Human Molecular Genetics, 2014, Vol. 23, No. 8 **1957–1963**
doi:10.1093/hmg/ddt588
Advance Access published on November 26, 2013

Quantifying rare, deleterious variation in 12 human cytochrome P450 drug-metabolism genes in a large-scale exome dataset

Adam S. Gordon[1], Holly K. Tabor[2,18], Andrew D. Johnson[6], Beverly M. Snively[7], Themistocles L. Assimes[8], Paul L. Auer[9], John P.A. Ioannidis[8], Ulrike Peters[9], Jennifer G. Robinson[10], Lara E. Sucheston[11], Danxin Wang[12], Nona Sotoodehnia[3,5], Jerome I. Rotter[14], Bruce M. Psaty[3,17], Rebecca D. Jackson[13], David M. Herrington[15], Christopher J. O'Donnell[6], Alexander P. Reiner[4,9], Stephen S. Rich[16], Mark J. Rieder[1], Michael J. Bamshad[1,2] and Deborah A. Nickerson[1,*], On Behalf of the NHLBI GO Exome Sequencing Project[†]

[1]Department of Genome Sciences, [2]Department of Pediatrics, [3]Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology and Health Services, [4]Department of Epidemiology and [5]Division of Cardiology, Department of Medicine, University of Washington, Seattle, WA, USA, [6]NHLBI Division of Intramural Research and NHLBI's Framingham Heart Study, Framingham, MA, USA, [7]Department of Biostatistical Sciences, Division of Public Health Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA, [8]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA, [9]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [10]Departments of Epidemiology & Medicine, University of Iowa College of Public Health, Iowa City, IA, USA, [11]Department of Cancer Prevention & Control, Roswell Park Cancer Institute, Buffalo, NY, USA, [12]Department of Pharmacology, College of Medicine, [13]Division of Endocrinology, Diabetes and Metabolism and Center for Clinical and Translational Science, The Ohio State University, Columbus, OH, USA, [14]Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute and Department of Pediatrics, Harbor-UCLA Medical Center, Torrance, CA, USA, [15]Department of Cardiology, Wake Forest University, Winston-Salem, NC, USA, [16]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA, [17]Group Health Research Institute, Group Health Cooperative, Seattle, WA, USA and [18]Treuman Katz Center for Pediatric Bioethics, Seattle Children's Research Institute, Seattle, WA, USA

Received November 11, 2013; Revised and Accepted November 15, 2013

The study of genetic influences on drug response and efficacy ('pharmacogenetics') has existed for over 50 years. Yet, we still lack a complete picture of how genetic variation, both common and rare, affects each individual's responses to medications. Exome sequencing is a promising alternative method for pharmacogenetic discovery as it provides information on both common and rare variation in large numbers of individuals. Using exome data from 2203 AA and 4300 Caucasian individuals through the NHLBI Exome Sequencing Project, we conducted a survey of coding variation within 12 *Cytochrome P450* (*CYP*) genes that are collectively responsible for catalyzing nearly 75% of all known Phase I drug oxidation reactions. In addition to identifying many polymorphisms with known pharmacogenetic effects, we discovered over 730 novel nonsynonymous alleles across the 12 *CYP* genes of interest. These alleles include many with diverse functional effects such as premature stop codons, aberrant splicesites and mutations at conserved active site residues. Our analysis considering both novel, predicted functional alleles as well as known, actionable *CYP* alleles reveals that rare, deleterious variation contributes markedly to the overall burden of pharmacogenetic alleles within the populations considered, and that the contribution of rare variation to this burden is over three times greater in AA individuals as compared with Caucasians. While most of these impactful alleles are individually rare, 7.6–11.7% of individuals interrogated in the study carry at least one newly described potentially deleterious alleles in a major drug-metabolizing *CYP*.

[*]To whom correspondence should be addressed. Tel: +1 2066857387; Fax: +1 2062216498; Email: debnick@uw.edu
[†]Full ESP author list located in the Supplementary Material.

INTRODUCTION

Genetic influences on drug action ('pharmacogenetics') have been studied directly for several decades, yet we still lack a comprehensive understanding of how genetic variation, both common and rare, affects an individual's responses to medications (1). Exome sequencing provides a promising new approach for accelerating pharmacogenetic discovery because it assesses both common (i.e. minor allele frequency (MAF) $>5\%$) and rare (MAF $< 1\%$) variation in virtually all genes in an individual at relatively low cost. To this end, exome sequencing can simultaneously capture variation across many genes with diverse roles in pharmacological pathways; these include the 'pharmacokinetic' proteins that catalyze drug-metabolism reactions, the proteins that influence drug absorption and excretion, and the 'pharmacodynamic' proteins that are the targets for drug action. A recent resequencing study of 202 drug-target genes in over 14 000 individuals revealed an excess of rare coding variation, 90% of that had not been previously identified. Ninety-five percent of variants had an MAF $< 0.5\%$, and ~75% of variants were present in only one or two individuals (2). The current study suggests that rare variation in drug-metabolism genes is also extensive, and this variation may explain, in part, the observed variation in overall drug response. While some drug-metabolism genes were included in the previous study, no members of the Cytochrome P450 family were included, despite their well-established connection to variation in drug efficacy and toxicity (3). In this report, we report on the variation and predicted functional implication of rare coding variants on a set of 12 Cytochrome P450 (*CYP*) genes.

The *CYP* genes are of particular interest because they catalyze oxidation reactions on a wide variety of drugs. While the human genome contains 57 *CYP* genes (4), a subset of just 12 genes (designated as *CYP*-12) are collectively responsible for ~75% of all known drug oxidation reactions (3) and most of these are already known to influence clinically important phenotypes such as the efficacy of clopidogrel and the maintenance dosing of warfarin (5,6). For example, *CYP2C9* encodes the enzyme that catalyzes the oxidation of warfarin. Two *CYP2C9* missense variants impair protein function such that individuals heterozygous for either variant require a lower dose of warfarin to achieve the same steady-state concentrations (7).

RESULTS

Using large-scale exome sequencing data generated by the NHLBI Exome Sequencing Project (ESP), we identified and characterized variation within the *CYP*-12 to define the full spectrum of variation (i.e. rare and common variants) that potentially shapes inter-individual differences in drug response. Specifically, we analyzed exome sequence data from 6503 individuals of AA ($n = 2203$) and European-American (EA; $n = 4300$) ancestry (8). This dataset is available to the public through the ESP Exome Variant Server (http://eversusgs.washington.edu/EVS/).

Across the CYP-12, 98.1% of coding sequence was covered with an average depth of 30× or greater. We discovered a total of 1006 unique variants in the *CYP*-12. This included 275 (27.3%) known and 731 (72.7%) novel variants compared with dbSNP (build 132, http://www.ncbi.nlm.nih.gov/projects/SNP/) of which 486 were missense variants and 42 were nonsense/splicesite variants or frameshifting insertion/deletion (indel) variants (Fig. 1). Compared with all other genes across the exome, the *CYP*-12 do not appear to have exceptionally more or less missense variation than most genes. For example, among the CYP-12, *CYP2A6*, *CYP2B6* and *CYP2D6* contain the most nonsynonymous variation and are the only *CYP*-12 genes in the top 20% of genes assessed. Additionally, the CYP-12 appear to be a representative sample of the nonsynonymous variation observed in among the 57 human Cytochrome P450s this gene family, spanning the range of diversity in these genes while containing neither the most nor the least diverse human CYPs (Supplementary Material, Fig. S2). We estimated the MAF of each *CYP* variant in EA and AA separately and the site frequency spectrum of known and novel *CYP* alleles (Fig. 2). Overall, the majority of variation in these genes are exceedingly rare in both AA and EA. Indeed, 474 (64.8%) of novel variants (177 in AA and 297 in EA) were found on only a single chromosome and only one novel variant had an MAF $> 2\%$. In addition to this novel variation, we identified

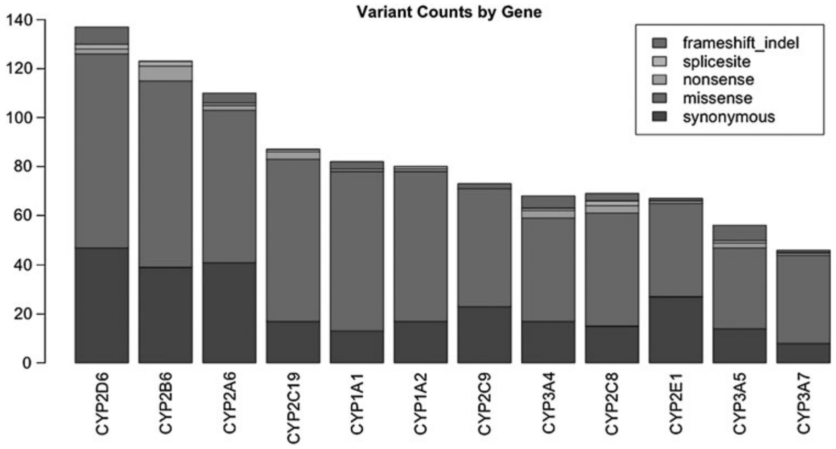

Figure 1. Distribution of exonic variation across the 12 drug-metabolizing *CYP* genes separated by variant consequence. Variant types (missense, nonsense, synonymous, splicesite, frameshift) were determined using SeattleSeq annotation. For genes that produce more than one known transcript (*CYP2D6*, *CYP2C8*, *CYP3A4*), annotation was based on the primary transcript.

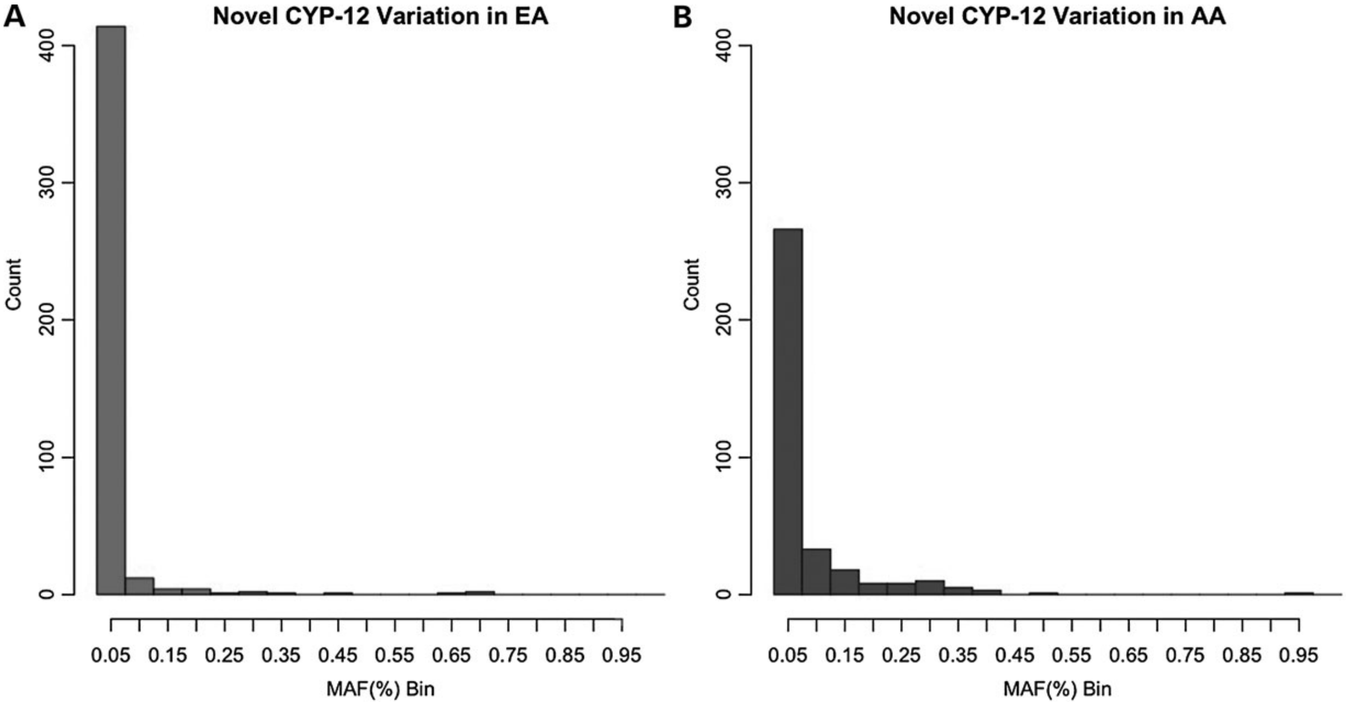

Figure 2. MAF for novel and known variants across the *CYP*-12 in EAs and AAs.

many known functional exonic variants across the *CYP*-12, including clinically relevant alleles such as CYP2C9*2, CYP2B6*6 and CYP2D6*4. Table 1 provides MAFs in both EA and AA for these and other functional variants, many of which have not been genotyped in a cohort as large as the ESP to date. However, while virtually all of the common exonic variants in the *CYP*-12 in AA and EA have been identified, exome sequencing revealed that most of the variants that are predicted to be functional are rare and yet to be discovered (Fig. 3).

Identifying putatively functional variation using prediction algorithms is challenging and each approach has its own strengths and weaknesses. In the *CYP*-12, PolyPhen2, SIFT and Condel (9) predict that most of the novel variants we found were functional. Yet, these algorithms also fail to accurately predict the effects of some CYP-12 variants recognized experimentally to be functional (Supplementary Material, Fig. S1, Supplementary Material, Table S1). Accordingly, to make functional predictions about the novel variants discovered in the *CYP*-12, we used a combination of orthogonal approaches that consider information on evolutionary, biochemical and structural constraint.

To estimate the evolutionary constraint of each missense variant, Genomic Evolutionary Rate Profiling (GERP) scores (10) were calculated for each variant. SNVs with GERP scores >3 are predicted to more likely affect protein function and thus be enriched for alleles with phenotypic effect (11). We also calculated a Grantham score (12) for each missense variant. The Grantham score assesses the "severity" of a substitution by comparing biochemical properties of each amino acid residue; missense variants with a Grantham score >100 are predicted to result in "damaging" substitutions (12). Finally, we used published crystallographic and mutagenic studies to manually annotate residues that have a critical role in overall enzyme structure and function. Missense variants with GERP scores >3 or Grantham scores >100 were considered putatively functional. Because of their highly predictable effect on protein structure, all nonsense and splicesite variants as well as frameshifting indels were considered putatively functional. Using these criteria, we identified 219 novel, rare, putatively functional variants including 180 missense variants, 21 nonsense/splicesite variants and 18 frameshifting indels. Accordingly, we estimated that ~30% (219/731) of the novel variants we found in the *CYP*-12 are predicted to be functional (Supplementary Material, Table S2). Our results are comparable to the previous study of variation in drug targets; ~90% of variants reported in sequenced drug targets were novel (CYP-12, this study: 72.7%), and 39.2% of novel drug-target variants were predicted to have a functional effect (CYP-12, this study: 30%) (2). In comparison, the previous study had both larger sample size ($n = 14\,002$) and a broader set of genes ($n = 202$), which likely contributed to the slightly higher values reported by the study.

The extent to which these rare, novel predicted function variants in the *CYP*-12 contribute to overall -rug metabolism phenotypes remains to be tested. However, each of these *CYP* genes participates in the metabolism of diverse pharmaceuticals; therefore, a functional variant in any one of these genes could affect a broad range of drug responses. To this end, we counted the number of individuals who harbored one or more putatively functional novel variants in the *CYP*-12 (Table 2). We found that 11.7% of AA and 7.6% of EA carry a predicted functional novel allele in at least one major drug-metabolizing *CYP* gene, and while most individuals have only a single putatively functional allele, 25 individuals carried two or more predicted functional alleles. Known common actionable alleles, defined by exonic CYP-12 alleles with evidence rated at "Level 1 evidence" by PharmGKB, are widespread; 36.8% of AA and 57.2% of EA

Table 1. CYP variants with a known effect on drug response found among ESP individuals, along with their MAF, in individuals of either EA ($n = 4300$) or AA ($n = 2203$) ancestry

Chromosome	Position	rsID	Allele	Gene	Star allele	Amino acid change	ESP AA MAF	ESP EA freq
10	96522463	rs28399504	G	*CYP2C19*	*4	M/V	0.000920	0.00518
10	96702047	rs1799853	T	*CYP2C9*	*2	R/C	0.0588	0.264
10	96741053	rs1057910	C	*CYP2C9*	*3	I/L	0.0276	0.129
10	96798749	rs10509681	C	*CYP2C8*	*3	K/R	0.0515	0.248
10	96818106	rs11572103	A	*CYP2C8*	*2	I/F	0.312	0.00370
10	96827030	rs11572080	T	*CYP2C8*	*3	R/K	0.0515	0.247
19	41512841	rs3745274	T	*CYP2B6*	*6	Q/H	0.259	0.491
19	41518221	rs28399499	C	*CYP2B6*	*16	I/T	0.119	0.00148
19	41522715	rs3211371	T	*CYP2B6*	*5	R/C	0.0599	0.228
22	42523610	rs59421388	T	*CYP2D6*	*29	V/M	0.183	0
22	42523943	rs16947	G	*CYP2D6*	*2	C/R	0.494	0.136
22	42526694	rs1065852	A	*CYP2D6*	*10	P/S	0.236	0.441
22	42524947	rs3892097	T	*CYP2D6*	*4	Splice-3′	0.072854	.190723

Variants were gathered from PharmGKB annotations of the 12 drug-metabolizing CYP genes.

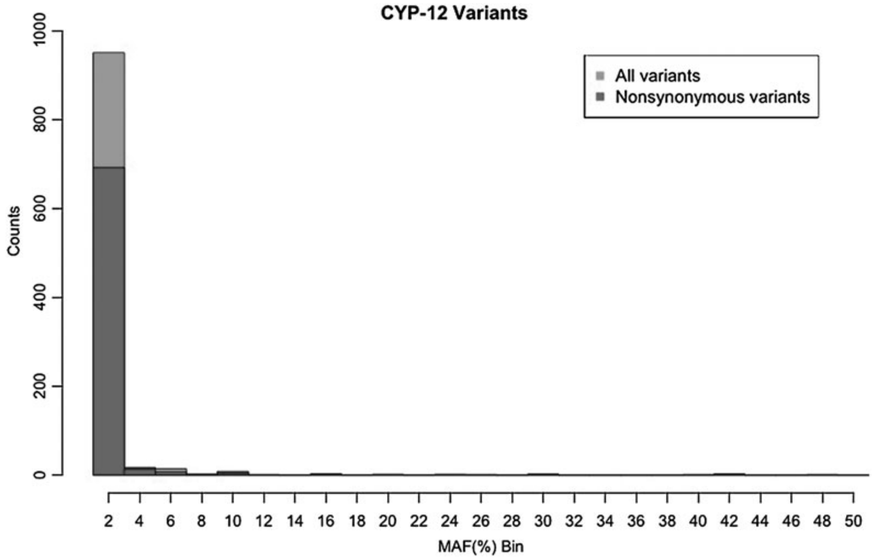

Figure 3. MAF for all *CYP*-12 variants as well as for only nonsynonymous variants (missense, nonsense, splicespite, indels).

carry at least one such allele. Moreover, if both novel predicted functional alleles and known exonic functional alleles are considered, 43.4% of AA and 59.3% of EA carried at least one putatively functional allele. Overall, 818 individuals (12.6%) had two or more predicted functional alleles in major drug-metabolizing *CYP* genes (Table 3). While the burden of potentially functional alleles increases in both populations when considering novel, rare variation, this increase is disproportionate in AA, whose individual burden increases by a factor of more than three compared with EA (+6.6 and +2.1%, respectively). This difference highlights the need for further studies of pharmacogenomic variation in admixed, understudied populations such as AAs. Because the data analyzed here are drawn from exome sequencing, we do not examine rare or common noncoding variation which could also contribute to overall drug response. Since there are several noncoding variants known to affect drug response in these genes (13), it is likely that our results are underestimates of the true burden of impactful *CYP*-12 variation within an individual.

DISCUSSION

To fully understand the effect of rare *CYP* variation on human drug metabolism and its clinical relevance, direct functional assessment and studies of genotype–phenotype relationships of each variant will be required. Our studies provide investigators with nearly 200 new high-priority candidate variants to test. Furthermore, some of the variants we identified have perhaps an even higher prior likelihood of being of clinical utility. For example, we identified 13 variants in *CYP2C9* that putatively affect its function and may, therefore, alter warfarin metabolism. These include variants predicted to disrupt known substrate-binding residues (Arg97Thr) (14), alter protein translation (Met1Val) and result in damaging substitutions at conserved sites (Pro363Leu; GERP = 3.51, Grantham = 98) (15). Previous *in vitro* mutagenesis studies have demonstrated that Arg97Thr is a loss-of-function variant that compromises heme cofactor binding, substrate specificity and overall protein stability (14,16). The functional effect of Met1Val has not been

Table 2. Amount of putative novel functional variation per CYP gene

Gene	Total number of putative functional variants	Number of individuals with putative functional variants AAs ($n = 2203$)	EAs ($n = 4300$)
CYP1A1	36	24	89
CYP1A2	21	19	18
CYP2A6	7	7	8
CYP2B6	14	11	26
CYP2C19	28	67	27
CYP2C8	22	32	41
CYP2C9	13	13	10
CYP2D6	21	40	39
CYP2E1	13	4	14
CYP3A4	19	9	17
CYP3A5	16	21	32
CYP3A7	9	11	5
Total	219	258	326

Columns 3 and 4 show the number of individuals that carry at least one allele of a candidate variant in the given gene.

Table 3. Burden of predicted functional CYP-12 variation per individual across the ESP6500 data, EA ($n = 4300$) or AA ($n = 2203$) ancestry

	Known (PharmGKB) only		Novel (ESP) only		Known (PharmGKB) and Novel (ESP)	
	EA	AA	EA	AA	EA	AA
4 alleles	0	0	0	1	6	3
3 alleles	5	0	0	1	44	13
2 alleles	527	100	10	13	591	161
1 allele	1928	712	231	207	1910	779
None	1840	1391	4059	1981	1749	1247

The number of individuals with 0, 1, 2, 3 or 4 predicted functional CYP-12 alleles. ‘Novel’ refers to predicted functional alleles discovered in ESP6500; ‘Known’ refers to exonic CYP-12 variants with “level 1 evidence” for functional association as reported by PharmGKB. ‘Known and Novel’ include both ESP and PharmGKB variants.

similarly studied, but can be inferred from an analogous Met1Val mutation in the highly homologous *CYP2C19*. This variant, *CYP2C19**4, is a well-characterized loss-of-function allele (17), and individuals carrying this allele along with a second loss-of-function allele are known to exhibit the clinically actionable clopidogrel ‘poor metabolizer’ phenotype (17). These data suggest that the *CYP2C9* Met1Val allele leads to a similar functional consequence due to the high homology between these genes and the nature of the mutation.

Only a fraction of phenotypic variance in warfarin maintenance dose is explained by the currently known variants, in *VKORC1* (25% of the variance) and *CYP2C9* (10% of the variance) (18). Accordingly, rare variants in *CYP2C9*, such as those identified herein, likely account for part of the variance that remains unexplained. While common variants such as those in *CYP2C9* were successfully identified through a GWAS approach, rare and private variants with an effect of drug phenotypes are unlikely to be identified through this method due to insufficient sequence coverage and study power. In recent years, the advent of next-generation sequencing has led to a rapid increase in the quantity of personal genomes and exomes available for analysis. This explosion of data is revealing that the vast majority of human variation is quite rare, and that this rare variation is enriched for variants predicted to alter gene function (8,19,20). As sequencing begins to enter the clinical environment, understanding the burden of rare and private variants on an individual’s genome is critical to the interpretation of personal genome data inherent to personalized medicine.

In summary, we discovered a large number of novel variants, nearly a third of which are predicted to be functional, in 12 *CYP* genes that affect the metabolism of ~75% of pharmaceuticals. Collectively 9% of individuals carry at least one of the novel predicted functional variant we found herein and together with known variants, 16.7% of individuals are predicted to carry a functional variant. As the pharmacokinetic reactions catalyzed by the CYP-12 represent only a small subset of the processes and pathways that collectively determine drug response, our findings likely represent an estimate of the lower bound of overall pharmacogenetic burden. Indeed, other large-scale studies of rare and private variation within pharmacodynamic genes (drug targets) reveal a similar abundance of rare, putatively functional variants that are individually rare, yet collectively common (2). Additionally, smaller-scale studies are revealing that this trend likely extends to genes involved in other aspects of drug response, including membrane transport (21). Given the results of our survey and others, we hypothesize that virtually every individual is likely to contain at least one predicted or known functional variant of pharmacogenetic relevance. Genotyping-based platforms that test only for known common variants are likely to misclassify as heterozygotes those individuals who are in fact compound heterozygotes, carrying both a common and a rare functional variant on separate haplotypes. Thus, while the potentially functional variation presented here is quite rare, and likely found in the homozygous state in very few individuals, we believe that consideration of such variation is necessary as clinical pharmacogenetic testing becomes increasingly popular. Although high-throughput techniques to experimentally evaluate the true functional effects of these variants are needed, our results indicate that rare variation, pervasive throughout these 12 critical genes, should be assessed and considered carefully in future pharmacogenetics work in both research and clinical settings. Understanding the phenotypic consequences of such rare variation will be a major next step forward in explaining the inter-individual variation in drug responses that have been observed for centuries and will provide better guidance for implementing personal genome sequencing at the clinical level.

MATERIALS AND METHODS

Study sample

The NHLBI ESP is a multi-center study to deeply sequence the exomes of individuals segregating a variety of heart, lung and blood disorders. The 6503 individuals used in the analysis were generated from samples ascertained from 20 different cohorts (detailed information of cohorts can be found in 19). Although these individuals are not a random sample, they were ascertained on a variety of distinct phenotypes such that cohort-specific effects are not expected to bias patterns of SNVs. Indeed, detailed analyses of a large subset ($n = 2440$) of these 6503 individuals found no systematic biases in patterns and characteristics of SNVs attributable to cohort or technical sources of variation. All study participants in each of the component studies provided written informed consent for the use of their DNA in studies aimed at identifying genetic risk variants for disease and for broad data sharing. Institutional certification was obtained for each sample to allow deposition of phenotype and genotype data in dbGaP and BAM files in the short-read archive.

Exome resequencing, variant calling and filtering

The processes of library construction, exome capture, sequencing and mapping were performed as previously described (19). SNVs were called using the UMAKE pipeline at University of Michigan, which allowed all samples to be analyzed simultaneously, both for variant calling and filtering. Briefly, we used BAM files summarizing BWA alignments generated at the University of Washington and the Broad Institute as input. These BAM files summarized alignments generated by BWA, refined by duplicate removal, recalibration and indel re-alignment. We excluded all reads that were not confidently mapped (Phred-scaled mapping quality <20) from further analysis. To avoid PCR artifacts, we clipped overlapping ends in paired reads. We then computed genotype likelihoods for exome targeted regions and 50 flanking bases, accounting for per base alignment quality using samtools. Variable sites and their allele frequencies were identified using a maximum-likelihood model, implemented in glfMultiples. These analyses assumed a uniform prior probability of polymorphism at each site. We used a support vector machine (SVM) classifier, which is a machine-learning algorithm, to separate likely true-positive and false-positive variant sites. SVM filtering started by collecting a series of features related to quality of each SNV, including overall depth, fraction of samples with coverage, fraction of reference bases in heterozygous individuals (allele balance), correlation of alternative alleles with strand and read position (strand and cycle bias), and inbreeding coefficient for each variant. SNVs that deviated significantly from expected values in three or more categories were flagged as likely false positives when training the SVM filter. SNVs at HapMap polymorphic sites and Omni 2.5 array polymorphic sites in the 1000 Genomes project data were flagged as likely true positives. After examining this training set, the SVM classifier was used to identify all likely false-positive sites, which were excluded from downstream analyses. A total of 1 908 614 SNVs passed the SVM filter, with an overall transversion to transition ratio (Ts/Tv) of 2.84.

After the initial SNV calls were generated, we re-examined the VCF files and applied filters considering total read depth, the number of individuals with coverage at the site, the fraction of variant reads in each heterozygote, the ratio of forward and reverse strand reads for reads carrying reference and variant alleles, and the average position of variant alleles along a read. Next, the SNV call set included variants that were called with posterior probability $>99\%$ (glfMultiples SNP quality >20), were at least 5 bp away from an indel detected in the 1000 Genomes Pilot Project, were targeted in at least 99% individuals and had a total depth across samples between 6823 and 6 823 000 ($\sim$1–1000 reads per sample at average). Sites where the read depth of the variant allele was $>65\%$ in heterozygotes or where the absolute squared correlation between allele (variant or reference) and strand (forward or reverse) was >0.15 were excluded. In order to obtain genotypes with high accuracy suitable for population genetics analyses, we further set individual genotype to missing data if it had quality (GQ) <30 and/or filtered depth (DP) <10. After such filtering, variants with $>10\%$ of missing genotypes across individuals were excluded from further analysis. A sample of 145 novel, singleton variants and 323 novel, non-singleton variants from across the exome were selected for validation via Sanger sequencing; 143/145 (99%) of the singleton variants and 316/323 (98%) of the non-singleton variants were validated (19).

Identification of related individuals and assignment of ancestry

In total, 6823 exomes were obtained from individuals who self-identified as EA ($n = 4419$), AA ($n = 2343$), and others (including Asian, Hispanic and Native American). To remove related individuals, we performed a KING analysis on the filtered data. Specifically, we performed LD pruning using PLINK to the variants with MAF $>5\%$. This resulted in 34 945 SNVs for the analysis. KING identifies kinship by pairwise comparisons across all individuals and is robust to population structure. Using the authors' guidelines for a 3rd degree relationship (i.e. first cousins), we used a kinship coefficient threshold of 0.04419. From this, we were able to form clusters of related individuals, with the majority of clusters consisting of two individuals. When all individuals were related to all other individuals in a cluster, we preferentially removed those with the greatest overall missingness. When these clusters had partial relationships (i.e. A is related to B and C but B and C are not related) then we preferentially removed those who would leave the largest number of samples. This resulted in the removal of 242 individuals. After removing these individuals, we repeated the KING analysis and found no kinships in the remaining dataset. Using the same filtered dataset from the KING analysis, we performed a principal component analysis (PCA) to infer genetic ancestry. Asian, Hispanic and Native American samples were removed from the analysis.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff and study participants in creating this resource for biomedical research.

Conflict of Interest statement. None declared.

FUNDING

Funding for GO ESP was provided by NHLBI grants RC2 HL-103010 (HeartGO), RC2 HL-102923 (LungGO) and RC2 HL-102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL-102925 (BroadGO) and RC2 HL-102926 (SeattleGO).

REFERENCES

1. Wang, L., McLeod, H.L. and Weinshilboum, R.M. (2011) Genomics and drug response. *N. Engl. J. Med.*, **364**, 1144–1153.
2. Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D. *et al.* (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*. 10.1126/science.1217876.
3. Evans, W.E. and Relling, M.V. (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. *Science*, **286**, 487–491.
4. Nelson, D.R., Zeldin, D.C., Hoffman, S.M.G., Maltais, L.J., Wain, H.M. and Nebert, D.W. (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics*, **14**, 1–18.
5. Paré, G., Mehta, S.R., Yusuf, S., Anand, S.S., Connolly, S.J., Hirsh, J., Simonsen, K., Bhatt, D.L., Fox, K.A.A. and Eikelboom, J.W. (2010) Effects of CYP2C19 genotype on outcomes of clopidogrel treatment. *N. Engl. J. Med.*, **363**, 1704–1714.
6. International Warfarin Pharmacogenetics ConsortiumKlein, T.E., Altman, R.B., Eriksson, N., Gage, B.F., Kimmel, S.E., Lee, M.-T.M., Limdi, N.A., Page, D., Roden, D.M. *et al.* (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Med.*, **360**, 753–764.
7. Meckley, L.M., Wittkowsky, A.K., Rieder, M.J., Rettie, A.E. and Veenstra, D.L. (2008) An analysis of the relative effects of VKORC1 and CYP2C9 variants on anticoagulation related outcomes in warfarin-treated patients. *Thromb. Haemost.*, **100**, 229–239.
8. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D.A. *et al.* (2012) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. 10.1038/nature11690.
9. González-Pérez, A. and López-Bigas, N. (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.*, **88**, 440–449.
10. Cooper, G.M., Stone, E.A. and Asimenos, G. NISC Comparative Sequencing Program, Green, E.D., Batzoglou, S. and Sidow, A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
11. Cooper, G.M., Goode, D.L., Ng, S.B., Sidow, A., Bamshad, M.J., Shendure, J. and Nickerson, D.A. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods*, **7**, 250–251.
12. Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.
13. Lane, S., Al-Zubiedi, S., Hatch, E., Matthews, I., Jorgensen, A.L., Deloukas, P., Daly, A.K., Park, B.K., Aarons, L., Ogungbenro, K. *et al.* (2012) The population pharmacokinetics of R- and S-warfarin: effect of genetic and clinical factors. *Br. J. Clin. Pharmacol.*, **73**, 66–76.
14. Dickmann, L.J., Locuson, C.W., Jones, J.P. and Rettie, A.E. (2004) Differential roles of Arg97, Asp293, and Arg108 in enzyme stability and substrate specificity of CYP2C9. *Mol. Pharmacol.*, **65**, 842–850.
15. Williams, P.A., Cosme, J., Ward, A., Angove, H.C., Matak Vinković, D. and Jhoti, H. (2003) Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature*, **424**, 464–468.
16. Davies, C., Witham, K., Scott, J.R., Pearson, A., DeVoss, J.J., Graham, S.E. and Gillam, E.M.J. (2004) Assessment of arginine 97 and lysine 72 as determinants of substrate specificity in cytochrome P450 2C9 (CYP2C9). *Drug Metab. Dispos.*, **32**, 431–436.
17. Scott, S.A., Sangkuhl, K., Gardner, E.E., Stein, C.M., Hulot, J.S., Johnson, J.A., Roden, D.M., Klein, T.E. and Shuldiner, A.R. Clinical Pharmacogenetics Implementation Consortium. (2011) Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450-2C19 (CYP2C19) genotype and clopidogrel therapy. *Clin. Pharmacol. Ther.*, **90**, 328–332.
18. Rettie, A.E. and Tai, G. (2006) The pharmocogenomics of warfarin: closing in on personalized medicine. *Mol. Interv.*, **6**, 223–227.
19. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.
20. Marth, G.T., Yu, F., Indap, A.R., Garimella, K., Gravel, S., Leong, W.F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X. *et al.* (2011) The functional spectrum of low-frequency coding variation. *Genome Biol.*, **12**, R84.
21. Ramsey, L.B., Bruun, G.H., Yang, W., Trevino, L.R., Vattathil, S., Scheet, P., Cheng, C., Rosner, G.L., Giacomini, K.M., Fan, Y. *et al.* (2012) Rare versus common variants in pharmacogenetics: SLCO1B1 variation and methotrexate disposition. *Genome. Res.*, **22**, 1–8. doi: 10.1101/gr.129668.111.
22. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.