplemental digital content 1, *http://links.lww.com/FPC/B140*). We observed statistically significantly higher F_{ST} values in ADME compared with non-ADME genes using NS variation (Mann–Whitney P-value $= 7.6 \times 10^{-11}$), but no significant difference using synonymous variation (Fig. 3), supporting the inference that selection of functional ADME alleles has influenced the population differentiation.

Quantification of genetic variation in core ADME genes

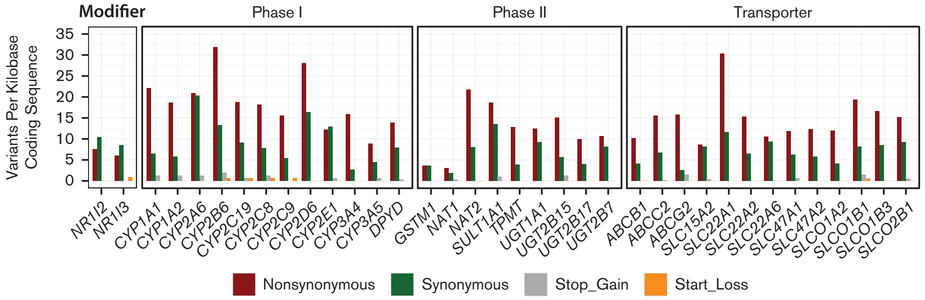
Core ADME genes [20] are commonly involved in drug ADME, and genetic variants in these genes could be relevant to many pharmaceutical clinical studies. (Supplemental digital content 7, *http://links.lww.com/FPC/B147*). We observed significantly higher variant density in core than extended ADME genes (Mann–Whitney $P = 5.0 \times 10^{-5}$, Fig. 2). There was considerable heterogeneity in sequence variation among core ADME genes, with densities of variation in consensus sequence ranging from five to 47 VPK across the 36 genes analyzed (Fig. 4). The distributions of NS variants on the basis of the continental populations were summarized for the core

Fig. 3



Variability in ADME variation across population. Box and whisker plots showing the distribution of F_{ST} values across the core ($N=38$) and the extended ($N=260$) ADME genes, and non-ADME ($N=15\,124$) genes calculated from variants in the consensus capture sequence (see Fig. S1, Supplemental digital content 1, *http://links.lww.com/FPC/B140* and Supplemental digital content 2, *http://links.lww.com/FPC/B141*). F_{ST} values represent Weir and Cockerham F_{ST} statistics as calculated using VCFtools [33,34]. Higher F_{ST} values for a given gene imply allele frequency differences across populations. *P*-values correspond to two-sample Wilcoxon (Mann–Whitney) test results. Upper and lower box hinges correspond to the 25th and 75th percentile; whiskers extend from hinges to the highest and lowest values within 1.5 times the interquartile range (IQR; distance between 25th and 75th percentile). Points beyond whiskers represent outliers (e.g. F_{ST} values greater/less than 1.5 × IQR).

Fig. 4



Heterogeneity in variants per kilobase (VPK) (core ADME genes). VPK coding sequence for 36 selected core ADME genes analyzed, stratified by variant type and gene category. Gene categories assigned per pharmADME.org categorizations [20]. Gene and variant type assigned as annotated by TabAnno [30].

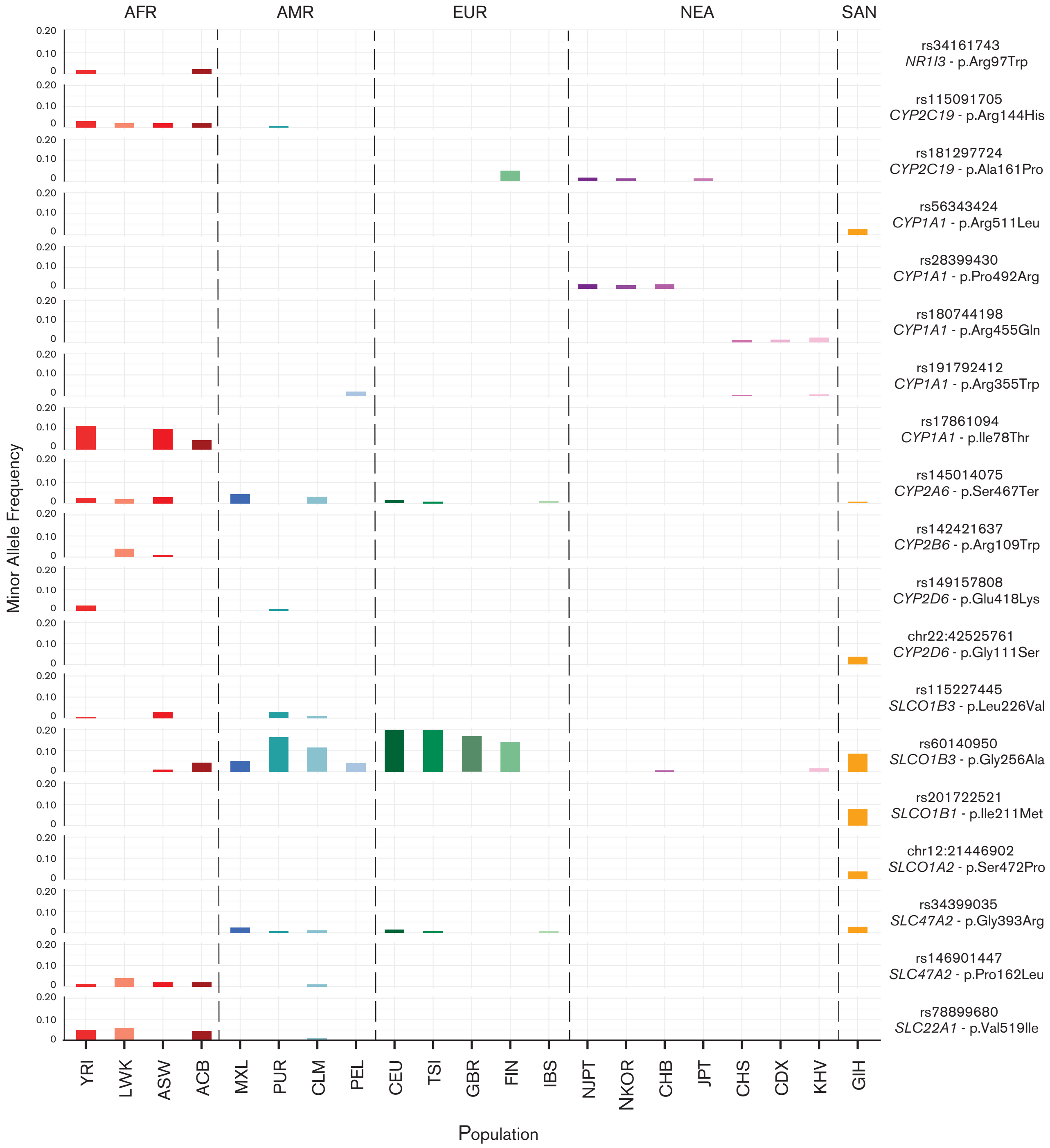
ADME genes (Table S6, Supplemental digital content 1, *http://links.lww.com/FPC/B140*).

While assessing whether all functionally relevant variants, common in geographically and ethnically diverse clinical studies, were captured by current genotyping arrays, we identified 40 variants in core ADME genes that fulfilled the criteria of (i) relatively common (≥2% MAF) in at least one of the 21 populations analyzed, (ii) predicted to be functionally deleterious (Condel score ≥0.47, stop gain, or start loss), and (iii) not included on the Affymetrix DMET Plus Array that is commonly used for ADME PGx studies (Supplemental digital content 8, *http://links.lww.com/FPC/B146*). Of the 40 variants, 13 were known ADME variants and the functional impact of 10 of the 13 variants was confirmed *in vitro* or *in vivo*. Nineteen of the 40 variants are shown in Fig. 5. The functional effects of these variants deserve further investigation and consideration for inclusion in ADME genotyping arrays and relevant pharmacokinetic studies.

Discussion

Previous analyses of ADME genetic variation have been limited to variants identified mostly from patients of northern European ancestry. The falling cost of NGS sequencing and the availability of exome-sequencing

Fig. 5



Common, potentially functional variants in core genes not captured by ADME genotyping arrays. Population allele frequencies for 19 potentially functional variants in core ADME genes that are not included in dbSNP and not captured by ADME genotyping arrays. Supplemental digital content 8 (*http://links.lww.com/FPC/B146*) contains a complete list of 40 variants proposed for consideration in future ADME genetic studies. Each bar represents the variant minor allele frequency in the corresponding population (*y*-axis scale: 0–20%). RS numbers are from dbSNP137; gene and amino acid change as annotated by TabAnno [30]. Individual population abbreviations are defined in Table S1 (Supplemental digital content 1, *http://links.lww.com/FPC/B140*). RS number, reference single nucleotide polymorphism cluster id.

data from reference samples from diverse ethnic populations provide an opportunity to expand our understanding of population-specific ADME variation. In this study, we analyzed ADME genetic variation in 21 global populations using commoditized exome capture of ADME genes in 1432 samples from Northeast Asians and 1000G Project, enriched for individuals of East Asian ancestry ($N=631$).

We evaluated the quality of ADME SNVs and indels and the possibility of CNV detection for ADME genes using exome sequencing data, and compared these variants with a compiled list of characterized ADME variants and variants contained on current ADME genotyping arrays. Our analyses characterized the allele frequency and predicted functionality of ADME variation in 21 populations; many variants were novel or uncharacterized, and therefore of potential interest to researchers carrying out clinical studies of ADME genes. We investigated patterns of variation in global populations and contrasted the variation in ADME and non-ADME genes. Our comprehensive characterization of ADME coding variation provides insight into potential evolutionary forces acting on these genes, and details the ADME variation of potential clinical relevance across a diverse range of populations.

We showed that the exome sequencing of ADME genes was modestly better than the exome as a whole by using commoditized NGS tools. The high quality of variant calls was confirmed by alternative genotyping/sequencing methods, PCA, and variant rediscovery. The exome-sequencing captured well and produced high quality variant calls for most ADME genes, and coverage for poorly captured genes should improve as capture technology evolves.

The comparison of variants between the current ADME genotyping panels and this study indicates the European sample bias present in the panels and highlights important candidates for future ADME genotyping panels and further functional exploration. Particularly, we identified 40 potentially functional variants (MAF $\geq 2\%$) in core ADME genes common in at least one of the populations analyzed that, to the best of our knowledge, are not included in pharmacogenetic studies profiling those core ADME genes. In addition, we show that many novel and uncharacterized population-specific ADME variants are likely deleterious. With 25% more variants per kilobase in ADME genes compared with non-ADME for the populations studied, it appears that ADME genes are under less selection than the rest of the genome, which is consistent with the results reported by Li *et al.* [14]. Our results indicate that this differentiation is driven by NS variation between ADME and non-ADME genes. Possible explanations for the higher rates of NS variation in ADME genes include weaker negative (purifying) selection, stronger or more frequent positive balancing selection, or differences in population-specific selection pressures. Although Li *et al.* highlighted greater than expected diversity in ADME genes across populations, we further illustrated the differences in population-specific variation observed across ADME genes. F_{ST} analyses in consensus regions show diversity between populations across both ADME and all exonic sequences. Analyses of variation within each ADME gene suggest that some of the genes may be under different selection pressures than others.

The SNV calling methods used in this study have been shown to result in low false-positive and false-negative rates. In contrast, the indel and CNV calling algorithms are less optimal for NGS short-read data, likely resulting in lower quality calls, an important consideration, given the known effects for ADME CNVs ranging from loss of function (e.g. *CYP2D6*5*) to increased activity (e.g. *CYP2D6*1* or **2xN*). Furthermore, given our initial focus on variation in Northeast Asians, the indel calling was performed at sites discovered using only NEAVA individuals, which might have resulted in an undercalling of non-Northeast Asian indels overall. Consequently, it would be desirable to improve the NGS-based CNV and indel calling algorithms to effectively and comprehensively study ADME variations and accurately report allele frequency for the variants located in the ADME genes with CNVs that occurred by a single platform.

In general, African populations showed the highest levels of genetic variation in both ADME and non-ADME genes, followed by Americans, Europeans, South Asians, and East Asians. Among East Asian populations, the patterns of predicted variation were consistent with ancestral bottleneck events in migration to these isolated territories. Although we observed more NS variants within ADME genes and greater differentiation in those variants among populations compared with non-ADME genes, these differences were modest. Populations in relatively close geographic proximity generally shared most ADME variants at similar frequencies. The analyses presented further support the extrapolation of results from pharmacokinetic studies carried out among historically and culturally related populations. In particular, our results suggest that ADME-related findings in any East Asian population may be of relevance to other East Asian populations. We therefore believe that data from this study may be useful in future pharmacokinetic studies for evaluating the potential impact of frequency differences of putatively functional ADME variants among populations.

In summary, this sequence-based analysis systematically and comprehensively characterized ADME coding variation in multiple populations, showed the potential utility and value of NGS for studying ADME variations, assessed the completeness of current ADME genotyping panels, and indicated that ADME genes have significantly more variants and are more variable among populations than non-ADME genes. Our comprehensive summary of ADME variation in diverse populations provides insights relevant for interpretation and generalization of association results found in these genes. Furthermore, as a high-throughput platform to study the ADME genes comprehensively, NGS showed an incontestable advantage for rare and/or novel variants identification, which may be useful for studying pharmacokinetic outliers or supporting safety-related case studies during drug development.

Acknowledgements

The authors thank Mike Lawson for his assistance integrating data from several online databases. We also wish to thank those who contributed genetic samples and made this research possible.

Conflicts of interest

This study was supported by GlaxoSmithKline (GSK). The authors, Zhengyu Xue, Elizabeth C. Harris, and Sophie L. Stocker, were previously employed by GSK and hold GSK stocks, and the research for this publication was performed while they were employed by GSK; the authors, Margaret G. Ehm, Annette S. Gross, Lon R. Cardon, Stephanie L. Chissoe, and Matthew R. Nelson, have an employment or leadership position with GlaxoSmithKline and hold GSK stocks. Gonçalo Abecasis received GSK research grants. For the remaining authors there are no conflicts of interest.

References

1 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**:265–269.

2 Ma MK, Woo MH, McLeod HL. Genetic basis of drug metabolism. *Am J Health Syst Pharm* 2002; **59**:2061–2069.

3 Stingl J, Viviani R. Polymorphism in CYP2D6 and CYP2C19, members of the cytochrome P450 mixed-function oxidase system, in the metabolism of psychotropic drugs. *J Intern Med* 2015; **277**:167–177.

4 Lee H, Hu M, Ho C, Wong C, Tomlinson B. Effects of polymorphisms in ABCG2, SLCO1B1, SLC10A1, and CYP2C9/19 on plasma concentrations of rosuvastatin and lipid response in Chinese patients. *Pharmacogenomics* 2013; **14**:1283–1294.

5 Kurose K, Sugiyama E, Saito Y. Population differences in major functional polymorphisms of pharmacokinetics/pharmacodynamics-related genes in Eastern Asians and Europeans: implications in the clinical trials for novel drug development. *Drug Metab Pharmacokinet* 2012; **27**:9–54.

6 Liao JK. Safety and efficacy of statins in Asians. *Am J Cardiol* 2007; **99**:410–414.

7 Birmingham BK, Bujac SR, Elsby R, Azumaya CT, Wei C, Chen Y, *et al.* Impact of ABCG2 and SLCO1B1 polymorphisms on pharmacokinetics of rosuvastatin, atorvastatin and simvastatin acid in Caucasian and Asian subjects: a class effect? *Eur J Clin Pharmacol* 2015; **71**:341–355.

8 Myrand SP, Sekiguchi K, Man MZ, Lin X, Tzeng RY, Teng CH, *et al.* Pharmacokinetics/genotype associations for major cytochrome P450 enzymes in native and first- and third-generation Japanese populations: comparison with Korean, Chinese, and Caucasian populations. *Clin Pharmacol Ther* 2008; **84**:347–361.

9 Inoue S, Howgate EM, Rowland-Yeo K, Shimada T, Yamazaki H, Tucker GT, *et al.* Prediction of in vivo drug clearance from in vitro data. II: potential inter-ethnic differences. *Xenobiotica* 2006; **36**:499–513.

10 Man M, Farmen M, Dumaual C, Teng CH, Moser B, Irie S, *et al.* Genetic variation in metabolizing enzyme and transporter genes: comprehensive assessment in 3 major East Asian subpopulations with comparison to Caucasians and Africans. *J Clin Pharmacol* 2006; **50**:929–940.

11 Sistonen J, Sajantila A, Lao O, Corander J, Barbujani G, Fuselli S. CYP2D6 worldwide genetic variation shows high frequency of altered activity variants and no continental structure. *Pharmacogenet Genomics* 2007; **17**:93–101.

12 Van der Weide J, Steijns LS. Cytochrome P450 enzyme system: genetic polymorphisms and impact on clinical pharmacology. *Ann Clin Biochem* 1999; **36**:722–729.

13 Gordon AS, Tabor HK, Johnson AD, Snively BM, Assimes TL, Auer PL, *et al.* Quantifying rare, deleterious variation in 12 human cytochrome P450 drug-metabolism genes in a large-scale exome dataset. *Hum Mol Genet* 2014; **23**:1957–1963.

14 Li J, Zhang L, Zhou H, Stoneking M, Tang K. Global patterns of genetic diversity and signals of natural selection for human ADME genes. *Hum Mol Genet* 2011; **20**:528–540.

15 Standford Human Genome Center. Human Genome Diversity Project. Available at: *http://www.hagsc.org/hgdp/*. [Accessed September 2013].

16 NCBI retiring HapMap Resource. International HapMap Project. Available at: *http://hapmap.ncbi.nlm.nih.gov/*. [Accessed September 2013].

17 Ramos E, Doumatey A, Elkahloun AG, Shriner D, Huang H, Chen G, *et al.* Pharmacogenomics, ancestry and clinical decision making for global populations. *Pharmacogenomics* 2013; **14**:217–222.

18 Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 2005; **15**:1496–1502.

19 Nelson MR, Wegmann D, Ehm MG, Kessner D St, Jean P, Verzilli C, *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14 002 people. *Science* 2012; **337**:100–104.

20 PharmADME. ADME gene list. Available at: *http://pharmaadme.org/joomla/index.php?option = com_content&task = view&id = 12&Itemid =27*. [Accessed September 2013].

21 1000 Genomoes (a deep catalog of human genetic variation). IGSR and the 1000 Genomes Project. Available at: *http://www.1000genomes.org/*. [Accessed September 2013].

22 Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, *et al.* An integrated map of genetic variation from 1092 human genomes. *Nature* 2012; **491**:56–65.

23 Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC known genes. *Bioinformatics* 2006; **22**:1036–1046.

24 González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 2013; **88**:440–449.

25 Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009; **25**:1754–1760.

26 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.

27 Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* 2015; **25**:918–925.

28 Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. *Genome Res* 2011; **21**:961–973.

29 Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 2012; **91**:597–607.

30 Zhan X, Liu D. TaSer (TabAnno and SeqMiner): a toolset for annotating and querying next-generation sequence data. arXiv:1306.5715; 2013.

31 McLaren W, Pritchard B, Rios D, Chen Y. Deriving the consequences of genomic variants with the ensembl API and SNP effect predictor. *Bioinformatics* 2010; **26**:2069–2070.

32 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**:559–575.

33 Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution* 1984; **38**:1358–1370.

34 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* The variant call format and VCFtools. *Bioinformatics* 2011; **27**:2156–2158.

35 Li Y, Dong M, Hua J. Simultaneous localized feature selection and model detection for gaussian mixtures. *IEEE Trans Pattern Anal Mach Intell* 2009; **31**:953–960.

36 Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012; **44**:955–959.

37 Ulloa AE, Chen J, Vergara VM, Calhoun V, Liu J. Association between copy number variation losses and alcohol dependence across African American

and European American ethnic groups. *Alcohol Clin Exp Res* 2014; **38**:1266–1274.

38 The Human Cytochrome P450 (CYP) Allele Nomenclature Database. Allele nomenclature for Cytochrome P450 enzymes. Available at: *http://www.cypalleles.ki.se/*. [Accessed September 2013].

39 Martis S, Mei H, Vijzelaar R, Edelmann L, Desnick RJ, Scott SA. Multi-ethnic cytochrome-P450 copy number profiling: novel pharmacogenetic alleles and mechanism of copy number variation formation. *Pharmacogenomics J* 2013; **13**:558–566.

40 Gaedigk A, Twist GP, Leeder JS. CYP2D6, SULT1A1 and UGT2B17 copy number variation: quantitative detection by multiplex PCR. *Pharmacogenomics* 2012; **13**:91–111.

41 Marenne G, Real FX, Rothman N, Rodríguez-Santiago B, Pérez-Jurado L, Kogevinas M, *et al*. Genome-wide CNV analysis replicates the association between GSTM1 deletion and bladder cancer: a support for using continuous measurement from SNP-array data. *BMC Genomics* 2012; **13**:326.

42 Xu S, Wang Y, Roe B, Pearson WR. Characterization of the human class Mu glutathione S-transferase gene cluster and the GSTM1 deletion. *J Biol Chem* 1998; **273**:3517–3527.

43 McLellan RA, Oscarson M, Alexandrie AK, Seidegård J, Evans DA, Rannug A, *et al*. Characterization of a human glutathione S-transferase mu cluster containing a duplicated GSTM1 gene that causes ultrarapid enzyme activity. *Mol Pharmacol* 1997; **52**:958–965.

44 Hebbring SJ, Moyer AM, Weinshilboum RM. Sulfotransferase gene copy number variation: pharmacogenetics and function. *Cytogenet Genome Res* 2008; **123**:205–210.

45 Ménard V, Eap O, Harvey M, Guillemette C, Lévesque E. Copy-number variations (CNVs) of the human sex steroid metabolizing genes UGT2B17 and UGT2B28 and their associations with a UGT2B15 functional polymorphism. *Hum Mutat* 2009; **30**:1310–1319.

46 Department of Molecular Biology and Genetics, Democritus University of Thrace. The database of arylamine N-acetyltransferases (NATs). Available at: *http://nat.mbg.duth.gr/*. [Accessed September 2013].

47 Pharmacogenomics Laboratory. UGT-glucuronosyltransferase alleles nomenclature. Available at: *https://www.pharmacogenomics.pha.ulaval.ca/ugt-alleles-nomenclature/*. [Accessed September 2013].

48 NHLBI Exome Sequencing Project (ESP). Available at: *http://evs.gs.washington.edu/EVS/*. [Accessed September 2013].

49 Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, *et al.* Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA* 2011; **108**:11983–11988.