Spring 5-15-2016

Genetic Imputation: Accuracy to Application

Shelina Raynell Ramnarine
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST LOUIS

Division of Biology and Biomedical Sciences
Human and Statistical Genetics

Dissertation Examination Committee:
Nancy Saccone, Chair
Arpana Agarwal
Barak Cohen
John Rice
Elisha Roberson

Genetic Imputation: Accuracy to Application

by
Shelina Raynell Stancia Ramnarine

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2016
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Dr. Nancy Saccone for giving me the opportunity to join her laboratory in 2011. None of the work in this dissertation would be possible without her guidance, support, vision, and resources. Dr. Saccone has also given me a tremendous amount of freedom. This has allowed me to grow as a scientist and a person. She has given me opportunities to present at national and international meetings, provided me instruction in grant-writing, and allowed me to participate in numerous activities during my tenure. I learned to think and write like a scientist and I will always cherish this training.

Next, I would like to thank Dr. Barak Cohen. I first participated in a research project as an undergraduate student doing a summer fellowship in his lab. His mentorship, guidance, encouragement, and support paved the way for me to pursue a PhD. Through my graduate tenure, Dr. Cohen has been a phenomenal mentor. I would not be the scientist I am today without his guidance and support. I will always appreciate the role he has played in my life.

I would also like to thank the members of my dissertation committee: Dr. Arpana Agarwal, Dr. John Rice, and Dr. Elisha Robertson. Their advice and feedback were instrumental and shaped my work.

I would like to thank Dr. Laura Bierut, Dr. Sarah Hartz, Dr. Robert Culverhouse, Dr. Li-Shiun Chen and Emily Olfson. Their input and guidance was instrumental in shaping this work.

A special acknowledgement is necessary for the participants in studies that are included in this work. Human genetics could not be conducted without research participants and I am thankful for those individuals who participated.

I would be remise if I did not thank several friends. Their support was invaluable to me completing this work: Latricia Wallace, Donell Carey PhD, Marina Cheung DPT, David Dadey, Jordan Atkins MD, Ervin Malakaj PhD, Ashley Macrander, Dana King, Shirin Farhadian DMD, Michelle Faits, Stephanie Ramos, Erica Smith, and Bryan Davis.

Last but certainly not least, I would like to thank my family. I would not be the person that I am today without them. My grandmother, Jurah Ramnarine, and father, Sheldon Ramnarine, have been an unbelievable support for me. I would like to thank my mother, Rhonda Thomas for all that she has done for me. Finally, thanks to my sisters, Jada Ramnarine, Jené Ramnarine, and Tessa Thomas, as well as my extended family.

Shelina Raynell Stancia Ramnarine

Washington University in St. Louis

May 2016

Abstract of the Dissertation

Genetic Imputation: Accuracy to Application

by

Shelina Raynell Stancia Ramnarine

Doctor of Philosophy in Biology and Biomedical Sciences

Human and Statistical Genetics

Washington University in St. Louis, 2016

Professor Nancy Saccone, Chair

Genotype imputation, the process of inferring genotypes for untyped variants, is used to identify and refine genetic association findings. This body of work focuses on assessing imputation accuracy and uses imputed data to identify genetic contributors to mentholated cigarette preference.

Inaccuracies in imputed data can distort the observed association between variants and a disease. Many statistics are used to assess accuracy; some compare imputed to genotyped data and others are calculated without reference to true genotypes. Prior work has shown that the Imputation Quality Score (IQS), which is based on Cohen's kappa statistic and compares imputed genotype probabilities to true genotypes, appropriately adjusts for chance agreement; however, it is not commonly used. To identify differences in accuracy assessment, we compared IQS with concordance rate, squared correlation, and accuracy measures built into imputation programs. Genotypes from the 1000 Genomes reference populations (AFR N = 246 and EUR N = 379) were masked to match the typed single nucleotide polymorphism (SNP)

coverage of several SNP arrays and were imputed with BEAGLE 3.3.2 and IMPUTE2 in regions associated with smoking behaviors. Additional masking and imputation was conducted for sequenced subjects from the Collaborative Genetic Study of Nicotine Dependence and the Genetic Study of Nicotine Dependence in African Americans (N =1,481 African Americans and N =1,480 European Americans). Our results offer further evidence that concordance rate inflates accuracy estimates, particularly for rare and low frequency variants. For common variants, squared correlation, BEAGLE R^2, IMPUTE2 INFO, and IQS produce similar assessments of imputation accuracy. However, for rare and low frequency variants, compared to IQS, the other statistics tend to be more liberal in their assessment of accuracy. IQS is important to consider when evaluating imputation accuracy, particularly for rare and low frequency variants. This work directly impacts the interpretation of association studies by improving our understanding of accuracy assessments of imputed variants.

Mentholated cigarettes are addictive, widely available, and commonly used, particularly by African American smokers. We aim to identify genetic variants that increase susceptibility to mentholated cigarette use in hopes of gaining biological insights into risk that may ultimately improve cessation efforts. We begin by pursuing hypothesis-driven candidate genes and regions (*TAS2R38, CHRNA5/A3/B4*, *CHRNB3/A6,* and *CYP2A6/A7*) and extend to a genome-wide approach. This study involves 1,365 African Americans and 2,206 European Americans (3,571 combined ancestry) nicotine dependent current smokers from The Collaborative Genetic Study of Nicotine Dependence (COGEND) and Transdisciplinary Tobacco Use Research Center (UW-TTURC). Analyses were conducted within each cohort, and meta-analysis was used to combine results across studies and across ancestral groups. We identified some suggestively associated

variants, although none meet genome wide significance. This study represents a new, important aspect to understanding menthol cigarette preference. Further work is necessary to better understand this smoking behavior in efforts to improve cessation.

Chapter 1: Introduction

Genetic analyses attempt to identify and elucidate relationships between genetic variants and phenotypes of interest. The identification of genetic factors, which are a substantial component of disease risk, is important towards improving the prevention, diagnosis, and treatment of disease (Manolio et al., 2009). Researchers first used microsatellites in family-based linkage studies, obtained from large pedigrees, in efforts to illuminate the genetic architecture of a specific phenotype. Changes in society have influenced family sizes making family based studies challenging. Data from large pedigrees are still used today but microsatellites have largely been replaced by single nucleotide polymorphisms (SNPs). Linkage studies were the cornerstone of human genetics. However, these studies often produced large associated regions which limit the ability to understand the biology of disease (Manolio et al., 2009). Candidate gene studies, focused on specific regions and involved small sample sizes, were the precursors to Genome Wide Association Analyses (GWAS) (Manolio et al., 2009).

GWAS have dominated but their success has been debated in the field (Visscher, Brown, McCarthy, & Yang, 2012). These studies offered great promise as they involved thousands of unrelated individuals and millions of markers (Manolio et al., 2009). Some believe that GWAS have been a success because hundreds of associated variants have been identified. Since most of these variants are of modest effect, in aggregate do not account for a large proportion of the heritability, and often are not clearly biologically related to common diseases, some researchers consider GWAS a failure (Manolio, 2010; Manolio et al., 2009; Visscher et al., 2012). Conversely, GWAS have provided many insights into the genetic architecture of Mendelian

disorders (Manolio et al., 2009). The common disease, common variant hypothesis and commercial SNP arrays are two underlying components of GWAS (Manolio et al., 2009). The common disease, common variant hypothesis is based on the idea that common diseases are caused by many common variants all with modest effect, as they have survived evolutionary selective pressures. Studies have shown that the aggregate effect of associated variants leave missing heritability (Manolio et al., 2009). Researchers are now pursing rare variants amongst other possibilities to gain a better understanding of disease. The common variant, rare disease hypothesis dictates that common diseases are caused by rare variants with larger effects. SNP arrays contain thousands or millions of SNPs and are believed to contain common variation in the human genome (Manolio et al., 2009). These arrays typically only contain variants with a minor allele frequency (MAF) greater than 5%, i.e. rare variants are excluded from traditional GWAS.

Gene-environment interactions are another potential source of missing heritability (Manolio et al., 2009). Phenotypes of interest, i.e. common human diseases or traits, are usually influenced by both genotypic and phenotypic factors. The environment has become an inclusive term representing all non-genetic factors: diet, exercise, even advertising, and other behaviors.

Linkage disequilibrium (LD) and population stratification make the identification of casual variants illusive at times. Crossing over and other forms of recombination influence the length of genomic regions that co-occur in populations, i.e. non-random allelic association or linkage disequilibrium (Teo, Small, & Kwiatkowski, 2010). These regions vary in length in different populations. Haplotypes are genetic markers that are present within a block marked by linkage disequilibrium. Tag SNPs are selected for SNP arrays with the idea that they define

haplotypes. Population stratification is that idea that allele frequencies vary in different populations due to ancestry. To this end, tag SNPs may not equally define haplotypes across populations necessitating the use of a greater number of variants or different variants when examining disease, hence the increased popularity of genetic imputation.

1.1 Imputation

Genotyping, or sequencing, is integral to analyzing variants for association with a trait; however, there may be missing data. Since cost and time prevents the wide use of whole genome sequencing in association studies and meta-analyses, imputation is a widely used tool in providing additional variants for testing. Genotype imputation, a method for inferring untyped genotypes, is an important tool for human genetic studies as it allows these inferred loci to be tested for association with phenotypes (B. L. Browning & Browning, 2009; B. Howie, Marchini, & Stephens, 2011; Li, Willer, Ding, Scheet, & Abecasis, 2010; Marchini & Howie, 2010; Purcell et al., 2007). Imputation programs exploit correlations between between typed and untyped genomic locations in haplotype blocks by matching similar haplotype patterns between reference and study samples (B. L. Browning & Browning, 2009; B. Howie et al., 2011; B. N. Howie, Donnelly, & Marchini, 2009; Marchini & Howie, 2010). Haplotype blocks are created by a lack or recombination in genomic regions of linkage disequilibrium which are regions which segregate together through inheritance. In association analyses, imputation can be used as a fine mapping tool to refine the region of association. In meta-analysis, imputation aligns multi-study data by extending each study's genotyping coverage to match a larger set (B. L. Browning & Browning, 2009; B. Howie et al., 2011; Marchini & Howie, 2010).

Genotyping arrays are the cornerstone of genetic imputation. These arrays have increased coverage over the years from thousands to millions of variants and are believed to capture common variation in the human genome. However, linkage disequilibrium and population stratification influence the ability of these variants to truly tag important haplotypes.

1.1.1 Imputation in Diverse Populations

Although race has been considered in genetic studies the concept is related to historical, social, and cultural influences (Garte, 2002). The differences in ancestry create linkage disequilibrium patterns and allele frequencies which are important to consider in the identification of biologically relevant disease determinants. Due to the historical focus on European ancestry populations in genetic studies, many of the approaches to disease risk identification and perceptions of specific causal variants may not be appropriate for all populations. There has been an effort to expand genetic tools to better understand the genetic architecture of disease in other populations. For example, companies have tried to optimize SNP arrays to consider population stratification as more diverse populations are being studied.

The 1000 Genomes Project is the newest form of large efforts to examine human disease in a variety of ancestral populations. The International HapMap, HapMap2 and HapMap3 were precursors to this project. The genomes of individuals from a variety of countries are contained in these studies: Yoruba from Ibadan, Nigeria; The CEU (Utah resides with northern and western European ancestry from the CEPH collection); JPT + CBH (Japanese from Tokyo, Japan and Chinese from Beijing, China) (Marchini & Howie, 2010). 1000 Genomes

has been expanded to include individuals from African Ancestry in Southwest US (ASW); Utah residents (CEPH) with Northern and Western European ancestry (CEU); Luhya in Webuye, Kenya (LWK); Yoruba in Ibadan, Nigeria (YRI); Iberian populations in Spain (IBS); Finnish from Finland (FIN); Toscani in Italia (TSI); British from England and Scotland (GBR). These individuals have been sequenced across their genomes which allow these data to be used as reference panels in imputation as the representation of a variety of ancestries improves the ability to infer missing genotypes.

Although more populations are being considered, imputation in African Americans remains a challenge. Hancock et al demonstrates that the imputation quality of low frequency variants is reduced when more closely related reference panels are considered (Hancock et al., 2012). Furthermore, Nelson et al found that SNP arrays do not represent common variants in African Americans as well as other groups since 75% of common variants are covered in all groups except African Americans (Nelson et al., 2013). Several studies have examined the ability of several reference panels, SNP arrays, and imputation programs to accurately infer imputed data in African Americans (Chanda et al., 2012; Hancock et al., 2012; Nelson et al., 2013).

1.1.2 Imputation Accuracy

Due to the widespread use of imputed data and a desire to understand the true relationship between genetic loci and a trait or disease of interest, it is important to assess true accuracy of inferred genotypes. Since variants are inferred based on correlations with typed variants, the genotype coverage of the SNP array used in imputation can influence accuracy. Although the decision of which commercially available SNP arrays is best varies based on

ancestral population and SNP coverage, generally, the array with the most variants that best tag haplotypes of interest should produce the best imputation quality scores. There are two classes of statistics used for accuracy estimation: (1) statistics which use both genotyped and imputed data and (2) statistics which use only imputed data for accuracy estimation. For the former, one must first obtain both genotyped and imputed data.

Since genotyped data is usually unavailable at imputed loci, evaluations of imputation accuracy that require true genotypes for comparison with imputed genotypes typically rely on either masking variants, or using a leave-one-out approach in which one individual is imputed using the remaining reference panel members. Researchers typically mask a percentage of variants from a commercially available SNP array and use the imputed data to compare with the genotyped data to provide some accuracy estimation (Chanda et al., 2012; Hancock et al., 2012; Lin et al., 2010; Sung et al., 2012). This method is limited since the imputation quality information does not inform general accuracy as it only applies to the masked variants. Duan et al 2013 used MACH-Admix software to create a database of variants and their imputation quality score using the leave-one-out method in which each person is removed individually from the reference panel and imputed using the other individuals in the reference panel given the genotype coverage of a commercially available SNP array. This method allows true genotypes to be compared with imputed genotypes for each left-out individual, and produces accuracy estimates of all variants except those on a SNP array. However, the imputation quality produced by this method is a specific to the reference panel used; the quality score produced may be higher or lower than the quality score for another study. Furthermore, this database only provides quality information as measured by squared Pearson correlation coefficient.

1.1.2.1 Accuracy Based on Genotyped and Imputed Data

Several statistics compare genotyped and imputed data when assessing accuracy: squared Pearson correlation coefficient (squared correlation), Imputation Quality Score (IQS), concordance rate. Although these statistics can be calculated using the most likely genotypes (best guess genotypes) or the posterior genotyped probabilities, it is preferable to use the genotyped probabilities in their calculation as the discrete classification of each individual's genotype does not consider the probabilistic nature of imputation.

Concordance rate is the proportion of matching genotypes divided by the total proportion of genotypes. Lin et al 2010 describes concordance rate as the gold standard for imputation quality assessment. Concordance rate was found to inflate accuracy for rare and low frequency variants due to chance concordance or chance agreement (Lin et al., 2010). Chance agreement causes a decrease in accuracy as MAF increases (Lin et al., 2010). Due to allele frequency, there is a low probability of the rare allele being present in the imputed sample; therefore, when the major allele is assigned this inference would be "correct" by chance. As the frequency of the allele decreases the accuracy rate increases e.g. 5% MAF is 90% accuracy. This inflation is increasingly problematic given that studies are becoming more interested in examining low frequency variants. IQS is based on Cohen's Kappa Coefficient and was introduced by Lin et al. 2010 to assess imputation accuracy. Squared correlation is the squared Pearson correlation coefficient which refers to the linear relationship between two variables (original and imputed dosage for each SNP). It estimates the ability of a linear model to depict the relationship between two variables.

1.1.2.2 Accuracy Based Only on Imputed Data

There are several statistics which use only imputed data to assess accuracy. Imputation efficiency is a measure of how confidently the imputation software was able to infer the individual's genotype. The efficiency is the number of individuals that have a probability greater than 0.9 (Lin et al., 2010) using the posterior genotype probabilities file. For each SNP, the number of people with a probability greater than 0.9 for one of the three genotype probabilities was recorded. That number divided by the total number of people is the SNP efficiency.

Of the statistics that use only imputed data, some are calculated by imputation software programs: Beagle R^2, Impute2 Info, and MACH r^2. Beagle R^2 or Allelic R^2 has good precision if the genotype probabilities are accurately calculated (B. L. Browning & Browning, 2009); however, these probabilities may not be accurately calculated for several reasons, e.g. poor LD between genotyped and imputed variants.

1.2 Smoking Behaviors and Nicotine Dependence

Smoking represents a massive public health burden. Smoking related diseases are a preventable cause of premature mortality particularly in youth and minorities resulting in cancer, cardiovascular disease, and pulmonary disease accounting for 1 in 5 deaths in the U.S. and 1 in 10 deaths worldwide (Benowitz, 2010). Population data have shown smoking rates to be approximately 19% in European Americans and 18% in African Americans (http://www.lung.org/stop-smoking/smoking-facts/tobacco-use-racial-and-ethnic.html).

There are several differences phenotypically in smoking behavior between African, and European Americans. African Americans smoke fewer cigarettes per day (CPD) (Luo et al., 2008; Okuyemi, Ebersole-Robinson, Nazir, & Ahluwalia, 2004; Stellman et al., 2003), start smoking regularly at a later age (McCarthy et al., 1995), are more likely to attempt to quit (but fail at quitting at higher rates) (Okuyemi et al., 2004), prefer to smoke menthol cigarettes (McCarthy et al., 1995; Muscat, Richie, & Stellman, 2002) and experience more smoking related illnesses (Okuyemi et al., 2004). Furthermore, differences in allele frequencies influence the genetic architecture of nicotine dependence in these populations.

Understanding the similarities and differences in the genetic architecture of smoking behavior in African and European Americans may lead to advances in the treatment and prevention of nicotine dependence. Cross population comparisons of associated regions are important to identifying signals that are unique to certain populations and others that are consistent across populations. Some of these studies have demonstrated the use of imputation to enhance the identification of signals in smoking associated regions (J. Z. Liu et al., 2010; Tobacco and Genetics Consortium, 2010).

Several genetic loci have been identified for smoking behavior through genome-wide association studies and meta-analyses (Bierut et al., 2007; J. Z. Liu et al., 2010; Saccone et al., 2010; Scott F. Saccone et al., 2007; Thorgeirsson et al., 2010; Tobacco and Genetics Consortium, 2010). Some genetic loci are genome wide significant (GWS) in European American individuals while others are suggestive in their association with smoking behavior in African Americans (Chen et al., 2012; David et al., 2012; S. F. Saccone et al., 2007). These previously identified variants are mostly contained in cholinergic nicotinic receptors on chromosomes 8 (*CHRNB3*-

CHRNA6), 15 (*CHRNA5-CHRNA3-CHRNB4*), and 19 (*CYP2A6-CYP2A7*) (Bierut et al., 2007; Chen et al., 2014; Chen et al., 2012; David et al., 2012; Saccone et al., 2010; Scott F. Saccone et al., 2007; Tobacco and Genetics Consortium, 2010). Genetic variants in *CHRNA5* on chromosome 15 are most strongly associated with smoking behavior (Chen et al., 2012; David et al., 2012; Rice et al., 2012).

1.2.1 Measures of Nicotine Dependence

Nicotine dependence can be classified by the six questions known as the Fagerström test for nicotine dependence (FTND) (Heatherton, Kozlowski, Frecker, & Fagerstrom, 1991). These questions are scored on a 10 point scale and nicotine dependent cases can be classified as those with an FTND ≥ 4 while controls can be classified as those with FTND ≤1. Using this measure, 60% of current smokers are nicotine dependent (Bierut, 2010). The questions, responses, and scoring are as follows:

Table 1: Fagerstrom Test for Nicotine Dependence Questions, Responses, Scores

Questions	Responses	Scores
How soon after you wake up do you smoke your first cigarette?	Within 5 minutes 6-30 minutes 1-60 minutes After 60 minutes	3 2 1 0
Do you find it difficult to refrain from smoking in places where it is forbidden (e.g., in church, at the library, in cinema, etc)?	Yes No	1 0
Which cigarette would you hate most to give up?	The first one in the morning All others	1 0
How many cigarettes per day do you smoke?	10 or less 11-20 21-30 31 or more	0 1 2 3

Do you smoke more frequently during the first hours after walking than during the rest of the day?	Yes No	1 0
Do you smoke when you are so ill that you are in bed most of the day?	Yes No	1 0

Cigarettes-per-day (CPD) is also used as a measure of nicotine dependence (Heatherton et al., 1991; Scott F. Saccone et al., 2007). Both CPD and FTND are necessary to examine since they are both standard assessments of nicotine addiction (Luo et al., 2008). CPD is a continuous measure of smoking behavior and a commonly available proxy for nicotine dependence in studies not focused on smoking behavior. Additionally, CPD and FTND predict cotinine levels (Luo et al., 2008). Both measures are also differentially associated with certain genetic variants (Rice et al., 2012). While these measures are correlated, they are not identical especially in African-Americans.

1.2.2 Mentholated Cigarettes

Mentholated cigarettes are popular and differently used by African and European Americans. These cigarettes were first marked to be used when one had the cold or a cough, which prevented the use of non-mentholated cigarettes (Gardiner, 2004). Advertising by brands such as Kool gave the impression that the cigarettes were healthier (Gardiner, 2004). The Federal Trade commission sued the company for false advertising but tobacco companies continued to market mentholated cigarettes as healthier (Gardiner, 2004). One in four cigarettes are classified as mentholated, containing menthol a component in peppermint oil (TPSAC;(Giovino et al., 2004). These cigarettes generally contain more tar and nicotine than

non-mentholated cigarettes (Gardiner, 2004; Muscat et al., 2002; Okuyemi et al., 2004). The cooling effects of these cigarettes are believed to contribute to their popularity.

There are differences in smoking behavior of mentholated cigarette smokers. African American menthol smokers have decreased quit rates compared to African American non-menthol smokers (Okuyemi et al., 2004). Menthol smokers are more likely to be African American, less educated, younger, and never married (Fagan et al., 2010; Okuyemi et al., 2004). Although smoking rates in African and European American populations are similar, the National Surveys on Drug Use and Health report that between 2004 and 2008 82.6% of Black/African Americans and 23.8% of Whites smoked mentholated cigarettes (Office of Applied Studies).

Mentholated cigarettes have been marketed differently to African Americans. Gardiner et al 2004 discusses the African Americanization of menthol cigarette use in the United States. The migration of African Americans from rural areas to urban cities led tobacco companies to begin to target this growing consumer market (Gardiner, 2004). The popularity of mentholated cigarettes have been linked to their increased use in the African American community and the rise and fall of some mentholated cigarette brands was predicated on their perception in this community. Gardiner et al discusses a rumor that the K in Kool represented the Ku Klux Klan which led to the decreased use of this brand by the African American community. The sponsorship of Civil Rights efforts by tobacco companies and the perceived health benefits further solidified the use of this product in the African American community (Gardiner, 2004). During the civil rights movement menthols were smoked because of their association as being the cigarette used by individuals who are brave, ambitious and daring (Gardiner, 2004).These cigarettes were advertised on television and promoted by prominent figures in the African

American community in magazines such as Ebony, a magazine targeting the African American community (Gardiner, 2004). Mentholated cigarettes still remain in high use amongst African Americans smokers.

Chapter 2: When Does Choice of Accuracy Measure Alter Imputation Accuracy Assessments?

2.1 Introduction

In genomic analyses high-quality data are crucial to accurate statistical inferences. Data accuracy can typically be assessed by different methods and measures.

Genetic imputation provides an informative scenario for examining how the use of different accuracy measures can influence the assessment of accuracy. Genotype imputation is a valuable tool in association studies and meta-analyses. This process infers "in silico" genotypes for untyped variants in a study sample by matching genotyped variants in the study to corresponding haplotypes in a comprehensively genotyped reference panel (B. L. Browning & Browning, 2009; S. R. Browning, 2006; B. Howie, Fuchsberger, Stephens, Marchini, & Abecasis, 2012; B. Howie et al., 2011; B. N. Howie et al., 2009; Li et al., 2010; E. Y. Liu, Li, Wang, & Li, 2013; Marchini & Howie, 2010). Therefore, imputation accuracy is influenced by haplotype frequencies in the reference panel (Hancock et al., 2012; Sung et al., 2012) and the typed single nucleotide polymorphism (SNP) coverage of the study sample (Johnson et al., 2013; Nelson et al., 2013). Once untyped variants are inferred, statistics that measure imputation accuracy are calculated to identify poorly imputed SNPs.

Imputation accuracy statistics can be classified into two types: (1) statistics that compare imputed to genotyped data and (2) statistics produced without reference to true genotypes. Concordance rate, squared correlation, and Imputation Quality Score (IQS) (Lin et al., 2010) are examples of the first type. Because imputed SNPs usually do not have genotyped data for comparison, statistics of the second type are usually provided by imputation programs and are commonly relied upon in practice. However, a direct comparison of imputed and

genotyped data can be made possible by masking a percentage of variants that were genotyped in the study sample (Chanda et al., 2012; Hancock et al., 2012; Shriner, Adeyemo, Chen, & Rotimi, 2010).

Lin et al (2010) introduced IQS, which is based on Cohen's kappa statistic for agreement (Lin et al., 2010). Because of chance agreement, concordance rate, i.e. the proportion of agreement, can lead to incorrect assessments of accuracy for rare and low frequency variants. IQS adjusts for chance agreement (Lin et al., 2010). Furthermore, Lin et al. (2010) used simulated data to show that requiring an IQS threshold > 0.9 removed all false positive association signals, while concordance rate > 0.99 still resulted in many false positives. Despite this evidence, IQS is not widely used in accuracy assessment.

This work builds upon previous studies by comparing IQS with commonly used accuracy measures - concordance rate, squared correlation, and built-in accuracy statistics - with the goal of identifying situations in which the choice of accuracy measure leads to differing assessments of accuracy. We compared imputed and genotyped data via masking, and used African-ancestry and European-ancestry populations to evaluate imputation accuracy in genomic regions associated with nicotine dependence and smoking behavior, some of which have also been implicated in lung cancer and chronic obstructive pulmonary disease (COPD).

2.2 Methods

We examined differences and similarities in accuracy assessment as measured by IQS, squared correlation, concordance rate and built-in accuracy statistics using: (1) 1000 Genomes as the sample and the reference, and (2) data from nicotine dependence studies as the sample and

1000 Genomes as the reference. Below we describe both approaches, beginning with analyses involving 1000 Genomes as the sample and the reference.

2.2.1 Masking and Imputation using 1000 Genomes Data

Because IQS adjusts for chance agreement (Lin et al., 2010), we used IQS as a benchmark for accuracy estimation. Calculating IQS, concordance rate, and squared correlation requires genotyped data for comparison with imputed data. We created a study sample for imputation by masking genotypes in the reference panel to mimic the typed SNP coverage of commercially available SNP arrays (Affymetrix – Affy 500 and Affy 6 as well as Illumina – Duo, Omni, and Quad matched by genomic position using Build 37.3/hg19). We used 1000 Genomes African (AFR) and European (EUR) continental reference panels with 246 and 379 individuals respectively (Table 2.1) (The 1000 Genomes Project Consortium, 2012). All data analyzed here are de-identified, publicly available data from the 1000 Genomes (1000G) project, which provides these data as a resource for the scientific community. Participants provided informed consent to the 1000G Project for broad use and broad data release in databases (The 1000 Genomes Project Consortium, 2010, 2012). We also have Washington University Human Research Protection Office approval for analyses of de-identified data.

The process of creating the study sample is described in Figure 2.1 and the numbers of typed variants are presented in Table 2.2. Figure 2.1 illustrates several key characteristics of our masking approach. The reference panel individuals were the same as the study sample individuals. Our approach is expected to give an upper bound on accuracy because of the ideal match between the reference panel and study sample; the “correct” haplotype for each

individual being imputed is present in the reference. Using population-specific reference panels (AFR and EUR) rather than a cosmopolitan reference panel maximizes the matching between the reference panel and study sample. Also, this design allowed us to compare accuracy estimates for variants not found on a SNP array. This sample data set was then imputed and the results were used to calculate accuracy statistics.

2.2.2 Imputation Programs

BEAGLE (version 3.3.2) (B. L. Browning & Browning, 2009; S. R. Browning, 2006) and IMPUTE2 (B. Howie et al., 2012; B. Howie et al., 2011; B. N. Howie et al., 2009) were used to obtain imputed genotype probabilities. We obtained the BEAGLE R^2 and IMPUTE2 INFO accuracy measures for each SNP; neither of these makes use of true genotypes. The BEAGLE R^2 and IMPUTE2 INFO accuracy measures are well established (Chanda et al., 2012; Marchini & Howie, 2010). BEAGLE R^2 approximates the squared correlation between the most likely genotype and the true unobserved allele dosage (B. L. Browning & Browning, 2009; S. R. Browning, 2006). IMPUTE2 INFO considers allele frequency as well as the observed and expected allele dosage (Chanda et al., 2012). We include their formulas for completeness, in Equation 1 and 2, Here g_n represents the observed dosage, e_n represents the expected allele dosage, and $\hat{\theta}$ represents the sample allele frequency for sample n at a particular SNP, where n ranges from 1 to N, the total number of individuals and $0 < \hat{\theta} < 1$. Additionally, z_n represents the genotype with the highest posterior probability from imputation, i.e. 0, 1, or 2 corresponding to the number of copies of the coded allele. Finally, $f_n = p_{n1} + 4p_{n2}$ where p_{nk} represents the imputed probability of the genotypic class k (0, 1, and 2) corresponding to the nth sample.

$$Equation\ 1.\ BEAGLE\ R^2 = \frac{[\sum_{n=1}^{N} g_n e_n - (1/N)(\sum_{n=1}^{N} g_n \sum_{n=1}^{N} e_n)]^2}{[\sum_{n=1}^{N} f_n - (1/N)(\sum_{n=1}^{N} e_n)^2][\sum_{n=1}^{N} z_n - (1/N)(\sum_{n=1}^{N} z_n)^2]}$$

$$Equation\ 2.\ IMPUTE2\ INFO = 1 - \frac{\sum_{n=1}^{N}(f_n - e_n^2)}{2N\hat{\theta}(1-\hat{\theta})}$$

Imputed probabilities produced by BEAGLE and the corresponding accuracy statistics showed variability, so we focus on these results. Analyses using IMPUTE2 were less informative in this matched sample-reference setting; this program appears to identify the matching individual in the reference and assign imputed data accordingly. The result was highly accurate imputation in this special context. Since we aim to compare concordance rate, squared correlation, and IQS in efforts to identify scenarios where these statistics produce similar or divergent conclusions regarding accuracy estimation, the variation produced by using BEAGLE for imputation allows us to address our question of interest.

2.2.3 Statistics that Compare Genotyped and Imputed Data

The imputed genotype probabilities produced by BEAGLE and IMPUTE2 were used to calculate concordance rate, squared correlation and IQS. These imputed genotype probabilities, one for each genotype class (e.g. AA, AB, or BB), are transformed to dosage values by multiplying by 0, 1 or 2 for each genotypic class. IQS is calculated from genotype probabilities while squared correlation uses dosage values. Note that a specific dosage value can correspond to multiple genotypic probabilities, but only one dosage value can result from a specific set of

genotypic probabilities. Although the most likely (best guess) genotype for each variant can be used to calculate these statistics, it is not recommended because the discrete classification of each individual's genotype does not consider the probabilistic nature of imputation (J. Zheng, Li, Abecasis, & Scheet, 2011).

The incorporation of the genotypic classes into the IQS calculation is represented in Table 2.1, where each cell is the sum of the genotype probabilities for each genotyped and imputed genotypic class combination. The IQS calculation is demonstrated in Equation 3. IQS considers both the observed proportion of agreement (concordance rate or P_o shown in Equation 4) as well as chance agreement (P_c in Equation 5). Concordance rate (P_o) is the sum of probabilities for each matching genotypic class divided by the total sum of all genotype probabilities. Chance agreement is evaluated as the sum of the products of the marginal frequencies. An IQS score of one indicates that the data matched perfectly, while a negative IQS score indicates that the SNP was imputed worse than expected by chance (Lin et al., 2010). Mathematically, the value of IQS will always be less than or equal to the value of concordance rate: $P_oP_c \leq P_c$, so $P_o - P_c \leq P_o - P_oP_c$, hence $(P_o-P_c)/(1-P_c) \leq (P_o-P_oP_c)/(1-P_c)$, which says that IQS $\leq P_o$. Some statistics can be confounded with Hardy-Weinberg equilibrium (HWE) if they assume HWE to calculate "expected" genotype counts (Shriner, 2013). IQS avoids this concern since it uses imputed and experimentally determined genotypes.

Table 2.1: Calculating concordance (P_0) and IQS from imputed genotype probabilities and actual genotypes. The table was created by summing over probabilities for all N individuals (n = 1 to N) in each cell with p_{ij_n} representing the probability that the nth individual has the

imputed genotype i and actual genotype j, where 1 corresponds to AA, 2 corresponds to AB, and 3 corresponds to BB. N_1 = number of individuals with AA actual genotype, N_2 = number of individuals with AB actual genotype, N_3 = number of individuals with BB actual genotype, and N = number of total individuals.

Table 2: Calculating concordance (P0) and IQS from imputed genotype probabilities and actual genotypes

		Actual			
		AA	AB	BB	Total
Imputed	AA	$\sum_{n=1}^{N} p_{11\_n}$	$\sum_{n=1}^{N} p_{12\_n}$	$\sum_{n=1}^{N} p_{13\_n}$	$\sum_{j=1}^{3}\sum_{n=1}^{N} p_{1j\_n}$
	AB	$\sum_{n=1}^{N} p_{21\_n}$	$\sum_{n=1}^{N} p_{22\_n}$	$\sum_{n=1}^{N} p_{23\_n}$	$\sum_{j=1}^{3}\sum_{n=1}^{N} p_{2j\_n}$
	BB	$\sum_{n=1}^{N} p_{31\_n}$	$\sum_{n=1}^{N} p_{32\_n}$	$\sum_{n=1}^{N} p_{33\_n}$	$\sum_{j=1}^{3}\sum_{n=1}^{N} p_{3j\_n}$
	Total	$\sum_{i=1}^{3}\sum_{n=1}^{N} p_{i1\_n} = N_1$	$\sum_{i=1}^{3}\sum_{n=1}^{N} p_{i2\_n} = N_2$	$\sum_{i=1}^{3}\sum_{n=1}^{N} p_{i3\_n} = N_3$	N

$$Equation\ 3. \qquad \mathrm{IQS} = \frac{Po - Pc}{1 - Pc}$$

$$Equation\ 4. \qquad Po = \frac{\sum_{n=1}^{N} p_{11_n} + \sum_{n=1}^{N} p_{22_n} + \sum_{n=1}^{N} p_{33_n}}{N}$$

$$Equation\ 5. \quad Pc = \frac{N_1 * \sum_{j=1}^{3}\sum_{n=1}^{N} p_{1j\_n} + N_2 * \sum_{j=1}^{3}\sum_{n=1}^{N} p_{2j\_n} + N_3 * \sum_{j=1}^{3}\sum_{n=1}^{N} p_{3j\_n}}{N^2}$$

Squared correlation is the square of the Pearson correlation coefficient between the imputed and genotyped dosage for each SNP. This is calculated using Equations 6-11 where x_i and y_j are the imputed and genotyped dosage values for the nth sample respectively. It represents the proportion of the variability in the imputed data that can be explained by the least squared regression model.

$$Equation\ 6. \qquad \mathrm{R}^2 = 1 - \frac{SSE}{SS_{yy}}$$

$$Equation\ 7. \qquad \mathrm{SS_{yy}} = \sum_{n=1}^{N} (y_i - \bar{y})^2$$

$$Equation\ 8. \qquad \mathrm{SSE} = SS_{yy} - \widehat{\beta_n}(SS_{xy})$$

$$Equation\ 9. \qquad \mathrm{SS_{xy}} = \sum_{n=1}^{N} (y_i - \bar{y})(x_j - \bar{x})$$

$$Equation\ 10. \qquad \widehat{\beta_n} = \frac{SS_{xy}}{SS_{xx}}$$

$$Equation\ 11. \qquad \mathrm{SS_{xx}} = \sum_{n=1}^{N} (x_j - \bar{x})^2$$

2.2.4 Evaluating Accuracy Across MAF and LD

Imputation accuracy is influenced by a variant's minor allele frequency (MAF) and linkage disequilibrium (LD) with genotyped variants (measured by pairwise squared correlation r^2). We examined imputation accuracy in relation to these properties. The MAFs used here were based on the allele frequencies found in the genotyped data. We will use the terminology "rare" to denote variants with MAF ≤ 1%; and "low frequency" to refer to variants with 1% < MAF ≤ 5%. For each imputed SNP, the genotyped SNP in the region with the highest LD was

used to define the maximum r^2_{LD} with a genotyped SNP (denoted by max r^2_{LD}). PLINK was used to generate the LD values (Purcell et al., 2007). Bins for maximum r^2_{LD} and MAF were defined in 0.01 increments (Lin et al., 2010). For each bin, the mean and one standard deviation of the values produced by each accuracy statistic were calculated.

2.2.5 Examining Regions Associated with Nicotine Dependence

We examined the imputation accuracy of two genomic regions known to be associated with nicotine dependence and smoking behavior. These regions were the nicotinic receptor subunit gene clusters on chromosome 15 (*CHRNA5-CHRNA3-CHRNB4*) and chromosome 8 (*CHRNB3-CHRNA6*) (Bierut et al., 2007; J. Z. Liu et al., 2010; Saccone et al., 2010; S. F. Saccone et al., 2007; Thorgeirsson et al., 2010; Tobacco and Genetics Consortium, 2010). These signals were identified through genome-wide association studies (GWAS) and meta-analyses for smoking behavior, with the chromosome 15 region being the most significantly associated. We imputed 3Mb on each chromosome: 2Mb regions used for analysis plus two 500Kb flanking buffer regions according to Build 37.3/hg19. We focused our analyses on polymorphic variants with dbSNP identifiers in each 2MB region.

2.2.6 Masking and Imputation in a Real Data Application using a Nicotine Dependence Sample

A comparison of accuracy statistics was also conducted using nicotine dependence data as the study samples (N=1,481 African Americans and N=1,480 European Americans who were sequenced) and 1000 Genomes as the reference. The study sample was masked and imputed

separately by race. This analysis provided a more conventional imputation scenario for comparison with the patterns found in the 1000 Genomes analyses.

The sequenced subjects in this applied analysis were from the Collaborative Genetic Study of Nicotine Dependence (COGEND) and the Genetic Study of Nicotine Dependence in African Americans (AAND). These studies are cross-sectional and contain extensive smoking behavior phenotypes in African Americans and European Americans (Bierut et al., 2007). These individuals were between the ages of 25-44 years old and were assessed for dependence as measured by the Fagerstrom Test for Nicotine Dependence (FTND) and cigarettes-per-day (CPD) (Luo et al., 2008). The study protocol was approved by the appropriate Institutional Review Boards and written informed consent was obtained from all subjects.

Center for Inherited Disease Research (CIDR) performed next-generation targeted sequencing on genomic regions previously associated with smoking behaviors, using COGEND and AAND DNA samples derived from blood. Genotypic data that passed initial quality control at CIDR were released to the Quality Assurance/Quality Control analysis team at the University of Washington Genetics Coordinating Center. These data had mean on-target coverage of 180X with more than 96% of on-target bases containing a depth greater than 20X. A total of 1,481 African Americans and 1,480 European Americans were used in the analysis.

These sequencing data were masked to match the typed SNP coverage of the Omni 2.5 SNP array in a 500kb region on chromosome 15. The cosmopolitan reference panel, composed of individuals from a variety of ancestries, was used for imputation since it has been shown to produce the best accuracy estimates (Hancock et al., 2012). The imputation was performed using BEAGLE and IMPUTE2 to evaluate whether observed trends in accuracy were consistent

across imputation programs. The imputed probabilities were compared to the masked sequencing data and accuracy statistics were calculated. We focused our analyses on polymorphic variants.

2.3 Results

We compared IQS with squared correlation, concordance rate, and BEAGLE R^2 to examine changes in accuracy assessment using 1000 Genomes as the study sample in Figs. 2-5. IQS is our benchmark because it adjusts for chance agreement, in contrast to concordance rate which inflates assessments of accuracy (Lin et al., 2010). We focus here on the results for the AFR reference population using Omni 2.5M typed coverage on chromosome 15 (13,442 imputed SNPs). We emphasize Omni 2.5 because it has the greatest genotype SNP coverage in the region (Table 2.4).

2.3.1 Results for 1000 Genomes Imputation with Matching Reference

Results produced using BEAGLE and the AFR reference population are shown. Results for different chromosomal regions and populations were similar and are shown in Figure 2.12-2.14.

To help interpret results that are displayed by MAF and max r^2_{LD} bin, S1 Fig. shows the number of imputed variants in each MAF bin in panel A and max r^2_{LD} bin in panel B. This figure indicates that most of the imputed variants were rare and low frequency variants. There were 6,480 (48.21%) rare and low frequency rsID SNPs in the AFR population. The bins ranged in size from 7 variants ($0.49 \geq MAF < 0.50$) to 2,371 variants ($0.01 \geq MAF < 0.02$).

2.3.2 Concordance Rate and BEAGLE R^2 Inflate Assessments of Accuracy for Rare Variants

Results show that the choice of statistic is important when examining the imputation accuracy of rare and low frequency variants. Figure 2.2 displays the mean accuracy and one standard deviation in each MAF bin, after imputing from Omni 2.5M coverage. IQS (Panel A) and squared correlation (Panel B) produced similar means and standard deviations in each bin, though this does not necessarily represent similarity of values for particular SNPs. For rare and low frequency variants, both concordance rate (Panel C) and BEAGLE R^2 (Panel D) produce inflated assessments of accuracy. The higher concordance rate and BEAGLE R^2 values could mislead a researcher into assuming that these variants were imputed well, and that accuracy is best measured using concordance rate and BEAGLE R^2. IQS and squared correlation also show low accuracy for rare variants using other SNP array coverages (Figure 2.8).

A MAF bin can have a wide range in accuracy values. Figure 2.2 shows variability within MAF bins across all MAF values. Standard deviations for IQS, squared correlation and BEAGLE R^2 can be sizeable for both rare and common variants (panels A, B and D); concordance rate does not reflect this as it classifies most variants as well imputed (panel C).

2.3.3 Rare and Low Frequency Variants can be Well Tagged but Poorly Imputed

We examined max r^2_{LD}, the maximum LD r^2 between imputed and genotyped SNPs, to understand the relationship between typed SNP coverage and imputation accuracy as measured by these accuracy statistics. Figure 2.3 displays the mean accuracy and one standard deviation in each max r^2_{LD} bin, after imputing from Omni 2.5M coverage, additional arrays are

in S3 Fig. Mean accuracy tends to increase with increasing max r^2_{LD}, as expected. For low to moderate max r^2_{LD}, we observed substantial variability in IQS as well as squared correlation and BEAGLE R^2 values; however, at high max r^2_{LD}, the variability decreases. IQS and squared correlation show a surprisingly wide standard deviation for variants in the highest max r^2_{LD} bin ($0.99 < \max r^2_{LD} \leq 1$) as well as the max r^2_{LD} bin $0.5 < \max r^2_{LD} \leq 0.51$. Upon investigation, we found that the variability was due to rare variants: after limiting to SNPs with MAF > 5%, these standard deviations were comparable to those of the other bins, Figure 2.10 This pattern suggests that even rare variants that are well tagged (as measured by max r^2_{LD}) can be poorly imputed.

2.3.4 Concordance Classifies Most Variants as Well Imputed

Concordance differs from IQS, squared correlation, and BEAGLE R^2 in that it indiscriminately classifies most variants as well imputed, across MAF (Figure 2.2) and r^2_{LD} bins (Figure 2.3). The results in Figure 2.2 and 2.3 support prior concerns regarding concordance rate (Lin et al., 2010) and led us to focus the rest of our evaluation on IQS, squared correlation, and BEAGLE R^2.

2.3.5 For Rare Variants, IQS and Squared Correlation Produce Different Assessments of Accuracy

Although squared correlation and IQS appeared similar overall in their assessment of imputation accuracy when examined using means and standard deviations by bin (Figure 2.2 and 2.3), further investigation showed that on an individual SNP level, these statistics produce divergent assessments of accuracy for rare and low frequency variants. We compared accuracy

estimates produced by IQS and squared correlation in Figure 2.4 for each SNP. Panel A shows results for all variants, and panel B displays results for variants with MAF > 5%. A comparison of these panels is useful to identify divergent trends for common variants versus rare and low-frequency variants. For most SNPs, IQS and squared correlation produced similar assessments of accuracy as seen by the many observations on and near the y=x line in panels A and B. This is consistent with the accuracy patterns observed for IQS and squared correlation in Figure 2.2 – 2.3. However, discrepancies in accuracy assessment do occur, with squared correlation generally being more liberal in assigning high accuracy compared to IQS. This is indicated by the sparseness of observations above the y=x line in panels A and B. The points below the y=x line indicate SNPs for which squared correlation values were higher than IQS**.** Panel B shows that widely discrepant values for IQS and squared correlation are attributable to rare and low frequency SNPs: filtering out SNPs with MAF ≤ 5% removes the widely discrepant observations.

To further examine trends in the discrepancies between these statistics, we subtracted squared correlation from IQS for each variant and displayed this result across all MAF values in Figure 2.11 Thus negative differences denote that squared correlation was greater than IQS (i.e. squared correlation more liberal) while positive differences indicate that IQS was greater than squared correlation. Large discrepancies occur over all MAF values with squared correlation tending to be higher than IQS, especially for SNPs with higher MAFs.

2.3.6 For Common Variants, IQS and BEAGLE R^2 Provide Similar Assessments of Accuracy

For common variants, BEAGLE R^2 produces a similar assessment of imputation accuracy as IQS, but BEAGLE R^2 can differ dramatically from squared correlation. In Figure 2.5, we

compared BEAGLE R^2 to IQS (panels A and C) and squared correlation to BEAGLE R^2 (panels B and D). For many variants, squared correlation and BEAGLE R^2 differ in accuracy assessment as seen by the variants above the y=x line in panel B. Although most of these variants are rare, there are still many common variants for which this trend is true (panel D). Large differences between IQS and BEAGLE R^2 occur mostly when rare variants are examined.

2.3.7 Results are Similar in Different Genomic Regions and Populations

Figure 2.2-2.5 displayed results for the AFR reference population and Omni 2.5M typed coverage in the chromosome 15 region. Results similar to those described above were also observed using the AFR reference on chromosome 8 (Figure 2.12) as well as using the EUR reference panel for chromosomes 15 and 8 (Figure 2.13 and 2.14. respectively). In particular, low IQS values do occur for rare variants that have high squared correlation or high BEAGLE R^2. The number of variants for each imputation subset can be found in Table 2.4.

2.3.8 Results are Consistent in Application to Nicotine Dependence Study Sample

Figure 2.6 shows results produced using African American individuals from the nicotine dependence data as the study sample and a 1000 Genomes cosmopolitan reference panel imputed using BEAGLE. These data show discrepancies in accuracy assessment between statistics. If IQS and squared correlation are compared, squared correlation tends to be similar or higher (i.e. more liberal) than IQS. In the applied scenario, we observed some variants with high IQS and low squared correlation (Figure 2.6, panel A, upper left quadrant), which was not observed for the upper bound values from the 1000 Genomes analysis (Figure 2.4, panel A);

however, these discrepancies are few, and mostly among rare and low frequency variants (see Figure 2.6, panel D). When comparing IQS to Beagle R^2, the applied scenario showed IQS to be similar to or less than Beagle R^2 (Figure 2.6, panel B), which recapitulates patterns seen in 1000 Genomes (Figure 2.5, panel A).

In European Americans, from the nicotine dependence data, we also observed these same patterns as in African Americans, with squared correlation's more liberal assignment of accuracy as compared to IQS, Figure 2.15 These results were also consistent using IMPUTE2 with African American and European American study samples, Figure 2.16 and 2.17. respectively. This confirms that these patterns are not limited to specific populations, chromosomes, or imputation programs.

2.4 Discussion

Genotype imputation is used to improve the density of genomic coverage and increase power by combining datasets (Winkler et al., 2014), in efforts to identify and refine genetic variants associated with disease. We investigated how assessment of imputation accuracy changes when concordance rate, squared correlation and BEAGLE R^2 are compared to IQS, focusing on two genomic regions associated with smoking behavior.

Results showed that the choice of accuracy statistic matters for rare variants more than for common variants. This is important given that researchers are increasingly interested in imputing rare and low frequency variants (E. Y. Liu et al., 2012; H.-F. Zheng, Ladouceur, Greenwood, & Richards, 2012; H.-F. Zheng et al., 2015). While it has been recognized that rare

variants are more difficult to impute accurately, our work here goes further by highlighting that choice of accuracy measure has an important role.

For common variants, squared correlation, IMPUTE2, and BEAGLE R^2 produce similar assessments of imputation accuracy as compared to IQS. For rare and low frequency variants, we observed varying assessments of accuracy compared to IQS. Our results also showed that discrepancies between IQS and squared correlation are most likely to occur at rare and low frequency variants, where squared correlation is more liberal in assigning higher accuracy as compared to IQS. An evaluation of nicotine dependence samples also showed discrepancies between IQS and squared correlation. We recommend calculating IQS to confirm imputation accuracy, especially for rare or low frequency variants.

The variability observed within a MAF or max r^2_{LD} bin is a reminder that not all variants that share the same MAF or max r^2_{LD} value can be imputed with the same level of accuracy. This is consistent with the expectation that the inference of untyped variants depends on haplotype block structure and not simply the pairwise relationships between the genotyped and untyped variants. For rare variants, high LD with a genotyped SNP may not guarantee high imputation accuracy. Still, overall, a high max r^2_{LD} usually implies high accuracy, as we observed increasing mean accuracy along with decreasing variability within max r^2_{LD} bins as max r^2_{LD} increases.

We applied this approach to genomic regions associated with our phenotype of interest, smoking behavior using an upper bound scenario and a nicotine dependence sample. Thus, one limitation is that rather than comprehensively examining the genome, we focused only on selected genomic regions. Furthermore we focused on certain populations (European and African ancestry). Nevertheless, different regions (on chromosome 8 and 15), different

imputation programs, and different populations showed similar overall patterns, suggesting that our observations are relevant throughout the genome and across multiple populations.

In our masking process using only the 1000 Genomes reference data, the reference panel individuals were the same as the study sample individuals, and our masked SNPs are not limited to a SNP array, making our approach different from the two most common masking processes. One common masking method removes the genotypes for a portion of markers (e.g. 10%) found amongst the typed variants on a study sample SNP array. This method can provide accuracy comparisons only for SNPs on the array. Our approach is able to provide accuracy assessments for SNPs not on the array.

Another commonly used masking method is the "leave-one-out" masking of a comprehensively genotyped reference panel, in which one individual is imputed using the remaining reference panel members. Our study design differed from the leave-one-out method since all individuals in the reference panel and study sample were the same. Our approach was expected to give an upper bound on accuracy because of the ideal match between the reference and study sample; the "correct" genotype for each individual at each variant was present in the reference panel.

Our results provide further evidence that concordance rate inflates accuracy estimates particularly for rare and low frequency variants (Asimit & Zeggini, 2010; Lin et al., 2010). These observations highlight a need to account for chance agreement not only when assessing imputation accuracy, but also more broadly in other situations for which concordance is traditionally used to assess accuracy, such as checking genotype agreement across duplicate samples (Rogers, Beck, & Tintle, 2014; Truong, 2015). Concordance rate will always produce a

value greater than or equal to IQS due to their mathematical relationship (see Methods for proof).

IQS is important to consider, as it is designed to identify variants for which imputation accuracy is better than can be expected by chance; accordingly, other measures were generally more liberal in assigning high accuracy. Our analyses indicate that especially for rare and low frequency variants, IQS may be important to avoid overly liberal assessments of imputation quality. In practice, IQS can be computed by the leave-one-out method. Databases that provide per-SNP "imputability," such as that created by Duan et al. (Duan, Liu, Croteau-Chonka, Mohlke, & Li, 2013), would have increased usefulness if they included IQS values. As imputation methodology continues to develop and reference panels become more comprehensive, we expect that imputation will become increasingly accurate. However, it will be important to take chance agreement into account when assessing this accuracy, and IQS provides a means to do so.

Table 2.2: Sub-populations in the BEAGLE and IMPUTE2 AFR and EUR reference panels.

Table 3: Sub-populations in the BEAGLE and IMPUTE2 AFR and EUR reference panels

AFR: 246 individuals	EUR: 379 individuals
61 African Ancestry in Southwest US (ASW)	85 Utah residents (CEPH) with Northern and Western European ancestry (CEU)
97 Luhya in Webuye, Kenya (LWK)	89 British from England and Scotland (GBR)
88 Yoruba in Ibadan, Nigeria (YRI)	98 Toscani in Italia (TSI)
	93 Finnish from Finland (FIN)
	14 Iberian populations in Spain (IBS)

Table 2.3: Numbers of SNPs in the 1000 Genomes study samples. Study sample variants were those found on each commercially available SNP array for the 2 MB chromosomal regions of interest. Only variants with dbSNP identifiers are listed in the number of variants in the reference panel column.

Table 4: Numbers of SNPs in the 1000 Genomes study samples

Number of Genotyped SNPs in Each Region					
Chromosome	Omni 2.5M	Affy 500	Affy 6	Duo	Quad
8	1669	255	531	960	611
15	2740	555	1105	1231	1970

Table 2.4: Polymorphic, imputed SNPs used in the comparison of accuracy measures. These variants were found in the 2 MB chromosomal regions of interest using 1000 Genomes as the study sample and were imputed using Omni 2.5 coverage.

Table 5: Polymorphic, imputed SNPs used in the comparison of accuracy measures

	AFR	EUR
Chr 8	10,149	6,753
Chr 15	12,290	8,464

Figure 2.1: General process for creating the study sample for imputation. The reference panel was masked to mimic a commercial SNP array, resulting in a study sample which contains the same individuals as the reference panel.

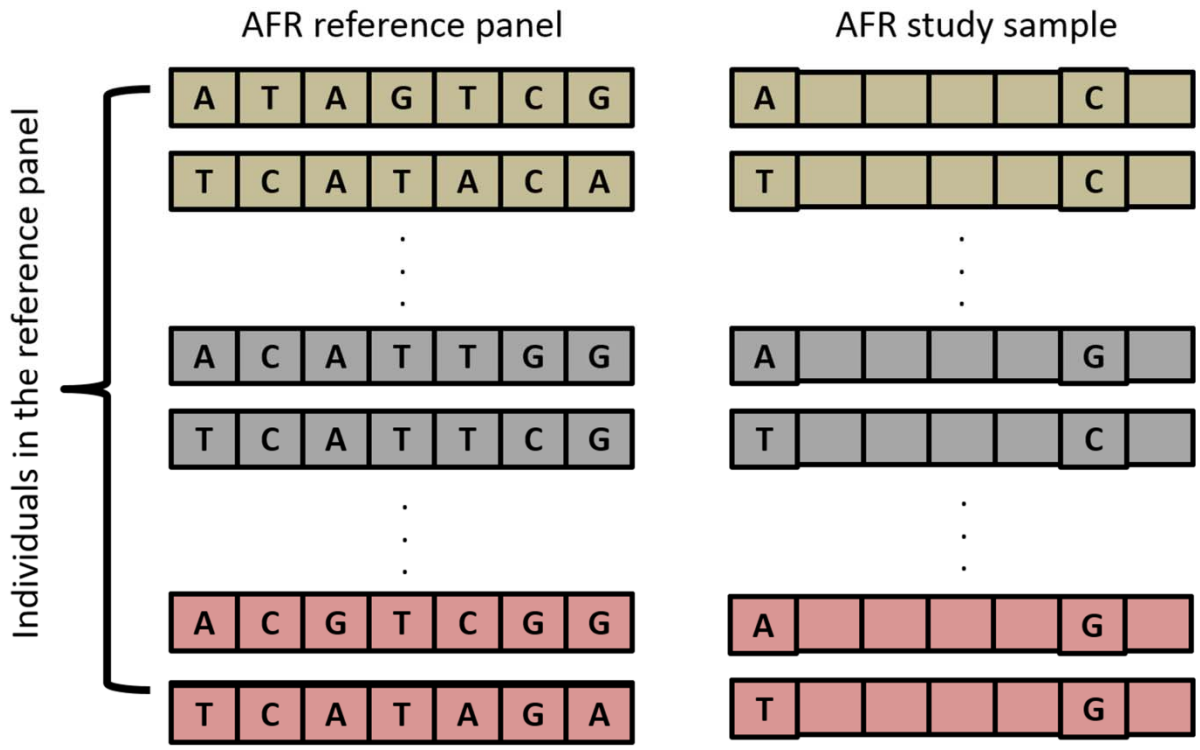

Figure 1: General process for creating the study sample for imputation

Figure 2.2: IQS, squared correlation, concordance rate, and BEAGLE R2 are shown in MAF bins. Mean accuracy of SNPs in each MAF bin (defined by 0.01 increments with N=13,442 variants total) is denoted by the red dots and the bars indicate one standard deviation (above and below the mean). These results are produced by using the 1000 Genomes AFR reference population as the study sample with Omni 2.5M typed coverage on chromosome 15.

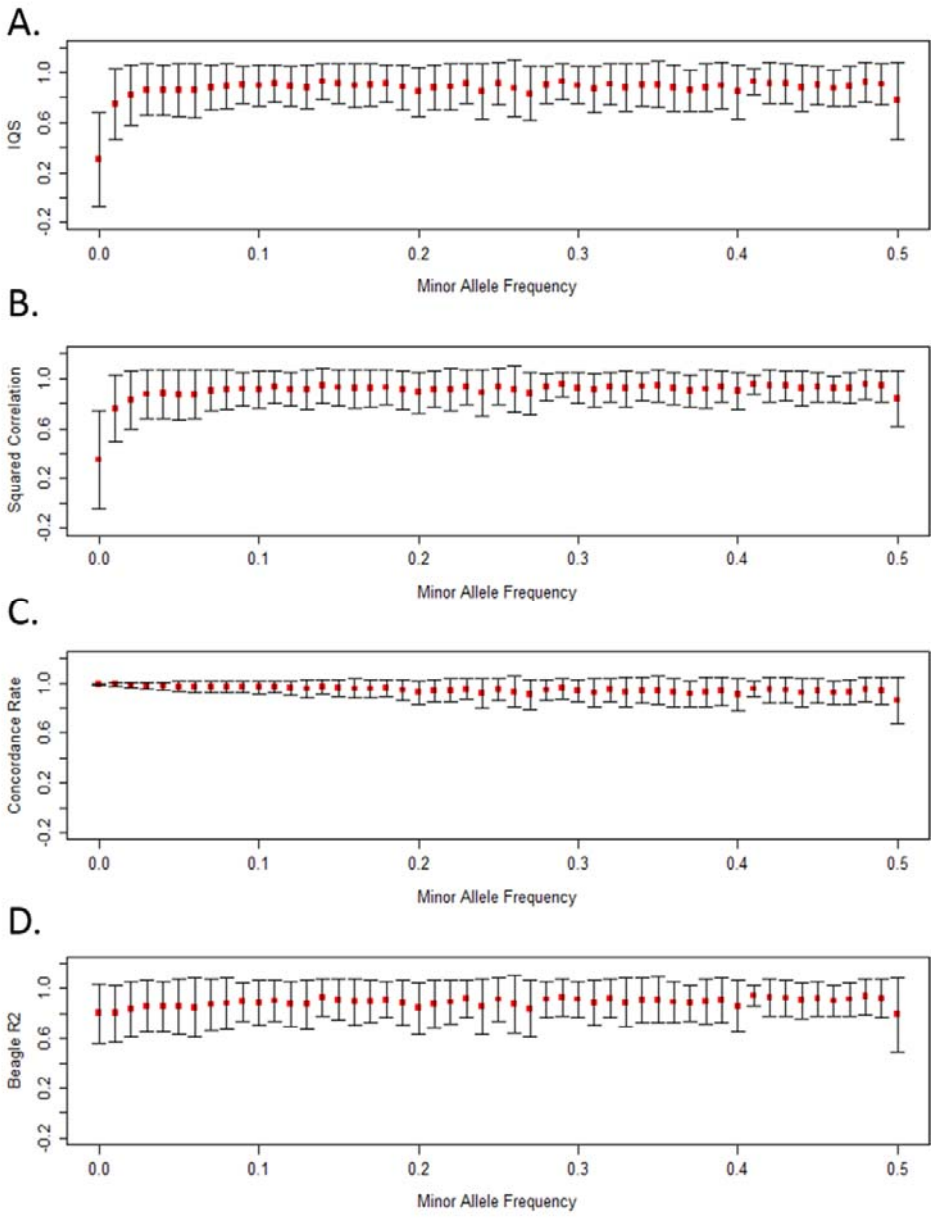

Figure 2: IQS, squared correlation, concordance rate, and BEAGLE R2 in MAF bins

Figure 2.3: IQS, squared correlation, concordance rate, and BEAGLE R2 are shown in max r2LD bins. Mean accuracy of SNPs in each MAF bin (defined by 0.01 increments with N=13,442 variants total) is denoted by the red dots and the bars indicate one standard deviation (above and below the mean). These results were produced by using the 1000 Genomes AFR reference population as the study sample with Omni 2.5M typed coverage on chromosome 15.

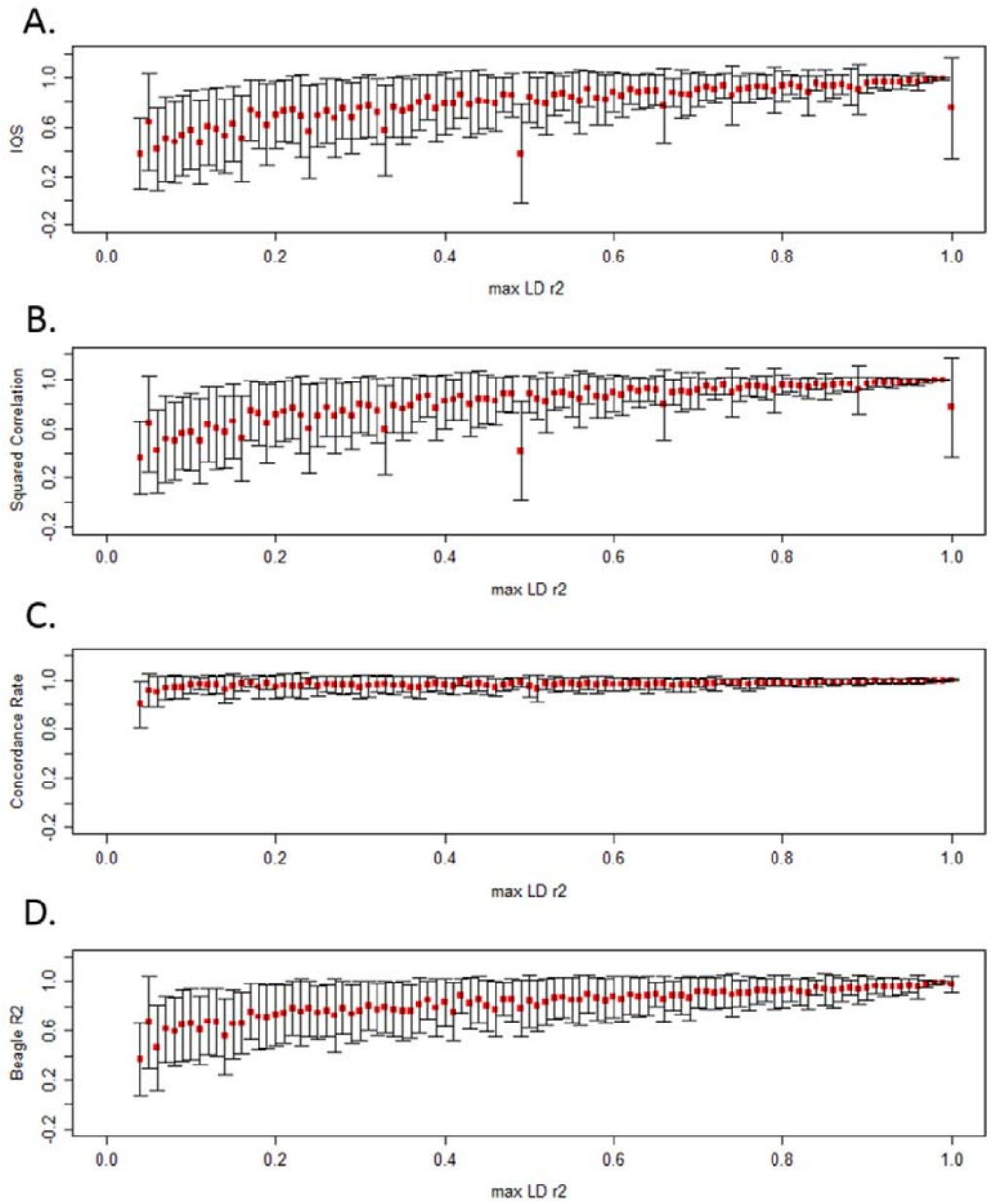

Figure 3: IQS, squared correlation, concordance rate, and BEAGLE R2 in max r2LD

Figure 2.4: Scatterplots of squared correlation and IQS. Data for all 13,442 variants are displayed in panel A, while the results for variants with MAF>5% (N = 6,480) are found in panel B. The line y = x is denoted in red.

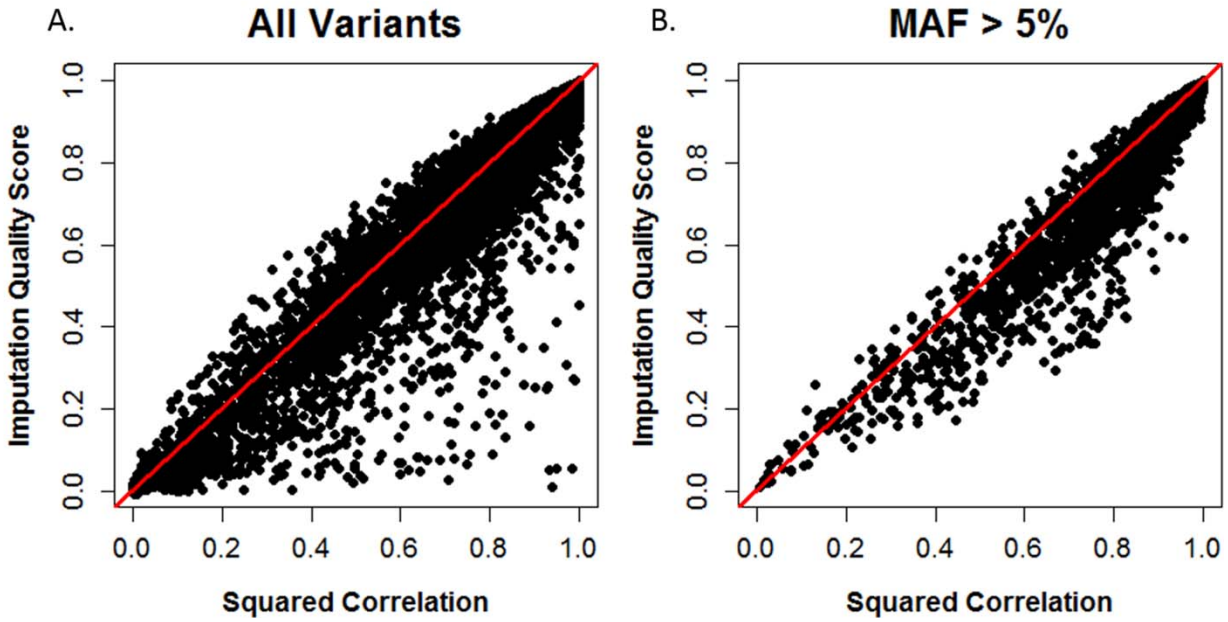

Figure 4: Scatterplots of squared correlation and IQS

Figure 2.5: Scatterplots of IQS, squared correlation, and BEAGLE R2. Panels A and B display all 13,442 variants, and panels C and D display variants with MAF>5% (N = 6,480). The line y = x is denoted in red.

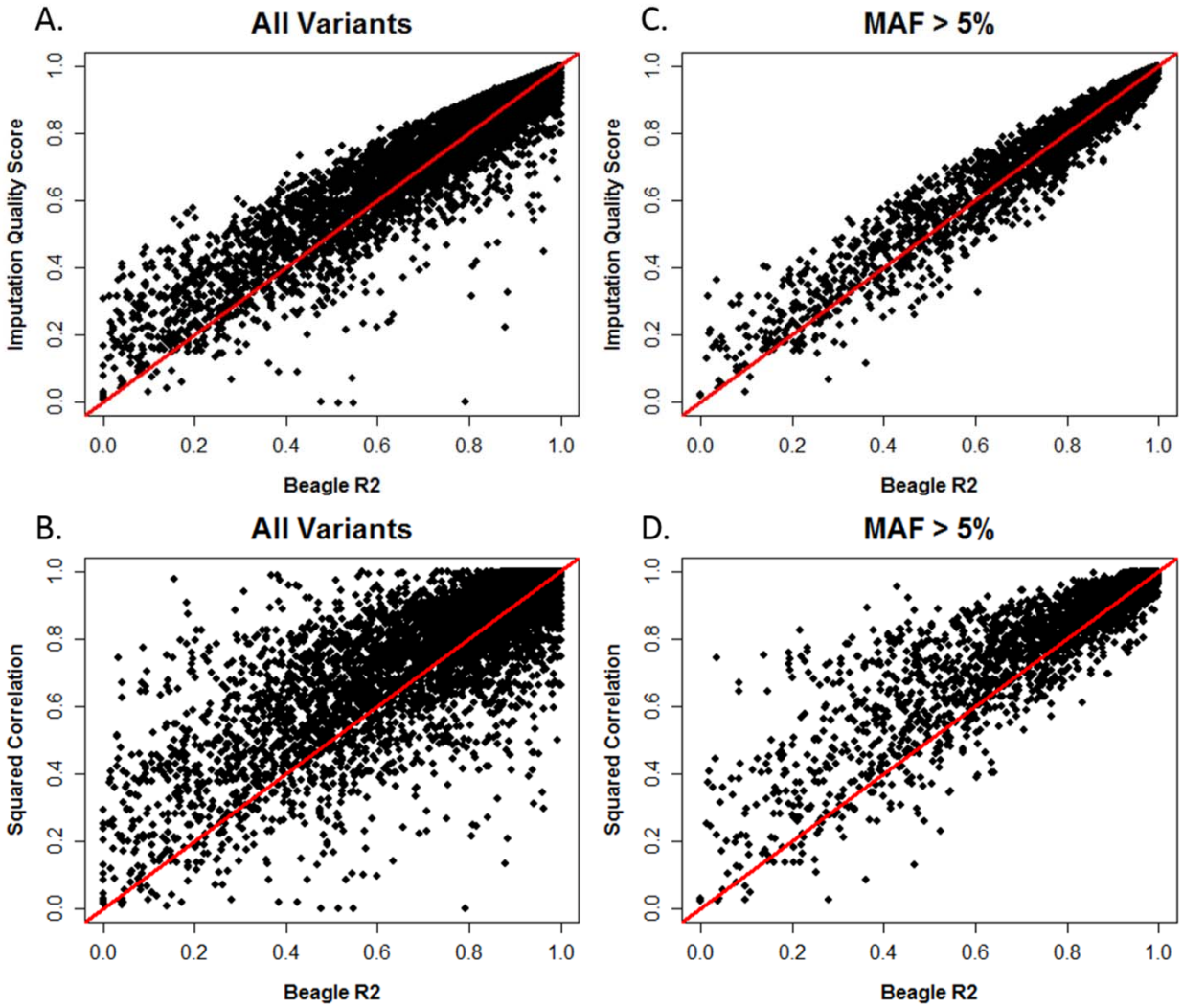

Figure 5: Scatterplots of IQS, squared correlation, and BEAGLE R2

Figure 2.6: Scatterplots of IQS, squared correlation, and BEAGLE R2 using the cosmopolitan reference panel and the African American nicotine dependence study sample for chromosome 15. Data for all 1,545 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 631) are found in panel D, E, and F. These results were generated using Omni SNP coverage. The line y = x is denoted in red.

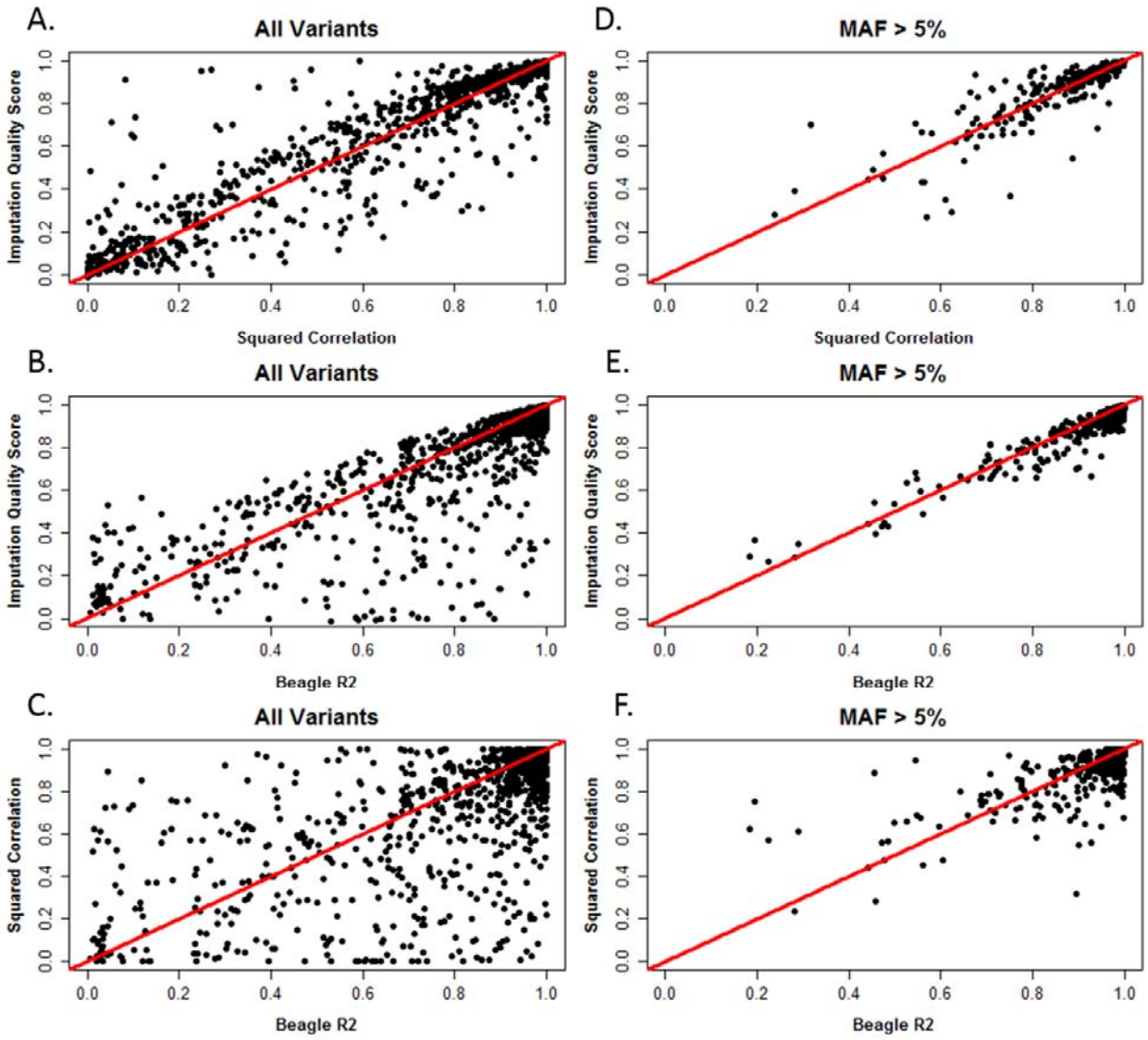

Figure 6: Scatterplots of IQS, squared correlation, and BEAGLE R2 using nicotine dependence samples

Figure 2.7: Mean numbers of polymorphic variants in each MAF (panel A) and max r2LD (panel B) bin. These results are for the AFR population on chromosome 15 (13,442 imputed SNPs).

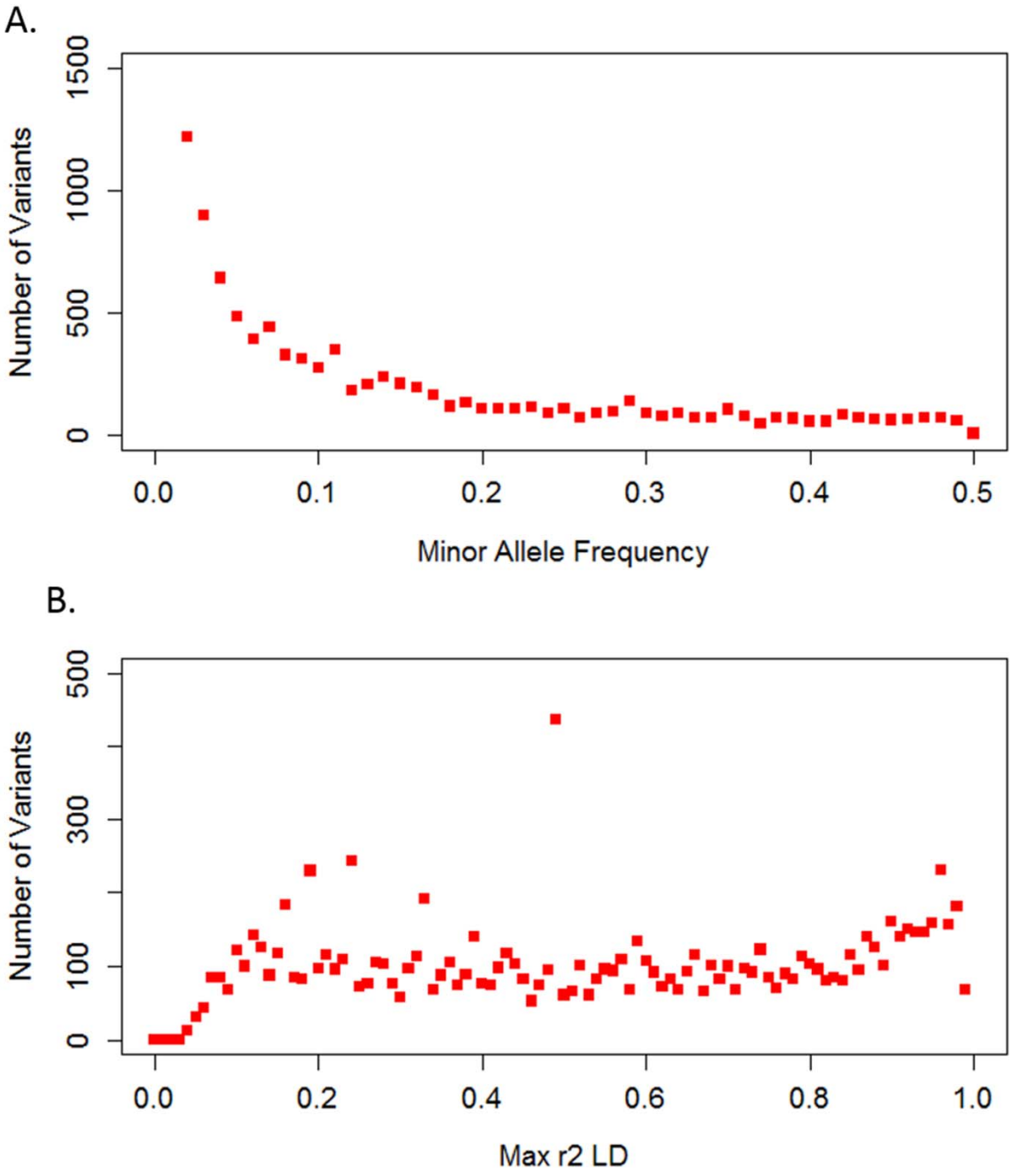

Figure 7: Mean numbers of polymorphic variants in each MAF and max r2LD bin

Figure 2.8: Average accuracy of all SNPs according to 0.01 incremental MAF bins for each accuracy measure using several typed SNP array coverages. These results were produced by using the 1000 Genomes AFR reference populations as the study samples for chromosome 15.

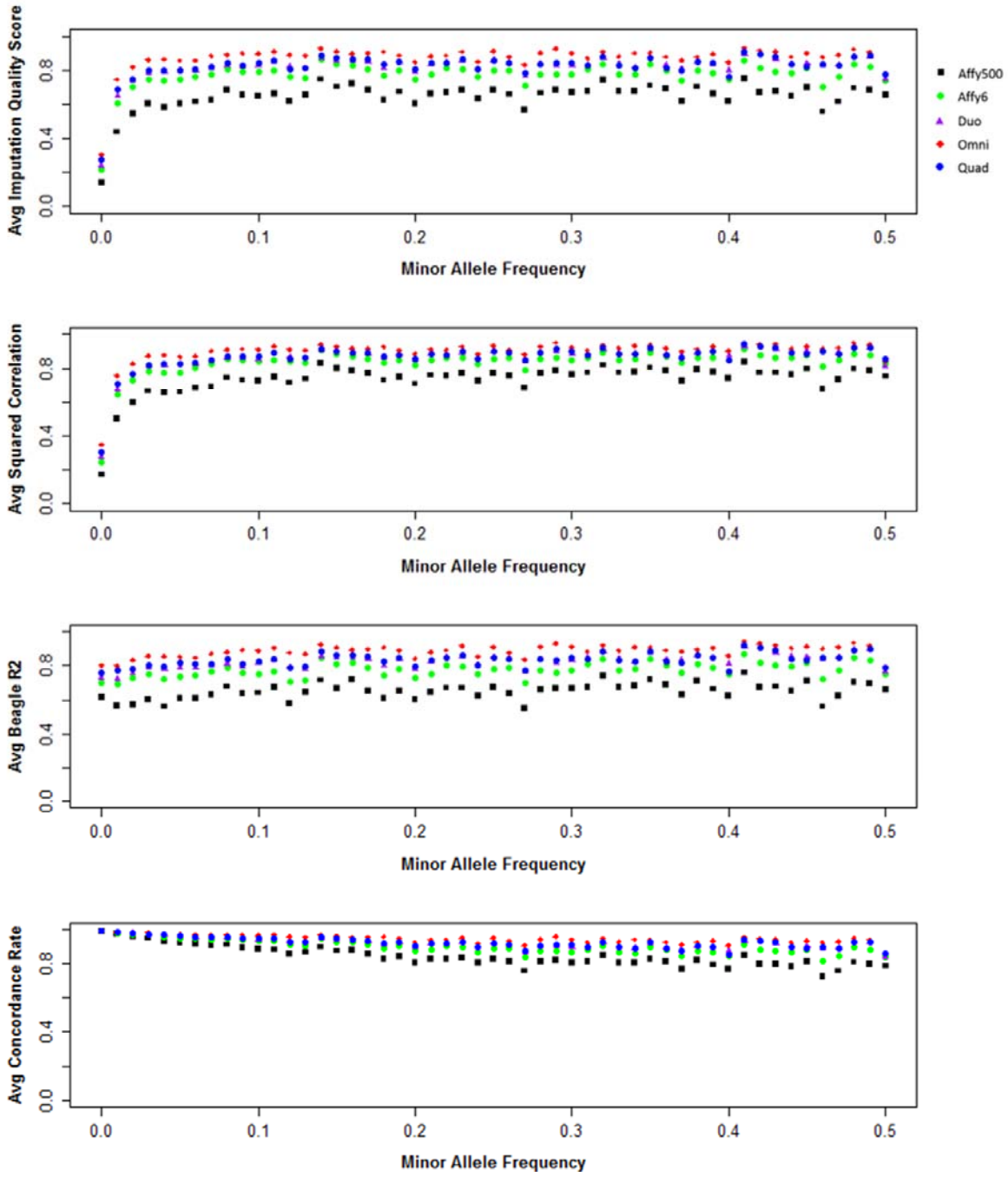

Figure 8: Average accuracy of all SNPs according to 0.01 incremental MAF bins for each accuracy measure using several typed SNP array coverages

Figure 2.9: Average accuracy of all SNPs in 0.01 incremental max r2LD bins for each accuracy measure using several typed SNP array coverages. These results were produced by using the 1000 Genomes AFR reference population as the study sample for chromosome 15.

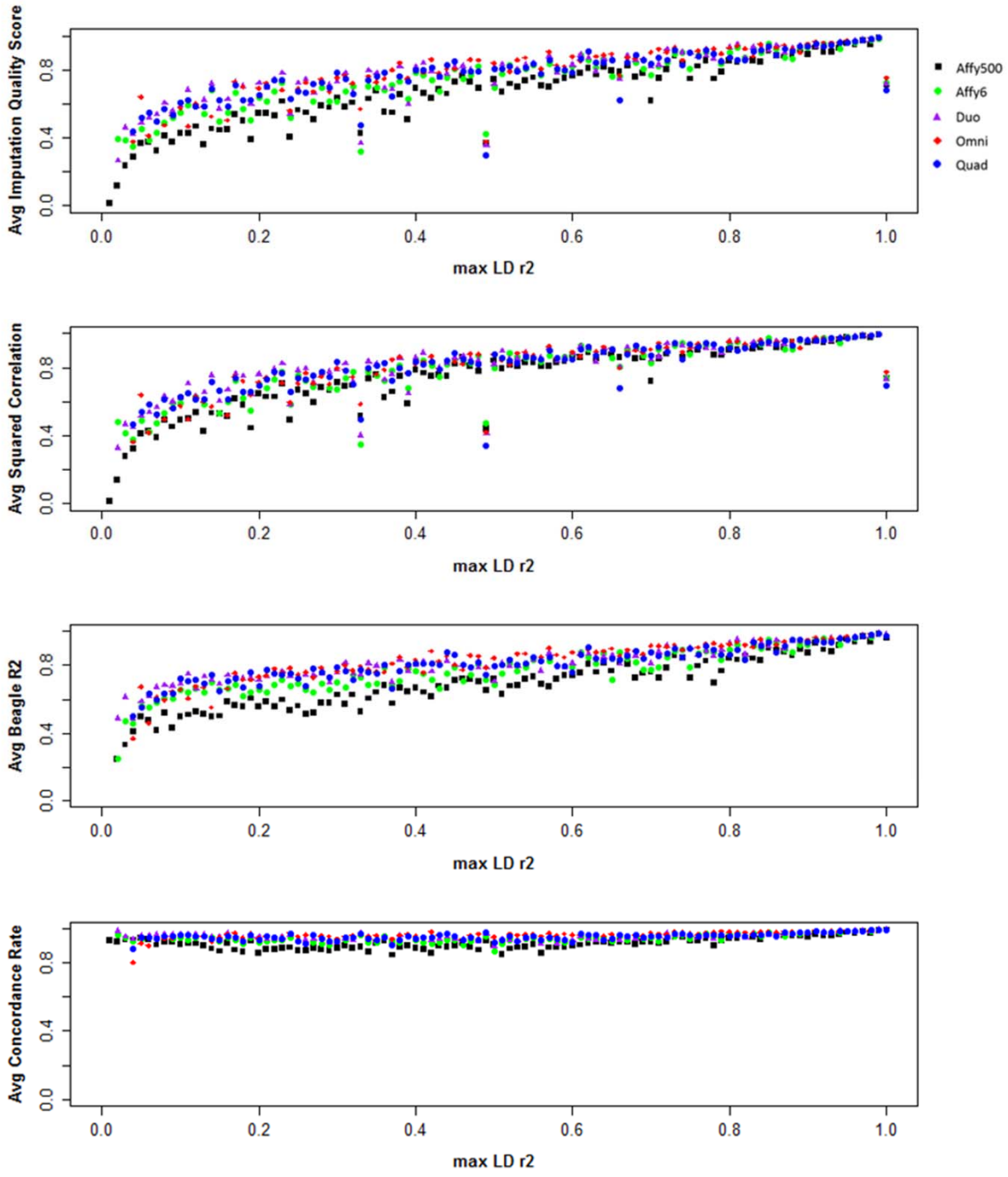

Figure 9: Average accuracy of all SNPs in 0.01 incremental max r2LD bins for each accuracy measure using several typed SNP array coverages

Figure 2.10: Accuracy scores produced by IQS, squared correlation, concordance rate and Beagle R2 for SNPs with MAF > 5% (N=6,480 SNPs) in max r2LD bins. Bins are defined by 0.01 increments. Mean accuracy is denoted by the red dots and the bars indicate one standard deviation (above and below the mean). These results were produced by using 1000 Genomes AFR reference population as the study sample with Omni 2.5M typed coverage on chromosome 15.

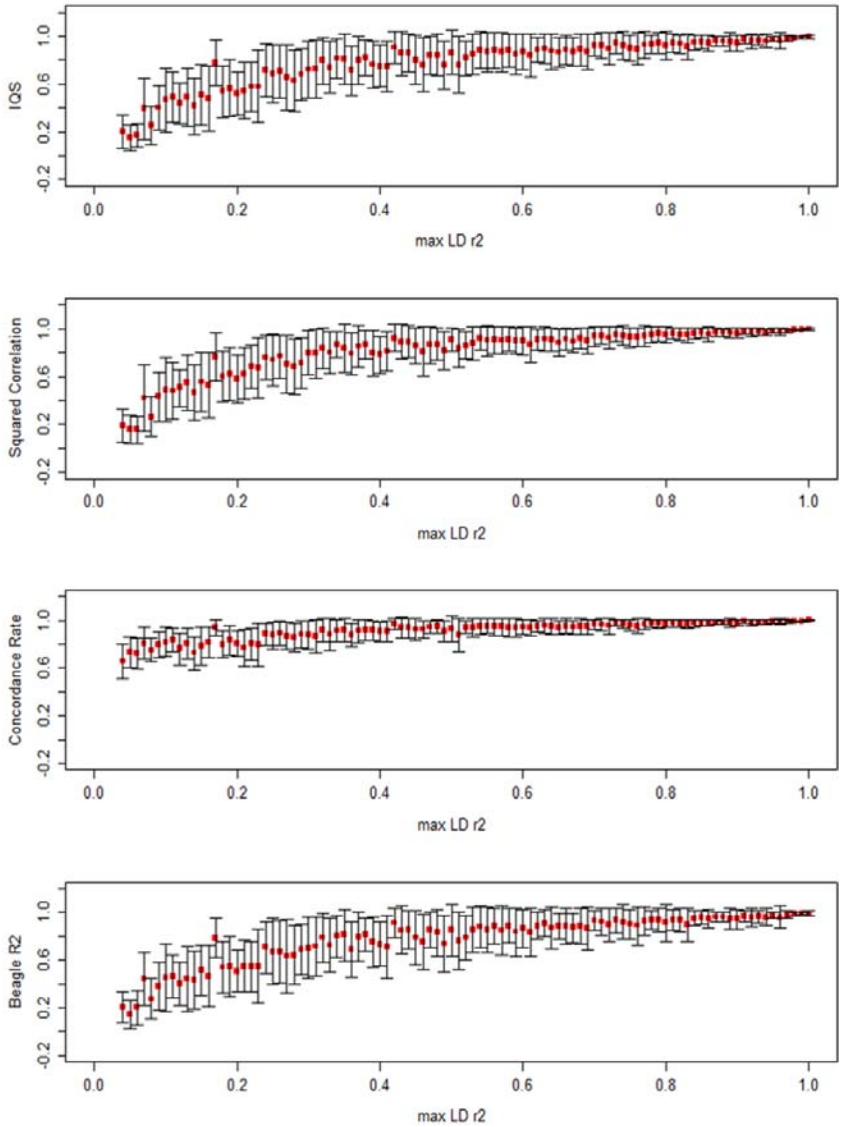

Figure 10: Accuracy scores produced by IQS, squared correlation, concordance rate and Beagle R2 for SNPs with MAF > 5% in max r2LD bins

Figure 2.11: Relationship between squared correlation and IQS by MAF. Squared correlation was subtracted from IQS for variants on chromosome 15 in the 1000 Genomes AFR reference population (N= 13,442 variants) as the study sample. Negative values indicate that the squared correlation score was higher while the positive values indicate that the IQS value was higher. The red line indicates the line y = 0.

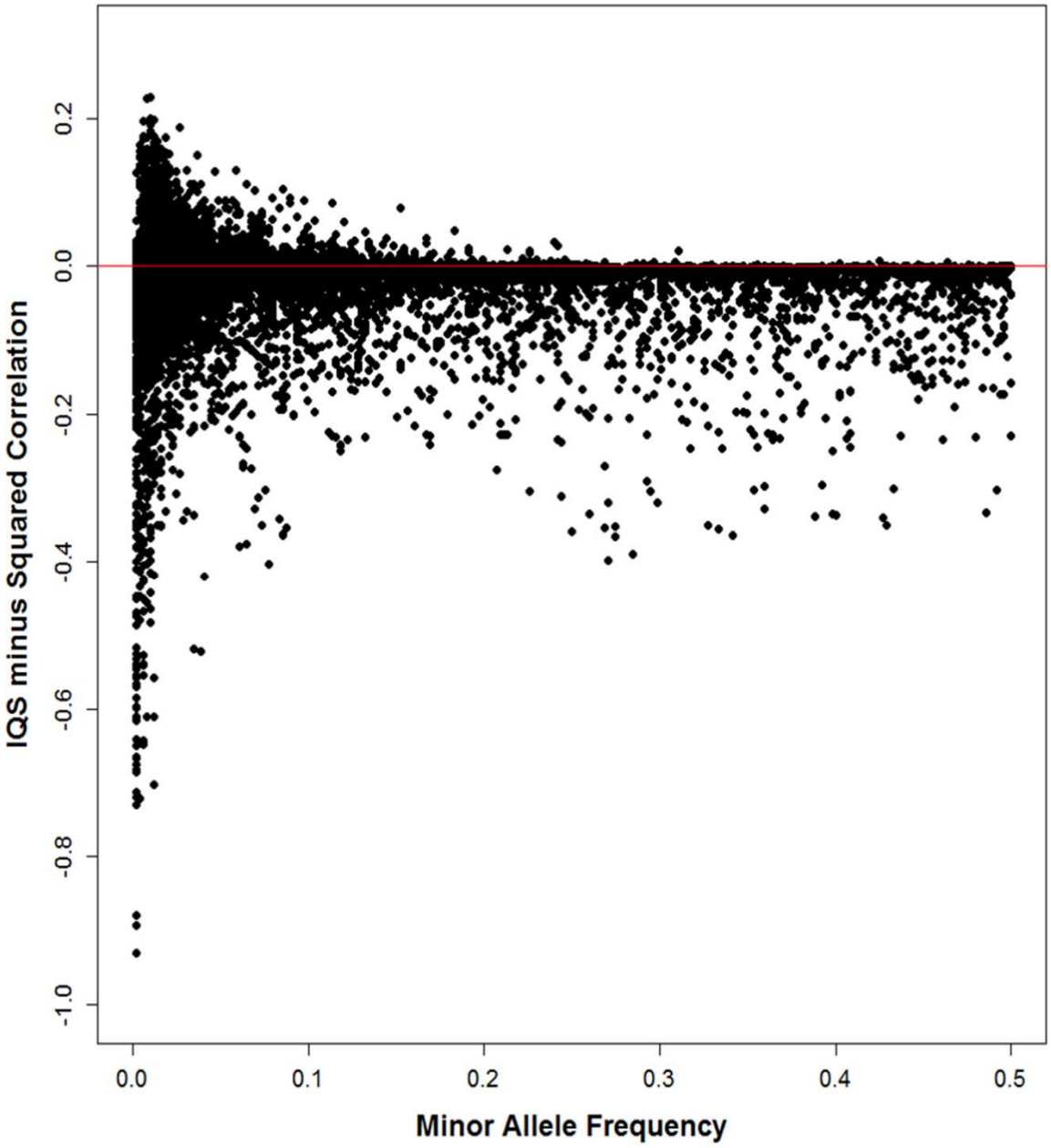

Figure 11: Relationship between squared correlation and IQS by MAF

Figure 2.12: Scatterplots of IQS, squared correlation, and BEAGLE R2 using the 1000 Genomes AFR reference panel as the study sample for chromosome 8. Data for all 10,937 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 4,533) are found in panel D, E, and F. These results were generated using Omni SNP coverage. The line y = x is denoted in red.

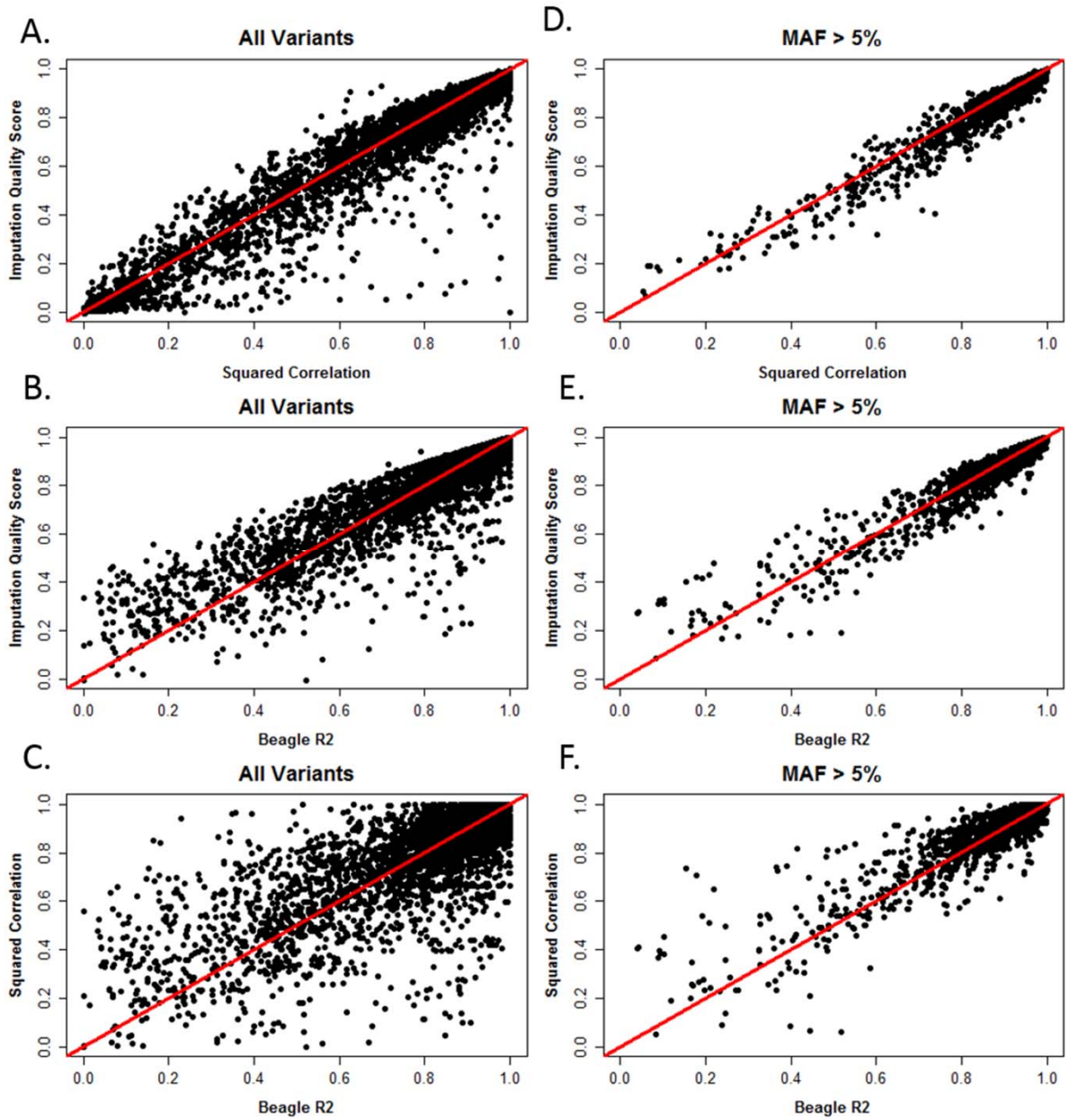

Figure 12: Scatterplots of IQS, squared correlation, and BEAGLE R2 using the 1000 Genomes AFR reference panel as the study sample for chromosome 8

Figure 2.13: Scatterplots of IQS, squared correlation, and BEAGLE R2 using the 1000 Genomes EUR reference panel as the study sample for chromosome 15. Data for all 9,401 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 4,627) are found in panel D, E, and F. These results were produced by using Omni SNP coverage. The line y = x is denoted in red.

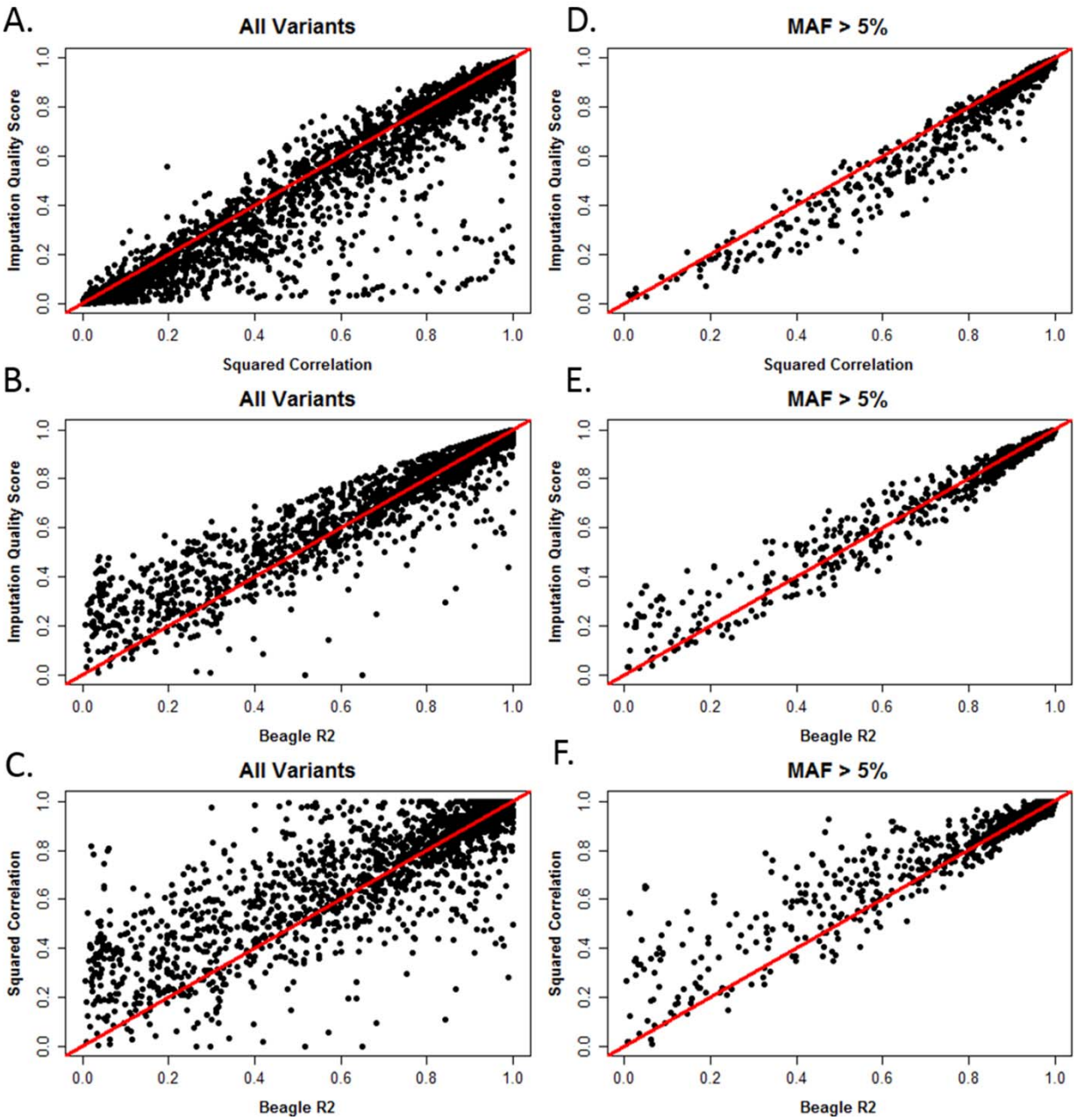

Figure 13: Scatterplots of IQS, squared correlation, and BEAGLE R2 using the 1000 Genomes EUR reference panel as the study sample for chromosome 15

Figure 2.14: Scatterplots of IQS, squared correlation, and BEAGLE R2 using the 1000 Genomes EUR reference panel as the study sample for chromosome 8. Data for all 7,401 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 1,903) are found in panel D, E, and F. These results were produced by using Omni SNP coverage. The line y = x is denoted in red.

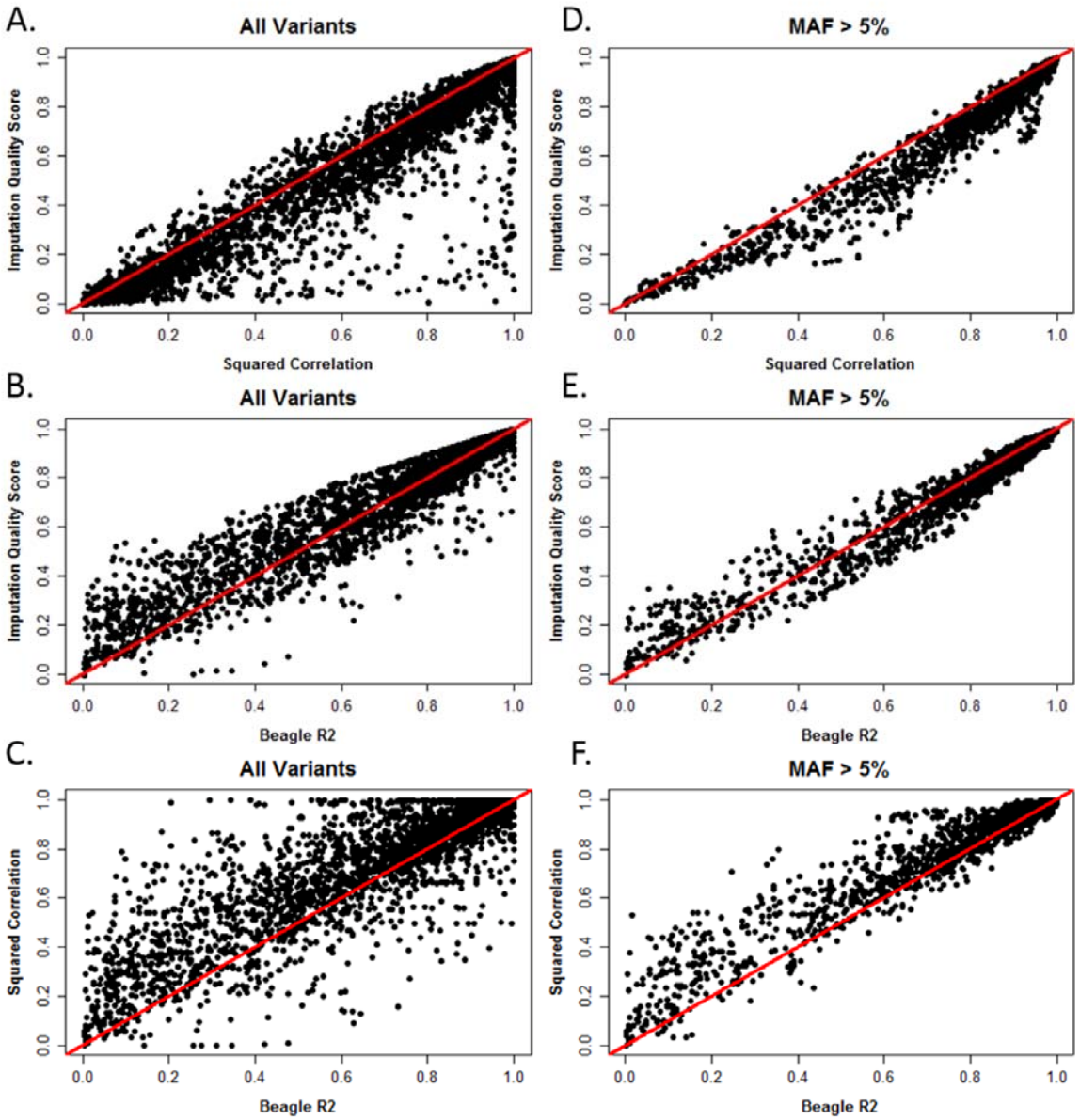

Figure 14: Scatterplots of IQS, squared correlation, and BEAGLE R2 using the 1000 Genomes EUR reference panel as the study sample for chromosome 8

Figure 2.15: Scatterplots of IQS, squared correlation, and BEAGLE R2 using the cosmopolitan reference panel and the European American nicotine dependence study sample for chromosome 15. Data for all 1,170 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 387) are found in panel D, E, and F. These results were produced by using Omni SNP coverage. The line y = x is denoted in red.

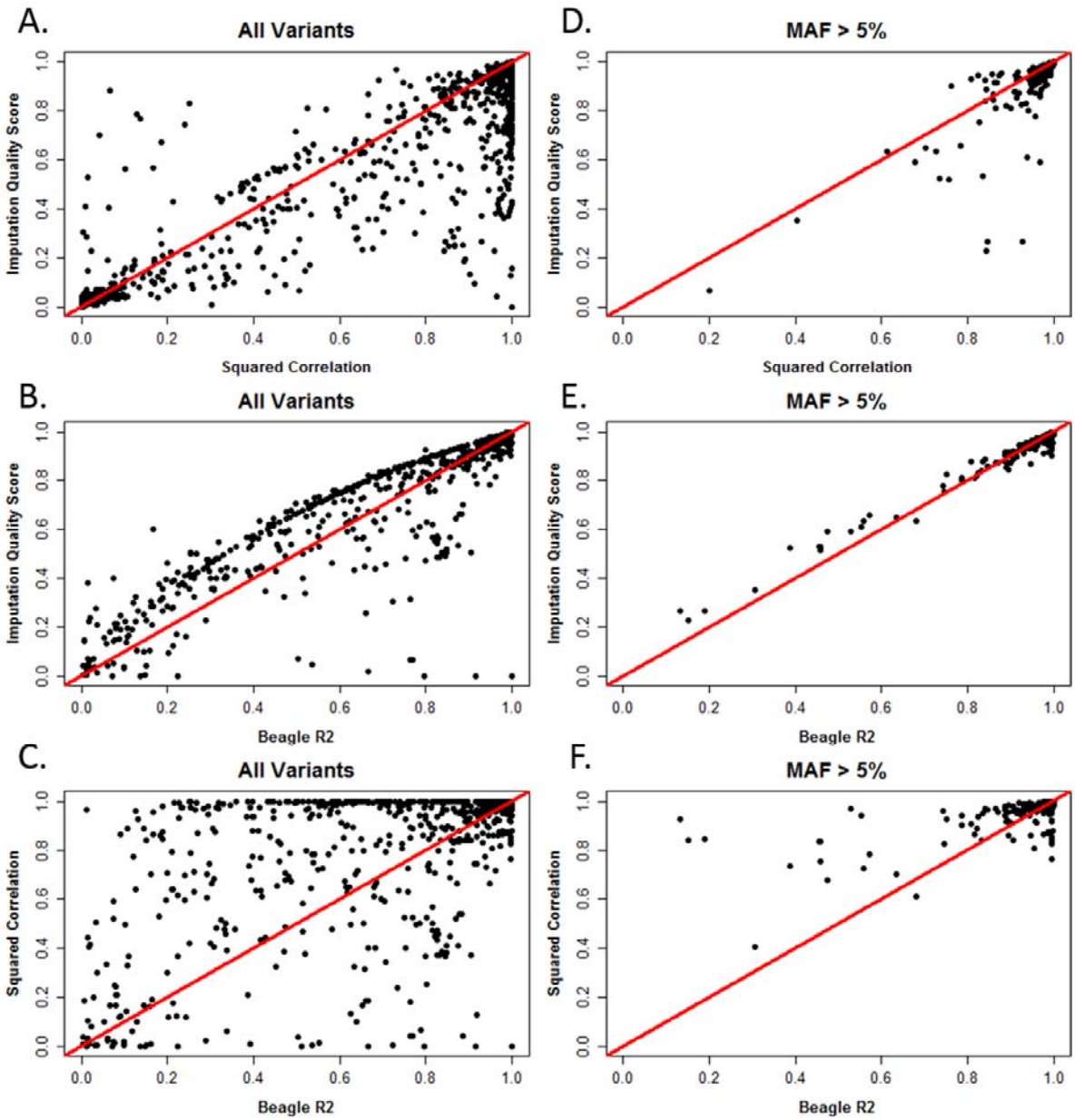

Figure 15: Scatterplots of IQS, squared correlation, and BEAGLE R2 using the European American nicotine dependence study sample for chromosome 15

Figure 2.16: Scatterplots of IQS, squared correlation, and IMPUTE2 INFO using the cosmopolitan reference panel and the African American nicotine dependence study sample for chromosome 15. Data for all 1,878 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 475) are found in panel D, E, and F. These results were generated using Omni SNP coverage. The line y = x is denoted in red.

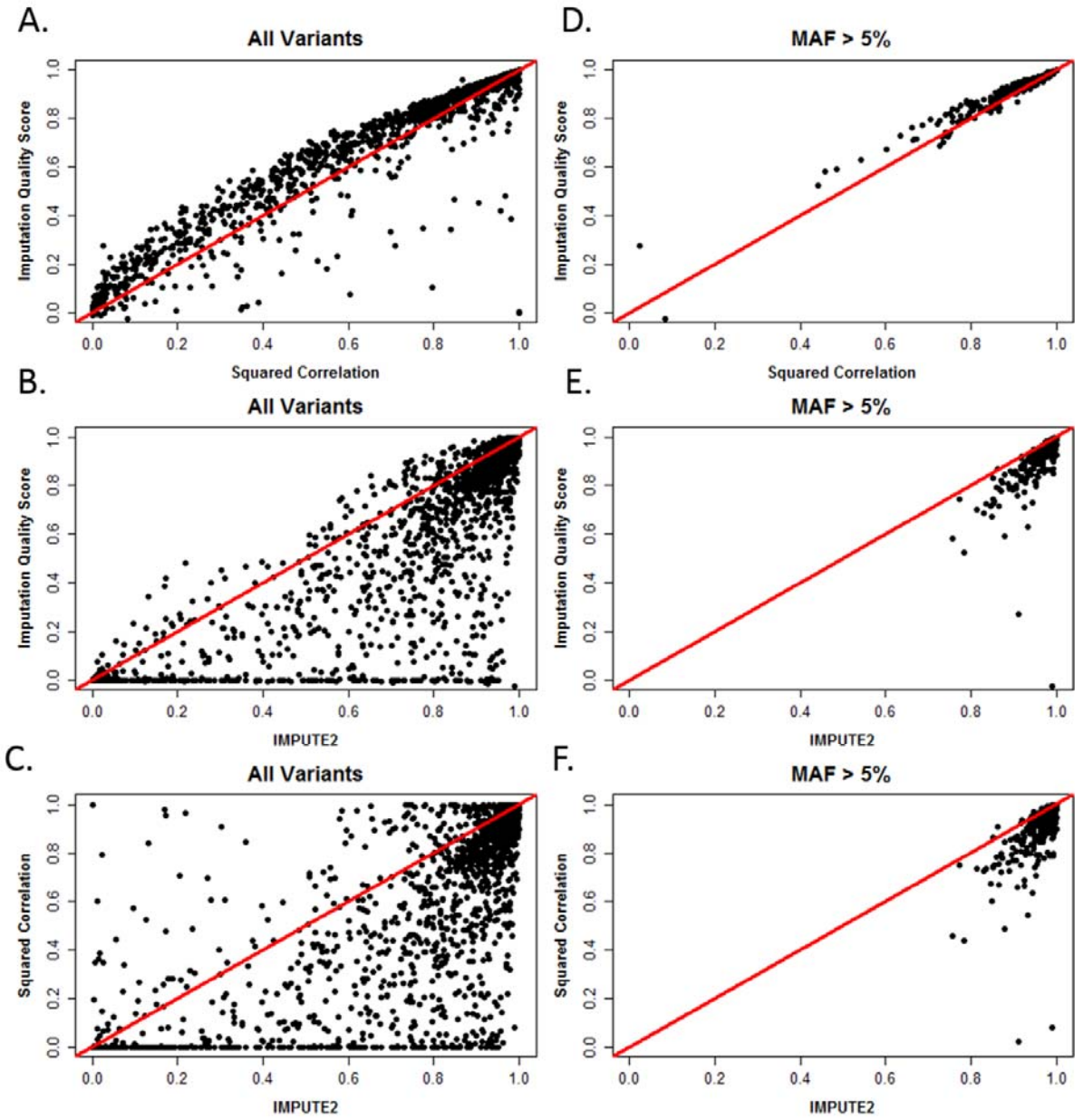

Figure 16: Scatterplots of IQS, squared correlation, and IMPUTE2 INFO using the African American nicotine dependence study sample for chromosome 15

Figure 2.17: Scatterplots of IQS, squared correlation, and IMPUTE2 INFO using the cosmopolitan reference panel and the European American nicotine dependence study sample for chromosome 15. Data for all 1,253 variants are displayed in panel A, B, and C while the results for variants with MAF>5% (N = 259) are found in panel D, E, and F. These results were generated using Omni SNP coverage. The line y = x is denoted in red.

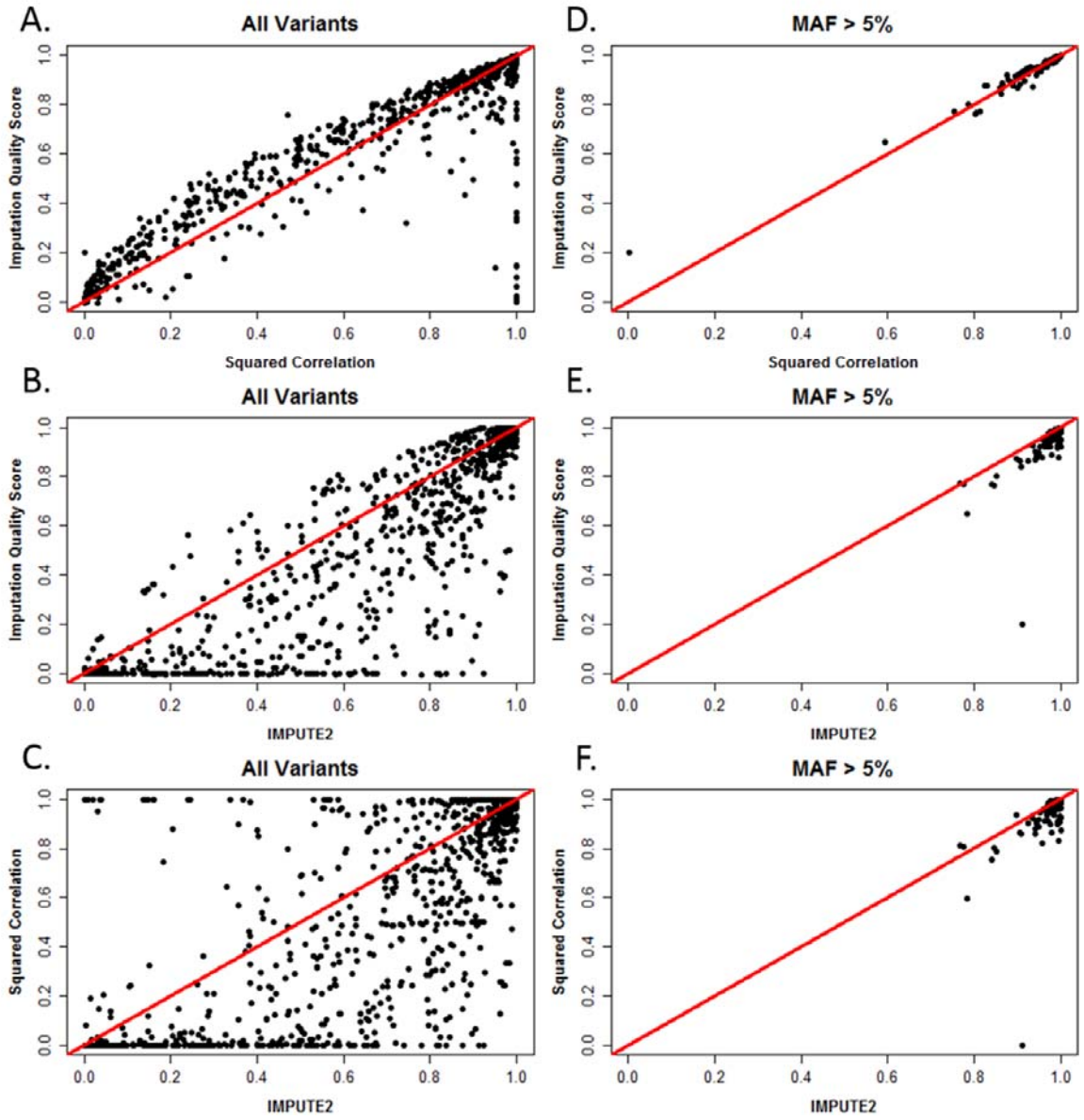

Figure 17: Scatterplots of IQS, squared correlation, and IMPUTE2 INFO using the European American nicotine dependence study sample for chromosome 15

Chapter 3: Assessing Genetic Influences on Mentholated Cigarette Preference in Nicotine Dependent Smokers

3.1 Introduction

Cigarette smoking remains the single greatest preventable cause of premature death in the US and world-wide (Benowitz, 2010). Mentholated cigarettes are widely available, with one in four cigarettes technically classified as mentholated (TPSAC;(Giovino et al., 2004). Furthermore, mentholated cigarette use varies between African American and European American smokers, with much higher rates in African Americans: 70% compared to 30% (Gardiner, 2004). Most studies of mentholated cigarette use are focused on the influence of social, demographic, or advertising factors. However, since genetic studies have been highly successful in identifying genetic contributions to nicotine dependence and smoking heaviness (Bierut et al., 2007; David et al., 2012; Scott F. Saccone et al., 2007; Tobacco and Genetics Consortium, 2010), we will investigate potential genetic influences on mentholated cigarette use.

Hypothesis driven candidate regions related to taste, nicotine dependence and nicotine metabolism may be associated with mentholated cigarette preference. The *TAS2R38* gene is a bitter taste receptor which accounts for 85% of the variability in bitter taste (TPSAC;(Mangold, Payne, Ma, Chen, & Li, 2008)). Variants in this gene can be used to classify individuals as tasters or non-tasters (Kim, Breslin, Reed, & Drayna, 2004; Mangold et al., 2008). Previous work has also shown that tasters are less likely to become heavy smokers (TPSAC;(Mangold et al., 2008)). Furthermore, non-tasters were associated with increased susceptibility to nicotine dependence (Mangold et al., 2008). However, the relationship between genetic variants in this gene and

mentholated cigarette use is unknown. The genetic variants most strongly associated with nicotine dependence are located in nicotinic receptors (Bierut, 2010; Bierut et al., 2007; Chen et al., 2014; Chen et al., 2012; S. F. Saccone et al., 2007). We aim to examine variants in nicotinic receptors especially those previously associated with nicotine dependence: *CHRNA5/A3/B4* and *CHRNB3/A6*.

CYP2A6 is associated with nicotine dependence (Chen et al., 2014) and is one of the key genes involved in nicotine metabolism (Benowitz, 2010). In humans, most nicotine, about 75%, is converted to cotinine through *CYP2A6* with 33 to 40% of the cotinine then converted to trans-3-hydroxycotinine which is excreted by the liver (Benowitz, 2010). African Americans covert nicotine to cotinine and metabolize cotinine more slowly than European Americans (Benowitz, Herrera, & Jacob, 2004; Benowitz et al., 1999). African Americans have higher concentrations of serum cotinine, a metabolite of nicotine, than whites (MacDougall, Fandrick, Zhang, Serafin, & Cashman, 2003; Pérez-Stable, Herrera, Jacob, & Benowitz, 1998). Menthol has been shown to slow the oxidation of nicotine to cotinine (Benowitz et al., 2004; MacDougall et al., 2003). These factors may work in aggregate in African American smokers; to this end, testing genetic variants in *CYP2A6* for association with mentholated cigarette use may elucidate this relationship.

We will use data for African Americans (N = 1,365) and European Americans (N = 2,206) in attempts to identify genetic variants that increase susceptibility to mentholated cigarette use. We begin in hypothesis driven candidate regions then expand to genome-wide analyses. These analyses will identify variants that are consistent or unique in African Americans and European Americans.

3.2 Methods

Several cohorts were used in these analyses: The Collaborative Genetic Study of Nicotine Dependence (COGEND) and Transdisciplinary Tobacco Use Research Center (UW-TTURC). We focus our analyses on nicotine dependent current smokers, since these individuals are at highest risk for health consequences. Nicotine dependent cases are defined as smokers with Fagerström Test for Nicotine Dependence (FTND) >= 4.

3.2.1 The Collaborative Genetic Study of Nicotine Dependence (COGEND)

3.2.1.1 Part 1

COGEND is a cross-sectional study of extensive smoking behavior phenotypes in African Americans and European Americans (Bierut et al., 2007) from Saint Louis, Missouri Detroit, Michigan, and Minneapolis, Minnesota. The study protocol was approved by the Institutional Review Boards at each site. All participants provided informed consent. Participants were recruited from 2002-2007. These individuals were between the ages of 25-44 years old and were assessed for dependence as measured by FTND and Cigarettes-per-day (CPD) (Luo et al., 2008). COGEND subjects were ascertained based on having smoked at least 100 cigarettes lifetime. Eligibility was based on FTND score, with nicotine dependent cases and non-dependent smoking controls defined by a FTND of 0 or 1. Mentholated cigarette use was measured by the question “Do you usually smoke mentholated cigarettes?”

From COGEND, there were 1,406 participants genotyped on the Illumina Human1M-Duo BeadChip and 1,480 COGEND participants genotyped on the Illumina HumanOmni2.5 BeadChip

(Laurie et al., 2010). PLINK was used for all SNP-level and participant-level QC steps (Purcell et al., 2007). In each subset, genotyped SNPs with call rate≥97% and Hardy Weinberg equilibrium (HWE) $P \geq 1x10^{-4}$ were retained using PLINK. This sample was imputed using only the genotyped SNPs in the intersection of two arrays to avoid bias in the association results (Johnson et al., 2013).

For imputation, there were 605,735 genotyped SNPs which passed initial SNP-level QC and were imputed using IMPUTE2 (B. Howie et al., 2012; B. Howie et al., 2011; Marchini & Howie, 2010). The imputation was performed on each ethnic group separately using 1,933 EAs (1,000 ND cases and 933 controls) and 704 AAs (459 ND cases and 245 controls) using the 1000 Genomes ALL reference panel.

3.2.1.2 Part 2

COGEND was expanded and additional recruitment was approved by the Institutional Review Board at Washington University prior to enrolling participants. All participants provided informed consent. Participants were recruited from the St. Louis metropolitan area from 2011-2014. The inclusion criteria remained the same for this additional recruitment. Mentholated cigarette use was measured by the question "Do you usually smoke mentholated cigarettes?"

This second part of COGEND was genotyped using a custom Illumina array combining the coverage of the Illumina HumanOmniExpress and the Illumina Exome arrays. The cleaned, analysis ready dataset of 2514 subjects and 950,847 variants required call rate ≥ 98% for variants and ≥ 99% for subjects. Gender and race/ethnicity checks were performed using Plink(Purcell et al., 2007) and Eigenstrat (Price et al., 2006), and sample-chromosome

combinations with chromosomal anomalies were removed. Filters were imposed for Hardy Weinberg equilibrium p-value < 10×10^{-4} in controls.

Imputation was performed by first pre-phasing with SHAPEIT (B. Howie et al., 2012) and then imputing with IMPUTE2. The 1000 Genomes ALL reference panel was used, following the same protocols as for COGEND Part 1.

3.2.1.3 Part 3

COGEND was extended with new recruitment that also included assessment of exhaled carbon monoxide level as a biomarker of smoking. Institutional Review Board approval was obtained at Washington-- University prior to enrolling participants, and all participants provided informed consent. Participants were recruited from the St. Louis metropolitan area between 2014 and 2015. All participants were current smokers as demonstrated by an exhaled carbon monoxide level $\geq$ 7 parts per million and self-reported smoking on $\geq$ 15 days during the past month. Participants were required to have smoked 100 cigarettes lifetime and be between the ages of 25-44 years.

Participants provided saliva samples for genetic analysis using 23andMe DNA collection kits. 23andMe is a privately held personal genomics and biotechnology company that produces high quality genetic data in CLIA-certified laboratories. The success rate of genotyping submitted saliva samples was 97%. In addition to data cleaning performed by 23andMe, we performed additional checks including individual sample quality, SNP quality, Hardy-Weinberg Equilibrium (HWE), duplicates, and relatedness across participants. We required at least a 98% call rate across all SNPs for a sample to be included in analyses. At a SNP level, we required at

least a 98% call rate for each SNP in the sample. Relatedness across participants was examined to make sure that our participants were independent. Setting thresholds of per sample call rate of 98%, minor allele frequency of 1% or greater, and HWE p value more than $10^{-10,}$, we had a final SNP set of 488,487 variants with a mean call rate of 99.89% per sample and average 99.79% call rate per SNP.

Imputation was performed using the same protocols as COGEND Parts 1 and 2. SHAPEIT was used for pre-phasing and IMPUTE2 for imputation, with the 1000 Genomes ALL panel.

3.2.2 University of Wisconsin Transdisciplinary Tobacco Use Research Center (UW-TTURC)

A randomized placebo-controlled smoking cessation trial was conducted at University of Wisconsin Transdisciplinary Tobacco Use Research Center (UW-TTURC) with individuals aged 18 years or older who smoked 10 or more CPD and were motivated to quit smoking. This study was approved by the University of Wisconsin-Madison IRB. Mentholated cigarette use was measured in UW-TTURC using the question "Do you smoke menthol cigarettes?"

The Center for Inherited Disease Research at Johns Hopkins University performed the genotyping of the UW-TTURC sample using the Illumina Omni2.5 microarray and the GENEVA Coordinating Center at the University of Washington led the data cleaning.

Imputation was performed for all autosomes and X chromosome using IMPUTE2 software with reference panels from the 1000 Genomes Project with reference to the February 2012 release of the 1000 Genomes ALL haplotype panel release version 3 (available at http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html). The

imputation was conducted separately by race. The study genotypes were pre-phased across whole chromosomes using the SHAPEIT2 program.

3.2.3 Statistical Analyses

We report results for imputed and genotyped variants present in all population specific analyses and have MAF ≥ 1% as well as an Info score ≥ 0.3 as reported from IMPUTE2. Furthermore, this study focused on the nicotine dependent current smokers for whom mentholated cigarette use was available. In the African American analyses, there were 14,445,263 variants in COGEND Part 1; 13,588,781 variants in COGEND Part 2; 14,268,121 variants in COGEND Part 3; and 14,189,019 variants in UW-TTURC. In the European American analyses, there were 8,045,217 variants in COGEND Part 1; 7,204,743 variants in the COGEND Part 2; 8,016,251 variants in COGEND Part 3; and 8,025,424 variants in UW-TTURC.

Logistic regression was used to test genetic variants for association with mentholated cigarette use (yes or no), in hypothesis-driven candidate regions as well as genome-wide. These analyses were conducted in PLINK (Purcell et al., 2007). For each cohort, the data were subset by race. Age, sex and the first two principal components were included as covariates in the individual study analyses. A meta-analysis, stratified by race, was also conducted to identify variants most strongly associated in each ancestral population: 1,365 African Americans and 13,460,165 variants as compared to 2,206 European Americans and 14,016,119 variants. We also conducted a meta-analysis that combined both races. The African, and European American combined meta-analysis consisted of 3,571 individuals and 6,281,925 variants. Manhattan and QQ plots were generated using R using the qqman package (Turner, 2014).

3.2.4 Hypothesis-driven Candidate Regions

There are three variants that define the haplotypes in the taste receptor, TAS2R38: rs713598, rs1726866, and rs10246969 (Mangold et al., 2008; Wang et al., 2007). This gene is located on chromosome 7. Of these three variants, rs1726866 distinguishes the taster haplotype from the non-taster haplotype.

We focused on specific genomic regions associated with nicotine dependence, smoking behavior, and nicotine metabolism: nicotinic receptor subunit gene clusters on chromosome 8p11 (*CHRNB3-CHRNA6*), chromosome 15q25 (*CHRNA5-CHRNA3-CHRNB4*), and chromosome 19q13 (*CYP2A6-CYP2A7*) (Bierut et al., 2007; David et al., 2012; S. F. Saccone et al., 2007; Tobacco and Genetics Consortium, 2010). We examined the relationship between all variants in these receptors and mentholated cigarette use. Additionally, within these gene clusters, we examined specific variants previously associated with nicotine dependence, including the *CYP2A6* gene because of the potential relationship of menthol use and nicotine metabolism.

3.2.5 Power Analysis

Power for Genetic Association Analyses (PGA) was used to estimate power to detect a genetic effect for a range of allele frequencies and effect sizes (Menashe, Rosenberg, & Chen, 2008). We estimated power for each of the race specific analyses as well as the combined race analysis. The African Americans power analysis contained 1,365 individuals of which 1,305 were cases. There were 2,206 European Americans of which 737 were cases. The combined analysis contained 3,571 individuals of which 2,042 were cases. Since most of the African Americans

were cases (that is, menthol cigarette users), there were only a limited number of controls for comparison, thus limiting the interpretation and power of African-American-only analyses.

3.3 Results

Here we present the preliminary results.

Table 3.1 contains demographic information about the COGEND and UW-TTURC cohorts. Overall, these studies are fairly similar in demographics. Table 1 shows that mentholated cigarette use is 90-97% in African Americans as compared to 30-39% in European Americans. This influences the case-control ratio in these samples. Because so few African American smokers are non-menthol smokers, conclusions about the influence of genetic variants on mentholated cigarette use are most appropriate in the European American or combined ancestry analyses. Mentholated cigarette use amongst African Americans in these studies is higher than in population samples while the 30% mentholated cigarette use in European Americans is consistent with population samples (Gardiner, 2004). The percentage of females varies in each study. This is of note since there have been studies showing an effect of sex on mentholated cigarette use in which women were shown to smoke mentholated cigarettes at a higher rate than men (TPSAC).

Figure 3.1 displays several power analyses. We estimated the effect size (relative risk (RR)) that is detectable with 80% power at a genome-wide significance level of 5×10^{-8}. We assumed an additive (1df) model and a range of allele frequencies. The prevalence varies based on which group is being examined: African Americans 0.70, European Americans 0.30, and combined ancestry 0.453 which is a weighted prevalence. Power varies by population group

and is most challenging for lower allele frequencies. For an allele with frequency 0.05, in the full, combined ancestry sample, we have power to detect a locus with RR = 1.4 (Figure 3.1, Combined Ancestry); in the two ancestry specific samples, we can detect down to RR = 1.68 in European Americans and only RR = 2.15 in African Americans (Figure 3.1). Thus in the combined sample, we have good power to detect common alleles with effect sizes similar to those reported in other GWAS. For rarer alleles, the combined ancestry sample retains some power to detect strong effects (RR = 1.9 for an allele frequency of 0.01). But in the ancestry specific samples, power decreases sharply for lower allele frequencies; for a 0.01 frequency allele, we can detect an RR of only 2.5 in European Americans and RR = 4.5 in African Americans (Figure 3.1). This study is under powered to detect rare variants with low effects.

A manhattan and QQ plot for the combined ancestry analysis is shown in Figure 3.2 and Figure 3.3 respectively. From these figures, we see no inflation (**Figure 3.3**) and suggestive signals on chromosome 7 and 12 (**Figure 3.2**). These suggestive signals can be seen in the European American meta-analysis on chromosome 7 and 12 in **Figure 3.4** and **Figure 3.5** respectively as well as in the African American meta-analysis in **Figures 3.6** and **3.7** respectively. All figures only depict variants with MAF ≥ 0.01%.

Manhattan and QQ plots for European Americans in each analysis are depicted in Figures 3.8-15. The QQ plots for European American analyses (**Figure 3.8, 3.10, 3.15, and 3.14**) show no inflation that would be indicative of false positives. In fact, they indicate some deflation of p-values compared to expected; this would be consistent with the fact that the set of SNPs tested (genotyped and imputed) includes a fair amount of correlation/redundancy. Manhattan plots for European American analyses (**Figure 3.9, 3.11, 3.13, and 3.15**) show no

genome wide significant associations. The QQ plot for the African American meta-analysis can be found in **Figure 3.16**. This plot shows little inflation.

3.4 Discussion

Smoking is a major public health burden as it causes cancer and cardiovascular disease amongst other illnesses (Fagan et al., 2010). There are several differences in smoking patterns between mentholated and non-mentholated cigarette smokers. Menthol smokers are more likely to be African American, less educated, younger, and never married (Fagan et al., 2010; Okuyemi et al., 2004). Interestingly, several studies have found that mentholated cigarette smokers smoke less cigarettes per day than non-mentholated cigarette smokers (Fagan et al., 2010). Additionally, mentholated cigarettes may be more harmful than non-mentholated cigarettes since they generally contain more nicotine and tar (Gardiner, 2004; Muscat et al., 2002; Okuyemi et al., 2004), can contain more carcinogenic components (McCarthy et al., 1995), and absorb higher levels of carcinogens (Muscat et al., 2002). To this end, identifying genetic variants that contribute to mentholated cigarette use has the potential to aid in attempts to improve smoking cessation and ultimately reduce mortality.

We conducted ancestry specific meta-analyses in addition to our combined races analysis. We identified several suggestive signals on chromosome 7 and 12. QQ plots for each European American analysis shows deflation which is evidence that there may not be false positives. This study is well powered to detect common variants but under powered to detect rare variants with low effects.

Mentholated cigarettes were first marked to be used when one had the cold or a cough and were perceived to be healthier (Gardiner, 2004). Mentholated cigarettes have historically been marketed differently to African Americans and this began with the migration of African Americans from rural areas to urban cities (Gardiner, 2004). The sponsorship of Civil Rights efforts by tobacco companies and the perceived health benefits further solidified the use of this product in the African American community (Gardiner, 2004). Mentholated cigarettes still remain in high use amongst African Americans smokers. Many studies have examined the influence of mentholated cigarettes in lung cancer risk as a disparity exists in lung cancer between African and European Americans. However, not casual associations have been identified.

This study adds to the current literature by examining genetic determinants of mentholated cigarette use in large diverse samples over a large quantity of variants. This examination of genetic influences represents a new, important aspect to understanding menthol cigarette preference. Ultimately, to more fully understand menthol cigarette preference and its health consequences, research combining socioeconomic, genetic, and environmental determinants is needed.

Table 3.1: Demographics of Current Nicotine Dependent Smokers

Table 6: Demographics of Current Nicotine Dependent Smokers

	African Americans (N=1,365)				European Americans (N=2,206)			
	COGEND Part 1 (N=437)	COGEND Part 2 (N=286)	COGEND Part 3 (N=517)	UW-TTURC (N=125)	COGEND Part 1 (N=952)	COGEND Part 2 (N=228)	COGEND Part 3 (N = 181)	UW-TTURC (N=845)
Females N (%)	275 (62.93)	143 (50.00)	169 (32.69)	86 (68.80)	507 (53.26)	131 (57.46)	55 (30.39)	484 (57.28)
Age Mean ± SD	36.45±5.95	34.72±5.75	33.99±5.58	47.30±9.55	36.95±5.33	34.16±5.84	34.12±5.44	45.70±11.07
Menthol N (%)	412 (94.28)	277 (96.85)	503 (97.29)	113 (90.40)	282 (29.62)	72 (31.58)	70 (38.67)	313 (37.04)

Figure 3.1. Power Analyses. Detectable effect at 80% power for genome-wide screen (alpha=$5x10^{-8}$) under an additive model. In all analyses LD D' = 1.0 and EDF = 1.

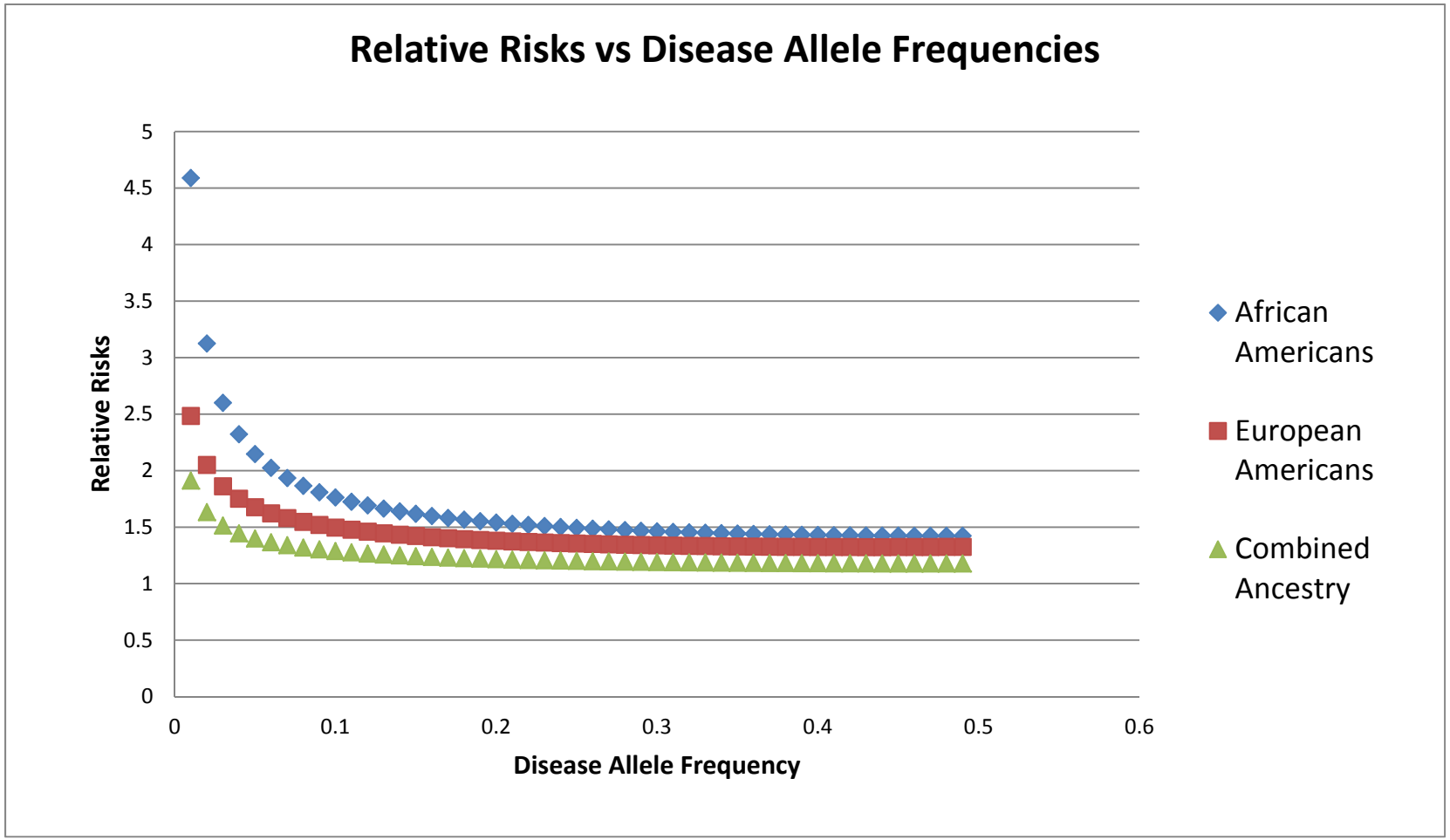

Figure 18: Power Analyses

Figure 3.2. Manhattan plot of association results predicting mentholated cigarette use in the Combined Ancestry Analysis

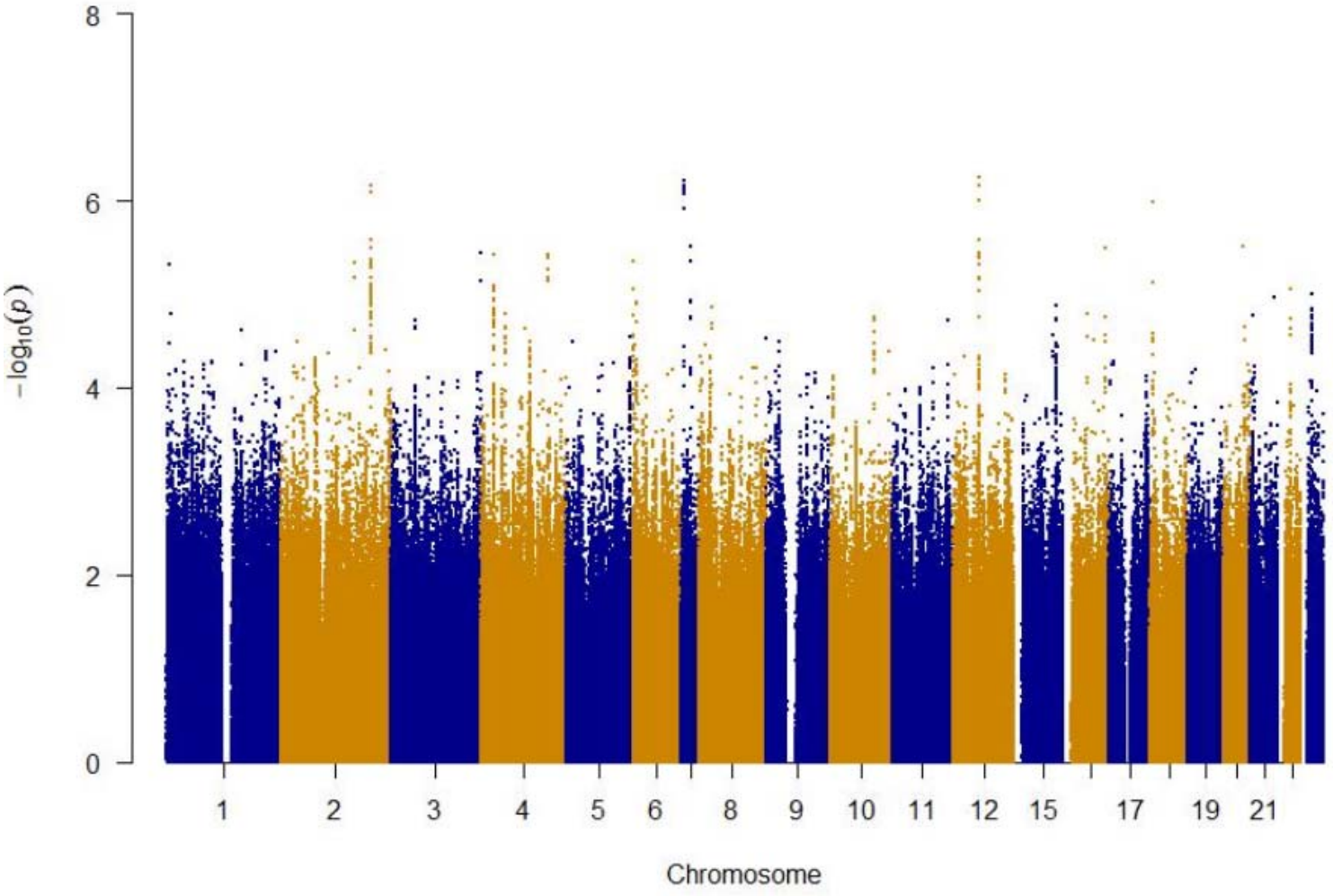

Figure 19: Manhattan Plot Combined Ancestry Analysis

Figure 3.3. QQ plot of association results predicting mentholated cigarette use in the Combined Ancestry Analysis

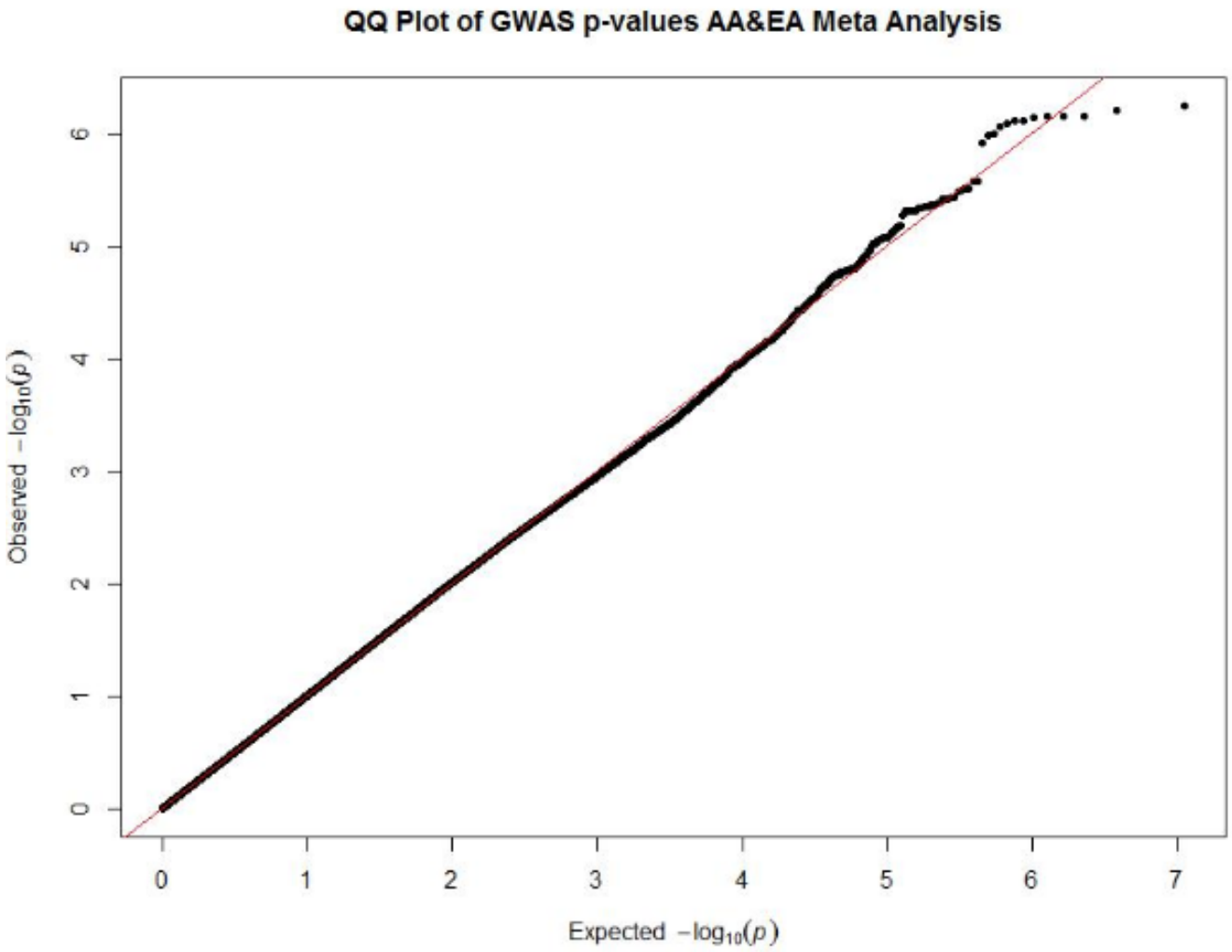

Figure 20: QQ Plot Combined Ancestry Analysis

Figure 3.4. Chromosome 7 plot of association results predicting mentholated cigarette use in European American Meta-Analysis

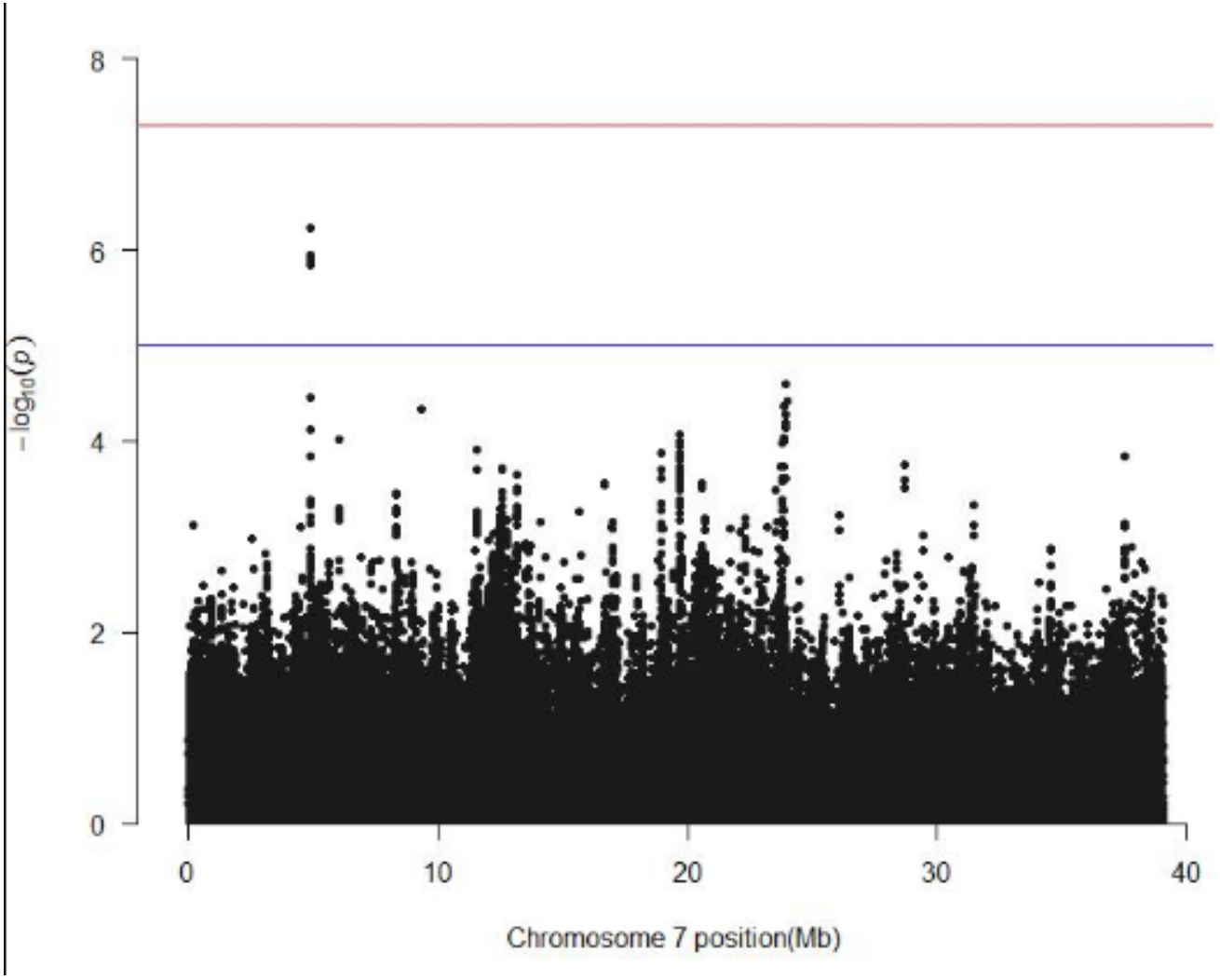

Figure 21: Chromosome 7 European American Meta-Analysis

Figure 3.5. Chromosome 12 plot of association results predicting mentholated cigarette use in European American Meta-Analysis

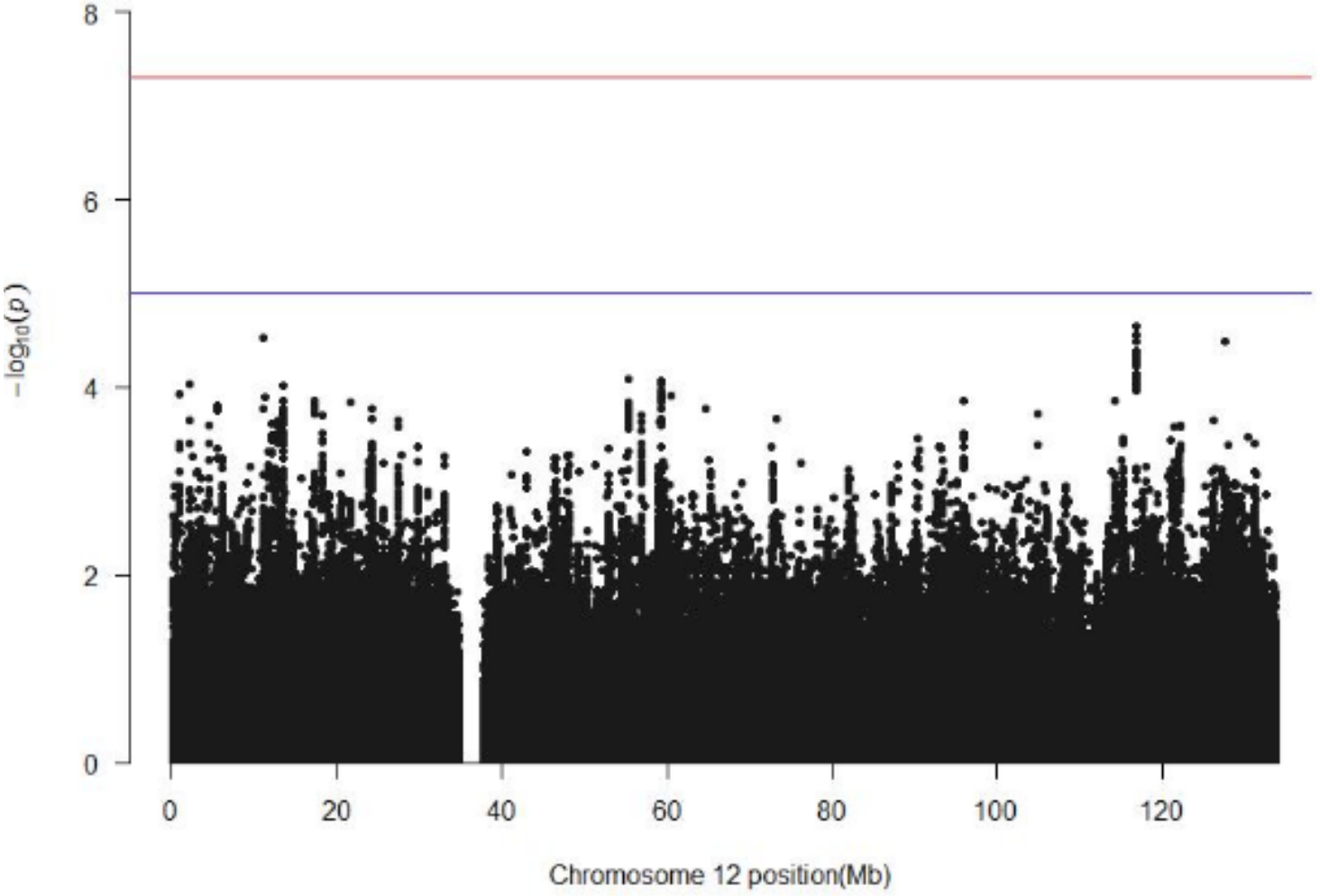

Figure 22: Chromosome 12 European American Meta-Analysis

Figure 3.6. Chromosome 7 plot of association results predicting mentholated cigarette use in African American Meta-Analysis

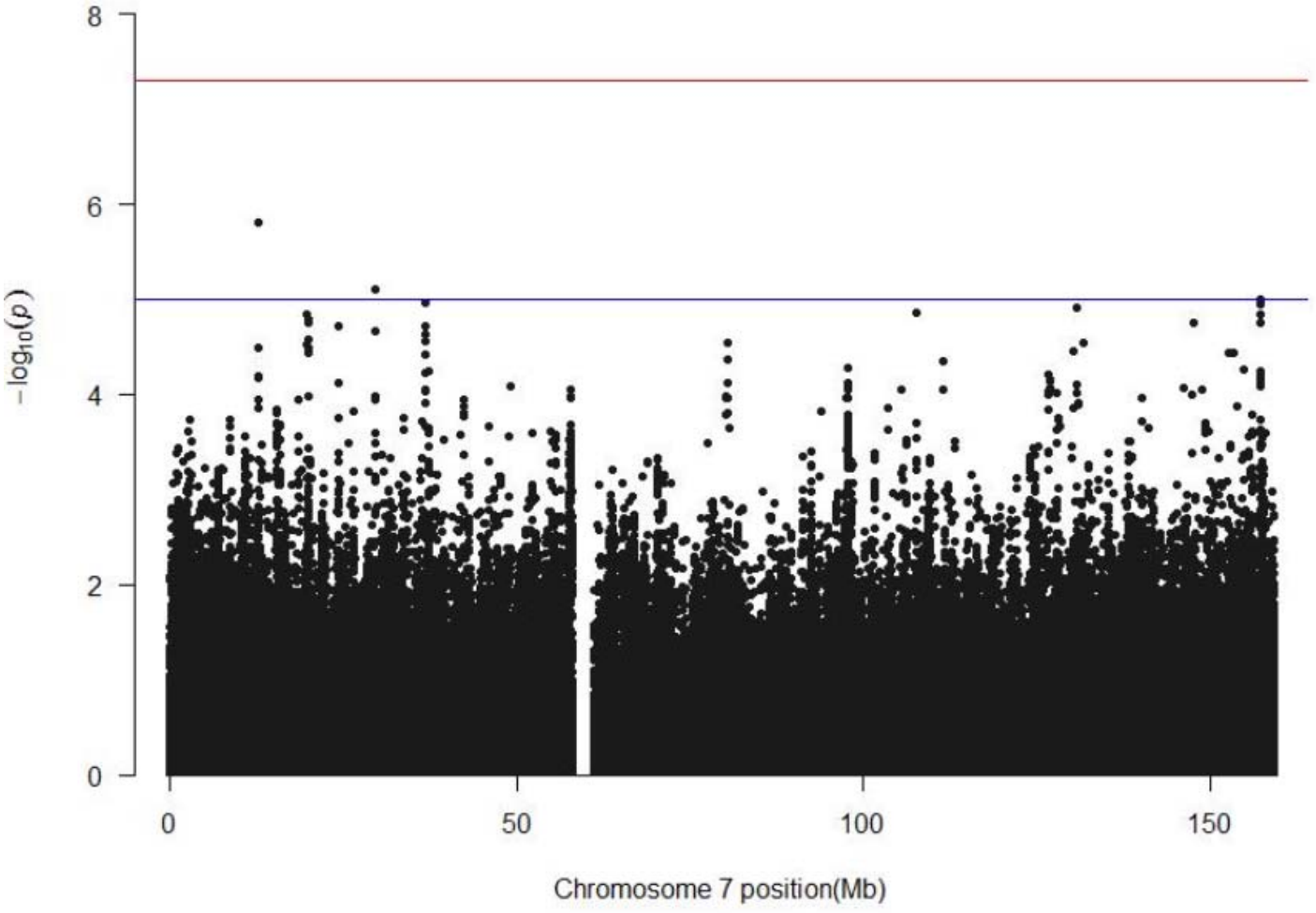

Figure 23: Chromosome 7 African American Meta-Analysis

Figure 3.7. Chromosome 12 plot of association results predicting mentholated cigarette use in African American Meta-Analysis

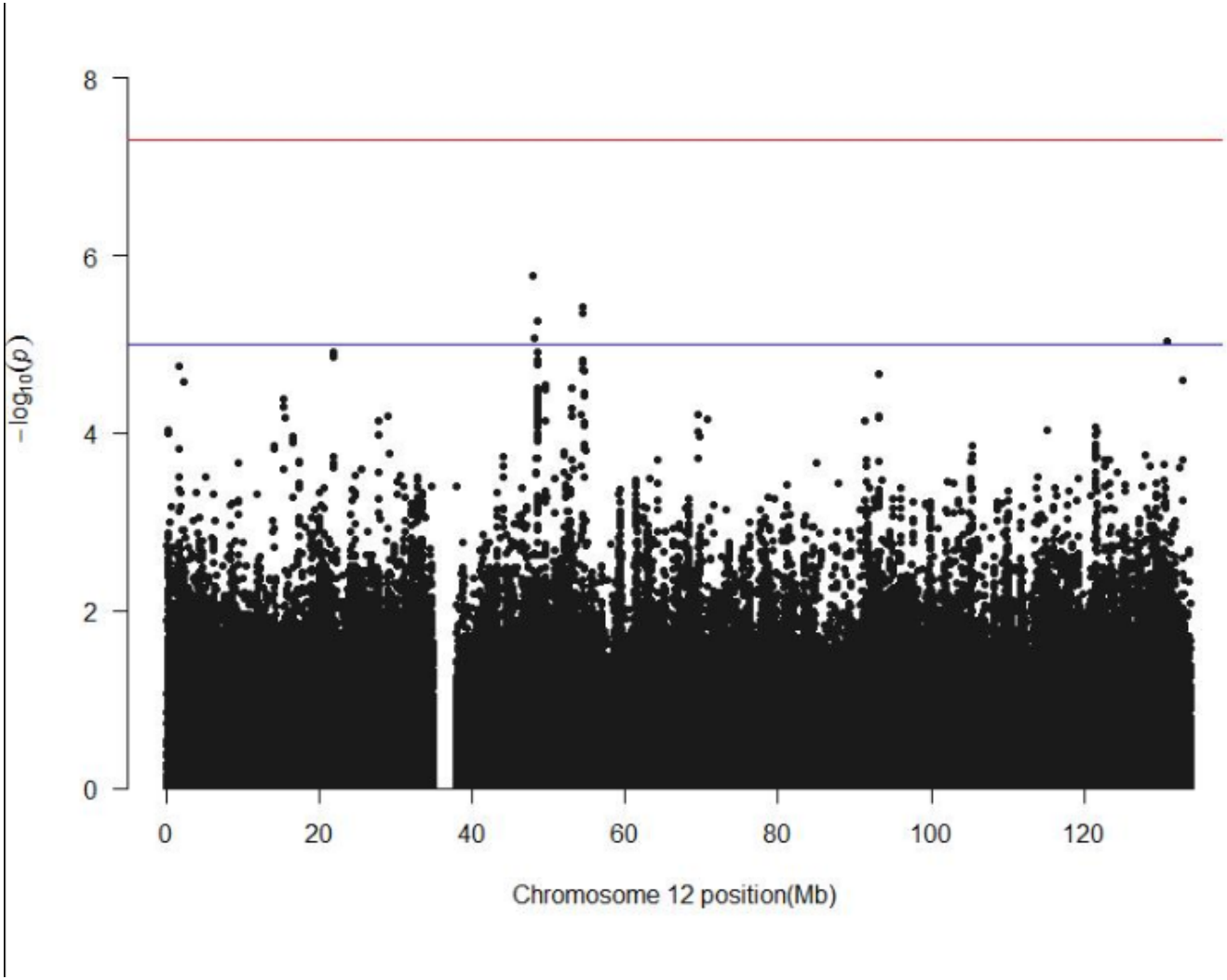

Figure 24: Chromosome 12 African American Meta-Analysis

Figure 3.8. QQ plot of association results predicting mentholated cigarette use in COGEND Part 1 European Americans

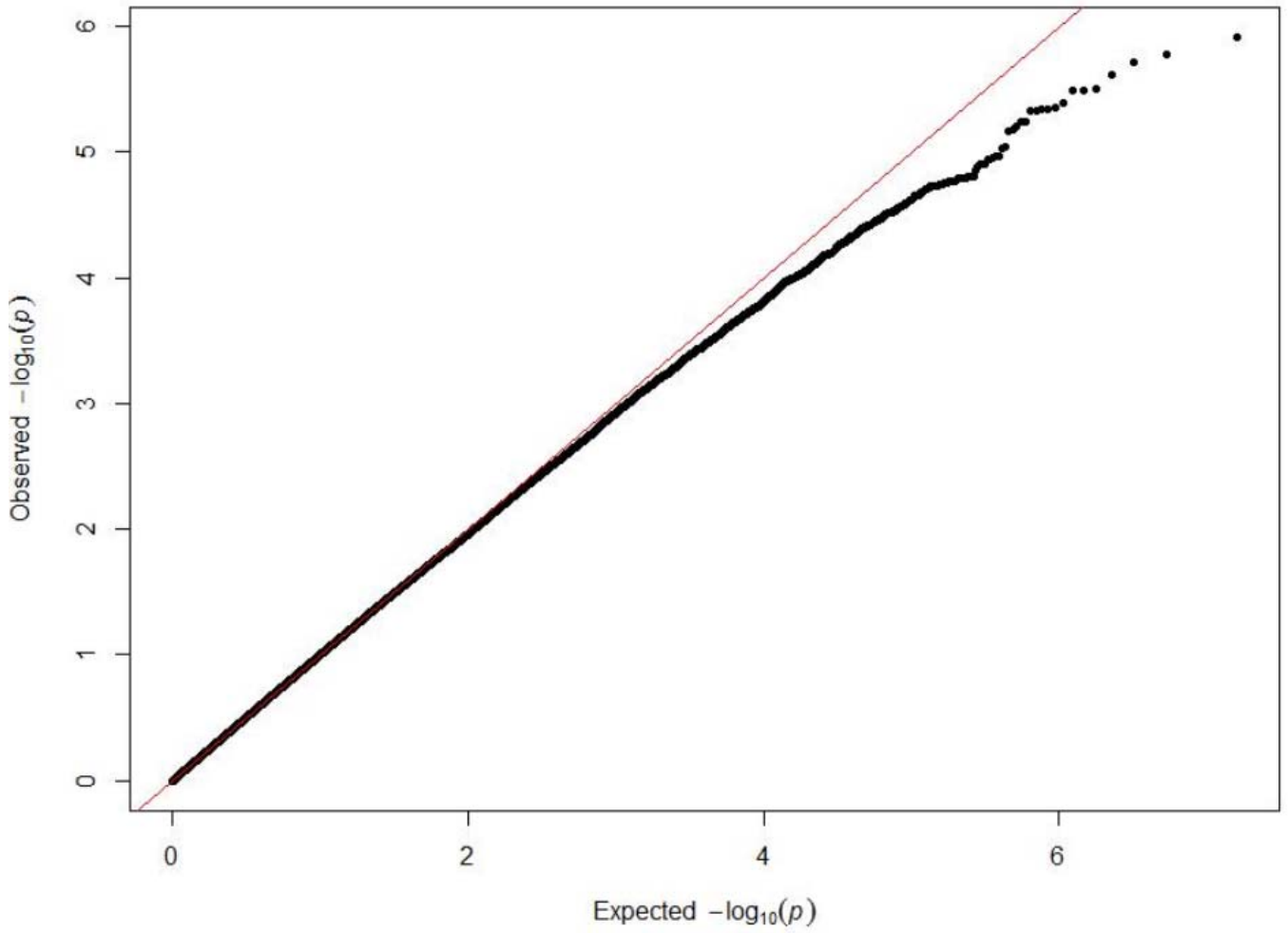

Figure 25: QQ Plot COGEND Part 1 EA Analysis

Figure 3.9. Manhattan plot of association results predicting mentholated cigarette use in COGEND Part 1 European Americans

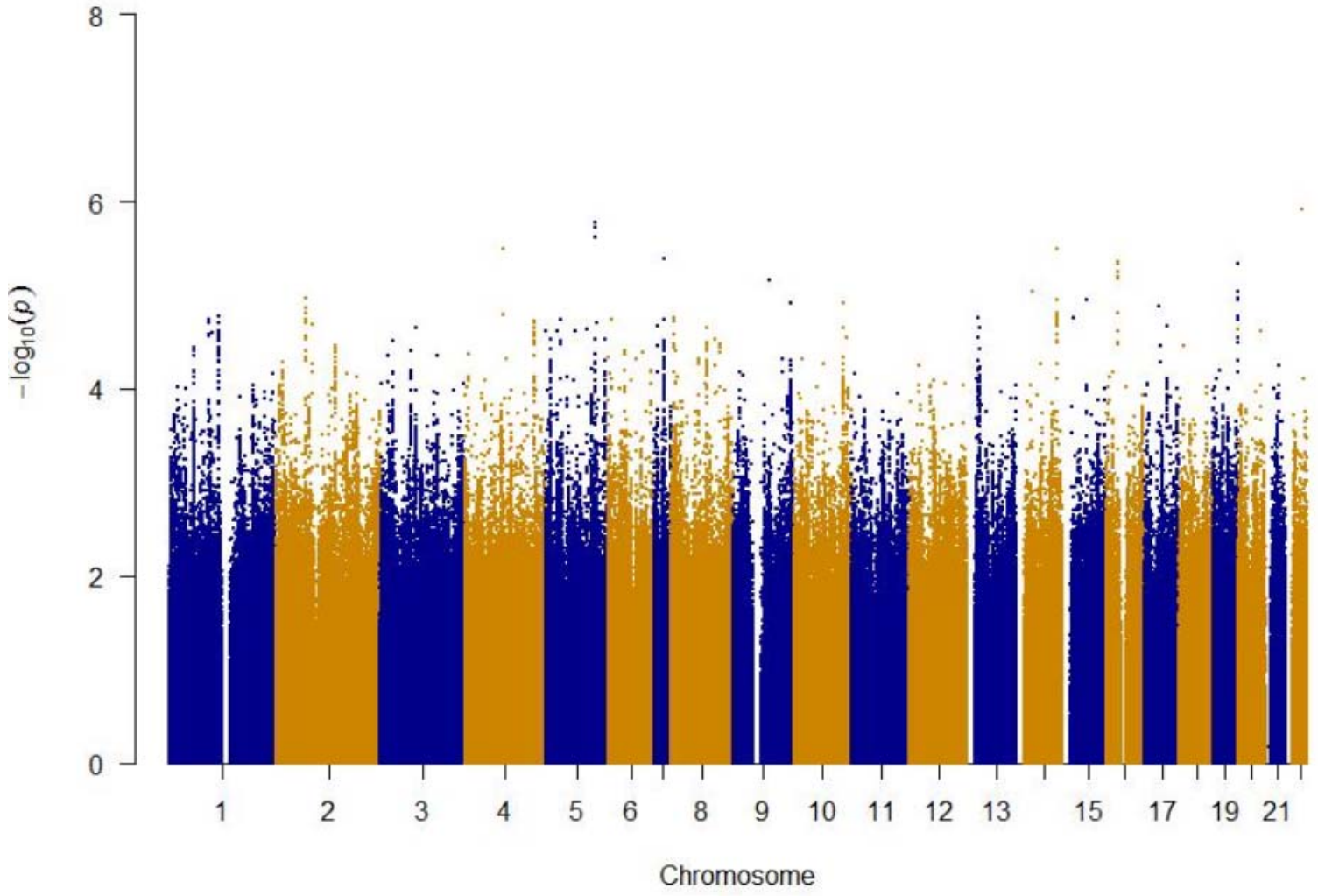

Figure 26: Manhattan Plot COGEND Part 1 EA Analysis

Figure 3.10. QQ plot of association results predicting mentholated cigarette use in COGEND Part 2 European Americans

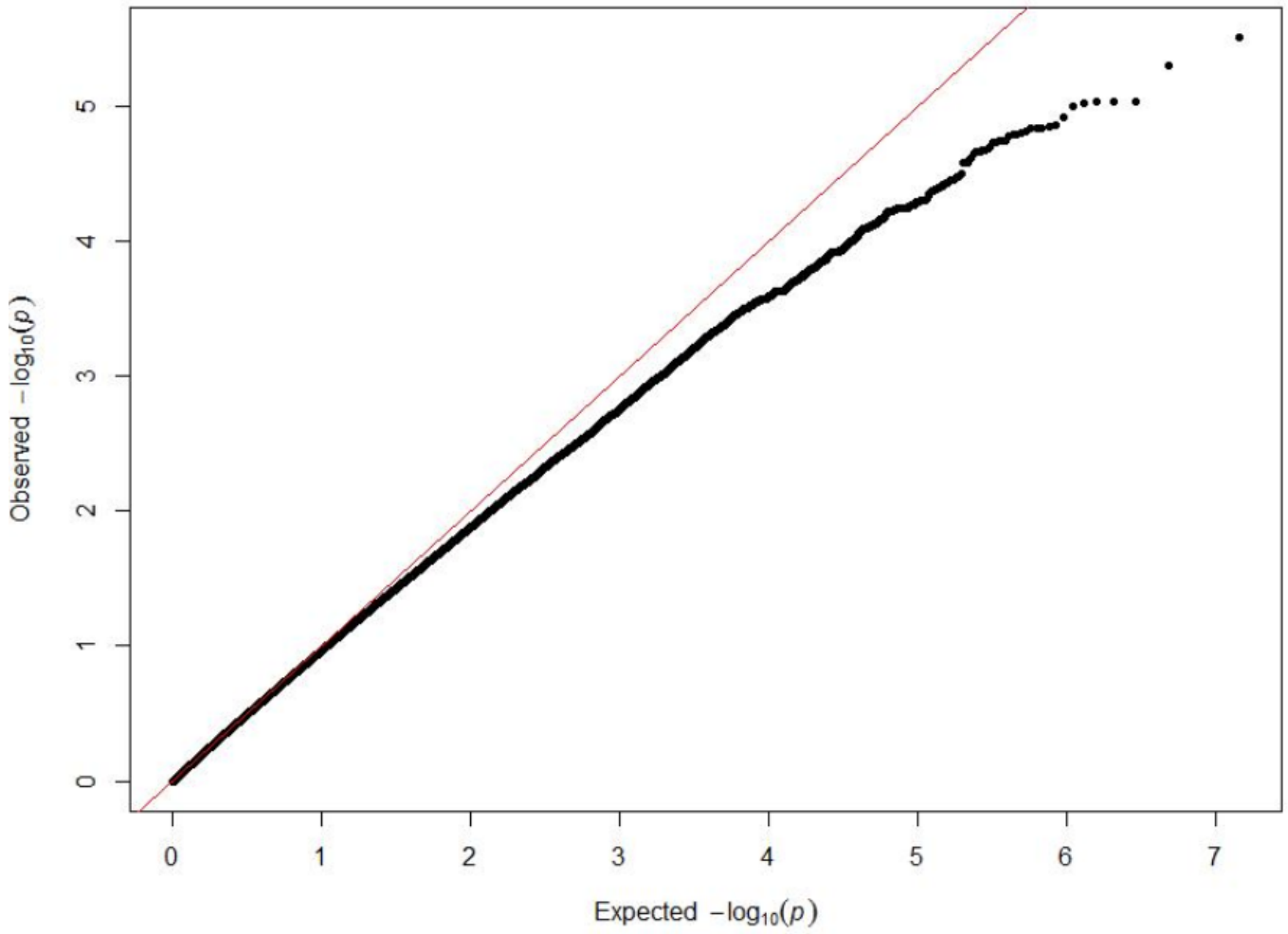

Figure 27: QQ Plot COGEND Part 2 EA Analysis

Figure 3.11. Manhattan plot of association results predicting mentholated cigarette use in COGEND Part 2 European Americans

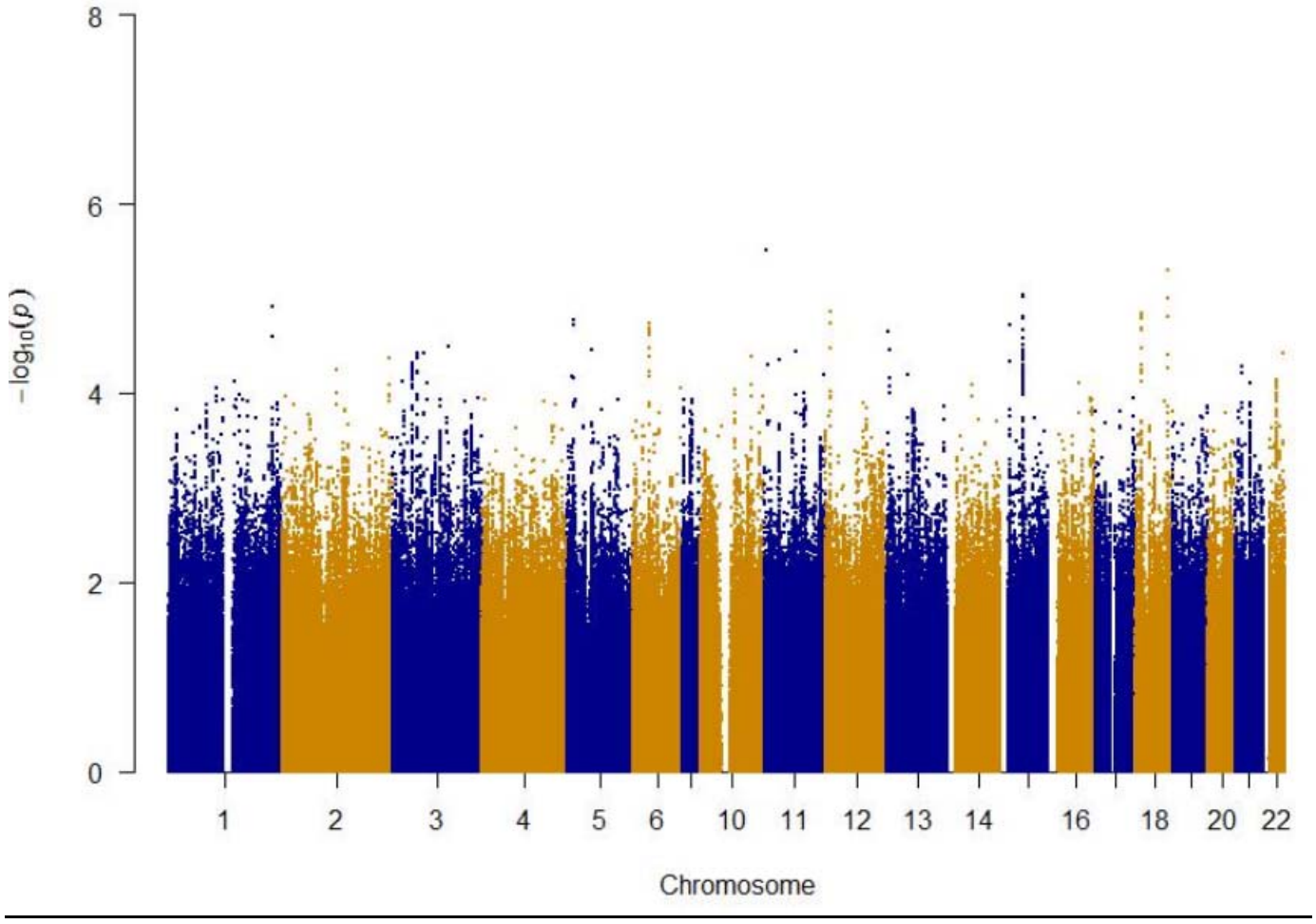

Figure 28: Manhattan Plot COGEND Part 2 EA Analysis

Figure 3.12. QQ plot of association results predicting mentholated cigarette use in COGEND Part 3 European Americans

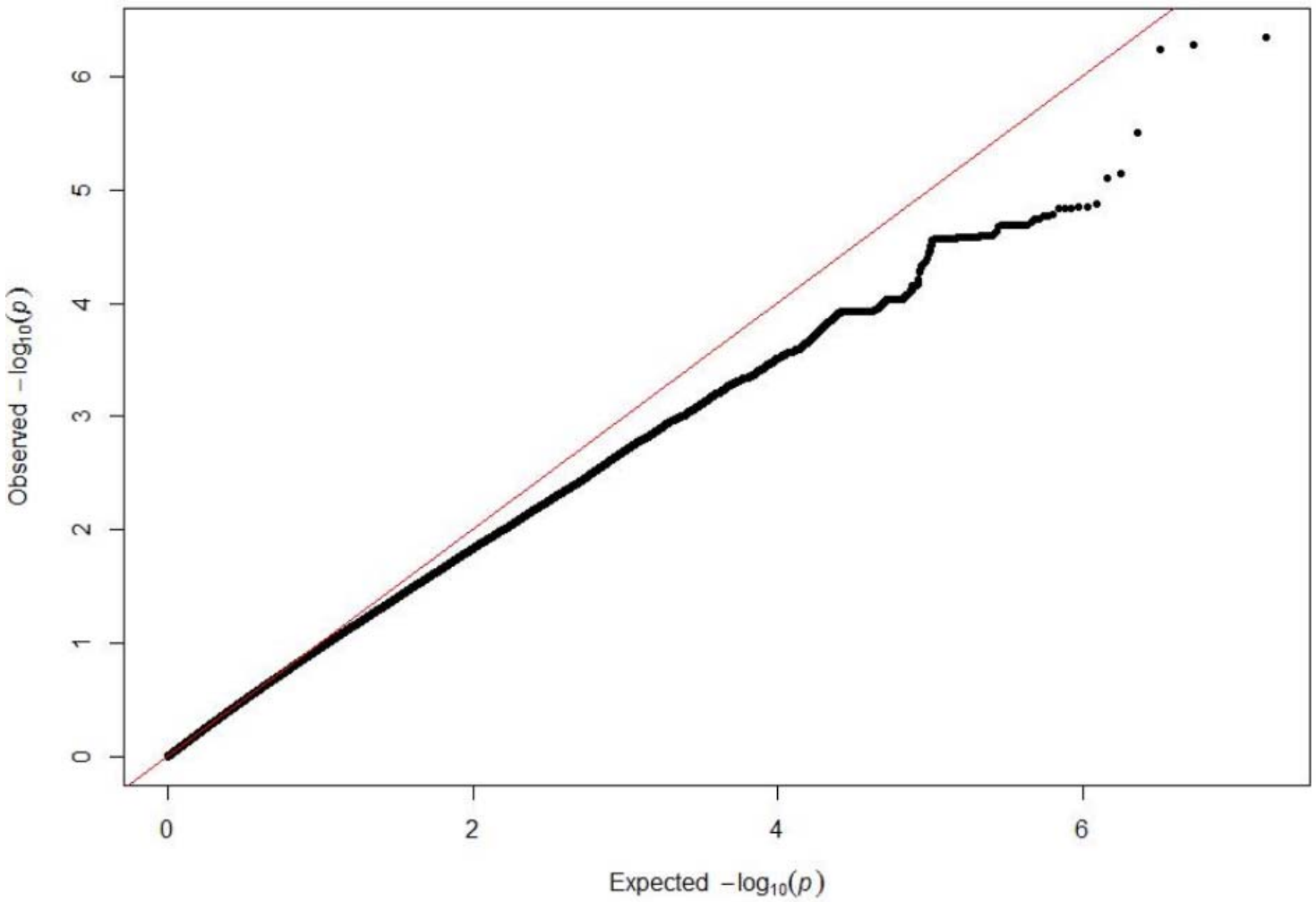

Figure 29: QQ Plot COGEND Part 3 EA Analysis

Figure 3.13. Manhattan plot of association results predicting mentholated cigarette use in COGEND Part 3 European Americans

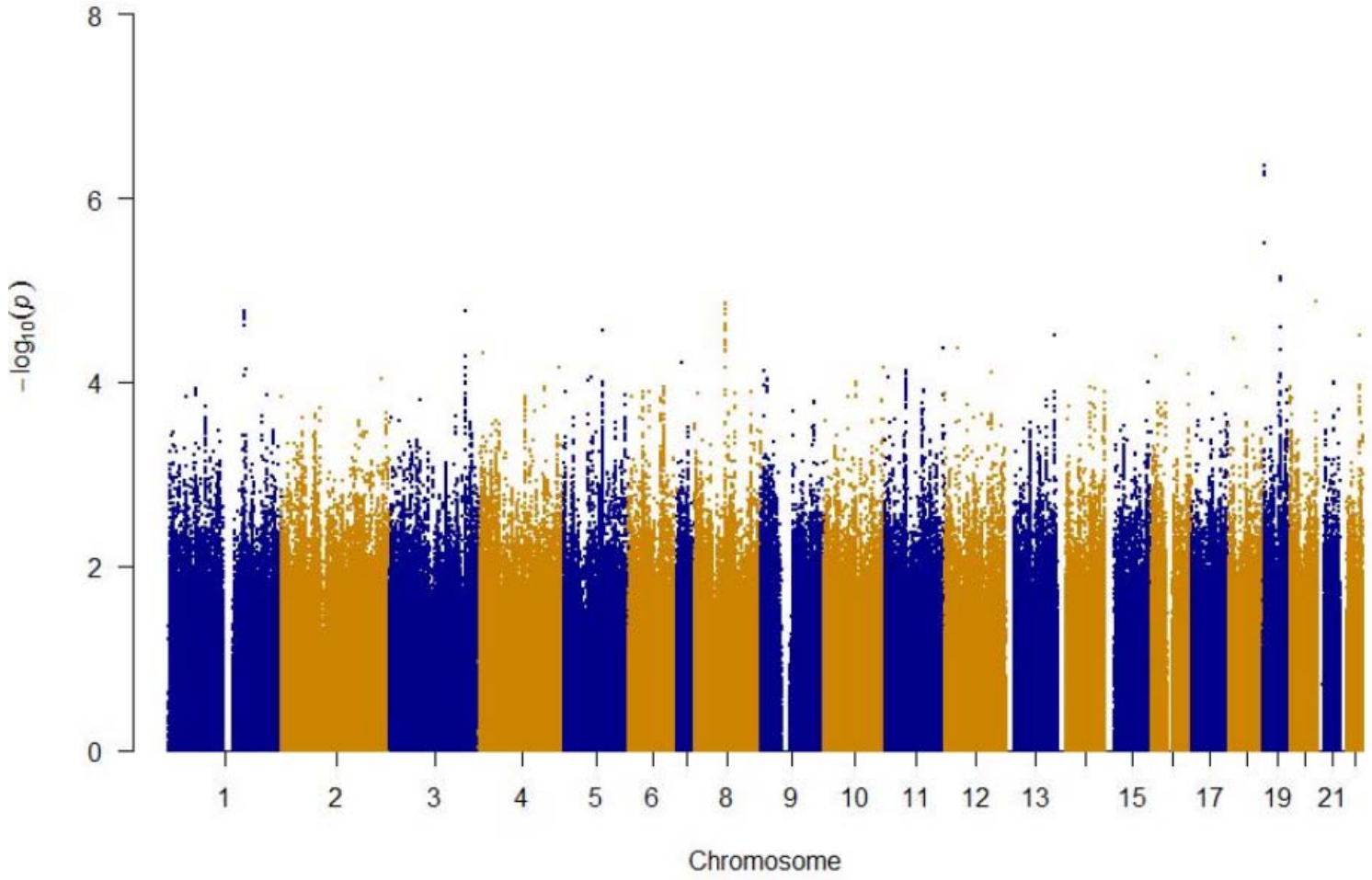

Figure 30: Manhattan Plot COGEND Part 3 EA Analysis

Figure 3.14. QQ plot of association results predicting mentholated cigarette use in UW-TTURC European Americans

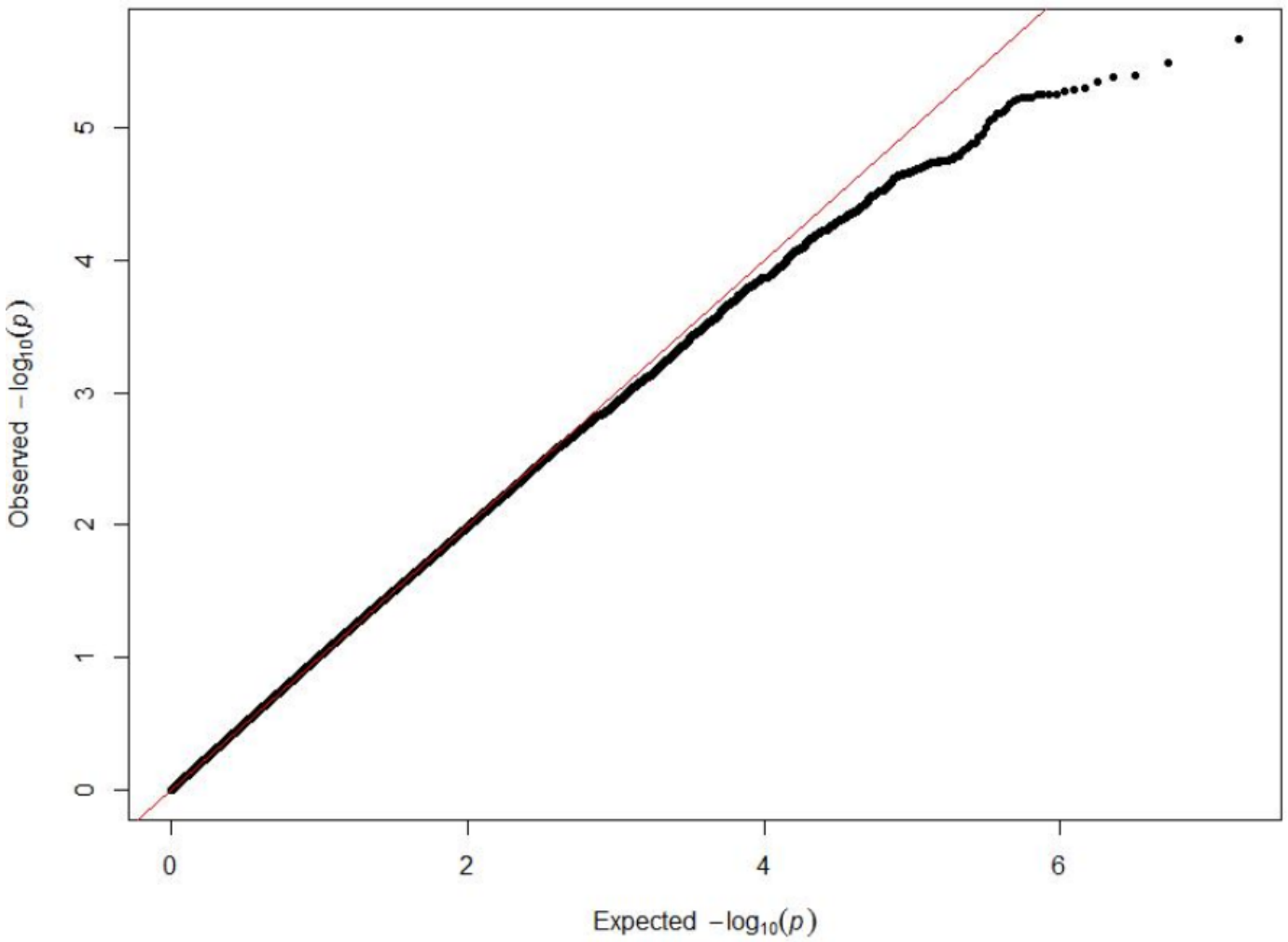

Figure 31: QQ Plot UW-TTURC EA Analysis

Figure 3.15. Manhattan plot of association results predicting mentholated cigarette use in UW-TTURC European Americans

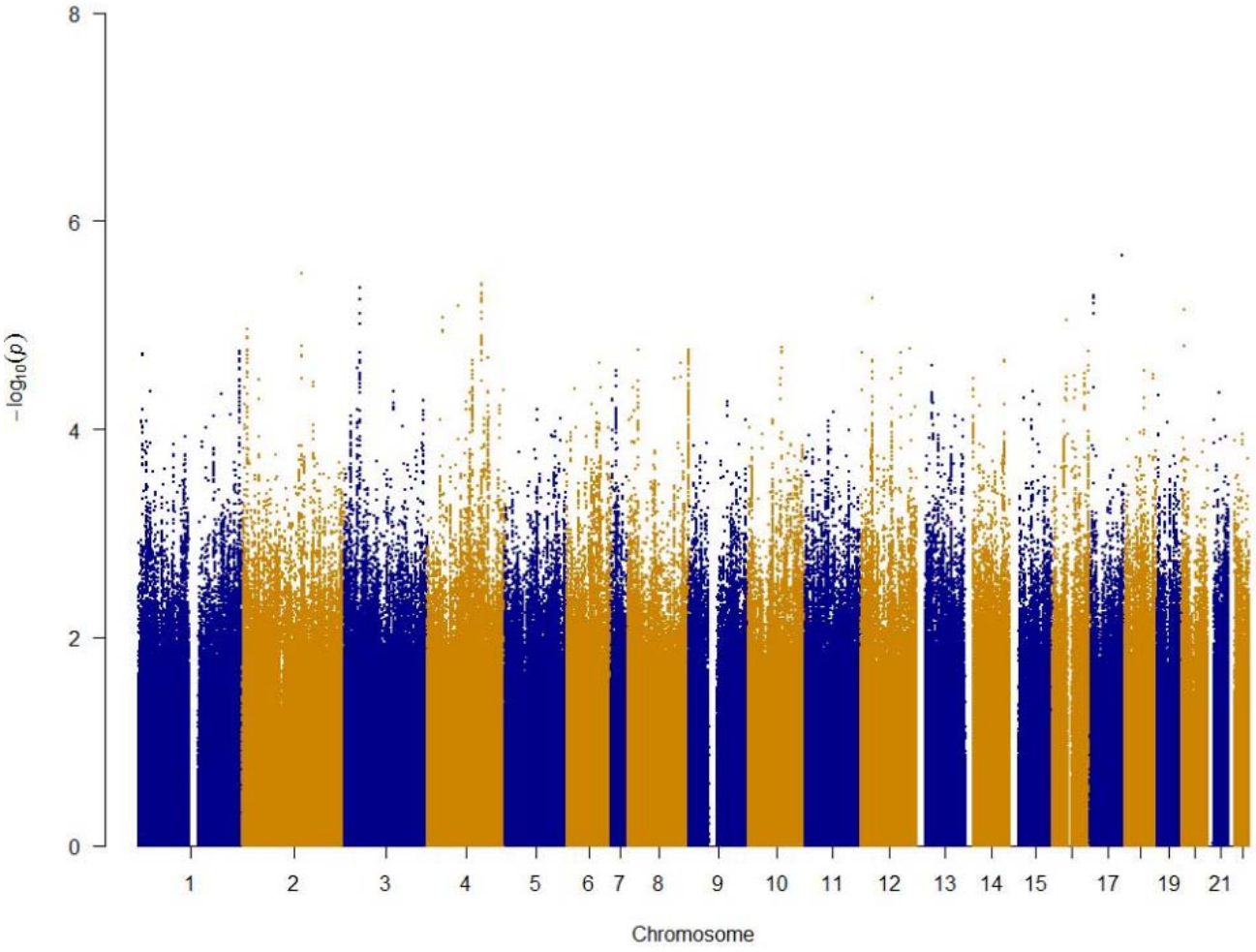

Figure 32: Manhattan Plot UW-TTURC EA Analysis

Figure 3.16. QQ plot of association results predicting mentholated cigarette use in African American Meta-Analysis

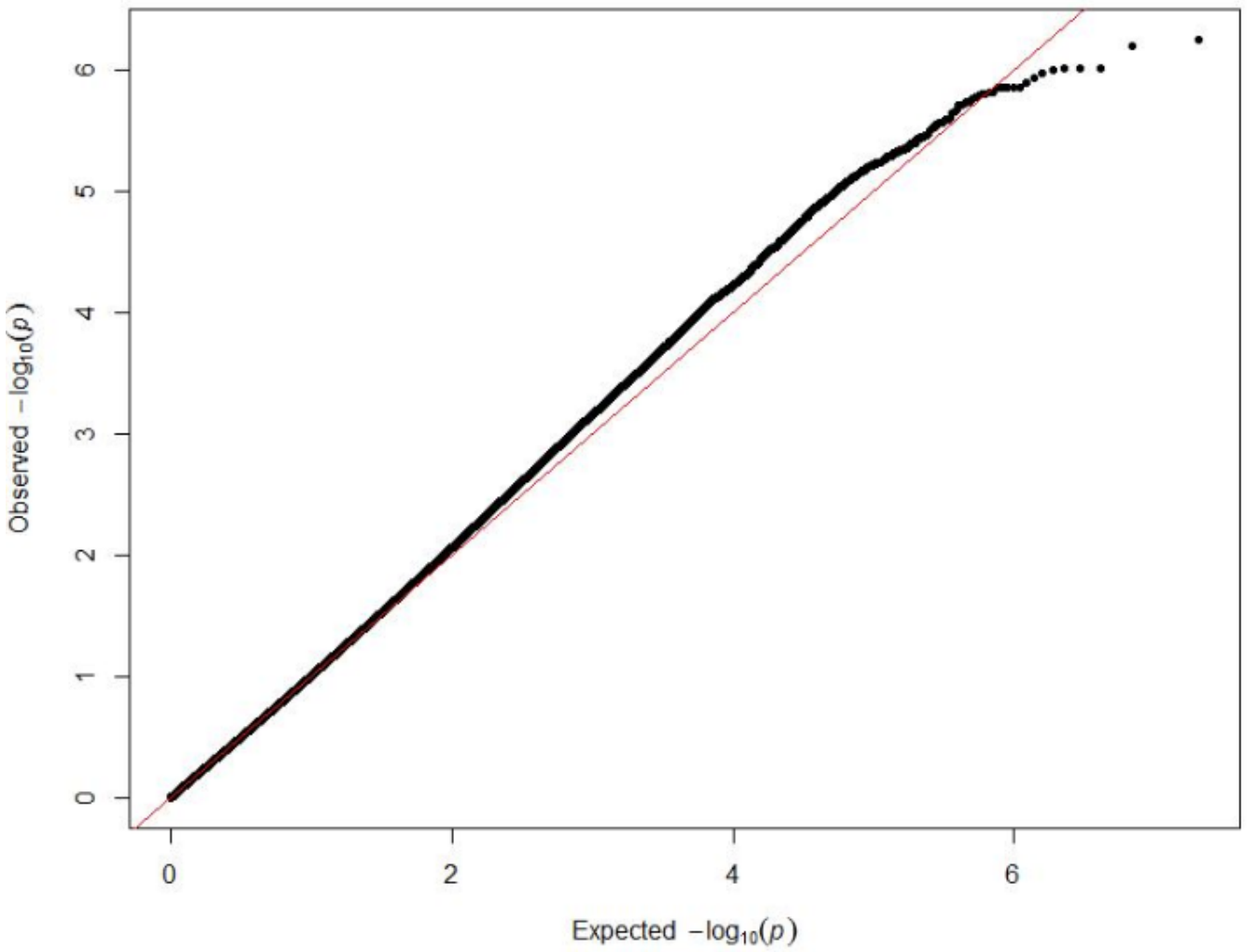

Figure 33: QQ Plot of AA Meta-Analysis

Chapter 4: Conclusions and Future Directions

This work considers imputation by examining imputation accuracy and using imputed data in a meta analysis. Chapter 3 of this work is in preparation for publication and Chapter 2 of this work has been published, see CV for citation.

4.1 Future Directions

As society and research changes, there are several areas which warrant study: imputation, genomic analyses in diverse populations, nicotine dependence in diverse populations, and electronic cigarettes. Imputation may be used for different purposes in the future. Genomic analyses in diverse populations need to be expanded and electronic cigarettes are changing smoking behaviors and nicotine dependence. These are discussed in subsequent sections.

4.1.1 Imputation

Although the cost of whole genome sequencing is decreasing, imputation will still be used in many analyses to come. However, as the field moves towards better understanding rare variants, the imputation quality of these data will be more important to consider. It has also been suggested that imputation can be used in scenarios of low sequencing coverage. While there are many uses for imputation, the limitations of reference panels and SNP arrays regarding diverse populations remain a concern.

4.1.2 Genomic Analyses in Diverse Populations

There is a disparity in the effort placed to understand European populations as compared to other populations. Bustamante et al 2011 reports that 96% of subjects in all GWAS studies were people of European descent. This commentary also expresses concern that this trend will likely continue in whole genome sequencing analyses (Bustamante, De La Vega, & Burchard, 2011). This disparity is of note as medical genomics and personalized medicine are becoming the standard of care. It was believed that findings from understanding European populations would be generalizable to other ancestries (Bustamante et al., 2011). However, in the case of rare variants, this may not the case. In particular, population-specific rare variants have been associated with various diseases, including nicotine dependence, that are not observed in other population groups (Olfson et al., 2015). Furthermore, studying genomic diversity in these populations is equitable and important irrespective of the implications in European populations.

Rare variants are usually population specific (Bustamante et al., 2011). Linkage disequilibrium contributes to population stratification. GWAS is predicated on the idea that the associated marker is casual or is linked to the causal variant. Since linkage disequilibrium patterns differ across populations, the relationship between the associated and casual marker may change depending on which population is being considered. African populations tend to contain greater genetic diversity and less linkage disequilibrium than other populations (Bustamante et al., 2011; Teo et al., 2010). Many researchers believe that the rare variants hold a significant proportion of the missing heritability in genomic analyses. Furthermore, the study of rare variants is increasingly important to treatment and pharmacological effects.

Factors that hinder the investigation and understanding of disease etiology in diverse populations include lack of expansive reference data and study designs that do not consider environmental factors in each population. Although 1000 Genomes created a large catalogue of genomic diversity, many other populations have not been sequenced. This lack of reference data hinders the study of disease in such populations. While studies need to consider the influence of environmental factors on disease in diverse populations (poverty, unequal access to care, lifestyle, and health-related cultural practices (Rotimi & Jorde, 2010)) other factors such as racism and internalization of negative racial bias are also paramount when examining some diseases. Chae et 2014 found accelerated biological aging amongst African American men due to racial discrimination and internalization of negative racial bias (Chae et al., 2014). With the advent of personalized medicine, more work is needed in the development of protocols to integrate cultural, socioeconomic, psychosocial, environmental and genetic factors in efforts to truly understand disease. A bias towards European populations exists and will persist unless major changes occur.

4.1.3 Nicotine Dependence in Diverse Populations

This bias towards studying European Americans persists in nicotine dependence studies (Bierut, 2010). The genetic variant most strongly associated with nicotine dependence is rs16969968 (Bierut, 2010; Tobacco and Genetics Consortium, 2010). This variant differs in allele frequency in European and African populations (Bierut, 2010). However, recently important strides have been made with several genetic investigations of nicotine dependence and other smoking behaviors in African Americans (Chen et al., 2012; David et al., 2012; Saccone et al.,

2010) and Asians (Chen et al., 2012). Understanding the genetic architecture of nicotine dependence in a variety of populations is an ongoing area of research.

4.1.4 Cessation, Nicotine Replacement, and Electronic Cigarettes

Nicotine replacement therapies (NRT) have been embraced as they positively influence cessation (Cahn & Siegel, 2011). Varenicline, buproprion, and NRT are the three main pharmacological approaches to smoking cessation (Bierut, 2010). Smoking cessation efforts are important as they are believed to decrease mortality and smoking related diseases/illnesses.

Recently, electronic cigarettes have gained in popularity. Electronic cigarettes are generally not considered to be a part of cessation efforts since they closely mimic the act of smoking (Cahn & Siegel, 2011). Electronic cigarettes were first created in 2003 (Odum, O'Dell, & Schepers, 2012) and introduced to the American markets in 2007; these products have been heavily marketed and purchased through the internet grossing millions of dollars (Noel, Rees, & Connolly, 2011). These battery powered devices vaporize a nicotine solution and avoid combustion (Cahn & Siegel, 2011).

Many believe that these products are safer than tobacco cigarettes (Cahn & Siegel, 2011; Etter, 2010; Trtchounian, Williams, & Talbot, 2010). The safety of these products cannot be comprehensively examined due to the wide variety in brands (Noel et al., 2011; Trtchounian et al., 2010). These products are banned in Australia, Brazil, Canada, Denmark, and Switzerland (Etter, 2010). Furthermore, there is variability in the amount of nicotine present in these products (Odum et al., 2012; Trtchounian et al., 2010). These products are often used for

smoking cessation although they are not FDA approved for this purpose (Odum et al., 2012). It is yet to be determined if electronic cigarettes are harmless and should not be regulated.

4.2 Summary

There are several important frontiers still to be explored regarding genetic influences on nicotine dependence and cessation, as discussed above. The work in this thesis has bridged some of the knowledge gaps to enable future work. In Chapter 2, we provided new understanding of how to assess imputation accuracy, which is being used in large-scale genetic studies of smoking and addictions. In addition, this work on imputation accuracy focused on imputation in African ancestry populations and thus should aid future genetic studies of understudied African Americans. Chapter 3 also addresses the need for genetic study in diverse populations by focusing on a smoking phenotype, menthol cigarette use, that is particularly prevalent in the minority population of African Americans. Our use of both European Americans and African Americans together affirms the value of diverse population cohorts in genetic studies.

References

Asimit, J., & Zeggini, E. (2010). Rare Variant Association Analysis Methods for Complex Traits. *Annual Review of Genetics, 44*(1), 293-308. doi: 10.1146/annurev-genet-102209-163421

Benowitz, N. L. (2010). Nicotine Addiction. *New England Journal of Medicine, 362*(24), 2295-2303. doi: doi:10.1056/NEJMra0809890

Benowitz, N. L., Herrera, B., & Jacob, P. (2004). Mentholated Cigarette Smoking Inhibits Nicotine Metabolism. *Journal of Pharmacology and Experimental Therapeutics, 310*(3), 1208-1215. doi: 10.1124/jpet.104.066902

Benowitz, N. L., Perez-Stable, E. J., Fong, I., Modin, G., Herrera, B. (1999). Ethnic Differences in N-Glucuronidation of Nicotine and Cotinine. *Journal of Pharmacology and Experimental Therapeutics, 291*(3), 1196-1203.

Bierut, L. J. (2010). Convergence of genetic findings for nicotine dependence and smoking related diseases with chromosome 15q24-25. *Trends in Pharmacological Sciences, 31*(1), 46-51. doi: http://dx.doi.org/10.1016/j.tips.2009.10.004

Bierut, L. J., Madden, P. A. F., Breslau, N., Johnson, E. O., Hatsukami, D. (2007). Novel genes identified in a high-density genome wide association study for nicotine dependence. *Human Molecular Genetics, 16*(1), 24-35. doi: 10.1093/hmg/ddl441

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet, 84*(2), 210-223. doi: S0002-9297(09)00012-3 [pii]

10.1016/j.ajhg.2009.01.005

Browning, S. R. (2006). Multilocus Association Mapping Using Variable-Length Markov Chains. *American Journal of Human Genetics, 78*(6), 903-913.

Bustamante, C. D., De La Vega, F. M., & Burchard, E. G. (2011). Genomics for the world. *Nature, 475*(7355), 163-165.

Cahn, Z., & Siegel, M. (2011). Electronic cigarettes as a harm reduction strategy for tobacco control: A step forward or a repeat of past mistakes[quest]. *J Public Health Pol, 32*(1), 16-31.

Chae, D. H., Nuru-Jeter, A. M., Adler, N. E., Brody, G. H., Lin, J. (2014). Discrimination, Racial Bias, and Telomere Length in African-American Men. *American Journal of Preventive Medicine, 46*(2), 103-111. doi: http://dx.doi.org/10.1016/j.amepre.2013.10.020

Chanda, P., Yuhki, N., Li, M., Bader, J. S., Hartz, A. (2012). Comprehensive evaluation of imputation performance in African Americans. *Journal of human genetics, 57*(7), 411-421. doi: 10.1038/jhg.2012.43

Chen, L.-S., Bloom, A. J., Baker, T. B., Smith, S. S., Piper, M. E. (2014). Pharmacotherapy Effects on Smoking Cessation Vary with Nicotine Metabolism Gene (CYP2A6). *Addiction (Abingdon, England), 109*(1), 128-137. doi: 10.1111/add.12353

Chen, L.-S., Xian, H., Grucza, R. A., Saccone, N. L., Wang, J. C. (2012). Nicotine Dependence and Comorbid Psychiatric Disorders: Examination of Specific Genetic Variants in the CHRNA5-A3-B4 Nicotinic Receptor Genes(). *Drug and Alcohol Dependence, 123S1*, S42-S51. doi: 10.1016/j.drugalcdep.2012.01.014

David, S. P., Hamidovic, A., Chen, G. K., Bergen, A. W., Wessel, J. (2012). Genome-wide meta-analyses of smoking behaviors in African Americans. *Transl Psychiatry, 2*, e119. doi: 10.1038/tp.2012.41

Duan, Q., Liu, E. Y., Croteau-Chonka, D. C., Mohlke, K. L., & Li, Y. (2013). A comprehensive SNP and indel imputability database. *Bioinformatics, 29*(4), 528-531. doi: 10.1093/bioinformatics/bts724

Etter, J.-F. (2010). Electronic cigarettes: a survey of users. *BMC Public Health, 10*(1), 1-7. doi: 10.1186/1471-2458-10-231

Fagan, P., Moolchan, E. T., Hart, J. A., Rose, A., Lawrence, D. (2010). Nicotine dependence and quitting behaviors among menthol and non-menthol smokers with similar consumptive patterns. *Addiction, 105*, 55-74. doi: 10.1111/j.1360-0443.2010.03190.x

Gardiner, P. S. (2004). The African Americanization of menthol cigarette use in the United States. *Nicotine & Tobacco Research, 6*(Suppl 1), S55-S65. doi: 10.1080/14622200310001649478

Garte, S. (2002). The racial genetics paradox in biomedical research and public health. *Public Health Reports, 117*(5), 421-425.

Giovino, G. A., Sidney, S., Gfroerer, J. C., O'Malley, P. M., Allen, J. A. (2004). Epidemiology of menthol cigarette use. *Nicotine & Tobacco Research, 6*(Suppl 1), S67-S81. doi: 10.1080/14622203710001649696

Hancock, D. B., Levy, J. L., Gaddis, N. C., Bierut, L. J., Saccone, N. L. (2012). Assessment of Genotype Imputation Performance Using 1000 Genomes in African American Studies. *PLoS One, 7*(11), e50610. doi: 10.1371/journal.pone.0050610

Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., & Fagerstrom, K.-O. (1991). The Fagerström Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *British Journal of Addiction, 86*(9), 1119-1127. doi: 10.1111/j.1360-0443.1991.tb01879.x

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet, 44*(8), 955-959. doi: http://www.nature.com/ng/journal/v44/n8/abs/ng.2354.html#supplementary-information

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype Imputation with Thousands of Genomes. *G3: Genes|Genomes|Genetics, 1*(6), 457-470. doi: 10.1534/g3.111.001198

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet, 5*(6), e1000529.

Johnson, E. O., Hancock, D. B., Levy, J. L., Gaddis, N. C., Saccone, N. L. (2013). Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Human genetics, 132*(5), 509-522. doi: 10.1007/s00439-013-1266-7

Kim, U.-K., Breslin, P. A. S., Reed, D., & Drayna, D. (2004). Genetics of Human Taste Perception. *Journal of Dental Research, 83*(6), 448-453. doi: 10.1177/154405910408300603

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol, 34*(6), 591-602. doi: 10.1002/gepi.20516

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. *Genetic epidemiology, 34*(8), 816-834. doi: 10.1002/gepi.20533

Lin, P., Hartz, S. M., Zhang, Z., Saccone, S. F., Wang, J. (2010). A New Statistic to Evaluate Imputation Reliability. *PLoS One, 5*(3), e9697. doi: 10.1371/journal.pone.0009697

Liu, E. Y., Buyske, S., Aragaki, A. K., Peters, U., Boerwinkle, E. (2012). Genotype Imputation of MetabochipSNPs Using a Study-Specific Reference Panel of ∼4,000 Haplotypes in African Americans From the Women's Health Initiative. *Genetic epidemiology, 36*(2), 107-117. doi: 10.1002/gepi.21603

Liu, E. Y., Li, M., Wang, W., & Li, Y. (2013). MaCH-Admix: Genotype Imputation for Admixed Populations. *Genetic epidemiology, 37*(1), 25-37. doi: 10.1002/gepi.21690

Liu, J. Z., Tozzi, F., Waterworth, D. M., Pillai, S. G., Muglia, P. (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet, 42*(5), 436-440. doi: http://www.nature.com/ng/journal/v42/n5/suppinfo/ng.572_S1.html

Luo, Z., Alvarado, G. F., Hatsukami, D. K., Johnson, E. O., Bierut, L. J. (2008). Race Differences in Nicotine Dependence in the Collaborative Genetic Study of Nicotine Dependence (COGEND). *Nicotine & Tobacco Research, 10*(7), 1223-1230. doi: 10.1080/14622200802163266

MacDougall, J. M., Fandrick, K., Zhang, X., Serafin, S. V., & Cashman, J. R. (2003). Inhibition of Human Liver Microsomal (S)-Nicotine Oxidation by (−)-Menthol and Analogues. *Chemical Research in Toxicology, 16*(8), 988-993. doi: 10.1021/tx0340551

Mangold, J. E., Payne, T. J., Ma, J. Z., Chen, G., & Li, M. D. (2008). Bitter taste receptor gene polymorphisms are an important factor in the development of nicotine dependence in African Americans. *Journal of Medical Genetics, 45*(9), 578-582. doi: 10.1136/jmg.2008.057844

Manolio, T. A. (2010). Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine, 363*(2), 166-176. doi: doi:10.1056/NEJMra0905980

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet, 11*(7), 499-511. doi: 10.1038/nrg2796

McCarthy, W. J., Caskey, N. H., Jarvik, M. E., Gross, T. M., Rosenblatt, M. R. (1995). Menthol vs nonmenthol cigarettes: effects on smoking behavior. *American Journal of Public Health, 85*(1), 67-72.

Menashe, I., Rosenberg, P. S., & Chen, B. E. (2008). PGA: power calculator for case-control genetic association analyses. *BMC Genetics, 9*, 36-36. doi: 10.1186/1471-2156-9-36

Muscat, J. E., Richie, J. P., & Stellman, S. D. (2002). Mentholated cigarettes and smoking habits in whites and blacks. *Tobacco Control, 11*(4), 368-371. doi: 10.1136/tc.11.4.368

Nelson, S. C., Doheny, K. F., Pugh, E. W., Romm, J. M., Ling, H. (2013). Imputation-Based Genomic Coverage Assessments of Current Human Genotyping Arrays. *G3: Genes|Genomes|Genetics, 3*(10), 1795-1807. doi: 10.1534/g3.113.007161

Noel, J. K., Rees, V. W., & Connolly, G. N. (2011). Electronic cigarettes: a new 'tobacco' industry? *Tobacco Control, 20*(1), 81. doi: 10.1136/tc.2010.038562

Odum, L. E., O'Dell, K. A., & Schepers, J. S. (2012). Electronic Cigarettes: Do They Have a Role in Smoking Cessation? *Journal of Pharmacy Practice, 25*(6), 611-614. doi: 10.1177/0897190012451909

Office of Applied Studies. *Results From the 2008 National Survey on Drug Use and Health: National Findings*. Rockville, MD: Substance Abuse and Mental Health Services Administration; 2009

Okuyemi, K. S., Ebersole-Robinson, M., Nazir, N., & Ahluwalia, J. S. (2004). African-American menthol and nonmenthol smokers: differences in smoking and cessation experiences. *Journal of the National Medical Association, 96*(9), 1208-1211.

Olfson, E., Saccone, N. L., Johnson, E. O., Chen, L. S., Culverhouse, R. (2015). Rare, low frequency and common coding variants in CHRNA5 and their contribution to nicotine dependence in European and African Americans. *Mol Psychiatry*. doi: 10.1038/mp.2015.105

Pérez-Stable, E. J., Herrera, B., Jacob, I. P., & Benowitz, N. L. (1998). NIcotine metabolism and intake in black and white smokers. *JAMA, 280*(2), 152-156.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet, 38*(8), 904-909. doi: http://www.nature.com/ng/journal/v38/n8/suppinfo/ng1847_S1.html

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, Manuel A R. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics, 81*(3), 559-575.

Rice, J. P., Hartz, S., Agrawal, A., Almasy, L., Bennett, S. (2012). CHRNB3 is more strongly associated with FTCD-based nicotine dependence than cigarettes per day: phenotype definition changes GWAS results. *Addiction (Abingdon, England), 107*(11), 2019-2028. doi: 10.1111/j.1360-0443.2012.03922.x

Rogers, A., Beck, A., & Tintle, N. L. (2014). Evaluating the concordance between sequencing, imputation and microarray genotype calls in the GAW18 data. *BMC Proceedings, 8*(Suppl 1), S22-S22. doi: 10.1186/1753-6561-8-s1-s22

Rotimi, C. N., & Jorde, L. B. (2010). Ancestry and Disease in the Age of Genomic Medicine. *New England Journal of Medicine, 363*(16), 1551-1558. doi: doi:10.1056/NEJMra0911564

Saccone, N. L., Culverhouse, R. C., Schwantes-An, T.-H., Cannon, D. S., Chen, X. (2010). Multiple Independent Loci at Chromosome 15q25.1 Affect Smoking Quantity: a Meta-Analysis and Comparison with Lung Cancer and COPD. *PLoS Genetics, 6*(8), e1001053. doi: 10.1371/journal.pgen.1001053

Saccone, S. F., Hinrichs, A. L., Saccone, N. L., Chase, G. A., Konvicka, K. (2007). Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum Mol Genet, 16*(1), 36-49. doi: 10.1093/hmg/ddl438

Saccone, S. F., Hinrichs, A. L., Saccone, N. L., Chase, G. A., Konvicka, K. (2007). Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Human Molecular Genetics, 16*(1), 36-49. doi: 10.1093/hmg/ddl438

Shriner, D. (2013). Impact of Hardy–Weinberg disequilibrium on post-imputation quality control. *Human genetics, 132*(9), 1073-1075. doi: 10.1007/s00439-013-1336-x

Shriner, D., Adeyemo, A., Chen, G., & Rotimi, C. N. (2010). Practical Considerations for Imputation of Untyped Markers in Admixed Populations. *Genetic epidemiology, 34*(3), 258-265. doi: 10.1002/gepi.20457

Stellman, S. D., Chen, Y., Muscat, J. E., Djordjevic, I. V., Richie Jr, J. P. (2003). Lung Cancer Risk in White and Black Americans. *Annals of Epidemiology, 13*(4), 294-302. doi: http://dx.doi.org/10.1016/S1047-2797(02)00420-9

Sung, Y. J., Gu, C. C., Tiwari, H. K., Arnett, D. K., Broeckel, U. (2012). Genotype Imputation for African Americans Using Data From HapMap Phase II Versus 1000 Genomes Projects. *Genetic epidemiology, 36*(5), 508-516. doi: 10.1002/gepi.21647

Teo, Y.-Y., Small, K. S., & Kwiatkowski, D. P. (2010). Methodological challenges of genome-wide association analysis in Africa. *Nat Rev Genet, 11*(2), 149-160.

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073. doi: http://www.nature.com/nature/journal/v467/n7319/abs/10.1038-nature09534-unlocked.html#supplementary-information

The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. doi: http://www.nature.com/nature/journal/v491/n7422/abs/nature11632.html#supplementary-information

Thorgeirsson, T. E., Gudbjartsson, D. F., Surakka, I., Vink, J. M., Amin, N. (2010). Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nature genetics, 42*(5), 448-453. doi: 10.1038/ng.573

Tobacco and Genetics Consortium. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet, 42*(5), 441-447. doi: http://www.nature.com/ng/journal/v42/n5/suppinfo/ng.571_S1.html

TPSAC

http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/TobaccoProductsScientificAdvisoryCommittee/UCM247689.pdf 1/7/2016

Trtchounian, A., Williams, M., & Talbot, P. (2010). Conventional and electronic cigarettes (e-cigarettes) have different smoking characteristics. *Nicotine & Tobacco Research, 12*(9), 905-912. doi: 10.1093/ntr/ntq114

Truong, L., Park, H. , Chang, S. , Ziogas, A. , Neuhausen, S. , Wang, S. , Bernstein, L. and Anton-Culver, H. (2015). Human Nail Clippings as a Source of DNA for Genetic Studies. *Open Journal of Epidemiology*(5), 41-50. doi: 10.4236/ojepi.2015.51006

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv*. doi: 10.1101/005165

Visscher, Peter M., Brown, Matthew A., McCarthy, Mark I., & Yang, J. (2012). Five Years of GWAS Discovery. *American Journal of Human Genetics, 90*(1), 7-24. doi: 10.1016/j.ajhg.2011.11.029

Wang, J. C., Hinrichs, A. L., Bertelsen, S., Stock, H., Budde, J. P. (2007). Functional Variants in TAS2R38 and TAS2R16 Influence Alcohol Consumption in High-Risk Families of African-American Origin. *Alcoholism: Clinical and Experimental Research, 31*(2), 209-215. doi: 10.1111/j.1530-0277.2006.00297.x

Winkler, T. W., Day, F. R., Croteau-Chonka, D. C., Wood, A. R., Locke, A. E. (2014). Quality control and conduct of genome-wide association meta-analyses. *Nat. Protocols, 9*(5), 1192-1212. doi: 10.1038/nprot.2014.071

http://www.nature.com/nprot/journal/v9/n5/abs/nprot.2014.071.html#supplementary-information

Zheng, H.-F., Ladouceur, M., Greenwood, C. M. T., & Richards, J. B. (2012). Effect of Genome-Wide Genotyping and Reference Panels on Rare Variants Imputation. *Journal of Genetics and Genomics, 39*(10), 545-550. doi: http://dx.doi.org/10.1016/j.jgg.2012.07.002

Zheng, H.-F., Rong, J.-J., Liu, M., Han, F., Zhang, X.-W. (2015). Performance of Genotype Imputation for Low Frequency and Rare Variants from the 1000 Genomes. *PLoS One, 10*(1), e0116487. doi: 10.1371/journal.pone.0116487

Zheng, J., Li, Y., Abecasis, G. R., & Scheet, P. (2011). A Comparison of Approaches to Account for Uncertainty in Analysis of Imputed Genotypes. *Genetic epidemiology, 35*(2), 102-110. doi: 10.1002/gepi.20552

Curriculum Vitae

Shelina Ramnarine
722 Westgate Ave, Apt A St. Louis, MO 63130 ● shelinasr@gmail.com ● 706.288.4091

SUMMARY OF QUALIFICATIONS

- PhD Candidate with focus in statistical genetics, genetic epidemiology, and biostatistics
- Experience in gene-environment interactions and health care disparities based on racial origin
- Excellent verbal and written communication skills as well as project management and leadership abilities

PROFESSIONAL EXPERIENCE

Washington University in St. Louis, St. Louis, MO 2010 – Present
Lab uses mathematical and statistical methods to study the genetics of complex traits in humans

- *PhD Candidate*
  - Primary research focus: Smoking behaviors, mentholated cigarettes, nicotine dependence, genotype imputation, diverse populations, and gene-environment interactions
    - Identified appropriate statistic for evaluating inferred data
    - Used statistical analyses to identify genetic variants that increase susceptibility to mentholated cigarette use
    - Used MACH, IMPUTE2, BEAGLE, PLINK, SAS, R, and Perl to impute and analyze imputed data
  - Secondary focus: Cancer epidemiology in Trinidad and Tobago
    - Provided analytic support for team in efforts to identify epidemiological landscape of breast cancer in Trinidad and Tobago population sample

- *Summer Research Scholar* 2009
  - Research focused on understanding gene-gene-environment interactions in yeast sporulation efficiency
  - Generated data that elucidated single nucleotide gene-gene-environment interactions

- *Summer Research Scholar* 2008
  - Research focus: Elucidation of the role of Hcm1 on the cell cycle in yeasts
  - Resulted in two first place oral presentation of the results prizes

The University of Georgia, Athens, GA

- *Research Assistant* 2009 - 2010
  - Research focus: Modeling disease transmission in the context of kin selection
    - Demonstrated that modern day allele frequency and disease transmission considering kin selection can be modeled using mathematics and statistics

- *Co-Graduate Teaching Assistant* 2008
  - Developed presentation and communication skills by presenting information to lab class of 16 students
  - Improved ability to explain concepts from different angles and cater to individual needs

EDUCATION

- PhD in Human and Statistical Genetics Jan 2016
  Washington University in St. Louis (Wash U), St Louis, MO

- B.S. in Biology and Statistics 2010
  University of Georgia (UGA), Athens, GA

PROFESSIONAL MEMBERSHIPS

- International Genetic Epidemiology Society 2012 – 2013
- Society for Research in Nicotine and Tobacco 2011 – 2012

PRESENTATIONS

- Improved Criteria for Identifying SNPs that Do Not Impute Well, Poster Presentation in International Genetic Epidemiology Society October 2012, Stevenson, WA

- Methods for Determining Upper Bounds of Imputation Accuracy: Application to Smoking-Associated Genomic Regions, Poster Presentation in Society for Research on Nicotine and Tobacco March 2012, Houston, TX

- Improved Criteria for Identifying SNPs that Do Not Impute Well, Poster Presentation in International Genetic Epidemiology Society October 2012, Stevenson, WA

- Speaker for the Weekly Interdisciplinary Science Colloquia Seminar Series hosted by MBRS-RISE and MARC U*STAR Programs at Morgan State University April 19 2012, Baltimore, MD

- Methods for Determining Upper Bounds of Imputation Accuracy: Application to Smoking-Associated Genomic Regions, Poster Presentation in Society for Research on Nicotine and Tobacco March 2012, Houston, TX

- Methods for Determining Upper Bounds of Imputation Accuracy: Application to Smoking-Associated Genomic Regions, Poster Presentation in Washington University Graduate Research Symposium, February 2012, Saint Louis, MO

- What is the Genetic Basis of Diversity?, 2nd place for Oral Presentation in Life Sciences Cellular and Molecular Biology at the Peach State Louis Stokes Alliance for Minority Participation Fourth Annual Fall Symposium and Research Conference, November 2009, Forth Valley, GA

- The Effect of Family Structure on the Spread of Infectious Disease, 2nd place for Oral Presentation in Mathematics at the Peach State Louis Stokes Alliance for Minority Participation Fourth Annual Fall Symposium and Research Conference, November 2009, Forth Valley, GA

- What is the Genetic Basis of Diversity?, Poster Presentation in Molecular Biological Sciences- Genetics at the Annual Biomedical Research Conference for Minority Students, November 2009, Phoenix, AZ

- Understanding the Cell Cycle: Cyclin-Dependent Kinase Phosphorylation of Hcm1, Poster Presentation The University of Georgia Center for Undergraduate Research Opportunities Symposium, April 2009, Athens, GA

- Understanding the Cell Cycle: Cyclin-Dependent Kinase Phosphorylation of Hcm1, 1st place for Oral Presentation in Life Sciences at the Florida Georgia Louis Stokes Alliance for Minority Participation EXPO, February 2009, Miami, FL

- Understanding the Cell Cycle: Cyclin-Dependent Kinase Phosphorylation of Hcm1, 1st place for Oral Presentation in Life Sciences Cellular and Molecular Biology at the Peach State Louis Stokes Alliance for Minority Participation Third Annual Fall Symposium and Research Conference, November 2008, Savannah, GA

SKILLS AND COMPETENCES

- Software and programming languages: Perl, R, and SAS
- GWAS analysis/tools and statistical genetics software: PLINK, MACH, IMPUTE2, BEAGLE
- Computer skills: Microsoft Word, Excel, PowerPoint

SELECTED LEADERSHIP AND SERVICE EXPERIENCES

- Entrepreneurial Venture: MediMeld *Chief Operations Officer* | Wash U 2014 – Present
- Biotechnology and Life Science Advising Group *Consultant* | Wash U 2014
- Co-Graduate Teaching Assistant | University of Georgia 2008
- Venture Café Youth Entrepreneurship March to May Boot Camp *Mentor* 2015
- International Graduate Student Association Career Development and Networking *Co-President* | Wash U 2012 – 2013

- Black Educational Support Team (BEST) *Counselor*| University of Georgia 2008 - 2010
- Louis Stokes Alliance for Minority Participation Advisory Board *Vice President* | UGA 2009 – 2010
- The Caribbean Student Association (CaribSA) *Vice President* | University of Georgia 2007 – 2008

AWARDS, GRANTS, AND HONORS

- Managing Science in the Biotech Industry *Course Participant* (1 of 40 from 591 applicants) 2015
- Scientist Mentoring and Diversity Program *Scholar* 2014 – 2015
- International Center for Career Development (Sponsored by Janssen, Amgen, and Onyx)
- Interdisciplinary Science Colloquia Seminar Series *Invited Speaker* 2012
- Morgan State University Baltimore, MD
- Advances in Genomics Research Summer Program *Scholar* (NHGRI) 2012
- NSF Graduate Research Fellowship Program *Honorable Mention* 2011

PUBLICATIONS

- Ramnarine S, Chen LS, Culverhouse R, Duan W, Hancock DB, Hartz S, Johnson E, Bierut L, Saccone N *Assessing Genetic Predisposition to Mentholated Cigarette Preference in Nicotine Dependent Smokers* [in preparation]

- Ramnarine S, Zhang J, Chen L-S, Culverhouse R, Duan W, et al. (2015) When Does Choice of Accuracy Measure Alter Imputation Accuracy Assessments? PLoS ONE 10(10): e0137601. doi: 10.1371/journal.pone.0137601

- Hartz SM, Olfson E, Culverhouse R, Cavazos-Rehg P, Chen LS, DuBois J, Fisher S, Kaphingst K, Kaufman D, Plunk A, Ramnarine S, Solomon S, Saccone NL, Bierut LJ Return of individual genetic results in a high-risk sample: enthusiasm and positive behavioral change. *Genetics in Medicine* 2014 August doi:10.1038/gim.2014.110

- Schwantes-An TH, Culverhouse R, Duan W, Ramnarine S, Rice JP, Saccone NL, 2013 Interpreting joint SNP analysis results: when are two distinct signals really two distinct signals? Genet Epidemiol 2013 April;37(3):301-9. doi: 10.1002/gepi.21712.

- Gerke J, Lorenz K, Ramnarine S, Cohen B, 2010 Gene–Environment Interactions at Nucleotide Resolution. PLoS Genet. 2010 September; 6(9): e1001144. doi: 10.1371/journal.pgen.1001144.