Tools and Challenges for the Implementation of Next-Generation Sequencing in Clinical Pharmacogenetics

Adam Gordon

A dissertation

submitted in partial fulfillment of the

requirements of the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Deborah Nickerson

Joshua Akey

Maitreya Dunham

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Tools and Challenges for the Implementation of Next-Generation Sequencing in Clinical Pharmacogenetics

Adam Gordon

Chair of the Supervisory Committee:
Professor Deborah Nickerson
Department of Genome Sciences

Understanding the genetic basis of an individual's response to therapeutic drugs (pharmacogenetics) is a unique area of research with significant translational impact for medicine. Known genetic variants with effects on important clinical phenotypes, including clopidogrel efficacy and warfarin maintenance dose, highlight the potential translational utility of pharmacogenetic analysis. Current strategies for clinical pharmacogenetic testing are primarily limited to genotyping of known, common variants. The emergence of next-generation sequencing offers a promising new tool to explore the links between drug response and genetic variation, both common and rare.

The focus of my dissertation has been the application of next-generation sequencing technology to pharmacogenetic research and implementation. First, using exome sequence data from thousands of individuals, I demonstrate that novel, deleterious variation is common in key drug metabolizing enzymes among individuals of European and African descent, despite each variant

being individually quite rare. I then use this same dataset to explore the inability of current pharmacogenetic nomenclature systems to accurately translate and represent results derived from exome sequencing. Finally, I present the development and testing of PGRNseq, a custom-capture platform designed for rapid, accurate detection of genetic variation within key pharmacogenes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Debbie Nickerson, for supporting me throughout my graduate career. Without her guidance I would not have discovered my passion for precision medicine. A tireless supporter of young investigators and grad students, Debbie provided me access both to high quality data and high-quality collaborators that are often inaccessible to early-career scientists such as myself. I greatly appreciate the opportunities she has given me, and my years in her lab have taught me a lot about how to be a P.I.

I would also like to thank everyone else in the Nickerson Lab, past and present, for creating an environment that at times felt more like a family than a group of coworkers. I've really enjoyed being your friend, colleague, and encyclopedia, and I'll wholeheartedly miss the camaraderie that I've come to know and love. Likewise, I'm also deeply thankful for the opportunity to attend graduate school in the Department of Genome Sciences, and I can't thank the faculty, staff, and students enough for establishing an environment where collaboration and good science are valued so highly. As I continue on in my career I can only hope to carry that spirit with me to any institution I find myself at.

Finally, none of this would be possible without the support of my family. Although being on opposite corners of the nation has been difficult at times, your love is always with me.

Chapter 1: Introduction

During their lifetime, the average American will use medication to treat most ailments they encounter. These medications, both prescribed and over-the-counter, come with recommended doses based on clinical trials. Despite these recommendations, Adverse Drug Reactions (ADRs), including overdoses and harmful drug interactions, are a leading cause of hospitalization and death in the US[1]. Although there are many factors that contribute to ADRs, genetic variability in drug targets, drug transporters, and drug metabolizing enzymes is thought to be responsible for up to 30% of all reported ADRs[2]. Clinical pharmacogenetics (PGx) research aims to describe the set of genetic variants that underlie the inter-individual variation observed in drug efficacy and toxicity, with a distinct focus on those variants that can guide clinical decision making to reduce ADRs and improve patient outcomes broadly for commonly prescribed pharmaceuticals.

1.1 Genetics of drug response and toxicity

Although the term 'pharmacogenetics' was first coined by Friedrich Vogel in 1959[3] genetic variation was first proposed as a contributing factor in adverse drug reactions by Arno Motulsky in 1957. In reviewing data from clinical trials on the efficacy of succinylcholine, a muscle relaxant, Motulsky speculated that "drug reactions…may be considered pertinent models for demonstrating the interaction of heredity and environment in the pathogenesis of disease.[4]" In the over 50 years that have passed since this initial hypothesis, studies aiming to describe the genetic underpinnings of variation in drug response have uncovered associations with genes throughout entire pharmacological pathways, including drug metabolism genes, drug transporters, and the drug targets themselves[5].

Pharmacogenetics of drug-metabolizing enzymes

Most pharmaceuticals are ingested in an inactive form, and only become active upon the addition of reactive and polar groups. This reaction, known as Phase I metabolism, is often catalyzed by a family of enzymes known as Cytochrome P450s (CYPs), responsible for oxidizing a variety of endogenous and xenobiotic compounds[6]. Although the human genome contains 57 different CYPs, over three-quarters of all known Phase I metabolism reactions are catalyzed by only 13 critical CYPs[5]. Accordingly, genetic variation in these 13 CYPs has been linked to response to a variety of their substrates: variation in *CYP2C9* is associated with both phenytoin[7] and warfarin response[8], variation in *CYP2D6* is associated with response to codeine[9] and tricyclic antidepressants[10]. Phase II reactions, or 'conjugations,' couple these active metabolites to charged species for the purpose of excretion. As the rate at which these reactions occur is directly related to the concentration of the active metabolite, genes responsible for catalyzing Phase II reactions, commonly transferases, have been associated with response to a variety of their substrates, including *TPMT* and thiopurines[11] as well as *UGT1A1* and irinotecan[12].

Pharmacogenetics of drug transport genes

Although some drugs cross cellular membranes via passive diffusion, proteins involved in the active transport of drugs play a critical role in the absorption, distribution, and elimination of many common pharmaceuticals. This includes proteins in the ABC family of efflux pumps, which enable the active transport of pharmaceuticals across GI membranes[13], and organic anion transporters in the SLC family, which are largely involved in intracellular hepatic transport of a

number of different pharmaceuticals[14]. Accordingly, variation in these genes that even subtly alters their ability to perform these functions can have profound effects on overall drug response and toxicity.

For example, *SLCO1B1* is responsible for mediating hepatic clearance of over 30 different endogenous and exogenous compounds, including simvastatin, the third most commonly prescribed drug in the U.S.[15]. Although there are many common statin-related side effects, skeletal muscle toxicity is the most common statin-related ADR, and statin-related myalgias are reported in an estimated 5% of all statin users[16]. As this toxicity is directly related to increased plasma concentrations of the drug, variation affecting SLCO1B1's transport efficiency are likely to alter both risk of this ADR and the overall therapeutic index of statins generally. Indeed, variation in *SLCO1B1*, both common and rare, has been associated with decreased transport function *in vitro*[17] and decreased drug clearance *in vivo*[18]; additional GWAS have confirmed the association between common variation and statin-induced myopathy, including one particular allele found to explain 60% of myopathy cases studied[19].

Pharmacogenetics of drug target genes

Just as variation in genes responsible for metabolizing and transporting drugs can affect response and toxicity, so too can variation within the genes encoding the drug targets themselves. Indeed, variation in over 25 different targets has been associated with variation in response to their respective drug[20]. Notable gene-drug pairs that fall into this category include *ACE* and response to ACE inhibitors[21], *ADRB2* and response to beta blockers[22], and *VKORC1* and warfarin maintenance dose[23]. Although the associations behind these drug-target gene pairs remain robust genome-wide, they also emphasize the need to include contributions from genetic

variation across the entire ADME pathway when modeling drug response within individuals. For example, while variants in both *CYP2C9* (a drug metabolizer) and *VKORC1* (a drug target) are independently associated with warfarin maintenance dose, combined analysis of both loci together reveal differential contributions to the overall warfarin dose variation: *VKORC1* polymorphisms explain approximately 25% of the population variance in stabilized warfarin dose, compared to the approximately 10% explained by variation in *CYP2C9* [24]. This disparity illustrates the utility of genome-wide approaches in pharmacogenetic research and implementation that are able to interrogate the full complement of variation within PGx loci using a single assay.

1.2 Pharmacogenetics in clinical practice

Clinical guidelines for 'actionable' genes

Although the past 5 decades of PGx research have uncovered dozens of associations between genetic variation and drug response phenotypes, only a subset are both robustly replicated and capable of altering clinical decision making regarding drug choice or dosing. As the evidence base behind these associations is constantly evolving, The Clinical Pharmacogenetics Implementation Consortium (CPIC) was formed to identify this critical subset, and to provide published, evidence-based, updated guidelines to aid in the implementation of PGx findings in clinical care [25]. To date, over 10 such guidelines have been published, with more in preparation as the evidence for clinical utility of additional gene/drug pairs continues to grow. Gene/drug pairs with an accompanying CPIC guide – the set of 'actionable' PGx loci—represent the minimum set that should be assessed for comprehensive, preemptive pharmacogenetic testing.

Current methods for assaying clinical PGx variation

Despite CPIC's standardization of the set of actionable loci, a variety of competing methods are being adopted to assay these loci, or a subset of them, by different clinical entities nationwide; this fragmentation in methodology hinders the harmonization of PGx-guided clinical decision support. For example, while both St. Jude Children's Research Hospital and Vanderbilt University Medical Center currently assay *TPMT* to help guide thiopurine dosing, they each employ entirely different platforms to generate these results, potentially leading to inconsistent interpretation of results between centers[26]. While these assays are generally centered around array-based genotyping, the exact targets differ substantially between platforms, and often the exact probe sequences are not disclosed. Additionally, these platforms are often supplemented with CNV-typing methods including qPCR and LR-PCR which may differ substantially between institutions.

These institutional differences can lead to significant consequences in the interpretation of PGx results. For example, a test result may indicate a patient carries only the reference haplotype for a key pharmacogene (e.g. *CYP2C9*1/*1*) is dependent on the suite of variants that were assessed: a **1/*1* result from an assay of only 3 common *CYP2C9* variants cannot be interpreted in the same context as a **1/*1* result from an assay which genotypes all known CYP2C9 alleles. As clinical PGx reports often do not indicate exactly which loci tested (and, similarly, which were not assayed), there is a need for new standard methods capable of rapid, comprehensive interrogation of many PGx loci, common and rare.

1.3 Next-generation sequencing

Next-generation sequencing (NGS) is becoming increasingly popular in clinical practice due to its declining cost and ability to both genotype known variants and discover novel variation across a large number of genomic loci. Briefly, NGS involves the preparation of a complex library of sample DNA which is subsequently amplified and densely arrayed onto a solid surface. This array is then subjected to several rounds of complementary base incorporation, which is monitored in parallel for all clusters on the array using fluorescent byproducts of base incorporation[27]. As whole-genome sequencing using this method remains relatively costly, capture-based methods for rapid interrogation of genomic loci are becoming widely adopted for clinical NGS; this includes both large target spaces such as the entire exome and smaller, custom-target assays that capture only a subset of genes or other genomic features. NGS methods are a promising avenue for clinical implementation of PGx due to their ability to accurately assay known variants and identify rarer, novel variants within known PGx targets that may play a role in overall drug response. As recent sequencing-based studies of individual genes or suites of genes continue to uncover substantial rare, deleterious variation within key PGx loci, platforms that genotype only known, common variants are becoming increasingly obsolete[28]. Although this technology can be a significant step forward for pharmacogenetics, there are substantial challenges facing clinical implementation of NGS inherent in the use of short reads, including erroneous variants derived from paralogous loci and difficulty detecting and genotyping Copy Number Variation (CNV) and other structural variants[29].

1.4 Dissertation aims

The general theme of this dissertation is to explore the advantages and pitfalls of using next-generation sequencing for clinical pharmacogenetics. Chapter 2 describes the extent of rare, deleterious variation across a set of key drug metabolizing enzymes using exome sequencing data from thousands of individuals. Chapter 3 explores how this exome data can be harmonized with existing pharmacogenetic nomenclature, with an eye towards identifying and quantifying misclassifications with potential clinical impact. Chapter 4 presents PGRNseq, a new custom capture, NGS-based tool for pharmacogenetic research and implementation, which retains high-throughput nature of exome sequencing without sacrificing the ability to assess known PGx alleles that lie outside coding regions. The final chapter summarizes this work and presents a vision for the future of NGS-based, preemptive PGx testing.

Chapter 2: Quantifying rare, deleterious variation in 12 human Cytochrome P450 drug-metabolism genes using large-scale exome data

This chapter was previously published as:

Gordon, A.S.; Tabor, H.K.; Johnson, A.D.; Snively, B.M.; Assimes, T.L.; Auer, P.L.; Ioannidis, J.P.; Peters, U.; Robinson, J.G.; Sucheston, L.E.; et al. Quantifying rare, deleterious variation in 12 human cytochrome p450 drug-metabolism genes in a large-scale exome dataset. *Hum. Mol. Gen.* 2014, 23, 1957–1963, doi:10.1093/hmg/ddt588.

2.1 Abstract

Although the study of genetic influences on drug response and efficacy ('Pharmacogenetics') has existed for over 50 years, we still lack a complete picture of how genetic variation, both common and rare, affects each individual's responses to medications. Previous studies of such variation using Genome Wide Association and resequencing approaches have had limited success, in part due to their incomplete characterization of coding variation that may have significant consequences on drug response. 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 African-American and 4300 Caucasian individuals through the NHLBI Exome Sequencing Project, we conducted a survey of coding variation within twelve 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 novel nonsynonymous variants in each of the target *CYP* genes. We constructed a list of putative functional variants that may play a role in overall drug metabolism using Genomic Evolutionary Rate Profiling (GERP), Grantham score, and literature review to assess

evolutionary, biochemical, and structural significance. This list includes variants with diverse functional effects such as premature stop codons, aberrant splice sites, and mutations at conserved active site residues. While these candidate variants are individually rare, 7.6-11.7% of individuals interrogated in the study carry at least one newly described potentially deleterious mutation in a major drug-metabolizing *CYP*.

2.2 Background

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[30]. 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.

The cytochrome P450 (*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[6], a subset of just 12 (*CYP*-12) of them are collectively responsible for ~75% of all known drug oxidation reactions[5]. Several reported *CYP* variants influence clinically-important phenotypes such as the efficacy of clopidogrel and the maintenance dosing of warfarin[31,32]. 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[33].

2.3 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 African-American (AA; n = 2203) and European-American (EA; n = 4300) ancestry[34]. Variants were identified and genotyped using the UMAKE pipeline (http://genome.sph.umich.edu/wiki/UMAKE) and subjected to an SVM-based filter based on quality, depth, and allele balance metrics to remove false positive calls due to sequencing errors and mismapped paralogous reads[35]. Small in/del variants were analyzed using the GATK variation discovery pipeline following the guidelines in the GATK best practices v4 (http://gatkforums.broadinstitute.org/discussion/1186/best-practice-variant-detection-with-the-gatk-v4-for-release-2-0). Variants passing these filters were subjected to additional filters for missingness (all sites with calls in <10% of samples were removed) and kinship (closely related individuals are removed). The final data set was then annotated with functional information using the SeattleSeq pipeline (http://snp.gs.washington.edu/SeattleSeqAnnotation134/) and is available through the ESP Exome Variant Server (http://evs.gs.washington.edu/EVS/). 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[36].

Across the CYP-12, 98.1% of coding sequence was covered with an average depth of 30X or greater. We discovered a total of 1006 unique variants in the *CYP*-12. This included 275 known and 731 novel variants compared to dbSNP (build 132, http://www.ncbi.nlm.nih.gov/projects/SNP/) of which 486 were missense variants and 42 were nonsense/splice site variants or frameshifting in/dels (Figure 2.1).

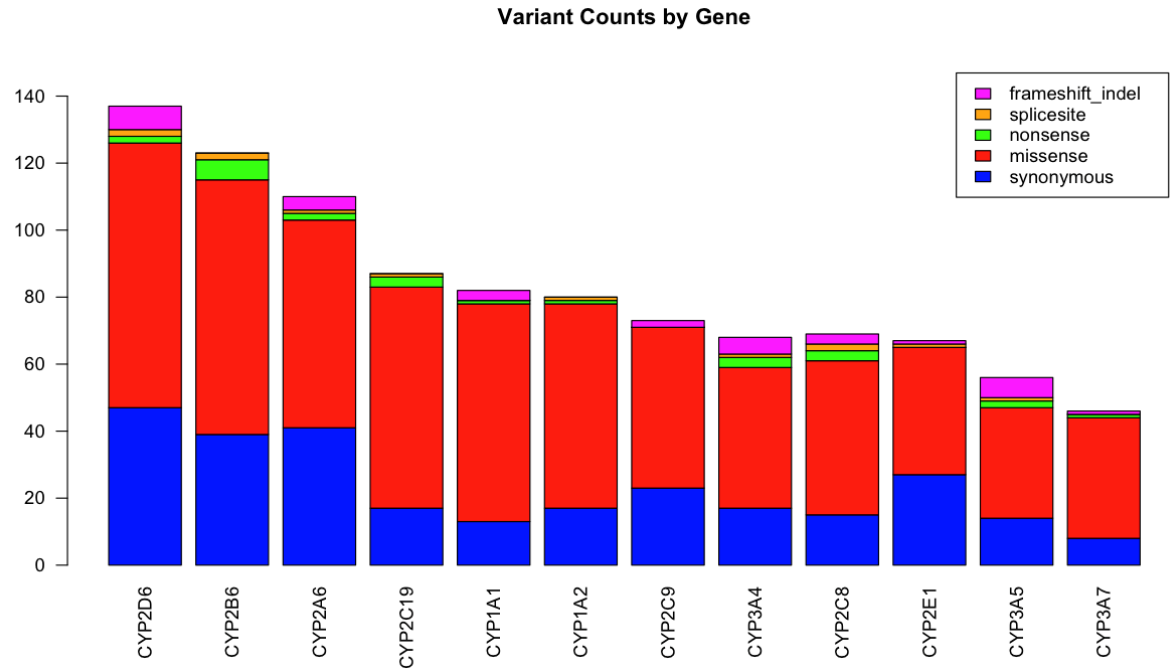

Figure 2.1 Distribution of exonic variation across 12 *CYP* genes separated by variant consequence. Variant types (missense, nonsense, synonymous, splicesite, frameshift) were determined using SeattleSeq annotation. For genes that produce more than one known transcript (*CYP2D6, CYP2C8, CYP3A4)*, annotation was based on the primary transcript.

We estimated the minor allele frequency of each *CYP* variant in EA and AA separately and the site frequency spectrum of known and novel *CYP* alleles (Figure 2.2). Overall, the majority of variation in drug-metabolizing *CYP*s is 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 a MAF > 2%.

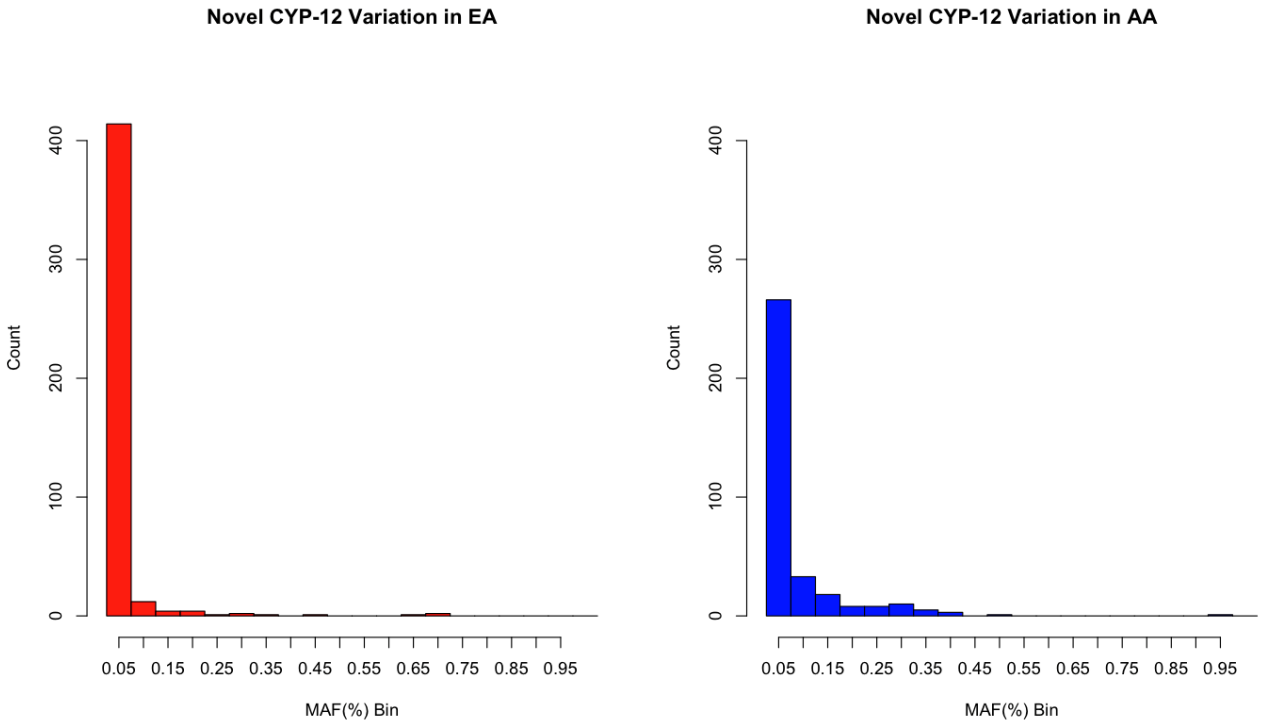

Figure 2.2 Minor allele frequency (MAF) for novel and known variants across the *CYP*-12 in European Americans and African Americans.

In addition to this novel variation, we identified many known functional exonic variants across the *CYP-12*, including clinically-relevant alleles such as CYP2C9*2, CYP2B6*6, and CYP2D6*4. Table 2.1 provides accurate allele frequencies 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 (Figure 2.3).

Table 2.1 *CYP* variants with a known effect on drug response found among ESP individuals
Minor Allele Frequencies (MAF) are for individuals of either European-American (EA,n = 4300) or African-American (AA, n = 2203) ancestry. Variants were gathered from PharmGKB annotations of the 12 drug-metabolizing *CYP* genes.

Chromosome	Position	rsID	Allele	Gene	Star Allele	Amino Acid change	ESP AA MAF	ESP EA MAF
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

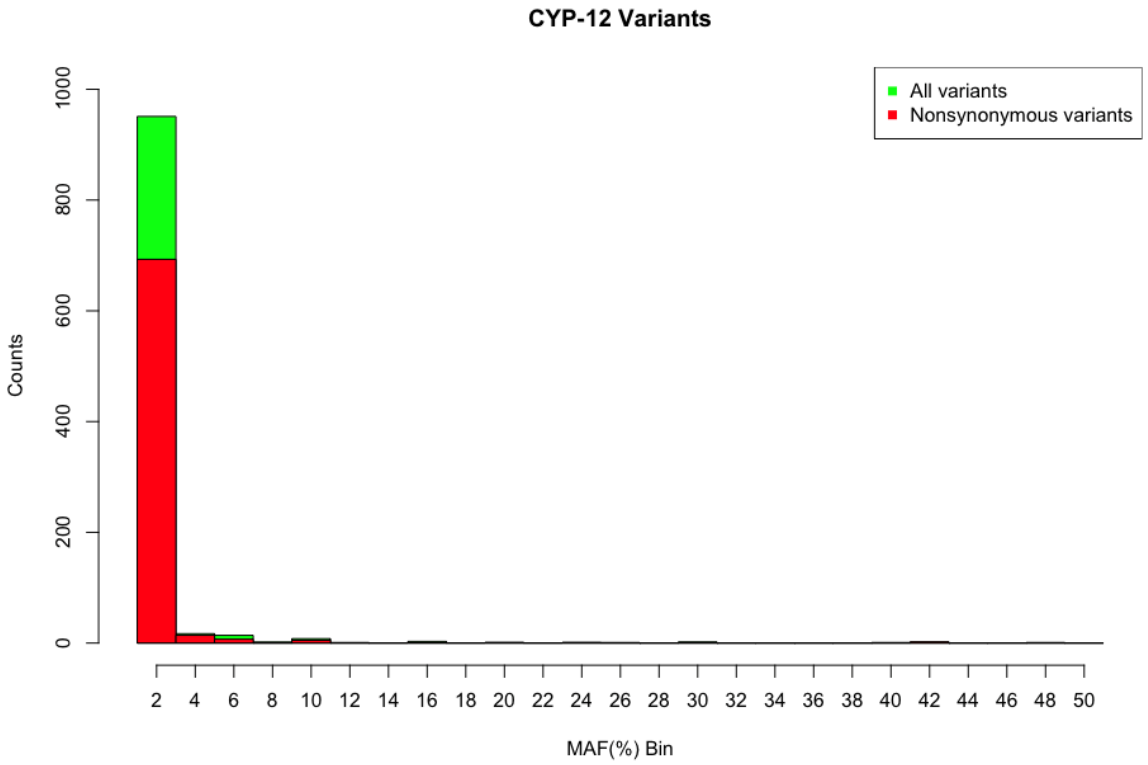

Figure 2.3 Minor allele frequency (MAF) for all *CYP*-12 variants as well as for only nonsynonymous variants (missense, nonsense, splice spite, in/dels).

Identifying putatively functional variation using prediction algorithms is challenging and each approach has its own strengths and weaknesses. In the *CYP*-12, PolyPhen2 and SIFT 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 (Figure 2.4). 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.

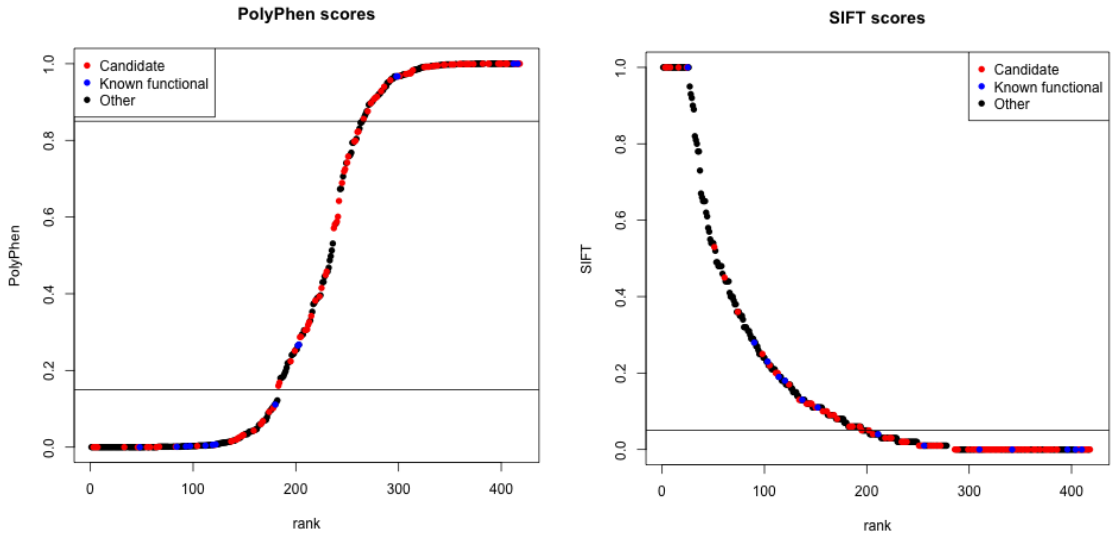

Figure 2.4: *CYP* variant effect prediction using common algorithms. We calculated PolyPhen2 and SIFT scores for candidate, non-candidate, and known functional *CYP* variants using ENSEMBL's VEP tool (http://uswest.ensembl.org/info/docs/variation/vep/index.html). The set of known functional *CYP* variants was gathered from PharmGKB and OMIM annotations of the 12 drug-metabolizing *CYP* genes.

To estimate the evolutionary constraint of each missense variant, Genomic Evolutionary Rate Profiling (GERP) scores [37] 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 [38]. We also calculated a Grantham score[39] 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[39]. Last, 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 splice-site variants as well as frameshifting in/dels were considered putatively functional. Using these criteria, we identified 219 novel, rare, putatively functional variants including 180

missense variants, 21 nonsense/splice-site variants, and 18 frameshifting in/dels. Accordingly, we estimated that approximately 30% (219/731) of the novel variants we found in the *CYP*-12 are predicted to be functional.

The extent to which these rare, novel predicted function variants in the *CYP*-12 contribute to overall drug metabolism phenotypes remains unclear. However, since each of these *CYP* genes participates in the metabolism of diverse pharmaceuticals, 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 2.2). We found that 11.7% of AA and 7.6% of EA carry a predicted functional novel variant in at least one major drug-metabolizing *CYP* gene, and while most individuals have only a single putatively functional variant, 42 individuals carried two or more predicted functional variants.

Table 2.2 Amount of putative novel functional variation per *CYP* gene. Columns 3 & 4 show the number of individuals that carry at least one allele of a candidate variant in the given gene.

Gene	Total number of putative functional variants	Number of individuals with putative functional variants	
		African-Americans (n=2203)	European-Americans (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

Moreover, if both novel predicted functional variants and known exonic functional variants are considered (Level 1 or 2 evidence for function in PharmGKB), 21.8% of AA and 14.1% of EA carried at least one putatively functional variant and 92 individuals (1.4%) had two or more predicted functional variants in major drug metabolizing *CYP* genes (Table 2.3). Because the data analyzed here are drawn from exome sequencing, this study does not examine rare or common noncoding variation which may contribute to overall drug response. However, as there are several noncoding *CYP*-12 variants that are known to affect drug response[40], our results are likely underestimates of the true individual burden.

Table 2.3 Burden of predicted functional CYP-12 variation per individual across the ESP data. Table shows the number of individuals with 1, 2, 3, 4, or no predicted functional CYP-12 variants. 'Novel' refers to potential functional alleles discovered in ESP; 'Known & Novel' includes both the new ESP variants and exonic CYP-12 variants with level 1 or 2 evidence for function.

	Number of individuals with X predicted functional CYP-12 variants			
	Novel (ESP) only		**Known (PharmGKB) & Novel (ESP)**	
	EA	**AA**	**EA**	**AA**
4 variants	1	2	2	3
3 variants	5	4	7	4
2 variants	18	13	42	34
1 variant	291	172	572	423
none	3985	2012	3677	1739

2.4 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 two hundred 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 thirteen 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)[41], alter protein translation (Met1Val), and result in damaging substitutions at conserved sites (Pro363Leu; GERP = 3.51, Grantham = 98)[42]. Only a small fraction of phenotypic variance in warfarin maintenance dose is explained by known variants, *VKORC1*, (25% of the variance), and *CYP2C9* (10% of the variance)[43]. Accordingly, rare variants in *CYP2C9*, such as those identified herein, likely account for part of the variance that remains unexplained.

In summary, we discovered a large number of novel variants, nearly a third of which are predicted to be functional, in twelve *CYP* genes that affect the metabolism of approximately 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. If our findings are indicative of patterns of rare variation in other genes involved in drug metabolism and response, we hypothesize that virtually every individual is likely to contain at least one predicted / known functional variant of pharmacogenetic relevance. Understanding the phenotypic consequences of such rare variation could be a major next step forward in explaining the inter-individual variation in drug responses

that have been observed since antiquity and provide better guidance for developing more personalized therapeutics.

2.5 Materials and Methods

Study Sample

The NHLBI Exome Sequencing Project (ESP) is a multi-center study to deeply sequence the exomes of individuals segregating a variety of heart, lung, and blood disorders. The 6,503 individuals used in the analysis were generated from samples ascertained from 20 different cohorts (detailed information of cohorts can be found in [36]). 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=2,440) of these 6,503 individuals found no systematic biases in patterns and characteristics of SNVs attributable to cohort or technical sources of variation1. 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[36]. 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 (BAQ) 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 5bp 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 to 6823000 (~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) less than 30 and/or filtered depth (DP) less than 10. After such filtering, variants with more than 10% of missing genotypes across individuals were excluded from further analysis.

Identification of related individuals and assignment of ancestry

In total, 6,823 exomes were obtained from individuals who self-identified as European American (EA, n=4,419), African American (AA, n=2,343) 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 minor allele frequency (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 data set. Using the same filtered data set 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.

Chapter 3: Evaluating the Use of Star Allele Nomenclature with High-Throughput Sequence Data

3.1 Introduction

Pharmacogenetic (PGx) testing is becoming progressively more common in clinical care as the number of available tests increases while sequencing costs continue to decline. Much of the decision support for these tests assigns a diplotype, often expressed as a pair of "star alleles", to each test result, translates these alleles into a predicted phenotype, and recommends a course of action to the provider. Therefore, the robust representation of test results is essential for deriving accurate clinical interpretations and recommendations. Although many of the current commercial testing platforms are based on genotyping approaches, next-generation sequencing (NGS) based methods are being quickly adopted for the detection of actionable PGx alleles. As NGS becomes the standard in clinical care, inconsistencies in the interpretation and reporting of PGx alleles could have severe impacts at the patient level. Indeed, nearly every individual is predicted to carry at least one PGx diplotype designated as "high-risk" by the Clinical Pharmacogenetic Implementation Consortium (CPIC)[26]. Recent studies of large-scale exome datasets have supported this prediction, revealing a significant burden of actionable PGx alleles in both European-American (EA, mean = 11.1 alleles) and African-American (AA, mean = 12.3 alleles) individuals[44]. Although striking, these findings are still likely underestimates of the true individual PGx burden due to the presence of rare, deleterious variants not accounted for by the current nomenclature system. Indeed, analysis of exome data reveals that 7-10% of individuals carry at least one undescribed, potentially deleterious rare allele in a major drug-metabolizing enzyme (See Chapter 2).

As the star allele nomenclature system was designed largely in the context of genotyping data, its ability to represent next-generation sequencing (NGS) results has not been systematically evaluated. In particular, applying the star nomenclature system to NGS data can lead to significant structural and semantic issues in the interpretation of results. Structurally, the current star system does not have the ability to name newly described alleles within a clinically-relevant timeframe, and the naming rules regarding synonymous variation is inconsistent or unclear. Additionally, star nomenclature fails to capture the type of variation described (i.e. single variant, multi-variant haplotype, or full gene deletion), and the star allele definition used in clinical testing to translate genetic into a specific diplotype is often undocumented or unclear. Perhaps most critical for NGS, however, is that a star allele designation often implies genotypes at sites that were non-interrogated. For example, *CYP2C9**3 is defined by a single missense variant (rs1057910); a clinical test report indicating a patient's diplotype is *3/*3 implies that that individual carries no other variation in this gene, even if only the one missense variant is typed. As NGS continues to uncover rare, deleterious variation not currently represented by the star system, these issues could lead to patient misclassification with significant potential clinical impact. Here we describe our efforts to quantify and categorize these naming errors across 5 key pharmacogenetic targets (*TPMT, CYP2C9, SLCO1B1, CYP3A5, CYP2C19)* using a large-scale, phased exome data set.

3.2 Methods

Exome data

We extracted genotype information for the coding regions of our 5 genes of interest from the NHLBI-Exome Sequencing Project (ESP) dataset, which consists of 4300 European-American and 2203 African-American individuals drawn from a variety of longitudal cohorts. Sample selection, sequencing pipeline, genotyping, and QC details have been previously described[34]. Sequence data was then phased for each gene individually using Beagle 4.0 with default parameters and no reference panel[45], resulting in 5 gene-specific, phased multisample VCFs representing our final dataset.

Allele definition tables and test data

In order to ensure accurate translation between NGS data and star allele names, we hand curated the allele definition table hosted on PharmGKB[46] for each gene of interest. For all 5 genes, we added missing hg19 coordinates to every site, and added rsIDs for all variants in the tables also present in dbSNP. We also encountered several redundancies and ambiguitites that have been fixed in our final, revised definition tables. These included, for example, inconsistent strand orientations, identical allele definitions with different names, and poorly represented tri-allelic sites. Where inconsistencies arose, we relied on the original publication describing each allele as listed on PharmGKB or the *CYP* Allele Nomenclature Database (http://www.cypalleles.ki.se). In order to test whether our definition tables could be used to accurately translate NGS data, we constructed a test VCF for each gene with simulated individual data representing each named star allele. These VCFs served as a ‘positive control’ in

our analysis to ensure that all defined alleles are represented properly in both the allele definition tables and the phased exome data itself.

Naming algorithm

In order to quantify naming errors, we developed an algorithm that attempts to match phased exome data with a defined star allele. The algorithm utilizes a global interpretation of allele definitions; that is, all sites not specified in the allele definition table, but present in the VCF, are assumed to match the hg19 reference base at that position for all named alleles. As our dataset is derived from exome sequencing, variants in the definition tables that fall outside of coding exons could not be assessed. In these cases, we collapsed ambiguous alleles into categories defined only by coding variants. For example, *CYP2C19*1B* and *CYP2C19*17* share coding variants, but are distinguished by noncoding variants lying outside the sequenced region; our algorithm reports any haplotype matching the shared coding variants as "*1B or *17." We tested our algorithm independently for each gene using our test VCFs described above to ensure all named alleles could be accurately detected. After reading in a gene-specific, phased multisample VCF, the algorithm reports the star allele designations for each haplotype, or reports all variants on any haplotype that does not match a named star allele. As phasing of very low frequency variants can be error-prone, the algorithm performs a secondary analysis on the data after flipping the phase of any variant observed in 2 or fewer individuals; this ensures that observed trends in the data are not driven largely by constitutive phasing errors.

3.3 Results

TPMT

TPMT is responsible for catalyzing the S-methylation of thiopurine drugs in addition to a variety of other aromatic compounds[47]. In the almost 30 years since variation in *TPMT* was first associated with mercaptopurine response[48] over 25 different *TPMT* alleles have been identified and assigned a star allele. Several of these alleles have been designated as actionable by CPIC due to their influence on thioguanine, azathiopurine, and mercaptopurine response[11]. Our expanded allele definition table for *TPMT* includes 60 variants: 42 from the original definition table and 18 variants present in the ESP data but absent from the original table. Of the 42 variants from the original definition table, 15 were observed in ESP, 23 were not observed, and 4 fell outside the sequenced region. At the haplotype level, we observed 13 of 42 haplotypes representing named alleles (31%) as well as 33 additional haplotypes that did not match a named allele; 2/3 of these haplotypes were unnamed due to the presence of additional variation not captured by the original definition table (Figure 3.1).

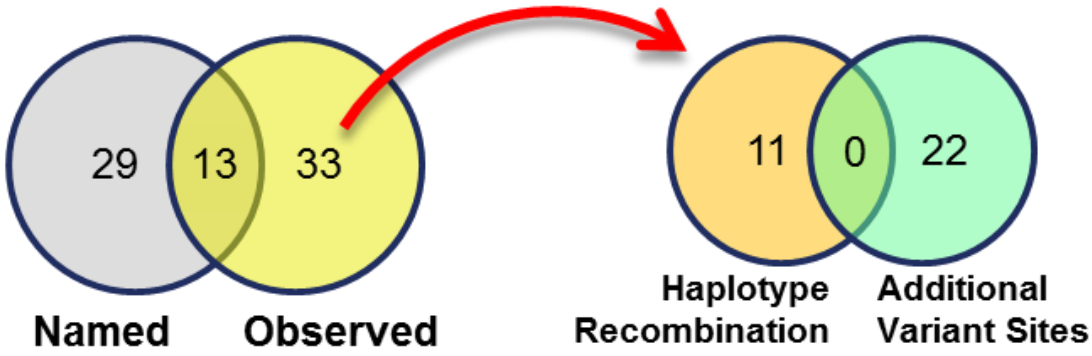

Figure 3.1 *TPMT* haplotypes observed in ESP.

Despite the presence of these unnamed haplotypes, at the sample level we find that *TPMT* variation is well captured by the star system, as only 1.6% of individuals in our dataset carry a *TPMT* allele that could not be named (Table 3.1, 3.2). Accordingly, all unnamed *TPMT* alleles observed in our dataset were quite rare, with frequencies <1% in all cases (Figure 3.2). Overall, our results indicate that *TPMT* variation in our dataset is generally well-captured by the star allele system

Table 3.1 Frequency of named *TPMT* haplotypes within ESP individuals

Allele Calls	% AA	% EA	Overall
Named (*)	98.0%	98.5%	98.4%
Unnamed (?)	2.0%	1.5%	1.6%

Table 3.2 Distribution of named and unnamed *TPMT* haplotypes at the individual level

Diplotypes	% AA	% EA	Overall
* /*	96.1%	97.1%	96.8%
*/?	3.9%	2.9%	3.2%
?/?	0%	0%	0%

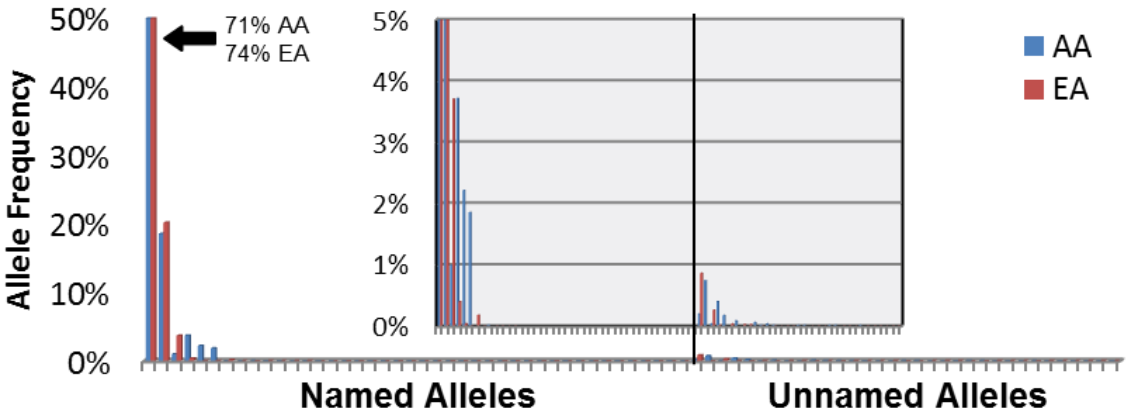

Figure 3.2 Frequency spectrum of named and unnamed *TPMT* haplotypes within ESP.

CYP2C9

A member of the Cytochrome P450 family, *CYP2C9* is responsible for the oxidation of many endogenous and xenobiotic compounds, including roughly 20% of all Phase I metabolized drugs[49]. Accordingly, studies of this gene have uncovered many associations between variation within *CYP2C9* and response to a variety of pharmaceuticals; of these, CPIC has determined that there are actionable associations between *CYP2C9* alleles and response to two common medications, phenytoin[7] and warfarin[8]. The original allele definition table for *CYP2C9* contains 54 variants, only 14 of which were present in ESP individuals; this translates to 58 named *CYP2C9* star alleles, of which only 13 (22%) were present in ESP. However, we also observed an additional 59 variants in our dataset that were not included in the original table, leading to 85 unique haplotypes that could not be assigned a star allele; the majority of these unnamed haplotypes are distinguished by rare variation not present in the definition table (Figure 3.3).

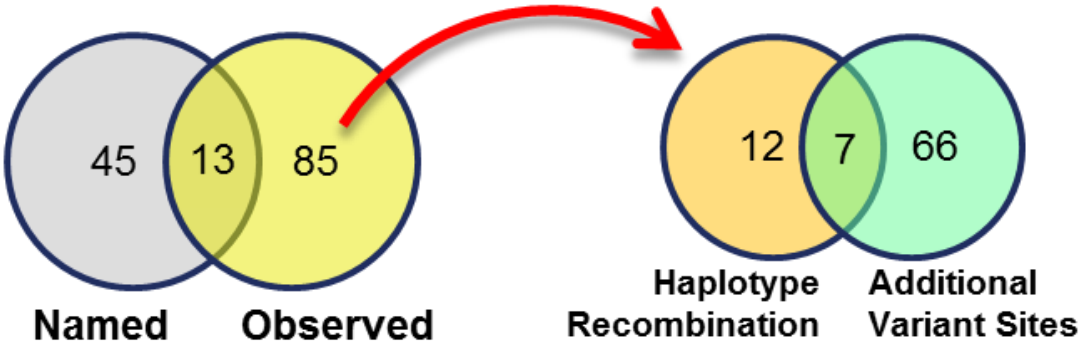

Figure 3.3 *CYP2C9* haplotypes observed in ESP.

At the individual level, we find that 9.5% of haplotypes overall could not be assigned a star allele. This value varies strikingly by ethnicity—13.4% of haplotypes from AA individuals could not be named, as opposed to 7.5% in EA (Table 3.3). This trend persists at the individual level, as 25.3% of AA carry at least one unnamed allele compared to only 14.5% of EA (Table 3.4). This disparity, coupled with the observation that unnamed haplotypes are driven by novel, rare variation, indicates that the original definition table does not accurately capture *CYP2C9* variation in undersequenced populations, particularly in individuals of African-American descent.

Table 3.3 Frequency of named *CYP2C9* haplotypes within ESP individuals

Allele Calls	% AA	% EA	Overall
Named (*)	86.6%	92.5%	90.5%
Unnamed (?)	13.4%	7.5%	9.5%

Table 3.4 Distribution of named and unnamed *CYP2C9* haplotypes at the individual level

Diplotypes	% AA	% EA	Overall
* /*	74.7%	85.4%	81.8%
*/?	23.8%	14.1%	17.4%
?/?	1.5%	0.4%	0.8%

In order to assess the potential clinical impact of these observations, we reprocessed our *CYP2C9* data after masking all genotypes not currently recommended for clinical testing (warfarindosing.org). This new dataset not only mimics the scope of what a clinical lab might test, but allows us to directly quantify misclassification errors by comparing the results of the masked and unmasked data. Across EA and AA, 100% of haplotypes assigned *3 in the masked data did not in fact match the canonical *3 definition in the unmasked data, indicating that the clinical population receiving a test result including a *3 haplotype is likely much more genetically heterogeneous than previously thought. Even more striking, however, is the disparity between EA and AA for alleles labeled *1 (i.e. the reference haplotype) in the masked data. In EA, only 3.4% of *1/*1 diplotypes in the masked data did not in fact match *1/*1 in the unmasked data. In AA, however, we find that 47.7% of *1/*1 diplotypes in the masked data are not in fact true *1/*1 when the data is unmasked (Table 3.5, Table 3.6). This observation is largely driven by many low-frequency AA variants not present in the definition table that, although individually rare, are collectively common.

Table 3.5 and Table 3.6 Misclassifications in *CYP2C9* when considering only clinically typed alleles. Reclassified haplotypes are those that did not match the strict allele definition when the full sequence data was unmasked.

	Correct	Re-classified	% Re-classified
*1/*1	1038	945	47.7%
*1/*2	82	26	24.1%
*1/*3	0	59	100.0%
*1/*5	24	21	46.7%
*2/*2	4	0	0.0%
*2/*3	0	1	100.0%
*2/*5	1	0	0.0%
*3/*3	0	1	100.0%
*3/*5	0	1	100.0%
	1149	1054	47.8%

AA

	Correct	Re-classified	% Re-classified
*1/*1	2692	96	3.4%
*1/*2	863	21	2.4%
*1/*3	0	447	100.0%
*1/*5	1	0	0.0%
*2/*2	74	1	1.3%
*2/*3	0	90	100.0%
*2/*5	0	0	-
*3/*3	0	15	100.0%
*3/*5	0	0	-
	3630	670	15.6%

EA

SLCO1B1

SLCO1B1 is membrane-bound anion transport protein involved in active transport of many diverse xenobiotic compounds, notably the statin family of pharmaceuticals[14]. As *SLCO1B1*-dependant transport is a key component of drug clearance in the liver, variation in *SLCO1B1* that may alter this function can affect drug response and toxicity; CPIC has designated that variation in *SLCO1B1* has an actionable association with simvastatin response[50]. The original allele definition table for *SLCO1B1* contains 31 unique variants that comprise 38 named star alleles. In our dataset, we observed 75 additional variants not present in the definition table, translating into 144 unique haplotypes that could not be assigned a star name. Unlike *TPMT* and *CYP2C9,* this allelic diversity is driven by recombination of known alleles as well as novel variation, often on the same haplotype (Figure 3.4).

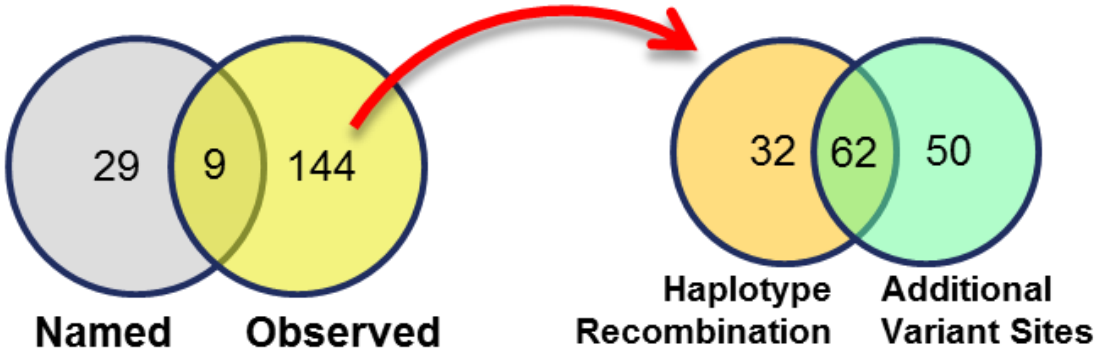

Figure 3.4 *SLCO1B1* haplotypes observed in ESP.

The large quantity of unnamed haplotypes we identified at this locus is further reflected at the individual level, as 83.9% of all ESP individuals possess at least one *SLCO1B1* allele that could not be named (Table 3.7, Table 3.8). Unlike *TPMT,* several of these unnamed alleles are found at intermediate frequency (1% < MAF < 5%) in EA or AA, highlighting the necessity of a more comprehensive nomenclature system for this gene that can accurately represent both new recombination of known haplotypes and novel variation (Figure 3.5).

Table 3.7 Frequency of named *SLCO1B1* haplotypes within ESP individuals

Allele Calls	% AA	% EA	Overall
Named (*)	61.2%	21.8%	35.2%
Unnamed (?)	38.8%	78.2%	64.8%

Table 3.8 Distribution of named and unnamed *SLCO1B1* haplotypes at the individual level

Diplotypes	% AA	% EA	Overall
* /*	38.4%	4.7%	16.1%
*/?	45.8%	34.1%	38.1%
?/?	15.9%	61.1%	45.8%

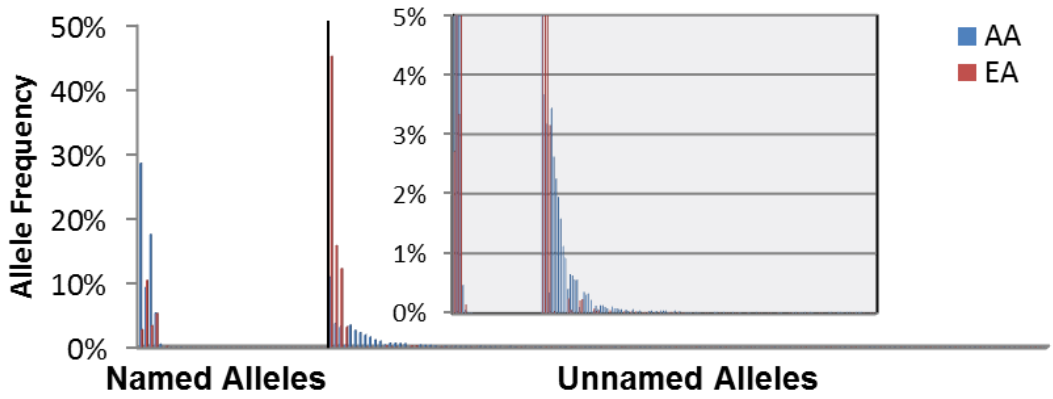

Figure 3.5 Frequency spectrum of named and unnamed *SLCO1B1* haplotypes within ESP

CYP3A5

CYP3A5, another key metabolizer of pharmaceuticals, is the most abundant Cytochrome P450 enzyme expressed in human liver and involved in the elimination of 37% of the 200 most commonly prescribed drugs in the U.S.[51]. Although a number of drug response phenotypes have been associated with genetic variation in *CYP3A5*, response to tacrolimus is considered currently to be the most clinically actionable association, and preemptive genetic testing for this gene-drug pair is becoming increasingly common[26]. The original allele definition table for this gene contains 22 variants that make up 25 unique named star alleles. In our dataset, we observed 6 of these 22 variants (27%) in addition to 52 other variants not present in the table; collectively these variants are arranged into 74 different haplotypes, only 6 of which (8%) could be assigned a star allele. Despite this diversity, these 6 haplotypes are very common among ESP individuals, representing 95.9% of all observed haplotypes (Table 3.9). However, like *CYP2C9*, there is a notable disparity between ethnicities in the accuracy of the star system at the individual level: 19.4% of African-Americans carry at least one allele that could not be named, compared to only 2% of European Americans (Table 3.10).

Table 3.9 Frequency of named *CYP3A5* haplotypes within ESP individuals

Allele Calls	% AA	% EA	Overall
Named (*)	90.0%	99.0%	95.9%
Unnamed (?)	10.0%	1.0%	4.1%

Table 3.10 Distribution of named and unnamed *CYP3A5* haplotypes at the individual level

Diplotypes	% AA	% EA	Overall
* /*	80.6%	97.9%	92.0%
*/?	18.8%	2.0%	7.7%
?/?	0.6%	0.0%	0.2%

Although this disparity is striking, *CYP3A5* also highlights some inherent limitations in using exome data for clinical PGx testing. *CYP3A5**3 is defined by rs776746, an intronic variant with a robust, well-known association with tacrolimus response due to the creation of a cryptic splice site that effectively inactivates the gene[52]. This variant lies outside the sequenced region of most currently available exome capture reagents, and cannot be typed by this method. Thus, exome data is unable to distinguish between 10 common subtypes of *1 and *3; the resulting ambiguous allele designation collectively represents 88% of all ESP haplotypes that could be named. The ambiguity inherent in using exome data for this target highlights the need for accurate clinical reporting of the method used to generate any PGx result; not only which regions were included, but also any critical allele that could not be tested.

CYP2C19

Also a member of the Cytochrome P450 family, *CYP2C19* oxidizes 10% of the 200 most commonly prescribed drugs in the U.S.[51], and variation within *CYP2C19* has been determined by CPIC to be actionably associated with both tricyclic antidepressants[10] and clopidogrel[53]. The original allele definition table for *CYP2C19* contains 63 named star alleles comprised of 64 unique variants. At the individual level, *CYP2C19* resembles *CYP2C9*: although only 3.3% of haplotypes overall could not be named, 11.8% of African-Americans carried at least one unnamed allele as opposed to only 3.6% of European-Americans (Table 3.11, Table 3.12). Like *CYP3A5*, exome data is inherently limited for this gene as it cannot differentiate between *1B and *17. Defined by two noncoding variants, *CYP2C19**17 is a gain-of-function allele leading to ultrarapid metabolism of *CYP2C19* substrates in homozygotes[54]. Thus, clinical reports of

CYP2C19 diplotype derived from exome sequence data should indicate that *17 was not assessed by the platform.

Table 3.11 Frequency of named *CYP2C19* haplotypes within ESP individuals

Allele Calls	% AA	% EA	Overall
Named (*)	94.0%	98.1%	96.7%
Unnamed (?)	6.0%	1.9%	3.3%

Table 3.12 Distribution of named and unnamed *CYP2C19* haplotypes at the individual level

Diplotypes	% AA	% EA	Overall
* /*	88.2%	96.4%	93.6%
*/?	11.6%	3.5%	6.3%
?/?	0.2%	0.1%	0.2%

3.4 Discussion

As the cost of next-generation sequencing continues to decline, exome, genome, and custom-target sequencing will quickly become the standard for clinical pharmacogenetic testing. Despite the rapid shift in the underlying technology used to generate these results, the classical star allele nomenclature system continues to be the standard for clinical reporting. The results presented here summarize the first attempt, to our knowledge, to evaluate the performance of the star allele nomenclature system as applied to NGS data. On a broad scale, our results indicate that the current star nomenclature system fails to capture a large extent of the variation present in the 5 critical pharmacogenes considered here. Our results are likely an underestimate of the true extent of misclassification due to the presence of intronic and noncoding variation inaccessible by exome sequencing. Therefore, whole-genome sequencing will almost certainly compound the issues presented here, especially for large genes primarily comprised of intronic space such as

SLCO1B1: 98% of *SLCO1B1*'s 108,603 bases are noncoding. Even considering only exonic variation, we observed a significant disparity between European-Americans and African-Americans in the star allele system's ability to classify haplotypes. Although many of the variants underlying these unnamed haplotypes are seen only in a handful of individuals, collectively these variants are quite common, especially in under-sequenced populations where much of the lower frequency variation remains yet to be discovered. As NGS-based testing expands into these populations, the star allele system, in which new alleles are often designated only by committee, will become increasingly inadequate for reporting PGx results within a clinically relevant timeframe. As clinical labs transition to NGS-based platforms, our findings argue that the currently adopted nomenclature system should transition as well.

Chapter 4: PGRNseq, a targeted capture sequencing panel for pharmacogenetic research and implementation

4.1 Abstract

While the costs associated with whole-genome and whole-exome next-generation sequencing continue to decline, they remain prohibitively expensive for large-scale studies of genetic variation. As an alternative, custom-target sequencing has become a common methodology based on its favorable balance between cost, throughput, and deep coverage. We have developed PGRNseq, a custom-capture panel of 84 genes with associations to pharmacogenetic phenotypes, as a tool to explore the relationship between drug response and genetic variation, both common and rare. We utilized a set of 32 diverse HapMap trios and 2 clinical cohorts to assess platform performance, accuracy, and ability to discover novel variation. We find that PGRNseq generates ultra-deep coverage data (mean = 496x) that is over 99.8% concordant with orthogonal datasets. Additionally, in our testing sets, PGRNseq identified many novel, rare variants of interest, underscoring its utility in both research and clinical settings.

4.2 Introduction

As next-generation sequencing costs continue to decrease, and rare variant analysis becomes an imperative, sequencing-based association analysis is developing as widely applied tool in human genetic analysis through whole exome and whole genome sequencing as well as the application of targeted sequencing panels. Indeed, these approaches have been successful in identifying novel associations between genetic variation and a range of traits including cardiovascular, psychiatric, and pharmacogenetic phenotypes[55-57]. However, sequencing full genomes or even full exomes of the tens of thousands of individuals needed for adequately powered association studies remains costly and time-consuming. Targeted high-throughput sequencing panels, which capture and sequence a small set of genomic targets to high depth, represent a middle-ground that maximizes throughput while maintaining the deep coverage characteristic of high-quality next generation sequencing (NGS) data. To date, targeted sequencing panels have been successfully deployed in both clinical research and diagnostics with applications as diverse as the mutational analysis of individuals with Lynch or polyposis syndrome[58], the detection of somatic mutation in lung cancer[59], and the molecular diagnosis of retinitis pigmentosa[60]. In this article we will discuss the process of creating and validating such a panel, focusing on 84 genes of pharmacogenetic importance, including many genes identified as actionable by the Clinical Pharmacogenetics Implementation Consortium (CPIC)[61].

Since its inception, pharmacogenetic research has identified many genes that play a role in drug response, and has shown that many variants within these genes contribute to overall variation in drug phenotypes[5]. These gene-phenotype pairs include drug-metabolizing enzymes (such as *CYP2C19* and clopidogrel response[53]), drug transporters (such as *UGT1A1* and irinotecan[62]) and specific drug targets (such as *VKORC1* and warfarin[8]). Many of these drug-

gene pairs are clinically actionable, and the number of clinical entities performing pharmacogenetic testing is increasing steadily. Despite this increasing popularity, most of the known variation within these pharmacogenes is common (i.e. MAF > 1%). Indeed, many existing platforms are currently used to genotype these targets in a high-throughput manner, such as the Affymetrix DMET+ array and the Illumina ADME assay which focus largely on common variation. However, initial large-scale, NGS-based studies have revealed that rare (i.e. MAF < 1%) deleterious variation is in fact collectively common across drug metabolizing enzyme and drug targets; though each individual variant can be vanishingly rare. In fact, 7-10% of individuals harbor such a variant (See Chapter 2). Additionally, rare variation in pharmacogenes has been directly linked to variation in drug response and to rare adverse events in the several cases that have been studied extensively to date[63,64]. Thus, it's clear that this category of variation is of importance as pharmacogenetic testing expands clinically. This necessitates collaborative efforts on the analysis of rare variation in pharmacogenes.

The Pharmacogenomics Research Network (PGRN) is a collaborative network formed in order to coordinate pharmacogenetic research and to collectively provide recommendations as to the clinical relevance of pharmacogenetic variation. Seeing an opportunity to facilitate large-scale sequencing studies of pharmacogenetic targets to assess both rare and common variation, as well as an opportunity to explore the clinical utility of NGS, the PGRN called on the network's 3 Deep Sequencing Resources (DSRs: Department of Genome Sciences, University of Washington (UW); The Genome Institute at Washington University (WashU); and the Human Genome Sequencing Center at Baylor College of Medicine (BCM-HGSC)) to develop a custom-capture panel centered on pharmacogenes of known interest. Here we present an overview of the design,

testing, and quality control of PGRNseq; the resulting panel is currently available for use by members of the pharmacogenetics community.

4.3 Results

General Platform Performance

Using a 24-plex capture strategy leads to an average coverage of 496X across the target space, demonstrating that PGRNseq can consistently generate ultra-deep sequencing data while maintaining the high throughput necessary for studies of large sample size (Table 4.1).

Table 4.1: Overall PGRNseq performance (HapMap96). PGRNseq summary statistics drawn from the BCM HapMap96 data.

Plex Level	Avg. # Reads (M)	Avg. Unique Aligned Gb	Avg. Mean Quality Score	Avg. % Q30 Bases	Avg. % Targets Hit	Avg. Coverage	Mean % target at > 20x	Mean % targets at > 40x
24	16.7	1.37	36.7	92.1	94.7	496x	94.8	93.4

At the single gene level, PGRNseq generates deep coverage data for the complete coding region for nearly every captured gene (Table 4.2). The major exceptions are the two MHC genes on the platform, *HLA-B* and *HLA-DQB3* despite the inclusion of all 8 alternative reference haplotypes in the design phase. As these genes are highly structurally polymorphic, they present a considerable challenge to assemble using short reads[65]. Therefore, SNV calls in this region were not considered further. Other areas of low coverage consist largely of noncoding regions distant from the coding regions, and in most cases these low coverage regions were also related to the presence of repetitive elements within the 2kb/1kb upstream/downstream design window.

Table 4.2: 84 Pharmacogenes of interest captured by PGRNseq and their overall performance. These genes were nominated and voted on by the PGRN community for inclusion in the final target. "Coding Plus" length indicates the number of base pairs that make up the gene's exons as well as 2kb upstream and 1kb downstream of the coding region. Function/Role annotations derived from PharmGKB. Per-gene coverage drawn from UW data.

Gene Symbol	Chromosome	Coding Plus length (bp)	Gene Function/Role	Mean HapMap96 coverage	Mean HapMap96 coverage (coding only)	% Coding bases >30X
ABCA1	9	9982	Target	350.24	487.28	100
ABCB1	7	7480	Absorption	286.73	397.34	100
ABCB11	2	7074	Absorption	335.36	415.32	100
ABCC2	10	7766	Absorption	355.52	457.97	100
ABCG1	21	5265	Absorption	392.68	566.58	98
ABCG2	4	5028	Absorption	250.18	372.38	100
ACE	17	7224	Target	378.92	521.51	96
ADRB1	10	4438	Target	252.14	275.46	79
ADRB2	5	4246	Target	356.08	538.5	100
AHR	7	5591	Metabolism	248.22	328.84	100
ALOX5	10	5081	Target	326.2	495.21	100
APOA1	11	3816	Target	370.24	417.82	100
ARID5B	10	6607	Disease	321.12	419.56	100
BDNF	11	3804	Target	311.84	460.9	100
CACNA1C	12	9936	Target	380.02	564.67	100
CACNA1S	1	8622	Target	439.26	686.3	100
CACNB2	10	5349	Target	271.8	384.62	100
CES1	16	4763	Metabolism	350.84	490.24	77
CES2	16	4920	Metabolism	346.27	567.4	100
COMT	22	3939	Metabolism	361	604.82	100
CRHR1	17	4300	Target	302.18	596.33	100
CYP1A2	15	4575	Metabolism	286.58	690.18	100
CYP2A6	19	6052	Metabolism	339.84	529.18	100
CYP2B6	19	4512	Metabolism	331.18	482.97	100
CYP2C19	10	5648	Metabolism	332.72	475.06	100
CYP2C9	10	4509	Metabolism	315.8	450.6	100
CYP2D6	22	4530	Metabolism	327.6	440.5	97
CYP2R1	11	4524	Metabolism	320.54	305.6	100
CYP3A4	7	4564	Metabolism	361.7	427.83	100
CYP3A5	7	4561	Metabolism	310.1	421.32	100
DBH	9	4854	Target	429.54	564.72	100
DPYD	1	6170	Excretion	292.67	350.6	100

DRD1	5	4345	Target	304.08	504.8	100
DRD2	11	4360	Target	390.34	618.09	100
EGFR	7	7009	Target	356.74	500.2	98
ESR1	6	5065	Target	318.08	415.87	99
FKBP5	6	4414	Target	315.66	431.94	100
G6PD	X	4690	Drug-induced Disease	269.4	415.12	96
GLCCI1	7	4676	Drug-induced Disease	289.12	285.14	76
GRK4	4	4801	Target	340.84	435.02	100
GRK5	10	4837	Target	337.76	578.7	100
HLA-B	6	3000	Toxicity	98.64	101.3	13
HLA-DQB3	6	3000	Toxicity	123.6	131.7	22
HMGCR	5	5743	Target	295.04	367.56	100
HSD11B2	16	4238	Metabolism	358.96	365.08	78
HTR1A	5	4273	Target	284.2	350.8	100
HTR2A	13	4428	Drug-induced Disease	333.02	430.06	100
KCNH2	7	6921	Drug-induced Disease	306.64	399.5	87
LDLR	19	5655	Target	318.74	596.7	100
MAOA	X	4644	Target	239.4	296.2	100
NAT2	8	3877	Metabolism/Excretion	286	333.16	100
NPPB	1	3417	Drug-induced Disease	348.22	443.64	100
NPR1	1	6186	Target	336.04	502.86	98
NR3C1	5	5421	Target	301.53	377.88	100
NR3C2	4	5955	Target	294.06	408.14	100
NTRK2	9	5664	Target	330.66	453.94	100
PEAR1	1	6202	Target	337.84	515.1	100
POR	7	5103	Drug-induced Disease	343.6	514.3	100
PTGIS	20	4543	Target	381.64	575.76	97
PTGS1	9	4844	Target	393.76	603.78	100
RYR1	19	18541	Drug-induced Disease	392.65	522.32	97
RYR2	1	18324	Drug-induced Disease	320.3	383.2	100
SCN5A	3	9255	Drug-induced Disease	400.44	579.4	100
SLC15A2	3	5278	Excretion	297.8	408.44	100
SLC22A1	6	4709	Excretion	337.75	551.72	100
SLC22A2	6	4712	Excretion	351.12	434.84	100
SLC22A3	6	4715	Excretion	304.1	331.77	88
SLC22A6	11	4732	Absorption	332.01	554.4	100
SLC47A1	17	4781	Absorption	334.68	538.28	100
SLC47A2	17	4877	Absorption	402.04	513.26	100
SLC6A3	5	4919	Target	371.69	665.48	100
SLC6A4	17	4945	Disease	313.64	566.32	100
SLCO1A2	12	5476	Absorption	265.88	299.14	100
SLCO1B1	12	5132	Absorption	260.12	290.98	100
SLCO1B3	12	5165	Absorption	249.76	286.68	100

SLCO2B1	11	5186	Absorption	359.58	586.51	100
TBXAS1	7	4657	Metabolism	351.4	470.22	100
TCL1A	14	3357	Disease	379.86	463.1	100
TPMT	6	3770	Metabolism	251.13	335.64	100
UGT1A1	2	4622	Excretion	322.58	289.92	98
UGT1A4	2	6806	Excretion	352.58	501.66	100
VDR	12	4316	Absorption	391.52	611.44	100
VKORC1	16	3504	Target	308.46	580.16	100
ZNF423	16	6887	Target	405.36	626.15	100

Quality and accuracy of PGRNseq variants

Although PGRNseq generates high-coverage data many targeted genes could be prone to erroneous variant calls due to sequence homology with other gene family members and the associated inappropriate capture and/or sequence read mismapping, or due to the presence of structural variants (SVs). Therefore, we assessed accuracy of PGRNseq genotype calling using orthogonal datasets as well as consistency with Mendelian inheritance. Across the HapMap96, we observed an average of 1325 total variants called per individual. See Table 4.3 for counts of variants per gene separated by variant type. Analysis of Mendelian inconsistencies within the 32 trios revealed that the majority of genes (63/82, 77%) did not contain any such errors (Table 4.3). Of the 19 genes that did, 17 contained 3 or fewer Mendelian errors, all of which were in noncoding regions at the edges of the target space. However, two genes, *CYP2A6* and *CYP2D6*, contained 10 or more Mendelian inconsistencies. These results were not unexpected as both genes have one or more neighboring pseudogenes with high homology, and both are known to harbor structural variants of functional consequence[66,67]. Indeed, many of these inconsistencies were found across multiple trios and located in regions of low unique mappability (Figure 4.1).

Table 4.3 Single Nucleotide Variants observed in PGRNseq genes across HapMap96 individuals. Annotations derived from SeattleSeq.

Gene Symbol	Chr	mendel_errors	Total_variant	dbSNP	novel	Novel missense	Novel nonsense/splice	Novel synonymous	Novel UTR	Novel intron	Novel near_gene	Known missense	Known nonsense/splice	Known synonymous	Known UTR	Known intron	Known near_gene
ABCA1	9	0	76	70	6	1	0	2	2	0	1	8	0	20	14	5	23
ABCB1	7	0	52	47	5	0	0	0	0	1	4	7	0	4	8	16	12
ABCB11	2	0	70	65	5	0	0	0	1	0	4	5	0	10	4	21	25
ABCC2	10	0	41	40	1	0	0	0	1	0	0	12	0	8	4	5	11
ABCG1	21	0	70	68	2	0	0	0	0	1	1	0	0	4	8	28	28
ABCG2	4	0	32	31	1	1	0	0	0	0	0	3	0	2	3	1	22
ACE	17	0	67	57	10	1	0	2	2	0	5	9	0	13	4	3	28
ADRB1	10	0	24	21	3	0	0	0	1	0	2	3	0	2	3	13	0
ADRB2	5	0	28	27	1	0	0	0	0	0	1	2	0	5	4	0	16
AHR	7	0	32	25	7	0	0	0	4	1	2	3	0	1	9	2	10
ALOX5	10	0	34	31	3	0	0	0	1	0	2	1	0	5	6	8	11
APOA1	11	0	23	20	3	0	0	0	0	0	3	0	0	1	0	1	18
ARID5B	10	1	31	26	5	1	0	0	2	0	2	2	0	3	9	0	12
BDNF	11	1	31	30	1	0	0	0	1	0	0	1	0	1	13	11	4
CACNA1C	12	1	68	59	9	2	0	4	3	0	0	4	0	12	18	22	3
CACNA1S	1	0	50	46	4	2	0	0	1	1	0	9	0	18	1	2	16
CACNB2	10	2	58	46	12	0	0	0	5	4	3	2	0	5	9	11	19
CES1	16	3	42	40	2	0	0	0	0	1	1	11	1	5	4	2	17
CES2	16	0	16	16	0	0	0	0	0	0	0	0	0	2	4	7	3
COMT	22	0	50	48	2	1	0	0	0	0	1	2	0	6	8	12	20
CRHR1	17	1	42	34	8	0	0	1	2	0	5	1	0	4	14	4	11
CYP1A2	15	0	20	16	4	1	0	1	2	0	0	4	0	1	2	4	5
CYP2A6	19	10	48	44	4	1	0	0	1	0	2	5	0	11	6	8	14
CYP2B6	19	3	55	52	3	0	0	0	2	0	1	8	1	5	16	7	15
CYP2C19	10	0	48	40	8	0	0	1	0	0	7	4	1	4	0	6	25
CYP2C9	10	0	47	38	9	3	0	0	0	2	4	5	0	3	4	7	19
CYP2D6	22	36	50	42	8	0	0	0	0	4	4	7	1	7	0	15	12
CYP2R1	11	0	15	9	6	0	0	0	0	2	4	0	0	2	0	3	4
CYP3A4	7	0	24	21	3	1	0	0	0	1	1	1	0	1	6	5	8
CYP3A5	7	0	44	40	4	0	0	0	0	3	1	2	0	2	3	17	16

DBH	9	0	48	43	5	0	0	0	1	0	4	8	0	7	5	0	23
DPYD	1	0	41	35	6	2	0	0	0	3	1	7	0	3	5	11	9
DRD1	5	0	27	21	6	0	0	0	2	0	4	0	0	2	8	0	11
DRD2	11	0	39	35	4	0	0	0	0	1	3	2	0	8	8	3	14
EGFR	7	0	83	77	6	0	0	1	1	1	3	2	0	11	3	18	43
ESR1	6	0	84	77	7	1	0	0	1	2	3	0	0	5	21	15	36
FKBP5	6	0	53	48	5	0	0	0	1	3	1	0	0	2	8	27	11
G6PD	X	0	23	18	5	0	0	2	1	0	2	2	0	3	1	3	9
GLCCI1	7	0	54	42	12	0	0	0	8	1	3	0	0	2	6	8	26
GRK4	4	0	40	32	8	0	0	0	1	7	0	7	0	4	4	12	5
GRK5	10	0	33	29	4	0	0	0	0	1	3	5	0	1	1	7	15
HMGCR	5	1	42	38	4	0	0	0	2	0	2	0	0	2	10	9	17
HSD11B2	16	0	24	22	2	0	0	0	1	0	1	1	0	4	6	0	11
HTR1A	5	0	18	14	4	0	0	0	0	0	4	2	0	1	0	0	11
HTR2A	13	0	54	46	8	1	0	0	2	0	5	3	0	3	12	0	28
KCNH2	7	0	44	37	7	0	0	0	0	2	5	2	0	5	2	9	19
LDLR	19	0	50	46	4	0	0	0	3	0	1	2	0	11	19	0	14
MAOA	X	0	15	10	5	0	0	0	1	0	4	0	0	3	2	1	4
NAT2	8	0	28	25	3	0	0	0	0	0	3	6	0	3	0	0	16
NPPB	1	0	38	29	9	0	0	2	3	0	4	1	0	2	0	0	26
NPR1	1	0	21	19	2	0	0	0	2	0	0	5	0	3	1	0	10
NR3C1	5	0	43	39	4	0	0	1	1	1	1	3	0	7	10	5	14
NR3C2	4	0	40	31	9	1	0	0	4	0	4	2	0	5	10	2	12
NTRK2	9	0	98	87	11	0	0	0	2	8	1	2	0	2	8	54	21
PEAR1	1	2	62	55	7	2	0	0	0	2	3	10	0	10	9	3	23
POR	7	1	60	48	12	3	0	1	1	3	4	3	0	6	7	22	10
PTGIS	20	0	30	29	1	0	0	0	1	0	0	1	0	8	11	0	9
PTGS1	9	0	63	59	4	0	0	0	2	0	2	6	0	5	18	4	26
RYR1	19	0	80	72	8	3	1	1	0	0	3	12	0	49	1	0	10
RYR2	1	0	65	54	11	3	0	3	1	1	3	7	0	19	5	11	12
SCN5A	3	0	75	72	3	0	0	0	0	0	3	9	0	14	17	12	20
SLC15A2	3	0	48	42	6	0	0	1	1	1	3	3	0	2	14	6	17
SLC22A1	6	0	34	30	4	1	0	0	1	0	2	9	0	5	0	6	10
SLC22A2	6	0	77	69	8	0	0	1	0	0	7	3	0	2	8	9	47
SLC22A3	6	0	47	38	9	0	0	0	6	0	3	1	0	4	18	0	15
SLC22A6	11	0	17	14	3	0	0	0	1	0	2	1	0	3	4	0	6
SLC47A1	17	0	28	23	5	1	0	0	2	0	2	2	0	4	5	3	9
SLC47A2	17	0	69	58	11	3	0	0	1	1	6	1	0	4	1	10	42
SLC6A3	5	1	49	43	6	0	0	1	2	0	3	0	0	6	19	1	17

SLC6A4	17	0	42	36	6	0	0	1	2	0	3	2	0	2	10	0	22
SLCO1A2	12	0	101	85	16	0	0	0	7	4	5	4	0	5	45	23	8
SLCO1B1	12	1	36	35	1	1	0	0	0	0	0	8	0	7	6	0	14
SLCO1B3	12	2	49	43	6	0	0	1	1	1	3	5	0	5	0	5	28
SLCO2B1	11	0	71	55	16	0	0	0	1	9	6	4	0	6	8	32	5
TBXAS1	7	0	54	47	7	0	0	1	2	2	2	10	0	4	3	23	7
TCL1A	14	0	43	38	5	0	0	1	1	0	3	2	0	0	5	5	26
TPMT	6	1	36	28	8	0	0	0	3	0	5	3	0	1	7	1	16
UGT1A1/4	2	1	120	111	9	0	0	0	0	9	0	2	0	0	8	85	16
VDR	12	1	58	53	5	1	0	0	1	0	3	1	0	3	20	5	24
VKORC1	16	0	26	25	1	0	0	1	0	0	0	1	0	0	1	9	14
ZNF423	16	1	34	26	8	0	0	3	4	0	1	4	0	11	0	1	10
TOTALS		70	3802	3337	465	38	1	33	109	84	200	297	4	451	584	706	1295

Figure 4.1 Mendelian inconsistency deriving from mis-mapped reads. A child with a T/T genotype is not consistent with a C/T father and a C/C mother. Reads containing a 'C' at this site likely derive from the pseudogenic *CYP2D7*, which, when aligned to *CYP2D6*, has a C at this site and the site 2bp downstream.

In addition to the quality checks inherent in the use of trio data, we chose a panel of HapMap samples in order to compare our results to those from other large sequencing or genotyping efforts, e.g. 1000 Genomes. To evaluate accuracy, we calculated the mean per-individual genotype concordance at various coverage cutoffs using 3 different datasets: HapMap 3.3 (n=96)[68], 1000 Genomes deeply sequenced trios (n=6)[69], and high-coverage exome data generated at UW through the NIEHS Environmental Genome Project (n=54)[70]. Generally, mean per-individual concordance was greater than 99% (Table 4.4). We noticed that the mean of 99.4% concordance with HapMap 3.3 was consistent across different depth cutoffs; and on closer analysis, we found that 2 noncoding sites were solely responsible for these discrepancies. We believe these sites may be difficult to type using chip-based genotyping, as genotypes from the three different sequencing datasets agree with the PGRNseq genotype at this site. In fact, the two sequencing-based comparison datasets (1000 Genomes deep trios and EGP exomes) showed mean per-individual concordance greater than 99.8%, indicating that the vast majority of PGRNseq-derived genotypes are accurate.

Table 4.4: PGRNseq per-individual concordance vs. orthogonal datasets. Sample sizes indicate number of overlapping samples between datasets. Concordance calculated for variant sites only with coverage at or above the thresholds listed in the column headers. Table values in parentheses indicate mean number of overlapping variants per individual from which the final percentage was derived.

Dataset	Concordance, coverage ≥10x (mean # overlap)	Concordance, coverage ≥20x (mean # overlap)	Concordance, coverage ≥30x (mean # overlap)	Concordance, coverage ≥50x (mean # overlap)
YRI Deep Trio (n=3)	99.9% (650)	99.9% (547)	100% (354)	100% (182)
CEU Deep Trio (n=3)	99.8% (554)	100% (497)	100% (337)	100% (137)
HapMap 3.3 (n=96)	99.4% (296)	99.4% (296)	99.4% (296)	99.4% (296)
EGP exomes (n=54)	100% (147)	100% (138)	100% (127)	100% (107)

We observed a similarly high accuracy when PGRNseq genotypes were compared to those derived from the Pharmacogenetics-specific genotyping platforms, Affy DMET+ and Illumina ADME. Specifically, the average per-individual concordance was 99.7% between PGRNseq genotypes and Affy DMET+ data (Table 4.5). When comparing the Illumina ADME data, 87 of the HapMap96 were 100% concordant across all overlapping sites. The remaining 9 individuals were all discordant for a single shared variant; upon closer examination, this variant (rs9282861) appears to be poorly genotyped by the Illumina ADME platform due to inaccurate cluster calling (Figure 4.2), and PGRNseq genotypes at this site are concordant with genotypes for these same individuals derived from HapMap 3.3 and from the other sequencing-based comparison sets. Based on these data, sequencing-based PGRNseq genotypes are exceedingly accurate.

Table 4.5 Per-individual concordance between PGRNseq and Affy DMET+ genotypes within the antiplatelet clinical testing cohort. Affy DMET+ data and PGRNseq data generated at BCM (N=96).

sample	DMET+ genotyes	count shared	number concordant	number discordant	concordance
P0011	1837	1804	1803	1	99.94457
P0012	1839	1808	1803	5	99.72345
P0020	1832	1800	1795	5	99.72222
P0024	1831	1800	1797	3	99.83333
P0030	1845	1810	1805	5	99.72376
P0033	1842	1810	1802	8	99.55801
P0035	1848	1818	1814	4	99.77998
P0036	1819	1790	1784	6	99.6648
P0037	1833	1801	1797	4	99.7779
P0038	1814	1782	1778	4	99.77553
P0041	1840	1809	1804	5	99.7236
P0043	1832	1803	1802	1	99.94454
P0074	1822	1793	1786	7	99.60959
P0075	1845	1815	1810	5	99.72452
P0079	1841	1810	1805	5	99.72376

P0080	1838	1803	1800	3	99.83361
P0081	1846	1813	1808	5	99.72421
P0096	1831	1799	1794	5	99.72207
P0111	1846	1815	1813	2	99.88981
P0126	1819	1790	1783	7	99.60894
P0127	1845	1814	1811	3	99.83462
P0132	1839	1809	1805	4	99.77888
P0133	1843	1813	1808	5	99.72421
P0134	1844	1813	1806	7	99.6139
P0138	1840	1809	1806	3	99.83416
P0144	1836	1805	1801	4	99.77839
P0147	1840	1809	1801	8	99.55777
P0148	1835	1805	1801	4	99.77839
P0149	1840	1809	1802	7	99.61305
P0150	1835	1802	1796	6	99.66704
P0174	1836	1803	1797	6	99.66722
P0179	1848	1818	1809	9	99.50495
P0180	1841	1812	1806	6	99.66887
P0181	1847	1817	1812	5	99.72482
P0183	1847	1817	1812	5	99.72482
P0193	1846	1813	1806	7	99.6139
P0198	1816	1786	1779	7	99.60806
P0200	1838	1808	1801	7	99.61283
P0201	1837	1808	1803	5	99.72345
P0204	1846	1817	1813	4	99.77986
P0205	1842	1811	1807	4	99.77913
P0209	1843	1809	1803	6	99.66833
P0213	1844	1814	1804	10	99.44873
P0221	1822	1794	1791	3	99.83278
P0273	1821	1791	1784	7	99.60916
P0279	1824	1795	1789	6	99.66574
P0288	1836	1805	1800	5	99.72299
P0289	1845	1814	1809	5	99.72437
P0294	1845	1814	1807	7	99.61411
P0295	1847	1817	1809	8	99.55971
P0302	1846	1816	1810	6	99.6696
P0305	1832	1799	1794	5	99.72207
P0308	1815	1783	1778	5	99.71957
P0315	1825	1796	1792	4	99.77728
P0340	1846	1813	1809	4	99.77937
P0341	1837	1806	1800	6	99.66777

P0373	1820	1790	1785	5	99.72067
P0374	1848	1817	1809	8	99.55971
P0375	1835	1804	1799	5	99.72284
P0378	1844	1815	1804	11	99.39394
P0379	1842	1812	1809	3	99.83444
P0385	1845	1815	1809	6	99.66942
P0386	1839	1806	1803	3	99.83389
P0387	1839	1807	1801	6	99.66796
P0388	1849	1819	1813	6	99.67015
P0389	1846	1815	1808	7	99.61433
P0390	1844	1814	1813	1	99.94487
P0426	1843	1811	1806	5	99.72391
P0434	1843	1810	1803	7	99.61326
P0441	1835	1803	1798	5	99.72268
P0442	1839	1810	1804	6	99.66851
P0482	1840	1807	1803	4	99.77864
P0494	1848	1818	1814	4	99.77998
P0495	1846	1814	1808	6	99.66924
P0509	1845	1810	1808	2	99.8895
P0513	1846	1811	1807	4	99.77913
P0514	1843	1812	1807	5	99.72406
P0540	1843	1811	1806	5	99.72391
P0541	1836	1805	1799	6	99.66759
P0570	1842	1811	1806	5	99.72391
P0571	1843	1810	1806	4	99.77901
P0581	1846	1815	1810	5	99.72452
P0582	1839	1808	1803	5	99.72345
P0584	1842	1810	1804	6	99.66851
P0619	1846	1815	1811	4	99.77961
P0622	1850	1819	1815	4	99.7801
P0641	1837	1806	1799	7	99.6124
P0670	1838	1806	1803	3	99.83389
P0674	1830	1801	1796	5	99.72238
X0027	1842	1811	1805	6	99.66869
X0534	1833	1800	1795	5	99.72222
X0585	1833	1799	1792	7	99.61089
X0638	1823	1790	1787	3	99.8324
X0651	1819	1792	1787	5	99.72098
X0722	1840	1808	1804	4	99.77876
mean	1838.211	1807.095	1801.926	5.168421	99.71403

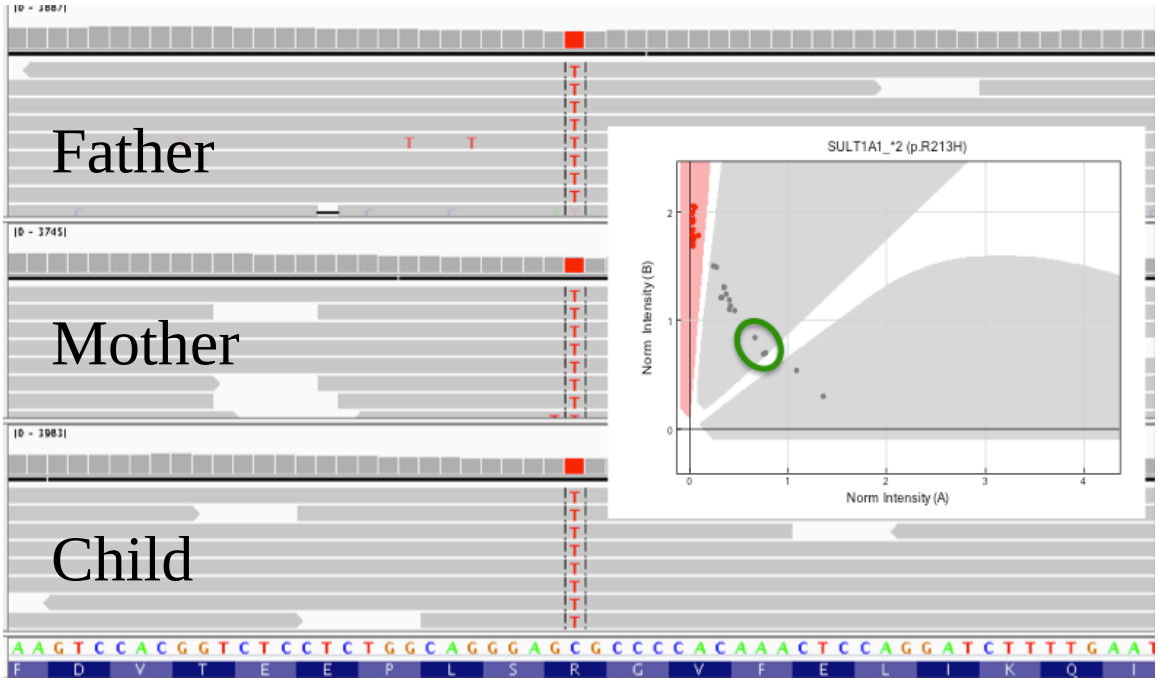

Figure 4.2 Genotyping error in Illumina ADME data due to incorrect cluster definitions. SULT1A1*2 is correctly genotyped by PGRNseq as homozygous non-reference in this trio, but all 3 are called as heterozygotes by Illumina ADME. Inset shows the raw Illumina ADME clustering data for this site, trio circled in green. Clustering data reveals that the incorrect genotypes derive from poor cluster boundary placement.

Novel variation in the HapMap96 and clinical cohorts

Across 82 genes on the panel (excluding the MHC genes), we identified an average of 45 variants per HapMap96 individual that were not present in dbSNP build 137. This value is similar to data from the liver cohort (mean = 35 novel variants per ind) and the antiplatelet cohort (mean = 55 novel variants per ind), which consisted largely of Caucasian individuals, and were less diverse than the trios in the HapMap96. Though the majority of novel variants were in noncoding regions, we identified several novel, potentially deleterious nonsense and missense variants (See Table 4.3) across both the HapMap96 and clinical cohorts. Examples include

clearly deleterious variants such as a novel nonsense allele in *RYR1*, a gene linked to dominant Malignant Hyperthermia[71], and potentially deleterious novel missense alleles scattered across genes of clinical importance such as *CYP2C9*, *SLCO1B1*, and *SLC22A1*. As the sample sizes of the testing cohorts are relatively small, we conclude that PGRNseq can identify many more novel alleles of interest when applied to large studies that are well-powered to detect rare variation associated with variation in drug response.

4.4 Discussion

As adverse drug reaction events are a significant cause of morbidity in the US, a platform that can accurately detect and genotype variants, both common and rare, that affect drug response has the potential to both deepen our understanding of these events as well as reduce their incidence. Recent large scale sequencing efforts have revealed that very rare, potentially deleterious variants are carried by nearly 1 in 10 individuals. Therefore, platforms like PGRNseq that can accurately detect such variation, as well as genotype common variants of known effect, are particularly well-suited for pharmacogenetic research and clinical implementation. To this end, we have developed a novel custom-target panel designed to capture 84 genes with known roles in drug metabolism and response phenotypes. Aimed at larger pharmacogenomics studies, PGRNseq strikes a favorable balance between low cost, the high throughput associated with chip-based genotyping and the ultra-deep coverage associated with NGS; additionally, the deep coverage inherent to quality NGS data enables the discovery of rare variation of potential clinical impact. Testing PGRNseq across multiple sample sets revealed the high accuracy of genotypes obtained from this platform. Finally, the quantity and types of novel variation discovered in our small testing sets demonstrates that PGRNseq is ideal for larger sequencing studies aimed at

assessing rare variation within pharmacogenomics targets. Indeed, several current studies are making use of the PGRNseq platform. These studies range from investigations into a specific drug response phenotypes, such as irinotecan response in cancer patients, to large, multi-institution sequencing efforts such as eMERGE-PGx, which is deploying PGRNseq prospectively across 9000 patients with linked medical records in order to discover rare pharmacogenetic alleles of interest and to explore the utility of clinical NGS implementation[72]. Overall, we believe PGRNseq will continue to be a valuable resource for any investigator interested in examining rare variation in targets of known pharmacogenomic importance.

Although PGRNseq is currently being deployed in studies such as those described above, we are also in the process of designing and testing a 'version 2.0' platform to expand our abilities to target variation within the complex regions on the platform: *CYP2D6, CYP2A6,* and *HLA-B.* For *CYP2D6* and *CYP2A6,* we intend to extend the probe design to include the full gene (introns included) as well as the neighboring pseudogenes in order to aid in the assembly of this complex region, particularly with longer read lengths. We have also identified for inclusion several variants that are well known to tag the two "actionable" *HLA-B* alleles (*57:01 and *15:02) and that can be accurately typed by NGS[73]. Finally, we also plan to fine-tune our coverage of non-coding space by focusing on regions of putative regulatory function and removing low-coverage, repetitive regions that happened to fall within the boundaries of the original design scheme.

In PGRNseq v2, we will also continue to explore the utility of using this technology for the discovery and genotyping of Copy Number Variation (CNV) present within the targeted gene set. Several of these genes, such as *CYP2D6, CYP2A6,* and *GSTM1* are known to harbor common CNVs of functional relevance[66,67,74]. As a proof of concept, we applied a read-depth-based CNV-finding algorithm[75] to PGRNseq HapMap96 data, and checked for calls consistent

with Mendelian inheritance. Though there were many calls, likely false positives, not consistent with Mendelian inheritance, we did observe known, common CNVs that were transmitted from parent to child (Figure 4.3). We aim to explore these methods further during development of the next version, as the redesigned complex regions will likely improve call quality.

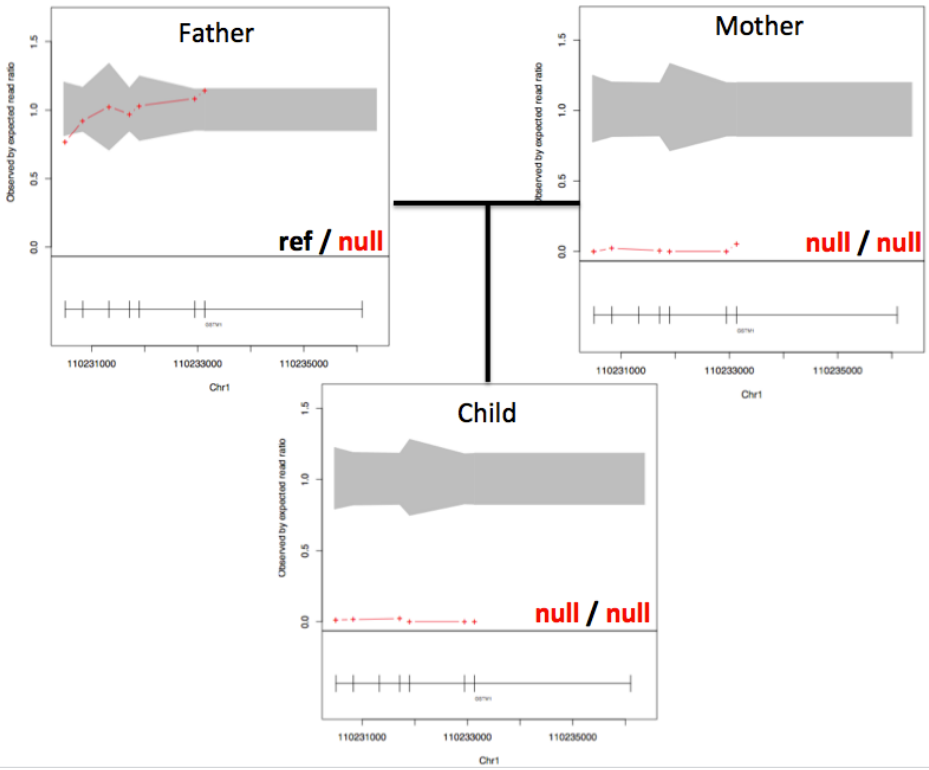

Figure 4.3 CNV Typing with PGRNseq ExomeDepth[75] detects common *GSTM1* deletion associated with drug toxicity. Calls consistent with both Mendelian inheritance and orthogonal Illumina ADME data

In addition to PGRNseq development, we feel the design and testing strategy utilized here is broadly applicable to the development of any custom-capture panel focusing on specific subsets of genomic targets. The use of the HapMap96 was essential in assessing optimum platform conditions, overall performance, and genotype accuracy due to the abundance of orthogonal data on these samples as well as the Mendelian inheritance analysis enabled by the use of trios, specifically. Though initially intended as a research tool, our experiences in the design and testing of this specific platform has led to an interest in pursuing the use of PGRNseq as a clinical test, and efforts to clinically validate this platform for certain actionable alleles are currently underway. As custom target platforms such as PGRNseq continue to demonstrate their efficacy as a research tool for the study of genetic variation, both rare and common, we believe that these very same platforms will become the standard for clinical sequencing, carving a translational niche for NGS in medicine ahead of clinical whole-genome sequence implementation.

4.5 Methods

PGRNseq Design

PGRN network members nominated genes for consideration in the design of the NGS platform. As one of the design criteria was to produce a panel that could be cost competitive with genotyping arrays, not all nominations could be included in the final list. Through multiple rounds of balloting and discussion, the group collectively decided on a final consensus list of 84 pharmacogenes for inclusion in the panel (Table 4.2). These pharmacogenes are functionally diverse and include drug-metabolizing enzymes, drug transporters, and drug targets. Although all 84 genes have some prior association with a pharmacogenetic trait, they range from those deemed clinically actionable by CPIC[61] at the time of voting to those about which little is known aside from strong preliminary association data. For the design of each of the 84 genes, we included all exons (based on all transcript models) as well as 2kb upstream and 1kb downstream of their untranslated regions (UTRs) in order to discover and assess nearby potential regulatory variation, which is already known to affect drug response in genes such as *VKORC1*[76]. In addition, the design also included probes to capture every site present on the Affy DMET+ array and the Illumina ADME assay, so that the sequencing platform would be backwards compatible with existing datasets, and as orthogonal platforms for PGRNseq quality control via genotype concordance. After submitting the final list of genomic coordinates to Nimblegen, we worked closely with their developmental team to generate a set of probes to capture these regions; the resulting set of SeqCap probes, known as PGRNseq, covers 968kb of the genome, which is highly scalable for large studies while maintaining high coverage.

PGRNseq Testing

In order to test different multiplexing strategies and assess the accuracy of the platform, the DSRs assembled a set of 96 HapMap samples of diverse ancestry (HapMap96). Since all samples have HapMap genotypes available, and some have 1000 Genomes sequencing data available, they represent a robust set to assess overall platform performance and concordance. Furthermore, these 96 samples consist of 32 trios, so analysis of Mendelian inheritance can reveal sites prone to false-positive calls due to mapping errors deriving from repetitive elements or from regions of high sequence homology; several genes on the platform are members of large gene families that can be prone to these errors. In addition to these samples, we also wanted to test PGRNseq performance on cohorts of actual patients with orthogonal data. We obtained two different clinical cohorts: 1) a set of 96 liver-derived patient samples[77], and 2) a separate set of 96 clinical samples collected for research into antiplatelet response for testing.

All sequencing was performed on the Illumina HiSeq 2000 instrument using paired-end, 100bp reads. Initially, all three DSR groups tested the HapMap96 using a variety of capture probe and sequencing lane multiplexing strategies (8-plex, 12-plex, 24-plex) to identify the maximum batch size that preserves the sequence read depth needed for high quality variant calls across the target set; with these criteria in mind the group settled on a 24-plex batch size. To compare performance, Illumina ADME genotypes and Affy DMET+ genotypes were generated for the HapMap96 at UW and BCM-HGSC, respectively. Clinical cohorts were sequenced at UW (liver samples) and BCM-HGSC (antiplatelet samples) using the same protocol as was used for the HapMap96 assays. At each site, raw sequencing reads were mapped to the hg19 reference genome using BWA, and variants called and filtered using GATK[78] and ATLAS[79].

Chapter 5: The Future of Pharmacogenetics in Clinical Practice

5.1 Research Summary

As next-generation sequencing continues to become the standard in patient care, clinical decision-making guided by results from NGS assays will become common practice. Although the field is rapidly transitioning from traditional genotyping-based methods to these newer technologies, their ability to detect and accurately represent all manner of pharmacogenetic variation has yet to be assessed. The work described here seeks to quantify this variation, assess its clinical representation with an eye towards broad-scale, preemptive pharmacogenetic testing. Chapter two uses large-scale exome data to describe the extent and qualities of rare, deleterious variation within key pharmacogenes that escapes detection by traditional genotyping assays. Chapter three places this variation in the context of known, actionable alleles by attempting to apply canonical pharmacogenetic nomenclature to genotypes derived from this exome dataset. With these issues in mind, chapter four introduces a new sequencing panel that marries the low cost and high throughput associated with genotyping assays with the accuracy and discovery ability inherent in whole-genome sequencing.

Although the work presented here is a step towards broad-scale implementation of clinical sequencing, there are still many challenges to overcome as this type of clinical test becomes increasingly common. Though the results presented in this dissertation are diverse in their scope, from nomenclature standardization issues to sequencing platform development, this work highlights three significant areas for future development as clinical sequencing for PGx expands: the need for larger, diverse cohorts; the need for high-throughput functional analysis of genomic variation; and the need for new methods to collect, interpret, and report

pharmacogenomics results focused on using Electronic Health Records to engage providers and the patients themselves.

5.2 Large-scale, multi-ethnic cohorts for pharmacogenomic discovery

Although considerable progress has been made regarding the discovery and implementation of clinically actionable pharmacogenetic variation, the majority of these variants are common, and many of the drug phenotype associations underlying this designation were identified in cohorts of Caucasian individuals. For example, PharmGKB lists 7 studies underlying the association between *SLCO1B1*5* (p.Val174Ala) and simvastatin response, an association ranked Level 1A by PharmGKB, denoting that this association not only has the highest level of evidence supporting it, but also that is currently implemented for testing at a PGRN site[46]. Though on the surface it seems there is little more to say regarding this association, 5 out of the 7 cited studies on PharmGKB consisted of exclusively Caucasian cohorts. Of the two remaining studies, one does not describe the 32 subjects' ancestries whatsoever[80] and the other is based on a cohort of 289 Caucasians, merely 22 African-Americans, and 43 individuals with ethnicity labeled simply as 'other'[81]. Despite the considerable effort to study this association in Caucasian individuals, these studies do little to address serious ethnic disparities in drug response phenotypes; minority individuals on statins hospitalized for coronary heart disease not only have significantly increased 1-year odds of rehospitalization or death, but they are also least likely to have health insurance to pay for the significant costs associated with any additional care[82].

The analyses presented here reveal that considerable variation in genes known to influence drug response traits has yet to be discovered, especially in under-studied and minority populations. Among ESP individuals, for example, 11.7% of African-American individuals carry

a rare, novel, potentially deleterious *CYP* variant, as opposed to only 7.6% of European-Americans; studies of other minority populations would likely reveal a similarly high burden of such variants. As exome sequencing data continues to aggregate rapidly, this prediction continues to hold true. The recently released ExAC database draws on exome sequencing data from 60,706 individuals of European, East Asian, South Asian, Latino, and African ancestry, none of whom were also part of ESP[83]; preliminary analysis of this dataset supports the need for larger cohorts of diverse ancestry. For example, rs374201833 (*CYP2C9* Arg97Thr), originally identified in a single ESP individual of European descent, affects a residue critical for substrate binding, and is predicted to alter enzyme activity[41,42]. Although no European individual in ExAC carried this variant, 13 individuals of South Asian descent and 1 Latino individual did. Further, a single South Asian individual carried a third allele at this same position, leading to a Lysine substitution at this residue that was previously unobserved, yet potentially equally as deleterious.

5.3 High throughput functional analysis of pharmacogenetic variation

Despite the utility in quantifying the extent and types of PGx variation in large, diverse cohorts, these studies can only suggest potential phenotypic consequences of the variation they seek to describe. As clinical PGx sequencing expands to millions of individuals of a variety of ethnicities, traditional techniques for assaying the effects of genetic variation on protein function are increasingly incapable of describing the functional consequences of such variation on a clinically-relevant timescale. Further, the algorithms designed to provide clinical decision support, those that interpret and report PGx test results, must be able to support reporting for all potential variants, even those that have yet to be observed. Despite the daunting nature of this

undertaking, recent advances in protein science and methodology will enable the high-throughput functional analysis necessary for the development of such reporting systems.

One method, Deep Mutational Scanning (DMS), is particularly ideal for rapid analysis of missense variation. Through the coupling of rounds of selection on a diverse library of protein variants with next-generation sequencing of input library DNA, this method can generate an activity landscape for nearly all possible single-amino-acid substitutions for a single enzyme[84]. Data generated via this method for a drug-metabolism enzyme such as CYP2C9 could lead to pre-existing clinical decision support for all possible substitutions; as the contributions of many common *CYP2C9* missense alleles to warfarin dose variation is known, DMS data for those alleles could be used to construct a model able to predict the effect of any allele, even before it is ever observed. In addition to this clinical utility, DMS can provide a deeper understanding of the precise biochemical effects underlying known associations, such as the catalytically impaired CYP2C9*3, defined by a seemingly benign substitution (p.Ile359Leu)[85]. As our understanding of the biology behind these common, actionable variants deepens, our ability to predict the functional consequences of similar variation in paralogous yet less well-studied genes expands in parallel, as much of the variability in warfarin dose has yet to be explained[86]. Although DMS-based approaches will significantly aid our ability to predict the enzymatic consequences of genetic variation, actual drug response phenotypes are complex traits not easily predicted by the effects of a single variant alone. Moving forward, this complexity necessitates a shift not only in how we ascertain sequence and genotype, but also how drug response phenotypes themselves are gathered and reported.

5.4 New approaches to collecting, interpreting, and reporting pharmacogenomic data

Novel methods such as next-generation sequencing provide unprecedented access to accurate genotype data drawn from many individuals. However, in order to truly understand the genetic architecture of drug response, new methods for the ascertainment and analysis of phenotype data are equally as important. As many PGx phenotypes are only observed after exposure to a certain pharmaceutical, separating true cases from true controls is essential. Electronic health records linked to genomic data are a rich source of phenotype information that will significantly improve on our ability detect pharmacogenetic associations. A central repository for health information throughout life, the EHR captures longitudinal data that is often inaccessible in traditional association studies, data that is often coded in standardized terms allowing for meta-analysis across many large cohorts[87]. Large multi-center cohorts with linked genomic data are a particularly promising avenue for pharmacogenetic research as they can retrospectively replicate previously described gene-drug associations as well as identify new associations inaccessible to smaller, single institution studies[88]. In addition to discovery, these cohorts provide the opportunity to implement preliminary clinical decision support systems to gauge their overall effectiveness in terms of cost, provider use, and patient outcome[26].

Although these large cohorts will be essential in unraveling the contributions of genetic variation to known drug response phenotypes, many adverse drug events are difficult to diagnose, not coded using a standardized nomenclature, and will require creative new approaches detect and understand. Despite these challenges, a new wave of crowd-based initiatives engaging directly with research participants are transforming the way adverse events are identified and reported. In the era when broadband internet is a public utility, many individuals turn to a search

engine for immediate access to medical information and symptomology. By applying natural language processing methods to de-identified search logs from 80 million users, one group was able not only to detect known adverse drug reactions at a rate comparable to the current FDA reporting system but also to discover novel drug reactions for future follow-up in clinical cohorts[89]. Though this method shows promise, search logs can only reveal an incomplete snapshot of an individual's symptomology, and lacks the ability to track these symptoms longitudinally. The increasing public interest in personal genomics and biometrics presents an incredible opportunity to directly engage with research participants to form an unprecedented cohort of millions of individuals, linking health records to genomic data to longitudinal biometric phenotypes. Indeed, the recently announced Precision Medicine Initiative is poised to make significant progress towards this ultimate goal by linking together existing biobanks into a million-patient consortium, the largest ever assembled[90]. In addition to top-down approaches such as this initiative, the coming years will see major advances in analysis of data collected from the bottom-up, data collected by research participants themselves. According to a recent survey, 69% of Americans track at least one health indicator, and 21% of these do so via some form of technology. Similarly, one in five smartphone owners owns at least one app related to health tracking, skewed heavily towards individuals 18-29[91]. To unify these efforts, Apple recently announced ResearchKit, a suite of developer tools designed specifically for the collection and reporting of biometric data using smartphones[92]. By empowering individuals in research studies to actively participate in data collection, we can discover new adverse events that are difficult to detect by standard methods and gain a clearer picture of how an individual's drug response varies over time on a much finer scale.

5.5 Pharmacogenomics and the future of precision medicine

Clinical implementation of next-generation sequencing for pharmacogenetics is poised to deliver on the promise of the human genome project: to use the information stored within our personal genomes to optimize treatment for all individuals. Although this goal is closer than ever before, the work presented in this dissertation highlights significant challenges that will need to be overcome before this potential is truly realized. Specifically, large cohorts of diverse ancestry are necessary to understand the true global spectrum of pharmacogenetic variation. However, the analysis of the massive amounts of data resulting from such an effort, and the reporting of results back to participants and patients, must be accompanied by new nomenclature and reporting systems, as current practices are insufficient to describe the complexity of pharmacogenetic variation at the exome or genome level. Until clinical genome sequencing becomes the standard of care, custom target panels such as PGRNseq are a valuable tool to make progress on these challenges. Predicting drug response variation and preventing adverse drug events is only one aspect of precision medicine, which generally seeks to use personal genomic, environmental, and lifestyle data to guide disease prevention and treatment. Thus, pharmacogenomics serves a test system for issues that will be faced across many aspects of precision medicine; as early pharmacogenomic implementation efforts continue adapt to the challenges outlined in this thesis presented by next-generation sequencing, the lessons learned in overcoming these barriers will serve as a valuable roadmap as precision medicine expands into all aspects of preventive and diagnostic care.

BIBLIOGRAPHY

1. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA*. 1998;**279**(15):1200-5.
2. Budnitz DS, Pollock DA, Weidenbach KN, Mendelsohn AB, Schroeder TJ, Annest JL. National surveillance of emergency department visits for outpatient adverse drug events. *JAMA*. 2006;**296**(15):1858-66.
3. Vogel F. Moderne problem der humangenetik. *Ergeb Inn Med U Kinderheilk*. 1959;**12**:52–125.
4. Motulsky AG. Drug reactions enzymes, and biochemical genetics. *J Am Med Assoc*. 1957;**165**(7):835-7.
5. Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science*. 1999;**286**(5439):487-91.
6. Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM, Nebert DW. Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics*. 2004;**14**(1):1-18.
7. Caudle KE, Rettie AE, Whirl-carrillo M, et al. Clinical pharmacogenetics implementation consortium guidelines for CYP2C9 and HLA-B genotypes and phenytoin dosing. *Clin Pharmacol Ther.* 2014;**96**(5):542-8.
8. Johnson JA, Gong L, Whirl-carrillo M, et al. Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. *Clin Pharmacol Ther*. 2011;**90**(4):625-9.
9. Crews KR, Gaedigk A, Dunnenberger HM, et al. Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update. *Clin Pharmacol Ther.* 2014;**95**(4):376-82.
10. Hicks JK, Swen JJ, Thorn CF, et al. Clinical Pharmacogenetics Implementation Consortium guideline for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants. *Clin Pharmacol Ther.* 2013;**93**(5):402-8.
11. Relling MV, Gardner EE, Sandborn WJ, et al. Clinical pharmacogenetics implementation consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing: 2013 update. *Clin Pharmacol Ther.* 2013;**93**(4):324-5.
12. Swen JJ, Nijenhuis M, De boer A, et al. Pharmacogenetics: from bench to byte--an update of guidelines. *Clin Pharmacol Ther.* 2011;**89**(5):662-73.
13. Hodges LM, Markova SM, Chinn LW, et al. Very important pharmacogene summary: ABCB1 (MDR1, P-glycoprotein). *Pharmacogenet Genomics.* 2011;**21**(3):152-61.
14. Oshiro C, Mangravite L, Klein T, Altman R. PharmGKB very important pharmacogene: SLCO1B1. *Pharmacogenet Genomics.* 2010;**20**(3):211-6.
15. Wilke RA, Ramsey LB, Johnson SG, et al. The clinical pharmacogenomics implementation consortium: CPIC guideline for SLCO1B1 and simvastatin-induced myopathy. *Clin Pharmacol Ther.* 2012;**92**(1):112-7.
16. Buettner C, Lecker SH. Molecular basis for statin-induced muscle toxicity: implications and possibilities. *Pharmacogenomics*. 2008;**9**(8):1133-42.

17. Tirona RG, Leake BF, Merino G, Kim RB. Polymorphisms in OATP-C: identification of multiple allelic variants associated with altered transport activity among European- and African-Americans. *J Biol Chem.* 2001;**276**(38):35669-75.

18. Niemi M, Schaeffeler E, Lang T, et al. High plasma pravastatin concentrations are associated with single nucleotide polymorphisms and haplotypes of organic anion transporting polypeptide-C (OATP-C, SLCO1B1). *Pharmacogenetics*. 2004;**14**(7):429-40.
19. Link E, Parish S, Armitage J, et al. SLCO1B1 variants and statin-induced myopathy--a genomewide study. *N Engl J Med.* 2008;**359**(8):789-99.
20. Evans WE, Mcleod HL. Pharmacogenomics--drug disposition, drug targets, and side effects. *N Engl J Med.* 2003;**348**(6):538-49.
21. Thorn CF, Klein TE, Altman RB. PharmGKB summary: very important pharmacogene information for angiotensin-converting enzyme. *Pharmacogenet Genomics.* 2010;**20**(2):143-6.
22. Litonjua AA, Gong L, Duan QL, et al. Very important pharmacogene summary ADRB2. *Pharmacogenet Genomics.* 2010;**20**(1):64-9.
23. Owen RP, Gong L, Sagreiya H, Klein TE, Altman RB. VKORC1 pharmacogenomics summary. *Pharmacogenet Genomics.* 2010;**20**(10):642-4.
24. Cooper GM, Johnson JA, Langaee TY, et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood*. 2008;**112**(4):1022-7.
25. Relling MV, Klein TE. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clin Pharmacol Ther.* 2011;**89**(3):464-7.
26. Dunnenberger HM, Crews KR, Hoffman JM, et al. Preemptive Clinical Pharmacogenetics Implementation: Current Programs in Five US Medical Centers. *Annu Rev Pharmacol Toxicol.* 2015;**55**:89-106.
27. Shendure J, Porreca GJ, Reppas NB, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*. 2005;**309**(5741):1728-32.
28. Nelson MR, Wegmann D, Ehm MG, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*. 2012;**337**(6090):100-4.
29. Lee H, Schatz MC. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics*. 2012;**28**(16):2097-105.
30. Wang L, Mcleod HL, Weinshilboum RM. Genomics and drug response. *N Engl J Med*. 2011;**364**(12):1144-53.
31. Paré G, Mehta SR, Yusuf S, et al. Effects of CYP2C19 genotype on outcomes of clopidogrel treatment. *N Engl J Med.* 2010;**363**(18):1704-14.
32. Klein TE, Altman RB, Eriksson N, et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med*. 2009;**360**(8):753-64.
33. Meckley LM, Wittkowsky AK, Rieder MJ, Rettie AE, Veenstra DL. An analysis of the relative effects of VKORC1 and CYP2C9 variants on anticoagulation related outcomes in warfarin-treated patients. *Thromb Haemost.* 2008;**100**(2):229-39.
34. Fu W, O'connor TD, Jun G, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. 2013;**493**(7431):216-20.

35. Mckenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;**20**(9):1297-303.
36. Tennessen JA, Bigham AW, O'connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012;**337**(6090):64-9.
37. Cooper GM, Stone EA, Asimenos G, et al, et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 2005;**15**(7):901-13.
38. Cooper GM, Goode DL, Ng SB, et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods*. 2010;**7**(4):250-1.
39. Grantham R. Amino acid difference formula to help explain protein evolution. *Science*. 1974;**185**(4154):862-4.
40. Lane S, Al-zubiedi S, Hatch E, et al. The population pharmacokinetics of R- and S-warfarin: effect of genetic and clinical factors. *Br J Clin Pharmacol*. 2012;**73**(1):66-76.
41. Dickmann LJ, Locuson CW, Jones JP, Rettie AE. Differential roles of Arg97, Asp293, and Arg108 in enzyme stability and substrate specificity of CYP2C9. *Mol Pharmacol*. 2004;**65**(4):842-50.
42. Williams PA, Cosme J, Ward A, Angove HC, Matak vinković D, Jhoti H. Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature*. 2003;**424**(6947):464-8.
43. Rettie AE, Tai G. The pharmocogenomics of warfarin: closing in on personalized medicine. *Mol Interv*. 2006;**6**(4):223-7.
44. Tabor HK, Auer PL, Jamal SM, et al. Pathogenic variants for Mendelian and complex traits in exomes of 6,517 European and African Americans: implications for the return of incidental results. *Am J Hum Genet.* 2014;**95**(2):183-93.
45. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet.* 2007;**81**(5):1084-97.
46. Whirl-carrillo M, Mcdonagh EM, Hebert JM, et al. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther.* 2012;**92**(4):414-7.
47. Woodson LC, Ames MM, Selassie CD, Hansch C, Weinshilboum RM. Thiopurine methyltransferase. Aromatic thiol substrates and inhibition by benzoic acid derivatives. *Mol Pharmacol.* 1983;**24**(3):471-8.
48. Lennard L, Van loon JA, Lilleyman JS, Weinshilboum RM. Thiopurine pharmacogenetics in leukemia: correlation of erythrocyte thiopurine methyltransferase activity and 6-thioguanine nucleotide concentrations. *Clin Pharmacol Ther.* 1987;**41**(1):18-25.
49. Lee CR, Goldstein JA, Pieper JA. Cytochrome P450 2C9 polymorphisms: a comprehensive review of the in-vitro and human data. *Pharmacogenetics*. 2002;**12**(3):251-63.
50. Ramsey LB, Johnson SG, Caudle KE, et al. The clinical pharmacogenetics implementation consortium guideline for SLCO1B1 and simvastatin-induced myopathy: 2014 update. *Clin Pharmacol Ther*. 2014;**96**(4):423-8.
51. Zanger UM, Turpeinen M, Klein K, Schwab M. Functional pharmacogenetics/genomics of human cytochromes P450 involved in drug biotransformation. *Anal Bioanal Chem.* 2008;**392**(6):1093-108.

52. Busi F, Cresteil T. CYP3A5 mRNA degradation by nonsense-mediated mRNA decay. *Mol Pharmacol*. 2005;**68**(3):808-15.
53. Scott SA, Sangkuhl K, Stein CM, et al. Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update. *Clin Pharmacol Ther.* 2013;**94**(3):317-23.
54. Ohlsson rosenborg S, Mwinyi J, Andersson M, et al. Kinetics of omeprazole and escitalopram in relation to the CYP2C19*17 allele in healthy subjects. *Eur J Clin Pharmacol.* 2008;**64**(12):1175-9.
55. Crosby J, Peloso GM, Auer PL, et al. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med.* 2014;**371**(1):22-31.
56. O'roak BJ, Vives L, Girirajan S, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*. 2012;**485**(7397):246-50.
57. Tammiste A, Jiang T, Fischer K, et al. Whole-exome sequencing identifies a polymorphism in the BMP5 gene associated with SSRI treatment response in major depression. *J Psychopharmacol.* 2013;**27**(10):915-20.
58. Pritchard CC, Smith C, Salipante SJ, et al. ColoSeq provides comprehensive lynch and polyposis syndrome mutational analysis using massively parallel sequencing. *J Mol Diagn*. 2012;**14**(4):357-66.
59. Kim EH, Lee S, Park J, Lee K, Bhak J, Kim BC. New lung cancer panel for high-throughput targeted resequencing. *Genomics Inform.* 2014;**12**(2):50-7.
60. Wang F, Wang H, Tuan HF, et al. Next generation sequencing-based molecular diagnosis of retinitis pigmentosa: identification of a novel genotype-phenotype correlation and clinical refinements. Hum Genet. 2014;133(3):331-45.
61. Caudle KE, Klein TE, Hoffman JM, et al. Incorporation of pharmacogenomics into routine clinical practice: the Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline development process. *Curr Drug Metab.* 2014;**15**(2):209-17.
62. Liu X, Cheng D, Kuang Q, Liu G, Xu W. Association of UGT1A1*28 polymorphisms with irinotecan-induced toxicities in colorectal cancer: a meta-analysis in Caucasians. *Pharmacogenomics J.* 2014;**14**(2):120-9.
63. Ramsey LB, Bruun GH, Yang W, et al. Rare versus common variants in pharmacogenetics: SLCO1B1 variation and methotrexate disposition. *Genome Res.* 2012;**22**(1):1-8.
64. Ramirez AH, Shaffer CM, Delaney JT, et al. Novel rare variants in congenital cardiac arrhythmia genes are frequent in drug-induced torsades de pointes. *Pharmacogenomics J.* 2013;**13**(4):325-9.
65. Major E, Rigó K, Hague T, Bérces A, Juhos S. HLA typing from 1000 genomes whole genome and whole exome illumina data. *PLoS ONE.* 2013;**8**(11):e78410.
66. Hicks JK, Swen JJ, Gaedigk A. Challenges in CYP2D6 phenotype assignment from genotype data: a critical assessment and call for standardization. *Curr Drug Metab.* 2014;**15**(2):218-32.
67. Mwenifumbo JC, Tyndale RF. Genetic variability in CYP2A6 and the pharmacokinetics of nicotine. *Pharmacogenomics*. 2007;**8**(10):1385-402.
68. Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.
69. Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;**491**(7422):56-65.

70. NIEHS Environmental Genome Project, Seattle, WA (URL: http://evs.gs.washington.edu/niehsExome/)
71. Kim JH, Jarvik GP, Browning BL, et al. Exome sequencing reveals novel rare variants in the ryanodine receptor and calcium channel genes in malignant hyperthermia families. *Anesthesiology*. 2013;**119**(5):1054-65.
72. Rasmussen-torvik LJ, Stallings SC, Gordon AS, et al. Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems. *Clin Pharmacol Ther*. 2014;**96**(4):482-9.
73. De bakker PI, Mcvean G, Sabeti PC, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet.* 2006;**38**(10):1166-72.
74. Monteiro TP, El-jaick KB, Jeovanio-silva AL, et al. The roles of GSTM1 and GSTT1 null genotypes and other predictors in anti-tuberculosis drug-induced liver injury. *J Clin Pharm Ther*. 2012;**37**(6):712-8.
75. Plagnol V, Curtis J, Epstein M, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*. 2012;**28**(21):2747-54.
76. Wang D, Chen H, Momary KM, Cavallari LH, Johnson JA, Sadée W. Regulatory polymorphism in vitamin K epoxide reductase complex subunit 1 (VKORC1) affects gene expression and warfarin dose requirement. *Blood*. 2008;**112**(4):1013-21.
77. Innocenti F, Cooper GM, Stanaway IB, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet*. 2011;**7**(5):e1002078.
78. Van der auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;**11**(1110):11.10.1-11.10.33.
79. Challis D, Yu J, Evani US, et al. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics*. 2012;**13**:8.
80. Pasanen MK, Neuvonen M, Neuvonen PJ, Niemi M. SLCO1B1 polymorphism markedly affects the pharmacokinetics of simvastatin acid. *Pharmacogenet Genomics*. 2006;**16**(12):873-9.
81. Voora D, Shah SH, Spasojevic I, et al. The SLCO1B1*5 genetic variant is associated with statin-induced side effects. *J Am Coll Cardiol*. 2009;**54**(17):1609-16.
82. Mochari-greenberger H, Liao M, Mosca L. Racial and ethnic differences in statin prescription and clinical outcomes among hospitalized patients with coronary heart disease. *Am J Cardiol.* 2014;**113**(3):413-7.
83. Exome Aggregation Consortium (ExAC), Cambridge, MA (URL: http://exac.broadinstitute.org)
84. Fowler DM, Stephany JJ, Fields S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat Protoc*. 2014;**9**(9):2267-84.
85. Kirchheiner J, Brockmöller J. Clinical consequences of cytochrome P450 2C9 polymorphisms. *Clin Pharmacol Ther*. 2005;**77**(1):1-16.
86. Schwarz UI, Ritchie MD, Bradford Y, et al. Genetic determinants of response to warfarin during initial anticoagulation. *N Engl J Med.* 2008;**358**(10):999-1008.

87. Mccarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics.* 2011;**4**:13.
88. Wilke RA, Xu H, Denny JC, et al. The emerging role of electronic medical records in pharmacogenomics. *Clin Pharmacol Ther.* 2011;**89**(3):379-86.
89. White RW, Harpaz R, Shah NH, Dumouchel W, Horvitz E. Toward enhanced pharmacovigilance using patient-generated data on the internet. *Clin Pharmacol Ther.* 2014;**96**(2):239-46.
90. Collins FS, Varmus H. A New Initiative on Precision Medicine. *N Engl J Med*. 2015, epub ahead of print
91. Fox S and Duggan M. "Tracking for Health." Pew Research Center, Washington D.C. (January 28, 2013). http://www.pewinternet.org/2013/01/28/tracking-for-health/
92. https://www.apple.com/researchkit/

VITA

Adam Gordon was born in Boca Raton, Florida. He graduated with Bachelor of Arts degree in Biology from the University of Chicago in 2009. He then moved westward to Seattle to pursue graduate studies in the laboratory of Dr. Deborah Nickerson at the University of Washington, earning his Doctor of Philosophy degree in Genome Sciences in 2015.