NEW APPROACHES IN UNDERSTANDING DRUG METABOLISM

By

STEVEN HART

Submitted to the graduate degree program in Department of Pharmacology, Toxicology and Therapeutics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chairperson: Xiao-bo Zhong, PhD

_____

Grace Guo, PhD

_____

J. Steven Leeder, PharmD/PhD

_____

Kenneth Peterson, PhD

_____

Curtis Klaassen, PhD

Date Defended: February 22nd, 2011

The Dissertation Committee for STEVEN HART

certifies that this is the approved version of the following dissertation:

NEW APPROACHES IN UNDERSTANDING DRUG METABOLISM

_____

Chairperson Xiaobo Zhong, PhD

Date approved: February 22nd, 2011

ACKNOWLEDGEMENTS

I am extremely grateful to have worked with two former lab members, Ye Li, PhD, and Kaori Nakamoto, BS. Ye and Kaori were the only two people in the lab for the first couple of years of my PhD work. Ye and Kaori were two friends with whom I could discuss the complexities of our research designs and give honest and frank feedback for future research ideas. This critical responsibility has now transferred to our new postdoctoral fellow Dan Li, PhD. Admittedly, her soft-spoken style often complements (and lessens) my sometimes boorish personality. Another soft-spoken supporter I have is my fellow graduate student Lai Peng. Lai always questions everything - even things that may seem trivial to me - however, I frequently don't know the answer, and am compelled to find the answer. I would also like to acknowledge Hans Tregear, our new research assistant.

I have also had much support from the main campus. Jianwen Fang, PhD, Research Assistant Professor, Information & Telecommunication Technology Center and Courtesy Assistant Professor, Electrical Engineering & Computer Science Department has let me remotely access his server architecture at KU, without which I would not have been able to perform many of my computational experiments. Emily Scott, PhD, an Associate Professor in the Department of Medicinal Chemistry had many discussions with me about structural impacts of POR mutations. Emily has also written some of my Letters of Recommendation. Finally, the Madison and Lila Self Graduate Fellowship provided my stipend and additional non-scientific

training. Thanks to Biopredic International for providing the HepaRG cells and necessary media.

I deeply appreciate the patience and leeway granted to me by members of my Dissertation Committee, especially Grace Guo. Not only did Grace write several of my letters of recommendation, but I also worked very closely with her lab for many years, and it was on her behalf that Jianwen offered access to his computer cluster. I have never met Jianwen. Steve Leeder introduced me to the concept of Principal Component Analysis which has now become an essential part of my process of reducing dimensionality of data, making it simpler to understand. Thank you to Ken Peterson for his our discussions about the roles of genetics versus epigenetics. Curt Klaassen and I have worked on several projects throughout my tenure, and I appreciate his insightful suggestions. Absolutely none of this would have been possible without the support of my mentor Xiaobo Zhong. Thank you for providing an excellent working atmosphere and for keeping my life in balance. Along these lines, I would like to thank the help I received from the departmental administrative staff including Rosa Meagher, Dorothy McGregor, and Cody Tully. Finally, thank you to my family for all your support.

ABSTRACT

Limitations in technology, such as DNA sequencing and appropriate model systems, have made it difficult to understand the genetic and non-genetic factors that influence the liver's role in metabolizing drugs. New approaches are required to overcome these limitations. In this Dissertation, we evaluate 3 such new approaches.

Our first new approach relates to the field of pharmacogenetics: using genetics to predict how a patient will respond to medication based on their genetic code. We looked for polymorphisms in a novel target gene, Cytochrome P450 Oxidoreductase (POR). Our results show a mutation in P450 reductase (L577P) that associates with decreased metabolism for 8 of 10 major drug metabolizing enzymes. However, even though we found a statistical association between POR polymorphism and drug metabolism, a wide range of variation in POR activity was still observed among the samples with the L577/ P577 genotype, making predicting POR activity solely on the basis of L577P genotype difficult.

POR represents only a single gene amongst the tens of thousands present in the human genome. To investigate the relationship between how genes and their products interact, a systems approach is necessary. Therefore, in our second new approach, we will characterize the transcriptome of our model system, the HepaRG cell line. We found that HepaRG cells globally transcribe genes at the levels more similar to human primary hepatocytes and human liver than HepG2 cells, particularly in genes encoding drug processing proteins.

Finally, I describe the third new approach: the use of next-generation DNA sequencing to understand hepatic drug response. This section contains two parts. First, we introduce methods that significantly decrease the false discovery rate of genotyping from RNA-Seq data. With these high fidelity SNPs, we were able to perform a genome-wide pharmacogenomic analysis on HepaRG cells. Second, we introduce a new program, called PRUNE, to more accurately quantify gene expression, and compare its performance to that of established programs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADH	alcohol dehydrogenase
AHR	aryl-hydrocarbon receptor
ALDH	aldehyde dehydrogenase
CAR or NR1I3	constitutive androstane receptor
CYP	cytochrome P450
Dex	Dexamethasone
DMSO	Dimethyl sulfoxide
FMO	Flavin monooxygenase
GST	glutathione *S*-transferase
NAT	*N*-acetyl transferase
PB	Phenobarbital
POR	cytochrome P450 Oxidoreductase
PXR or NR1I2	pregnane x receptor
Rif	Rifampicin
SNP	single nucleotide polymorphism
SULT	sulfotransferase
UGT	UDP glucuronosyltransferase
XDH	Xanthine dehydrogenase
FP	False Positive

TP	True Positive
FN	False Negative
TN	True Negative
SNS	Sensitivity
SPC	Specificity
ACC	Accuracy
PCN	Precision
FDR	False Discovery Rate

CHAPTER 1. INTRODUCTION

1.1. Theme and Scope

The following Dissertation is presented in the following three main sections, each with its own dedicated chapter:

1. **A New Target Gene for Pharmacogenetics:** Cytochrome P450 Oxidoreductase (POR)
2. **A New Cell Model for Studying Drug Metabolism and Liver Toxicity:** HepaRG Cells as a Surrogate for Primary Human Hepatocytes
3. **A New Tool for Understanding Drug Response:** Next-generation-based mRNA Sequencing in Pharmacogenomics

Chapter 2 will focus on our contribution to the field of pharmacogenetics: using genetics to predict how a patient will respond to medication. The *status quo* for the field of pharmacogenetics and genomics is to identify rare alleles in well-known and well-characterized pharmacogenetic genes from large populations. In contrast to this, and thus representing the ***first new approach***, we chose to look for polymorphisms in a novel target gene, Cytochrome P450 Oxidoreductase (*POR*). Chapter 2 begins with a brief history of pharmacogenetic research, define several key terms that will be used throughout this Dissertation, and describe the approach, results, and significance for identifying several novel mutations in the *POR* gene.

Chapter 3 will focus on the use of HepaRG cell line as a valuable newer *in vitro* cell model for studies of drug metabolism and liver toxicity in drug discovery. The ***second new approach*** was to assess on the total gene expression profile by microarrays, whereby were able to show that HepaRG are more correlated to the gene expression of both primary hepatocytes

and liver tissue than the commonly-used hepatocyte cell line, HepG2. Primary human hepatocytes still are considered the 'gold standard' for understanding human hepatocyte function, but HepaRG may be a suitable alternative in many cases.

In chapter 4, we describe the ***third new approach***: the use of next-generation mRNA sequencing to understand drug response, in the context of hepatic gene expression. Because the first mRNA-Seq experiments were published less than two years ago, much of the analytical dogma required for interpretation of the data has not been established. We propose, and gauge the performance of, analytical pipelines for genotyping mRNA-Seq and how to relate those genotypes to pharmacologically relevant information. We also developed a new software program to accurately quantitate gene expression levels and compare its performance to an established method.

1.2. A New Target Gene for Pharmacogenetics: Cytochrome P450 Oxidoreductase (POR)

1.2.1. Pharmacogenetic studies of pharmacokinetic genes

1.2.1.1. Variation in drug metabolism and clinical outcomes

Adverse drug reactions cause about 100,000 deaths and 2 million serious events, and is responsible for 5-7% of hospital admissions in the United States of America each year (Lazarou et al., 1998; Gandhi et al., 2003; Dormann et al., 2004), indicating a need to improve drug safety. Current medical practice uses standard protocols to select drugs and doses to treat human disease and other maladies. Standard drug dosing protocols are based on the assumption that most patients respond homogeneously to drugs. However, significant inter-individual variation in drug response exists in the general population with respect to both what a drug does to the body

(pharmacodynamics: interactions between a drug, its target, and downstream effects) and what the body does to the drug (pharmacokinetics: absorption, distribution, metabolism, and excretion of a drug from the body, ADME). The duration and intensity of pharmacological action of a drug is influenced by both its pharmacokinetic and pharmacodynamic parameters.

As an example of how pharmacokinetic parameters can affect efficacy and safety, think about the rate at which the body metabolizes xenobiotics (especially the intestines and liver). If the drug is metabolized too fast, it may not reach the required effective concentration to act on its target. Conversely, if the drug is metabolized too slowly, then it may accumulate, leading to higher, and potentially dangerous, drug concentrations. Based on the rate for metabolizing a specific drug, an individual can be classified as an ultrafast metabolizer, extensive metabolizer, intermediate metabolizer, or poor metabolizer for that drug. The majority of individuals in the general population are defined as extensive metabolizers. Standard protocols are designed for optimal therapeutic efficacy and minimal adverse drug reactions for extensive metabolizer patients. However, if a drug is primarily inactivated by metabolic enzymes, ultrafast metabolizers have a higher ability to inactivate the drug than extensive metabolizers and may not benefit from the drug at the standard dose. Conversely, poor metabolizers have a lower ability to inactivate drugs than extensive metabolizers and may experience undesirable, even fatal, adverse drug reactions resulting from increased plasma concentrations, leading to more on- and off-target effects. Dosage should be adjusted to optimize the safety and efficacy of prescription drugs based on each individual's drug metabolic rates.

Many factors influence the drug metabolic rates, including environmental factors (diet and environmental toxicants), physiological factors (age and gender), pathological factors (liver, kidney, or heart diseases), and genetic factors (genetic polymorphism). Of these factors, genetic

polymorphism may be one of the most important factors. For some drugs, genetic polymorphism contributes to more than 50% of the variation in drug response (Ozdemir et al., 2000; Sanderson et al., 2005; Wadelius and Pirmohamed, 2007). Pharmacogenetic studies on drug metabolic pathways attempt to determine which genetic polymorphisms can influence drug metabolic rates, and how to adjust the dose for individual patients based on their genetic makeup.

1.2.1.2. Pharmacogenetic studies related to drug metabolism pathways

Drugs are usually metabolized by specialized enzymatic systems to convert lipophilic chemical compounds into more readily excreted polar products. The metabolism occurs primarily in liver hepatocytes and epithelial cells of the small intestine, but also in other tissues, such as lungs, kidneys, and skin. Figure 1.1 illustrates a process of drug metabolism by the specialized xenobiotic response systems in a cell. Drugs are usually absorbed by the small intestine and distributed to the liver. The drugs must diffuse through the lipid bilayer cellular membrane or be transported by uptake membrane transporters. Lipophilic drugs are then biotransformed into polar products by phase I and phase II reactions. Phase I reactions usually precedes phase II, although not necessarily. If the metabolites of phase I reactions are sufficiently polar, they may be readily excreted. However, many phase I products are not eliminated rapidly and undergo a subsequent phase II reaction. The polar products are secreted out of the cells by efflux membrane transporters into the elimination system. Drug import and export by membrane transporters are sometime referred to as phase III reactions (though this nomenclature is not widely accepted). The phase I, II, and III reactions are coordinately regulated by similar mechanisms mediated by nuclear receptors (Honkakoski and Negishi, 2000;

Xu et al., 2005; Nakata et al., 2006), such as pregnane x receptor (PXR) and constitutive androstane receptor (CAR). These nuclear receptors can be activated by drugs as ligands. The activated nuclear receptors act as xenosensors to interact with xenobiotic response elements in the promoters of phase I, II, and III genes to stimulate gene transcription.

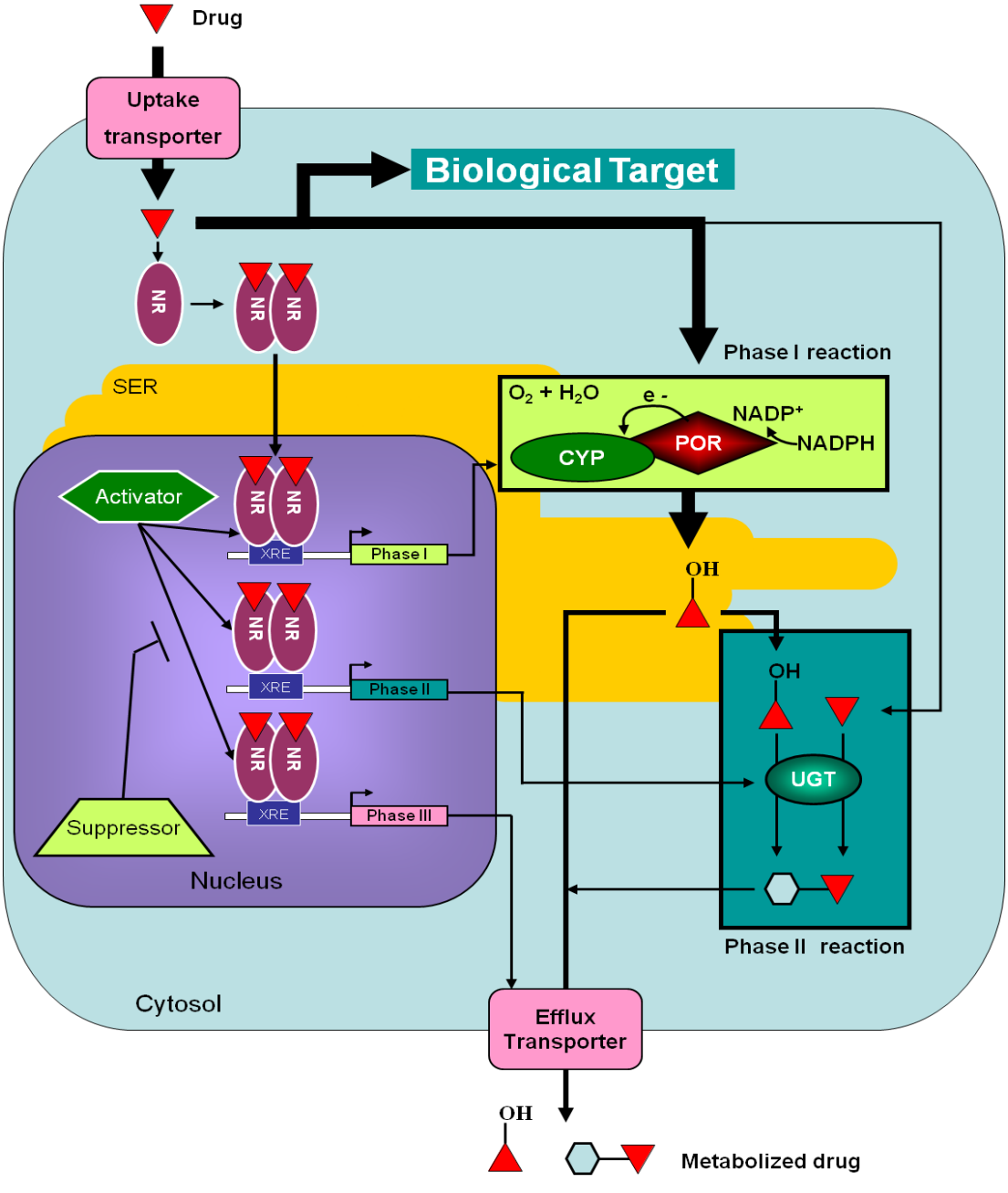


Figure 1.1. Graphic representation of drug metabolic processes in a drug metabolizing cell. Abbreviations: NR, nuclear receptor; XRE, xenobiotic response element; CYP, cytochrome P450; POR, cytochrome P450 oxidoreductase; UGT: uridine 5'-diphospho-glucuronosyltransferase; SER: smooth endoplasmic reticulum.

Phase I reactions include oxidation, reduction, hydrolysis, cyclization, and decyclization. The most extensive pharmacogenetic studies of phase I reactions have focused on the microsomal cytochrome P450 (CYP) superfamily of genes. Humans have 50 microsomal CYP genes (plus 7 additional non-microsomal), of which 15 are involved in drug metabolism for about 80% of prescription drugs (Evans and Relling, 1999; Lander et al., 2001). Pharmacogenetic and clinical aspects of CYP polymorphisms have been the subject of many reviews (Daly, 2004; Ingelman-Sundberg et al., 2007; Plant, 2007). The impact of null alleles resulting from genetic polymorphisms, which can affect CYP gene expression and/or enzyme activity, have been well characterized for some CYP isoforms, such as CYP2A6, CYP2B6, CYP2C9, CYP2C19, CYP2D6, and CYP3A5. Based on whether one or both chromosomes carry the null alleles or have multiple copies of the normal CYP genes, phenotypes of drug metabolism (i.e. ultrafast, extensive, or poor metabolizer) can be predicted with a certain degree of accuracy (Gaedigk et al., 1999; Lundqvist et al., 1999; Zanger et al., 2004). However, null alleles have not been identified in all of the CYP genes, for example CYP1A2 and CYP2E1. Null mutations in CYP3A4 are rare, with only one group observing 4 family members with a heterozygous *CYP3A4*20* allele (Westlind-Johnsson et al., 2006). CYP3A4 is the most abundant CYP isoform (up to 30% of total CYP content) in the liver and small intestine (Wrighton and Stevens, 1992) and plays a major role in the biotransformation of ~40-50% of currently prescribed drugs (Evans and Relling, 1999). Significant variation in CYP3A4-mediated drug metabolism exists in the general population, and this may be the reason for differences in therapeutic efficacy and toxicity for many drugs administrated at a standard dose (Schellens et al., 1988; Renwick et al., 1998). Genetics are likely the cause for these differences (Ozdemir et al., 2000), however most identified CYP3A4 polymorphisms have allele frequencies

of <1% in the general population, which cannot explain such wide variation in CYP3A4-catalyzed drug metabolism, the exception being *CYP3A4*1B* (Rebbeck et al., 1998; Walker et al., 1998). This indicates that CYP enzymes may not be the only factors for varied drug responses. In many cases, variations in drug response may be due to polygenic or epigenetic factors that remain to be elucidated - not only within the CYP superfamily, but also in other players involved in drug metabolism.

The other players include (1) phase II enzymes, which further metabolize phase I reaction products; (2) phase III membrane transporters, which control the amount of drugs in each cell, or (3) nuclear receptors, which determine the amount of phase I, II, and III proteins available to act on the drugs (Figure 1.1). These players are critical to drug metabolism as they can influence metabolic rates. Phase II reactions essentially conjugate drugs with hydrophilic moieties, such as glucuronic acid, glutathione, sulfonates, and amino acids. Phase II reactions are catalyzed by UDP-glucuronosyltransferases (UGTs), glutathione *S*-transferases (GSTs), sulfotransferases (STs), *N*-acetyltransferases (NATs), and amino acid *N*-acyl transferases. Clinically significant adverse drug reactions have been reported in patients carrying defective UGT (Burchell et al., 2000), GST (Roy et al., 2001), and NAT alleles (Huang et al., 2002). Pharmacogenetic aspects of phase II enzymes have been extensively reviewed elsewhere (Guillemette, 2003; de Jong et al., 2006; Nowell and Falany, 2006; Lo and Ali-Osman, 2007).

From the phase III perspective, genetic polymorphisms in the membrane transporter genes of multidrug resistance 1 *P*-glycoprotein (Lamba et al., 2006) and organic anion transporter proteins (Tirona et al., 2001; Michalski et al., 2002; Nozawa et al., 2002; Letschert et al., 2004) influence P450-catalyzed drug metabolism either by inhibiting proper maturation and localization of the transporters, or by changing substrate specificity. In either case, this alters the

amount of drug that can get into the cells where the Cytochrome P450s and other metabolic machinery can metabolize them. Additionally, genetic polymorphisms in nuclear receptor genes of PXR (Koyano et al., 2004; Lamba et al., 2005) and CAR (Ikeda et al., 2005; Lamba et al., 2005) have been identified and shown to alter downstream target CYP gene expression.

Undoubtedly, pharmacogenetic studies of phase II enzymes, phase III transporters, and nuclear receptors will continue to elucidate mechanisms by which variation occurs for drug metabolism. However, a great need still exists to identify markers to better predict drug response for drugs that are predominantly oxidized by the CYP superfamily of enzymes. Because genetic explanations for some aberrant CYP-mediated metabolism remain enigmatic, polymorphisms in their functional partners may contribute to this altered metabolic function. One important partner is cytochrome P450 oxidoreductase (POR), which is the sole electron donor for all microsomal CYP enzymes.

1.2.2. Cytochrome P450 Oxidoreductase (POR)

1.2.2.1. Physiologic functions of POR

NADPH-cytochrome P450 oxidoreductase (POR or CYPOR) is also known as NADPH-cytochrome P450 reductase (CPR), P450 reductase (P450R), NADPH-hemoprotein oxidoreductase, and NADPH-ferrihemoprotein oxidoreductase. POR was first identified as an NADPH-specific cytochrome *c* reductase (Horecker, 1950). Later studies showed that POR is not the physiological enzyme for cytochrome *c* reduction in mitochondria (Williams and Kamin, 1962), rather it is located on the smooth endoplasmic reticulum where it donates electrons to several oxygenase enzymes, as depicted in Figure 1.2. These oxygenase enzymes include: (1)

heme oxygenases (Schacter et al., 1972), which catalyze the degradation of heme to bilirubin; (2) squalene monooxygenase (or squalene epoxidase) (Ono and Bloch, 1975), the first oxygenation step and rate-limiting enzyme to reduce squalene to 2,3-oxidosqualene in sterol biosynthesis; (3) 7-dehydrocholesterol reductase (Nishino and Ishibashi, 2000), the enzyme that reduces the C7-C8 double bond of 7-dehydrocholesterol in the cholesterol biosynthesis pathway; (4) cytochrome *b5* (Enoch and Strittmatter, 1979), which supports the fatty acid desaturase and elongase for metabolism of fatty acids; (5) microsomal CYP monooxygenases (Vermilion JL, 1981), heme-containing proteins that catalyze biosynthesis of steroid hormones, cholesterol, and bile acids, as well as metabolism of vitamins, steroids, and more than 80% of current prescription drugs.

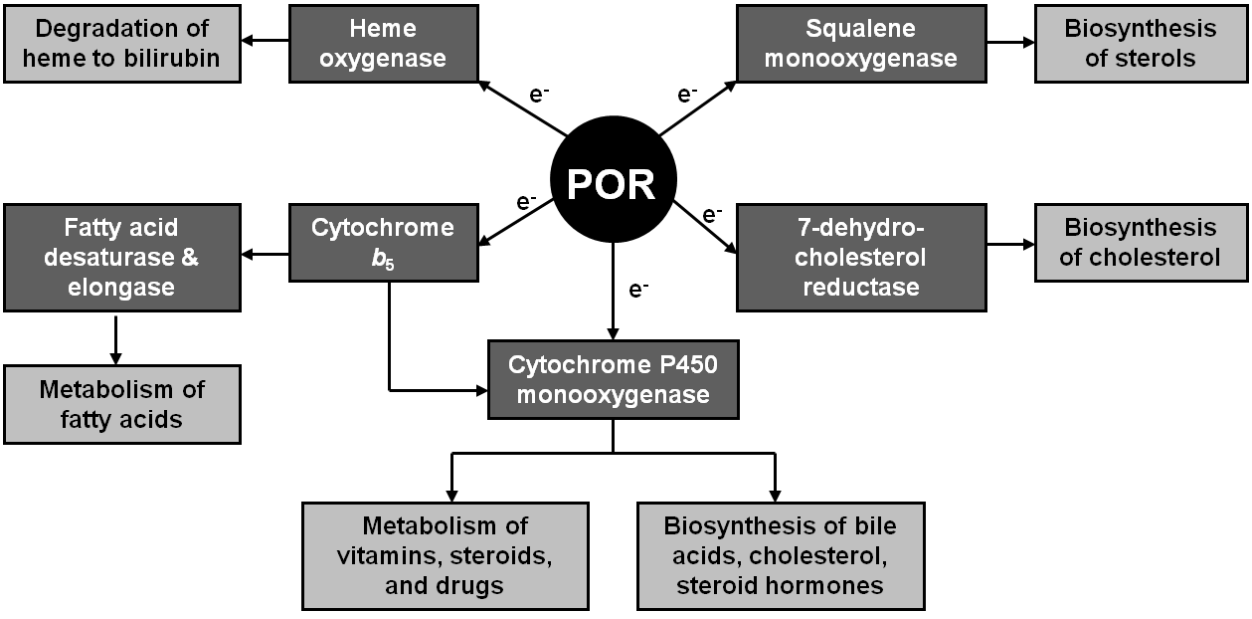


Figure 1.2. Involvement of POR as electron donor in various physiological functions.

The general catalytic cycle scheme for electron transfer from POR to CYP enzymes is shown in Figure 1.3. Briefly, CYPs containing ferric (Fe^{3+}) heme iron can initiate catalysis by either accepting the drug into the CYP's catalytic pocket, and then reducing to ferrous (Fe^{2+}) state by one electron reduction from POR, or first reducing to ferrous (Fe^{2+}) state by POR, and then binding to a drug (Guengerich and Johnson, 1997). Regardless of either initiating sequence, after the first electron reduction, a molecular oxygen binds ferrous iron, and then a second electron must be donated by POR, or, in some situations, cytochrome b_5 (Hildebrandt and Estabrook, 1971; Noshiro et al., 1981). For a long time, cytochrome b_5 was thought to only be able to donate one of the two electrons needed for P450-mediated catalysis, although this idea has been challenged recently (Finn et al., 2008). Subsequent steps include the introduction of a proton, cleavage of the O-O bond, abstraction of the hydrogen, product formation, and release. Typical and atypical CYP-catalyzed reactions have been thoroughly reviewed by Guengerich (2001).

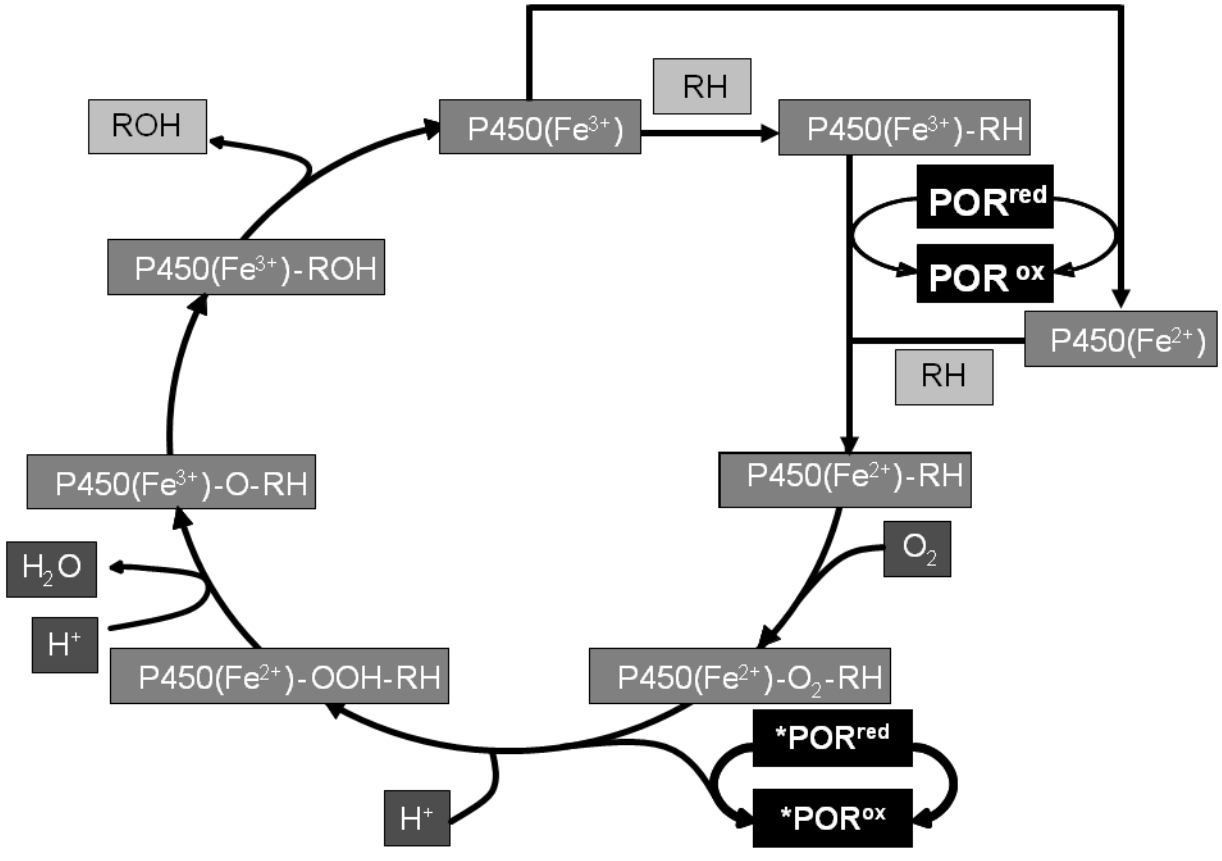


Figure 1.3. Electron donations by POR in a catalytic cycle of cytochrome P450-mediated drug oxidation. P450 = Cytochrome P450, Fe = heme iron, RH = substrate (drug), ROH = oxidized substrate (drug). *For some CYP isozymes, cytochrome b5 can act as an electron donor.

Not only is POR necessarily interesting for CYP activation, but the process by which electrons are prepared for transfer is also quite unique. The crystal structure of the rat POR protein (Wang et al., 1997) shows that the POR protein contains five structural domains, a transmembrane anchoring domain, a hinge domain, and three binding domains to the cofactors of nicotinamide adenine dinucleotide phosphate reduced form (NADPH), flavin adenine dinucleotide (FAD), and flavin mononucleotide (FMN), independently. Through a series of macromolecular motions, NADPH initially binds to POR and donates a hydride ion to FAD (Hubbard et al., 2001). Once reduced by the two electrons, FAD is converted to the dihydroquinone state (FADH2). Electrons are then transferred sequentially from FAD to FMN through inter-domain electron transfer (Gutierrez et al., 2001), during which time the FMN cycles between the semiquinone (FMNH•) and dihydroquinone (FMNH2) states. This action is proposed to be gated by a tryptophan residue in the C-terminus of the POR protein (Gutierrez et al., 2002) that sterically impedes electron flow during the macromolecular shifting in primary binding events, such as NADPH (Gutierrez et al., 2002; Grunau et al., 2006), and also potentially secondary binding events, such as POR-CYP interactions (Sue Masters and Marohnic, 2006). More detailed information on the kinetics of electron transfer can be found elsewhere (Gutierrez et al., 2003).

The possibility of POR as a potential rate-limiting step in CYP-mediated drug metabolism was initially considered in 1969 (Gigon et al., 1969; Ullrich, 1969), but it was generally disregarded as a rate-limiting step because all CYP isoforms would be catalytically activated at the same rate (Guengerich, 2001). However, this assumption did not consider that the POR gene could be polymorphically expressed and POR protein might have genetic variations. The question remained as to if a biological system containing limited or functionally

disrupted POR has altered physiological functions controlled by microsomal CYPs. This question has been tested in animal models.

Recently, POR knock-out mice have been engineered, which are embryonically lethal, giving rise to multiple developmental defects such as neural tube, cardiac, eye, and limb abnormalities, general growth retardation, and vascular defects (Shen et al., 2002; Otto et al., 2003). These irregularities are thought to be caused by a number of factors, including elevated retinoic acid levels due to the loss of *Cyp26* activity (Otto et al., 2003), which normally metabolizes all-trans retinoic acids into hydroxyretinoic acids and 4-oxoretinoic acids for elimination. However, diets with low retinoic acids can only partially rescue the phenotypes (Otto et al., 2003; Ribes et al., 2007). In the case of liver-specific deletion of the POR gene, mice are reproductively and morphologically normal, but they show a profound decrease in the metabolism of steroids and drugs (Gu et al., 2003; Henderson et al., 2003). Interestingly, hepatocyte-specific POR knockout mice demonstrate a compensatory 5-fold increase in total CYP content, indicating a negative feedback pathway regulating CYP gene expression (Henderson et al., 2003). Compensatory changes are also observed in extrahepatic tissues such as the ileum, jejunum, and colon in the mouse model (Mutch et al., 2007).

Since engineered disruption of the POR gene has shown significant influence on physiological functions in mice, it begs the question, "What would happen in a human?"

1.2.2.2. Cause of a human disease, POR deficiency, by severe POR mutations

The gene encoding the human POR is genetically polymorphic. Located on chromosome 7q11.2 (Shephard et al., 1989), the POR gene (GeneID 5447 in the National Center for Biotechnology Information database, NCBI) is a 71753-bp gene (NT 007933) containing 16

exons that transcribes a 2509-bp mRNA (NCBI NM_000941.2) and encodes an 82-kDa membrane-bound protein with 680 amino acids (NCBI NP_000932.3). Currently, the NCBI dbSNP database (build 128) has reported ~320 single nucleotide polymorphisms (SNPs) in the 72-kb genomic region (4.4 SNPs per 1 kb, higher than 0.8 SNPs per 1 kb, an estimate of the average density of SNPs in human genome (Zhao et al., 2003)). Fifteen of these SNPs are located in the exonic regions, in which 8 are synonymous and 7 are nonsynonymous. Five of the SNPs, rs10262966 (G5G), rs1135612 (P129P), rs2228104 (A485A), rs1057868 (A503V), and rs1057870 (S572S), have minor allele frequencies of more than 10% in at least one examined ethnic population (http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?locusId=5447&chooseRs=all).

Genetic mutations in the POR gene cause an autosomal recessive genetic disease, P450 oxidoreductase deficiency (Huang et al., 2005; Flück and Miller, 2006; Arlt, 2007; Krone et al., 2007). Its clinical phenotypes include ambiguous genitalia, congenital adrenal hyperplasia, Antley-Bixler syndrome, and polycystic ovary syndrome. These phenotypes typically link to abnormal steroid profiles with accumulation of steroid metabolites. Molecular genetic analyses of these phenotypes in 1985 first focused on steroid 17α-hydroxylase (CYP17) and steroid 21-hydroxylase (CYP21), which are CYPs involved in steroid metabolism. Deficiencies of CYP17 and CYP21 enzyme activities were observed in patients with glucocorticoid deficiency, skeletal dysplasia, and Antley-Bixler syndrome, but no mutations could be identified in the CYP17 or CYP21 genes (Adachi et al., 1999; Reardon et al., 2000). These findings suggested that a defect may exist in functional partners that interact with these CYP enzymes. Miller (1986) first hypothesized that the mutations might be in the CYP electron donor, POR, at three years before the POR gene was cloned (Shephard et al., 1989).

Flück *et al.* (2004) first reported five missense POR mutations (A287P, R457H, V492E, C569Y, and V608F) and a splicing mutation in an initial study with four patients who had disordered steroidogenesis and Antley-Bixler syndrome. Later, Arlt *et al.* (2004) identified another POR missense mutation (Y181D) in three patients who had congenital adrenal hyperplasia, and also confirmed three POR mutations (A287P, R457H, and C569Y), originally described by Flück *et al.* (2004). Furthermore, in a study with a larger patient sample size (32 individuals), Huang *et al.* (2005) identified and characterized additional missense and frameshift mutations found in patients or in other SNP databases (A115V, T142A, Q153R, P228L, M263V, R316W, G413S, Y459H, A503V, G504R, G539R, L565P, R616X, V631I, and F646del). In that study, fifteen of nineteen patients having abnormal genitalia and disordered steroidogenesis were homozygous or heterozygous for POR mutations that eliminated or dramatically decreased POR activity. The R457H mutation was found at a high allele frequency (62.5%) in a study with 10 Japanese patients from 8 families with Antley-Bixler syndrome, abnormal genitalia, and impaired steroidogenesis (Fukami et al., 2005). Four other mutations were also identified in these patients: a missense mutation (Y578C), a silent transition (G5G), a 1-bp insertion (I444fs), and a 24-bp deletion (L612_W620delinsR). A distinct new disease, POR deficiency, was defined (Huang et al., 2005; Flück and Miller, 2006; Arlt, 2007; Krone et al., 2007). So far, all POR deficiency associated mutations were found only once in a single patient, except A287P, R457H, and C569Y. Interestingly, the G5G polymorphism was found to be associated with an increased risk of breast cancer in African Americans (Haiman et al., 2007).

Locations of these mutations can be seen in the artistic rendition of the three-dimensional POR structure (Figure 1.4). Site-mutagenesis studies confirmed that the mutations located in the NADPH, FAD, and FMN domains had the most severe influence on electron flow within POR

and its catalytic ability to cytochrome *c* reduction and CYP-catalyzed oxidations (Huang et al., 2005). Although the most significant decreases in activity were observed in these cofactor-binding domains, the consequences of these mutations may (e.g. Y459H and V492E; (Marohnic et al., 2006)) or may not necessarily (F646del; (Huang et al., 2005)) influence cofactor binding.

POR deficiency is a very rare genetic disease. Mutations causing the POR deficiency may not be common in the general population. However, it is unclear whether genetic polymorphisms in the POR gene affect P450-catalyzed drug metabolism. Therefore, in chapter 2, we will discuss the identification of known and novel polymorphisms in the *POR* gene and assess their impact on drug metabolism.

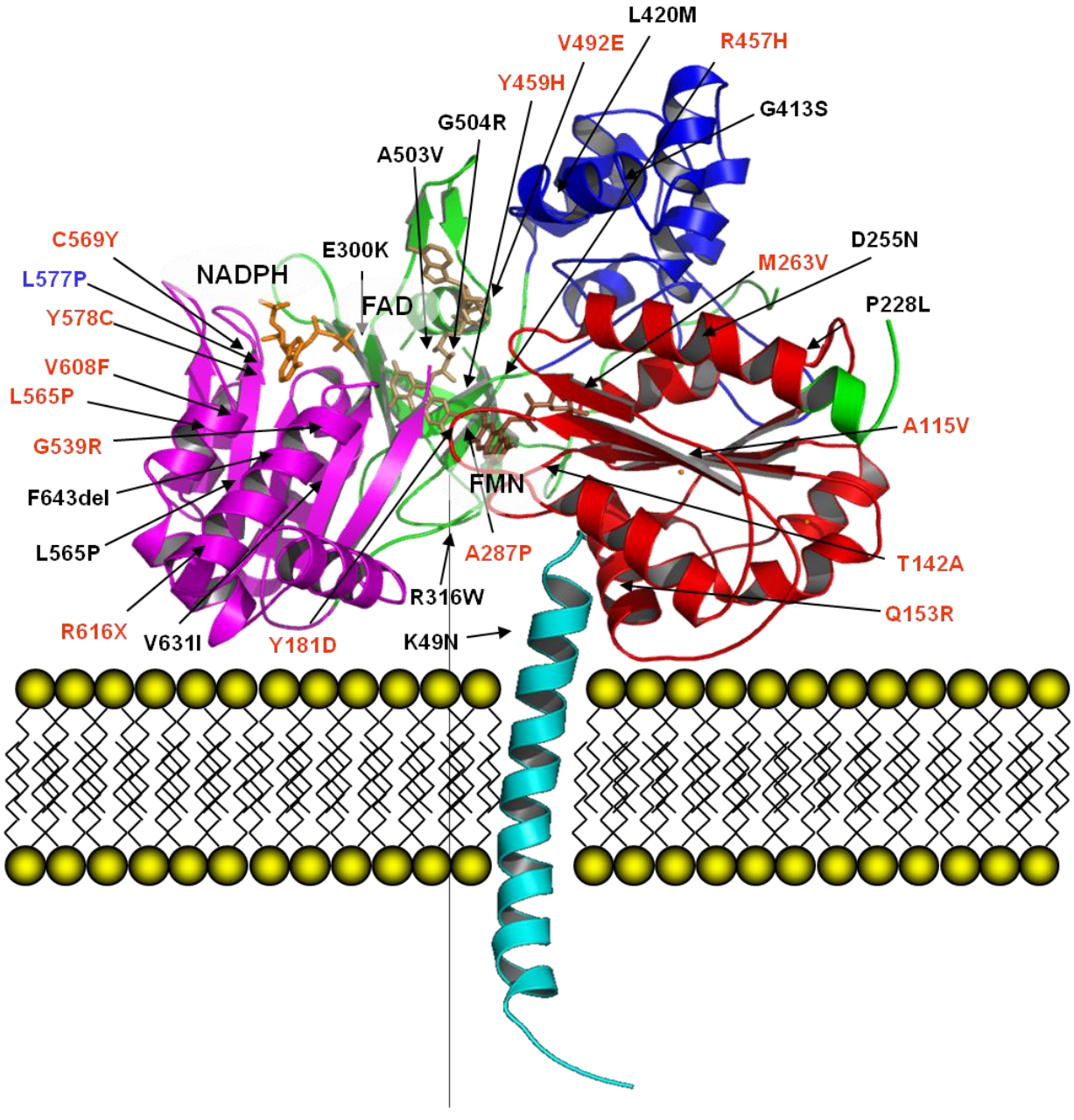


Figure 1.4. Locations of POR mutations in three-dimensional structure of the POR protein. The NADPH-binding domain is colored in light purple, FAD-binding domain is green, hinge domain is dark blue, FMN-binding domain is red, and transmembrane domain is blue-green. All reported non-synonymous mutations in POR are labeled by color. Red letters indicate mutations

or polymorphisms found in patients with POR deficiency. Blue letters are polymorphisms associated with decreased drug metabolism.

Despite our rapid understanding of how genetic polymorphisms affect POR function, *it likely represents a fractional contribution of genetically-mediated altered drug metabolism*. POR represents only a single gene amongst the tens of thousands present in the human genome. Tens of thousands of genes are capable of producing over 100,000 different proteins, many of which can interact with one another directly or indirectly, alone or in combination. As noted earlier, many genes can have genetic polymorphisms that result in altered drug metabolism, not just POR. Complicating things even more, there may be situations where combinations of genetic variation in different genes only result in an altered drug-response phenotype when both polymorphisms are present. To investigate these relationships, a systems approach is necessary for studying how genes, proteins, and genetic variation interact with one another. The first step in understanding the complexities of these interactions is to define and utilize a model system that closely resembles the target cell, tissue, or organ of interest. Therefore, the next section of this Dissertation will be the characterization of a model system, the HepaRG cell line.

1.3. A New Cell Model for Studying Drug Metabolism and Liver Toxicity: HepaRG Cells

1.3.1. What are HepaRG cells and why do we need them?

An ideal drug candidate for a clinical trial is expected to present the desired functional response on the highly selective target molecule, to have a distinct mechanism of effect, adequate bioavailability and biodistribution, and most importantly to pass the formal toxicity evaluation to demonstrate the risks for human participants in the clinical studies. To achieve the desired candidates, modern drug discovery has mapped out the road with well-delineated milestones, including selection of drug target, identification of lead compounds, establishment of

pharmacokinetic parameters, and the toxicity testing (Helfti, 2008). The broad availability of chemical libraries and automatic screening technologies has made it possible to identify the candidates for human disease targets at a rapid pace (Bajorath, 2002).

In contrast, the metabolic transformation and toxic effects of drug candidates vary from species to species, which renders interpreting metabolism studies conducted in animals very complex and challenging tasks (Olson et al., 2000). Liver becomes the center of attention because it is the principal organ involved in the biotransformation of xenobiotics and the most relevant systemic toxicity-target organ. A variety of *in vitro* human preparations, including cellular (tissue slices, suspensions and primary human hepatocyte, hepatic cell lines) and subcellular (S9 fractions, liver microsomes, recombinant enzymes) systems were established for subverting this purpose. Of these, only primary human hepatocytes - which account for 60% of liver and produce 90% of the total hepatic proteins - are able to express the entire hepatic metabolic machinery for drug metabolic studies and consequently to mimic the diverse mechanism of toxicity occurring in liver (Guillouzo, 1998). Thus, primary human hepatocytes in culture are the current system of choice for studying drug metabolism.

The widespread use of primary human hepatocytes, however, is limited by their restricted availability, functional instability with time in culture, limited lifespan and growth potential, large variability in CYP activities, and in magnitude of the responsiveness to prototypical inducers which is often greater than that *in vivo* (Guillouzo and Guguen-Guillouzo, 2008). In addition, the elaborate procedures used to isolate hepatocyte from liver samples/slices cause stress to the cells, rendering it difficult to use large batches of pre-characterized inducible cyropreserved human hepatocytes for high throughput screening tests (LeCluyse, 2001;

Rodriguez-Antona et al., 2002; Richert et al., 2006). There is still an urgent need to develop better prediction models to assess the metabolic and toxic effects of drug candidates.

To overcome these difficulties, researchers have been searching for new human liver cell lines. Currently used human liver cell lines are generally derived from hepatic tumors. Unfortunately, most of them have altered gene expression profiles, which lack most liver-specific functions. In particular, cytochrome P450 gene expression and enzyme activities are usually very low or undetectable in these human liver cells. For example, HepG2 cells, the most frequently used human liver cell line, express many CYP genes at very low levels (Sassa et al., 1987). Although some CYP genes, such as CYP1A1 and CYP3A7, are expressed in HepG2 cells (Ogino et al., 2002), these P450 members are fetal-specific and not expressed in most adult livers. These changes in gene expression may have happened in HepG2 cells after they were derived from the liver tissue of a differentiated hepatocellular carcinoma or may represent a developmental phenotype.

Recently, a new human liver cell line, HepaRG, has become available (Gripon et al., 2002). Although this cell line is derived from a female hepatocarcinoma patient, unlike other human liver cell lines, HepaRG cells express many drug processing genes at similar levels compared to primary human hepatocytes under a certain culture condition (Aninat et al., 2006). These drug processing genes encode phase I drug metabolizing enzymes (CYP1A2, 2B6, 2C9, 2E1, and 3A4), phase II enzymes (UDP glucuronosyltransferase 1 family, polypeptide A1, UGT1A1; glutathione *S*-transferase alpha 1, GSTA1; GSTA4, and GSTM1), gene regulatory proteins (aryl-hydrocarbon receptor, AHR; pregnane x receptor, PXR; constitutive androstane receptor, CAR), liver-specific proteins (albumin, haptoglobin, and aldolase B), as well as alpha-fetoprotein, glutathione-related enzymes (γ-glutamylcysteine synthase regulatory subunit, γ-

glutamylcysteine synthase catalytic subunit, glutathione synthase, and glutathione reductase), and thioredoxin. The activities of several Phase I and Phase II drug metabolizing enzymes were also comparable between HepaRG and freshly isolated human hepatocytes (Aninat et al., 2006). HepaRG cells also respond to PXR, CAR, and AhR activators, resulting in induction of CYP1A1, CYP1A2, CYP2B6, CYP2C8, CYP2C9, CYP2C19, and CYP3A4 *in vitro* (Lambert et al., 2009a; Lambert et al., 2009c).

HepaRG cells can maintain a proliferative status in an undifferentiated culture media for several weeks at sub-confluency. At confluence, and with the addition of a differentiation-inducing culture medium, HepaRG cells are capable of differentiating into biliary epithelial cells and hepatocytes (Gripon et al., 2002). The genes encoding liver-specific factors, drug-metabolizing enzymes, transporters, and transcription factors are stably expressed over a multi-week culture period.

In Chapter 3, we show the microarray gene expression profiles of HepaRG cells as compared to the HepG2 cell line, primary hepatocytes, and human liver under basal conditions. Although not perfect, HepaRG cells are more representative of the basal gene expression profile observed *in vivo* to human primary hepatocytes than HepG2.

The main interest in our lab, however, is not to understand how the liver functions under basal conditions, since this is a very broad concept. Rather, our aim is to focus in on how liver responds to exogenous stimuli, such as exposure to drugs. Therefore, the next logical step would be to treat HepaRG cells with drugs to see how they respond at the gene expression level. But the resolution of microarrays is limited. For instance, one can only expect 3-4 orders of magnitude for quantitation, which makes seeing induction of gene expression for highly expressed genes difficult. In the same sense, high background levels make it difficult to

accurately estimate expression of lowly-expressed genes. Also, the scope of information that can be extracted from arrays is limited, i.e. one can only get gene-level quantitation from probes represented on the array. That is why we used the most advanced tool for analyzing gene expression - mRNA-Seq, the details of which are described below.

1.4. A New Tool For Understanding Drug Response: Next-Generation-Based mRNA Sequencing For Pharmacogenomics

1.4.1. Next-Gen DNA Sequencing using Illumina Technology

The first publication of Illumina DNA sequencing technology was in 2006 when it was called Solexa sequencing (Fedurco et al.). To perform Illumina sequencing, libraries first need to be constructed. These libraries can consist of any fragment of DNA from any organism, as long as it also has Illumina's patented primers ligated to either end. These primers serve as a template for bridge PCR, which is essentially like doing PCR in place on a solid-phased substrate. In this manner, a single DNA can be clonally amplified generating a cluster of identical sequence fragments. Several million clusters can be amplified at discrete locations within each of eight lanes on a single flow-cell. After cluster generation, the amplicons are then directionally linearized so that step-wise incorporation of fluorescently nucleotides incorporates the same nucleotide in all fragments in each cluster. After one base is incorporated, four lasers are used to excite the flourophores in each cluster. Each of the nucleotides is labeled in such a way that they emit different fluorescence when excited at different wavelengths (e.g. A is green, C is blue, etc). An image is taken to show the x and y coordinate positions of each cluster, which

will glow their respective color. Then, the terminators are reversed on the fluorescent nucleotides and the next base is incorporated, imaged, etc.

Even more refinements of the sequencing chemistry have resulted in the application of paired-end sequencing. Paired-end differs from single-end in that after all cycles have been completed, a new primer is added to the flow cell to amplify clusters from the alternate orientation. So even though the number of fragments remains the same (the density of the clusters is fixed, so new clusters are not being added), the amount of sequence generated can double by repeating the same number of cycles as the other single-end run. Since both reads will originate from the same x and y coordinates in each lane of the flow cell, one can easily compare reading of the same fragment from two different orientations.

The usefulness of paired-end sequencing is two-fold. First, twice as many reads are sequenced - and in most cases, the more data the better. Second, sequencing from both ends retains positional information that will aid in genomic mapping. Fragments that are clonally amplified in the individual clusters are of a fixed 200-300 bp in length. A 2 × 36 bp paired-end run not only results in 72 bp of actual sequencing, but because the sequence of the first and last 36 bp is known, and it is known that those two sequences are separated by ~ 200 bp, one could infer the remaining non-sequenced portion of the fragment by looking at a reference genome to find where these two 36 bp fragments separated by ~200 bases, and assume that sequence reflects the unknown.

1.4.2. mRNA-Seq Library Prep for Defining and Quantifying Transcriptome

The typical strategy for performing mRNA-Seq experiments begins with library construction. Starting with 1 μg of total RNA, mRNA is isolated using polyA selection. The

mRNA is then fragmented and randomly primed for reverse transcription followed by second-strand synthesis to create double-stranded cDNA fragments. Ends are repaired and modified to produce blunt ends. Next, an 'A'- base is added to the blunt ends followed by ligation to Illumina Paired-End Sequencing adapters. These adapters contain unique sequencing primer hybridization sites. Additional sequences complementary to the oligonucleotides in the flow cell are added to the adapter sequences with tailed PCR primers. This is followed by gel-based size selection, purification, and amplification by PCR to create libraries for cluster generation. These libraries get hybridized to the oligos on the flow cell, which then is mounted in a cBot, an instrument that allows the bridge amplification to occur on the flow cell to generate clusters. Then, the flow cell is transferred to the sequencing instrument where it is sequenced cycle-by-cycle as described above.

1.4.3. Data Analysis for mRNA-Seq

What does data analysis mean? That actually becomes a complicated question when dealing with Illumina-based mRNA sequencing. For an example, a single mRNA-Seq experiment can yield up to 6 dimensions worth of data, compared to 1 dimension of microarray (fluorescence intensity levels representing gene expression). Not 6 *times* the amount of data, but 6 different *types* of data. With a single wet-lab experiment, one can quantify gene expression, discover and annotate new genes, identify transcript-level isoforms, quantitate different isoforms, discover fusion genes, and identify coding SNPs. Each of these features often has different requirements, which further complicates the analytical processing. For example, sequencing ~ 10 million single-end 36 bp reads is generally sufficient to quantify gene expression, but quantitating transcripts generally requires 100-150 million 2 × 75 bp reads. So, if an

experimental design for such transcript level analysis will be conducted, then gene expression will be no problem, however the reverse is not true. If one is to sequence only 25 million 1 × 36 bp, then gene expression can be calculated but very few transcripts can be annotated correctly. Therefore, ***the most critical element to an mRNA-Seq experiment is an optimal experimental design***. Data can always be computationally discarded, but cannot be computationally created, so in this case less is not more.

Although many of the data analysis procedures require different programs, assumptions, filtering parameters, and other factors to consider, all of the current methods have the same first step in data analysis. All of the reads from the instrument must be mapped to the genome. To map reads to the genome, one requires three elements: a genome to map the reads to, a tool to align the reads to that genome, and a FASTQ file containing the sequence reads and their quality metrics. The FASTQ file is described in more detail below.

1.4.4. FASTQ format

For each single-end read from a given cluster, the images are overlaid from each sequencing cycle so that a linear sequence can be deduced. The base call from each cycle for each cluster is not to be thought of as a discrete variable (i.e. A, C T, G), rather it is a probability based event (e.g. there is a 99.99% chance that this base is 'A'). In order to retain the base-calling and probabilities for each of the clusters, Illumina sequencing software output data in FASTQ format. FASTQ is a text-based format for storing both a biological sequence and its corresponding quality scores using a single character, which is necessary to reduce the file size. This is simple for the base calls - 'C' for cytosine - but is a little more complex for storing

probabilities. To overcome this, engineers have merged tricks from computer science and biology. Rather than interpret probabilities as very small numbers, biologists have introduced the PHRED scale (Ewing et al., 1998). The PHRED scale is simply , where P_e is the probability of making an error. So from our example before, if there is a 99.99% chance that this base is 'A', then the probability of making an error is $\frac{1}{10{,}000}$. Therefore, the PHRED-scaled probability of making an error would be $-10 * log\left(\frac{1}{10{,}000}\right) = 40$. The trick from computer science is to encode those PHRED-scores in a single ASCII character. In the computer world, each letter has an associated ASCII value attributed to it. For instance, 'b' is equivalent to 98, whereas 'B' is 66. There probably is some method to this madness, but for the sake of this discussion, just accept that each key represents a value. Converters for these values can be readily found on the internet.

The default output of Illumina software is an Illumina 1.3+ FASTQ variant, which encodes PHRED scores with an ASCII offset of 64. This means that the Illumina 1.3+ FASTQ variant uses different characters to represent base-qualities than standard FASTQ files (which are given by other sequencing platforms). To fix this problem, the ASCII value minus 64 will give you the correct PHRED-scaled probability. So in our previous case we showed 'B' is equivalent to 66. If this were the true value, we could reshape the PHRED calculation to $10^{\frac{PHRED}{-10}} = P_e$ and we would find that 66 is equivalent to an error 2.5 times in 10 million samples. However, if the FASTQ file is the Illumina 1.3+ FASTQ variant, then the actual probability would be 66 - 64 = 2, or 63 times in 100 samples. The difference in these two values are stark, one says that this is a highly accurate base call, while the other says it couldn't be much worse. Therefore it is critical

to know what type of FASTQ file is being analyzed so as to interpret probabilities and data quality.

1.4.5. Problems with current data analysis methods

1.4.5.1. There is no established dogma

Unlike microarrays, there is no standard approach to analysis yet. This is highlighted in the many different approaches that existing analysis tools take to analyze and interpret data. For instance, Alexa-Seq (Griffith et al., 2010) measures gene and exon expression at the junction level, whereas Cufflinks (Trapnell et al., 2010) tries to assemble all reads into transcripts and then report expression level by estimating the number of reads belonging to a given transcript. In this section, we will discuss two major flaws in current sequence analysis pipelines, namely the way reads are counted for each gene and SNP calling from RNA-Seq. We have developed new methods that make improvements in correcting for these limitations that will be discussed later.

Often times, output from data analysis tools report gene expression in terms of **R**eads (or **F**ragments) **P**er **K**ilobase of exon per **M**illion fragments mapped (RPKM or FPKM). The term FPKM is used only in cases where paired-end reads are used, since two reads are present for every fragment sequenced. In this way, RPKM are a more intuitive way of quantitating gene expression than using raw tag counts because one needs to account for depth of sequencing and length of transcript. Adjusting for the depth of sequencing is required to account for differences between sequenced libraries. For instance, if you have a 2.5 kb long transcript and 5 kb of sequence covering that transcript in condition A, and 5.2 kb in condition B, then one might

assume that there is no difference in transcript level. If one then finds out that 50 million reads were sequenced in condition A, but 60 million reads were sequenced in condition B, then the outcome will likely be different. In condition A, 5 kb sequence of a 2.5 kb transcript from a library of 50 M reads yields an RPKM of 40. In condition B, 5.2 kb sequence of a 2.5 kb transcript from a library of 60 M reads yields an RPKM of 34.7. Now it is apparent that the transcript level in condition B is much lower than condition A even though the raw coverage would suggest otherwise at the gene level.

Transcript length is also important to factor into gene expression so that relative levels within the same library can be compared to one another. In this case, assume transcript A and transcript B both contain 5 kb of sequence data. Without factoring in transcript length, one would falsely assume that these transcripts are expressed at equivalent levels. If transcript A is 5 kb long and transcript B is 1.5 kb long, then one would expect more reads in transcript A because there are more places in transcript A to map to (5/1.5 = 3.33 × more).

1.4.5.2. No tool can accurately quantify gene expression levels when multiple genes share same exons

DNA has two strands: the sense (a.k.a. forward ,+, or Watson) and the nonsense (a.k.a. reverse, -, Crick) strands. DNA can be interpreted by reading the nucleotides in a 5'- to 3'- fashoin. Consider the following sequence:

$$+ \quad AGGTCA$$

$$- \quad TCCAGT$$

The + strand reads AGGTCA, whereas the - strand reads TGACCT. The directionality of reading DNA sequence allows genes to be coded on either the + or - strands. Therefore, it is

possible for a gene on the + strand to overlap with a different gene coded on the - strand. This is to say that parts of these overlapping genes share the same chromosomal coordinates. In the event of counting the number of reads assigned to a given gene that overlaps with another, different programs can yield different results. HT-Seq (Anders *et al.*, unpublished) ignores reads lying in genes sharing chromosomal coordinates, whereas BEDtools (Quinlan and Hall, 2010) will actually count them twice, which ultimately decreases RPKM measurements.

Although the unit is conceptually easy to understand, the calculation of RPKM is also not always consistent by different tools. For example, Cufflinks (Trapnell et al., 2010) and ERANGE (Mortazavi et al., 2008) compute RPKM differently. Cufflinks uses standard annotated exon models to count reads belonging to a gene and the length of those exons are summed to give the per kb exon model information in the calculation of RPKM. ERANGE however, tries to first build new exons and then uses those sizes and counts within them into total reads and exon length, thereby influencing RPKM (Pepke et al., 2009).

1.4.5.3. High false discovery rates in SNP calling from RNA-Seq

Another major problem with RNA-Seq analysis is that there are no current tools specifically designed to call SNPs from RNA-Seq datasets. All current SNP-calling tools are based on whole genome or targeted enrichment-based sequencing. The assumptions required for accurately genotyping these experiments are quite different from the assumptions need to genotype RNA-Seq data. Unlike sequencing genomic DNA, the levels of RNA sequenced varies significantly, which is dependent on gene expression. This causes two problems. First, coverage is non-uniform. Due to paralogous genes, some reads (especially short 36bp reads) are able to be mapped to more than one place in the genome. This is particularly true if one allows flexibility

in the mapping algorithm by allowing x number of mismatches in each read - which by definition must exist if one wants to identify variations. For example if the entirety of a gene was covered about 10 ×, but one small area had 20 × - it might be possible to infer that this inconsistency in mapping to genomic DNA (gDNA) is a result of similar sequences throughout the genome - and the 'SNP' it was calling was actually the correct base in its paralogous gene. This is not the case, however when dealing with an unspecified level of expression for isoforms of paralogous genes. If one was sequencing gDNA, one would expect a 1:1 ratio for those paralogues, but it would be impossible to determine such a ratio from cDNA.

Yet another major problem when calling cSNPs is that most SNP-calling tools require (or prefer) a maximum coverage cut-off, with the idea being that if this paralogous gene mapping got out of control, then SNPs with unrealistic coverage should be discarded. However, RNAs can be expressed and many thousand times more or less than other mRNAs, so such an arbitrary calculation cannot be made.

Despite these drawbacks, others have already attempted to genotype coding SNPs (cSNPs) using RNA-Seq in human samples. Morin *et al.*, (2008) used 15 million 31 bp single end reads to profile HeLa cells. Reads were aligned to the genome using MAQ (Li et al., 2008), after building a synthetic splice junction database (since MAQ does not perform split-read alignments). In their study, Morin et al., (2008) were able to genotype 5,928 SNPs, 38% of which had not been previously reported. In another study, Chepelev *et al.,* (2009) compared the cSNP profile of Jurkat T-Cells and $CD4^+$ T-cells. This time, 27 M 30 bp single-end reads were aligned to the human genome using ELAND (again with no split-reads). Only uniquely mapped reads were used and potential PCR artifacts were removed. 12,000 cSNPs were identified in Jurkat cells while 10,000 SNPs were called in the non-tumorigenic CD4+ T-cell sample. 39% of

the Jurkat cSNPs and 28% of the CD4+ T-cell sample were novel. It is highly suspect that 28% of the total cSNPs from a healthy individual are reported as novel. Although the total number of SNPs detected from mRNA-Seq is far less from cSNPs identified from whole genome sequencing (~12,000 compared to about 75,000), the entire genome set (cSNPs, introns, intergenic, etc) from all genomes sequenced to date is around 18%. One would logically assume that exons are under higher selection pressure and therefore should harbor fewer new mutations than less conserved intergenic regions. Nevertheless, Chepelev *et al.*, (2009) note that they detected only 40% of the total homozygous cSNPs and only 14% of the heterozygous cSNPs. In another study, CD4+ T-cells from 4 individuals were evaluated for allele-specific expression (Heap et al., 2010). The authors used 20 M 2 × 45 sequencing fragments, but did not report the total number of SNPs identified or which were novel.

The most comprehensive analysis to date compares exome and transcriptome of PBMCs (Cirulli et al., 2010). Exome sequencing is a targeted sequencing approach similar to whole genome sequencing, except that probes designed to bind exons enrich those regions over background genomic DNA. However, Cirulli et al. took a different approach. They sequenced the entire genome, but compared only the exome to mRNA-Seq. In this way, one is able to get the sequences from all genes targeted without those genes being necessarily expressed. 980 M reads that were either 2 × 75bp or 1 × 75bp were sequenced from the exome and aligned by BWA (Li and Durbin, 2009). These data were then compared to 81 M reads from the transcriptome that were either 1 × 75 or 1 × 68, due to machine errors. These reads were aligned by TopHat (Trapnell et al., 2009), a split-read aligner. Although the authors were able to call 40,605 cSNPs, only 19,504 SNPs were present in cDNA and exome DNA, with a 6% novel rate. This suggests at least two major problems. First, only 6% of SNPs were novel, which means that

filtering of SNPs is too stringent, since one should expect around 20%. Second, only about 50% of the SNPs identified by RNA-Seq were validated by exome sequencing. ***This suggests that optimization of SNP-calling tools from RNA-Seq is necessary to decrease the high degree of false positives***.

Given the problems such as reporting reads in read-count level and the high false positive rate for SNP-detection described above, there exists a critical need to optimize RNA-Seq data analysis pipelines to get the most accurate information from these highly expensive information-rich experiments. Therefore, the goal of Chapter 4 is to improve existing tools for quantitating gene expression and SNP-calling. To perform such experiments and remove as much variability as possible, the framework of the experimental design will be based on understanding how HepaRG cells transcriptionally respond to drug treatment.

CHAPTER 2. A NEW TARGET GENE FOR PHARMACOGENETICS

Chapter 2.1. Reprinted with permission from the Japanese Society for the Study of Xenobiotics

Chapter 2.2. Reprinted with permission from Wolters Kluwer Health

2.1. Novel SNPs in Cytochrome P450 Oxidoreductase

2.1.1 Abstract

Cytochrome P450 oxidoreductase (POR) is the single flavoprotein which donates electrons to the microsomal cytochrome P450 enzymes for oxidation of their substrates. In this study, we sequenced all 15 exons and the surrounding intronic sequences of POR in 100 human liver samples to identify novel and confirm known genetic polymorphisms in POR. Thirty-four single nucleotide polymorphisms (SNPs) were identified including 9 in the coding exons (5 synonymous and 4 nonsynonymous), 20 in the intronic regions, and 5 in the 3'-UTR. Of these, 9 were novel SNPs, including three nonsynonymous SNPs, SNH313003 (817733G>C; K49N), SNH313020 (848661C>A; L420M), and SNH313029 (849577T>C; L577P) with minor allele frequencies of 0.005, 0.045, and 0.020, respectively. We also confirmed a previously reported non-synonymous SNP rs1057868 (A503V) as well as five synonymous SNPs (G5G, T29T, P129P, S485S, and S572S) all with allele frequencies similar to those previously reported. Structurally, these polymorphisms occur in different regions: SNH313003 (K49N) in the amino–terminal tail, SNH313020 (L420M) in the connecting domain, SNH313029 (L577P) in the NADPH-binding domain, and rs1057868 (A503V) in the FAD binding domain.

2.1.2. Introduction

The human genome contains 57 genes encoding cytochrome P450 (CYP) enzymes, of which 7 are Type I and 50 are Type II. Type I enzymes, found in the mitochondria, receive electrons from NADPH through ferredoxin reductase; whereas, Type II enzymes, located in the endoplasmic reticulum, receive electrons from NADPH through a single protein, cytochrome P450 oxidoreductase (POR). Type II P450 enzymes perform a variety of functions: 20 are involved in metabolism of steroid, fatty acids, and bile acids, 25 have "orphan" classification, and 15 catalyze drug metabolism (Guengerich, 2006). These drug-metabolizing Type II P450 enzymes are the primary enzymes, metabolizing more than 80% of current prescription drugs (Evans and Relling, 1999). In some P450 enzymes, single nucleotide polymorphisms (SNPs) have significant influence on drug metabolic rates which are critical for drugs with narrow therapeutic indices, such as warfarin and phenytoin (Schwarz, 2003; Thomas et al., 2004). However, unlike polymorphisms in a P450 enzyme which can only affect the expression level, activity, or structure of that particular enzyme, polymorphisms in a universal P450 enzyme co-factor could influence the metabolism of all drugs catalyzed by the P450 enzymes. Cytochrome P450 oxidoreductase (NP_000932.3) is a co-factor for all microsomal Type II P450 enzymes (Porter and Coon, 1991). POR transfers electrons from NADPH to a microsomal Type II P450 enzyme for oxidation of its substrates. Because there is no other electron donor for the microsomal Type II P450 system, POR is essential for drug metabolism.

The necessity of POR has been supported in the mouse by targeted gene disruption during development giving rise to different embryonically lethal phenotypes (Shen et al., 2002). Liver specific conditional knock-out studies have shown reproductively normal mice, but with a

severely diminished capacity for hepatic drug metabolism and accumulation of hepatic lipids (Gu et al., 2003; Henderson et al., 2003).

The human POR gene is quite polymorphic. Currently, the NCBI dbSNP database has reported ~320 SNPs in the 32-kb POR gene (http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?locusId=5447&chooseRs=all). Genetic polymorphisms in the POR gene have recently been associated to an autosomal recessive genetic disease, "P450 oxidoreductase deficiency" (Arlt et al., 2004; Flück et al., 2004; Huang et al., 2005; Flück and Miller, 2006) with clinical phenotypes of ambiguous genitalia, congenital adrenal hyperplasia, the skeletal malformation Antley-Bixler syndrome, and polycystic ovary syndrome. However, it is unclear whether or not genetic polymorphisms in the POR gene in the general population affect P450-catalyzed drug metabolism. Here we report the identification of 34 SNPs in the exons and surrounding introns of the POR gene in 100 human liver samples. Nine of these SNPs are novel.

2.1.3. Materials and Methods

DNA samples were isolated from 100 liver tissue lysates purchased from XenoTech (Lenexa, Kansas, USA) with ChargeSwitch® gDNA Mini Tissue kit (Invitrogen, Carlsbad, California, USA). This cohort consisted of samples from 77 Caucasians, 10 African-Americans, 10 Hispanics, and 3 Asians. The livers were initially harvested for transplantation purposes, but were not utilized and subsequently donated for research for varying reasons. Target DNA molecules were amplified by PCR. Forward and reverse primers were designed by DS Gene Software (Accelrys, Cambridge, UK). The primer sequences and PCR product sizes are listed in **Table 2.1.1**. The selected primer sequences were synthesized by Integrated DNA Technologies

(Coralville, Iowa, USA). PCR reactions were performed at cycling conditions of 95°C for 10 min, 40 cycles of 94°C for 15 seconds, 60°C for 30 seconds, and 72°C for 45 seconds, followed by 72°C for 5 min, with Go Taq Polymerase purchased from Promega (Madison, Wisconsin, USA). PCR products were purified with the Pre-sequencing Kit provided by USB (Cleveland, Ohio, USA). DNA sequencing reactions were carried out using BigDye Terminator V 3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, California, USA) with the forward primers. For dye terminator removal, PERFORMA® DTR Gel Filtration cartridges (Edge BioSystems, Gaithersburg, Maryland, USA) were used, and sequences were analyzed with a 3130 DNA Analyzer (Applied Biosystems, Foster City, California, USA).

Table 2.1.1 Primers used for PCR amplification and sequencing of the POR exons.

Amplified Region	Forward Primer (5' to 3')	Reverse Primer (5' to 3')	Fragment Size (bp)
Exon 1	GTGCAGTGACCATTTCCTG	AAACAGCAGATAGAAAAGGGC	625
Exon 2	ACAAAGCTGGAATGTCCCC	GCAGCTCTGTGAGATTTACC	505
Exon 3	TGACAGTGAGAAGCAAGTCC	GTTTGGTTTGGGAGATGTGG	580
Exon 4	TAACACGGGTGACCTTGTC	AGGAGAGGGTCTCACAAGTG	503
Exon 5	TCTTCAGTGGCCCAGTGTTC	ACCCAGCGACATAAACCCAG	518
Exon 6	CCCTGCCAGTTTTGCTTTTC	TTGAACCTAGCCACAGAGCC	533
Exon 7	TTCTCCCAGATGGAAGCCTG	GCAGAGTAAGGTGGCTAAGTG	571
Exon 8,9	GAGAGCCCTTGATGTAACCG	GCCTAAGCAGAAGCTCAACC	571
Exon 10	TGCCTCTGATGAGGACTTCC	GTACAGCTCCTAAGAGACACG	500
Exon 11	ACTACCTGGACATCACCAAC	ATGCTGAGAATCTCACAAGC	557
Exon 12	TACTCCATCGCCTCATCCTC	AAGCCTATGAAGGGTGCCAC	606
Exon 13	TGTGGAGTACGAGACCAAGG	TTAGCAGGTGCTGGACGTAG	567
Exon 14	ACCCTTCATAGGCTTCATCC	AAGGTGTTCTGCACATCCC	573
Exon 15	GATGTGCAGAACACCTTCTAC	TCTACTCACACAATACCAGGC	521

2.1.4. Results & Discussion

Genetic polymorphisms were identified by sequencing PCR amplicons from the exons and approximately 100 bp flanking intronic sequences. The sequences were compared to the reference contig for human POR from the NCBI database (NT_007933.14). We found 34 SNPs in the exons and the surrounding introns of the POR gene in the 100 human liver samples, including 9 novel SNPs (**Table 2.1.2**). As expected, we did not observe any of the missense or frameshift mutations (T142A, Q153R, Y181D, M263V, A287P, R457H, Y459H, V492E, G539R, L565P, C569Y, Y578C, V608F, R616X, F646del, I444fs, and L612W620delinsR), which have been associated with POR deficiency. Three novel nonsynonymous SNPs, 817733G>C (K49N), 848661C>A (L420M), and 849577T>C (L577P) were found with minor allele frequencies of 0.005, 0.045, and 0.020, respectively (**Figure 2.1**.1). Six other novel SNPs were observed in intron 7 (846457G>A, 846539C>T), intron 12 (848803T>A, 848832C>T), and the 3'-UTR (850151G>A, 850197G>A) all with minor allele frequencies less than 0.010. Five silent mutations (G5G, T29T, P129P, S485S, and S572S) and a missense mutation (A503V) were confirmed in these liver samples with similar allele frequencies reported in the NCBI dbSNP database (**Table 2.1.2**).

Table 2.1.2. Summary of polymorphisms detected in this study

SNP ID	NCBI SNP ID	Location in gene	DNA Position NT_007933.14	Nucleotide change	mRNA Position NM_000941.2	Amino acid change NP_000932.3	Minor Allele Frequency
SNH_313001	rs10262966	Exon 2	817601	ATGATCAACATGGGA>GGACTCCCACGTGGA	97	G5G	0.045
SNH_313002	rs412952381	Exon 2	817673	CTTTTCAGCATGACC>AGACATGATTCTGTTT	169	T29T	0.010
SNH_313003	SNP1	Exon 2	817733	TTCCTCTTCAGAAAG>CAAAAAAGAAGAAGTC	229	K49N	0.005
SNH_313004	rs1135612	Exon 5	843953	CTGAGCAGCCTGCCA>GGAGATCGACAACGC	469	P129P	0.215
SNH_313005	rs2286819	Intron 6	845039	GGTGGGGTCGGGGCA>GTGCCTGGCACCAGG			0.070
SNH_313006	rs2286820	Intron 7	845084	GCCTCCCCTGAGCCA>GCTCCCCCTCTCCTC			0.010
SNH_313007	SNP2	Intron 7	846457	CCCTGCTTCTTGTCG>ATATGTACCTGGGAC			0.005
SNH_313008	SNP3	Intron 7	846539	GGACTGACCCCTGCC>TGCTTCCCGGCCTCA			0.010
SNH_313009	rs41299517	Intron 7	845954	GGGCAGACGGCTCTA>GTGGCCACTGGTGCA			0.030
SNH_313010	rs3815455	Intron 8	846032	CACCAGACCCCGTGC>TCCCGAGTGGGTGTG			0.225
SNH_313011	rs13223707	Intron 8	847046	TGTGCAACCAGAAGC>GGTCCTTGGAGACGG			0.035
SNH_313012	rs13240147	Intron 8	847059	GCGTCCTTGGAGACA>GGAGACTCAGATCAA			0.055
SNH_313013	rs41301394	Intron 8	847079	TCAGATCAAAGCCCC>TGGCCGCTCACTGTG			0.180
SNH_313014	rs4732514	Intron 10	848274	GGGGCACCTGTTGCC>TGCAGAGCTGGCCCA			0.170
SNH_313015	rs6971082	Intron 10	848304	GGTGTCACCCCCTCC>TCGCCGCAGCCACCC			0.010
SNH_313016	rs4732515	Intron 10	848305	GTGTCACCCCCTCCC>TGCCGCAGCCACCCA			0.060
SNH_313017	rs4732516	Intron 10	848358	CAAGTCCTGCCTGTC>GTCTTCCCTGCAGAG			0.045
SNH_313018	rs2286822	Intron 11	848564	AAGGTGCGCCCCCTC>TAGCCCCCGCAACCT			0.365
SNH_313019	rs2286823	Intron 11	848572	CCCCCTCAGCCCCCA>GCAACCTCCGCCCCG			0.380
SNH_313020	SNP4	Exon 12	848661	GCAAGGAGCTGTACC>ATGAGCTGGGTGGTG	1340	L420M	0.045
SNH_313021	SNP5	Intron 12	848803	CTCATCCTCCAAGGT>AGAGGGCCGGCACTG			0.005
SNH_313022	SNP6	Intron 12	848832	CCCTGCCAGCCACAC>TGCTGGAGGCCCAGC			0.010
SNH_313023	rs41301427	Intron 12	848833	CCTGCCAGCCACACG>ACTGGAGGCCCAGCC			0.095
SNH_313024	rs2302431	Intron 12	849139	GGCAAGGGCCTCGGC>TTTGGCGGTGGAGCT			0.066
SNH_313025	rs2302432	Intron 12	849140	GCAAGGGCCTCGGCG>TTGGCGGTGGAGCTC			0.035
SNH_313026	rs2228104	Exon 13	849229	TACGAGACCAAGGCT>CGGCCGCATCAACAA	1537	S485S	0.076
SNH_313027	rs1057868	Exon 13	849282	GGCCAAGGAGCCTGC>TCGGGGAGAACGGCG	1590	A503V	0.217
SNH_313028	rs1057870	Exon 14	849563	GGCTGCCGCCGCTCG>AGATGAGGACTACCT	1798	S572S	0.306
SNH_313029	SNP7	Exon 14	849577	GGATGAGGACTACCT>CGTACCGGGAGGAGC	1812	L577P	0.020
SNH_313030	rs41302345	3'-UTR	850126	CTGTAATCAGCTCTC>TCTGGCTCCCTCCCG	2176		0.005
SNH_313031	SNP8	3'-UTR	850151	CCCGTAGTCTCCTGG>AGTGTGTTTGGCTTG	2201		0.005
SNH_313032	SNP9	3'-UTR	850197	AGGCCCAGTGACAAA>GGACTCCTCTGGGCC	2246		0.005
SNH_313033	rs41302348	3'-UTR	850282	CAGCCCAGGGCCTGC>GATGGGGGCACCGGG	2332		0.010
SNH_313034	rs17685	3'-UTR	850381	CTCACTGGAAATCAC>TGTGGAGGGGCTGGG	2431		0.195

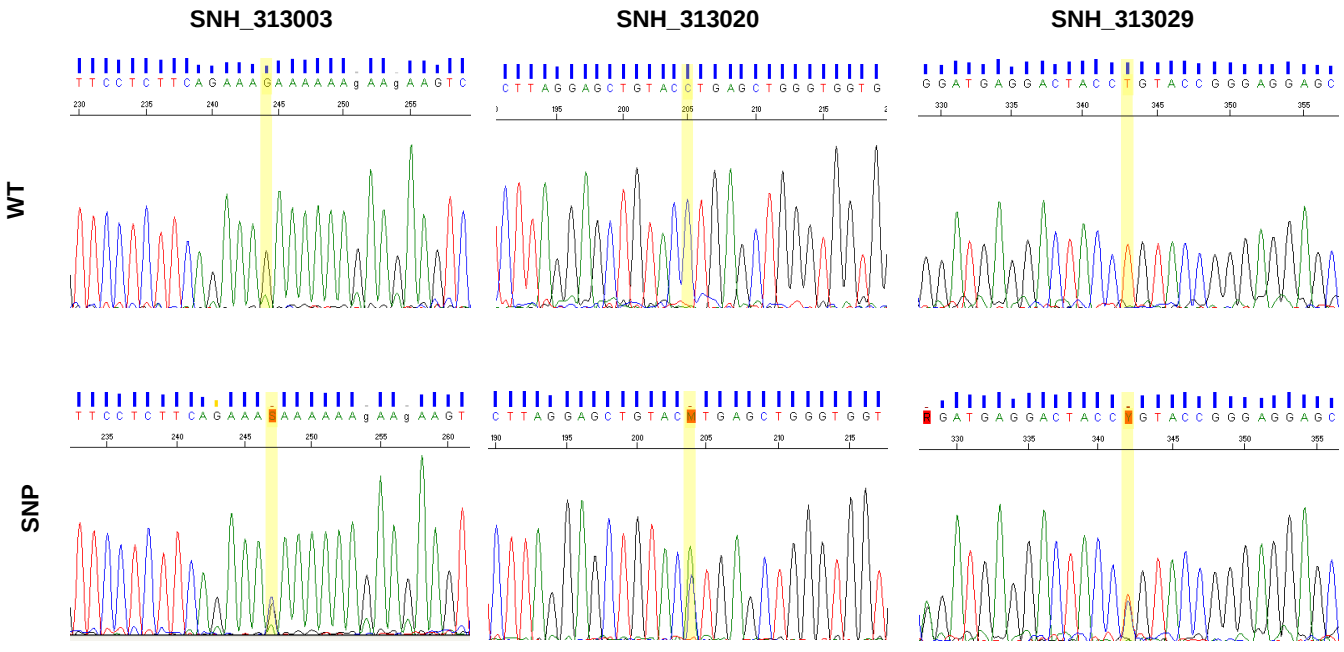


Figure 2.1.1. Electropherograms (sense strands) for the three novel non-synonymous SNPs: SNH313003 (817733G>C; K49N), SNH313020 (848661C>A; L420M), and SNH313029 (849577T>C; L577P). Wild type homozygous alleles are shown in the top row while the bottom row shows heterozygous mutants. The SNP of interest is highlighted in yellow.

Based on the crystal structure of rat POR protein (Wang et al., 1997), which is 92% identical to the human POR, we predicted the locations of all identified POR polymorphisms with respect to the different functional domains. We found one sample with a heterozygous G>C transversion at 817733, resulting in a missense mutation K49N. This missense mutation is located in the amino-terminal tail of POR. This terminal region is responsible for anchoring POR into the endoplasmic reticulum or into the plasma membrane (Osada et al., 2002) and is important for proper electron transfer function (Black et al., 1979; Bonina et al., 2005). We found nine heterozygotes for a C>A transversion at 848661, which results in a leucine to methionine change at amino acid 420 (L420M). This amino acid lies in the connecting domain. This domain is responsible for efficient electron transport (Wang et al., 1997). Four heterozygotes were found with a T>C transition at 849577, leading to an amino acid change L577P, which is located in the NADPH-binding domain of POR (Wang et al., 1997). An amino acid substitution in this region could potentially affect binding kinetics for NADPH. Once NADPH releases its electrons, it becomes $NADP^+$, an unusable metabolite. POR must release $NADP^+$ in order to bind NADPH for electron cycling to continue. Neighboring mutations (G539R, L565P, C569Y, Y578C, and V608F) in the NADPH binding domain have been identified in POR deficiency patients (Arlt et al., 2004; Flück et al., 2004; Fukami et al., 2005). All these mutations showed significantly reduced POR activity in cytochrome *c* reduction assays (Huang et al., 2005). Twenty-nine samples were heterozygous for a C>A transversion at 849282 (rs1057868) and 6 were homozygous. This A503V mutation was first uploaded to the NCBI dbSNP database in 2000, and was confirmed by the NIEHS Environmental Genome Project in 2006. This mutation occurs in the FAD binding domain of POR. An *in vitro* experiment showed

that cytochrome c reduction by POR carrying this mutation was decreased 31% compared to wild type (Huang et al., 2005).

2.2. Genetic polymorphisms in cytochrome P450 oxidoreductase influence microsomal P450-catalyzed drug metabolism

2.2.1. Abstract

Cytochrome P450 oxidoreductase (POR) is the only flavoprotein that donates electrons to all microsomal P450 enzymes, which catalyze biosynthesis of steroids, fatty acids, and bile acids, as well as metabolism of more than 80% of prescription drugs. Although mutations of POR have been identified in several disease states with disordered steroidogenesis, effects of polymorphisms on drug metabolism in the general population are unclear. In this report, we performed a comprehensive study to correlate POR polymorphisms with POR gene expression, POR activity, and P450-catalyzed drug metabolism. A set of human liver samples (n=99) were used in this study. POR polymorphisms were identified by sequencing the exons and surrounding introns of the POR gene and mRNA levels were quantified by branched DNA technology. POR activity was quantified by quantifying cytochrome *c* reduction in liver microsomes and activities of ten drug-metabolizing P450 enzymes were quantified by HPLC methods with drugs known to be specific for each enzyme. Of the 34 polymorphisms identified in this cohort, four polymorphisms changed an amino acid: K49N, L420M, A503V, and L577P. L577P likely resulted in alpha helix changes, possible disruption of the NADPH interaction, and decreased POR activity (p=0.003) and several drug-metabolizing P450 activities. We also found an intronic polymorphisms rs41301427, which is associated with altered POR, but not P450 activities. Polymorphisms in the POR gene can affect POR and P450-catalyzed drug oxidation.

2.2.2. Introduction

Microsomal P450 enzymes are heme-containing proteins that catalyze biosynthesis of steroids, fatty acids, and bile acids (Guengerich, 2004), as well as metabolism of more than 80% of prescription drugs (Evans and Relling, 1999). In the last few decades, pharmacogenomic studies have revealed that genetic polymorphisms and/or mutations can affect P450-catalyzed drug metabolism in various ways. Importantly, the distinction between mutation and a polymorphism has only to do with how frequent the allele is in a given population. To be classed as a polymorphism, the least common allele must have a frequency of 1% or more in the population, otherwise, the allele is regarded as a mutation. One way either of these variations in a P450 enzyme can affect the metabolic rates for drugs oxidized by that P450 enzyme. For example, mutations in CYP2C9 can decrease warfarin metabolism, leading to hemorrhagic complications (Rettie and Tai, 2006; Wadelius and Pirmohamed, 2007). Second, because several P450 enzymes share the same mechanisms for activation, suppression, and regulation, then genetic polymorphisms in co-activators, co-suppressors, or regulators may affect metabolic capacity of P450 enzymes, which could in turn influence a larger set of drugs. Genetic polymorphisms in the nuclear receptor pregnane x receptor (PXR) (Lamba et al., 2005) and the membrane transporter multidrug resistance 1 (MDR1) (Lamba et al., 2006) are two such examples. PXR mutations have been shown in vivo to decrease midazolam clearance (He et al., 2006). MDR1 variations have been associated with drug response to anthracyclines and taxanes (Kafka et al., 2003). Third, all microsomal P450 enzymes require co-factors for their functions. Genetic polymorphisms in the co-factor genes may influence metabolic rates of all P450-catalyzed drugs. Cytochrome P450 oxidoreductase (POR) is one such co-factor.

POR is the only flavoprotein that donates electrons to microsomal P450 enzymes (Porter and Coon, 1991). Oxidation of drugs by the P450s requires two sequential one-electron donations, but the source of these electrons comes from nicotinamide adenine dinucleotide phosphate (NADPH), which gives up a pair of electrons. POR compensates for this discrepancy by stabilizing the one-electron reduced form of the flavin cofactors of flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN). Electrons pass from NADPH through FAD to FMN in the POR protein. Following a conformational change, the FMN binding domain of the POR interacts with the oxidation/reduction-partner binding site of the P450 enzymes so that electrons reach the P450 heme iron (Fe^{3+}) to achieve catalysis of drug oxidation.
It is reasonable to assume that disruption of electron flow in the POR protein would have destructive effects on oxidation of drugs by all microsomal P450 enzymes. This assumption has been supported by studies in animal models. POR knock-out mice are embryonically lethal, giving rise to multiple developmental defects (Shen et al., 2002; Otto et al., 2003). Mice with liver-specific deletion of POR are reproductively and morphologically normal, but they show a profound decrease of capabilities in the metabolism of steroids and drugs (Gu et al., 2003; Henderson et al., 2003; Wu et al., 2003).

The gene encoding human POR is quite genetically polymorphic. Located on chromosome 7q11.2 (Shephard et al., 1989), the POR gene (GeneID 5447 in the National Center for Biotechnology Information database, NCBI) is a 71753-bp gene (NT 007933) containing 15 exons that transcribe a 2509-bp mRNA (NCBI NM_000941.2) and encodes an 82-kDa membrane-bound protein with 680 amino acids (NCBI NP_000932.3). Currently, the NCBI dbSNP database has reported ~320 SNPs in the 72-kb genomic region (4.4 SNP per 1 kb, higher than 0.8 SNP per 1 kb, an estimate of the average density of SNPs in human genome (Zhao et al.,

2003)). Fifteen of these SNPs are located in the exonic regions, in which 8 are synonymous and 7 are nonsynonymous. Five of the SNPs, rs10262966 (G5G), rs1135612 (P129P), rs2228104 (A485A), rs1057868 (A503V), and rs1057870 (S572S), have minor allele frequencies of more than 10% in at least one examined ethnic population (http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?locusId=5447&chooseRs=all).

Genetic polymorphisms in the POR gene have recently been associated with an autosomal recessive genetic disease, P450 oxidoreductase deficiency (Arlt et al., 2004; Flück et al., 2004; Fukami et al., 2005; Huang et al., 2005; Flück and Miller, 2006; Krone et al., 2007). Its clinical phenotypes include ambiguous genitalia, congenital adrenal hyperplasia, Antley-Bixler syndrome, and polycystic ovary syndrome. These phenotypes typically link to abnormal steroid profiles with accumulation of steroid metabolites. Molecular genetic analyses first focused on steroid 17α-hydroxylase (CYP17) and steroid 21-hydroxylase (CYP21), which are P450 enzymes involved in steroid metabolism. Deficiencies of CYP17 and CYP21 were observed in patients with glucocorticoid deficiency, skeletal dysplasia and Antley-Bixler syndrome, but no mutations in the CYP17 and CYP21 genes could be identified (Adachi et al., 1999; Reardon et al., 2000). These findings suggested a defect in a cofactor that interacts with these P450 enzymes. Flück *et al.* (2004) first reported five missense POR mutations (A287P, R457H, V492E, C569Y, and V608F) and a splicing mutation in an initial study with four patients who had disordered steroidogenesis and Antley-Bixler syndrome. Later Arlt *et al.* (2004) identified another POR missense mutation (Y181D) in three patients who had congenital adrenal hyperplasia, and also reported three POR mutations (A287P, R457H, and C569Y) originally described by Flück *et al* (2004). Furthermore, in a study with a larger patient sample size (32 individuals), Huang *et al.* (2005) identified additional missense and frameshift mutations

(A115V, T142A, Q153R, P228L, M263V, R316W, G413S, Y459H, A503V, G504R, G539R, L565P, R616X, V631I, and F646del) in the POR gene and recognized a distinct new disease: POR deficiency. In that study, fifteen of nineteen patients having abnormal genitalia and disordered steroidogenesis were homozygous or heterozygous for POR mutations that eliminated or dramatically decreased POR activity. The R457H mutation was found at very high allele frequency (62.5%) in a study with 10 Japanese patients from 8 families with Antley-Bixler syndrome, abnormal genitalia, and impaired steroidogenesis (Fukami et al., 2005). Four other mutations were also identified in these patients: a missense mutation (Y578C), a silent transition (G5G), a 1-bp insertion (I444fs), and a 24-bp deletion (L612_W620delinsR). The mutations of Y181D, A287P, R457H, V492E, and V608F also significantly increased cytotoxicity in cultured Chinese hamster ovary cells induced by paraquat, a widely used herbicide (Han et al., 2006), and mitomycin C, a highly active anticancer prodrug (Wang et al., 2007).

POR deficiency is a very rare genetic disease. Mutations causing the POR deficiency may not be common in the general population. However, it is unclear whether genetic polymorphisms in the POR gene affect P450-catalyzed drug metabolism. Recently, we identified novel SNPs in the POR gene in subjects without POR deficiency (Hart et al., 2007). In this report, we performed a comprehensive study to establish correlations of genetic polymorphisms in the POR gene with POR gene expression, POR activity, and POR-assisted P450 activities using a set of human liver tissue samples. Our data suggest that genetic polymorphisms in the POR gene may influence P450-catalyzed drug metabolism.

2.2.3. Materials and Methods

2.2.3.1. Human livers

Human liver tissue samples (n=99) were purchased from XenoTech LLC (Lenexa, Kansas, USA) in the form of three 5 mL lysates in DNA, RNA, and microsome isolation buffers. The samples were acquired by XenoTech through the Midwest Transplant Network (Westwood, Kansas, USA), the National Disease Research Interchange (Philadelphia, Pennsylvania, USA) and the Anatomical Gift Foundation (Woodbine, Georgia, USA). Livers were initially harvested for transplantation purposes, but were not used for various reasons and subsequently were donated for research. The livers were cooled immediately after procurement with a cold perfusion solution and frozen within 1 to 36 hours. All liver samples were tested for, and declared free of infectious agents, including human immunodeficiency virus (HIV), hepatitis B (HBV), and hepatitis C (HCV). Demographic information such as gender, age, ethnicity, and confounding factors are listed in **Table 2.2.1**.

Table 2.2.1. Demographic information of confounding factors in the human liver cohort

Confounding factors	**Distributions (n)**
Gender	Male (59) and female (40)
Ethnicity	Caucasian (77), African-American (9), Hispanic (10), and Asian (3)
Age	Year 0-1, (4); year 1-18, (7); year 18-45, (27); year 45-60, (39); and year older than 60, (22)
Smoking	Non-smoker (61) and smoker (38)
Alcohol drinking	Non-drinker (47), drinker (51), and unknown (1)
Reason for death	Anoxia (18), aortic aneurysm (1), cerebrovascular aneurysm (61), head trauma (14), myocardial infarction (3), and motor vehicle accident (2).
CMV infection	Negative (67), positive (30), and not determined (2)

CMV: cytomegalovirus

2.2.3.2. POR and P450 activities

P450 enzyme profiles of this liver cohort were characterized by XenoTech LLC. Liver microsomes were prepared by using differential ultracentrifugation (Dayer et al., 1987). The rate of cytochrome *c* reduction by liver microsomes was determined spectrophotometrically based on a previously described method (Phillips and Langdon, 1962) with some modifications (Pearce et al., 1996). The reaction was conducted in a 1 ml solution with 50 µM cytochrome *c*, 100 µM NADPH, and ~50 µg liver microsomal protein at room temperature for 10 min. The rate of cytochrome *c* reduction was determined from the rate of increase in absorbance at 550 nm by reduced form of cytochrome *c* with a DW2C dual beam spectrophotometer (SLM-Aminco, Urbana, IL). P450 enzyme activities were determined by measuring the rates of the following reactions with a spectrofluorometer or High-Performance Liquid Chromatography (HPLC) according to previously described procedures: 7-ethoxyresorufin *O*-dealkylation (CYP1A2) (Pearce et al., 1996), coumarin 7-hydroxylation (CYP2A6) (Pearce et al., 1996), *S*-mephenytoin *N*-demethylation (CYP2B6) (Pearce et al., 1996; Ko et al., 1998), paclitaxel hydroxylation (CYP2C8) (Robertson et al., 2000), diclofenac 4'-hydroxylation (CYP2C9) (Robertson et al., 2000), *S*-mephenytoin 4'-hydroxylation (CYP2C19) (Pearce et al., 1996), dextromethorphan *O*-demethylation (CYP2D6) (Pearce et al., 1996), chlorzoxazone 6-hydroxylation (CYP2E1) (Pearce et al., 1996), testosterone 6β-hydroxylation (CYP3A4/5) (Pearce et al., 1996), and lauric acid 12-hydroxylation (CYP4A9/11) (Pearce et al., 1996). With the substrate concentrations and amount of liver microsomal protein used in each reaction (**Table 2.2.2**), the probe drugs are considered to be specific for each P450 enzyme.

Table 2.2.2. The POR and P450 enzyme activities in the human liver cohort

Enzyme and assay	Substrate concen-tration[#] (μM)	Liver microsomal protein (μg) used in each reaction	Enzyme activity (pmol/mg protein/min)			
			Highest	Lowest	Mean	SD
POR: Cytochrome *c* reduction	**50**	**502**	**343000**	**505**	**177382**	**53185**
CYP1A2: 7-ethoxyresorufin *O*-dealkylation	10	100	258	NPD	51	45
CYP2A6: Coumarin 7-hydroxylation	50	100	7310	5	984	1167
CYP2B6: *S*-mephenytoin *N*-demethylation	400	400	1280	10	107	160
CYP2C8: Paclitaxel 6a-hydroxylation	10	50	2040	NPD	341	340
CYP2C9: Diclofenac 4'-hydroxylation	100	50	5870	127	1810	867
CYP2C19: *S*-mephenytoin 4'-hydroxylation	400	400	895	NPD	101	141
CYP2D6: Dextromethorphan *O*-demethylation	80	500	1160	18	308	192
CYP2E1: Chlorzoxazone 6-hydroxylation	500	200	11500	201	2109	1505
CYP3A4/5: Testosterone 6β-hydroxylation	250	200	20300	NPD	3421	3670
CYP4A9/11: Lauric acid 12-hydroxylation	100	400	4350	NPD	1699	818

#The substrate concentration is near the 10 Km for the reaction and has been shown to be appropriate for metabolite formation.

NPD: no catalyzed product detected.

SD: standard deviation.

2.2.3.3. Sequencing the POR gene

Genomic DNA was isolated from liver tissue using the ChargeSwitch® gDNA Mini Tissue Kit from Invitrogen (Carlsbad, California, USA), following the manufacturers protocol. Exonic regions of POR were amplified by PCR from the genomic DNA using forward and reverse primers designed by DS Gene Software (Accelrys, Cambridge, UK). Primer sequences and PCR product sizes have been previously reported by our lab (Hart et al., 2007), that flank the exons and ~200bp worth of introns. Primers were synthesized by Integrated DNA Technologies (Coralville, Iowa, USA) and the subsequent PCR reactions were performed using Go Taq® DNA Polymerase (Promega, Madison, Wisconsin, USA), with cycling conditions of 95°C for 3 min, 40 cycles of 94°C for 15 sec, 60°C for 30 sec, and 72°C for 45 sec, followed by 72°C for 5 min. PCR products were purified with the Pre-sequencing Kit provided by USB (Cleveland, Ohio, USA). DNA sequencing reactions were carried out using BigDye Terminator V 3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, California, USA) with the forward primers. For dye terminator removal, PERFORMA® DTR Gel Filtration cartridges (Edge BioSystems, Gaithersburg, Maryland, USA) were used, and sequences were analyzed with a 3130 DNA Analyzer (Applied Biosystems, Foster City, California, USA).

2.2.3.4. POR gene expression

Total RNA was isolated from each liver tissue with TRIZOL® reagent (Invitrogen) following the manufacturer's protocols. Human POR mRNA levels were assessed by branched DNA technique (Hartley and Klaassen, 2000; Czerwinski et al., 2002) with a Quantigene expression kit (Bayer, Walpole, Massachusetts, USA) as described in the manufacturer's

protocol. The specific oligonucleotide probe sets (capture extenders, label extenders, and blockers) were designed to have a melting temperature around 65°C using the probe Designer software, version 1.0 (Bayer, Emeryville, California, USA). A luminescent readout was measured with a Quantiplex 320 Luminometer (Chiron Corp., Emeryville, California, USA).

2.2.3.5. Prediction of protein structure changes

Amino acid sequences from POR homologs of human (NCBI NP_000932.3), rat (NP_113764.1), frog (*Xenopus*, AAH59318.1), fruit fly (*Drosophila*, NP_477158.1), and yeast (NP_596046.1) were aligned and displayed by ClustalW (Thompson et al., 1994). Molecular modeling of the identified non-synonymous mutations were performed with ESyPred3D (Lambert et al., 2002) using the crystalline rat *Por* structure (PDB:1AMO) chain 'A' as a template. The rat *Por* protein shares 92.1% identity and 96.3% similarity with the human POR in alignment by the Needle EMBOSS pairwise tool (Needleman and Wunsch, 1970). Models were visualized and displayed using PyMOL (Warren). Protein secondary structure and membrane topology for the POR mutations were predicted by the Quick2D analysis tool with PSIPRED (Jones, 1999) and MEMSAT2 (Jones et al., 1994).

2.2.3.6. Statistical analysis

We used multiple linear regressions to assess the effect of POR SNPs on POR enzyme activity. Age, gender, ethnicity, reason for death, smoking history, drinking history, and cytomegalovirus infection were included in the model to adjust for potential confounding effects.

The same linear regression was also applied to assess the association between POR gene expression and POR enzyme activity.

2.2.4. Results

2.2.4.1. Interindividual variation of POR enzyme activity

Interindividual variation of the POR enzyme activity was observed in this liver cohort. Although cytochrome *c* has been shown not to be the natural substrate of POR as initially thought (Horecker, 1950; Williams and Kamin, 1962), cytochrome *c* reduction is still used by many researchers in the field to quantify POR activity (Reardon et al., 2000; Flück et al., 2004; Huang et al., 2005). Cytochrome *c* reduction was also used to quantify the POR activity in this study. **Figure 2.2.1** shows the histogram of POR activity with a mean of 177 nmole/mg protein/min and a standard deviation of 53. There were 75 individuals who had POR activity within one standard deviation (SD) around the mean, which was considered as normal POR activity. There were 10 individuals who had POR activity between -1SD and -2SD, which was considered low POR activity. One individual had extremely low POR activity (<-2SD). In contrast, 10 individuals had high POR activity (between +1SD and +2SD) and 3 had extremely high POR activity (>+2SD). Between +2SD and -2SD, there is about a 4-fold difference.

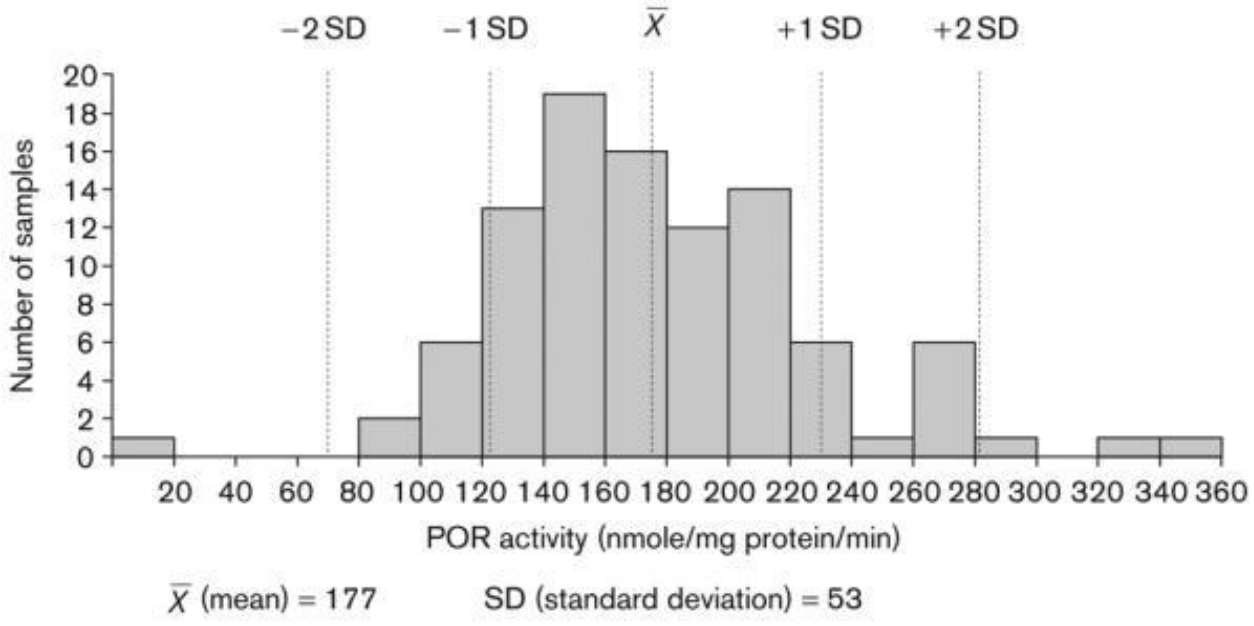


Figure 2.2.1. Distribution of POR activity quantified by measuring cytochrome *c* reduction in the liver cohort.

2.2.4.2. Correlation of enzyme activities between POR and P450 enzymes

Significant associations between the activity of POR and activities of most drug-metabolizing P450 enzymes were observed in this study. Activities of ten P450 enzymes were quantified in the 99 liver microsomes using probe drugs known to be specific for each enzyme. **Table 2.2.2** lists descriptive statistics for each P450 enzyme activity. Pearson's correlations between POR activity and a P450 enzyme activity together with the *p* value (testing correlation equals 0) are summarized in **Table 2.2.3**. Seven of the P450 enzyme activities (CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2E1, CYP3A4/5, and CYP4A9/11) correlated with POR activity at significant levels of $p<0.001$. CYP2D6 correlated with POR activity at a significant level of $p<0.05$. Only CYP1A2 and CYP2C19 did not significantly correlate with POR ($p>0.05$). As examples, **Figure 2.2.2** shows the scatter plots of enzyme activity of POR versus CYP4A9/11 ($p<0.001$), CYP2D6 ($p=0.043$), and CYP2C19 ($p=0.460$).

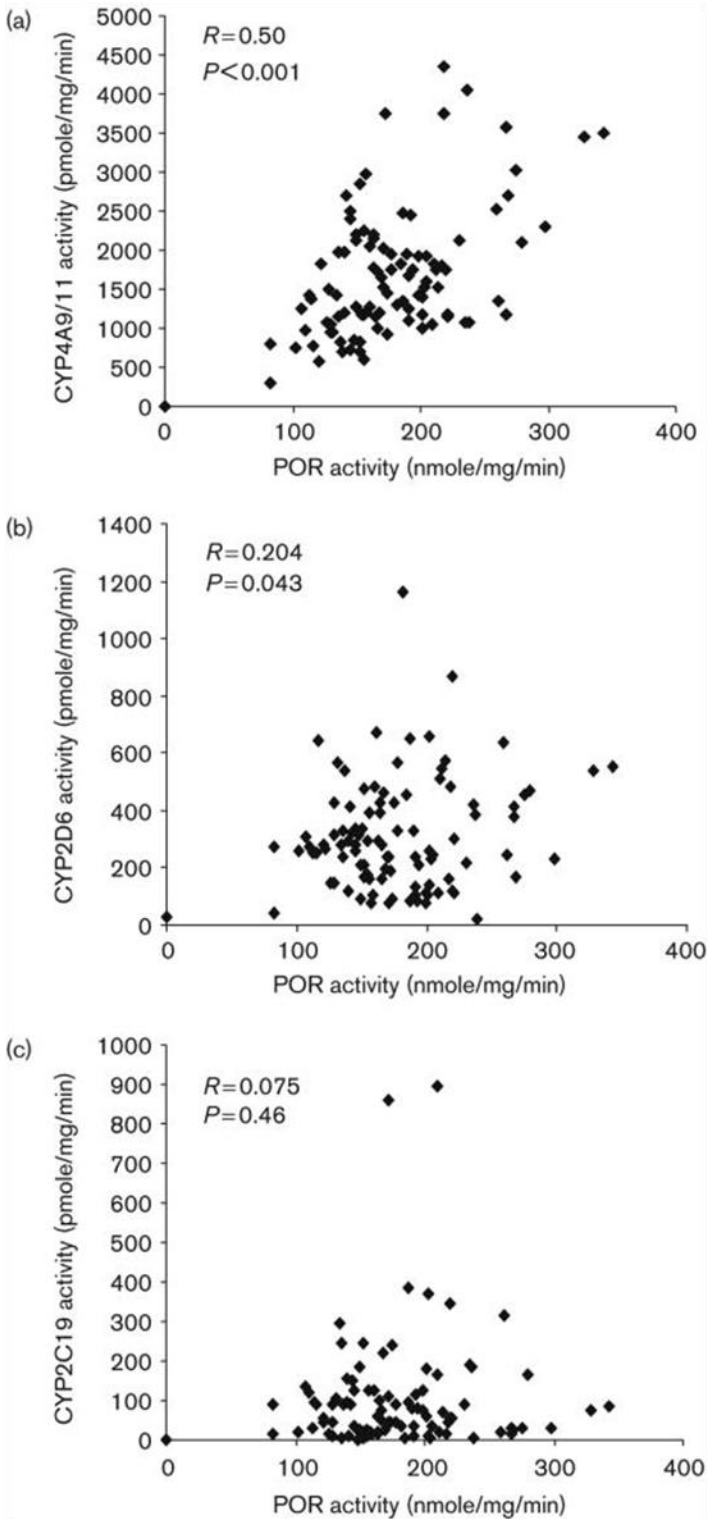

(a)
R=0.50
P<0.001
CYP4A9/11 activity (pmole/mg/min)
POR activity (nmole/mg/min)
(b)
R=0.204
P=0.043
CYP2D6 activity (pmole/mg/min)
POR activity (nmole/mg/min)
(c)
R=0.075
P=0.46
CYP2C19 activity (pmole/mg/min)
POR activity (nmole/mg/min)

Figure 2.2.2. Scatter plots of POR activity versus P450 activity of CYP4A9/11 (A), CYP2D6 (B), and CYP2C19 (C). Pearson's correlations (R) and corresponding *p*-values are listed in the upper left portion of each plot.

Table 2.2.3. Pearson's correlation between POR activity and P450 activity

POR vs. P450 enzyme	**Pearson's correlation**	***p*-value**
CYP1A2	0.102	0.316
CYP2A6***	0.396	<0.001
CYP2B6***	0.354	<0.001
CYP2C8***	0.437	<0.001
CYP2C9***	0.482	<0.001
CYP2C19	0.075	0.46
CYP2D6*	0.204	0.043
CYP2E1***	0.414	<0.001
CYP3A4/5***	0.406	<0.001
CYP4A9/11***	0.5	<0.001

*p<0.05, **p<0.01, ***p<0.001

2.2.4.3. Identification of genetic polymorphisms in the POR gene

Genetic polymorphisms were identified in the exons and surrounding introns (~7.7 kb genomic sequences) in the POR gene by sequencing PCR products amplified from genomic DNAs isolated from the 99 human liver samples. Thirty-four single nucleotide polymorphisms (SNPs) were identified in these areas. Of these, 20 were in the introns, 5 in the 3'-UTR, and 9 in the exons (**Table 2.2.4**). Nine of the 34 SNPs were novel polymorphisms recently reported for the first time (Hart et al., 2007). Of the 9 exonic polymorphisms, 5 were previously reported synonymous polymorphisms (G5G, T29T, P129P, S485S, and S572S) and 4 were nonsynonymous polymorphisms resulting in amino acid changes at K49N, L420M, A503V, and L577P that had minor allele frequencies of 0.005, 0.045, 0.219, and 0.020, respectively. As expected, we did not observe any of the missense or frameshift mutations (T142A, Q153R, Y181D, M263V, A287P, R457H, Y459H, V492E, G539R, L565P, C569Y, Y578C, V608F, R616X, F646del, I444fs, and L612W620delinsR) that have been associated with POR deficiency. All SNPs but rs2286816 in intron 6 had a Hardy-Weinberg *p* value greater than 0.001.

Table 2.2.4. Genetic polymorphisms in the POR gene identified in the liver cohort

SNP ID	rs10262966	rs412952381	SNP1*	rs1135612	rs2286819	rs2286820	SNP2*	SNP3*	rs41299517	rs3815455	rs13223707	rs13240147	rs41301394	rs4732514	rs6971082	rs4732515	rs4732516
Location	Exon 2	Exon 2	Exon 2	Exon 5	Intron 6	Intron 7	Intron 7	Intron 7	Intron 7	Intron 8	Intron 8	Intron 8	Intron 8	Intron 10	Intron 10	Intron 10	Intron 10
Amino Acid change	G5G	T29T	K49N	P129P													
Major allele	A	G	G	A	A	G	G	C	A	C	G	A	C	T	C	C	G
Minor allele	G	A	C	G	G	A	A	T	G	T	C	G	T	C	T	T	C
Samples with genotype	99	99	99	99	99	99	99	99	99	99	99	99	99	99	99	93	96
Allele Frequency (%)																	
Major allele	95.5	99	99.5	78.8	91.9	99	99.5	99	97	77.3	96.5	94.4	81.8	82.8	99	93.5	95.3
Minor allele	4.5	1	0.5	21.2	8.1	1	0.5	1	3	22.7	3.5	5.6	18.2	17.2	1	6.5	4.7
Genotype																	
Homo major allele (n)	92	97	98	63	88	97	98	97	93	57	94	91	66	73	97	84	89
Homo minor allele (n)	2	0	0	6	5	0	0	0	0	3	0	3	3	6	0	3	2
Heterozygous (n)	5	2	1	30	6	2	1	2	6	39	5	5	30	20	2	6	5
Observed Hetero (%)	5.1	2	1	30.3	6.1	2	1	2	6.1	39.4	5.1	5.1	30.3	20.2	2	6.5	5.2
Predicted Hetero (%)	8.7	2	1	33.4	13.1	2	1	2	5.9	35.1	6.8	10.5	29.8	28.4	2	11.4	8.8
HW P-value	0.02	1	1	0.48	0.0009	1	1	1	1	0.38	0.21	0.002	1	0.02	1	0.004	0.02

SNP ID	rs2286822	rs2286823	SNP4*	SNP5*	SNP6*	rs41301427	rs2302431	rs2302432	rs2228104	rs1057868	rs1057870	SNP7*	rs41302345	SNP8*	SNP9*	rs41302348	rs17685
Location	Intron 11	Intron 11	Exon 12	Intron 12	Intron 12	Intron 12	Intron 12	Intron 12	Exon 13	Exon 13	Exon 14	Exon 14	3'-UTR	3'-UTR	3'-UTR	3'-UTR	3'-UTR
Amino Acid change			L420M						S485S	A503V	S572S	L577P					
Major allele	C	A	C	T	T	G	C	G	C	C	G	T	T	G	G	C	G
Minor allele	T	G	A	A	C	A	T	T	T	T	A	C	C	A	A	G	A
Samples with genotype	99	99	99	99	99	99	98	98	98	99	97	99	99	99	99	99	99
Allele Frequency (%)																	
Major allele	64.1	62.6	95.5	99.5	99	90.4	93.4	96.4	92.3	78.1	69.1	98	99.5	99.5	99.5	99	80.3
Minor allele	35.9	37.4	4.5	0.5	1	9.6	6.6	3.6	7.7	21.9	30.9	2	0.5	0.5	0.5	1	19.7
Genotype																	
Homo major allele (n)	47	45	90	98	97	82	88	91	86	63	63	95	98	98	98	97	63
Homo minor allele (n)	19	21	1	0	0	2	3	0	3	6	0	0	0	0	0	0	3
Heterozygous (n)	33	33	8	1	2	15	7	7	9	29	34	4	1	1	1	2	33
Observed Hetero (%)	33.3	32.3	9.1	1	2	15.2	7.1	7.1	9.2	29.6	35.1	4	1	1	1	2	33.3
Predicted Hetero (%)	46	46.8	8.7	1	2	17.4	12.4	6.9	14.1	34.3	42.7	4	1	1	1	2	31.6
HW *p* -value	0.01	0.004	1	1	1	0.42	0.01	1	0.02	0.27	0.11	1	1	1	1	1	0.9

* Novel SNPs identified in this study. HW *p* -value: Hardy-Weinberg *p* -value. *p* value is significant when $p < 0.001$.

2.2.4.4. Prediction of functional influence of the POR mutations

Four nonsynonymous SNPs, which result in the amino acid changes at K49N, L420M, A503V, and L577P, were identified in this liver cohort. K49N resides in the amino-terminal tail, L420M in the connecting domain, A503V in the FAD binding domain, and L577P in the NADPH binding domain. A503V has been reported in the NCBI dbSNP database and has moderate influence on POR activity (69% of wild type) (Huang et al., 2005). We first reported K49N, L420M, and L577P (Hart et al., 2007). To predict the potential influence of K49N, L420M, and L577P on POR functions, we performed a series of modeling analyses to establish conservation of residues and influence on secondary structures and 3 dimensional (3D) configurations, and interactions with cofactors.

The amino acids are conserved at varying degrees: Human POR protein was aligned with homologues from the rat, frog, fruit fly, and yeast (**Figure 2.2.3**). The rationale is that if an amino acid is highly conserved, then a change would likely have a greater impact on POR function because there would have been an evolutionary force to retain the amino acid. The K49 amino acid is located in a FRKKKEE motif that is conserved in human, rat, and frog, but not in fruit fly or yeast. The L420 is also conserved among the human, rat, and frog in a LYLSWVVE motif, but almost no conservation is seen in fruit fly or yeast. Unlike the previous two amino acids, the L577 is highly conserved in a DYLYR motif. In the fruit fly, the leucine is replaced by an isoleucine, an isomeric form of leucine. Because this residue is so highly conserved, it is predicted that a change at L577 may influence POR function.

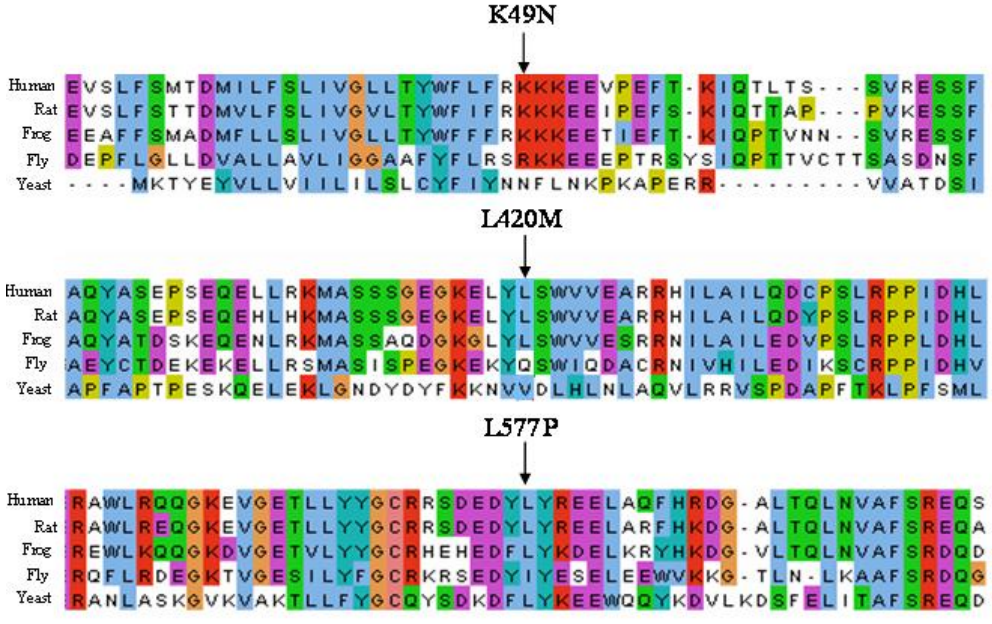


Figure 2.2.3. Alignment fragments of amino acid sequences of POR from five species. Representative amino acid sequences are human (NCBI NP_000932.3), rat (NP_113764.1), frog (AAH59318.1), fruit fly (NP_477158.1), and yeast (NP_596046.1). Alignment was made by ClustalW and shown in default ClustalX color scheme.

The amino acid changes are predicted to affect helix or beta sheet formation: Given the sequence conservation throughout evolution, we sought predictive modeling to (1) identify secondary structures in areas where amino acid changes occur, and (2) predict the outcome of the changed amino acids on protein secondary structure. PSIPRED (Jones, 1999) was used to predict helix or beta sheet structure and MEMSAT2 (Jones et al., 1994) was applied to identify transmembrane domains. First, the full-length POR amino acid sequence with 100% major alleles was uploaded to the Quick2D analysis tool. The major K49 allele was predicted to occur in a random coil directly adjacent to an alpha helical region from S16 to F47 and a transmembrane domain spanning from M31 to F47 (**Figure 2.2.4A** top row). The L420 major allele was predicted to occur in the middle of an alpha helix that forms between K416 and E425 (**Figure 2.2.4B**, top row). The L577 major allele was found in an alpha helix beginning at Y576 and ending at R587 (**Figure 2.2.4C**, top row). When the major alleles of K49, L420, and L577 were replaced by the minor alleles of N49, M420, and P577, the Quick2D analysis revealed some disparities between major and minor allele simulations in the prediction (**Figure 2.2.4A**, **4B**, and **4C** bottom rows). The L420M simulations did not predict any functional changes, but the K49N made F26 involving in an alpha helix from E17 to F47 and V9-D10 forming a β-sheet. In our prediction, the L577P replacement prevented the Y576 residue from participating in the alpha helix.

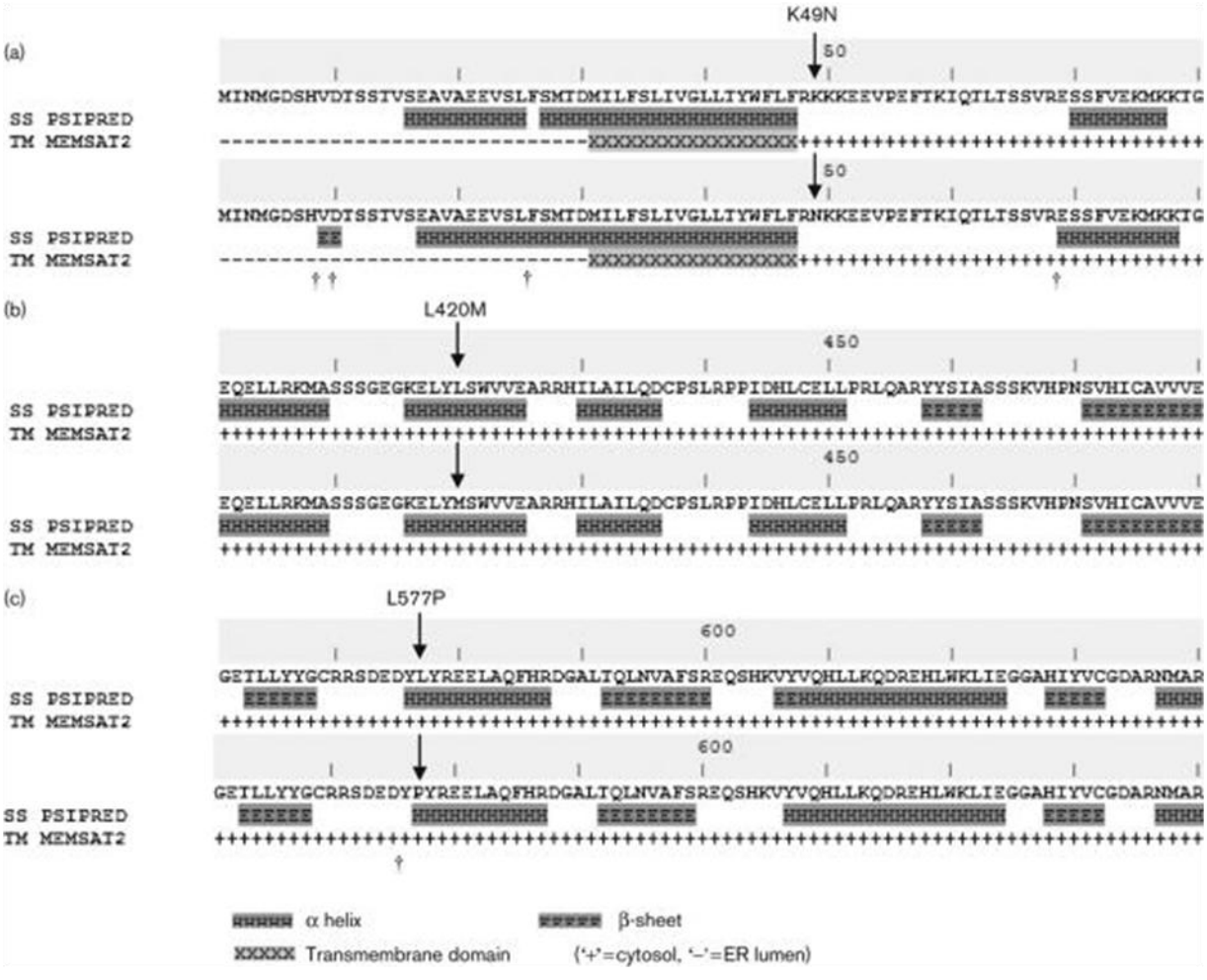


Figure 2.2.4. Predicted secondary structures and membrane topology for K49N (A), L420M (B), and L577P (C). † indicates difference between minor alleles and major alleles. Protein alpha helix and beta sheet were predicted by SS PSIPRED and transmembrane helices were predicted by TM MEMSAT2-D. Arrows indicate the positions of an amino acid change.

The amino acid changes are predicted to affect interaction with cofactors: A 3D structure of the human POR protein was predicted with the ESyPred3D modeling server and visualized with PyMOL, using the fully crystallized rat *Por* structure (PDB: 1AMO chain 'A') as a template (Wang et al., 1997). Crystallography methods used to generate the rat *Por* structure involve trypsinolysis, which truncates the protein at I53 (human V54). Thus, our homology model begins at V67 and the K49N polymorphism could not be modeled and evaluated for its impact on structure and function. **Figure 2.2.5A** shows a 3D structure of the human POR homolog and interactions with its cofactors of NADPH, FAD, and FMN predicted by the EsyPred3D server. Although the L420M polymorphism occurs in the K416- E425 alpha helix, compared to the L420 residue (**Figure. 2.2.5B**), the M420 does not disrupt the helix formation (**Figure. 2.2.5C**). Neither the L420 nor M420 residues participate in hydrophilic, hydrophobic, or hydrogen bonding with FAD or any other cofactor**. Figure 2.2.5D** shows hydrogen bonding between the backbone nitrogen of L577 and a water molecule which likely also dynamically bonds to several other surrounding residues and stabilizes NADPH binding. When L577 is replaced with P577 (**Figure. 2.2.5E**), a hydrogen bond between the backbone nitrogen of P577 and the water molecule no longer forms due to the conformation and properties of proline. Such change may destabilize NADPH binding.

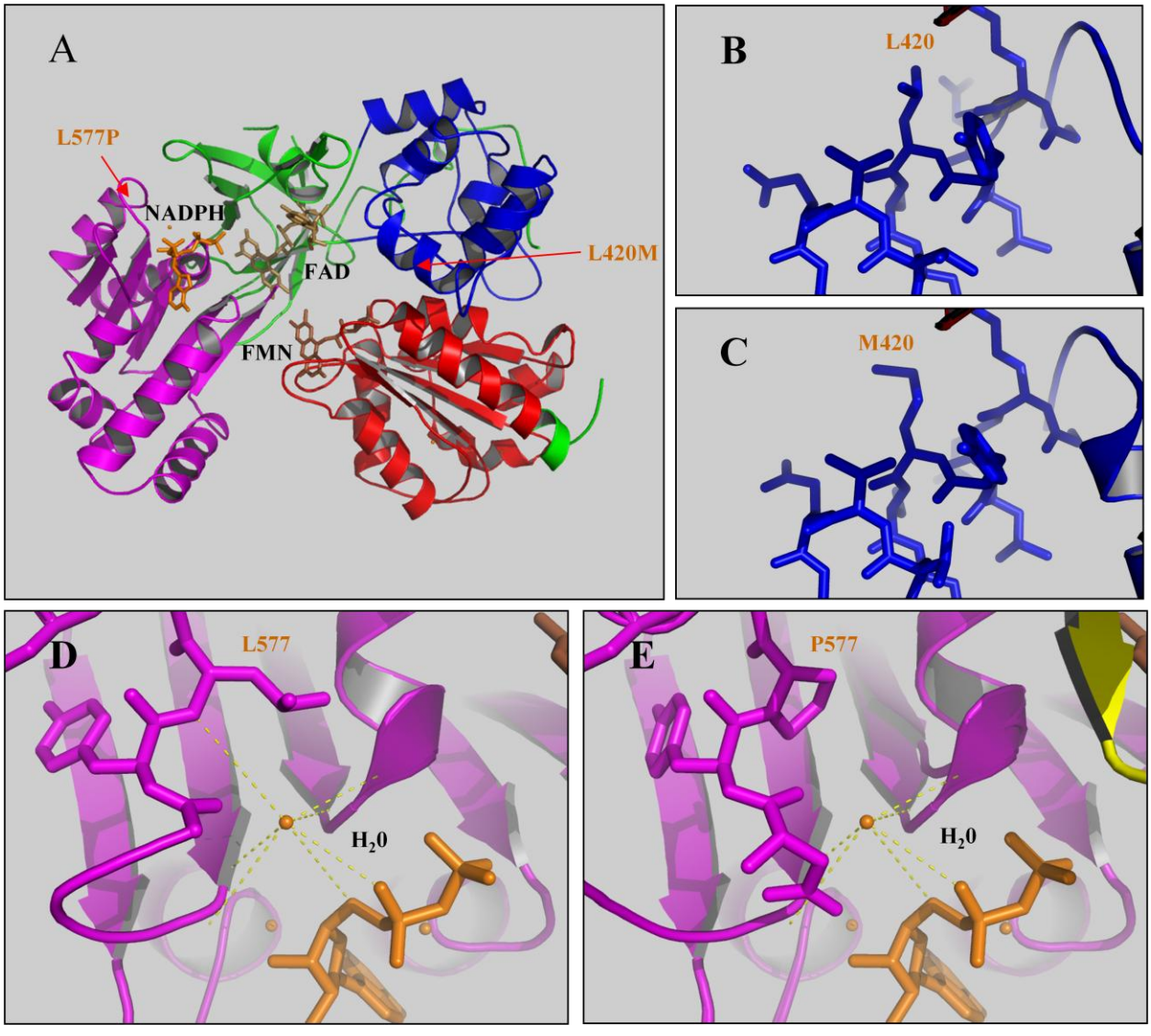

A
L577P
NADPH
FAD
L420M
FMN
B
L420
C
M420
D
L577
H20
E
P577
H20

Figure 2.2.5. Effects of the polymorphisms on POR structure. (A) EsyPred3d predicted 3D structure using rat *Por* (PDB: 1AMO chain 'A') as a template and displayed by PyMOL. Functional domains are distinguished by coloration. The FMN-binding domain is in red, FAD-binding domain is green, the flexible hinge domain is colored blue, and the NADPH-binding domain is pink. FAD, FMN, and NADPH cofactors are tan, brown, and orange, respectively. (B and C) Magnification around wild-type L420 and mutant M420 residues (respectively) shows this mutation occurs in an exposed region that is not involved in interactions with FAD or other cofactors. (D) The backbone nitrogen of residue L577 is potentially involved in stabilizing a water molecule that interacts with several other residues and NADPH. (E) In the case of the P577 mutation, its backbone nitrogen is unable to participate in this interaction with the water molecule due to the unique structure of proline.

2.2.4.5. Correlation of SNPs in the POR gene with POR activity

SNPs in the POR gene had significant influence on the POR activity in this liver cohort. Assuming a dominant genetic model, we performed a multiple linear regression analysis on each SNP adjusting for gender, age, and ethnicity, reason for death, CMV infection, smoking, and drinking. As an example, **Table 2.2.5a** summarizes the regression coefficients, standard error, *t*-values, and *p*-values for SNP7, which causes the L577P amino acid change. Gender, age, ethnicity, CMV infection, reason for death, smoking, and drinking were not associated with POR activity ($p>0.05$). After adjusting for all confounding factors, we found that samples with L577/P577 had significantly lower POR activity than samples with L577/L577 (coefficient estimate=-101.152 and $p=0.003$). **Table 2.2.5b** summarizes the effects of all 34 POR SNPs on POR activity after adjusting for all the above-mentioned confounders. In addition to SNP7, samples with GA and AA genotypes of rs41301427 (a G>A change in intron 12 with a minor allele frequency of 0.096) was associated with decreased POR activity compared to samples with GG genotype (coefficient estimate=-32.409 and $p=0.030$) without influence from the confounders.

Table 2.2.5a. Association of SNP7 (L577P) with POR activity after adjusting for possible confounders

Factor	Coefficient Estimate	Std. Error	*t*-value	*p*-value
Intercept	176.331	36.679	4.807	0
Gender M vs. F	5.698	12.18	0.468	0.641
Age	-0.658	0.366	-1.799	0.076
Ethnicity AA vs. A	26.271	34.492	0.762	0.449
Ethnicity C vs. A	26.754	31.321	0.854	0.396
Ethnicity H vs. A	56.797	35.211	1.613	0.111
Death AA vs. A	59.706	55.452	1.077	0.285
Death CVA vs. A	9.483	17.911	0.529	0.598
Death HT vs. A	-10.238	20.234	-0.506	0.614
Death MI vs. A	10.241	36.448	0.281	0.779
Death MVA vs. A	-60.224	40.609	-1.483	0.142
CMV positive vs. negative	-5.779	12.976	-0.445	0.657
Smoker yes vs. no	-0.5	12.459	-0.04	0.968
Alcohol yes vs. no	9.773	12.277	0.796	0.428
L577/P577 vs. L577/L577**	**-101.152**	**32.776**	**-3.086**	**0.003**

Ethnicity: A: Asian, AA: African-American, C: Caucasian, H: Hispanic.

Death: A: Anoxia, AA: Aortic aneurysm, CVA: Cerebrovascular aneurysm, HT: Head trauma, MI: Myocardial infarction. MVA: Motor vehicle accident.

** Significant level $p<0.01$.

Table 2.2.5b. Association of POR SNPs with POR activity[1]

SNP	Coefficient Estimate	Std. Error	*t*-value	*p*-value
rs10262966	39.997	25.69	1.557	0.124
rs412952381	53.519	41.036	1.304	0.196
SNP1	41.628	55.631	0.748	0.457
rs1135612	18.7	12.465	1.5	0.138
rs2286819	-4.208	20.492	-0.205	0.838
rs2286820	-43.708	40.479	-1.08	0.284
SNP2	72.665	58.322	1.246	0.217
SNP3	66.196	40.19	1.647	0.104
rs41299517	26.616	26.629	1	0.321
rs3815455	-15.296	12.055	-1.269	0.208
rs13223707	0.332	26.344	0.013	0.99
rs13240147	4.096	22.362	0.183	0.855
rs41301394	-6.677	12.923	-0.517	0.607
rs4732514	-16.875	13.353	-1.264	0.21
rs6971082	79.195	39.998	1.98	0.051
rs4732515	11.014	23.103	0.477	0.635
rs4732516	5.153	26.652	0.193	0.847
rs2286822	13.34	11.823	1.128	0.263
rs2286823	9.86	11.961	0.824	0.412
SNP4	34.617	19.824	1.746	0.085
SNP5	74.489	56.321	1.323	0.19
SNP6	-25.56	49.912	-0.533	0.595
rs41301427*	**-32.409**	**14.685**	**-2.207**	**0.03**
rs2302431	2.129	20.504	0.104	0.918
rs2302432	-7.039	23.962	-0.294	0.77
rs6950661	1.008	19.355	0.052	0.959
rs1057868	-2.467	13.634	-0.181	0.857
rs1057870	6.38	12.217	0.522	0.603
SNP7**	**-101.152**	**32.776**	**-3.086**	**0.003**
rs41302345	131.3	67.637	1.941	0.056
SNP8	131.3	67.637	1.941	0.056
SNP9	86.951	56.308	1.544	0.127
rs41302348	-11.587	46.209	-0.251	0.803
rs17685	-17.435	11.956	-1.458	0.149

[1]Adjusted for age, gender, ethnicity, reason for death, smoking history, drinking history, and cytomegalovirus infection. *$p<0.05$, **$p<0.01$

2.2.4.6. Correlation of SNPs in the POR gene with P450 activities

POR SNP7 had significant influence not only on the POR activity but also on most P450 enzyme activities. The influence of L577P is shown in Figure 2.2.6. Four samples (L099, L078, L059, and L080) had heterozygous genotype of SNP7 encoding heterozygous L577/P577 POR protein. The POR activity in the 4 samples with heterozygous L577/P577 was significantly lower than that in the 95 samples with homozygous L577/L577 (p=0.003, **Figure 2.2.6A**). For 10 drug-metabolizing P450 enzymes, the corresponding activities were lower for CYP2A6 and CYP2E1 at significant levels of p<0.01, lower for CYP2B6, CYP2C9, CYP3A4, and CYP4A9 at significant levels of p<0.05, and lower for CYP1A2 at p=0.09, but not significantly lower for CYP2C8, CYP2C19, and CYP2D6 (**Figure 2.2.6B**) when Student's *t*-tests were applied. Although rs41301427 correlated to decreased POR activity (coefficient estimate=-32.409 and p=0.030), Student's *t*-tests revealed this SNP did not influence all P450 activities (data are not shown).

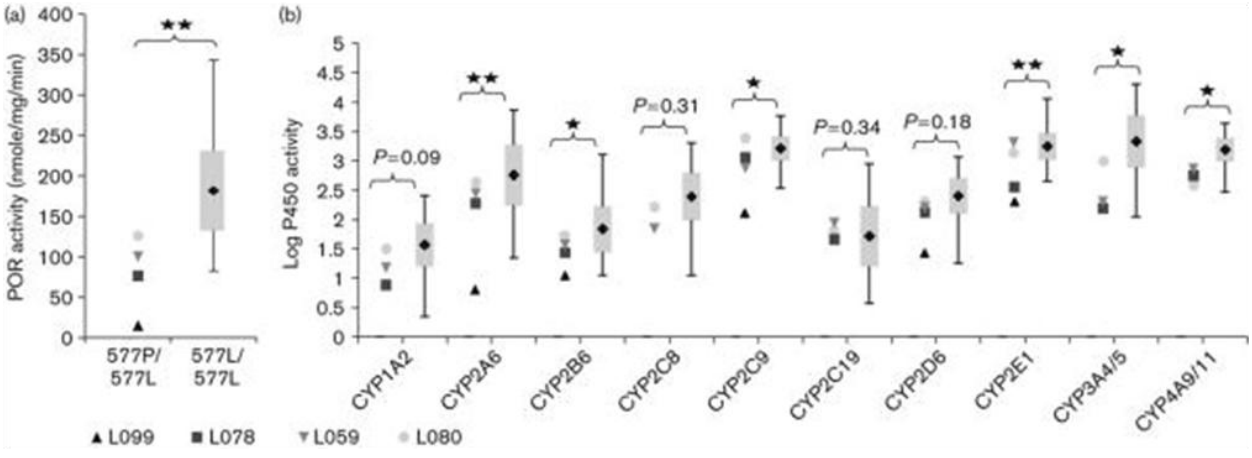


Figure 2.2.6. Effect of the L577P amino acid change on POR and P450 activities.

A) Comparison of the POR activity between samples (n=4) with L577/P577 (left) and samples (n=95) with L577/L577 (right). B) Comparison of the P450 activities between L577/P577 and L577/L577 samples using Student's *t*-tests. To better meet normality assumption, logarithm transformation was applied to all P450 enzyme activities. The samples with no detected enzyme activities therefore were removed for being undefined to take logarithm transformation. There are 1 for CYP1A2, 2 for CYP2C8, 1 for CYP2C19, 1 for CYP3A4/5, and 1 for CYP4A9/11. A black diamond and a grey bar in each right column represent the mean and standard deviation of the enzyme activity in 95 samples with L577/L577. * $p \leq 0.05$, ** $p \leq 0.01$.

2.2.4.7. Correlation of POR gene expression with POR activity

The POR activity was significantly associated with POR gene expression at mRNA levels in this liver cohort. The POR mRNA levels were quantified by branch DNA (bDNA) technology. A housekeeping gene, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), was simultaneously quantified by bDNA technology. Relative POR mRNA levels were defined by normalizing with GAPDH. Inter-individual variations of POR mRNA were observed in this liver cohort with a mean POR/GAPDH ratio of 0.195 and a standard deviation of 0.109. Correlation analysis of POR mRNA with POR activity was performed using the same multiple linear regression adjusting for all previously mentioned confounders. None of the confounders had a significant effect on correlation between POR mRNA and POR activity in this liver cohort (**Table 2.2.6**). A significant association (coefficient estimate=123.653 and p=0.041) was found between POR mRNA and POR activity.

Table 2.2.6. Association of POR mRNA level with POR activity after adjusting for possible confounders.

Factor	**Coefficient Estimate**	**Std. Error**	***t*-value**	***p*-value**
Intercept	225.483	45.175	4.991	4.21×10^{-6}
Gender M vs. F	10.246	13.036	0.786	0.434
Age	-0.47	0.385	-1.218	0.227
Ethnicity AA vs. A	32.475	35.765	0.908	0.366
Ethnicity C vs. A	24.908	32.306	0.771	0.443
Ethnicity H vs. A	49.679	36.936	1.345	0.183
Death AA vs. A	11.164	62.369	0.179	0.858
Death CVA vs. A	-12.372	19.392	-0.638	0.525
Death HT vs. A	-22.85	21.618	-1.057	0.294
Death MI vs. A	-28.25	35.535	-0.795	0.429
Death MVA vs. A	-59.316	41.95	-1.414	0.162
CMV positive vs. negative	5.725	13.376	0.428	0.7
Smoker yes vs. no	1.128	12.966	0.087	0.931
Alcohol yes vs. no	10.455	12.703	0.823	0.413
POR mRNA vs. POR activity*	**123.653**	**59.391**	**2.082**	**0.041**

Ethnicity: A: Asian, AA: African-American, C: Caucasian, H: Hispanic.

Death: A: Anoxia, AA: Aortic aneurysm, CVA: Cerebrovascular aneurysm, HT: Head trauma, MI: Myocardial infarction, MVA: Motor vehicle accident.

* significant ($p<0.05$)

Four samples (L099, L078, L059, and L080), which have the same heterozygous genotype of SNP7 and can produce both wild type L577 and mutant P577 POR proteins in their livers, had varied POR activities from 505 to 139,000 pmol/mg protein/min (**Table 2.2.7**). If the average POR activity (181,000 pmol/mg protein/min) in the samples (n=95) with the wild type genotype of L577/L577 is considered as 100%, POR activity in L099, L078, L059, and L080 is 0.2%, 45%, 61%, and 77% of the average, respectively. An increasing trend from extremely low in L099 to close to the average in L080 maintains same in almost all examined P450 enzymes with *p* value all less than or close to 0.1 in a linear regression analysis (**Table 2.2.7**). This trend was also observed between the POR activity and the POR mRNA level (p=0.083). The liver of L099 was from a 58-year old Caucasian man who did not smoke or drink, was not infected by CMV, and died as a result of myocardial infarction. The liver was removed after 5 hours of death and recorded having normal morphology. His total P450 protein content and GAPDH mRNA were normal. His extremely low POR activity (0.2%) and either no detectable or less than 10% of the average P450 activities may suggest a combinational effect of P577 mutation and low POR gene expression (9% of the average POR mRNA level). The influence of P577 mutation on POR and P450 activities in the sample L080 might be overcome by high POR gene expression (240% of the average POR mRNA level).

Table 2.2.7. Comparison of POR activity with P450 activities and POR mRNA level in wild type samples with L577/L577 and mutant samples with L577/P577

Enzyme activity	**95 samples with wild type L577/L577**	**Sample L099 L577/P577**	**Sample L078 L577/P577**	**Sample L059 L577/P577**	**Sample L080 L577/P577**	**Linear Regression**
(pmole/mg protein/min)	Activity mean (%)	Activity (%)	Activity (%)	Activity (%)	Activity (%)	*p* value[#]
POR activity	181000 (100)	505 (0.2)	82000 (45)	110000 (61)	139000 (77)	**0**
CYP1A2	52 (100)	NPD	8 (16)	15 (29)	33 (63)	***p*< 0.001**
CYP2A6	1016 (100)	6 (0.6)	209 (21)	306 (30)	402 (40)	**0.036**
CYP2B6	110 (100)	10 (9.1)	29 (26)	40 (36)	48 (43)	**0.006**
CYP2C8	350 (100)	NPD	NPD	82 (24)	186 (53)	**0.036**
CYP2C9	1826 (100)	127 (7.0)	1540 (84)	1350 (74)	2750 (151)	**0.032**
CYP2C19	100 (100)	NPD	91 (91)	119 (119)	95 (95)	0.078
CYP2D6	312 (100)	25 (8.0)	274 (88)	270 (87)	292 (94)	0.071
CYP2E1	2146 (100)	201 (9.4)	358 (17)	2400 (112)	1900 (89)	0.12
CYP3A4/5	3504 (100)	NPD	207 (6)	268 (8)	1800 (51)	**0.014**
CYP4A9/11	1726 (100)	NPD	805 (47)	987 (57)	710 (41)	0.087
POR mRNA	0.195 (100)	0.018 (9.2)	0.279 (143)	0.128 (66)	0.468 (240)	0.083

NPD: no product detected

 # p value for linear regression between POR activity and each P450 activity or between POR activity and POR mRNA. Due to abnormal distribution of P450 activities in this liver cohort, the linear regression was conducted with logarithm transformation of the P450 activities. Samples with NPD are considered as 0 after logarithm in the linear regression.

2.2.5. Discussion

We have observed weak, yet significant associations between POR activity and most drug-metabolizing P450 enzyme activities in the study samples. This finding suggests that these P450 enzymes are sensitive to the amount of POR available. This phenomenon was also observed in microsomal P450 enzymes that are involved in steroid hormone biosynthesis (P450C17 and P450C21) (Yanagibashi and Hall, 1986; Lin et al., 1993).

We observed less significant correlations between POR activity and CYP2D6, CYP2C19, or CYP1A2 activity in the study population, which may be due to the known genetic polymorphisms in these P450 enzyme genes. Pharmacogenomic aspects of CYP2D6 and CYP2C19 have been well documented (Daly, 2004; Ingelman-Sundberg, 2005; Eichelbaum et al., 2006; Padol et al., 2007). To date, more than 60 alleles for CYP2D6 and 20 alleles for CYP2C19 have been named by the Human Cytochrome P450 Allele Nomenclature Committee (http://www.cypalleles.ki.se/). Genetic polymorphisms in CYP2D6 and CYP2C19 are found in all ethnic populations (Ozawa et al., 2004), though some are specific for only one or a few populations. When individuals carry genetic mutations that decrease their CYP2D6 or CYP2C19 activity, these individuals in an examined population will likely have low CYP2D6 or CYP2C19 activities regardless of how high their POR activity is. In Figure 2.2.2B and 2C, there are numerous individuals who have normal range, or even high POR activity, but very low CYP2D6 or CYP2C19 activity. These samples are from subjects potentially carrying genetic mutations in CYP2D6 and CYP2C19 genes, which need to be analyzed.

Our data show that the POR gene is quite polymorphic in the study population. Within the 7.7 kb genomic area covering all POR exons and surrounding introns, we identified 34 SNPs, in which 9 are common with a minor allele frequency of >10%, 11 are rare with a minor allele

frequency of <1% and 14 are between (1-10%) (Table 2.2.4). The common SNPs include three exonic polymorphisms (rs1135612 A>G P129P, rs1057863 C>T A503V, and rs1057870 G>A S572S), but only rs1057863 results in an amino acid change, from alanine to valine, at position 503. A503V is a conservative change in an unstructured loop of the FAD binding domain. The amino acid replacement from alanine to valine at 503 results in a minor decrease in cytochrome *c* reduction and P450C17 hydroxylase (Huang et al., 2005). Because of the functional importance of the POR protein, we would not expect there to be common SNPs existing in the POR gene that could significantly affect POR functions. All the common SNPs (>10%) identified in the study samples are not associated with decreased POR activity.

Naturally existing POR mutations have been identified in POR deficiency patients (Arlt et al., 2004; Flück et al., 2004; Fukami et al., 2005; Huang et al., 2005), including the missense or frameshift mutations of T142A, Q153R, Y181D, M263V, A287P, R457H, Y459H, V492E, G539R, L565P, C569Y, Y578C, V608F, R616X, F646del, I444fs, and L612W620delinsR. Site-mutagenesis experiments demonstrated that mutations in the FMN (Q153R, Y181D), FAD (A287P, R457H, Y459H, V492E), and NADPH (G539R, L565P, C569Y, Y578C, V608F, R616X) binding domains had the most influence on POR functions (Huang et al., 2005; Marohnic et al., 2006). Some mutations ablated virtually all measurable POR activity and caused serious human diseases. However, these mutations occur at very low frequency. As expected, none of these POR deficiency mutations were detected in the study population.

Three novel nonsynonymous polymorphisms, which result in amino acid changes at K49N, L420M, and L577P, are identified in this study. K49N resides in the amino-terminal tail, L420M in the connecting domain, and L577P in the NADPH binding domain. Molecular modeling predicts that L420M do not change secondary or 3D structure of POR protein, but

replacement of P577 prevents Y576 from participating in an alpha helix and may disrupt hydrogen bonding with a water molecule, destabilizing NADPH binding in POR 3D configuration. The 3D configuration is very important for electron flow from NADPH through POR to P450. Crystal structure of rat *Por* (Wang et al., 1997) shows that *Por* contains two distinct regions, one containing the NADPH-binding site and the FAD binding domain, and the other containing the FMN domain that eventually interacts with the redox-partner binding site of P450. NADPH attaches to the NADPH-binding site where it releases an electron to FAD and becomes $NADP^+$. The NADPH-binding site and FAD-binding domain must be within 4 Å for this reaction to proceed (Wang et al., 1997). Kinetics of NADPH binding, electron pass, and $NADP^+$ release in POR are important for both POR and P450 enzyme functions. Genetic mutations in the NADPH-binding site (G539R, L565P, Y578C, and V608F), which decrease cytochrome *c* reduction and P450 (CYPC17 and CYPC21) activities, have been identified in the POR deficiency patients (Fukami et al., 2005; Huang et al., 2005). In this report, we demonstrate that L577P, which is adjacent to Y578C, correlates with decreased POR activity (cytochrome *c* reduction) and influences most drug-metabolizing P450 enzymes, except CYP1A2, CYP2D6, CYP2C19, and CYP2C8 (**Figure 2.2.6**). Less correlation in CYP2D6 and CYP2C19 may be again due to genetic polymorphisms in these P450 genes. We have also shown that the POR mutation that decreases POR activity do not necessarily lead to the POR deficiency disease. The mechanisms as to how L5577P decreases POR activity merits further investigation.

L577P has a minor allele frequency of 0.02 in the study population. Its allele frequency in the general population needs to be investigated. We expect that individuals carrying the homozygous mutation of P577/ P577 have decreased ability to metabolize almost all drugs primarily catalyzed by microsomal P450 enzymes. However, this hypothesis needs to be

confirmed in future studies. It is difficult to predict POR activity in an individual who carries heterozygous L577/ P577, because one copy of the POR gene encodes normal POR and another encodes mutant POR. Large inter-individual variations of POR activity were observed in 4 samples with heterozygous L577/ P577 in this study (**Table 2.2.7**). Sample L099 had extremely low POR activity (<mean-2SD), samples L078 and L059 had low POR activity (between -2SD and -1SD), and sample L080 had normal POR activity (between -1SD and +1SD). In these samples, most P450 enzyme activities correlated with POR activity.

The variation in POR activity among the samples with the same L577/ P577 genotype lead us to hypothesize that the interindividual variations of POR activity are not solely dependent on genotype but may partially be due to variations of POR gene expression either among or within individuals. Among individuals in a group, total POR expression may vary considerably, as POR mRNA varied 26-fold in our four samples with the P577 mutation. Additionally, imbalanced allelic variation of POR gene expression may exist between the L577 and P577 alleles in the individuals. Imbalanced allelic variation of gene expression is a common phenomenon existing in more than 50% of human genes (Yan et al., 2002; Lo et al., 2003). The expression ratio of normal to mutant alleles may vary from person to person. An individual who expresses higher levels of the normal allele will ultimately have a greater enzyme activity than one who expresses more of the mutant allele. Therefore, quantification of POR gene expression from the normal L577 allele will provide more accurate correlation with the POR activity in comparison to quantification of total POR mRNA levels from both L577 and P577 alleles. Genetic polymorphisms in the promoter, 5'-untranslated region, 3'-untranslated region, and introns, all can affect gene expression at levels of transcription, splicing, translation, RNA stability or protein modification and contribute to imbalanced allelic gene expression. A

comprehensive analysis of genetic polymorphisms in the entire POR gene is needed to further understand influence of genetic polymorphisms on gene expression and functions of POR.

An intronic polymorphism, rs41301427, was found to be associated with a decrease in POR activity (minor allele frequency 9.5%). The mechanisms for the decrease of POR activity are unclear, although it is possible to speculate that it may interfere with pre-mRNA splicing or affect mRNA stability. However, the degree to which it associates with decreasing POR activity (coefficient estimate = -32.409 and p=0.030) is much less significant than the L577P (coefficient estimate = -101.152 and p=0.003). Such a small change may help to explain why rs41301427 did not influence P450 activities.

Overall, our data demonstrate several findings that may be important for the study of drug metabolism. First, cytochrome P450 activities are significantly correlated to POR activity, suggesting that POR can be a rate-limiting step in P450-mediated catalysis. Second, we have predicted the impact of novel nonsynonymous polymorphisms identified in the study population on protein structure and 3D configuration. Finally, we have identified and characterized several known and novel polymorphisms, including L577P, which decreases POR activity and subsequent P450-catalyzed drug metabolism, but is not associated with the POR deficiency disease. Such discoveries highlight the importance of POR on drug metabolism and warrant further investigation.

CHAPTER 3. A NEW CELL MODEL FOR STUDYING DRUG METABOLISM AND LIVER TOXICITY

Chapter 3.1. Reprinted with permission from the American Society for the Pharmacology and Experimental Therapeutics

3.1. A comparison of whole genome gene expression profiles of HepaRG cells and HepG2 cells to primary human hepatocytes and human liver tissues

3.1.1. Abstract

HepaRG cells, derived from a female hepatocarcinoma patient, are capable of differentiating into biliary epithelial cells and hepatocytes. Importantly, differentiated HepaRG cells are able to maintain activities of many xenobiotic metabolizing enzymes and expression of the metabolizing enzyme genes can be induced by xenobiotics. The ability of these cells to express and be able to induce xenobiotic-metabolizing enzymes is in stark contrast to the frequently used HepG2 cells. The previous studies have mainly focused on a set of selected genes; therefore, it is of significant interest to know the extent of similarity of gene expression at whole genome levels in HepaRG cells and HepG2 cells compared to primary human hepatocytes and human liver tissues. To accomplish this objective, we used Affymetrix U133 Plus 2.0 arrays to characterize the whole genome gene expression profiles in triplicate biological samples from HepG2 cells, and HepaRG cells (undifferentiated and differentiated cells), freshly isolated primary human hepatocytes, and frozen liver tissues. After using similarity matrix, principal components, and hierarchical clustering methods, we found that HepaRG cells globally transcribe genes at the levels more similar to human primary hepatocytes and human liver than HepG2 cells. Particularly, many genes encoding drug processing proteins are transcribed at a more similar level in HepaRG cells than in HepG2 cells compared to primary human hepatocytes and liver samples. The transcriptomic similarity of HepaRG with primary human hepatocytes is encouraging for use of HepaRG cells in the study of drug metabolism, hepatotoxicology, and hepatocyte differentiation.

3.1.2. Introduction

Drug-induced liver injury is one of the leading causes for the failure of drug approval and the withdrawal of approved drugs from the market (Lasser et al., 2002). Animal models are frequently used to identify potentially hazardous drugs for liver injury. However, more than 50% of drugs which induce liver injury in human clinical trials are not hepatotoxic to animals (Olson et al., 2000). Therefore, human liver cells are needed for more accurate *in vitro* screening of drug toxicity.

Freshly isolated primary human hepatocytes are currently the "gold standard" as *in vitro* human liver cells for understanding the pathways and mechanisms influencing drug metabolism and disposition as well as hepatotoxicity (LeCluyse et al., 2000; Luo et al., 2004; Kato et al., 2005). These cells, however, are fraught with difficulties, including their scarce and unpredictable availability, limited growth potential, differences in batch to batch preparation, short life-span, and propensity to undergo early and variable phenotypic alterations. CYP expression decreases quickly over time, likely due to the adaptation of cells to the culture environment (LeCluyse, 2001; Rodriguez-Antona et al., 2002). Additionally, basal gene expression in freshly isolated primary human hepatocytes is also distinctively different from one culture to another, which can introduce additional bias (Richert et al., 2006).

To overcome these difficulties, researchers have been searching for human liver cell lines for a long time. Currently used human liver cell lines are generally derived from hepatic tumors. Unfortunately, most of them have altered gene expression profiles that lack most liver-specific functions. In particular, P450 gene expression and enzyme activities are usually very low or undetectable in these human liver cells. For example, HepG2 cells, the most frequently used human liver cell line, express many CYP genes at very low levels (Sassa et al., 1987). Although

some CYP genes, such as CYP1A1 and CYP3A7, are expressed in HepG2 cells (Ogino et al., 2002), these P450 members are fetal-specific and not expressed in most adult livers. These facts suggest that many changes in gene expression have happened in HepG2 cells after they were derived from the liver tissue of a differentiated hepatocellular carcinoma, or that they represent more of a developmental phenotype.

Recently, a new human liver cell line, HepaRG, is available. Although this cell line is also derived from a female hepatocarcinoma patient, unlike other human liver cell lines, HepaRG cells express many drug processing genes at similar levels compared to primary human hepatocytes under a certain culture condition (Aninat et al., 2006). These drug processing genes encode phase I drug metabolizing enzymes (CYP1A2, 2B6, 2C9, 2E1, and 3A4), phase II enzymes (UDP glucuronosyltransferase 1 family, polypeptide A1, UGT1A1; glutathione *S*-transferase alpha 1, GSTA1; GSTA4, and GSTM1), gene regulatory proteins (aryl-hydrocarbon receptor, AHR; pregnane x receptor, PXR; constitutive androstane receptor, CAR), liver-specific proteins (albumin, haptoglobin, and aldolase B), as well as alpha-fetoprotein, glutathione-related enzymes (γ-glutamylcysteine synthase regulatory subunit, γ-glutamylcysteine synthase catalytic subunit, glutathione synthase, and glutathione reductase), and thioredoxin. The activities of several phase I and phase II drug metabolizing enzymes were also comparable between HepaRG and freshly isolated human hepatocytes (Aninat et al., 2006). HepaRG cells also respond to PXR, CAR, and AhR activators, resulting in induction of CYP1A1, CYP1A2, CYP2B6, CYP2C8, CYP2C9, CYP2C19, and CYP3A4 *in vitro* (Kanebratt and Andersson, 2008b; Lambert et al., 2009a; Lambert et al., 2009b; Lambert et al., 2009c).

HepaRG cells can maintain a proliferative status in an undifferentiated culture media for several weeks at sub-confluency. At confluence, and with the addition of a differentiation-

inducing culture medium, HepaRG cells are capable of differentiating into biliary epithelial cells and hepatocytes (Gripon et al., 2002). The genes encoding liver-specific factors, drug-metabolizing enzymes, transporters, and transcription factors are stably expressed over a multi-week culture period. Given the stable expression of these liver enriched factors over a long time in culture and the activity of several drug-metabolizing enzymes, HepaRG cells have been touted as surrogates to primary human hepatocytes for drug metabolism and disposition studies (Guillouzo et al., 2007; Hewitt et al., 2007).

The ability of the HepaRG cells to express and competently respond to drug-metabolizing gene inducing agents is in stark contrast to the frequently used human liver cell line, HepG2. Although HepaRG cells and HepG2 cells have been compared to human primary hepatocytes and liver tissues for their gene expression and enzyme activities of drug metabolism, the studies were done in a limited set of genes. Therefore, it is of significant interest to know the extent of similarity of gene expression at whole genome levels of HepaRG cells and HepG2 cells compared to primary human hepatocytes and human liver tissues.

3.1.3. Materials and Methods

3.1.3.1. Human liver tissues, primary human hepatocytes, and cultured cells

Human liver tissues. Three different human liver tissue samples were provided by XenoTech, LLC (Lenexa, Kansas). These subjects were Caucasian females at ages of 42, 56, and 60, respectively. The female samples were chosen because HepaRG cell line was derived from a female patient. The samples were acquired by XenoTech through the Midwest Transplant Network (Westwood, Kansas). The livers were cooled immediately after procurement with a cold perfusion solution and frozen within 6 h.

Primary human hepatocytes. Three different primary human hepatocyte samples were provided by Biopredic International (Rennes, France). The primary human hepatocytes were isolated from livers donated by three Caucasian female patients undergoing resection for primary or secondary tumors at ages of 54, 65, and 76, respectively. The hepatocytes were isolated by collagenase perfusion of histologically normal liver fragments and seeded overnight hepatocyte monolayers in seeding medium. After two days culture in short term culture medium, total RNA was isolated from hepatocytes monolayer with Trizol. All liver fragments were not infected by hepatitis B, hepatitis C, and HIV1 viruses.

HepaRG cells. HepaRG cells were obtained from Biopredic International (Rennes, France). The cells in the original culture dish were detached by gentle trypsinization and seeded at 1×10^5 undifferentiated cells/cm^2 (high density) in hepatocyte wash medium (Invitrogen Corporation, Carlsbad, CA) supplemented with additives for growth media (Biopredic International, Rennes, France). The cells were incubated at 37°C and 5% CO_2. The medium was renewed every 3 days. After incubation for 14 days, the undifferentiated HepaRG cells were

induced to differentiate with additives for differentiation media (Biopredic International, Rennes, France) for another 14 days. That medium was also renewed every 3 days.

HepG2 cells. HepG2 cells were obtained from the ATCC Cell Biology Collection (Manassas, Virginia). The cells were seeded at 1×10^5 cells/cm^2 in 75 cm^2 flask with a minimum essential medium supplemented with 10% FBS (Sigma, St. Louis, Missouri) and 5% of penicillin/streptomycin (Sigma, St. Louis, Missouri). The cells were incubated at 37°C and 5% CO_2. The cells were used at the time of 90-100% confluence.

3.1.3.2. Total RNA isolation

Total RNAs of the liver tissues, primary human hepatocytes, undifferentiated and differentiated HepaRG cells, and HepG2 cells were isolated by using Invitrogen TRIZOL reagent (Invitrogen, Carlsbad, California) following the manufacturer's instructions. RNA quality, quantity, and integrity were analyzed by utilizing the Agilent Bioanalyzer 2100 (Agilent Technologies, Amstelveen, The Netherlands).

3.1.3.3. Whole genome gene expression

The whole genome gene expression profiles of the adult liver tissues, primary human hepatocytes, differentiated and undifferentiated HepaRG cells, and HepG2 cells were determined by the Microarray Core Facility at the University of Kansas Medical Center using Affymetrix U133 Plus 2.0 arrays (Affymetrix, Santa Clara, California) in triplicate biological replicates of each sample type. The target preparation, library labeling, hybridization, post wash, and signal scanning were performed based on the Affymetrix manufacturer's instructions.

3.1.3.4. Microarray data analysis.

The raw microarray data in CEL files are available in the Gene Expression Omnibus with accession number GSE18269 at http://www.ncbi.nlm.nih.gov/geo/. The microarray data were normalized using GC-RMA (GeneChip Robust Multichip Average) algorithm (Wu et al., 2004) implemented in the R package *affylmGUI* (Wettenhall et al., 2006) when any two or more data sets are compared. "Present", "Marginal" and "Absent" calls were made in R using the MAS5 algorithm in the *affy* package (Irizarry et al., 2003). A linear model was used to average data among three replicate arrays and also look for variability among them. A probe was removed if it did not correspond to a mapped gene or not register at least 2 "Present" calls in triplicate data sets by the MAS5 algorithm for all five groups. The remaining probes, hereafter defined as quality filtered probes, were used for further analysis. The similarity of whole genome gene expression profiles of HepaRG cells and HepG2 cells compared to human liver tissue and primary human hepatocytes was analyzed by similarity matrix, principal components, and hierarchical clustering. The similarity matrix between any two sets of the data was presented by the Pearson product-moment correlation coefficient (r) value which measured the strength of the linear relationship between two sets of variables. Principal component analysis was applied to identify similarity and differences of the whole genome gene expression profiles among the different samples. In addition to the whole genome gene expression profiles, some liver specific functional pathways, particularly the drug processing pathways were also compared among the different samples.

3.1.3.5. Pathway analysis

To define significant pathway differences during differentiation of HepaRG cells, we used the Functional Annotation Clustering Tool in DAVID (Glynn et al., 2003). Gene lists were made of Affymetrix IDs where the average replicate difference was greater or less than a $\log_2$ value of 1. Each gene list was uploaded using Affymetrix IDs and run against a background containing only the quality filtered probe sets. The Group Enrichment Score, which represents the geometric mean (in $\log_2$ scale) of member's *p*-values in a corresponding annotation cluster, was used to rank biological significance. Thus, the top ranked annotation groups most likely have consistent lower *p*-values for their annotation members. For a pathway or process to be defined, the Enrichment Score was set at 2.

3.1.4. Results

Whole genome gene expression profiles of HepG2 cells, undifferentiated and differentiated HepaRG cells, primary human hepatocytes, and human liver tissues were generated by using Affymetrix U133 Plus 2.0 arrays in triplicate samples. A total of 54,675 probe sets existed on each array. After the probes which did not correspond to a mapped gene or not register at least 2 "Present" calls in the triplicate data sets for all five groups were removed, the remaining quality filtered 30,849 probe sets were selected for similarity and difference analysis by similarity matrix, principal components, and hierarchical clustering methods. Hybridization signal intensities in log2 scale on the 30,849 probe sets among the tested samples were presented in Supplemental Table 3.S1 (Hart et al., 2010).

3.1.4.1. Similarity Matrix Analysis.

A similarity matrix was constructed for each pairwise comparison of any two sets of the data (**Figure 3.1A**). The Pearson product-moment correlation coefficient (*r*) was used to represent the strength of the linear relationship between any two sets of variables. The relative higher *r* values (0.949-0.996) were found between any two replicates in each type of the five groups with a range between 0.991-0.995 in HepG2 cells, 0.995-0.996 in undifferentiated HepaRG cells, 0.989-0.995 in differentiated HepaRG cells, 0.979-0.984 in primary human hepatocytes, and 0.949-0.971 in human livers. The relative lower *r* values (0.768-0.937) were observed between any two sets of the data from the different types of samples. The *r* values in each group were then averaged to represent similarity of whole genome gene expression between any two groups of the samples (**Figure 3.1B**). The highest *r* value (0.966) was found between

undifferentiated and differentiated HepaRG cells. The second highest *r* value (0.920) existed between primary human hepatocytes and human livers. The *r* value between human primary hepatocytes and undifferentiated HepaRG (0.887) or between human primary hepatocytes and differentiated HepaRG (0.891) was higher than the *r* value between human primary hepatocytes and HepG2 (0.813). Similarly, the *r* value between human livers and undifferentiated HepaRG (0.883) or between human livers and differentiated HepaRG (0.881) was higher than the *r* value between human livers and HepG2 (0.791).

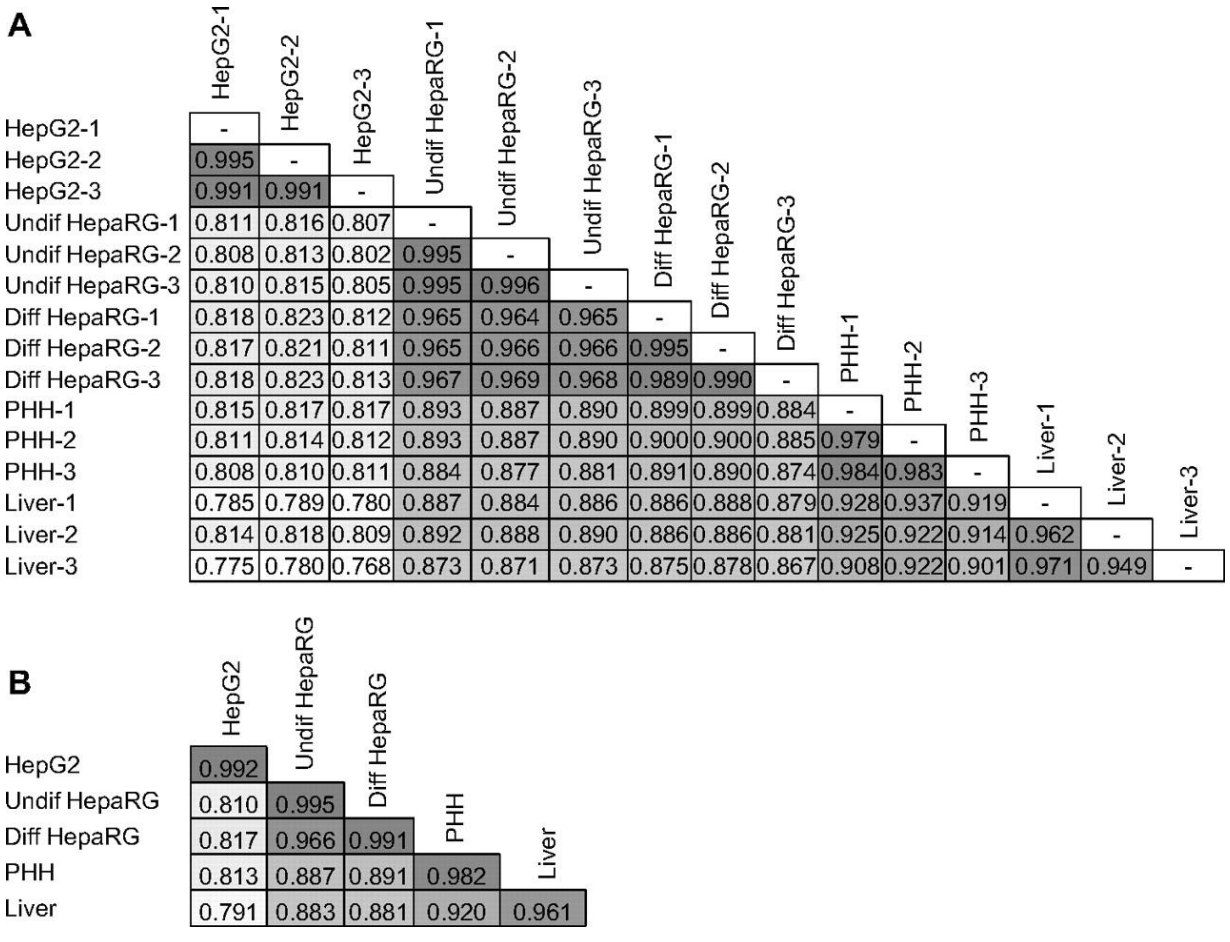


Figure 3.1. A. Similarity matrix of gene expression profiles for each pairwise comparison of HepG2 cells (HepG2-1, -2, -3), undifferentiated HepaRG cells (Undif HepaRG-1, -2, -3), differentiated HepaRG cells (Diff HepaRG-1, -2, -3), primary human hepatocytes (PHH-1, -2, -3), and human liver tissues (Liver-1, -2, -3). The number in each column represents Pearson's product-moment correlation coefficient *r* value. **B**. Average correlation coefficient *r* values for each type of the biological replicates within each group as well as between two groups. Data based on 30,849 probe sets passing a quality filtering test. The background colors in each column indicate different levels of the *r* values.

We further examined the number of probes with signal intensities different by more than 2 fold between any two sets of the data. **Figure 3.2** shows that about 10% of the total quality filtered 30,849 probe sets expressed differently between undifferentiated and differentiated HepaRG cells. The number of probes increased to approximate 22% between human liver tissues and primary human hepatocytes. The differentially expressed probes were 26-28% between HepaRG cells and human liver tissues or primary human hepatocytes. However, up to about 37-39% of the probes were differently expressed between HepG2 cells and human liver tissues or primary human hepatocytes or HepaRG cells.

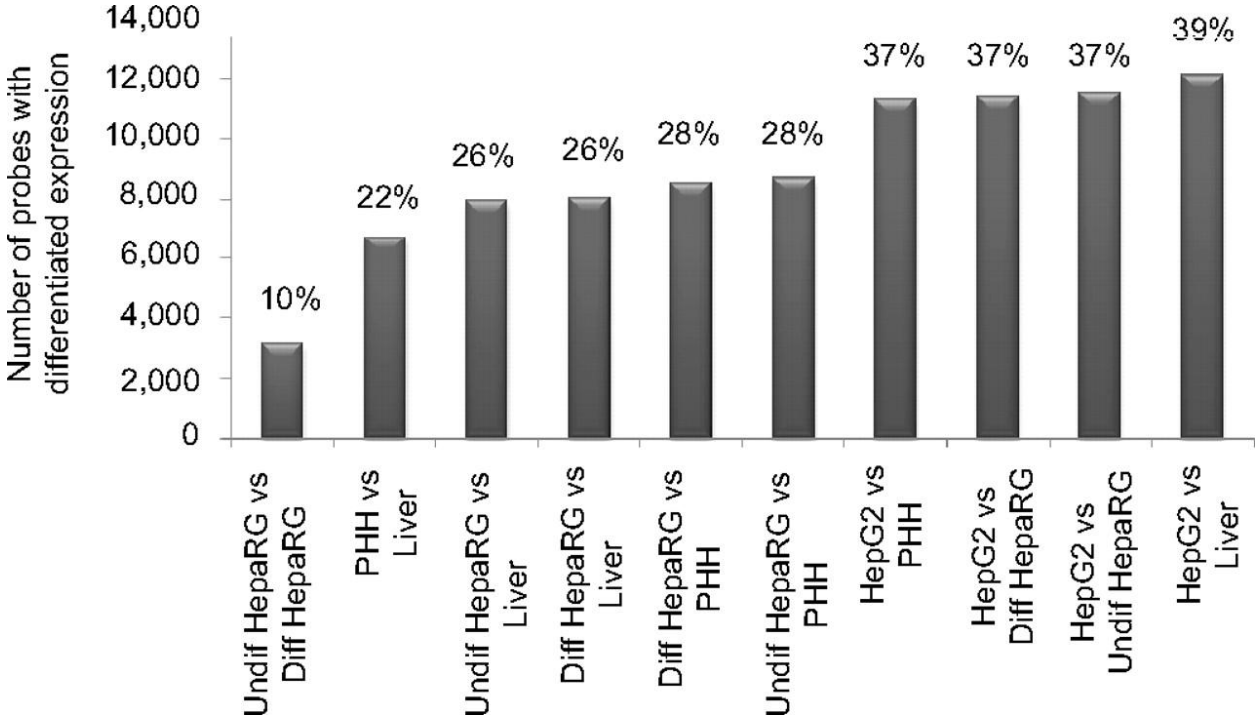


Figure 3.2. Numbers and percentages of probe sets with differential gene expression by more than two fold between any two groups of the samples. The comparison was based on average signal intensities on each set of the probes from three replicates in each group of the samples. PHH: primary human hepatocytes.

We further characterized the 10% of probes with expression levels different by more than 2 fold between undifferentiated and differentiated HepaRG cells. The probes were listed in Supplemental Table 3.S2. A total of 1321 probes had 2-fold higher signal intensities in differentiated HepaRG cells than in undifferentiated HepaRG cells. Pathways analysis indicated that the up-regulated genes included xenobiotic and steroid metabolism, cell cycle genes, DNA replication and repair, and nuclear and ER proteins. Another 1831 probes had 2-fold lower signal intensities in differentiated HepaRG cells than in undifferentiated HepaRG cells. The down-regulated genes during HepaRG differentiation were involved in developmental processes, extracellular signaling, actin binding, and amino acid metabolism.

3.1.4.2. Principal Components Analysis

The similarity and differences in whole genome gene expression among HepG2 cells, undifferentiated and differentiated HepaRG cells, primary human hepatocytes, and human liver tissues were further highlighted by Principal Components Analysis (PCA). The intensities of the quality filtered 30,849 probe sets were first log2 transformed. Three replicate sets of the data were averaged and then used in the PCA analysis. The first three principal components (PC1, PC2, and PC3), which account for most of the variability, were plotted in three dimensions in **Figure 3.3**. HepG2, undifferentiated and differentiated HepaRG, primary human hepatocytes, and human livers contributed nearly equal to the variations in PC1 (93.0%). HepG2 made a major contribution to the variations in PC2 (3.3%). Undifferentiated and differentiated HepaRG cells contributed the majority of the variations in PC3 which only counted 2.1% of total components.

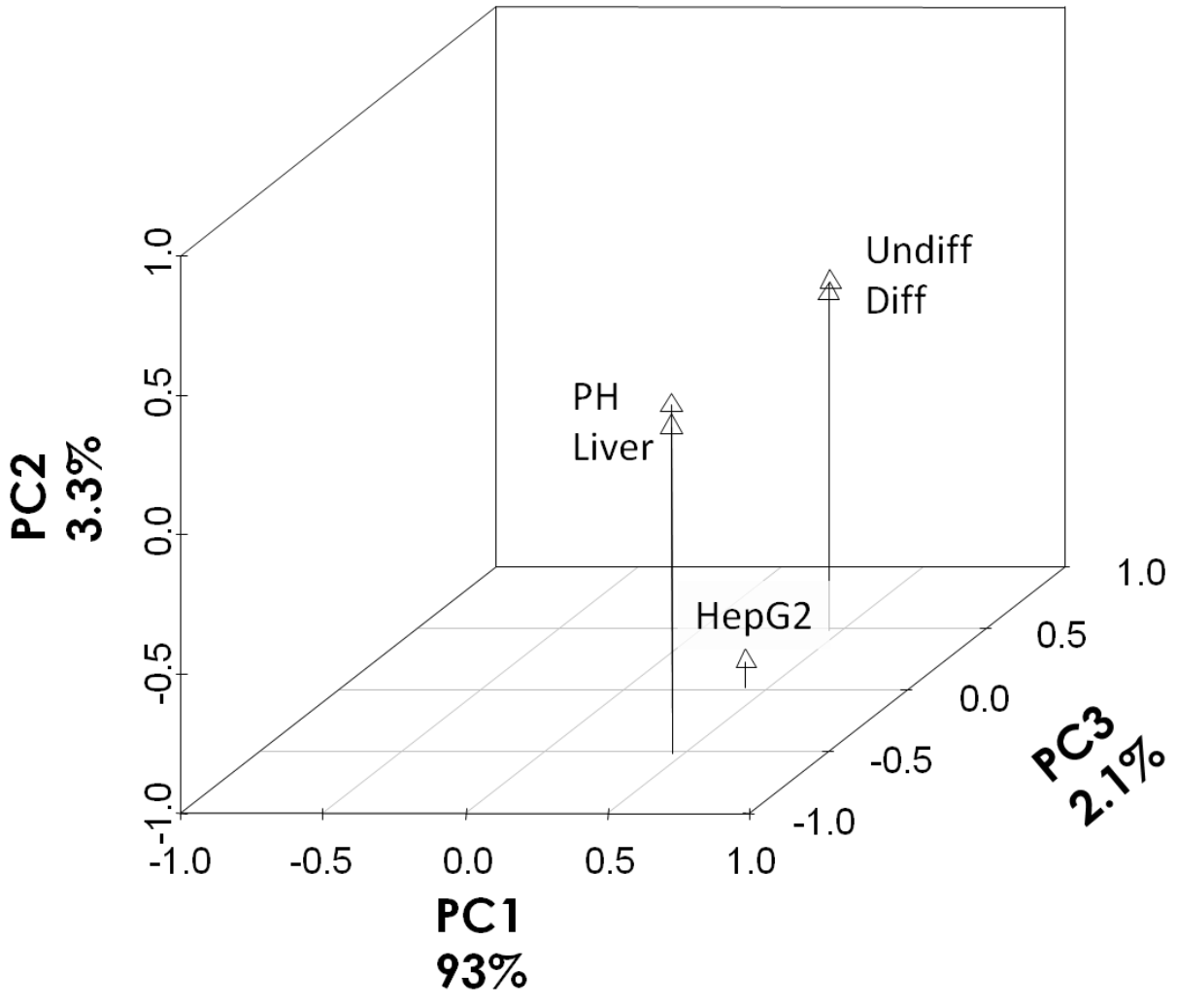


Figure 3.3. Principal components analysis on variations of gene transcription among HepG2 cells (HepG2), undifferentiated HepaRG cells (Undif HepaRG), differentiated HepaRG cells (Diff HepaRG), primary human hepatocytes (PHH), and liver tissues (Liver). For the 30,849 probes passing quality filtering, the relative contribution of the variance is shown by the first three principal components plotted in three dimensions.

3.1.4.3. Hierarchical Clustering Analysis

Hierarchical clustering of gene expression data is a more intuitive way to analyze the many different possible combinations of differentially expressed genes. **Figure 3.4A** shows a two way clustering diagram of five groups of the triplicate samples based on the intensities of the quality filtered 30,849 probe sets on the arrays. The data showed again that the relationship within each group was closer than the relationship between different groups. Within each group, undifferentiated HepaRG cells had the least variation, whereas liver had the biggest variations. Between different groups, differentiated and undifferentiated HepaRG cells are more closely related to primary human hepatocytes than human liver tissues. HepG2 cells have the farther clustering distances to all other groups.

We further selected 115 genes annotated as being involved in xenobiotic metabolism, including the genes encoding phase I and phase II metabolizing enzymes and membrane transporters (a gene list and average signal intensities in log2 scale from the three replicate sets are provided in Supplemental Table 3.S3). The average signal intensities on the probes annotated to the selected genes were clustered in **Figure 3.4B** shown in the groups of Phase I enzymes (ADHs, ALDHs, CYPs, and FMOs), Phase II enzymes (GSTs, NATs, SULTs, and UGTs), and transporters (ABCBs, ABCCs, ABCGs, and SLCOs). A similar clustering pattern as in the whole genome analysis was observed in the phase I and phase II metabolizing enzymes, indicated that HepG2 cells are the most dissimilar to the rest of the groups. A comparison of the expression values of these drug response genes among HepG2 cells, differentiated HepaRG cells, and primary human hepatocytes is also shown in Supplemental Figure S1 with means and standard deviations (Hart et al., 2010). Differences of the expression levels between HepG2 cells and primary human hepatocytes as well as between differentiated HepaRG cells and

primary human hepatocytes were determined by a student *t* test. Overall, differences in gene expression from drug processing genes between HepG2 and primary human hepatocytes is much bigger than between HepaRG and primary human hepatocytes. For examples, all drug metabolizing cytochrome P450 genes are expressed at a significantly different level of ***$p<0.001$, except CYP3A43 at **$p<0.01$, in HepG2 than in primary human hepatocytes (Figure S1-A) (Hart et al., 2010), but only CYP2D6 is expressed at a significantly different level of ***$p<0.001$ in HepaRG than in primary human hepatocytes and CYP1A2, CYP2A6, and CYP2C8 are at a significantly different level of **$p<0.01$. A similar situation is also found for many other drug response genes, such as ADH1A, ADH1B, ADH1C, ADH4, ALDH1L1, ALDH1L2, ALDH9A1, NAT1, NAT2, GSTA1, GSTA3, GSTK1, SULT1A1, SULT1A2, SULT2A1, UGT1A1, UGT1A6, UGT2B4, UGT2B15, UGT2B17, UGT3A1, ABCB1, ABCB4, ABCC10, and SLCO2B1.

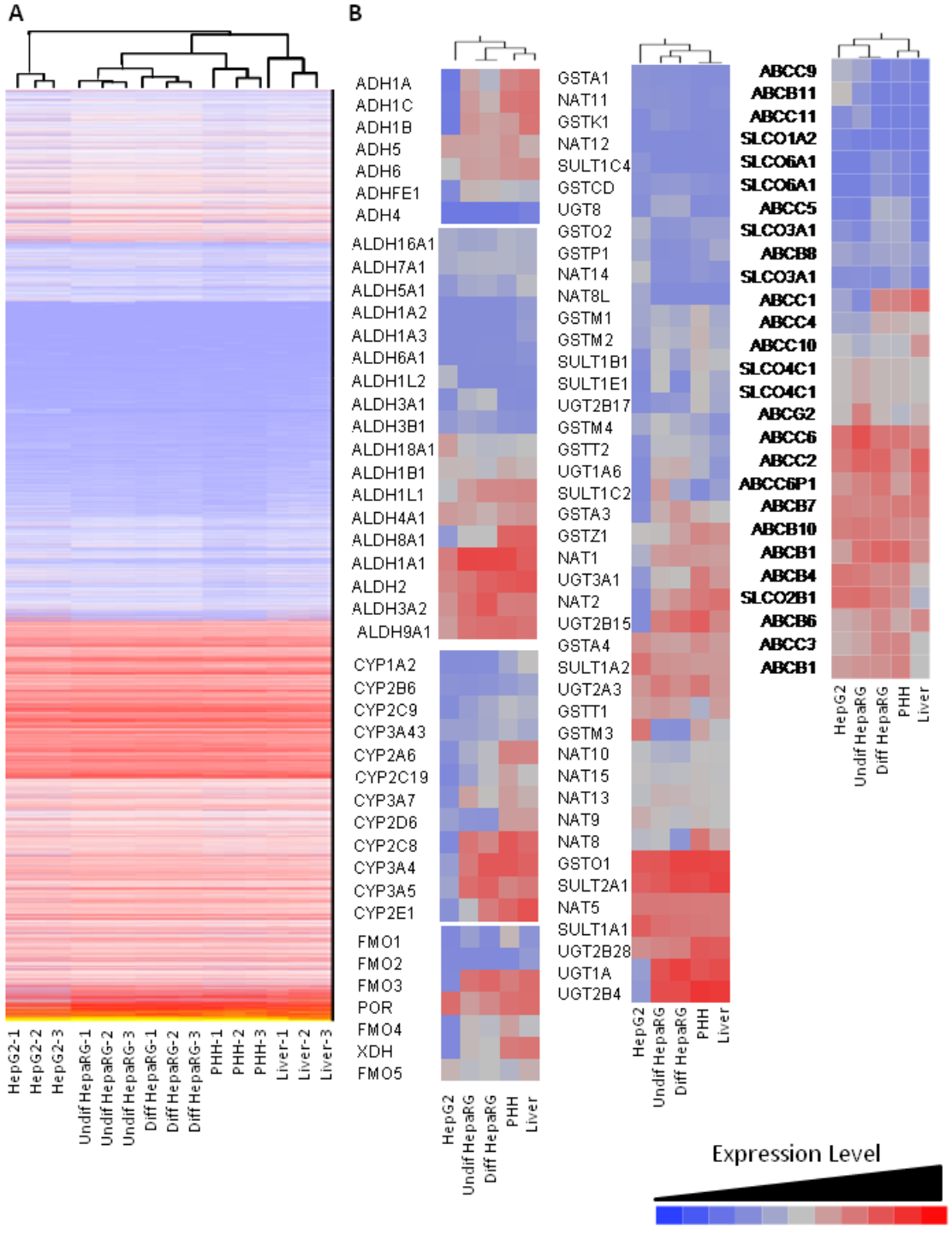

A
HepG2-1
HepG2-2
HepG2-3
Undif HepaRG-1
Undif HepaRG-2
Undif HepaRG-3
Diff HepaRG-1
Diff HepaRG-2
Diff HepaRG-3
PHH-1
PHH-2
PHH-3
Liver-1
Liver-2
Liver-3
B
ADH1A
ADH1C
ADH1B
ADH5
ADH6
ADHFE1
ADH4
ALDH16A1
ALDH7A1
ALDH5A1
ALDH1A2
ALDH1A3
ALDH6A1
ALDH1L2
ALDH3A1
ALDH3B1
ALDH18A1
ALDH1B1
ALDH1L1
ALDH4A1
ALDH8A1
ALDH1A1
ALDH2
ALDH3A2
ALDH9A1
CYP1A2
CYP2B6
CYP2C9
CYP3A43
CYP2A6
CYP2C19
CYP3A7
CYP2D6
CYP2C8
CYP3A4
CYP3A5
CYP2E1
FMO1
FMO2
FMO3
POR
FMO4
XDH
FMO5
HepG2
Undif HepaRG
Diff HepaRG
PHH
Liver
GSTA1
NAT11
GSTK1
NAT12
SULT1C4
GSTCD
UGT8
GSTO2
GSTP1
NAT14
NAT8L
GSTM1
GSTM2
SULT1B1
SULT1E1
UGT2B17
GSTM4
GSTT2
UGT1A6
SULT1C2
GSTA3
GSTZ1
NAT1
UGT3A1
NAT2
UGT2B15
GSTA4
SULT1A2
UGT2A3
GSTT1
GSTM3
NAT10
NAT15
NAT13
NAT9
NAT8
GSTO1
SULT2A1
NAT5
SULT1A1
UGT2B28
UGT1A
UGT2B4
HepG2
Undif HepaRG
Diff HepaRG
PHH
Liver
ABCC9
ABCB11
ABCC11
SLCO1A2
SLCO6A1
SLCO6A1
ABCC5
SLCO3A1
ABCB8
SLCO3A1
ABCC1
ABCC4
ABCC10
SLCO4C1
SLCO4C1
ABCG2
ABCC6
ABCC2
ABCC6P1
ABCB7
ABCB10
ABCB1
ABCB4
SLCO2B1
ABCB6
ABCC3
ABCB1
HepG2
Undif HepaRG
Diff HepaRG
PHH
Liver
Expression Level

Figure 3.4. A. Hierarchical clustering analysis of gene expression for HepG2 cells (HepG2-1, -2, -3), undifferentiated HepaRG cells (Undif HepaRG-1, -2, -3), differentiated HepaRG cells (Diff HepaRG-1, -2, -3), primary human hepatocytes (PHH-1, -2, -3), and human liver tissues (Liver-1, -2, -3). The clustering is based on the 30,849 probes passing quality filtering. **B.** Hierarchical clustering analysis of expression of phase I drug metabolizing enzyme genes (ADHs, ALDHs, CYPs, and FMOs), phase II drug metabolizing enzyme genes (GSTs, NATs, SILTs, and UGTs), and membrane transporter genes (ABCBs, ABCCs, ABCGs, and SLCOs). The clustering is based on average signal intensities from the three replicates in each group of the samples.

HepaRG cells were derived from a human hepatocarcinoma liver tissue. Abnormality of the karyotype in HepaRG cells has been identified with a trisomic chromosome 7 and a translocated chromosome from 22 to 12 (Gripon et al., 2002). Here, we examined whether the karyotype abnormality has any influence on gene expression. Gene expression profiles across each chromosome were compared between differentiated HepaRG cells and primary human hepatocytes. Among all chromosomes, only chromosome 7 had a significant higher gene expression level ($p<0.001$ in a *t*-test) in differentiated HepaRG cells compared to primary human hepatocytes (**Figure 3.5A**). The rest of the chromosomes, including the translocated chromosome 22 (**Figure 3.5B**, $p=0.084$ in a *t*-test), did not show significant difference of gene expression between differentiated HepaRG cells and primary human hepatocytes.

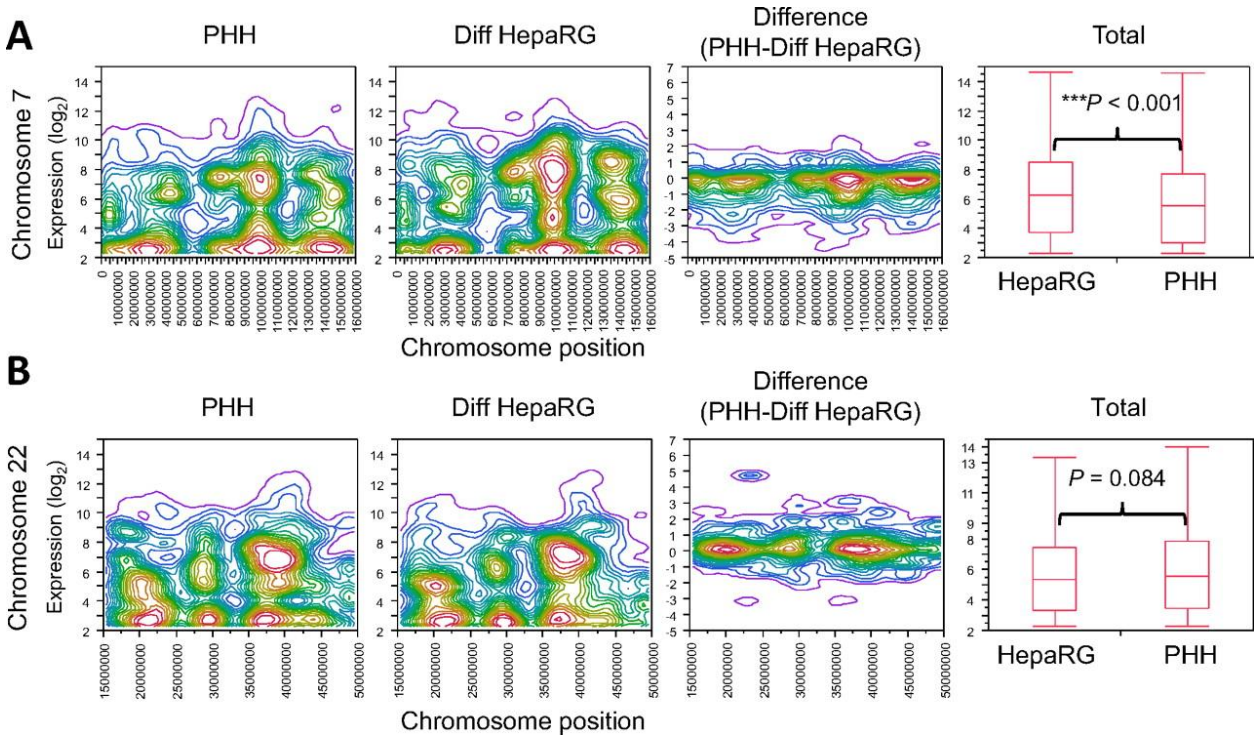


Figure 3.5. Comparison of gene expression profiles across chromosome 7 (**A**) and 22 (**B**) between primary human hepatocytes (PHH) and differentiated HepaRG cells (Diff HepaRG). In the first two panels, the Y axis represents the signal intensity levels in $\log_2$ scale and the × axis indicates the genomic positions along the chromosomes where the microarray probes are located, for PHH and Diff HepaRG, respectively. The third panel was constructed by subtracting PHH values from Diff HepaRG signal values for each probe on the chromosome and then clustered as in the first two panels. For example, if the expression of a given probe was 5 in PHH and 3 in Diff HepaRG, then that value would be 5-3=2. Negative values indicate probes with lower expression in PHH than Diff HepaRG. Blue areas represent few data points, whereas red areas indicate more data points. In the forth panel, signal intensity levels from all probes on each chromosome are compared between Diff HepaRG and PHH with means and standard deviations. The differences are tested by a *t*-test. ***$p<0.001$ indicates that a significant difference is identified between the two sets of samples.

3.1.5. Discussion

The current study used Affymetrix gene expression arrays to establish genome-wide gene expression profiles of HepaRG cells at both undifferentiated and differentiated stages and compared the genome-wide gene expression profiles of HepaRG cells and HepG2 cells with human primary hepatocytes and human liver tissues using similarity matrix, principal components, and hierarchical clustering methods. The comparison was also done for many drug processing genes. These analyses conclude that the mRNA content in HepaRG cells more accurately reflects primary human hepatocytes and human liver tissues than HepG2 cells. The similarity matrix analysis shows the relative high *r* values (0.949-0.996) between any two replicates within a same type of the samples. These high *r* values indicate that the gene expression profiles generated by the Affymetrix gene expression arrays are highly reproducible. It is not surprising that the *r* values between the replicate samples in the cultured HepG2 and HepaRG cells were higher than in primary human hepatocytes and human liver samples, because the cultured cells consisted of a highly homogenous cell population with little environment-mediated perturbations, but primary human hepatocytes and human liver samples came from different individuals in which their gene expression could be influenced by many factors which cannot be controlled in the experiment. Particularly, the *r* values within the liver tissue samples (0.949-0.971) are relatively lower than the *r* values within other groups, indicating that a certain degree of variations exists among the individual liver samples, which may be caused by interindividual variations or mRNA quality of the liver tissues.

When similarity is compared between different groups of samples, differentiated HepaRG and undifferentiated HepaRG cells are still highly similar ($r = 0.966$ in **Figure 3.1B**)

with only a small proportion of the probes (~10% in **Figure 3.2**) expressed differentially by more than 2 fold. A list of differentially expressed genes and pathways can be found in Supplemental Table 3.S2 (Hart et al., 2010). Drug processing genes are the most significantly up-regulated genes during HepaRG differentiation, including many phase I enzymes (such as CYP2C9, CYP2C19, CYP2E1, and CYP3A4) and phase II enzymes (for example, UDP-glucosyltransferases and glycosyltransferase). Although both undifferentiated and differentiated HepaRG cells are very similar to primary human hepatocytes at whole genome gene expression, differentiated HepaRG cells express xenobiotic processing genes more similar to primary human hepatocytes than undifferentiated HepaRG cells.

Similarity is also high between primary human hepatocytes and liver tissues (r=0.920 in Figure 3.1B), but with about 20% of probes expressed differentially (**Figure 3.2**). Although hepatocytes are the major types of cells in liver, making up to 70-80% of the mass of the liver, liver also consists of several other types of cells, such as cholangiocytes, endothelial cells, hepatic stellate cells, and kupffer cells, which have different gene expression profiles than hepatocytes. Other factors which can influence the measurement of gene expression in human livers are the procedures for harvest, treatment, and storage of liver tissue samples. Therefore, freshly isolated primary human hepatocytes should be considered as the key reference for comparison of gene expression between the *in vitro* cultured liver cells and the *in vivo* liver cells.

The similarity levels represented by the Pearson product-moment correlation coefficient (r) value and number of the non-differentially expressed probes were higher between HepaRG cells and primary human hepatocytes or liver tissues than between HepG2 cells and primary human hepatocytes or liver tissues, whereas the differences were lower between HepaRG cells and primary human hepatocytes or liver tissues than between HepG2 cells and primary human

hepatocytes or liver tissues. These data indicated that HepaRG cells expressed genes at a genome-wide level were more similar to primary human hepatocytes and human livers than HepG2 cells.

The above conclusion is also supported by principal component analysis, which confirms that the variations in gene expression at a whole genome level are contributed mainly from HepG2 cells compared to HepaRG cells, primary human hepatocytes, and liver tissues. Hierarchical clustering analysis also shows that the association of gene expression at genome levels is closer in each type of the groups than between different types of the groups (**Figure 3.4A**). Within each type of the groups (differentiated/undifferentiated HepaRG, primary hepatocytes, etc), undifferentiated HepaRG cells have the closest association, whereas liver tissue samples have the least association. Between different groups, undifferentiated HepaRG and differentiated HepaRG cells are close each other, and primary human hepatocytes and liver tissues are close each other. Then, HepaRG cells are closer to primary human hepatocytes and liver tissues than HepG2 cells. When a set of genes involved in drug processing, including many phase I enzymes, phase II enzymes, and transporters were selected for a clustering analysis (**Figure 3.4B**), the gene expression profiles are also more similar between HepaRG cells and human primary hepatocytes or liver tissues than between HepG2 cells and human primary hepatocytes or liver tissues. These findings are in agreement with several previous studies (Aninat et al., 2006; Le Vee et al., 2006; Richert et al., 2006; Kanebratt and Andersson, 2008a; Kanebratt and Andersson, 2008b). When the differences in expression levels among HepG2, HepaRG, and primary human hepatocytes were compared in the major drug processing gene families, such as CYPs, ADHs, ALDHs, FMOs, NATs, GSTs, SULTs, UGTs, ABCBs, ABCCs,

ABCGs, and SLCOs, the differences between HepG2 and primary human hepatocytes are much larger than the differences between HepaRG and primary human hepatocytes for most genes.

In conclusion, we used a high-throughput genome-wide approach to define gene transcriptional profiles of HepaRG cells at both differentiated and undifferentiated stages and compared the gene expression profiles of HepaRG cells and HepG2 cells with human primary hepatocytes and liver tissues. We found gene transcription levels in HepaRG cells have a much higher level of similarity to human primary hepatocytes and liver tissues in comparison to HepG2 cells; the most commonly used cultured cells for studying liver biology. The transcriptomic similarity of HepaRG with human primary hepatocytes is encouraging for use of the HepaRG cells in the study of drug metabolism, hepatotoxicology, and hepatocyte differentiation in the future. These sets of data can also serve as a database for researchers who want to compare expression levels of any genes in HepaRG cells, HepG2 cells, primary human hepatocytes, and human liver tissues.

The current study highlights the similarity of gene transcription between HepaRG cells and human primary hepatocytes or liver tissues in comparison with HepG2 cells. The high similarity at mRNA levels between HepaRG cells and human primary hepatocytes or liver tissues does not necessarily mean that the similarity occurs also at protein levels. Studies of genome-wide protein levels require high-throughput protein arrays which are not available yet. It is also worth noting that because a trisomic chromosome 7 exists in HepaRG cells, genes located on chromosome 7 may have higher expression levels in HepaRG cells than in primary human hepatocytes due to the extra copy of chromosome 7. This factor should be taken into consideration when an experimental design in the use of HepaRG cells involves genes located on chromosome 7.

CHAPTER 4. A NEW TOOL FOR PHARMACOGENOMICS.

4.1. SNP analysis from mRNA-Seq Data: a framework for implementing mRNA-Seq in Pharmacogenomics

4.1.1. Abstract

mRNA-Seq is an attractive tool for use in pharmacogenomics. In a single experiment, one can theoretically quantify gene expression, identify and quantitate transcript variants on a genome-wide scale, and discover genetic variation that may underpin a phenotypic response. This represents an advantage over other assays such as whole genome or whole exome sequencing because with those approaches one could only discover genetic variation. However, given the complexity of the data and the relative novelty of the technology, it is important to first assess the efficacy of RNA-Seq for pharmacogenomics. In this study, we assessed the technical performance of current SNP-calling methods compared to genome-wide Affymetrix SNP arrays to yield measures of accuracy and performance. We found a high rate of false positives (FDR = 35%) were called when no filtering was applied to the sequencing dataset. The false discovery rate was only marginally improved (FDR = 26%) by a low-stringency filter that required at least 3-fold coverage and a minimum SNP quality of 25. We then used these data, combined with the SNP array data, to train a logistic regression model to reduce the number of false positives, which greatly decreased the false discovery rate (FDR = 5%). The reduction in the FDR concomitantly decreased the sensitivity to 26%, but increased the specificity to 99%. No effect of RNA-Seq mapping alignment programs was observed if similar filtering conditions were applied to each. Once SNP-calling parameters were optimized, we called 5,012 SNPs in the RNA from HepaRG cells and annotated those SNPs as either being known or predicted to have a significant effect on particular drugs. 321 SNPs are predicted to alter protein structure/function, and 24 SNPs have known associations with drug-response phenotypes. In summary, we have

generated an accurate model (70%) for genotyping SNPs from mRNA-Seq data, and have laid the foundation for using mRNA-Seq in pharmacogenomics.

4.1.2. Introduction

There are two traditional ways to perform pharmacogenetic research, namely forward and reverse genetics. In forward genetics, one starts with a phenotype of interest and moves toward finding the gene and mutations contributing to the phenotype, whereas reverse genetics starts with a particular gene and assays the effect of its disruption. In terms of human studies, reverse genetics is useful for *in vitro* assays, but forward genetic screens are appropriate for *in vivo* studies. The problem with the forward screen is that one must *a priori* identify a candidate gene that is suspected to be involved in the absorption, distribution, metabolism, or excretion of a drug. Then, primers surrounding the exons of that gene are designed to amplify those regions by PCR from hundreds of genomic DNAs, and the amplicons are sequenced using the Sanger method. Ideally, one would hope to discover enough genetic variation in that gene of interest and try to link individuals with different genotypes with some phenotypic measurement of drug response. Using this approach, if the genetic variation lies outside the regions of the candidate gene that were selected to be sequenced (and is not in linkage disequilibrium with the causative gene/SNP), no association can be established. Alternatively, no association can be established if the effect is cumulative with other genes because they were not sequenced in the original assay. One would have no way of knowing this and may come to an incorrect conclusion that variations in that gene are not responsible for the observed phenotype.

Advances in technology have literally revolutionized the field in just a few years, which is rapidly making this approach obsolete. Now the community is moving toward a more

comprehensive approach: next-generation sequencing. Many chemical, methodological, and throughput differences set next-generation apart from Sanger sequencing. One can now sequence more genes, faster, and cheaper than ever before on an unprecedented scale. Rather than sequencing one template, next-generation sequencing instrument sequence millions of templates simultaneously. With new technology, however comes new problems and one has to understand the technology before one can understand the problems. With understanding of the principles of the experiment, one can optimize the performance to generate volumes of highly accurate data. Failure to acknowledge the principles can and will lead to spurious, non-reproducible conclusions.

Currently, it is still not feasible to sequence human genomes on a large scale for pharmacogenomics analysis (currently the reagent cost alone per sample hovers around $10,000), although this obstacle will inevitably be overcome. Certain derivatives of whole genome sequencing, such as RNA-Seq (Wang et al., 2009), offer effective alternatives to whole genome sequencing and provide additional information that whole genomes cannot. In a single experiment, one can quantify gene expression with digital resolution and genotype all expressed coding exons on a massive scale. RNA-Seq is highly accurate for quantifying expression levels (Mortazavi et al., 2008; Nagalakshmi et al., 2008), and is highly reproducible for both technical and biological replicates (Cloonan et al., 2008; Nagalakshmi et al., 2008), with greater sensitivity than microarrays (Wang et al., 2009). RNA-Seq has also been proposed as an effective tool to discover genetic polymorphisms in coding regions (Morin et al., 2008; Chepelev et al., 2009). However, several challenges exist in data analysis.

In RNA-Seq experiments, one of the first and most critical steps is alignment. Millions of sequence fragments must be mapped to the genome, that is to say that each fragment of DNA

sequenced (a.k.a. reads) must be assigned to chromosomal coordinates in order to "know" where that piece of DNA actually belongs. The problem however, is that in order to align these sequenced fragments, they must uniquely match the reference genome or transcriptome (generally with less than 2 mismatches). This is by no means a simple task, and is a very interesting and rapidly advancing field of research (Pepke et al., 2009; Trapnell and Salzberg, 2009; Bryant et al., 2010; Wang et al., 2010). Infomatically, there are many difficulties. First, if aligning to a reference transcriptome, such as with the ERANGE package (Mortazavi et al., 2008), one will not be able to identify novel exons or transcripts, including any underlying sequence information that may be contributing to those isoforms. Secondly, when mapping directly to a genome, a read can sometimes map to more than one chromosomal coordinate, making it difficult to distinguish which particular genomic location it arises from. Thirdly, if reads are mapped directly to the genome, the majority of reads spanning splicing junctions will not be mapped unless that particular program explicitly allows for split-read mapping. Two of such read aligners, TopHat (Trapnell et al., 2009) and SOAPals [unpublished] , are able to directly map reads to the genome allowing for split-reads. Given the differences in the underlying algorithms of these programs, the question remains as to whether or not those differences can significantly affect the outcomes of downstream genotyping programs.

To study the effectiveness of RNA-Seq for pharmacogenomics, we present the current study. Our goals are 1) to determine the effect of RNA-Seq mapping alignment programs on genotyping, 2) assess the technical performance of current SNP-calling methods, 3) improve this performance, and 4) to provide a starting point going from generating data to understanding the biological consequences.

The first step in a line of many to achieve these goals, was to select HepaRG cells as a model system for the current study. The mRNA content of HepaRG has been shown to replicate that of primary human hepatocytes and liver tissue better than the commonly used HepG2 cell (Hart et al., 2010), and this is supported by activities of several phase I and phase II drug metabolizing enzymes (Aninat et al., 2006), and thus are ideal for drug metabolism studies. Because genes encoding drug metabolizing enzymes are expressed in these cells, and those would be critical targets of genotyping by RNA-Seq, this model is well suited for our analyses. Finally, we need to use a cell line with a stable karyotype to keep the genomic content essentially constant, we that way we can reproduce the genetic variations we observe as many times as necessary. In other words, we are decreasing the amount of heterogeneity in our system to test the accuracy of our statistical and algorithmic assumptions rather than on introduced biological variation.

4.1.3. Methods

4.1.3.1. HepaRG Cell culture

HepaRG cells and culture medium were provided by Biopredic International. The undifferentiated HepaRG cells were seeded at 0.2 million cells/well in 6-well plates, maintained in the growth medium for two weeks, and then cultured in the differentiation medium containing 2% dimethyl sulfoxide (DMSO) for two more weeks to obtain differentiated HepaRG cells (Hart et al., 2010). The differentiated HepaRG cells were incubated with serum-free growth medium for 48 h. Total RNA from HepaRG cells was prepared using TRIzol reagent according to the manufacturer's protocol.

4.1.3.2. Illumina Sequencing & Mapping

We generated 2 × 36 bp sequence reads using the Illumina platform following the manufacturer's recommended protocol (Genpathway, San Diego, CA). The paired-end reads were mapped to the human genome (hg19) in-house using TopHat or SOAPals with the following parameters. For TopHat, we specified the maximum number of multi-hits to 1, the inner-mate pair distance to 40 and solexa1.3 quality metrics. For SOAPals we set the maximum number of mismatches for a one-segment alignment to 2, and then post-filtered the reads to only contain one possible mapping location. We then used the soap2sam.pl script to convert the SOAP format to the widely adopted SAM format specifications (Li et al., 2009) so comparisons could be made. Paired-end RNA-Seq data from six HepaRG samples were used. For each sample, the data were aligned and filtered as described above, and reads of the same sequence (i.e. likely PCR artifacts) were removed. These data sets were used to compare the performance of different read aligners. For the final analysis, the unique data from each sample were combined together into one SAM file and converted to BAM (the binary equivalent to SAM, to improve computational speed) for subsequent analyses. A general scheme for the data analysis was shown in Figure 4.1.1.

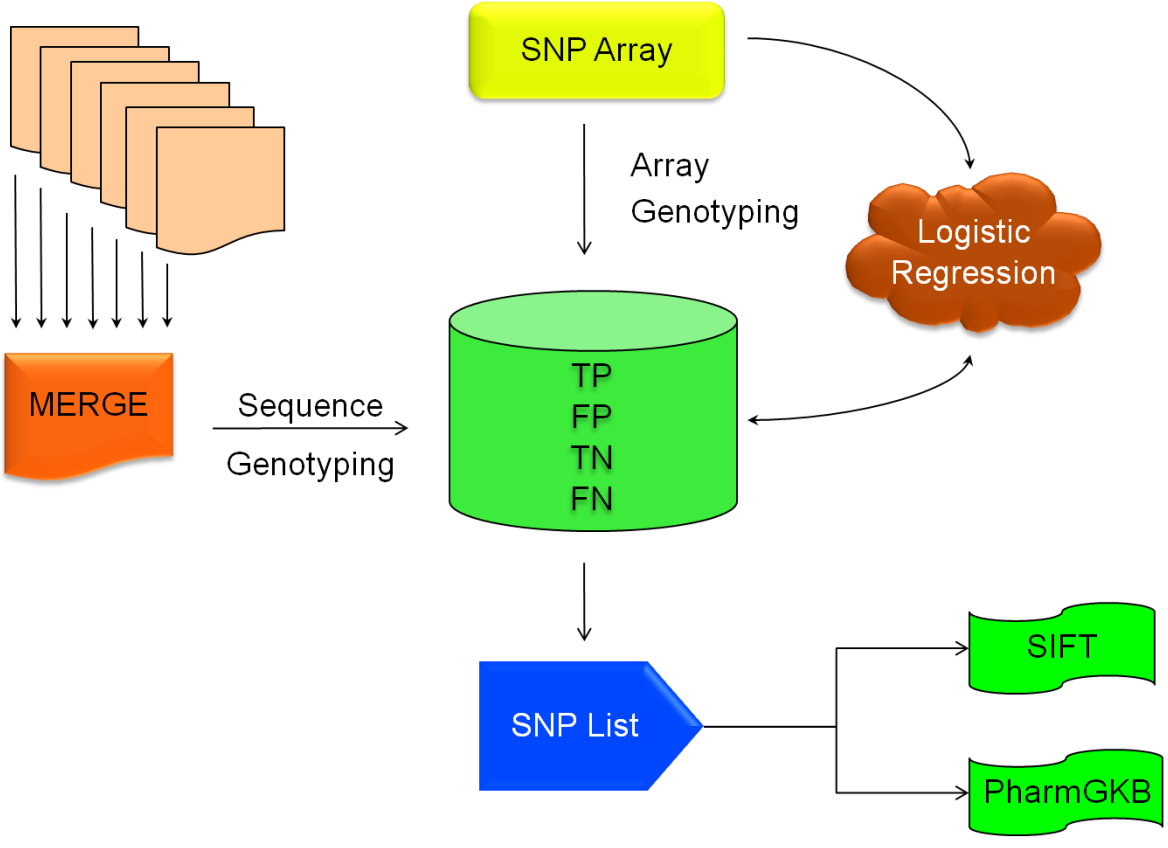


Figure 4.1.1. Experimental scheme for calling SNPs from RNA-Seq reads. Six lanes of RNA-Seq data were filtered to remove PCR duplicates and then merged into a single file (MERGE). Next, genotyping of the DNAs was performed and compared to results from Affymetrix 6.0 SNP arrays to calculate True or False (Sequencing matches array data or not) and Positive or Negative (SNP or wild type). A logistic regression model was implemented to improve SNP calling performance before generating a final list of SNPs that could be further annotated by SIFT or PharmGKB.

4.1.3.3. Affymetrix SNP and CNV Analysis

As a control method for genotyping SNPs and copy number variations in HepaRG cells, we first determined the genotypes on the Affymetrix 6.0 SNP arrays. SNPs and copy number variations were determined using the Affymetrix 6.0 SNP array (processed at Washington University [St. Louis, MO]). Data were analyzed in-house using the Affymetrix Genotyping console (version 4.0), following the manufacturers specifications.

4.1.3.4. SNP calling performance

To assess the performance of the genotyping methods, the following guidelines were used. If a nucleotide called by the sequencing agreed with the genotyping array, then that position would be treated as a TRUE event, with all others being FALSE. If the genotype of the sequencing data was different from the reference allele, then that position would be treated as POSITIVE for a SNP with all others being NEGATIVE. Sensitivity (SNS) is defined as true positives (TP) divided by the sum of true positives and false negatives (FN). As such, specificity (SPC) is equivalent to TN / (FP + TN), accuracy (ACC) is (TP + TN) / (T + F), precision (PCN) is TP / (TP + FP), false discovery rate (FDR) = FP / (TP + FP), false positive rate (FPR) = FP/(T + F), negative predictive value (NPV) = TN / (TN + FN), and false negative rate (FNR) = FN / (TP + FN) (Lu et al., 2004). In other words, SNS is the percentage of SNPs correctly identified as being a SNP (i.e. power); SPC is the percentage of wild-type alleles correctly identified as being wild-type; ACC means the percentage of predictions that are correct; and PCN is the percentage of SNP predictions that are correct.

4.1.3.5. SNP Analysis

SNPs were called using SAMtools pileup with no filter, a minimum SNP score of 25 and 3-fold coverage (25 × 3), or a refinement of the 25 × 3 using a logistic regression model. In the pileup format, each line represents a genomic position, consisting of chromosome name, coordinate, reference base (the expected nucleotide), consensus base (the experimentally derived nucleotide), consensus quality (CNSq, a measure of confidence for the consensus call), SNP quality (SNPq, Phred-scaled probability of the consensus being identical to the reference) root mean square (RMS) mapping quality, the number of reads sequenced, and the actual bases called for each read along with information as to if the base was located at the end of the read. We ignored base calls that were different from the reference if they occur at the end of a read because those are most likely due to splicing-induced mapping errors (see Figure 4.1.5). The logistic regression model was performed using the built-in R function *glm*. SNPs passing through the 25 × 3 filter from sequencing, that were also called on the array, were further described as being a FP or not.

We built an additive logistic regression model using consensus base quality, SNP quality, and percent of reference allele (i.e. the percentage of reads at a single position that was the same as the reference allele). If the probability of being a FP was greater than 90%, then that SNP was not allowed to pass through the filter. The probability of being a FP was calculated using the following formula:

$$\frac{1}{1 + e^{-z}},$$

where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

or

$$z = 2.673 + 0.021(CNSq) - 0.022(SNPq) - 3.463(\%\ \text{Reads matching the reference})$$

The values for the beta parameters were optimized from cross-referencing SNP calls from the genotyping array for known SNPs. Each parameter was significant at $p < 10^{-5}$. The x parameters come directly from the SAMtools pileup for the unknown SNPs (i.e. those called by sequencing and not represented on the array). Using this formula, we expect that our model will produce false positives 5% of the time (see *Results).*

We used SIFT (Kumar et al., 2009) to evaluate the functional consequences of identified SNPs. Most of the SNPs identified resulted from the genomic alignment of reads spanning different exons, and were present only due to the imperfections of the alignment to the reference genome. Therefore, we used the SIFT coding filter to remove SNPs from these splice sites to give us only cSNPs (SNPs in cDNA regions). We used SIFT to evaluate the functional consequences of identified SNPs. When available, mutations were predicted to either be tolerated or damaging substitutions and if they changed an amino acid. Known mutations were then cross referenced with all Variant Annotations from the PharmGKB website (downloaded 6/6/10)(Klein et al., 2001).

4.1.4. Results

4.1.4.1. Affymetrix SNP array

More than 600,000 genomic positions passed quality control checks from the array. The distribution of genotypes was shown in Figure 4.1.2, with 18% homozygous non-reference (i.e.

not matching the reference allele), 67% heterozygous, and 15% were homozygous wild type. One percent of the probes were unable to produce genotype calls. Copy number variation in these cells confirmed that these cells were derived from a female patient as well as the known cytogenetic abnormalities of trisomy chromosome 7 and deletion of the p arm of chromosome 12. An interesting and novel result was that we also saw an extra copy of chromosome 2, which had not been previously reported for these cells. Genotypes from this array were used to assess the performance metrics for SNP quality filtering.

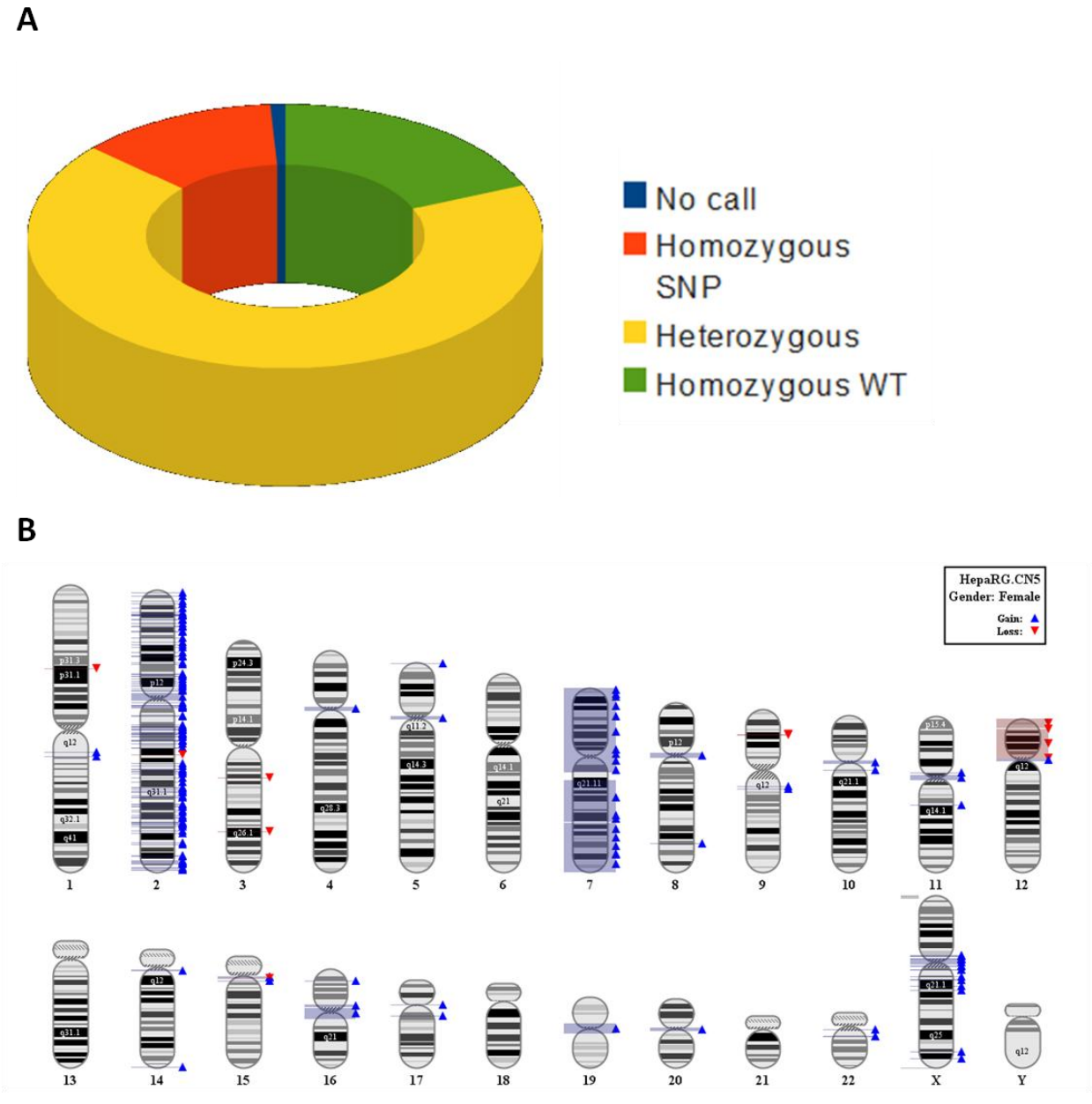


Figure 4.1.2. A) Genotypes of SNPs passing quality filtering from Affymetrix 6.0 Array. B) Karyogram representation of copy number variation in HepaRG.

4.1.4.2. Effect of Read Aligners on Genotype Calling

To assess the impact of read alignment programs on SNP calling performance, we called SNPs from all 6 samples independently using the SAMtools pileup. Average performance metrics for each aligner was given in Figure 4.1.3. SOAPals was able to genotype more true positives and true negatives than TopHat, however total performance metrics such as SNS, SPC, ACC, PCN, and FDR were similar, with essentially no difference in performance (all values differed by less than 1%). This suggested that the type of aligner, given comparable alignment parameters, did not significantly affect the performance of genotyping in RNA-Seq datasets.

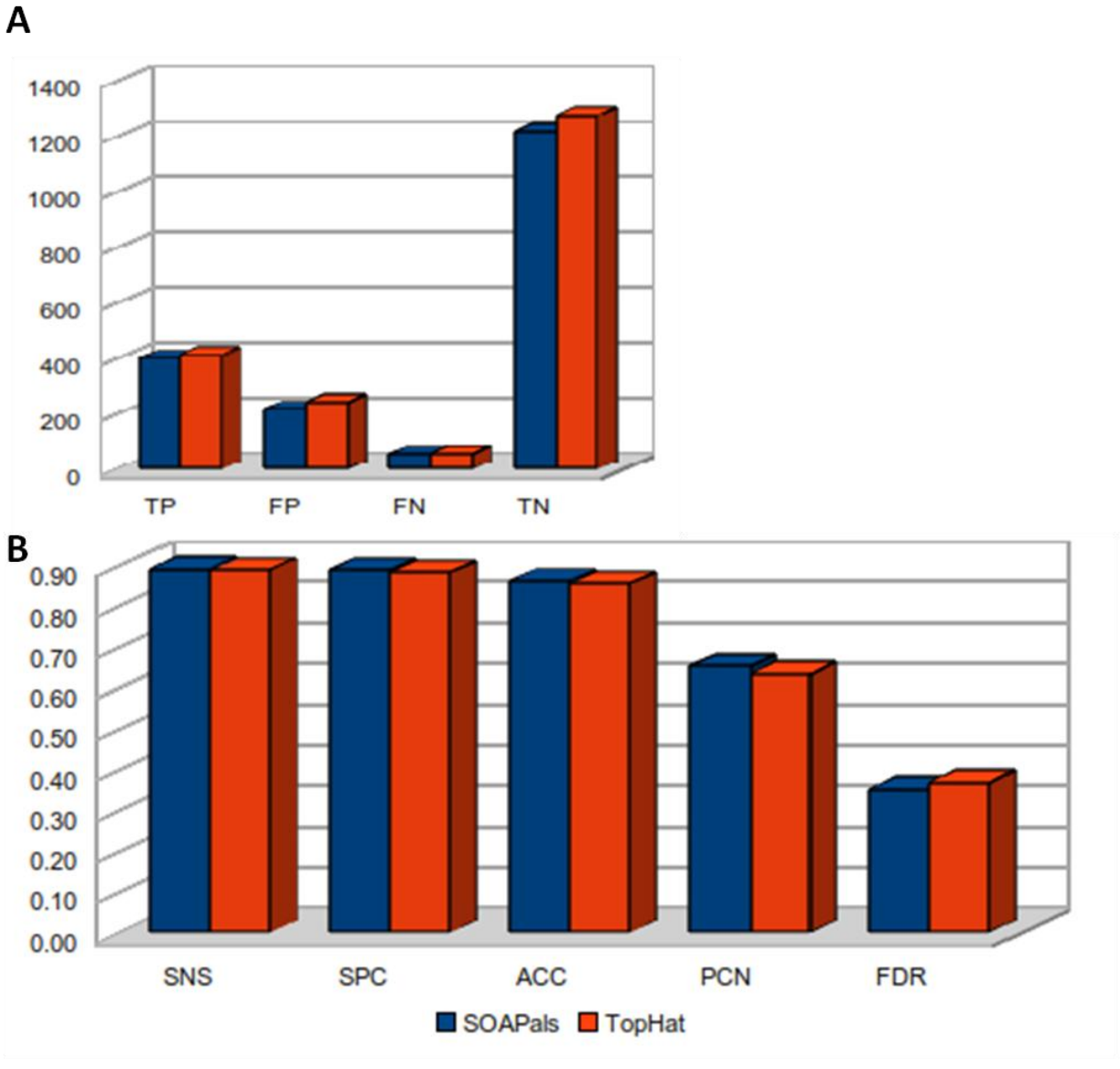


Figure 4.1.3. Performance metrics of split-read aligners TopHat and SOAPals. (A) TP: true positive; FP: false positive; FN: false negative; TN: true negative; (B) SNS: sensitivity; SPC: specificity; ACC: accuracy; PCN; precision; FDR: false discovery rate.

4.1.4.3. Performance metrics

When we combined the six RNA-Seq datasets from TopHat alignment results, SAMtools pileup found 1,882 out of the 7,077 cSNPs that were detected and passed QC filter on the Affymetrix array were expressed in HepaRG. These 1,822 SNPs served to define the true genotypes of the HepaRG cSNPs. We then compared three SNP filtering methods: SAMtools pileup no filter, SAMtools pileup 25 × 3, or SAMtools pileup 25 × 3 plus a logistic regression model to decrease the number of FP (see *Methods*, Figure 4.1.4). By implementing a basic filter (25x3), only marginal improvements to genotyping performance were observed. SNS dropped from 89% to 64% and a concomitant increase in SPC from 85% to 91%. The FDR, however, remained high at 26%. At a higher cost of sensitivity, the specificity and precision increased, and importantly, the false discovery rate dropped from 35% to 5%. A drop in accuracy was observed but it was complemented by an increase in precision (from 65% to 95%).

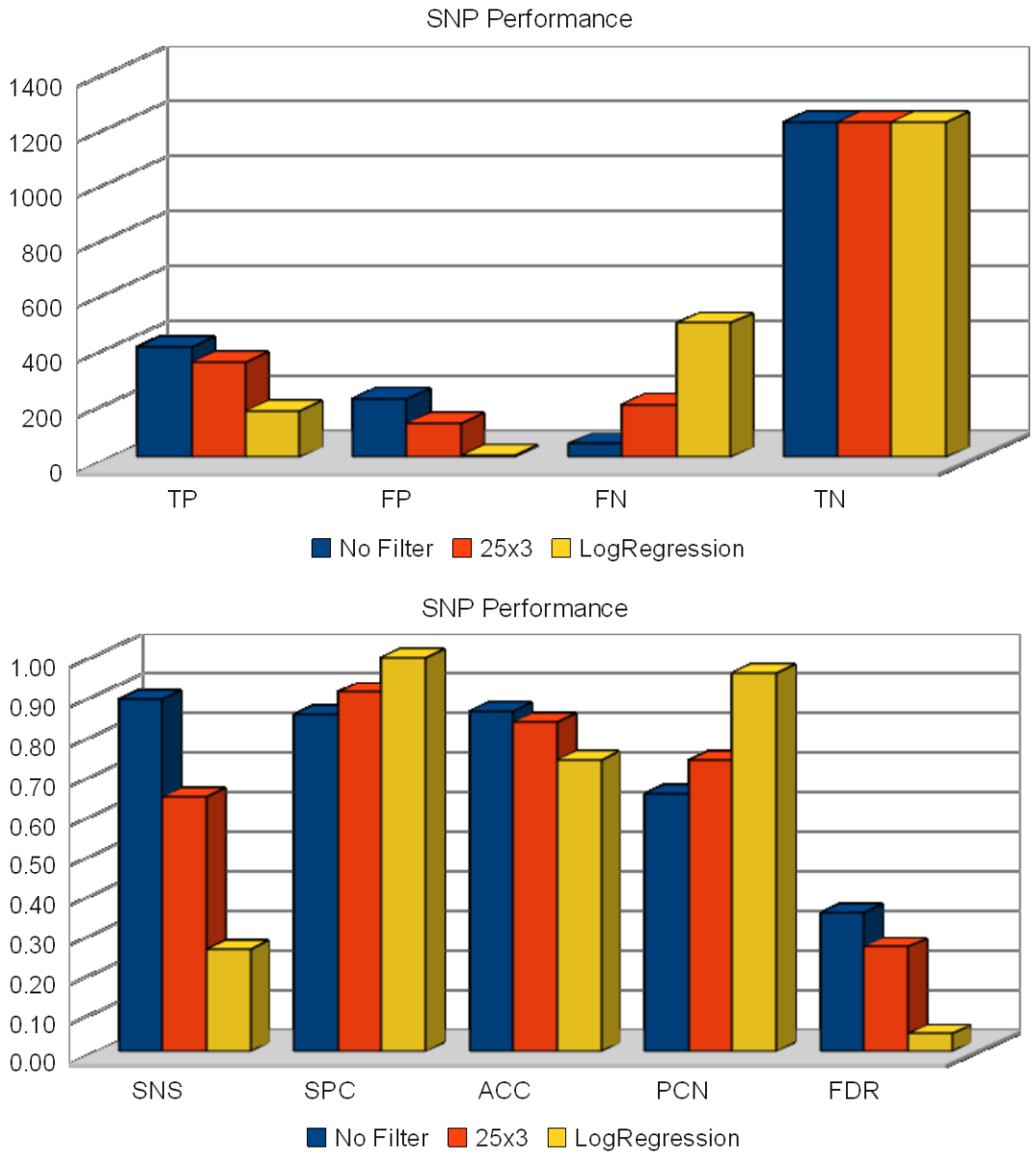


Figure 4.1.4. Performance metrics for TopHat with three different filtering options. TP: true positive; FP: false positive; FN: false negative; TN: true negative; SNS: sensitivity; SPC: specificity; ACC: accuracy; PCN; precision; FDR: false discovery rate.

4.1.4.4. Pharmacogenomics of HepaRG

4.1.4.4.1. SIFT

By applying the pileup 25 × 3 plus logistic regression model to all the sequencing datasets, we were able to identify 5,012 cSNPs in HepaRG cells. Of these, functional predictions by SIFT could be made for 2,283 polymorphisms (Table 4.1.1). 85% of the SIFT predicted SNPs were unlikely to be damaging to protein function, but the remaining 15% (321) were potentially damaging. We were able to cross-reference the coding SNPs with dbSNP130 to determine whether an individual SNP had previously been reported. Only 18% of the 5,012 SNPs were novel, which was surprising given that the origin of this cell line is from a hepatoma, and cancers are notorious for inducing somatic nucleotide changes. Previous mRNA-Seq studies in HeLa cells revealed 5,928 SNPs of which 38% were not in dbSNP (Morin et al., 2008), although the authors also noted a high degree of false positives in their dataset as well. In Jurkat T-cells, Chepelev *et al.* (2009) identified 12,176 variations of which 39% were novel. These indicated that the mutation rate of HepaRG cells was relatively low compared to other cell lines, and most of the genetic variations were likely polymorphisms rather than mutations.

Table 4.1.1. Results from SIFT.

Number of input		100%	5,012
Non-coding SNPs		1%	(22 out of 5012)
Coding variants		99%	(4990 out of 5012)
Coding variants predicted		45%	(2283 out of 4990)
	Tolerated	85%	(1962 out of 2283)
	Damaging	15%	(321 out of 2283)
Non-synonymous		46%	(2316 out of 4990)
Synonymous		54%	(2497 out of 4990)
Novel		18%	(914 out of 5012)

4.1.4.4.2. PharmGKB

PharmGKB (Thorn et al., 2010) serves as a major repository for collecting and curating information linking genetic variation to altered drug response. We took advantage of the variant annotations present in PharmGKB to rapidly discern any known pharmacogenetic implication for SNPs identified by RNA-Seq. This approach was also recently used in the whole genome sequencing of patients with a family history of vascular disease and early sudden death (Ashley et al., 2010). Table 4.1.2 shows a condensed version of the variant annotation report on HepaRG cells. Several variants were observed in the *ABCB1* gene that encodes the multi-drug resistance transporter MDR1. Figure 4.1.5 shows an example of the RNA-Seq data alignment results. The reference coordinates were present in the 1st line, followed by the reference base, and the consensus base. The remaining lines showed the alignment of individual reads. The reference base was from GRCh37/hg19 genome assembly, whereas the consensus base was derived from the sequences present in HepaRG. At position "A", the SNP rs2032582 causes a serine to threonine change at amino acid position 893 in the *ABCB1* gene (a.k.a. MDR1, Pgp). This particular base was sequenced 147 times with 146 individual reads calling the "C" allele, indicating a homozygous polymorphism. This and other SNPs of HepaRG in the *ABCB1* gene were present in the PharmGKB database (Table 4.1.2). Closer inspection of the functionality of this SNP revealed its impact was obscure. For example, both positive (Yamauchi et al., 2002) and negative (Mai et al., 2004) findings regarding its impact on pharmacokinetics and drug response phenotype for tacrolimus administration had been reported. Importantly, we noted that the PharmGKB annotation of this polymorphism was an alanine, but the reference genome codon would be serine. It was not known why this difference exists, but it was not due to different genomic assemblies (PharmGKB annotations use hg18 coordinates, whereas this study uses hg19

coordinates). Also notice a "T" polymorphism in the consensus sequence (position "B"). This particular SNP was at an exon boundary that was marked by a precipitous decline in sequence coverage, and it did not pass our filtering quality control. This SNP calling was most likely a result of aligning cDNA reads to genome space; that is, if one allows for up to two mismatches in the alignment (as is the case in many experiments), several polymorphisms will be identified by mistake in and around the splice site of exons. Therefore, when analyzing RNA-Seq data for genetic variations, it is imperative to remove SNPs that lie in splice sites or introns, unless that intron is aberrantly retained in the cDNA, and the entire intron is sequenced to a sufficient depth. Other polymorphisms linked with altered pharmacokinetics for several drug types were listed in Table 4.1.2.

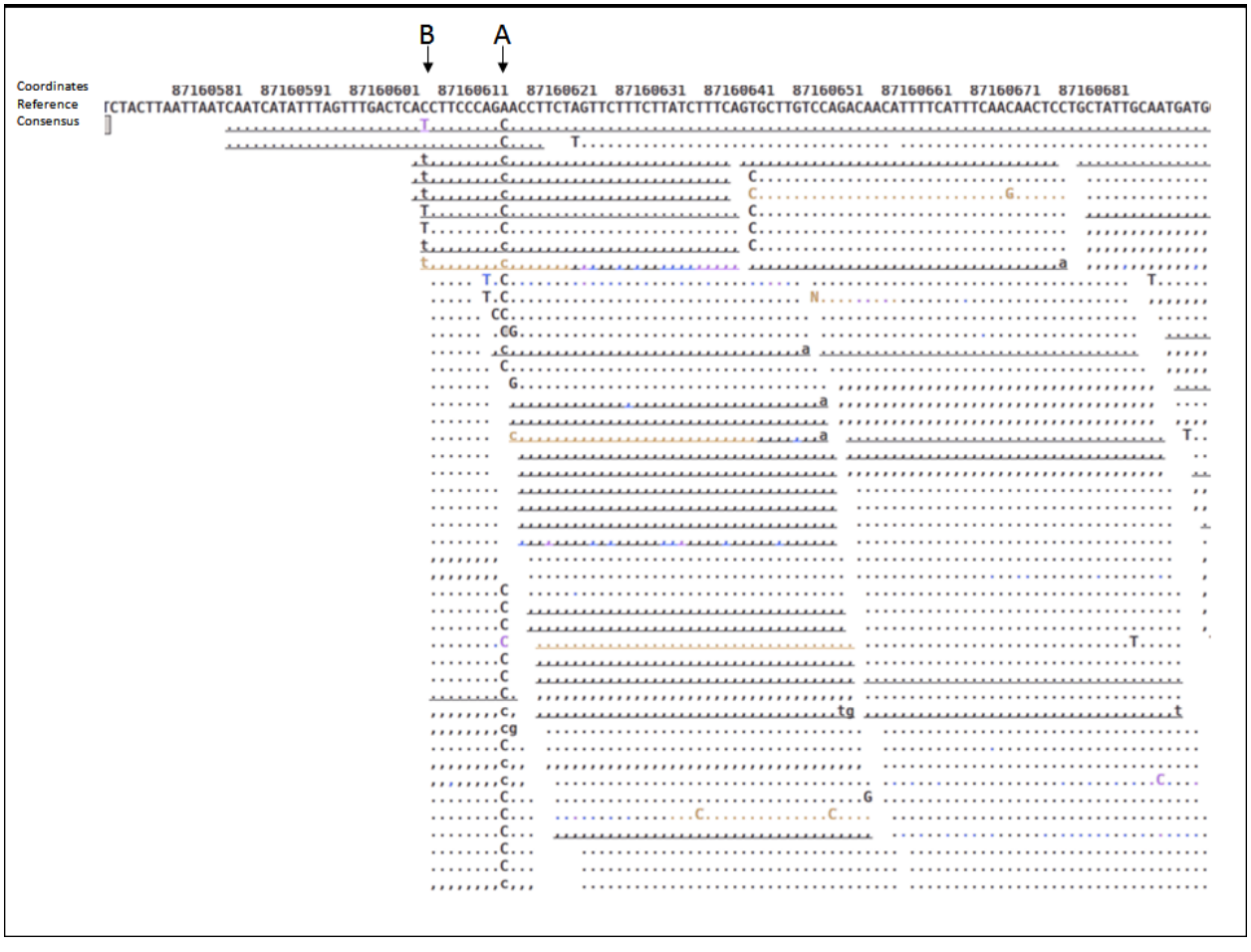


Figure 4.1.5. Representative example of sequencing alignment. The reference coordinates were present in the 1st line, followed by the reference base, and the consensus base. The remaining lines showed the alignment of individual reads.

Table 4.1.2. Results from PharmGKB

Gene	dbSNP ID	Position	AA Change	Drugs influenced by change	PubMed ID
ABCB1	rs2032582	chr7:86998554	A893S/T	atorvastatin	18851956
ABCB1	rs1045642	chr7:86976581	I1145I	atorvastatin	18851956
ABCB1	rs1128503	chr7:87017537	G412G	cyclosporine; digoxin; verapamil; vinblastine	12781336
ABCB1	rs2032582	chr7:86998554	A893S/T	cytarabine; idarubicin	16331627
ABCC1	rs35605	chr16:16069520	L562L	irinotecan; SN-38	19940846
ADRB2	rs1042713	chr5:148186633	R16G	atenolol; verapamil	18615004
CBR3	rs8133052	chr21:36429371	C4Y	doxorubicin	18551042
CYP2C8	rs10509681	chr10:96788739	K399R	amodiaquine	18855526
CYP2C8	rs11572080	chr10:96817020	R139K	amodiaquine	18855526
DCTD	rs4742	chr4:184052682	V105V	gemcitabine	15224082
DPYD	rs1801265	chr1:98121473	T85C	fluorouracil	17848752
F4	rs6025	chr1:167785673	R534R	drotrecogin alfa; tamoxifen	PGKB
HNF1A	rs2464196	chr12:119919810	S486D		18439552
HSPA1L	rs2227956	chr6:31886251	T493M	carbamazepine	16538175
LRP2	rs2075252	chr2:169719231	K4094E	cisplatin	17457342
MTHFR	rs4846051	chr1:11777044	F435F	methotrexate	16439441
NAT2	rs1208	chr8:18302596	R268K	clonazepam	19356010
SLC1A1	rs2228622	chr9:4554432	T138T	clozapine; olanzapine; risperidone	19884611
SLC28A1	rs2242048	chr15:83279414	Q456Q	gemcitabine	15224082
SLCO1B1	rs2306283	chr12:21221005	N130D	Associated with decreased pravastatin plasma AUC.	PGKB
SLCO1B1	rs11045819	chr12:21221080	P155T	fluvastatin	18781850
SLCO1B3	rs7311358	chr12:20907027	M233I	cyclosporine; mycophenolate mofetil; sirolimus; tacrolimus	19890249
SLCO1B3	rs4149117	chr12:20902747	S112A	mycophenolate mofetil; mycophenolic acid; sirolimus; tacrolimus	19890249
TGFB1	rs1800470	chr19:46550761	L10P		11740340

4.1.5. Discussion

4.1.5.1. Limitations of RNA-Seq

RNA-Seq holds significant promise for pharmacogenomics, however it does present inherent limitations, most notably if a gene is not expressed it will not be sequenced. This could prevent the identification of very important genetic changes. For example, CYP2D6 is not expressed in this cell line. To ascertain whether there is a genetic origin for the failed expression of CYP2D6, we used the conventional sequencing strategy from Gaedigk et al (2005) (data not shown). We found three polymorphisms in the *CYP2D6* gene: rs16947, rs1135840, and rs5030656. The rs16947 and rs1135840 SNPs resulted in amino acid changes R296C and S486T, respectively. These two genotypes corresponded to the *CYP2D6*2* haplotype. The rs5030656 SNP resulted in a frame shift K281del responsible for the *CYP2D6*9* haplotype. If one is not careful, the absence of these mutations in a final SNP list generated from RNA-Seq might lead to faulty assumptions about CYP2D6 genotype and associated phenotype, although in this case we also used the expression information to flag this gene as not expressed and therefore were not expecting to identify mutations in the final SNP list. SNP calling from RNA-Seq reads shows low sensitivity and a high false negative rate.

The use of RNA-Seq for pharmacogenomics, like all assays, has its advantages and limitations. Clearly, the sheer volume of sequence information is much more cost-effective than traditional Sanger sequencing, and one is not dependent *a priori* on knowledge of whether a particular gene is likely a contributing genetic factor to the response being measured. Rather, a sequencing-based survey of expressed genes can offer new ways to understand complex genotype/phenotype relationships. Increasing read length for sequencing and more advanced

algorithms for identifying genetic structure in individuals and populations are rapidly advancing our ability to analyze and interpret sequencing data, and will ultimately be a necessary tool in every geneticist's toolbox. In the meantime, care should be taken to interpret SNP-calling from RNA-Sequencing results.

Here, we show that many improvements can be made to SNP-calling algorithms that will be advantageous for RNA-Seq data analysis. The performance of SNP-calling was improved by filtering out highly probable false positives, but at the same time increased the false negative prediction rate. As with the example of *CYP2D6*, the absence of SNP data is not indicative of wild-type, rather it may be due to the lack of mRNA expression or more than 2 base differences between one's sample and the reference genome being used. An alternative strategy to looking at cSNPs is to perform exome sequencing (Ng et al., 2010), which relies on hybridization of probes corresponding to exonic DNA sequences and then sequencing the interacting sequences. GC-rich exons do not hybridize well and are poor templates with current sequencing methods. Also, 17-23% of coding sequences in the RefSeq database are not targeted by commercial providers including insulin, ABO blood group genes, and many others [http://www.genomeweb.com/sequencing/current-whole-exome-capture-products-omit-important-genes-nci-researchers-find]. Plus, this approach is not able to discern alternative splicing, quantitate gene expression, or identify cases of RNA-editing. Regardless of its current difficulties, the power and unique data characteristics it provides make RNA-Seq an attractive tool that will be used more and more by many investigators.

4.2. mRNA-Seq Analysis of HepaRG treated with drugs

4.2.1. Abstract

The simplest way to calculate gene expression from mRNA-Seq data is to count the number of sequencing reads that lie in the genomic coordinates of a particular gene. However, 14% of human genes overlap in genomic coordinates due to them being coded for on both the sense and antisense strands. Assigning reads to the correct genes to these overlapping coordinates therefore requires some adjustment. In this subchapter, we present a custom program, PRUNE, to address this limitation. We use *in silico* and *in vitro* methods to validate our program. For the *in silico* approach, we simulated an mRNA-sequencing experiment with 10 million paired end reads and compared the true gene expression level to gene expression estimates from PRUNE ($r^2 = 0.742$) and another commonly used program, Cufflinks ($r^2 = 0.708$). For the *in vitro* approach, we compared gene expression measurements and differential expression analysis between biological replicates in which very few genes should be differentially expressed. Cufflinks and its differential expression tool CuffDiff called 37% of genes as differentially expressed, whereas PRUNE read allocation followed by DESeq differential expression testing found no genes differentially expressed between the replicates. Because PRUNE is more accurate in estimating gene expression and its coupling with DESeq produced no false positives, we used this strategy to induce gene expression in HepaRG cells using well-known nuclear receptor agonists (rifampicin, phenobarbital, and dexamethasone), and describe the HepaRG transcriptional response.

4.2.2. Introduction

Unlike microarrays, there is no standard approach to analysis of mRNA-Seq. This is highlighted in the many different approaches that existing analysis tools take to analyze and interpret data. Alexa-Seq (Griffith et al., 2010) , Cufflinks (Trapnell et al., 2010) ERANGE (Mortazavi et al., 2008), HT-Seq (Anders and Huber, 2010), and BEDtools (Quinlan and Hall, 2010) are some of the available tools. As warned by the authors, Alexa-Seq is computationally intensive software that requires highly paralleled computing and several ancillary programs that need to be managed by a dedicated network administrator. Cufflinks and ERANGE do not report expression in units of raw tag counts; rather they use the RPKM unit (**R**eads **P**er **K**ilobase of exon per **M**illion fragments mapped). The problem with RPKM is that it is heavily influenced by a small fraction of highly expressed genes (Bullard et al., 2010). Furthermore, there are inherent differences in how these programs compute RPKM. Cufflinks uses only standard annotated exon models to count reads belonging to a gene, but ERANGE includes new exon sizes and counts within them into total reads and exon length, thereby influencing RPKM (Pepke et al., 2009). HT-Seq ignores reads lying in genes sharing chromosomal coordinates, whereas BEDtools will actually count them twice. These discrepancies cause serious concern since approximately 14% of human genes in the RefSeq database fit this rule, usually one 5'-UTR overlapping with a 3'-UTR from a gene on the opposite strand. For this reason, the purpose of this subchapter is to improve the quantitation of gene expression by developing a software program that will export read-level data to a given gene without duplicating the read counts or discarding ambiguous reads in exons that are overlapping in their genomic coordinates.

In this study, we will examine the transcriptional response of HepaRG cells to prototypical nuclear receptor agonists using mRNA-Seq. RNA-Seq is highly accurate for

quantifying expression levels (Mortazavi et al., 2008; Nagalakshmi et al., 2008), and is highly reproducible for both technical and biological replicates (Cloonan et al., 2008; Nagalakshmi et al., 2008), with greater sensitivity than microarrays (Wang et al., 2009). Because this is a stable cell line, genetic heterogeneity is removed, as have been reported from one human hepatocyte population to another (Madan et al., 2003), which simplifies our understanding of the transcriptional response to these drugs.

Some nuclear receptors, such as CAR and PXR, are ligand-dependent transcription factors that act as xenosensors to respond to environmental perturbations. They are especially important in liver where they are master regulators of hepatic development and function. Pharmaceutical compounds, and other xeno- and endobiotics can serve as nuclear receptor ligands to induce expression of drug metabolizing enzyme gene expression such as Cytochromes P450. Cytochrome P450 enzymes, which metabolize many drugs, are induced until drug concentration can longer induce gene expression through the nuclear receptors, at which time the cytochrome P450s return to basal expression state.

Unlike other hepatocyte cell lines, HepaRG cells express nuclear receptors. We treated the hepatoma cell line HepaRG with prototypical inducers of drug metabolizing genes. Rifampicin (Rif), Phenobarbital (PB), and Dexamethasone (Dex) are all potent xenobiotics that induce hepatocyte gene expression *in vivo*. The mRNA content of HepaRG have been shown to replicate that of primary human hepatocytes and liver tissue better than the commonly used HepG2 cell (Hart et al., 2010), and this is supported by activity of several phase I and phase II drug metabolizing enzymes (Aninat et al., 2006). HepaRG cells have been shown to respond appropriately to Rif and PB, resulting in induction of CYP1A1, CYP1A2, CYP2B6, CYP2C8,

CYP2C9, CYP2C19, and CYP3A4 *in vitro* (Kanebratt and Andersson, 2008b; Lambert et al., 2009a; Lambert et al., 2009b; Lambert et al., 2009c).

4.2.3. Materials and Methods

4.2.3.1. Chemicals.

Rifampicin, phenobarbital, dexamethasone and DMSO were purchased from Sigma-Aldrich (St. Louis, MO). Rifampicin, dexamethasone and phenobarbital were dissolved in DMSO to obtain stock solutions of 20 mM, 300 mM and 1500 mM, respectively. The SuperScript III First-Strand Synthesis System for reverse transcription-polymerase chain reaction (PCR) and TRIzol were obtained from Invitrogen (Carlsbad, CA).

4.2.3.2. HepaRG Cell culture.

HepaRG cells and culture medium were provided by Biopredic International. The undifferentiated HepaRG cells were seeded at 0.2 million cells/well in 6-well plates, maintained in the growth medium for two weeks, and then cultured in the differentiation medium containing 2% dimethyl sulfoxide (DMSO) for two more weeks to obtain the differentiated HepaRG cells . The differentiated HepaRG cells were incubated with serum-free growth medium for 24 hours, then incubated with solvent control (0.1% DMSO), rifampicin (10 µM), phenobarbital (750 µM) and dexamethasone (200 µM), respectively, for 24 h. Total RNA from HepaRG cells was prepared using TRIzol reagent according to the manufacturer's protocol.

4.2.3.3. Illumina Sequencing, Mapping, and Read Counting

Total RNA from HepaRG cells was used for RNA-Seq. The RNA-Seq experiments were performed by Genpathway Inc. (San Diego, CA, USA). Briefly, a population of poly (A)+ RNA was selected and converted to a library of cDNA fragments (175 to 225 bp) with adaptors attached to both ends using an Illumina mRNA-Seq sample preparation kit (San Diego, CA, USA). The quality of the library preparation was confirmed by analysis on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The cDNA fragments were then sequenced on an Illumina Genome Analyzer IIx to obtain 36-bp sequences from both ends. Reads were mapped to the reference human genome hg19 using TopHat v.1.1.2 (Trapnell et al., 2009) with the following options: --solexa1.3-quals, --mate-inner-distance 30, -G (RefSeq.GTF (hg19)), and --no-novel-juncs. Then, Cufflinks and CuffDiff were used for transcript assembly using the same reference GTF file, minimum mapping value of 10, and the quantile normalization option.

To overcome the problem of assigning reads to a given gene without duplicating the read counts or discarding ambiguous reads in exons that are overlapping in their genomic coordinates, we created a custom program. The following paragraph describes what the program, called PRUNE, actually does. Using a RefSeq GTF annotation file, overlapping exons from each transcript for each gene are collapsed with other exons on the same strand into a single feature to create a non-redundant set of exon features (using parts of BEDtools code (Quinlan and Hall, 2010)). Each non-redundant exon is then categorized as being unique (Uex) or non-unique (NUex) to a given gene (Figure 4.1). Then, the number of times fragments were sequenced within the genomic intervals for Uex and NUex are counted and summed them for each gene, separately. To allocate reads lying in non-unique exons, the unique reads per base (URpB) for the overlapping genes are calculated and summed. The proportions of each gene's contribution

to the summed URpB is then used to allocate reads in NUex (NUex.rc). To calculate gene expression, the sum of the read counts from Uex and NUex.rc is calculated for each gene. The formula is given here:

$$\Gamma_i = \mu_i + \eta_{ij}\left(\frac{\mu_i}{\mu_i + \mu_j}\right)$$

where μ represents $\frac{the\ number\ of\ reads\ in\ the\ unique\ exons}{the\ sum\ of\ the\ exon\ lengths}$

and η represents $\frac{the\ number\ of\ reads\ in\ the\ non-unique\ exons}{the\ sum\ of\ the\ exon\ lengths}$

To calculate gene expression level counts, Γ, for gene *i* that overlaps with gene *j*, the number of reads in the non-unique exon, η_{ij}, is multiplied by the ratio of the number of sequence counts per base in the unique exons, μ, of gene *i* to ratio of the number of sequence counts per base in the unique exons of gene *j*. This ratio determines the proportion of reads in the non-unique exons that belong to gene *i*. Therefore, the total counts for gene *i* is the proportion of reads in the non-unique exons that belong to gene *i* plus the number of reads in all unique exons in gene *i*. These read counts were used for statistical analysis in DESeq,

4.2.3.4. Statistics

To test for differential expression, the read counts from genes (allocated as described above) were used as input for the DESeq R package (Anders and Huber, 2010). Unlike other tools for significance testing of RNA-Seq data, DESeq considers both biological and technical variance in determining whether a gene is differentially expressed. In the case of phenobarbital

and dexamethasone treatment, only one replicate is available, so we use the variance implied from the DMSO or rifampicin treated samples (whichever is greater) as a surrogate parameter. Then means and variances from each sample are compared to one another and *p*-values are adjusted for multiple testing as described (Benjamini and Hochberg, 1995). Significant pathways were interrogated for preferentially-induced transcripts using the Functional Annotation Clustering Tool in DAVID (Dennis et al., 2003)

4.2.3.5. Simulating mRNA-Seq

To assess the accuracy of PRUNE and Cufflinks, we simulated RNA-Seq experiments using the Flux Simulator (http://flux.sammeth.net/simulator.html). Flux Simulator provides an in silico production of the experimental pipelines for RNA-Seq, adopting a set of parameters. We set NB_MOLECULE (total number of RNA molecules in the sample) to be 50 million, reverse-transcribed cDNA molecule range from 500 to 5500 bp and paired end 36bp reads. Flux Simulator produced 10 million RNA-Seq reads from the UCSC hg19 mRNA isoform annotation (refGene). The simulated reads were run on Cufflinks/CuffDiff and PRUNE/DESeq.

4.2.4. Results

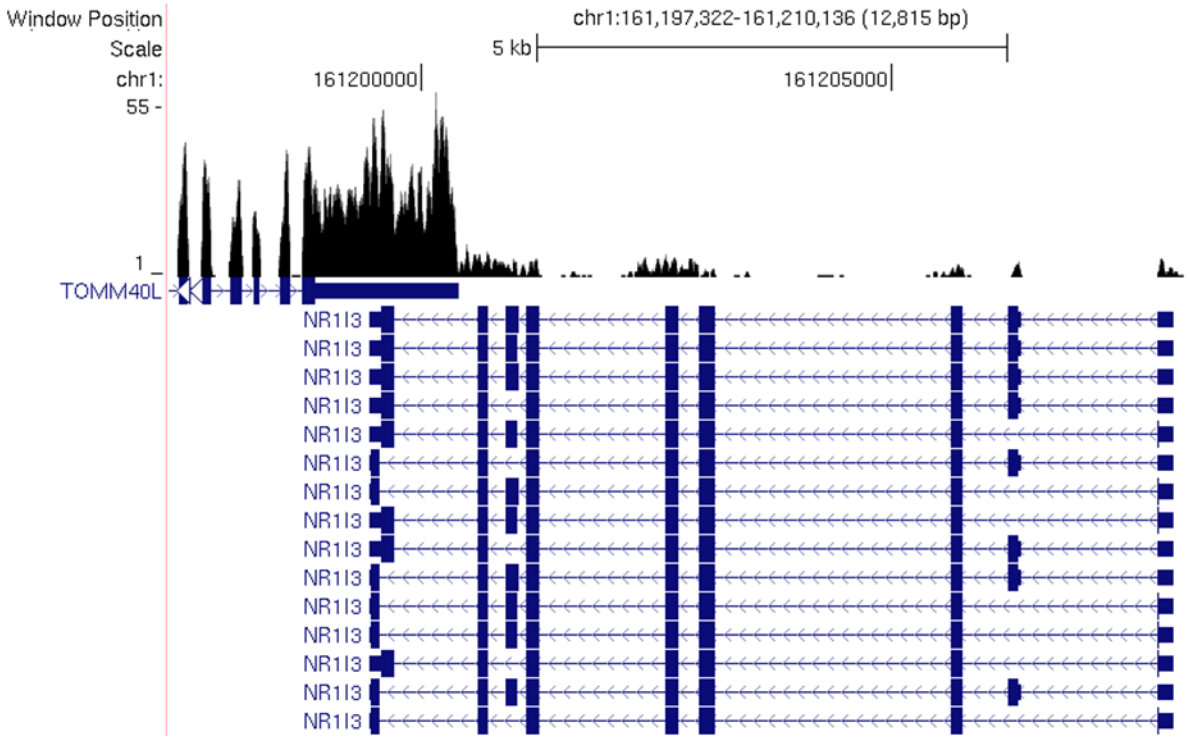
4.2.4.1. Calculation of gene expression

An example of difficulty assigning reads to individual genes is shown in Figure 4.17A. Here, the nuclear hormone receptor *NR1I3* (a.k.a. *CAR*) overlaps in genomic coordinates with *TOMM40L*. Clearly, *TOMM40L* is expressed at a greater level than *NR1I3*, however, most tools count the number of reads that are found within the genomic coordinates of a given gene. Two

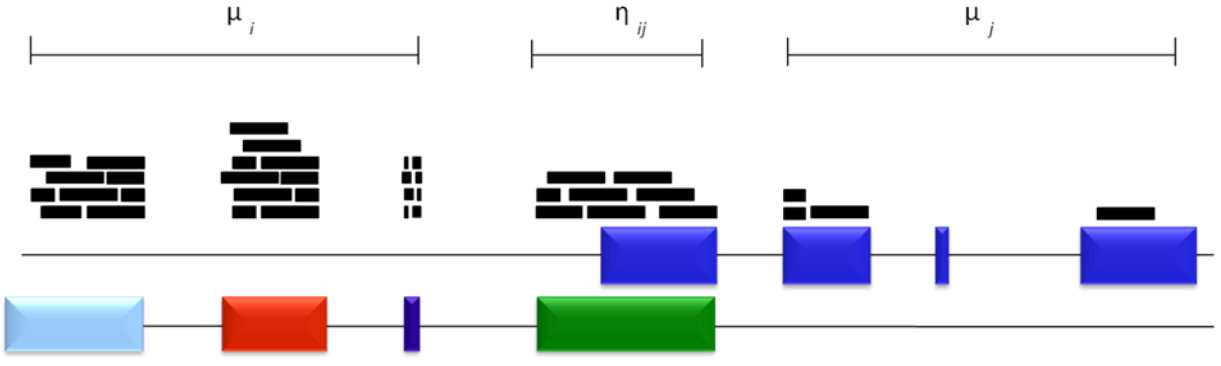
possible scenarios will occur: either the reads will get counted twice (once for *NR1I3* and once for *TOMM40L*), or the ambiguous reads will be discarded. Either way, the expression level of *NR1I3* is distorted. If the reads are discarded, then the assessment of gene expression will be lower for *NR1I3,* but if the reads are counted, then *NR1I3* will be reported as having higher expression than it actually does.

To solve the problem, consider they toy example in Figure 4.16B. For each gene, we distribute the reads in the contested exon based on the proportion of unique reads per base in each intersecting gene's unique exonic regions. In this way, each read will only be counted once, and the final counting of reads aligning to exons can be calculated. Reads not aligning to annotated exons were discarded.

A



B



$$\Gamma_i = \mu_i + \eta_{ij}\left(\frac{\mu_i}{\mu_i + \mu_j}\right)$$

Figure 4.16. A) Example of difficulty assigning reads to individual genes. Here, the nuclear hormone receptor *NR1I3* (a.k.a. *CAR*) overlaps in genomic coordinates with *TOMM40L*. Clearly, *TOMM40L* is expressed at a greater level than *NR1I3*, however, most tools count the number of reads that are found within the genomic coordinates of a given gene. B) To allocate reads in overlapping exons, we propose a formula to allocate the reads relative to the proportional expression of reads from unique exons in the overlapping genes. μ_i, unique reads per base for gene i; μ_j, unique reads per base for gene j; η_{ij}, non-unique reads lying in coordinates shared by genes i & *j*; Γ_i, read counts allocated to Gene *i*.

Table 4.9. Sequencing metrics

	DMSO_1	DMSO_2	Rif_1	Rif_2	PB	Dex
Total Sequenced	45,261,906	70,807,932	67,209,866	71,941,568	61,016,010	70,566,746
Total Mapped to Genome (PE-only)	17,336,128	47,815,646	48,343,164	46,271,990	41,923,880	46,192,616
Estimated From PRUNE	12,146,826	32,973,645	33,336,921	31,088,398	28,012,049	31,549,424
Estimated From PRUNE (nonZero All Samples)	12,142,748	32,954,453	33,314,822	31,065,813	27,994,169	31,529,914
Estimated From PRUNE (Normalized)	**27,384,866**		**25,224,810**		**27,127,968**	**27,129,632**

4.2.4.2. Comparing calculations of gene expression

To validate to the optimal functionality of PRUNE, we used FluxSimulator to generate 10 million artificial sequencing reads with a known origin and compared those data to what was observed from Cufflinks of PRUNE gene-level quantitation. The results are shown in Figure 4.17. The r^2 value for Cufflinks was 0.708, whereas it was 0.742 from PRUNE, indicating prune estimates are more reflective of the actual counts than Cufflinks.

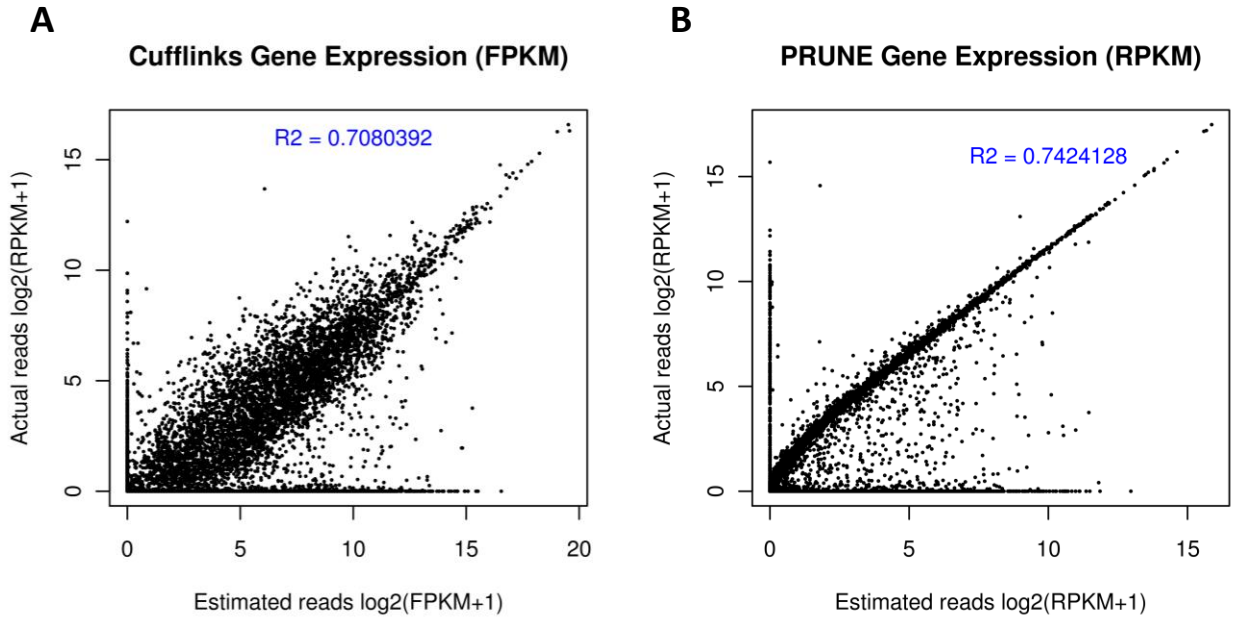


Figure 4.17. Simulation comparison between Cufflinks FPKM Data (A) and PRUNE RPKM Data (B). The x-axis is the values reported by each program, whereas thy Y axis reports the actual number reads. Correlation coefficients (r^2) are also given.

4.2.4.3. Significance testing for transcript and gene expression

Between 45 and 72 million reads were sequenced for each sample (average = 64,467,338; standard deviation = 10,213,244 (see Table 4.9). Between 17M and 48M reads were mapped to the genome while retaining their paired-end properties. Then all reads in all exons were added from all conditions to determine sequencing depth. Genes were removed if they did not have at least one read in all samples. Because depth varied considerably and raw tag counts are more accurate for differential expression testing than RPKM values (Oshlack and Wakefield, 2009), we required a normalization scheme that maintained the expression measurements in units of read counts. Simply normalizing the expression level by the length of the gene will remove the bias for expression level but also introduces a bias toward giving longer genes an increased likelihood of being differentially expressed. Thus, we used a linear scaling method to compare tag counts between samples. DESeq was used to scale each library (Figure 4.18), capture the mean and variance within and among replicate samples, and to determine differential expression.

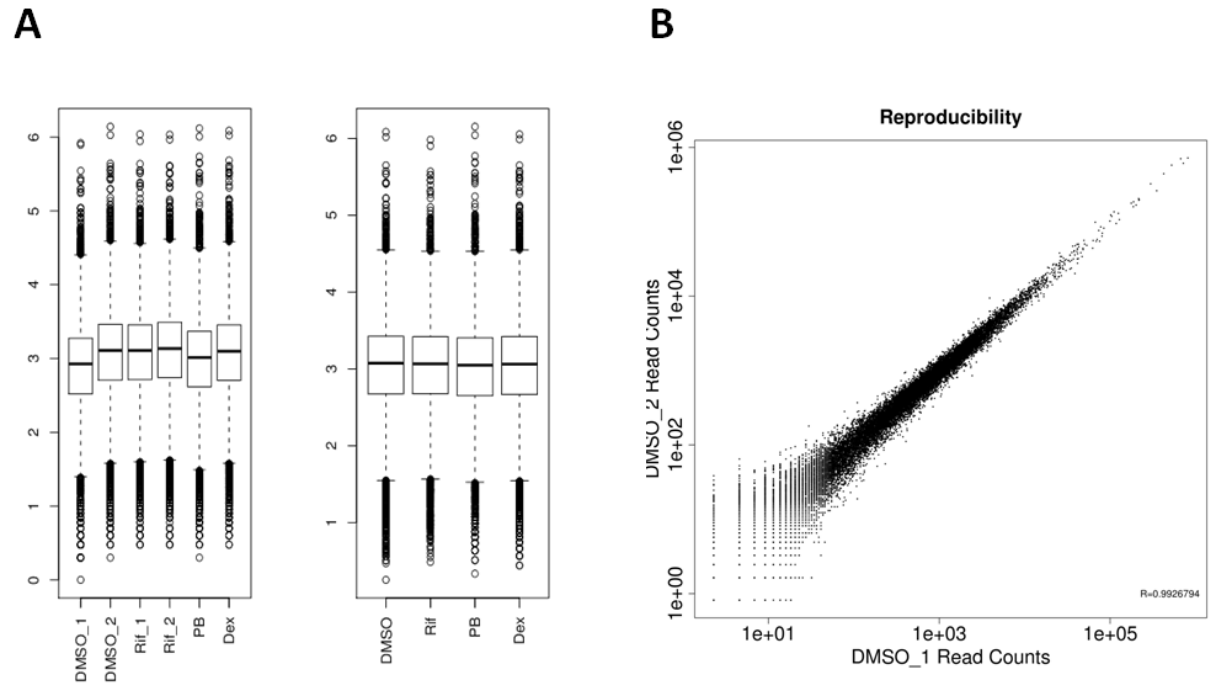


Figure 4.18. Replication and Normalization of mRNA-Seq Gene Expression. A) The left panel shows raw gene expression levels in all 6 samples, while the right panel shows the gene expression levels for the four treatment groups normalized for sequencing depth. Replicate samples were merged during the normalization procedure. B) Assessment of the reproducibility of the raw tag count data. Using read count expression measurements for each gene as a metric, the reproducibility of our experiments was assessed. Even though the sequence depth for DMSO_1 is much lower than DMSO_2 (Figure 4.A, left panel), there is still a strong correlation between gene expression levels: R=.99.

4.2.4.4. Differential Gene Expression

Because PRUNE represents a new method for quantifying gene expression, we compared our results with that of popular software: Cufflinks and its differential expression testing suite CuffDiff (Trapnell et al., 2010). We evaluated the false discovery rate of each pipeline by comparing the DMSO_1 sample against the DMSO_2 sample, under the assumption that no (or very few) genes would be differentially expressed because they were cultured under similar conditions (Figure 4.19). Of the 15,432 genes reported by Cufflinks to have expression values greater than 0 in both DMSO replicates, CuffDiff was able to test the expression of 7,663 genes, 2,852 of which were classified as being differentially expressed. The remaining genes could not be tested by CuffDiff. This represents a false discovery rate of 37%. Using PRUNE read allocation followed by DESeq, we were able to detect and test the expression 21,974 genes, none of which were differentially expressed between the DMSO replicates. Similar results were observed for Rifampicin replicates (data not shown). Even though the same alignment data were used to estimate the expression of genes and their respective fold changes, the reported levels are strikingly different. Figures 4.17 and 4.19 attempt to explain why such a disagreement exists between the two analysis tools. In Figure 4.17, the gene expression values reported by Cufflinks show much more orthogonal variation then PRUNE. If gene expression is not accurately assigned, it will ultimately affect downstream statistical interpretation. Second, as shown in Figure 4.19, the normalization scheme in CuffDiff seems to be inadequate. Again, using the DMSO-treated replicates as examples, there is a considerable amount of fold change for lowly expressed genes (with FPKMs < 0). Perhaps more importantly is shown in the right panels of Figure 4.19, the fold change between the DMSO samples is not symmetrical about 0 in the

Cufflinks/CuffDiff group, but is in the PRUNE/DESeq set. This bias of representation in one sample leads to the higher fold change, and thus lower *p* values. The cause of this bias is unknown, but it is likely due to inadequate normalization.

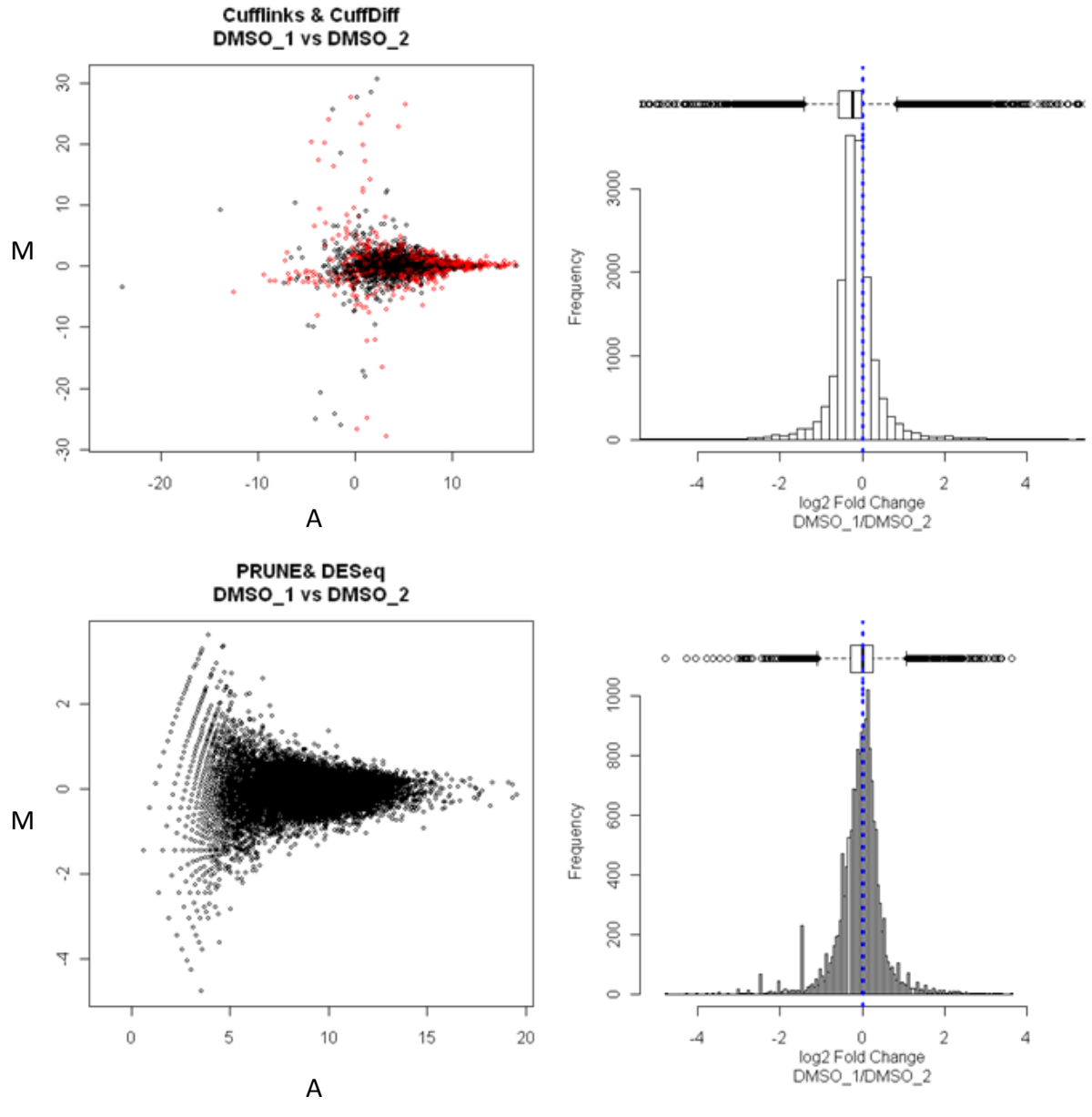


Figure 4.19. DMSO_1 versus DMSO_2 treatment using Cufflinks FPKM Data (top panels) or PRUNE read count data (bottom panels). For the left panels, the X-axis represents the average expression value between DMSO and Rif, while the Y-axis denotes the fold change in the two conditions. 'M', log2 fold change; 'A', log2 average expression. The right panels are the log2 fold change distributions from the data in the left panels. The blue dotted line marks where the

log2 fold change is 0. Boxplots are shown above each histogram. The line inside the boxplot demarcates the mean log2 fold change.

Not surprisingly, if we continue to test the differential expression for DMSO replicates versus Rifampicin replicates, we also observe the same bias (Figure 4.20). In the PRUNE dataset, *CYP3A4* is clearly the most significantly influenced gene - showing induction of about 100-fold. No other gene induced more than *CYP3A4* in the PRUNE set.

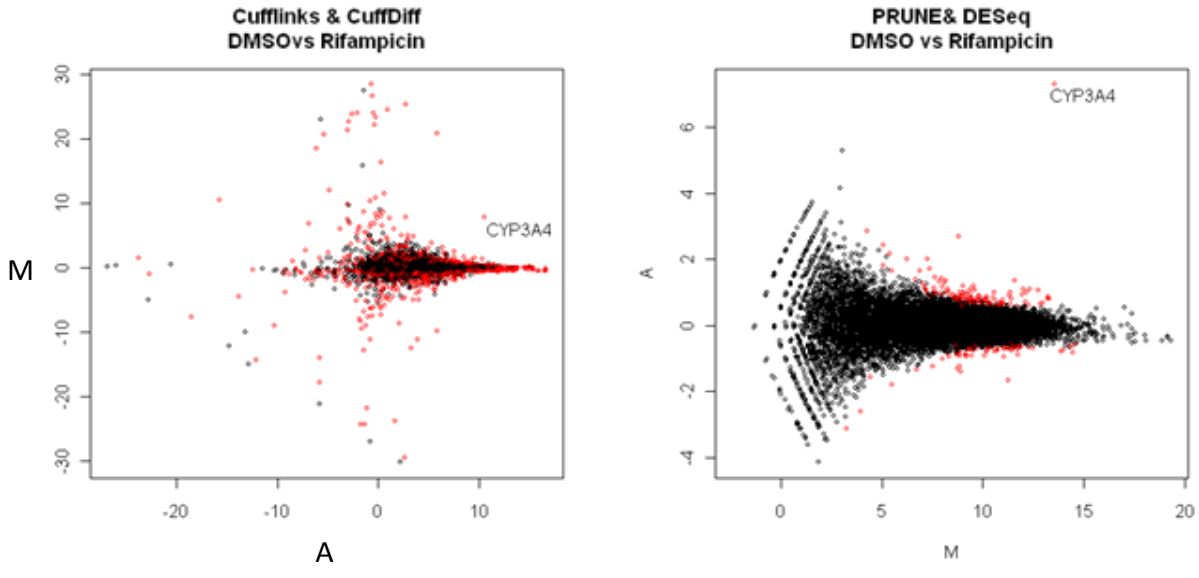


Figure 4.20. MA plots from DMSO versus rifampicin treatment using Cufflinks FPKM Data (left panel) or PRUNE Data (right panel). For both figures, the X-axis represents the average expression value between DMSO and Rif, while the Y-axis denotes the fold change in the two conditions. 'M', log2 fold change; 'A', log2 average expression. Red circles denote where $p <$ 0.05.

Now that we have an accurate estimation of differential expression, we can use the data from PRUNE analysis for further study. The number of differentially expressed genes in each treatment was 212, 237, and 65 for rifampicin, phenobarbital, and dexamethasone, respectively. The overlaps of these gene alterations can be seen in a 3-way Venn diagram (Figure 4.21). Pathway analysis revealed several pathways differentially expressed, in particular drug metabolism and cell cycle. Most of the pathways are functionally similar and are due to the induction of cytochrome P450 enzymes; the prototypical target genes of nuclear receptors. As expected, *ABCB1*, *CYP3A4*, and *UGT1A1* genes were induced by all 3 NR agonists. These represent major pathways for efflux transport, phase I metabolism, and phase II metabolism, respectively. *CYP2E1* is repressed 3-fold by Rifampicin treatment.

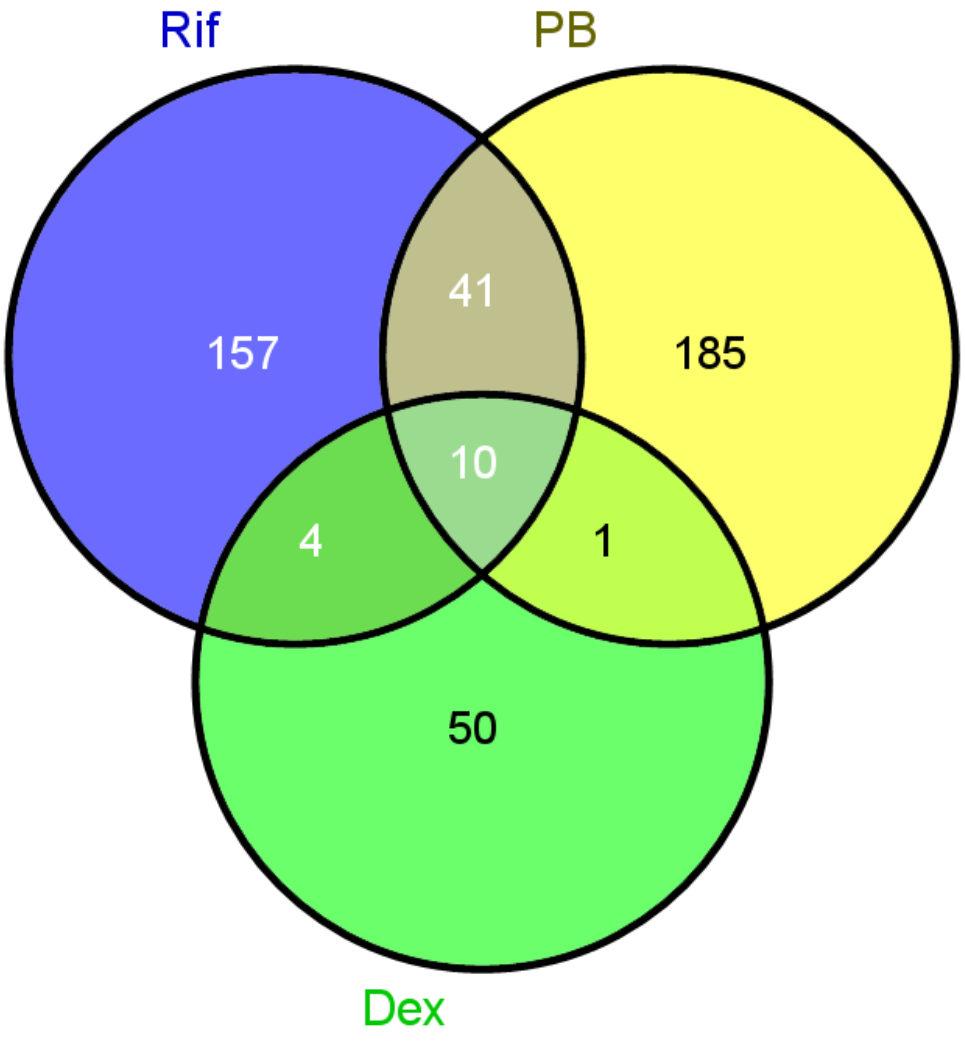


Figure 4.21. 3-Way Venn diagram showing genes with significantly differential expression between treatment groups.

If we look at genes specifically induced by a given NR agonist, we see several pathways of interest. For example, genes that were only induced by Rifampicin treatment were enriched for cell cycle genes, suggesting that Rifampicin promotes cell division (see Table 4.10).

Table 4.1. Genes involved in the induction of cell cycle progression that are induced by Rif only. Units of expression are reported in RPKM. FC: fold change

Gene	Description	DMSO	Rif	FC
ANLN	anillin, actin binding protein	15.5	29.0	1.88
ASPM	asp (abnormal spindle) homolog, microcephaly associated	3.5	7.5	2.14
BARD1	BRCA1 associated RING domain 1	1.9	4.1	2.20
BLM	Bloom syndrome, RecQ helicase-like	0.5	1.3	2.78
BRCA1	breast cancer 1, early onset	1.8	3.7	2.09
BRCA2	breast cancer 2, early onset	0.7	1.5	2.13
BUB1	budding uninhibited by benzimidazoles 1 homolog (yeast)	8.1	14.1	1.75
BUB1B	budding uninhibited by benzimidazoles 1 homolog beta	6.0	10.3	1.73
CDCA2	cell division cycle associated 2	2.2	4.4	2.01
CDK1	cell division cycle 2, G1 to S and G2 to M	5.1	10.6	2.09
CENPE	centromere protein E, 312kDa	1.8	3.9	2.18
CENPF	centromere protein F, 350/400ka (mitosin)	6.4	12.8	2.01
CIT	citron (rho-interacting, serine/threonine kinase 21)	2.5	4.5	1.78
CLSPN	claspin homolog (Xenopus laevis)	1.2	2.2	1.87
E2F7	E2F transcription factor 7	1.2	2.7	2.18
EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	14.7	23.6	1.60
FANCA	Fanconi anemia, complementation group A	0.9	1.6	1.84
FANCD2	Fanconi anemia, complementation group D2	1.8	3.6	1.96
GAS2L3	growth arrest-specific 2 like 3	1.5	3.4	2.26
KIF11	kinesin family member 11	4.5	7.8	1.74
KIF20B	kinesin family member 20B	2.6	5.2	1.99
KNTC1	kinetochore associated 1	1.7	4.2	2.43
MCM8	minichromosome maintenance complex component 8	1.7	3.5	2.06
MKI67	antigen identified by monoclonal antibody Ki-67	9.2	19.6	2.13
MLL	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila)	3.5	5.9	1.70
NCAPG	non-SMC condensin I complex, subunit G	3.2	6.1	1.92
NDC80	NDC80 homolog, kinetochore complex component (S.	4.7	9.2	1.96
RBL1	retinoblastoma-like 1 (p107)	1.1	2.4	2.08
RIF1	RAP1 interacting factor homolog (yeast)	5.3	9.2	1.73
SGOL2	shugoshin-like 2 (S. pombe)	3.8	7.0	1.82
SPAG5	sperm associated antigen 5	11.5	18.8	1.63
TPX2	TPX2, microtubule-associated, homolog (Xenopus laevis)	14.5	23.5	1.62
TTK	TTK protein kinase	3.5	6.4	1.86
TTN	Titin	0.1	0.2	1.94
ZWINT	ZW10 interactor	12.0	21.1	1.75

Most surprisingly, phenobarbital treatment repressed more genes (n = 203) than it activated (n = 34), including major transcriptional regulators C*EBP*β and *CEBPδ*. Others also showed similar numbers of differentially expressed genes. Lambert *et al.* (2009a) reported 128 genes as differentially expressed by PB in HepaRG cells using microarrays. Unlike our findings, they suggested that only 49 of the 128 genes were repressed, while the other 79 were induced (albeit in relatively small fold change). This is likely due to the increased sensitivity and specificity of RNA-Seq over arrays because we had the most concordance with the microarray when the expression values were significantly higher than background (data not shown). Interestingly, the genes that it down-regulates include cell death genes, meaning that similar to Rifampicin, PB may act as a tumor promoter - though in a transcriptionally distinct way, by repressing the repressors of cell death (Table 4.11).

Table 4.2. Genes involved in cell death are repressed by PB treatment. Units of expression are reported in RPKM. FC: fold change

Gene	Description	DMSO	Dex	FC
BOK	BCL2-related ovarian killer	23.8	10.6	0.44
FKBP8	FK506 binding protein 8, 38kDa	96.1	47.7	0.50
HSPB1	heat shock 27kDa protein-like 2 pseudogene; heat shock 27kDa protein 1	223.6	111.0	0.50
LRDD	leucine-rich repeats and death domain containing	15.9	8.0	0.50
MAP1S	microtubule-associated protein 1S	10.3	4.2	0.41
MAP3K11	mitogen-activated protein kinase kinase kinase 11	40.2	20.4	0.51
MFSD10	major facilitator superfamily domain containing 10	15.2	7.0	0.46
MRPL41	mitochondrial ribosomal protein L41	116.6	60.2	0.52
NME3	non-metastatic cells 3, protein expressed in	26.8	11.8	0.44
PHLDA3	pleckstrin homology-like domain, family A, member 3	76.5	37.0	0.48
PPP1R13L	protein phosphatase 1, regulatory (inhibitor) subunit 13 like	17.1	8.3	0.49
SCRIB	scribbled homolog (Drosophila)	19.4	9.8	0.50
SHARPIN	SHANK-associated RH domain interactor	25.1	11.6	0.46
TNFSF9	tumor necrosis factor (ligand) superfamily, member 9	13.8	5.9	0.43
TSPO	translocator protein (18kDa)	129.6	52.4	0.40

In the Dex treatment group, pathways for steroid and bile acid synthesis were markedly up-regulated. 11 out of 17 genes involved in this pathway were significantly induced only by dexamethasone treatment. The genes (and their location within the pathway) are shown in Figure 4.22.

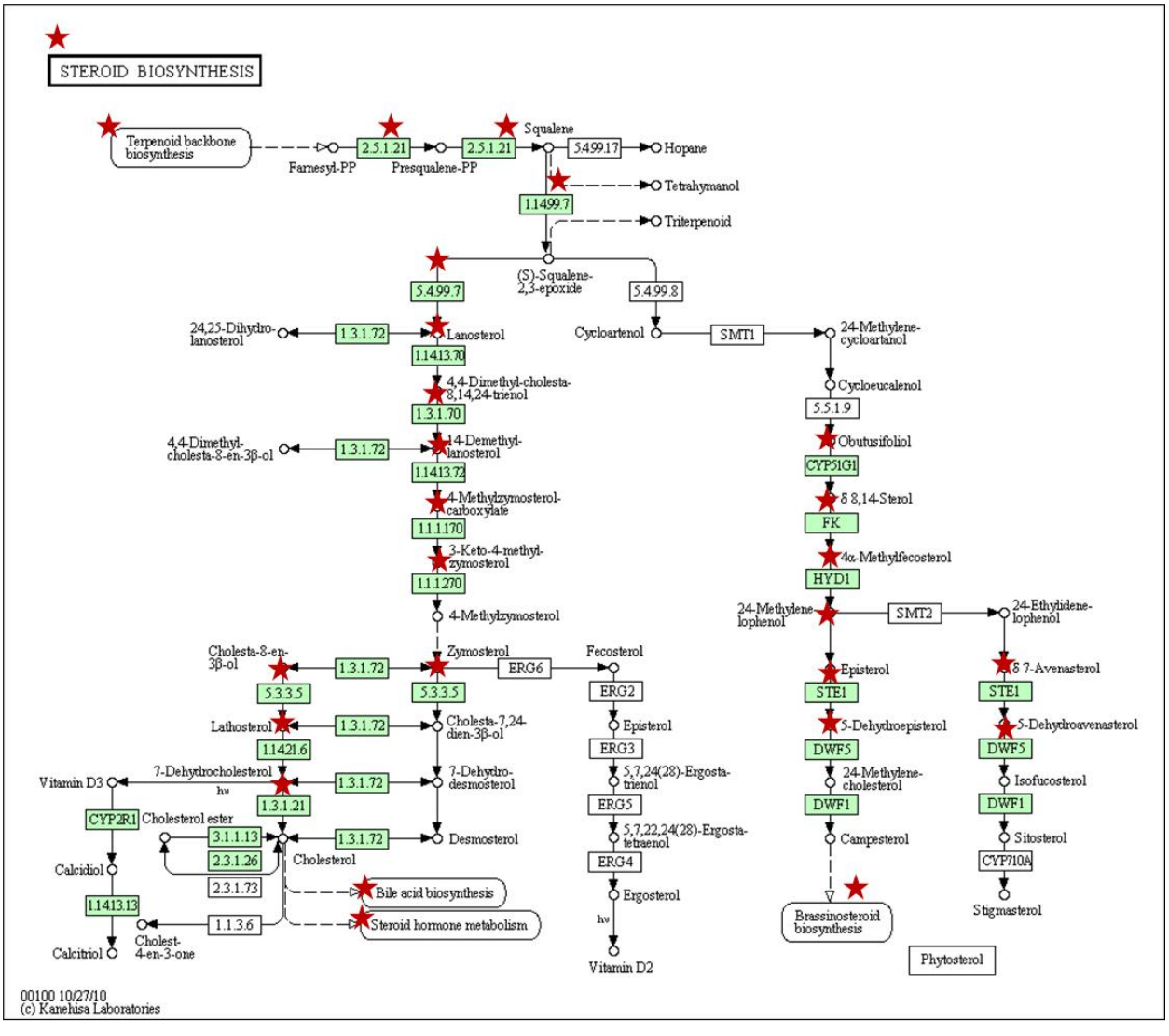


Figure 4.22. KEGG pathway for steroid biosynthesis. Red stars indicate genes that are induced by Dexamethasone.

4.2.4.4. Differential Transcript Expression

Although CuffDiff reports several types of differential expression (i.e. isoform usage between/among conditions, differential promoter use, differential coding sequences, etc), we cannot use those information because they rely on gene expression levels to quantify transcripts. We have already shown in Figure 4.18 that Cufflinks - the precursor program to CuffDiff - is less accurate for quantifying gene expression under the conditions of our experimental design. The 37% false discovery rate we identified at the gene level would only be propagated to an even higher rate if we were to look at differential expression of transcripts because 1) there are significantly more transcripts than genes, and 2) expression measurements are separated between different isoforms of the same gene, in effect decreasing total gene expression and nearing limits of detection.

PRUNE does not consider transcript-level expression, but can serve as an alternative starting point for such analysis.

4.2.5. Discussion

It is important to address the issue of genomic mapping for genes with overlapping exons, since significant alteration in gene expression may be present for only one of those genes, and the effects of this making is highlighted by the *UGT1A1* gene. *UGT1A1* contains 5 exons; four of those exons are shared with others from the *UGT1A* gene family (e.g. *UGT1A3-10*), and is known to be induced by Rif and PB in HepaRG cells (Anthérieu et al., 2010). In RNA-Seq data, this gene is only significantly induced when we consider and correct for the properties of overlapping genes, because most of the read counts come from the highly expressed *UGT1A6*

gene. Because of this, the relative fold change for *UGT1A1* compared to the total reads in the remaining four exons of the gene get diluted to a point where the fold induction is minimal and thus does not reach statistical significance. So far, PRUNE is the only program that we are aware of capable of allocating reads correctly to the *UGT1A* gene family.

4.2.5.1. How do these data compare to current methods?

We added two major concepts to consider for RNA-Seq Experiments. First, we introduced PRUNE, a set of shell scripts that can perform read-count allocation for genes with overlapping exons. This is an improvement over existing tools like HT-Seq (Anders *et al.,* unpublished) and BEDtools (Quinlan and Hall, 2010) which disregard or count reads twice, respectively. The other major concept we introduce is to test the empirical FDR for your tool of interest. We tested these tools by comparing replicate samples to one another. One should be able to assume that no (or very few) genes should be differentially expressed between replicate conditions. Here, we showed that to not be the case with Cufflinks and CuffDiff, which showed an empirical false discovery rate of 37%, but could not identify any false positives with PRUNE and DESeq.

4.2.5.2. What did we learn that we didn't already know?

Biologically speaking, our data suggest that both Rifampicin and Phenobarbital are regulators of cell fate in HepaRG cells. Rifampicin induced expression of many cell cycle regulated genes, whereas PB decreased expression of several repressors of cell death. In fact, we show that PB down-regulates more genes than it activates suggesting that it is more of a

transcriptional repressor than a transcriptional activator. Finally, we showed that many genes in the cholesterol/bile-acid synthesis pathway are induced by Dexamethasone.

CHAPTER 5. FINAL THOUGHTS

5.1. Chapters and their approaches

Each of these chapters is divided in such a way that they each contain at least one new approach to help understand mechanism involved in hepatic drug response and biotransformation pathways. However, even with the new approaches discussed throughout this text, many questions remain.

5.2. Opinions on future promise and pitfalls of these new approaches

5.2.1. POR

POR represents a novel target for pharmacogenetic research. Only in the past few years has POR been appreciated as a contributor to altered steroidogenesis pathways. However, steroidogenesis is not the only pathway affected. Several groups have confirmed that mutations in POR can affect its activity and disrupt interactions with several CYP isoforms. Importantly, investigators are discovering novel polymorphisms and are learning how they affect CYP-catalyzed drug metabolism. These studies may help to establish correlation of genetic polymorphisms in the POR gene and variation of drug metabolism through CYP-catalyzed oxidation. Additionally, the discovery of functional polymorphisms in the general population may provide a pharmacogenetic marker for drugs primarily metabolized by a one electron reduction from POR. Of course, before POR becomes a pharmacogenetic predictor, extensive studies are required for clinical investigations into whether or not POR polymorphisms associate with decreased metabolic activity *in vivo* and if screening for these polymorphisms would be

efficacious in a clinical setting.

Several challenges to understanding POR function and contribution to drug metabolism exist in current research. First, POR activity is typically quantified by the reduction of cytochrome *c*, a non-physiological substrate. Its affect on electron donation to CYP enzymes need to be simultaneously investigated with CYP catalysis assays. Second, effects of POR polymorphisms on CYP activities may vary markedly between different CYP enzymes, so in determining the consequences of POR mutations, several CYP isoforms should be tested to provide more robust confirmation of altered protein activity. Third, because the amino-terminal tail of POR is necessary for physiological activity, experiments to characterize functional polymorphisms should use a full length construct embedded into membranes instead of a partial construct that have been used in many studies. Finally, very little is known about how POR is regulated in humans. Understanding the regulation of this gene will be important knowledge if POR becomes a viable pharmacogenetic marker, because increased gene expression can have the ability to abrogate functionally deficient polymorphisms in heterozygous samples by simultaneously increasing the expression of the wild-type allele. Further research is needed to overcome these obstacles and to address the clinical significance of POR polymorphisms.

Soon after our 2008 publication in *Pharmacogenetics and Genomics*, Miller's group published a similar study that sequenced POR genes of 842 healthy individuals belonging to four different ethnic groups (Huang et al., 2008). Some of the genotypes we observed in our study were also found by them. Later, that same group showed that the A503V supports normal activity by CYP1A2 and CYP2C19 (Agrawal et al., 2008). Our study directly led Gomes *et al.* (2009) to do a similar study with 150 human liver microsomes - identifying other POR

polymorphisms that link to CYP activity. Similarly to our findings, they also show that POR can be more limiting for some CYPs than others.

These and other studies prompted the Human CYP Allele Nomenclature Chair and Committee to devise a system for the designation of POR alleles that follows the guidelines for CYP allelic star (CYP*) nomenclature (http://www.cypalleles.ki.se/criteria.htm)(Sim et al., 2009). The POR allele nomenclature web page (http://www.cypalleles.ki.se/por.htm) was launched in September 2008, listing 35 different alleles. On this POR web page, the alleles are presented together with their corresponding nucleotide and amino acid changes, and the phenotypic consequences observed by *in vitro* and *in vivo* studies. Three polymorphisms we identified are now named by that committee: *POR*25, *26, & *27* (K49N, L420M, L577P, respectively).

Studies are still ongoing to identify relationships between POR polymorphisms and CYP activity. Recently, Miller's group showed the A503V POR polymorphism can decrease the activity of CYP3A4 by 67% (Agrawal et al., 2010), and will continue to explore the pharmacogenomics of POR at least until 2012 when that grant will expire (5R01GM073020-07).

5.2.2. HepaRG Cells

A major obstacle in drug discovery is the lack of an adequate model to predict xenobiotic metabolism, drug-drug interactions, and hepatotoxicity of drug candidates. Poor correlation exists between animals and humans regarding drug-drug interactions and drug-induced toxicity, mainly caused by the significant species difference in metabolic transformation and toxic effects of drug candidates. Primary human hepatocytes are approved by FDA and are the "gold

standard" model for CYP-mediated drug-drug interactions, whereas cyropreserved primary human hepatocytes are widely accepted to be advantageous for short-term biotransformation studies (Silva et al., 1999). However, limited availability and short lifespan are fatal flaws in these two models, and consequently make them not feasible to be used in high-throughput systems for drug candidate screening. Our work has shown HepaRG cells are more similar to *in vivo* hepatocytes that the commonly used cell line HepG2 at the level of gene expression. Therefore, it is logical to hypothesize that they can also predict hepatotoxicity and metabolism.

The introduction of HepaRG cell line into drug metabolism, drug-drug interaction, and hepatotoxicity fields constitutes a breakthrough in drug discovery field. Because HepaRG cells mimic biological performance in gene expression and functional activities of liver specific proteins, they exhibit the consistent intrinsic clearance capacity and responsiveness to well-defined inducers to a comparable level to that seen in primary human hepatocytes. The stability of HepaRG cells in long-term cultivation provides the base for sub-chronic and chronic exposure to chemicals for drug safety evaluation. Aninat *et al.* (2006) were the first to claim HepaRG could act as surrogates for primary human hepatocytes in the context of drug metabolism. This claim was a bit premature because it was based only on mRNA content and protein activity from just a handful of genes. We expanded on this claim by demonstrating the transcriptome wide similarity between HepaRG and primary hepatocytes.

As a stable hepatic cell line, HepaRG cells are applicable for widespread usage as an *in vitro* cell model for high-throughput screening, however many improvements can be made. First, they do not express *CYP2D6*. Current work in our lab is aiming to introduce a stably expressed *CYP2D6* cDNA to overcome this problem. If one wants to use these cells to identify

genes that metabolize a specific drug, several deletions of each P450 can be made so that a panel of HepaRG cells could be available, each one lacking a member of the Cytochrome P450 family. In the same context, genetic variants of many different cytochromes P450 could be introduced, thereby creating a cell-based "population" that could be used for pharmacogenetic research.

5.2.3. mRNA-Seq

Scientific discovery is an evolving process. New ideas, techniques, and perspectives continually change the dynamic of our understanding of how biological systems really work, especially with regard to new algorithms for RNA-Seq experiments. The instrumentation for such experiments is only four years old. The first RNA-Seq experiment was performed only *two* years ago, and many of the algorithms for making sense of the data are still undergoing development. Already, new methods in sample preparation are incorporating strand-specific information - which is why PRUNE was initially developed. PRUNE still has an advantage in this respect though. There are cases such as the *UGT1A* family where strand information will not solve the problem, because the gene family members are all on the same strand. Until longer read lengths are possible (especially when they are long enough to span the entire cDNA), PRUNE remains a viable option.

There are few analysis options available for quantitating gene expression and no single package is capable of extracting all the information contained in a single mRNA-Seq experiment. Although we made PRUNE to remedy the issue of assigning sequencing reads to overlapping genomic coordinates, many difficulties still exist, including accurate SNP-calling and transcript-level quantitation.

No clear consensus has been made on how to best define transcript isoforms from short-read data, a prerequisite for quantitation. New methods, such as *de novo* assembly should overcome these issues may overcome this issue in the near future, however this strategy also has its limitations. Recently, Trans-Abyss, a *de novo* assembly strategy, was compared to reference-based assemblers such as Cufflinks, but measures of sensitivity and specificity were unimpressively different by the two methods (Robertson et al., 2010). Moreover, the computational cost of the *de novo* assembler was high 370 CPU hours for Trans-Abyss compared to 12 CPU hours for cufflinks. Regardless, these issues are well known and new methods are in development to overcome their algorithmic limitations.

Genotyping from mRNA-Seq is the second difficult task. Our work and that of others show evidence of a high false discovery rate in the sequencing data (Morin et al., 2008; Chepelev et al., 2009; Cirulli et al., 2010). One way to overcome the high number of false positives is to train a logistic regression model using a set of known SNPs - as was the case in Chapter 4. However, cost to sensitivity was great even though our specificity was high. This means that when a SNP is called using our logistic regression model, then it is highly likely to be a real SNP, but we only catch < 30% of all the true SNPs in the dataset. Therefore, more algorithms are needed to have both high sensitivity and specificity.

Currently, we are working on a project to leverage expertise from the whole genome resequencing community. That community is more developed than the RNA-Seq community and have experienced and overcome first-hand some of the problems with short-read sequencing. For example, one way to decrease the false discovery rate in mRNA Seq data would be to recalibrate quality values. Recalibration is simply modifying the quality values output from the

sequencing instrument to values based on empirical data. A pre-calibrated file could contain many *reported* Q25 bases, (or $P_e = 10^{\frac{20}{-10}} = 0.003$ = 3 mistakes in 1000 total reads) which seems good. However, it may be that these bases *actually* mismatch the reference at a 1 in 100 rate empirically, so they are actually Q20. These higher-than-empirical quality scores provide false confidence in the base calls resulting in higher false discovery rates. Moreover, base mismatches with the reference occur at the end of the reads more frequently than at the beginning and differences in dinucleotide calling quality. These re-calibration adjustments are an absolute requirement for large-scale DNA sequencing studies such as the 1000 Genomes Project (Durbin et al., 2010), Pediatric Cancer Genome Project (http://www.pediatriccancergenomeproject.org/site/), and the Cancer Genome Atlas (http://cancergenome.nih.gov). However, using the tools from whole genome or whole exome sequencing projects, such as those listed, require different assumptions about the data than those from mRNA-Seq, many of which have yet to be worked out. Therefore, tools will need to continue to be developed to convert the knowledge gained from these sequencing projects into the mRNA-Seq realm. Another way to improve SNP calling is to adjust instrument-specific error profiles. Sequencing errors for the Illumina are dependent on the local sequence context of the base being read, the position of the base in the read, etc. which can all result in increasing the number of false positive SNP calls (Dohm et al., 2008; Erlich et al., 2008).

5.3. Final thoughts

From new genetic targets to advances in resources and technology, the field of pharmacogenomics is moving rapidly. The application of whole genome sequencing is likely to

become part of routine clinical practice. When this happens, researchers will not need to use candidate gene approaches as we have in chapter 2. The combination of electronic medical records with every patient's genome to identify genetic variants without *a priori* bias, will allow rapid significant medical advancements.

Once genotypes or haplotypes are inferred as having clinically significant impact, cell lines such as HepaRG could be genetically modified to recapitulate the haplotype to mechanistically explain genotype-phenotype relationships. As mentioned in chapter 3, our group is already modifying the genetic structure of HepaRG by introducing *CYP2D6*1*. Perhaps it will not be long before hundreds of mutations are introduced in HepaRG so that it would be possible to do a cell-based population study. Whole genome sequencing will give us a detailed schematic of nuclear DNA organization, but it cannot tell us the outcomes of that organization, e.g. is the gene is expressed or if RNA editing is involved. Therefore, there will be a continued need for mRNA-Seq to complement whole genome DNA sequencing. The computational algorithms for interpreting mRNA-Seq, as discussed in chapter 4, are continually improving and will continue to improve in the near future. Given these recent advances, this is an exciting time to be a pharmacogenomicist.

REFERENCES

Adachi M, Tachibana K, Asakura Y, Suwa S and Nishimura G (1999) A male patient presenting with major clinical symptoms of glucocorticoid deficiency and skeletal dysplasia, showing a steroid pattern compatible with 17alpha-hydroxylase/17,20-lyase deficiency, but without obvious CYP17 gene mutations. *Endocr J* **46**:285-292.

Agrawal V, Choi JH, Giacomini KM and Miller WL (2010) Substrate-specific modulation of CYP3A4 activity by genetic variants of cytochrome P450 oxidoreductase. *Pharmacogenet Genomics* **20**:611-618.

Agrawal V, Huang N and Miller WL (2008) Pharmacogenetics of P450 oxidoreductase: effect of sequence variants on activities of CYP1A2 and CYP2C19. *Pharmacogenet Genomics* **18**:569-576.

Anders S and Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* **11**:R106.

Aninat C, Piton A, Glaise D, Le Charpentier T, Langouet S, Morel F, Guguen-Guillouzo C and Guillouzo A (2006) Expression of cytochromes P450, conjugating enzymes and nuclear receptors in human hepatoma HepaRG cells. *Drug Metab Dispos* **34**:75-83.

Anthérieu S, Chesné C, Li R, Camus S, Lahoz A, Picazo L, Turpeinen M, Tolonen A, Uusitalo J, Guguen-Guillouzo C and Guillouzo A (2010) Stable Expression, Activity, and Inducibility of Cytochromes P450 in Differentiated HepaRG Cells. *Drug MetabDispos* **38**:516-525.

Arlt W (2007) P450 oxidoreductase deficiency and Antley-Bixler syndrome. *Rev Endocr Metab Disord* **8**:301-307.

Arlt W, Walker EA, Draper N, Ivison HE, Ride JP, Hammer F, Chalder SM, Borucka-Mankiewicz M, Hauffa BP, Malunowicz EM, Stewart PM and Shackleton CH (2004) Congenital adrenal hyperplasia caused by mutant P450 oxidoreductase and human androgen synthesis: analytical study. *Lancet* **363**:2128-2135.

Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, Pushkarev D, Neff NF, Hudgins L, Gong L, Hodges LM, Berlin DS, Thorn CF, Sangkuhl K, Hebert JM, Woon M, Sagreiya H, Whaley R, Knowles JW, Chou MF, Thakuria JV, Rosenbaum AM, Zaranek AW, Church GM, Greely HT, Quake SR and Altman RB (2010) Clinical assessment incorporating a personal genome. *Lancet* **375**:1525-1535.

Bajorath J (2002) Integration of virtual and high-throughput screening. *Nat Rev Drug Discov* **1**:882-894.

Benjamini Y and Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Royal Stat Society Series B* **57**:289-300.

Black SD, French JS, Williams CH, Jr. and Coon MJ (1979) Role of a hydrophobic polypeptide in the N-terminal region of NADPH-cytochrome P-450 reductase in complex formation with P-450LM. *Biochem Biophys Res Commun* **91**:1528-1535.

Bonina TA, Gilep AA, Estabrook RW and Usanov SA (2005) Engineering of proteolytically stable NADPH-cytochrome P450 reductase. *Biochemistry (Mosc)* **70**:357-365.

Bryant DW, Jr., Shen R, Priest HD, Wong WK and Mockler TC (2010) Supersplat--spliced RNA-seq alignment. *Bioinformatics* **26**:1500-1505.

Bullard JH, Purdom E, Hansen KD and Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**:94.

Burchell B, Soars M, Monaghan G, Cassidy A, Smith D and Ethell B (2000) Drug-mediated toxicity caused by genetic deficiency of UDP-glucuronosyltransferases. *Toxicol Lett* **112-113**:333-340.

Chepelev I, Wei G, Tang Q and Zhao K (2009) Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res* **37**:e106.

Cirulli ET, Singh A, Shianna KV, Ge D, Smith JP, Maia JM, Heinzen EL, Goedert JJ and Goldstein DB (2010) Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biol* **11**:R57.

Cloonan N, Forrest AR, Kolle G, Gardiner BA, Faulkner GF and *al. e* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5**:613-619.

Czerwinski M, Opdam P, Madan A, Carroll K, Mudra DR, Gan LL, Luo G and Parkinson A (2002) Analysis of CYP mRNA expression by branched DNA technology. *Methods Enzymol* **357**:170-179.

Daly AK (2004) Pharmacogenetics of the cytochromes P450. *Curr Top Med Chem* **4**:1733-1744.

Dayer P, Kronbach T, Eichelbaum M and Meyer UA (1987) Enzymatic basis of the debrisoquine/sparteine-type genetic polymorphism of drug oxidation. Characterization of bufuralol 1'-hydroxylation in liver microsomes of in vivo phenotyped carriers of the genetic deficiency. *Biochem Pharmacol* **36**:4145-4152.

de Jong FA, de Jonge MJ, Verweij J and Mathijssen RH (2006) Role of pharmacogenetics in irinotecan therapy. *Cancer Lett* **234**:90-106.

Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**:P3.

Dohm JC, Lottaz C, Borodina T and Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**:e105.

Dormann H, Neubert A, Criegee-Rieck M, Egger T, Radespiel-Troger M, Azaz-Livshits T, Levy M, Brune K and Hahn EG (2004) Readmissions and adverse drug reactions in internal medicine: the economic impact. *J Intern Med* **255**:653-663.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME and McVean GA (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**:1061-1073.

Eichelbaum M, Ingelman-Sundberg M and Evans WE (2006) Pharmacogenomics and individualized drug therapy. *Annu Rev Med* **57**:119-137.

Enoch HG and Strittmatter P (1979) Cytochrome b5 reduction by NADPH-cytochrome P-450 reductase. *J Biol Chem* **254**:8976-8981.

Erlich Y, Mitra PP, delaBastide M, McCombie WR and Hannon GJ (2008) Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods* **5**:679-682.

Evans WE and Relling MV (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**:487-491.

Ewing B, Hillier L, Wendl MC and Green P (1998) Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Res* **8**:175-185.

Fedurco M, Romieu A, Williams S, Lawrence I and Turcatti G BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research* **34**:e22.

Finn RD, McLaughlin LA, Ronseaux S, Rosewell I, Houston JB, Henderson CJ and Wolf CR (2008) Defining the in Vivo Role for cytochrome b5 in cytochrome P450 function through the conditional hepatic deletion of microsomal cytochrome b5. *J Biol Chem* **283**:31385-31393.

Flück CE and Miller WL (2006) P450 oxidoreductase deficiency: a new form of congenital adrenal hyperplasia. *Curr Opin Pediatr* **18**:435-441.

Flück CE, Tajima T, Pandey AV, Arlt W, Okuhara K, Verge CF, Jabs EW, Mendonca BB, Fujieda K and Miller WL (2004) Mutant P450 oxidoreductase causes disordered steroidogenesis with and without Antley-Bixler syndrome. *Nat Genet* **36**:228-230.

Fukami M, Horikawa R, Nagai T, Tanaka T, Naiki Y, Sato N, Okuyama T, Nakai H, Soneda S, Tachibana K, Matsuo N, Sato S, Homma K, Nishimura G, Hasegawa T and Ogata T (2005) Cytochrome P450 oxidoreductase gene mutations and Antley-Bixler syndrome with abnormal genitalia and/or impaired steroidogenesis: molecular and clinical studies in 10 patients. *J Clin Endocrinol Metab* **90**:414-426.

Gaedigk A, Bhathena A, Ndjountche L, Pearce RE, Abdel-Rahman SM, Alander SW, Bradford LD, Rogan PK and Leeder JS (2005) Identification and characterization of novel sequence variations in the cytochrome P4502D6 (CYP2D6) gene in African Americans. *Pharmacogenomics J* **5**:173-182.

Gaedigk A, Gotschall RR, Forbes NS, Simon SD, Kearns GL and Leeder JS (1999) Optimization of cytochrome P4502D6 (CYP2D6) phenotype assignment using a genotyping algorithm based on allele frequency data. *Pharmacogenetics* **9**:669-682.

Gandhi TK, Weingart SN, Borus J, Seger AC, Peterson J, Burdick E, Seger DL, Shu K, Federico F, Leape LL and Bates DW (2003) Adverse drug events in ambulatory care. *N Engl J Med* **348**:1556-1564.

Gigon PL, Gram TE and Gillette JR (1969) Studies on the rate of reduction of hepatic microsomal cytochrome P-450 by reduced nicotinamide adenine dinucleotide phosphate: effect of drug substrates. *Mol Pharmacol* **5**:109-122.

Gomes AM, Winter S, Klein K, Turpeinen M, Schaeffeler E, Schwab M and Zanger UM (2009) Pharmacogenomics of human liver cytochrome P450 oxidoreductase: multifactorial analysis and impact on microsomal drug oxidation. *Pharmacogenomics* **10**:579-599.

Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou YC, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJ, Tai IT and Marra MA (2010) Alternative expression analysis by RNA sequencing. *Nat Methods* **7**:843-847.

Gripon P, Rumin S, Urban S, Le Seyec J, Glaise D, Cannie I, Guyomard C, Lucas J, Trepo C and Guguen-Guillouzo C (2002) Infection of a human hepatoma cell line by hepatitis B virus. *Proc Natl Acad Sci U S A* **99**:15655-15660.

Grunau A, Paine MJ, Ladbury JE and Gutierrez A (2006) Global effects of the energetics of coenzyme binding: NADPH controls the protein interaction properties of human cytochrome P450 reductase. *Biochemistry* **45**:1421-1434.

Gu J, Weng Y, Zhang QY, Cui H, Behr M, Wu L, Yang W, Zhang L and Ding X (2003) Liver-specific deletion of the NADPH-cytochrome P450 reductase gene: impact on plasma cholesterol homeostasis and the function and regulation of microsomal cytochrome P450 and heme oxygenase. *J Biol Chem* **278**:25895-25901.

Guengerich FP (2001) Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem Res Toxicol* **14**:611-650.

Guengerich FP (2004) Cytochrome P450: what have we learned and what are the future issues? *Drug Metab Rev* **36**:159-197.

Guengerich FP (2006) Cytochrome P450s and other enzymes in drug metabolism and toxicity. *Aaps J* **8**:E101-111.

Guengerich FP and Johnson WW (1997) Kinetics of ferric cytochrome P450 reduction by NADPH-cytochrome P450 reductase: rapid reduction in the absence of substrate and variations among cytochrome P450 systems. *Biochemistry* **36**:14741-14750.

Guillemette C (2003) Pharmacogenomics of human UDP-glucuronosyltransferase enzymes. *Pharmacogenomics J* **3**:136-158.

Guillouzo A (1998) Liver cell models in in vitro toxicology. *Environ Health Perspect* **106 Suppl 2**:511-532.

Guillouzo A, Corlu A, Aninat C, Glaise D, Morel F and Guguen-Guillouzo C (2007) The human hepatoma HepaRG cells: a highly differentiated model for studies of liver metabolism and toxicity of xenobiotics. *Chem Biol Interact* **168**:66-73.

Guillouzo A and Guguen-Guillouzo C (2008) Evolving concepts in liver tissue modeling and implications for in vitro toxicology. *Expert Opin Drug Metab Toxicol* **4**:1279-1294.

Gutierrez A, Grunau A, Paine M, Munro AW, Wolf CR, Roberts GC and Scrutton NS (2003) Electron transfer in human cytochrome P450 reductase. *Biochem Soc Trans* **31**:497-501.

Gutierrez A, Lian LY, Wolf CR, Scrutton NS and Roberts GC (2001) Stopped-flow kinetic studies of flavin reduction in human cytochrome P450 reductase and its component domains. *Biochemistry* **40**:1964-1975.

Gutierrez A, Paine M, Wolf CR, Scrutton NS and Roberts GC (2002) Relaxation kinetics of cytochrome P450 reductase: internal electron transfer is limited by conformational change and regulated by coenzyme binding. *Biochemistry* **41**:4626-4637.

Haiman CA, Setiawan VW, Xia LY, Le Marchand L, Ingles SA, Ursin G, Press MF, Bernstein L, John EM and Henderson BE (2007) A variant in the cytochrome p450 oxidoreductase gene is associated with breast cancer risk in African Americans. *Cancer Res* **67**:3565-3568.

Han JF, Wang SL, He XY, Liu CY and Hong JY (2006) Effect of genetic variation on human cytochrome p450 reductase-mediated paraquat cytotoxicity. *Toxicol Sci* **91**:42-48.

Hart SN, Li Y, Nakamoto K, Subileau EA, Steen D and Zhong XB (2010) A comparison of whole genome gene expression profiles of HepaRG cells and HepG2 cells to primary human hepatocytes and human liver tissues. *Drug Metab Dispos* **38**:988-994.

Hart SN, Li Y, Nakamoto K, Wesselman C and Zhong XB (2007) Novel SNPs in cytochrome P450 oxidoreductase. *Drug Metab Pharmacokinet* **22**:322-326.

Hartley DP and Klaassen CD (2000) Detection of chemical-induced differential expression of rat hepatic cytochrome P450 mRNA transcripts using branched DNA signal amplification technology. *Drug Metab Dispos* **28**:608-616.

He P, Court MH, Greenblatt DJ and von Moltke LL (2006) Human Pregnane X Receptor: Genetic Polymorphisms, Alternative mRNA Splice Variants, and Cytochrome P450 3A Metabolic Activity. *J Clin Pharmacol* **46**:1356-1369.

Heap GA, Yang JH, Downes K, Healy BC, Hunt KA, Bockett N, Franke L, Dubois PC, Mein CA, Dobson RJ, Albert TJ, Rodesch MJ, Clayton DG, Todd JA, van Heel DA and Plagnol V (2010) Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet* **19**:122-134.

Helfti FF (2008) Requirements for a lead compound to become a clinical candidate. *BMC Neuroscience* **9**.

Henderson CJ, Otto DM, Carrie D, Magnuson MA, McLaren AW, Rosewell I and Wolf CR (2003) Inactivation of the hepatic cytochrome P450 system by conditional deletion of hepatic cytochrome P450 reductase. *J Biol Chem* **278**:13480-13486.

Hewitt NJ, Lechon M, Houston JB, Hallifax D, Brown HS, Maurel P, Kenna JG, Gustavsson L, Lohmann C, Skonberg C, Guillouzo A, Tuschl G, Li AP, LeCluyse E, Groothuis GMM and Hengstler JG

(2007) Primary Hepatocytes: Current Understanding of the Regulation of Metabolic Enzymes and Transporter Proteins, and Pharmaceutical Practice for the Use of Hepatocytes in Metabolism, Enzyme Induction, Transporter, Clearance, and Hepatotoxicity Studies. *Drug Metab Rev* **39**:159-234.

Hildebrandt A and Estabrook RW (1971) Evidence for the participation of cytochrome b 5 in hepatic microsomal mixed-function oxidation reactions. *Arch Biochem Biophys* **143**:66-79.

Honkakoski P and Negishi M (2000) Regulation of cytochrome P450 (CYP) genes by nuclear receptors. *Biochem J* **347**:321-337.

Horecker B (1950a) Triphosphopyridine nucleotide-cytochrome c reductase in liver. *J Biol Chem* **183**:593-605

http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?locusId=5447&chooseRs=all

Huang N, Agrawal V, Giacomini KM and Miller WL (2008) Genetics of P450 oxidoreductase: sequence variation in 842 individuals of four ethnicities and activities of 15 missense mutations. *Proc Natl Acad Sci U S A* **105**:1733-1738.

Huang N, Pandey AV, Agrawal V, Reardon W, Lapunzina PD, Mowat D, Jabs EW, Van Vliet G, Sack J, Flück CE and Miller WL (2005) Diversity and function of mutations in p450 oxidoreductase in patients with Antley-Bixler syndrome and disordered steroidogenesis. *Am J Hum Genet* **76**:729-749.

Huang YS, Chern HD, Su WJ, Wu JC, Lai SL, Yang SY, Chang FY and Lee SD (2002) Polymorphism of the N-acetyltransferase 2 gene as a susceptibility risk factor for antituberculosis drug-induced hepatitis. *Hepatology* **35**:883-889.

Hubbard PA, Shen AL, Paschke R, Kasper CB and Kim JJ (2001) NADPH-cytochrome P450 oxidoreductase. Structural basis for hydride and electron transfer. *J Biol Chem* **276**:29163-29170.

Ikeda S, Kurose K, Jinno H, Sai K, Ozawa S, Hasegawa R, Komamura K, Kotake T, Morishita H, Kamakura S, Kitakaze M, Tomoike H, Tamura T, Yamamoto N, Kunitoh H, Yamada Y, Ohe Y, Shimada Y, Shirao K, Kubota K, Minami H, Ohtsu A, Yoshida T, Saijo N, Saito Y and Sawada J (2005) Functional analysis of four naturally occurring variants of human constitutive androstane receptor. *Mol Genet Metab* **86**:314-319.

Ingelman-Sundberg M (2005) Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *Pharmacogenomics J* **5**:6-13.

Ingelman-Sundberg M, Sim SC, Gomez A and Rodriguez-Antona C (2007) Influence of cytochrome P450 polymorphisms on drug therapies: Pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacol Ther* **116**:496-526.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B and Speed TP (2003) Summaries of Affymetrix GeneChip probe level data. *Nucl. Acids Res.* **31**:e15-.

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**:195-202.

Jones DT, Taylor WR and Thornton JM (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33**:3038-3049.

Kafka A, Sauer G, Jaeger C, Grundmann R, Kreienberg R, Zeillinger R and Deissler H (2003) Polymorphism C3435T of the MDR-1 gene predicts response to preoperative chemotherapy in locally advanced breast cancer. *Int J Oncol* **22**:1117-1121.

Kanebratt KP and Andersson TB (2008a) Evaluation of HepaRG cells as an in vitro model for human drug metabolism studies. *Drug Metab Dispos* **36**:1444-1452.

Kanebratt KP and Andersson TB (2008b) HepaRG cells as an in vitro model for evaluation of cytochrome P450 induction in humans. *Drug Metab Dispos* **36**:137-145.

Kato M, Chiba K, Horikawa M and Sugiyama Y (2005) The quantitative prediction of in vivo enzyme-induction caused by drug exposure from in vitro information on human hepatocytes. *Drug Metab Pharmacokinet* **20**:236-243.

Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, Lin Z, Liu Y, Liu S, Oliver DE, Rubin DL, Shafa F, Stuart JM and Altman RB (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J* **1**:167-170.

Ko JW, Desta Z and Flockhart DA (1998) Human N-demethylation of (S)-mephenytoin by cytochrome P450s 2C9 and 2B6. *Drug Metab Dispos* **26**:775-778.

Koyano S, Kurose K, Saito Y, Ozawa S, Hasegawa R, Komamura K, Ueno K, Kamakura S, Kitakaze M, Nakajima T, Matsumoto K, Akasawa A, Saito H and Sawada J (2004) Functional characterization of four naturally occurring variants of human pregnane X receptor (PXR): one variant causes dramatic loss of both DNA binding activity and the transactivation of the CYP3A4 promoter/enhancer region. *Drug Metab Dispos* **32**:149-154.

Krone N, Dhir V, Ivison HE and Arlt W (2007) Congenital adrenal hyperplasia and P450 oxidoreductase deficiency. *Clin Endocrinol (Oxf)* **66**:162-172.

Kumar P, Henikoff S and Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**:1073-1081.

Lamba J, Lamba V and Schuetz E (2005) Genetic variants of PXR (NR1I2) and CAR (NR1I3) and their implications in drug metabolism and pharmacogenetics. *Curr Drug Metab* **6**:369-383.

Lamba J, Strom S, Venkataramanan R, Thummel KE, Lin YS, Liu W, Cheng C, Lamba V, Watkins PB and Schuetz E (2006) MDR1 genotype is associated with hepatic cytochrome P450 3A4 basal and induction phenotype. *Clin Pharmacol Ther* **79**:325-338.

Lambert C, Leonard N, De Bolle X and Depiereux E (2002) ESyPred3D: Prediction of proteins 3D structures. *Bioinformatics* **18**:1250-1256.

Lambert CB, Spire C, Claude N and Guillouzo A (2009a) Dose- and time-dependent effects of phenobarbital on gene expression profiling in human hepatoma HepaRG cells. *Toxicol App Pharmacol* **234**:345-360.

Lambert CB, Spire C, Renaud M-P, Claude N and Guillouzo A (2009b) Reproducible chemical-induced changes in gene expression profiles in human hepatoma HepaRG cells under various experimental conditions. *Toxicology in Vitro* **23**:466-475.

Lambert CB, Spire C, Renaud M-P, Claude N and Guillouzo A (2009c) Reproducible chemical-induced changes in gene expression profiles in human hepatoma HepaRG cells under various experimental conditions. *Toxicol in Vitro* **23**:466-475.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**:860-921.

Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SM and Bor DH (2002) Timing of new black box warnings and withdrawals for prescription medications. *JAMA* **287**:2215-2220.

Lazarou J, Pomeranz BH and Corey PN (1998) Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* **279**:1200-1205.

Le Vee M, Jigorel E, Glaise D, Gripon P, Guguen-Guillouzo C and Fardel O (2006) Functional expression of sinusoidal and canalicular hepatic drug transporters in the differentiated human hepatoma HepaRG cell line. *Eur J Pharm Sci* **28**:109-117.

LeCluyse E, Ajay M, Geraldine H, Kathy C, Ryan D and Andrew P (2000) Expression and regulation of cytochrome P450 enzymes in primary cultures of human hepatocytes. *J BiochemMol Toxicol* **14**:177-188.

LeCluyse EL (2001) Human hepatocyte culture systems for the in vitro evaluation of cytochrome P450 expression and regulation. *Eur J Pharm Sci* **13**:343-368.

Letschert K, Keppler D and Konig J (2004) Mutations in the SLCO1B3 gene affecting the substrate specificity of the hepatocellular uptake transporter OATP1B3 (OATP8). *Pharmacogenetics* **14**:441-452.

Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078-2079.

Li H, Ruan J and Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**:1851-1858.

Lin D, Black SM, Nagahama Y and Miller WL (1993) Steroid 17 alpha-hydroxylase and 17,20-lyase activities of P450c17: contributions of serine106 and P450 reductase. *Endocrinology* **132**:2498-2506.

Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH and Lee MP (2003) Allelic variation in gene expression is common in the human genome. *Genome Res* **13**:1855-1862.

Lo HW and Ali-Osman F (2007) Genetic polymorphism and function of glutathione S-transferases in tumor drug resistance. *Curr Opin Pharmacol* **7**:367-374.

Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C and Eisner R (2004) Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* **20**:547-556.

Lundqvist E, Johansson I and Ingelman-Sundberg M (1999) Genetic mechanisms for duplication and multiduplication of the human CYP2D6 gene and methods for detection of duplicated CYP2D6 genes. *Gene* **226**:327-338.

Luo G, Guenthner T, Gan LS and Humphreys WG (2004) CYP3A4 induction by xenobiotics: biochemistry, experimental methods and impact on drug discovery and development. *Curr Drug Metab* **5**:483-505.

Madan A, Graham RA, Carroll KM, Mudra DR, Burton LA, Krueger LA, Downey AD, Czerwinski M, Forster J, Ribadeneira MD, Gan LS, LeCluyse EL, Zech K, Robertson P, Jr., Koch P, Antonian L, Wagner G, Yu L and Parkinson A (2003) Effects of prototypical microsomal enzyme inducers on cytochrome P450 expression in cultured human hepatocytes. *Drug Metab Dispos* **31**:421-431.

Mai I, Perloff ES, Bauer S, Goldammer M, Johne A, Filler G, Budde K and Roots I (2004) MDR1 haplotypes derived from exons 21 and 26 do not affect the steady-state pharmacokinetics of tacrolimus in renal transplant patients. *Br J Clin Pharmacol* **58**:548-553.

Marohnic CC, Panda SP, Martasek P and Masters BS (2006) Diminished FAD binding in the Y459H and V492E Antley-Bixler syndrome mutants of human cytochrome P450 reductase. *J Biol Chem* **281**:35975-35982.

Michalski C, Cui Y, Nies AT, Nuessler AK, Neuhaus P, Zanger UM, Klein K, Eichelbaum M, Keppler D and Konig J (2002) A naturally occurring mutation in the SLC21A6 gene causing impaired membrane localization of the hepatocyte uptake transporter. *J Biol Chem* **277**:43058-43063.

Miller WL (1986) Congenital adrenal hyperplasia. *N Engl J Med* **314**:1321-1322.

Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S and Marra M (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45**:81-94.

Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**:621-628.

Mutch DM, Klocke B, Morrison P, Murray CA, Henderson CJ, Seifert M and Williamson G (2007) The Disruption of Hepatic Cytochrome P450 Reductase Alters Mouse Lipid Metabolism. *J Proteome Res* **6**:3976-3984.

Nagalakshmi U, Wang Z, Waern F, Shou C, Raha D, Gerstein M and al. e (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**:1344-1349.

Nakata K, Tanaka Y, Nakano T, Adachi T, Tanaka H, Kaminuma T and Ishikawa T (2006) Nuclear receptor-mediated transcriptional regulation in Phase I, II, and III xenobiotic metabolizing systems. *Drug Metab Pharmacokinet* **21**:437-457.

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**:443-453.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J and Bamshad MJ (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**:30-35.

Nishino H and Ishibashi T (2000) Evidence for requirement of NADPH-cytochrome P450 oxidoreductase in the microsomal NADPH-sterol Delta7-reductase system. *Arch Biochem Biophys* **374**:293-298.

Noshiro M, Ullrich V and Omura T (1981) Cytochrome b5 as electron donor for oxy-cytochrome P-450. *Eur J Biochem* **116**:521-526.

Nowell S and Falany CN (2006) Pharmacogenetics of human cytosolic sulfotransferases. *Oncogene* **25**:1673-1678.

Nozawa T, Nakajima M, Tamai I, Noda K, Nezu J, Sai Y, Tsuji A and Yokoi T (2002) Genetic polymorphisms of human organic anion transporters OATP-C (SLC21A6) and OATP-B (SLC21A9): allele frequencies in the Japanese population and functional analysis. *J Pharmacol Exp Ther* **302**:804-813.

Ogino M, Nagata K and Yamazoe Y (2002) Selective suppressions of human CYP3A forms, CYP3A5 and CYP3A7, by troglitazone in HepG2 cells. *Drug Metab Pharmacokinet* **17**:42-46.

Olson H, Betton G, Robinson D, Thomas K, Monro A, Kolaja G, Lilly P, Sanders J, Sipes G, Bracken W, Dorato M, Van Deun K, Smith P, Berger B and Heller A (2000) Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul Toxicol Pharmacol* **32**:56-67.

Ono T and Bloch K (1975) Solubilization and partial characterization of rat liver squalene epoxidase. *J Biol Chem* **250**:1571-1579.

Osada M, Imaoka S, Sugimoto T, Hiroi T and Funae Y (2002) NADPH-cytochrome P-450 reductase in the plasma membrane modulates the activation of hypoxia-inducible factor 1. *J Biol Chem* **277**:23367-23373.

Oshlack A and Wakefield MJ (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* **4**:14.

Otto DM, Henderson CJ, Carrie D, Davey M, Gundersen TE, Blomhoff R, Adams RH, Tickle C and Wolf CR (2003) Identification of novel roles of the cytochrome p450 system in early embryogenesis: effects on vasculogenesis and retinoic Acid homeostasis. *Mol Cell Biol* **23**:6103-6116.

Ozawa S, Soyama A, Saeki M, Fukushima-Uesaka H, Itoda M, Koyano S, Sai K, Ohno Y, Saito Y and Sawada J (2004) Ethnic differences in genetic polymorphisms of CYP2D6, CYP2C19, CYP3As and MDR1/ABCB1. *Drug Metab Pharmacokinet* **19**:83-95.

Ozdemir V, Kalow W, Tang BK, Paterson AD, Walker SE, Endrenyi L and Kashuba AD (2000) Evaluation of the genetic component of variability in CYP3A4 activity: a repeated drug administration method. *Pharmacogenetics* **10**:373-388.

Padol S, Yuan Y, Thabane M, Padol IT and Hunt RH (2007) Clinical impact of CYP2C19 polymorphism on the action of proton pump inhibitors: a review of a special problem. *Int J Clin Pharmacol Ther* **45**:188; author reply 189-190.

Pearce RE, McIntyre CJ, Madan A, Sanzgiri U, Draper AJ, Bullock PL, Cook DC, Burton LA, Latham J, Nevins C and Parkinson A (1996) Effects of freezing, thawing, and storing human liver microsomes on cytochrome P450 activity. *Arch Biochem Biophys* **331**:145-169.

Pepke S, Wold B and Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6**:S22-32.

Phillips AH and Langdon RG (1962) Hepatic triphosphopyridine nucleotide-cytochrome c reductase: isolation, characterization, and kinetic studies. *J Biol Chem* **237**:2652-2660.

Plant N (2007) The human cytochrome P450 sub-family: transcriptional regulation, inter-individual variation and interaction networks. *Biochim Biophys Acta* **1770**:478-488.

Porter TD and Coon MJ (1991) Cytochrome P-450. Multiplicity of isoforms, substrates, and catalytic and regulatory mechanisms. *J Biol Chem* **266**:13469-13472.

Quinlan AR and Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841-842.

Reardon W, Smith A, Honour JW, Hindmarsh P, Das D, Rumsby G, Nelson I, Malcolm S, Ades L, Sillence D, Kumar D, DeLozier-Blanchet C, McKee S, Kelly T, McKeehan WL, Baraitser M and Winter RM (2000) Evidence for digenic inheritance in some cases of Antley-Bixler syndrome? *J Med Genet* **37**:26-32.

Rebbeck TR, Jaffe JM, Walker AH, Wein AJ and Malkowicz SB (1998) Modification of clinical presentation of prostate tumors by a novel genetic variant in CYP3A4. *J Natl Cancer Inst* **90**:1225-1229.

Renwick AB, Mistry H, Ball SE, Walters DG, Kao J and Lake BG (1998) Metabolism of Zaleplon by human hepatic microsomal cytochrome P450 isoforms. *Xenobiotica* **28**:337-348.

Rettie AE and Tai G (2006) The pharmocogenomics of warfarin: closing in on personalized medicine. *Mol Interv* **6**:223-227.

Ribes V, Otto DM, Dickmann L, Schmidt K, Schuhbaur B, Henderson C, Blomhoff R, Wolf CR, Tickle C and Dolle P (2007) Rescue of cytochrome P450 oxidoreductase (Por) mouse mutants reveals functions in vasculogenesis, brain and limb patterning linked to retinoic acid homeostasis. *Dev Biol* **303**:66-81.

Richert L, Liguori MJ, Abadie C, Heyd B, Mantion G, Halkic N and Waring JF (2006a) Gene expression in human hepatocytes in suspension after isolation is similar to the liver of origin, is not affected by hepatocyte cold storage and cryopreservation, but is strongly changed after hepatocyte plating. *Drug Metab Dispos* **34**:870-879.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA and Birol I (2010) De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**:909-912.

Robertson P, DeCory HH, Madan A and Parkinson A (2000) In vitro inhibition and induction of human hepatic cytochrome P450 enzymes by modafinil. *Drug Metab Dispos* **28**:664-671.

Rodriguez-Antona C, Donato MT, Boobis A, Edwards RJ, Watts PS, Castell JV and Gomez-Lechon MJ (2002a) Cytochrome P450 expression in human hepatocytes and hepatoma cell lines: molecular mechanisms that determine lower expression in cultured cells. *Xenobiotica* **32**:505-520.

Rodriguez-Antona C, Donato MT, Boobis A, Edwards RJ, Watts PS, Castell JV and Gomez-Lechon MJ (2002b) Cytochrome P450 expression in human hepatocytes and hepatoma cell lines: molecular mechanisms that determine lower expression in cultured cells. *Xenobiotica* **32**:505-520.

Roy B, Chowdhury A, Kundu S, Santra A, Dey B, Chakraborty M and Majumder PP (2001) Increased risk of antituberculosis drug-induced hepatotoxicity in individuals with glutathione S-transferase M1 'null' mutation. *J Gastroenterol Hepatol* **16**:1033-1037.

Sanderson S, Emery J and Higgins J (2005) CYP2C9 gene variants, drug dose, and bleeding risk in warfarin-treated patients: a HuGEnet systematic review and meta-analysis. *Genet Med* **7**:97-104.

Sassa S, Sugita O, Galbraith RA and Kappas A (1987) Drug metabolism by the human hepatoma cell, Hep G2. *Biochem Biophys Res Commun* **143**:52-57.

Schacter BA, Nelson EB, Marver HS and Masters BS (1972) Immunochemical evidence for an association of heme oxygenase with the microsomal electron transport system. *J Biol Chem* **247**:3601-3607.

Schellens JH, Soons PA and Breimer DD (1988) Lack of bimodality in nifedipine plasma kinetics in a large population of healthy subjects. *Biochem Pharmacol* **37**:2507-2510.

Schwarz UI (2003) Clinical relevance of genetic polymorphisms in the human CYP2C9 gene. *Eur J Clin Invest* **33 Suppl 2**:23-30.

Shen AL, O'Leary KA and Kasper CB (2002) Association of multiple developmental defects and embryonic lethality with loss of microsomal NADPH-cytochrome P450 oxidoreductase. *J Biol Chem* **277**:6536-6541.

Shephard EA, Phillips IR, Santisteban I, West LF, Palmer CN, Ashworth A and Povey S (1989) Isolation of a human cytochrome P-450 reductase cDNA clone and localization of the corresponding gene to chromosome 7q11.2. *Ann Hum Genet* **53**:291-301.

Silva JM, Day SH and Nicoll-Griffith DA (1999) Induction of cytochrome-P450 in cryopreserved rat and human hepatocytes. *Chem Biol Interact* **121**:49-63.

Sim SC, Miller WL, Zhong XB, Arlt W, Ogata T, Ding X, Wolf CR, Flück CE, Pandey AV, Henderson CJ, Porter TD, Daly AK, Nebert DW and Ingelman-Sundberg M (2009) Nomenclature for alleles of the cytochrome P450 oxidoreductase gene. *Pharmacogenet Genomics* **19**:565-566.

Sue Masters B and Marohnic CC (2006) Cytochromes P450--a family of proteins and scientists-understanding their relationships. *Drug Metab Rev* **38**:209-225.

Thomas FJ, McLeod HL and Watters JW (2004) Pharmacogenomics: the influence of genomic variation on drug response. *Curr Top Med Chem* **4**:1399-1409.

Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**:4673-4680.

Thorn CF, Klein TE and Altman RB (2010) Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics* **11**:501-505.

Tirona RG, Leake BF, Merino G and Kim RB (2001) Polymorphisms in OATP-C: identification of multiple allelic variants associated with altered transport activity among European- and African-Americans. *J Biol Chem* **276**:35669-35675.

Trapnell C, Pachter L and Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**:1105-1111.

Trapnell C and Salzberg SL (2009) How to map billions of short reads onto genomes. *Nat Biotechnol* **27**:455-457.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ and Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**:511-515.

Ullrich V (1969) On the hydroxylation of cyclohexane in rat liver microsomes. *Hoppe Seylers Z Physiol Chem* **350**:357-365.

Vermilion JL BD, Massey V, Coon MJ. (1981) Separate roles for FMN and FAD in catalysis by liver microsomal NADPH-cytochrome P-450 reductase. *J Biol Chem* **256**:266-277.

Wadelius M and Pirmohamed M (2007) Pharmacogenetics of warfarin: current status and future challenges. *Pharmacogenomics J* **7**:99-111.

Walker AH, Jaffe JM, Gunasegaram S, Cummings SA, Huang CS, Chern HD, Olopade OI, Weber BL and Rebbeck TR (1998) Characterization of an allelic variant in the nifedipine-specific element of CYP3A4: ethnic distribution and implications for prostate cancer risk. Mutations in brief no. 191. Online. *Hum Mutat* **12**:289.

Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, Macleod JN, Chiang DY, Prins JF and Liu J (2010) MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*.

Wang M, Roberts DL, Paschke R, Shea TM, Masters BS and Kim JJ (1997) Three-dimensional structure of NADPH-cytochrome P450 reductase: prototype for FMN- and FAD-containing enzymes. *Proc Natl Acad Sci U S A* **94**:8411-8416.

Wang SL, Han JF, He XY, Wang XR and Hong JY (2007) Genetic variation of human cytochrome p450 reductase as a potential biomarker for mitomycin C-induced cytotoxicity. *Drug Metab Dispos* **35**:176-179.

Wang Z, Gerstein M and Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**:57-63.

Warren L The PyMOL Molecular Graphics System. DeLano Scientific. San Carlos, California, http://www.pymol.org.

Westlind-Johnsson A, Hermann R, Huennemeyer A, Hauns B, Lahu G, Nassr N, Zech K, Ingelman-Sundberg M and von Richter O (2006) Identification and characterization of CYP3A4*20, a novel rare CYP3A4 allele without functional activity. *Clin Pharmacol Ther* **79**:339-349.

Wettenhall JM, Simpson KM, Satterley K and Smyth GK (2006) affylmGUI: a graphical user interface for linear modeling of single channel microarray data. *Bioinformatics* **22**:897-899.

Williams CH, Jr. and Kamin H (1962) Microsomal triphosphopyridine nucleotide-cytochrome c reductase of liver. *J Biol Chem* **237**:587-595.

Wrighton SA and Stevens JC (1992) The human hepatic cytochromes P450 involved in drug metabolism. *Crit Rev Toxicol* **22**:1-21.

Wu L, Gu J, Weng Y, Kluetzman K, Swiatek P, Behr M, Zhang QY, Zhuo X, Xie Q and Ding X (2003) Conditional knockout of the mouse NADPH-cytochrome p450 reductase gene. *Genesis* **36**:177-181.

Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F and Spencer F (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* **99**:909.

Xu C, Li CY and Kong AN (2005) Induction of phase I, II and III drug metabolism/transport by xenobiotics. *Arch Pharm Res* **28**:249-268.

Yamauchi A, Ieiri I, Kataoka Y, Tanabe M, Nishizaki T, Oishi R, Higuchi S, Otsubo K and Sugimachi K (2002) Neurotoxicity induced by tacrolimus after liver transplantation: relation to genetic polymorphisms of the ABCB1 (MDR1) gene. *Transplantation* **74**:571-572.

Yan H, Yuan W, Velculescu VE, Vogelstein B and Kinzler KW (2002) Allelic variation in human gene expression. *Science* **297**:1143.

Yanagibashi K and Hall PF (1986) Role of electron transport in the regulation of the lyase activity of C21 side-chain cleavage P-450 from porcine adrenal and testicular microsomes. *J Biol Chem* **261**:8429-8433.

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

Zhao Z, Fu YX, Hewett-Emmett D and Boerwinkle E (2003) Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene* **312**:207-213.

APPENDIX

Drug Metabolism and Pharmacokinetics

Date: 12-13-2010

To: DMPK Editorial Office

The Japanese Society for the Study of Xenobiotics (JSSX)

c/o International Medical Information Center

35 Shinanomachi, Shinjuku-ku, Tokyo 160-0016 JAPAN

E-mail: JSSX@imic.or.jp Fax:+81-3-5361-7091

From:

Steven Hart

E-mail: shart3@kumc.edu

I am preparing a paper entitled:

Novel Approaches In Understanding Drug Response

To appear in a book / magazine / journal / proceedings / other (mark one) entitled:

Dissertation

To be published by:

The University of Kansas

Material:

Novel SNPs in cytochrome P450 oxidoreductase.

Hart SN, Li Y, Nakamoto K, Wesselman C, Zhong XB.

Drug Metab Pharmacokinet. 2007 Aug;22(4):322-6.

(The space below is reserved for Drug Metabolism and Pharmacokinetics)

We hereby grant you permission to reprint free of charge in print, subject to the following conditions:

1. Reproduction is confirmed for the purpose for which permission is hereby given, for one-time use only.

2. Appropriate credit to our publication must appear on the copy of your journal.

Approved by: President, JSSX  / Editor-in-Chief, DMPK 

Date:

December 14, 2010

Steven Hart
University of Kansas Medical Center
3901 Rainbow Blvd., MS1018
Kansas City, KS 66160

Email: shart3@kumc.edu

Dear Mr. Hart:

This is to grant you permission to include the following article in your thesis entitled "Novel Paradigms for Understanding Drug Response":

> Steven N. Hart, Ye Li, Kaori Nakamoto, Eva-anne Subileau, David Steen, and Xiao-bo Zhong, A Comparison of Whole Genome Gene Expression Profiles of HepaRG Cells and HepG2 Cells to Primary Human Hepatocytes and Human Liver Tissues, *Drug Metab Dispos* June 2010 38:988-994

On the first page of each copy of this article, please add the following:

> Reprinted with permission of the American Society for Pharmacology and Experimental Therapeutics. All rights reserved.

In addition, the original copyright line published with the paper must be shown on the copies included with your thesis.

Sincerely yours,

Richard Dodenhoff
Journals Director

American Society for Pharmacology and Experimental Therapeutics