Linkage, Association, And Haplotype Analysis: A Spectrum Of Approaches To Elucidate The Genetic Influences Of Complex Human Traits

Amy Elizabeth Webb

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biomedical Engineering.

Chapel Hill
2010

Approved By:
Kirk Wilhelmsen
Kari North
David Lalush
Morgan Giddings
Terry Magnuson

Abstract

Amy Elizabeth Webb
Linkage, Association, And Haplotype Analysis: A Spectrum Of Approaches To Elucidate The Genetic Influences Of Complex Human Disease
Under the direction of Kirk Wilhelmsen

The goal of human genetics is to identify genetic variants that influence a certain trait with the intent to provide a better understanding of the biology behind that trait. As technologies and statistical methods towards this goal have developed, there has been a change in the approaches to identify trait-causing variants. The three projects reported here cover a range of approaches. Early studies focused on family-based data, using linkage analysis to find regions of the genome shared by members with similar trait values. This approach was used to confirm the involvement of CYP2E1 with the level of response to alcohol in sibling pairs with an alcoholic parent. With the advent of high through-put genotyping panels, the field of human genetics has shifted to population-based association studies that seek to find variants that correlate with a trait. This approach was used to search for regions of the genome that infer risk for Pick's disease, a spectrum of heterogeneous dementia diseases, and to reproduce the association with MAPT, a gene with known disease-causing mutations. Haplotype based analysis approaches have emerged to improve the analysis of genomic data. A novel algorithm for haplotype based analysis was developed to identify long haplotypes shared in a population based on genotypes from genome-wide association data and was found to be very accurate when predicting haplotypes within the shared regions. Together, these three projects represent the past, present, and future of the study of human genetics.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ALS – Amyloid Lateral Sclerosis

CBD – Corticobasal Degeneration

CYP2E1 – Cytochrome P450 2e1

FTD – Frontotemporal Dementia

GWAS – Genome-wide Association Study

LCR – Low Copy Repeat

LD – Linkage Disequilibrium

LOD – Log of Odds

LRA – Level of Response to Alcohol

MAF – Minor Allele Frequency

MAPT – Microtubule Associated Protein Tau

PSP – Progressive Supranuclear Palsy

SNP – Single Nucleotide Polymorphism

Chapter 1 – Introduction

The introduction to this dissertation outlines the purpose and significance of this study, defines the problem, and introduces the three projects covered in the chapters to follow.

1.1 Purpose of the Study

The study of the genetics of complex human traits seeks to identify genetic variants that predispose to the trait. Researchers expect that the identification of these genetic variants that predispose to disease traits will elucidate the pathogenesis of disease revealing possible targets for treatment. Additionally, an understanding of disease provides insight into the genetics of normal biological processes.

Mendelian traits have simple patterns of inheritance (e.g. autosomal dominant, sex-linked recessive) and are influenced by a small number of genetic variants that have large effects. In contrast, complex human traits have complex patterns of inheritance due to the effect of multiple genetic variants and gene-environment interactions. Mendelian and complex traits can be vanishingly rare or common, but in the absence of a simple pattern of inheritance, the possibility of a genetic cause for a rare trait is generally not obvious. Typically in complex traits, an accumulation of variants combined with environmental exposures are needed to increase the risk for the trait to a level sufficient for that trait to occur. A continuum exists between complex and Mendelian traits in the number of causal variants and the effect size of the causal markers. A trait that is common, with a larger number of causal variants that each contribute a small effect would have a complex pattern of inheritance, while a trait with a single causal marker that contributes the entire effect would have a Mendelian pattern of inheritance. These patterns of disease characteristics can be used to optimize the study design and statistical approach used to identify possible genetic variants.

To understand the genetic causes of a disease, genetic variants must be identified that occur in diseased or affected individuals. Some studies take a candidate gene approach to identify these genetic variants, where previous evidence, such as the biology underlying the trait or previous genetic studies, is used to target the search on variants or genetic regions most likely to have a direct effect on disease. However, for most diseases we do not completely understand the molecular processes leading to disease, making it difficult to identify possible variants likely to increase the risk for disease. A genome-wide approach instead considers randomly chosen markers at a high density across the genome. By systematically searching throughout the genome for regions with variants that are correlated with a trait or disease, we can identify genes likely to have an effect on the trait without requiring a previous hypothesis.

The purpose of this dissertation was to search for genetic regions that increase risk for neurological phenotypes, specifically dementia and alcoholism, and to evaluate the development of a statistical application that seeks to aid in the identification of regions that influence a trait by considering haplotypes, or combinations of markers inherited together. Specific issues addressed by this study include the differences between linkage and association analysis, locus heterogeneity, linkage disequilibrium, and haplotype phase inference.

1.2 Overview and approaches to the problem

Through the three projects that make up this dissertation, various statistical techniques were applied to different types of genetic data to find regions of the genome that affect a trait of interest. The first two projects focus on techniques used to find regions of the genome more likely to occur in individuals with a certain trait or with extreme values of a trait. A trait, often referred to as a phenotype, is a physical characteristic that can be reliably observed or measured. It can be quantitative as in the case of a continuous measurement of height or qualitative as in the case of a dichotomous disease status. The current dissertation addresses both kinds of traits—a project looking at the spectrum of responses to alcohol and a separate project concerning the disease status of various subtypes of dementia. The final project deals with haplotype phase inference, a technique created as an intermediate step in the identification of regions containing possible causal variants able to deal with problems that obscure localization. The combination of marker alleles on a single chromosome is referred to as a haplotype. Haplotype phase inference seeks to statistically determine this combination of alleles from phase unknown SNP genotype data.

Two approaches used to find correlation between a genetic variant and a phenotypic trait were applied to the projects in this dissertation. The first uses linkage analysis which identifies regions that are shared "identical by descent" or inherited from a common ancestor. Regions shared more often in related individuals with similar trait values than expected by chance are likely to have an influence on the trait. The second approach uses association analysis to look for regions shared “identical by state.” Identity by state means the genotypes are the same in a population-based context but there is not enough information contained in a single genotype to determine whether it was inherited from a common ancestor. Regions shared more often in a population with a certain phenotype when compared to a population without the phenotype are likely to have an influence on the trait. These two approaches differ with regards to the study population, the type of markers measured, the statistical calculation approach, the localization of a causal variant, and the power to detect an effect.

1.2.1 Linkage Analysis – study design and statistical approach

A family based approach is utilized in linkage analysis to look for genetic regions shared between family members with the same trait ultimately identifying regions “linked” with a trait. In a region containing a variant that has a direct effect on a trait, affected family members should share more genotypes identical by descent in the area surrounding that region than expected from chance based on their level of relatedness. For example, we expect two siblings should share 50% of their genome by chance alone. So if concordantly affected sibling pairs show enrichment for sharing in a particular region, something in that region likely has an effect on the trait. For quantitative traits, siblings with less difference in their trait should share more alleles identical by descent in the region surrounding a causal locus and siblings with greater difference should share less. Chromosomal regions where there is a correlation between the chromosome sharing and trait sharing are said to show evidence for linkage and the evidence for linkage can be aggregated across families. DNA sequence variations, often called markers, are used to identify chromosomal regions. Markers are said to be informative in a family if the chromosomes in that family can be distinguished by the sequence variations interrogated by the marker. One class of sequence variants commonly used for linkage analysis has been the microsatellite marker. Microsatellites have short tandem repeats that commonly occur throughout the genome. These markers are highly polymorphic with multiple allelic forms that can easily be traced through families.

While many statistical approaches are available for linkage analysis, in the linkage project that makes up part of this dissertation, a multipoint variance component approach implemented through the computer package SOLAR[2] was used to find regions of the genome correlated with a quantitative trait. Variance component linkage analysis is able to simplify the characterization of a trait by partitioning out the components that affect the trait. The trait is modeled based on the linear function of $Yi = \mu+\beta_j\nu_{ij}+g_i+e_i$ where μ is the population average of the trait, β a regression coefficient for the jth covariate, ν is the value of the jth covariate, g represents the additive genetic effects, and e the unmeasured environmental effects.[1,3] The last two parameters can be estimated through the variance-covariance matrix represented by $\Omega=\Sigma\Pi\sigma^2_{qi}+2\Phi\sigma^2_g+I\sigma^2_e+\sigma^2_{cov}$ where σ^2_{qi} is the additive genetic variance, Π is the estimated number of markers shared identical by descent, σ^2_g is the variance attributable to residual additive genetic factors, Φ represents the kinship matrix, σ^2_e represents the environmental factors, and σ^2_{cov} is the variance due to covariates.[1,3] An added benefit to the variance component approach is opportunity for the inclusion of covariates and the possibility of identifying a covariate that can account for all of the variance in the trait and be classified as causal. A covariate is a predictive variable that can influence a phenotype often independently from genotypic influences. Covariates such as gender or age are included in the model to correct for the influences of these variables. In variance component linkage analysis, the covariate parameter is modeled first, so if all of the variance can be attributable to an included covariate, there is no variance left for the other terms and the LOD score will be reduced to zero. SOLAR considers a multipoint measurement of identity by descent, where markers from the entire chromosome influence the calculation but markers closest to the location of calculation are weighted more highly. This provides more information than the standard two point calculation of IBD. The evidence for linkage in a region is based on the calculation of a LOD score which compares the likelihood of a model assuming linkage with the basic polygenic model with no linkage.[3]

1.2.2 Association Analysis – study design and statistical approaches

Association analysis has quickly become a more suitable alternative to linkage analysis for investigation into the genetics of complex human traits due to advances in multiplex, high throughput genotyping technology and an improved understanding of common human population variation as a result of the International HapMap project.[7] Association analysis seeks to identify genetic changes that are identical by state, focusing on comparing an unrelated population of individuals with a trait (cases) with an unrelated population of individuals without a trait (controls). If the measured allele of a certain variant occurs at a statistically different frequency in the population of cases compared with the population of controls, the variant is said to be "associated" with the trait. The main genetic variant used in these studies is the single nucleotide polymorphism (SNP) which represents a single nucleotide change in the sequence of DNA. By definition, a polymorphism is a genetic variant that has reached a relatively high frequency in a population (minor allele frequency of greater than 1%). In practice, SNP genotyping is limited when the minor allele frequency is less than 5% because the probability of detecting association is very low and often leads to spurious results. SNPs are biallelic, meaning there are two forms present for each SNP in a population, often simplified to an A and B allele. With only two forms, it is impossible to predict whether the alleles from a single SNP shared by two individuals are inherited from a common ancestor. A single SNP provides less information than a microsatellite and many more are needed to provide enough power to detect a trait causing locus. To make up for the low information content in a SNP, high throughput genotyping technology has made it very cost effective to genotype markers throughout the genome with a much denser coverage than allowed by microsatellites.

Two approaches were used to identify genetic regions associated with a trait. The first was to apply a mixed linear model to genotypes from sibling pairs. Since the dataset was family based, a correlation exists between related individuals but the mixed model method was able to account for the correlated family structure.[10] This type of method considers both the correlation within families and the correlation between families to make statistical inferences about the genetic effects of a continuous trait. This method was implemented through the PROC MIXED command in SAS.[24] The approach is similar to the variance component approach of linkage analysis where there exists unknown random variables that can influence the variability of the trait. The mixed model fits the data to a linear model of $y=X\beta+Z\gamma+\varepsilon$ where y is the observed data, β represents the fixed effects parameters, γ represents random-effects parameters, and ε represents unknown random error.

The second approach to measuring the level of association of genetic factors was through the calculation of a Fisher exact test to compare the genotype counts of AA, AB, and BB between populations of cases and controls. The test is used to determine whether the allele frequency in the cases is significantly different from the controls. The Fisher exact test was chosen over a chi squared test due to the possibility of low allele counts for markers with small minor allele frequencies. For the genome-wide association described in this dissertation, the fisher exact test for a two by three contingency table was approximated through the use of permutations.

Genotype determination for the Pick's disease project were made based on genotype calls generated using probe intensity levels from the Affymetrix genotyping chip. A signal is generated after the hybridization of labeled DNA fragments with the complementary probes on the genotyping chip providing an intensity for each genotype measurement. The signal intensity is normalized to correct for both the variation between features on a single chip and variation across different chips containing different samples ultimately generating a measurement corresponding to the amount of each allele in each sample. These intensity measurements for the two alleles are transformed into Contrast (asinh(K(S_a-S_b)/(S_a+S_b))/asinh(K)) and Strength log(S_a+S_b) which represent genotype and brightness respectively. The constant K is termed the stretch factor and is used to increase the distance between genotype clusters creating a balance in the variability between the three clusters and allowing for improved differentiation. The cluster membership of each sample is determined with the Mahalanobis distance defined as sqrt[$(x-\mu)^2\Sigma^{-1}(x-\mu)$] measured between each point and the cluster center. The confidence score or call quality for each genotype measurement is determined based on the ratio between the closest cluster and the next to closest cluster. This approach to genotype determination is implemented in the BRLMM algorithm: Bayesian Robust Linear Model with a Mahalanobis distance classifier.

1.2.3 Genome-wide analysis – why should it work?

In every diploid organism, two copies of each marker are typically present. One is inherited from the maternal chromosome and one from the paternal chromosome. The combination of SNPs on a single chromosome is referred to as a haplotype. If every marker was truly independent, each SNP would need to be genotyped in order to capture all of the variation within the genome. With 14 million validated SNPs cataloged in dbSNP[5] this is not only unreasonable but also unnecessary. There is statistical correlation between nearby SNPs since they are inherited together on a chromosome, referred to as linkage disequilibrium or LD. Markers located closely together on a chromosome will more often be inherited together, while markers farther apart are more likely to be separated by recombination. As a result, markers located in a region surrounding a causal variant will show association to the trait of interest even if the causal variant is not genotyped.[6] This association will wane as the distance from the causal variant increases as there is less correlation due to the recombination that occurs though generations.

Single marker association (considering a single SNP at a time) takes advantage of linkage disequilibrium to find an association with markers that may or may not have been directly genotyped. The length of a region that is correlated or in complete linkage disequilibrium depends on the number of generations that have passed from the nearest common ancestor. For a linkage study focusing on sibling pairs, that nearest common ancestor would be the parent, only one generation away. But for association analysis, the nearest common ancestor may be hundreds to thousands of generations away. Since shorter genetic regions are correlated this allows for a finer localization of the trait causing locus. So after the identification of an associated marker, the region that could possibly contain a causal variant is smaller and easier to systematically search through than a region provided by linkage analysis. Patterns of LD in many world populations have been catalogued with the International HapMap project[7] allowing for the creation of efficient SNP genotyping panels that can predict or "tag" nearby common variants based on the localized LD patterns inherent in each population.[8]

Although markers or regions can be shown to be "linked" or "associated" with a trait, it provides little indication regarding causality. Genome-wide analysis takes advantage of the linkage disequilibrium as described above to locate genetic regions that could contain causal variants. Most markers used in these studies have no effect on the amino acid sequence of a protein due to either the redundancy of the amino acid codons or that they lie in intergenic or intronic regions with no obvious connection with protein expression. More often, the markers identified in genome-wide studies do not play a role in the trait, but instead represents genetic variation in the surrounding region that increases risk.[6]

Perhaps the most important question regarding genome-wide analysis is not "Why should it work?" but instead "Why does it not work?" The reality is that even though many studies are performed with the intention of understanding the genetic influences of disease, most of these studies provide inconsistent results because of unaccounted heterogeneity and poor power. The final chapter of this dissertation details many of the problems facing genome-wide studies and why so many of them present conflicting results.

1.2.4 Haplotype Analysis – The best of both worlds

A number of algorithms have been created to apply statistical methods of linkage and association to different types of genotype data to solve genetic problems. One option for obtaining more information from genetic data is to consider multiple genotype measurements together. While many genetic association studies focus on one marker at a time, these markers are inherited as a unit and interact in complex ways. A haplotype refers to the set of alleles for nearby SNPs that are inherited together on a chromosome and the haplotype phase refers to the determination of a haplotype or the placement of alleles together along a chromosome. The third project in this dissertation was performed to evaluate a novel algorithm created to determine haplotype phase on genome-wide association datasets and to understand the accuracy of this new method compared to standard haplotype phase inference programs.

Haplotype based analysis seeks to find an association between the ancestral haplotype harboring a causal variant inherited by individuals with a trait of interest.[9] When a mutation arises in an individual that causes a certain trait, it is contained on a chromosome and creates a new chromosome length haplotype. As it is passed through to further generations of the population, the full length of the chromosome will be eroded away due to recombination. After many generations, individuals with that trait caused by the mutation started in the first individual, or founder, will have varying lengths of the original haplotype (or founder chromosome) surrounding the variant, assuming the mutation survives selection.[9] The intersection of these haplotypes will map the trait-causing variant, usually to a finer region than by single marker analysis improving the probability of localizing a causal variant.

Haplotype analysis combines the population based approach of association analysis with the search for regions identical by descent of linkage analysis. While SNP genotyping represents common variation in the genome, haplotypic analysis allows for the consideration of untyped rare variants including those with low frequency or recent mutations that could be hidden on a haplotype and not be “tagged” well by a single SNP on conventional genotyping panels.[11] Considering multiple markers together as a haplotype provides a great deal of benefit to genetic association studies not only in terms of the power to detect an association with a trait, but also with the possibility for providing useful insights of the evolutionary history of human populations and complex patterns of linkage disequilibrium allowing for an understanding of natural selection, recombination rates, and patterns of migration.[8,12]

1.2.5 Looking ahead

When viewed together, the three main studies that make up this project represent the natural progression in the study of complex human disease, addressing issues relating to the ever-changing technological advances allowing for the measurement of more genetic units in larger sets of individuals and ways to understand the effect of multiple genetic units inherited as a haplotype. Throughout the history of human genetics, early studies used linkage analysis to investigate sparse maps to identify variants that cosegregate with a trait through a limited number of generations within families. As time progressed, the field has moved towards the consideration of dense genetic maps measured in large unrelated populations to understand the common causes of common disease through association analysis. Going into the future, there is already a trend towards whole genome sequencing to allow for the consideration of all common and rare variants. New scientific advances in genotype technology provide more information to add to our understanding of complex human traits, but also create more problems related to data management and ways to deal with confounding effects. The last chapter of this dissertation provides a discussion about the future trends in the study of complex human disease.

1.3 Research Objectives

The research objectives that constitute this dissertation are divided into three distinct projects. While there is little overlap between the subject matters and analysis approaches, all three projects represent a continuum of the overarching methods for the study of the genetics of complex human disease.

1.3.1 The investigation of CYP2E1 with the level of response to alcohol

A genome-wide linkage study was performed to search for regions of the genome that confer risk to alcoholism as measured by the level of response to alcohol after an alcohol challenge. Results from the genome-wide study led to the consideration of CYP2E1, a gene with known involvement with the metabolism of ethanol. To further understand the relationship between CYP2E1 and the level of response to alcohol, both linkage and association analyses were applied to a combined map of microsatellite and SNP markers. Variance component linkage analysis supported the linkage shown at the end of chromosome 10 from the original study. However the addition of a second set of samples reduced the significance of the linkage signal. An investigation of possible locus heterogeneity led to the discovery of a single family with unreliable phenotype data that was responsible for the reduction of signal. Association analysis was performed on the SNPs genotyped in CYP2E1. The best evidence for association came from a marker upstream of the CYP2E1 promoter. Combined linkage and association was performed by including this associated marker as a covariate in variance component linkage, but this analysis was unable to definitively implicate the marker as a causal variant.

1.3.2 The association of the MAPT region with Pick's complex diseases

A genome-wide association analysis was performed on a number of different, but related neurodegenerative diseases collectively referred to as tauopathies due to the aggregation of tau proteins commonly found in diseased neurons. Mutations had previously been identified in affected cases in the gene that encodes the tau protein, MAPT. The MAPT gene is located on an inversion on chromosome 17 resulting in a high degree of linkage disequilibrium between markers across the inverted interval limiting the number of possible haplotypes. A single haplotype, H1, was found to be overrepresented in certain subtypes of disease (specifically CBD and PSP). The current genome-wide association study was able to replicate the overrepresentation of the H1 haplotype in PSP and CBD cases when compared to controls and show a constant high level of association for the inverted region for both PSP and PSP combined with CBD. This positively replicated association provides increased confidence for the results generated from the GWAS in other regions of the genome (whole genome results not reported at this time).

1.3.3 The evaluation of a novel algorithm for haplotype phase inference

A novel haplotype phase inference algorithm called Convergent Haplotype Association Tagging, or CHAT, was created to determine the haplotype phase of a population of unrelated individuals genotyped for genome-wide association. This new algorithm bases phase inference on the identification of subsets of individuals that share a region of the genome identical by descent allowing for the generation of a consensus haplotype for each region of sharing. The complementary haplotype for each individual in the subset can be easily identified and added to the set of known haplotypes. The algorithm was created to outperform existing packages that tend to perform poorly across long regions and across recombination hotspots. Reported is the evaluation of the novel phase inference algorithm compared to three publicly available phase inference packages. Each algorithm was applied to simulated datasets created under different test conditions to understand how each program performs in regards to selection, degree of linkage disequilibrium, sample and marker size, and the imputation of missing genotypes. CHAT demonstrated an improved single site error rate compared to the alternative haplotype phase inference algorithms and an improved switch error compared to ENT when considering a dataset with a large number of samples. CHAT performed best with a larger number of samples but needs improvement in coverage to be able to compete with current haplotype phase inference programs and be practical for haplotype-based association mapping.

1.4 Summary of the Chapter

Chapter one introduced the research topic of this dissertation by briefly describing the problem facing the study of the genetics of complex human disease, discussing the techniques used throughout this dissertation for finding regions of the genome that confer risk to disease, and providing an overview of the three research problems covered in the following chapters.

This dissertation will be organized into five chapters. Chapter one presented the relevant background and introduced the three research projects. Chapter two describes the combined linkage and association analysis performed as a follow-up to understanding the involvement of the gene CYP2E1 with the level of response to alcohol and thus the risk for alcoholism. Chapter three describes the results from a genome-wide association study on a number of related neurodegenerative diseases, focusing on the region containing MAPT which shows association with the two diseases that more commonly include aggregation of the MAPT protein product. Chapter four describes the evaluation of a novel haplotype phase inference algorithm that bases phase inference on the identification of long haplotypes shared in a population and compares the accuracy of this algorithm with standard haplotype phase inference programs. Chapter five provides a final discussion of the three research projects focusing on the limitations facing the detection of genetic variants that cause disease and future directions in the field of the genetics of complex human disease.

1.5 References

1: Almasy L, Blangero J. Human QTL linkage mapping. Genetica. 2009 Jun;136(2):333-40. Epub 2008 Jul 31. Review.

2: Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 1998 May;62(5):1198-211.

3: Broeckel U, Maresso K, Martin LJ. Linkage analysis for complex diseases using variance component analysis: SOLAR. Methods Mol Med. 2006;128:91-100.

4: BRLMM:an Improved Genotype Calling Method for the GeneChip Human Mapping 500K Array Set. White Paper from Affymetrix.com. Apr 2006

5: dbSNP online

6: Gusev A, Măndoiu II, Paşaniuc B. Highly scalable genotype phasing by entropy minimization. IEEE/ACM Trans Comput Biol Bioinform. 2008 Apr-Jun;5(2):252-61.

7: International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789-96.

8: Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, Donnelly P; International HapMap Consortium. A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet. 2006 Mar;78(3):437-50. Epub 2006 Jan 26.

9: Morris RW, Kaplan NL. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol. 2002 Oct;23(3):221-33.

10: Overview: Mixed Procedure. SAS online users manual.

11: Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev. 2009 Jun;19(3):212-9. Epub 2009 May 28. Review.

12: Zhao H, Pfeiffer R, Gail MH. Haplotype analysis in population genetics and association studies. Pharmacogenomics. 2003 Mar;4(2):171-8. Review.

Chapter 2 - The investigation of CYP2E1 in relation to the level of response to alcohol through a combination of linkage and association analysis

2.1 Abstract

A low level of response to alcohol during an individual's early experience with alcohol is associated with an increased risk for alcoholism. A family-based genome-wide linkage analysis using sibling pairs that underwent an alcohol challenge where the level of response to alcohol was measured with the Subjective High Assessment Scale (SHAS) implicated the 10q terminal region. *CYP2E1*, a gene known for its involvement with ethanol metabolism, maps to this region. Variance component multipoint linkage analysis was performed on a combined map of single nucleotide polymorphism (SNP) and microsatellite data. To account for the heterogeneity evident in the dataset, a calculation assuming locus heterogeneity was made using the HLOD (heterogeneity LOD) score. Association between SNP marker allele counts and copy number and SHAS scores were evaluated using a mixed model regression. Linkage analysis detected significant linkage to *CYP2E1* which was diminished due to apparent locus heterogeneity traced to a single family with extreme phenotypes. In retrospect, circumstances recorded during testing for this family suggest that

their phenotype data are likely to be unreliable. Strong allelic associations were detected for several *CYP2E1* polymorphisms and the SHAS score. DNA sequencing from families that contributed the greatest evidence for linkage did not detect any changes directly affecting the primary amino acid sequence. With the removal of a single family, combined evidence from microsatellites and SNPs offer significant linkage between the level of response to alcohol and the region on the end of chromosome 10. Combined linkage and association indicate that sequence changes in or near *CYP2E1* affect the level of response to alcohol providing a predictor of risk for alcoholism. The absence of coding sequence changes indicates that regulatory sequences are responsible. Implicating *CYP2E1* in the level of response to alcohol allows inferences to be made about how the brain perceives alcohol.

2.2 Introduction

While a number of phenotypic factors can affect the risk for alcoholism, one of the most studied endophenotypes is an individual's level of response to alcohol during their early experience with alcohol.[11] The level of response to alcohol can be reliably measured with the Subjective High Assessment Scale (SHAS) during an alcohol challenge or by the Self-Rating of the Effects of Alcohol (SRE) which uses recall to establish the number of drinks required to reach an effect. Children of alcoholics have a greater risk for alcoholism when they have a lower level of response.[32,35,36] A low level of response established early in an individual's drinking career can lead to higher future drinking levels.[8,11,42] Populations at historically higher risk for alcoholism, such as Native American or Korean, need to consume larger amounts of alcohol to become intoxicated[9,24,43] compared to those with lower risk[28] who exhibit a more intense level of response to alcohol. Several studies have implicated genes showing association with the level of response to alcohol (*GABA*, *5-HT*, and *KCNMA1*).[4,7,34] The evidence proving association for these genes is weak by current standards that have been developed as a consequence from the technological advances enabling genome-wide association studies. Even though these genes may affect the level of response to alcohol to some degree, it is possible that these reported associations reflect the typical reporting bias seen in candidate gene studies.

Initially, data were collected from 139 sibling pairs.[46] Variance component analysis found a significant LOD (Log of Odds) score peak of 3.2 for the SHAS score at the 10q terminal region. Of the genes located at 10qter, *CYP2E1* has a known involvement with ethanol metabolism. The CYP2E1 enzyme metabolizes ethanol and acetaminophen, as well as many toxicologic and carcinogenic compounds and can be induced by ethanol and nicotine.[39] In the second stage of the study, when 99 newly collected sibling pairs were added,[34] the peak at 10qter was significantly diminished. As will be described in this paper, it was initially assumed that the diminishment was due to locus heterogeneity, but ultimately the reduced evidence for linkage was explained by a single family with extreme and unreliable phenotypes.

Most of the ethanol that is consumed is oxidized by the liver using alcohol dehydrogenase (ADH). At the high concentrations associated with chronic alcohol consumption, metabolism of ethanol to acetaldehyde increases while the subsequent conversion into acetate is decreased, leading to even higher levels of acetaldehyde. It was shown in rats that chronic consumption reduced the oxidation of acetaldehyde in the liver, thus providing an explanation for the high blood acetaldehyde levels measured after chronic use in human subjects.[19] Acetaldehyde is toxic and highly reactive,[50] binding to nearby proteins thus creating an antibody response, decreased DNA repair, and glutathione depletion ultimately reducing the ability of the liver to clear free radicals.[20] As a result from the oxidation, NAD^+ is reduced with the addition of an electron to form NADH[50] used by mitochondria for ATP synthesis. At high concentrations, ethanol is oxidized by ADH at a higher rate leading to an increase in the $NADH/NAD^+$ ratio.[50]

CYP2E1 is part of the Microsomal Ethanol Oxidizing System (MEOS) accounting for up to 10% of ethanol oxidation in the liver.[39] Once the ADH pathway becomes saturated due to high ethanol concentrations, the MEOS pathway activity increases.[20] By the MEOS pathway, CYP2E1 metabolizes ethanol and other substrates into toxic metabolites creating free radicals in the form of reactive oxygen (O_2) intermediates creating oxidative stress leading to liver damage. CYP2E1 uses O_2 to oxidize ethanol to aldehyde and NADPH to $NADP^+$. While generally used biosynthetically, NADPH can be regenerated from $NADP^+$ with the conversion of NADH to NAD^+. In the absence of NADPH, oxidation of ethanol to aldehyde by CYP2E1 results in superoxides.[50] An excess of reduced NADH in addition to the increased activity of hydrogen shuttles in mitochondria, results in an increased intake of electrons leading to an increase of superoxide anions.[37] The increased creation of Reactive Oxygen Species, or ROS, as a result from the shift in cellular redox state, coupled with the reduced ability to clear these free radicals, due to the increase in acetaldehyde, is thought to be a major driving force in the development of alcohol related liver disease.

The catalase pathway can oxidize ethanol in conjunction with hydrogen peroxide generating systems, such as NADPH oxidase.[50] The catalase pathways plays a larger roles in the oxidation of ethanol in the brain, where little ADH oxidation occurs.[50] In a study by Vasiliou et al.,[40] it was found that animals with a knockout of either catalase or CYP2E1 were more sensitive to the sedative effects of ethanol than control, wild-type animals. The study found that CYP2E1 did not contribute significantly to ethanol clearance in the brain, but was instead involved with ethanol processing in the brain affecting sensitivity.

High ethanol concentrations can interfere with the ability of CYP2E1 to metabolize other substrates due to competition from the shared oxidation pathway leading to reduced drug clearance and elevated drug concentrations.[39] The interaction of certain drugs with alcohol will lead to a long-lasting, enhanced drug effect, often leading to overdose. A similar relationship is thought to exist with nicotine. It has been shown that smokers have a more rapid ethanol clearance than non-smokers, suggesting a biological basis for the correlation of tobacco and alcohol consumption seen in alcoholics.[33]

A number of polymorphisms in *CYP2E1* have been tested in relation to alcoholism and a number of related disorders, including many types of cancer, with varying, often conflicting, results. Carriers of the c2 allele of CYP2E1*5B have increased risk for alcoholic liver disease and are more likely to consume excessive amounts of alcohol possibly due to the higher transcriptional levels of CYP2E1 seen with this allele.[10,30,44] Variants in the gene have been implicated in the increased risk of different types of cancer relating to the respiratory and digestive systems.[5,21,49]

Due to the previously described relationship between *CYP2E1* and the metabolism of ethanol and positive linkage results concerning the level of response to alcohol in relation to alcoholism, a number of single nucleotide polymorphisms (SNP) were genotyped in the *CYP2E1* to further elucidate the gene’s role in alcohol response. Both genotype and copy number were tested for association with the level of response to alcohol as measured by the

SHAS questionnaire. Combined linkage and association analysis was performed to determine whether a single marker or haplotype could account for the linkage signal seen at 10qter.

2.3 Methods

2.3.1 Alcohol Challenge

The data collection protocol was approved by the Human Subjects Protection Committee at the University of California in San Diego and used written, informed consent. The design for the alcohol challenge is fully described in the initial report by Wilhelmsen et al..[46] Male and female subjects ranging in age from 18 to 29 years old were recruited from a population of college students. Chosen sibling pairs reported having an alcohol dependent parent, but were not alcohol dependent themselves. The siblings included 43.7% males and 56.3% females. They had an average age of 22.4 years and 14.2 years of education. 72.2% were Caucasian, 20.0% were Hispanic, and 7.8% were African-American. For 85.0% of the subjects, the alcohol-dependent parent was the father, whereas for 4.4% it was the mother, in 4.0% it was both parents, and in 6.6% the more intensive interview revealed that neither parent met full criteria for dependence.

To measure each participant's response to alcohol, the Subjective High Assessment Scale (SHAS) questionnaire was administered. For the challenge, each subject was given 8 minutes to consume a 20% by volume solution of 95% ethanol, at 0.75 ml/kg for women and 0.9 ml/kg for men. Baseline levels for each score (SHAS, body sway, and breath alcohol level) were measured prior to the challenge and then were measured at multiple set time intervals throughout the 3 hour challenge. Ultimately the changes in SHAS score and body sway at 1 hour after the challenge were used as phenotypes for the genome-wide genetic analysis. Genotyping was performed on 811 microsatellite markers across the genome.

2.3.2 Taqman Genotyping

Genomic DNA was extracted from whole peripheral blood samples. Genotyping was performed on 10 SNPs with Taqman genotyping assays using locus specific PCR primers and fluorescent allele specific probes designed by Applied Biosystems. Standard Taqman protocol was followed and endpoint amplification intensity was measured by the 7900 ABI Sequence Detector. The position of the genotyped markers in relation to *CYP2E1* can be found in Figure 2.1. The HapMap Consortium reported three major haplotypes in the Caucasian population, as seen in Figure 2.1, which could be distinguished by the initial two SNPs that were genotyped. Table 2.1 lists the names and positions of the genotyped SNPs.

2.3.3 Copy number analysis

The copy number of *CYP2E1* was determined for each sample in quadruplicate through the amplification of both a probe specific to *CYP2E1* and a standard probe by real-time PCR using the standard Gene Dosage protocol provided by Applied Biosystems.[22] Preliminary amplification showed the two probes used for analysis had different efficiencies of amplification, which was corrected by a standard dilution curve added to each plate. The fold increase after n number of cycles was calculated by $(\text{efficiency})^n$ and the ratio of this increase between the target and reference genes provided copy number. Standard copy number quantification assumes equal amplification efficiencies, but this is not always a valid assumption. Correcting for even small differences in amplification efficiencies leads to less variability among the quadruplicate samples and lowered standard error in overall copy number determination.

2.3.4 CYP2E1 resequencing

Index cases from the 96 families with the greatest evidence for linkage to the 10q terminal region were selected for resequencing. Each coding sequence exon was resequenced using primers from Applied Biosystems using the standard provided procedure.

2.3.5 Linkage allowing for heterogeneity

A map of the positions of the genotyped SNPs relative to the microsatellite markers was created with Fastlink. Variance component methods were used to recalculate LOD scores using SOLAR v4.0.7 with the identity by descent provided through pedigree information and estimating multipoint identity by descent sharing probabilities.[2] Variance component linkage analysis uses correlation in the phenotype to partition out variance between relative pairs into the effects of the genes in the region of interest, additive genetic effects of other genes, and non-shared environmental variance. The trait is modeled based on the linear function of $Yi = \mu+\beta_j\nu_{ij}+g_i+e_i$ where μ is the population average of the trait, β a regression coefficient for the jth covariate, ν is the value of the jth covariate, g represents the additive genetic effects, and e the unmeasured environmental effects.[2] The last two parameters can be estimated through the variance-covariance matrix represented by $\Omega=\Sigma\Pi\sigma^2_{qi}+2\Phi\sigma^2_g+I\sigma^2_e+\sigma^2_{cov}$ where σ^2_{qi} is the additive genetic variance, Π is the estimated number of markers shared identical by descent, σ^2_g is the variance attributable to residual additive genetic factors, Φ represents the kinship matrix, σ^2_e represents the environmental factors, and σ^2_{cov} is the variance due to covariates.[2] The evidence for linkage in a region is based on the calculation of a LOD score which compares the likelihood of a model assuming linkage with the basic polygenic model with no linkage.[2] To account for the heterogeneity evident in the dataset, a calculation assuming locus heterogeneity was made using the HLOD (heterogeneity LOD) score to identify the cause of the lowered peak at the end of chromosome 10 observed after the addition of samples to the dataset.

2.3.6 Association analysis

Association between SNP marker allele counts and copy number and SHAS scores were evaluated using a mixed model regression through the SAS statistical package testing for statistical inferences using a generalization of the standard linear model. The mixed model fits the data to a linear model of $y=X\beta+Z\gamma+\varepsilon$ where y is the observed data, β represents the fixed effects parameters with design matrix X, γ represents random-effects parameters with design matrix Z, and ε represents unknown random error. Family ID and marker genotype were used as classification variables and the effects of copy number and genotype were modeled against the SHAS score. Genotypes were classified based in the count of the minor frequency allele.

2.3.7 Combined Linkage and Association

Combined linkage and association analysis using SOLAR v4.0.7 was performed to include identity by state information similar to the approach used by Almasy et al.[1] where the variance in the SHAS score that cannot be accounted for by the covariate parameter based on the number of minor alleles was decomposed into the standard variance components. To see whether the linkage signal could be explained by the allele effects of a single SNP, each SNP marker was individually tested by including the number of minor alleles as a covariate. Polygenic covariate screening was used to calculate the significance level after the inclusion of any particular SNP and multipoint analysis was used to calculate the multipoint LOD score for the SHAS score.

By combining linkage and association approaches, a disease loci position can be confined to a region finer than linkage analysis alone and avoid false positive association results due to admixture. Assuming the linkage is not over-estimated, if a measured variant is the actual functional variant affecting the phenotype and no other variants nearby confer any additional risk, linkage analysis conditional on the genotype of such a variant should provide no evidence for linkage. However if the suspected variant is in some degree of linkage disequilibrium with the actual causal variant, the evidence for linkage will be reduced proportional to the degree of LD.

2.4 Results

With a combined map of microsatellite and SNP markers the dataset was reanalyzed, as described in the original linkage study,[34,46] using SOLAR. When divided into the two stages of sample sets from the previous study, significant linkage was found in stage 1 samples with a peak LOD score of 3.14; however, when combining the initial 139 sibling pairs with the additional 99 sibling pairs in stage 2, the linkage signal lowered to a peak LOD score of 1.61. The LOD score plot can be seen in Figure 2.2.

Multipoint linkage analysis allowing for locus heterogeneity was performed for the SHAS phenotype using microsatellite and SNP data from 10qter. The family specific heterogeneity score (α) and LOD score was inspected for each family. Most of the families had α scores of 0.99, one family stood out with an α of 0.37 indicating poor support for linkage to the region. The individual pedigree LOD score for this family alone was -0.97 accounting for most of the reduction in LOD score seen between stage 1 and stage 2. Based on the level of alleles shared between siblings in this family, it was expected that these siblings should have more similar SHAS scores. The sibling pair had a large difference in phenotype despite having inherited the same chromosome Identical By Descent from both of their parents (IBD of 2). While one sibling reported a SHAS score of 26.75 (z = 2.147), the other reported a SHAS score of 4 (z = -0.7161).

Comments from the observers of the study propose phenotyping error as the cause of the extreme phenotype difference. The more sensitive sibling felt nauseated during the challenge. This is a common response reported by many subjects, and the SHAS score for this individual is around the same value reported by other individuals with a similar response. The other sibling, with SHAS scores indicating they were insensitive to alcohol, had a blood alcohol level close to the predicted value, indicating that they were appropriately dosed. This sibling fainted briefly during blood draw. The researchers involved in testing these subjects take special care to avoid fainting and report that it only happens 2% of the time. Although the subject woke up quickly and admittedly felt fine, this fainting spell could have contributed to the low response indicated by a low SHAS score. Therefore it is likely that the reduction in LOD score by the inclusion of the discordant family is due to a phenotyping error. While the inclusion of this family has dramatic effect on the linkage analysis it has a negligible effect on the association analysis. Due to irregularity from the reported testing, the whole family was removed from analysis. When the single family was removed, the original LOD score peak was reestablished with the maximum LOD score for 10qter at 3.40.

SNP genotypes could not be called automatically using the manufacturer supplied program due to copy number differences which affected the allele signal intensities causing samples to fall between the heterozygous and homozygous genotypes. Genotype calls were made manually, assuming that the intermediate samples were either AAB or ABB depending on the location relative to the heterozygote cluster. The best estimate of copy number and genotype was made by integrating the real-time PCR measured copy number, Taqman derived genotype and pedigree structure. While the majority of subjects (85%) have 2 copies, 11% had 3 copies, and 4% had 1 copy. A small number (<2%) were considered to have greater than 3 copies based on real time PCR, but these measurements were not considered plausible based on Taqman derived genotype and pedigree information.

Mixed model regression analysis which controlled for the relatedness of subjects within families was used to investigate the association between copy number genotype and level of response to alcohol. Copy number had little effect on the level of response to alcohol. The presence of at least one copy of a relatively rare allele for several SNPs is associated with a more intense response to alcohol. The best evidence for association was found for the first three markers, which lie near the beginning of the gene near the promoter region, when considering genotype alone. The SNP rs10776687 showed the greatest evidence for association with a p value of 0.007 and an odds ratio of 2.893 (1.476-5.669 95% CI). Odds ratios for all other SNPs were not significant. In this case, copy number was ignored and all genotypes were assigned as biallelic. When considering copy number and genotype together, none of the markers were significant. Copy number alone (1, 2, or 3) was not significantly associated. Interestingly, when modeling the effect of genotype and copy number on the number of cigarettes smoked per day, copy number of *CYP2E1* was associated with this smoking phenotype with an average p value of 0.014. Association p values for the regression analysis are shown in Table 2.2.

Intuitively we would expect that inclusion of a causal SNP as a covariate would reduce the residual linkage due to IBD to zero,[1] but this is an area of active research. After each SNP marker was separately tested in the variance component model, it was found that inclusion of the number of minor alleles of any single SNP was not able to explain all variation in the SHAS phenotype. When considering the dataset after removing family 44, the peak LOD score was 3.36. The marker that lowered this LOD score the most (by 1.21 LOD units) when included as a covariate was rs10776687, the marker located closest to the linkage peak. Other SNPs lowered the LOD score by lesser amounts, as seen in the Table 2.3 below. Combined linkage and association analysis indicates that no single locus tested is likely to be the only causal allele. When testing haplotypes, none of the three haplotypes were able to completely account for the linkage signal.

In addition to testing SNP markers to determine whether a single SNP could directly influence the SHAS phenotype, a number of smoking and drinking phenotypes were analyzed by including them as covariates when modeling their relationship with SHAS. Three of these phenotypes (average number of cigarettes per day, the maximum amount drank in one day, and the average number of drinks consumed on days the subject consumed alcohol) improved the overall peak LOD score compared to the model without covariates. The other tested phenotypes (number of days the subject drank in the last week, amount consumed 24 hours prior to challenge, number of days of smoking in a month, and number of cigarettes smoked on days where the subject smoked), lowered the peak LOD score presumably because these phenotypes are correlated with the SHAS score. All LOD scores are shown in the Table 2.4 below.

To examine whether previously unidentified missense sequence changes could be responsible for the association detected, coding sequence exons were resequenced for 96 index cases from the families with the strongest evidence of linkage. No missense changes were found and no novel polymorphisms were identified.

2.5 Discussion

Alcoholism is a complex disease with potentially many genetic influences. Investigators have tried to minimize heterogeneity by choosing a narrowly defined phenotype such as the level of response to alcohol that was measured with the SHAS score in this study. Strong evidence for linkage to 10qter was observed for the SHAS score in subjects from an alcohol challenge only after the removal of one family that retrospectively should not have been included in the analysis. After linkage was found at the end of chromosome 10, several SNPs genotyped in *CYP2E1* were associated through mixed model regression with the level of response to alcohol as reported by the SHAS score. Copy number did not appear to affect the SHAS score even though copy number differences were found between individuals across the *CYP2E1* gene.

When considered separately, linkage was found over the region containing *CYP2E1* and SNPs from the gene were found to be associated with the SHAS score. If inclusion of the causal variant as a covariate in variance component analysis always reduces the residual LOD score to zero, we were unable to implicate a causal variant directly influencing the level of response to alcohol in the families studied. It is possible that an unidentified polymorphism nearby could play a causal role in the level of response to alcohol as the degree of signal reduction is largest in the marker closest to the linkage peak and then declines for markers farther away. Another possibility is that a single marker cannot account for the entire linkage signal because many markers in the region play a role in the response. Instead of a single polymorphism causing variation, combinations of polymorphisms across the region may work together to contribute to the variation seen in our dataset. Support from the heterogeneity LOD score calculation showing that all families showed evidence for linkage combined with the independently derived association analysis imply that the LOD score peak was not over-estimated. It still can be concluded the regulatory sequences near *CYP2E1* appear to play a role in the level of response to alcohol.

Variance component linkage analysis for the level of response to alcohol was significantly affected by including covariates for recent drinking and smoking behavior. Since *CYP2E1* expression is inducible by alcohol and nicotine,[15] this further supports the role of *CYP2E1* in level of response to alcohol. *CYP2E1* represents a metabolic intersection between these substances of abuse.[33] It was initially surprising that while an association was not found between the level of response to alcohol and copy number of *CYP2E1*, an association was found between nicotine use and copy number. Studies have shown that neither ethanol nor nicotine increase the level of CYP2E1 mRNA in rat hepatic tissue.[12] Ethanol likely changes the activity of CYP2E1 by interacting with the active site leading to increased protein stabilization and reduced clearance by degradation. Given that the induction of CYP2E1 by nicotine requires multiple doses and does not interact with the catalytic function of *CYP2E1*, it is thought that the mechanism behind nicotine induction is not through protein stabilization.[26,33] Since the molar concentrations are vastly different it is unlikely that both nicotine and alcohol could stabilize CYP2E1 by the same mechanism. Since ethanol and nicotine likely induce CYP2E1 through different mechanisms, these two drugs may have an additive effect on CYP2E1 induction and function.[33] The ability to induce CYP2E1 activity by nicotine, but not alcohol, could be dependent on basal transcription rates that could be affected by gene copy number.

The four polymorphisms commonly tested in *CYP2E1*, CYP2E1*5B (c2), CYP2E1*6 (C), CYP2E1*1B (A1), and CYP2E1*1D (1C), have been found to be associated with alcoholism and related disorders in a number of studies. Several of these variants are rare in the Caucasian population (see below). Carriers of the c2 allele of *5B have often been found to have increased risk for alcoholic liver disease.[10] possibly due to the increased tendency to consume excessive amounts of alcohol.[30] The C allele of *6 was shown to be associated with the predisposition for alcoholism in Japanese men.[14] The A1 allele of *1B was found to have a significantly higher allele frequency in alcoholics than in nonalcoholic individuals from a Mexican Indians population.[27] The 1D variant allele was shown to be associated with elevated CYP2E1 activity after alcohol consumption.[25] For every association found with *CYP2E1* variants, a number of studies found no association between the variants and alcohol consumption or risk of alcoholism which could be due to differing phenotype categorization or population allele frequencies.[6,13,29,31,41]

Of the markers measured in the current study, the most associated SNP with level of response to alcohol, rs10776687, is in complete linkage disequilibrium (LD) with the c1 allele of CYP2E1*5B, rs2031920, implying that this marker is associated with the level of response to alcohol as well. A homozygous genotype of the minor allele c2 of CYP2E1*5B is associated with an increase in gene transcription.[44] Another marker, rs2515641, is in complete LD with rs2070676, also known as CYP2E1*1B.

As *CYP2E1* is involved with the metabolism of many carcinogenic compounds, it is not surprising that variants in the gene have been implicated in a number of different types of cancer. The generation of ROS as a result of CYP2E1 oxidation will lead to the creation of lipid peroxidation products such as 4-hydroxynonenal which reacts with DNA to form DNA adducts leading to highly mutagenic cells resistant to apoptosis.[6] The metabolism of procarcinogens by CYP2E1 commonly found in alcohol, tobacco, and industrial chemicals can be enhanced through chronic ethanol.[3]

While a number of *CYP2E1* variants have been analyzed in relation to cancer development, CYP2E1*5B is most often considered. Many of these associations are enhanced by alcohol or nicotine intake which further supports the role of *CYP2E1* in the metabolism of these substances. The c1/c1 genotype of the CYP2E1*5B variant increased risk of hepatocellular carcinoma in smokers from a Taiwanese population[49] and oral cavity cancer in heavy smokers from Caucasians and African Americans populations.[21] Conversely other studies have found evidence for the minor c2 allele leading to an increased risk of hepatocellular carcinoma in ethanol users with chronic liver disease and oral cavity cancer in combination with heavy drinking.[5] Others have found no association between the CYP2E1*5B variant and the same types of cancer including a number of studies for hepatocellular carcinoma.[17,18,47] Many *CYP2E1* association studies did not detect an association because the c2 risk allele is rare in Caucasians (2-3%)[17,29] and African Americans (0.3-7%),[17,23,48] but much more common for Asian (24-30%)[18,38,45] and Mexican American populations (15%).[48]

In aggregate it appears that alleles that increase CYP2E1 expression increase level of response to alcohol and risk for cancer, presumably by allowing the activation of procarcinogens or the production of ROS. Previous evidence for the involvement of *CYP2E1* with alcohol metabolism and the incidence of several alcohol related cancers, strongly supports the conclusion that *CYP2E1* alleles are associated with the level of response to alcohol and ultimately the development of alcohol use disorders. With multiple lines of evidence linking *CYP2E1* to alcohol intake and subsequent outcomes, this gene can be an important predictor of risk for alcoholism and provide us with a better understanding of how the brain perceives alcohol. Drugs that affect the expression of this gene and, subsequently, the perception of alcohol, could reduce intoxication or limit consumption and thus moderate the development of alcoholism.

2.7 References

1: Almasy L, Williams JT, Dyer TD, Blangero J. Quantitative trait locus detection using combined linkage/disequilibrium analysis. Genet Epidemiol. 1999;17 Suppl 1:S31-6.

2: Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 1998 May;62(5):1198-211.

3: Bailey SM, Cunningham CC. Contribution of mitochondria to oxidative stress associated with alcoholic liver disease. Free Radic Biol Med. 2002 Jan 1;32(1):11-6. Review.

4: Barr CS, Newman TK, Becker ML, Champoux M, Lesch KP, Suomi SJ, Goldman D, Higley JD. Serotonin transporter gene variation is associated with alcohol sensitivity in rhesus macaques exposed to early-life stress. Alcohol Clin Exp Res. 2003 May;27(5):812-7.

5: Bouchardy C, Hirvonen A, Coutelle C, Ward PJ, Dayer P, Benhamou S. Role of alcohol dehydrogenase 3 and cytochrome P-4502E1 genotypes in susceptibility to cancers of the upper aerodigestive tract. Int J Cancer. 2000 Sep 1;87(5):734-40.

6: Carr LG, Yi IS, Li TK, Yin SJ. Cytochrome P4502E1 genotypes, alcoholism, and alcoholic cirrhosis in Han Chinese and Atayal Natives of Taiwan. Alcohol Clin Exp Res. 1996 Feb;20(1):43-6.

7: Dick DM, Plunkett J, Wetherill LF, Xuei X, Goate A, Hesselbrock V, Schuckit M, Crowe R, Edenberg HJ, Foroud T. Association between GABRA1 and drinking behaviors in the collaborative study on the genetics of alcoholism sample. Alcohol Clin Exp Res. 2006 Jul;30(7):1101-10.

8: Ehlers CL, Garcia-Andrade C, Wall TL, Cloutier D, Phillips E. Electroencephalographic responses to alcohol challenge in Native American Mission Indians. Biol Psychiatry. 1999 Mar 15;45(6):776-87.

9: Garcia-Andrade C, Wall TL, Ehlers CL. The firewater myth and response to alcohol in Mission Indians. Am J Psychiatry. 1997 Jul;154(7):983-8.

10: Grove J, Brown AS, Daly AK, Bassendine MF, James OF, Day CP. The RsaI polymorphism of CYP2E1 and susceptibility to alcoholic liver disease in Caucasians: effect on age of presentation and dependence on alcohol dehydrogenase genotype. Pharmacogenetics. 1998 Aug;8(4):335-42.

11: Heath AC, Madden PA, Bucholz KK, Dinwiddie SH, Slutske WS, Bierut LJ, Rohrbaugh JW, Statham DJ, Dunne MP, Whitfield JB, Martin NG. Genetic differences in alcohol sensitivity and the inheritance of alcoholism risk. Psychol Med. 1999 Sep;29(5):1069-81.

12: Howard LA, Micu AL, Sellers EM, Tyndale RF. Low doses of nicotine and ethanol induce CYP2E1 and chlorzoxazone metabolism in rat liver. J Pharmacol Exp Ther. 2001 Nov;299(2):542-50.

13: Itoga S, Harada S, Nomura F. Polymorphism of the 5'-flanking region of the CYP2E1 gene: an association study with alcoholism. Alcohol Clin Exp Res. 2001 Jun;25(6 Suppl):11S-5S.

14: Iwahashi K, Ameno S, Ameno K, Okada N, Kinoshita H, Sakae Y, Nakamura K, Watanabe M, Ijiri I, Harada S. Relationship between alcoholism and CYP2E1 C/D polymorphism. Neuropsychobiology. 1998 Nov;38(4):218-21.

15: Joshi M, Tyndale RF. Induction and recovery time course of rat brain CYP2E1 after nicotine treatment. Drug Metab Dispos. 2006 Apr;34(4):647-52. Epub 2006 Jan 24.

16: Kato S, Onda M, Matsukura N, Tokunaga A, Tajiri T, Kim DY, Tsuruta H, Matsuda N, Yamashita K, Shields PG. Cytochrome P4502E1 (CYP2E1) genetic polymorphism in a case-control study of gastric cancer and liver disease. Pharmacogenetics. 1995;5 Spec No:S141-4.

17: Kato S, Shields PG, Caporaso NE, Hoover RN, Trump BF, Sugimura H, Weston A, Harris CC. Cytochrome P450IIE1 genetic polymorphisms, racial variation, and lung cancer risk. Cancer Res. 1992 Dec 1;52(23):6712-5.

18: Lee HS, Yoon JH, Kamimura S, Iwata K, Watanabe H, Kim CY. Lack of association of cytochrome P450 2E1 genetic polymorphisms with the risk of human hepatocellular carcinoma. Int J Cancer. 1997 May 29;71(5):737-40.

19: Lieber CS. Microsomal ethanol-oxidizing system (MEOS): the first 30 years (1968-1998)--a review. Alcohol Clin Exp Res. 1999 Jun;23(6):991-1007. Review.

20: Lieber CS. Cytochrome P-4502E1: its physiological and pathological role. Physiol Rev. 1997 Apr;77(2):517-44. Review.

21: Liu S, Park JY, Schantz SP, Stern JC, Lazarus P. Elucidation of CYP2E1 5' regulatory RsaI/Pstl allelic variants and their role in risk for oral cancer. Oral Oncol. 2001 Jul;37(5):437-45.

22: Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001 Dec;25(4):402-8.

23: London SJ, Daly AK, Cooper J, Carpenter CL, Navidi WC, Ding L, Idle JR. Lung cancer risk in relation to the CYP2E1 Rsa I genetic polymorphism among African-Americans and Caucasians in Los Angeles County. Pharmacogenetics. 1996 Apr;6(2):151-8.

24: Luczak SE, Elvine-Kreis B, Shea SH, Carr LG, Wall TL. Genetic risk for alcoholism relates to level of response to alcohol in Asian-American men and women. J Stud Alcohol. 2002 Jan;63(1):74-82.

25: McCarver DG, Byun R, Hines RN, Hichme M, Wegenek W. A genetic polymorphism in the regulatory sequences of human CYP2E1: association with increased chlorzoxazone hydroxylation in the presence of obesity and ethanol intake. Toxicol Appl Pharmacol. 1998 Sep;152(1):276-81.

26: Micu AL, Miksys S, Sellers EM, Koop DR, Tyndale RF. Rat hepatic CYP2E1 is induced by very low nicotine doses: an investigation of induction, time course, dose response, and mechanism. J Pharmacol Exp Ther. 2003 Sep;306(3):941-7. Epub 2003 May 15.

27: Montano Loza AJ, Ramirez Iglesias MT, Perez Diaz I, Cruz Castellanos S, Garcia Andrade C, Medina Mora ME, Robles Díaz G, Kershenobich D, Gutierrez Reyes G. Association of alcohol-metabolizing genes with alcoholism in a Mexican Indian (Otomi) population. Alcohol. 2006 Jun;39(2):73-9.

28: Monteiro MG, Klein JL, Schuckit MA. High levels of sensitivity to alcohol in young adult Jewish men: a pilot study. J Stud Alcohol. 1991 Sep;52(5):464-9.

29: Pastorelli R, Bardazzi G, Saieva C, Cerri A, Gestri D, Allamani A, Airoldi L, Palli D. Genetic determinants of alcohol addiction and metabolism: a survey in Italy. Alcohol Clin Exp Res. 2001 Feb;25(2):221-7.

30: Pirmohamed M, Kitteringham NR, Quest LJ, Allott RL, Green VJ, Gilmore IT, Park BK. Genetic polymorphism of cytochrome P4502E1 and risk of alcoholic liver disease in Caucasians. Pharmacogenetics. 1995 Dec;5(6):351-7.

31: Plee-Gautier E, Foresto F, Ferrara R, Bodénez P, Simon B, Manno M, Berthou F, Lucas D. Genetic repeat polymorphism in the regulating region of CYP2E1: frequency and relationship with enzymatic activity in alcoholics. Alcohol Clin Exp Res. 2001 Jun;25(6):800-4.

32: Pollock VE. Meta-analysis of subjective sensitivity to alcohol in sons of alcoholics. Am J Psychiatry. 1992 Nov;149(11):1534-8.

33: Schoedel KA, Tyndale RF. Induction of nicotine-metabolizing CYP2B1 by ethanol and ethanol-metabolizing CYP2E1 by nicotine: summary and implications. Biochim Biophys Acta. 2003 Feb 17;1619(3):283-90. Review.

34: Schuckit MA, Wilhelmsen K, Smith TL, Feiler HS, Lind P, Lange LA, Kalmijn J. Autosomal linkage analysis for the level of response to alcohol. Alcohol Clin Exp Res. 2005 Nov;29(11):1976-82.

35: Schuckit MA, Smith TL, Kalmijn J, Tsuang J, Hesselbrock V, Bucholz K. Response to alcohol in daughters of alcoholics: a pilot study and a comparison with sons of alcoholics. Alcohol Alcohol. 2000 May-Jun;35(3):242-8.

36: Schuckit MA, Smith TL. An 8-year follow-up of 450 sons of alcoholic and control subjects. Arch Gen Psychiatry. 1996 Mar;53(3):202-10.

37: Seitz HK, Stickel F. Molecular mechanisms of alcohol-mediated carcinogenesis. Nat Rev Cancer. 2007 Aug;7(8):599-612. Review.

38: Tan W, Song N, Wang GQ, Liu Q, Tang HJ, Kadlubar FF, Lin DX. Impact of genetic polymorphisms in cytochrome P450 2E1 and glutathione S-transferases M1, T1, and P1 on susceptibility to esophageal cancer among high-risk individuals in China. Cancer Epidemiol Biomarkers Prev. 2000 Jun;9(6):551-6.

39: Tanaka E, Terada M, Misawa S. Cytochrome P450 2E1: its clinical and toxicological role. J Clin Pharm Ther. 2000 Jun;25(3):165-75. Review.

40: Vasiliou V, Ziegler TL, Bludeau P, Petersen DR, Gonzalez FJ, Deitrich RA. CYP2E1 and catalase influence ethanol sensitivity in the central nervous system. Pharmacogenet Genomics. 2006 Jan;16(1):51-8.

41: Vidal F, Lorenzo A, Auguet T, Olona M, Broch M, Gutiérrez C, Aguilar C, Estupiñà P, Santos M, Richart C. Genetic polymorphisms of ADH2, ADH3, CYP4502E1 Dra-I and Pst-I, and ALDH2 in Spanish men: lack of association with alcoholism and alcoholic liver disease. J Hepatol. 2004 Nov;41(5):744-50.

42: Volavka J, Czobor P, Goodwin DW, Gabrielli WF Jr, Penick EC, Mednick SA, Jensen P, Knop J. The electroencephalogram after alcohol administration in high-risk men and the development of alcohol use disorders 10 years later. Arch Gen Psychiatry. 1996 Mar;53(3):258-63.

43: Wall TL, Johnson ML, Horn SM, Carr LG, Smith TL, Schuckit MA. Evaluation of the self-rating of the effects of alcohol form in Asian Americans with aldehyde dehydrogenase polymorphisms. J Stud Alcohol. 1999 Nov;60(6):784-9.

44: Watanabe J, Hayashi S, Kawajiri K. Different regulation and expression of the human CYP2E1 gene due to the RsaI polymorphism in the 5'-flanking region. J Biochem. 1994 Aug;116(2):321-6.

45: Watanabe J, Hayashi S, Nakachi K, Imai K, Suda Y, Sekine T, Kawajiri K. PstI and RsaI RFLPs in complete linkage disequilibrium at the CYP2E gene. Nucleic Acids Res. 1990 Dec 11;18(23):7194.

46: Wilhelmsen KC, Schuckit M, Smith TL, Lee JV, Segall SK, Feiler HS, Kalmijn J. The search for genes related to a low-level response to alcohol determined by alcohol challenges. Alcohol Clin Exp Res. 2003 Jul;27(7):1041-7.

47: Wong NA, Rae F, Simpson KJ, Murray GD, Harrison DJ. Genetic polymorphisms of cytochrome p4502E1 and susceptibility to alcoholic liver disease and hepatocellular carcinoma in a white population: a study and literature review, including meta-analysis. Mol Pathol. 2000 Apr;53(2):88-93.

48: Wu X, Shi H, Jiang H, Kemp B, Hong WK, Delclos GL, Spitz MR. Associations between cytochrome P4502E1 genotype, mutagen sensitivity, cigarette smoking and susceptibility to lung cancer. Carcinogenesis. 1997 May;18(5):967-73.

49: Yu MW, Gladek-Yarborough A, Chiamprasert S, Santella RM, Liaw YF, Chen CJ. Cytochrome P450 2E1 and glutathione S-transferase M1 polymorphisms and susceptibility to hepatocellular carcinoma. Gastroenterology. 1995 Oct;109(4):1266-73.

50: Zakhari S. Overview: how is alcohol metabolized by the body? Alcohol Res Health. 2006;29(4):245-54. Review.

Figure 2.1 Location of genotyped SNPs in relation to *CYP2E1* on chromosome 10.

Figure 2.1 The top of the figure shows the position on chromosome 10 with each SNP location indicated by triangles. The middle part of the figure shows the position of *CYP2E1* with exons represented as yellow rectangles and introns as the lines between. At the bottom, phased haplotypes derived from the HapMap Caucasian (CEU) population are shown. Each vertical block represents a SNP genotyped in HapMap. Not all of these markers were genotyped in the study, so vertical black lines through the haplotype figure indicate actual genotyped SNPs.

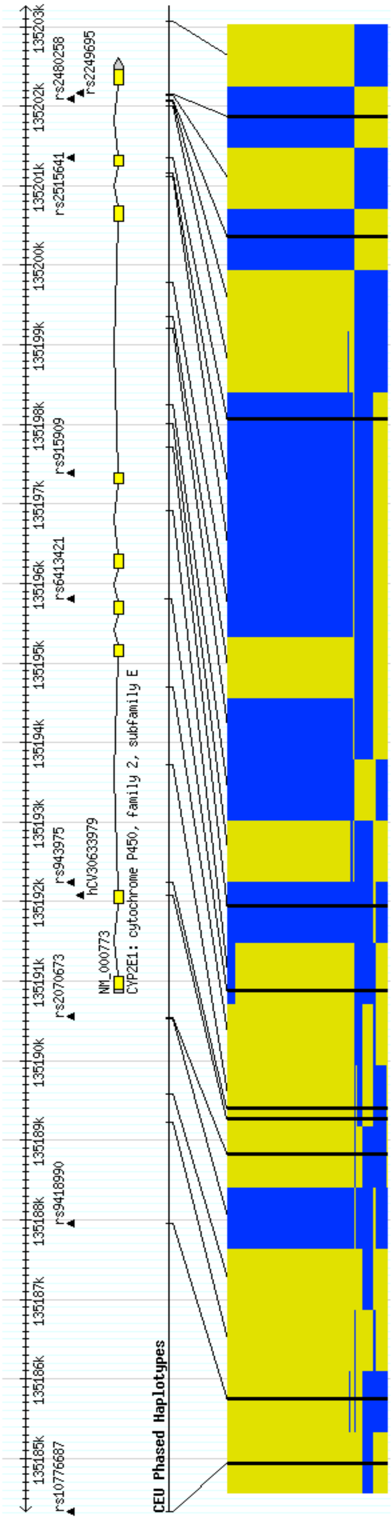

135185k
135186k
135187k
135188k
135189k
135190k
135191k
135192k
135193k
135194k
135195k
135196k
135197k
135198k
135199k
135200k
135201k
135202k
135203k
rs10776687
rs9418990
rs2070673
rs943975
hCV30633979
rs6413421
rs915909
rs2515641
rs2480258
rs2249695
NM_000773
CYP2E1: cytochrome P450, family 2, subfamily E
CEU Phased Haplotypes

Figure 2.1

Figure 2.2 LOD score plot showing linkage in the region surrounding *CYP2E1*.

Figure 2.2 Shows the LOD score plot highlighting the linkage near *CYP2E1*. The line labeled as First Set represents the initial 139 sibling pairs. The line labeled as Second Set represents the complete set of 238 sibling pairs. Once the family with questionable phenotypes was removed, the strength of the linkage signal was restored to the level provided by the First Set samples. Locations of microsatellite markers and SNPs are shown on the X-axis.

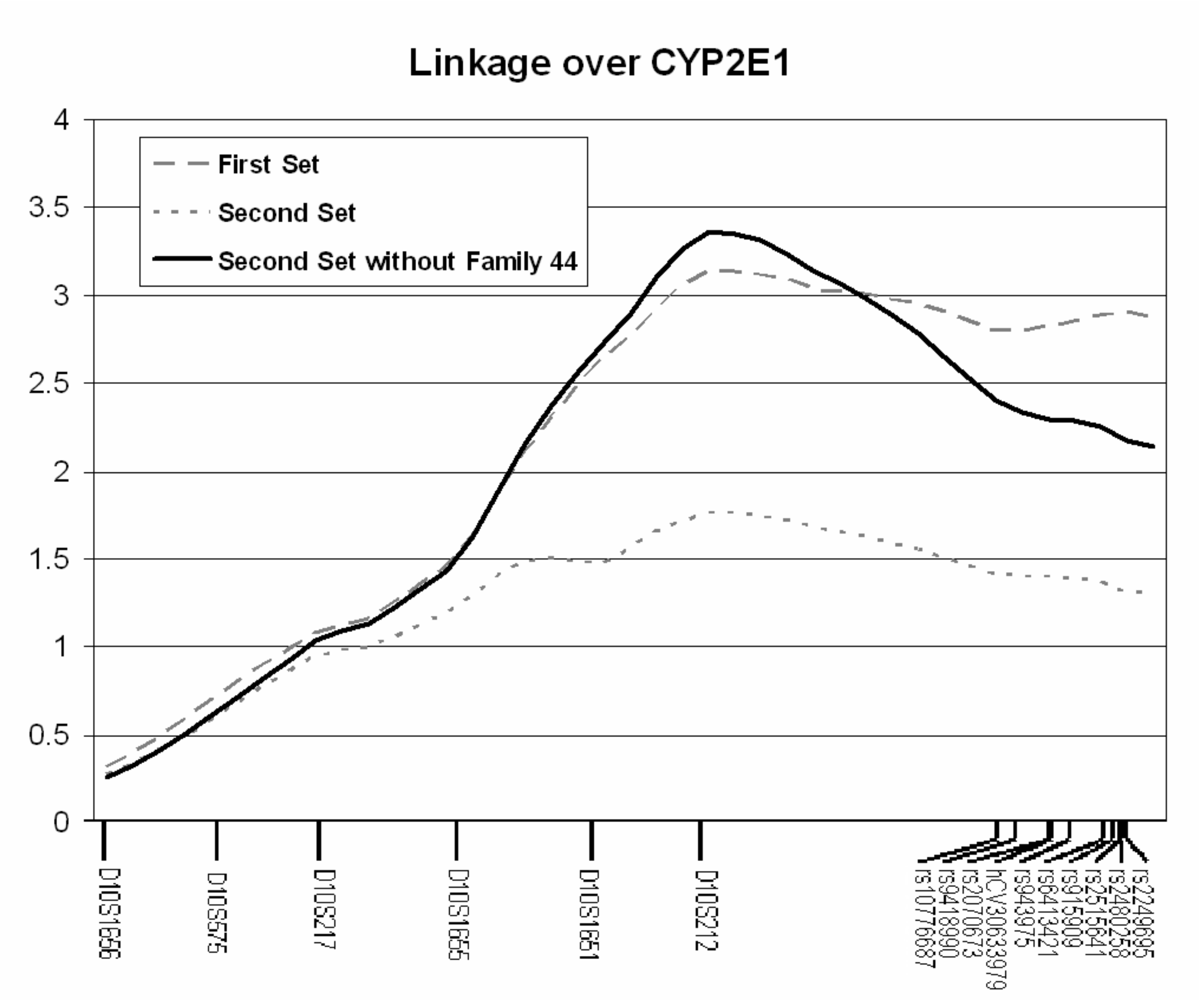


Table 2.1 Translation of identification values for genotyped SNPs and position

rs ID	ABI Assay ID	Build 129 position (bp)	Other names
rs10776687	hCV2431881	135184332	
rs9418990	hCV2431878	135187956	
rs2070673	hCV2431871	135190557	CYP2E1*7_-333T>A
--	hCV30633979	135192024	CYP2E1*2,g.1132G>A
rs943975	hCV7468406	135192250	
rs6413421	hCV25594214	135195801	
rs915909	hCV7468401	135197387	CYP2E1_6498C>T(I321I)
rs2515641	hCV16026002	135201352	CYP2E1_10463T>C(F421F)
rs2480258	hCV2431850	135202090	
rs2249695	hCV2431848	135202158	

Table 2.1 A listing of the various identification names for the SNPs genotyped in the study based on the Applied Biosystems ID. Included under "other names" are names commonly used for specific markers.

Table 2.2 Association p values for logistic regression analysis between the SHAS score and *CYP2E1* genotype alone or genotype considering copy number.

	Genotype	Genotype considering copy number	Minor allele frequency
rs10776687	**0.007**	0.103	0.056
rs9418990	**0.024**	0.125	0.244
rs2070673	**0.015**	0.077	0.238
hCV30633979	0.215	0.253	0.004
rs943975	0.182	0.274	0.131
rs6413421	0.133	0.123	0.057
rs915909	0.058	0.081	0.007
rs2515641	0.45	0.261	0.176
rs2480258	**0.04**	0.187	0.252
rs2249695	**0.024**	0.139	0.268

Table 2.2 Logistic regression was used to test for association between the genotyped SNPs and the SHAS quantity representing the level of response to alcohol. P-values < 0.05 are in bold. The three markers near the 3' end and two from the 5' end were most associated with the level of response to alcohol. The best evidence for association came from the initial SNP,

rs10776687. None of the SNPs were associated when the genotype call was made considering copy number.

Table 2.3 Results of variance component linkage analysis for combined linkage and association.

Covariate	LOD score	covar p	Variance
None	3.36		
rs10776687	2.15	4.25E-04	0.0531
rs9418990	2.46	5.65E-03	0.021
rs2070673	2.54	5.04E-03	0.0311
rs943975	2.23	8.74E-02	0.0105
rs2515641	2.88	5.30E-02	0.0076
rs2480258	2.42	1.26E-02	0.0268
rs2249695	2.54	9.58E-03	0.02

Table 2.3 Combined linkage and association analysis showed that a single marker was unable to account for all of the variation in the signal. This was accomplished by adding each SNP individually into the model as a covariate. With no covariates, the LOD score was 3.36. The SNP that lowered the score the most when added as a covariate was rs10776687and was able to explain 5.3% of the variance in the SHAS score.

Table 2.4 Results of variance component linkage analysis with the inclusion of several drinking and smoking covariates

	covariate LOD	variance
number of days in the last week where subject drank	2.19	0.078
amount consumed in the last 24 hours	2.82	0.028
days smoking per month in previous 6 months	3.03	0.046
cigarettes per day, on smoking days in previous 6 months	3.27	0.036
Average number of cigarettes per day	3.37	0.026
maximum amount drank in one day	3.47	0.084
Average number of drink on days they drank	3.77	0.066

Table 2.4 A number of smoking and drinking phenotypes were analyzed by including them as covariates when modeling their relationship with SHAS. Three of these phenotypes (average number of cigarettes per day, the maximum amount drank in one day, and the

average number of drinks consumed on days the subject drank) improved the overall peak LOD score compared to the model without covariates.

Chapter 3 - The role of the *Tau* gene region chromosome inversion in PSP, CBD and related disorders

3.1 Abstract

A genome wide association scan was performed to search for variants that confer susceptibility to 4 tauopathies and clinically related disorders. This paper focuses on the results from an inverted region of chromosome 17 that contains the *MAPT* gene. A total of 231 samples were genotyped on the GeneChip 500K Affymetrix SNP arrays. Missing or untyped SNPs were imputed with IMPUTE from the Chiamo suite. Genotypes of cases and controls were compared with a Fisher exact test on a marker by marker basis. Haplotypes were determined by the visual inspection of genotypes. Cases of PSP, CBD, FTD, and FTD with amyotrophy were collected from an unrelated Caucasian population. Unaffected individuals from the same population were used as controls. The samples included in the study were collected by the Memory and Aging Center at UCSF or by KCW. For the comparison between any particular disease and controls, the association was constant across the interval. Significant associations were seen for both PSP and PSP combined with CBD. Of the two haplotypes seen in the region, the H1 haplotype was overrepresented in PSP and

CBD cases when compared to controls. The association found in these tauopathies across this interval on chromosome 17 further supports the theory that one or more susceptibility loci in this region is affecting susceptibility specifically to PSP and CBD. Since the markers are highly correlated and the association is seen across the whole region, it is difficult to narrow down a disease causing variant or even a possible candidate gene. However considering the pathology of these diseases and the involvement of tau mutations seen in familial forms, the MAPT gene represents the most likely cause driving the association.

3.2 Introduction

The Pick Complex refers to a spectrum of diseases with a variety of overlapping clinical and pathological features, due to a related genetic etiology. A common, though not ubiquitous, overlapping feature of these diseases is the presence of tau protein inclusions, or aggregates. Thus, the genetics and brain histochemistry of the gene that encodes for tau, microtubule associated protein tau (*MAPT*), provides a compelling reason for thinking that patients with these clinically and pathologically diverse findings should be thought of as a contiguous group. These diseases are characterized clinically by cognitive, behavioral, and movement defects. This study focused on four diseases in the spectrum where tau histochemistry and genetics are believed to be critical—progressive supranuclear palsy (PSP), corticobasal degeneration (CBD), and frontotemporal dementia (FTD) with or without amyotrophy.

The clinical signs and symptoms observed in patients with these diseases are correlated with the anatomic distribution of neuronal loss, which can be quite variable. There are several patterns of inclusions of insoluble proteins in affected individuals, but there is only limited correlation between inclusion type and clinical symptomatology. Pick Complex diseases can be accompanied by tau inclusions, ubiquitin inclusions, or no inclusions at all.[15] Pick Complex diseases that contain tau inclusions are collectively referred to as tauopathies.

Many families with inherited tauopathies have been linked to the same genomic region, and collectively, these families are said to be affected with frontotemporal dementia and parkinsonism linked to chromosome 17 (FTDP-17).[8] *MAPT* was considered to be a likely candidate gene in this region for its involvement in FTD with tau inclusions, and subsequently, many *MAPT* mutations have been identified in affected individuals. A variety of Tau mutations have been identified that affect protein function by either creating changes in level of translated protein or by alternative RNA splicing, which may upset the interaction between tau and microtubules, allowing unbound and abnormally phosphorylated tau to polymerize into inclusions.[5] Different tau mutations alter biochemical properties of the gene product, but these mutations do not necessarily predict the exact clinical nature of the disease. The same mutation in affected individuals, even in the same family, may result in a different age of onset, combination of symptoms, and clinical diagnosis.[13] The variable morphology of accumulated tau proteins could be explained by the wide range of mutations that have been found in these diseases. In various tauopathies, the inclusions may differ based on the ratios of particular isoforms and the physical location of accumulation. At least 40 MAPT mutations have been identified in patients with FTD and related diseases.[12] Tau inclusions, usually without *MAPT* mutations, are part of the pathologic definition of CBD and PSP while cases of FTD are often seen without tau mutations or tau inclusions. Another set of cases with FTDP-17 that contain ubiquitin inclusions, but no tau inclusions, was linked to the same region on chromosome 17.[9] Further gene resequencing of this set of cases led to the discovery of mutations in the progranulin gene from this region, which are responsible for many cases of FTDP-17.[9]

In its natural state, *MAPT* works to stabilize microtubule formation and regulate transport along microtubules.[13] Dysfunctional tau proteins can interrupt axonal transport by reducing the cell's ability to control microtubule formation, ultimately leading to neuron dysfunction and death.[8] Normally, the tau protein is located in axons, but in diseased cells it will relocate to the cell body and form insoluble hyperphosphorylized fibrillary inclusions.[15] This hyperphosphorylation of tau may lead to a loss of microtubule affinity and a resistance to proteases, leading to aggregation.[2] Six major isoforms are produced in the adult human brain through the alternative splicing of exons 2, 3, and 10.[13] The 6 isoforms can be divided into 2 groups, depending on the number of microtubule binding domains. Alternative splicing of exon 10 will lead to four repeat (4R) binding domains or three repeat (3R) binding domains.[13] The number of binding domains affects the binding of tau to tubulin; 4R tau will bind stronger and assemble more efficiently than 3R tau.[13] A reduction in binding efficiency may increase the amount of unbound tau in the neuron leading to aggregation, although increased binding may have an equally damaging effect.[13] An accumulation of unbound tau may result if any isoform fails to function, creating insoluble inclusions.[8] Inclusions found in affected individuals may contain all 6 isoforms in equal amounts or different ratios of selected isoforms. Many mutations disrupt the splicing of exon 10, leading to unequal ratios of 3R and 4R ratios. Tau deposits in PSP and CBD are predominantly 4R, where deposits in FTD contain equal levels of 4R and 3R.[10,15]

The region containing the *MAPT* gene has been shown to be genetically complex due to an inversion commonly found in Caucasian populations. There are three highly homologous low copy repeats (LCRs) that flank the region.[3] The two LCRs telomeric of *MAPT*, LCRs B, and C are inverted relative to the centromeric LCR A.[3] LCR A and LCR B flank the *MAPT* haplotype, suggesting that the inversion was caused by non-allelic homologous recombination. [3] Figure 3.1 shows the structure of MAPT in relation to these LCRs. Extensive genotyping across the interval identified two haplotypes in almost complete disequilibrium.[14] These haplotypes are commonly referred to as the H1 and H2 haplotypes. Recombination within the inverted segment between carriers of the H1 and H2 haplotype would result in a Robertsonian translocation. The high degree of disequilibrium in this region suggests that recombination has been suppressed or that there was a selection against recombinant chromosomes prior to the inversion becoming established in the Caucasian population.[1] A study on the expression of tau in Alzheimer patients found that one variant of the H1 haplotype led to an increase in overall tau levels and specifically an increase in 4R tau creating an imbalance of isoforms.[7] Similar changes in expression could be found in these diseases.

There is a locus in or near the *MAPT* gene that clearly affects susceptibility to PSP and CBD. Conrad et al established that common variations in the *MAPT* gene affect susceptibility to PSP.[2] They reported that the a0 allele of a dinucleotide repeat marker located in intron 9 of *MAPT* is observed in 57% of control chromosomes compared to 95.5% of PSP cases.[2] The a0 allele was also shown to be overrepresented in CBD chromosomes.[4] Other tauopathies have a less certain association with *MAPT* region polymorphisms.[2] The a0 allele is not believed to be biologically relevant to the disease process, but is instead in linkage disequilibrium with some other polymorphism.[2] The a0 allele is inherited with the H1 haplotype, so it is not surprising that the H1 haplotype is also overrepresented in PSP cases.[1] It is uncertain whether the increased risk for PSP and CBD is associated with a specific common variation of the haplotype or a rare mutation that is found on same chromosomes with the H1 haplotype.[2]

Despite the varied clinical features that are used to categorize the different diseases, there is quite a bit of overlap, suggesting that there could be a shared underlying biochemical abnormality resulting from the altered expression of tau.[4] In order to explore this possibility, we performed a high density association scan looking for markers that may confer susceptibility to several different tauopathies. In this report, we focus on the markers contained in the region including and surrounding tau. Using our data, and genotypes imputed using Hapmap, it was shown that a significant association exists across the entire inverted interval on chromosome 17 for PSP and CBD cases.

3.3 Methods

3.3.1 Sample Collection and preparation

The samples included in the study were collected by the Memory and Aging Center at UCSF or by KCW. Cases of PSP, CBD, FTD, and FTD with amyotrophy were collected from an unrelated Caucasian population. Cases of FTD met Neary criteria and PSP met Litvan criteria. While all cases were clinically confirmed, only 46 had pathological confirmation of disease. None of the cases have known tau or progranulin mutations. Unaffected individuals from the same population were used as controls. DNA was isolated from whole blood using the Puregene kit (Gentra Systems). The number of patients used for each diagnosis in this study can be found in Table 3.1. The number of cases per gender can be found in Table 3.2. The average age was 73 for controls and 67 for cases with an average

age of onset of 60 years. All subjects participated with informed consent procedures approved by the UCSF and UNC Human Subjects Institutional Review Boards.

3.3.2 Genotyping

Genotyping was performed using the GeneChip 500K Affymetrix SNP arrays using the protocol provided by Affymetrix. The BRLMM algorithm was used to make genotyping calls. Acceptable genotypes had confidence scores less than 0.5. Any call that did not meet this threshold was removed from further analysis.

3.3.3 Analysis

The genotypes of cases versus controls were compared using a Fisher exact test to determine whether the allele frequency in the cases was significantly different from the controls. Markers that were considered to be out of Hardy Weinberg Equilibrium were excluded from analysis. There was no population stratification detected when tested with Eigenstrat.[11] Genotype calls made by the BRLMM algorithm were used to infer the rest of the known Hapmap markers in the area based on correlation using the program Impute from the Chiamo suite.[6] Imputed genotypes were considered acceptable with a posterior probability greater than 0.8, and markers were included in association tests if the call rate was greater than 80%.

3.4 Results

Genome wide, the average sample call rate was 95% and the average single nucleotide polymorphism (SNP) call rate was 92% on the Affymetrix 500K platform. Less than 1% of SNPs were out of Hardy Weinberg equilibrium and 1.5% of SNPs were monomorphic.

From the 326 SNPs typed in the MAPT region ranging from approximately 40.4 Mb to 42.5 Mb on chromosome 17, we attempted to impute an additional 4,845 HapMap SNPs. After eliminating imputed polymorphisms with a posterior probability lower than 0.8, 1,477 SNPs remained. Of these, 60 SNPs were monomorphic and 68 were not in Hardy Weinberg equilibrium. Any marker with a sample call rate less than 80% was removed. Genotypes for 1,169 genotyped and imputed SNPs were used to explore the region near the *MAPT* gene for allelic association with PSP, CBD, FTD, and FTD with amyotrophy.

Figure 3.2a shows a plot of the probability that the cases and controls have equivalent genotype frequencies for each of the typed or imputed SNPs that met inclusion criteria for each of the disease classification models tested. All of the significant associations observed are within the boundaries of the chromosomal inversion that distinguish the H1 and H2 haplotype. While there are some clear exceptions, the majority of markers across the inversion for any given comparison fall within a constant range of probabilities across the interval. The most striking associations observed are for PSP alone or combined with CBD versus controls across the entire region of the chromosomal inversion. Rarely a marker from other comparisons will reach a nominally significant association, but these events are rare and not constant across the inverted interval. Inspection of the raw allele specific hybridization intensity for these markers does not robustly distinguish between genotype clusters and are not considered to be significant associations. The region where allelic association is detected clearly defines the inversion interval boundaries. Figure 3.2b shows the genotypes for all samples across the region of interest in the following order: control, FTD, FTD with amyotrophy, CBD, and PSP. Each row represents an individual and each column represents a marker. Known genes are indicated as lines above the genotypes. The samples were sorted based on diagnosis and haplotype similarity. Two distinct haplotypes can be identified in this figure consistent with previous designation of the H1 and H2 haplotype.

Table 3.3 shows the counts for the three haplotype combinations (H1/H1, H1/H2, H2/H2) for each category of diagnosis. Very few heterozygous haplotypes, and no homozygous H2 haplotypes, were seen in either PSP or CBD. When compared to controls using a fisher exact test, only PSP and PSP/CBD were significantly different. This confirms that the H1 haplotype is overrepresented in PSP and CBD cases when compared to controls, while both FTD and FTD with amyotrophy had H1 levels in the same proportion as controls as seen in Table 3.4 which gives the percentage of H1 and H2 haplotypes in each group.

3.5 Discussion

The association found in these tauopathies across this interval on chromosome 17 further supports the theory that one or more susceptibility loci in this region is affecting susceptibility specifically to PSP and CBD. Since the markers are highly correlated and the association is seen across the whole region, it is difficult to narrow down a disease causing variant or even a possible candidate gene. However considering the pathology of these diseases and the involvement of tau mutations seen in familial forms, the MAPT gene represents the most likely cause driving the association.

While all of the diseases in the current study are part of what is referred to as tauopathies, not all of the diseases were highly associated with this region. The hypothesis leading to this study was that similar diseases were caused by mutations in genes controlled by similar biochemical pathways. So while the mutation causing any particular presentation of one of these diseases may be located in different genes, the pathological outcome is the same. There were other significantly associated regions in our genome wide scan which may be affecting susceptibility to the diseases which were not as highly associated with the tau region. There was no association seen between FTD samples and this region even though cases with FTD can have tau inclusions and mutations of the tau gene have been found in affected families. The lack of association may be influenced by our small sample size and the heterogeneity of sporadic FTD. Most cases of FTD do not have tau mutations or inclusion, but cases of PSP and CBD are almost always accompanied by tau inclusions.

The odds ratio was calculated to determine the risk associated with a particular haplotype. Controls have a 7 fold greater odds (95% confidence interval 2.08-25.36) of having an H2 haplotype, on either one or both chromosomes, compared to PSP and a 4 fold greater odds (95% confidence interval 1.16-15.10) when compared to CBD. When CBD and PSP are considered together, the odds ratio is in the middle with a value of 5.7 (95% confidence interval 2.24-14.62). This suggests that the H2 haplotype provides some protection from the PSP and CBD diseases. This proposed protective allele is considered to be significant for PSP, CBD, and PSP+CBD since none of the confidence intervals go below 1.

Imputation filled in missing genotypes and genotypes for markers not included in the Affymetrix chip 500K sets. This gave a fuller picture of the association in the region. Imputation methods are useful for association studies since they combine information from genotyped markers with existing datasets such as Hapmap. Testing a larger number of markers across the genome provides a finer grid for association. However in an area of high linkage disequilibrium, as in the chromosome 17 inversion, the true disease causing variant cannot be distinguished from the surrounding markers even with the extra imputed genotypes. Imputation added little to our association due to the strong LD in the region that was implicated. The regions flanking this inversion are much more thoroughly evaluated and there is little reason to investigate these flanking regions further. Filtering the data for call rate and posterior probability removed noise and most false positives that were detected using unfiltered data. The genotype calling software which used to impute genotypes resulted in more stringent and reliable genotyping calls.

The results from this association study provides strong evidence that a susceptibility locus in the MAPT gene region is related to certain Pick Complex diseases, but the high degree of linkage disequilibrium in the region makes it difficult to draw conclusions about the exact location of the locus. To our knowledge, previous studies have reported results from candidate gene studies focusing on the tau gene. We instead looked at the entire region and found an association with the entire inversion region with no evidence that any part of the region is more important than any other. The genotypes are constant across the inversion due to the high level of linkage disequilibrium, but outside of the inversion they become highly variable with no identifiable pattern. This is also supported by the constant level of association that drops off at the boundaries of the inversion. The inversion is likely a recent event since it is only found in Caucasian populations. While a specific cause cannot be determined, something in the inversion is likely affecting expression of the tau gene and ultimately disease status. The inversion, or more specifically the H2 haplotype, appears to offer some protection against PSP and CBD.

3.7 References

1: Baker M, Litvan I, Houlden H, Adamson J, Dickson D, Perez-Tur J, Hardy J, Lynch T, Bigio E, Hutton M. Association of an extended haplotype in the tau gene with progressive supranuclear palsy. Hum Mol Genet. 1999 Apr;8(4):711-5.

2: Conrad C, Andreadis A, Trojanowski JQ, Dickson DW, Kang D, Chen X, Wiederholt W, Hansen L, Masliah E, Thal LJ, Katzman R, Xia Y, Saitoh T. Genetic evidence for the involvement of tau in progressive supranuclear palsy. Ann Neurol. 1997 Feb;41(2):277-81.

3: Cruts M, Rademakers R, Gijselinck I, van der Zee J, Dermaut B, de Pooter T, de Rijk P, Del-Favero J, van Broeckhoven C. Genomic architecture of human 17q21 linked to frontotemporal dementia uncovers a highly homologous family of low-copy repeats in the tau region. Hum Mol Genet. 2005 Jul 1;14(13):1753-62. Epub 2005 May 11.

4: Josephs KA, Petersen RC, Knopman DS, Boeve BF, Whitwell JL, Duffy JR, Parisi JE, Dickson DW. Clinicopathologic analysis of frontotemporal and corticobasal degenerations and PSP. Neurology. 2006 Jan 10;66(1):41-8.

5: Kertesz A. Pick Complex: an integrative approach to frontotemporal dementia: primary progressive aphasia, corticobasal degeneration, and progressive supranuclear palsy. Neurologist. 2003 Nov;9(6):311-7. Review.

6: Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007 Jul;39(7):906-13. Epub 2007 Jun 17.

7: Myers AJ, Pittman AM, Zhao AS, Rohrer K, Kaleem M, Marlowe L, Lees A, Leung D, McKeith IG, Perry RH, Morris CM, Trojanowski JQ, Clark C, Karlawish J, Arnold S, Forman MS, Van Deerlin V, de Silva R, Hardy J. The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts. Neurobiol Dis. 2007 Mar;25(3):561-70. Epub 2006 Dec 15.

8: Neary D, Snowden J, Mann D. Frontotemporal dementia. Lancet Neurol. 2005 Nov;4(11):771-80. Review.

9: Pittman AM, Fung HC, de Silva R. Untangling the tau gene association with neurodegenerative disorders. Hum Mol Genet. 2006 Oct 15;15 Spec No 2:R188-95. Review.

10: Pittman AM, Myers AJ, Duckworth J, Bryden L, Hanson M, Abou-Sleiman P, Wood NW, Hardy J, Lees A, de Silva R. The structure of the tau haplotype in controls and in progressive supranuclear palsy. Hum Mol Genet. 2004 Jun 15;13(12):1267-74. Epub 2004 Apr 28.

11: Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies.

12: Rademakers R, Melquist S, Cruts M, Theuns J, Del-Favero J, Poorkaj P, Baker M, Sleegers K, Crook R, De Pooter T, Bel Kacem S, Adamson J, Van den Bossche D, Van den Broeck M, Gass J, Corsmit E, De Rijk P, Thomas N, Engelborghs S, Heckman M, Litvan I, Crook J, De Deyn PP, Dickson D, Schellenberg GD, Van Broeckhoven C, Hutton ML. High-density SNP haplotyping suggests altered regulation of tau gene expression in progressive supranuclear palsy. Hum Mol Genet. 2005 Nov 1;14(21):3281-92. Epub 2005 Sep 29.

13: Rademakers R, Cruts M, van Broeckhoven C. The role of tau (MAPT) in frontotemporal dementia and related tauopathies. Hum Mutat. 2004 Oct;24(4):277-95. Review.

14: Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, Desnica N, Hicks A, Gylfason A, Gudbjartsson DF, Jonsdottir GM, Sainz J, Agnarsson K, Birgisdottir B, Ghosh S, Olafsdottir A, Cazier JB, Kristjansson K, Frigge ML, Thorgeirsson TE, Gulcher JR, Kong A, Stefansson K. A common inversion under selection in Europeans. Nat Genet. 2005 Feb;37(2):129-37. Epub 2005 Jan 16.

15: Williams DR. Tauopathies: classification and clinical update on neurodegenerative diseases associated with microtubule-associated protein tau. Intern Med J. 2006 Oct;36(10):652-60. Review.

16. NIEHS SNPs. NIEHS Environmental Genome Project, University of Washington, Seattle, WA (URL: http://egp.gs.washington.edu) [9,2007]

Figure 3.1 Tau Gene Region

Figure 3.1 shows the structure of the Tau gene, indicating the locations of microtubule binging domains and flanking LCRs.

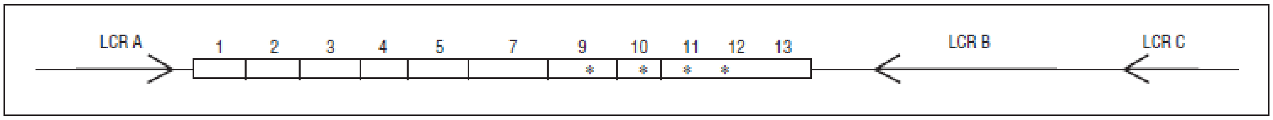


Figure 3.2 Association and haplotypes across the interval

a) Figure 3.2a shows a plot of the negative log of the p value from the comparisons between cases and controls. Known genes are represented as lines at the top of the figure.

b) Figure 3.2b shows the genotypes for all samples across the region of interest. Each row corresponds to a sample. The samples were sorted based on diagnosis and haplotype similarity. Samples with mostly blue or major alleles have the H1 haplotype while samples with mostly yellow or minor alleles have the H2 haplotype. Samples with mostly red or heterozygote alleles are H1/H2. The figure was created using the *NIEHS SNPs Visual* Genotypes program.[14]

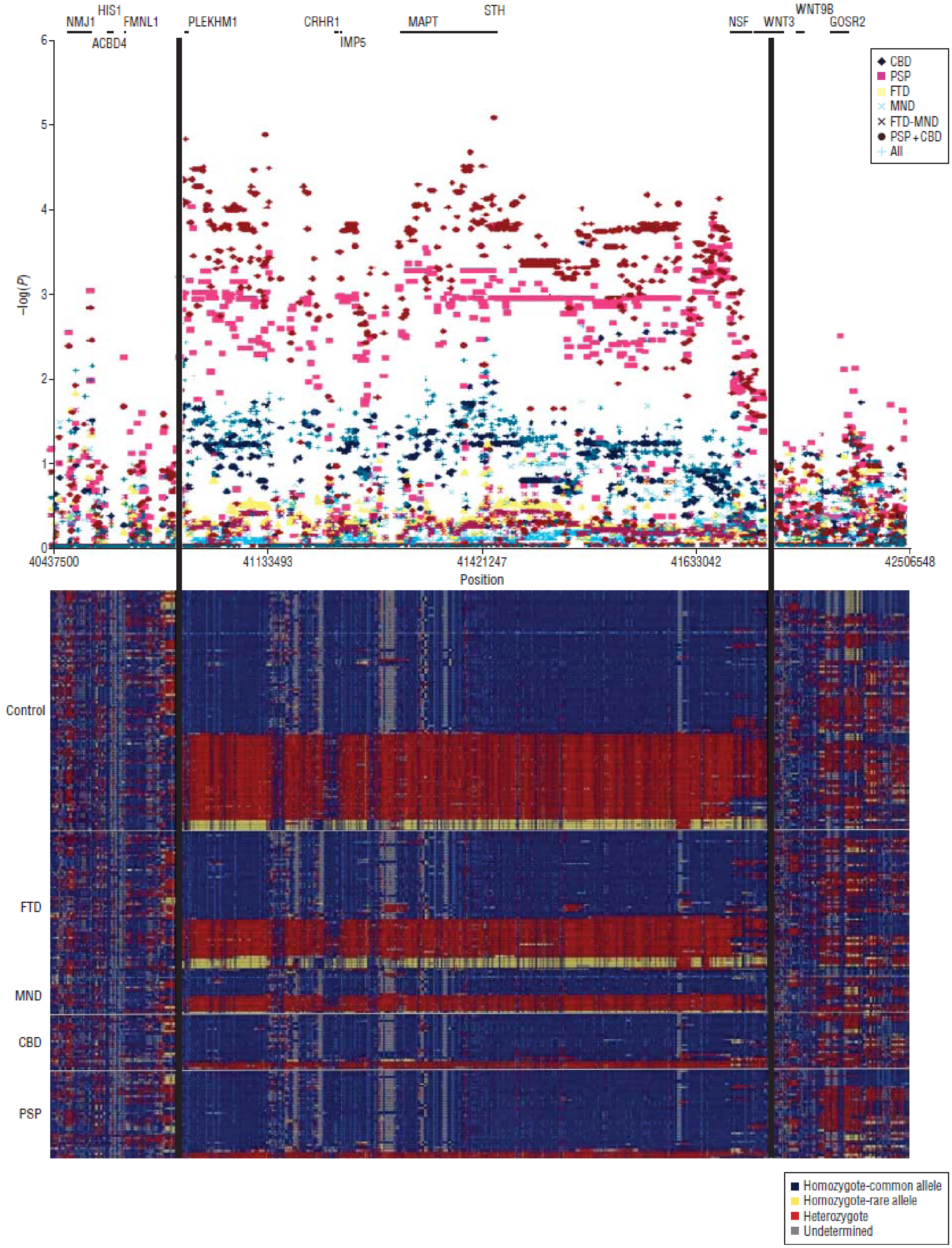


Figure 3.2

Table 3.1 Number of Samples included per Diagnosis

Table 3.1 shows the number of samples genotyped for each diagnosis type.

Diagnosis	**Count**
Control	98
FTD	56
PSP	36
CBD	23
ALS	18

Table 3.2 Gender of Patients in Study

Table 3.2 shows the breakdown of samples by gender

	Male	**Female**
FTD	32	24
PSP	20	16
ALS	11	7
CBD	9	14
Control	42	56

Table 3.3 Comparison of Haplotypes

Table 3.3 shows the p value and odds ratio based on haplotype counts for each diagnosis.

	H1/H1	H1/H2	H2/H2	p value	Odds Ratio	Confidence Interval
Control	59	35	4			
PSP	33	3	0	0.0011	7.27	2.08-25.36
CBD	19	3	0	0.061	4.19	1.16-15.10
FTD	37	15	4	0.4617	1.29	0.65-2.55
MND	11	6	1	1	1.04	0.37-2.91
PSP+CBD	52	6	0	0.0002	5.73	2.24-14.62
ALL	100	27	5	0.0311	2.07	1.17-3.64

Table 3.4 Percentage of Haplotypes

Table 3.4 shows the percentage of samples of each diagnosis with the H1 haplotype.

	Control	PSP	CBD	FTD	MND	PSP+CBD	ALL
H1	153	69	41	89	28	110	227
H2	43	3	3	23	8	6	37
total	196	72	44	112	36	116	264
%H1	0.781	0.958	0.932	0.795	0.778	0.948	0.860

Chapter 4 - The evaluation of Convergent Haplotype Association Tagging: a novel algorithm for haplotype phase inference

4.1 Abstract

Current approaches for the identification of genetic influences of complex disease typically focus on the effect of one genetic variant at a time. Consecutive variants along a chromosome are inherited together as a haplotype and knowledge of this haplotype can be very beneficial to genetic analysis. The statistical prediction of haplotype, referred to as haplotype phase inference, has proven difficult when applied to the ambiguous genotypes created by conventional genome-wide SNP genotyping methods. A new approach, Convergent Haplotype Association Tagging or CHAT, was created to search for subsets of a population that share a long haplotype and to phase haplotypes based on the identified sharing. In order to test the performance of the haplotype phase inference capabilities of CHAT, comparisons were made with three publicly available haplotype phasing programs: ENT, HaploRec, and Beagle. Performance comparisons were based on two calculations of accuracy: the single site error rate, which measures the percentage of heterozygous loci incorrectly phased according to the true sequence, and the switch accuracy, which measures the number of recombinations needed to regain the true sequence. The programs were applied to simulated data generated to mimic world populations. CHAT made very accurate

haplotype predictions especially when applied to a sample set with a large number of individuals; however, due to the nature of the algorithm it is only able to improve the haplotype phase for regions with haplotype sharing.

4.2 Background

4.2.1 Haplotype-based association analysis

Current approaches for the identification of genetic influences of complex disease typically focus on the effect of one genetic variant or marker at a time and assume each marker is independent. Consecutive markers along a chromosome are actually inherited together and knowledge of this configuration can be very beneficial to genetic analysis. The combination of alleles for nearby markers on a single chromosome is referred to as a haplotype. In single marker association analysis, we assume that a genotyped marker will be able to represent the variation of an untyped, causal variant. Unfortunately, many association studies are unsuccessful at identifying a reproducible causal variant. Considering multiple genotyped markers improves the chance that the untyped variant can be captured, especially if the variant is rare in the sampled population due to selection or recent mutation.[1] Haplotype based association analysis seeks to find association between a trait and an ancestral haplotype harboring a variant that influences the trait. Individuals with the trait affected by the same causal variant will have varying amounts of the ancestral haplotype containing that variant. The intersection of these haplotypes can map the trait-causing variant to a smaller chromosomal region with greater certainty than by single marker analysis. In this way, haplotypes can provide more information and improve the power to detect a variant associated with the trait.

Although knowledge of haplotypes is advantageous for genetic analysis of complex human disease, haplotype based analysis has been limited. Current genotyping platforms used in genome-wide association studies provide only the genotype at each position, either heterozygous or homozygous for any particular allele, with no indication about the placement of each allele on a chromosome in relation to other genotyped markers. For this type of data, the haplotype phase, or set of markers together on a chromosome, must be predicted after genotype assignments have been determined. Haplotype phase inference refers to the identification of haplotypes in genotyped samples by determining which alleles are inherited together on a single chromosome. Two copies of each marker are typically present in every diploid organism—one inherited from the maternal chromosome and one from the paternal chromosome. Complete inherited parental chromosomes are broken up by recombination, so that the grand parental origin for a region will vary across the length of the chromosome. In a small region, markers are less likely to be separated by recombination and more likely to be inherited together. Given this correlation between nearby markers on a chromosome, referred to as linkage disequilibrium or LD, markers located closely together will more often be inherited as a single haplotype, while markers farther apart are more likely to be separated by recombination.

4.2.2 Genetic approaches to haplotype phase inference

Early genetic studies were family-based linkage analyses which considered the sharing of alleles between relatives often using sibling pairs or parent child trios. Using family genotypes simplifies phase inference, but phase is ambiguous if all family members are heterozygous. The recruitment of family members can be difficult, especially for late onset diseases, and increases the genotyping costs. With the transition to large scale case-control association studies, the phase inference capabilities provided by genotypes of relatives have been lost.

As an alternative to family based haplotype phase determination, a variety of molecular genotyping methods are available to determine phase by separating the chromosomes and genotyping each directly. These include single-molecule dilution, long-range allele specific PCR, diploid to haploid conversion, pyrosequencing, rolling circle amplification, etc.[11,12] Compared to SNP genotyping, molecular haplotyping methods are expensive, low throughput, and often unreliable.[11,12]

Both molecular genotyping and family based methods are unacceptable for haplotype phase inference of large case-control association studies. Effective haplotype analysis requires a quick, reliable, and cost effective method to phase millions of ambiguous genotypes created by conventional genome-wide SNP genotyping methods for large groups of unrelated individuals.

4.2.3 Statistical approaches to haplotype phase inference

Early statistical based haplotype phase inference algorithms were created for a small number of markers and samples. The earliest described algorithm for haplotype phase determination was a maximum parsimony approach by Clark.[4] While Clark's method is straight-forward and able to handle a potentially large number of markers, the algorithm depends on the identification of one completely unambiguous individual in the dataset. As the number of markers increases, the likelihood of finding an individual with no more than one heterozygous genotype becomes vanishingly small. More recently, the standard for haplotype phase inference has been PHASE v2.0[13] which uses a Bayesian approach and fits model parameters based on Markov Chain Monte Carlo. The algorithm uses an approximate coalescent prior probability assuming that haplotypes from the same population cluster together due to shared ancestry. While the program can deal with larger sets of genotypes by focusing on short subsets, it is still not practical for dense SNP data and large sample sets typical for GWAS.

Contemporary statistical algorithms have been developed to handle genome-wide association data. The three programs chosen for comparison in the current study were ENT, HaploRec, and Beagle.

ENT[7] attempts to maximize the likelihood of phasing using a count-based estimation of haplotype frequencies. It is capable of phasing long stretches of genotypes by using an overlapping window and batched implementation where a section of previously phased haplotypes is included in the model to aid in phasing the neighboring section of genotypes. The algorithm iteratively changes the phase of an unknown set of haplotypes until the calculation of entropy is minimized.

HaploRec[5] is a likelihood expectation maximization based method that considers the local regularities observed between haplotypes. The expectation maximization, or EM, algorithm was first used for haplotype phasing by Exoffier and Slatkin.[6] Population haplotype frequencies are initially estimated and iteratively updated to maximize the log-likelihood function to estimate an updated set of genotype frequencies. The frequencies are iteratively updated until the frequency converges.[6] The overall probability of the haplotypes is derived from the probability of local fragments. This method uses long variable fragment sizes based on a fragment threshold and takes advantage of information contained in long maps.

Beagle[2] uses a localized haplotype cluster model to define a hidden Markov model to determine the most likely phased haplotype for each individual. The haplotype is modeled as a string of sites with a finite number of states at each site, with an emission probability for any given state to transition to another state at the next site.[1,2] The Viterbi algorithm is applied to the hidden Markov model to determine the single most likely phased haplotype.[1] The inclusion of localization avoids sampling irregularities across long regions leading to false correlation between distant markers. Observed haplotypes are grouped into clusters depending on similarity allowing the model to adapt based on the data.

While these algorithms vary on their accuracy and efficiency, they tend to perform poorly across recombination hotspots, meaning the localized haplotypes may be phased correctly when considered individually, but placed on the wrong chromosome strand when combined with haplotypes from flanking regions.

4.2.4 CHAT: A new option for haplotype phase inference

A new algorithm for haplotype based analysis, known as Convergent Haplotype Association Tagging or CHAT, is currently under development in the Wilhelmsen lab. This algorithm searches for subsets of a population that have inherited a long shared haplotype, a CHAT, harboring a mutation from a common ancestor. Figure 4.1 illustrates the inheritance of an ancestral haplotype harboring a disease-causing mutation. As described earlier, identifying haplotypes implicitly can be difficult given the phase unknown genotype data provided from genome-wide association data. As an alternative, a pair-wise comparison of samples determines whether it is possible for each pair to share a haplotype at any given location. The only case when sharing is not possible is when two individuals are homozygous for the alternative alleles of a SNP (AA vs BB). The assumed haplotype in the region of sharing can be determined from the consensus haplotype of the individuals identified.

CHAT evaluates a pair-wise comparison of all subjects with genotype data for the potential for long shared haplotypes starting at each seed location. The distribution of observed sharing is assumed to be the sum of a distribution due to what would be seen by chance (which can be modeled as a Gaussian distribution) and the distribution of sharing of the rare pairs of individuals that have a long shared haplotype. CHAT models the combined distribution given the length of potential sharing and a prior probability of sharing that estimate the probability that the subjects share a long haplotype.

The length of the shared haplotype and the allele frequencies of specific alleles found on the haplotype are used to calculate the Pi-SMOR statistic, a measurement of the likelihood that the haplotype is inherited identical by descent or from a common ancestor. Pi-SMOR is the cumulative sum for markers in the putatively shared chromosome segment of the negative log of the single marker odds ratio of the probability of the observed genotypes assuming identity by descent of greater than 1 to the probability of the observed genotypes assuming identity by descent of zero. This measure reflects the uniqueness and length of a putative long shared haplotype. The Pi-SMOR statistic was developed to overcome the problem that an individual with a long string of heterozygous markers has the potential to share a haplotype with many other samples for the segment.

Using all the putative long shared haplotypes that individual X shares with others, a consensus haplotype for individual X can be predicted. CHAT uses the putative long shared haplotypes that involve subject X with the largest PiSMOR statistic to set phase and allows data from additional putative long shared haplotypes to further resolve the phase as long as they are parsimonious with the previously phased haplotype. After phasing subject X and other samples that putatively share a haplotype with subject X, CHAT tests whether the solutions are consistent. We describe this as a transitive test. We have observed that the most common reason why the phase solution for subject X (and the other subjects that putatively share a haplotype with subject X) are incompatible is that one of the putative long shared haplotypes is a false positive. By iteratively phasing subjects, and performing a transitive test to remove presumed false positive putative long shared haplotypes, CHAT converges on the most parsimonious haplotypes across long shared haplotypes. CHAT has the capability to use local entropy minimization to infer the haplotype of remaining chromosome segments as nearly as efficiently as other commonly used phasing programs.

The goal of the current study was to test the performance of CHAT on the haplotype phase inference of simulated data sets generated under a number of conditions and compare the performance to other haplotype phase inference programs in regards to the overall accuracy across the entire simulated region and the ability to minimize the localized haplotype effect.

4.3 Methods

4.3.1 Data Simulation

Fregene was used to simulate data with known haplotype phase.[3,8] Fregene is a forward data simulator that generates sequence data for large populations under various conditions allowing for control over selection, population level changes, rates of mutation, and patterns of recombination. The program tracks the sequence level changes of a population as mutations and recombination events arise through the creation of generations by simulated random mating. Each new individual in a subsequent generation is created from two random sequences from the previous generation.

The Population C dataset[3] available from the Fregene website was chosen for analysis as it is more likely to achieve the high level of complexity found in a real dataset than any user generated dataset. It was created with the intention of mimicking the patterns of genetic variation found in major human populations worldwide.[3] A series of events were simulated to mimic the creation of African, European, and Asian populations. (It should be noted that while these populations are referred to as "African," "European," or "Asian," they do not represent the true population of the same name and could not be considered as a subset of that population.) The Fregene website provides a set of simulations assuming neutral selection and another that includes selection. Throughout the simulations, mutations were allowed to occur at a rate of 1.5×10^{-8} and recombination at a rate of 1.1×10^{-8}.

To simulate the creation of current world population, a number of steps were taken to mimic the history of actual human populations. Figure 4.2 summarizes the data simulation. To begin, a population of 25K sequences is created in "Africa" and allowed to evolve for 125K generations. The African population further expands to 48K sequences and continues for another 17K generations. From this set, 4K sequences leave Africa, termed the Out Of Africa population, and the remaining bottleneck to 380 sequences. The African and Out of

Africa populations expand to 48K and 15.4K respectively and evolve for 3.5K generations. The Out of Africa population experiences a bottleneck to 1.3K sequences and splits into a "European" population of 320 sequences and an "Asian" population of 1K sequences. The European and Asian populations expand to 15.4K sequences and evolve for 2K generations. During this time, migration is allowed to occur at a rate of $0.8x10^{-5}$ between the Asian and African populations and $3.2x10^{-5}$ between the European and African populations. Finally each population evolves independently until each reaches 50K sequences.

4.3.2 Data Sampling and Dataset Creation

Sample,[3] a companion program for Fregene, was used to sample from the simulated sequence-level population data to obtain genotype data. For simplicity of comparison, sampling was performed on the African neutral selection population, the European neutral selection population, and the European positive selection population. African and European populations with neutral selection were chosen to compare differences in performance related to the level of linkage disequilibrium in a population. The European populations with neutral or positive selection were chosen to compare the effect of selection. The standard sampled population size was 1000 individuals. An additional dataset was created with 2000 individuals from the European positive selection set to understand the effect of sample size. Ultimately four sets were generated for comparison: African with neutral selection, European with neutral selection, European with positive selection, and European with positive selection and a large sample size.

4.3.3 Haplotype Phase Inference

Fregene generated genotype data was transformed into formats appropriate for each haplotype phase inference algorithm. ENT version 1.0.2 was applied to each dataset using default operating conditions suggested for optimal use, where free and locked window sizes are automatically selected and batching is included. Beagle version 3.1 was applied to each dataset using default operating conditions. Under these conditions, 4 haplotype pairs were sampled for each subject during each iteration and 10 iterations were applied per run. When HaploRec was applied to each dataset, a window size of 1000 markers and an overlap of 250 markers was chosen so that the program could complete the analysis given the available computational resources. Other operating conditions were default, where a variable order Markov Model with smoothed probabilities was used, iterations continued until the likelihood was unchanged, and a sequential pruning strategy was applied that builds haplotypes along the chromosome one marker at a time. CHAT was applied to the datasets under standard operating conditions. The operating conditions chosen for each program may not be the optimal conditions for accuracy, but default conditions were chosen in each case to understand the baseline accuracy levels.

4.3.4 Haplotype Phase Comparisons

To understand the accuracy of haplotype phase inference provided by each program, the experimentally phased haplotypes were compared to the real haplotype provided from Fregene. For performance comparisons, three measurements of accuracy were calculated. First, the single site error rate, described by Stephens and Donnelly,[14] provided a measurement of how well each program could recreate the whole phased chromosome. The single site error rate was calculated as the number of incorrectly placed alleles divided by total number of heterozygous markers averaged across samples. Second, the switch error rate was calculated as the number of contiguous heterozygous sites correctly phased or the number of switches needed to regain the original chromosome. The switch error is able to show how well each program performs across sites of recombination. A related measurement described by Lin[9] is the switch accuracy calculated as (het-1-sw)/(het-1). Like the switch error used in the present study, the switch accuracy represents the number of switches needed between neighboring pairs of heterozygous sites to regain the true haplotype sequence and is roughly equivalent to 1-switch error.

4.4 Results

The amount of change for each average error comparison between sample sets is shown in table 4.1 which displays the fold change for each comparison. Statistically significant changes are indicated in bold and were determined through a t test comparing the error measurements generated by each dataset.

4.4.1 Single Site Error Rate

With regards to single site error rate, CHAT performed better overall compared to all other programs. Beagle performed better compared to ENT and HaploRec which both had single site error rate near 50%, not much better than chance. There was wide variation in the individual sample single site error rates for Beagle and CHAT. ENT and HaploRec were more precise in their inaccuracy. Figure 4.3 shows a histogram of the average single site error rate for each program.

Single site error rates ranged between 0.16-0.19. Compared to the other haplotype phase inference programs, CHAT had a consistently lower single site error rate. CHAT demonstrated a significantly lower single site error rate with the African sample compared to the European sample and performed better with a dataset with more samples.

When considering program specific analysis, Beagle performed better than ENT and HaploRec with regards to single site error rate with average values ranging from 0.25-0.30. When considering the different datasets, the single site error rate was better for the African sample compared to the European sample, better with neutral selection than with positive selection, and better with a larger number of samples. While ENT performed poorly overall in regards to the single site error rate with average values ranging from 0.46-0.47. HaploRec performed slightly better than ENT with single site error rates ranging from 0.40-0.43. The single site error rate was lower for the European sample compared to the African sample and lower with neutral selection than with positive selection.

4.4.2 Switch Error Rate

When considering the switch error rate, Beagle performed better than all other programs with average rates between 0.01-0.03. Again, ENT had the highest switch error rates ranging between 0.25-0.36. CHAT (0.06-0.12) and HaploRec (0.08-0.18) had similar rates, with CHAT performing slightly better. Figure 4.4 shows a histogram of the average switch error rate for each program.

All four programs showed similar trends when considering individual dataset comparisons. With one exception, all programs had a significantly lower switch error for the African sample compared to the European sample, significantly lower with neutral selection than with positive selection, and significantly lower with a larger dataset. The one exception pertains to ENT which had a lower switch error rate for the European sample compared to the African sample.

4.5 Discussion

CHAT displayed a higher level of accuracy when compared to all other haplotype phase inference programs with regards to single site error rate. However when considering switch error, of which CHAT was predicted to perform better, CHAT was able to outperform both ENT and HaploRec. An increase in sample size consistently showed a significant improvement in the performance of all haplotype phasing programs, although the change was small for ENT and HaploRec. A larger sample size provides a more complete sampling of the overall population and as a result improves the representation of rare haplotypes. Specifically for CHAT, a larger dataset increases the enrichment of sharing between individuals. It is notable that in the large dataset nearly 30% of samples had a perfectly predicted haplotype phase configuration for the markers that were chosen for inclusion. Although the markers included in the phase determination was limited, an insurmountable number of individuals were phased for these markers with absolute accuracy.

CHAT was able to have a high degree of accuracy when phasing because it was selective when choosing markers to include in the finished haplotype phase configuration. The publicly available haplotype phase inference programs determine haplotype phase for all markers entered into the program; however, CHAT does not attempt to determine the haplotype phase for all of the markers. This is not surprising since CHAT was set up to determine phase only for markers covered by the shared haplotypes. For the African sample an average of 39% of markers were phased per sample, for the European sample with neutral selection 35.4%, for the European sample with positive selection 18.2%, and for the larger European sample 18.6%.

Figure 4.5 shows the number of samples for each marker for which phased haplotype data was successfully generated by CHAT. For the first 100 kb and the last 50 kb, phase was determined for around a hundred individuals. Phase inference may be difficult for these regions because there is less overlap at the beginning and end of the sequence. In the middle, most markers reach the upper limit of phased individuals with other markers scattered between phasing for 900 to 650 individuals. Around 900 kb, there is a reduction in coverage as the number of successfully phased individuals drops to a range between 850 and 550 individuals. The genomic region covered by this reduction has a lower marker density which is likely influencing the drop in coverage. This graph reveals that marker coverage is poor in regions of low marker density and at either end of the chromosome.

As shown by the graph, the maximum number of phase determined individuals never reaches the maximum of 1000. For the datasets with neutral selection, phase was determined for all samples for at least some of the markers. For the datasets with positive selection, CHAT provided phased haplotypes for 90% of the samples: 89% for the European sample with positive selection and 92% for the larger European sample with positive selection.

It was postulated that markers left unphased by CHAT were likely rare in the population and not contained in the common haplotypes easily identified as shared between individuals. To improve the coverage of CHAT, an additional filter was applied to the dataset with the larger sample size to remove markers with a minor allele frequency (MAF) less than 5%. A marker with a low minor allele frequency could be a recent mutation or simply not represented well in the subset of the population sampled from the original, larger population. In a real dataset, markers with low minor allele frequencies could also be due to genotyping error. Regardless of the cause, exclusion of these rare markers can reduce the specificity required for the identification of shared haplotypes by reducing the rare haplotypes represented in the sample and allowing more individuals to be identified with a common haplotype. The MAF filter greatly improved coverage and moderately improved phasing accuracy. While 8% of samples were completely unphased when considering the full range of MAF, only 1.5% of samples remained unphased after the application of the MAF filter. The average percentage of phased markers per individual also improved after the application of a MAF filter. When applied to the complete dataset, an average of 81.4% of markers per sample were unphased. After the MAF filter, an average of 60.5% of markers per sample were unphased. This indicates that nearly a quarter of the original unphased genotypes had low minor allele frequencies meaning that they were rare in the population and likely not included in the haplotypes shared between individuals in the population. The removal of markers with low MAF improved both the single site error rate and switch error by 1.3 fold. However with regards to both measures, the improvement did not change the performance rankings of CHAT in relation to the other phasing programs.

When comparing the performance on the various datasets within programs, there was improved accuracy for the African dataset compared to the European as well as for the neutral and compared to positive selection seen nearly consistently across haplotype phasing programs. Haplotype phase inference should be more difficult in an African population compared to European since there should be greater diversity and more complex, shorter regions of linkage disequilibrium. But this also results in a higher number of polymorphic sites and thus a denser genetic map than the younger European population. Likewise, positive selection will lead to less variation in haplotypes present in a population, but more polymorphic sites can develop under neutral selection. This improved accuracy is likely driven by the density of markers found in each dataset. It is logical to assume that a denser set of markers will provide a more complete picture of haplotype diversity and better frequency predictions leading to increased accuracy overall.

The main goal of this study was to understand the accuracy of the haplotype phasing capabilities of a novel algorithm Convergent Haplotype Association Tagging, or CHAT. CHAT is unique compared to other haplotype phase inference programs because it restricts the search space of possible haplotype configurations by identifying haplotypes shared in a population that were inherited from a common ancestor. When considering single site error rate, or how many markers were placed together on the same chromosome, CHAT consistently outperformed other phasing programs applied to the same data. For switch error, or how flanking segments of the chromosome are placed, CHAT performed better when compared to ENT for the African set and the large sample set.

Clearly the strength of CHAT lies in the ability to accurately predict the haplotype phase in regions covering long shared haplotypes for datasets with a large number of samples. CHAT is very accurate with regards to single site error rate and moderately accurate concerning switch error on the markers that it does choose to phase. To improve coverage, it is important to filter markers with low minor allele frequencies and to maximize the number of samples. It is probable that coverage will improve with larger sample sizes given the trends comparing 1000 and 2000 samples; however, more complete coverage will likely come from reduced stringency in sharing or through an alternative method that can be applied to fill in the unphased gaps. While the current implementation of CHAT is unable to meet the complete coverage provided by current haplotype phase inference algorithms, it is possible to apply the entropy minimization technique implemented in ENT to determine phase for the markers left out by CHAT analysis. It is likely that the high accuracy phasing across shared haplotypes provided by CHAT combined with the moderate accuracy provided by entropy minimization would provide both high accuracy predictions and better coverage.

4.6 References

1: Browning SR. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet. 2008 Dec;124(5):439-50. Epub 2008 Oct 11. Review.

2: Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007 Nov;81(5):1084-97. Epub 2007 Sep 21.

3: Chadeau-Hyam M, Hoggart CJ, O'Reilly PF, Whittaker JC, De Iorio M, Balding DJ. Fregene: simulation of realistic sequence-level data in populations and ascertained samples. BMC Bioinformatics. 2008 Sep 8;9:364.

4: Clark AG. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol. 1990 Mar;7(2):111-22. Review.

5: Eronen L, Geerts F, Toivonen H. HaploRec: efficient and accurate large-scale reconstruction of haplotypes. BMC Bioinformatics. 2006 Dec 22;7:542.

6: Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol. 1995 Sep;12(5):921-7.

7: Gusev A, Măndoiu II, Paşaniuc B. Highly scalable genotype phasing by entropy minimization. IEEE/ACM Trans Comput Biol Bioinform. 2008 Apr-Jun;5(2):252-61.

8: Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ. Sequence-level population simulations over large genomic regions. Genetics. 2007 Nov;177(3):1725-31. Epub 2007 Oct 18.

9: Lin S, Cutler DJ, Zwick ME, Chakravarti A. Haplotype inference in random population samples. Am J Hum Genet. 2002 Nov;71(5):1129-37. Epub 2002 Oct 17.

10: Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, Donnelly P; International HapMap Consortium. A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet. 2006 Mar;78(3):437-50. Epub 2006 Jan 26.

11: Niu T. Algorithms for inferring haplotypes. Genet Epidemiol. 2004 Dec;27(4):334-47. Review.

12: Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet. 2002 Jan;70(1):157-69. Epub 2001 Nov 26. Erratum in: Am J Hum Genet. 2006 Jan;78(1):174.

13: Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet. 2005 Mar;76(3):449-62. Epub 2005 Jan 31.

14: Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet. 2003 Nov;73(5):1162-9. Epub 2003 Oct 20.

15: Zhao H, Pfeiffer R, Gail MH. Haplotype analysis in population genetics and association studies. Pharmacogenomics. 2003 Mar;4(2):171-8. Review.

Table 4.1 Fold change between comparisons

The amount of difference between comparisons. Values less than one indicate the first set had lower error rates, while values greater than one indicate the second had lower error rates.

	Single Site Error Rate					**Switch Error**			
	Beagle	**ENT**	**Haplo**	**CHAT**		**Beagle**	**ENT**	**Haplo**	**CHAT**
			Rec					**Rec**	
African vs European	**0.95**	**1.02**	**1.07**	**0.9**		**0.79**	**1.04**	**0.86**	**0.69**
neutral vs positive	**0.91**	1	**0.94**	0.98		**0.38**	**0.71**	**0.5**	**0.73**
sample size small vs large	**1.22**	1	1	**1.17**		**1.82**	**1.01**	**1.04**	**1.23**
Compared to CHAT									
African	**1.54**	**2.87**	**2.61**	--		**0.17**	**4.27**	**1.25**	--
European neutral	**1.45**	**2.53**	**2.18**	--		**0.15**	**2.85**	**1.01**	--
European positive	**1.57**	**2.48**	**2.28**	--		**0.28**	**2.93**	**1.47**	--
large sample	**1.5**	**2.92**	**2.68**	--		**0.19**	**3.58**	**1.75**	--

Table 4.2 Significance Testing

Provides the p values, as calculated through a student t test, used to determine significance in the comparisons between programs as seen in table 4.1.

	Single Site Error Rate					**Switch Error**			
	Beagle	**ENT**	**Haplo**	**CHAT**		**Beagle**	**ENT**	**Haplo**	**CHAT**
			Rec					**Rec**	
African vs European	**4.32E-02**	**1.15E-18**	**6.21E-29**	**7.25E-03**		**3.31E-19**	**1.25E-20**	**1.26E-49**	**1.21E-20**
neutral vs positive	**3.89E-05**	3.10E-01	**1.13E-18**	7.09E-01		**1.19E-225**	**0.00E+00**	**0.00E+00**	**2.97E-15**
sample size small vs large	**3.81E-04**	8.01E-01	6.11E-01	**9.47E-16**		**3.44E-134**	**3.40E-03**	**5.93E-07**	**1.00E-21**
Compared to CHAT									
African	**5.21E-35**	**0.00E+00**	**4.63E-299**	--		**2.71E-123**	**0.00E+00**	**1.11E-15**	--
European neutral	**2.19E-32**	**0.00E+00**	**2.90E-232**	--		**2.34E-170**	**0.00E+00**	6.22E-01	--
European positive	**5.26E-49**	**7.79E-259**	**2.60E-218**	--		**1.65E-100**	**0.00E+00**	**8.98E-49**	--
large sample	**1.10E-97**	**0.00E+00**	**0.00E+00**	--		**1.55E-131**	**0.00E+00**	**3.93E-246**	--

Figure 4.1 Inheritance of Ancestral Haplotype

Illustrates the inheritance of an ancestral haplotype harboring a mutation after several generations. The initial chromosome represents the ancestral chromosome containing a trait-causing mutation. After several generations, individuals that have inherited the mutation also contain some amount of ancestral haplotype surrounding the mutation.

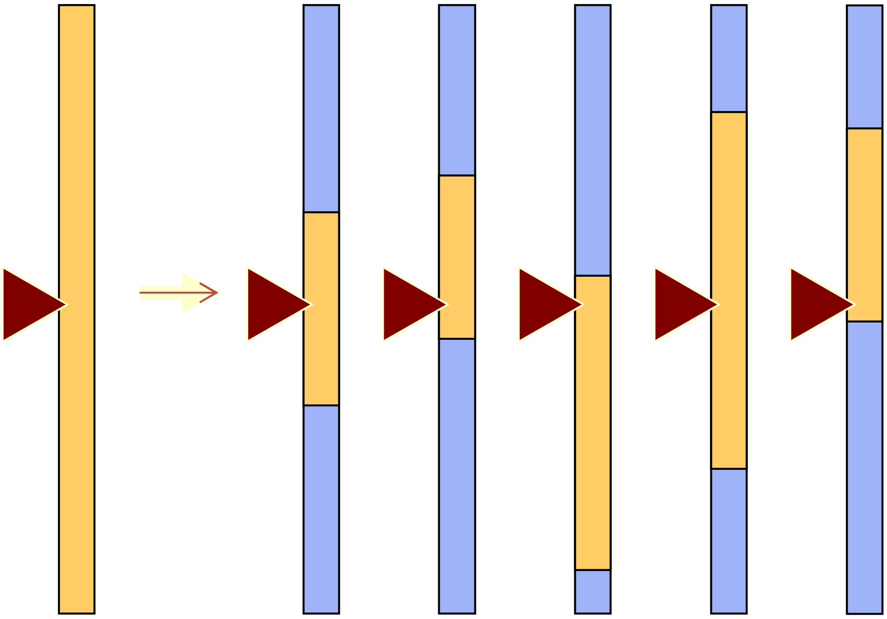


Figure 4.2 Data Simulation

Diagram of Fregene generated data simulation. A population of 25K sequences is created in "Africa" and allowed to evolve for 125K generations. The African population further expands to 48K sequences and continues for another 17K generations. From this set, 4K sequences leave Africa, termed the Out Of Africa population, and the remaining bottleneck to 380 sequences. The African and Out of Africa populations expand to 48K and 15.4K respectively and evolve for 3.5K generations. The Out of Africa population experiences a bottleneck to 1.3K sequences and splits into a "European" population of 320 sequences and an "Asian" population of 1K sequences. The European and Asian populations expand to 15.4K sequences and evolve for 2K generations. During this time, migration is allowed to occur at a rate of 0.8×10^{-5} between the Asian and African populations and 3.2×10^{-5} between the European and African populations. Finally each population evolves independently until each reaches 50K sequences.

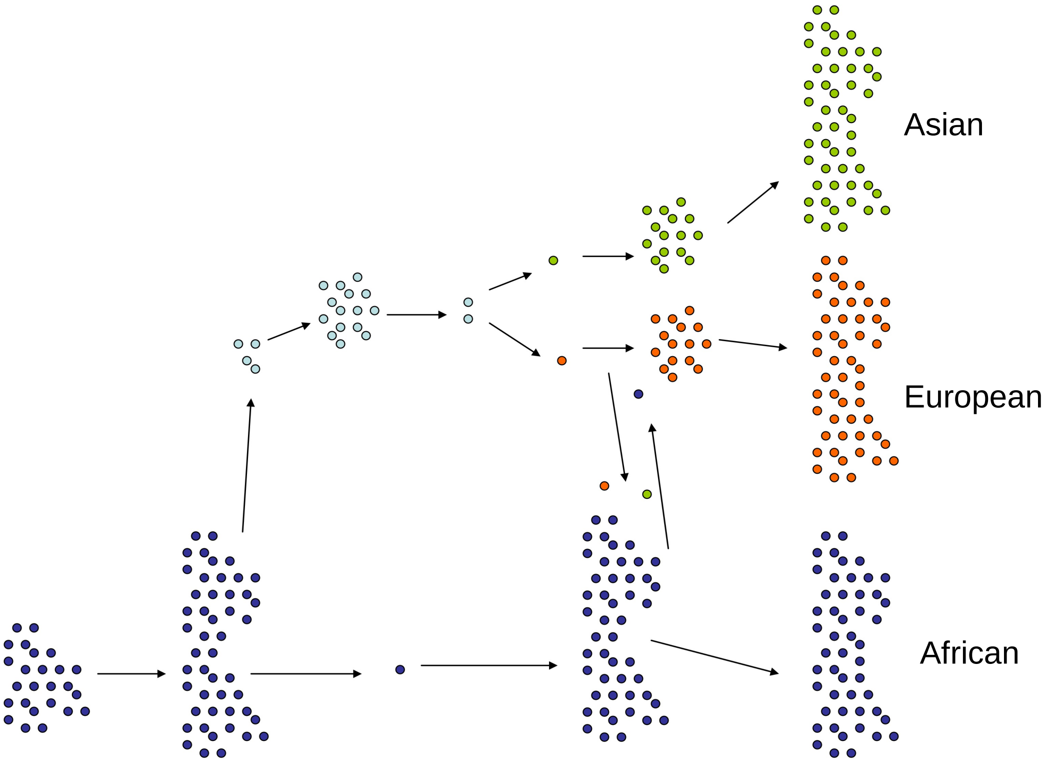


Figure 4.3 Single Site Error Rate

Average single site error rates for the tested datasets for each haplotype phase inference algorithm. Error bars represent the standard error of the measurement.

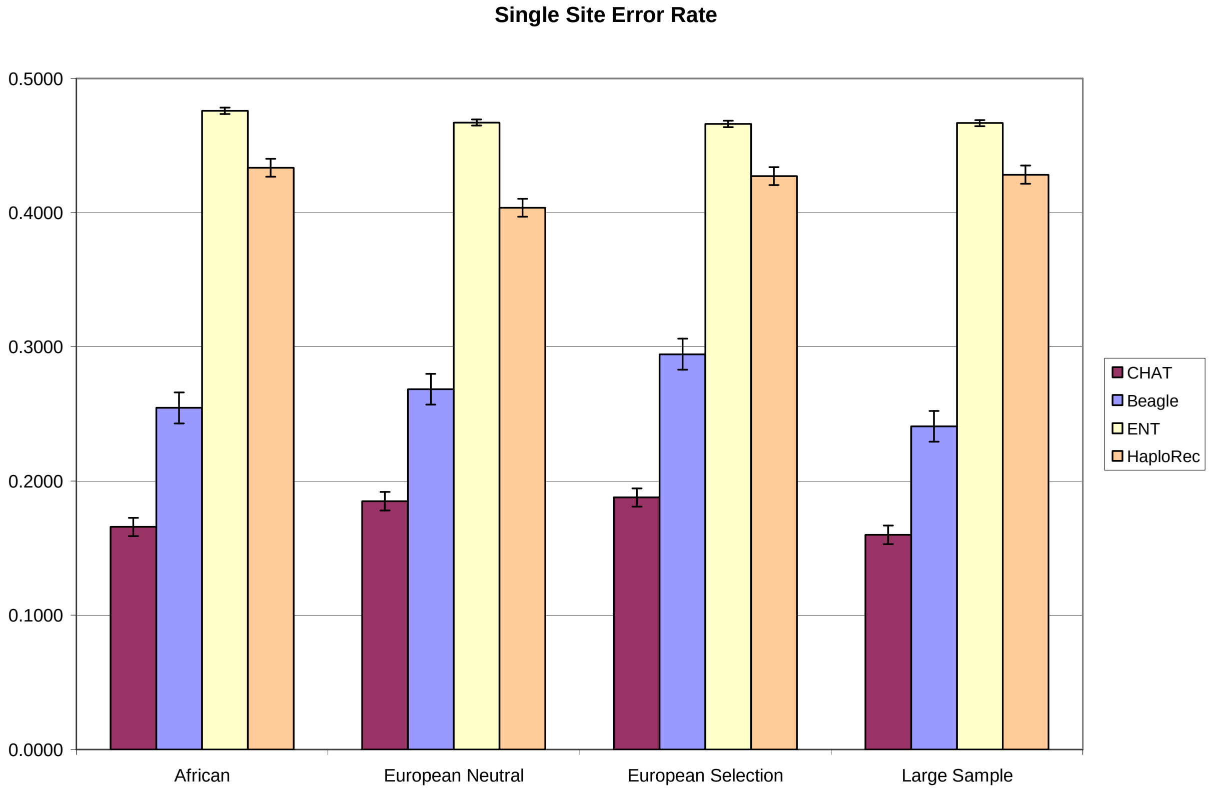


Figure 4.4 Switch Error Rate

Average switch error rates for the tested datasets for each haplotype phase inference algorithm. Error bars represent the standard error of the measurement.

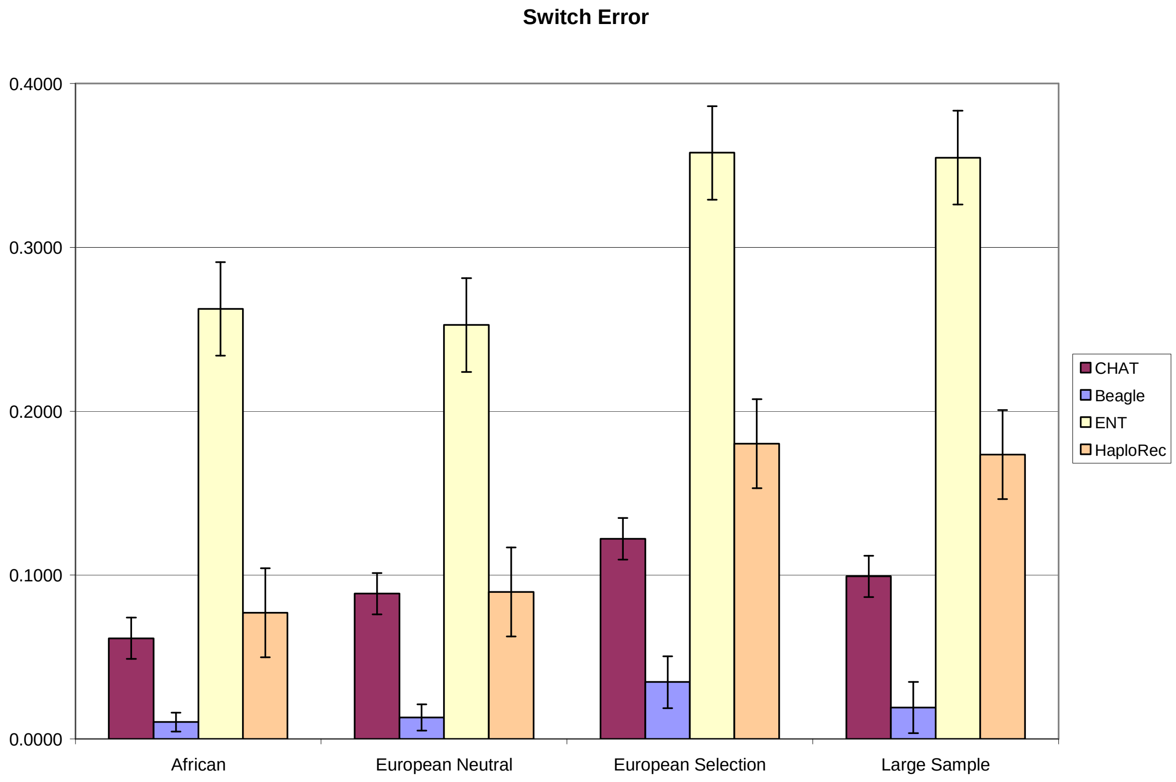
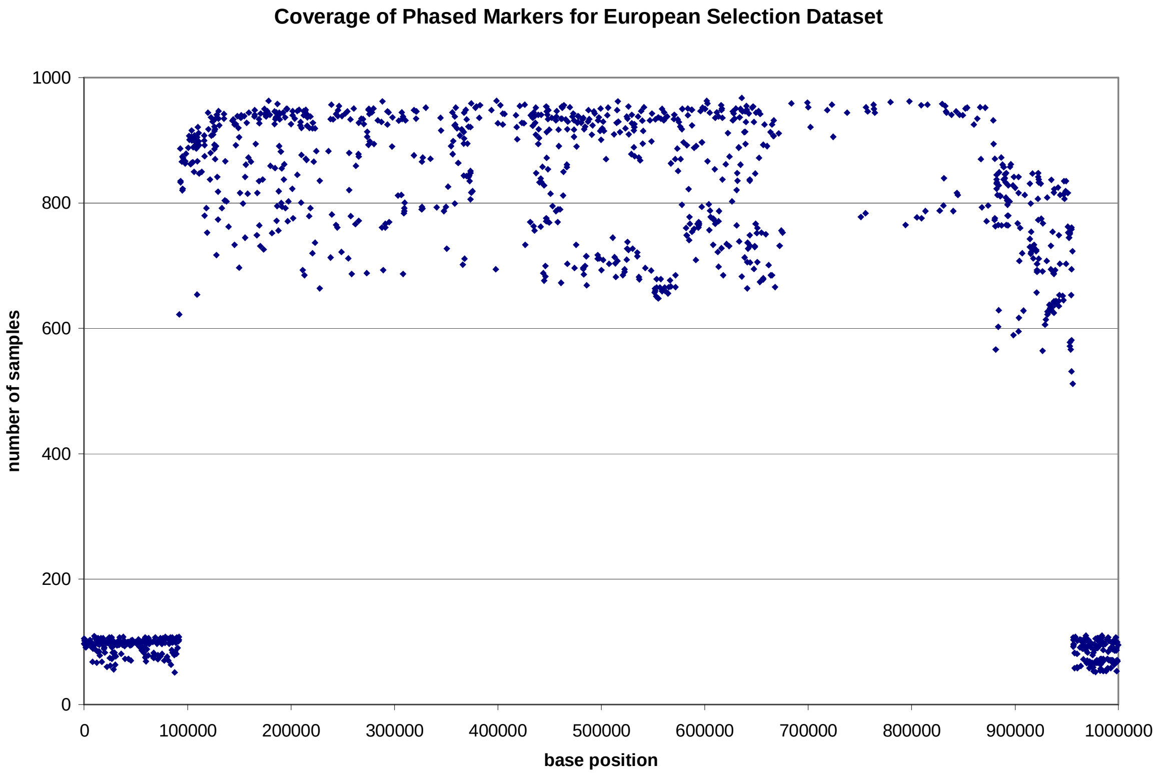


Figure 4.5 Coverage

Number of samples for which phase was predicted through CHAT for any given position.



Chapter 5 – Conclusions, Limitations, and Future Directions

The fifth and final chapter of this dissertation will review the conclusions from each project. Also included is a discussion of some of the major limitations encountered in the study of complex disease and how each relates to the current projects. At the end, future directions for each project will be suggested as well as some predictions on the future of the genetics of complex human disease as a whole.

5.1 Project Specific Conclusions

5.1.1 The investigation of *CYP2E1* with the level of response to alcohol

Combined linkage and association analysis of the gene *CYP2E1* was able to reinforce the implication that this gene, known for its metabolism of ethanol, affects the level of response to alcohol. A genome wide linkage scan of sibling pairs with an alcoholic parent originally suggested the involvement of *CYP2E1*. Variance component linkage analysis with the inclusion of a number of SNPs located in *CYP2E1* confirmed the suspected linkage between this region at the end of chromosome 10 and the level of response to alcohol. The reduced evidence for linkage after the addition of sibling pairs was attributable to a single family with unreliable phenotype measurements. A number of SNPs genotyped from *CYP2E1* were found to be associated with the level of response to alcohol through a mixed model regression, but copy number was not found to be associated. So either the level of response to alcohol is not affected by carrying multiple copies of the risk allele, or individuals with copy number changes did not carry the minor allele due to a low minor allele frequency (MAF = 0.056). Testing for linkage while simultaneously modeling association allows for the confirmation of a causal variant or a variant in complete LD with a causal variant. However tests showed that the most associated SNP could not be conclusively ascribed as a causal variant.

It is likely that there are multiple causal variants in CYP2E1 in different families affecting the level of response to alcohol or that the single causal variant was not included in the SNPs chosen for genotyping in the gene. If the latter is true, it is likely that the causal variant is upstream of the most associated SNP due to the trend in LOD score reduction. The LOD score was reduced the most by the SNP that showed the best evidence for association, and the amount of reduction decreased for markers downstream on the chromosome. The most associated SNP is located upstream from *CYP2E1*, so this unidentified causal variant could be located even farther upstream and have some kind of regulatory effect on the gene. Although a causal variant could not be identified, evidence from linkage, association, and knowledge of biological pathways indicate that changes in or near *CYP2E1* regulate the activity or expression of this gene, thus affecting how the brain perceives alcohol leading to differences in the response to alcohol.

5.1.2 The association of the *MAPT* region with Pick’s complex diseases

A genome-wide association was performed on four different sub-types of Pick’s disease. While a very large number of SNPs were found to be associated with different diseases, the results reported focus on the inverted region on chromosome 17 surrounding the gene Microtubule Associated Protein Tau, or *MAPT*. The association of this region with Progressive Supranuclear Palsy (PSP) and Corticobasal Degeneration (CBD) supports previously reported associations of this gene with these diseases. PSP and CBD most commonly include aggregates of tau, the protein product of *MAPT*. The study was also able to replicate the overrepresentation of the H1 haplotype in this region in PSP and CBD. Either the H1 haplotype contains causal variants increasing risk, or the alternative H2 provides some protection from PSP and CBD but not for FTD. Due to the inversion surrounding the gene, it is difficult to narrow down an exact causal variant.

5.1.3 The evaluation of a novel algorithm for haplotype phase inference

A novel algorithm called CHAT was created to determine the haplotype phase of an unrelated set of individuals for genome-wide genotype data. The current study aimed to understand the accuracy of haplotype phase inference on datasets simulated under different conditions as well as compared to publicly available haplotype phase inference programs. CHAT showed significant improvement regarding the single site error rate when compared to the other haplotype phase inference programs for all datasets. Regarding switch error, CHAT was able to outperform ENT and HaploRec. The coverage of haplotype phase inference performed by CHAT was very selective and directly dependent on the identification and overlap of shared haplotypes. Although up to 30% of samples had haplotype phase configurations with perfect accuracy for the dataset with the larger number of samples, CHAT was not able to determine the phase for the complete length of the region for all samples. The strength of CHAT lies in the ability to accurately predict the haplotype phase in regions covering long shared haplotypes for datasets with a large number of samples. CHAT is very accurate with regards to single site error rate and moderately accurate concerning switch error on the markers that it does choose to phase. While the current implementation of CHAT is unable to meet the complete coverage provided by current haplotype phase inference algorithms, it is possible to apply the entropy minimization technique implemented in ENT to determine phase for the markers left out by CHAT analysis. It is likely that the high accuracy phasing across shared haplotypes provided by CHAT combined with the moderate accuracy provided by entropy minimization would provide both high accuracy predictions and better coverage.

5.2 Limitations and Challenges of Genetic Analysis

The success of the studies presented here, as well as other studies of the genetics of complex human traits, is dependent on the ability to maximize the power to detect a genetic effect and to minimize the occurrence of false positive findings. Many genome-wide association studies are not powered well enough to detect variants with large effects that can be distinguished from the noise of false positive associations. While false positives can obscure the identification of true positives, overly strict exclusion criteria can be just as detrimental when true positives are eliminated along with the false positives. What follows is a discussion of the many considerations to maximize power and achieve a balance between true and false positives. Attempts to maximize power can be divided into six categories: sample considerations, phenotype considerations, the linkage disequilibrium or correlation between markers in a population, control for false positive results, replication of positive findings, and the genetic component of a phenotype or effect size.

5.2.1 Sample Considerations

The power to detect a genetic influence for a trait increases with the inclusion of more samples. The minimum number of samples needed to glean the most power from a genetic study and to minimize genotyping costs can be estimated based on features of the causal variant to be detected, including allele frequency, penetrance, and magnitude of effect, a measurement of the amount of risk provided by a variant. There is some uncertainty to this calculation because features of the unknown causal locus must be estimated as well. But more often the sample size of a study is limited based on the availability of suitable affected individuals, as was the case with the Pick's disease GWAS which focused on rare neurodegenerative diseases.

There is a tendency to collect large sample sets and combine samples genotyped at multiple research centers in order to maximize the power to detect variants with smaller effects. While this does increase the overall power, genotypes and phenotypes may not be consistent across research centers leading to biases resulting from the added heterogeneity. The *CYP2E1* alcoholism project is one example where the intent to increase the power to detect an effect by the addition of extra samples was detrimental to the overall evidence for linkage. Including more samples in a study will increase the chance for heterogeneity leading to the apparent association of regions unrelated to disease status thus increasing the chance for false positives or, as seen in the CYP2E1 project, obscuring the evidence for a genetic region that may actually affect disease—a false negative. Heterogeneity can occur through incorrect sample definitions, either by poorly defined cases and controls or by differences in population ancestry.

Through the identification of appropriate cases and controls heterogeneity can be minimized. For a qualitative trait of case/control, the status must be explicitly defined and easily differentiated. The choice of phenotype will be discussed in detail later, but the phenotype should be easy to measure and reliably consistent. Heterogeneity can be minimized if the cases share a specific subtype of a phenotype. Controls should come from the same population as the cases and be at risk for disease, but must be excluded from the disease or trait of interest. In studies of substance abuse, like alcoholism, the controls must have had exposure to the addictive substance but not be addicted. For a rare disease such as Pick's complex it is unlikely that any control individual will be affected but it is important for age-related diseases like dementia be matched for age. The same is true for gender and any other variable unrelated to a genetic effect that could have an influence on the trait, like environmental exposures.

Another consideration for reducing heterogeneity is the effect of population ancestry. Unlike linkage analysis, which controls for ancestral background internally through the inclusion of family members, association analysis is highly vulnerable to population stratification creating apparent correlations with variants unrelated to the trait. Population stratification occurs when the ancestral composition differs between cases and control. If the frequency of a specific allele is also different between the populations found in the sample, that allele could show a false association simply due to population effects. The samples in the association study of Pick's Disease were screened for European ancestry and population stratification was undetected by Eigenstat, a program that looks for stratification in GWAS data. The effects of population stratification are minimized by restricting samples to a single continent of origin. But stratification may even exist in a single continent if samples come from different regional areas introducing more cryptic differences. One way to reduce the effects of cryptic population stratification is to choose a homogeneous population isolate with a more recent common ancestor. In a homogeneous population most mating occurs within the population and genetic variants affecting a trait are more likely to come from the same mutation and thus be located on the same haplotype.

Genetic heterogeneity can also occur when multiple variant lead to the same phenotypic outcome. Methods to minimize the heterogeneity created by the choice of phenotype will be discussed in the next section.

5.2.2 Phenotype Considerations

A search for regions of the genome that affects a trait will perform poorly if the phenotype is not well defined. The *CYP2E1* linkage study illustrates the importance of accurately reported phenotypes. Skewed phenotype scores from one family were enough to introduce heterogeneity into the sample and obscure the genetic effect. This addresses the importance of accurate phenotype measurements that can be reliably measured. Complex human disease can be very heterogeneous; some diseases are defined by the occurrence of any number of symptoms. For example, in the association analysis for Pick's disease, each subtype of disease is defined by a number of clinical and pathological features, but the exact set of features can vary between individuals. A trait needs to have a clear definition and be easily determined. The different subtypes of Pick's disease can be pathologically confirmed, but are often misdiagnosed at the clinical level.

The choice of phenotype is also a concern for heterogeneity minimization. Often the phenotype used in GWAS is a dichotomous classification of disease status. Focusing instead on a particular symptom, or endophenotype, of a disease can reduce heterogeneity and make a genetic cause easier to identify. The CYP2E1 project focused on a quantitative trait, so there were no case/control determinations. Instead, sample collection was restricted to individuals that had an alcoholic parent, thus increasing that individual's own risk for alcoholism. Focusing on individuals with a known high risk for alcoholism reduces the heterogeneity that could be encountered in a study of the risk for alcoholism in the general population. Additionally, a specific measurement such as the level of response to alcohol is more likely to be affected by a similar genetic cause in different individuals compared to considering the broad definition of alcoholism.

5.2.3 Linkage Disequilibrium Considerations

The power to detect an untyped causal variant depends on the level of correlation between the genotyped markers and the causal variant. The number of samples needed to detect an unmeasured causal marker is proportional to the level of linkage disequilibium, or r^2 a measure of correlation, between the typed marker and the untyped causal marker. The likelihood of measuring a marker in linkage disequilibrium with a trait causing locus can be improved by increasing the density of genotyped markers or by choosing a population with a recent common ancestor. In such a homogeneous population, the regions in linkage disequilibrium will be larger, so fewer markers will be required to cover all of the variation. Alternatively, the power to locate the exact genetic effect increases with the number of generations to a common ancestor due to the smaller size of region in linkage disequilibrium.

The power of a genetic study increases with the inclusion of more distantly related pairs. Association maximizes this power by comparing very distantly related individuals.

Compared to association based analysis, linkage analysis can map a trait to a very large region since linkage analysis is based on the inheritance of chromosomes from a very recent ancestor. In the case of the sibling pairs used in the current study, that common ancestor would be the parents. As a result, the comparison is made after only one meiosis and one chance for recombination leading to very large regions in LD. This is illustrated in the linkage study to find regions that affect the level of response to alcohol where the linkage peak covered a large region. While *CYP2E1* provides the most logical evidence for involvement with our trait, there is not always such an obvious candidate gene in a region of linkage. Association analysis provides a way to map a trait locus to a much finer region because the regions of LD are smaller. Association analysis can be considered as an extreme version of linkage analysis. It is often assumed in association that the samples collected from a common population are unrelated. In reality, every individual in a population can be traced back to a common ancestor. With many more generations, more recombination can occur, dividing the chromosomal region, or haplotype, from the common ancestor into much smaller pieces. The length of haplotypes shared in a population from a common ancestor or the extent of linkage disequilibrium depends of the number of generations that have passed and the interrelatedness of the population. An older population will have very short regions of LD while a younger, less heterogeneous population will have longer regions of LD.

A densely mapped set of markers will increase the power to detect a genetic effect because it increases the probability that a genotyped marker is likely to have a sufficient correlation with the causal variant. Regions of linkage disequilibrium in a linkage study are much larger requiring only a few markers to represent the genome. In the genome-wide linkage scan to search for regions of the genome affecting the level of response to alcohol, only 811 microsatellite markers were genotyped across the entire genome. A much denser map is required for genome-wide association studies because the biallelic SNP contains less information and the regions of linkage disequilibrium are smaller. A more densely genotyped map will increase the likelihood that a measured allele will be in linkage disequilibrium with the causal allele as it will be able to capture more of the common variation in the genome. Similarly, the required number of markers to capture the variation in a population is dependent on the number of generations to the common ancestor. An older population requires more markers to represent the larger number of LD regions.

When the Pick's Disease project was performed the Affymetrix genotyping chip only included 500 thousand markers, but current genotyping technology has allowed for the inclusion of up to a million markers. Additionally, using the patterns of linkage disequilibrium among common SNPs captured by the International HapMap project,[6] it is possible to increase the density of markers in a study by predicting, or imputing, markers not originally included on the genotyping chip. Imputation was performed for the Pick's disease project, but the prediction of untyped markers provided limited benefit because the region had nearly perfect correlation due to the high degree of linkage disequilibrium.

5.2.4 Data Processing Considerations

An overabundance of false positive results in a study will limit the power to identify an actual causal variant. Careful study design through sample and phenotype guidelines, as described earlier, seeks to reduce false positives but is not enough to prevent all occurrences. Through improved genotype calling algorithms and extensive quality control, researchers aim to minimize the prevalence of false positive results.

The method of genotyping and subsequent data processing has a large influence on the success of a genetic study, especially for a genome-wide association study. Because so many markers are being measured, even a small error rate will lead to a sizable excess of false positives. It is very important that genotype calls be as accurate as possible. In a genome-wide association study, genotype calls for each marker are made automatically using a genotype calling software. Genotype calls for the Pick's disease project were generated by the BRLMM algorithm which categorizes genotype based on the distance of each sample measured from the center of the three predicted clusters.[1] The algorithm underlying BRLMM was explained in detail in chapter 1 of this dissertation. At the time of the Pick's disease study, BRLMM was the best available genotype calling method. While the BRLMM algorithm was an improvement over previous genotype calling software, experience revealed that many markers are genotyped poorly. The algorithm often misclassifies homozygous individuals as heterozygous, especially at low intensity levels or when clusters overlap. Experience has shown that systematic bias or batch effects in the genotyped samples will lead to poorly called genotypes. Batch effects can occur due to any difference that could lead to biased genotype calls and improper associations. These effects can come from differences in sample handling prior to genotyping, differences between plates of samples, or genotypes that were generated at different research centers. After the determination of genotype, possible genotyping errors are removed through stringent quality control although experience has shown this is not always the case.

Many quality control criteria were applied to the Pick's disease dataset to reduce the chance for false positive associations. The confidence score given by the genotype calling software for each genotype is used to distinguish acceptable genotypes from questionable genotypes. A call rate is calculated for each sample and each marker and individual samples or markers are removed if too many fail to pass the predetermined threshold rate. Additionally, a minor allele threshold is set to remove markers that are rare or monomorphic in the genotyped population. Markers with low minor allele frequencies are especially susceptible to errors in genotyping and often lead to over-inflated measures of significance (extremely low p values). Hardy Weinburg Equilibrium describes the allele frequencies expected from a stable population and is a calculation comparing the measured allele frequencies compared to the expected. Markers that do not meet criteria for Hardy Weinburg Equilibrium are likely to be a result of unreliable, biased genotype calls and are removed from analysis. Often the best way to exclude poorly called markers is to visually inspect the probe intensity genotype plots and manually make genotype designations. This can be easily performed with small scale Taqman genotyping efforts, but is nearly impossible when dealing with half a million markers at a genome-wide scale.

5.2.5 Replication

Even after stringent quality control and data processing, not all false positive results can be removed. After any genome-wide scan for genetic variants, replication is required to confirm the positive results and increase the likelihood of the identification of biologically appropriate variants. With so many statistical tests performed in an association analysis combined with the tendency for genomic errors and heterogeneity to cause biases, there is an increased potential for false positive findings, or associations that occur by chance, interspersed with true positive associations. The most accepted correction for multiple testing is through the Bonferoni correction which divides alpha, the predetermined significance threshold level for a single test or how often a test should reach significance by chance, by the number of independent tests performed. This method is often described as overly conservative because SNP genotypes are not completely independent due to linkage disequilibrium. Even so, the commonly accepted threshold for genome-wide significance is $5x10^{-8}$ which corrects for one million SNPs. The results from the Pick's disease project did not reach genome-wide significance under this criterion, but many of the unreported associations did reach that level. Considering so many different markers in this region surrounding MAPT independently show the same level of association and knowledge of the involvement of the gene with these diseases, it is certain that the association seen in the inversion across chromosome 17 is not a false positive.

Many reported findings have been difficult to reproduce. Either these studies are reporting false positives or the reproduction approach is not optimal. For the best likelihood of replication, an independent sample should be taken from the same population, the same phenotypes should be measured, and the same markers should be genotyped. A true reproduction will implicate the same SNP for the same allele in the same direction (meaning the same allele increases risk or provides protection).[11] Often associations found in one population do not translate to other populations due to differences in ancestry and allele frequencies.[5] When looking at variants from *CYP2E1* that confer risk for alcohol related phenotypes, there is a lot of discrepancy between studies because they draw their samples from different populations with different allele frequencies and they focus on different phenotypes. This is discussed in detail in Chapter 2.

For a replication study, often the most significantly associated markers are chosen for genotyping in an alternate set of samples. Errors in genotyping and biases in sample sets more often lead to extremely significant associations, while the findings most likely to be robust across replication samples are less impressive.[11] Often the markers with the best evidence for association, or most extreme p values, are chosen for replication. However, markers with less extreme association levels are more likely to replicate because they are less likely to be false positive results. Experience from the Pick's disease dataset and another GWAS not discussed in this dissertation has shown that these studies generate an overabundance of highly associated markers suggesting high false positive rates. In these studies, markers with the most inflated association p values have very unlikely genotype counts and visual inspection of probe assay plots reveal that unreliable genotype call have been made.

5.2.6 Effect Size (or the Genetic Component of a Trait)

To be successful, a trait considered in a genetic study must have a large genetic component. Obviously, it should be easier to ascribe a genetic cause to a trait caused by larger genetic effect. The heritability of a trait refers proportion of the trait variance attributable to genetic effects and can be measured through twin studies by the concordance between dizygotic twins compared to monozygotic twins. The level of response to alcohol was chosen as the phenotype for the *CYP2E1* linkage study because evidence from previous studies showed that the trait was highly heritable and consistent in families where children of alcoholics had lower levels of response to alcohol than control individuals. Additionally, the sibling relative risk for a trait or the disease risk for the sibling of an affected individual can be an indication of the feasibility of a study. The sibling relative risk is calculated as the ratio between the sibling risk and the overall population risk. A disease like FTD occurs in families fairly often but has a vary small overall population risk, so the sibling relative risk is very large especially compared to other more common dementia related diseases like Alzheimer's Disease. Alternatively diseases like PSP and CBD occur sporadically thus have a vanishingly small sibling relative risk suggesting that a genetic cause for these diseases would be more difficult to identify.

A large proportion of the heritability for most common, complex diseases has been left unexplained by identified genetic variants.[5] Association studies were created under the assumption that common genetic variants are likely to cause common diseases and these common variants with smaller effect sizes could be detected in a population better than by linkage. Association has increased power to detect variants that have a small effect of a trait than linkage.[4] However, association analysis only has power to detect common causal variants that are in linkage disequilibrium with genotyped SNPs. The detection of rare variants with larger effect sizes is technically feasible but would require unattainably high sample sizes. Typically linkage analysis has greater power to detect rare variants with large effects for rare diseases since these traits tend to be found in multiple affected members of a family. But linkage is less successful than association analysis for complex traits influenced by multiple alleles in different genes contributing only a small amount to the overall risk. Haplotype analysis, especially as constructed through CHAT, provides a bridge between linkage and association since it can detect rare variants with small effects for a relatively rare trait.

The Common Disease Common Variant, CDCV, hypothesis assumes that many common SNPs with small effect sizes and low penetrance could be detected to affect a common trait in a population if enough individuals were genotyped.[3,7] Current genotyping platforms are able to capture up to 80% of markers in the Caucasian population with minor allele frequencies greater than 0.05 but misses any markers with rarer allele frequencies.[3] The majority of markers found to be associated by GWAS have small effect sizes (1.1-1.5) and explain only a small proportion of the estimated heritability or underling genetic cause (5-10%).[10,12] It is likely that the CDCV hypothesis is not entirely correct and the rest of the heritability may be explained by rare variants, copy number changes, and structural variation, as well as interactions between genes or between genes and environment. To fully understand the genetic heritability of complex traits will require the investigation of these variants by deep sequencing or new genotyping platforms that can better capture rarer genetic changes.

Variants able to explain more of the heritability with large effect sizes are expected to be rare and not well represented by common variation. This is logical considering a mutation with a large effect on disease would likely be deleterious and be minimized in a population by selection since it would decrease reproductive fitness.[2] Additionally stabilizing selection would seek to minimize the extremes of the trait caused by variants with larger effects.[10] The best option for capturing rare variants with large effect sizes will be with sequencing either through targeted regions likely to harbor mutations or through a whole genome approach. The 1000 genomes project plans to sequence a thousand individuals with the hope of capturing rarer genetic variation with minor allele frequencies between 1-5%.[3]

5.3 Project Specific Future Directions

5.3.1 The investigation of *CYP2E1* with the level of response to alcohol

It seems evident that sequence changes in or around *CYP2E1* are affecting the expression or activity of this gene. To understand how *CYP2E1* is changed in response to chronic alcohol intake, the redox pair compound ratios could be measured in different mouse strains that respond to alcohol differently. During *CYP2E1* mediated metabolism, oxygen recruits an electron from hydrogen and NADPH resulting in $NADP^+$. The redox pair ratio compares the levels of NADP and NADPH in the brain, thus providing an indication of how much ethanol has been metabolized by *CYP2E1*. Similar compounds can be measured for the alcohol dehydrogenase pathway allowing for the comparison of pathway activity. To fully understand the sequence variants that affect the activity of *CYP2E1,* functional variants need to be identified either by sequencing or further SNP genotyping upstream of *CYP2E1* to find regions that correlate with expression.

A single causal variant may not be driving the evidence for linkage at the end of chromosome 10. Since linkage analysis seeks to find regions of a chromosome that co-segregate within a family, the specific variants can vary between families as long as they fall within the same region. So there could be allelic heterogeneity among the families, where a small number of variants in the gene could be independently affecting the differences in the level of response to alcohol. It is not fully understood what effect this would have on the identification of a causal variant as performed in this study. It would be helpful to generate simulated datasets with multiple causal variants to understand how the measured genotypes affect the reduction of the LOD score when included as a covariate.

A number of additional variables were available for the study of the level of response to alcohol. During the alcohol challenge, blood alcohol levels and body sway were also measured. It would be possible to look at the interaction of blood alcohol level and the level of response to alcohol. A linkage based study is family based and as a result is immune to population stratification. Linkage analysis compares only the allele sharing and trait difference between family members, not across families. So even though the sample set was made up of different ethnicities, the genetic differences affecting the response to alcohol in the families were centralized to the region at the end of chromosome 10. It would have been beneficial to include ethnicity as a covariate in association analysis to see if correcting for different genetic backgrounds increased our evidence for association. The inclusion of families in the association analysis accounts for the within family variation and is likely protected from the effects of stratification. It would also have been interesting to consider whether the alcoholic parent, mother or father, had any effect on the level of response to alcohol.

The current dataset had phenotype data related to nicotine use, however this data was sparse since values were not reported for the entire sample. Even so, it was shown that correcting for the average number of cigarettes smoked per day was able to increase the evidence for linkage. Given the known interaction between alcoholism and nicotine use combined with the evidence of an enhanced effect during the metabolism of ethanol and nicotine by *CYP2E1*, it would be interesting to further investigate the effect of nicotine use on the level of response to alcohol and activity of *CYP2E1* in a set of individuals who drink and smoke excessively.

5.3.2 The association of the *MAPT* region with Pick's complex diseases

This association found between the inversions containing *MAPT* can be considered a true positive due to the well documented involvement of the gene with these diseases, providing some level of plausibility for other positive associations found in the genome-wide study. These other associations are not reported here because extensive data cleaning and replication is needed. Replication is a necessary follow-up for any association study to minimize errors due to false positives and cryptic biases in a data set. There were some compelling possibilities for positive association, with many that reach genome-wide significance, in genes involved with microtubules, endosomal trafficking, and angiogenesis. But they will not be discussed here in detail.

It is likely that the H1 haplotype is harboring mutations in the *MAPT* gene leading to disease. Imputation was used to predict the genotypes of markers not included in the study to provide a more complete picture of the common variation in the dataset. However, due to the high degree of linkage disequilibrium across the inversion, imputation provided no additional information about the association of this. Sequencing is required to understand the rare variants that can distinguish different sub-haplotypes of H1 and H2. The inversion is only present in the Caucasian population, so as an alternative, the study could be performed in a different population to provide a clearer picture of the region and possible causal variants although ascertainment of enough samples in an alternate population could be difficult due to the rarity of the disease.

There are several known isoforms of tau and these isoforms are present in different ratios in different subtypes of Pick’s disease. The main two types of isoforms depend on the alternative splicing of exon 10 resulting in three or four repeat binding domains. Tau protein aggregates in PSP and CBD are predominantly made up of four repeat domains. It would be informative to investigate possible splice variants in the gene leading to the overabundance of one type of isoform.

5.3.3 The evaluation of a novel algorithm for haplotype phase inference

Analysis for this project was performed on a sequence of 1 Mb. Compared to a standard full chromosome that would be provided by GWAS data, this is rather small. To test data of that scale, it would be useful to apply the same performance criteria to the 10 Mb dataset provided through Fregene. Parameters for computational job parsing would need to be optimized for the successful completion of analysis for the larger scale dataset, but it would be expected that CHAT would be able to predict the haplotype phase of a 10Mb dataset with better accuracy and coverage.

CHAT was not created as a haplotype phasing algorithm and that is not its main purpose. CHAT was created to identify long haplotypes shared by a subset of individuals with a certain trait in a population with the assumption that such a haplotype could harbor an ancestrally inherited causal mutation. To test the performance of CHAT on the identification of trait-causing variants, simulated data can be generated by Fregene and Sample with case/control status and causal variants.

5.4 The Future of the Genetics of Complex Human Disease

Future genome-wide association studies need to be well planned and designed with more power to detect variants with small effects. The most controllable way to increase power is through sample size. Current studies include thousands of individuals and there is a trend for collaborations between large centers to combine genotypes into an even larger multi-center pool of data. But care should be taken to ensure consistency between sites regarding both genotype and phenotype measurements. As described earlier, systematic bias in either type of measurement will lead to biased associations. There is also a trend of focusing on subtypes of disease or intermediate symptoms, often referred to as endophenotypes. This approach is helpful when there is complex heterogeneity in the disease definition and avoids the loss of information that results from the use of a disease endpoint as a phenotype. This also allows for the comparison of symptoms that overlap across different diseases that could have similar biological causes.

Many GWAS have identified variants that could be involved with diseases and some of these have been replicated. When examining the association results from a variety of diseases, up to 80% of associated SNPs lie in intergenic or intronic regions with no obvious effect on expression.[9,13] It is important now to determine the exact causal variants and understand how these causal variants create changes that lead to disease. For the majority of identified variants, the effect that these variants have on expression leading to disease has not been obvious and detection of functional variants for most identified associations has been difficult.[5,9] It is likely that the variants are tagging unknown causal variants which may be nearby or at some distance. While linkage disequilibrium does decline with distance, it is complex and can span hundreds of kilobases.[8] One SNP may tag a large number of markers that are interspersed among other, untagged SNPs.[8] So extensive genotyping in a large region around associated markers is needed to identify probable causal SNPs. The best option for considering all variants in a region would be deep sequencing.

As mentioned briefly in the introduction, the field of the genetics of complex disease has been moving towards whole genome or whole exome sequencing to identify specific variants that increase risk for a disease or trait. With next-generation sequencing technology, whole genome sequencing has become both time and cost effective. Whole genome sequencing addresses the problem of uncaptured rare variants seen in association studies. It also bypasses the reliance on linkage disequilibrium between the genotyped variant and causal variant. But the creation of massive genome-wide sequence level data presents a wide range of challenges regarding bioinformatics, data mining, and data management. Just because the data can be created, does not mean we know what to do with it. New applications must be created to efficiently sift through the data, identify variants, and make calculations. Computational and statistical approaches that were appropriate for GWAS will likely not scale well for whole genome applications and more powerful computing resources will be needed. Data management and storage will be a particular challenge. There need to be ways to not only store the sequence data and compile results from many overlapping reads, but to retrieve that data and present it in a usable way.

The immediate goal of the study of the genetics of complex human disease is to find genetic variants that increase risk for disease. The long term goal is to apply this information to help treat or predict disease. Genes containing risk variants can be targets for drug treatment which can alter expression to either up-regulate or down-regulate a gene as needed. Variants can be used to screen for individuals that are at risk for disease so that those individuals can modify their behavior or seek preventative help. Genetic variants can be used to choose the best treatment for a disease—which drug or what dosage to minimize side effects. But to be good predictors, genetic variants need to be easily measured, highly accurate, cost effective, and have an effect size large enough to actually confer enough risk to matter. A genetic variant with an odds ratio of 1.1 will not increase an individual's risk for disease much compared to the general public risk level. A variant would also need to be relatively common in the general population, or it would be rarely seen and not practical for clinical use. A better option may be to use multiple genetic variants together as a biomarker to predict whether an individual will develop that disease based on their set of combined risk alleles.

5.5 Final Thoughts

The purpose of this dissertation was to apply various statistical and computational techniques to aid in the understanding of the genetics of complex human disease and the assessment of genome-wide datasets. The presented projects demonstrate that even though regions of the genome that increase risk for a trait can be readily discovered the identification of the specific variants that directly influence a trait has proven difficult. As demonstrated throughout the dissertation, to be effective in the identification of risk variants, future studies need to increase the power to detect an effect, reduce heterogeneity caused by differences in populations or genotypes, minimize the chance of false positives, and increase the amount of information provided by genetic polymorphisms by using dense maps and combinations of markers as haplotypes.

5.6 References

1: BRLMM:an Improved Genotype Calling Method for the GeneChip Human Mapping 500K Array Set. White Paper from Affymetrix.com. Apr 2006

2: Cho JH. Genome-wide association studies: present status and future directions. Gastroenterology. 2010 May;138(5):1668-1672.e1. Epub 2010 Mar 16. Review.

3: Psychiatric GWAS Consortium Coordinating Committee, Cichon S, Craddock N, Daly M, Faraone SV, Gejman PV, Kelsoe J, Lehner T, Levinson DF, Moran A, Sklar P, Sullivan PF. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. Am J Psychiatry. 2009 May;166(5):540-56. Epub 2009 Apr 1. Review.

4: Cordell HJ, Clayton DG. Genetic association studies. Lancet. 2005 Sep 24-30;366(9491):1121-31.

5: Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet. 2009 Apr;10(4):241-51. Review.

6: International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789-96.

7: Lander ES. The new genomics: global views of biology. Science. 1996 Oct 25;274(5287):536-9.

8: Lawrence R, Evans DM, Morris AP, Ke X, Hunt S, Paolucci M, Ragoussis J, Deloukas P, Bentley D, Cardon LR. Genetically indistinguishable SNPs and their influence on inferring the location of disease-associated variants. Genome Res. 2005 Nov;15(11):1503-10.

9: Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010 Jul 8;363(2):166-76. Review.

10: Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747-53. Review.

11: Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA. 2008 Mar 19;299(11):1335-44. Erratum in: JAMA. 2008 May 14;299(18):2150.

12: Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev. 2009 Jun;19(3):212-9. Epub 2009 May 28. Review.

13: Xiong M, Guo SW. Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet. 1997 Jun;60(6):1513-31.