Variation in Smoking and Nicotine Metabolism Among American Indians: Novel Influences on In Vivo and In Vitro Nicotine Metabolism Phenotypes

by

Julie-Anne Tanner

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Pharmacology and Toxicology University of Toronto

© Copyright by Julie-Anne Tanner 2017
Variation in Smoking and Nicotine Metabolism Among American Indians: Novel Influences on In Vivo and In Vitro Nicotine Metabolism Phenotypes

By Julie-Anne Tanner

Doctor of Philosophy

Department of Pharmacology and Toxicology

University of Toronto 2017

Abstract

Tobacco use and associated disease risk vary widely among different populations and ethnic/racial groups. While American Indian and Alaska Native populations collectively have the highest prevalence of smoking, recent studies have highlighted disparities in smoking and disease risk across different American Indian/Alaska Native populations. The Northern Plains American Indian tribal populations of South Dakota have a much higher smoking prevalence, smoke more cigarettes per day, and have higher lung cancer risk than the Southwest American Indian tribe of Arizona. Underlying reasons behind the differences in smoking and lung cancer risk between these tribes remain unknown. We sought to biochemically characterize the level of tobacco exposure among smokers and non-smokers in the Northern Plains and Southwest tribes, using cotinine as a biomarker. We demonstrated that despite both tribes being relatively light smoking populations, secondhand tobacco smoke exposure is high among non-smokers, highlighting a potential source of increased risk for tobacco-related disease. As variation in the rate of nicotine metabolism (i.e. CYP2A6 activity) has been associated with differences in smoking behaviours and lung cancer risk in other ethnic/racial groups, we then evaluated variation in the CYP2A6 gene and the rate of nicotine metabolism in the Northern Plains and
Southwest tribes. We identified distinct patterns of \textit{CYP2A6} genetic variation across the two tribes, and an overall faster rate of nicotine metabolism in the Northern Plains compared to the Southwest tribe. We then used a human liver bank to explore additional sources (both genetic and non-genetic) of variation in CYP2A6 activity. We identified several independent and shared sources of variation that could translate to variable hepatic clearance of nicotine, and may impact smoking behaviour. Lastly, we searched for \textit{CYP2A6} gain of function genetic variants in order to identify sources of faster CYP2A6 activity. Through investigation of novel uncharacterized variation at the \textit{CYP2A6} locus using next-generation sequencing, we identified seven novel high frequency \textit{CYP2A6} genetic variants and a five-SNP diplotype, which were associated with faster CYP2A6 activity \textit{in vitro} and \textit{in vivo}. Overall, our findings provide insight into the patterns of tobacco exposure, \textit{CYP2A6} genetic variation, and nicotine metabolism rate in American Indian tribal communities, and expand on our knowledge of genetic and non-genetic sources of variation in nicotine metabolism rate. These findings could improve our understanding of lung cancer risk in these American Indian populations, and help to inform future tobacco cessation efforts.
ACKNOWLEDGEMENTS

To the many people who have supported me throughout my PhD – thank you! There are a few people in particular to whom I would like to extend special thanks.

First, I would like to thank my supervisor, Dr. Rachel Tyndale, for her guidance and invaluable expertise. Under your mentorship, I have been able to grow as a scientist, writer, presenter, and independent thinker. I am grateful to have been pushed outside of my comfort zone, as this has shown me that there are no limits to my capabilities. I truly appreciate the opportunity to learn from you and collaborate with you, and I thank you for your encouragement along the way.

I would also like to acknowledge the collaborators with whom I have been fortunate enough to work. I am thankful to Dr. Jeffrey Henderson, Dr. Kenneth Thummel, and Dr. Steven Scherer for their insight and advice, stemming from their extensive experience and accomplishments in the fields of public health, pharmacology, and genomics.

I also thank my PhD supervisory committee members, Dr. Denis Grant and Dr. Christian Hendershot, for the helpful ideas and feedback that they have provided throughout this entire process. They have both challenged and encouraged me to expand on the interpretations of my data. I was lucky to have them as committee members.

I would also like to extend my thanks to the entire Tyndale lab. Qian Zhou, Ewa Hoffman, and Dr. Zhao Bin were a key part of my learning to conduct quality experimental work within the lab. All members of the Tyndale lab have been a great support system throughout my PhD, but I have to give a special thank you to Dr. Meghan Chenoweth and (soon-to-be) Dr. Doug McMillan for ensuring that there was never a dull moment and for allowing me to sit between you and distract you for five whole years.

In addition to the support received from the Tyndale lab and research community, my friends and family have been a constant source of support and love. Gill and Hélène, you have been along for the ride since long before I embarked on becoming a scientist, and I could not have completed this journey without your friendship and belly laughs.
To my soon-to-be family, the Martelli’s, thank you for welcoming me in to your family and for always showing interest in my work, even if it was a tad confusing. I have always appreciated your support and look forward to boring you with more scientific tidbits in the future.

Of course, I want to thank my parents for making all of this possible. Saying “thank you” just doesn’t seem sufficient. You have made so many sacrifices for Lindsay and me, and have always put our happiness and education first. Mom, you have picked me up when I have stumbled and cheered me on when I needed that extra boost. Dad, you have shown me the meaning of hard work and integrity. Thank you both for the huge role you played in getting me here. Thank you to my sister, Lindsay, for being my partner in crime since birth and for always looking out for me. I also want to thank Gram for encouraging me to learn more about the world around me.

Lastly, I must say thank you to my better half, someone who inspires me to be my best self every day. Mike, you have showered me in love and support for nearly nine years. You are one of the most hard working and passionate people that I know, and saying “yes” to you was one of the smartest (and easiest) decisions I have ever had to make. We have created an incredible life together with our little Bear, and I can’t wait take on this next chapter with you.
# Table of Contents

ACKNOWLEDGEMENTS........................................................................................................ iv  
TABLE OF CONTENTS........................................................................................................ vi  
LIST OF PUBLICATIONS................................................................................................... xi  
LIST OF TABLES............................................................................................................. xiii  
LIST OF FIGURES............................................................................................................ xv  
LIST OF ABBREVIATIONS.............................................................................................. xix  
LIST OF APPENDICES.................................................................................................... xxi  
STATEMENT OF RESEARCH PROBLEM........................................................................ xxii  
MAIN RESEARCH OBJECTIVES......................................................................................... xxv  

1 GENERAL INTRODUCTION......................................................................................... 1  

1.1 Epidemiology of Tobacco Use............................................................................. 1  

1.1.1 Prevalence........................................................................................................ 1  

1.1.2 Health Consequences....................................................................................... 1  

1.1.2.1 Lung Cancer................................................................................................ 3  

1.1.2.2 Chronic Obstructive Pulmonary Disease.................................................. 5  

1.1.2.3 Cardiovascular Disease............................................................................. 6  

1.1.2.4 Other Health Effects.................................................................................. 6  

1.1.3 Vulnerable Populations..................................................................................... 8  

1.1.3.1 Low Income and Socioeconomic Status................................................... 8  

1.1.3.2 Mental Illness and Substance Use Disorders........................................... 9  

1.1.3.3 Racial Minorities...................................................................................... 11  

1.2 Interethnic Tobacco Use and Associated Disease........................................... 13  

1.2.1 Trends in Whites............................................................................................. 13  

1.2.1.1 Prevalence and Characteristics of Tobacco Use...................................... 13  

1.2.1.2 Tobacco-Related Disease Risk.................................................................. 16  

1.2.2 Trends in African Americans.......................................................................... 16  

1.2.2.1 Prevalence and Characteristics of Tobacco Use...................................... 16  

1.2.2.2 Tobacco-Related Disease Risk.................................................................. 18  

1.2.3 Trends in Asian Americans............................................................................ 18
1.2.3.1 Prevalence and Characteristics of Tobacco Use .......................... 18
1.2.3.2 Tobacco-Related Disease Risk .............................................. 19
1.2.4 Trends in American Indians/Alaska Natives ............................... 20
  1.2.4.1 Prevalence and Characteristics of Tobacco Use ......................... 20
  1.2.4.3 Ceremonial Tobacco Use .................................................. 21
  1.2.4.4 Tobacco-Related Disease Risk ............................................ 22
  1.2.4.5 Education and Research Toward Health (EARTH) Study .............. 23

1.3 Pharmacology of Smoking ......................................................... 24
  1.3.1 Nicotine Pharmacodynamics .................................................. 24
    1.3.1.1 Nicotine Neuropharmacology ........................................... 24
    1.3.2.2 Peripheral Effects of Nicotine ........................................ 25
  1.3.2 Nicotine Pharmacokinetics ................................................... 26
    1.3.2.1 Absorption and Distribution .......................................... 26
    1.3.2.2 Metabolism and Clearance ............................................ 26

1.4 Biomarkers of Tobacco Use and Nicotine Metabolism ..................... 30
  1.4.1 Carbon Monoxide ............................................................. 30
  1.4.2 Cotinine ................................................................. 30
  1.4.3 Total Nicotine Equivalents ............................................... 31
  1.4.4 The Nicotine Metabolite Ratio ........................................... 32

1.5 Genetic and Non-Genetic Variation Associated with Nicotine Metabolism .......................... 33
  1.5.1 Cytochrome P450 2A6 (CYP2A6) ............................................ 33
    1.5.1.1 CYP2A6 Genetic Variation and Nicotine Metabolism ............. 33
    1.5.1.2 Interethnic Variation in CYP2A6 and Nicotine Metabolism ...... 38
    1.5.1.3 Unidentified CYP2A6 Genetic Variation .......................... 40
    1.5.1.4 Non-Genetic Influences on CYP2A6 Activity ..................... 41
  1.5.2 Additional Enzymes Involved in Nicotine Metabolism .................. 44
    1.5.2.1 Cytochrome P450 Oxidoreductase (POR) ............................ 44
    1.5.2.2 Aldo-Keto Reductase 1D1 (AKR1D1) ............................... 44

1.6 Associations of CYP2A6 and Nicotine Metabolism with Smoking Behaviours and Disease .................................................. 45
  1.6.1 Smoking Initiation and Dependence ..................................... 45
  1.6.2 Smoking Quantity .......................................................... 47
5 CHAPTER 4: NOVEL CYP2A6 DIPLOTYPES IDENTIFIED THROUGH NEXT-GENERATION SEQUENCING ARE ASSOCIATED WITH IN VITRO AND IN VIVO NICOTINE METABOLISM

5.1 Abstract..........................................................................................................................142
5.2 Introduction.....................................................................................................................143
5.3 Methods..........................................................................................................................145
5.4 Results............................................................................................................................150
5.5 Discussion.......................................................................................................................169
5.6 Significance to Thesis......................................................................................................172

6 GENERAL DISCUSSION....................................................................................................173

6.1 Summary of Research Findings.....................................................................................173

6.2 What Mediates the Low Smoking Quantities in the Northern Plains and Southwest Tribal Populations?..............................................................................................174

6.2.1 Genetic Impacts...........................................................................................................174
  6.2.1.1 CYP2A6 and Nicotine Metabolism.................................................................174
  6.2.1.2 Nicotinic Acetylcholine Receptors.................................................................180
6.2.2 Additional Impacts.......................................................................................................182

6.3 Origins and Genetic Diversity of the Northern Plains and Southwest Tribes...184

6.3.1 Ancestry.....................................................................................................................184
6.3.2 Genetic Divergence.......................................................................................................185
6.3.3 Future Research: Establishing Genetic Divergence of the Northern Plains and Southwest Tribal Populations......................................................................................188

6.4 Sources of Elevated Nicotine Metabolism in the Northern Plains......................190

6.4.1 Genetic Sources...........................................................................................................190
  6.4.1.1 Gene Duplications.............................................................................................196
  6.4.1.2 Transcriptional Regulation...............................................................................197
  6.4.1.3 Post-Transcriptional Regulation .................................................................200
6.4.2 Future Research: Determining Constant and Transient Sources of Variable Nicotine Metabolism in the Northern Plains and Southwest .................................................202

6.5 Using Functional CYP2A6 Diplotypes for Prediction of the NMR.....................204

6.5.1 Future Research: Which of the Novel SNPs are Functional?.................................205
6.5.2 Estimation and Generalizability of Genetic Risk and Phenotype Predictions from CYP2A6 Diplotypes.............................................................................................................207
6.6 Conclusions.........................................................................................................................208

7 BIBLIOGRAPHY......................................................................................................................210

8 APPENDICES..........................................................................................................................243
LIST OF PUBLICATIONS

Research Articles in Thesis


Additional Manuscripts from Doctoral Work Attached in Appendix

RESEARCH ARTICLES


BOOK CHAPTER

Conference Oral Presentations


Conference Poster Presentations


**LIST OF TABLES**

Table 1 | Comparison of tobacco use characteristics and tobacco-related disease risk across ethnic/racial groups .................................................................................................................. 14

Table 2 | Characteristics, functional impact, and frequencies of *CYP2A6* genetic variants (MAF>1%, functionally significant variants only) .................................................................................. 36

Table 3 | Questionnaire assessments and participant classification ............................................. 62

Table 4 | Linear regression analyses of cotinine levels (ng/ml) among Northern Plains and Southwestern smokers and non-smokers .................................................................................................................. 67

Table 5 | Characteristics of study populations in the Northern Plains (South Dakota) and the Southwest (Arizona) .......................................................................................................................... 70

Table 6 | Comparison of Northern Plains (n=318) and Southwestern (n=172) *CYP2A6* allele frequencies, and a summary of frequencies from previously studied populations of different ethnic backgrounds .......................................................................................... 89

Table 7 | Linear regression analyses of the nicotine metabolite ratio among Northern Plains and Southwestern smokers .................................................................................................................................................. 97

Table 8 | Linear regression analyses of the nicotine metabolite ratio among wild-type (*1A/*1A, *1A/*1B, *1B/*1B genotypes) Northern Plains and Southwestern smokers combined into one sample. .................................................................................................................................................. 98

Table 9 | Associations between *CYP2A6* genotype and activity (NMR) with smoking behaviors among Northern Plains and Southwestern tribal populations, with smokers defined by self-report .................................................................................................................................................. 100

Table 10 | Associations between *CYP2A6* genotype and activity (NMR) with smoking behaviors among Northern Plains and Southwestern tribal populations, with smokers defined as having COT>10 .................................................................................................................................................. 101

Table 11 | Optimized MS/MS parameters for CYP2A6 and POR surrogate peptide quantification .................................................................................................................................................. 115

Table 12 | Associations between POR SNPs rs17148944, rs2868177, and rs1057868 with POR mRNA, POR protein, and CYP2A6 activity .................................................................................................................................................. 129

Table 13 | Linear regression analysis of CYP2A6 mRNA expression (FPKM values) ........ 130
Table 14 | Linear regression analysis of CYP2A6 protein levels (pmol/mg) .......................... 132

Table 15 | Linear regression analysis of CYP2A6 enzyme activity (cotinine formation from nicotine, nmol/min per milligram) ................................................................. 134

Table 16 | Linear regression analysis of CYP2A6 enzyme activity (7-OH coumarin formation from coumarin, nmol/min per milligram) ............................................................ 135

Table 17 | N=21 novel SNPs of interest based on bioinformatics analyses, with the top N=7 novel SNPs of interest bolded ................................................................. 151

Table 18 | Summary of the seven novel CYP2A6 SNPs of interest ................................. 152

Table 19 | Summary of effect sizes (Cohen’s d) and P values for seven novel SNPs with CYP2A6 phenotypes (enzyme activity, protein levels, and mRNA expression) in White liver donors ......................................................................................................................... 155

Table 20 | Seven-SNP phased diplotypes predicted in donors in the liver bank .................. 159

Table 21 | Seven-SNP phased diplotypes predicted in treatment-seeking smokers in PNAT trial .................................................................................................................................................. 161

Table 22 | Five-SNP phased diplotypes predicted in donors in the liver bank ..................... 163

Table 23 | Five-SNP phased diplotypes predicted in treatment-seeking smokers in PNAT trial .................................................................................................................................................. 164

Table 24 | Calculated genetic distance between NP and SW American Indian populations, Alaska Natives, and other Asian ancestry populations at the CYP2A6 locus .......................... 187

Table 25 | Comparison of the mean NMR for different CYP2A6 genotype groups across the NP and SW tribes .................................................................................................................. 193
LIST OF FIGURES

Figure 1 | Top ten leading causes of death Worldwide (A), in Canada (B), and in the United States (C)…………………………………………………………………………………………………………………………….. 2

Figure 2 | Prevalence of current smoking is higher among populations of lower socioeconomic status and with mental health and/or substance use disorders in the United States…………….. 8

Figure 3 | Prevalence of current smoking among different ethnic/racial groups in the United States……………………………………………………………………………………………………………………………… 12

Figure 4 | Lung cancer incidence rate (per 100,000 population) among different ethnic/racial groups in the United States……………………………………………………………………………………………………………….. 12

Figure 5 | The major pathways of nicotine metabolism and clearance, with proportion (%) estimates of nicotine and metabolites recovered in urine………………………………………. 27

Figure 6 | Schematic of the CYP2A6 gene and location/distribution of frequent (MAF>1%), functionally significant genetic variants………………………………………………………………………….. 34

Figure 7 | Potential genetic and non-genetic contributors to CYP2A6 activity and nicotine clearance………………………………………………………………………………………………………………… 42

Figure 8 | The nicotine metabolite ratio is associated with end-of-treatment quit rates for the following therapies: (A) Nicotine patch (retrospective analysis), (B) Bupropion (retrospective analysis), (C) Varenicline (prospective analysis) ………………………………………………………………………………………………………. 51

Figure 9 | Association between self-reported smoking status and cotinine levels (ng/ml) among Northern Plains and Southwest smokers and non-smokers ………………………………………….. 64

Figure 10 | Association of cotinine levels (ng/ml) and number of cigarettes smoked per day among Northern Plains (NP), Southwest (SW), and Caucasian (C) smokers ………………….. 65

Figure 11 | Proportion of Northern Plains and Southwest smokers and non-smokers who had cotinine levels equal to or higher than 15 ng/ml ……………………………………………………………….. 66

Figure 12 | (A) Association between cotinine levels (ng/ml) and ceremonial traditional tobacco use among Northern Plains (NP) smokers and non-smokers. (B) Association between the number of cigarettes smoked per day and traditional tobacco use among NP smokers. (C) Association between cotinine levels and traditional tobacco use among Southwest (SW) smokers and non-smokers. (D) Association between the number of cigarettes smoked per day and traditional tobacco use among SW smokers ……………………………………………….. 69
Figure 13 | (A) Association between cotinine levels (ng/ml) and allowing smoking in the home among Northern Plains (NP) smokers and non-smokers. (B) Association between the number of cigarettes smoked per day and allowing smoking in the home among NP smokers. (C) Association between cotinine levels and allowing smoking in the home among Southwest (SW) smokers and non-smokers. (D) Association between number of cigarettes per day and allowing smoking in the home among SW smokers ............................... 71

Figure 14 | (A) Association between cotinine levels (ng/ml) and the number of close friends who smoke among Northern Plains (NP) smokers and non-smokers. $P$ value is based on the Mann-Whitney test. (B) Association between the number of cigarettes smoked per day and the number of close friends who smoke among NP smokers. $P$ value is based on the Kruskal-Wallis test. (C) Association between cotinine levels and the number of close friends who smoke among Southwest (SW) smokers and non-smokers. $P$ values are based on the Mann-Whitney test. (D) Association between the number of cigarettes smoked per day and the number of close friends who smoke among SW smokers ............................... 72

Figure 15 | Comparison of CYP2A6 activity (NMR) between the wild-type CYP2A6 genotype groups *1A/*1A, *1A/*1B, and *1B/*1B ................................................................. 91

Figure 16 | Association between CYP2A6 genotype and CYP2A6 activity (NMR) among smokers .................................................................................................................. 92

Figure 17 | Comparison of CYP2A6 activity (NMR) among the two tribal populations of smokers (NP, Northern Plains; SW, Southwest) ................................................................. 94

Figure 18 | Frequency distributions of the CYP2A6 activity ratio (NMR) among the two tribal populations of smokers (NP, Northern Plains; SW, Southwest), i.e. the number of instances in which NMR has taken each of its possible values ................................................................. 95

Figure 19 | (A) Correlation of two measures of CYP2A6 enzyme activity (the velocity of cotinine formation from nicotine versus the velocity of 7-OH-coumarin formation from coumarin, nmol/min per milligram). (B) Correlation between CYP2A6 mRNA levels (FPKM values, fragments per kilobase per million reads) and CYP2A6 protein levels (pmol/mg). (C) Correlation between CYP2A6 protein levels and CYP2A6 enzyme activity (cotinine formation from nicotine). (D) Correlation between CYP2A6 protein levels (pmol/mg) and CYP2A6 enzyme activity (7-OH coumarin formation from coumarin, nmol/min/ per milligram)........ 119

Figure 20 | Association of gender with (A) CYP2A6 mRNA levels (FPKM values), (B) CYP2A6 protein levels (pmol/mg), (C) CYP2A6 enzyme activity (cotinine formation from nicotine, nmol/min per milligram), and (D) CYP2A6 enzyme activity (7-OH coumarin formation from coumarin, nmol/min per milligram) ................................................................. 120
Figure 21 | Correlation of age with (A) CYP2A6 mRNA levels (FPKM values), (B) CYP2A6 protein levels (pmol/mg), (C) CYP2A6 enzyme activity (cotinine formation from nicotine, nmol/min per milligram), and (D) CYP2A6 enzyme activity (7-OH coumarin formation from coumarin, nmol/min per milligram) …………………………………………………………….. 122

Figure 22 | Association of CYP2A6 genotype with (A) CYP2A6 mRNA levels (FPKM values), (B) CYP2A6 protein levels (pmol/mg), (C) CYP2A6 enzyme activity (cotinine formation from nicotine, nmol/min per milligram), and (D) CYP2A6 enzyme activity (7-OH coumarin formation from coumarin, nmol/min per milligram) ……………………………….. 124

Figure 23 | (A) Correlation between AKR1D1 mRNA expression (FPKM values) and (A) CYP2A6 mRNA expression (FPKM values), (B) CYP2A6 protein levels (pmol/mg), (C) CYP2A6 activity (cotinine formation from nicotine, nmol/min per milligram), and (D) CYP2A6 activity (7-OH coumarin formation from coumarin, nmol/min per milligram). ……….. 126

Figure 24 | (A) Correlation between POR protein levels (pmol/mg) and CYP2A6 activity (cotinine formation from nicotine, nmol/min per milligram). (B) Correlation between POR protein levels (pmol/mg) and CYP2A6 activity (7-OH coumarin formation from coumarin, nmol/min per milligram). (C-E) Association between POR SNP rs17148944 and (C) POR mRNA expression (FPKM values), (D) POR protein levels (pmol/mg) and (E) CYP2A6 activity (cotinine formation from nicotine, nmol/min per milligram) ……………………………….. 128

Figure 25 | Association of seven novel SNPs with CYP2A6 enzyme activity (rate of cotinine formation from nicotine, nmol/min/mg) among N=327 White liver donors. (A) rs57837628, (B) rs7260629, (C) rs7259706, (D) rs150298687, (E) rs56113850, (F) rs28399453, (G) rs8192733 …………………………………………………………………………………………….. 154

Figure 26 | LD plot for seven novel SNPs and established variants (*2, *4, *9, *12); r^2 values (number) and D’ (shading) were derived using Haploview ……………………………….. 157

Figure 27 | LD plot for seven novel SNPs; r^2 values (number) and D’ (shading) were derived using Haploview ……………………………………………………………………………………………………….. 158

Figure 28 | Comparison of CYP2A6 enzyme activity of the full reference diplotype group and top 5 most frequent five-SNP CYP2A6 diplotypes ……………………………………………………………………………………………………….. 167

Figure 29 | Comparison of CYP2A6 enzyme activity of the full reference diplotype group and top 5 most frequent seven-SNP CYP2A6 diplotypes ……………………………………………………………………………………………………….. 168

Figure 30 | (A) Association of number of cigarettes smoked per day with cotinine levels. (B) Approach for predicting NP and SW total nicotine equivalents (TNE). (C) Association of CPD
and predicted TNE for NP and SW smokers, and known TNE for Alaska Native and White smokers.

Figure 31 | Trends of smoking and socioeconomic status across the general United States, Hispanics/Latinos in the U.S., and the Northern Plains and Southwest American Indian populations

Figure 32 | Mechanisms of increasing CYP2A6 mRNA levels via genetic variation. (A) Mechanism of CYP2A6 gene duplications. (B) Mechanisms of transcriptional and post-transcriptional regulation of CYP2A6 mRNA levels

Figure 33 | Observed and predicted associations between the nicotine metabolite ratio (NMR) and the allele frequencies of two novel 5’ CYP2A6 SNPs across different ethnic/racial groups
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3’-UTR</td>
<td>3’-Untranslated region</td>
</tr>
<tr>
<td>3HC</td>
<td>Trans-3’-hydroxycotinine</td>
</tr>
<tr>
<td>AI/AN</td>
<td>American Indian/Alaska Native</td>
</tr>
<tr>
<td>AKR1D1</td>
<td>Aldo-keto reductase 1D1</td>
</tr>
<tr>
<td>BMI</td>
<td>Body mass index</td>
</tr>
<tr>
<td>Bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>CAR</td>
<td>Constitutive androstane receptor</td>
</tr>
<tr>
<td>CDC</td>
<td>US Centers for Disease Control and Prevention</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary deoxyribonucleic acid</td>
</tr>
<tr>
<td>C/EBP</td>
<td>CCAAT-enhancer-binding protein</td>
</tr>
<tr>
<td>CHD</td>
<td>Coronary heart disease</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence interval</td>
</tr>
<tr>
<td>CO</td>
<td>Carbon monoxide</td>
</tr>
<tr>
<td>COPD</td>
<td>Chronic obstructive pulmonary disease</td>
</tr>
<tr>
<td>COT</td>
<td>Cotinine</td>
</tr>
<tr>
<td>CPD</td>
<td>Cigarettes per day</td>
</tr>
<tr>
<td>CNS</td>
<td>Central nervous system</td>
</tr>
<tr>
<td>CVD</td>
<td>Cardiovascular disease</td>
</tr>
<tr>
<td>CYP</td>
<td>Cytochrome P450</td>
</tr>
<tr>
<td>CYP2A6</td>
<td>Cytochrome P450 2A6</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>EARTH</td>
<td>Education and Research Toward Health Study</td>
</tr>
<tr>
<td>FMO</td>
<td>Flavin-containing monoxygenase</td>
</tr>
<tr>
<td>fMRI</td>
<td>Functional magnetic resonance imaging</td>
</tr>
<tr>
<td>FTND</td>
<td>Fagerström test of nicotine dependence</td>
</tr>
<tr>
<td>Gs²</td>
<td>Genetic distance</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome wide association study</td>
</tr>
<tr>
<td>HNF4α</td>
<td>Hepatocyte nuclear factor 4α</td>
</tr>
<tr>
<td>hnRNPA1</td>
<td>heterogeneous nuclear ribonucleoprotein A1</td>
</tr>
<tr>
<td>HONC</td>
<td>Hooked on nicotine checklist</td>
</tr>
<tr>
<td>IBD</td>
<td>Identity-by-descent</td>
</tr>
<tr>
<td>Kb</td>
<td>Kilobase</td>
</tr>
<tr>
<td>LC/MS-MS</td>
<td>Liquid chromatography-tandem mass spectrometry</td>
</tr>
<tr>
<td>LD</td>
<td>Linkage disequilibrium</td>
</tr>
<tr>
<td>MAF</td>
<td>Minor allele frequency</td>
</tr>
<tr>
<td>mFTQ</td>
<td>Modified version of the Fagerström tolerance questionnaire</td>
</tr>
<tr>
<td>MRE</td>
<td>miRNA recognition element</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger ribonucleic acid</td>
</tr>
<tr>
<td>nAChR</td>
<td>Nicotinic acetylcholine receptor</td>
</tr>
<tr>
<td>NMR</td>
<td>Nicotine metabolite ratio</td>
</tr>
<tr>
<td>NNAL</td>
<td>4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol</td>
</tr>
<tr>
<td>NNK</td>
<td>4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone</td>
</tr>
<tr>
<td>NNN</td>
<td>N-nitrosonornicotine</td>
</tr>
<tr>
<td>NP</td>
<td>Northern Plains</td>
</tr>
<tr>
<td>NRT</td>
<td>Nicotine replacement therapy</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>NSCLC</td>
<td>Non small cell lung cancer</td>
</tr>
<tr>
<td>OCT</td>
<td>Organic cation transporters</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PGC-1α</td>
<td>peroxisome proliferator-activated receptor-γ coactivator 1α</td>
</tr>
<tr>
<td>POR</td>
<td>NADPH-cytochrome P450 oxidoreductase</td>
</tr>
<tr>
<td>PNAT</td>
<td>Pharmacogenetics of Nicotine Addiction Treatment</td>
</tr>
<tr>
<td>PXR</td>
<td>Pregnane X receptor</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SEER</td>
<td>Surveillance, Epidemiology, and End Results Program</td>
</tr>
<tr>
<td>SES</td>
<td>Socioeconomic status</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SCLC</td>
<td>Small cell lung cancer</td>
</tr>
<tr>
<td>SUD</td>
<td>Substance use disorder</td>
</tr>
<tr>
<td>SW</td>
<td>Southwest</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>TNE</td>
<td>Total nicotine equivalents</td>
</tr>
<tr>
<td>TRICL</td>
<td>Transdisciplinary Research in Cancer of the Lung</td>
</tr>
<tr>
<td>UDP</td>
<td>Uridine diphosphate</td>
</tr>
<tr>
<td>UGT</td>
<td>UDP-glucuronosyltransferase</td>
</tr>
<tr>
<td>U.S.</td>
<td>United States</td>
</tr>
<tr>
<td>Vcf</td>
<td>Variant call format</td>
</tr>
<tr>
<td>v.s.</td>
<td>Versus</td>
</tr>
<tr>
<td>wGRS</td>
<td>Weighted genetic risk score</td>
</tr>
</tbody>
</table>
LIST OF APPENDICES

Appendix A: Nicotine metabolite ratio (3-Hydroxycotinine/Cotinine) in plasma and urine by different analytical methods and laboratories: implications for clinical implementation

Appendix B: Does coffee consumption impact on heaviness of smoking?

Appendix C: CYP2D6 and CYP2A6 biotransform dietary tyrosol into hydroxytyrosol

Appendix D: Pharmacogenetics of nicotine and associated smoking behaviors

Appendix E: Gender differences in the correlation of CYP2A6 phenotypes in a human liver bank
STATEMENT OF RESEARCH PROBLEM

Tobacco use, killing approximately 6 million people annually, is the leading cause of preventable disease and death worldwide. Although the majority of smokers who are aware of tobacco’s health consequences have a desire to quit, a number of factors influence smoking cessation success, including social, cultural, and biological factors. Tobacco use and associated disease risk vary widely among different populations and ethnic/racial groups. Overall smoking prevalence among adults in the United States was 15% in 2015, but ranged from 7% among Asians to 22% among American Indian/Alaska Natives (AI/ANs). Despite having the highest prevalence of smoking, AI/AN populations are largely understudied. Recently, disparities in smoking and disease risk have been highlighted across different American Indian populations, with the Northern Plains (NP) populations exhibiting a very high smoking prevalence (50%) compared to the Southwest (SW) population (14%). The NP tribal population also smokes on average more cigarettes per day (13 versus 7) and presents much higher lung cancer incidence and mortality rates (>6 times higher), compared to the SW tribe. As underlying reasons behind smoking and lung cancer risk disparities between American Indian tribes remain unknown, further assessment of tobacco use and exposure is warranted, including the investigation of biological markers of exposure, and influences on smoking.

A validated biological measure of tobacco use/exposure is nicotine’s primary metabolite, cotinine. The relatively long half-life (~16 hours) and strong association of cotinine with self-reported cigarettes per day and lung cancer risk underscores the utility of cotinine as a biomarker of tobacco use/exposure. Assessment of cotinine levels in smokers and non-smokers in these tribal populations could highlight patterns of tobacco use and passive smoke exposure, which may help inform future public health and smoke-free policies in tribal communities.

The level and amount of tobacco use can be influenced by genetic factors, which have been assessed in populations of different ethnic/racial backgrounds (i.e. Caucasians, African Americans, Asians, Alaska Natives). Characterization of genetic components of smoking has not occurred in the NP and SW tribal populations. Specifically, the metabolism rate of the primary psychoactive component of cigarette smoke, nicotine, which is primarily dependent on the activity of the hepatic enzyme CYP2A6, is largely influenced by variation in the CYP2A6 gene. Nicotine metabolism (often determined by a biomarker of CYP2A6 activity: 3’-hydroxycotinine/cotinine, known as the nicotine metabolite ratio, NMR) is associated with smoking quantity, tobacco dependence, cessation, and lung cancer risk among smokers.
Despite the consistent association of CYP2A6 genotype with nicotine metabolism within ethnic/racial groups, patterns of nicotine metabolism rate and CYP2A6 genetic variation are inconsistent across different ethnic/racial populations. Considering the tribes’ dissimilar smoking prevalence, consumption, and lung cancer trends, it is plausible that the NP and SW tribes exhibit unique CYP2A6 genetic variation and differences in the rate of nicotine metabolism. Knowledge of these genetic and metabolic tribal patterns may help inform strategies to reduce the high risk of smoking and lung cancer, particularly in the NP population.

Although the rate of nicotine metabolism (measured as the NMR) is significantly associated with CYP2A6 genotype in all ethnic/racial populations studied, a portion of variation in nicotine metabolism remains unaccounted for. Both genetic (e.g. CYP2A6, CYP2B6, FMO, UGT) and non-genetic (e.g. age, sex, race/ethnicity, body mass index, cigarettes per day) factors contribute to the rate of nicotine clearance, with many factors specifically influencing CYP2A6 activity; therefore, there is a need to further characterize the independent and combined impact of genetic and non-genetic factors on CYP2A6 phenotypes. The rate of nicotine metabolism is highly heritable (60-80%), and all top NMR Genome Wide Association Study (GWAS) hits occur at the CYP2A6 locus; this, along with the highly polymorphic CYP2A6 gene (>40 variants characterized to date), suggests that there is novel uncharacterized variation at this gene locus. However, due to the high sequence homology of CYP2A6, CYP2A7, and CYP2A13, a gene-specific high-throughput sequencing approach would be required to selectively identify novel CYP2A6 genetic variation. As the NMR has been successfully used to prospectively tailor smoking cessation treatment (nicotine replacement therapy vs. varenicline), identifying and characterizing novel genetic variants associated with the NMR, using high-throughput sequencing technologies, will aid in the development of CYP2A6 genotype models for personalized prediction of response to cessation pharmacotherapies.

The studies included in this thesis aim to provide insight into the patterns of tobacco exposure, CYP2A6 genetic variation, and nicotine metabolism rate in American Indian tribal communities, which may advance our ability to inform public tobacco use regulations and smoking cessation approaches in culturally distinct and understudied populations. Subsequent investigations expand on the literature describing CYP2A6 genetic variation and nicotine metabolism. Exploration of additional genetic and non-genetic sources of variation in nicotine
metabolism rate, and characterization of novel CYP2A6 genetic variation, will be important for the development and improvement of genetically-tailored personalized cessation approaches.
MAIN RESEARCH OBJECTIVES

Substantial variation in smoking prevalence occurs across different populations and ethnic/racial groups in North America. American Indian and Alaska Native (AI/AN) populations have the highest smoking prevalence of all ethnic/racial groups in the United States, with a similarly high lung cancer risk. Two distinct American Indian tribal populations (Northern Plains, NP, and Southwest, SW) exhibit very different smoking prevalence and lung cancer risk, however reasons for these differences remain unknown. Therefore, the first objective was to biochemically characterize smokers and non-smokers from the NP and SW American Indian tribes, using cotinine as a biomarker of tobacco use/exposure, and to assess factors associated with tobacco use/exposure in these tribal populations.

Many factors are known to influence smoking and lung cancer risk, including several genetic elements. Variability in the rate of nicotine metabolism (i.e. CYP2A6 activity), which is associated with differences in smoking behaviours and lung cancer risk, is largely mediated by variation at the CYP2A6 gene locus. Due to the notable differences in smoking prevalence, quantity, and lung cancer risk/incidence between the NP and SW tribal populations, the second objective of this thesis was to compare patterns of CYP2A6 genetic variation and rate of nicotine metabolism between the NP and SW, and to compare the overall rate of nicotine metabolism across these tribal populations and previously studied Caucasian, African American, and Alaska Native populations.

Through analyses of CYP2A6 among these two AI/AN tribes, we found that NP smokers exhibited a much faster overall rate of nicotine metabolism compared to SW smokers, and compared to previously studied Caucasian, African American, and Alaska Native populations. The elevated rate of nicotine metabolism could not be fully explained by variation in the frequency of the investigated CYP2A6 genetic variants, suggesting there are additional sources of variation in nicotine metabolism. Therefore, we explored other sources of variation in CYP2A6 activity. The third objective of this thesis was to quantify the independent and shared influence of genetic and non-genetic factors on CYP2A6 phenotypes in a human liver bank, including CYP2A6 mRNA and protein expression, and in vitro enzyme activity.

While genetic and non-genetic factors collectively accounted for more than 75% of the variation in CYP2A6 enzyme activity in the human liver bank, additional variation in CYP2A6 activity and other phenotypes remained unaccounted for. Further, only a small proportion
(<1%) of the variation in CYP2A6 activity was attributed to established CYP2A6 genetic variants. Considering the high heritability estimates for nicotine metabolism (60-80%), and the highly polymorphic CYP2A6 locus, additional unknown CYP2A6 genetic variation is probable. In order to extend the investigation beyond the established CYP2A6 alleles in the human liver bank, the fourth objective of this thesis was to use a novel CYP2A6-specific sequencing method to identify functionally significant novel variation at the CYP2A6 gene locus, and quantify the additional portion of variation in CYP2A6 activity that is accounted for by these novel variants. Additionally, we aimed to establish the haplotype structure of novel CYP2A6 variants and assess the association of diplotypes with in vitro CYP2A6 activity in a human liver bank, and in vivo CYP2A6 activity (the NMR) in a clinical trial population.
1 GENERAL INTRODUCTION

1.1 Epidemiology of Tobacco Use

1.1.1 Prevalence

Global smoking prevalence is greater than 1 billion, with approximately 18% of the world’s population reporting regular smoking in 2015 (Jha et al., 2015). While overall global smoking prevalence has decreased between the years 1980 and 2012, total cigarette consumption has increased as a result of population growth (Ng et al., 2014). In Canada, smoking prevalence is approximately 18%, ranging from 14 to 62% across the provinces (lowest: British Columbia, highest: Nunavut) (Statistics Canada, 2016a). Similar trends are observed in the United States (U.S. Department of Health and Human Services), as approximately 15% of the population are smokers, although this ranges from 9 to 26% across the states (lowest: Utah, highest: Kentucky) (Centers for Disease and Prevention, 2015). Smoking prevalence typically peaks between ages 20 to 40 and declines in older ages, with less than 10% of people age 65 and older being smokers in Canada and the U.S. (Jamal et al., 2016, Statistics Canada, 2016b). Current regular smoking is more common among men than women (31% of men and 6% of women globally; 21% of men and 15% of women in Canada; 17% of men and 14% of women in the U.S.) (Ng et al., 2014, Jamal et al., 2016, Statistics Canada, 2016b). Smoking prevalence is also higher among individuals with low income or socioeconomic status (26% below the poverty line versus 14% at or above the poverty line) (Jamal et al., 2016), individuals with mental illness and substance use disorders (SUDs) (36% of individuals with mental illness versus 21% without mental illness) (Centers for Disease and Prevention, 2013), and some racial minorities (22% among American Indian/Alaska Natives versus 17% among Whites) (Jamal et al., 2016).

1.1.2 Health consequences

The main smoking-attributable diseases, lung cancer, cardiovascular disease (CVD), and chronic obstructive pulmonary disease (COPD), are among the top five leading causes of death worldwide, in Canada, and in the U.S. (Fig. 1). The associations of tobacco use with these, and
other, deadly health consequences are described below. This section will focus on North American smoking and disease trends.

Figure 1 | Top ten leading causes of death Worldwide (A), in Canada (B), and in the United States (C). The ★ symbols signify tobacco-related diseases. Adapted from reports by

1.1.2.1 Lung Cancer

Tobacco smoke is composed of over 4000 chemicals, with more than 90 that are classified as harmful or potentially harmful by the U.S. Food and Drug Administration, including more than 70 that are carcinogenic (FDA, 2012). Smoking is the primary risk factor for lung cancer as approximately 90% of lung cancer cases can be attributed to active or passive tobacco smoke exposure (Alberg et al., 2007). A causal link between smoking and lung cancer was first suggested in 1950 when striking trends in tobacco use and increasing lung cancer incidence were identified (Doll and Hill, 1950, Levin et al., 1950). Currently lung cancer is the most common cause of cancer death globally, resulting in 1.7 million deaths in 2015 (World Health Organization, 2017b). The high mortality risk of this disease is illustrated by the low 5-year survival rate, estimated at 17.7% (National Cancer Institute, 2017). With increased education on the harms of smoking, availability of smoking cessation aids, and tobacco control (e.g. indoor smoking bans) smoking prevalence in Canada and the U.S. has decreased. As trends in lung cancer incidence historically reflect smoking patterns from twenty years prior, recent declines in smoking should eventually result in similar decreases in lung cancer mortality. Continuing efforts to increase smoking cessation should result in future declines in lung cancer incidences and death.

Quitting smoking at any age reduces a person’s risk for lung cancer (Office of the Surgeon General, 1990), and improves lung cancer prognosis (Parsons et al., 2010). Former smokers exhibit less than half the relative risk for lung cancer compared to current smokers (Khuder, 2001). Although not as effective as quitting, reduced smoking quantity also reduces lung cancer risk relative to heavier smoking levels. For example, compared to non-smokers, lighter smokers (1-19 cigarettes per day, CPD) had relative risks of 3.4 and 4.3 for small cell carcinoma and squamous cell carcinoma, respectively, which increased to 6.2 and 8.9 for smokers consuming 20-29 CPD, and 11.3 and 18.3 for smokers of 30 or more CPD (Khuder, 2001). Nonetheless, lung cancer risk models predict a stronger impact of smoking duration than quantity (Doll and Peto, 1978), as risk increases exponentially with increasing smoking duration but levels off when reaching high smoking quantities (Bach et al., 2003). Additionally, smoking duration has been strongly linked to lung cancer mortality rates,
regardless of sex, birth cohort, or type of cigarettes smoked (Mannino et al., 2001). These findings suggest that, although lowering cigarette exposure (e.g. dose) reduces lung cancer risk, complete, early cessation should be the primary goal for reducing lung cancer risk and mortality among smokers.

Although lung cancer is considerably less prevalent among non-smokers than smokers, passive smoke exposure is associated with dose- and duration-dependent increases in lung cancer risk and mortality (Office of the Surgeon General, 2006). This risk is further influenced by age; passive smoke exposure that occurs during the period of lung growth and expansion (i.e. at less than 25 years of age) is associated with higher risk for lung cancer, compared to after this maturation period (Asomaning et al., 2008). The majority of passive smoke exposure occurs in the home or historically in the workplace, however other sites of exposure have included public arenas such as restaurants and casinos (Office of the Surgeon General, 2006). Implementation of additional smoke-free policies is associated with substantially decreased passive smoke exposure. For example, 88% of non-smokers over three years old were passively exposed to tobacco smoke (plasma cotinine, a biomarker of nicotine exposure, > 0.05 ng/ml) in 1988-1991 compared to only 25% in 2011-2012 (Homa et al., 2015, Centers for Disease and Prevention, 2010).

Despite the proven associations of tobacco smoke and other environmental exposures with lung cancer risk, the majority of smokers do not get lung cancer (e.g. approximately 17% of male and 12% of female smokers develop lung cancer in Canada, estimated based on rates from 1987-1989) (Villeneuve and Mao, 1994), suggestive of a potential genetic component of lung cancer risk. Investigation of familial lung cancer risk has shown that, particularly among non-smokers with a younger disease onset (40-59 years old), there was a 6-fold increased risk of lung cancer among first-degree relatives (Schwartz et al., 1996).

The different histological types of lung cancer, which are all associated with smoking, include non-small cell (NSCLC) and small cell lung cancer (SCLC). NSCLC is the more common form and consists of the following subtypes: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. As of 2010, adenocarcinoma was the most common subtype among lung cancer patients in the U.S. (41%), followed by squamous cell (21%), SCLC (13%), and large cell (2%) (Lewis et al., 2014). Smoking is more strongly associated (i.e. higher odds ratio) with
the increased risk for squamous cell lung cancer and SCLC, compared to adenocarcinoma and large cell carcinoma (Khuder, 2001).

1.1.2.2 Chronic Obstructive Pulmonary Disease

Chronic Obstructive Pulmonary Disease (COPD) is an irreversible debilitating respiratory condition that develops gradually, with symptoms typically appearing at 40 to 50 years of age (World Health Organization, 2017c). In 2015, over 3 million people died from COPD globally making it the fourth leading cause of death, behind heart disease, stroke, and lower respiratory infections (World Health Organization, 2017a). In people 30 years of age and older the global prevalence of COPD cases has increased substantially from 227 million cases in 1990 to 384 million cases in 2010, likely due to sustained exposure to risk factors and population aging (i.e. general global increases in life-expectancy) (Adeloye et al., 2015).

Similar to lung cancer, the leading risk factor for COPD is tobacco smoke exposure (Chen and Mannino, 1999). The predicted mechanism of COPD development from smoking is through inflammation in central and peripheral airways and lung parenchyma due to exposure to constituents of cigarette smoke (Saetta, 1999). There is a 50% chance that smokers will develop COPD during their lifetime; COPD prevalence was 14% in an overall sample of 1500 subjects, whereas COPD prevalence was 50% among elderly subjects in this sample (Lundback et al., 2003). The association of smoking and COPD is highlighted by notable differences in COPD prevalence among current, former, and never smokers in the U.S. in 2011: 13%, 7%, and 3%, respectively (Centers for Disease and Prevention, 2012). Additionally, there is a linear relationship between smoking duration and prevalence of COPD (Liu et al., 2015). As is the case with lung cancer, smoking cessation improves the prognosis for an individual with COPD (Anthonisen et al., 1994, Xu et al., 1992). Although complete cessation is the primary goal, intermittent periods of cessation, involving failed quit attempts and smoking relapse, are still associated with a slower rate of decline in lung function in smokers with COPD, compared to continued smoking (Murray et al., 1998). Moreover, when comparing intermittent and continuous smokers who consumed the same cumulative amount of cigarettes, intermittent smokers experienced a lower loss of pulmonary function (Murray et al., 1998). In addition to the direct impact of smoking cessation on active smokers, this also limits passive exposure to tobacco smoke among non-smokers and other smokers, which itself is associated with increased risk for COPD (Eisner et al., 2005).
1.1.2.3 Cardiovascular Disease

Cardiovascular disease (CVD) is a broad term encompassing several different types of diseases, including coronary heart disease (CHD) and stroke. CVD is the leading cause of global death, accounting for approximately 17.5 million deaths worldwide in 2013, increasing from approximately 12.5 million in 1990 (O’Rourke, 2015); 80% of those deaths resulted from heart attack or stroke (World Health Organization, 2016). One of the principle risk factors for CVD, including heart attack and stroke, is tobacco smoking (Freund et al., 1993), with 1 of every 10 CVD deaths worldwide attributable to smoking (Ezzati et al., 2005). Smoking increases CVD risk via several mechanisms, including the promotion of atherosclerosis, thrombosis, and inflammation (Office of the Surgeon General, 2004).

Traditional risk factors and inflammatory markers of CVD can be used to quantify smoking-related CVD risk and the impact of reduced smoking quantity or complete cessation. Inflammatory markers of CVD risk include increases in C-reactive protein, white blood cells, and fibrinogen, and decreases in albumin; levels of these markers return to near-normal levels among former smokers after approximately five years of smoking cessation (Bakhru and Erlinger, 2005). However, a longitudinal study suggests 20 years of cessation may be required for CVD risk to reach that of a never smoker (Shields, 2013). Decreasing smoking quantity or intensity may also reduce a smoker’s risk for CVD, as illustrated by the positive association of smoking intensity and levels of inflammatory markers (Bakhru and Erlinger, 2005).

Passive tobacco smoke exposure increases CHD risk by approximately 25% (relative risk 1.25) among non-smokers, compared to non-smokers not exposed to tobacco smoke (He et al., 1999). Further, greater risk was observed for non-smokers exposed to smoke from 20 or more CPD (relative risk 1.31), compared to passive smoke exposure from 1-19 CPD (relative risk 1.23) (He et al., 1999). Implementation of smoking bans has been effective in reducing smoke exposure and CVD risk in high-income countries (Meyers et al., 2009), however, there is still a strong need for smoke-free policies and assessment of resulting CVD benefits in low- and middle-income countries (Olasky et al., 2012).

1.1.2.4 Other Health Effects

In addition to the major tobacco-related diseases described above, smoking is associated with disease in nearly every organ system in the body. Tobacco smoking was responsible for 24%
of all deaths in developed countries in 1990, and for 35% of deaths among people ages 35-69 years (Sasco et al., 2004). Although the lungs are the first point of contact for tobacco constituents and chemicals, toxicants are then carried via the blood to all organs in the body; therefore in addition to lung cancer, smoking has been causally linked to cancers of the urinary tract (relative risk, RR, of 3, for smokers compared to non-smokers), oral cavity (RR 4-5), pharynx (RR 1.5-5), esophagus (RR 1.5-5), larynx (RR 10), pancreas (RR 2-4), nasal cavity and paranasal sinuses (RR 1.5-2.5), stomach (RR 1.5-2), liver (RR 1.5-2.5), kidney (RR 1.5-2), uterine cervix (RR 1.5-2.5), as well as myeloid leukemia (RR 1.5-2) (Sasco et al., 2004). The continuation of smoking after a cancer diagnosis can worsen prognosis; smoking can result in treatment interruptions for patients using radiation and chemotherapies due to increased symptom/side effect burden (Peppone et al., 2011). In addition, smoking impacts the rate of clearance and therefore exposure to drug treatments through induction of drug metabolizing enzymes by constituents of tobacco smoke (O'Malley et al., 2014). Smoking can also reduce immune system function (Holt, 1987, Sopori, 2002), adding to complications that can arise during cancer treatment.

Reproductive health, pregnancy, and early childhood health are also influenced by smoking. Fertility is reduced among female smokers compared to non-smokers, with current smokers exhibiting fertility rates that are 28% lower than non-smokers (Baird and Wilcox, 1985). Among women seeking reproductive assistance, current smoking was associated with lower peak estradiol levels, total oocytes, embryos obtained, and pregnancy rates, compared to never smoking (Van Voorhis et al., 1996). The risk of miscarriage is also increased due to active or passive smoke exposure (relative risk ratios of 1.2 and 1.1, respectively), with the risk being proportional to number of cigarettes smoked per day by active smokers (Pineles et al., 2014). The risk for preterm birth is also cigarette dose-dependent, with heavier smoking being associated with higher risks (Wisborg et al., 1996, Kyrlund-Blomberg et al., 2005). Maternal smoking is associated with infant cleft lip and palate, increasing the relative risk by 34% and 22%, respectively, compared to not smoking during pregnancy (Little et al., 2004). Other maternal smoking-attributable health consequences among offspring include childhood asthma (Jaakkola and Gissler, 2004) and sudden infant death syndrome (Wisborg et al., 2000).

Collectively, health burdens of tobacco smoking are associated with annual economic costs of $326 Billion U.S. dollars in the United States (Xu et al., 2015, Office of the Surgeon General,
2014) and $1436 Billion U.S. dollars globally (Goodchild et al., 2017). Economic costs reflect both direct healthcare costs and indirect costs of reduced productivity.

1.1.3 Vulnerable populations

1.1.3.1 Low Income and Socioeconomic Status

Smoking prevalence and related mortality are inversely related to socioeconomic status (SES). In Canada and the U.S., there are clear trends of decreasing smoking prevalence among those with higher annual household income (smoking prevalence: Canada, 26% and 13% among annual income groups of less than $15,000 versus $80,000 and above, respectively; U.S., 26% and 14% among groups below versus groups at or above poverty level, respectively) (Fig. 2) (Physicians for a smoke-free Canada, 2005, Jamal et al., 2016). Similarly, education level, another measure of SES, is related to smoking prevalence. In Canada, 14% of people with a post-secondary degree reported smoking, compared to 22% with a high school diploma (but no further education) (Physicians for a smoke-free Canada, 2005). In the U.S., the difference is even more prominent, with 7% of people with an undergraduate degree reporting regular smoking compared to 20% of people with a high school diploma (Jamal et al., 2016).

Figure 2 | Prevalence of current smoking is higher among populations of lower socioeconomic status and with mental health and/or substance use disorders in the United States. General U.S., poverty level, education, and serious psychological distress data are from a 2015 survey of adults aged ≥18. Past month mental illness and/or SUD data are from a 1991-
1992 survey of subjects ages 15-54. Adapted from reports by CDC and Lasser and colleagues (Lasser et al., 2000, Jamal et al., 2016)

Several factors likely contribute to the higher smoking prevalence among low SES populations, including, but not limited to, greater risk for initiation, higher tobacco dependence, and reduced adherence to cessation aids (Hiscock et al., 2012). Lower SES is associated with greater risk of first cigarette use and progression to regular smoking (Gilman et al., 2003). Once regular smoking has been established, levels of tobacco consumption, using saliva cotinine levels as an objective measure of tobacco exposure, are higher among economically disadvantaged smokers compared to those living in more affluent households (Fidler et al., 2008). There is an inverse association between smoking quantity and cessation success (Cohen et al., 1989), therefore higher tobacco consumption is one factor that may contribute to lower quitting rates and greater overall smoking prevalence in low SES populations. Additionally, adherence to smoking cessation interventions (behavioural counseling and pharmacotherapies), which is positively associated with cessation success (Clarke et al., 1993, Lam et al., 2005), is lower among smokers of lower SES (Burns and Levinson, 2008, Nevid et al., 1996). Poorer adherence can result, in part, from the cost of therapies (Bonevski et al., 2011). A potentially promising approach for increasing cessation in low SES populations may be to raise the price of tobacco products (Pisinger et al., 2011, Vangeli and West, 2008). As higher mortality rates among lower SES populations are largely attributable to smoking-related diseases, such as lung cancer, COPD, and CVD (Steenland et al., 2004), optimizing cessation treatment strategies among these vulnerable populations will be important in reducing smoking amount and prevalence and the associated health consequences.

1.1.3.2 Mental Illness and Substance Use Disorders

The co-occurrence of one or more forms of mental illness and substance use disorders (SUDs) is well established (Regier et al., 1990). Tobacco use disorder, one type of SUD, is associated with mental illness and other SUDs (Kalman et al., 2005). The term “mental illness” or “mental disorder” encompasses numerous psychological conditions, including, but not limited to, major and mild depression, bipolar disorder, panic disorder, phobias, generalized anxiety disorder, antisocial personality, conduct disorder, and nonaffective psychosis (includes schizophrenia, schizophreniform disorder, schizoaffective disorder, delusional disorder, and atypical psychosis) (World Health Organization, 1993). SUDs are diagnosed based on
recurrent use of one or more substances that cause clinically and functionally significant impairment; these substances include tobacco, alcohol, cannabis, stimulants, hallucinogens, and opioids.

Compared to the general U.S. population, smoking prevalence is 2 to 4 times higher among those with mental illness and/or SUDs, with smoking being most prevalent among alcohol, cocaine, and opioid users, and individuals with schizophrenia (Kalman et al., 2005). In a study of more than 4000 subjects in the U.S., for which the definition of mental illness included alcohol and drug abuse/dependence, 41% of subjects with mental illness in the past month were current smokers, whereas prevalence was only 23% among subjects with no past month mental illness (Fig. 2) (Lasser et al., 2000). Moreover, smoking prevalence was positively associated with number of lifetime mental illness diagnoses, and quit rates were only 31% in subjects with past month mental illness compared to 43% among subjects without past month mental illness (Lasser et al., 2000). Among current smokers, the number of cigarettes smoked per day increases with the level of psychological distress (Lawrence et al., 2009), suggesting individuals with higher levels of psychological distress have increased exposure to toxicants in cigarette smoke. Tobacco-related mortality is 54% higher among those with mental illness and SUDs compared to the general population (Bandiera et al., 2015), and among mentally ill individuals, the predominant cause of death was tobacco-related illness (Druss et al., 2011).

A combination of environmental, social, and biological factors appears to mediate the high smoking prevalence and quantity, and poor cessation rates among subjects with mental illness and/or SUDs. Of particular interest is the etiology of the comorbidity of smoking and mental illness and/or SUDs. These disorders are often characterized by neurobiological abnormalities. Nicotine, the primary psychoactive component of cigarette smoke, modulates the release of the following neurotransmitters, which have been implicated in the comorbidity of smoking and mental illness/SUDs: dopamine, norepinephrine, serotonin, acetylcholine, endogenous opioid peptides, glutamate, γ-aminobutyric acid, and endocannabinoids (Kalman et al., 2005).

Accordingly, nicotine administration has been shown to improve neurocognitive deficits associated with mental illness and SUDs, thus smokers may self-medicate in order to reduce some of the negative symptoms of these disorders (Sacco et al., 2004). As there is a complex relationship between smoking and mental illness/SUDs, several additional explanations for the comorbidity have also been suggested. For example, preexisting genetic factors may mediate
the vulnerability for both smoking and mental illness/SUDs (Brady and Sinha, 2005). Additionally, common environmental stressors may promote smoking behaviour and the onset of other SUDs or psychiatric symptoms (Kalman et al., 2005).

1.1.3.3 Racial Minorities

Patterns of tobacco use and associated disease risk vary widely across populations of different ethnic/racial backgrounds. In 2015, smoking prevalence ranged from 7 to 22% across several ethnic/racial populations in the U.S.: 7% of Asians, 10% of Hispanics, 17% of Whites, 17% of African Americans, 20% of persons of multiple-races, and 22% of American Indians/Alaska Natives (AI/AN) were current cigarette smokers (Fig. 3) (Jamal et al., 2016). The lowest declines in smoking prevalence from 2005 to 2015 were among multiracial and African American populations (Jamal et al., 2016). Recent cessation was less prevalent among African American (5%) relative to White (7%) smokers, while quitting was reported among 8% of Hispanic and 17% of Asian smokers (data not available for AI/AN smokers) (Babb et al., 2017). With respect to smoking cessation, ethnic/racial minorities may be at a greater disadvantage than Whites, who were more likely to receive advice from a health professional to quit, and were more likely to use counseling and/or medication to help with quitting (Babb et al., 2017). Additional disparities are evident in smoking-related disease risk, such that lung and bronchus cancer incidence and mortality are higher among African American than White smokers, followed by AI/ANs, Asians/Pacific Islanders, and Hispanics (Fig. 4) (Horner et al., 2009). Furthermore, in a multi-ethnic cohort study of Whites, African Americans, Japanese Americans, Native Hawaiians, and Latinos, African Americans exhibited the highest lung cancer risk, even after accounting for level of smoking among current and former smokers (Haiman et al., 2006). Described differences in smoking behaviours and disease risk may result from interplay of cultural, environmental, and biological factors. More comprehensive discussion of smoking and related disease among individual ethnic/racial groups is provided in the next section (1.2 Interethnic Tobacco Use and Associated Disease).
Figure 3 | Prevalence of current smoking among different ethnic/racial groups in the United States. White, African American, Asian, Hispanic, and AI/AN data are from a 2015 survey of adults aged ≥18. Alaska Native data are from a 2008-2010 survey of adults aged ≥18. Northern Plains and Southwest American Indian data are from a 1997-1999 survey of subjects ages 15-54. Adapted from reports by CDC, Rohde and colleagues, and Nez Henderson and colleagues (Jamal et al., 2016, Rohde et al., 2013, Nez Henderson et al., 2005).

Figure 4 | Lung cancer incidence rate (per 100,000 population) among different ethnic/racial groups in the United States. Incidence rates for White, African American, Asian/Pacific Islander, Hispanic, and AI/AN groups are 2002-2006 data from the Surveillance, Epidemiology, and End Results (SEER) Program. Alaska Native, Northern Plains, and Southwest American Indian data are from a 1999-2004 survey. Adapted from reports by SEER, and Bliss and colleagues (Horner et al., 2009, Bliss et al., 2008).
1.2 Interethnic Tobacco Use and Associated Disease

A detailed comparison of all ethnic/racial groups is provided in Table 1. The following sections will primarily focus on smoking and related disease risk among individual ethnic/racial groups in the United States, compared to Caucasian/White Americans as a reference group.

1.2.1 Trends in Whites

1.2.1.1 Prevalence and Tobacco Use Characteristics

The majority of studies of smoking and associated risks have been conducted in White populations. The changing prevalence of smoking has been well documented for White smokers in the U.S., with a reported 42% and 17% prevalence of White current smokers in 1965 and 2015, respectively (Jamal et al., 2016, Giovino, 2002).

Important factors that contribute to smoking prevalence in a population are the age and likelihood of smoking initiation. The majority of White smokers reported initiating smoking before the age of 18 (Trinidad et al., 2004). A longitudinal study of smoking in youth showed that maternal smoking, problem behaviour, and peer pressure were positive predictors of smoking initiation among youth of all ethnic/racial groups, whereas low level of positive parenting was an additional predictor of initiation among White youth (Griesler et al., 2002). A role for genetics in smoking initiation has also been reported among Whites (Bares et al., 2016); the association of one important genetic factor, CYP2A6, with smoking initiation will be discussed in section 1.6.1 of the General Introduction of this thesis.

Once regular smoking has commenced, White smokers report heavier smoking (i.e. consume more CPD) on average compared to other ethnic/racial groups – a trend that has persisted over time, despite overall reductions in average CPD (Schoenborn et al., 2013). White smokers aged 25 and older have reduced their average level of tobacco consumption from 23 to 20 CPD (serum cotinine change from 235 to 208 ng/ml) from 1988-1994 to 1999-2002 (O'Connor et al., 2006). Accordingly, 18% of White adults reported heavy smoking (>25 CPD) in 2000, down from 28% in 1974 (Giovino, 2002). Consumption was further reduced to an average of 17 CPD among White smokers aged 18 and older in 2008-2010 (Schoenborn et al., 2013).
Table 1 | Comparison of (A) tobacco use characteristics and (B) tobacco-related disease risk across ethnic/racial groups in the United States. The study year(s) (i.e. timeframe in which the data were collected) are indicated in brackets for each category and ethnic/racial group.

<table>
<thead>
<tr>
<th>Subpopulations</th>
<th>White American (Non-Hispanic)</th>
<th>African American</th>
<th>Asian American</th>
<th>American Indian/Alaska Native Subpopulations</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Alaska Native</td>
<td>Northern Plains</td>
</tr>
<tr>
<td>Urinary TNE levels (6-item), nmol/mg creatinine</td>
<td>53 (2010)</td>
<td>45 (2010)</td>
<td>N/A b</td>
<td>61 (2008-2010)</td>
<td>N/A b</td>
</tr>
<tr>
<td>Smoking duration, years</td>
<td>18-19 (1999-2012)</td>
<td>20-21 (1999-2012)</td>
<td>N/A b</td>
<td>N/A b</td>
<td>N/A b</td>
</tr>
<tr>
<td>Subpopulations</td>
<td>References</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alaska Native</td>
<td>Horner et al. (2009), Kelly et al. (2014), Bliss et al. (2008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northern Plains</td>
<td>Plescia et al. (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southwest</td>
<td>Horner et al. (2009), Plescia et al. (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### (B) Tobacco-Related Disease Risk a

<table>
<thead>
<tr>
<th>Disease Risk</th>
<th>White American (Non-Hispanic)</th>
<th>African American</th>
<th>Asian American</th>
<th>American Indian/Alaska Native Subpopulations</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14 (1999-2009)</td>
<td></td>
</tr>
<tr>
<td>COPD relative risk</td>
<td><strong>1.0</strong> (ref group) (1978-1985)</td>
<td><strong>0.9</strong> (1978-1985)</td>
<td><strong>0.4</strong> (1978-1985)</td>
<td>N/A b</td>
<td>N/A b</td>
</tr>
</tbody>
</table>

a. All data presented are population-level data, with the exception of plasma cotinine and urinary TNE levels, which are reported from studies of select groups of smokers.
b. N/A: Data not available for this category
c. % Quit: the proportion of the ever-smoking population who have quit smoking. This was calculated by dividing the number of former smokers by the number of ever smokers (i.e. former smokers/ever smokers)d. Data reported for the combined AI/AN populations, as separate data were not available for each AI/AN subpopulation. Individuals in this combined group identified themselves as “American Indian/Alaska Native”, therefore this group may represent members of Alaska Native, Northern Plains, Southwest, or other AI/AN tribes.
e. Abbreviations: TNE, total nicotine equivalents; COPD, chronic obstructive pulmonary disease.
With respect to smoking duration, White female and male former smokers reported smoking for 18 and 19 years, respectively (Jones et al., 2016). Cessation prevalence (i.e. proportion of ever smokers who are former smokers) among Whites was approximately 50% in 2000, increasing fairly consistently from 25% in 1965 (Giovino, 2002).

1.2.1.2 Tobacco-Related Disease Risk

As discussed in section 1.1.2, the major smoking-related diseases are lung cancer, chronic obstructive pulmonary disease (COPD), and cardiovascular disease (CVD). Among Whites, incidence of, and mortality from, lung and bronchus cancer was 64 and 54 per 100,000 population, respectively, in 2002-2006 (Horner et al., 2009). With respect to COPD, prevalence was 6% among Whites in the U.S. in 2011 (Centers for Disease and Prevention, 2012). Smoking also increases CVD risk among Whites, with an estimated 67% and 169% higher risk in current smoking White men and women, respectively, compared to never smokers of the same sex (Huxley et al., 2012). Additionally, mortality from coronary heart disease and stroke, two types of CVD, among Whites was 132 and 42 per 100,000 population, respectively, in 2006 (Keenan et al., 2011).

1.2.2 Trends in African Americans

1.2.2.1 Prevalence and Tobacco Use Characteristics

In 2015, 17% of African Americans in the U.S. reported current smoking (Jamal et al., 2016), declining from 46% in 1965 (Giovino, 2002). The majority of African American current and former smokers reported initiating smoking in young adulthood, at 18 years and older (Trinidad et al., 2004, Taylor et al., 1997). Once progressing to regular smoking, African Americans are lighter smokers compared to Whites; only 4% of African American smokers reported heavy smoking (≥ 25 CPD) in 2000 (Giovino, 2002). More specifically, African American adult smokers consumed on average 11 CPD in 2008-2010 (Schoenborn et al., 2013) and averaged fewer than 9 CPD in a study of smokers aged 21-28 (Branstetter et al., 2015). However, CPD may not be an entirely accurate measure of tobacco consumption in African American smokers, as they exhibit greater average plasma cotinine levels than Whites, despite smoking fewer CPD (O'Connor et al., 2006). African Americans also have higher urinary total nicotine equivalents (TNE), a measure of total nicotine intake, compared to Whites, when
controlling for number of CPD (Park et al., 2016). This suggests that there may be ethnic differences in smoking topography, and/or in the rate of metabolism of nicotine and cotinine (discussed in section 1.3.2.2). African Americans tend to smoke cigarettes more intensely, which can be achieved through greater puff volume or duration, than White smokers, resulting in greater nicotine or cotinine levels per cigarette (Perez-Stable et al., 1998, Benowitz et al., 2011, Shiffman et al., 2014). One possible explanation underlying the more intensive smoking among African Americans may be the greater use of menthol cigarettes; 74% of African American smokers consume mentholated cigarettes compared to only 21% of White smokers (Lawrence et al., 2010). Smoking menthol cigarettes may promote deeper or prolonged inhalation, due to the soothing effect of menthol in the throat, however the scientific literature does not support this postulation (Werley et al., 2007). Additionally, higher intensity smoking may primarily occur among lighter smoking African Americans (< 10 CPD); a study of heavier smoking African Americans (mean 18 CPD) and Whites (mean 20 CPD) found no significant differences in plasma cotinine or urinary TNE between the two ethnic/racial groups (St Helen et al., 2013a). Moreover, African American smokers exhibited only slightly higher cotinine/CPD and TNE/CPD compared to Whites, suggestive of little difference in smoking intensity (St Helen et al., 2013a).

With regard to smoking duration, African American females and males smoke for approximately 21 and 20 years, respectively (Jones et al., 2016). Prevalence of smoking cessation has increased from 16% in 1965 to 38% in 2000 among African Americans, although cessation remains lower than in White populations (Giovino, 2002). Similarly, recent quitting was lowest among African Americans compared to all other ethnic/racial groups in 2015, despite a higher proportion of African American smokers reporting interest in quitting and making quit attempts in the past year, compared to White smokers (Babb et al., 2017). Less cessation may result, in part, from the prominent use of mentholated cigarettes by African American smokers, as this has been associated with lower cessation rates among African Americans but not Whites (Stahre et al., 2010). However, the findings on menthol cigarette use and smoking cessation are inconsistent (Blot et al., 2011). Quitting can be improved in African Americans through the use of smoking cessation pharmacotherapies; nicotine replacement therapy (NRT) and bupropion are associated with higher quit rates compared to placebo (Robles et al., 2008). However, African Americans may be less likely to use cessation
pharmacotherapies, for example, they are less likely than Whites to report the use of NRT for smoking cessation (Fu et al., 2008).

1.2.2.2 Tobacco-Related Disease Risk

Lung cancer risk is highest among African American smokers compared to smokers of other ethnic/racial backgrounds; incidence and mortality rates from lung/bronchus cancer are 75 and 60 per 100,000 population, respectively, among African Americans (2002-2006 study) (Horner et al., 2009). This higher risk is consistent across sexes, age groups, histological subtypes, and levels of cigarette consumption (Haiman et al., 2006). Greater exposure to secondhand tobacco smoke and tobacco-specific carcinogens likely contribute to some of the lung cancer risk disparity in African American smokers (Homa et al., 2015, Park et al., 2015). Despite menthol cigarettes potentially contributing to more intensive smoking, no relationship was observed between menthol cigarette use and lung cancer risk among African American smokers in a systematic review (Lee, 2011).

Unlike lung cancer, COPD prevalence is similar among African American and White populations (6% in 2011) (Centers for Disease and Prevention, 2012). However, despite a similar prevalence, African Americans may be more susceptible to COPD, as they exhibit an earlier disease onset (Foreman et al., 2011). In addition, compared to White smokers with COPD, African American smokers with COPD reported fewer pack-years and started smoking at older ages (Chatila et al., 2004).

Disparities in CVD risk are also apparent among African Americans compared to Whites. Death rates for coronary heart disease and stroke were highest in African Americans (162 and 62 per 100,000, respectively, in 2006) (Keenan et al., 2011). Smoking was associated with a 72% and 136% increased risk for CVD risk among African American men and women, respectively (Huxley et al., 2012).

1.2.3 Trends in Asian Americans

1.2.3.1 Prevalence and Tobacco Use Characteristics

Relative to White populations in the U.S., Asian Americans exhibit lower smoking prevalence, with 7% currently smoking in 2015, decreasing from 13% in 2005 (Jamal et al., 2016). However, there is variation in smoking prevalence within ethnic/racial subgroups of Asian
Americans, with the prevalence of 30-day cigarette use ranging from 8-20% (Chinese, 8%; Filipino, 13%; Japanese, 10%; Asian Indian, 8%; Korean, 20%; Vietnamese, 16%, in 2010-2013) (Martell et al., 2016).

More than 65% of Asian Americans initiate regular smoking in young adulthood, between ages 18 and 25, as opposed to in their youth (Trinidad et al., 2004). Language spoken at home and native language skills are predictors of Asian American youth smoking. In a California study of Asian American youth, speaking English at home and having equal or better English language skills, compared to native language, was associated with a greater likelihood of ever smoking (Chen et al., 1999). Adult Asian American daily smokers are generally lighter smokers compared to White smokers, smoking on average 11 CPD (2008-2010) (Schoenborn et al., 2013). In addition to low levels of cigarette consumption, Chinese Americans have lower levels of nicotine intake per cigarette than White smokers, suggesting a lower smoking intensity and presumably less exposure to harmful tobacco constituents (Benowitz et al., 2002). Quit rates are similar among Asian and White American smokers, with approximately 47% of Asian American ever smokers having quit smoking (2002-2003) (Chae et al., 2006). However, smoking cessation varies among Asian American subgroups, with the lowest quit rates occurring in South Asian, Korean, and Vietnamese smokers, and the highest among Filipino, Chinese, and Japanese smokers in California (An et al., 2008).

1.2.3.2 Tobacco-Related Disease Risk

Tobacco-related disease risk is lower among Asian than White Americans, however cancer and CVD remain the leading causes of death among Asian Americans (Hastings et al., 2015). Lung cancer incidence and mortality rates are 39 and 26 per 100,000 population (Horner et al., 2009). Japanese Americans have lower risks for all types of lung cancer compared to White Americans (Haiman et al., 2006), with low cigarette consumption likely mediating a large portion of this reduced risk. Accordingly, Japanese American smokers exhibit significantly lower urinary levels of the tobacco-specific carcinogen NNAL (4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol), relative to White American smokers (Park et al., 2015). Lung cancer risk remains lower among Japanese, compared to White, Americans, regardless of amount of smoking, suggesting additional contribution from other factors (Haiman et al., 2006).
CVD risk is also lowest among Asian Americans, with coronary heart disease and stroke death rates of 77 and 37 per 100,000, respectively, in 2006 (Keenan et al., 2011). Among different Asian American subpopulations, overall CVD mortality rates (per 100,000 population) ranged from 88 to 169 and 72 to 112 in males and females, respectively; rates were considerably higher in White Americans (249 and 159 among males and females, respectively) (Hastings et al., 2015). Following a similar trend, relative risk for COPD was lower among Chinese, Filipino, and Japanese American populations, but higher among South Asian Americans, compared to White Americans, when controlling for age, sex, body mass index, education, alcohol intake, and smoking (Tran et al., 2011).

1.2.4 Trends in American Indians/Alaska Natives

1.2.4.1 Prevalence and Tobacco Use Characteristics

Among ethnic/racial groups in the U.S., American Indians and Alaska Natives (AI/AN) have the highest smoking prevalence. As of 2015, 22% of AI/ANs were current smokers, down from 32% in 2005 (Jamal et al., 2016). However, because AI/AN populations are grouped together in most national surveys in the U.S., vast differences in tobacco use and disease trends among different AI/AN subpopulations and tribal groups can be overlooked.

Smoking trends among Alaska Natives do not reflect the generalized AI/AN data. Reports from 2008-2010 indicate that 41% of Alaska Natives are current smokers (Fig. 3), and 68% are ever smokers, which is significantly higher than the smoking prevalence for non-Native Alaska residents (i.e. non-AI/AN residents of Alaska; 17% are current and 45% are ever smokers) (Rohde et al., 2013). Additionally, approximately 18% of Alaska Natives use smokeless tobacco (Smith et al., 2010). Contributing to the high prevalence of tobacco use, more than 75% of Alaska Natives initiate use of commercial cigarettes and smokeless tobacco by age 18 (Smith et al., 2010). Following initiation, these smokers consume on average less than 10 CPD, and are therefore described as light smokers (Rohde et al., 2013, Smith et al., 2010). However, similar to what has been observed among African American light smokers, Alaska Natives appear to smoke cigarettes more intensely; levels of both urinary TNE and plasma cotinine, two measures of nicotine intake, are comparable between Alaska Native light smokers and White heavy smokers (Zhu et al., 2013a). Although Alaska Native and non-Native smokers are similarly interested in quitting smoking (77% in each group), Alaska Native smokers have less
success, with only 12% of Alaska Natives quitting smoking in the past year relative to 24% of non-Natives (Rohde et al., 2013). Further, the quit ratio (former/ever smokers) for Alaska Natives is 39%, compared to 62% for non-Natives. The widespread tobacco use and a perceived lack of access to cessation interventions (Renner et al., 2004), may partially explain the lack of cessation success among Alaska Natives.

Unlike Alaska Natives, American Indian tribal populations are distributed across multiple States, and most are understudied with respect to smoking and tobacco-related diseases. Two unrelated and culturally distinct American Indian tribal populations, the Northern Plains and Southwest tribal groups, exhibit very different smoking prevalence, patterns of tobacco use, and tobacco-related disease risk. The Northern Plains tribal population refers to several related tribal groups that will be collectively referred to as the Northern Plains tribe. The Northern Plains tribe is located in South Dakota and has a high current smoking prevalence of 50%, whereas prevalence in the Southwest tribe, located in Arizona, is only 14% (1997-1999, tribal members aged 15-54), which is even lower than smoking prevalence among White Americans (Fig. 3) (Jamal et al., 2016, Nez Henderson et al., 2005). In the Northern Plains, 61% of men and 58% of women initiated smoking by age 18, compared to just 32% of men and 23% of women in the Southwest tribe (Nez Henderson et al., 2009). The amount of tobacco consumption is also higher among smokers in the Northern Plains than in the Southwest tribes, who smoke approximately 13 and 7 CPD, respectively (Eichner et al., 2010). These are both therefore relatively light smoking populations, although true levels of consumption have yet to be verified through measurement of urinary TNE or plasma cotinine, as has been performed for Alaska Natives, described above. Following the trend of greater smoking risk in the Northern Plains, quitting prevalence (former/ever smokers) among this tribal group is 29%, compared to 59% in the Southwest (Nez Henderson et al., 2005). A combination of cultural and genetic factors likely contributes to these large differences in smoking prevalence and behaviours between these two American Indian tribes; understanding the underlying mechanisms may be useful for designing future smoking interventions for these groups.

1.2.4.2 Ceremonial Tobacco Use

Use of ceremonial or traditional tobacco by AI/AN populations predates the European colonization of North America. Historically, traditional tobacco was used ceremonially for prayer, or as an offering, medicine, or educational tool (Daley et al., 2006, Linton, 1924).
However, not all AI/AN tribes use ceremonial tobacco, or use it to the same extent; for example, tobacco is not used for ceremonial or traditional purposes among Alaska Natives (Daley et al., 2006). Traditional tobacco is used more extensively in the Northern Plains than in the Southwest tribe, with the Sacred Pipe being an important component of the Northern Plains tribal members’ spiritual philosophy (Nez Henderson et al., 2009, Kunitz, 2016).

Although referred to as “tobacco”, most traditional tobacco does not in fact contain nicotine or portions of the actual tobacco plant, but rather is composed of varieties of dried leaves and barks (Cutler, 2002b, Margalit et al., 2013). In the Northern Plains, traditional tobacco comes in the form of caŋsása, which is composed of red willow or red osier dogwood (Cornus sericea) (Margalit et al., 2013). However, among members of the Northern Plains and other tribes, non-nicotinic traditional tobacco is often used in combination with, or fully replaced by, commercial forms of tobacco, including cigarettes or smokeless tobacco (Nez Henderson et al., 2009, Margalit et al., 2013, Forster et al., 2008, Unger et al., 2006). Traditional forms of tobacco can be difficult and inconvenient to obtain, compared to the ease of purchasing commercial cigarettes (Yazzie et al., 2016). Additionally, AI/AN populations have been targeted by commercial tobacco advertisements. For example, AI/AN imagery has been used for Natural American Spirit cigarettes or Red Man chewing tobacco, which may contribute to the increased use of commercial tobacco for ceremonial purposes (Unger et al., 2006).

1.2.4.3 Tobacco-Related Disease Risk

General estimates of AI/AN lung cancer incidence and mortality are 45 and 40 per 100,000 population, respectively, which is lower than estimates in White Americans (Horner et al., 2009). However, there is considerable heterogeneity in disease risk among different AI/AN populations (Fig. 4). Alaska Natives have significantly higher lung cancer incidence rates than Whites; lung cancer incidence rates (per 100,000 population) were 115 and 78 for Alaska Native men and women, respectively, compared to rates of 72 and 54 among White men and women, respectively (2007-2011) (Kelly et al., 2014). Lung cancer incidence rates in the Northern Plains tribal population exceed that of Alaska Natives, and far exceed Southwest rates; lung cancer incidence was 104 per 100,000 in the Northern Plains, compared to 15 per 100,000 in the Southwest tribe (1999-2004) (Bliss et al., 2008). Similarly, lung cancer mortality rates were 94 and 14 per 100,000 in the Northern Plains and Southwest, respectively (1999-2009) (Plescia et al., 2014).
CVD is another important smoking-related disease among AI/AN populations. Overall, the age-adjusted death rate was 97 per 100,000 and 29 per 100,000 for heart disease and stroke, respectively, among AI/ANs (2006) (Keenan et al., 2011). Although data were not available for the Northern Plains and Southwest tribes specifically, coronary heart disease death rates were reported for the states in which the tribes are found; death rates were highest in South Dakota, the location of the Northern Plains tribal population, and in North Dakota, Wisconsin, and Michigan (Casper, 2005). Death rates from heart disease were low in California and Florida, as well as in Arizona, the home of the Southwest tribe (Casper, 2005). CVD is the leading cause of death among AI/ANs (2007-2009) (Indian Health Service, 2014).

The prevalence of COPD is highest among AI/ANs compared to other U.S. ethnic/racial groups. Overall, 10% of AI/ANs have COPD (2013) (Wheaton et al., 2015); to our knowledge, regional or tribal breakdowns for COPD prevalence have not been reported. Additionally, the rate of COPD-attributable deaths among AI/ANs has increased from 1999 to 2010 (for adults age 25 and older), but not among other ethnic/racial groups (Ford et al., 2013).

1.2.4.4 Education and Research Towards Health (EARTH) Study

Primary reasons for disparities in the prevalence of chronic diseases across AI/AN populations and tribal groups are largely unknown. For this reason, a large National Cancer Institute-funded cohort study of lifestyle factors, cultural factors, and disease associations was initiated in four populations of AI/ANs. This study, named the Education and Research Towards Health (EARTH) study, began in 2001 and Slattery et al. (2007) have described the methods in detail. Briefly, participants were recruited in Alaska, the Northern Plains, the Southwest, and from the Navajo Nation. Eligible participants were 18 years or older, AI/AN ancestry, eligible for health care through the Indian Health Service, residing in the study area, not pregnant, not actively undergoing cancer treatment, physically and mentally able to provide consent, complete surveys, and undergo medical tests, and able to complete the interview in English or tribal language. Medical testing consisted of blood pressure, height, weight, waist and hip circumference, and serum lipid and glucose levels. Surveys and interviews gathered information on diet, health, physical activity, and lifestyle factors, including smoking. Overall, 5207 participants were enrolled, from the Northern Plains (N=3553) and the Southwest (N=1654). Vast differences in smoking patterns were observed between the Northern Plains and Southwest, as described above, which highlighted a need for further investigation of
smoking in these tribes. A follow up study of Northern Plains (N=426) and Southwest (N=210) subjects, resampled from the original EARTH cohort, was conducted in which genetic and non-genetic components of cigarette smoking were investigated. Some of these findings are presented and discussed in Chapters 1 and 2 of this thesis.

1.3 Pharmacology of Smoking

1.3.1 Nicotine Pharmacodynamics

1.3.1.1 Nicotine Neuropharmacology

Nicotine is the main psychoactive component in cigarette smoke, and is a principal factor in mediating tobacco dependence. Upon inhalation of cigarette smoke, nicotine is actively transported to the brain following absorption in the lungs. Nicotine readily crosses the blood-brain barrier and binds central nicotinic acetylcholine receptors (nAChRs) (Liu et al., 2012, Picciotto et al., 1998), which are ligand-gated ion channels, composed of alpha (α2-10) and beta (β2-4) subunits. Different combinations of nAChR subunits form receptor subtypes that elicit distinct responses to agonist binding (Wooltorton et al., 2003). Therefore, a variety of downstream signaling effects result from nicotine-mediated activation of nAChRs, including release of the neurotransmitters dopamine, serotonin, norepinephrine, acetylcholine, γ-aminobutyric acid, glutamate, and endorphins (Benowitz, 2008). The main nAChR subtype that has been implicated in nicotine’s reward pathway is the α4β2 receptor; this subtype is highly sensitive to nicotine and mediates nicotine-stimulated dopamine release (Picciotto et al., 1998, Fenster et al., 1997).

The stimulation of dopamine release following nicotine-mediated activation of nAChRs is thought to be a fundamental rewarding effect of smoking (Corrigall et al., 1992). As evidence of this, rats with lesions in their mesolimbic dopamine system exhibit a significant decrease in nicotine self-administration compared to those with an intact dopamine system (Corrigall et al., 1992). Central dopamine release provides smokers with an initial and rapid feeling of pleasure and reward (Adinoff, 2004), providing near immediate positive reinforcement from cigarette smoke inhalation. The rapidity of reward following cigarette smoke intake, in part, underlies the addiction liability of smoking; this allows a smoker to associate their feeling of reward with the drug taking behaviour (Henningfield and Keenan, 1993). However, nicotine-mediated
stimulation of dopamine release is not the only mechanism of positive feedback from smoking; non-pharmacological effects can also contribute to smoking reinforcement (Rose et al., 2010). Over time, smoking-related cues and other environmental stimuli become associated with nicotine-seeking (in rodents) or smoking behaviour (in humans) (Caggiula et al., 2002). These non-nicotinic factors can contribute to nicotine dependence and help to maintain smoking behaviour by acting as conditioned cues (Conklin et al., 2008).

In addition to acquiring positive reinforcement from smoking, smokers also strive to avoid the negative consequences (i.e. withdrawal symptoms) associated with smoking abstinence. Symptoms of nicotine withdrawal include tobacco cravings, anxiety, stress, negative and depressed mood, irritability, difficulty concentrating, increased hunger, and insomnia (Hughes and Hatsukami, 1986). Regular smoking is associated with neuroadaptation that results in tolerance and withdrawal. Smokers, at typical levels of smoking (i.e. multiple CPD), maintain nearly maximal α4β2 nAChR occupancy throughout the day, which leads to receptor desensitization (closing of ligand-gated ion channel) (Brody et al., 2006). This is followed by post-transcriptional upregulation of nAChRs with a high affinity for nicotine (Perry et al., 1999, Marks et al., 1992), and eventually nicotine tolerance develops from chronic exposure (Pidoplichko et al., 1997). It has been hypothesized that during periods of smoking abstinence, craving and withdrawal occur due to the lack of nAChR occupancy and the recovery of receptors to an active state (Benowitz, 2009). Maintaining the desensitized state of nAChRs through smoking, a form of negative reinforcement, is one way in which smokers avoid symptoms of withdrawal.

1.3.1.2 Peripheral Effects of Nicotine

The rapid absorption of nicotine in the lungs facilitates its transport through the bloodstream, which is associated with peripheral physiological effects, in addition to the central effects, discussed above. Notable peripheral consequences of nicotine intake occur in the cardiovascular system. Smoking is associated with increased blood pressure, heart rate, and myocardial contractility due to increased sympathetic activity (Haass and Kubler, 1997). These cardiovascular effects have been attributed to nicotine-mediated stimulation of peripheral epinephrine and norepinephrine release from sympathetic nerves and adrenal medulla (Haass and Kubler, 1997). Increases in heart rate and blood pressure are observed following
consumption of a single cigarette, and are maintained if smoking persists (Groppelli et al., 1992).

1.3.2 Nicotine Pharmacokinetics

1.3.2.1 Absorption and Distribution

Upon inhalation of cigarette smoke, nicotine, a weak base with a $\text{pK}_a$ of 8, is transported via particulate matter to the lungs (Fowler, 1954). The high surface area in the lung, and the alkaline pH of lung fluid (pH 7.4), help facilitate the absorption of nicotine into the bloodstream (Hukkanen et al., 2005). Approximately 1 to 2 mg of inhaled nicotine is absorbed from a single cigarette, resulting in arterial blood concentrations of approximately 20 to 60 ng/ml, which corresponds to an inhaled bioavailability of 80 to 90% (Armitage et al., 1975, Benowitz and Jacob, 1984, Gori and Lynch, 1985). The rise in blood concentration of nicotine during the process of smoking a cigarette is rapid, peaks immediately following consumption of the cigarette, and is followed by a rapid decline in blood nicotine concentration due to tissue distribution (Benowitz et al., 1988). With a steady state volume of distribution of 1 to 3 L/kg, nicotine is widely distributed to body tissues (Svensson, 1987). Nicotine reaches the brain in approximately 10 to 20 seconds. The liver, kidney, spleen, and lung are the primary tissues in which nicotine is detected (Urakawa et al., 1994). Nicotine’s extensive tissue distribution, and subsequent slow release, contributes to the accumulation of blood nicotine concentrations throughout the day among smokers consuming multiple cigarettes. Daily smokers experience peaks (19-50 ng/ml) and troughs (10-37 ng/ml) in blood nicotine concentrations, but over time, trough levels increase as a result of tissue accumulation and slow release from these tissues (Benowitz et al., 1982, Hukkanen et al., 2005).

1.3.2.2 Metabolism and Clearance

The total clearance of nicotine is approximately 1200 ml/min, the vast majority (95%) of which is attributed to nonrenal (metabolic) clearance (Hukkanen et al., 2005). The primary site of nicotine metabolism is the liver, such that approximately 70% of nicotine is extracted from blood during passage through the liver (Hukkanen et al., 2005). The main enzymes responsible for nicotine metabolism are cytochrome P450s (CYPs; 70-80% of nicotine metabolism) (Benowitz and Jacob, 1994), flavin-containing monooxygenase 3 (FMO3; 4-7%), and uridine
diphosphate-glucuronosyltransferases (UGTs; 3-5%) (Benowitz et al., 1994, Byrd et al., 1992) (Fig. 5).

Figure 5 | The major pathways of nicotine metabolism and clearance, with proportion (%) estimates of nicotine and metabolites recovered in urine. The bolded arrow represents the predominant pathway of nicotine metabolism (70-80% of absorbed nicotine is metabolized by CYP2A6). Adapted from Hukkanen et al. (2005) and Tanner et al. (2015a).

The rate of nicotine inactivation and clearance is primarily dependent on the activity of the CYP2A6 enzyme. Cotinine is the main metabolite of nicotine, with 70 to 80% of nicotine being converted to cotinine; CYP2A6 accounts for approximately 90% of this metabolism (Messina et al., 1997). This metabolic conversion occurs in two steps. First, nicotine is metabolized to the nicotine-Δ1′(5′)-iminium ion by CYP2A6 (the rate-limiting step), followed by the metabolism of this intermediate metabolite to cotinine by the cytosolic enzyme, aldehyde oxidase (Brandange and Lindblom, 1979). The aldehyde oxidase enzyme has not been shown to interact with inhibitors in vivo, and the gene encoding this enzyme is not known
to be polymorphic, therefore we do not anticipate variation in aldehyde oxidase activity influencing the rate of nicotine’s conversion to cotinine (Beedham et al., 2003). Cotinine is then further metabolized to trans-3’hydroxycotinine (3HC) in a process that is entirely mediated by the CYP2A6 enzyme (Nakajima et al., 1996a). The half-lives of nicotine, and its two primary metabolites, cotinine and 3HC, are approximately 1 to 2, 16 to 18, and 6 to 7 hours, respectively (Benowitz and Jacob, 2001, Benowitz and Jacob, 2000, Zevin et al., 1997). The long half-life of cotinine, and the formation dependence of 3HC from cotinine, results in stable relative levels of these metabolites over time in regular smokers. For these reasons, the ratio of these metabolites (3HC/COT, also referred to as the nicotine metabolite ratio, NMR), in the saliva, plasma, or urine of smokers, can be used as a biomarker of CYP2A6 enzymatic activity, and a measure of the rate of nicotine clearance (Dempsey et al., 2004). The NMR is discussed in further detail in section 1.4.4 of this thesis.

The remaining portion of nicotine metabolism to cotinine (approximately 10%) is predominantly mediated by CYP2B6 (Al Koudsi and Tyndale, 2010, Yamazaki et al., 1999). Whereas CYP2A6 is almost exclusively expressed in the liver (Koskela et al., 1999), CYP2B6 is detectable in multiple tissues, including the liver, kidney, intestine, brain, and lung (Gervot et al., 1999, Miksys et al., 2003). Although CYP2B6 plays a minor role in overall nicotine metabolism and clearance, it may modulate central nicotine metabolism thereby limiting nicotine’s duration of action at nAChRs. Other CYPs may play minor roles in tissue-specific metabolism of nicotine to cotinine, such as CYP2A13 in the lung (Bao et al., 2005).

Nicotine is also metabolized to nicotine N’-oxide by FMO enzymes, however only 4 to 7% of nicotine dose recovered in the urine is in the form of nicotine N’-oxide (Benowitz et al., 1994, Byrd et al., 1992). Although nicotine can be metabolized by multiple isoforms of FMOs, FMO3, prevalent in the liver, is responsible for the bulk of the metabolic inactivation of nicotine to nicotine N’-oxide (Zhang and Cashman, 2006, Cashman et al., 1992). The FMO1 enzyme, which is more efficient than FMO3 in the N’-oxidation of nicotine in vitro, is expressed only extrahepatically, in the lung, small intestine, and at highest abundance in the kidney (Zhang and Cashman, 2006, Hinrichs et al., 2011).

The final class of primary mediators of nicotine metabolism are the UGT enzymes. Multiple UGT isoforms contribute to the N-glucuronidation of nicotine, which is then excreted in the urine (Benowitz et al., 1994, Byrd et al., 1992). There is evidence for the involvement of
UGT1A1, 1A4, 1A9, 2B7, and 2B10 in the glucuronidation of nicotine (Kaivosaari et al., 2007, Nakajima et al., 2002, Kuehl and Murphy, 2003), with UGT2B10 playing the most prominent role due to its high affinity for nicotine ($K_m$ 0.29 mM) and high level of expression in human liver, relative to other UGTs (Izukawa et al., 2009, Kaivosaari et al., 2007). Cotinine is also $N$-glucuronidated by UGT2B10 (Chen et al., 2007), with cotinine $N$-glucuronide accounting for 15 to 17% of total nicotine derivatives recovered from urine (Benowitz et al., 1994, Byrd et al., 1992). In contrast, 3HC is $O$-glucuronidated primarily by UGT2B17, and the $O$-glucuronide conjugate constitutes 8 to 9% of total nicotine and metabolites recovered in urine (Benowitz et al., 1994, Byrd et al., 1992, Chen et al., 2012a).

Renal clearance accounts for approximately 5% of total nicotine clearance, which corresponds to an average clearance level of 35 to 90 ml/min. However, the renal excretion of nicotine is largely dependent on urine pH, as nicotine is a weak base and undergoes glomerular filtration, tubular secretion, and reabsorption (Hukkanen et al., 2005). For example, renal nicotine elimination rate was approximately 200% higher among individuals treated with ammonium chloride, which resulted in urine acidification (pH 4.5), compared to placebo treatment (pH 5.6) (Benowitz and Jacob, 1985). This was associated with a 40% increase in total nicotine clearance following ammonium chloride treatment, compared to placebo treatment (Benowitz and Jacob, 1985). Nicotine is ionized in acidic urine, resulting in ion trapping, preventing reabsorption from the urine. Conversely, the alkalization of urine (pH 6.7) through treatment with sodium bicarbonate is associated with a nearly 80% decrease in renal elimination, but only a 1% decrease in total nicotine clearance, compared to placebo (Benowitz and Jacob, 1985). Nicotine will be in an unionized form in alkaline urine, resulting in greater reabsorption from the urine; however, due to the low contribution (5%) of renal elimination to overall nicotine clearance, this corresponds to only a slight decrease in total clearance. As an additional mediator of renal elimination, nicotine may be actively transported across renal epithelium. Organic cation transporters (OCTs), which actively secrete cations at the basolateral membrane (Motohashi and Inui, 2013), can facilitate nicotine transport into a human carcinoma cell line (Zevin et al., 1998), and may be involved in the renal active transport of nicotine (Urakami et al., 1998).
1.4 Biomarkers of Tobacco Use and Nicotine Metabolism

1.4.1 Carbon Monoxide

Exhaled carbon monoxide (CO) is a commonly used, relatively inexpensive, and non-invasive biomarker of smoking. The measurement of expired breath CO is often employed in clinical trials in which it is necessary to differentiate smokers from non-smokers, as it is a more reliable indicator of recent smoking than self-report (Stookey et al., 1987). An expired CO cutoff of greater than 8-10 ppm typically indicates current active smoking, with anyone below this level considered to not have smoked recently (i.e. in the past several hours) (SRNT Subcommittee on Biochemical Verification, 2002). The half-life of CO in sedentary individuals is 4 hours, and in those who are physically active, this is reduced to just 2 hours (SRNT Subcommittee on Biochemical Verification, 2002). This means a physically active smoker who has smoked as recently as approximately 6 hours prior to this measurement may exhibit CO levels below the cutoff for active smoking (SRNT Subcommittee on Biochemical Verification, 2002). An additional caveat of using CO as an indicator for smoking is that it is not tobacco smoke-specific; therefore, expired CO may not be useful in distinguishing between light smokers and non-smokers exposed to environmental sources of combustion. Exposure of non-smokers to combustible reactions, including cooking, burning fossil fuels (for example, automobile emissions), and secondhand smoke, can result in CO exposure levels similar to light or intermittent smokers (Cunnington and Hormbrey, 2002). Recently, a more stringent CO cutoff of 3 ppm has been suggested to reduce the misclassification of smokers, as compared to when using the 8 or 10 ppm cutoffs (Cropsey et al., 2014).

1.4.2 Cotinine

Cotinine is a biomarker of smoking status and quantity that can be measured in the saliva, plasma, or urine of smokers. As discussed in section 1.3.2.2, cotinine is the primary metabolite of nicotine, derived from CYP2A6-mediated nicotine metabolism (Messina et al., 1997). Consequently, cotinine is a nicotine-specific biomarker. Cotinine’s long half-life (16-18 hours) (Benowitz and Jacob, 2000) allows for its detectable and accurate measurement several hours after smoking. Plasma cotinine cutoffs for active smoking can range from 1-15 ng/ml (Jarvis et al., 1987, Benowitz et al., 2009), with the 15 ng/ml cutoff being frequently used, as suggested
by the SRNT Subcommittee on Biochemical Verification (2002). More recent work supports the use of 3 ng/ml as a more stringent cotinine cutoff (Benowitz et al., 2009). Higher cotinine cutoffs were likely used in the past due to greater secondhand tobacco smoke exposure, prior to the widespread implementation of indoor smoking bans (Sanders-Jackson et al., 2013).

Average plasma cotinine levels among White, African American, and Alaska Native smokers have been reported to be approximately 165, 180, and 170 ng/ml, respectively (Benowitz et al., 2011, Zhu et al., 2013a). However, this varies among smokers, primarily depending on their level of tobacco consumption. Hence, in addition to assessing smoking status, cotinine can be used as a measure of the level of tobacco consumption; there is a robust association between self-reported CPD and plasma cotinine level among smokers, particularly in heavy smokers (Heatherton et al., 1989). Among individuals who have slower CYP2A6 enzymatic activity, which can result from genetic and environmental influences (described in section 1.5 of this thesis), plasma cotinine may provide an overestimation of smoking quantity (Zhu et al., 2013c). Cotinine formation and clearance are both mediated by the CYP2A6 enzyme. However, compared to cotinine formation, the rate of cotinine clearance is more dependent on CYP2A6 activity; cotinine formation is more highly dependent on liver blood flow, in addition to CYP2A6 activity. This suggests that, among individuals with slow CYP2A6 activity, cotinine can accumulate at a faster rate than it is cleared, resulting in an overestimation of smoking and tobacco exposure. Similarly, cotinine can inaccurately estimate tobacco consumption due to differences in glucuronidation activity. Smokers with lower levels of UGT2B10-mediated cotinine glucuronidation exhibit higher salivary and plasma cotinine levels, despite similar levels of tobacco consumption (i.e. adjusting for self-reported CPD and urinary TNE) (Murphy et al., 2017). Overall, cotinine is a superior biomarker of smoking status and quantity than CO, but the gold standard measurement of tobacco exposure is urinary TNE, described in the following section.

1.4.3 Total Nicotine Equivalents

The molar sum of nicotine and all its metabolites can be quantified in the urine of smokers and is collectively referred to as total nicotine equivalents (TNE). Metabolites included in this measurement are cotinine, 3HC, nicotine N-glucuronide, cotinine N-glucuronide, 3HC O-glucuronide, nicotine-N-oxide, cotinine-N-oxide, and nornicotine, in addition to urinary
nicotine (9-item TNE), or just urinary nicotine, cotinine, 3HC, and their respective glucuronide conjugates (6-item TNE) (Benowitz et al., 1994, Scherer et al., 2007). Following smoking and nicotine patch administration, approximately 88% and 98% of the respective systemic nicotine doses can be accounted for by urinary TNE (Benowitz et al., 1994). Like cotinine, TNE is a nicotine-specific biomarker, for which measurement can be conducted non-invasively, and it is highly associated with the level of nicotine consumption and self-reported CPD (Benowitz et al., 2010, Lowe et al., 2009). TNE accounts for the number of cigarettes smoked per day as well as puffing behaviour, and it is measured at steady-state nicotine levels, which are achieved by regular smokers. TNE is not directly impacted by variation in the rate of nicotine metabolism, as the primary metabolism pathways are accounted for, and it is therefore superior to cotinine as a biomarker of smoking levels. Consequently, TNE can be used as a measure of tobacco consumption in studies of smokers with different rates of CYP2A6 enzymatic activity (Zhu et al., 2013a).

1.4.4 The Nicotine Metabolite Ratio

As discussed previously, the major nicotine metabolic pathway is the inactivation to cotinine, followed by further metabolism of cotinine to 3HC, processes which are 90% and 100% mediated by the CYP2A6 enzyme, respectively (Messina et al., 1997, Nakajima et al., 1996a). The ratio of these primary metabolites, 3HC/COT (the nicotine metabolite ratio, NMR), is used as a phenotypic biomarker of CYP2A6 enzymatic activity and the rate of nicotine metabolism (Dempsey et al., 2004). The rate of nicotine metabolism, and thus the NMR, varies widely across individuals and is associated with variable smoking behaviour (discussed in section 1.6 of this thesis). The NMR is highly correlated with overall rate of nicotine clearance (Dempsey et al., 2004), which results from the primary role of CYP2A6 in nicotine’s metabolism (70-80%), and the subsequent major contribution of metabolism to overall nicotine clearance (95% metabolic vs. 5% renal clearance) (Hukkanen et al., 2005). Further evidence for the NMR as a biomarker of CYP2A6 activity is the lack of 3HC made by subjects with the CYP2A6*4/*4 genotype (full gene deletion), indicating that there is no conversion of cotinine to 3HC in these CYP2A6 null individuals (Dempsey et al., 2004). The half-lives of cotinine and 3HC are 16 to 18 and 6 to 7 hours, respectively. The formation dependence of 3HC, coupled with the longer relative half-life of cotinine versus 3HC, contributes to the stability of the NMR over time in regular smokers (Mooney et al., 2008, St Helen et al., 2013b) irrespective of sampling time of
day (Lea et al., 2006), and across a 44-week repeated sampling period (St Helen et al., 2012). Measurements of the NMR in different biological matrices (saliva, plasma, blood, and urine) are also consistent (correlations ranging from r=0.76-0.95) (St Helen et al., 2012), as is the measurement of the NMR between laboratories (Appendix A) (Tanner et al., 2015b), further contributing to the utility of this biomarker. Altogether, these findings indicate that the metabolic clearance rate of nicotine in a regular smoker can be reliably and reproducibly estimated by measuring the NMR through obtaining a single biological sample.

Although the NMR is generally consistent within a smoker, this phenotype is highly variable across different smokers (Dempsey et al., 2004). The NMR is influenced by factors that alter the expression and/or function of the CYP2A6 enzyme, including genetic variation in the CYP2A6 gene, and non-genetic and environmental factors that can induce or inhibit the CYP2A6 enzyme; these will be described in section 1.5.

1.5 Genetic and Non-Genetic Variation Associated with Nicotine Metabolism

1.5.1 Cytochrome P450 2A6 (CYP2A6)

1.5.1.1 CYP2A6 Genetic Variation and Nicotine Metabolism

There is interindividual variability in CYP2A6 enzyme activity, and therefore variation in the rate of nicotine metabolism between smokers. According to monozygotic versus dizygotic twin analyses, the NMR phenotype is highly heritable, with genetic influences accounting for 60 to 80% of NMR variation (Loukola et al., 2015, Swan et al., 2009). The primary predictor of nicotine metabolism rate is variation at the CYP2A6 gene locus (Loukola et al., 2015). The CYP2A6 gene is located on chromosome 19 and comprises 9 exons, which span approximately 7 kb of the genome (Yates et al., 2016). More than 40 genetic variants have been characterized at this highly polymorphic gene locus, with the findings for each allele summarized at http://www.cypalleles.ki.se/cyp2a6.htm. Although there are more than 40 CYP2A6 variant alleles, not all are functionally relevant (i.e. causal) with respect to altering nicotine metabolism. The common (minor allele frequency, MAF >1%) variants that alter CYP2A6 mRNA or protein expression and/or enzyme activity are described below, and are visually depicted in Fig. 6.
**Types of Genetic Variation:**

- Regulatory Region Variant (*9, *1B)
- Full Gene Duplication/Deletion Crossover Point (*1X2, *4)
- Two-Nucleotide Deletion (*20)
- Crossover Point for CYP2A7 and CYP2A6 Hybrid (*12)

---

**Figure 6 | Schematic of the CYP2A6 gene and location/distribution of frequent (MAF>1%) genetic variants that alter CYP2A6 enzyme function.** Black boxes depict the nine exons, and connecting lines are introns. Lines extending from exons at the 5’ and 3’ end of the gene represent flanking non-coding sequence. Coloured dots signify different types and locations of CYP2A6 genetic variants. Adapted from (Tanner et al., 2015a, Yates et al., 2016).

The majority of the functionally significant CYP2A6 genetic variants result in a decrease in CYP2A6 expression and/or activity; however, the CYP2A6*1X2A and *1X2B (gene duplications) and *1B alleles are the main exceptions. The CYP2A6 gene duplications result from an unequal crossover of CYP2A6 and the adjacent CYP2A7 gene during recombination (Rao et al., 2000, Fukami et al., 2007). CYP2A6*1X2A and *1X2B, which have breakpoints in intron 8 and 5.2-5.6 kb downstream of CYP2A6, respectively, are associated with faster nicotine metabolism, compared to the wild-type CYP2A6*1A allele (Rao et al., 2000, Fukami et al., 2007). The CYP2A6*1B allele is characterized by a 58 bp gene conversion with CYP2A7, occurring in the CYP2A6 3’-untranslated region (3’-UTR) (Mwenifumbo et al., 2008b). This variant is associated with greater in vivo nicotine metabolism (Mwenifumbo et al., 2008b), possibly resulting from increased CYP2A6 mRNA stability and thus higher protein levels, relative to the wild-type (*1A/*1A) CYP2A6 (Wang et al., 2006); however, no differences in CYP2A6 mRNA, protein, or activity were observed between CYP2A6*1A/*1A and CYP2A6*1A/*1B or *1B/*1B genotype groups in a human liver bank (Al Koudsi et al., 2010).

CYP2A6 genetic variants that have MAFs >1% in one or more populations, and are associated with lower CYP2A6 expression and/or activity, include *2, *4, *5, *7, *9, *10, *12, *17, *18, *20, *21, *23, *24, *25, *28, and *35. The majority of these CYP2A6 polymorphisms are nonsynonymous single nucleotide polymorphisms (SNPs) present in exons, whereas the *4, *9,
*12, and *20 variants are deletion, non-coding, or hybrid alleles. CYP2A6*4 is a full gene deletion resulting from an unequal crossover with CYP2A7. Similar to the duplication variant, there are multiple forms of the *4 variant, resulting from different crossover points (Mwenifumbo et al., 2010). The resulting lack of CYP2A6 mRNA expression is associated with a complete loss of CYP2A6 activity in individuals with two copies of the CYP2A6*4 allele (Oscarson et al., 1999b, Mwenifumbo et al., 2010). CYP2A6*9 is a SNP present in the TATA box of the CYP2A6 promoter region, which also results in decreased CYP2A6 mRNA expression and subsequently lower enzyme activity, but to a lesser degree than *4 (Pitarque et al., 2001). A crossover between the CYP2A6 and CYP2A7 genes produces the CYP2A6*12 hybrid allele, which is characterized by exon 1 and 2 from CYP2A7 and exons 3-9 from CYP2A6; this results in a 10-amino acid substitution and decreased CYP2A6 activity (Oscarson et al., 2002). The CYP2A6*20 allele possesses a two-nucleotide deletion in exon 4 that results in a frame shift, a truncated CYP2A6 protein, and substantially reduced CYP2A6 activity (Fukami et al., 2005a).

Of the nonsynonymous coding region SNPs, *5 (G479V), *7 (I471T), *18 (Y392F), *21 (K476R), *23 (R203C), *24 (V110L and N438Y), *25 (F118L), *28 (N418D and E419D), and *35 (N438Y) are associated with moderately decreased CYP2A6 enzyme activity (Oscarson et al., 1999a, Xu et al., 2002, Fukami et al., 2005b, Mwenifumbo et al., 2008a, Ho et al., 2008, Al Koudsi et al., 2009), whereas *2 (L160H), *10 (I471T and R485L), and *17 (V365M) result in a substantial decrease or complete loss of CYP2A6 activity toward nicotine, similar to *4 and *20 (Hadidi et al., 1997, Oscarson et al., 1998, Xu et al., 2002, Fukami et al., 2004). The functional impact and frequencies of each CYP2A6 variant in different ethnic/racial populations are summarized in Table 2.
Table 2 | Characteristics, functional impact, and frequencies of CYP2A6 genetic variants (MAF>1%, functionally significant variants only). Adapted from Tanner et al. (2015a), Appendix D.

<table>
<thead>
<tr>
<th>CYP2A6 Genetic Variant</th>
<th>rs ID</th>
<th>CYP2A6 Region</th>
<th>Genetic Impact</th>
<th>Functional Impact on CYP2A6 (^a)</th>
<th>Allele Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>White</td>
<td>AA</td>
</tr>
<tr>
<td>*1B</td>
<td>N/A</td>
<td>3’-UTR</td>
<td>58 bp gene conversion with CYP2A7</td>
<td>Increased mRNA stability</td>
<td>28-35</td>
</tr>
<tr>
<td>*1X2A and B</td>
<td>N/A</td>
<td>intron 8 and 5.2-5.6 kb 3’</td>
<td>CYP2A6 gene duplications</td>
<td>Increased mRNA expression</td>
<td>0-1.7</td>
</tr>
<tr>
<td>*2</td>
<td>rs1801272</td>
<td>Exon 3</td>
<td>Nonsynonymous, L160H</td>
<td>Substantially decreased enzyme activity</td>
<td>1.1-5.3</td>
</tr>
<tr>
<td>*4</td>
<td>N/A</td>
<td>N/A</td>
<td>CYP2A6 gene deletion</td>
<td>No mRNA expression</td>
<td>0.1-4.2</td>
</tr>
<tr>
<td>*5</td>
<td>rs5031016</td>
<td>Exon 9</td>
<td>Nonsynonymous, G479V</td>
<td>Decreased enzyme activity</td>
<td>0-0.3</td>
</tr>
<tr>
<td>*7</td>
<td>rs5031017</td>
<td>Exon 9</td>
<td>Nonsynonymous, I471T</td>
<td>Decreased enzyme activity</td>
<td>0-0.3</td>
</tr>
<tr>
<td>*9</td>
<td>rs28399433</td>
<td>5’</td>
<td>Promoter SNP, interrupts TATA box</td>
<td>Decreased mRNA expression</td>
<td>5.2-8.0</td>
</tr>
<tr>
<td>*10</td>
<td>rs5031017, rs28399468</td>
<td>Exon 9</td>
<td>Nonsynonymous, I471T, R485L</td>
<td>Inactive enzyme</td>
<td>0</td>
</tr>
<tr>
<td>*12</td>
<td>N/A</td>
<td>N/A</td>
<td>Exons 1-2 from CYP2A7, exons 3-9 from CYP2A6, 10 amino acid substitution</td>
<td>Decreased enzyme activity</td>
<td>0-0.3</td>
</tr>
</tbody>
</table>
Table 2 continued…

<table>
<thead>
<tr>
<th>CYP2A6 Genetic Variant</th>
<th>rs ID</th>
<th>CYP2A6 Region</th>
<th>Genetic Impact</th>
<th>Functional Impact on CYP2A6&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Allele Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>White</td>
</tr>
<tr>
<td>*17</td>
<td>rs28399454</td>
<td>Exon 7</td>
<td>Nonsynonymous, V365M</td>
<td>Substantially decreased enzyme activity</td>
<td>0</td>
</tr>
<tr>
<td>*18</td>
<td>rs1809810</td>
<td>Exon 8</td>
<td>Nonsynonymous, Y392F</td>
<td>Decreased enzyme activity</td>
<td>1.1-2.1</td>
</tr>
<tr>
<td>*20</td>
<td>N/A</td>
<td>Exon 4</td>
<td>Two-nucleotide deletion, frame shift, truncated protein</td>
<td>Substantially decreased protein levels</td>
<td>0</td>
</tr>
<tr>
<td>*21</td>
<td>rs6413474</td>
<td>Exon 9</td>
<td>Nonsynonymous, K476R</td>
<td>Decreased enzyme activity</td>
<td>0-2.3</td>
</tr>
<tr>
<td>*23</td>
<td>rs56256500</td>
<td>Exon 4</td>
<td>Nonsynonymous, R203C</td>
<td>Decreased enzyme activity</td>
<td>0</td>
</tr>
<tr>
<td>*24</td>
<td>rs72549435, rs143731390</td>
<td>Exon 2, 9</td>
<td>Nonsynonymous, V110L, N438Y</td>
<td>Decreased enzyme activity</td>
<td>0</td>
</tr>
<tr>
<td>*25</td>
<td>rs28399440</td>
<td>Exon 3</td>
<td>Nonsynonymous, F118L</td>
<td>Decreased enzyme activity</td>
<td>0</td>
</tr>
<tr>
<td>*28</td>
<td>rs28399463</td>
<td>Exon 8</td>
<td>Nonsynonymous, N418D, E419D</td>
<td>Decreased enzyme activity</td>
<td>–</td>
</tr>
<tr>
<td>*35</td>
<td>rs143732390</td>
<td>Exon 9</td>
<td>Nonsynonymous, N438Y</td>
<td>Decreased enzyme activity</td>
<td>0</td>
</tr>
</tbody>
</table>

<sup>a</sup> Functional impact toward nicotine metabolism (nicotine C-oxidation)
The decrease- and loss-of-function CYP2A6 genetic variants are typically grouped together into an overall reduced nicotine metabolizer group, or into intermediate (decrease-of-function) and slow (substantial decrease-of-function or loss-of-function) metabolizer groups, which have been extensively characterized, both in vitro and in vivo. For example, in a study assessing CYP2A6 phenotypes using liver tissue from 67 human donors, CYP2A6 protein levels, determined from the measurement of immunoreactivity by Western blotting, were significantly lower among donors with one or more decrease- or loss-of-function CYP2A6 genetic variants (i.e. reduced metabolizers) compared to CYP2A6*1/*1 donors (i.e. normal metabolizers, including CYP2A6*1A/*1A, *1A/*1B, and *1B/*1B genotypes) (Al Koudsi et al., 2010). CYP2A6 enzymatic activity (nicotine C-oxidation in human liver microsomes) was correspondingly lower among liver donors who were CYP2A6 genotype reduced versus normal metabolizers. In multiple clinical trials, and in studies of different ethnic/racial groups, the NMR was lower among smokers who were CYP2A6 genotype reduced nicotine metabolizers, compared to smokers who were CYP2A6 genotype normal metabolizers (Ho et al., 2009b, Malaiyandi et al., 2006, Binnington et al., 2012, Park et al., 2016).

1.5.1.2 Interethnic Variation in CYP2A6 and Nicotine Metabolism

Different ethnic/racial groups exhibit distinct patterns of CYP2A6 genetic variation, as summarized in Table 2. Some CYP2A6 alleles are more common, or have thus far only been observed in specific ethnic/racial populations. For example, the CYP2A6*7 allele has been found predominantly in Asian populations (Mwenifumbo et al., 2005), and, to date, CYP2A6*17, *20, *23-25, and *28 have been identified in people of African descent (Nakajima et al., 2006, Ho et al., 2008, Ho et al., 2009b). Despite differences in allele frequencies, the impact of each allele on CYP2A6 activity is similar across ethnic/racial populations. For example, a similar decrease in CYP2A6 enzyme activity of 33% and 39%, respectively, is observed in White and African American smokers with the CYP2A6*1/*9 genotype compared to CYP2A6*1/*1 (wild-type genotype) (Park et al., 2016).

Asian and African American populations tend to have higher frequencies of CYP2A6 decrease- or loss-of-function genetic variants, and consequently exhibit slower overall rates of nicotine metabolism, compared to White populations (Schoedel et al., 2004, Nakajima et al., 2006, Mwenifumbo et al., 2005). For example, compared to Whites, Japanese American smokers have, on average, half the rate of CYP2A6 enzymatic activity (measured by urinary total
3HC/free COT) (Park et al., 2016). Similarly, after chewing nicotine gum, Japanese non-smokers exhibit nearly half the average CYP2A6 activity (measured by plasma cotinine/nicotine) compared to White non-smokers, consistent with a much higher frequency of CYP2A6 reduce-of-function genetic variants (overall frequency of reduce-of-function alleles in each population: 50% in Japanese Americans, 12% in Whites) and CYP2A6 genotype reduced metabolizers (overall frequency of individuals with reduce-of-function CYP2A6 genotypes: 72% of Japanese Americans, 21% of Whites) (Nakajima et al., 2006). These trends also extend to other Asian populations. Following infusions of deuterium-labeled nicotine and cotinine, Chinese American smokers exhibit significantly lower total and non-renal clearance of both compounds, compared to White smokers (Benowitz et al., 2002). Likewise, salivary NMR was significantly lower in Asian than in White adolescents (ages 13-17) (Rubinstein et al., 2013).

Among African Americans, there are also more CYP2A6 genotype reduced metabolizers compared to in White populations (overall frequency of individuals with reduce-of-function CYP2A6 genotypes: 38% of African Americans, 21% of Whites) (Nakajima et al., 2006). Consistent with these higher frequencies, African Americans infused with deuterium-labeled nicotine and cotinine exhibit significantly lower non-renal clearance of cotinine compared to Whites (Perez-Stable et al., 1998), suggestive of slower metabolism, which is largely mediated by CYP2A6. This is supported by a separate observation of lower overall CYP2A6 enzyme activity (measured by plasma NMR) among African American compared to White smokers (Ross et al., 2016). Compared to Whites, African American adolescent (ages 13-17) smokers also have significantly slower CYP2A6 activity (lower salivary NMR) (Rubinstein et al., 2013).

Compared to Whites, Alaska Natives also have higher frequencies of certain reduce-of-function CYP2A6 genetic variants (higher *4, *9, and *10 allele frequencies), and these variants are similarly associated with lower CYP2A6 activity among Alaska Natives, as in other ethnic/racial groups (Binnington et al., 2012). The CYP2A6 gene deletion (the *4 allele) is present at a frequency of nearly 15% in Yupik Alaska Native people, compared to approximately 1% in Whites. Similarly, the frequency of the CYP2A6*10 haplotype was 2% in this population, but has not been reported to occur in Whites. Despite the higher frequencies of these slower CYP2A6 activity alleles, Alaska Native smokers exhibit faster overall CYP2A6 activity (measured by plasma NMR) compared to White smokers. This difference is further
increased when comparing CYP2A6 activity among Alaska Native and White subjects with the CYP2A6*/*1 (wild-type) genotype (i.e. excluding all subjects who possess any reduce-of-function CYP2A6 genetic variants). The observed elevated rate of nicotine metabolism is not accounted for by known gain-of-function CYP2A6 genotypes, and therefore other factors (novel genetic variation, dietary inducers, environmental exposures, etc.) may be contributing to the faster CYP2A6 activity in this population.

Parallel comparisons cannot be made between White and American Indian smokers as CYP2A6 genotype and the rate of nicotine metabolism has not been assessed among American Indian populations. This has been investigated and described in Chapter 2 of this thesis (section 4).

1.5.1.3 Unidentified CYP2A6 Genetic Variation

As previously discussed, the NMR is highly heritable, such that 60 to 80% of the variation in the phenotype is attributable to genetic variation, and all top NMR GWAS hits (i.e. most highly significant genome-wide significant signals) occur at or near the CYP2A6 locus (within the CYP2A6 gene, several kb downstream of CYP2A6, or in the CYP2A6-CYP2A7 intergenic region) (Loukola et al., 2015, Swan et al., 2009, Baurley et al., 2016). The established CYP2A6 genetic variants (i.e. some of those that were described above) and the top independent SNPs in the GWAS accounted for 20 to 30% of this NMR variation, indicating that a substantial portion of the genetic contribution to the NMR remains uncharacterized (Loukola et al., 2015).

Considering the highly polymorphic nature of the CYP2A6 gene locus (>40 variants characterized to date), it is highly likely that there is additional functional genetic variation within this gene.

Recent advancements in sequencing technology will allow for the further characterization of the CYP2A6 gene locus. It has historically been difficult to sequence the CYP2A6 locus, as CYP2A6 shares approximately 95% sequence homology with the CYP2A7 and CYP2A13 genes (Fernandez-Salguero et al., 1995). This has limited the discovery and characterization of novel CYP2A6 genetic variants. Consequently, individuals are typically characterized for CYP2A6 using an accurate, but time-consuming, allele-specific genotyping approach, which identifies the genotype for a variant of interest, but does not identify novel variants. This two-step method involves targeted amplification of the CYP2A6 gene, followed by targeting the specific allele of interest in a second DNA amplification (Wassenaar et al., 2016). The approach
ensures selective genotyping calls for CYP2A6, as opposed to the highly homologous adjacent CYP2A7 or CYP2A13 genes. Next-generation sequencing advancements, with high-throughput capabilities and improved CYP2A6 specificity, will be useful in further developing a comprehensive profile of CYP2A6 genetic variation.

1.5.1.4 Non-Genetic Influences on CYP2A6 Activity

Although CYP2A6 genetic variation is the primary mediator of CYP2A6 enzyme activity and nicotine metabolism, CYP2A6 activity is ultimately determined by a combination of genetic and non-genetic factors (Fig. 7). A person’s age may impact their rate of nicotine clearance, however evidence suggests that this may not be a CYP2A6 metabolic effect. A lower rate of nonrenal nicotine clearance has been observed in subjects aged 65 and older, compared to subjects in the 22-43 age group (Molander et al., 2001). However, no association of age and CYP2A6 protein or enzyme activity was observed in a human liver bank (N=67 livers) (Al Koudsi et al., 2010), suggesting that the lower nicotine clearance in elderly subjects may be a product of physiological effects of aging, such as reduced liver blood flow. Likewise, despite having a longer nicotine half-life than adults, neonates exposed to tobacco smoke exhibited similar cotinine half-lives to adults (Dempsey et al., 2000). This again suggests that the age disparities in nicotine clearance may result from potential age-related differences in liver blood flow (Gow et al., 2001). Nicotine has a higher extraction ratio than cotinine, and nicotine’s clearance is therefore more sensitive to changes in liver blood flow, whereas cotinine clearance is more sensitive to changes in intrinsic clearance and thus to changes in CYP2A6 enzymatic activity (Hukkanen et al., 2005).

There are also sex differences in the rate of nicotine metabolism. Among women and men who were administered deuterium-labeled nicotine and cotinine, multiple measures of CYP2A6 activity (nicotine clearance, cotinine clearance, fractional conversion of nicotine to cotinine, and the NMR) were significantly higher in the women (Benowitz et al., 2006). In this same study, women who were taking oral contraceptives that contain estrogen exhibited greater nicotine metabolism rate compared to women who were not using oral contraceptives, or were using estrogen-free therapies. Lastly, there was no difference in the rate of nicotine metabolism between males and menopausal or post-menopausal women in this study. A similar relationship was observed in a separate investigation of female CYP2A6 activity (the NMR) with and without the use of estrogen-containing oral contraceptives or hormone replacement.
therapies; women taking the birth control pill or using hormone replacement therapy had 20% and 30% higher CYP2A6 activity, respectively, compared to women who were not taking these drugs (Chenoweth et al., 2014a). Together, these findings suggest that estrogen is involved in differences in the rate of nicotine metabolism. The mechanism is believed to involve estrogen binding to the estrogen receptor, a nuclear hormone receptor that, upon activation, relocates to the nucleus and binds to an estrogen response element located several kb upstream of the CYP2A6 gene (Higashi et al., 2007). This results in the transactivation of the CYP2A6 promoter, and the induction of transcription of the gene. In support of this, female liver donors exhibit higher CYP2A6 mRNA and protein levels than male donors in a human liver bank (N=67 livers) (Al Koudsi et al., 2010).

In addition to oral contraceptives and hormone replacement therapies, several medications and drugs may alter the rate of nicotine metabolism and clearance. Phenobarbital, dexamethasone, and rifampin are associated with CYP2A6 induction (Rae et al., 2001, Donato et al., 2000, Maurice et al., 1991), potentially via processes mediated by constitutive androgen receptor (CAR), pregnane X receptor (PXR), and peroxisome proliferator-activated receptor-γ coactivator 1α (PGC-1α) signaling (Itoh et al., 2006, Moore et al., 2000). Conversely, several drugs act as inhibitors of CYP2A6. For example, methoxsalen (8-methoxysporalen) is a mechanism-based CYP2A6 inhibitor (Koenigs et al., 1997). In addition, tobacco smoking is associated with reduced nicotine metabolism and clearance, as reflected by a 14% and 36% increase in nicotine clearance after 4 and 7 days of smoking abstinence, respectively, compared with overnight abstinence (Benowitz and Jacob, 2000, Lee et al., 1987). Although the active substance from tobacco smoke involved in CYP2A6 inhibition has not been identified, there is evidence suggesting that nicotine may play a role in the inhibition of CYP2A6 through the down-regulation of CYP2A6 mRNA and protein expression. This has been demonstrated in monkeys following 21 days of nicotine treatment (Schoedel et al., 2003). Further, nicotine has been shown to irreversibly bind and inhibit CYP2A6 activity in vitro, which is possibly mediated by the nicotine-Δ1'-(5')-iminium ion, the metabolite formed by CYP2A6 prior to conversion of this intermediate to cotinine by aldehyde oxidase (von Weymarn et al., 2006, von Weymarn et al., 2012).

Components of the diet can act to either induce or inhibit CYP2A6. Hakooz and Hamdan (2007) demonstrated that, at very large quantities (500 mg for 6 days), broccoli consumption is associated with greater metabolism of 1,7-dimethylxanthine to 1,7-dimethylurate (two metabolites of caffeine), which can be used as a surrogate measure of CYP2A6 activity. Conversely, grapefruit juice and menthol are inhibitors of CYP2A6. Specifically, consumption of grapefruit juice, compared to water, is associated with decreased conversion of nicotine to cotinine, lower nonrenal nicotine clearance, and an increase in nicotine and cotinine renal clearance (Hukkanen et al., 2006). Menthol, which is used to add flavour to certain foods, toothpastes, cigarettes, and other products, has been shown to inhibit in vitro nicotine and coumarin metabolism in human liver microsomes (MacDougall et al., 2003). This is supported by in vivo assessments in which the rate of nicotine metabolism to cotinine was slower among individuals who smoked mentholated cigarettes for one week, compared to individuals who smoked nonmentholated cigarettes over the same time period (Benowitz et al., 2004).
Furthermore, CYP2A6 expression is upregulated in several hepatic pathological states, including viral and parasitic infections, cirrhotic liver, and steatotic liver (Satarug et al., 1996, Kirby et al., 1996, Fisher et al., 2009). The mechanisms remain unknown, however in vitro evidence suggests that this may be mediated by high affinity binding of the cytosolic hnRNPA1 (heterogeneous nuclear ribonucleoprotein A1) protein to the 3'-UTR of CYP2A6, stabilizing the mRNA and increasing overall CYP2A6 protein expression (Christian et al., 2004).

1.5.2 Additional Enzymes Involved in Nicotine Metabolism

1.5.2.1 Cytochrome P450 Oxidoreductase

The cytochrome P450 oxidoreductase (POR) enzyme is vital to the catalytic function of all drug-metabolizing cytochrome P450s (CYPs), including CYP2A6. POR is a membrane-bound hepatic enzyme that donates electrons from NADPH to CYPs during their catalytic cycle (Hu et al., 2012). The fundamental role of POR in CYP enzyme function is illustrated by the substantial decrease in CYP activity in hepatic POR-deficient mice (Gu et al., 2003, Henderson et al., 2003). Similar to CYP2A6, the POR enzyme is encoded by a polymorphic gene, which in turn impacts the function of CYP enzymes (Gomes et al., 2009, Lv et al., 2016). Interestingly, the impact of POR genetic variants on different CYP isoforms is not always consistent. For example, the Y181D and A287P POR variants have been associated with complete abolition of CYP3A4 activity in vitro, while CYP2B6 activity was only moderately reduced (Chen et al., 2012c). Variation in POR contributes to variation in CYP2A6 enzyme activity such that CYP2A6 normal metabolizers (i.e. individuals who do not possess any reduce-of-function CYP2A6 genetic variants) who possess the A503V POR variant exhibit faster CYP2A6 activity (measured by the NMR) compared to individuals without this variant (Chenoweth et al., 2014b). This suggests potential interplay between POR and CYP2A6 that warrants further investigation.

1.5.2.2 Aldo-Keto Reductase

Another enzyme that may play a role in variable CYP2A6 enzyme activity is aldo-keto reductase 1D1 (AKR1D1). AKR1D1 is involved in the synthesis of bile acids and the reduction of some steroid hormones, which have been implicated in the transcriptional regulation of
CYPs via activation of nuclear hormone receptors, such as PXR and CAR (Schuetz et al., 2001, Lee et al., 2009, Rizner and Penning, 2014). PXR and CAR have both been shown to regulate the transcription of CYP2A6 (Itoh et al., 2006, Maglich et al., 2003), providing a possible mechanism for an association between AKR1D1 and CYP2A6 expression. Through the generation of a CYP regulatory subnetwork, AKR1D1 was identified as an important regulator of hepatic CYP expression (Yang et al., 2010). Follow up in vitro work using human hepatocytes indicated that the overexpression of AKR1D1 was associated with greater expression of multiple CYPs, including CYP2B6, CYP2C8, CYP2C9, CYP2C19, and CYP3A4 (Chaudhry et al., 2013). Likewise, the knockdown of AKR1D1 expression resulted in decreased CYP expression levels. In the same study, the AKR1D1 3’-UTR SNP rs1872930, which was associated with greater AKR1D1 mRNA expression in human liver donors, was associated with higher CYP2B6, CYP2C8, CYP2C9, CYP2C19, and CYP3A4 mRNA expression and in vitro enzyme activity. CYP2A6 was not included in the array of CYPs that were investigated in this study, therefore the impact of variable AKR1D1 on CYP2A6 expression and function remains to be determined.

1.6 Associations of CYP2A6 and Nicotine Metabolism with Smoking Behaviours and Disease

As discussed in section 1.3.1 of this thesis, nicotine is the primary psychoactive component of cigarette smoke, responsible for eliciting reward from smoking and for eventual withdrawal during smoking abstinence. The duration of nicotine’s action in the CNS, which is primarily mediated by the rate of metabolism and clearance by the CYP2A6 enzyme, is important in the modulation of smoking behaviours. The multiple stages of smoking, and the impact of CYP2A6, will be outlined in this section.

1.6.1 Smoking Initiation and Dependence

The majority of smokers first try smoking and transition to regular smoking during adolescence (O'Loughlin et al., 2014). Smoking initiation is influenced by many factors, including parental or household smoking (Eisenberg and Forster, 2003), friend or peer smoking (Goldade et al., 2012), impulsivity (O'Loughlin et al., 2014), and genetic factors, with twin studies reporting approximately 35 to 60% heritability for smoking initiation (Vink et al., 2005, Broms et al., 2006). One important genetic factor influencing smoking behaviours is CYP2A6 genetic
variation, through effects on nicotine metabolism rates. Adolescents with reduced CYP2A6 activity, inferred from CYP2A6 genotype, are at a greater risk for transitioning to nicotine dependence (Karp et al., 2006, O’Loughlin et al., 2004, Chenoweth et al., 2016, Olfson et al., 2016), compared to adolescents with normal CYP2A6 activity. One potential explanation for these observations is that slower nicotine metabolizers have different initial smoking experiences and prolonged reinforcement from nicotine, as a consequence of a greater duration of nicotine exposure in the CNS, or due to greater accumulation of nicotine in the body. For example, among novice smokers, subjective reactions to an initial smoking attempt were more positive (according to the Pleasurable Smoking Experiences scale) among individuals with CYP2A6 genotype intermediate rates of nicotine metabolism, relative to slow or normal nicotine metabolism (Cannon et al., 2016). However, a separate investigation demonstrated that initial smoking experience does not mediate the relationship between CYP2A6 genotype and progression to nicotine dependence (Audrain-McGovern et al., 2007). Nonetheless, once regular smoking has been acquired, slow nicotine metabolizers smoke less (O’Loughlin et al., 2004, Chenoweth et al., 2016), increase more gradually in their level of nicotine dependence (Audrain-McGovern et al., 2007), and are less likely to be smokers (Schoedel et al., 2004), compared to normal metabolizers. This suggests that although their risk for initiation and transition to dependent smoking is higher, the overall smoking risk among slow metabolizers is lower than among normal metabolizers.

Similar to smoking initiation, nicotine dependence (based on the Fagerström Test for Nicotine Dependence, FTND) is heritable, with approximately 75% of the variation in dependence being attributable to genetic factors (Vink et al., 2005). The FTND consists of six questions that assess (1) time to first cigarette, (2) difficulty in abstaining from smoking, (3) preference for morning cigarette, (4) number of cigarettes smoked per day, (5) time of day when most actively smoking, and (6) smoking when ill (Heatherton et al., 1991). Among adult regular smokers, nicotine dependence is associated with CYP2A6 genotype and the rate of nicotine metabolism, however findings are inconsistent. In several investigations, smokers with CYP2A6 reduce-of-function genotypes or slower CYP2A6 activity (lower NMR quartiles) exhibited lower FTND scores compared to faster nicotine metabolizers (Kubota et al., 2006, Wassenaar et al., 2011, Sofuoglu et al., 2012). Conversely, multiple studies indicated no difference in FTND among different CYP2A6 genotype or NMR groups (Johnstone et al., 2006, Schnoll et al., 2009, Ho et al., 2009b). Because the FTND measure is largely determined
by smoking quantity (CPD), differences may not be detectable in light smoking populations where smoking intensity plays a larger role than the number of CPD (Ho et al., 2009b).

Additionally, gender differences may contribute to the lack of association between FTND and CYP2A6 in some studies. Schnoll and colleagues (2014) demonstrated that a higher NMR was associated with greater FTND scores among men, but not women. This suggests that nicotine dependence may be more strongly mediated by non-nicotinic factors among women than among men, which is supported by studies of nicotine and non-nicotinic mediated reinforcement and reward in men and women (reviewed by Perkins, 2009).

1.6.2 Smoking Quantity

Smokers can regulate their tobacco and thus nicotine intake in order to maintain relatively constant circulating nicotine levels throughout the day through altering smoking topography or the number of cigarettes smoked (McMorrow and Foxx, 1983). This occurs to a different extent among smokers, likely due to differences in cigarette craving, which is negatively associated with levels of nicotine in the blood (Jarvik et al., 2000). Accordingly, variation in the rate of nicotine metabolism and clearance, primarily mediated by CYP2A6 genetic variation, is associated with differences in smoking quantity.

1.6.2.1 Heavy Smokers

CYP2A6 variation is associated with differences in both the number of cigarettes smoked per day and smoking topography (puff volume, duration, and velocity) among heavier smokers (i.e. those who consume ≥10 CPD). Multiple studies have shown that smokers with reduce-of-function CYP2A6 genotypes smoke fewer CPD than those with wild-type genotypes (Rao et al., 2000, Ariyoshi et al., 2002, Fujieda et al., 2004, Wassenaar et al., 2011). Similar relationships have been demonstrated for NMR and CPD (Benowitz et al., 2003). Smokers with the full CYP2A6 gene deletion (*4/*4 genotype) smoke significantly fewer CPD compared to smokers with one or two active functional copies of the gene (Ariyoshi et al., 2002). This relationship also extends beyond the gene deletion. In a meta-analysis of eighteen studies, from multiple ethnic/racial populations, that each assessed the association between CYP2A6 genotype and CPD, reduced metabolizers (individuals with one or more of the CYP2A6 alleles *2, *4, *7, *9, *10, *12, *17, *20, *23, *24, *25, *26, *27, *28, *35) smoked
significantly fewer CPD compared to normal nicotine metabolizers (individuals with the 
*CYP2A6*/*1/*1 genotype) (Pan et al., 2015).

In some cases, expected differences in CPD are not observed across different *CYP2A6* genotype or activity groups. This could be due to CPD being a poor measure of consumption, or could result from differences in smoking intensity; the relationship between smoking quantity and CYP2A6 is often seen when biochemical measures of smoking quantity (plasma cotinine, breath CO) or intensity (CO/cigarette) are used. For example, while smokers possessing the *CYP2A6* gene duplication have been shown to smoke fewer CPD than smokers with just two functional copies of the gene (*CYP2A6*/*1/*1), they smoke each cigarette more intensely, as evidenced by a CO/cigarette ratio double that of wild-type smokers (Rao et al., 2000). Similarly, when assessing the association between NMR quartiles and smoking quantity, only subtle differences in CPD were observed between the top (i.e. fastest) and bottom two quartiles (approximately 22 vs. 19 CPD) (Strasser et al., 2011). However, there was a significant stepwise increase in total puff volume and NNAL (a tobacco-specific carcinogen) exposure with each NMR quartile (Strasser et al., 2011). A similar association was observed between *CYP2A6* genotype group and smoking intensity, measured as mean and total puff volume (Strasser et al., 2007). In cases where titration of nicotine dose via the number of CPD or smoking intensity is not evident, it may be due to lower levels of nicotine dependence among these smokers. Schoedel and colleagues (2004) were able to detect differences in CPD only among smokers who met DSM-IV criteria for tobacco dependence, a finding that was mirrored among nicotine-dependent (based on mFTQ) youth smokers (Audrain-McGovern et al., 2007). Alternatively, determination of smoking quantity may call for the use of a more robust biomarker, such as urinary TNE. Among lighter smokers (<10 CPD), TNE highlights differences in smoking quantity, by *CYP2A6* genotype or NMR, that were not evident when using other measures of consumption, as described in more detail in the following section (Zhu et al., 2013a).

1.6.2.2 Light Smokers

Smokers who consume on average fewer than 10 CPD are described as light smokers. While there is evidence that heavy smokers adjust their nicotine intake through varying both their CPD and smoking topography, only the latter occurs among light smokers. In a study of adult African American light smokers (<10 CPD), self-reported CPD did not differ across *CYP2A6*
genotype groups or NMR quartiles (Ho et al., 2009a). Similarly, among Alaska Native light smokers (<10 CPD), there were no significant differences in CPD or chews per day (smokeless tobacco) based on CYP2A6 activity (slower vs. faster NMR group) (Zhu et al., 2013a). However, urinary TNE levels were significantly lower for both smokers and smokeless tobacco users in the lower NMR group, indicating that slower nicotine metabolizers had lower levels of tobacco consumption than faster nicotine metabolizers. Therefore, these light smokers appear to regulate their smoking quantity and nicotine intake by adjusting their smoking topography (e.g. intensity), as opposed to CPD, which was detectable by the more sensitive smoking biomarker, TNE. Additionally, through the measurement of urinary TNE, Zhu et al. (2013a) showed that, despite smoking on average <10 CPD, Alaska Natives had mean TNE levels (61 nmol/mg creatinine) similar to that of heavy smokers of other ethnic/racial groups (>20 CPD, TNE: 58-87 nmol/mg creatinine) (Le Marchand et al., 2008).

The impact of CYP2A6 on TNE (i.e. nicotine dose) is further demonstrated in a multi-ethnic cohort study of more than 2000 smokers (Park et al., 2016), where the NMR accounted for 8% of the variation in TNE in the combined sample of different ethnic/racial groups (Park et al., 2016). Among African Americans alone, who are typically light smoking populations, the contribution of NMR to TNE was higher, at 12%. Together this suggests that, even when controlling for CPD, there is still variation in smoking quantity (as captured by TNE) that is, in part, mediated by CYP2A6 activity. This study did not report the percent contribution of NMR to variation in TNE without including CPD in the TNE model.

1.6.3 Smoking Cessation

Smokers with different CYP2A6 genotypes and rates of nicotine metabolism vary in their ability to quit smoking with and without cessation aids, such that slow nicotine metabolizers are more likely to quit smoking compared to normal metabolizers. Disparities between nicotine metabolizer groups may result from less severe craving and withdrawal symptoms (Kubota et al., 2006, Sofuoglu et al., 2012), fewer ingrained smoking behaviours due to less smoking (Pan et al., 2015), and/or weaker conditioned responses to smoking cues (Tang et al., 2012, Falcone et al., 2016) among smokers with slower CYP2A6 genotypes or activity phenotypes. Fluctuations in blood nicotine levels during the day are likely less drastic among smokers with slower rates of nicotine metabolism and clearance. This may result in fewer surges in brain
nicotine levels via smoking, ultimately facilitating a lower level of functional connectivity between rewarding brain regions compared to what is observed for smokers who metabolize and clear nicotine rapidly (Li et al., 2017).

1.6.3.1 Pharmacotherapies

Without the use of pharmacological smoking cessation aids, CYP2A6 genotype slow metabolizers have a greater likelihood of spontaneous quitting compared to smokers with normal genotypes (Gu et al., 2000, Chenoweth et al., 2013). The improved cessation success among slower nicotine metabolizers has been further investigated with relation to cessation pharmacotherapies. A widely-used smoking cessation treatment is nicotine replacement therapy (NRT), which comes in the form of nicotine patch, gum, lozenges, nasal spray, and inhaler. Compared to faster nicotine metabolizers (higher NMR), slower metabolizers exhibit higher quit rates following 8 weeks of treatment with nicotine patch (Lerman et al., 2006, Schnoll et al., 2009) (Fig. 8).

Another smoking cessation pharmacotherapy is bupropion, which inhibits dopamine and norepinephrine reuptake and functions as a weak nAChR antagonist (Warner and Shoaib, 2005). Unlike nicotine, bupropion is not metabolized by CYP2A6. In contrast to the NRT clinical trials, a study by Chen and colleagues (2014) showed, as expected, that there was no association between CYP2A6 genotype and quitting using bupropion therapy, with a similar decrease in smoking relapse observed among CYP2A6 genotype slow and fast nicotine metabolizers. In a separate smoking cessation clinical trial, end-of-treatment quit rates decreased with increasing NMR quartiles in the placebo group, whereas quit rates again did not differ according to CYP2A6 activity (NMR quartiles) for subjects taking bupropion; however, bupropion was associated with significantly improved quit rates compared to placebo (34% vs. 10%) in the highest NMR quartile (i.e. fastest nicotine metabolizers) (Patterson et al., 2008) (Fig. 8). No significant differences in quit rates from bupropion versus placebo were observed in the three lower NMR quartiles.

The abovementioned analyses of CYP2A6 and quitting were conducted retrospectively, whereas a recently completed phase 3 clinical trial demonstrated the utility of the NMR as a predictive biomarker of smoking cessation outcome (Lerman et al., 2015). In the Pharmacogenetics of Nicotine Addiction Treatment (PNAT2) clinical trial (NCT0131001),
participants were randomized to treatment groups (placebo, nicotine patch, or varenicline) using prospective stratification based on their pretreatment NMR. Varenicline is a partial agonist, and competes with nicotine binding at α4β2 nAChRs, reducing nicotine-evoked dopamine release, the primary reward mechanism of smoking (Garrison and Dugan, 2009). At end-of-treatment (11 weeks) and six-month follow-up, slower nicotine metabolizers in the PNAT2 trial exhibited no difference in quit rates between nicotine patch or varenicline, whereas faster metabolizers exhibited significantly higher quit rates on varenicline compared to nicotine patch (Lerman et al., 2015) (Fig. 8). Varenicline (versus placebo) associated side effects were also more severe for slow versus faster metabolizers. Collectively, these data suggest that slower nicotine metabolizers may benefit most from NRT, because of a safer side effect profile and lower cost, whereas faster metabolizers may benefit from bupropion or varenicline.

**Figure 8 |** The nicotine metabolite ratio is associated with end-of-treatment quit rates for the following therapies: (A) Nicotine patch (retrospective analysis), (B) Bupropion (retrospective analysis), (C) Varenicline (prospective analysis). The mean NMRs for each quartile in part A are: Q1 0.19, Q2 0.33, Q3 0.45, Q4 0.75. The mean NMRs for each quartile in part B are: Q1 0.17, Q2 0.29, Q3 0.41, Q4 0.90. In part C, participants with NMR<0.31 were classified as slow metabolizers, and those with NMR≥0.31 were fast metabolizers. Adapted from (Lerman et al., 2006, Lerman et al., 2015, Patterson et al., 2008)
1.6.4 Lung Cancer

CYP2A6 likely contributes to lung cancer risk among smokers via both indirect and direct mechanisms. Smoking fewer CPD is associated with lower lung cancer risk (Khuder, 2001), resulting from lower exposure to harmful carcinogens, such as the tobacco-specific nitrosamines N-nitrosonornicotine (NNN) and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) (Hecht, 1998). As slower nicotine metabolism is associated with smoking fewer CPD (Pan et al., 2015), and lower smoking intensity (Strasser et al., 2011), CYP2A6 activity variation may indirectly influence lung cancer risk through its impact on smoking quantity. However, when controlling for cigarette consumption (cigarettes per day and/or cigarette pack-years), CYP2A6 genotype reduced metabolizers, relative to normal metabolizers, continue to exhibit decreased odds of having lung cancer, in both White and African American populations (Wassenaar et al., 2011, Wassenaar et al., 2015). Interestingly, among African American smokers, this was true for male, but not female, smokers (Wassenaar et al., 2015). Additionally, a significant relationship has been demonstrated between functionally significant CYP2A6 genetic variants and lung cancer in a recent GWAS (Transdisciplinary Research in Cancer of the Lung (TRICL) consortium), even when controlling for smoking duration and quantity (Patel et al., 2016). The top genome-wide significant SNPs were located within or near the CYP2A6 gene and remained genome-wide significant after conditioning on known functional CYP2A6 genetic variants. Several of the top SNPs remained significant when adjusting for smoking status and cigarette pack-years. These findings highlight the possibility for a more direct role of CYP2A6 in lung cancer risk.

The CYP2A6 enzyme metabolically activates procarcinogenic tobacco-specific nitrosamines NNN and NNK via α-hydroxylation (Wong et al., 2005, Kushida et al., 2000). Therefore, smokers with slower CYP2A6 activity will have less activation of tobacco-specific nitrosamines, decreasing exposure to these activated lung carcinogens. This notion is supported by the measurement of higher levels of NNN, the parent pro-carcinogen, suggesting lower levels of NNN bioactivation, among smokers who were CYP2A6 genotype and phenotype reduced compared to normal metabolizers, even when controlling for smoking quantity (urinary TNE) (Zhu et al., 2013a). Thus, being a slower CYP2A6 metabolizer can reduce both tobacco consumption and procarcinogen activation, resulting in lower lung cancer risk.
This may suggest that women exhibit greater lung cancer risk than men, considering their faster CYP2A6 activity (described previously in section 1.5.1.4). However, lung cancer risk is greater in males than females (Lewis et al., 2014). The impact of gender on CYP2A6 activity is modest (Chenoweth et al., 2014a) and there are many additional risk and protective factors for lung cancer, including gender differences in smoking prevalence and quantity, which are greater among men versus women (Ng et al., 2014, Jamal et al., 2016). For example, fast (compared to slow) nicotine metabolizers, according to CYP2A6 genotype, exhibit ORs of 1.26-1.60 for lung cancer risk, whereas smoking 1-19 CPD and >20 CPD (compared to not smoking) are associated with ORs of 3-4 and 6-18, respectively (Wassenaar et al., 2011, Khuder, 2001).
1.7 STATEMENT OF RESEARCH HYPOTHESES

In Chapter 1, “Relationships between smoking behaviors and cotinine levels among two American Indian populations with distinct smoking patterns”, we hypothesized that cotinine levels would be positively correlated with the number of cigarettes smoked per day among smokers in both the Northern Plains (NP) and Southwest (SW) American Indian tribal populations. We further hypothesized that cotinine levels would be higher among smokers versus non-smokers in both tribes, and that the NP smokers would have higher cotinine levels than the SW smokers. We speculated that, in both tribes, cotinine levels would not be associated with traditional tobacco use, but would be associated with self-report measures of secondhand tobacco exposure, including smoking in the home and having friends who smoke.

In Chapter 2, “CYP2A6 and nicotine metabolism among two American Indian tribal groups differing in smoking patterns and risk for tobacco-related cancer”, we hypothesized that the NP tribe would have a lower frequency of CYP2A6 genotype reduced metabolizers than the SW tribe, and that CYP2A6 genotype would be associated with the rate of nicotine metabolism (nicotine metabolite ratio, NMR) in smokers from both tribes. We also hypothesized that NP smokers would have a faster rate of nicotine metabolism (higher NMR) compared to SW smokers.

In Chapter 3, “Predictors of variation in CYP2A6 mRNA, protein, and enzyme activity in a human liver bank: influence of genetic and non-genetic factors”, we hypothesized that all three in vitro CYP2A6 phenotype measures (mRNA, protein, activity) would be correlated, select CYP2A6 genotypes would be associated with variation in CYP2A6 mRNA expression and protein levels, and CYP2A6 genotype would be associated with CYP2A6 enzyme activity. We also hypothesized that the following non-genetic factors would be associated with and account for additional variation in CYP2A6 phenotypes: gender, age, liver disease, AKR1D1 mRNA, and POR protein.

In Chapter 4, “Novel CYP2A6 diplotypes identified through next-generation sequencing are associated with in vitro and in vivo nicotine metabolism”, we hypothesized that novel genetic variants identified at the CYP2A6 gene locus would be associated with variation in in vitro CYP2A6 phenotypes (mRNA, protein, activity). We further postulated that CYP2A6
diplotypes, comprised of a combination of novel SNPs, would be associated with \textit{in vitro} and \textit{in vivo} CYP2A6 enzyme activity.
2 CHAPTER 1: RELATIONSHIPS BETWEEN SMOKING BEHAVIORS AND COTININE LEVELS AMONG TWO AMERICAN INDIAN POPULATIONS WITH DISTINCT SMOKING PATTERNS

Julie-Anne Tanner, Jeffrey A. Henderson, Dedra Buchwald, Barbara V. Howard, Patricia Nez Henderson, and Rachel F. Tyndale

This is a pre-copyedited, author-produced version of an article accepted for publication in Nicotine & Tobacco Research following peer review. The version of record (Tanner et al. Relationships between smoking behaviors and cotinine levels among two American Indian populations with distinct smoking patterns. Nicotine & Tobacco Research, 2017) is available online at: https://doi.org/10.1093/ntr/ntx114.

JAH, DB, BVH, and RFT designed the overall study. JAH, DB, and BVH recruited participants. JAT conducted exploratory analyses of the biomarker data (cotinine levels in the tribes), which indicated high levels of cotinine among non-smokers. Based on this, JAT determined experimental questions and designed a study to compare cotinine levels between the tribes and assess determinants of cotinine levels in each tribe. JAT performed all statistical analyses and modeling. All authors interpreted the results. JAT and RFT wrote the manuscript.
2.1 Abstract

**Introduction:** Smoking prevalence, cigarettes per day (CPD), and lung cancer incidence differ between Northern Plains (NP) and Southwest (SW) American Indian populations. We used cotinine as a biomarker of tobacco smoke exposure to biochemically characterize NP and SW smokers and non-smokers, and to investigate factors associated with variation in tobacco exposure.

**Methods:** American Indians (N=636) were recruited from two different tribal populations (NP and SW) as part of a study conducted as part of the Collaborative to Improve Native Cancer Outcomes P50 project. For each participant, a questionnaire assessed smoking status, CPD, secondhand smoke exposure, and traditional ceremonial tobacco use; plasma and/or salivary cotinine was measured.

**Results:** Cotinine levels were (mean ±95% CI) 81.6 ± 14.1 and 21.3 ± 7.3 ng/ml among NP smokers and non-smokers, respectively, and 44.8 ± 14.4 and 9.8 ± 5.8 ng/ml among SW smokers and non-smokers, respectively. Cotinine levels correlated with CPD in both populations (P<0.0001). Cotinine >15 ng/ml was measured in 73.4% of NP smokers and 47.8% of SW smokers, and in 19.0% of NP non-smokers and 10.9% of SW non-smokers. Ceremonial traditional tobacco use was associated with higher cotinine among NP smokers only (P=0.004). Secondhand smoke exposure was associated with higher cotinine among NP non-smokers (P<0.02). More secondhand smoke exposure was associated with smoking more CPD in both populations (P=0.03-0.29). Linear regression modeling mirrored these findings.

**Conclusions:** High prevalence of smoking in the Northern Plains and high cotinine levels among non-smokers in both regions highlights the tribal populations’ risk for tobacco-related disease.

**Implications:** There is a high prevalence of smoking in Northern Plains American Indians. Among Northern Plains and Southwest non-smokers, relatively high cotinine levels, representative of high tobacco exposure, suggest considerable exposure to secondhand smoke. It is critical to highlight the extent of secondhand smoke exposure among the Northern Plains and Southwest American Indians, and to enhance efforts to initiate smoke-free policies in tribal communities, which are not subject to state-level polices.


2.2 Introduction

Commercial tobacco use, the leading cause of preventable disease and death in the United States (U.S. Department of Health and Human Services), is more prevalent in adult American Indians and Alaska Natives (AI/ANs) than in other U.S. populations. For cigarette smoking, prevalence and patterns of use vary substantially across tribal populations. The Northern Plains (NP) tribal members of South Dakota have a smoking prevalence of ~50% compared to ~14% among Southwest (SW) tribal members in Arizona (Nez Henderson et al., 2005). NP tribal members who smoke consume an average of 13 cigarettes per day (CPD), compared to 7 CPD in SW tribal populations (Nez Henderson et al., 2005, Eichner et al., 2010). As smoking is a risk factor for many health conditions, including lung cancer and cardiovascular disease, the distinctive rates of smoking reported for these two regions are consistent with tribal members’ different rates of lung cancer incidence (104/100,000 in NP vs. 15/100,000 in SW) and mortality (94/100,000 in NP vs. 14/100,000 in SW) (Bliss et al., 2008, Plescia et al., 2014). Despite these smoking patterns, most AI/AN adults report a desire to quit smoking, and have made quit attempts (Jamal et al., 2014, Gohdes et al., 2002, Forster et al., 2007).

AI/AN tribes differ in their consumption of commercial tobacco and use of traditional tobacco for spiritual and ceremonial purposes (Nez Henderson et al., 2005, Linton, 1924, Cutler, 2002a). Ceremonial traditional tobacco generally does not contain nicotine, the major psychoactive component of tobacco (Margalit et al., 2013). However, nicotine-containing commercial tobacco can be used in combination with traditional tobacco, and this practice varies across tribes and age groups (Forster et al., 2008, Nez Henderson et al., 2009, Unger et al., 2006). The shift toward use of commercial tobacco for ceremonial purposes resulted, in part, from convenience of obtaining commercial compared to traditional tobacco, and in response to commercial cigarette advertisements; for further information please see the Black Hills Center for American Indian Health videos (e.g., https://www.youtube.com/watch?v=l_gYL_oJxhg&spfreload=5, last accessed May 29 2017).

Secondhand smoke inhalation, another potential source of nicotine and carcinogen exposure, can increase lung cancer risk by 20%-30% in non-smokers (Eriksen et al., 1988). Smoke-free policies have been implemented in public places throughout the U.S. (Centers for Disease and Prevention, 2011), and as a result, rates of exposure to secondhand smoke and associated
chemical additives have fallen substantially in most U.S. communities (Jensen et al., 2010). However, secondhand exposure in tribal populations is poorly understood; studies of the effectiveness of smoke-free policies have not typically included residents of AI/AN reservations (Gonzalez et al., 2013, Sanders-Jackson et al., 2013). As sovereign nations, AI/AN tribes are not legally subject to state-level smoke-free policies (Nez Henderson et al., 2009). Therefore, it is critical to establish the extent of secondhand smoke exposure among understudied NP and SW American Indians, particularly for efforts to initiate smoke-free policies in tribal communities.

American Indian populations in the NP and SW likely experience differential exposure to tobacco smoke through different sources, both actively and passively. This variation might affect their risk for tobacco-related disease. One way to quantify tobacco smoke exposure and lung cancer risk is to measure levels of cotinine, the primary metabolite of nicotine (Nakajima et al., 1996b), in smokers and non-smokers. The utility of cotinine as a biomarker of tobacco smoke exposure derives from its relatively long half-life of ~16 hours (Benowitz and Jacob, 1994). Plasma cotinine levels are robustly associated with self-reported CPD and lung cancer risk among active smokers (Boffetta et al., 2006, Connor Gorber et al., 2009). Cotinine levels above a given threshold have also been used to differentiate active smoking from secondhand tobacco smoke exposure; potential cutpoints range from 1-15 ng/ml of plasma cotinine (Jarvis et al., 1987, Benowitz et al., 2009). Although cotinine >15 ng/ml has been used in the past as a cutpoint for active smoking (Jarvis et al., 1987), recent work advocates a more stringent threshold of 3 ng/ml (Benowitz et al., 2009). Nonetheless, the higher threshold is likely more relevant for assessing passive exposure in AI/AN communities, given the absence of smoke-free policies on most reservations. For the present investigation, we elected to assess the proportion of smokers and non-smokers with cotinine >15 and 3 ng/ml.

No previous studies in NP or SW tribal communities have used biomarkers to investigate tobacco smoke exposure. Thus, the relationship of cotinine with smoking levels, as well as other potential sources of variability in tobacco smoke exposure, has never been explored in these communities. The aim of the present study was to assess cotinine levels and investigate factors associated with variation in cotinine among NP and SW smokers and non-smokers, using survey measures of exposure. In particular, this study assessed commercial cigarette
consumption, use of ceremonial traditional tobacco products, and potential sources of passive
smoke exposure.

2.3 Methods

Study Design

The data presented here were collected for a cross-sectional study entitled “Topography and
Genetics of Smoking and Nicotine Dependence in American Indians,” one of five major
research studies conducted by the Collaborative to Improve Native Cancer Outcomes P50
program project. To protect the confidentiality of the participating communities, geographic
descriptors, consistent with previous publications and approval by tribal review boards, were
assigned instead of tribal names. Participating tribal populations in each region (NP versus
SW) were culturally and linguistically unrelated to those in the other region, and had
substantially different historical experiences. The SW group is urban, while the NP group
resides predominantly on rural reservations. NP participants were recruited from a random
subset of participants in an earlier study of community health (Slattery et al., 2007). The
original study from which participants were re-sampled represented approximately one third of
all adults in the participating reservations and communities (Duncan et al., 2009). SW
participants were recruited by using respondent driven sampling among American Indian
friends and family members of tribal participants in an earlier randomized clinical trial in the
greater Phoenix metropolitan area (Howard et al., 2008); both NP and SW subsamples were
stratified by sex and smoking status. The data presented stem from a secondary analysis of
cotinine levels in the NP and SW tribal populations, therefore a priori power analyses were not
performed to calculate the minimum sample size required in each population.

Data Collection

From 2012-2014, all participants completed a questionnaire either via an interviewer or self-
administered, which assessed, in part, demographics, tobacco use, nicotine dependence, and
social influences on smoking. Biological samples (blood or saliva) were also collected from all
participants. The majority of the participants provided a blood sample for the cotinine analysis
(97%), with the remaining 3% providing a saliva sample. The individuals who provided saliva
samples were from both tribal populations (N=12 from NP, N=9 from SW). There is generally
strong agreement between saliva and plasma/blood measures of cotinine (St Helen et al., 2012), and the same cotinine cutpoint is suggested to differentiate active versus secondhand smoke exposure in saliva and plasma (SRNT Subcommittee on Biochemical Verification, 2002). The final cohort comprised 636 American Indians aged 20-88 years, with 426 in the NP and 210 in the SW. Ethical approval for all study procedures was obtained from the review boards of the Great Plains Indian Health Service, the University of Toronto, the University of Washington, MedStar Health Research Institute, and appropriate tribal entities. All participants provided informed consent.

**Measures**

Smoking status, CPD, ceremonial traditional tobacco use, and two measures of secondhand smoke exposure were assessed, as outlined in Table 3. For smoking status, former and never smokers were aggregated in a single non-smoking group, within each tribal population, distinguishing them from participants who reported current cigarette smoking. To determine the amount of current active tobacco consumption, participants reported either CPD or cigarettes per month (CPMo). Data on CPMo were divided by 30 for consistency with CPD data. Within each tribal population, and within smoking status groups, participants who formerly or never used ceremonial traditional tobacco were also aggregated in a single non-traditional tobacco user group, being separate from participants who reported current traditional tobacco use. Our two measures of secondhand smoke exposure were 1) allowing smoking in the home and 2) having friends who smoke. Each measure of secondhand smoke was analyzed independently. Either plasma or salivary cotinine was measured in biological samples collected from all participants by using liquid chromatography tandem mass spectrometry; analytic methods have been previously described (Tanner et al., 2015b).

**Statistical Analyses**

As these are two distinct American Indian populations and smoking patterns differ between the NP and SW, we have chosen to analyze each tribal population separately, allowing us to determine independent relationships between cotinine levels and CPD, ceremonial traditional tobacco use, and second hand tobacco smoke exposure in each tribal population. Cotinine levels and CPD were non-normally distributed, indicating the use of nonparametric statistical tests. The correlations of cotinine levels and CPD within each tribal population were
### Table 3 | Questionnaire assessments and participant classification.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Survey Question(s)</th>
<th>Responses</th>
<th>Classification</th>
</tr>
</thead>
</table>
| Smoking Status                  | “Have you smoked at least 100 cigarettes in your **entire life**?”
                                  |                       | Yes/No                | Yes/Yes=Smoker
                                  | “Have you smoked at least 1 cigarette in the last month?”
                                  |                       | Yes/No                | Yes/No=Former Smoker
                                  |                                                                        |                       | No/No=Never Smoker               |
| Cigarettes Per Day              | “On average, about how many cigarettes do you smoke?”
                                  | “I smoke __ cigarettes per DAY” or “I don’t smoke daily. I only smoke __ cigarettes per MONTH” |                       | N/A                              |
| Ceremonial Traditional Tobacco Use | “Have you ever used traditional Native forms of tobacco (not commercial, such as canšaša or red willow bark, kinnikinnick, are forms of traditional Native tobacco not processed with chemicals)?” | Yes, currently        | Traditional Tobacco User        |
                                  |                                                                        | Yes, but not now      | Former Traditional Tobacco User |
                                  |                                                                        | No                    | Never Traditional Tobacco User  |
| Passive Smoke Exposure          | “Are your home’s residents or visitors allowed to smoke in your home?”
                                  | “Of your closest three friends, how many of them smoke?”
                                  | Yes/No                | Yes=Passive Exposure
                                  |                                                                        |                       | No=Non-Exposed                   |
                                  |                                                                        | 0, 1, 2, or 3         | 1, 2, or 3=Passive Exposure
                                                            |                       | 0=Non-Exposed                   |

Determined by Spearman correlations. Chi-squared tests were used to determine differences between the proportion of smokers and non-smokers who had mean cotinine levels equal to or higher than our predetermined cutpoints, 3 and 15 ng/ml, both of which have been used to differentiate secondhand from active smoke exposure. We used Mann-Whitney and Kruskal-Wallis tests to analyze the association of smoking status, traditional tobacco use, and our two measures of passive smoke exposure with cotinine levels and CPD in each smoking status group. We ran separate linear regression models for NP smokers, NP non-smokers, SW smokers, and SW non-smokers to calculate the percentage of variation in cotinine levels attributable to each variable in each group. Model variables included CPD (except for non-smokers), traditional tobacco use, smoking in the home, and number of friends who smoked. Cotinine was the output variable. Analyses were conducted with GraphPad Prism (v6.0) and SPSS (v22), and statistical tests were considered significant for $P < 0.05$. 

62
2.4 Results

Cotinine levels in the Northern Plains

NP smokers had cotinine levels (mean±95% CI) of 81.6±14.1 ng/ml, and NP non-smokers had cotinine levels of 21.3±7.3 ng/ml (Fig. 9). The latter value exceeds both cutpoints (3 and 15 ng/ml) used for secondhand smoke exposure (Jarvis et al., 1987, SRNT Subcommittee on Biochemical Verification, 2002). Previous studies indicated that NP smokers consumed an average of 13 CPD (Eichner et al., 2010). However, smokers in the present study reported consuming ~7 CPD, consistent with the relatively low cotinine levels observed among the smokers. This measure was positively correlated with cotinine levels among smokers (n=130, Spearman r=0.50, P<0.0001; Fig. 10). This result is consistent with previous findings in other racial and ethnic groups, including Caucasians (St Helen et al., 2013a) (Fig. 10). CPD accounted for 19.01% of the variation in cotinine levels in our linear regression model (Table 4). Only 84% and 73% of NP smokers had cotinine levels ≥ 3 and 15 ng/ml, respectively, whereas a relatively high proportion of NP non-smokers (28% and 19%) had cotinine levels ≥ 3 and 15 ng/ml (Fig. 11). Thus, NP non-smokers had substantial exposure to tobacco smoke, and NP smokers smoked at relatively low levels.
Figure 9 | Association between self-reported smoking status and cotinine levels (ng/ml) among Northern Plains and Southwest smokers and non-smokers. *P* values are based on Mann-Whitney tests. The number of participants included in each analysis was determined by available data. Two Northern Plains participants were excluded because they had no data on cotinine levels.
Figure 10 | Association of cotinine levels (ng/ml) and number of cigarettes smoked per day among Northern Plains (NP), Southwest (SW), and Caucasian (C) smokers. Inset graphs depict the correlation and linear regression of cotinine and cigarettes smoked per day among NP and SW smokers. *P* and Spearman r values are based on Spearman correlation tests or Mann-Whitney tests. The number of participants included in each analysis was determined by available data. Nine NP smokers were excluded because they did not report number of cigarettes smoked per day. All error bars represent 95% confidence intervals (CI) except those referring to cigarettes smoked per day by Caucasians; these error bars represent the interquartile range. Caucasian data were taken from St Helen et al. (2013a).
Figure 11 | Proportion of Northern Plains and Southwest smokers and non-smokers who had cotinine levels equal to or higher than 15 ng/ml. Using 3 ng/ml as the cut point, these values were 84.2% and 28.2% in Northern Plains smokers and non-smokers ($P<0.0001$), and 74.4% and 27.7% in Southwest smokers and non-smokers ($P<0.0001$), respectively. $P$ values are based on Chi-squared tests. The number of participants included in each analysis was determined by available data. Two Northern Plains participants were excluded because they had no data on cotinine levels.
## Table 4 | Linear regression analyses of cotinine levels (ng/ml) among Northern Plains and Southwestern smokers and non-smokers.

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>95% confidence interval for B</th>
<th>Beta</th>
<th>Variation in cotinine attributable to each variable a</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Northern Plains Smokers</strong> (n=129 included in model) b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cigarettes per day</td>
<td>5.61</td>
<td>3.66 to 7.56</td>
<td>0.45</td>
<td>19.01%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ceremonial traditional tobacco use</td>
<td>38.51</td>
<td>4.70 to 72.31</td>
<td>0.18</td>
<td>2.99%</td>
<td>0.03</td>
</tr>
<tr>
<td>Residents and/or visitors smoke at home</td>
<td>9.88</td>
<td>-16.41 to 36.17</td>
<td>0.06</td>
<td>0.32%</td>
<td>0.46</td>
</tr>
<tr>
<td>How many of 3 closest friends smoke (0 vs. 1-3)</td>
<td>30.01</td>
<td>-116.89 to 176.90</td>
<td>0.03</td>
<td>0.10%</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>Northern Plains Non-Smokers</strong> (n=282 included in model) c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ceremonial traditional tobacco use</td>
<td>-2.04</td>
<td>-24.39 to 20.31</td>
<td>-0.01</td>
<td>0.01%</td>
<td>0.86</td>
</tr>
<tr>
<td>Residents and/or visitors smoke at home</td>
<td>4.60</td>
<td>-13.12 to 22.32</td>
<td>0.03</td>
<td>0.09%</td>
<td>0.61</td>
</tr>
<tr>
<td>How many of 3 closest friends smoke (0 vs. 1-3)</td>
<td>22.72</td>
<td>3.95 to 41.48</td>
<td>0.14</td>
<td>1.99%</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Southwestern Smokers</strong> (n=90 included in model) d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cigarettes per day</td>
<td>5.83</td>
<td>3.55 to 8.11</td>
<td>0.49</td>
<td>22.00%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ceremonial traditional tobacco use</td>
<td>-21.17</td>
<td>-50.29 to 7.95</td>
<td>-0.14</td>
<td>1.77%</td>
<td>0.15</td>
</tr>
<tr>
<td>Residents and/or visitors smoke at home</td>
<td>9.63</td>
<td>-21.65 to 40.93</td>
<td>0.06</td>
<td>0.32%</td>
<td>0.54</td>
</tr>
<tr>
<td>How many of 3 closest friends smoke (0 vs. 1-3)</td>
<td>17.97</td>
<td>-21.70 to 57.63</td>
<td>0.09</td>
<td>0.69%</td>
<td>0.37</td>
</tr>
<tr>
<td><strong>Southwestern Non-Smokers</strong> (n=120 included in model) e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ceremonial traditional tobacco use</td>
<td>-1.02</td>
<td>-17.80 to 15.76</td>
<td>-0.01</td>
<td>0.01%</td>
<td>0.90</td>
</tr>
<tr>
<td>Residents and/or visitors smoke at home</td>
<td>-3.40</td>
<td>-14.64 to 7.85</td>
<td>-0.06</td>
<td>0.30%</td>
<td>0.55</td>
</tr>
<tr>
<td>How many of 3 closest friends smoke (0 vs. 1-3)</td>
<td>-3.34</td>
<td>-15.17 to 8.49</td>
<td>-0.05</td>
<td>0.27%</td>
<td>0.58</td>
</tr>
</tbody>
</table>

a. The variation in cotinine levels attributable to each variable is determined by: (Part Correlation)² x 100
b. NP smokers: R²=0.272, P<0.001. R² indicates the proportion of variance in cotinine levels (27.2%) explained by this model.
c. NP non-smokers: R²=0.022, P>0.05. R² indicates the proportion of variance in cotinine levels (2.1%) explained by this model.
d. SW smokers: R²=0.275, P<0.001. R² indicates the proportion of variance in cotinine levels (27.5%) explained by this model.
e. SW non-smokers: R²=0.006, P>0.05. R² indicates the proportion of variance in cotinine levels (0.6%) explained by this model.
Association between cotinine levels and ceremonial traditional tobacco use in the Northern Plains

We found no association between ceremonial traditional tobacco use and cotinine levels among NP non-smokers ($P=0.85; \textbf{Fig. 12a}$), suggesting that traditional tobacco was not a source of nicotine exposure. Nonetheless, cotinine levels were higher among NP smokers who used traditional tobacco than those who did not ($P=0.004; \textbf{Fig. 12a}$). This finding might be explained by the fact that NP smokers who used traditional tobacco also smoked more commercial CPD than NP smokers who did not ($P=0.13; \textbf{Fig. 12b}$).

In linear regression models, CPD ($P<0.001$) and traditional tobacco use ($P=0.03$) were significant independent predictors of cotinine levels among NP smokers (\textbf{Table 4}). However, a similar linear regression model indicated that traditional tobacco use was not a significant predictor of cotinine levels among NP non-smokers ($P=0.77; \textbf{Table 4}$).
Figure 12 | (A) Association between cotinine levels (ng/ml) and ceremonial traditional tobacco use among Northern Plains (NP) smokers and non-smokers. (B) Association between the number of cigarettes smoked per day and traditional tobacco use among NP smokers. (C) Association between cotinine levels and traditional tobacco use among Southwest (SW) smokers and non-smokers. (D) Association between the number of cigarettes smoked per day and traditional tobacco use among SW smokers. All $P$ values are based on Mann-Whitney tests. The number of participants included in each analysis was determined by available data. Two NP participants had no data on cotinine levels, nine NP smokers did not report number of cigarettes smoked per day, and two NP smokers had no data on traditional tobacco use, so all were excluded from this analysis.
Association between cotinine levels and passive smoke exposure in the Northern Plains

Table 5 summarizes tribal measures of secondhand smoke exposure. We investigated the possibility that secondhand smoke exposure influenced cotinine levels in the NP population sample by using two items in the participant survey: “Are your home’s residents or visitors allowed to smoke in your home?” and “Of your closest three friends, how many of them smoke?” NP non-smokers who allowed smoking in the home and had close friends who smoked had higher cotinine levels than those who did not ($P=0.01$ and $P<0.0001$, respectively; Fig. 13a and 14a). Similar relationships between cotinine levels and secondhand smoke exposure were observed among NP smokers, although the differences were not significant ($P=0.39$; Fig. 13a and 14a). NP smokers who allowed smoking in the home and had friends who smoked also consumed more CPD than those who did not ($P=0.07$ and $P=0.06$, respectively; Fig. 13b and 14b). Linear regression modeling suggested that, among NP non-smokers, having friends who smoked was a significant independent predictor of elevated cotinine levels ($P=0.02$; Table 4). However, neither indicator of passive smoke exposure had a significant effect on cotinine levels among NP smokers ($P>0.45$; Table 4).

Table 5 | Characteristics of study populations in the Northern Plains (South Dakota) and the Southwest (Arizona)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Northern Plains (n=426)</th>
<th>Southwest (n=210)</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Male</td>
<td>47.2 (42.5-51.9)</td>
<td>56.7 (49.9-63.2)</td>
<td>0.02 $^a$</td>
</tr>
<tr>
<td>Mean age, years</td>
<td>49.9 (48.5-51.3)</td>
<td>43.3 (41.5-45.0)</td>
<td>$&lt;0.0001$ $^b$</td>
</tr>
<tr>
<td>% Smokers</td>
<td>32.6 (28.4-37.2)</td>
<td>42.9 (36.4-49.6)</td>
<td>0.01 $^a$</td>
</tr>
<tr>
<td>Mean cigarettes per day $^c$</td>
<td>7.2 (6.1-8.4)</td>
<td>3.7 (2.5-4.9)</td>
<td>$&lt;0.0001$ $^b$</td>
</tr>
<tr>
<td>% Traditional tobacco users $^d$</td>
<td>14.0 (11.0-17.6)</td>
<td>19.5 (14.6-25.4)</td>
<td>0.07 $^a$</td>
</tr>
<tr>
<td>% Allowing smoking in the home</td>
<td>29.6 (25.4-34.1)</td>
<td>16.2 (11.8-21.8)</td>
<td>0.0003 $^a$</td>
</tr>
<tr>
<td>% Having ≥1 friend who smokes $^e$</td>
<td>86.8 (83.2-89.7)</td>
<td>67.1 (58.4-74.9)</td>
<td>$&lt;0.0001$ $^a$</td>
</tr>
</tbody>
</table>

$^a$ $P$ value based on Chi-squared test
$^b$ $P$ value based on Mann-Whitney test
$^c$ Nine NP participants did not provide information on cigarettes per day.
$^d$ Three NP participants did not provide information on traditional tobacco use.
$^e$ Two NP participants did not provide information on the number of friends who smoked.
Figure 13 | (A) Association between cotinine levels (ng/ml) and allowing smoking in the home among Northern Plains (NP) smokers and non-smokers. (B) Association between the number of cigarettes smoked per day and allowing smoking in the home among NP smokers. (C) Association between cotinine levels and allowing smoking in the home among Southwest (SW) smokers and non-smokers. (D) Association between number of cigarettes per day and allowing smoking in the home among SW smokers. All $P$ values are based on Mann-Whitney tests. The number of participants included in each analysis was determined by available data. Two NP participants had no data on cotinine levels and nine NP smokers did not report number of cigarettes smoked per day, so all were excluded from this analysis.
Figure 14 | (A) Association between cotinine levels (ng/ml) and the number of close friends who smoke among Northern Plains (NP) smokers and non-smokers. P value is based on the Mann-Whitney test. (B) Association between the number of cigarettes smoked per day and the number of close friends who smoke among NP smokers. P value is based on the Kruskal-Wallis test. (C) Association between cotinine levels and the number of close friends who smoke among Southwest (SW) smokers and non-smokers. P values are based on the Mann-Whitney test. (D) Association between the number of cigarettes smoked per day and the number of close friends who smoke among SW smokers. P value is based on the Kruskal-Wallis test. The number of participants included in each analysis was determined by available data. Two NP participants had no data on cotinine levels, nine NP smokers did not report number of cigarettes smoked per day, and two NP smokers had no data on number of friends who smoke, so all were excluded from this analysis.
Cotinine levels in the Southwest

Findings for the SW sample were similar to those for the NP. SW smokers had relatively low cotinine levels (mean±95% CI) of 44.8±14.4 ng/ml, yet SW non-smokers had mean cotinine levels above the more stringent cutpoint (3 ng/ml) for a non-actively-smoking population (9.8±5.8 ng/ml; Fig. 9). Although previous studies have reported that SW smokers consumed an average of 7 CPD (Eichner et al., 2010), SW smokers in the present study reported consuming ~4 CPD, consistent with the low cotinine levels observed. CPD was positively correlated with cotinine levels among smokers (n=90, Spearman r=0.56, P<0.0001; Fig. 10). Again, the relationship between cotinine and CPD was consistent with findings for other racial and ethnic groups (St Helen et al., 2013a) (Fig. 10). CPD accounted for 22.0% of the variation in cotinine levels in our linear regression model (Table 4). Approximately 74% and 48% of SW smokers had cotinine levels ≥3 and 15 ng/ml, respectively, while 28% and 11% of SW non-smokers had cotinine levels ≥3 and 15 ng/ml, respectively (Fig. 11).

Association between cotinine levels and ceremonial traditional tobacco use in the Southwest

In the SW sample, we found no association between cotinine levels and ceremonial traditional tobacco use among smokers or non-smokers (P=0.68 and P=0.53, respectively; Fig. 12c). Unlike NP smokers, SW smokers who used traditional tobacco did not smoke more CPD than those who did not (P=0.44; Fig. 12d). Linear regression modeling for SW smokers suggested that CPD was a significant independent predictor of cotinine levels (P<0.001; Table 4), whereas for SW smokers and non-smokers, traditional tobacco use was not a predictor of cotinine levels (P=0.21 and P=0.88, respectively; Table 4).

Association between cotinine levels and passive smoke exposure in the Southwest

Compared to household smoking in the general U.S. population (17% in all households, 9% in households with no adult smokers, 54% in households with ≥1 adult smoker) (King et al., 2016), NP household smoking consisted of 30% in all households, 22% among non-smokers, and 45% among smokers, and in the SW smoking in the home occurred in 16% of all households, 12% among non-smokers, and 29% among smokers. We found no significant association between indicators of secondhand smoke exposure and cotinine levels among SW smokers or non-smokers (smokers, P=0.40 and P=0.20, respectively; non-smokers, P=0.57 and
$P=0.13$, respectively; **Fig. 13c and 14c**). SW smokers who allowed smoking in the home and had friends who smoked also smoked more CPD than those who did not ($P=0.03$ and $P=0.29$, respectively; **Fig. 13d and 14d**). Measures of passive smoke exposure were not significant independent predictors of cotinine levels among smokers or non-smokers, as indicated by linear regression analysis ($P>0.39$; **Table 4**).
2.5 Discussion

Tribal cotinine levels, cigarette consumption, and smoking prevalence

We report several key findings related to tobacco smoke exposure among American Indian populations. NP and SW smokers exhibited relatively low average cotinine levels of 81.6 and 44.8 ng/ml, respectively, compared to 165-180 ng/ml among U.S. smokers of Caucasian, African American, and Alaska Native descent (St Helen et al., 2013a, Strasser et al., 2011, Zhu et al., 2013a). Among smokers in both populations, cotinine levels were correlated with CPD (NP r=0.50, SW r=0.56), and CPD accounted for approximately the same degree of variation in cotinine levels (NP 19.01%, SW 22.00%), confirming previous reports of the association between cotinine levels and self-reported CPD (Connor Gorber et al., 2009). The high prevalence of smoking in the NP tribal population (50%) (Nez Henderson et al., 2005), and the association between CPD and cotinine levels, suggest widespread exposure to nicotine and carcinogens across the NP population. For a non-smoking population, NP non-smokers had very high cotinine levels (21.3 ng/ml), with 28% at levels above the more stringent cutpoint (3 ng/ml), and 19% above the more modest cutpoint (15 ng/ml). In contrast, smoking prevalence was considerably lower in the SW tribal population, and SW non-smokers had lower average cotinine levels (9.8 ng/ml). Nevertheless, this value still exceeded the more stringent cutpoint for secondhand smoke exposure.

Although CPD and cotinine levels were relatively low among smokers in both tribal populations, lung cancer incidence remains high in the NP (Bliss et al., 2008, Plescia et al., 2014). This incidence is likely related to the high prevalence of smoking in this region (Nez Henderson et al., 2005), as well as to genetic factors (Wassenaar et al., 2011). Additionally, even light smokers are at elevated risk for lung cancer (Khuder, 2001), as are non-smokers exposed to secondhand smoke (Eriksen et al., 1988). Therefore, we suggest that a combination of risk factors contributes to the elevated lung cancer incidence and mortality rates reported for the NP (Bliss et al., 2008, Plescia et al., 2014).

Traditional tobacco use

We found no association between ceremonial traditional tobacco use and variation in cotinine levels among SW smokers, SW non-smokers, or NP non-smokers. While a similar proportion
of participants in each regional sample used traditional tobacco (NP 14%, SW 20%), such use is common only in the NP (Nez Henderson et al., 2005). Our null result supports general knowledge of genuine ceremonial traditional tobacco, which does not contain nicotine and thus does not elevate cotinine levels. However, among NP smokers, traditional tobacco users had higher cotinine levels and reported smoking more CPD than non-users, suggesting that elevated cotinine levels among NP smokers reflect more consumption of commercial cigarettes. Yet we also found that both CPD and traditional tobacco use independently contributed to variation in cotinine levels, indicating that higher commercial tobacco consumption does not fully account for elevated cotinine levels among NP smokers. Therefore NP smokers may consume commercial and traditional tobacco in combination. This practice can occur in all age groups, but studies have suggested that AI/AN adolescents may be a vulnerable population, although this was not investigated in the current study. Specifically, Forster et al. (2008) demonstrated that 39% of AI/AN adolescents (ages 11-18 years) living in Minneapolis and St. Paul areas reported using commercial tobacco for ceremonial purposes, compared to 24% who use native tobacco for ceremony. Additionally, Unger et al. (2006) showed that adolescents report observing both homegrown and commercial tobacco smoking at ceremonies and events. Smoking a mixture of commercial and traditional tobacco results in exposure to nicotine and tobacco-specific carcinogens, elevating the risk of lung cancer and cardiovascular disease, health disparities that are already present in the NP tribal population (Bliss et al., 2008, Plescia et al., 2014, Veazie et al., 2014, Mowery et al., 2015). Accordingly, we advise community efforts to work with traditional healers and tribal leaders to educate tribal members on risks of blending commercial tobacco products with traditional tobacco.

Secondhand smoke exposure

We observed higher cotinine levels among NP, but not SW, non-smokers who allowed smoking in the home and had friends who smoked, suggesting that these two factors increase passive exposure to tobacco smoke. These results are consistent with previous findings that non-smokers with friends who smoke have cotinine levels 1.5 times higher than non-smokers without such friends (Etter et al., 2000). The general absence of smoke-free policies in AI/AN communities means that smokers and non-smokers often share the same environments. For example, in the largely urban SW tribal population, participants benefited from smoke-free workplaces, whereas the reservation-based NP tribal populations did not. More secondhand
smoke exposure among NP non-smokers increases the community risk of lung cancer, stroke, and cardiovascular disease, while also increasing symptoms of nicotine dependence among younger non-smokers (Eriksen et al., 1988, Belanger et al., 2008, Pitsavos et al., 2002). Additionally, secondhand smoke exposure is a negative predictor of smoking cessation among smokers (Eng et al., 2014, Homish and Leonard, 2005, Okechukwu et al., 2010).

The same two factors also affected smokers in both regions, such that NP and SW smokers who allowed smoking in the home and had friends who smoked also exhibited higher cotinine levels and smoked more CPD than those who did not. Smoking more CPD likely accounts for these elevated cotinine levels, since CPD was a significant independent predictor of cotinine in our linear regression models, whereas passive smoke exposure was not.

**Tailored smoking cessation approaches**

Culturally tailored, population-specific smoking cessation interventions may be beneficial in AI/AN communities. Interventions validated in heavy smokers are not necessarily effective for light smokers, as observed in two unsuccessful clinical trials of smoking cessation therapy among African Americans light smokers where compliance was also low (Ahluwalia et al., 2006, Cox et al., 2012). Thus improving compliance, which increases the likelihood of successfully quitting (Faseru et al., 2013, Nollen et al., 2006), and culturally tailoring cessation, such as incorporating AI/AN imagery in intervention materials, disseminating information on traditional tobacco and its spiritual use, and providing pharmacotherapy and counseling by AI/AN practitioners (Daley et al., 2006, D'Silva et al., 2011, Choi et al., 2011), may be approaches which could be tested to enhance smoking cessation among these populations.

**Limitations**

This study has several limitations. First, there are limitations associated with using cotinine as a biomarker of tobacco smoke exposure. Although cotinine is superior to carbon monoxide as a biomarker of tobacco smoke exposure, primarily because of its specificity to nicotine (SRNT Subcommittee on Biochemical Verification, 2002), measurement of total nicotine equivalents in urine is more sensitive than cotinine, especially among light or intermittent smokers. Additionally, cotinine levels in smokers and non-smokers are sensitive only to tobacco smoke exposure occurring a few days before measurement. Therefore, our approach assessed only
recent tobacco smoke exposure. Despite these concerns, we found a robust correlation between cotinine levels and self-reported CPD, suggesting that cotinine was a useful biomarker of tobacco smoke exposure for study purposes.

Second, although we assessed several measures of secondhand smoke exposure, this was not an exhaustive evaluation. Other potential sources of secondhand exposure should be addressed by future research in AI/AN communities. The scope of this study also did not include the assessment of other tobacco products, aside from commercial cigarettes and traditional tobacco, including non-combustibles. Non-combustibles, such as electronic cigarettes, were not in widespread use at the time of evaluation (2012-2014). Moreover, NP and SW cotinine levels were correlated with CPD, suggesting that tobacco exposure in these populations is associated with cigarette consumption, and not the use of other tobacco products, such as non-combustibles; however, the observed correlations do not preclude the association of cotinine levels with the use of other products. Additionally, we did not independently determine the composition of the forms of ceremonial traditional tobacco used by study participants. This study also did not assess the impact of body mass index on cotinine, which has previously exhibited a weak negative correlation (Ho et al., 2009a, Perez-Stable et al., 1995, Chenoweth et al., 2014a). Furthermore, sample sizes limited our statistical power. Analyses indicated that the direction of effects and the association between cotinine levels and tobacco smoke exposure were similar in the NP and SW samples. However, specific findings differed in significance, likely because of a lack of analytic power resulting from insufficient sample sizes, particularly in the SW. Additionally, weighting was not conducted to correct for possible clustering due to respondent driven sampling. Finally, participants in the current study were individuals who had participated in an earlier large epidemiological study that was designed to characterize the lifestyle, dietary, environmental, and cultural factors associated with cancer in adult American Indians. It is possible that being involved in the EARTH study increased participants’ awareness of, and education on, the health risks of smoking, contributing to reduced smoking quantities, and thus lower than anticipated measures of CPD in the current study (NP 7 vs. 13, and SW 4 vs. 7 CPD) compared to EARTH.

Conclusions

Although overall cotinine levels were lower among these two tribal populations compared to smokers in other U.S. ethnic and racial groups, smoking prevalence (Nez Henderson et al.,
2005) and secondhand smoke exposure were high in both populations, especially the NP. Given the causal relationship of smoking and secondhand smoke inhalation with lung cancer and cardiovascular disease, our results suggest that these are at risk tribal populations. In both tribal populations, our findings support the implementation of stricter indoor smoking bans and culturally tailored tobacco cessation programs.
2.6 Significance to Thesis

The findings from this chapter contribute to the literature in three ways. First, these data demonstrate that cotinine is associated with self-reported CPD in the Northern Plains and Southwest tribal populations, expanding on existing literature on cotinine as a smoking biomarker into understudied light smoking (<10 CPD) populations. This further validates the use of this biomarker in future studies of tobacco use and exposure.

Second, this study identifies the use of commercial tobacco for ceremonial purposes by Northern Plains smokers. Traditional tobacco use was a significant independent predictor of cotinine levels among Northern Plains, but not Southwest, smokers, suggesting that Northern Plains smokers substitute nicotine-containing commercial tobacco for non-nicotinic traditional tobacco products during ceremonies. The implications of this are such that the tobacco users, and non-users in their immediate vicinity, are being exposed to harmful tobacco smoke, increasing their risk for tobacco-related diseases. Further, this suggests that tobacco users are exposed to more nicotine, increasing their chance of tobacco dependence. As lung cancer risk and smoking prevalence are already high in the Northern Plains population, identifying this source of carcinogen and nicotine exposure may help in focusing future efforts to reduce smoking and disease risk.

Lastly, this chapter reveals the high levels of secondhand tobacco smoke exposure among non-smokers in both tribal populations. Approximately 1 out of every 4 non-smokers in the Northern Plains and Southwest had secondhand smoke exposure above the cotinine cutpoint of 3 ng/ml. Two sources of secondhand smoke exposure in the Northern Plains, but not the Southwest, were allowing smoking in the home and having close friends who smoke. Collectively, this chapter supports work toward increased implementation of indoor smoking bans. It would also be useful to assess tribal knowledge regarding the harms of secondhand tobacco smoke exposure, and the use of commercial tobacco for ceremonial purposes, both of which result in exposure to tobacco-specific carcinogens. These investigations would help determine the necessity and inform the content of educational materials addressing these harms.
3 CHAPTER 2: CYP2A6 AND NICOTINE METABOLISM AMONG TWO AMERICAN INDIAN TRIBAL GROUPS DIFFERING IN SMOKING PATTERNS AND RISK FOR TOBACCO-RELATED CANCER

Julie-Anne Tanner, Jeffrey A. Henderson, Dedra Buchwald, Barbara V. Howard, Patricia Nez Henderson, and Rachel F. Tyndale


JAH, DB, BVH, and RFT designed the overall study. JAH, DB, and BVH recruited participants. JAT determined experimental questions based on reviews of the literature. JAT conducted all genotyping and merged and maintained the study database. JAT conducted genotype-phenotype analyses within and between tribes, performing all statistical analyses and modeling. All authors interpreted the results. JAT and RFT wrote the manuscript.
3.1 Abstract

**Objectives:** The Northern Plains (NP) and Southwest (SW) American Indian populations differ in their smoking patterns and lung cancer incidence. We aimed to compare \textit{CYP2A6} genetic variation and \textit{CYP2A6} enzyme activity (representative of the rate of nicotine metabolism) between the two tribal populations, as these have previously been associated with differences in smoking, quitting, and lung cancer risk.

**Methods:** American Indians (N=636) were recruited from two different tribal populations (NP in South Dakota, SW in Arizona) as part of a study conducted as part of the Collaborative to Improve Native Cancer Outcomes P50 project. A questionnaire assessed smoking-related traits and demographics. Participants were genotyped for \textit{CYP2A6} genetic variants *1B, *2, *4, *7, *9, *12, *17, and *35. Plasma and/or saliva samples were used to measure nicotine’s metabolites cotinine and 3’-hydroxycotinine and determine \textit{CYP2A6} activity (3’-hydroxycotinine/cotinine, i.e. the nicotine metabolite ratio, NMR).

**Results:** The overall frequency of genetically reduced nicotine metabolizers, those with \textit{CYP2A6} decrease- or loss-of-function alleles, was lower in the NP compared to the SW (\(P=0.0006\)). \textit{CYP2A6} genotype was associated with NMR in both tribal groups (NP \(P<0.001\), SW \(P=0.04\)). Notably, the rate of nicotine metabolism was higher in NP compared to SW smokers (\(P=0.03\)), and in comparison to other ethnic groups in the USA. Of the variables studied, \textit{CYP2A6} genotype was the only variable to significantly independently influence NMR among smokers in both tribal populations (NP \(P<0.001\), SW \(P=0.05\)).

**Conclusions:** Unique \textit{CYP2A6} allelic patterns and rates of nicotine metabolism among these American Indian populations suggest different risks for smoking, and tobacco-related disease.
3.2 Introduction

Patterns of commercial tobacco use in the USA differ among ethnic groups, with cigarette smoking being more prevalent in American Indian and Alaska Native (AI/AN) populations compared to other USA populations (Centers for Disease and Prevention, 2004). However, between different AI/AN tribal populations, prevalence and smoking patterns are also variable, for example between the Northern Plains (NP) and Southwestern (SW) tribal populations (Nez Henderson et al., 2005, Eichner et al., 2010). Despite most AI/ANs wanting to quit smoking, and making quit attempts (Jamal et al., 2014, Gohdes et al., 2002, Forster et al., 2007), the prevalence of smoking is 50% in the NP tribal population of South Dakota while only being 14% in the SW tribal population of Arizona (Nez Henderson et al., 2005) and cigarette consumption is higher in the NP relative to the SW (13 vs. 7 cigarettes per day, CPD) (Eichner et al., 2010). Consistent with this, rates of lung cancer incidence and mortality, to which cigarette smoking has been causally linked, are more than six times higher in the NP compared to the SW tribal population (Bliss et al., 2008, Plescia et al., 2014). The underlying reasons for these large differences in tribal smoking and disease risk, whether biological and/or social, are unknown.

One biological influence on smoking is the activity of the hepatic CYP2A6 (cytochrome P450 2A6) enzyme, which varies substantially between individuals and ethnicities (Malaiyandi et al., 2005). Nicotine, the primary psychoactive component present in cigarette smoke, is metabolically inactivated by CYP2A6 (Benowitz and Jacob, 1994) with the major pathway of inactivation being conversion of nicotine to cotinine (COT) (Messina et al., 1997). COT is further metabolized to trans-3'-hydroxycotinine (3HC) exclusively by CYP2A6 (Nakajima et al., 1996a). Interindividual variation in CYP2A6 enzymatic activity, and thus the rate of nicotine metabolism and clearance, largely results from genetic variation in CYP2A6, the highly polymorphic gene encoding the CYP2A6 enzyme (variants characterized to date found at http://www.cypalleles.ki.se/cyp2a6.htm). Regular daily smokers alter their levels and intensity of smoking to titrate their nicotine intake such that they maintain desired nicotine levels in the body (McMorrow and Foxx, 1983). Accordingly, CYP2A6 genetic variation and altered rates of nicotine metabolism are associated with differences in smoking behaviors (Malaiyandi et al., 2005). Smokers with slower rates of nicotine metabolism, those possessing decrease and/or loss-of-function CYP2A6 alleles (referred to as reduced metabolizers), often
have lower tobacco consumption, dependence, and difficulty quitting smoking compared to faster metabolizers (Schoedel et al., 2004, Schnoll et al., 2009, Wassenaar et al., 2011). Among heavy smokers (≥10 CPD), genetically reduced metabolizers smoke fewer CPD compared to smokers with wild-type CYP2A6 genotypes (Schoedel et al., 2004). However, in lighter smoking populations (<10 CPD), measurement of CPD is not a sensitive measure of tobacco consumption (Zhu et al., 2013a, Ho et al., 2009b); while light smokers may consume the same number of CPD, they can differ substantially in their smoking topography (puff volume, duration, velocity), and thus overall tobacco consumption (Bridges et al., 1990). Measuring urinary total nicotine equivalents (TNE) is a considerably more precise biomarker of nicotine dose compared to CPD and is the current gold standard for assessing intake (Benowitz et al., 1994, Scherer et al., 2007). TNE was lower among AN light smokers with decrease-of-function CYP2A6 genotypes, and lower rates of nicotine metabolism, despite no differences in CPD (Zhu et al., 2013a).

In addition to CYP2A6 genotype, enzyme activity itself is associated with smoking behaviors (Lerman et al., 2015, Falcone et al., 2011). CYP2A6 enzymatic activity can be determined in smokers using a ratio of nicotine’s metabolites 3HC/COT (nicotine metabolite ratio, NMR), which is highly correlated to the rate of nicotine metabolism and clearance (Dempsey et al., 2004). The NMR is a validated biomarker of CYP2A6 activity due to its stability over time and measurement consistency among heavy and light smokers (Lea et al., 2006, Mooney et al., 2008). Previous studies have confirmed the strong concordance between CYP2A6 genotype and NMR in multiple populations and ethnic groups, suggesting similar impacts on smoking and lung cancer (Ho et al., 2009b, Binnington et al., 2012).

Lung cancer risk is also associated with CYP2A6 genetic variation. Smoking fewer CPD, which has been observed among genetically reduced compared to normal nicotine metabolizers (Schoedel et al., 2004), is associated with a lower risk of lung cancer (Khuder, 2001); however the relationship between CYP2A6 and lung cancer risk remains even when controlling for cigarette consumption (Wassenaar et al., 2014). One reason for this is that CYP2A6 can also metabolically activate procarcinogenic tobacco-specific nitrosamines; thus, being a slower CYP2A6 metabolizer can reduce both tobacco consumption and procarcinogen activation, resulting in lower lung cancer risk (Wong et al., 2005).
Although the impact of individual CYP2A6 alleles/genotypes on the rate of nicotine metabolism has been consistent across ethnicities (Ho et al., 2009b, Binnington et al., 2012, Malaiyandi et al., 2006), patterns of CYP2A6 genetic variation and nicotine metabolism vary substantially between different ethnic groups. (Mwenifumbo et al., 2005, Nakajima et al., 2006, Benowitz et al., 2002, Perez-Stable et al., 1998, Binnington et al., 2012). To date, no studies have investigated CYP2A6 genetic variation and characterized associations between CYP2A6 genotype and the rate of enzymatic activity, among AI populations, let alone compared two independent tribal groups with contrasting patterns of smoking. Given the distinct patterns of CYP2A6 genetic variation among different ethnic groups, it is plausible that the NP and SW tribal populations exhibit unique patterns of variation at this locus. Therefore, the aim of the current study was to compare CYP2A6 genetic variation and the rate of nicotine metabolism between the NP and SW tribal populations.

3.3 Participants and Methods

Study design

Members of AI tribal groups were recruited between 2012 and 2014 in South Dakota and Arizona for a study entitled ‘Topography and Genetics of Smoking and Nicotine Dependence in American Indians’, one of five major research studies conducted by the Collaborative to Improve Native Cancer Outcomes. Recruitment in the NP occurred from a random subset of participants in an earlier study of community health (Slattery et al., 2007). Recruitment in the SW was conducted using respondent-driven sampling among American Indian friends and family members of tribal participants in an earlier randomized clinical trial in the greater Phoenix metropolitan area (Howard et al., 2008). Both NP and SW subsamples were stratified by sex and smoking status. The final cohort comprised N=636 American Indians aged 20-88 years, with N=426 in the NP and N=210 in the SW groups. The urban SW and the predominantly rural NP tribal populations are culturally and linguistically distinct, and have considerably different historical experiences (Sonneborn, 2007).

Data were collected by trained research staff who were tribal members. Biological samples (blood or saliva) were collected from all participants to be used for genotyping and phenotyping. Personal interviews were conducted, and questionnaires were administered to all participants. Ethical approval for all study procedures was obtained from the institutional
review boards of the Great Plains Area Indian Health Service, the University of Toronto, the University of Washington, and MedStar Health Research Institute, and passed through individual tribal approval processes. All participants provided informed consent.

**Measures**

Data were obtained on age, sex, BMI, smoking status, duration of smoking, CPD, nicotine dependence scores (Fagerström Test for Nicotine Dependence, FTND; Hooked On Nicotine Checklist, HONC), and ceremonial traditional tobacco use. Levels of COT and 3HC were measured in blood or saliva samples using liquid chromatography tandem mass spectrometry, as described previously (Tanner et al., 2015b). Our phenotypic measure of CYP2A6 activity, NMR, was defined as the ratio of nicotine’s metabolites, 3HC/COT (Dempsey et al., 2004).

Relatedness was determined for all participants based on self-report and genetic analyses. When nuclear families were enrolled, only data on parents (founders) were used to determine allele frequencies; when sibships were enrolled, only one randomly chosen sibling was used. In addition to self-report, relatedness was also assessed by identity-by-descent (IBD) values. These data showed an excellent concordance with the self-reported family data. For individuals who had enrolled as independent, we assessed potential cryptic relatedness, and excluded one member of each pair with IBD values of 0.25 or greater. IBD values were computed using SNP & Variation Suite (Golden Helix, Inc., Bozeman, MT, USA). Only unrelated participants were included in the comparison of CYP2A6 gene variant frequencies between the two tribal populations. Only self-reported smokers with COT levels above 10 ng/ml have been included in the NMR analyses to ensure smokers had sufficient COT levels that 3HC was quantifiable and formation dependent (St Helen et al., 2012, Tanner et al., 2015b). NMR is used as a biomarker of CYP2A6 activity in current regular smokers only as steady state levels of cotinine and 3HC, achieved through regular cigarette smoking, are necessary for accurate measurement of CYP2A6 activity; this is not a feasible measurement of CYP2A6 activity in non-smokers in this study (St Helen et al., 2013b).

**CYP2A6 Genotyping Assays**

DNA was extracted from saliva or blood, and DNA from N=634 participants (NP N=426, SW n=208) was genotyped successfully for the following CYP2A6 alleles: *1B, *2, *4, *7, *9, *12, *17, and *35. Genotyping was performed using a two-step allele-specific PCR approach, or an
allele-specific TaqMan single nucleotide polymorphism genotyping assay (Applied Biosystems) and real-time PCR as described previously (Wassenaar et al., 2016). The CYP2A6 alleles *2, *4, *7, *9, *12, *17, and *35 have been associated with a decrease or loss of CYP2A6 enzyme activity and rates of nicotine metabolism (Al Koudsi et al., 2009, Xu et al., 2002, Ho et al., 2009b), whereas the *IB allele is associated with faster nicotine metabolism (Mwenifumbo et al., 2008b). We have grouped participants as ‘normal’ or ‘reduced’ CYP2A6 metabolizers on the basis of the predicted impact of their CYP2A6 genotype. Normal metabolizers had no CYP2A6 decrease-of-function or loss-of-function genetic variants, whereas reduced metabolizers possessed one or more copies of these alleles.

Statistical analyses

NMR was non-normally distributed; thus, nonparametric statistical tests were used. $\chi^2$-tests were used to determine Hardy-Weinberg equilibrium, and to compare the frequencies of CYP2A6 genetic variants and the overall proportion of subjects who were genetically reduced metabolizers between the NP and SW tribal populations. We used Kruskal-Wallis and Dunn’s multiple comparisons tests to analyze the association of CYP2A6 genotype with NMR. Mann-Whitney tests were conducted to compare NMR between genetically normal versus reduced metabolizers within each tribal population and to compare overall NMR between the NP and SW populations.

Linear regression models were run among smokers with COT above 10 ng/ml to determine the percentage of variation in NMR (the output variable) accounted for. Our first model included the variables CYP2A6 genotype, sex, age, BMI, CPD, ceremonial traditional tobacco use, and tribal population (i.e. NP or SW). We also ran separate regression models for each tribal population, in which model variables included CYP2A6 genotype and BMI. In each model, CYP2A6 genotype was coded as 0 for individuals with no known CYP2A6 reduced activity alleles (i.e. *1A/*1A, *1A/*1B, *1B/*1B genotypes), -1 for those possessing one decrease-of-function allele (i.e. *1/*9 and *1/*12 genotypes), and -2 for those possessing one or more loss-of-function or two decrease-of-function alleles (i.e. *1/*2, *1/*4, and *9/*9 genotypes), as done previously (Chenoweth et al., 2013). A linear regression model of NMR was also run in wild-type (CYP2A6*1/*1, i.e. individuals who do not have any known CYP2A6 reduced activity genetic variants) smokers only to assess the impact of the increase-of-function CYP2A6*1B genotype and tribal population. Additionally, a linear regression model was used
to test if CYP2A6*1B genotype interacts with tribal population to influence NMR. The single predictors (CYP2A6*1B genotype and tribal population) were entered into block 1, and the interaction term (CYP2A6*1B genotype x tribal population) was entered into block 2. In both models, the CYP2A6 genotype was coded as 0 for individuals with the *1A/*1A genotype, 1 for those with the *1A/*1B genotype, and 2 for those with the *1B/*1B genotype (Mwenifumbo et al., 2008b). Analyses were conducted with GraphPad Prism (version 6.0; La Jolla, CA, USA) and SPSS (version 22; Armonk, NY, USA), and statistical tests were considered significant for P value less than 0.05.

3.4 Results

 Frequencies of CYP2A6 alleles and genetically reduced nicotine metabolizers

CYP2A6 genotype frequencies within each tribal population did not deviate significantly from Hardy-Weinberg equilibrium (P>0.05). The two NP and SW tribal populations showed distinct frequencies of CYP2A6 variant alleles from one another and compared to other ethnic groups (Table 6). The frequency of the gain-of-function *1B allele was significantly higher in the NP than the SW (NP 69.7% vs. SW 61.6%, P=0.01), whereas the prevalence of the decrease-of-function *9 allele was significantly lower in the NP compared to the SW (NP 11.9% vs. SW 20.9%, P=0.0002). Several alleles that have been found in other ethnic groups were not present in the NP and SW tribal populations (NP *7, *17, *35; SW *7, *17). Next we categorized individuals into ‘normal’ or ‘reduced’ CYP2A6 activity groups on the basis of genotype, with reduced metabolizers defined as individuals who have any decrease-of-function or loss-of-function alleles; this included the following genotypes: *1/*2, *1/*4, *1/*9, *9/*9, *1/*12, and *1/*35. A smaller proportion of the NP tribal population were CYP2A6 genotype-reduced nicotine metabolizers compared to the SW tribal population (NP 27.5% vs. SW 41.8%, P=0.0006).
Table 6 | Comparison of Northern Plains (n=318<sup>a</sup>) and Southwestern (n=172<sup>a</sup>) CYP2A6 allele frequencies, and a summary of frequencies from previously studied populations of different ethnic backgrounds.

<table>
<thead>
<tr>
<th>CYP2A6 allele</th>
<th>Activity</th>
<th>Allele Frequency (%)&lt;sup&gt;c,d&lt;/sup&gt;</th>
<th>Alaska Native</th>
<th>White</th>
<th>African American</th>
<th>Japanese</th>
<th>Chinese</th>
<th>Korean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Northern Plains</td>
<td>Southwest</td>
<td>P value&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*1B</td>
<td>Increase</td>
<td>69.7</td>
<td>61.6</td>
<td>0.01</td>
<td>65.3</td>
<td>26.7–35.0</td>
<td>11.2–18.2</td>
<td>25.6–54.6</td>
</tr>
<tr>
<td>*2</td>
<td>Inactive</td>
<td>0.3</td>
<td>0.6</td>
<td>0.53</td>
<td>0.4</td>
<td>1.1–5.3</td>
<td>0–1.1</td>
<td>0</td>
</tr>
<tr>
<td>*4</td>
<td>Inactive</td>
<td>1.6</td>
<td>0.3</td>
<td>0.07</td>
<td>14.5</td>
<td>0.13–4.2</td>
<td>0.5–2.7</td>
<td>17.0–24.2</td>
</tr>
<tr>
<td>*7</td>
<td>Inactive</td>
<td>0</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>0–0.3</td>
<td>0</td>
<td>6.3–12.6</td>
</tr>
<tr>
<td>*9</td>
<td>Decrease</td>
<td>11.9</td>
<td>20.9</td>
<td>0.0002</td>
<td>8.9</td>
<td>5.2–8.0</td>
<td>5.7–9.6</td>
<td>19.0–20.7</td>
</tr>
<tr>
<td>*12</td>
<td>Decrease</td>
<td>0.3</td>
<td>0.3</td>
<td>0.95</td>
<td>0.4</td>
<td>0–3.0</td>
<td>0–0.4</td>
<td>0–0.8</td>
</tr>
<tr>
<td>*17</td>
<td>Inactive</td>
<td>0</td>
<td>0</td>
<td>–</td>
<td>0</td>
<td>0</td>
<td>7.1–10.5</td>
<td>0</td>
</tr>
<tr>
<td>*35</td>
<td>Decrease</td>
<td>0</td>
<td>0.3</td>
<td>0.17</td>
<td>0</td>
<td>0</td>
<td>2.5–2.9</td>
<td>0.8</td>
</tr>
</tbody>
</table>

<sup>a</sup> Only unrelated tribal participants have been included in this analysis. Relatedness was determined based on self-report and genetic analyses.

<sup>b</sup> P values compare NP and SW allele frequencies and are based on χ²-tests.

<sup>c</sup> Allele frequency data for Alaska Native, White, African American, Japanese, Chinese, and Korean populations as cited in Tanner et al. (2015)

<sup>d</sup> Genotype frequencies did not deviate significantly from Hardy-Weinberg equilibrium (P>0.05 for each genotypes in each tribe). The P values that were significant (<0.05) were bolded to highlight statistical significance.
Associations of *CYP2A6* genotype with CYP2A6 activity (nicotine metabolite ratio)

We assessed the rate of CYP2A6 activity, determined by the NMR, among smokers to determine if the functional impact of the change in each variant *CYP2A6* allele was consistent with the impact observed in other populations, and in *in vitro* expression systems (Al Koudsi et al., 2010). The *CYP2A6*/*1B allele had previously been associated with an increased rate of overall nicotine clearance and NMR *in vivo* (Mwenifumbo et al., 2008b), and we observed higher NMR among smokers in both the NP (*P* = 0.02) and SW (*P* = 0.002) tribal populations who possessed the *1B* allele relative to the wild-type *1A* allele (Fig. 15; three-group comparison across *1A/*1A, *1A/*1B, and *1B/*1B genotypes). As many decrease-of-function and loss-of-function alleles exist on a variety of *1A* or *1B* backbones, we have grouped these as the *1/*1 wild-type group for comparisons with the *CYP2A6* genotype-reduce activity group (a compilation of all *CYP2A6* decrease-of-function and loss-of-function alleles).

The *CYP2A6* genotype was associated with CYP2A6 activity such that smokers with one or more decrease-of-function or loss-of-function alleles showed lower mean NMR compared with the *1/*1 reference group (NP *P*< 0.0001, SW *P* = 0.04; Fig. 16). We also observed the presence of a gene-dose effect with the prevalent decrease-of-function *9* allele in both the NP and SW tribal populations. In the NP tribal population, relative to the wild-type group (*1/*1 genotype), those with one copy of the *9* allele (*1/*9 genotype) had 63.4% CYP2A6 activity and those with two copies of *9* (*9/*9 genotype) had 29.6% CYP2A6 activity (Fig. 16). Similarly, in the SW tribal population, smokers with one and two copies of *9* exhibited 84.0% and 52.7% CYP2A6 activity, respectively (Fig. 16).
Figure 15 | Comparison of CYP2A6 activity (NMR) between the wild-type CYP2A6 genotype groups *1A/*1A, *1A/*1B, and *1B/*1B. Individuals included in this analysis do not have any other tested variants. $P$ values for multiple group comparisons (across all three genotypes) are based on Kruskal-Wallis tests. $P$ values comparing between groups are based on Dunn’s multiple comparisons tests. One outlier, 1.5 SDs from the mean, was excluded from the plot for illustration purposes, but was included in the statistical analyses (was one of four members of the Northern Plains *1A/*1A group, NMR=1.37). CI, confidence interval; NMR, nicotine metabolite ratio.
Figure 16 | Association between CYP2A6 genotype and CYP2A6 activity (NMR) among smokers. The *1/*1 group represents the normal metabolizers (those who do not have any known CYP2A6 decrease-of-function or loss-of-function alleles; for simplicity, the wild-type *1A and *1B alleles were assessed generically as *1 for these analyses). The reduced metabolizer group (RM) combines all individuals who possess one or more decrease-of-function or loss-of-function alleles (i.e. the following genotypes: *1/*2, *1/*4, *1/*9, *9/*9, and *1/*12). P values for multiple group comparisons (across all genotypes) are based on Kruskal-Wallis tests. P values comparing *1/*1 and RM groups are based on Mann-Whitney tests. NMR, nicotine metabolite ratio.
Comparison of the rate of nicotine metabolism across populations

We compared the rate of nicotine metabolism (i.e. CYP2A6 activity), determined by NMR, between NP and SW smokers, and with additional populations of different ethnic backgrounds. Among all smokers (wild-type and CYP2A6 genotype-reduced metabolizers combined), NP smokers had a faster mean rate of nicotine metabolism relative to SW smokers ($P=0.03$; Fig. 17a and 18a). NP smokers also had a higher mean rate of nicotine metabolism compared to smokers from other ethnic groups, including AN, White, and African American smokers, whose NMR had been assessed previously (Binnington et al., 2012, Lerman et al., 2006, Ho et al., 2009b). When analyses were restricted only to wild-type smokers (those who do not possess any known CYP2A6 decrease-of-function or loss-of-function alleles) to avoid the confound of higher or lower frequencies of CYP2A6 genotype-reduced nicotine metabolizers between populations, the rate of nicotine metabolism remained higher among NP smokers compared to SW smokers ($P=0.003$), and compared to smokers from other ethnic groups (Figs. 17b and 18b). In addition, the rate of nicotine metabolism was again higher among NP relative to SW smokers when analyses were restricted to the *1B/*1B group ($P=0.02$; Fig 15).
Figure 17 | Comparison of CYP2A6 activity (NMR) among the two tribal populations of smokers (NP, Northern Plains; SW, Southwest). (a) All smokers included in analyses, irrespective of the CYP2A6 genotype. (b) Smokers with wild-type genotypes; *1A/*1A, *1A/*1B, and *1B/*1B genotypes only (CYP2A6 reduced metabolizer genetic variants were excluded). P values are based on Mann-Whitney tests. Other American smoking populations included for visual comparison (AN, Alaska Native; C, White; AA, African American). CYP2A6 genotype and NMR data from other American populations taken from the following sources: AN (Binnington et al., 2012); C (Lerman et al., 2006); AA (Ho et al., 2009b). CI, confidence interval; NMR, nicotine metabolite ratio.
Figure 18 | Frequency distributions of the CYP2A6 activity ratio (NMR) among the two tribal populations of smokers (NP, Northern Plains; SW, Southwest), i.e. the number of instances in which NMR has taken each of its possible values. (a) All smokers were included in analyses, regardless of CYP2A6 genotype. $P=0.03$ for NP vs. SW mean NMR (Mann-Whitney test). (b) Smokers with wild-type genotype; *1A/*1A, *1A/*1B, and *1B/*1B genotypes only (CYP2A6 reduced metabolizer (RM) genetic variants were excluded). $P=0.003$ for NP vs. SW mean NMR (Mann-Whitney test).
Variables that influence the rate of nicotine metabolism

Using linear regression modeling, we investigated the proportion of variation in the rate of nicotine metabolism (NMR) that was attributable to different predictor variables among smokers in each tribal population, while controlling for the influence of other variables. Both tribal populations were included in a model of NMR in which CYP2A6 genotype, sex, age, ceremonial traditional tobacco use, BMI, CPD, and tribal population (i.e. NP and SW) were assessed as predictor variables of NMR. Only CYP2A6 genotype (P<0.001), BMI (P=0.03), and tribal population (P=0.06) each independently accounted for more than 1% of the variation in NMR (Table 7a). In a second model we assessed predictors of NMR separately in the NP and SW tribal populations, including only the variables CYP2A6 genotype and BMI (Table 7b and c). These models indicated that, of the variables studied in both tribal populations, CYP2A6 genotype was the only significant independent predictor of NMR among smokers, accounting for 19.71% (P<0.001) and 8.12% (P=0.05) of the variation in NMR in the NP and SW tribal populations, respectively (Table 7).

As we have observed that NP smokers have a higher overall NMR compared to SW smokers, we aimed to determine if this observation was independent of the effect of the increase-of-function CYP2A6*1B allele, which was expressed at a higher frequency in the NP compared to the SW tribal population. We ran a linear regression model for NMR among wild-type smokers only in which CYP2A6*1B genotype and tribal population were included as predictor variables. Both variables independently accounted for \( \geq 6\% \) of the variation in NMR (CYP2A6*1B genotype, 7.34\%, \( P=0.0006 \); tribal population, 6.00\%, \( P=0.01 \); Table 8a). In addition, using linear regression analyses, we assessed if there was a significant interaction between CYP2A6*1B genotype and tribal population in influencing NMR. The interaction term (genotype x tribal population) was not significant (\( P=0.72 \), model R\(^2\) change=0.001; Table 8b), suggesting that the relationship between CYP2A6*1B genotype and NMR is similar in both tribal populations.
Table 7 | Linear regression analyses of the nicotine metabolite ratio among Northern Plains and Southwestern smokers.

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>95% CI for B</th>
<th>Beta</th>
<th>Variation in NMR accounted for by each variable a</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Both Tribal Populations Combined (n=150 included in model) c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2A6 Genotype b</td>
<td>0.18</td>
<td>0.11 to 0.25</td>
<td>0.38</td>
<td>14.52%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sex (male=1, female=0)</td>
<td>-0.04</td>
<td>-0.14 to 0.05</td>
<td>-0.07</td>
<td>0.45%</td>
<td>0.37</td>
</tr>
<tr>
<td>Age (years)</td>
<td>0.0001</td>
<td>-0.003 to 0.004</td>
<td>0.02</td>
<td>0.03%</td>
<td>0.82</td>
</tr>
<tr>
<td>Traditional Tobacco user (yes=1, no=0)</td>
<td>-0.01</td>
<td>-0.12 to 0.10</td>
<td>-0.02</td>
<td>0.03%</td>
<td>0.81</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>-0.008</td>
<td>-0.01 to -0.001</td>
<td>-0.17</td>
<td>1.35%</td>
<td>0.03</td>
</tr>
<tr>
<td>CPD</td>
<td>0.003</td>
<td>-0.004 to 0.01</td>
<td>0.07</td>
<td>0.45%</td>
<td>0.36</td>
</tr>
<tr>
<td>Tribal population (NP=1, SW=2)</td>
<td>-0.10</td>
<td>-0.20 to 0.002</td>
<td>-0.15</td>
<td>2.05%</td>
<td>0.06</td>
</tr>
<tr>
<td>b. Northern Plains Smokers (n=108 included in model) d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2A6 Genotype b</td>
<td>0.21</td>
<td>0.13 to 0.30</td>
<td>0.45</td>
<td>19.71%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>-0.007</td>
<td>-0.02 to 0.001</td>
<td>-0.16</td>
<td>2.40%</td>
<td>0.07</td>
</tr>
<tr>
<td>c. Southwestern Smokers (n=48 included in model) e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2A6 Genotype b</td>
<td>0.11</td>
<td>0.0001 to 0.22</td>
<td>0.29</td>
<td>8.12%</td>
<td>0.05</td>
</tr>
<tr>
<td>Body Mass Index</td>
<td>-0.005</td>
<td>-0.02 to 0.005</td>
<td>-0.14</td>
<td>1.99%</td>
<td>0.33</td>
</tr>
</tbody>
</table>

CI, confidence interval; CPD, cigarettes per day; NMR, nicotine metabolite ratio; NP, Northern Plains; SW, Southwest.

a. Variation in NMR accounted for by each variable is determined by: (part correlation)^2 x 100.
b. CYP2A6 genotype was coded as 0 for individuals with no known CYP2A6 reduced activity alleles (i.e. *1A/*1A, *1A/*1B, *1B/*1B genotypes), -1 for those possessing one decrease-of-function allele (i.e. *1/*9 and *1/*12 genotypes), and -2 for those possessing one or more loss-of-function or two decrease-of-function alleles (i.e. *1/*2, *1/*4, and *9/*9 genotypes).
c. Both tribal populations combined: R^2=0.222, P<0.001. R^2 indicates the proportion of variance in NMR levels (22.2%) that is explained by this model.
d. NP smokers: R^2=0.231, P<0.001. R^2 indicates the proportion of variance in NMR levels (23.1%) that is explained by this model.
e. SW smokers: R^2=0.092, P=0.11. R^2 indicates the proportion of variance in NMR levels (9.2%) that is explained by this model.

The P values that were significant (<0.05) were bolded to highlight statistical significance.
Table 8 | Linear regression analyses of the nicotine metabolite ratio among wild-type (*1A/*1A, *1A/*1B, *1B/*1B genotypes) Northern Plains and Southwestern smokers combined into one sample.

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>95% CI for B</th>
<th>Beta</th>
<th>Variation in NMR accounted for by each variable (^a)</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a. Original Model</strong> (n=97) (^d)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CYP2A6^{*1B}) Genotype (^b)</td>
<td>0.14</td>
<td>0.04 to 0.23</td>
<td>0.27</td>
<td>7.34%</td>
<td><strong>0.006</strong></td>
</tr>
<tr>
<td>Tribal Population (NP=1, SW=2)</td>
<td>-0.17</td>
<td>-0.30 to 0.04</td>
<td>-0.25</td>
<td>6.00%</td>
<td><strong>0.01</strong></td>
</tr>
<tr>
<td><strong>b. Original Model Plus Interaction Term “Genotype x Tribe”</strong> (n=97) (^e)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CYP2A6^{*1B}) Genotype (^b)</td>
<td>0.09</td>
<td>-0.20 to 0.37</td>
<td>0.17</td>
<td>0.32%</td>
<td>0.55</td>
</tr>
<tr>
<td>Tribal Population (NP=1, SW=2)</td>
<td>-0.23</td>
<td>-0.56 to 0.11</td>
<td>-0.33</td>
<td>1.64%</td>
<td>0.19</td>
</tr>
<tr>
<td>Genotype x Tribe Interaction (^c)</td>
<td>0.04</td>
<td>-0.16 to 0.23</td>
<td>0.13</td>
<td>0.12%</td>
<td>0.72</td>
</tr>
</tbody>
</table>

CI, confidence interval; CPD, cigarettes per day; NMR, nicotine metabolite ratio; NP, Northern Plains; SW, Southwest.

a. Variation in NMR accounted for by each variable is determined by: (part correlation)\(^2\) x 100.
b. \(CYP2A6\) genotype was coded as 0 for individuals with *1A/*1A genotype, 1 for those possessing *1A/*1B genotype, and 2 for those possessing *1B/*1B genotype.
c. \(CYP2A6^{*1B}\) genotype x tribal population interaction term

d. Original Model: \(R^2=0.142, P=0.001. R^2\) indicates the proportion of variance in NMR levels (14.2\%) that is explained by this model.
e. Original Model + Interaction Term: \(R^2\) change=0.001, \(P=0.72.\) The \(R^2\) change tells you the degree of change in your model with the addition of the interaction term (genotype x tribe).

The \(P\) values that were significant (<0.05) were bolded to highlight statistical significance.
Associations between smoking behaviors and CYP2A6 genotype and activity

Relationships between CYP2A6 genotype and NMR and smoking behaviors were assessed. There was no difference in the proportion of current and former smokers who were CYP2A6 genotype-reduced metabolizers in either the NP or SW tribal populations (NP current smokers 31.7%, former smokers 26.7%, $P=0.37$; SW current smokers 44.9%, former smokers 41.5%, $P=0.71$). There was no association between the following smoking behaviors and CYP2A6 genotype or NMR among either NP or SW self-reported smokers: mean number of CPD, FTND score, HONC score, and duration of smoking (Table 9). Parallel analyses were carried out in which smokers were defined by the more stringent cut point of COT above 10 ng/ml; these analyses yielded similar results (Table 10).
Table 9 | Associations between CYP2A6 genotype and activity (NMR) with smoking behaviors among Northern Plains and Southwestern tribal populations, with smokers defined by self-report.

<table>
<thead>
<tr>
<th>Smoking Behavior</th>
<th>CYP2A6 Genotype</th>
<th>NMR Tertiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>Reduced</td>
</tr>
<tr>
<td>A. Northern Plains</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean CPD (± 95% CI)</td>
<td>6.8 (5.4-8.2)</td>
<td>8.2 (6.1-10.4)</td>
</tr>
<tr>
<td>Mean FTND (± 95% CI)</td>
<td>2.1 (1.6-2.5)</td>
<td>1.8 (1.2-2.5)</td>
</tr>
<tr>
<td>Mean HONC (± 95% CI)</td>
<td>4.7 (4.0-5.4)</td>
<td>4.1 (3.1-4.9)</td>
</tr>
<tr>
<td>Duration of smoking, years (± 95% CI)</td>
<td>22.1 (20.1-24.0)</td>
<td>24.3 (21.2-27.3)</td>
</tr>
<tr>
<td>B. Southwest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean CPD (± 95% CI)</td>
<td>4.4 (2.5-6.3)</td>
<td>2.8 (1.4-4.3)</td>
</tr>
<tr>
<td>Mean FTND (± 95% CI)</td>
<td>1.6 (1.0-2.3)</td>
<td>2.1 (1.4-2.8)</td>
</tr>
<tr>
<td>Mean HONC (± 95% CI)</td>
<td>3.3 (2.4-4.2)</td>
<td>4.1 (2.9-5.3)</td>
</tr>
<tr>
<td>Duration of smoking, years (± 95% CI)</td>
<td>19.5 (16.9-22.0)</td>
<td>20.1 (17.0-23.2)</td>
</tr>
</tbody>
</table>

a. Smoking status was determined for each analysis based on self-report.
b. Only smokers who have COT>10 were included in NMR analyses.
c. $P$ values based on Mann-Whitney tests.
Table 10 | Associations between *CYP2A6* genotype and activity (NMR) with smoking behaviors among Northern Plains and Southwestern tribal populations, with smokers defined as having COT >10.

<table>
<thead>
<tr>
<th>Smoking Behavior</th>
<th>CYP2A6 Genotype a</th>
<th>NMR Tertiles a,b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>Reduced</td>
</tr>
<tr>
<td>Mean CPD (± 95% CI)</td>
<td>8.4 (6.6-10.1)</td>
<td>8.5 (6.2-10.8)</td>
</tr>
<tr>
<td>Mean FTND (± 95% CI)</td>
<td>2.4 (1.8-3.0)</td>
<td>2.0 (1.2-2.7)</td>
</tr>
<tr>
<td>Mean HONC (± 95% CI)</td>
<td>5.2 (4.4-6.0)</td>
<td>4.2 (3.3-5.1)</td>
</tr>
<tr>
<td>Duration of smoking, years (± 95% CI)</td>
<td>22.1 (20.1-24.0)</td>
<td>24.3 (21.2-27.3)</td>
</tr>
</tbody>
</table>

A. Northern Plains

B. Southwest

| Mean CPD (± 95% CI)               | 5.4 (3.4-7.5)    | 3.7 (0.9-6.4)                          | 5.3 (3.3-7.3)      | 3.5 (0.7-6.4)  | 0.08      |
| Mean FTND (± 95% CI)              | 1.9 (1.1-2.7)    | 2.3 (1.2-3.3)                          | 2.3 (1.5-3.0)      | 1.6 (0.6-2.6)  | 0.31      |
| Mean HONC (± 95% CI)              | 3.5 (2.3-4.8)    | 4.1 (2.2-6.0)                          | 3.7 (2.5-4.9)      | 3.9 (2.0-5.9)  | 0.95      |
| Duration of smoking, years (± 95% CI) | 19.5 (16.9-22.0) | 20.1 (17.0-23.2)                      | 23.0 (17.0-29.0)  | 21.7 (17.5-26.0)| 0.66      |

a. Smoking status was determined for each analysis such that participants who have COT >10 were designated smokers, and all remaining participants (regardless of self-reported smoking status) were excluded.
b. The NMR tertile data presented are the same as Table 9, but were included here to allow for comparison with this CYP2A6 genotype data.
c. P values based on Mann-Whitney tests.
3.5 Discussion

We have shown that the NP and SW tribal populations are genetically distinct at the CYP2A6 locus, with the NP population including a lower frequency of CYP2A6 genotype-reduced metabolizers relative to the SW tribal population, while also showing distinct individual allele frequencies. In both cohorts, CYP2A6 genotype was associated with our phenotypic measure of CYP2A6 activity and rate of nicotine metabolism, NMR. Notably, NP smokers showed a faster overall rate of nicotine metabolism compared with smokers in the SW tribal population and other ethnic groups, even among the CYP2A6 wild-type subgroup. The CYP2A6 genotype and NMR were not significantly associated with smoking behaviors in either tribal population.

The variety of CYP2A6 alleles that were investigated in this study (*1B, *2, *4, *7, *9, *12, *17, *35) was representative of those with a functional impact and those which were prevalent among different ethnic groups (Ho et al., 2009b, Binnington et al., 2012, Xu et al., 2002). For example, as illustrated in Table 6, common CYP2A6 alleles (frequency >1%) among different ethnic populations include the following: Whites *1B, *2, *4, *9, *12; African Americans *1B, *2, *4, *9, *17, *35; Asians *1B, *4, *7, *9; Alaska Natives *1B, *4, *9 (Tanner et al., 2015a). Choosing to investigate this variety of alleles allowed us to compare CYP2A6 genetic variation between the two populations, and characterize the tribal populations with respect to other ethnic groups. We had expected that the pattern of CYP2A6 allele frequencies among both the NP and SW tribal populations would most resemble that of Asian populations, as it is postulated that the NP and SW tribal populations have common Asian ancestral origins (Raghavan et al., 2015). However, the CYP2A6*7 allele, which is observed almost exclusively in those of Asian descent, was not present in either tribal population, and both the NP and SW have much lower frequencies of the CYP2A6*4 allele that is common in Asians (see Table 6). Additionally, we observed significantly different CYP2A6 allele frequencies between the NP and SW tribal populations, suggesting that, if these two populations share common ancestry, there have been incidences of genetic divergence between these two populations. There are several potential explanations for their unique genetic patterns. As each tribal population may have been established by a relatively small sample of original ancestors, a founder effect could have contributed toward the genetic divergence between tribal populations as well as from their ancestors (Templeton, 2008). In addition, because the tribal populations have been geographically separated, factors such as lack of gene flow, random genetic drift, and selective
pressures could contribute to development of these unique allele frequency patterns (Bamshad and Wooding, 2003, Fumagalli and Sironi, 2014).

Not only do the NP and SW tribal populations exhibit different patterns of CYP2A6 allele frequencies, but more specifically, the NP population includes fewer CYP2A6 genotype-reduced nicotine metabolizers. Considering the strong association that we observed between the CYP2A6 genotype and activity, this suggests that NP smokers will have a faster rate of nicotine metabolism. As previously stated, faster metabolism has been associated with heavier smoking and higher risk of lung cancer (Schoedel et al., 2004, Schnoll et al., 2009, Wassenaar et al., 2011, Wassenaar et al., 2014), consistent with characteristics of the NP compared with the SW tribal population (Eichner et al., 2010, Bliss et al., 2008, Plescia et al., 2014).

Phenotypic measurements of CYP2A6 activity also confirmed that the NP tribal population has an elevated rate of nicotine metabolism relative to the SW tribal population and compared with AN, White, and African American populations. This observation was present in the overall population of smokers, but also when excluding smokers possessing known CYP2A6 decrease-of-function or loss-of-function alleles. The latter finding suggests that the observed elevated rate of nicotine metabolism among NP smokers was not a result of a lower frequency of reduced metabolizers, but rather other factors are contributing toward this, as discussed below. Faster nicotine metabolism may contribute toward an elevated risk for smoking, and tobacco-related disease in the NP relative to the SW as well as in comparison to other ethnic populations in the USA.

To investigate potential contributors to NMR, we performed linear regression modeling in the tribal populations. The CYP2A6 genotype and BMI were the only significant independent contributors to NMR in the combined population sample. However, in this model, tribal population itself trended toward significance, which suggests that there may be unaccounted for differences between the regional tribal populations that explain the observed higher NMR among the NP compared to the SW tribal population. When we investigated each tribal population separately, we found that the CYP2A6 genotype was the only significant independent contributor to NMR in both the NP and SW tribal populations. Therefore, we investigated whether a possible explanation for the high NMR in the NP tribal population could be their relatively high frequency of the CYP2A6*1B allele (69.7%) compared with the *1B allele frequency in the SW tribal population (61.6%) and other ethnic groups (ANs 65%,
Whites 32%, African Americans 14%) (Binnington et al., 2012, Mwenifumbo and Tyndale, 2007). CYP2A6*1B was associated with higher NMR in both tribal populations, and in previous studies, with the mechanism believed to be through improving CYP2A6 mRNA stabilization, leading to increased CYP2A6 enzyme expression and activity (Mwenifumbo et al., 2008b, Wang et al., 2006). However, as NP smokers possessing the *1B/*1B genotype had significantly higher NMR compared with SW smokers with the same genotype, this suggests that the tribal differences in NMR are not resulting from the higher frequency of *1B in the NP compared to the SW tribal population. As further evidence of this, our linear regression model run only in wild-type smokers indicated that the CYP2A6*1B genotype and tribal population were significant independent predictors of NMR. This suggests that, independent of CYP2A6*1B genotype, the NP tribal population shows higher NMR than the SW tribal population. We also confirmed that there was no interaction between the CYP2A6*1B genotype and tribal population in influencing NMR, suggesting that the relationship between the CYP2A6*1B genotype and NMR is similar in both tribal populations and therefore other factors, unrelated to CYP2A6*1B, are contributing to tribal differences in NMR.

Considering the highly polymorphic nature of the CYP2A6 gene (http://www.cypalleles.ki.se/cyp2a6.htm), there may be novel CYP2A6 genetic variants present in the NP tribal population that act to increase CYP2A6 gene and protein expression, and/or enzyme activity. For example, in the NP population there may be undetected genetic variation present at the 5’ promoter region of the CYP2A6 gene that interrupts the binding site for an inhibitory transcription factor, such as C/EBPβ, or creates a binding site for an activating transcription factor, such as HNF4α, thus decreasing or increasing CYP2A6 levels (Pitarque et al., 2005). This may result in greater CYP2A6 promoter activity and thus gene transcription. Variation in regions 3’ of the CYP2A6 gene, similar to the *1B allele, may contribute toward increased mRNA stability and thus increased protein expression (Wang et al., 2006), contributing toward higher NMR among NP relative to SW smokers. The significantly higher frequency of CYP2A6*9 in the SW compared to NP supports the concept of ethnicity, or tribal group, specific genetic variation and variant distribution. Assessment of both common and rare variation at the CYP2A6 locus, and in upstream and downstream sequence, by whole CYP2A6 gene sequencing would improve our understanding of the inter-tribe differences in CYP2A6 and nicotine metabolism. Additionally, environmental factors can also modulate CYP2A6 activity and the rate of nicotine metabolism. For example, CYP2A6 is inducible by dietary and
pharmacological agents such as broccoli, rifampin, and phenobarbital (Hakooz and Hamdan, 2007, Rae et al., 2001, Itoh et al., 2006), and therefore potentially by other dietary components of the NP tribal population.

Implications of having a faster rate of nicotine metabolism, as seen in the NP tribal population, include having greater tobacco consumption, nicotine dependence scores, difficulty quitting smoking, and lung cancer risk (Schoedel et al., 2004, Schnoll et al., 2009, Wassenaar et al., 2011, Wassenaar et al., 2014). Greater activation of tobacco-specific procarcinogens, such as nitrosamines, because of faster CYP2A6 activity, may explain the higher lung cancer incidence and mortality rates in the NP compared with the SW (Bliss et al., 2008, Plescia et al., 2014, Wong et al., 2005), however our study did not specifically assess this.

When investigating smoking behaviors, we observed that NP smokers consume more CPD (7 vs. 4), have higher nicotine dependence scores (FTND 2.0 vs. 1.8, HONC 4.5 vs. 3.7), and have longer durations of smoking (23 vs. 20 years) (unpublished data from current study; authors JAT, JAH, DB, BVH, PNH, RFT) compared with SW smokers. However, we did not observe an association between the CYP2A6 genotype and NMR with smoking behaviors in either tribal population. This highlights some limitations of our study. First, this lack of association may have resulted from limited statistical power to detect smoking behavior differences between CYP2A6 genotype or NMR groups due to low sample sizes, particularly of the smokers. The sample size calculations were based on a separate candidate-gene study, conducted as part of the broader ‘Topography and Genetics of Smoking and Nicotine Dependence in American Indians’ study. Additionally, we were limited by available biomarkers and measures of tobacco consumption. The use of a more sensitive biomarker of tobacco consumption, such as TNE, may allow us to detect subtle differences in smoking among these two light smoking populations in the future. Moreover, titration of nicotine intake by CPD according to rate of nicotine metabolism may not be observed in these AI/AN populations, similarly to African American smokers (Ross et al., 2016, Ho et al., 2009b), but titration may be detected with TNE as measure of consumption, as found among Alaska Native smokers (Zhu et al., 2013a).
**Conclusion**

We have identified distinct patterns of CYP2A6 genetic variation and rates of nicotine metabolism between two AI/AN tribal populations with different smoking patterns and risk for lung cancer. Given the association between faster nicotine metabolism and greater smoking and lung cancer risk, these findings suggest that CYP2A6 may contribute toward an elevated risk for smoking, and tobacco-related disease in the NP compared with the SW and compared with other ethnic populations in the USA. Future work should focus on determining the cause(s) of the faster nicotine metabolism in the NP and assess any causal relationships between their increased rate of metabolism and lung cancer risk and differences in smoking behaviors including cessation, which may in turn inform strategies to mitigate the elevated risk (Lerman et al., 2015, Wassenaar et al., 2011).
3.6 Significance to Thesis

This chapter adds to the existing literature on racial/ethnic variation in nicotine metabolism rate. This is the first study to compare CYP2A6 genetics and nicotine metabolism between these two American Indian tribal populations, and it contributes two key findings to the literature. First, this chapter highlights the distinct patterns of CYP2A6 genetic variation across the two tribes. The second major contribution of this work is the identification of an overall faster rate of nicotine metabolism in the Northern Plains compared to the Southwest population. Together, these findings suggest that differences in CYP2A6 activity may contribute to the tribes’ different smoking and lung cancer risks. In a number of other ethnicities, faster CYP2A6 genotypes and phenotypes have been associated with increased tobacco consumption, which leads to higher exposure to tobacco-specific carcinogens. Moreover, faster CYP2A6 is associated with higher exposure to these chemicals, even when controlling for the level of tobacco consumption, due to greater activation of procarcinogenic tobacco-specific nitrosamines.

Knowledge obtained from this study could eventually be used to help inform cessation strategies in these populations. A recent smoking cessation clinical trial showed that the NMR can be used to optimize cessation success by informing the selection of the appropriate treatment. As the Northern Plains have a higher frequency of faster nicotine metabolizers, this implies that more individuals in this population may benefit from using varenicline to support cessation, whereas cessation may be improved among the many slower metabolizers in the Southwest through the use of the nicotine patch.

The findings outlined in this chapter support future work aimed at investigating novel sources of variation in nicotine metabolism, which may account for the overall faster rate of nicotine metabolism in the Northern Plains relative to the Southwest tribe.
4 CHAPTER 3: PREDICTORS OF VARIATION IN CYP2A6 mRNA, PROTEIN, AND ENZYME ACTIVITY IN A HUMAN LIVER BANK: INFLUENCE OF GENETIC AND NON-GENETIC FACTORS

Julie-Anne Tanner, Bhagwat Prasad, Katrina G. Claw, Patricia Stapleton, Amrit Chaudhry, Erin G. Schuetz, Kenneth E. Thummel, and Rachel F. Tyndale

Reprinted with permission of the American Society for Pharmacology and Experimental Therapeutics: Journal of Pharmacology and Experimental Therapeutics 360(1): 129-139. Copyright © 2016 by the American Society for Pharmacology and Experimental Therapeutics. All rights reserved.

RFT, KET, and EGS designed the overall study. KGC conducted RNA-sequencing to quantify CYP2A6, POR, and AKR1D1 mRNA. BP quantified CYP2A6 and POR protein. Based on the available literature, JAT identified remaining questions regarding sources of variation in CYP2A6 activity and nicotine metabolism rate, and determined experimental questions. JAT conducted the CYP2A6 genotyping, and optimized and performed in vitro CYP2A6 activity assays to determine the rates of nicotine and coumarin metabolism. JAT created a study database, and used this to perform modeling and statistical analyses. All authors interpreted the results. JAT and RFT wrote the manuscript.
4.1 Abstract

Cytochrome P450 2A6 (CYP2A6) metabolizes several clinically relevant substrates, including nicotine, the primary psychoactive component in cigarette smoke. Smokers vary widely in their rate of inactivation and clearance of nicotine, altering numerous smoking phenotypes. We aimed to characterize independent and shared impact of genetic and non-genetic sources of variation in CYP2A6 mRNA, protein, and enzyme activity in a human liver bank (n=360). For the assessment of genetic factors, we quantified levels of CYP2A6, cytochrome P450 oxidoreductase (POR), and aldo-keto reductase 1D1 (AKR1D1) mRNA, and CYP2A6 and POR proteins. CYP2A6 enzyme activity was determined through measurement of cotinine formation from nicotine and 7-hydroxycoumarin formation from coumarin. Donor DNA was genotyped for CYP2A6, POR, and AKR1D1 genetic variants. Non-genetic factors assessed included gender, age, and liver disease. CYP2A6 phenotype measures were positively correlated to each other (r values ranging from 0.47-0.88, P<0.001). Female donors exhibited higher CYP2A6 mRNA expression relative to males (P<0.05). Donor age was weakly positively correlated with CYP2A6 protein (r=0.12, P<0.05) and activity (r=0.20, P<0.001). CYP2A6 reduce-of-function genotypes, but not POR or AKR1D1 genotype, were associated with lower CYP2A6 protein (P<0.001) and activity (P<0.01). AKR1D1 mRNA was correlated with CYP2A6 mRNA (r=0.57, P<0.001), protein (r=0.30, P<0.001), and activity (r=0.34, P<0.001). POR protein was correlated with CYP2A6 activity (r=0.45, P<0.001). Through regression analyses, we accounted for 17% (P<0.001), 37% (P<0.001), and 77% (P<0.001) of the variation in CYP2A6 mRNA, protein, and activity, respectively. Overall, several independent and shared sources of variation in CYP2A6 activity in vitro have been identified, which could translate to variable hepatic clearance of nicotine.
4.2 Introduction

Cytochrome P450 2A6 (CYP2A6) metabolizes several clinically relevant substrates, including nicotine, tegafur, letrozole, efavirenz, valproic acid, and pilocarpine (Kiang et al., 2006, Messina et al., 1997, Endo et al., 2007, Ikeda et al., 2000, Murai et al., 2009, di Iulio et al., 2009). Nicotine metabolism by CYP2A6 is of interest as nicotine is the primary psychoactive component in cigarette smoke and the main source of tobacco dependence (Stolerman and Jarvis, 1995). Smokers can vary widely in their rate of inactivation and clearance of nicotine (Malaiyandi et al., 2005), with variations in the rate of nicotine metabolism associating with differences in smoking behaviors, cessation, and response to cessation pharmacotherapies (Schoedel et al., 2004, Schnoll et al., 2009, Patterson et al., 2008). The major pathway of nicotine inactivation is its conversion to cotinine, primarily catalyzed by CYP2A6 (Messina et al., 1997).

Genetic and non-genetic factors contribute to variation in CYP2A6 enzyme activity and rate of nicotine metabolism. The CYP2A6 gene, encoding the CYP2A6 enzyme, is highly polymorphic, and CYP2A6 genetic variation is associated with variable rates of nicotine metabolism in vitro and in vivo (Al Koudsi et al., 2010, Binnington et al., 2012), and accordingly with differences in smoking behavior (Wassenaar et al., 2011, Chen et al., 2014). A significant proportion of variation in CYP2A6 activity can be attributed to CYP2A6 genetic variation (heritability estimates of 60-80%) (Swan et al., 2009, Loukola et al., 2015), however unaccounted for variation remains even after additional factors (age, sex, race/ethnicity, body mass index, cigarettes per day, and total nicotine equivalents) are controlled for (Park et al., 2016, Chenoweth et al., 2014a). Unidentified CYP2A6 genetic variation, or variation in other genes regulating CYP expression or function, may account for the 35% missing variation.

Aldo-keto reductase 1D1 (AKR1D1), an enzyme involved in bile acid synthesis (Schuetz et al., 2001, Lee et al., 2009), has been identified as a potential regulator of CYP activity (Yang et al., 2010). AKR1D1 single nucleotide polymorphism (SNP) rs1872930, associated with higher AKR1D1 mRNA expression, is associated with increased expression and activity of several CYPs in vitro (Chaudhry et al., 2013). To our knowledge, the relationship between rs1872930 and CYP2A6 has not yet been investigated. NADPH-cytochrome P450 oxidoreductase (POR), an enzyme that donates electrons to CYPs during their catalytic cycle (Hu et al., 2012), also
contributes to variation in multiple CYP activities, including those of CYP2A6 (Gomes et al., 2009, Chenoweth et al., 2014b, Lv et al., 2016).

Combined effects of CYP2A6, AKR1D1, and POR genetic variation should also be considered. POR SNP rs1057868 interacts with CYP2A6 genotype, with rs1057868 only associating with faster CYP2A6 activity among individuals not possessing known CYP2A6 reduced-function genetic variants (Chenoweth et al., 2014b). Owing to potential shared effects of each genetic factor on CYP2A6 activity, we aimed to elucidate both the independent and combined influences of CYP2A6, POR, and AKR1D1 genetic variation on CYP2A6 mRNA expression, protein levels, and enzyme activity. We also aimed to investigate relationships between POR and AKR1D1 expression, independent of genotype, with variation in CYP2A6 expression and activity.

Several non-genetic factors, including gender, age, and liver disease, have been associated with altered CYP2A6 expression, function, and nicotine pharmacokinetics. Here we will examine independent and combined impacts of these variables in a large human liver bank (n=360) with extensive CYP2A6 phenotyping. Relative to men, females had higher microsomal CYP2A6 protein levels in a smaller human liver bank (n=67) (Al Koudsi et al., 2010), and women smokers exhibited greater CYP2A6-mediated nicotine metabolism (Benowitz et al., 2006), consistent with estrogen-mediated induction of CYP2A6 transcription (Higashi et al., 2007, Benowitz et al., 2006). Individuals 65 and older have been shown to have lower nonrenal nicotine clearance than those aged 22-43 (Molander et al., 2001); however, there was no association between age and CYP2A6 protein or activity when investigated in a smaller human liver bank (n=67) (Al Koudsi et al., 2010). Nonalcoholic fatty liver disease was associated with higher CYP2A6 mRNA expression and enzyme activity in vitro (Fisher et al., 2009).

Despite several known sources of variation in CYP2A6 activity, substantial variation remains to be characterized (Swan et al., 2009). Considering the clinical relevance of CYP2A6, it is important to further characterize causes of variability in CYP2A6 activity and thus the rate of nicotine metabolism, for its utility in tailoring smoking cessation treatment (Lerman et al., 2015). We aimed to assess independent and shared impacts of CYP2A6, POR, and AKR1D1 genotypes and levels, as well as gender, age, and liver disease, on CYP2A6 mRNA expression, protein expression, and enzyme activity in a large human liver bank.
4.3 Materials and Methods

Chemicals and reagents

Lodoacetamide, dithiothreitol, and sequencing-grade trypsin were purchased from ThermoFisher Scientific (Pierce Protein Biology, Rockford, IL). Ammonium bicarbonate was purchased from Acros Organics (Geel, Belgium). Sodium deoxycholate (98% purity) was obtained from MP Biomedicals (Santa Ana, CA). Synthetic light peptides for CYP2A6 and POR quantification were procured from New England Peptides (Boston, MA), with purity established by amino acid analysis. Heavy stable isotope labeled amino acids, \([^{13}C_6^{15}N_2]\)-lysine and \([^{13}C_6^{15}N_4]\)-arginine, were purchased ThermoFisher Scientific (Pierce Protein Biology, Rockford, IL). Liquid chromatography-mass spectrometry-grade acetonitrile (99.9% purity) and formic acid (≥99.5% purity) were purchased from Fischer Scientific (Fair Lawn, NJ). (−)-Nicotine hydrogen tartrate, (−)-cotinine, coumarin, 7-hydroxycoumarin, and 4-hydroxycoumarin were purchased from Sigma-Aldrich (St. Louis, MO); chemical structures illustrated in Emami and Dadashpour (2015). Nicotine-D4 and cotinine-D3 were purchased from Toronto Research Chemicals (Toronto, ON); chemical structures illustrated in Hukkanen et al. (2005).

Human liver bank

Human liver tissue samples are from two liver banks: 1) the St. Jude Liver Resource at the St. Jude Children’s Research Hospital (Memphis, TN, USA) (n=295), and 2) the University of Washington Human Liver Bank (Seattle, WA, USA) (n=65). The St Jude Liver Resource human liver tissues were obtained through the Liver Tissue Cell Distribution System, Minneapolis, MN, and Pittsburgh, PA, which was funded by National Institutes of Health contract # HHSN276201200017C. Details on the selection of the livers and investigator blinding for sample analyses have been described previously (Shirasaka et al., 2015). Age, gender, and ethnicity were known for most (≥90%) of the liver donors. The donors ranged in age from 0-87 years (mean 40 years, SD ± 22 years). Of the donors with known gender, 58% were male. The liver bank consists of 92% Caucasian, 3% African American, <1% Asian, <1% Hispanic, and 5% unknown ethnicity donors. Cause of death, medications used, and liver pathology was known for less than 50% of donors. Smoking status was unknown for >88% of
donors and therefore was not assessed as a predictor of CYP2A6 phenotypes in the present study

**CYP2A6, POR, and AKR1D1 mRNA quantification**

**RNA Isolation**

Liver RNA was isolated and purified using a NucleoSpin® miRNA kit (Macherey-Nagel, Duren, Germany; Clonetech Labs, Mountain View, CA), according to manufacturer’s protocol. Briefly, ~30 mg liver tissue was combined with 4°C Lysis buffer, homogenized using a TissueLyser LT (Qiagen, Valencia, CA), and allowed to sit at room temperature for 5 minutes. The solution was then added to a column. After centrifugation to bind the large RNA to the column, the column was treated with an rDNase solution at room temperature for at least 15 minutes. Meanwhile, the flow-through containing the small RNA was treated, in order to precipitate out the protein. The small RNA was then bound to a new column. Following three wash steps of each column, the resulting large and small RNA were each eluted, quantitated and bioanalyzed for quality control using a Bioanalyzer 2100 (Agilent, Santa Clara, CA). Only RNA with an RNA integrity score greater or equal to 7.0 was submitted for sequencing.

**TruSeq Stranded mRNA Preparation**

Next-generation sequencing libraries were prepared from 1.25 µg of total RNA in an automated, high-throughput format using the TruSeq Stranded mRNA kit (Illumina, San Diego, CA). All the steps required for sequence library construction have been automated and performed on a Sciclone NGSx Workstation (Perkin Elmer, Waltham, MA). During library construction, ribosomal RNA was depleted by means of a poly-A enrichment and first and second strand cDNA syntheses were performed. Each library was then uniquely barcoded using the Illumina adapters and amplified using a total of 13 cycles of PCR. After amplification and cleanup, library concentrations were quantified using the Quant-it dsDNA Assay (Life Technologies/ThermoFisher Scientific, Carlsbad, CA). Libraries were subsequently normalized and pooled based on Agilent 2100 Bioanalyzer results (Agilent Technologies, Santa Clara, CA). Pooled libraries were size selected using a Pippin Prep (Sage Science, Beverly, MA) and then balanced by mass and pooled in batches of 96 with a final pool concentration of 2-3 nM for sequencing on the HiSeq 2500.
Read Processing and Analysis Pipeline

The Northwest genome sequencing laboratory (University of Washington, Seattle, WA) processing pipeline included the following elements: 1) base calls generated in real-time on the HiSeq or NextSeq instrument; 2) Illumina RTA-generated BCL files converted to FASTQ files; 3) custom scripts developed in-house and used to process the FASTQ files and to output de-multiplexed FASTQ files by lane and index sequence; 4) sequence read and base quality checked using the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/); 5) sequences aligned to hg19 with reference transcriptome Ensembl v67 (http://www.ensembl.org) using Tophat (Johns Hopkins University Center for Computational Biology) (Kim et al., 2013) followed by mate-fixing; and 6) custom scripts for quality assessment generate metrics. All aligned read data were subject to the following steps: 1) lane level bam data files were merged using the Picard MergeSamFiles tool and suspected PCR duplicates were marked, not removed, in the alignment files using the Picard MarkDuplicates tool (http://broadinstitute.github.io/picard/); 2) local realignment performed around indels, and base quality score recalibration was run using GATK tools (Broad Institute) (McKenna et al., 2010); 3) variant detection performed with the GATK Unified Genotyper version 2.6.5 (DePristo et al., 2011); (4) aligned data were used for isoform assembly and quantitation with Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/) (Kim et al., 2013, Trapnell et al., 2013); genomic features were quantitated with featureCounts (Walter and Eliza Hall Institute, Parkville Victoria, Australia) (Liao et al., 2014); and 5) gene-specific quantitation data were used for further analysis.

CYP2A6 and POR protein quantification

Simultaneous quantification of CYP2A6 and POR was carried out using a LC-MS/MS proteomics method (Prasad and Unadkat, 2014). The surrogate peptides were selected and light and heavy peptides (Table 1) containing labeled $^{13}C_6^{15}N_2$-lysine or $^{13}C_6^{15}N_4$-arginine residues were procured. Liver microsomal samples were diluted to 2 mg/ml, and 40 μg microsomal protein was digested as described before (Shuster et al., 2014). Briefly, microsomal protein was denatured and reduced with 4 μl of 100 mM dithiothreitol, 10 μl of sodium deoxycholate (2.6 % w/v) and 10 μl of ammonium bicarbonate buffer (100 mM) at 95°C for 5 min. The denatured protein was then alkylated by 4 μl of 200 mM iodoacetamide at room temperature. The digestion was performed by addition 10 μl of trypsin (protein:trypsin
ratio, 25:1) at 37°C for 22 hours. The reaction was quenched by the addition of 20 μl of peptide internal standard cocktail (prepared in 50% acetonitrile in water containing 0.1% formic acid) and 10 μl of the neat solvent, i.e., 50% acetonitrile in water containing 0.1% formic acid. The samples were vortexed and centrifuged at 3500 × g for 5 min. The calibration curves were generated using serial dilutions of light peptide standard in phosphate buffer (50 mM Kpi, 0.25 M sucrose, 10 mM EDTA, pH 7.4) to replace microsomal sample.

Triple-quadrupole LC-MS instrument (Agilent 6460A) coupled to an Agilent 1290 Infinity LC system (Agilent Technologies), in ESI positive ionization mode was used for quantification. 2 μg of the trypsin digest was injected onto the column (Kinetex 1.7 μ, C18 100A; 100 × 2.1 mm, Phenomenex, Torrance, CA). Mobile phase and gradient program were exactly the same as described before (Shuster et al., 2014). Surrogate light and heavy (internal standards) peptides were monitored using instrument parameters provided in Table 11. The LC-MS/MS data were processed using MassHunter (Agilent Technologies) and Skyline (University Of Washington) software.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Peptide</th>
<th>Precursor Ion (m/z)</th>
<th>Product Ion (m/z)</th>
<th>Fragmentor (V)</th>
<th>Collision Energy (eV)</th>
<th>LLOQ (fmol on-column)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP2A6</td>
<td>GTGGANIDPTFFLSR</td>
<td>777.1</td>
<td>982.5</td>
<td>130</td>
<td>22</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>GTGGANIDPTFFLSR</td>
<td>777.1</td>
<td>867.2</td>
<td>130</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>781.9</td>
<td>992.5</td>
<td>130</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>CYP reductase</td>
<td>FAVFGLGNK</td>
<td>476.9</td>
<td>734.3</td>
<td>100</td>
<td>9</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>FAVFGLGNK</td>
<td>476.9</td>
<td>635.3</td>
<td>100</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

Underlined residues represent the internal standards with the labeled [13C615N2]-lysine or [13C615N4]-arginine.
CYP2A6 enzyme activity assays

Human liver microsomes were prepared, and total protein concentrations were quantified, as described previously (Shirasaka et al., 2015). CYP2A6 enzyme activity was determined by quantifying the rate of metabolism of two known substrates of this enzyme, nicotine and coumarin. Linear conditions for the rate of nicotine metabolism (i.e. the rate of cotinine formation from nicotine) were established for the following assay conditions: 0.5 mg/ml microsomal protein, 50 µM Tris-HCl buffer (pH 7.4), 30 µM nicotine, 1 mM NADPH, 10 µl cytosol (source of aldehyde oxidase), and water to a final volume of 100 µl, for an incubation time of 20 minutes at 37°C. The reaction was terminated with 20 µl of 20% Na₂CO₃, and 20 ng of nicotine-D₄ and cotinine-D₃ internal standards were added. Samples were extracted and analyzed using LC-MS/MS, as described previously (Jacob et al., 2011, Craig et al., 2014).

In order to allow for adequate detection of 7-hydroxycoumarin and avoid substrate depletion, linear conditions for the rate of coumarin metabolism (i.e. the rate of 7-hydroxycoumarin formation from coumarin, a very rapidly metabolized CYP2A6 substrate) were established at multiple incubation times and protein concentrations using a number of substrate concentrations. For donors who exhibited a relatively slow rate of *in vitro* nicotine metabolism (cotinine formation velocity of <0.1 nmol/min/mg), 0.05 mg/ml microsomal protein and 15 min incubation was used. For donors exhibiting intermediate rates of nicotine metabolism (cotinine formation velocity of 0.1-0.3 nmol/min/mg), the assay was adjusted to 0.02 mg/ml microsomal protein for a 10 minute incubation, and for donors with fast rates of nicotine metabolism (cotinine formation velocity of >0.3 nmol/min/mg), 0.01 mg/ml microsomal protein for a 7 min incubation was used. All other assay conditions were identical among slow, intermediate, and fast metabolizers, including 50 mM Tris-HCl buffer (pH 7.4), 2 µM coumarin, 1 mM NADPH, and water to a final volume of 200 µl with an incubation at 37°C. The reactions were terminated with 40 µl of trichloroacetic acid (20% w/v), and 25 ng of 4-hydroxycoumarin internal standard was added. Samples were extracted and analyzed using high performance liquid chromatography, as described previously (Li et al., 1997), with minor modifications. Limits of quantification for nicotine, cotinine, coumarin, and 7-hydroxycoumarin were 1 ng/ml, 1 ng/ml, 50 ng/ml, and 10 ng/ml, respectively.
**CYP2A6, POR, and AKR1D1 genotyping**

DNA was extracted using the DNeasy tissue kit. DNA from all donors was genotyped for the CYP2A6 alleles *2, *4, *9, and *12, whereas DNA from African American and unknown ethnicity donors were also genotyped for CYP2A6 *17, *20, *23, *25, *28, and *35, and Asian and unknown ethnicity donors were additionally genotyped for CYP2A6 *7, *8, and *10, as these alleles have zero to extremely low frequencies among Caucasians (Mwenifumbo and Tyndale, 2007). Genotyping for the CYP2A6 alleles was conducted using a two-step allele-specific polymerase chain reaction approach, except for CYP2A6*2, which was genotyped using a TaqMan SNP genotyping assay (Applied Biosystems) and real-time polymerase chain reaction; CYP2A6 genotyping approaches have been described in detail previously (Wassenaar et al., 2016). All donors were genotyped for POR SNPs rs17148944, rs2868177, rs1057868 and AKR1D1 SNP rs1872930 using allele-specific TaqMan SNP genotyping assays (Applied Biosystems).

**Statistical Analyses**

The following data were non-normally distributed, and therefore non-parametric statistical tests were used: CYP2A6 mRNA, AKR1D1 mRNA, POR mRNA, CYP2A6 protein, POR protein, nicotine metabolism, and coumarin metabolism. Correlations were determined via Spearman rank correlations. We used the Mann Whitney test to analyze the association of gender, liver disease, and AKR1D1 genotype with CYP2A6 mRNA, CYP2A6 protein, and CYP2A6 activity. The Kruskal-Wallis test was used to determine associations between CYP2A6 genotype and CYP2A6 mRNA, CYP2A6 protein, and CYP2A6 activity, as well as between POR SNP genotypes and POR mRNA, POR protein, and CYP2A6 activity. We ran separate linear regression models for CYP2A6 mRNA, CYP2A6 protein, and CYP2A6 activity to calculate the proportion of variation in each that is accounted for by each variable in the model. The liver bank is comprised of samples collected and processed at two different research sites; the phenotype means were different between sites. For the purposes of illustrating the combined data we have corrected for overall site differences using the following conversion factor: mean phenotype measure (CYP2A6 mRNA, AKR1D1 mRNA, POR mRNA, CYP2A6 protein, POR protein, rate of nicotine metabolism, and rate of coumarin metabolism) from University of Washington site divided by the mean phenotype measure from St. Jude site, which was multiplied by each sample from the University of Washington site (the smaller liver
bank). Findings are illustrated using the site-corrected measures. Additional regression models were run in which the original, non-corrected phenotype data was included, along with the liver bank site (University of Washington or St. Jude) as a covariate. Both versions of each model (corrected vs. uncorrected data with site as covariate) produced very similar results, indicating that the conversion factor had a minimal effect but allowed for combining the data sets for illustration purposes. Analyses were conducted with GraphPad Prism (v6.0) and SPSS (v22), and statistical tests were considered significant for $P<0.05$.

4.4 Results

Moderate to strong correlation between CYP2A6 mRNA, protein, and activity

The two *in vitro* measures of CYP2A6 activity, velocity of cotinine formation from nicotine, and velocity of 7-hydroxycoumarin formation from coumarin, were strongly positively correlated (Spearman $r=0.86$, $P<0.001$; Fig. 19a). The rate of nicotine metabolism serves as the primary measure of CYP2A6 activity in all analyses, and all coumarin metabolism data has been included as an additional measure of enzyme activity. CYP2A6 mRNA and protein expression were moderately positively correlated (Spearman $r=0.47$, $P<0.001$; Fig. 19b). CYP2A6 protein was strongly positively correlated with both measures of CYP2A6 activity (nicotine metabolism: Spearman $r=0.88$, $P<0.001$; Fig. 19c; coumarin metabolism: Spearman $r=0.81$, $P<0.001$; Fig. 19d). All correlations were similar among males and females, and among wild-type (*CYP2A6 *1/*1*) donors and the whole sample (donors with wild-type and variant genotypes included); the rate of nicotine and coumarin metabolism were strongly correlated in females ($r=0.86$, $P<0.001$) and males ($r=0.86$, $P<0.001$); CYP2A6 mRNA and CYP2A6 protein were moderately correlated in females ($r=0.53$, $P<0.001$) and males ($r=0.42$, $P<0.001$); CYP2A6 protein and CYP2A6 enzyme activity (cotinine formation) were strongly correlated in females ($r=0.88$, $P<0.001$) and males ($r=0.88$, $P<0.001$) (Appendix E); CYP2A6 mRNA and protein for the wild-type donors only were moderately correlated ($r=0.50$, $P<0.001$), CYP2A6 mRNA and enzyme activity (cotinine formation from nicotine) were moderately correlated ($r=0.49$, $P<0.001$), and CYP2A6 protein and enzyme activity were strongly correlated ($r=0.87$, $P<0.001$). The mean, standard deviation, and range for each CYP2A6 phenotype measure were as follows: CYP2A6 mRNA $373 \pm 481$ (0.2-4940) FPKM values, CYP2A6 protein $22.5 \pm 19.4$ (0.0-121.1) pmol/mg, CYP2A6 activity (nicotine
metabolism) 0.10 ± 0.10 (0.0007-0.55) nmol/min/mg, and CYP2A6 activity (coumarin metabolism) 0.52 ± 0.68 (0.01-4.51) nmol/min/mg.

Figure 19 | (A) Correlation of two measures of CYP2A6 enzyme activity (the velocity of cotinine formation from nicotine versus the velocity of 7-OH-coumarin formation from coumarin, nmol/min per milligram). (B) Correlation between CYP2A6 mRNA levels (FPKM values, fragments per kilobase per million reads) and CYP2A6 protein levels (pmol/mg). (C) Correlation between CYP2A6 protein levels and CYP2A6 enzyme activity (cotinine formation from nicotine). (D) Correlation between CYP2A6 protein levels (pmol/mg) and CYP2A6 enzyme activity (7-OH coumarin formation from coumarin, nmol/min/ per milligram). Values for r and P were determined on the basis of Spearman correlations.
Impact of non-genetic factors on CYP2A6 mRNA, protein, and activity

Gender was associated with differences in CYP2A6 mRNA expression, protein levels, and enzyme activity. Female liver donors exhibited higher CYP2A6 mRNA levels compared to males ($P<0.05$; Fig. 20a). Although not statistically significant, there was a trend for higher CYP2A6 protein ($P<0.1$; Fig. 20b) and enzyme activity ($P<0.1$ for both nicotine and coumarin metabolism; Fig. 20c and Fig. 20d) among females relative to males.

![Graphs showing gender differences in CYP2A6 mRNA, protein, and enzyme activity](image)

**Figure 20** | Association of gender with (A) CYP2A6 mRNA levels (FPKM values), (B) CYP2A6 protein levels (pmol/mg), (C) CYP2A6 enzyme activity (cotinine formation from nicotine, nmol/min per milligram), and (D) CYP2A6 enzyme activity (7-OH coumarin formation from coumarin, nmol/min per milligram). *P values of $<0.05$ on the basis of Mann Whitney tests.
There was no association between donor age and CYP2A6 mRNA expression (Spearman \( r=0.04, P>0.1; \textbf{Fig. 21a} \)), while age was weakly positively correlated with both CYP2A6 protein expression (Spearman \( r=0.12, P<0.05; \textbf{Fig. 21b} \)) and enzyme activity (nicotine metabolism: Spearman \( r=0.20, P<0.001; \textbf{Fig. 21c} \); coumarin metabolism: Spearman \( r=0.13, P<0.05; \textbf{Fig. 21d} \)).

Liver disease was defined as being positive for at least one of the following conditions: hepatitis, liver injury, biliary atresia, cirrhosis, fat accumulation, fibrosis, or hepatoma. There was no association between liver disease state and CYP2A6 mRNA (normal vs. disease, means=383 and 299 FPKM values, respectively, \( P>0.1 \)), protein (normal vs. disease, means=21.6 and 20.0 pmol/mg, respectively, \( P>0.1 \)), or enzyme activity (normal vs. disease, nicotine means=0.10 and 0.09 nmol/min/mg, respectively, coumarin means=0.55 and 0.47 nmol/min/mg, respectively, \( P>0.1 \) for both nicotine and coumarin metabolism).
Figure 21 | Correlation of age with (A) CYP2A6 mRNA levels (FPKM values), (B) CYP2A6 protein levels (pmol/mg), (C) CYP2A6 enzyme activity (cotinine formation from nicotine, nmol/min per milligram), and (D) CYP2A6 enzyme activity (7-OH coumarin formation from coumarin, nmol/min per milligram). Values for $r$ and $P$ determined on the basis of Spearman correlations.
CYP2A6 genotype is associated with CYP2A6 protein and enzyme activity, but not CYP2A6 mRNA

The CYP2A6 genetic variants *2, *4, *9, *10, *12, and *17 were identified among one or more liver donors, while *7, *8, *20, *23, *25, *28, or *35 were not. Liver donors were grouped according to their CYP2A6 genotype (e.g. *1/*9), where wild-type (*1/*1) donors were those who did not possess any tested CYP2A6 genetic variant, and the “all variants” group included all donors that possessed one or more CYP2A6 genetic variant. CYP2A6 mRNA expression did not differ across all genotypes (P>0.1; Fig. 2a), and there was no difference between the wild-type and the all variants groups (P>0.1; Fig. 2a). There was an apparent gene-dose effect on CYP2A6 mRNA expression with increasing copies of the CYP2A6*9 allele (mean FPKM values: *1/*1 386, *1/*9 326, *9/*9 213); this TATA box variant has previously been associated with decreased CYP2A6 transcription (Pitarque et al., 2001). Although no individual genotypes were significantly different with respect to CYP2A6 mRNA, protein, or activity, CYP2A6 genotype was associated with differences in CYP2A6 protein and activity such that liver donors possessing one or more CYP2A6 genetic variant exhibited lower CYP2A6 protein expression (P<0.001; Fig. 2b) and enzyme activity (P<0.01 for both nicotine and coumarin metabolism; Fig. 2c and Fig. 2d) relative to the wild-type CYP2A6*1/*1 donors. Additionally, there was a gene-dose effect on the CYP2A6 phenotypes for CYP2A6*9 (CYP2A6 protein: *1/*1 24.3, *1/*9 20.8, *9/*9 8.9 pmol/mg; CYP2A6 activity (nicotine metabolism): *1/*1 0.11, *1/*9 0.09, *9/*9 0.03 nmol/min/mg; Fig. 2b-c) and CYP2A6*2 (CYP2A6 protein: *1/*1 24.3, *1/*2 13.0, *2/*2 3.0 pmol/mg; CYP2A6 activity (nicotine metabolism): *1/*1 0.11, *1/*2 0.06, *2/*2 0.01 nmol/min/mg; Fig. 2b-c). A wide degree of variation in CYP2A6 mRNA (0.9-4940 FPKM values), protein (0.0-121.1 pmol/mg), and enzyme activity (nicotine metabolism: 0.002-0.55 nmol/min/mg; coumarin metabolism: 0.02-4.51 nmol/min/mg) was observed in the wild-type genotype group.
Figure 22 | Association of CYP2A6 genotype with (A) CYP2A6 mRNA levels (FPKM values), (B) CYP2A6 protein levels (pmol/mg), (C) CYP2A6 enzyme activity (cotinine formation from nicotine, nmol/min per milligram), and (D) CYP2A6 enzyme activity (7-OH coumarin formation from coumarin, nmol/min per milligram). Horizontal lines represent the mean for each genotype. Three data points exceed the y-axis limit for CYP2A6*1/*1 group in (A), two data points exceed the y-axis limit for CYP2A6*1/*1 group in (B), two data points exceed the y-axis limit for CYP2A6*1/*1 group in (C), and five data points exceed the y-axis limit in (D); all points were included in the mean and statistical tests. **P<0.01, ***P<0.001, on the basis of Mann Whitney tests.
AKR1D1 mRNA expression, but not genotype, is associated with CYP2A6 mRNA, protein, and activity

Based on work by Chaudhry and colleagues (Chaudhry et al., 2013, Yang et al., 2010), we investigated the correlation between AKR1D1 mRNA expression and CYP2A6 mRNA, protein, and enzyme activity, and further the association between the AKR1D1 SNP rs1872930 with AKR1D1 mRNA expression, CYP2A6 mRNA, protein, and enzyme activity. AKR1D1 mRNA expression was moderately correlated with CYP2A6 mRNA expression (Spearman r=0.57, \(P<0.001\); Fig. 23a), protein levels (Spearman r=0.30, \(P<0.001\); Fig. 23b), and enzyme activity (nicotine metabolism: Spearman r=0.34, \(P<0.001\); Fig. 23c; coumarin metabolism: Spearman r=0.30, \(P<0.001\); Fig. 23d). There was no difference in the association between AKR1D1 mRNA and CYP2A6 mRNA, protein, or enzyme activity in females versus males; AKR1D1 mRNA and CYP2A6 mRNA were moderately correlated in females (r=0.49, \(P<0.001\)) and males (r=0.60, \(P<0.001\)); AKR1D1 mRNA and CYP2A6 protein were weakly to moderately correlated in females (r=0.26, \(P<0.001\)) and males (r=0.34, \(P<0.001\)); AKR1D1 mRNA and CYP2A6 enzyme activity were moderately correlated in females (r=0.30, \(P<0.001\)) and males (r=0.34, \(P<0.001\)). There was no association between rs1872930 and AKR1D1 mRNA (TT vs TC genotype, means=33.8 and 32.5 FPKM values, respectively, \(P>0.1\)), CYP2A6 mRNA (TT vs TC genotype, means=363 and 389 FPKM values, respectively, \(P>0.1\)), CYP2A6 protein (TT vs TC genotype, means=22.1 and 23.2 pmol/mg, respectively, \(P>0.1\)), or CYP2A6 activity (TT vs TC genotype, nicotine means=0.10 and 0.10 nmol/mg/min, respectively, coumarin means= 0.51 and 0.52 nmol/mg/min, respectively, \(P>0.1\) for both nicotine and coumarin metabolism). The mean, standard deviation, and range of AKR1D1 mRNA expression were follows: 33.7 ± 32.3 (0.1-152) FPKM values.
Figure 23 | (A) Correlation between AKR1D1 mRNA expression (FPKM values) and (A) CYP2A6 mRNA expression (FPKM values), (B) CYP2A6 protein levels (pmol/mg), (C) CYP2A6 activity (cotinine formation from nicotine, nmol/min per milligram), and (D) CYP2A6 activity (7-OH coumarin formation from coumarin, nmol/min per milligram). Values for r and P values determined on the basis of Spearman correlations.
POR protein levels, but not genotype, are positively correlated with CYP2A6 activity

POR protein and CYP2A6 activity were moderately correlated (r=0.45, \textit{P}<0.001 for both nicotine and coumarin metabolism; \textbf{Fig. 2a-b}). Based on the literature (Gomes et al., 2009, Chenoweth et al., 2014b, Lv et al., 2016), we investigated the association of three POR SNPs (rs17148944, rs2868177, rs1057868) with POR mRNA expression, POR protein levels, and, due to the observed association between POR protein and CYP2A6 activity, also with CYP2A6 enzyme activity. \textit{POR} SNP rs2868177 was associated with higher POR mRNA expression (AA, AG, GG genotype means=202, 221, and 240 FPKM values, respectively, \textit{P}<0.05), however there were no other associations between POR genotypes and POR and CYP2A6 phenotypes (\textit{P} values provided in \textbf{Table 12}). Associations between \textit{POR} SNP rs17148944 and POR mRNA, POR protein, and CYP2A6 activity are shown in \textbf{Fig 2c-e} as representative examples. There was a weak correlation between AKR1D1 mRNA and POR mRNA expression (Spearman \textit{r}=0.17, \textit{P}<0.01) and POR protein levels (Spearman \textit{r}=0.21, \textit{P}<0.001), suggesting minimal overlap in the influence of POR and AKR1D1 on CYP2A6 activity. The mean, standard deviation, and range for each POR phenotype measure were as follows: POR mRNA expression 215 ± 89 (95-795) FPKM values, and POR protein 23.7 ± 12.3 (1.2-67.3) pmol/mg.
Figure 24 | (A) Correlation between POR protein levels (pmol/mg) and CYP2A6 activity (cotinine formation from nicotine, nmol/min per milligram). (B) Correlation between POR protein levels (pmol/mg) and CYP2A6 activity (7-OH coumarin formation from coumarin, nmol/min per milligram). Values for r and P determined on the basis of Spearman correlations. (C-E) Association between POR SNP rs17148944 and (C) POR mRNA expression (FPKM values), (D) POR protein levels (pmol/mg) and (E) CYP2A6 activity (cotinine formation from nicotine, nmol/min per milligram). P values for C-E all >0.1, on the basis of Kruskal-Wallis tests.
<table>
<thead>
<tr>
<th>POR mRNA</th>
<th>rs17148944</th>
<th>rs2868177</th>
<th>rs1057868</th>
</tr>
</thead>
<tbody>
<tr>
<td>POR protein</td>
<td>$P=0.50$</td>
<td><strong>$P=0.02$</strong></td>
<td>$P=0.32$</td>
</tr>
<tr>
<td>CYP2A6 activity (nicotine metabolism)</td>
<td>$P=0.75$</td>
<td>$P=0.11$</td>
<td>$P=0.32$</td>
</tr>
<tr>
<td>CYP2A6 activity (coumarin metabolism)</td>
<td>$P=0.99$</td>
<td>$P=0.67$</td>
<td>$P=0.67$</td>
</tr>
</tbody>
</table>

Table 12 | Associations between POR SNPs rs17148944, rs2868177, and rs1057868 with POR mRNA, POR protein, and CYP2A6 activity.
A significant proportion of variation in CYP2A6 mRNA, protein, and activity is accounted for by genetic and non-genetic predictors

Using linear regression analyses, we assessed the individual and combined contribution of genetic and non-genetic variables to variation in CYP2A6 mRNA, protein, and enzyme activity. In each regression model, we included only those variables that were significantly \((P<0.05)\), or trending \((P<0.1)\) toward being, associated with each CYP2A6 phenotype in univariate analyses, or variables that have been previously associated with CYP2A6 mRNA, protein, or enzyme activity in the literature. All models presented here are derived from phenotype data corrected for liver bank site differences, as illustrated above and described in the Methods. We have compared this to models in which uncorrected phenotype data was used with the addition of site as a covariate and found only minor insignificant differences between the two modeling approaches.

When modeling influences on CYP2A6 mRNA expression in the human livers, AKR1D1 mRNA expression was the only significant independent contributor to variation in CYP2A6 mRNA, accounting for 16.0\% of this variation \((P<0.001; \textbf{Table 13})\); neither gender or genotype were significant predictors despite both estrogen and CYP2A6*9 working at a transcriptional level (Pitarque et al., 2001; Higashi et al., 2007). Overall, this model accounted for 17\% of the variation in CYP2A6 mRNA expression \((R^2=0.17, P<0.001; \textbf{Table 13})\).

\textbf{Table 13 | Linear regression analysis of CYP2A6 mRNA expression (FPKM values) \textsuperscript{a,b}.} \\

<table>
<thead>
<tr>
<th>Predictor variable</th>
<th>(B)</th>
<th>(\text{Beta})</th>
<th>95% CI</th>
<th>% variation accounted for</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{CYP2A6*9} Genotype\textsuperscript{d}</td>
<td>-117.9</td>
<td>-0.09</td>
<td>-264.8 to 25.2</td>
<td>0.8%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>Gender (M=1, F=0)</td>
<td>-12.0</td>
<td>-0.01</td>
<td>-118.5 to 98.5</td>
<td>0.01%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>AKR1D1 mRNA (FKPM)</td>
<td>6.0</td>
<td>0.4</td>
<td>4.3 to 7.6</td>
<td>16.0%</td>
<td>\textit{&lt;0.001}</td>
</tr>
</tbody>
</table>

\textsuperscript{a.} Cases excluded pairwise \\
\textsuperscript{b.} \(R^2=0.17, P<0.001\) \\
\textsuperscript{c.} % variation accounted for by each variable is determined by: (Part Correlation)\(^2\) x 100 \\
\textsuperscript{d.} \textit{CYP2A6*9} genotype coded as 0 for \textit{CYP2A6*1/*1} genotype, 1 for \textit{CYP2A6*1/*9} genotype, and 2 for \textit{CYP2A6*9/*9} genotype.
In order to investigate factors that influence CYP2A6 protein levels, we ran two versions of the regression model, one without and one with CYP2A6 mRNA included as a predictor variable, in order to identify factors that influence CYP2A6 protein independently or as a byproduct of their influence on CYP2A6 mRNA. In model 1, genotype, AKR1D1 mRNA levels, and age were all significant predictors (Table 14). When CYP2A6 mRNA was added (model 2), the overall proportion of variation in CYP2A6 protein that we were able to account for increased from 17% ($R^2=0.17$, $P<0.001$) to 37% ($R^2=0.37$, $P<0.001$; Table 14). Additionally, as expected the $CYP2A6^*9$ allele was no longer a significant independent predictor of CYP2A6 protein, with the proportion of variation accounted for by this genetic variant decreasing from 1.3% ($P<0.05$) to 0.5% ($P>0.1$; Table 14). There was a substantial decrease in the contribution of AKR1D1 mRNA to variation in CYP2A6 protein when CYP2A6 mRNA was included as a predictor in the model. The proportion of variation accounted for by AKR1D1 mRNA decreased from 9.1% ($P<0.001$) to 1.0% ($P<0.05$; Table 14). CYP2A6 mRNA independently accounted for 20.0% of the variation in CYP2A6 protein levels ($P<0.001$; Table 14).
Table 14 | Linear regression analysis of CYP2A6 protein levels (pmol/mg)\(^a\).

<table>
<thead>
<tr>
<th>Predictor variable</th>
<th>B</th>
<th>Beta</th>
<th>95% CI</th>
<th>% variation accounted for (^d)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model 1(^b)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2A6*2 Genotype(^e)</td>
<td>-11.8</td>
<td>-0.2</td>
<td>-20.0 to -3.7</td>
<td>2.7%</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>CYP2A6*9 Genotype(^e)</td>
<td>-6.1</td>
<td>-0.1</td>
<td>-12.2 to -0.08</td>
<td>1.3%</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>CYP2A6*12 Genotype(^e)</td>
<td>-11.4</td>
<td>-0.1</td>
<td>-23.2 to 0.5</td>
<td>1.2%</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>Gender (M=1, F=0)</td>
<td>-2.2</td>
<td>-0.06</td>
<td>-6.8 to 2.3</td>
<td>0.3%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>AKR1D1 mRNA (FPKM)</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1 to 0.3</td>
<td>9.1%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td>0.1</td>
<td>0.1</td>
<td>0.01 to 0.2</td>
<td>1.6%</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>CYP2A6 mRNA (FPKM)</td>
<td>0.07</td>
<td>0.1</td>
<td>0.001 to 0.1</td>
<td>1.0%</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>CYP2A6*2 Genotype(^e)</td>
<td>-13.6</td>
<td>-0.2</td>
<td>-20.8 to -6.4</td>
<td>3.5%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CYP2A6*9 Genotype(^e)</td>
<td>-3.8</td>
<td>-0.07</td>
<td>-9.1 to -1.5</td>
<td>0.5%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>CYP2A6*12 Genotype(^e)</td>
<td>-8.0</td>
<td>-0.08</td>
<td>-18.4 to 2.3</td>
<td>0.6%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>Gender (M=1, F=0)</td>
<td>-2.1</td>
<td>-0.05</td>
<td>-6.1 to 1.9</td>
<td>0.3%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>AKR1D1 mRNA (FPKM)</td>
<td>0.07</td>
<td>0.1</td>
<td>0.001 to 0.1</td>
<td>1.0%</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Age</td>
<td>0.1</td>
<td>0.1</td>
<td>0.01 to 0.2</td>
<td>1.2%</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>CYP2A6 mRNA (FPKM)</td>
<td>0.02</td>
<td>0.5</td>
<td>0.02 to 0.02</td>
<td>20.0%</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

\(^a\) Cases excluded pairwise

\(^b\) \(R^2=0.17, \ P<0.001\)

\(^c\) \(R^2=0.37, \ P<0.001\)

\(^d\) % variation accounted for by each variable is determined by: (Part Correlation)^2 \times 100

\(^e\) CYP2A6*2, CYP2A6*9, and CYP2A6*12 genotypes coded as 0 for CYP2A6*1/*1 genotype, 1 for CYP2A6*1/*2, CYP2A6*1/*9, or CYP2A6*1/*12 genotypes, and 2 for CYP2A6*2/*2 or CYP2A6*9/*9 genotypes.
Lastly, we quantified the contribution of predictor variables to CYP2A6 enzyme activity. We ran three versions of this model: (1) model 1, (2) model 2 with the addition of CYP2A6 mRNA as a predictor, and (3) model 3 with the further addition of CYP2A6 mRNA and CYP2A6 protein as predictors. Overall, each model accounted for 35%, 46%, and 77% of the variation in CYP2A6 enzyme activity, respectively (nicotine metabolism, $R^2=0.35, 0.46, 0.77, P<0.001$; Table 15; coumarin metabolism, $R^2=0.26, 0.32, 0.57, P<0.001$; Table 16). In model 1 genotype, AKR1D1 mRNA, age, and POR protein were all significant predictors (Table 15). We again observed decreased contributions of both the CYP2A6*9 allele and AKR1D1 mRNA toward CYP2A6 enzyme activity when CYP2A6 mRNA was added (model 2). In the third model, with CYP2A6 protein included as a predictor variable, the independent contribution of all other variables decreased below 1%, except for CYP2A6 protein, which independently contributed to 31.9% of the variation in nicotine (nicotine metabolism $P<0.001$; Table 15, coumarin metabolism 25.0% $P<0.001$; Table 16). Aside from CYP2A6 protein, POR protein was the only remaining significant independent predictor of CYP2A6 enzyme activity, accounting for 0.6% (nicotine metabolism, $P<0.05$; Table 15; coumarin metabolism, 0.8% $P<0.05$; Table 16).
Table 15 | Linear regression analysis of CYP2A6 enzyme activity (cotinine formation from nicotine, nmol/min per milligram) 

<table>
<thead>
<tr>
<th>Predictor variable</th>
<th>B</th>
<th>Beta</th>
<th>95% CI</th>
<th>% variation accounted for</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2A6*2 Genotype</td>
<td>-0.04</td>
<td>-0.1</td>
<td>-0.077 to -0.001</td>
<td>1.0%</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>CYP2A6*9 Genotype</td>
<td>-0.03</td>
<td>-0.1</td>
<td>-0.059 to -0.003</td>
<td>1.2%</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>CYP2A6*12 Genotype</td>
<td>-0.05</td>
<td>-0.09</td>
<td>-0.105 to 0.004</td>
<td>0.8%</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>Gender (M=1, F=0)</td>
<td>0.004</td>
<td>0.02</td>
<td>-0.02 to 0.03</td>
<td>0.03%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>AKR1D1 mRNA (FPKM)</td>
<td>0.001</td>
<td>0.2</td>
<td>&lt;0.001 to 0.001</td>
<td>3.3%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td>0.001</td>
<td>0.2</td>
<td>&lt;0.001 to 0.001</td>
<td>3.5%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>POR Protein</td>
<td>0.004</td>
<td>0.5</td>
<td>0.003 to 0.005</td>
<td>21.5%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Model 1 + CYP2A6 mRNA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2A6*2 Genotype</td>
<td>-0.05</td>
<td>-0.1</td>
<td>-0.08 to -0.01</td>
<td>1.5%</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>CYP2A6*9 Genotype</td>
<td>-0.02</td>
<td>-0.08</td>
<td>-0.047 to 0.005</td>
<td>0.5%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>CYP2A6*12 Genotype</td>
<td>-0.04</td>
<td>-0.07</td>
<td>-0.09 to 0.01</td>
<td>0.4%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>Gender (M=1, F=0)</td>
<td>0.003</td>
<td>0.02</td>
<td>-0.02 to 0.02</td>
<td>0.03%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>AKR1D1 mRNA (FPKM)</td>
<td>&lt;0.001</td>
<td>0.07</td>
<td>&lt;0.001 to 0.001</td>
<td>0.4%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>Age</td>
<td>0.001</td>
<td>0.2</td>
<td>&lt;0.001 to 0.001</td>
<td>2.6%</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>POR Protein</td>
<td>0.003</td>
<td>0.4</td>
<td>0.003 to 0.004</td>
<td>14.0%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CYP2A6 mRNA (FPKM)</td>
<td>&lt;0.001</td>
<td>0.4</td>
<td>&lt;0.001 to &lt;0.001</td>
<td>10.0%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Model 1 + CYP2A6 mRNA + CYP2A6 protein</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2A6*2 Genotype</td>
<td>0.03</td>
<td>0.009</td>
<td>-0.02 to -0.03</td>
<td>0.008%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>CYP2A6*9 Genotype</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>-0.02 to 0.02</td>
<td>&lt;0.001%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>CYP2A6*12 Genotype</td>
<td>0.007</td>
<td>0.01</td>
<td>-0.03 to 0.04</td>
<td>0.01%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>Gender (M=1, F=0)</td>
<td>0.009</td>
<td>0.05</td>
<td>-0.003 to 0.022</td>
<td>0.2%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>AKR1D1 mRNA (FPKM)</td>
<td>&lt;0.001</td>
<td>0.02</td>
<td>&lt;0.001 to &lt;0.001</td>
<td>0.04%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>Age</td>
<td>&lt;0.001</td>
<td>0.03</td>
<td>&lt;0.001 to &lt;0.001</td>
<td>0.06%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>POR Protein</td>
<td>0.001</td>
<td>0.09</td>
<td>&lt;0.001 to 0.001</td>
<td>0.6%</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>CYP2A6 mRNA (FPKM)</td>
<td>&lt;0.001</td>
<td>0.4</td>
<td>&lt;0.001 to &lt;0.001</td>
<td>0.1%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>CYP2A6 Protein</td>
<td>0.004</td>
<td>0.8</td>
<td>0.004 to 0.005</td>
<td>31.9%</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

a. Cases excluded pairwise
b. \( R^2 = 0.35, P < 0.001 \)
c. \( R^2 = 0.46, P < 0.001 \)
d. \( R^2 = 0.77, P < 0.001 \)
e. % variation accounted for by each variable is determined by: (Part Correlation)\(^2\) x 100
f. CYP2A6*2, CYP2A6*9, and CYP2A6*12 genotypes coded as 0 for CYP2A6*1/*1 genotype, 1 for CYP2A6*1/*2, CYP2A6*1/*9, or CYP2A6*1/*12 genotypes, and 2 for CYP2A6*2/*2 or CYP2A6*9/*9 genotypes.
Table 16 | Linear regression analysis of CYP2A6 enzyme activity (7-OH coumarin formation from coumarin, nmol/min per milligram) \(^a\)

<table>
<thead>
<tr>
<th>Predictor variable</th>
<th>B</th>
<th>Beta</th>
<th>95% CI</th>
<th>% variation accounted for (^e)</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model 1 (^b)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2A6*2 Genotype (^f)</td>
<td>-0.2</td>
<td>-0.08</td>
<td>-0.47 to 0.07</td>
<td>0.6%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>CYP2A6*9 Genotype (^f)</td>
<td>-0.2</td>
<td>-0.08</td>
<td>-0.35 to 0.05</td>
<td>0.6%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>CYP2A6*12 Genotype (^f)</td>
<td>-0.3</td>
<td>-0.08</td>
<td>-0.6 to 0.2</td>
<td>0.6%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>Gender (M=1, F=0)</td>
<td>0.03</td>
<td>0.02</td>
<td>-0.7 to 0.1</td>
<td>0.04%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>AKR1D1 mRNA (FPKM)</td>
<td>0.003</td>
<td>0.1</td>
<td>&lt;0.001 to 0.005</td>
<td>1.6%</td>
<td>(&lt;0.05)</td>
</tr>
<tr>
<td>Age</td>
<td>0.004</td>
<td>0.1</td>
<td>0.001 to 0.008</td>
<td>1.9%</td>
<td>(&lt;0.05)</td>
</tr>
<tr>
<td>POR Protein</td>
<td>0.02</td>
<td>0.4</td>
<td>0.02 to 0.03</td>
<td>18.0%</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td><strong>Model 1 + CYP2A6 mRNA (^c)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2A6*2 Genotype (^f)</td>
<td>-0.2</td>
<td>-0.09</td>
<td>-0.50 to 0.03</td>
<td>0.9%</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>CYP2A6*9 Genotype (^f)</td>
<td>-0.1</td>
<td>-0.06</td>
<td>-0.29 to 0.09</td>
<td>0.3%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>CYP2A6*12 Genotype (^f)</td>
<td>-0.2</td>
<td>-0.06</td>
<td>-0.6 to 0.2</td>
<td>0.3%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>Gender (M=1, F=0)</td>
<td>0.03</td>
<td>0.02</td>
<td>-0.1 to 0.2</td>
<td>0.04%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>AKR1D1 mRNA (FPKM)</td>
<td>0.001</td>
<td>0.04</td>
<td>-0.002 to 0.003</td>
<td>0.1%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>Age</td>
<td>0.004</td>
<td>0.1</td>
<td>&lt;0.001 to 0.007</td>
<td>1.4%</td>
<td>(&lt;0.05)</td>
</tr>
<tr>
<td>POR Protein</td>
<td>0.02</td>
<td>0.4</td>
<td>0.02 to 0.03</td>
<td>12.7%</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td>CYP2A6 mRNA (FPKM)</td>
<td>&lt;0.001</td>
<td>0.3</td>
<td>&lt;0.001 to 0.001</td>
<td>5.6%</td>
<td>(&lt;0.001)</td>
</tr>
<tr>
<td><strong>Model 1 + CYP2A6 mRNA + CYP2A6 protein (^d)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP2A6*2 Genotype (^f)</td>
<td>0.06</td>
<td>0.02</td>
<td>-0.2 to 0.3</td>
<td>0.05%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>CYP2A6*9 Genotype (^f)</td>
<td>0.02</td>
<td>0.01</td>
<td>-0.1 to 0.2</td>
<td>0.01%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>CYP2A6*12 Genotype (^f)</td>
<td>0.04</td>
<td>0.01</td>
<td>-0.3 to 0.3</td>
<td>0.008%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>Gender (M=1, F=0)</td>
<td>0.06</td>
<td>0.05</td>
<td>-0.05 to 0.18</td>
<td>0.2%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>AKR1D1 mRNA (FPKM)</td>
<td>&lt;0.001</td>
<td>0.003</td>
<td>-0.002 to 0.002</td>
<td>&lt;0.001%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>Age</td>
<td>&lt;0.001</td>
<td>-0.002</td>
<td>-0.003 to 0.003</td>
<td>&lt;0.001%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>POR Protein</td>
<td>0.006</td>
<td>0.1</td>
<td>&lt;0.001 to 0.01</td>
<td>0.8%</td>
<td>(&lt;0.05)</td>
</tr>
<tr>
<td>CYP2A6 mRNA (FPKM)</td>
<td>&lt;0.001</td>
<td>-0.01</td>
<td>&lt;0.001 to &lt;0.001</td>
<td>0.01%</td>
<td>&gt;0.1</td>
</tr>
<tr>
<td>CYP2A6 Protein</td>
<td>0.03</td>
<td>0.7</td>
<td>0.02 to 0.03</td>
<td>25%</td>
<td>(&lt;0.001)</td>
</tr>
</tbody>
</table>

- a. Cases excluded pairwise
- b. \(R^2=0.26, P<0.001\)
- c. \(R^2=0.32, P<0.001\)
- d. \(R^2=0.57, P<0.001\)
- e. % variation accounted for by each variable is determined by: (Part Correlation)\(^2\) x 100
- f. CYP2A6*2, CYP2A6*9, and CYP2A6*12 genotypes coded as 0 for CYP2A6*1/*1 genotype, 1 for CYP2A6*1/*2, CYP2A6*1/*9, or CYP2A6*1/*12 genotypes, and 2 for CYP2A6*2/*2 or CYP2A6*9/*9 genotypes.
4.5 Discussion

In a large human liver bank, we have identified several genetic and non-genetic factors that contribute to variation in CYP2A6 mRNA, protein, and ultimately CYP2A6 enzyme activity, which is an important determinant of the rate of nicotine metabolism. We confirmed that our CYP2A6 phenotype measures were correlated with, and predictive of one another. Using regression models we were able to account for over 75% of the variation in CYP2A6 activity, as well as assessing the novel roles of AKR1D1 and POR in these CYP2A6 phenotypes.

Consistent with the literature, female gender was associated with higher CYP2A6 mRNA expression, relative to males donors in the liver bank, in univariate analyses with similar trends for protein levels, and enzyme activity. However, gender was not a significant independent predictor of any CYP2A6 phenotype based on regression modeling. It is possible that other factors are blunting the relationship between gender and CYP2A6 in our models, or that there is a relatively large impact of pre- and post-menopausal women in this analysis, where previous data suggests similar levels of activity (Benowitz et al., 2006). 20.1% and 42.5% of females in this liver bank are below 16 years and above 50 years old, respectively.

The age of liver donors was positively, but weakly, associated with CYP2A6 protein levels and enzyme activity (Figure 21). It is unclear if this relationship denotes a true age effect on CYP2A6, or if this increase results from unknown covariates, for example greater inducer exposure among older donors. For example, there is an overall increase in polypharmacy with increasing age (Hajjar et al., 2007), and several drugs are known CYP2A6 inducers, such as phenobarbital, dexamethasone, or rifampin (Maurice et al., 1991, Rae et al., 2001).

Additionally, there may be dietary differences between age groups, with elderly individuals consuming more CYP2A6-inducing foods, or a role for estrogen in increasing levels over puberty (Hakooz and Hamdan, 2007, Higashi et al., 2007). However, when modeling predictors of CYP2A6 phenotypes, age was a significant independent predictor of both CYP2A6 protein and activity, even when accounting for gender, suggesting that a female puberty effect is not responsible for the observed positive association between age and CYP2A6 protein levels and enzyme activity.
CYP2A6 genotype was associated with reduced CYP2A6 protein and enzyme activity in both univariate and regression analyses. We observed a step-wise decrease in CYP2A6 mRNA expression, protein levels, and enzyme activity with increasing copies of the CYP2A6*9 allele (i.e. from *1/*1 to *1/*9 to *9/*9). CYP2A6*9 is a SNP present in the TATA box of the CYP2A6 promoter and is associated with decreased CYP2A6 transcription (Pitarque et al., 2001), consistent with our results. There was also a step-wise decrease in CYP2A6 protein and enzyme activity associated with CYP2A6*2, which is a nonsynonymous SNP in exon 3, resulting in the failure of the enzyme to incorporate heme necessary for catalytic function, resulting in a less stable enzyme, vulnerable to degradation (Yamano et al., 1990). Both the CYP2A6*2 and CYP2A6*9 alleles were significant independent predictors of CYP2A6 protein levels and enzyme activity in our regression models. However, when CYP2A6 mRNA was included as a covariate in both the protein and activity models, the impact of the *9 allele on each phenotype decreased, and was no longer significant, suggesting that *9 exhibits a secondary effect on CYP2A6 protein and activity, via its direct influence on CYP2A6 mRNA. Conversely, the *2 allele remains a significant predictor of both CYP2A6 protein and activity when CYP2A6 mRNA was added to the model, suggesting a mechanism independent of CYP2A6 mRNA expression. Within the wild-type (*1/*1) group, which possesses no tested CYP2A6 genetic variants, we still observed a large range in CYP2A6 mRNA, protein, and enzyme activity. This implies that there are additional uncharacterized sources of variation, which may be due to environmental exposures, or as a result of undetected genetic variation present at the CYP2A6 gene locus or potentially other regulatory loci.

AKR1D1 has been implicated as being a potential regulator of the expression of CYPs (Chaudhry et al., 2013), and our findings support a relationship between AKR1D1 and CYP2A6 mRNA expression. AKR1D1 mRNA expression was associated with increasing CYP2A6 mRNA, protein, and enzyme activity, while the specific genetic variation (AKR1D1 SNP rs1872930) was not. AKR1D1 mRNA levels accounted for a significant proportion (16%) of the variation in CYP2A6 mRNA expression (Table 13). AKR1D1 was also a significant predictor of CYP2A6 protein levels and enzyme activity, however the proportion of variation in CYP2A6 protein and activity accounted for by AKR1D1 mRNA decreased more than 8-fold when CYP2A6 mRNA was added to each model. This suggests that the relationship between AKR1D1 mRNA and CYP2A6 phenotypes is largely due to the influence of AKR1D1 mRNA on CYP2A6 mRNA expression. The association between AKR1D1 on CYP2A6 mRNA
expression may result from the role of AKR1D1 in bile acid synthesis and/or the reduction of steroid hormones, which can act as ligands of nuclear hormone receptors. For example AKR1D1 is responsible for 5β-reduction of progesterone, resulting in the activation of nuclear hormone receptor PXR (Bertilsson et al., 1998). Bile acids can activate nuclear hormone receptors, such as PXR and CAR, which can regulate the transcription of several genes, including CYPs (Schuetz et al., 2001). These data confirm a significant role for AKR1D1 on the regulation of multiple CYPs, suggesting a broader role for it in altering drug metabolism. Whether the variation contributes to altered nicotine pharmacokinetics and subsequent smoking behaviours remains to be tested.

POR protein was positively correlated with CYP2A6 activity, remaining a significant independent predictor of CYP2A6 activity when accounting for other sources of variation (Figure 24A, Tables 15 and 16). There was no association between the POR genotypes tested and CYP2A6 enzyme activity. The role of POR as an electron donor to CYPs during the process of substrate metabolism is consistent with the observed positive correlation of POR and CYP2A6 enzyme activity. However, the contribution of POR to CYP2A6 activity decreased from 21.5% to 14.0% to less than 1% when CYP2A6 mRNA and CYP2A6 protein were added to the models, respectively (Table 15). This suggests a potential role of POR in regulating mRNA or protein levels, or that POR and CYP2A6 may have a common regulatory feature, as opposed to a direct influence on CYP2A6 enzyme activity.

Although this liver bank was extensively characterized, this study did not assess several genetic and environmental variables that have been associated with CYP2A6 and nicotine metabolism and may account for a portion of the 23% missing variation in CYP2A6 activity. For example, we have focused on the assessment of common (MAF>1%) established CYP2A6 genetic variants, however the majority of genetic variants found in pharmacogenes, including CYP2A6, are thought to be rare (MAF<1%) (Kozyra et al., 2016), suggesting that a substantial portion of the 23% unidentified variation in CYP2A6 activity may be due to unknown CYP2A6 rare variants. Variation in other genes regulating CYP expression or function, including AKRs or nuclear hormone receptors, may also, in part, account for the 23% missing variation. Several enzymes in the AKR1 family, including AKR1C1-1C4, may be associated with variation in CYP2A6 expression due to their contribution to metabolism and biosynthesis of steroids (estrogen, testosterone, progesterone, unavailable in this study), steroid hormones, and bile
acids, ultimately controlling concentrations of active ligands at nuclear receptors, ligand occupancy, and trans-activation of receptors (as reviewed by Rizner and Penning, 2014). Cigarette smoking, an environmental factor, is associated with a reduction in nicotine clearance in vivo, such that nicotine clearance increases 14% after 4 days of abstinence in regular smokers (Benowitz and Jacob, 2000, Benowitz and Jacob, 1993). Similarly, grapefruit juice reduces nicotine’s metabolism to cotinine, mediated by CYP2A6, by 15% (Hukkanen et al., 2006).

In conclusion, we have identified several sources of variation in CYP2A6 activity in vitro that could translate to variable hepatic clearance of several clinically important drugs, namely nicotine. We were able to account for 77% of the variation in CYP2A6 activity in our models. Unaccounted for variation may result from unknown genetic variation at the highly polymorphic CYP2A6 gene locus and/or in additional regulatory genes, as well as environmental factors. Characterizing sources of variation in CYP2A6 activity is important, as variation in the rate of nicotine metabolism and clearance can ultimately influence smoking behavior, and affect a smoker’s ability to quit smoking.
4.6 Significance to Thesis

This chapter expands on the findings of a previous study using a smaller human liver bank (N=67), adding to the literature of genetic and non-genetic contributions to nicotine metabolism. This is the largest *in vitro* study that assesses sources of unaccounted for variation in CYP2A6 phenotypes. This chapter demonstrates the association of *CYP2A6* genotype, CYP2A6 mRNA, CYP2A6 protein, gender, age, AKR1D1, and POR with variation in CYP2A6 enzyme activity. Although many of the factors described in this chapter exhibit a minor contribution to the variation in CYP2A6 phenotypes on their own, their collective 77% contribution to variation in *in vitro* CYP2A6 activity could translate to altered hepatic nicotine clearance, which is associated with different smoking behaviors and tobacco-related disease risk. Determining sources of variation, both genetic and non-genetic, will aid in the ability to predict CYP2A6 activity and in turn, understand the impact on clinical outcomes, such as smokers’ response to different cessation pharmacotherapies. Knowledge, provided by this study, of the still substantial missing variation in CYP2A6 activity suggests that future research is required to characterize remaining variation, possibly focusing on novel *CYP2A6* genetic variation due to the highly polymorphic nature of this gene locus.
5 CHAPTER 4: NOVEL CYP2A6 DIPLOTYPES IDENTIFIED THROUGH NEXT-GENERATION SEQUENCING ARE ASSOCIATED WITH IN VITRO AND IN VIVO NICOTINE METABOLISM


Manuscript submitted to Pharmacogenetics and Genomics.

RFT, AZZ, and SES designed the overall study. SES, VK, JH, HVD, and DMM conducted the CYP2A6 sequencing (library preparation, sequence capture, sequencing, and sequence analysis). JAT participated in the optimization of the CYP2A6 sequencing approach through consultation with SES and his team. JAT performed CYP2A6 genotyping for established variants (described in the previous chapter) and ran in vitro CYP2A6 activity assays (described in the previous chapter). JAT analyzed the sequencing data (vcf files) and identified established and novel CYP2A6 genetic variants. To characterize the functional significance of novel genetic variants, JAT performed bioinformatics analyses using online tools, and analyzed associations of novel variants with CYP2A6 phenotypes in the liver bank. JAT used PHASE and Haploview software to compute linkage disequilibrium, haplotype structure, and diplotypes for novel genetic variants. JAT conducted statistical analyses and modeling. All authors interpreted the results. JAT and RFT wrote the manuscript.
5.1 Abstract

Objective: Smoking patterns and cessation rates vary widely across smokers and can be influenced by variation in rates of nicotine metabolism (i.e. cytochrome P450 2A6, CYP2A6, enzyme activity). There is high heritability of CYP2A6-mediated nicotine metabolism (60-80%) due to known and unidentified genetic variation in the CYP2A6 gene. We aimed to identify and characterize additional genetic variants at the CYP2A6 gene locus.

Methods: A new CYP2A6-specific sequencing method was used to investigate genetic variation in CYP2A6. Novel variants were characterized in a White human liver bank that has been extensively phenotyped for CYP2A6. Linkage and haplotype structure for the novel SNPs were assessed. The association between novel five-SNP diplotypes and nicotine metabolism rate was investigated.

Results: Seven high frequency (MAFs ≥6%) non-coding SNPs were identified as important contributors to CYP2A6 phenotypes in a White human liver bank (rs57837628, rs7260629, rs7259706, rs150298687, rs56113850, rs28399453, and rs8192733), accounting for two times more variation in in vitro CYP2A6 activity relative to the four established functional CYP2A6 variants that are frequently tested in Whites (CYP2A6*2, *4, *9, and *12). Two pairs of novel SNPs were in high linkage disequilibrium, allowing us to establish five-SNP diplotypes that were associated with CYP2A6 enzyme activity (rate of nicotine metabolism) in vitro in the liver bank and in vivo among smokers.

Conclusion: The novel five-SNP diplotype may be useful to incorporate into CYP2A6 genotype models for personalized prediction of nicotine metabolism rate, cessation success, and response to pharmacotherapies.
5.2 Introduction

Smoking patterns and rates of smoking cessation vary widely among smokers. Variation in the rate of nicotine metabolism and clearance is primarily mediated by the hepatic enzyme CYP2A6 (cytochrome P450 2A6) (Messina et al., 1997, Nakajima et al., 1996a). Variation in nicotine metabolism is associated with differences in smoking quantities, dependence, and cessation such that smokers who exhibit a slower rate of nicotine metabolism often smoke less, have lower nicotine dependence scores, and are more likely to quit smoking (Wassenaar et al., 2011, Patterson et al., 2008, Chenoweth et al., 2013, Sofuoglu et al., 2012). In addition, the rate of nicotine metabolism can be used to predict therapeutic response to smoking cessation pharmacotherapies, including behavioral counseling, nicotine replacement therapy, bupropion, and varenicline (Lerman et al., 2015, Patterson et al., 2008, Schnoll et al., 2009). The relationship between the nicotine metabolism rate and smoking quantity stems from smokers’ ability to titrate their nicotine intake from cigarettes in order to maintain consistent nicotine blood levels throughout the day and avoid withdrawal (McMorrow and Foxx, 1983). Nicotine dependence and cessation may be modulated by nicotine metabolism rate such that normal, compared to slow, nicotine metabolizers have greater fluctuation of blood nicotine concentrations, which may result in enhanced reward, altered strength of functional conductivity between rewarding brain regions, and increased conditioned responses to smoking cues (Sofuoglu et al., 2012, Tang et al., 2012, Li et al., 2016).

Heritability estimates suggest that 60-80% of the variation in nicotine metabolism is attributable to genetic influences, with known CYP2A6 genetic variants accounting for approximately 20-30% of the variation (Loukola et al., 2015, Swan et al., 2009). Only a small contribution arises from environmental or non-genetic factors; dietary or pharmacological agents, gender and estrogen consumption, age, body mass index, alcohol consumption, and smoking account for approximately 8% of the variation (Hakooz and Hamdan, 2007, Rae et al., 2001, Itoh et al., 2006, Tanner et al., 2017, Chenoweth et al., 2014a). CYP2A6 gene locus polymorphisms are associated with differences in nicotine metabolism in vitro and in vivo (Tanner et al., 2017, Al Koudsi et al., 2010, Binnington et al., 2012), and correspondingly with smoking behaviors (Wassenaar et al., 2011, Chen et al., 2014). In a recent GWAS of CYP2A6 activity, all three independent genome-wide significant signals were within or near the CYP2A6 locus (Loukola et al., 2015). The CYP2A6 gene is highly polymorphic, with more
than 40 genetic variants characterized to date (http://www.cypalleles.ki.se/cyp2a6.htm). Thus it is likely that additional genetic variation in CYP2A6 remains unaccounted for, contributing to the high heritability of the nicotine metabolism phenotype.

Part of the difficulty in fully characterizing CYP2A6 genetic variation is the high sequence homology of CYP2A6 with CYP2A7 and CYP2A13 (approximately 95% homologous) (Fernandez-Salguero et al., 1995). Consequently, CYP2A6 genotyping typically requires a two-step approach that first ensures specificity to the CYP2A6 gene followed by characterization of the specific variant of interest (Wassenaar et al., 2016). However, in order to perform large-scale evaluations of the CYP2A6 locus for identification of novel genetic variation, there is a need for sequencing approaches that both account for high homology with other genes and are high-throughput. In this study we describe a new method for sequencing this gene, examine novel and existing CYP2A6 genetic variation, and characterize seven novel high frequency single nucleotide polymorphisms (SNPs) with respect to CYP2A6 enzyme activity in a human liver bank in vitro assessment, and using an in vivo phenotype measure.
5.3 Methods

Overview

Using CYP2A6-specific next-generation sequencing, we identified and characterized novel CYP2A6 genetic variants. Sequencing was performed at the CYP2A6 locus and surrounding 5’ and 3’ region (hg38 40843538-40852095) for White (N=42) and African American (N=170) subjects. The sequencing sample set consisted of DNA from subjects from a human liver bank (N=43) (Tanner et al., 2017), a pharmacokinetic study (N=57) (Mwenifumbo et al., 2007), and two smoking cessation clinical trials (KIS2, N=26; KIS3, N=86) (Ho et al., 2009b, Cox et al., 2012). We prioritized potentially functional novel SNPs and characterized the association of each SNP with CYP2A6 phenotypes in a human liver bank previously characterized for CYP2A6 (CYP2A6 sequenced, mRNA, protein, and enzyme activity quantified). Finally, we assessed the linkage and haplotype structure of novel SNPs, and the impact of resulting diplotypes on in vitro and in vivo CYP2A6 activity, in the liver bank and in a population of treatment-seeking smokers from a smoking cessation clinical trial (Pharmacogenetics of Nicotine and Addiction Treatment clinical trial (PNAT), NCT01314001) (Lerman et al., 2015).

CYP2A6-Specific Sequencing

DNA from N=212 White and African American subjects was sequenced at the CYP2A6 locus, using CYP2A6-specific sequencing, in order to thoroughly interrogate SNPs at this locus for further testing. Library preparation, sequence capture, sequencing on Illumina MiSeq, and sequence analysis were conducted according to the following protocol: DNA samples were used to construct Illumina paired-end pre-capture libraries using Beckman robotic workstations (Biomek FX and FXp models). Between 80 and 500 ng of DNA per sample was used as input for library preparation. Briefly, DNA in a 70 μl volume was sheared into fragments of approximately 500-700 base pairs in a Covaris plate with the E220 system (Covaris, Inc. Woburn, MA). Fragmented DNA was purified using 0.8X AMPure XP beads (Beckman, Cat. No. A63882) followed by end-repair (NEBNext End-Repair Module; Cat. No. E6050L), A-tailing (NEBNext® dA-Tailing Module; Cat. No. E6053L) and ligation of the Illumina multiplexing PE adaptors utilizing 96 different 9-bp barcode sequences was performed using the ExpressLink™ T4 DNA Ligase (A custom product from LifeTech). Pre-capture Ligation Mediated-PCR (LM-PCR) was performed for 6-8 cycles using the Library Amplification
Readymix containing KAPA HiFi DNA Polymerase (Kapa Biosystems, Inc., Cat # KK2612). Universal primer IMUX-P1.0 and IMUX-P3.0 were used in the PCR amplification. AMPure XP beads (0.8X) were used to purify reaction contents after each enzymatic reaction. Libraries were quantified and their sizes estimated using the Fragment Analyzer capillary electrophoresis system (Advanced Analytical Technologies, Inc.). The complete protocol and oligonucleotide sequences are accessible from the HGSC website (https://www.hgsc.bcm.edu/sites/default/files/documents/Protocol-Illumina_Whole_Exome_Sequencing_Library_Preparation-KAPA_Version_BCM-HGSC_RD_03-20-2014.pdf).

For capture enrichment, pools of libraries (~32 ng/library and 47 libraries/pool) were prepared and 1.5 μg of DNA/pool was used for hybridization in solution to a custom capture design (260 Kb – Roche Nimblegen). Human COT1 DNA and xGen Universal Blocking oligonucleotides (Integrated DNA Technologies) were added to the hybridization to block both repetitive genomic sequences and the adaptor sequences. Hybridization was done at 47°C for 3 days. Post-capture LM-PCR amplification was performed using the Library Amplification Readymix containing KAPA HiFi DNA Polymerase (Kapa Biosystems, Inc., Cat # KK2612) with 12 cycles of amplification. After the final AMPure XP bead purification, quantity and size of the capture library was analyzed using the Agilent Bioanalyzer 2100 DNA Chip 7500. The enriched library pools (47 libraries/pool) were sequenced on MiSeq using the MiSeq Reagent Kit v2 (300-cycles) to generate 2X 150 bp reads. On average, for these samples, 92.7 Mb of mapped sequence data (88 Mb unique with 208.5 X sequence coverage) across the target region was generated. On an average, 99.4% bases in the design were covered at ≥20x depth.

Initial sequence analysis was performed using the HGSC Mercury analysis pipeline (Reid et al., 2014) (https://www.hgsc.bcm.edu/content/mercury and available in the cloud through DNANexus http://blog.dnanexus.com/2013-10-22-run-mercury-variant-calling-pipeline/). Briefly, the .bcl files produced on-instrument were transferred and used to run Illumina CASAVA software to de-multiplex pooled samples and generate sequence reads and base-call confidence values. The resulting fastq files were then mapped to the Human reference genome (GRCh37) using the Burrows-Wheeler aligner (BWA (Li and Durbin, 2009), http://bio-bwa.sourceforge.net/). The resulting BAM (Li et al., 2009) (binary alignment/map) files underwent quality recalibration, sorting, duplicate read marking and realignment to improve indel calling using GATK (DePristo et al., 2011). Variant calling was performed using
the Atlas-2 suite (Challis et al., 2012) (https://www.hgsc.bcm.edu/software/atlas-2), and the resulting variant call format (vcf (Danecek et al., 2011)) files were annotated using Cassandra (https://www.hgsc.bcm.edu/software/cassandra), an ANNOVAR (Wang et al., 2010)-based tool providing deep annotation including multiple gene models and a broad array of variant/mutation databases.

Novel CYP2A6 SNPs identified in the region (hg38) 40843538-40852095 (3′-UTR to 2kb 5′ of CYP2A6) were analyzed for potential impact on CYP2A6 using online publicly available bioinformatics tools, including ensembl genome browser (Yates et al., 2016), GTEx Portal (Consortium, 2013), RegulomeDB (Boyle et al., 2012), RegRNA 2.0 (Chang et al., 2013), HaploReg v4.1 (Ward and Kellis, 2012), RESCUE-ESE (Fairbrother et al., 2004), and FASESS (Wang et al., 2004). A final subset of novel SNPs predicted to impact CYP2A6 were further investigated in a human liver bank (described below).

**Liver Bank CYP2A6 PGRN-Sequencing and Genotyping**

The CYP2A6 locus was previously sequenced in the full human liver bank using PGRN-Seq (Gordon et al., 2016), an earlier sequencing platform prior to further optimization for CYP2A6. DNA was isolated from liver samples, as described previously (Shirasaka et al., 2015), and candidate genes were sequenced using the PGRN-Seq platform developed for the Pharmacogenomics Research Network by Gordon and colleagues (Gordon et al., 2016). For CYP2A6, sequence coverage included exons 1-9 and the 3′ and 5′ untranslated regions.

Sequence data was filtered using VCFtools (Danecek et al., 2011). All SNPs had a minimum read depth >50. SNPs were annotated with SeattleSeq Annotation, briefly described by Ng et al. (Ng et al., 2009). These CYP2A6 sequence data were one mechanism used to genotype SNP calls in the liver bank for the novel SNPs identified by CYP2A6-specific sequencing (described above). The majority of SNP genotype calls in the liver bank were readily available from PGRN-seq data; calls were then verified using traditional genotyping approaches (described below).

Genotyping for established CYP2A6 genetic variants common among White populations (CYP2A6*2, *4, *9, *12) was performed using two-step allele-specific polymerase chain reaction genotyping or TaqMan SNP genotyping assay (Applied Biosystems/ThermoFisher Scientific), as reported previously (Tanner et al., 2017, Wassenaar et al., 2016). Concordance between the genotype and PGRN-Seq calls for the established SNPs CYP2A6*2 (rs1801272)
and *9 (rs28399433) was 99% and 98%, respectively; CYP2A6*4 and *12 are copy number and hybrid variants not identified by this sequencing method. Next, we verified liver bank PGRN-Seq SNP calls with direct genotyping (TaqMan or SYBR Green SNP genotyping assays) for the seven novel SNPs that exhibited significant associations with CYP2A6 phenotypes. Concordance of genotyping and PGRN-Seq SNP calls was assessed (5/7 SNPs detected by PGRN-Seq), and ranged from 84-97%. For subsequent genotype-phenotype analyses, calls for the seven novel SNPs were made using PGRN-Seq data (i.e. data from an earlier sequencing platform) unless concordance was <90% (N=3) or was not available (N=2), in which case TaqMan or SYBR Green genotype data was used.

**Quantification of In Vitro CYP2A6 Phenotypes**

Genotype-phenotype associations for novel SNPs were assessed in White subjects in the full liver bank dataset (N=332). Measurement of CYP2A6 mRNA, protein, and enzyme activity in the liver bank have been described previously (Chapter 3) (Tanner et al., 2017). Briefly, CYP2A6 mRNA expression was quantified using RNA sequencing (values expressed as fragments per kilobase per million reads, FPKM). CYP2A6 protein was quantified using a liquid chromatography-tandem mass spectrometry (LC-MS/MS) proteomics method. *In vitro* CYP2A6 enzyme activity was determined by quantifying the rate of nicotine metabolism to cotinine in liver microsomes, using 30μM nicotine for this assay, approximated based on nicotine’s Km.

**Clinical Trial CYP2A6 GWAS and Genotyping**

A genome-wide association study (GWAS) of nicotine metabolite ratio (NMR) in the PNAT trial was used to ascertain SNP genotype calls for the seven novel SNPs (GWAS methodology described in detail by Chenoweth et al. 2017, manuscript in preparation). Briefly, genome-wide SNP genotyping was performed using the Illumina HumanOmniExpressExome-8 v1.2 array (Illumina, San Diego, CA). Genotypes were imputed using IMPUTE2 software, and variants with a quality score >0.85 and minor allele frequency >1% were included in analyses. In addition, samples were genotyped for the same established CYP2A6 genetic variants (CYP2A6*2, *4, *9, *12) (Wassenaar et al., 2016). The GWAS and genotyping SNP calls were 99.8% and 99.2% concordant for *2 and *9, respectively, supporting the concordance of GWAS SNP calls for the novel variants.
Assessing Linkage and Haplotype Structure of Novel SNPs

Using Haploview (Barrett et al., 2005), we estimated the linkage disequilibrium (LD) and haplotype structure of the seven novel SNPs separately in the liver bank and PNAT trial. Analyses were restricted to White subjects not possessing any previously established \textit{CYP2A6} genetic variants (i.e. \textit{CYP2A6*2, *4, *9, *12}). For SNP pairs in high LD ($r^2 \geq 0.80$) in PNAT (the larger study population), one SNP from each pair was selected to include in a simplified haplotype, resulting in a final haplotype comprised of five SNPs. Using the PHASE program (Stephens and Donnelly, 2003, Stephens et al., 2001), five-SNP haplotypes were phased (8360 bp genomic distance) in order to predict five-SNP diplotypes in both studies.

Clinical Trial CYP2A6 Phenotyping

\textit{In vivo} CYP2A6 activity was measured in blood at baseline for treatment-seeking smokers in the PNAT trial by quantifying the ratio of nicotine’s metabolites, \textit{trans}-3’-hydroxycotinine/cotinine, a measure referred to as the NMR, using LC-MS/MS (Lerman et al., 2015), as described previously (Tanner et al., 2015b, Dempsey et al., 2004). Limits of quantification for each metabolite were 1 ng/ml of blood.

Statistical Analyses

We assessed the association of each of the seven novel SNPs with CYP2A6 mRNA, protein, and enzyme activity in the liver bank using Kruskal Wallis and Dunn’s Multiple Comparisons tests or the Mann-Whitney test. The same tests were used to evaluate the association of the seven- and five-SNP diplotypes with CYP2A6 activity in the liver bank and PNAT trial. We ran nested linear regression models for CYP2A6 activity to calculate the additional proportion of variation that is accounted for by the seven SNPs, or the simplified five-SNP haplotype, relative to what was accounted for by the established \textit{CYP2A6} genetic variants alone. Genotypes were the only predictors included in these models. Analyses were performed using GraphPad Prism (v.6.0, LaJolla, CA) and SPSS (v.23; IBM, Armonk, NY), and statistical tests were considered significant for $P<0.05$. 
5.4 Results

**Seven novel SNPs identified through sequencing**

We identified N=229 distinct SNPs total (N=38 5’, N=46 exon, N=133 intron, and N=12 3’-UTR CYP2A6 SNPs) in the region (hg38) 40843538-40852095. Following bioinformatics analyses, N=21 novel SNPs were predicted to impact CYP2A6, including nonsynonymous SNPs, SNPs with a predicted association with CYP2A6 mRNA levels, and SNPs located in an miRNA binding site, an exonic splicing enhancer or silencer site, or a regulatory region (Table 17). Of the N=21 novel SNPs, N=13 were sufficiently common to be used in analyses of relationships with *in vitro* CYP2A6 activity in the liver bank (described below); the remaining N=8 novel SNPs were rare (minor allele frequency, MAF<1% in 1000 Genomes) and we were not powered to investigate these in the liver bank. Of the N=13 common SNPs, N=7 were significantly associated with CYP2A6 activity in the liver bank (described below), and were therefore prioritized for further investigation. Subjects possessing any of the SNPs identified through sequencing (i.e. rare or non-significant SNPs) were included in all of the following analyses.
Table 17 | N=21 novel SNPs of interest based on bioinformatics analyses, with the top N=7 novel SNPs of interest bolded.

<table>
<thead>
<tr>
<th>rs ID</th>
<th>CYP2A6 region</th>
<th>Bioinformatics evidence (tool used)</th>
<th>Liver bank P value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs57837628</td>
<td>5'</td>
<td>Associated with increased CYP2A6 expression in liver tissue (GTEx Portal)</td>
<td>AA vs. AG: 0.002, AA vs. GG: 0.0001</td>
</tr>
<tr>
<td>rs7260629</td>
<td>5'</td>
<td>Associated with increased CYP2A6 expression in liver tissue (GTEx Portal)</td>
<td>TT vs. TG: 0.23, TT vs. GG 0.01</td>
</tr>
<tr>
<td>rs7259706</td>
<td>5'</td>
<td>Associated with increased CYP2A6 expression in liver tissue (GTEx Portal)</td>
<td>CC vs. CT: 0.31, CC vs. TT: 0.02</td>
</tr>
<tr>
<td>rs150298687</td>
<td>5'</td>
<td>Associated with increased CYP2A6 expression in liver tissue (GTEx Portal)</td>
<td>TT vs. TC: 0.02, TT vs. CC: 0.004</td>
</tr>
<tr>
<td>rs144279119</td>
<td>Exon 1</td>
<td>Missense (G44A; ensembl); located in exonic splicing enhancer region (RESCUE-ESE)</td>
<td>N/A, too rare in Whites (MAF&lt;1%)</td>
</tr>
<tr>
<td>rs145308399</td>
<td>Exon 2</td>
<td>Missense, possibly damaging (E97K; ensembl); located in exonic splicing enhancer region (RESCUE-ESE)</td>
<td>N/A, too rare in Whites (MAF&lt;1%)</td>
</tr>
<tr>
<td>rs53582503</td>
<td>Exon 3</td>
<td>Synonymous (ensembl), located in exonic splicing enhancer region (RESCUE-ESE)</td>
<td>N/A, too rare in Whites (MAF&lt;1%)</td>
</tr>
<tr>
<td>rs28399441</td>
<td>Exon 3</td>
<td>Synonymous; present in potential regulatory region (ensembl)</td>
<td>N/A, too rare in Whites (MAF&lt;1%)</td>
</tr>
<tr>
<td>rs57133587</td>
<td>Exon 4</td>
<td>Missense (R190L; ensembl)</td>
<td>N/A, too rare in Whites (MAF&lt;1%)</td>
</tr>
<tr>
<td>rs145393137</td>
<td>Exon 4</td>
<td>Missense (R203H; ensembl)</td>
<td>N/A, too rare in Whites (MAF&lt;1%)</td>
</tr>
<tr>
<td>rs757324688</td>
<td>Exon 5</td>
<td>Missense (M275L; ensembl)</td>
<td>N/A, too rare in Whites (MAF&lt;1%)</td>
</tr>
<tr>
<td>rs200267449</td>
<td>Exon 7</td>
<td>Missense (A347T; ensembl)</td>
<td>CC vs. CT: 0.34</td>
</tr>
<tr>
<td>rs28399462</td>
<td>Exon 8</td>
<td>Synonymous; present in potential regulatory region (ensembl)</td>
<td>GG vs. GA: 0.29</td>
</tr>
<tr>
<td>rs145014075</td>
<td>Exon 9</td>
<td>Nonsense, resulting in generation of premature stop codon (ensembl)</td>
<td>GG vs. GT: 0.26</td>
</tr>
<tr>
<td>rs140179029</td>
<td>Intron 1</td>
<td>Likely to affect protein binding, score 2b (RegulomeDB)</td>
<td>N/A, too rare in Whites (MAF&lt;1%)</td>
</tr>
<tr>
<td>rs56113850</td>
<td>Intron 4</td>
<td>Associated with higher CYP2A6 expression in liver tissue (GTEx Portal); located in an miRNA binding site (miR-4743; RegRNA2.0)</td>
<td>TT vs. TC: 0.03, TT vs. CC: 0.001</td>
</tr>
<tr>
<td>rs28399453</td>
<td>Intron 6</td>
<td>Associated with higher CYP2A6 expression in liver tissue (GTEx Portal); located in an miRNA binding site (miR-765; RegRNA2.0)</td>
<td>GG vs. GA &amp; AA: 0.04</td>
</tr>
<tr>
<td>rs55971367</td>
<td>Intron 6</td>
<td>Associated with lower CYP2A6 expression in liver tissue (GTEx Portal); located in an miRNA binding site (miR-766-5p, miR-4739; RegRNA2.0)</td>
<td>GG vs. GA: &gt;0.99, GG vs. AA: &gt;0.99</td>
</tr>
<tr>
<td>rs28399482</td>
<td>3’-UTR</td>
<td>Likely to affect protein binding, score 2b (RegulomeDB)</td>
<td>GG vs. GC: 0.61, GG vs. CC: 0.26</td>
</tr>
<tr>
<td>rs28399481</td>
<td>3’-UTR</td>
<td>Likely to affect protein binding, score 2b (RegulomeDB)</td>
<td>CC vs. CG: 0.99, CC vs. CG: 0.07</td>
</tr>
<tr>
<td>rs8192733</td>
<td>3’-UTR</td>
<td>Associated with increased CYP2A6 expression in liver tissue (GTEx Portal); likely to affect protein binding, score 2b (RegulomeDB)</td>
<td>GG vs. GC: 0.001, GG vs. CC: &lt;0.0001</td>
</tr>
</tbody>
</table>

<sup>a</sup> P value for association with in vitro CYP2A6 enzyme activity in the human liver bank
Of the seven prioritized SNPs, four are located 5’ of CYP2A6 (rs57837628, rs7260629, rs7259706, rs150298687), two are located in introns (rs56113850, rs28399453), and one in the 3’-UTR (rs8192733) (Table 18). The seven SNPs are common in White populations in 1000 Genomes (6–70%), the human liver bank (7–72%), and the PNAT trial (7–69%).

Table 18 | Summary of the seven novel CYP2A6 SNPs of interest

<table>
<thead>
<tr>
<th>SNP rs ID</th>
<th>CYP2A6 region</th>
<th>Hg38 position</th>
<th>Allele change</th>
<th>1000 Genome</th>
<th>Liver Bank</th>
<th>PNAT Trial</th>
<th>Predicted Function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>African Allele Freq</td>
<td>European Allele Freq</td>
<td>White Allele Freq</td>
<td>White Allele Freq</td>
<td></td>
</tr>
<tr>
<td>rs57837628</td>
<td>5’</td>
<td>40852005</td>
<td>A&gt;G</td>
<td>17% (G)</td>
<td>54% (G)</td>
<td>51% (G)</td>
<td>49% (G)</td>
</tr>
<tr>
<td>rs7260629</td>
<td>5’</td>
<td>40851727</td>
<td>T&gt;G</td>
<td>71% (G)</td>
<td>69% (G)</td>
<td>72% (G)</td>
<td>69% (G)</td>
</tr>
<tr>
<td>rs7259706</td>
<td>5’</td>
<td>40851715</td>
<td>C&gt;T</td>
<td>73% (T)</td>
<td>70% (T)</td>
<td>69% (T)</td>
<td>69% (T)</td>
</tr>
<tr>
<td>rs150298687</td>
<td>5’</td>
<td>40851439</td>
<td>T&gt;C</td>
<td>46% (C)</td>
<td>58% (C)</td>
<td>63% (C)</td>
<td>58% (C)</td>
</tr>
<tr>
<td>rs56113850</td>
<td>Intron 4</td>
<td>40847202</td>
<td>T&gt;C</td>
<td>39% (C)</td>
<td>59% (C)</td>
<td>57% (C)</td>
<td>56% (C)</td>
</tr>
<tr>
<td>rs28399453</td>
<td>Intron 6</td>
<td>40845495</td>
<td>G&gt;A</td>
<td>0% (A)</td>
<td>6% (A)</td>
<td>7% (A)</td>
<td>7% (A)</td>
</tr>
<tr>
<td>rs8192733</td>
<td>3’-UTR</td>
<td>40843645</td>
<td>G&gt;C</td>
<td>23% (C)</td>
<td>48% (C)</td>
<td>47% (C)</td>
<td>47% (C)</td>
</tr>
</tbody>
</table>

a. Function predicted according to bioinformatics tools (e.g. GTEx Portal)
Association of seven novel SNPs with CYP2A6 phenotypes

All seven novel SNPs were significantly associated with higher in vitro CYP2A6 activity, each exhibiting a gene-dose effect (P values 0.31−<0.0001; Fig. 25, Table 19). Six of the seven SNPs were significantly associated with higher CYP2A6 protein levels (P values 0.49−<0.0001), with the exception of rs28399453 (P=0.11, same direction of effect) (Table 19). None of the seven SNPs significantly altered CYP2A6 mRNA expression (all P>0.11), however all indicated a similar direction of effect.

In order to determine if the novel SNPs collectively accounted for any additional variation in CYP2A6 activity in the liver bank, we performed nested linear regression modeling. The previously established CYP2A6 genetic variants (*2, *4, *9, *12) alone accounted for 2.1% of the variation in CYP2A6 activity (model 1: R^2=0.021, P=0.009). When the seven novel SNPs were added to the model, each as individual predictors, they accounted for an additional 4.5% of CYP2A6 activity variation (model 2: R^2=0.066, P=0.005), relative to the established variants (model 1 vs. model 2: R^2 change=0.045, P=0.03).
Figure 25 | Association of seven novel SNPs with CYP2A6 enzyme activity (rate of cotinine formation from nicotine, nmol/min/mg) among N=327 White liver donors. (A) rs57837628, (B) rs7260629, (C) rs7259706, (D) rs150298687, (E) rs56113850, (F) rs28399453, (G) rs8192733. P values derived from Kruskal Wallis and Dunn’s Multiple Comparisons tests for most SNPs, or Mann-Whitney tests for rs28399453. All White liver bank donors were included in analyses.
<table>
<thead>
<tr>
<th>SNP rs ID</th>
<th>Genotype groups compared</th>
<th>CYP2A6 enzyme activity (nmol/min/mg), N=327</th>
<th>CYP2A6 protein levels (pmol/mg), N=320</th>
<th>CYP2A6 mRNA expression (FPKM values), N=265</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Effect size</td>
<td>P value</td>
<td>Effect size</td>
</tr>
<tr>
<td>rs57837628</td>
<td>AA vs. AG</td>
<td>0.53</td>
<td>0.002</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>AA vs. GG</td>
<td>0.67</td>
<td>0.0001</td>
<td>0.74</td>
</tr>
<tr>
<td>rs7260629</td>
<td>TT vs. TG</td>
<td>0.44</td>
<td>0.23</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>TT vs. GG</td>
<td>0.65</td>
<td>0.01</td>
<td>0.68</td>
</tr>
<tr>
<td>rs7259706</td>
<td>CC vs. CT</td>
<td>0.06</td>
<td>0.31</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>CC vs. TT</td>
<td>0.30</td>
<td>0.02</td>
<td>0.28</td>
</tr>
<tr>
<td>rs150298687</td>
<td>TT vs. TC</td>
<td>0.48</td>
<td>0.02</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>TT vs. CC</td>
<td>0.60</td>
<td>0.004</td>
<td>0.62</td>
</tr>
<tr>
<td>rs56113850</td>
<td>TT vs. TC</td>
<td>0.40</td>
<td>0.03</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>TT vs. CC</td>
<td>0.59</td>
<td>0.001</td>
<td>0.66</td>
</tr>
<tr>
<td>rs28399453</td>
<td>GG vs. GA &amp; AA</td>
<td>0.39</td>
<td>0.04</td>
<td>0.28</td>
</tr>
<tr>
<td>rs8192733</td>
<td>GG vs. GC</td>
<td>0.44</td>
<td>0.001</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>GG vs. CC</td>
<td>0.65</td>
<td>&lt;0.0001</td>
<td>0.80</td>
</tr>
</tbody>
</table>
Linkage and haplotype structure for seven novel SNPs

Due to the high frequency of the seven novel SNPs and their similar impact on CYP2A6, we assessed linkage and haplotype structure. Haploview (Barrett et al., 2005) was used to predict $r^2$, D’, and haplotype blocks in liver donors and PNAT treatment-seeking smokers. The seven novel SNPs were not in LD with established CYP2A6 genetic variants (*2, *4, *9, *12) in either the liver bank or PNAT trial ($r^2$ values=0.00–0.07; Fig. 26). Further analyses were restricted to subjects without established CYP2A6 genetic variants. In the liver bank, of the seven novel SNPs, two SNP pairs were in high LD ($r^2$>0.8, D’>0.9): (1) rs57837628 and rs8192733, and (2) rs56113850 and rs150298687, with the rs7260629 and rs7259706 pair falling just below threshold ($r^2$=0.79) (Fig. 27a). In the PNAT trial, two SNP pairs were in high LD: (1) rs57837628 and rs8192733, (2) rs7260629 and rs7259706 (Fig. 27b). SNP rs28399453 was not in LD with any of the other six SNPs in either study population ($r^2$ values=0.03–0.07).

Due to high linkage among the novel SNPs, we chose to exclude one SNP from each of the high LD SNP pairs that were present in PNAT, the larger study population, (i.e. rs57837628 and rs8192733; rs7260629 and rs7259706) and investigate CYP2A6 phenotype associations for a simplified set of five SNPs: rs28399453, rs56113850, rs150298687, rs7260629, and rs57837628. We then quantified the proportion of variation in CYP2A6 activity in the liver bank additionally accounted for by the five SNPs, relative to the established CYP2A6 variants alone, using regression modeling. The five novel SNPs accounted for an additional 3.9% of the variation in CYP2A6 activity (model 2: $R^2$=0.060, $P$=0.003) compared to the established variants (model 1: $R^2$=0.021, $P$=0.009; model 1 vs. model 2: $R^2$ change=0.039, $P$=0.02).
Figure 26 | LD plot for seven novel SNPs and established variants (*2, *4, *9, *12); $r^2$ values (number) and D’ (shading) were derived using Haplovie. Analyses were restricted to White subjects. (A) Human liver bank. (B) Treatment-seeking smokers from PNAT.
Figure 27 | LD plot for seven novel SNPs; $r^2$ values (number) and $D'$ (shading) were derived using Haploview. Analyses were restricted to White subjects who do not possess any established $CYP2A6$ genetic variants. (A) Human liver bank. (B) Treatment-seeking smokers from PNAT.
Five-SNP diplotypes in the liver bank and PNAT trial

The PHASE software (Stephens and Donnelly, 2003, Stephens et al., 2001) was used to predict seven- and five-SNP diplotypes. Subjects possessing established CYP2A6 genetic variants (*2, *4, *9, *12) were excluded from analyses. Summaries of seven-SNP diplotypes in the liver bank and PNAT have been provided in Table 20 and Table 21, respectively.

### Table 20 | Seven-SNP phased diplotypes predicted in donors in the liver bank.

<table>
<thead>
<tr>
<th>Dipotype group #</th>
<th>Full Diplotype</th>
<th>N in liver bank</th>
<th>CYP2A6 activity (nicotine metabolism, nmol/min/mg), mean of diplotype group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>G G T T C T A / G G T T C T A</td>
<td>16</td>
<td>0.06</td>
</tr>
<tr>
<td>2</td>
<td>G G T T C T A / G G C C T G G</td>
<td>54</td>
<td>0.13</td>
</tr>
<tr>
<td>3</td>
<td>G G T T T G A / G G C C T G G</td>
<td>46</td>
<td>0.09</td>
</tr>
<tr>
<td>4</td>
<td>G G C C T G A / G G C C T G G</td>
<td>20</td>
<td>0.12</td>
</tr>
<tr>
<td>5</td>
<td>G G T T C T A / C A C C T G G</td>
<td>19</td>
<td>0.10</td>
</tr>
<tr>
<td>6</td>
<td>G G T T C T A / C A C C T G G</td>
<td>13</td>
<td>0.12</td>
</tr>
<tr>
<td>7</td>
<td>G G T T C T A / G G T T T G A</td>
<td>12</td>
<td>0.16</td>
</tr>
<tr>
<td>8</td>
<td>G G T T C T A / G G C C T G A</td>
<td>54</td>
<td>0.05</td>
</tr>
<tr>
<td>9</td>
<td>G G C C T G G / G G C C T G G</td>
<td>46</td>
<td>0.08</td>
</tr>
<tr>
<td>10</td>
<td>G G T T C T A / G G C C T G G</td>
<td>19</td>
<td>0.06</td>
</tr>
<tr>
<td>11</td>
<td>G G C C T G A / G G C C T G A</td>
<td>20</td>
<td>0.12</td>
</tr>
<tr>
<td>12</td>
<td>G G T T C T A / G G C C C G G</td>
<td>20</td>
<td>0.10</td>
</tr>
<tr>
<td>13</td>
<td>G G T T C T A / G G C C C G G</td>
<td>19</td>
<td>0.09</td>
</tr>
<tr>
<td>14</td>
<td>G G T T C T A / G G T T C G A</td>
<td>18</td>
<td>0.16</td>
</tr>
<tr>
<td>15</td>
<td>G G T T T G A / C A C C T G G</td>
<td>17</td>
<td>0.06</td>
</tr>
<tr>
<td>16</td>
<td>G G C C T G A / C A C C T G G</td>
<td>16</td>
<td>0.33</td>
</tr>
<tr>
<td>17</td>
<td>G G T T C G A / G G C C T G A</td>
<td>15</td>
<td>0.33</td>
</tr>
<tr>
<td>18</td>
<td>G G T T C T A / G G C C T G A</td>
<td>14</td>
<td>0.16</td>
</tr>
<tr>
<td>19</td>
<td>G G T T T G A / G G C C T G A</td>
<td>13</td>
<td>0.16</td>
</tr>
<tr>
<td>20</td>
<td>G G T T T G A / G G T T T G A</td>
<td>12</td>
<td>0.16</td>
</tr>
<tr>
<td>21</td>
<td>C A C C T G G / C A C C T G G</td>
<td>11</td>
<td>0.04</td>
</tr>
<tr>
<td>22</td>
<td>C G C C C G G / C G C C C G G</td>
<td>10</td>
<td>0.04</td>
</tr>
<tr>
<td>23</td>
<td>C G C C T G A / C A C C T G G</td>
<td>9</td>
<td>0.04</td>
</tr>
<tr>
<td>24</td>
<td>C G C C T G A / G G C C T G A</td>
<td>8</td>
<td>0.04</td>
</tr>
<tr>
<td>25</td>
<td>G G C C C G G / G G C C T G G</td>
<td>7</td>
<td>0.04</td>
</tr>
<tr>
<td>26</td>
<td>G G C C T G A / C A C C T G G</td>
<td>6</td>
<td>0.04</td>
</tr>
<tr>
<td>27</td>
<td>G G C C C G G / C G C C C G G</td>
<td>5</td>
<td>0.04</td>
</tr>
<tr>
<td>28</td>
<td>C G C C T G A / C A C C T G G</td>
<td>4</td>
<td>0.04</td>
</tr>
<tr>
<td>29</td>
<td>G G T C T G A / C G C C T G G</td>
<td>3</td>
<td>0.04</td>
</tr>
<tr>
<td>30</td>
<td>G G T C T G A / G G C C T G A</td>
<td>2</td>
<td>0.04</td>
</tr>
<tr>
<td>31</td>
<td>G G T C T G A / C A C C T G G</td>
<td>1</td>
<td>0.04</td>
</tr>
<tr>
<td>32</td>
<td>G G T C T G A / C G C C T G G</td>
<td>1</td>
<td>0.04</td>
</tr>
<tr>
<td>33</td>
<td>G G T T C G A / C A C C T G G</td>
<td>1</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Allele 1</td>
<td>Allele 2</td>
<td>Frequency</td>
</tr>
<tr>
<td>---</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>34</td>
<td>G G T T C G A / C G C C C G G</td>
<td>1</td>
<td>0.11</td>
</tr>
<tr>
<td>35</td>
<td>G G T T C G A / C G C C T G G</td>
<td>1</td>
<td>0.12</td>
</tr>
<tr>
<td>36</td>
<td>G G T T C T A / C G C C T T G</td>
<td>1</td>
<td>0.06</td>
</tr>
<tr>
<td>37</td>
<td>G G T T C T A / G G C C T T G</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>38</td>
<td>G G T T C T A / G G T T T T A</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>39</td>
<td>G G T T C T G / G G T C T G G</td>
<td>1</td>
<td>0.30</td>
</tr>
<tr>
<td>40</td>
<td>G G T T T G A / C G C C T G A</td>
<td>1</td>
<td>0.17</td>
</tr>
<tr>
<td>41</td>
<td>G G T T T G A / C G T T C T A</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>42</td>
<td>G G T T T G A / G G T C T G G</td>
<td>1</td>
<td>0.19</td>
</tr>
<tr>
<td>43</td>
<td>G G T T T T A / C G C C T G G</td>
<td>1</td>
<td>0.20</td>
</tr>
</tbody>
</table>

a. Full reference diplotype is bolded
Table 21 | Seven-SNP phased diplotypes predicted in treatment-seeking smokers in PNAT trial.

<table>
<thead>
<tr>
<th>Diplootype Group #</th>
<th>Full Diploype</th>
<th>N in PNAT</th>
<th>CYP2A6 activity (nicotine metabolite ratio, NMR), mean of diplotype group</th>
<th>Corresponding liver bank diplotype group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference</strong></td>
<td><strong>GGTTCTA / GGTCTCTA</strong></td>
<td><strong>40</strong></td>
<td><strong>0.30</strong></td>
<td><strong>Reference</strong></td>
</tr>
<tr>
<td>1</td>
<td>CGCCCTGA / CGCCCTGG</td>
<td>101</td>
<td>0.56</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>CGCCCTGA / CGCCCTGG</td>
<td>116</td>
<td>0.41</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>CGCCCTGA / CGCCCTGG</td>
<td>42</td>
<td>0.46</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>CGCCCTGA / CGCCCTGG</td>
<td>51</td>
<td>0.56</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>CGCCCTGA / CACCTGG</td>
<td>37</td>
<td>0.49</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>CGCCCTGA / CACCTGG</td>
<td>29</td>
<td>0.42</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>17</td>
<td>0.41</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>28</td>
<td>0.44</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>7</td>
<td>0.47</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>3</td>
<td>0.29</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>10</td>
<td>0.44</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>5</td>
<td>0.48</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>1</td>
<td>0.39</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>8</td>
<td>0.54</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>10</td>
<td>0.54</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>2</td>
<td>0.26</td>
<td>16</td>
</tr>
<tr>
<td>17</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>7</td>
<td>0.43</td>
<td>17</td>
</tr>
<tr>
<td>18</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>3</td>
<td>0.45</td>
<td>18</td>
</tr>
<tr>
<td>19</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>1</td>
<td>0.41</td>
<td>19</td>
</tr>
<tr>
<td>20</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>1</td>
<td>0.52</td>
<td>20</td>
</tr>
<tr>
<td>21</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>2</td>
<td>0.21</td>
<td>21</td>
</tr>
<tr>
<td>22</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>1</td>
<td>0.31</td>
<td>22</td>
</tr>
<tr>
<td>23</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>3</td>
<td>0.24</td>
<td>23</td>
</tr>
<tr>
<td>24</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>22</td>
<td>0.53</td>
<td>24</td>
</tr>
<tr>
<td>25</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>12</td>
<td>0.46</td>
<td>25</td>
</tr>
<tr>
<td>26</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>12</td>
<td>0.44</td>
<td>26</td>
</tr>
<tr>
<td>27</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>11</td>
<td>0.46</td>
<td>27</td>
</tr>
<tr>
<td>28</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>10</td>
<td>0.60</td>
<td>28</td>
</tr>
<tr>
<td>29</td>
<td>CGCCCTGA / GGTCTCTA</td>
<td>5</td>
<td>0.52</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>First Sequence</td>
<td>Second Sequence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---------------------</td>
<td>-------------------</td>
<td>---</td>
<td>-----</td>
</tr>
<tr>
<td>30</td>
<td>G GCCCTGA / CGCTTGG</td>
<td>5</td>
<td>0.54</td>
<td>N/A</td>
</tr>
<tr>
<td>31</td>
<td>CGCCTGG / CGCTCTG</td>
<td>4</td>
<td>0.67</td>
<td>N/A</td>
</tr>
<tr>
<td>32</td>
<td>GGTCTTA / GGCTTGA</td>
<td>4</td>
<td>0.45</td>
<td>N/A</td>
</tr>
<tr>
<td>33</td>
<td>CGCCTGA / CGCCTGG</td>
<td>2</td>
<td>0.39</td>
<td>N/A</td>
</tr>
<tr>
<td>34</td>
<td>GGCCTGA / GGCCTGG</td>
<td>2</td>
<td>0.46</td>
<td>N/A</td>
</tr>
<tr>
<td>35</td>
<td>GGTGTA / GGTTTGG</td>
<td>2</td>
<td>0.35</td>
<td>N/A</td>
</tr>
<tr>
<td>36</td>
<td>GGTCTTA / GGTTTGG</td>
<td>2</td>
<td>0.33</td>
<td>N/A</td>
</tr>
<tr>
<td>37</td>
<td>GGTGTGG / CACCTGG</td>
<td>1</td>
<td>0.35</td>
<td>N/A</td>
</tr>
<tr>
<td>38</td>
<td>CACCTGG / CACCTGG</td>
<td>1</td>
<td>0.46</td>
<td>N/A</td>
</tr>
<tr>
<td>39</td>
<td>GGTCTTA / CACCTGG</td>
<td>1</td>
<td>0.53</td>
<td>N/A</td>
</tr>
<tr>
<td>40</td>
<td>CGCCTGG / CGCTTGG</td>
<td>1</td>
<td>0.56</td>
<td>N/A</td>
</tr>
<tr>
<td>41</td>
<td>GGTCTTA / CGCCTGG</td>
<td>1</td>
<td>0.08</td>
<td>N/A</td>
</tr>
<tr>
<td>42</td>
<td>CGCCTGG / CACCTGG</td>
<td>1</td>
<td>0.69</td>
<td>N/A</td>
</tr>
<tr>
<td>43</td>
<td>CGCCTGG / CGCCTGG</td>
<td>1</td>
<td>0.30</td>
<td>N/A</td>
</tr>
<tr>
<td>44</td>
<td>CGCCTGG / CGCTCTG</td>
<td>1</td>
<td>0.48</td>
<td>N/A</td>
</tr>
<tr>
<td>45</td>
<td>GGCCTGA / CGCCTGG</td>
<td>1</td>
<td>0.67</td>
<td>N/A</td>
</tr>
<tr>
<td>46</td>
<td>CGCCTGA / CACCTGG</td>
<td>1</td>
<td>0.50</td>
<td>N/A</td>
</tr>
<tr>
<td>47</td>
<td>CGCCTGA / CGCCTGA</td>
<td>1</td>
<td>0.13</td>
<td>N/A</td>
</tr>
<tr>
<td>48</td>
<td>CGCCTGG / CGCCTGG</td>
<td>1</td>
<td>0.69</td>
<td>N/A</td>
</tr>
<tr>
<td>49</td>
<td>CGCCTGG / CGTCTTA</td>
<td>1</td>
<td>0.66</td>
<td>N/A</td>
</tr>
<tr>
<td>50</td>
<td>GGTCTTA / CGCTCTG</td>
<td>1</td>
<td>0.36</td>
<td>N/A</td>
</tr>
<tr>
<td>51</td>
<td>CGCTTGG / CGCTTGG</td>
<td>1</td>
<td>0.54</td>
<td>N/A</td>
</tr>
<tr>
<td>52</td>
<td>CGCTTGG / CGTCTTA</td>
<td>1</td>
<td>0.54</td>
<td>N/A</td>
</tr>
<tr>
<td>53</td>
<td>GGTCTTA / CGCTTGG</td>
<td>1</td>
<td>0.43</td>
<td>N/A</td>
</tr>
<tr>
<td>54</td>
<td>GGTCTTA / CGCTTGG</td>
<td>1</td>
<td>0.33</td>
<td>N/A</td>
</tr>
<tr>
<td>55</td>
<td>GTCTGA / GGTCTTA</td>
<td>1</td>
<td>0.22</td>
<td>N/A</td>
</tr>
<tr>
<td>56</td>
<td>GGTCTTA / CGTCTTA</td>
<td>1</td>
<td>0.14</td>
<td>N/A</td>
</tr>
<tr>
<td>57</td>
<td>GGTCTTA / GGTCTTA</td>
<td>1</td>
<td>0.49</td>
<td>N/A</td>
</tr>
<tr>
<td>58</td>
<td>GGTCTTA / CGCCTGG</td>
<td>1</td>
<td>0.12</td>
<td>N/A</td>
</tr>
<tr>
<td>59</td>
<td>GGTCTTA / GGTCTTA</td>
<td>1</td>
<td>0.61</td>
<td>N/A</td>
</tr>
</tbody>
</table>
With respect to the simplified group of five novel SNPs, there were 25 distinct diplotypes total in the liver bank (Table 22) and 40 diplotypes in PNAT (Table 23). The full reference diplotype (GTTTA/GTTTA; reference allele designated according to dbSNP (Sherry et al., 2001)) was present in 6.6% and 6.4% of samples from the subset of the liver bank and PNAT, respectively. Of note, not all diplotypes observed in the liver bank appeared in PNAT; diplotypes #12, 16-21, and 23-24 (Table 22) were not found among PNAT subjects (Table 23). Furthermore, 24 additional diplotypes were present in PNAT, which were not identified in the liver bank (diplotypes #16-39, Table 23).

Table 22 | Five-SNP phased diplotypes predicted in donors in the liver bank.

<table>
<thead>
<tr>
<th>Diplotype Group #</th>
<th>Full Diplotype</th>
<th>N in liver bank</th>
<th>CYP2A6 activity (nicotine metabolism, nmol/min/mg), mean of diplotype group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>G T T T A / G T T T A</td>
<td>17</td>
<td>0.07</td>
</tr>
<tr>
<td>1</td>
<td>G C C G G / G C C G G</td>
<td>64</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>G T T T A / G C C G G</td>
<td>60</td>
<td>0.09</td>
</tr>
<tr>
<td>3</td>
<td>G T T G A / G C C G G</td>
<td>22</td>
<td>0.12</td>
</tr>
<tr>
<td>4</td>
<td>G C C G A / G C C G G</td>
<td>17</td>
<td>0.10</td>
</tr>
<tr>
<td>5</td>
<td>G T T T A / G T T G A</td>
<td>15</td>
<td>0.08</td>
</tr>
<tr>
<td>6</td>
<td>G C C G G / A C C G G</td>
<td>14</td>
<td>0.11</td>
</tr>
<tr>
<td>7</td>
<td>G T T T A / A C C G G</td>
<td>12</td>
<td>0.14</td>
</tr>
<tr>
<td>8</td>
<td>G T T T A / G C C G A</td>
<td>8</td>
<td>0.08</td>
</tr>
<tr>
<td>9</td>
<td>G T T G A / G C C G A</td>
<td>5</td>
<td>0.06</td>
</tr>
<tr>
<td>10</td>
<td>G T T G A / A C C G G</td>
<td>4</td>
<td>0.03</td>
</tr>
<tr>
<td>11</td>
<td>G C C G A / A C C G G</td>
<td>3</td>
<td>0.31</td>
</tr>
<tr>
<td>12</td>
<td>G T T G A / G T T G A</td>
<td>2</td>
<td>0.04</td>
</tr>
<tr>
<td>13</td>
<td>G T T T A / G C C T G</td>
<td>2</td>
<td>0.08</td>
</tr>
<tr>
<td>14</td>
<td>A C C G G / A C C G G</td>
<td>1</td>
<td>0.03</td>
</tr>
<tr>
<td>15</td>
<td>G C C G A / G C C G A</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>16</td>
<td>G C T G A / A C C G G</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>17</td>
<td>G C T G A / G C C G G</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>18</td>
<td>G C T T A / G C C G G</td>
<td>1</td>
<td>0.04</td>
</tr>
<tr>
<td>19</td>
<td>G T C G A / G C C G A</td>
<td>1</td>
<td>0.23</td>
</tr>
<tr>
<td>20</td>
<td>G T C G A / G C C G G</td>
<td>1</td>
<td>0.03</td>
</tr>
<tr>
<td>21</td>
<td>G T C G G / A C C G G</td>
<td>1</td>
<td>0.12</td>
</tr>
<tr>
<td>22</td>
<td>G T C G G / G C C G G</td>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>23</td>
<td>G T T G A / G T C G G</td>
<td>1</td>
<td>0.19</td>
</tr>
<tr>
<td>24</td>
<td>G T T T G / G T C G G</td>
<td>1</td>
<td>0.30</td>
</tr>
</tbody>
</table>

a. Full reference diplotype is bolded
Table 23 | Five-SNP phased diplotype predicted in treatment-seeking smokers in PNAT trial.

<table>
<thead>
<tr>
<th>Diploptype Group #</th>
<th>Full Diplotype</th>
<th>N in PNAT</th>
<th>CYP2A6 activity (nicotine metabolite ratio, NMR), mean of diplotype group</th>
<th>Corresponding liver bank diplotype group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>G T T T A / G T T T A</td>
<td>41</td>
<td>0.3</td>
<td>Reference Group</td>
</tr>
<tr>
<td>1</td>
<td>G C C G G / G C C G G</td>
<td>110</td>
<td>0.55</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>G T T T A / G C C G G</td>
<td>123</td>
<td>0.40</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>G T T G A / G C C G G</td>
<td>43</td>
<td>0.46</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>G C C G A / G C C G G</td>
<td>55</td>
<td>0.55</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>G T T T A / G T T G A</td>
<td>20</td>
<td>0.38</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>G C C G G / A C C G G</td>
<td>38</td>
<td>0.49</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>G T T T A / A C C G G</td>
<td>29</td>
<td>0.42</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>G T T T A / G C C G A</td>
<td>30</td>
<td>0.43</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>G T T G A / G C C G A</td>
<td>8</td>
<td>0.42</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>G T T G A / A C C G G</td>
<td>8</td>
<td>0.54</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>G C C G A / A C C G G</td>
<td>10</td>
<td>0.54</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>G T T T A / G C T G</td>
<td>5</td>
<td>0.48</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>A C C G G / A C C G G</td>
<td>3</td>
<td>0.45</td>
<td>14</td>
</tr>
<tr>
<td>14</td>
<td>G C C G A / G C C G A</td>
<td>11</td>
<td>0.41</td>
<td>15</td>
</tr>
<tr>
<td>15</td>
<td>G T C G G / G C C G G</td>
<td>2</td>
<td>0.21</td>
<td>22</td>
</tr>
<tr>
<td>16</td>
<td>G C T G G / G C C G G</td>
<td>23</td>
<td>0.54</td>
<td>N/A</td>
</tr>
<tr>
<td>17</td>
<td>G T T T A / G C T G G</td>
<td>13</td>
<td>0.45</td>
<td>N/A</td>
</tr>
<tr>
<td>18</td>
<td>G C C G G / A C T G G</td>
<td>12</td>
<td>0.46</td>
<td>N/A</td>
</tr>
<tr>
<td>19</td>
<td>G T T G G / G C C G G</td>
<td>11</td>
<td>0.46</td>
<td>N/A</td>
</tr>
<tr>
<td>20</td>
<td>G C C T G / G C C G G</td>
<td>10</td>
<td>0.60</td>
<td>N/A</td>
</tr>
<tr>
<td>21</td>
<td>G C C G A / A C T G G</td>
<td>6</td>
<td>0.51</td>
<td>N/A</td>
</tr>
<tr>
<td>22</td>
<td>G C T G G / G C C G A</td>
<td>5</td>
<td>0.54</td>
<td>N/A</td>
</tr>
<tr>
<td>23</td>
<td>G C T T G / G C C G G</td>
<td>4</td>
<td>0.67</td>
<td>N/A</td>
</tr>
<tr>
<td>24</td>
<td>G T T T A / G C T G A</td>
<td>4</td>
<td>0.45</td>
<td>N/A</td>
</tr>
<tr>
<td>25</td>
<td>G T T G A / G T T G G</td>
<td>2</td>
<td>0.35</td>
<td>N/A</td>
</tr>
<tr>
<td>26</td>
<td>G T T T A / G T T G G</td>
<td>2</td>
<td>0.33</td>
<td>N/A</td>
</tr>
<tr>
<td>27</td>
<td>G C C T G / A C C G G</td>
<td>1</td>
<td>0.69</td>
<td>N/A</td>
</tr>
<tr>
<td>28</td>
<td>G T T G G / A C C G G</td>
<td>1</td>
<td>0.35</td>
<td>N/A</td>
</tr>
<tr>
<td>29</td>
<td>A C T G G / A C T G G</td>
<td>1</td>
<td>0.46</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>------------------</td>
<td>---</td>
<td>-------</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>G T T T A / A C T G G</td>
<td></td>
<td>0.53</td>
</tr>
<tr>
<td>31</td>
<td></td>
<td>G C C T G / G C C G A</td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>32</td>
<td></td>
<td>G C C T G / G C C T G</td>
<td></td>
<td>0.30</td>
</tr>
<tr>
<td>33</td>
<td></td>
<td>G C T T G / G C C T G</td>
<td></td>
<td>0.48</td>
</tr>
<tr>
<td>34</td>
<td></td>
<td>G C T G A / G C T G A</td>
<td></td>
<td>0.61</td>
</tr>
<tr>
<td>35</td>
<td></td>
<td>G C T G G / G C T G G</td>
<td></td>
<td>0.54</td>
</tr>
<tr>
<td>36</td>
<td></td>
<td>G T T G A / G C T G G</td>
<td></td>
<td>0.43</td>
</tr>
<tr>
<td>37</td>
<td></td>
<td>G T T G G / G C T G G</td>
<td></td>
<td>0.33</td>
</tr>
<tr>
<td>38</td>
<td></td>
<td>G T T T A / G C T T G</td>
<td></td>
<td>0.36</td>
</tr>
<tr>
<td>39</td>
<td></td>
<td>G T T T A / G T C G A</td>
<td></td>
<td>0.22</td>
</tr>
</tbody>
</table>

a. Full reference diplotype is bolded
Association of five-SNP diplotypes and CYP2A6 activity in vitro and in vivo

We compared the rate of \textit{in vitro} CYP2A6 activity (rate of cotinine formation from nicotine) of the top 5 most frequent diplotypes in the liver bank (diplotypes #1-5, Table 22) to that of the reference diplotype (Table 22), which does not possess the variant form of any of the five SNPs (Fig. 28a). Each of the 5 most frequent diplotypes had higher mean CYP2A6 activity compared to the reference diplotype ($P=0.23$). Additionally, when collapsed together into one group, the 5 most frequent diplotypes collectively had non-significantly higher mean CYP2A6 activity compared to the reference diplotype ($P=0.28$), which was also observed for the remaining diplotypes combined (diplotypes #6-24, Table 22; $P=0.28$), and when all variant diplotypes were grouped together (diplotypes #1-24; $P=0.26$).

To increase our power for assessing associations between diplotypes and CYP2A6 activity (N=252 subjects analyzed in the liver bank), we extended these analyses to the PNAT sample (N=536). In PNAT, we compared the mean NMR between the reference diplotype and the top 5 most frequent diplotypes that were found in the liver bank (diplotypes #1-5, Table 23; Fig. 28b). Compared to the full reference diplotype, diplotypes #1-4 (Table 23) each had significantly higher mean NMR ($P$ values $0.008-<0.0001$). Diploptotype #5, which only possesses the variant form of one of the five SNPs, exhibited non-significantly higher mean NMR ($P>0.99$), relative to the reference diplotype. The combined diplotype groups #1-5, #6-24, and #1-24 had significantly higher mean NMR compared to the reference diplotype (all $P<0.0001$). Parallel analyses conducted with the full seven-SNP diplotype illustrated very similar patterns of association for diplotypes with CYP2A6 \textit{in vitro} and \textit{in vivo} activity (Fig. 29).
A  

Liver Donors

B  

Treatment-Seeking Smokers

Figure 28 | Comparison of CYP2A6 enzyme activity of the full reference diplotype group and top 5 most frequent five-SNP CYP2A6 diplotypes. Data were restricted to White subjects, and to those without established CYP2A6 genetic variants. Enzyme activity for the total group (i.e. before splitting into diplotype groups) has been provided as a reference in each figure (far left). The diplotype group numbers refer to Tables 22 and 23. (A) Data from the human liver bank. In vitro CYP2A6 activity was measured as the rate of cotinine formation (nmol/min/mg) from nicotine (Tanner et al., 2017). (B) Data from the PNAT trial. In vivo CYP2A6 activity was measured using the ratio of nicotine’s metabolites (3HC/COT, nicotine metabolite ratio, NMR) (Tanner et al., 2015b, Dempsey et al., 2004). Note: The N=25 diplotype groups found in the liver bank (Table 22) were compared to the same diplotype groups in PNAT, to maintain consistency of the comparison; it should be noted however that some low frequency diplotype groups in the liver bank were not found in PNAT (diplotypes #12, 16-21, and 23-24 from Table 22), and likewise several low frequency diplotype groups in PNAT were not found in the liver bank (diplotypes #16-39; Table 23). P values are based on Kruskal Wallis and Dunns Multiple Comparison tests, or Mann-Whitney tests.

SNP IDs that make up the CYP2A6 diplotype (in order):
rs28399453 (G>A), rs56113850 (T>C), rs150298687 (T>C), rs7260629 (T>G), rs57837628 (A>G)

Diplotypes
Reference: G T T T A / G T T T A
Group #1: G C C C G / G C C C G
Group #2: G T T T A / G C C C G
Group #3: G T T G A / G C C C G
Group #4: G C C G A / G C C C G
Group #5: G T T T A / G T T G A

Table 2

<table>
<thead>
<tr>
<th>Diplotype Group</th>
<th>NMR</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>0.05</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Group #1</td>
<td>0.12</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Group #2</td>
<td>0.05</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Group #3</td>
<td>0.23</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Group #4</td>
<td>0.28</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Group #5</td>
<td>0.26</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Table 22

<table>
<thead>
<tr>
<th>Diplotype Group</th>
<th>COTinine Formation (nmol/min/mg)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>0.05</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Group #1</td>
<td>0.12</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Group #2</td>
<td>0.05</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Group #3</td>
<td>0.23</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Group #4</td>
<td>0.28</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Group #5</td>
<td>0.26</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

167
Figure 29 | Comparison of CYP2A6 enzyme activity of the full reference diplotype group and top 5 most frequent seven-SNP CYP2A6 diplotypes. Data were restricted to White subjects, and to those without established CYP2A6 genetic variants. Enzyme activity for the total group (i.e. before splitting into diplotype groups) has been provided as a reference in each figure (far left). The diplotype group numbers refer to Tables 20 and 21. (A) Data from the human liver bank. In vitro CYP2A6 activity was measured as the rate of cotinine formation (nmol/min/mg) from nicotine (Tanner et al., 2017). (B) Data from the PNAT trial. In vivo CYP2A6 activity was measured using the ratio of nicotine’s metabolites (3HC/COT, nicotine metabolite ratio, NMR) (Tanner et al., 2015b, Dempsey et al., 2004). Note: The N=44 diplotype groups found in the liver bank (Table 20) were compared to the same diplotype groups in PNAT, to maintain consistency of the comparison; it should be noted however that some low frequency diplotype groups in the liver bank were not found in PNAT (diplotypes #14, 17, 20, 22-24, 27-31, 33-39, and 42-43 from Table 20), and likewise several diplotype groups in PNAT were not found in the liver bank (diplotypes #24-59; Table 21). P values are based on Kruskal Wallis and Dunns Multiple Comparison tests, or Mann-Whitney tests.
5.5 Discussion

We have identified seven novel SNPs, and a resulting simplified five-SNP diplotype, at the *CYP2A6* locus, which significantly impact CYP2A6 activity assessed in both a human liver bank *in vitro* and using the NMR *in vivo*. All seven novel SNPs characterized here were non-coding variants, present in regulatory and intronic regions of the *CYP2A6* gene, with MAFs >1% in European ancestry populations (liver bank: 7−72%, PNAT: 7−69%). Non-coding variants are largely understudied, despite important roles in regulating transcription, chromatin state, splicing, and epigenetic modifications; this work further highlights the utility of full gene or genome sequencing in characterizing complex traits and diseases, such as smoking.

For the novel SNPs characterized here that were in high linkage disequilibrium (LD), allowing for incorporation of only one of the two SNPs into the simplified five-SNP diplotype, LD was similarly moderate-to-high in a GWAS of the NMR in a Finnish population (rs57837628 and rs8192733 $r^2=0.73$, rs7260629 and rs7259706 $r^2=0.98$) (Loukola et al., 2015). Likewise, associations of *CYP2A6* diplotypes with *in vitro* and *in vivo* rate of nicotine metabolism were consistent for the seven- and five-SNP diplotypes (see Figs. 28 and 29), and there was little change in the amount of variation in CYP2A6 enzyme activity accounted for by all seven SNPs compared to only five SNPs (6.6 vs. 6.0%). This supports the use and assessment of the five-SNP diplotype, as opposed to all seven SNPs, in CYP2A6 phenotype association studies and in future genotype-prediction algorithms. As we were able to capture significantly more of the variation in CYP2A6 activity with the five-SNP diplotype (additional 3.9% compared to established variants alone), we may be able to use genotype to tailor smoking cessation. As feasibility of genomics-based treatment optimization increases, implementation of pre-emptive genotyping will require further assessment of both coding- and non-coding genetic variation, and incorporation of genotypes into functionally significant diplotypes, as demonstrated here.

Our *in vitro* and *in vivo* functional characterization of novel SNPs expands on previous GWAS’s that have identified significant hits at *CYP2A6*. Consistent with our prioritization, six of the seven SNPs in our investigation reached genome-wide significance (rs28399453 did not, $P=6.42E-06$) in the Finnish GWAS of the NMR, with rs56113850 being the top hit (most highly significant, $P=5.77E-86$) (Loukola et al., 2015). rs56113850 was also the top hit for association with NMR in a multi-ethnic cohort (MEC) study in which it was globally
significantly associated with higher CYP2A6 activity \((P=1.19E-50)\), and it was the second most significant SNP associated with lung cancer risk in the TRICL consortium GWAS \((P=5.78E-11)\) \((Patel\ et\ al.,\ 2016)\). Similarly, rs56113850 was the top ranked European American SNP in a separate GWAS of NMR \((P=3.81E-10)\) \((Baurley\ et\ al.,\ 2016)\). SNPs rs57837628 and rs8192733 were also globally significant in the MEC \((P=6.84E-37\) and \(2.78E-08\), respectively), and were associated with increased lung cancer risk in the TRICL consortium GWAS \((P=4.01E-10\) and \(2.10E-9\), respectively) \((Patel\ et\ al.,\ 2016)\). The remaining SNPs, rs7260629, rs7259706, rs150298687, and rs28399453, were not significant in the MEC or TRICL consortium GWAS’s \((Patel\ et\ al.,\ 2016)\). Lack of genome-wide significance of rs28399453 in the Finnish GWAS of NMR may result from lower allele frequency (liver bank: 7\%, PNAT: 7\%, 1000 Genomes: 6\% for European ancestry) relative to the other six novel SNPs (liver bank: 47\–72\%, PNAT: 47\–69\%, 1000 Genomes: 48\–70\% for European ancestry), and in the MEC study due to even lower allele frequencies across populations of other ethnic origins (0, 0, and 2\% among African, Asian, and Latin American ancestry populations, respectively, from 1000 Genomes).

We also identified novel exon variation in CYP2A6 through sequencing \((N=46\) exon SNPs); only \(N=6\) missense and \(N=1\) nonsense SNPs were found, all of which were rare \((MAF<1\%\) in 1000 genomes). This finding is consistent with the majority of CYP genes where 93\% and 83\% of coding region SNPs are rare \((MAF<1\%)\) or very rare \((MAF<0.1\%)\), respectively \((Fujikura\ et\ al.,\ 2015)\). Due to the infrequency of nonsynonymous coding region SNPs in the current study, we were not powered to characterize functional impact in the liver bank \((Table\ 17)\). Functional investigation of these SNPs may require the use of an \textit{in vitro} CYP2A6 cDNA expression system \((Al\ Koudsi\ et\ al.,\ 2009)\), in combination with the BioBin software, which improves power to detect phenotype associations by combining rare variants \((Moore\ et\ al.,\ 2016)\). Further, while we were able to narrow down our list of SNPs identified through sequencing from \(N=229\) to \(N=21\) using bioinformatics tools, it is likely that we have missed one or more impactful SNPs due to the limitations of these predictive tools. As technologies continue to develop and we learn more about the CYP2A6 gene, the functional significance of these \(N=229\) SNPs could be reassessed.

In addition to a lack of power for investigating rare coding variation, a study limitation is the assessment of White populations only. Frequency of CYP2A6 genetic variation differs
substantially across ethnic groups (Tanner et al., 2015a), suggesting that our results may not extend to other ethnic populations. Additionally, due to haplotype and LD structure heterogeneity across different populations (Tishkoff and Verrelli, 2003), the utility of the simplified five-SNP diplotype for predicting CYP2A6 activity and nicotine metabolism should be reassessed independently among other ethnic groups.

**Conclusions**

Seven novel high frequency CYP2A6 SNPs were identified as important contributors to CYP2A6 phenotypes, accounting for approximately two times more variation in CYP2A6 activity, compared to established CYP2A6 variants alone. Due to high LD between two pairs of the seven novel SNPs, we established a functional five-SNP haplotype, which, when phased to derive diplotypes, was associated with CYP2A6 enzyme activity *in vitro* and *in vivo* in a human liver bank and PNAT clinical trial, respectively. Considering the association of CYP2A6 enzyme activity with smoking behavior and cessation outcomes, it will be important to determine the utility of incorporating the five-SNP diplotype into CYP2A6 genotype models for predicting cessation success and response to pharmacotherapies.
5.6 Significance to Thesis

This chapter extends the findings of chapter 3, and adds to the literature in three ways. First, this study demonstrated the functional significance of seven novel (i.e. previously uncharacterized) high frequency CYP2A6 genetic variants. This is the first time that these variants have been characterized with respect to CYP2A6 mRNA expression, protein levels, and *in vitro* enzyme activity. Second, the novel SNPs accounted for an additional 4.5% of the total missing variation in *in vitro* CYP2A6 enzyme activity. As the seven genetic variants were located in regions 5’, intronic, or 3’ of the CYP2A6 gene, this underscores the importance of non-coding genetic variation in the modulation of CYP2A6 enzyme activity. Future studies should increasingly focus on the assessment of functionally significant non-coding genetic variation.

Lastly, this chapter established the moderate-to-high linkage disequilibrium between several of the novel genetic variants. Through haplotype and diplotype analyses, this study established novel and functionally significant five-SNP CYP2A6 diplotypes, associated with both *in vitro* and *in vivo* rates of nicotine metabolism. The demonstrated associations imply that the novel five-SNP diplotypes may be clinically relevant for incorporation into future personalized medicine prediction algorithms to improve cessation success for smokers. The NMR has been proven to be a predictive biomarker for cessation treatment in a clinical trial of smokers; therefore, as these novel diplotypes exhibit a significant impact on NMR, they may be clinically relevant for cessation treatment.

This study provides support for future work investigating the functional significance of all seven novel SNPs, and the simplified five-SNP diplotypes, in populations of other ethnic/racial backgrounds.
6 GENERAL DISCUSSION

6.1 Summary of Research Findings

The studies presented in this thesis have investigated smoking, tobacco exposure, and biological factors associated with smoking in the NP and SW tribal populations. We also assessed predictors of nicotine metabolism rate, including novel genetic sources of variation, in a human liver bank. Chapter 1 investigated the active and passive exposure to tobacco smoke via the measurement of the established biomarker of nicotine intake, cotinine. This work showed that both tribal populations have low levels of smoking, on the basis of self-reported CPD and plasma/saliva cotinine measurements. Conversely, our findings suggest that prevalence of passive tobacco smoke exposure was relatively high in both the NP and SW. We further identified several factors that contribute to tobacco exposure (cotinine levels) in smokers and non-smokers in these tribes. Based on our observations of distinct smoking quantities and levels of tobacco exposure, along with previous reports of very different smoking prevalence and lung cancer incidence between the two tribes, we then investigated genetic and biological factors that may impact smoking behaviours in the NP and SW. In Chapter 2, we compared $CYP2A6$ genetic variation and the rate of nicotine metabolism between the two tribes, and identified distinct tribal genetic patterns, and a significantly faster rate of nicotine metabolism in the NP than the SW, and compared to other ethnic/racial groups. In order to assess factors that may contribute to variation in the rate of nicotine metabolism, Chapter 3 investigated both genetic and non-genetic predictors of variation in CYP2A6 activity in a human liver bank. We accounted for more than 75% of the variation in in vitro CYP2A6 activity. Extending this work in Chapter 4, we identified seven high-frequency (European ancestry) novel SNPs at the $CYP2A6$ gene locus, which were in moderate-to-high LD, and were associated with CYP2A6 activity. Further, we identified a functionally significant five-SNP diplotype, which was associated with both in vitro and in vivo nicotine metabolism. In the following sections, we will discuss mechanistic explanations for these findings, describe broad implications of this work, and propose areas for future research.
6.2 What Mediates the Low Smoking Quantities in the Northern Plains and Southwest Tribal Populations?

Despite the high prevalence of smoking in the NP (approximately 50%) (Nez Henderson et al., 2005), we have demonstrated that smokers in this tribe, and in the SW tribe, are light smokers, on the basis of both self-reported CPD (NP 7, SW 4) and cotinine levels (NP 82 ng/ml, SW 45 ng/ml). Further, there are individuals in both tribal populations who self-reported smoking several CPD, but who exhibited no detectable cotinine (Figure 10, thesis Chapter 1), further demonstrating that there are many individuals who do not even smoke daily, but rather smoke more intermittently, as 18% and 47% of smokers in the NP and SW, respectively, reported intermittent (non-daily) smoking. The explanation(s) for their patterns of light smoking are unknown. Here we will focus on genetic factors that may contribute to smoking quantity in both tribes, while also briefly expanding on an additional non-genetic impact, socioeconomic status.

6.2.1 Genetic Impacts

6.2.1.1 CYP2A6 and Nicotine Metabolism

As outlined in section 1.6.2 of the General Introduction, CYP2A6 genotype and the NMR are associated with smoking quantities, determined based on self-reported CPD or using biochemical markers of consumption. Smokers with a faster rate of nicotine metabolism, according to genotype or the NMR phenotype, smoke more CPD (Pan et al., 2015), or smoke each cigarette more intensely (Rao et al., 2000), compared to smokers with a slower rate of nicotine metabolism. Smokers titrate their quantity of nicotine intake such that they maintain sufficient levels throughout the day to avoid cravings and withdrawal symptoms (McMorrow and Foxx, 1983).

The light smoking observed in both the NP and SW tribes appears to contradict this theory of regulating smoking levels according to nicotine metabolism rate. Considering the very high rate of nicotine metabolism that was observed among NP smokers, relative to the SW and other ethnic/racial groups, we would expect to also see relatively high levels of smoking in this tribe. Instead, according to self-reported CPD and cotinine, both tribes appear to consist of relatively light smokers. However, there are several factors to consider when interpreting CPD and
cotinine as measures of tobacco consumption. First, as mentioned in section 1.6.2 of the General Introduction, self-reported CPD may not be an accurate measure of nicotine intake in all smokers. For light smokers (<10 CPD) who smoke their cigarettes relatively intensely, this self-report measure will underestimate consumption (Zhu et al., 2013a). One caveat to using cotinine as a biomarker of smoking quantity is the non-linear relationship between cotinine formation and removal. Nicotine is metabolized to cotinine by CYP2A6, and CYP2A6 further metabolizes cotinine to 3’-hydroxycotinine (Messina et al., 1997, Nakajima et al., 1996a). However, as introduced in section 1.4.2 of the General Introduction, variation in CYP2A6 metabolic activity has a larger impact on cotinine’s removal than on its formation. Due to differences in the intrinsic clearance of nicotine (higher) and cotinine (lower), faster CYP2A6 activity increases cotinine’s metabolism and removal more than cotinine’s formation, which is more dependent on changes in liver blood flow as opposed to metabolic activity (Zhu et al., 2013c). Whereas cotinine may accumulate in slow metabolizers, levels will be diminished in faster metabolizers. This suggests that the relatively low overall cotinine levels among NP smokers may be underestimating their smoking quantity as a result of their overall faster CYP2A6 activity (high NMR).

It is possible that CPD and cotinine are not appropriate measures of consumption in these light smoking (i.e. NP and SW) and fast metabolizing (i.e. NP) populations, and rather a more specific and robust biomarker, such as urinary TNE, would be more representative of their level of consumption. This has previously been demonstrated in a light smoking (according to CPD) Alaska Native population. Alaska Natives smoke an average of 7 CPD and exhibit a similarly high rate of nicotine metabolism (NMR) as NP smokers (Fig. 17 in Chapter 2, section 3.4). However, through the measurement of urinary TNE, Zhu et al. (2013a) showed that Alaska Native smokers have a similar level of nicotine consumption as White heavy smokers, suggesting that they smoke each cigarette more intensely. This study also showed that Alaska Native smokers titrated their intake (determined by TNE) according to their rate of nicotine metabolism. Considering the similarities between the NP and SW tribes and Alaska Native smokers, the same may be true for NP and SW smokers; they may consume relatively low numbers of CPD, while adjusting their smoking topography such that they inhale from each cigarette more deeply or take more puffs per cigarette.
In order to assess this, we would ideally measure urinary TNE among NP and SW smokers. Although TNE would have been the optimal measure for quantifying nicotine intake and tobacco exposure among NP and SW smokers and non-smokers, the measurement of urinary TNE was not planned, funded, nor approved by ethics review boards and IRBs; the study participants did not consent to the collection of 24-hour urine samples, which would be required for the assessment of urinary TNE. This is an important future direction that will be described in more detail below. Based on available data from the current study (Chapter 1), we can compare the mean CPD and mean cotinine levels across NP, SW, and Alaska Native light smokers, and White heavy smokers. If the tribes exhibit a similarly high smoking intensity as Alaska Native smokers, and therefore an overall level of tobacco consumption similar to White heavy smokers, we would expect to observe consistent cotinine levels across all four populations. Instead we see that, despite smoking the same number of CPD and having similar rates of nicotine metabolism, Alaska Natives have cotinine levels more than 2X greater than NP smokers, and almost 4X greater than SW smokers. In fact, the relationship between cotinine and CPD among the NP and SW smokers is consistent with that observed among White heavy smokers; the tribes exhibit a cotinine/CPD ratio similar to Whites. The low cotinine levels in NP smokers do not appear to correspond to their high rate of nicotine metabolism, whereas Alaska Native smokers exhibit both high cotinine levels and high NMR (Fig. 30A).

Together these data suggest several things. First, the faster CYP2A6 activity, and resulting faster cotinine removal, in the NP may not account for their low cotinine levels; Alaska Natives, who exhibit similarly high CYP2A6 activity and low CPD, have high cotinine levels that reflect their high tobacco consumption (according to TNE). A future study that includes the measurement of TNE in the NP and SW would be useful in confirming the overall low levels of tobacco consumption in these tribes. Based on the described cotinine analyses across the four populations (Fig. 30A), and using available TNE (6-item) and CPD data for Alaska Native and White smokers, we have predicted the relationship between TNE and CPD (indicative of smoking intensity) in the NP and SW tribes (Fig. 30B). Tribal TNE levels were predicted using their known average cotinine levels; we fit the NP and SW cotinine levels to the regression model derived from the known cotinine and TNE levels of Alaska Native and White smokers. These predictions suggest that there is a similar relationship between TNE and CPD among NP and SW smokers as was observed for White smokers (Fig. 30C); the intensity
at which NP and SW subjects smoke their cigarettes (i.e. nicotine intake per cigarette) appears to be similar to the smoking intensity of White heavy smokers and appears much lower than Alaska Native light smokers.

Figure 30 | (A) Association of number of cigarettes smoked per day with cotinine levels. (B) Approach for predicting NP and SW total nicotine equivalents (TNE). (C) Association of CPD and predicted TNE for NP and SW smokers, and known TNE for Alaska Native and White smokers. Dotted lines in A and C represent the regression line for NP, SW, and Whites (excluding Alaska Natives). TNE for NP and SW are theoretical values that were estimated according to the model depicted in part B (see orange arrows); known mean NP and SW cotinine levels were fit to the regression line derived from the established relationship between TNE and cotinine levels in Alaska Natives and Whites. It was assumed that TNE increases proportionately with cotinine levels. Cotinine, CPD, and TNE data for Alaska Native and White smokers are from studies by Zhu et al. (2013a) and St Helen et al. (2013a), respectively.
However, as we have not actually measured TNE in the NP and SW, we cannot rule out the possibility that the NP have more moderate, as opposed to low, levels of tobacco consumption, and that the high NMR among NP smokers may be contributing to their low cotinine levels; the fast cotinine metabolism in the NP may result in the appearance of being light smokers. If NP smokers metabolized cotinine more slowly (i.e. lower NMR) they may have appeared to have higher tobacco consumption, when using cotinine as a biomarker. In the same respect, Alaska Natives, who also have high NMR, may have appeared to have even higher tobacco consumption if their NMR was lower, according to cotinine levels; however, because TNE has been measured among Alaska Native smokers, we know their level of tobacco consumption is similar to White heavy smokers. This again supports estimating tobacco consumption through measurement of urinary TNE in both the NP and SW tribes.

A second inference that can be made from these data is that neither NP nor SW smokers appear to titrate their nicotine intake according to their rate of nicotine metabolism. Considering the NP population’s significantly faster rate of nicotine metabolism relative to the SW, we would expect NP smokers to exhibit much higher overall CPD and cotinine levels than SW smokers, similar to what is observed among the fast metabolizing Alaska Native smokers; this would be indicative of a level of tobacco consumption that is somewhat proportional to their rate of nicotine metabolism and clearance. However, the similar CPD and cotinine levels of these two populations suggest that NP smokers are not adjusting their tobacco intake proportionally to their nicotine metabolism rate. It is possible that tobacco intake is moderately higher in the NP than the SW, such that some titration of nicotine levels is occurring, but the NP fast metabolism has them appearing to be lighter smokers; however, based on these data, the NP smokers still appear to smoke less than the similarly fast metabolizing Alaska Native population. As a future direction, we propose measuring urinary TNE in both populations, and comparing this between slow and fast metabolizers, or CYP2A6 genotype groups, as we are unable to confirm the NP and SW subjects’ titration of nicotine intake based on the cotinine data. Although their tobacco consumption, according to CPD and cotinine, does not correspond to their overall rate of nicotine metabolism, there may be subtle differences in smoking quantity within each tribe, according to CYP2A6 activity.

The apparent lack of association between CYP2A6 and smoking observed in the NP and SW tribes may stem from their relatively low levels of nicotine dependence, according to the
FTND (Tables 9 and 10 in Chapter 2, section 3.4). On average, NP and SW smokers obtained FTND scores of 2.0 and 1.8, respectively. FTND scores of 1-2 denote low nicotine dependence, followed by scores of 3-4 for low-to-moderate dependence, 5-7 for moderate dependence, and ≥8 for high dependence. Schoedel et al. (2004) and Audrain-McGovern et al. (2007) have shown, in both adults and adolescents, that smokers did not titrate their nicotine intake if they were not dependent (according to DSM-IV criteria and the modified Fagerstrom Tolerance Questionnaire, mFTQ). Therefore, if NP and SW smokers are not reaching a moderate-to-high level of nicotine dependence, they may not be titrating their nicotine intake to sustain constant daily levels.

Alternatively, the FTND may not be a suitable measure of tobacco dependence for NP and SW smokers. The FTND is generally a poor measure of nicotine dependence in light smokers, due to the emphasis on CPD in the FTND score. In the FTND, the question assessing CPD scores individuals as follows: smoking <10 CPD is a score of 0, 11-20 CPD is 1, 21-30 CPD is 2, and ≥31 CPD is 3. In the NP and SW, 82% and 92% of smokers, respectively, scored 0 on this question. It was even common for NP and SW smokers to report light and intermittent smoking, with approximately 18% and 47% of NP and SW smokers, respectively, reporting smoking fewer than 30 cigarettes per month, and therefore an average of less than one CPD. In African American light smokers, Mwenifumbo and Tyndale (2011) similarly showed that 70% of smokers scored 0 on this question, but that the question could be rescored in order to better reflect that population’s light smoking patterns. This rescaling more than doubled the number of smokers who met criteria for dependence. Further, the agreement between FTND and other measures of nicotine dependence was poor (DSM-IV) or fair (ICD-10) in African American light smokers (Mwenifumbo and Tyndale, 2011). This suggests that additional or modified tests should be considered when quantifying a light smoking population’s level of nicotine dependence. The mFTQ that is used among adolescent smokers may be more appropriate for the light smoking NP and SW populations. The questions on the mFTQ are similar to the FTND, however the scoring of the mFTQ differs such that it allows for a greater variety of responses, due to more response options to choose from, and it also takes into account whether smokers inhale or not (Prokhorov et al., 1996).

In the studies described in Chapters 1 and 2 (thesis sections 2 and 3), the HONC was also used as a measure of nicotine dependence for NP and SW smokers. The HONC, which is typically
used in adolescents, measures the loss of autonomy over tobacco use, and any score above 0 suggests that a smoker is dependent on nicotine, with the severity increasing positively with score, to a maximum of 10 (DiFranza et al., 2002). Unlike results for the FTND, NP and SW smokers scored moderately on the HONC (mean scores of 4.5 and 3.7, respectively). Like the mFTQ, the HONC may provide a better representation of dependence in the light smoking NP and SW populations. Despite the tribes’ moderate nicotine dependence, according to the HONC, they do not appear to be titrating their smoking levels according to their rate of nicotine metabolism, although this should be confirmed by measuring TNE.

6.2.1.2 Nicotinic Acetylcholine Receptors

As the light level of smoking among NP and SW smokers does not appear to be mediated by CYP2A6 genetic variation and variable nicotine metabolism rate, an alternative possibility could be genetic variation affecting components of the CNS response to nicotine. Variation in genes encoding nicotinic acetylcholine receptor (nAChR) subunits has been implicated in both tobacco consumption and nicotine dependence. The risk allele (A) of a commonly studied nonsynonymous variant present in the gene encoding the nAChR α5 subunit (CHRNA5 gene, SNP rs16969968, D398N) is associated with smoking more CPD and higher FTND scores (Breetvelt et al., 2012, Chen et al., 2009). Similarly, the rs1051730 A allele, which is in LD with the rs16969968 SNP, is associated with regular smoking among adolescents, who tend to be lighter smokers, similar to the NP and SW populations (Ducci et al., 2011). It is hypothesized that smokers with the rs16969968 A allele smoke more CPD in order to compensate for a lower level of nAChR activation, relative to smokers who possess the GG genotype. The risk allele (A) of the CHRNA5 rs16969968 SNP is associated with reduced in vitro α4β2α5 nAChR response to the nicotine agonist epibatidine, as demonstrated by a decrease in intracellular calcium levels following epibatidine binding, relative to the G allele (Bierut et al., 2008).

Whereas reduced nAChR response to nicotine may promote higher levels of smoking, higher affinity binding of nicotine to nAChRs, or sustained downstream response following receptor activation, may be associated with less smoking, as seen in the NP and SW tribes. It is possible that there are one or more genetic variants that are prevalent in these tribal populations, which enable smokers to acquire adequate or prolonged nicotine-mediated nAChR activation and downstream signaling, while smoking fewer CPD than smokers without this genetic variation.
For example, the frequency and functional impact of genetic variants in the $CHRNA5$-$A3$-$B4$ gene cluster differ across ethnic/racial groups. The rs16969968 SNP has an allele frequency of 36% in European ancestry populations, compared to 3%, 2%, and 3% in Alaska Native, Asian, and African ancestry populations, respectively (Zhu et al., 2013b, Ward and Kellis, 2012). Likely resulting from a much lower allele frequency, rs16969968 was not significantly associated with tobacco consumption in Alaska Native smokers (Zhu et al., 2013b), whereas this SNP has been associated with a 1-3 CPD increase in European ancestry smokers (Breetvelt et al., 2012, Wassenaar et al., 2011).

Sequencing key nAChR subunit genes, such as the chromosome 15 $CHRNA5$-$A3$-$B4$ gene cluster, in NP and SW subjects may result in the discovery of novel genetic variants that are more common in populations of select ethnic/racial backgrounds, similar to the $CYP2A6*17$ allele that has to date only been found in populations of African descent. Types of $CHRN$ genetic variation that could result in decreased tobacco consumption would include variants that increase downstream dopamine release and reinforcement from smoking through (1) increased expression of nAChR subunits, (2) increased affinity for nicotine, or (3) increased speed or duration of nAChR-mediated activation of dopaminergic neurons. For example, non-coding $CHRNB2$ genetic variation could impact transcription factor binding, mRNA stability, or translational efficiency resulting in increased expression of the high affinity nAChR $\beta2$ subunit, which plays a crucial role in dopaminergic activation following nicotine binding (Picciotto et al., 1995, Mameli-Engvall et al., 2006). Using the CPD and cotinine data that we have generated, one could evaluate the association between novel genetic variants and smoking quantity. That said, the relatively small sample of smokers in both tribal populations (NP N=139, SW N=90) and their overall light smoking suggest that in order to be powered to detect differences in smoking quantity, urinary TNE would need to be used as a more sensitive measure of variation in consumption. For example, in N=163 Alaska Native smokers, a nearly significant ($P=0.05$) association between the $CHRNA3$ SNP rs578776 (G allele) and greater tobacco consumption was observed, indicated by higher TNE levels (Zhu et al., 2013b). Of note, in the same study, cotinine levels were also higher among Alaska Native smokers with one or two copies of the G allele, compared to smokers with the A allele.

Variation occurring downstream of the nAChRs could also influence the extent or duration of reward from smoking. For example, sustained dopamine release could be a result of decreased
dopamine reuptake from the synapse, resulting from genetic variation in the gene encoding the dopamine transporter (*SLC6A3*) (Heinz et al., 2000). Variable dopamine receptor activation may also contribute to smoking levels in these tribes. A variable number of tandem repeats (VNTM) polymorphism in the dopamine receptor subtype 4 gene (*DRD4*) is associated with decreased receptor coupling with adenylyl cyclase, and presumably less downstream signaling following receptor activation (Asghari et al., 1995). The 7-repeat version of this genetic variant is also associated with smoking more CPD among male adolescent smokers of European descent compared to other genotypes at this locus (Laucht et al., 2005), although results are inconsistent across studies (Shields et al., 1998). This highlights several biologically plausible mechanisms of decreased smoking in the NP and SW tribes, and suggests that sequencing several candidate genes could be a next step in the investigation of smoking in these populations.

Sequencing at the *CYP2A6* gene locus will be described as a future direction in the upcoming section 6.4.

6.2.2 Additional Impacts

Non-genetic factors can also influence smoking levels, and may contribute to light smoking in the NP and SW tribes. As outlined in section 1.1.3.1 of the General Introduction, smoking prevalence is negatively associated with socioeconomic status (SES), such that prevalence is higher among individuals in the U.S. who have lower income and education (Jamal et al., 2016). Conversely, in developed countries, smokers of low compared to high SES are more likely to be lighter smokers, based on CPD (Hiscock et al., 2012). However, economically disadvantaged smokers have higher cotinine levels that those living in affluent households (Fidler et al., 2008), so it appears that these individuals may smoke fewer CPD, but smoke each cigarette more intensely. Among NP and SW participants sampled in the EARTH study (see section 1.2.4.4 of the General Introduction), 30% and 29%, respectively, reported an annual household income of more than $15,000, compared to 88% in the general U.S. population (United States Census Bureau, 2015). Similarly, 34% and 38% of NP and SW subjects in EARTH were unemployed, while unemployment was less than 5% in the general U.S. population (United States Department of Labor, 2017). The proportion of NP and SW subjects in EARTH who graduated high school was also lower than the general U.S.,
particularly in the SW tribe: 71% and 48% of the NP and SW, respectively, were high school graduates. In the United States in 2015, approximately 88% of adults 25 and older had graduated high school (Ryan and Bauman, 2016). This indicates the overall low SES of these two tribal populations. If we compare patterns of SES and smoking quantity across the NP, SW, and general U.S. populations, we see a relatively consistent trend of fewer CPD with lower household income, employment, and education (Fig. 31). These data suggest that the relatively low SES of the NP and SW populations may, in part, contribute to their low levels of tobacco consumption.

Figure 31 | Trends of smoking and socioeconomic status across the general United States, Hispanics/Latinos in the U.S., and the Northern Plains and Southwest American Indian populations. Hispanic/Latino SES data are provided for reference to another U.S. low SES population. U.S. and Hispanic/Latino CPD data are based on 2008-2010 survey in adults age≥18. Data compiled from EARTH study and (United States Census Bureau, 2015, United States Department of Labor, 2017, Ryan and Bauman, 2016, Schoenborn et al., 2013).

In order to more specifically assess the association between smoking quantity and SES in these tribes, we could reassess the measures of SES in the current cohort of NP and SW subjects. As a first line of evidence, independent analyses could be run to determine the strength and direction of correlation between measures of SES (income, employment, and education) and smoking quantity (CPD, cotinine, and TNE). As smoking quantity is likely a consequence of interplaying genetic and environmental factors, we propose investigating the independent and shared contribution of the multiple factors discussed throughout section 6.2, using linear
regression modeling. Using TNE as a measure of tobacco consumption, we could quantify the proportion of variation in TNE that is accounted for by: NMR (or CYP2A6 genotype), novel CHRN genetic variation, annual household income, employment status, and highest level of education attained. Knowledge of the major factor(s) that contributes to the level of tobacco consumption in the NP and SW tribal populations would provide insight for developing more targeted strategies to reduce and eliminate smoking. For example, these data may highlight the utility of further research into tailoring cessation treatment according to CYP2A6 genotype or the NMR, or create an argument for government support of increased education and employment in these tribes.

6.3 Origins and Genetic Diversity of the Northern Plains and Southwest Tribes

In Chapter 2 (thesis section 3) we compared NP and SW CYP2A6 genetic variation and identified distinct genetic differences between the two related tribes. Therefore, in the following section, we will discuss theories of NP and SW genetic ancestry, provide an estimation of the tribes’ level of genetic divergence, and propose areas of future research in determining ancestry and the genetic divergence of the NP and SW tribes.

6.3.1 Ancestry

All modern human populations are believed to have originated from a founding population in Africa, according to archeological and genetic evidence (Manica et al., 2007). Following the dispersal of human populations from Africa across the continents, subpopulations were established. American Indians, Alaska Natives, and Canadian Native Americans are believed to have originated from a subpopulation of East Asian ancestry that migrated via Siberia and the Bering Land Bridge (now the Bering Strait) to present day Alaska approximately 20-30 thousand years ago (Raghavan et al., 2015). Some studies support the theory that a single migration wave resulted in the establishment of American Indians/Alaska Natives, while others suggest that this migration occurred in multiple waves, and that there were therefore multiple founding populations of American Indians (Raghavan et al., 2015, Reich et al., 2012, Schroeder et al., 2009).
A founder effect likely occurred during the establishment of American Indians. A founder effect results from the establishment of a new population by a small sample of the original larger population and is characterized by a sharp decrease in genetic variation. The new smaller founding population would likely only carry a portion of the genetic variation that existed in the original larger population. Evidence of a founder’s effect occurring during the establishment of American Indians has been presented in studies of mitochondrial DNA (mtDNA) from North, Central, and South American Indians (Schurr et al., 1990, Wallace et al., 1985). Based on the presence of rare Asian ancestral mtDNA markers in these samples of American Indians, and the much lower degree of variability in their mtDNA compared to Asians, it is suggested that American Indians share common ancestral origins and were founded by one or more small Asian-ancestry populations (Schurr et al., 1990, Wallace et al., 1985). It is then possible that American Indian subpopulations diverged from this founding population, and distinct geographically-isolated tribes were established. In support of this, DNA analyses of Native American and Siberian groups suggest that following the expansion and dispersal of founding populations throughout the Americas, there was little gene flow after populations had split (Reich et al., 2012). It is not known whether different founding populations established the NP and SW tribes, or, in the case that the same population founded both tribes, when they subsequently diverged from one another. However, it can be postulated that genetic divergence has occurred between the tribes, as evidenced by the findings in Chapter 2 (thesis section 3, discussed below), and due to their geographic isolation.

6.3.2 Genetic Divergence

Despite likely sharing common American Indian ancestry, the NP and SW subjects in the current study exhibited distinct patterns of CYP2A6 genetic variation, indicative of tribal genetic diversity at this gene locus. Specifically, in Chapter 2 we observed that the NP population had a significantly lower frequency of CYP2A6-genotype reduced metabolizers compared to the SW, along with evidence of differences in specific allele frequencies between the tribes (CYP2A6*1B, *4, *9), which in some cases did not reach significance, likely due to the relative small numbers studied. Both tribes were in Hardy-Weinberg equilibrium for each allele. Additionally, the tribes exhibit significantly different degrees of CYP2A6 heterozygosity. Observed tribal heterozygosity was 49% in the NP, compared to 59% in the SW (P=0.02, Chi-squared test); percent heterozygosity was calculated as the number of
individuals who were heterozygous at the \(CYP2A6\) locus divided by the total number of individuals sampled (\(\times\) 100). The expected heterozygosities were 50\% and 58\% for the NP and SW tribes, respectively, and were calculated based on the approach described by Nei (1978), shown below:

\[
\text{Expected Heterozygosity} = \left[1 - \sum (\text{allele } i \text{ frequency})^2\right] \times 100
\]

Where \(\Sigma\) denotes the summation over the variant \(CYP2A6\) alleles, with \(i = *1B, *2, *4, *7, *9, *12, *17,\) and *35. This again indicates that the tribes do not deviate from Hardy-Weinberg equilibrium (similar observed and expected heterozygosities within tribes), but the different heterozygosities between the tribes further suggest genetic divergence between the NP and SW populations.

The tribes’ genetic divergence can also be assessed through estimation of their genetic distance. Using an approach outlined by Balakrishnan and Sanghvi (1968) and Latter (1973), genetic distance \((G_s^2)\) can be calculated as follows:

\[
G_s^2 = \sum \frac{(\text{population A allele } i \text{ frequency} - \text{population B allele } i \text{ frequency})^2}{\sqrt{2}(\text{population A allele } i \text{ frequency} + \text{population B allele } i \text{ frequency})}
\]

Where \(\Sigma\) denotes the summation over the variant \(CYP2A6\) alleles, with \(i = *1B, *2, *4, *7, *9, *12, *17,\) and *35. Based on this approach, and using \(CYP2A6\) allele frequencies from Chapter 2 (thesis section 3), we calculated a genetic distance of 0.09 between the NP and SW populations. Smaller genetic distances suggest a closer relatedness between the two populations and a more recent shared common ancestor, compared to populations with larger genetic distances. In order to put this into perspective among other shared ancestry populations, the genetic distance was also calculated for Alaska Natives, Chinese, and Japanese populations using previously published \(CYP2A6\) allele frequencies, for \(CYP2A6\) alleles that overlap with those investigated in the study of the NP and SW populations (Chapter 2, thesis section 3) (Schoedel et al., 2004, Binnington et al., 2012) (Table 24). The NP and SW tribes appear to be the most closely related populations, with the smallest genetic distance of all compared groups, whereas the NP tribe and the Japanese population have the greatest genetic distance. The rank orders of relatedness for the two tribes are as follows: the NP tribe is most closely related to the
SW tribe, followed by Chinese, Alaska Native, and lastly Japanese populations. The same is true for the SW tribe, which is most closely related to the NP tribe, then to Chinese, Alaska Native, and Japanese populations. These data suggest that the NP and SW tribes are more closely related to each other than to other Asian ancestry populations, and are more closely related than other Asian Ancestry populations.

Table 24 | Calculated genetic distance between NP and SW American Indian populations, Alaska Natives, and other Asian ancestry populations at the CYP2A6 locus. Genetic distance \( (G_S^2) \) values calculated based on previously described method (Latter, 1973, Balakrishnan and Sanghvi, 1968). CYP2A6 alleles included in the NP vs. SW vs. Alaska Native comparisons are: *1B, *2, *4, *7, *9, *12, *17, and *35. The CYP2A6*17 and *35 alleles were not included in any comparisons with or between Chinese or Japanese populations, as data were not available for these populations. Alaska Native, Chinese, and Japanese data compiled from (Schoedel et al., 2004, Binnington et al., 2012).

<table>
<thead>
<tr>
<th></th>
<th>Northern Plains</th>
<th>Southwest</th>
<th>Alaska Native</th>
<th>Chinese</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern Plains</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southwest</td>
<td>0.09</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alaska Native</td>
<td>0.22</td>
<td>0.38</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>0.20</td>
<td>0.23</td>
<td>0.21</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>0.65</td>
<td>0.64</td>
<td>0.33</td>
<td>0.25</td>
<td>0</td>
</tr>
</tbody>
</table>

Based on the assessment of CYP2A6 genetic variation, there is evidence for genetic divergence between the NP and SW tribes, although to a lesser degree than across different Asian ancestry populations, as expected given the longer geographic isolation among Asians (Stoneking and Delfin, 2010). Genetic diversity has only been assessed at a single gene locus, and, as the CYP2A6 gene is highly polymorphic, these findings may not extend to other regions of the genome. Consequently, these findings may not be representative of the tribes’ full diversity, relatedness, or genetic divergence. In order to fully characterize the genetic divergence of these two populations, we would need to assess multiple loci within the genome.

The array of CYP2A6 alleles that were genotyped in the NP and SW tribes has allowed us to assess similarities between different ethnic/racial populations at this gene locus; these CYP2A6
genetic variants were selected as representative alleles from each major ethnic/racial group to gain insight into whether there appeared to be a relationship with this ancestral ethnicity, and to determine whether more alleles of higher prevalence in this ethnic/racial group should have been assessed. For example, if CYP2A6*17 would have been identified in the NP or SW, this may suggest similarities with African Ancestry populations, as this allele has predominantly been identified in African populations, and we could further examine other CYP2A6 alleles commonly found in these populations, such as CYP2A6*23, *24, and *25.

The classification of the tribes’ ancestry and genetic divergence is important because American Indian/Alaska Native populations are often investigated as a single sample in studies of smoking and disease risk; for example studies conducted by the CDC (Jamal et al., 2016), as well as summaries of research reported in the U.S. Surgeon General reports (U.S. Department of Health and Human Services, 1998), refer to all American Indian tribes and Alaska Native populations collectively as American Indians/Alaska Natives. It is essential to clarify the validity of combining individual tribes for genetic association analyses or to determine if, based on unique ancestry and genetic divergence, American Indian and Alaska Native tribes, specifically the NP and SW, should be studied as distinct populations.

6.3.3 Future Research: Establishing Genetic Divergence of the NP and SW Tribal Populations

Inferring genetic ancestry is important for genetic association studies in order to reduce population stratification that could result in biased genotype-outcome associations, or to identify and characterize susceptibility variants that have distinct ancestral distributions in admixed populations. With respect to CYP2A6 genetic diversity between the tribes, the NP and SW may exhibit distinct LD and haplotype structures at this gene locus, which, if the tribes were assessed as a single sample, could result in misleading associations, or lack of an association, between CYP2A6 genotype and outcome measures, such as nicotine metabolism, smoking behaviour, cessation success, or disease risk. For example, the CYP2A6 SNPs rs56113850 (intron 4) and rs57837628 (~2kb 5’), which were described in Chapter 4 of this thesis (section 5), differ in their level of LD across different ethnic/racial groups, according to 1000 Genome data (Ward and Kellis, 2012). The two SNPs are in high LD in Asian ancestry populations (r^2=0.81, D’=0.91), moderate LD in European ancestry populations (r^2=0.69, D’=0.92), and low LD in African ancestry populations (r^2=0.23, D’=0.83); LD structure has
not yet been determined for the NP and SW tribes. For demonstration purposes, we can designate rs56113850 the causal SNP and rs57837628 a tag SNP (it is not yet known which is causal, or if both are functionally significant). If one were to study the association of these SNPs with a particular outcome in Asian ancestry populations, it would not matter which SNPs were genotyped, as both SNPs occur together (due high LD) and will each be similarly associated with the outcome. However, if one only genotypes the tag SNP rs57837628 in African ancestry populations, who exhibit low LD for rs56113850 and rs57837628, you would not capture the causal effect of the rs56113850 SNP, and may not see an association between genotype and outcome. In African populations, rs56113850 and rs57837628 should be investigated independently so that the functional significance of both SNPs can be determined. Further, in African ancestry populations, the rs57837628 tag SNP may instead be in high LD with a different causal SNP, highlighting the difficulty of using the same tag SNP for a phenotype across genetically diverse populations.

These principles could also extend more broadly to other genes and clinical outcomes. For example, Alaska Natives exhibit a distinct LD pattern at the CHRNA5-A3-B4 gene cluster, compared to European Americans (Zhu et al., 2013b). A 30 kb region of high LD was identified in this Alaska Native population, with some variant alleles exhibiting similar patterns of LD as seen in European American populations (rs16969968 and rs1051730) and others showing very different levels of LD (rs7163730 and rs578776). The SNP rs578776 is similarly associated with higher nicotine intake in Alaska Native and European ancestry smokers, but not in Asian or African American populations, again highlighting the importance of identifying genetically distinct populations within the study sample (Chen et al., 2012b).

One approach that is often used to differentiate populations based on ancestry, which could be extended to the NP and SW American Indian populations, is the use of Ancestry Informative Markers (AIMs). AIMs are a subset of unlinked SNPs that are present at substantially different frequencies across populations. As AIMs differentiate individuals within a particular study population who have distinct ancestral origins, they can be used to adjust for population stratification that can result from different proportions of genetically similar populations in cases and controls (Royal et al., 2010). Further research is necessary to establish appropriate AIMs for different populations of American Indians. Previous efforts to identify and establish AIMs across different ethnic/racial populations have included American Indians collectively as
a single broad group (Kosoy et al., 2009). This could be problematic if there are distinct differences between tribes, as the data presented in this thesis data has begun to suggest for the NP and SW. Of note, in an effort to establish AIMs across European American, West African, Amerindian, and East Asian populations, Kosoy et al. (2009) showed that of the four continental population groups investigated, Amerindians had the lowest within-group relatedness, according to estimated co-ancestry coefficients (Fst). The Fst, another measure of genetic distance, was more than 2-fold higher within Amerindians than within East Asians, 10-fold higher than within West Africans, and 1000-fold higher than within European Americans. This suggests that American Indian tribal populations are the most genetically diverse of the major population ethnic/racial groups assessed, and that AIMs may need to be adjusted such that they differentiate American Indian tribes, and not just American Indians as a broad group. As this study has focused solely on CYP2A6 genetic variation and divergence between two distinct American Indian tribes, future studies could expand this to multiple loci, focusing on establishing tribe-specific AIMs, and assessing their utility in accounting for differences in population structure in future genetic association studies.

### 6.4 Sources of Elevated Nicotine Metabolism in the Northern Plains

In Chapter 2 (thesis section 3) we identified a much higher NMR, indicative of a faster rate of nicotine metabolism, among NP compared to SW smokers. Here we will discuss evidence for the contribution of genetic factors to the high NMR in the NP, while proposing future investigations into sources of variation in the nicotine metabolism rate in the NP and SW tribes.

#### 6.4.1 Genetic Sources

As discussed in the previous section, the NP and SW tribes are genetically distinct from one another at the CYP2A6 locus, which may help to explain the significantly higher NMR, a measure of CYP2A6 activity, in the NP than in the SW. However, the NP’s lower frequency of individuals who possess reduce-of-function CYP2A6 genetic variants, compared to the SW, did not entirely account for the tribe’s higher rate of nicotine metabolism. This is in contrast to findings from African American and White smokers, where the higher frequency of CYP2A6 reduced activity genetic variants among African American populations, compared to Whites,
appears to account for their overall lower NMR compared to White populations; when excluding smokers who possess known \textit{CYP2A6} reduced activity genetic variants, African American and White smokers had similar NMRs (Fig. 17 from Chapter 2, section 3.4). The ability to account for some of the ethnic/racial differences in the rate of nicotine metabolism among Whites and African Americans, through their differences in \textit{CYP2A6} genetic variation, suggests that the NP population may possess novel or undetected genetic variation at the \textit{CYP2A6} locus that contributes to their high NMR. Further, the \textit{CYP2A6} gene is highly polymorphic (http://www.cypalleles.ki.se/cyp2a6.htm), supporting the hypothesis that there are additional functionally significant genetic variants occurring within this gene locus in all ethnic/racial groups. Based on the genetic diversity between different ethnic/racial groups (discussed in the previous section) and their differences in \textit{CYP2A6} allele frequencies (Table 6 in Chapter 2, section 3.4), it is likely that the NP population exhibits distinct frequencies of novel \textit{CYP2A6} genetic variants. Specifically, we hypothesize that there are higher frequencies of \textit{CYP2A6} genetic variants that are associated with increased \textit{CYP2A6} expression or enzyme activity among the NP population compared to the SW and other ethnic/racial populations.

Considering that all \textit{CYP2A6} exon variation identified to date is associated with a decrease, or no change, in expression or metabolic function (Tanner et al., 2015a), and that \textit{CYP2A6} mRNA expression and enzyme activity are moderately positively correlated phenotypes (demonstrated in the human liver bank, Chapter 3, section 4.4), we propose that non-coding genetic variation that causes increased \textit{CYP2A6} expression is responsible for the higher NMR in the NP. This could manifest in several ways. One possibility is that there are several different low frequency genetic variants present in the NP (i.e. in just a few individuals) that pull up the mean NMR. This would produce a wide spread of variation in the NMR within each genotype group, as opposed to causing an overall shift for the majority of individuals. In support of this, we see a slight increase in the within-group NMR variation among \textit{CYP2A6}*1/*1 genotype individuals in the NP compared to \textit{CYP2A6}*1/*1 genotype individuals in the SW (coefficient of variation, CV 51% in NP and 45% in SW; calculated from data in Fig. 16. in Chapter 2, section 3.4); however, the opposite trend is observed among \textit{CYP2A6}*1/*9 genotype groups in the two tribes (CV 51% in NP and 66% in SW). These inconsistent trends suggest that rare genetic variants are not increasing the within-group NMR variation in \textit{CYP2A6}*1/*9 genotype individuals in the NP, but may be contributing to the higher NMR among \textit{CYP2A6}*1/*1 genotype individuals in the NP compared to \textit{CYP2A6}*1/*1
genotype individuals in the SW. Upon further examination of the NMR in the NP
*CYP2A6*/*1/*1 group, we see that more than 69% of *CYP2A6*/*1/*1 genotype NP smokers have
an NMR above the average for *CYP2A6*/*1/*1 genotype smokers from SW, White, and African
American populations (average NMRs of 0.48, 0.44, and 0.43, respectively) (Fig. 17 in
Chapter 2, section 3.4). This indicates a more widespread general increase in the NMR in the
NP relative to other ethnic/racial groups, and also that if genetic variation is contributing to this
high NMR, it is likely present at a high frequency in the NP population, and not driven by a
few individuals.

Based on these observations, another possibility is that a single common genetic variant,
especially a novel version of the wild-type allele, is present in the NP population, and is
mostly absent from the SW. If accurate, we might expect to see a somewhat global increase in
nicotine metabolism in all NP subjects of all *CYP2A6* genotype groups compared to SW
subjects of the same *CYP2A6* genotype, with the exception of individuals with full *CYP2A6* deletions (*CYP2A6*/*4/*4) or two copies of a loss-of-function allele (i.e. *CYP2A6*I/*10/*10). For
example, the *CYP2A6*I/*9 individuals in the NP would exhibit a higher average NMR
compared to the *CYP2A6*I/*9 individuals in the SW. Instead, the findings from Chapter 2
illustrate that the increased NMR in the NP is not consistent across all *CYP2A6* genotypes, but
rather appears to be primarily mediated by the *CYP2A6*I/*1/*1 individuals (including
*CYP2A6*I/*1A and *IB). The *CYP2A6*I/*1/*1 individuals in the NP have higher average NMR
than the *CYP2A6*I/*1/*1 individuals in the SW (Fig. 16 in Chapter 2, section 3.4), whereas,
subjects who possess the *CYP2A6*I/*9 genotype exhibit on average the same NMR across the
tribes, which is also the case when comparing individuals with the *CYP2A6*I/*9/*9 genotype
across tribes (Table 25).
Table 25 | Comparison of the mean NMR for different CYP2A6 genotype groups across the NP and SW tribes.

<table>
<thead>
<tr>
<th>Northern Plains</th>
<th>Mean NMR</th>
<th>Southwest</th>
</tr>
</thead>
<tbody>
<tr>
<td>All CYP2A6 Genotypes</td>
<td>&gt;</td>
<td>All CYP2A6 Genotypes</td>
</tr>
<tr>
<td>CYP2A6*1/*1</td>
<td>&gt;</td>
<td>CYP2A6*1/*1</td>
</tr>
<tr>
<td>CYP2A6*1A/*1A</td>
<td>&gt;</td>
<td>CYP2A6*1A/*1A</td>
</tr>
<tr>
<td>CYP2A6*1A/*1B</td>
<td>&gt;</td>
<td>CYP2A6*1A/*1B</td>
</tr>
<tr>
<td>CYP2A6*1B/*1B</td>
<td>&gt;</td>
<td>CYP2A6*1B/*1B</td>
</tr>
<tr>
<td>CYP2A6 RMs a</td>
<td>=</td>
<td>CYP2A6 RMs a</td>
</tr>
<tr>
<td>CYP2A6*1/*9</td>
<td>=</td>
<td>CYP2A6*1/*9</td>
</tr>
<tr>
<td>CYP2A6*9/*9</td>
<td>=</td>
<td>CYP2A6*9/*9</td>
</tr>
</tbody>
</table>

a. CYP2A6 Reduced Metabolizers (RMs) are individuals who possess one or more reduce-of-function genetic variant

There are several possible explanations for the higher NMR across different CYP2A6 genotype groups in the NP relative to the SW. First, a genetic variant that increases the expression of the CYP2A6 gene may be associated with higher NMR when it occurs alone or in combination with other increase-of-function genetic variants, such as CYP2A6*1B; however, when this variant occurs with a reduce-of-function variant, such as CYP2A6*9, the effect of increased CYP2A6 gene expression may be subdued or dominated by the impact of the reduce-of-function variant. Evidence from studies of the highly variable CYP2D6 enzyme support this possibility. Individuals with more than two copies of a normal functioning CYP2D6 gene (i.e. a duplication) exhibit higher CYP2D6 activity compared to CYP2D6*1/*1 individuals (i.e. no gene duplication); whereas, individuals with more than two copies of the reduced activity CYP2D6*10 variant (i.e. CYP2D6*10/*10X2) exhibit similar CYP2D6 activity as those with just two copies of the *10 allele (i.e. CYP2D6*10/*10) (Ishiguro et al., 2004). Therefore, the impact of a genetic variant that increases gene expression may not be sufficient to overcome the impact of a decreased activity variant. One caveat to translating this to the CYP2A6 findings is that CYP2D6*10 is a nonsynonymous variant that impacts the function of the CYP2D6 enzyme (Yokota et al., 1993, Sakuyama et al., 2008), whereas the CYP2A6*9 allele is a non-coding variant present in the TATA box of the CYP2A6 promoter that decreases CYP2A6 transcription (Pitarque et al., 2001). There may be a different effect of an increased expression variant (CYP2D6 duplication) occurring with a decreased activity variant.
(CYP2D6*10) than for an increased expression variant (CYP2A6 duplication) occurring with a decreased expression variant (CYP2A6*9).

Alternatively, a genetic variant that causes an increase in CYP2A6 gene expression may not be in LD with any of the tested CYP2A6 genetic variants, and therefore may not occur together with these variants in the NP population. As a result, this novel genetic variant would only be present and increase CYP2A6 expression, and thus the NMR, among NP subjects who have the CYP2A6*1/*1 genotype and do not possess any established CYP2A6 genetic variants. For example, among Alaska Natives, the reduce-of-function CYP2A6*2 and CYP2A6*12 alleles were not in LD with the increase-of-function CYP2A6*1B allele (r²=0, D’=0 for both) (Binnington et al., 2012). A similar phenomenon may occur in the NP population, which may explain the observed tribal differences in NMR, such that the effect size (difference in NMR) was larger when comparing CYP2A6*1/*1 genotype NP smokers to CYP2A6*1/*1 genotype SW smokers, than when comparing the overall populations of NP and SW smokers (all CYP2A6 genotypes included). However, if this were the case, we would still expect to see an increase in the NMR in the CYP2A6*1/*9 group in the NP compared to the same genotype group in the SW, as only one of the two alleles possesses the *9 variant, therefore the novel genetic variant could occur on the *1 allele. As illustrated in Table 25, the CYP2A6*1/*9 groups in the NP and the SW tribes exhibit similar NMRs.

As genetic variation influencing CYP2A6 gene expression is a plausible explanation for the observed high NMR in the NP, the following section will discuss specific examples of genetic variation both within the CYP2A6 gene (cis-acting) and in other genes (trans-acting) that may influence CYP2A6 expression, and how they may relate to the findings in the NP and SW tribal populations (different mechanisms are visually depicted in Fig. 32).
Figure 32 | Mechanisms of increasing CYP2A6 mRNA levels via genetic variation. (A) Mechanism of CYP2A6 gene duplications. (B) Mechanisms of transcriptional and post-transcriptional regulation of CYP2A6 mRNA levels. In part B, each mechanism is labeled by number, from 1-7. (1) A 5’ CYP2A6 genetic variant results in increased affinity binding for an activating transcription factor, promoting transcription. (2) A 5’ CYP2A6 genetic variant results in decreased binding affinity for an inhibitory transcription factor thus reducing the inhibition of gene transcription. (3) A distal CYP2A6 genetic variant indirectly stimulates gene transcription by promoting an open chromatin conformation, allowing transcriptional/regulatory machinery access to bind to DNA. (4) The AKR1D1 enzyme is
involved in the 5β reduction of bile acid intermediates, resulting in the synthesis of bile acids, which act as endogenous ligands of nuclear hormone receptors, such as PXR and CAR, which activate CYP transcription. (5) A 3’-UTR CYP2A6 genetic variant results in increased affinity binding of proteins that stabilize mRNA, preventing degradation. (6) A 3’-UTR CYP2A6 genetic occurs within a miRNA response element (MRE), resulting in less miRNA binding and less mRNA degradation. Similarly, lower expression of miRNAs could result in less mRNA degradation. (7) Less DNA methylation, a possible form of epigenetic modification, results in greater access of transcriptional/regulatory machinery to bind to DNA. (Rao et al., 2000, Fukami et al., 2007, Pai et al., 2015, Shi et al., 2016, Li et al., 2013, Xu et al., 2005, Rizner and Penning, 2014, Chaudhry et al., 2013, Wang et al., 2006, Nakano et al., 2015, Martinowich et al., 2003).

6.4.1.1 Gene Duplications

As described in section 1.5.1.1 of the General Introduction, two versions of a CYP2A6 gene duplication (*1X2A and *1X2B) have been observed to date, each with different crossover points with the adjacent highly homologous CYP2A7 gene (Rao et al., 2000, Fukami et al., 2007). A higher frequency of CYP2A6 gene duplications in the NP relative to the SW and other ethnic/racial groups, which would presumably result in greater levels of the CYP2A6 transcript and higher CYP2A6 enzyme activity relative to the wild-type CYP2A6 (2 copies of the gene), may explain the NP tribe’s higher NMR. However, a subset of NP and SW subjects from our study (Chapters 1 and 2) who had high NMR were genotyped for both currently identified CYP2A6 duplications, and none were positive for either variant. We may not have been able to detect CYP2A6 duplications in the tribes because the genotyping assays that were used targeted the specific crossover points of each duplication (Wassenaar et al., 2016). It is possible that an additional version of the CYP2A6 duplication, with a different crossover breakpoint, could still be present in the NP population. Therefore, to further investigate the presence of CYP2A6 duplications in the NP population, we could use a quantitative genotyping approach that allows for the comparison of the relative amounts of PCR products, such as the quantitative multiplex PCR method that has previously been used for the detection of CYP2D6 copy number variation (Gaedigk et al., 2012). On the other hand, if we consider the very low frequencies, or lack, of the established CYP2A6 duplications in other ethnic groups, including Asian populations (1.5% allele frequency) (Martis et al., 2013) who likely share common ancestry with these American Indian tribes, it seems unlikely that the NP population would possess a gene duplication at a sufficient frequency to account for their widespread higher NMR.
However, looking again to CYP2D6, the frequency of CYP2D6 duplications varies widely across different ethnic/racial populations (allele frequencies 0.5-29%) (Johansson and Ingelman-Sundberg, 2008), arguing for the possibility of novel high frequency CYP2A6 gene duplications in the NP population.

Additionally, there are other forms of genetic variation that may be found at relatively high frequencies, which may contribute to the NP higher rate of nicotine metabolism (described in the following sections).

6.4.1.2 Transcriptional Regulation

Another plausible source of the higher NMR in the NP relative to the SW and other ethnic/racial groups may be through increased CYP2A6 mRNA transcription, which can occur via a variety of mechanisms (Fig. 32). There are several examples of high frequency genetic variants located 5’ of the CYP2A6 gene that may be associated with higher CYP2A6 transcription, and could be expressed at the highest frequency in the NP population, compared to other populations. For example, four novel SNPs, identified in Chapter 4 of this thesis (section 5), are located approximately 1-2 kb upstream of CYP2A6 and are associated with higher in vitro CYP2A6 enzyme activity (rs57837628, rs7260629, rs7259706, and rs150298687; Fig. 25 in Chapter 4, section 5.4). As these non-coding SNPs were present at high frequencies in the White samples that were investigated (human liver bank and PNAT clinical trial), it is likely that the SNPs are also present in the NP population, albeit at different allele frequencies.

It could be hypothesized that, based on the high NMR of the NP population, these individuals have higher frequencies of these novel SNPs, relative to the SW population and other ethnic/racial groups. This hypothesis is further supported by the higher frequencies of two of the four increase-of-function SNPs (rs57837628 and rs150298687; allele frequencies obtained from the 1000 Genomes dataset) in faster nicotine metabolizing (higher NMR) European ancestry populations compared to Asian ancestry populations, who have lower NMRs (overall NMR and CYP2A6*1/*1 genotypes only) (Baurley et al., 2016, Park et al., 2016). The two SNPs that are the exception to this are rs7259706 and rs7260629, which are present at nearly the same frequency in European and Asian ancestry populations, according to 1000 Genomes; nonetheless, this does not rule out the possibility that rs7259706 and rs7260629 are present at
even higher frequencies in the NP than other ethnic/racial groups. For example, the Asian ancestry population that founded the NP population could have possessed these SNPs at an even higher frequency than other Asians, resulting in uncharacteristically high frequencies in the NP.

Using rs57837628 and rs150298687 allele frequencies reported from the 1000 Genomes Project, and the average NMR (overall NMR, i.e. all CYP2A6 genotypes included) reported among White and Asian smokers (Ward and Kellis, 2012, Baurley et al., 2016, Lerman et al., 2006), we have crudely predicted the frequencies of these SNPs in the NP and SW tribes (Fig. 33). We have chosen not to include African Americans in this analysis, as the difference in White and African American populations’ NMRs can be accounted for by the higher frequency of CYP2A6 reduced activity genetic variants among African American populations, compared to Whites (as described previously, i.e. Fig. 17 in Chapter 2, section 3.4). The analyses of rs57837628 and rs150298687 among White and Asian populations suggest that these SNPs would be expressed at highest frequencies in the fast nicotine-metabolizing NP population, compared to the other ethnic/racial groups. Obviously, many assumptions are being made in this analysis, with a large assumption being that these two SNPs have a consistent impact on NMR in different ethnic/racial groups, and also that the SNPs are principally responsible for the ethnic/racial differences in NMR. This is likely not the case, as other CYP2A6 genetic variants associated with decreased NMR are also present at different frequencies across the different ethnic/racial groups. However, this analysis serves to demonstrate a scenario in which these CYP2A6 genetic variants are predictive of the rate of nicotine metabolism, and highlights the potential of future research into the presence and functional significance of these SNPs in the NP tribe. An additional caveat is that the NMR data and allele frequency data were each derived from two separate datasets, for both European and Asian ancestry populations; these data were plotted under the assumption that allele frequencies would show little variation across two populations of the same ethnic/racial background.
Figure 33 | Observed and predicted associations between the nicotine metabolite ratio (NMR) and the allele frequencies of two novel 5’ CYP2A6 SNPs across different ethnic/racial groups. (A) The association between allele frequencies and NMR in European and Asian populations. (B) A representative example of the approach used to predict NP and SW rs150298687 allele frequency using known population NMRs. (C) The predicted association between allele frequencies and NMR across all four ethnic/racial groups. The allele frequencies for European and Asian ancestry populations are as reported in the 1000 Genomes dataset. NP and SW allele frequencies were extrapolated based on the observed relationship between allele frequency and NMR in the European and Asian groups (part B). Data were compiled from (Ward and Kellis, 2012, Baurley et al., 2016, Lerman et al., 2006).
The mechanism of the association of these SNPs with increased CYP2A6 activity has not yet been determined, but may include one of the following explanations. Genetic variation occurring 5’ of the CYP2A6 gene (i.e. in the promoter region or further upstream) can alter transcription factor binding motifs, and these SNPs in particular may be located in distal enhancer regions involved in long-range activation of transcription (Serfling et al., 1985). Genetic variation 5’ of the gene can have a direct stimulatory effect on gene transcription by increasing the affinity of an activating transcription factor for binding to the DNA sequence, or this can indirectly stimulate transcription by promoting an open chromatin state (i.e. DNA is accessible for binding) (Pai et al., 2015, Shi et al., 2016). Alternatively, genetic variation may interrupt the binding motif for an inhibitory transcription factor, which may reduce the suppression of gene transcription (Li et al., 2013).

Additionally, genetic variation in trans-regulatory elements may influence CYP2A6 transcription. For example, variation in elements functioning earlier in the regulatory pathway could impact CYP2A6 transcription, possibly through mediation of transcription factor synthesis or degradation. AKR1D1, described in the General Introduction of this thesis (section 1.5.2.2), plays a role in the reduction of steroid hormones that are involved in the transcriptional regulation of CYPs (Rizner and Penning, 2014). In Chapter 3 (thesis section 4) we demonstrated that AKR1D1 mRNA and CYP2A6 mRNA are moderately positively correlated, suggesting that genetic variation in the AKR1D1 gene that results in greater expression of the AKR1D1 enzyme could be associated with increased CYP2A6 transcription. Therefore, we could reasonably hypothesize that AKR1D1 mRNA expression is higher among NP subjects, relative to SW subjects and individuals of different ethnic/racial backgrounds, and that this may be associated with one or more genetic variants in the AKR1D1 gene.

6.4.1.3 Post-Transcriptional Regulation

CYP2A6 mRNA levels can also be impacted post-transcriptionally by genetic variation 3’ of CYP2A6. Similar to the CYP2A6*1B allele, variation in the 3’-UTR can result in the stabilization of mRNA, potentially through increased interaction or binding of the 3’-UTR with one or more proteins (Wang et al., 2006). This increase in CYP2A6 mRNA appears to translate to increased enzyme activity, as Mwenifumbo et al. (2008b) showed, in a human pharmacokinetic study, that individuals with the CYP2A6*1B/*1B genotype had higher NMR compared to those with the CYP2A6*1A/*1A genotype. As discussed in Chapter 2 (thesis
section 3), the NP participants have a higher frequency of the $CYP2A6*1B$ allele compared to the SW, however this did not account for their faster rate of nicotine metabolism (Table 6, Chapter 2, section 3.4). Nevertheless, it is possible that another functionally significant 3’-UTR genetic variant may be present in the NP population at a higher frequency than in the SW and other ethnic/racial groups. For example, the novel 3’-UTR SNP rs8192733, described in Chapter 4, was associated with increased hepatic $CYP2A6$ mRNA expression according to the GTEx Portal tool, which provides expression quantitative trait loci (eQTL) associations for SNPs in several different body tissues, such as the liver (Consortium, 2013). There was not a significant association between rs8192733 and $CYP2A6$ mRNA expression in the White human liver bank (Table 19, Chapter 4, section 5.4), however the direction of effect was consistent with an increase in $CYP2A6$ mRNA levels, and donors with one or two copies of this 3’-UTR SNP had significantly higher $CYP2A6$ protein levels and in vitro enzyme activity, compared to donors not positive for this SNP. It is possible that we were underpowered to detect a significant difference in $CYP2A6$ mRNA, as expression data was available for a smaller sample of donors (N=265) compared to the protein (N=320) and activity (N=327) data. We hypothesize that the rs8192733 3’-UTR SNP is associated with increased $CYP2A6$ mRNA stabilization and greater enzyme expression, and that this SNP is present at a significantly higher frequency in the NP population compared to other ethnic/racial populations.

Another mechanism by which $CYP2A6$ mRNA expression could be increased in the NP population is through altered miRNA binding. We would expect that, relative to other ethnic/racial groups, members of the NP tribe exhibit less miRNA binding 3’ of $CYP2A6$, therefore reducing miRNA-mediated degradation of $CYP2A6$ mRNA (Nakano et al., 2015). The novel 3’-UTR SNP rs8192733 was not predicted to be located in an miRNA binding site, according to the RegRNA 2.0 tool, which can be used to identify functional RNA motifs (Chang et al., 2013); however, additional novel 3’ genetic variants may be common in the NP population. Furthermore, expression of miRNAs that interact with the $CYP2A6$ 3’-UTR may differ between ethnic/racial groups. For example, NP subjects may have lower expression of miR-126*, an miRNA which has been associated with decreased $CYP2A6$ mRNA expression in vitro (Nakano et al., 2015).

Gene expression can also be altered by epigenetic modifications, which can be heritable and/or result from environmental exposures, and can accumulate with age (Christensen et al., 2009,
Gibney and Nolan, 2010, Fraga et al., 2005). It is possible that age-related differences in CYP2A6 gene expression may occur due to increases in DNA methylation with age, which can decrease the accessibility of regulating factors to DNA (Keshet et al., 1985, Razin and Cedar, 1991). However, age was not a significant predictor of the NMR in either the NP or SW tribe in our study (Table 7 in Chapter 2, section 3.4), which may indicate that epigenetic drift is not a major source of CYP2A6 gene expression or NMR differences between the tribes. That said, epigenetic modifications that are not associated with age could differ between the tribes, such that the NP tribe has greater CYP2A6 gene expression, possibly resulting from less DNA methylation in the 5’ of 3’ regions of CYP2A6 (Martinowich et al., 2003).

6.4.2 Future Research: Determining Constant and Transient Sources of Variable Nicotine Metabolism in the Northern Plains and Southwest

As discussed in Chapter 2 (thesis section 3.5) tribal differences in exposure to environmental inducers of CYP2A6 may also be important determinants of nicotine metabolism rate. For example, dietary components (broccoli) and medications (oral contraceptives, hormone replacement therapy, phenobarbital, rifampin, dexamethasone) induce CYP2A6. Therefore, a future study should investigate multiple genetic and non-genetic sources of variation in the NMR in the NP and SW tribes. First, a pharmacokinetic study could be conducted to validate that the NMR accurately reflects the NP and SW smokers’ rate of nicotine clearance. Thus far, we have operated under the assumption that CYP2A6 is the main enzyme involved in the NMR and nicotine clearance, based on previous studies in other ethnic/racial populations. However, it would be important to verify that the roles of other minor pathways of nicotine metabolism and clearance (Fig. 5, in section 1.3.2.2 of the General Introduction) are consistent in the NP and SW tribes as in other ethnic/racial groups. Therefore, participants would be administered infusions of deuterium-labeled nicotine and cotinine, as performed previously by Benowitz et al. (2006), and the total and non-renal clearance of both compounds would be quantified from plasma and urinary measurements of nicotine and cotinine. We hypothesize that plasma NMR would be strongly positively correlated with total and non-renal clearance of both compounds due to the major role of CYP2A6 in all three measures.

Additionally, blood samples taken from study participants would be used for DNA sequencing. DNA from each participant would be sequenced at the CYP2A6 locus using the CYP2A6-specific next-generation sequencing approach described in Chapter 4 of this thesis (section
5.3). In addition to investigating novel genetic variants that are identified through sequencing, we would assess the presence of the seven novel SNPs, described in Chapter 4, among all NP and SW participants. The LD of the seven SNPs would be assessed independently in each tribe to determine if linkage and haplotype structure is similar between tribes, and this would also be compared to what was observed for White populations. If we observe similar patterns of LD as White populations, the functional significance of the simplified five-SNP diplotype would be assessed in each tribe (i.e. association with the NMR). Additionally, the functional significance of all other novel CYP2A6 genetic variants would be investigated using bioinformatics tools.

As these populations are likely genetically distinct from other ethnic/racial groups, they may possess genetic variants that are not present or frequent in previously studied populations of different ancestral origins (i.e. Whites, African Americans, Asians). The NP and SW participants’ DNA would also be genotyped for AKR1D1 and POR genetic variants that have previously been associated with variation in the expression or function of other CYPs (e.g. AKR1D1 SNP rs1872930 or POR SNP rs1057868) (Chaudhry et al., 2013, Chenoweth et al., 2014b), and may therefore impact CYP2A6 activity.

Participants would be surveyed, via interview and questionnaire, on specific components of their diet, medications that they are currently taking, how many cigarettes they smoke per day, as well as on demographic factors including their BMI, gender, and age – factors for which previous studies have provided evidence of an association with CYP2A6 expression or activity. We would independently assess the association of each factor with the NMR. Several studies have identified a modest negative correlation between the NMR and BMI among smokers, and a modest association of gender with NMR, suggesting that these factors may contribute in part to the higher NMR in the NP. However, the average BMI across smokers in the two tribes did not differ (NP: 32, SW: 31), and was similarly higher in female compared to male smokers in both tribal populations (NP: males 31, females 33; SW: males 30, females 32). Further, our regression models of predictors of the NMR indicated that BMI accounted for a similar proportion of the variation in both tribes (2%). This suggests that BMI does not account for the differences in NMR between the NP and SW tribes.

Finally, all variables (genetic and non-genetic) that were significantly associated with CYP2A6 activity (i.e. the NMR) in independent analyses would be included in regression models of
NMR. This allows for the quantification of the independent and shared contribution of each variable to the NMR.

Findings from this study would help address the question of what factors are differentially influencing the rate of nicotine metabolism in the NP and SW tribal populations, and whether there is a stronger contribution of constant (genetic) or transient (environmental) factors. Additionally, this investigation may provide insight into genetic variation that regulates CYP2A6 enzyme activity through the identification and characterization of novel CYP2A6 genetic variants.

6.5 Using Functional CYP2A6 Diplotypes for Prediction of the NMR

In Chapter 4 of this thesis (section 5) we identified seven novel SNPs at the CYP2A6 locus that are associated with increased CYP2A6 activity: rs57837628, rs7260629, rs7259706, rs150298687, rs56113850, rs28399453, and rs8192733. Considering the high frequencies and functional significance of these SNPs in White populations (human liver bank and PNAT trial), they may play an important role in smokers’ response to cessation pharmacotherapies and ability to quit smoking. However, six of the seven SNPs are in moderate-to-high LD in White populations, therefore we need to first determine the mechanism by which each SNP increases CYP2A6 activity. This would allow us to differentiate between SNPs that are functionally significant, and SNPs that just tag another functional variant. Determining the independent functional significance of each of the novel SNPs will be important when developing CYP2A6 genotype algorithms for the prediction of the NMR, a phenotype that can be used prospectively to inform choice of cessation pharmacotherapy (Lerman et al., 2015). We cannot assume that each of the novel SNPs are functionally significant or causal to changes in the NMR, as this may overestimate the CYP2A6 genotype effect on the NMR and result in the development of an inaccurate phenotype prediction algorithm. Further, establishing the impact of CYP2A6 genetic variants on the NMR will be necessary in order to progress toward including these SNPs in treatment guidelines, for example when preparing Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for CYP2A6 genotype and cessation therapy. Additionally, it will be important to establish the generalizability of CYP2A6 genotype algorithms, which may include one or more of these novel SNPs, for predicting the NMR across different ethnic/racial populations.
Here we will discuss an experimental approach for elucidating mechanisms by which each SNP is associated with higher CYP2A6 activity, and the utility of these novel SNPs and established CYP2A6 genetic variants for predicting the NMR in the NP, SW, and other ethnic/racial populations.

6.5.1 Future Research: Which of the Novel SNPs are Functional?

All seven novel SNPs, when assessed individually, were associated with higher CYP2A6 mRNA expression (non-significantly), CYP2A6 protein levels, and enzyme activity (Table 19 in Chapter 4, section 5.4). This, along with the location of these SNPs in non-coding regions of CYP2A6, suggests that they are impacting CYP2A6 transcription and thus expression and activity. In order to investigate the effect of each SNP on CYP2A6 expression, we propose the use of luciferase reporter assays. This technique can be used to investigate the impact of genetic variants located in multiple regions of the gene on gene expression, including 5’, intron, and 3’-UTR genetic variants.

Using this method, the CYP2A6 promoter DNA sequence would be cloned upstream of the firefly luciferase gene in an expression vector. Additionally in this expression vector, the region of the CYP2A6 gene possessing the SNP of interest would be cloned upstream of the CYP2A6 promoter sequence (for 5’ and intron SNPs), or downstream of the firefly luciferase gene (for 3’-UTR SNP), as described previously (Wang et al., 2011, Wang et al., 2006, Pitarque et al., 2001). Several versions of the reporter construct would be derived, with some versions containing just one of the seven novel SNPs, and other versions containing combinations of the SNPs. This would allow for the assessment of independent and combined effects of the novel SNPs on CYP2A6 transcription. Reporter plasmids and a transfection efficiency control (Renilla luciferase construct) would be transfected into cells. Approximately 48 hours post-transfection, luciferase activity (luminescence) would be measured using a commercially available luciferase activity kit; this protocol typically involves the addition of luciferin and necessary cofactors, and the subsequent quantification of luminescence produced by the activity of the luciferase enzyme, as this serves as an indicator of the level of expression of the firefly luciferase gene. Luminescence produced from each version of the reporter construct, which should be correlated with the activity of the CYP2A6 regulatory elements added to the construct, would be compared to that of the control construct, which would contain CYP2A6*1 DNA sequence only (without any of the novel SNPs).
In the White PNAT trial population, we observed that two of the SNP pairs were in high LD (rs57837628 and rs8192733, rs7260629 and rs7259706; Fig. 27 in Chapter 4, section 5.4); therefore, it is possible that some, but not all, of the seven SNPs will independently associate with greater luciferase activity, relative to the control construct. For SNPs in high LD, where just one of the SNPs is causal (i.e. impacting CYP2A6 expression), we may only observe a difference in luciferase activity for one of the two SNPs. Alternatively, both SNPs may be functionally significant, or the SNPs may be benign on their own but occur in LD with other causal SNPs that are not one of the novel seven SNPs. Although it is possible for all seven novel SNPs to independently influence CYP2A6 expression, evidence from findings in Chapter 4 may suggest otherwise. We showed that the simplified five-SNP diplotype, which excluded one SNP from each of the high LD SNP pairs, in combination with the established CYP2A6 genetic variants (CYP2A6*2, *4, *9, and *12), accounted for a similar proportion of variation in in vitro CYP2A6 enzyme activity (6.0%) as the full seven-SNP diplotype (6.6%). This suggests that the high LD SNPs provide little, if any, additional impact on CYP2A6 activity. The 0.6% difference in CYP2A6 activity accounted for could be attributed to the two SNPs in the high LD SNP pairs, or this may be a product of variation in other factors (genetic or non-genetic) that were not assessed in these models. Investigation of the impact of the full seven-SNP and five-SNP diplotypes in other ethnic/racial populations may help to further characterize the functional significance of these SNPs and diplotypes.

SNPs that are shown to influence CYP2A6 expression, according to the in vitro investigations described above, should be genotyped in additional populations, and assessed for their impact on the NMR. For example, as suggested in the previous section, we propose that the frequency of these SNPs and their association with the NMR be assessed in the NP and SW tribal populations. This would provide insight into whether or not the findings from Chapter 4, which were derived from studies of White populations, can extend to other ethnic/racial groups.

In the case that some or all of the SNPs are not associated with any differences in luciferase activity compared to control, this does not mean that the variants have no effect on CYP2A6. Rather, the novel SNPs may be in LD with other causal SNPs that are not located within the CYP2A6 DNA sequence that was cloned into the expression vector. As we continue to learn more about variation at the CYP2A6 gene locus, for example through next-generation sequencing and GWAS studies, we could identify SNPs that are causal for variation in
CYP2A6 expression and function, which may be in LD with one or more of the seven novel SNPs described here.

6.5.2 Estimation and Generalizability of Genetic Risk and Phenotype Predictions from CYP2A6 Diplotypes

Based on knowledge of which CYP2A6 alleles are functional, including the previously established genetic variants (Table 2 in section 1.5.1.1 of the General Introduction) and the functional novel variants described in Chapter 4 (thesis section 5), models using CYP2A6 genotype to predict the NMR can be developed. For example, Loukola et al. (2015) performed a GWAS for the NMR and developed weighted genetic risk scores (wGRS) based on the predicted impact of their top genome-wide significant hits on the NMR. The wGRS was derived from the sum of the number of alleles for each SNP (0, 1, or 2 copies of each allele), weighted by their effect size determined by the GWAS meta-analysis. This approach, incorporating the established CYP2A6 genetic variants and the functional novel SNPs described in Chapter 4, should be assessed for its ability to predict the NMR and smoking phenotypes across multiple ethnic/racial populations.

Based on previous findings from studies of CYP2D6 and from the data presented in this thesis, ethnicity is an important consideration when developing genetic risk scores to predict the NMR. In the assessment of activity scores for predicting CYP2D6 activity, Gaedigk et al. (2008) showed that Whites with activity scores in the range of 1-2 had higher CYP2D6 activity compared to African Americans with scores in the same range. This resulted from significantly different CYP2D6 activities among Whites and African Americans who possessed CYP2D6*2-containing genotypes; CYP2D6*1/*2 genotype Whites had higher CYP2D6 activity compared to CYP2D6*1/*2 genotype African Americans, with the same relationship observed for White and African American individuals with the CYP2D6*2/*2 genotype. Likewise, our data suggest that the established CYP2A6 genetic variants (Table 2 in section 1.5.1.1 of the General Introduction) do not fully account for ethnic/racial differences in the NMR, as the mean NMR of CYP2A6*1/*1 genotype groups (excluding all known established CYP2A6 alleles) varied widely between the NP, SW, Alaska Native, White, and African American populations (Fig. 17 in Chapter 2, section 3.4). Therefore, if all individuals with the CYP2A6*1/*1 genotype, regardless of ethnic/racial background, were assigned the same
CYP2A6 activity score (or wGRS), this score would correspond to different mean NMRs in distinct ethnic/racial populations.

This highlights the utility of incorporating the novel high-frequency CYP2A6 SNPs into models of CYP2A6 activity. The novel five-SNP diplotypes account for an additional 4% of variation in CYP2A6 activity among CYP2A6*1/*1 genotype White liver donors, and these SNPs may account for an even higher proportion of variation in CYP2A6 activity in the NP, or other ethnic/racial populations. In support of this, we have observed wider NMR variability within the CYP2A6*1/*1 genotype group for NP smokers (coefficient of variation, CV 51%; calculated from data in Fig. 16, in Chapter 2, section 3.4) compared to White smokers (CV 44%; calculated from data in Fig. 29, in Chapter 4, section 5.4). Considering this greater variability, and the different CYP2A6 allele frequencies across different ethnic/racial groups, the functionally significant SNPs that comprise the novel five-SNP diplotype may account for additional variation in NMR in the NP population than in White populations. Alternatively, as discussed in section 6.4, the variation in the NMR of NP CYP2A6*1/*1 genotype group may reflect the presence of additional CYP2A6 SNPs, distinct from the novel SNPs characterized in Chapter 4.

Considering the ethnic/racial differences in genetic diversity and LD patterns (Tishkoff and Verrelli, 2003), including ethnicity/race and additional CYP2A6 genetic variants, many of which are yet to be discovered and characterized, as variables in prediction algorithms is likely to improve the performance of NMR predictions.

6.6 Conclusions

Our findings contribute to the existing literature on smoking patterns in ethnic/racial minority populations. Smoking quantity and tobacco exposure had not previously been investigated in the Northern Plains (NP) and Southwest (SW) American Indian populations using a tobacco-specific biomarker. We have demonstrated that smokers in these tribes have low daily tobacco consumption and do not appear to compensate for light smoking by adjusting their smoking topography, based on their low plasma cotinine levels. Despite the light smoking by current smokers, secondhand smoke exposure is high among non-smokers. This further highlights an important area of public health concern for these tribal populations, as both active and passive smoke exposure are risk factors for tobacco-related diseases such as lung cancer.
We have expanded on the literature on ethnic/racial CYP2A6 genetic variation and variable rate of nicotine metabolism, which have been shown to contribute to differences in smoking behaviours and cessation. The findings presented here suggest that the NP and SW tribes are genetically distinct at the CYP2A6 gene locus, which may be indicative of previous genetic divergence between the American Indian populations. Future studies should focus on the characterization of Ancestry Informative Markers (AIMs) for different American Indian tribes, to improve our ability to account for differences in population structure in future genetic association studies. Using a biomarker of the rate of nicotine metabolism, the nicotine metabolite ratio (NMR), we have demonstrated that the metabolism rate is high among NP smokers, compared to SW smokers, and compared to smokers of other ethnic/racial backgrounds. As higher NMR has previously been shown to be associated with more smoking, less success quitting, and greater lung cancer risk, these findings support the consideration of the high NMR in future studies of smoking and lung cancer risk in the NP population. Further, these findings highlight areas for future research, including investigation of genetic and environmental factors that contribute to variation in the NMR in the NP and SW tribes.

We have identified additional genetic and non-genetic sources of variation in the rate of nicotine metabolism in a human liver bank. For the first time, we characterized the functional significance of seven novel high frequency CYP2A6 genetic variants, with respect to CYP2A6 mRNA expression, protein levels, and enzyme activity. The identification and characterization of novel CYP2A6 SNPs may help improve our ability to predict the rate of nicotine metabolism, and further, improve cessation outcomes by personalizing the choice of pharmacotherapy. Future studies should focus on optimizing the prediction of the NMR, through incorporation of variables such as CYP2A6 genotype, ethnicity/race, and non-genetic factors into phenotype prediction models.
BIBLIOGRAPHY


GAEDIGK, A., TWIST, G. P. & LEEDER, J. S. 2012. CYP2D6, SULT1A1 and UGT2B17 copy number variation: quantitative detection by multiplex PCR. Pharmacogenomics, 13, 91-111.


IZUKAWA, T., NAKAJIMA, M., FUJIWARA, R., YAMANAKA, H., FUKAMI, T.,
Quantitative analysis of UDP-glucuronosyltransferase (UGT) 1A and UGT2B
expression levels in human livers. *Drug Metab Dispos*, 37, 1759-68.

JAAKKOLA, J. J. & GISSLER, M. 2004. Maternal smoking in pregnancy, fetal development,

JACOB, P., 3RD, YU, L., DUAN, M., RAMOS, L., YTURRALDE, O. & BENOWITZ, N. L.
2011. Determination of the nicotine metabolites cotinine and trans-3'-hydroxycotinine
in biologic fluids of smokers and non-smokers using liquid chromatography-tandem
mass spectrometry: biomarkers for tobacco smoke exposure and for phenotyping
cytochrome P450 2A6 activity. *J Chromatogr B Analyt Technol Biomed Life Sci*, 879,
267-76.

JAMAL, A., AGAKU, I. T., O’CONNOR, E., KING, B. A., KENEMER, J. B. & NEFF, L.


JARVIK, M. E., MADSEN, D. C., OLMSTEAD, R. E., IWAMOTO-SCHAAP, P. N., ELINS,

JARVIS, M. J., TUNSTALL-PEDOE, H., FEYERABEND, C., VESEY, C. & SALOOJEE, Y.
1987. Comparison of tests used to distinguish smokers from nonsmokers. *Am J Public
Health*, 77, 1435-8.

JENSEN, J. A., SCHILLO, B. A., MOILANEN, M. M., LINDGREN, B. R., MURPHY, S.,
exposure in nonsmoking hospitality workers before and after a state smoking ban.

JHA, P., MACLENNAN, M., CHALOUPKA, F. J., YUREKLI, A.,
RAMASUNDARAHAETTIGE, C., PALIPUDI, K., ZATONKSI, W., ASMA, S. &
GUPTA, P. C. 2015. Global Hazards of Tobacco and the Benefits of Smoking
Cessation and Tobacco Taxes. In: GELBAND, H., JHA, P.,
SANKARANARAYANAN, R. & HORTON, S. (eds.) *Cancer: Disease Control

JOHANSSON, I. & INGELMAN-SUNDBERG, M. 2008. CNVs of human genes and their

JOHNSTONE, E., BENOWITZ, N., CARGILL, A., JACOB, R., HINKS, L., DAY, I.,
MURPHY, M. & WALTON, R. 2006. Determinants of the rate of nicotine metabolism

Differences in Duration of Smoking among Former Smokers in the National Health and
Nutrition Examination Surveys (NHANES). *Nicotine Tob Res*.

KAIVOSAARI, S., TOIVONEN, P., HESSE, L. M., KOSKINEN, M., COURT, M. H. &
FINEL, M. 2007. Nicotine glucuronidation and the human UDP-


LOUKOLA, A., BUCHWALD, J., GUPTA, R., PALVIAINEN, T., HALLFORS, J., TIKKANEN, E., KORHONEN, T., OLLIKAINEN, M., SARIN, A. P., RIPATTI, S.,


serum and saliva cotinine in African American compared to White smokers. *Cancer Epidemiol Biomarkers Prev.*


abuse. Results from the Epidemiologic Catchment Area (ECA) Study. JAMA, 264, 2511-8.


STATISTICS CANADA 2016b. Smokers, by age group and sex. Table 105-0501 and Catalogue no. 82-221-X. http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/health73b-eng.htm.


238


WORLD HEALTH ORGANIZATION 2016. Cardiovascular Diseases Fact Sheet. 


WORLD HEALTH ORGANIZATION 2017b. Cancer Fact Sheet. 

WORLD HEALTH ORGANIZATION 2017c. Chronic Respiratory Diseases: Causes of COPD. 


APPENDIX A:

Nicotine metabolite ratio (3-Hydroxycotinine/Cotinine) in plasma and urine by different analytical methods and laboratories: implications for clinical Implementation

Full Text:


**Background**: The highly genetically variable enzyme CYP2A6 metabolizes nicotine to cotinine (COT) and COT to trans-3’-hydroxycotinin (3HC). The nicotine metabolite ratio (NMR, 3HC/COT) is commonly used as a biomarker of CYP2A6 enzymatic activity, rate of nicotine metabolism, and total nicotine clearance; NMR is associated with numerous smoking phenotypes, including smoking cessation. Our objective was to investigate the impact of different measurement methods, at different sites, on plasma and urinary NMR measures from ad libitum smokers.

**Methods**: Plasma (n = 35) and urine (n = 35) samples were sent to eight different laboratories, which used similar and different methods of COT and 3HC measurements to derive the NMR. We used Bland–Altman analysis to assess agreement, and Pearson correlations to evaluate associations, between NMR measured by different methods.

**Results**: Measures of plasma NMR were in strong agreement between methods according to Bland–Altman analysis (ratios, 0.82–1.16) and were highly correlated (all Pearson r > 0.96, P < 0.0001). Measures of urinary NMR were in relatively weaker agreement (ratios, 0.62–1.71) and less strongly correlated (Pearson r values of 0.66–0.98, P < 0.0001) between different methods. Plasma and urinary COT and 3HC concentrations, while weaker than NMR, also showed good agreement in plasma, which was better than that in urine, as was observed for
NMR.

**Conclusions:** Plasma is a very reliable biologic source for the determination of NMR, robust to differences in these analytical protocols or assessment site.

**Impact:** Together this indicates a reduced need for differential interpretation of plasma NMR results based on the approach used, allowing for direct comparison of different studies.
APPENDIX B:

Does coffee consumption impact on heaviness of smoking?

Full Text:


**Background and aims:** Coffee consumption and cigarette smoking are strongly associated, but whether this association is causal remains unclear. We sought to: 1) determine whether coffee consumption causally influences cigarette smoking, 2) estimate the magnitude of any association, and 3) explore potential mechanisms.

**Design:** We used Mendelian randomization (MR) analyses of observational data, using publicly available summarised data from the Tobacco and Genetics (TAG) consortium, individual level data from the UK Biobank, and in vitro experiments of candidate compounds.

**Setting:** The TAG consortium includes data from studies in several countries. The UK Biobank includes data from men and women recruited across England, Wales and Scotland.

**Participants:** The TAG consortium provided data on N ≤ 38,181 participants. The UK Biobank provided data on N = 8,072 participants.

**Measurements:** In MR analyses, the exposure was coffee consumption (cups/day) and the outcome was heaviness of smoking (cigarettes/day). In our in vitro experiments we assessed the effect of caffeic acid, quercetin, and p-coumaric acid on the rate of nicotine metabolism in human liver microsomes and cDNA-expressed human CYP2A6.

**Findings:** Two-sample MR analyses of TAG consortium data indicated that heavier coffee consumption might lead to reduced heaviness of smoking (beta -1.49, 95% CI -2.88 to -0.09). However, in vitro experiments found the compounds investigated are unlikely to significantly inhibit the rate of nicotine metabolism following coffee consumption. Further MR analyses in UK Biobank found no evidence of a causal relationship between coffee consumption and heaviness of smoking (beta 0.20, 95% CI -1.72 to 2.12).
**Conclusions:** Amount of coffee consumption is unlikely to have a major causal impact on amount of cigarette smoking. If it does influence smoking, this is not likely to operate via effects of caffeic acid, quercetin, or p-coumaric acid on nicotine metabolism. The observational association between coffee consumption and cigarette smoking may be due to smoking impacting on coffee consumption, or confounding.
APPENDIX C:

CYP2D6 and CYP2A6 biotransform dietary tyrosol into hydroxytyrosol

Full Text:


The dietary phenol tyrosol has been reported to be endogenously transformed into hydroxytyrosol, a potent antioxidant with multiple health benefits. In this work, we evaluated whether tyrosine hydroxylase (TH) and cytochrome P450s (CYPs) catalyzed this process. To assess TH involvement, Wistar rats were treated with α-methyl-L-tyrosine and tyrosol. Tyrosol was converted into hydroxytyrosol whilst α-methyl-L-tyrosine did not inhibit the biotransformation. The role of CYP was assessed in human liver microsomes (HLM) and tyrosol-to-hydroxytyrosol conversion was observed. Screening with selective enzymatic CYP inhibitors identified CYP2A6 as the major isoform involved in this process. Studies with baculosomes further demonstrated that CYP2D6 and CYP3A4 could transform tyrosol into hydroxytyrosol. Experiments using human genotyped livers showed an interindividual variability in hydroxytyrosol formation and supported findings that CYP2D6 and CYP2A6 mediated this reaction. The dietary health benefits of tyrosol-containing foods remain to be evaluated in light of CYP pharmacogenetics.
APPENDIX D:
Pharmacogenetics of Nicotine and Associated Smoking Behaviors

Full Text:


This chapter summarizes genetic factors that contribute to variation in nicotine pharmacokinetics and nicotine's pharmacological action in the central nervous system (CNS), and how this in turn influences smoking behaviors. Nicotine, the major psychoactive compound in cigarette smoke, is metabolized by a number of enzymes, including CYP2A6, CYP2B6, FMOs, and UGTs, among others. Variation in the genes encoding these enzymes, in particular CYP2A6, can alter the rate of nicotine metabolism and smoking behaviors. Faster nicotine metabolism is associated with higher cigarette consumption and nicotine dependence, as well as lower quit rates. Variation in nicotine's CNS targets and downstream signaling pathways can also contribute to interindividual differences in smoking patterns. Binding of nicotine to neuronal nicotinic acetylcholine receptors (nAChRs) mediates the release of several neurotransmitters including dopamine and serotonin. Genetic variation in nAChRs, and in transporter and enzyme systems that leads to altered CNS levels of dopamine and serotonin, is associated with a number of smoking behaviors. To date, the precise mechanism underpinning many of these findings remains unknown. Considering the complex etiology of nicotine addiction, a more comprehensive approach that assesses the contribution of multiple gene variants, and their interaction with environmental factors, will likely improve personalized therapeutic approaches and increase smoking cessation rates.
APPENDIX E

Gender differences in the correlation of CYP2A6 phenotypes in a human liver bank

(A) Correlation of two measures of CYP2A6 enzyme activity (the velocity of cotinine formation from nicotine versus the velocity of 7-OH-coumarin formation from coumarin,
nmol/min per milligram) in male liver donors. (B) Correlation of two measures of CYP2A6 enzyme activity in female liver donors. (C) Correlation between CYP2A6 mRNA levels (FPKM values, fragments per kilobase per million reads) and CYP2A6 protein levels (pmol/mg) in male liver donors. (D) Correlation between CYP2A6 mRNA levels and CYP2A6 protein levels in female liver donors. (E) Correlation between CYP2A6 protein levels and CYP2A6 enzyme activity (cotinine formation from nicotine) in male liver donors. (F) Correlation between CYP2A6 protein levels and CYP2A6 enzyme activity (cotinine formation from nicotine) in female liver donors. Values for r and P were determined on the basis of Spearman correlations. The data illustrated in these figures do not account for other variables that may influence CYP2A6 phenotypes, such as CYP2A6 genotype, age, AKR1D1 mRNA levels, and POR protein levels, but these factors have been included in regression models of variation in CYP2A6 phenotypes (Chapter 3, Tables 13-16).