1-1-2011

New methods for discovering common and rare genetic variants in human disease

Peng Lin
Washington University in St. Louis

Recommended Citation

Lin, Peng, "New methods for discovering common and rare genetic variants in human disease" (2011). *All Theses and Dissertations (ETDs)*. 606.
https://openscholarship.wustl.edu/etd/606

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Human and Statistical Genetics

Dissertation Examination Committee
John Rice, Chair
Anne Bowcock, Co-chair
Arpana Agrawal
Charles Gu
Christina Gurnett
Nancy Saccone

NEW METHODS FOR DISCOVERING COMMON AND RARE GENETIC VARIANTS IN HUMAN DISEASE

By

Peng Lin

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

August 2011

Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

NEW METHODS FOR DISCOVERING COMMON AND RARE GENETIC VARIANTS IN HUMAN DISEASE

By
Peng Lin

Doctor of Philosophy in Biology and Biomedical Sciences
Human and Statistical Genetics
Washington University in St. Louis, 2011

Professor John Rice, Chair
Professor Anne Bowcock, Co-chair

Since the discovery of Mendel’s laws, one of the most challenging problems in genetic research has been to locate and characterize genetic variants that cause human disease. Although thousands of disease-associated genetic variants have been discovered, many remain unknown. New methods are needed to facilitate the discovery process. Here, we present new methodology to improve detection of these genetic variants for genotyping imputation, Copy Number Variations (CNV) and sequencing data.

Currently, imputation is widely used to evaluate the evidence for association at genetic markers that are not directly-genotyped. However, imputation can be problematic especially when a genetic variant has low minor allele frequency. We present a new statistic, the imputation quality score, developed to better differentiate well-imputed and poorly-imputed SNPs. It is particularly useful for SNPs with low minor allele frequency and datasets that are genotyped on different platforms.

CNV calling, on the other hand, is not reliable. We developed a statistical method for estimating sensitivity and positive predictive rate, and evaluated the relative performance of CNV calling on a genome wide scale. We found that the positive predictive rate increases with the number of probes and the size of CNVs. We also noticed that CNVs reported by multiple programs have a higher reproducibility rate and positive predictive rate. This method was applied to the dataset from the Study of Addiction: Genetics and Environment. Our analysis revealed that CNVs in 6q14.1 (P= 1.04 x10^{-6}) and 5q13.2 (P= 3.37 x10 $^{-4}$) are significantly associated with alcohol dependence after adjusting for multiple tests. Evidence also suggested that CNVs at 5q13.2 increase the risk for alcohol dependence by lowering conscientiousness, or more specifically, self-discipline.

As genetics is looking towards the future with sequencing data, improved methods are needed for rare variants. By taking advantage of the simulation data from the Genetic Analysis Workshop, we integrated both the collapsing method and the family data method in an attempt to increase power for rare variants. We concluded that this combinational approach offers a substantial power boost for certain causal genes, and is therefore worth further investigation.

By focusing on improving the interpretation of data from imputation, CNV calling and sequencing, our work parallels the development of genetic research over the past few years, provides a direction for on-going methods development, and will be useful for future research endeavors.

Acknowledgements

This work was supported in part by funding for the Study of Addiction: Genetics and Environment (SAGE). SAGE was funded through the NIH Genes, Environment and Health Initiative [GEI] (U01 HG004422). Support for collection of datasets and samples was provided by U10 AA008401, P01 CA089392, R01 DA013423 and R01 DA019963. This research is also supported by U.S. Public Health Service (USPHS) grants U24 MH68457, HHSN271200477471C and NIH grant R01 GM031575.

I would like to express my sincere gratitude to the following individuals for their mentorship: Dr. John Rice for acting as my thesis mentor; Dr. Anne Bowcock for acting as my thesis co-mentor. During my PhD training, they continually gave me support, guidance, advice and inspiration. In every sense, this work would not have been possible without them. I also would like to thank Drs. John Rice and Anne Bowcock for their encouragement at some of the most important moments of my life. Words cannot express my heartfelt gratitude and appreciation to them.

I also would like to thank Drs. Nancy Saccone, Christina Gurnett, Charles Gu and Arpana Agrawal for serving on my thesis committee. They offered me invaluable advice and contributed important insights for this work. I would like to acknowledge the instruction and guidance from Drs. Laura Bierut, Sarah Hartz, Alison Goate, Mike Lovett, Scott Saccone and Rosalind Neuman. They offered me assistance in many different ways and facilitated my scientific training.

I am also thankful for interactions with my colleagues. I would like to thank Jen-Chyong Wang from the Goate lab for her assistance with experiments related to this work. I thank the present and former members of the Rice lab, the Bowcock lab and the Saccone lab for their friendship and collaboration, with particular gratitude to Nan Rochberg, Sharon Ryan, Bill Howells, Fengxian Wang, Zhehao Zhang and Michael Hamm.

I would also like to acknowledge many members of the Human and Statistical Genetics program, with particular gratitude to Kristina Hedtkamp for helping me in many practical aspects of my training; and my fellow classmates for their support and friendship.

Finally, I would like to dedicate this thesis to my parents and my wife Mi Cai for their love.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

A BRIEF REVIEW OF METHODS IN STATISTICAL GENETICS RESEARCH

The hereditary nature of different phenotypes lies secretly in the genome. The goal of genetic research is to identify these secret genetic components that affect these different phenotypes. Since Mendel first discovered the law of inheritance, the rapid development of technology has revolutionized technologies we can use to study genetics, and consequently the statistical methods to process the information generated by these new technologies.

One of the most important research areas in human genetics in the last century has been the effort to link various traits and diseases to a relatively-large region in the human genome. This would not have been possible without the discovery of Restriction Fragment Length Polymorphisms (RFLPs), Variable Number Tandem Repeat (VNTR) and Microsatellite (or simple sequence repeats) [1, 2]. Due to limited number of these genetic markers and also partly due to technical challenges, these genetic markers can only cover an extremely small fraction of total variants in the entire genome. Fortunately, it was known at that time that the recombination frequency between the genetic marker and a particular trait is correlated with the physical distance between the known genetic marker and the underlying gene for the particular trait. Linkage thus maps the position of underlying genes relative to known genetic markers in terms of recombination frequency. The LOD score (logarithm of odds), developed by Newton E. Morton, is a statistical test often used for linkage analysis in human. The LOD score compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data by chance. By convention, a LOD score greater than 3.0 is considered strong evidence for linkage, whereas smaller LOD scores indicate that linkage is less likely. Linkage has proven to be successful, numerous trait and disease genes loci have been found, and several new methods based on linkage have been developed [3].

More recently, the development of DNA microarray has made genome wide association studies possible [4]. Rather than <10,000 genetic markers used in linkage studies, these studies typically genotype 100,000–1,000,000 variants in each of the individuals being studied. As of today, some commercial DNA microarrays can genotype more than 2,500,000 Single Nucleotide Polymorphisms (SNPs). The total number of SNPs that can be genotyped in a DNA microarray is still far less than the total number of genetic variants in the human genome, but SNPs can tag surrounding blocks of ancient DNA (haplotypes). This property is often described in the term of linkage disequilibrium (LD). This huge number of SNPs provides a relatively sufficient coverage for the entire human genome, which underlies the rational of genome wide association studies. In 2005, the first widely replicable genome wide association study reported association between the complement factor H (CFH) gene and age-related macular degeneration (AMD) [5]. This discovery was the earliest of its kind in part because variation at CFH has a large effect—greater than fourfold—on AMD risk. In 2007, the Wellcome Trust Case Control Consortium carried out genome-wide association studies for the diseases of coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder, and hypertension[4]. This study was one of the largest studies at that time, including more than 14,000 cases of seven common diseases and 3,000 shared controls. To date, more than 1200 genome wide association studies have been conducted, over 200 diseases and traits have been examined, and about 4000 SNP associations have been found [6].

However, it has been estimated that more than 10 million common SNPs are likely to exist in the human genome [7]. Therefore, a typical genome wide association study can only examine a very small fraction of these common SNPs, and millions of known common variants have to be ignored. As described previously, SNPs can tag surrounding genetic variants; therefore, it is also possible to infer genotypes at unobserved SNPs based on multiple surrounding genetic variants in a procedure commonly known as imputation, leading to improved power. Imputation typically needs a reference panel that includes a large number of genetic variants. Study samples genotyped are compared to this reference panel and shared haplotype are then guessed and assigned with different probabilities. Missing genotypes for each study sample can be predicted by matching reference haplotypes [8]. A recent type 2 diabetes study showed that the accuracy of imputation is high. In this study, the investigators compared imputed genotypes generated *in silico* with experimental genotypes generated in the lab. Their results showed excellent concordance between genotype calls, with an overall concordance rate of >98% between genotyped and imputed SNPs [9]. Since there is almost no cost associated with imputation, and the benefit to have more *in silico* genotyped SNPs is obvious, imputation has become very popular and paves the way for meta-analysis that combines dozens of different studies.

Most variants found in genome wide association studies so far confer relatively small increments in risk (1~1.5 fold), and explain only a small proportion of heritability—the portion of phenotypic variance in a population attributable to additive genetic factors [10]. For example, the estimated heritability for human height is over 80%, so far at least 40 loci have been found, yet they together explain only about 5% of phenotypic variance [11]. Many investigators question how the remaining, 'missing' heritability can be explained. Several potential answers are offered. Some investigators suggest that copy number variation (CNV) may contribute a substantial amount to the heritability, while some others propose that the missing heritability may lie in rare variants, which cannot be well captured by current genome wide association studies.

Copy Number Variations usually refer to duplications or deletions of a particular segment in the human genome. Evidence has shown that CNVs can change the expression level of genes in or near these CNV regions[12]. In the past, CNVs are usually discovered by hypothesis-driven lab experiments at a limited number of target regions. The advent of genome-wide association studies (GWAS) has led to the possibility of discovering CNVs across the genome. Many CNV detection programs have been developed for this purpose, including CNVpartition, PennCNV, and QuantiSNP. Many studies have identified CNVs that may be associated with diseases [13-19]. Now it is not uncommon practice that many research groups will search CNVs after completing genome wide association studies.

Different from the common disease common variant hypothesis, the rare variant hypothesis proposes that a significant proportion of total variation may be due to the effects of a large number of low frequency genetic variants [20]. Some rare variants may be individually rare, but cumulatively large in number in the population, while others may be extremely private, and only exist in certain families. These rare variants cannot be tagged or imputed by surrounding common SNPs that are genotyped in genome wide association studies. The only way to find these rare variants is to do high-throughput sequencing, which is not economically feasible until just one or two years ago.

IMPUTATION ACCURACY

In statistics, agreement statistic describes the degree of agreement among raters. A large number of statistics are designed to assess agreement. Different statistics are often appropriate under certain conditions. One of the most commonly used statistics is concordance rate, which is the percentage of agreed paired measures among all paired measures (Table 1.1).

$$Concordance\ rate = \frac{The\ number\ of\ agreed\ pairs}{The\ total\ number\ of\ pairs}$$

Concordance rate is simple and straight forward, and gives a quick answer to evaluate agreement. Some other statistics are more applicable in a more complicated situation, for example joint-probability of agreement, Cohen's kappa and the related Fleiss' kappa, inter-rater correlation, concordance correlation coefficient and intra-class correlation [21]. Many methods for genetic research, including linkage analysis, genome wide association studies and imputation, depend heavily on the agreement statistic, because it addresses an important issue − the data quality. For example, in linkage analysis, genotyping errors are spotted by comparing genotypes of children and parents [22]. In genome wide association studies, duplicate samples are often included in a study design to assess SNP genotyping accuracy [23]. For imputation, the imputed genotypes are often compared to experimental genotypes generated in the lab [9].

More specifically, the commonly used concordance rate in imputation is defined by the proportion of correctly classified genotypes, or equivalently by the discrepancy rate between imputed and observed genotypes [24]. Evidence has shown that imputation has a high concordance rate and provides a cost efficient way to improve power [9]. But as the current research trend shifts toward genetic variants with low allele frequency, the commonly used concordance rate is not sufficient to evaluate imputation reliability, and can be misleading and even wrong in some scenarios. Table 1.1 shows that the occurrence of a positive reading ("+") is 1 in 10. The positive readings in Test 1 and Test 2 do not match with each other. However, according to the formula, the concordance rate is 0.8. This concordance rate is largely due to chance agreement of negatives.

This phenomenon becomes more serious when the frequency of positive readings continues to decrease. Because of chance agreement, an uncommon SNP has a higher apparent concordance rate than a common SNP. By randomly assigning the two alleles of a rare SNP, using only the MAF (<5%), an apparent accuracy greater than 90% can be reached if it is measured by the concordance rate [25, 26]. A related issue in statistical genetics is the difficulty of combining datasets that are genotyped on different platforms. Different platforms have different sets of SNPs and imputation is necessary to predict SNPs that are not genotyped in one platform or the other. If the concordance rate is used for quality control, the following analysis will report an enormous number of false positives [26].

THE RELIABILITY OF CNV DETECTION USING SNP GENOTYPING ARRAY

Copy Number Variations are a prevalent form of genetic variation. They are known as duplications or deletions of a particular segment of DNA. The comprehensive identification and validation of CNVs would greatly benefit the genetic research of human disease. Previous experimental studies for CNV detection were mainly performed by microarray comparative genomic hybridization (array- CGH) [27]. However, array-CGH has limited sensitivity and poor resolution and is highly subject to the variation of existing CNVs among unrelated samples that are used as a reference. The advancement of SNP microarray makes it more suitable for genome wide CNV detection. A typical DNA microarray provides total fluorescent intensity signals at each probe and the relative ratio of the fluorescent signals between two alleles at each SNP probe. These signals are often referred as the “log R Ratio” and “B Allele Frequency” respectively [28]. CNVs thus can be readily discovered by examining the pattern change of log R Ratio and B Allele Frequency.

Many methods are available to identify CNVs from SNP microarray. Conventional methods discover CNVs by averaging log R Ratio of probes in a sliding window, while more sophisticated methods involve the hidden Markov model to assign different CNV modes across the genome. Similar to rare SNPs, the majority of CNVs are not common in the population. Therefore, CNVs called by these methods typically have an apparent high concordance rate whereas other more appropriate measures indicate less confidence in CNV calling. In our study, we have compared CNVs from duplicate samples. Despite a concordance rate as high as > 98%, only about half of CNVs can be reproduced in the duplicate samples [29].

Due to poor reliability, CNV calls require further experimental validation. Quantitative PCR is the most commonly used technique to validate CNVs. However, it is not economically feasible − at least at this time − to validate all CNV calls across the genome by experiments. Without experimental validation, these CNV calls have no experiment validated results to be compared with and thus it is difficult to evaluate calling reliability. Analysis based on these results may cause many false positives. This thesis aims to find part of the solutions for this challenge.

THE CHALLENGE OF DISCOVERING RARE VARIANTS

Genome-wide association studies have identified many common genetic variants that contribute to complex diseases, but together they can only explain a small fraction of total variation. Many researchers believe that a large number of genetic variants contributing to disease susceptibility are yet to be discovered [30], and these genetic variants are very likely to be genetic variants with low frequency [20].

Association analyses involving rare variants are not as easy as analyses involving common variants. Power analysis has shown that in a standard case control study design, the sample size needed to detect an association with a single rare variant dwarfs the sample size of any current genome wide association studies (Figure 1.1).

Several methods have been proposed to overcome the power issue. One strategy involves collapsing sets of rare variants into a single group, and then compares their collective frequency between cases and controls [31]. Another strategy takes the family-based approach, because in theory, individuals sharing the same rare variant can be more easily recruited in the family-based study design. Currently, many investigators have ongoing research in this area and there is no consensus among the research community on what is the best approach.

CONCLUSION

The field of statistical genetics has evolved rapidly in the past few years. New technologies are introduced, and accordingly new methods are developed. The traditional methods have proved to be extremely successful in genetic research, but they are not sufficient to address challenging issues involved with the latest discoveries.

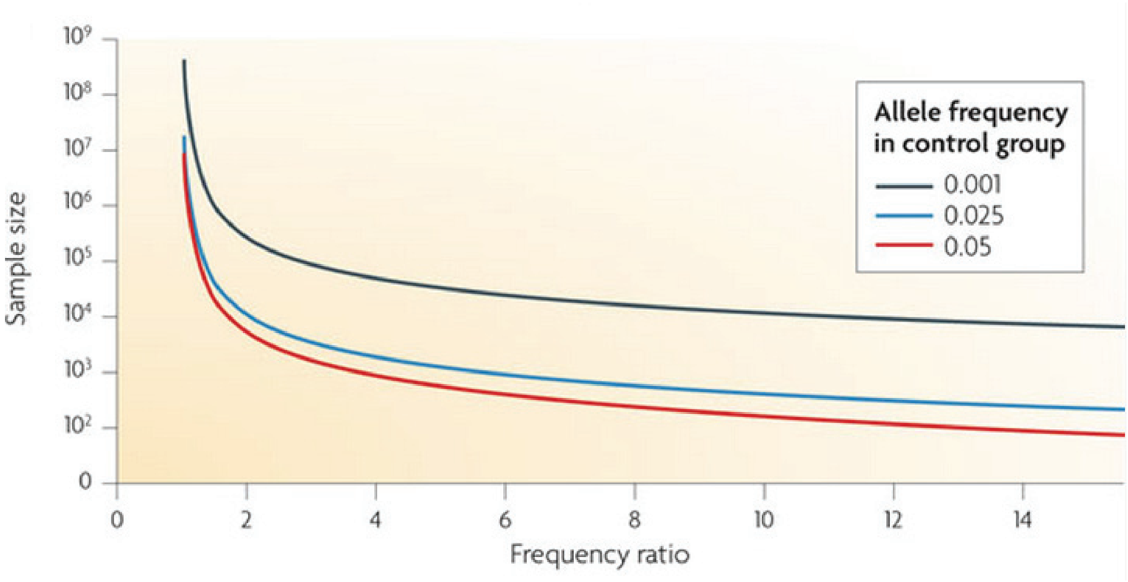
This work will mainly focus on imputation and CNV studies, partly due to the fact that imputation and CNV studies have a higher error rate and therefore require more sophisticated methods. It is also partly because most part of this research was done at the time frame when imputation and CNV studies were in a rapid development period. The development of new methods for imputation and CNV studies will be discussed in Chapter 2 and Chapter 3 respectively. In Chapter 4, this work will cover the application of these new methods in real datasets to facilitate the searching process for genetic variants that are associated with alcohol dependence and further link alcohol dependence with the “Big Five” factors in personality. Both of alcohol dependence and personality have been proved to have a strong genetic component. One goal of this work is also to cast light on the solutions for future genetic research. As whole genome sequencing will become economically feasible in the years ahead, this work also covers a new method for rare variants in the coming age of “common disease, rare variant.”

Table 1.1 - Concordance rate

	1	2	3	4	5	6	7	8	9	10
Test 1	+	-	-	-	-	-	-	-	-	-
Test 2	-	+	-	-	-	-	-	-	-	-

Concordance rate is defined as the percentage of agreed paired measures among all paired measures. The concordance rate between Test 1 and Test 2 is 0.8.

Figure 1.1 - The sample size required to detect rare variants



Adapted by permission from Macmillan Publishers Ltd: Nature reviews [30], copyright 2010

The x-axis indicates the ratio of the frequency of the allele in the case versus control groups. The y-axis shows the sample size required to detect rare variants when type I error rate is set to 10^{-9}. These lines with different colors indicate the allele frequency in control group.

Chapter 2: A New Statistic to Evaluate Imputation Reliability

*This chapter is adapted from: Lin P et al. (2010) A New Statistic to Evaluate Imputation Reliability. PLoS ONE 5(3): e9697. doi:10.1371/journal.pone.0009697.

ABSTRACT

As the amount of data from genome wide association studies grows dramatically, many interesting scientific questions require imputation to combine or expand datasets. However, there are two situations for which imputation has been problematic: (1) polymorphisms with low minor allele frequency (MAF), and (2) datasets where subjects are genotyped on different platforms. Traditional measures of imputation cannot effectively address these problems.

We introduce a new statistic, the imputation quality score (IQS). In order to differentiate between well-imputed and poorly-imputed single nucleotide polymorphisms (SNPs), IQS adjusts the concordance between imputed and genotyped SNPs for chance. We first evaluated IQS in relation to minor allele frequency. Using a sample of subjects genotyped on the Illumina 1M array, we extracted those SNPs that were also on the Illumina 550K array and imputed them to the full set of the 1M SNPs. As expected, the average IQS value drops dramatically with a decrease in minor allele frequency, indicating that IQS appropriately adjusts for minor allele frequency. We then evaluated whether IQS can filter poorly-imputed SNPs in situations where cases and controls are genotyped on different platforms. Randomly dividing the data into "cases" and "controls", we extracted the Illumina 550K SNPs from the cases and imputed the remaining Illumina 1M SNPs. The initial Q-Q plot for the test of association between cases and controls was grossly distorted (λ=1.15) and had 4016 false positives, reflecting imputation error. After filtering out SNPs with IQS < 0.9, the Q-Q plot was acceptable and there were no longer false positives. We then evaluated the robustness of IQS computed independently on the two

halves of the data. In both European Americans and African Americans the correlation was > 0.99 demonstrating that a database of IQS values from common imputations could be used as an effective filter to combine data genotyped on different platforms.

IQS effectively differentiates well-imputed and poorly-imputed SNPs. It is particularly useful for SNPs with low minor allele frequency and when datasets are genotyped on different platforms.

INTRODUCTION

Genome-wide association studies (GWAS) represent a powerful approach to the identification of genetic variants involved in common human diseases[32]. GWAS use commercial SNP microarrays to genotype large numbers of genetic markers. However, SNP microarrays currently can only genotype up to one million of the 9–10 million common SNPs in the assembled human genome [7]. In addition, for a typical case-control design, several thousand cases and several thousand controls may be needed for adequate power to detect associations[33]. With little cost, imputation can boost power both by increasing SNP coverage and by combining samples from similar studies. Based on haplotypes from the International HapMap project[14], imputation infers untyped variants from known genotypes. The inference uses one of several model-based methods, and the resulting imputed SNPs can be tested for association with a phenotype [34]. The power of this method has been demonstrated in the literature where several groups have found novel causal genes [35-38].

There are two situations where researchers avoid imputation due to increased error in imputation: (1) SNPs with minor allele frequency less than 1% [4, 35, 39], and (2) association studies where cases and controls are genotyped on different platforms. Imputation accuracy, calculated for each SNP as the proportion of genotypes correctly classified, is the gold standard for evaluating the quality of imputation. Unfortunately, it is an inadequate filter in both of these circumstances. For the majority of SNPs, imputation programs such as IMPUTE [34], MACH[39], and BEAGLE[40], have very high imputation accuracy [34, 40-42]. However, the use of imputation accuracy in low frequency SNPs to evaluate imputation quality can be misleading. When the minor allele frequency of a SNP is less than 5%, a program could randomly assign the two alleles to the sample only using the minor allele frequency and achieve more than 90% accuracy. Although SNPs with low minor allele frequencies (MAF<5%) are referred to as uncommon SNPs, they represent more than 30% of SNPs in the HapMap Phase II CEU population, and this proportion is even higher in African populations[7]. This problem assessing imputation accuracy in lower frequency SNPs means that a large part of the genome will not be adequately interrogated using imputation.

The second problematic situation for imputation is where cases and controls are genotyped on different platforms. This is problematic because imputation error can vary between cases and controls, causing increased rates of false positives in association studies. There is no known method for effectively filtering the poorly imputed SNPs from the well imputed SNPs on different platforms. Although this situation has been avoided by researchers, it is an important application. Large studies such as Wellcome Trust and the NIMH GAIN samples use common controls that could be used in other studies to gain power [32, 43]. But, if the primary datasets were genotyped on a different platform, imputation is necessary.

In order to assess the reliability of imputation with an emphasis on the less common SNPs and an interest in evaluating data imputed from different platforms, we introduce a new statistic, the imputation quality score (IQS). Partly motivated by Cohen's statistic *Kappa* to quantify rater agreement[44], IQS takes chance agreement into account and thus controls for allele frequencies. In this paper, we introduce IQS, demonstrate its value in situations of low minor allele frequencies, and demonstrate how it can be used to improve the type I error rate when cases and controls are genotyped on different platforms.

MATERIALS AND METHODS

Ethics statement

De-identified data from the Study of Addiction: Genetics and Environment (SAGE) were analyzed for the research reported in this manuscript. SAGE consists of existing data from three genetic studies of addiction: the Collaborative Study on the Genetics of Alcoholism (COGA), the Collaborative Genetic Study of Nicotine Dependence (COGEND), and the Family Study of Cocaine Dependence (FSCD). All participants in COGA, COGEND and FSCD provided written informed consent for genetic studies and agreed to share their DNA and phenotypic information for research purposes. The institutional review boards at all data collection sites granted approval for data collected from COGA, COGEND and FSCD to be used for the Study of Addiction: Genetics and Environment. Specifically, approval was obtained from the Washington University Human Research Protection Office (for COGA, COGEND and FSCD), the State University of New York Downstate Medical Center Institutional Review Board (COGA), the University of Connecticut Health Center Human Subjects Protection Office (COGA), the Indiana University Research Compliance Administration (COGA), the University of California, San Diego Human Research Protections Program (COGA), the Howard University Institutional Review Board (COGA), The University of Iowa Human Subjects Office (COGA), and the Henry Ford Health System Institutional Review Board (COGEND). The second dataset was obtained from the National Institute of Mental Health Center for Collaborative Genetic Studies on Mental Disorders (http://www.nimhgenetics.org/) and was also de-identified.

Methods

The computation of IQS requires the posterior probabilities of AA, AB and BB as output by the imputation program. For one SNP genotyped on N individuals, the probabilities can be readily tabulated into a 3 × 3 table where each cell, n_{ij}, represents the number of individuals with true genotype *i* and imputed genotype *j* (Table 2.1). Note, in this scenario, n_{ij} may not be an integer due to imputation probabilities being reported rather than imputed genotypes.

We define the observed proportion of agreement (P_o) as:

$$P_o = \frac{\sum_i n_{ii}}{n_{..}}$$

The observed proportion of agreement can be used to evaluate imputation reliability. But, like imputation accuracy and average maximum posterior probability, it can overestimate reliability for uncommon SNPs because it is not adjusted for "chance" agreement.

IQS adjusts for allele frequency by subtracting "chance" agreement from the "observed" agreement. Similar to P_o, "chance" agreement (P_c) is computed as the sum of the products of marginal frequencies that would occur if genotypes are called at random using the same marginal rates:

$$P_c = \frac{\sum_i n_{i.} n_{.i}}{n_{..}^2}$$

IQS is then computed by subtracting the chance agreement from the observed agreement and dividing by the maximum possible value of the numerator. The value of one indicates a perfect match, and negative values indicate that the imputation program performed worse than chance.

$$IQS = \frac{P_o - P_c}{1 - P_c}$$

In addition, the calculation of IQS can be expanded to evaluate non-random error. When cases and controls are genotyped on different platforms (e.g., cases genotyped on the Affymetrix array and controls genotyped on the Illumina array), some SNPs are not genotyped in either array but are imputed from their respective arrays. This imposes non-random errors on the imputed genotypes. In particular, if we combine these imputed genotypes together, it will inflate false positive rates. IQS can take this into account by incorporating marginal frequencies into the calculation. For instance, if imputation from the Illumina array reports that for a particular SNP, the probabilities of AA, AB and BB are a_1, a_2, a_3, and imputation from the Affymetrix array reports that the probabilities for the three genotypes are b_1, b_2, b_3, then n_{ij} in the calculation of P_o becomes

$$n_{ij} = a_i b_j$$

In this scenario, IQS provides a useful criterion to exclude unacceptable SNPs imputed from different sources.

Data and imputation

The first dataset was collected as part of SAGE, one study in the Gene Environment Association (GENEVA) project (http://genevastudy.org/). Samples were genotyped on the Illumina Human 1M array at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University. The Illumina 1M array has a total of 1,049,008 probes as SNP assays. All SNPs with a genotype call rate <98% were removed, as were SNPs with a Hardy-Weinberg exact p value $<1x10^{-4}$. Additional data cleaning procedures were applied to ensure the highest possible data quality, including using HapMap controls, detection of gender and chromosomal anomalies, hidden relatedness, population structure, batch effects, Mendelian error detection, and duplication error detection[23]. The composition of the remaining project samples in terms of self-identified ethnicity is 2597 European Americans and 1264 African Americans, confirmed by principal component analysis. Among the 1,049,008 SNPs, 948,658 SNPs (90%) passed data cleaning procedures.

The second dataset consists of controls from the National Institute of Mental Health Center for Collaborative Genetic Studies on Mental Disorders (http://www.nimhgenetics.org/). A total of 418 subjects (controls) were genotyped using both the Affymetrix GeneChip Mapping 500K Array Set and the Illumina HumanHap 550K Array set and passed all cleaning procedures. All individuals in this study were European Americans with no evidence of heterogeneity, verified by principal component analysis[45]. All SNPs with a genotype call rate <95% were removed, as well as SNPs with a Hardy-Weinberg exact p value $<1x10^{-5}$. After quality control, 447,250 autosomal

SNPs were retained from the Affymetrix 500K array, and 527,095 autosomal SNPs were retained from the Illumina 550K array.

Imputation from each array to Hapmap SNPs was done by the program IMPUTE (https://mathgen.stats.ox.ac.uk/impute/impute_v0.5.html) [34]. European Americans were imputed using the CEU reference panel (HapMap release 22 - NCBI Build 36 dbSNP b126). African Americans were imputed separately using the YRI reference panel (HapMap release 22 - NCBI Build 36 dbSNP b126). We omitted sex chromosomes in this study because of the complication of imputation on these chromosomes. The Illumina 1M array contains a small number of strand-ambiguous A/T C/G SNPs. Although Illumina provides strand information about those SNPs, we still found a few inconsistencies compared with the reference panel. In order to make sure that all SNPs were reported on the same strand, all strand-ambiguous A/T and C/G SNPs (5583 in total, 0.5% of all Illumina 1M SNPs) were excluded from the comparison.

Imputation efficiency is calculated as the proportion of genotypes that had a maximum posterior probability greater than 0.9, as recommended by IMPUTE.

Association tests were done by the program SNPTEST with the “-proper” option[34]. With this option, SNPTEST runs a logistic regression based on the probability of genotype rather than dichotomous genotype, allowing the uncertainty of the imputation to be factored into the consideration [46].

Statistical estimates of imputation quality

Both IQS and imputation accuracy compare true genotypes to imputed genotypes. Given that imputation is designed to infer unknown genotypes, one purpose of this paper was to use IQS to evaluate statistics that measure the quality of imputation *without* knowing the true genotype. The two statistics most commonly used for this purpose are the variance ratio (rsq_hat in MACH)[39] and the imputed information score (PROPER_INFO in SNPTEST) [34]. The variance ratio for a particular SNP is a ratio of the empirically observed variance (based on the imputation) to the expected binomial variance p(1-p), where p is the minor allele frequency[46]. As the amount of information available to impute the SNP decreases, the empirically observed variance decreases and the variance ratio approaches zero. The product of the variance ratio and sample size defines the 'effective sample size'. Similarly, the imputed information score is a measure of genotype information content, which is related to the effective sample size (power) for the genetic effect being estimated [32, 34, 46]. Although computed using a different approach, the information score is analogous to the variance ratio. For example, a SNP with an imputed information score of 0.75 indicates that the imputed SNP genotypes are equivalent to a dataset with 75% of the full sample size with precisely known genotypes.

RESULTS

The Illumina 1M array covers all of the SNPs on the Illumina 550K array. We started with all SAGE subjects genotyped on the Illumina 1M array and extracted the 545,966 SNPs that are present on the Illumina 550K SNPs. We used these Illumina 550K SNPs to impute to the full Illumina 1M array. We imputed 262,864 autosomal SNPs in 2597 European Americans (EA), and 304,425 autosomal SNPs in 1264 African Americans (AA). We compared imputed SNPs to the genotyping results from the Illumina 1M array. The remaining SNPs could not be evaluated due to the absence of those SNPs in either the Illumina 1M array or reference panel.

The imputation results are given in Table 2.2. The mean IQS is lower than the mean accuracy in both EA and AA. There are cases where IQS is negative, indicating that imputation did worse than chance in assigning genotypes. In this situation, 95% of the minor allele frequencies lie between 0 and 0.058, 95% of the chance agreement rates lie between 0.78 and 1, and the imputation accuracy is below chance agreement with 95% of the values between 0.81 and 1. These are strong examples of how imputation accuracy can be misleading when “chance” contributes so strongly to the proportion of agreement.

A second notable result is that the quality of imputation in AA is markedly lower than in EA. This is seen in the decreased efficiency by nearly ten percentage points, and decrease in mean IQS by nearly twelve percentage points. This is likely due to two factors. First, African Americans have more diverse haplotypes and more uncommon alleles. Second, there is non-negligible difference between African Americans and the YRI reference panel, which was clearly reflected by Eigenstrat population structure analysis [45]. Interestingly, imputation accuracy is nearly the same for EA as for AA, again highlighting how imputation accuracy can overestimate the quality of imputation.

The relationship between IQS and imputation accuracy with respect to minor allele frequency is seen in Figure 2.1. Although imputation accuracy increases with decreased minor allele frequency, IQS drops dramatically with decreased minor allele frequency. Because it is known that low minor allele frequency decreases the quality of imputation, many studies drop SNPs with minor allele frequency less than 1%. According to this plot, this practice would still retain SNPs with an average IQS score of 88%, and would eliminate some well-imputed SNPs.

We then evaluated the effectiveness of IQS in the situation where cases and controls are genotyped on different platforms. We randomly divided the SAGE data into two subgroups labeled “cases” and “controls”. In “cases”, original genotypes were retained for SNPs on the Illumina 550K array; and then imputation was performed to obtain the full Illumina 1M array. In “controls”, original genotypes were retained for all SNPs on the Illumina 1M array. This process is equivalent to combining cases genotyped by the Illumina 550K array and controls genotyped by the Illumina 1M array.

We tested genetic association of all the 1M SNPs with the cases and controls. A Quantile-Quantile Plot (Q-Q plot) is shown in Figure 2.2. By comparing the distribution of observed P values against the theoretical model distribution of expected P values, Q-Q plots are used in genome wide association studies to assess the inflation of false positive rates [47]. In randomized data without type I error arising from population stratification or some other artifact, the Q-Q plot should be a 45 degree line. To ensure that our random division of the data did not result in population stratification, we constructed a Q-Q plot based on the true genotypes, which was normal as expected (λ=1.03) (Figure 2.2 A). However, the Q-Q plot of imputed SNPs compared to genotyped SNPs is greatly distorted (λ=1.15), suggesting that combining imputed SNPs with genotyped SNPs without other quality control is problematic (Figure 2.2 B). Therefore, the observed distortion was due to imputation error and the statistically skewed SNPs (Figure 2.2 B) are false positives. We then filtered the imputed data by removing all SNPs with IQS $\leq$ 0.9, retaining 76% of the imputed SNPs, and dramatically improving the Q-Q Plot (λ=1.04) (Figure 2.2 C). The Q-Q plot remained grossly distorted even when the filter was changed to an imputation accuracy of > 99%, retaining 72% of the SNPs, although λ improved to 1.05 (Figure 2.2 D). Although this is a very strict value for imputation accuracy, the Q-Q plot clearly shows there is significant type I error.

A more practical way of evaluating this approach is to look at the false positive rate. Specifically, although no SNPs are associated with case/control status based on the true genotypes, there were 4016 imputed SNPs that reach genome-wide significance ($p<5\times10^{-8}$). The IQS filter >0.9 eliminated all the false positive SNPs, but the imputation accuracy filter >0.99 still retained 759 false-positive SNPs. Based on these results, IQS is better for discriminating between well-imputed SNPs and poorly-imputed SNPs.

Although IQS can serve as an effective filter to minimize the use of poorly-imputed SNPs, the computation of IQS requires a sample that was both imputed and genotyped for the SNPs of interest. This is impractical in most situations. A secondary goal of this paper is to determine whether there are ways to evaluate imputation quality without knowing the true genotypes.

The two common methods for filtering imputed data are to combine a minor allele frequency threshold with either the imputed information score >0.3 ~ 0.5 (PROPER_INFO in SNPTEST) [39, 46, 48-50] or the variance ratio >0.3 (rsq_hat in MACH) [36, 39, 48-53]. We calculated these two statistics for our data and compared these filters to IQS (Table 2.3). After filtering by these statistics, the type I error inflation decreases. In the AA sample, IQS also acts as an effective filter and can be cautiously approximated by a combination of MAF and either the imputed information score or the variance ratio (Table 2.4). Unfortunately, even in the most conservative situation, over three thousand false positives remain. Therefore this is an ineffective approach for filtering poorly-imputed SNPs.

Filtering on MAF differences between the Hapmap and the study genotypes is another possible approach to control false positives. In Table 2.3 and Table 2.4, we provided results filtered by MAF difference at 0.01, 0.1 and 0.2 for European Americans and African Americans, respectively. Filtering by MAF difference of 0.01 resulted in a reduction of false positives, but retained less than 25% of the SNPs. In contrast, filtering with a MAF difference of 0.1 or 0.2 retained many false positives.

A second method for using IQS without directly genotyping would be to develop a database of common imputations in common populations that records IQS scores for each SNP. To test the practicality of this approach, we randomly divided the data into two groups and tested the robustness of the IQS score for the SNPs imputed from the Illumina 550K array to the Illumina 1M array in both EA and AA. Because small changes in the denominator of IQS ($1-P_c$) will dramatically affect the value of the statistic when MAF is small, we included only SNPs with MAF>0.01. Figure 2.3 plots the IQS scores in both populations. The correlation in EA is 0.99519 and the correlation in AA is 0.99020, indicating that IQS is robust for the same imputation in a relatively homogeneous population.

We further tested whether the set of hard-to-impute SNPs compiled from the first group can be used to filter the imputed data in the second group. We applied a similar procedure as in Figure 2.1. We randomly divided the second group into cases and controls. Cases were genotyped on the Illumina 550K array and the remaining Illumina 1M SNPs were imputed. Controls were genotyped on the Illumina 1M array. Figure 2.4 shows that the QQ plot can be adjusted to normal by IQS calculated from the first group. This implies that the development of a database of IQS scores for standard imputations would allow researchers to use data genotyped on different platforms and filter out potential false positives.

In order to confirm these results in a different dataset, we replicated the study in European American subjects genotyped on two different platforms, Affymetrix 5.0 array and Illumina 550K array. All subjects were controls from the National Institute of Mental Health Center for Collaborative Genetic Studies on Mental Disorders. We randomly divided about 400 individuals into two subgroups labeled “cases” and “controls” in a similar manner as above. “Cases” were genotyped by the Affymetrix 5.0 array and “controls” were genotyped by the Illumina 550K array. In the replication, we also expanded our investigation to include those SNPs that were not genotyped in either array, but were imputed from their respective arrays. In fact, we had genotype data from both platforms. No genome wide significant SNPs were found. Therefore, if there were any significant SNPs in this simulation, they should be false positives. The result was similar with inflation of Type I error that is effectively filtered by IQS, whereas filtering by MAF and either the imputed information score or the variance ratio continue to have many false positive values (Table 2.5).

DISCUSSION

There are two situations in which imputation is avoided[46]: (1) SNPs with low minor allele frequency and (2) cases and controls genotyped on different platforms. The statistics previously used for measuring the accuracy of imputation are inadequate for evaluating the quality of imputation due to their dependence on marginal SNP frequency. Specifically, imputation accuracy, a measure of the concordance rate between the imputed and observed genotypes for each SNP, dramatically over-estimates reliability when minor allele frequencies are low and does not address the inflation of false positive rates arising from imputation error due to random agreement. We developed IQS to more precisely estimate imputation error, effectively filtering imputation error in these two problematic situations. We showed that IQS is a more appropriate measure to evaluate imputation reliability because it adjusts for "chance" agreement, and filtering by IQS eliminates the inflation of the false positive rate arising from imputation error.

It is important to note that the traditional genome inflation factor λ is not an ideal indicator of potential problems related to imputation quality. In our studies, we noticed that λ is not dramatically different from 1, in contrast to the extent that the Q-Q plot is distorted (Figure 2.2 B D). The reason is that λ reflects systematic inflation on all SNPs while the distortion of the Q-Q plot in our studies is due to a small number of poorly-imputed SNPs. However, problems with this limited number of SNPs (less than 0.5% of total SNPs) can be dramatic and lead to pronounced false positive P values that exceed genome wide significance.

We also would like to emphasize that we are dealing with the extreme situation when cases and controls are genotyped on different platforms. The elevated false positive rates are not explicitly reported in the literature, as most groups do not have this problem because of the study design. But many groups have noticed it. In a recent paper by de Bakker[46], the author noted "the dangers of combining cases genotyped on one platform and controls genotyped on another" (Page 124). In the GENEVA consortium, there is a consensus that genotypes imputed from one array should not be combined with imputed genotypes from another array.

The reasons for the false positives are very complicated. Among the 4016 genome wide significant SNPs, most of them have low R square with other available SNPs. It is difficult to correctly assign their values based on related haplotypes, and they therefore tend to receive the allele frequency from the reference panel.

Filtering by the difference between the reference and the estimated minor allele frequency can effectively remove some genome wide significant SNPs. Of the 4016 genome wide falsely-significant SNPs, 3120 (77.7%) SNPs are removed by removing those SNPs whose minor allele frequency difference is greater than 0.01.However, there are still 832 (21% of the 4016 SNPs) that have passed the filter. Most of the 832 remaining SNPs share one character: they tend to have very low minor allele frequency (MAF median =0.00096). Imputation tends to over-assign the major genotype to the imputed SNPs, resulting in different allele frequency and therefore inflating the P value. However, to filter by MAF difference at 0.01 is not an acceptable option. Most SNPs are correctly predicted even if the minor allele frequency is different. When we tried to remove all SNPs whose minor allele frequency difference was greater than 0.01, 583,456 of the total 788,944 available SNPs (74%) were removed. Most of these SNPs were correctly predicted even if minor allele frequency was different. This is because imputation does not assign predicted genotype based on minor allele frequency, but rather on haplotype modeling.

The typical methods for filtering poorly-imputed SNPs are using either the variance ratio or the imputed information score combined with minor allele frequency. Imputation quality is especially important in a study that combines genotypes from different platforms. Therefore, we increased our thresholds for variance ratio and the imputed information score in Table 2.3, Tables 2.4 and 2.5. But these measures were ineffective in this extreme situation. However, IQS may be used as an effective filter to combine data genotyped on different platforms.

Because IQS requires direct genotyping for evaluation, it is not a practical statistic for directly evaluating imputation in the case where imputation is used to screen for associations as a proxy for genotyping. However, IQS was shown to be a robust measure of imputation for specific imputations (from one standard platform to another) and within a broad population (tested in both EA and AA).

Generally speaking, different populations have different linkage disequilibrium structures and different allele frequencies that lead to different IQS values. A mixture of different populations will make the IQS sensitive to the ratio of population mixture. Therefore, as in general association studies, a mixture of different populations should be avoided. However, African Americans have a unique and relatively stable genetic structure. The IQS score from African Americans is stable in our study and is useful to filter out poorly imputed SNPs.

Based on this theory, a database can be constructed and used to filter future imputations and to avoid false positive associations. In order to advance the development of this database, we have posted IQS scores for imputation from Illumina 550K to Illumina 1M for CEPH on the website of the NIMH Center (http://www.nimhgenetics.org/). We envision this as a dynamic database to be updated when new datasets include subjects genotyped on multiple platforms. We will further provide IQS scores for various array combinations when the genotype data of 6,000 controls typed on both the Affymetrix 6.0 and Illumina 1M array are available in the near future [24]. The future database will include IQS scores for the following imputations: (1) from Affymetrix 6.0 to Illumina 1M, (2) from Illumina 1M to Affymetrix 6.0, (3) from Illumina 300K to Affymetrix 6.0 plus Illumina 1M, (4) from Illumina 550K to Affymetrix 6.0 plus Illumina 1M, and (5) from Affymetrix 5.0 to Affymetrix 6.0 plus Illumina 1M. Although genotyping will be ultimately required to confirm associations, using IQS as a filter will decrease the amount of false positive findings that arise, making follow up of positive associations practical.

As genome wide association studies move toward rare variants, over-estimation of the quality of imputation due to chance concordance of uncommon alleles will be more common. In addition, imputation will and should be used to analyze increasingly complex data structures. IQS can be used as an accurate evaluation of imputation quality enabling researchers to examine low allele frequencies and complex data structures.

Table 2.1 - Marginal cross classification of the genotypes used for the computation of IQS

Imputed Genotypes	True genotypes			
	AA	**AB**	**BB**	**Total**
AA	n_{11}	n_{12}	n_{13}	$n_{1.}$
AB	n_{21}	n_{22}	n_{23}	$n_{2.}$
BB	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

IQS adjusts for minor allele frequency by comparing observed frequencies to expected frequencies.

Table 2.2 - Summary of evaluation measures for European American and African American samples

Evaluation Measures		**European Americans**	**African Americans**
	No. of imputed SNPs	260908	304425
Imputation Accuracy	Efficiency %	94.5	85.1
	Mean %	98.8	97.1
	Range %	0.0~100.0	0.0~100.0
	Inter-quartile %	98.8~99.9	96.3~99.5
Imputation Quality Score (IQS)	Mean %	90.2	78.3
	Range %	-9.1~100	-7.9~100
	Inter-quartile %	90.7~99.2	68.4~94.3

Table 2.3 Comparison of empirical evaluations of imputation quality to IQS in European Americans

False positives n (retained %)	Minor Allele frequency		
	>0.01	>0.05	>0.10
IQS > 0.9	0 (89.47%)	0 (83.90%)	0 (72.92%)
No filter	3120 (96.63%)	2331 (89.48%)	1775 (77.47%)
Proper_info >0.5	3093 (96.62%)	2329 (89.48%)	1775 (77.47%)
Proper_info >0.7	2726 (96.32%)	2080 (89.28%)	1571 (77.31%)
Proper_info >0.9	1392 (94.16%)	1032 (87.67%)	805 (76.06%)
Variance Ratio >0.3	1869 (96.22%)	1526 (89.27%)	1234 (77.33%)
Variance Ratio >0.5	1226 (95.65%)	928 (88.89%)	770 (77.04%)
Variance Ratio >0.7	789 (94.57%)	514 (88.12%)	390 (76.47%)
Variance Ratio >0.9	498 (90.40%)	253 (85.00%)	153 (74.14%)
MAF difference <0.01	267 (22.89%)	120 (19.63%)	76 (15.60%)
MAF difference <0.1	2516 (95.11%)	1739 (87.97%)	1191 (75.94%)
MAF difference <0.2	2952 (96.57%)	2168 (89.42%)	1615 (77.38%)

The sample is based on 2,597 European Americans that were randomized to cases and controls. Cases used genotypes from the Illumina 550K platform and were imputed to the 1M platform and controls were genotyped on the 1M platform. Genome-wide significance is set as $p<5\times10^{-8}$. There were 792,563 SNPs available. False positives refer to the absolute number of SNPs that reached genome-wide significance despite the filter. The retained percentage is the proportion of SNPs that passed the filter.

Table 2.4 - Comparison of empirical evaluations of imputation quality to IQS in African Americans

False positives n (retained %)	Minor Allele frequency		
	>0.01	>0.05	>0.10
IQS > 0.9	1 (76.82%)	1 (70.69%)	1 (61.99%)
No filter	8634 (99.35%)	7671 (90.77%)	5537 (77.72%)
Proper_info >0.5	8620 (99.35%)	7665 (90.77%)	5537 (77.72%))
Proper_info >0.7	8087 (98.83%)	7254 (90.37%)	5205 (77.43%)
Proper_info >0.9	3715 (90.72%)	3243 (83.52%)	2058 (72.45%)
Variance Ratio >0.3	4728 (98.32%)	4185 (90.04%)	3530 (77.30%)
Variance Ratio >0.5	3059 (96.68%)	2524 (88.66%)	1976 (76.33%)
Variance Ratio >0.7	2092 (93.09%)	1570 (85,49%)	1054 (73.97%)
Variance Ratio >0.9	1518 (81.03%)	1020 (74.97%)	557 (65.59%)
MAF difference <0.01	348 (14.05%)	241 (12.44%)	159 (9.97%)
MAF difference <0.1	6288 (90.53%)	5331 (81.89%)	3227 (68.81%)
MAF difference <0.2	8461 (99.22%)	7495 (90.57%)	5363 (77.46%)

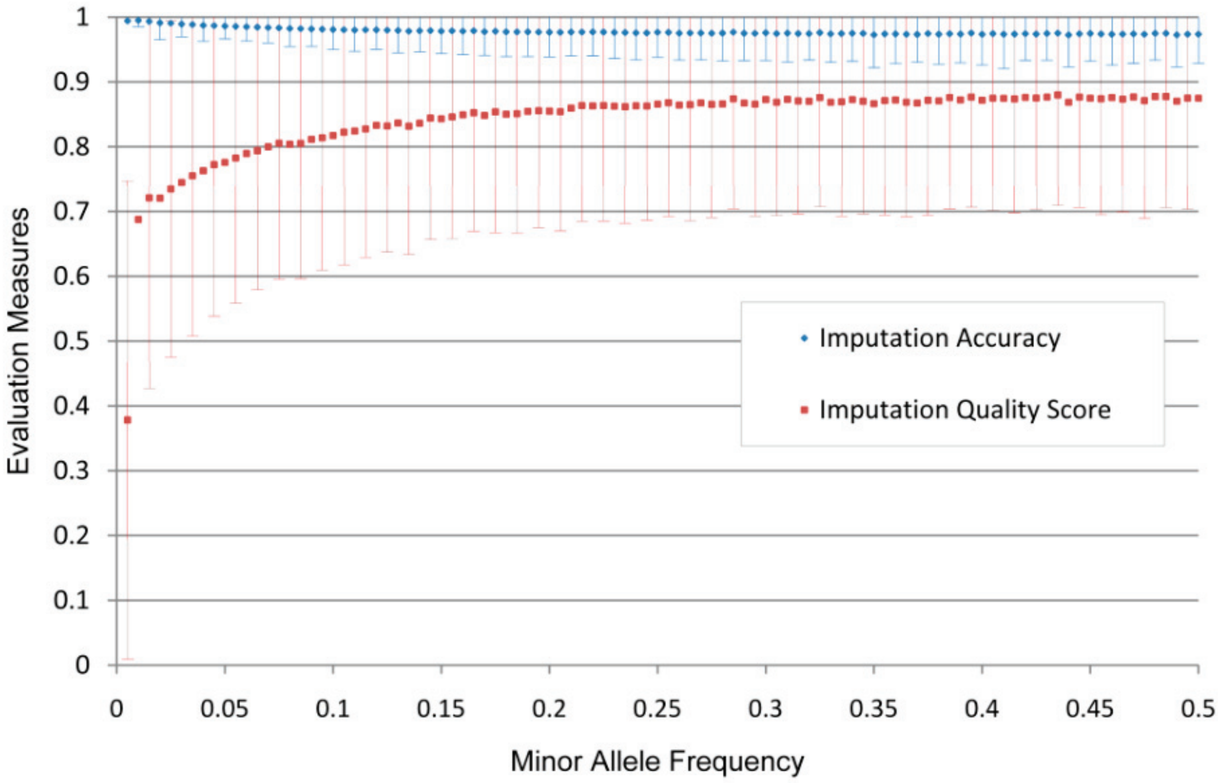
Sample is based on 1264 African Americans that were randomized to cases and controls. Cases used genotypes from the Illumina 550K platform and were imputed to the 1M platform and controls were genotyped on the 1M platform. Genome-wide significance is set as $p<5\times10^{-8}$. There were 837,001 SNPs available. False positives refer to the absolute number of SNPs that reached genome-wide significance despite the filter. The retained percentage is the proportion of SNPs that passed the filter.

Table 2.5 - Comparison of empirical evaluations of imputation quality to IQS in combining Affymetrix 5.0 and Illumina 550K SNPs

False positives n (Retained %)	Minor Allele frequency		
	>0.01	>0.05	>0.10
No filter	2047 (97.09%)	1536 (85.55%)	1107 (72.97%)
IQS > 0.5	63 (87.5%)	63 (78.48%)	60 (67.30%)
IQS > 0.7	2 (80.52%)	2 (72.78%)	2 (62.64%)
IQS > 0.9	0 (62.99%)	0 (57.62%)	0 (49.87%)
Proper_info >0.5	1550 (94.43%)	1280 (84.12%)	979 (71.89%)
Proper_info >0.7	1300 (90.46%)	1063 (81.23%)	818 (69.74%)
Proper_info >0.9	635 (77.49%)	479 (70.43%)	338 (60.98%)
Variance Ratio >0.3	1657 (95.40%)	1285 (84.34%)	937 (72.05%)
Variance Ratio >0.5	1138 (92.43%)	904 (82.12%)	673 (70.33%)
Variance Ratio >0.7	729 (87.17%)	562 (77.90%)	434 (66.93%)
Variance Ratio >0.9	427 (73.31%)	318 (65.58%)	257 (56.39%)

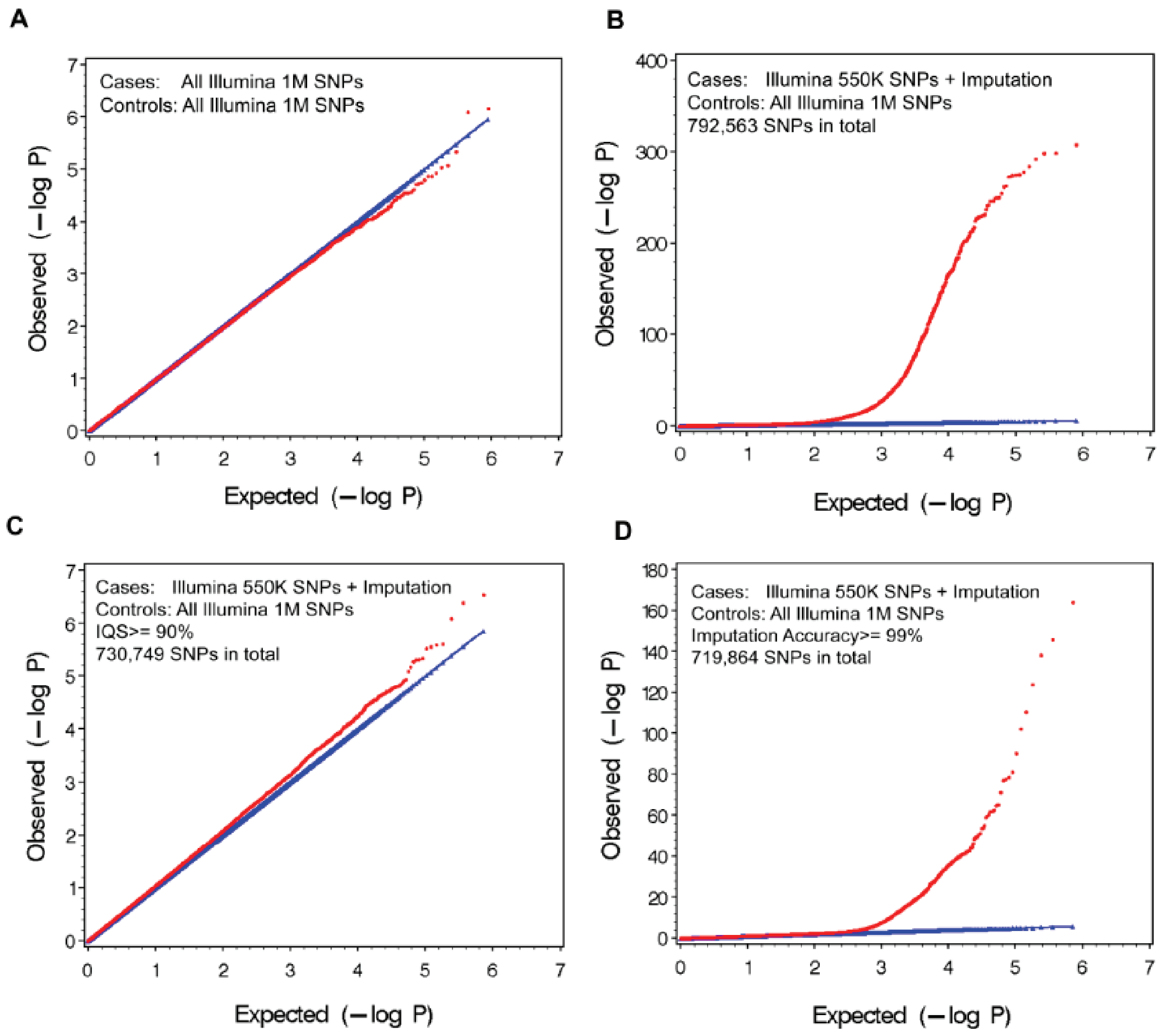
Sample is based on 418 healthy European Americans from NIMH. Cases used genotypes from the Affymetrix 5.0 platform and were imputed to the Illumina 550 platform and controls were genotyped on the Illumina 550 platform and imputed to the Affymetrix 5.0 platform. Genome-wide significance is set as $p<5\times10^{-8}$. There were 2,553,465 SNPs available (including Hapmap SNPs). False positives refer to the absolute number of SNPs that reached genome-wide significance despite the filter. The retained percentage is the proportion of SNPs that passed the filter.

Figure 2.1 - The means of IQS and imputation accuracy within each minor allele frequency interval



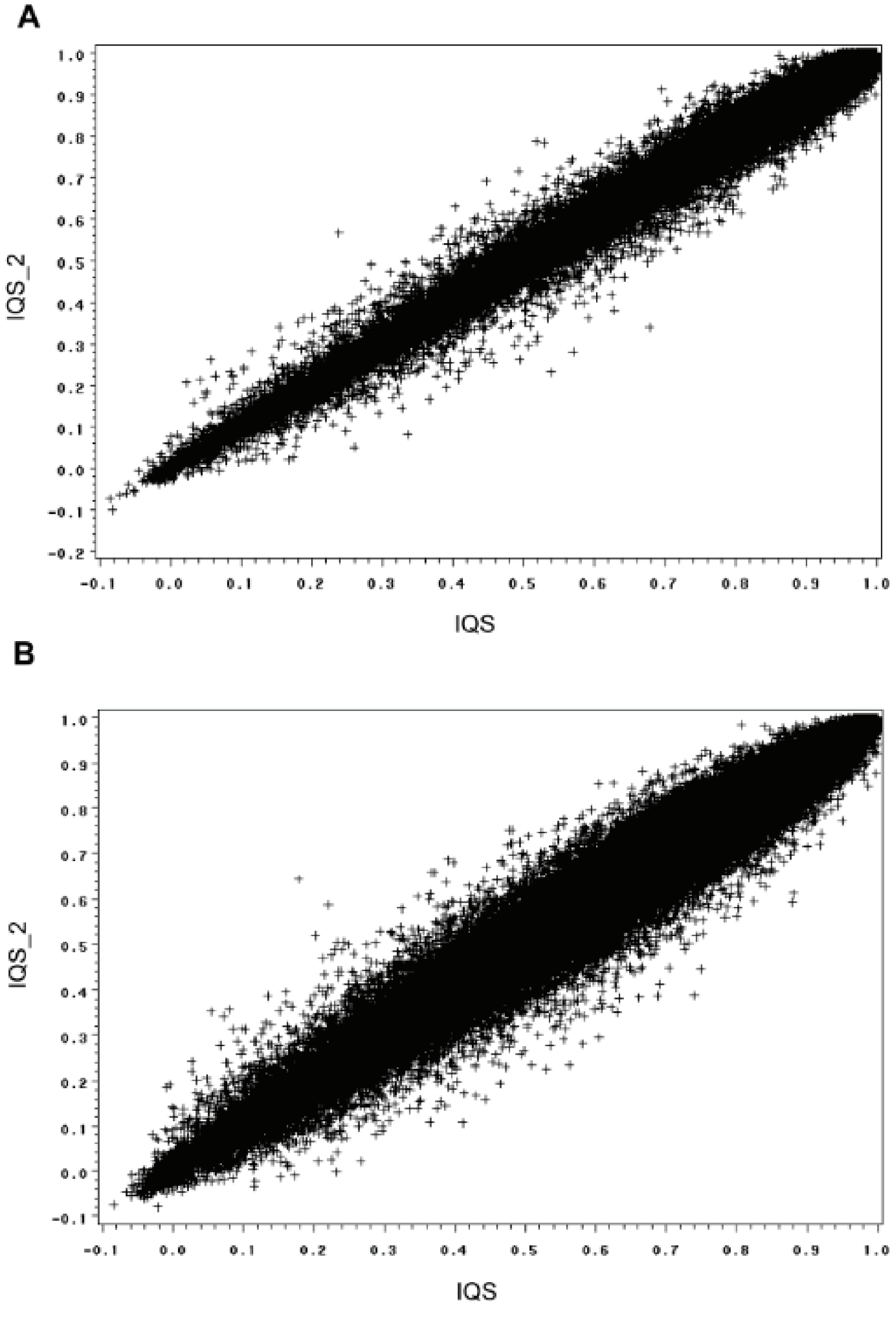
IQS adjusts for chance agreement. As the minor allele frequency approaches 0, the difference between IQS and imputation accuracy increases. The standard deviation is shown for every other point.

Figure 2.2 - The Q-Q plots based on randomly dividing data into cases and controls



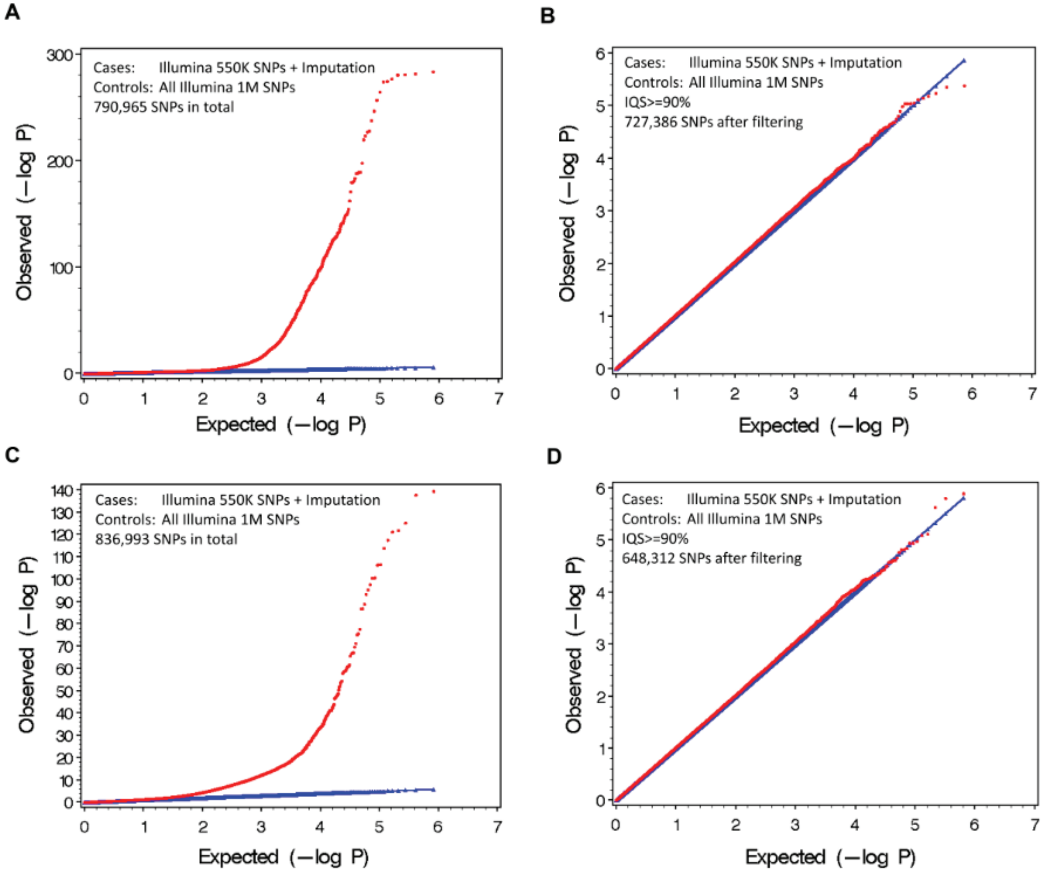
Samples were divided randomly into cases and controls. (A) All Illumina 1M SNPs are directly genotyped indicating there is no population stratification or other non-random factors in cases and controls. (B) Cases were genotyped on the Illumina 550K array and the remaining Illumina 1M SNPs were imputed. (C) An IQS filter (IQS>0.9) was applied, retaining 92% of the SNPs. (D) An imputation accuracy filter (>0.99) was applied, retaining 91% of the SNPs.

Figure 2.3 - Evaluation of the robustness of IQS score



European Americans (A) and African Americans (B) datasets were split in half and Illumina 550K SNPs were imputed to Illumina 1M SNPs. IQS score for the two halves of the data were plotted against each other. SNPs with minor allele frequency less than 0.01 were excluded to avoid zero in the denominator.

Figure 2.4 - A database of IQS can be used to filter poorly-imputed SNPs



The set of hard-to-impute SNPs compiled from one dataset can be used to filter the imputed data in another dataset. (A) Cases were European Americans genotyped on the Illumina 550K array and the remaining Illumina 1M SNPs were imputed. Controls were European Americans genotyped on the Illumina 1M array. The QQ plot was shown for the 790,965 available SNPs. (B) An IQS filter (IQS>0.9) was applied, retaining 92% of the SNPs. IQS was calculated from an independent dataset. (C) A similar QQ plot for African Americans. Cases were genotyped on the Illumina 550K array and the remaining Illumina 1M SNPs were imputed. Controls were genotyped on the Illumina 1M array. The QQ plot was shown for the 836,993 available SNPs. (D) An IQS filter (IQS>0.9) was applied, retaining 78% of the SNPs. IQS was calculated from an independent dataset.

Chapter 3: Copy Number Variation Accuracy in Genome Wide Studies

[*] This chapter is adapted from: Lin P et al. (2011) Copy Number Variation Accuracy in Genome Wide Studies. Human Heredity. In Press

ABSTRACT

Copy Number Variations (CNVs) are a major source of variation between individuals and are a potential risk factor in many diseases. Numerous diseases have been linked to deletions and duplications of these chromosomal segments. Data from genome-wide association studies (GWAS) and other microarrays may be used to identify CNVs by several different computer programs, but the reliability of the results has been questioned.

To help researchers reduce the number of false positive CNVs that need to be followed up with laboratory testing, we evaluated the relative performance of CNVpartition, PennCNV and QuantiSNP, and developed a statistical method for estimating sensitivity and positive predictive value of CNV calls and tested it on 96 duplicate samples in our dataset.

We found that the positive predictive rate increases with number of probes in the CNV and the size of the CNV, with the highest positive predicted rates in CNVs of at least 500kb and at least 100 probes. Our analysis also indicates that identifying CNVs reported by multiple programs can greatly improve the reproducibility rate and the positive predicted rate.

Our methods can be used by investigators to identify CNVs in genome-wide data with greater reliability.

INTRODUCTION

Copy Number Variations (CNVs) are duplications or deletions of a particular segment of an individual's genome. Over the past 10 years, evidence has accumulated that CNVs play an important role in disease [13-19]. It is hypothesized that a CNV changes the expression level of genes in or near those regions, leading to various phenotypes, including disease[12]. Therefore, CNVs constitute a major source of inter-individual variation that could contribute to common disorders and complex traits[54]. The advent of genome-wide association studies (GWAS) has led to the possibility of discovering CNVs across the genome. So far, many CNV detection programs have been developed for this purpose, including CNVpartition, PennCNV, and QuantiSNP.

However, despite the obvious scientific importance of understanding the role that CNVs play in human disease, there is some controversy regarding the use of GWAS data to detect CNVs. First, a recent study suggested that disease-related CNVs detected from GWAS data are well tagged by SNPs, and, therefore, CNVs do not add further information[55][55][55][55][55][55][55](Rice, Rochberg et al. 1992). Second, there is evidence that different methods for identifying CNVs from GWAS data report different results, even when applied to the same array data[56].

To address the first controversy, although many common CNVs that are well-typed in a microarray can be tagged by SNPs[55][55][55][55][55][55][55](Rice, Rochberg et al. 1992), there are at least three reasons why testing the association between a trait and CNVs remains important. First, CNVs may be the true causative variant of the trait, and will therefore show a stronger association than a SNP tag. For example, the copy number of the salivary amylase gene (*AMY1*) is correlated positively with salivary amylase protein level[12]. Second, the number of common CNV loci is limited, and therefore, the typical GWAS significance level of $p<5x10^{-8}$ is overly conservative [57]. After adjusting for multiple tests in GWAS, SNPs tagging associated CNVs are unlikely to be statistically significant at this stringent threshold, although they would be significant in a setting where only CNVs were tested. Third, de novo CNVs are not well-tagged by SNPs. In addition, tagging a recurrent CNV by multiple SNPs demands heavy computation. Thus, despite the potential for some CNVs to be tagged by SNPs, many researchers continue to look for CNVs in GWAS data [58].

The second controversy with localizing CNVs is the imprecision of estimation. Methodologies for measurement of CNVs in GWAS microarrays continue to evolve, leading to the varied results mentioned above. Currently, most methods that make use of SNP microarray data to detect CNVs depend on LogR ratio and b-allele frequency from microarray data. One simple and straightforward method draws LogR ratio and b-allele frequency as the Y axis and chromosome position as the X axis. When a deletion or duplication occurs, the pattern of LogR ratio and b-allele frequency will change accordingly [59]. However, this method requires extremely high data quality and necessitates investigators spot pattern changes. Subsequently, more sophisticated methods of identifying CNVs have attempted to adjust undesirable microarray artifacts, such as genomic wave [60], and build a mathematical model to detect CNVs from those data. Numerous programs have been written for this purpose. The most widely used are

CNVPartition (http://www.illumina.com/software/illumina_connect.ilmn), PennCNV[28] and QuantiSNP [61]. Although all three programs use standard statistics from the observed data to estimate the location of CNVs, they use different iterative mathematical methods. CNVPartition uses a likelihood-based algorithm, PennCNV implements a hidden Markov model (HMM), and QuantiSNP uses an Objective Bayes Hidden-Markov Model (OB-HMM). A detailed comparison of these different algorithms can be found in Dellinger et al' study[56]. These three programs have helped many studies find putative disease-related CNVs[62-66]. Moreover, several recent studies have used SNP microarray data to study the characteristics of CNVs[59, 67]. However, there is evidence that the varied algorithms identify different CNVs even with the same data, questioning the reliability of using these programs to detect CNVs [56].

Although laboratory confirmation is necessary to validate CNVs derived from SNP array platforms[14, 57, 63-66], it is not economically feasible to validate all CNVs in a genome-wide scale, especially for the purpose of estimating measurement accuracy. Here, using duplicates in a GWAS sample, we develop an algorithm to better evaluate the accuracy of CNVs predicted by several CNV calling algorithms for GWAS data. Whether a CNV that is called the first time can be confirmed the second time is restricted by sensitivity and specificity. This gives some insight about CNV calling accuracy to investigators wishing to evaluate CNVs found in SNP microarray data that might be associated with disease.

METHODS

Data and quality control

The dataset was collected as part of the Study of Addiction: Genetics and Environment (SAGE) [68]. SAGE is part of the Gene Environment Association Studies(GENEVA) project (http://genevastudy.org/) [58]. All participants in SAGE provided written informed consent for genetic studies and agreed to share their DNA and phenotypic information for research purposes. The institutional review boards at all data collection sites granted approval for the use of the data. In this study, all samples were de-identified and only subjects who consented to health research were included.

Samples were genotyped on the Illumina Human 1M array at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University. The Illumina 1M array has a total of 1,072,820 probes, of which 23,812 are "intensity-only". Data cleaning procedures included using HapMap controls, detection of gender mis-annotation and chromosomal anomalies, cryptic relatedness, population structure, batch effects, Mendelian error detection, and duplication error detection [23, 68]. In this study, 107 study subjects were genotyped in duplicate on the Illumina 1M array. These subjects were selected randomly from the study sample for the purpose of assessing genotyping accuracy. The mean of the SNP calling discordance rate between the duplicates was 0.02%. These duplicates were further compared against each other to determine the accuracy of CNV calling.

CNV calling

We used three common programs to call CNVs: CNVPartition, PennCNV, and QuantiSNP. We also implemented a procedure to adjust genomic waves when we called CNVs by PennCNV and QuantiSNP[60]. Both PennCNV and QuantiSNP report data quality control measures. In order to pass the quality control, a subject and its replication need to be considered as good quality by both PennCNV and QuantiSNP. After quality control, 96 subjects and their replications passed these filters. CNVpartition does not provide any quality control information for individual subjects. We also removed all CNV calls with Log Bayes Factor(LBF) less than 10, which is recommended by QuantiSNP (See supplementary materials for more details).

PennCNV reports log R ratio standard deviation (LRR_SD), B allele frequency drift (BAF_Drift) and waviness factor (WF) for quality control. We used the following criteria to look for good samples (http://www.yale.edu/state/Pipeline.htm).

1. LRR_SD < 0.28
2. BAF_Drift <0.01
3. WF > -0.05 and < 0.05

If a subject satisfied these criteria, it was considered a sample with good quality. Similarly, QuantiSNP reports B allele frequency outliers (BAFout), LogR standard outliers (LogRout), B allele frequency standard deviation (BAFstd) and LogR standard deviation (LogRstd) for each chromosome. The following criteria were used to determine good quality.

1. BAFout ≤ 0.1

2. LogRout ≤ 0.1
3. BAFstd ≤ 0.2
4. LogRstd ≤ 0.4

For any subject, if one or more autosomal chromosomes did not satisfy these criteria, the sample was considered poor quality.

In order to pass quality control, a sample and its replication had to be considered as good quality by both PennCNV and QuantiSNP. After quality control, 96 subjects and their replication data passed the filter. CNVpartition does not provide any quality control information for individual subjects.

In these samples, we identified 2348 potential regions across the genome for deletions and 851 potential regions for duplications. For any particular potential region, at least one of these 96 subjects had a duplication or deletion in this region. We restricted our study to only these potential regions.

Each program also reports a confidence score based on different mathematical models. The confidence score is a positive number representing the likelihood that there is a CNV at that region, with a higher number representing a greater probability of a CNV in that region. The confidence scores for the three programs are calculated differently and are on different scales. CNVPartition uses a likelihood-based method to compute the confidence score (http://www.illumina.com/software/illumina_connect.ilmn). QuantiSNP computes a Bayes Factor by comparing the evidence of the region containing deletions or duplications to that of having two copies, and reports the log of the Bayes Factor as the confidence [61]. PennCNV reports an experimental confidence score that is not well documented[28]. These confidence scores allow users to filter out CNV regions that are likely to be false positives. Due to variability of the confidence score distributed among the three programs, we converted the confidence scores within each program into percentiles and used them as covariates for modeling.

Comparative statistics

These CNV calls are then compared against each other among duplicate samples. Concordance is defined as the percentage of regions that have been consistent in the existence or absence of CNVs between duplicate samples. However, this measure is misleading, because a large percentage agreement is the chance agreement of negatives.

In addition to the concordance rate, we reported the reproducibility rate. We define a CNV as being reproduced when the percentage of overlap of these two CNVs is greater than 30% of the region where the two CNVs cover. The reproducibility rate is defined as, the percentage of CNVs that can be reproduced at time point 2 among CNVs that are discovered at time point 1.

Statistical modeling

Whether a CNV discovered at the first time can be confirmed at the second time is restricted by sensitivity and specificity. In turn, this information can be used to estimate sensitivity and specificity. Using a model derived in previous work [55, 69, 70], we calculated CNV sensitivity and positive predicted rate with logistic regression parameters derived from CNV characteristics. All CNVs called by any program or more than one program were used to fit the model. We also added the consistency rate--the number of programs reporting a CNV at a particular locus--as a covariate. The mathematical model allows us to estimate the cumulative probability of being true for a set of CNVs with similar characteristics, and thus avoids the issue of testing whether a particular CNV is true or not.

The model is derived from our previous work on psychiatric disorder diagnoses [55, 70]. For some psychiatric disorders, the symptoms are not stable, and similar to CNV calling, the diagnoses may not be accurate. Whether a positive diagnosis at time point 1 can be confirmed at time point 2 is determined by sensitivity and specificity. And this information can be used to estimate sensitivity and specificity *per se*. Similarly in this study, we used the result of CNV calling from replication samples to estimate sensitivity and specificity. We realized that testing the validity of each CNV is unfeasible on a genome wide scale. We therefore constructed a mathematical model to estimate the probability of being true at a larger scale.

For the mathematical model, we defined the reproducibility as being reproduced by any of the three algorithms; however, there are arguably several valid ways to do this. Some may require a region shared by one, two, and all of the three algorithms. We decided to calculate the maximum coverage of all available algorithms: it could be a direct report if only one algorithm was available, maximum coverage of two if two were available and maximum coverage of three if three were available. Then we tested whether the shared region of the maximum coverage was over 30% of the total coverage. We chose this approach because: 1. this issue is overcomplicated and needs to be simplified, and 2. the boundary of CNV is hard to define, thus any report may be treated as evidence of a CNV.

We modeled the positive predicted rate for duplication and deletion separately. We let *T* denote true state and O_i denote observed state, at time *i* (*i*=1,2). *T* and O_i take the value 1 for "presence" and 0 for "absence" of CNVs (duplications or deletions). The sensitivity *p* and specificity *q* are

$$p = \Pr(O_i = 1 \mid T = 1)$$

$$q = \Pr(O_i = 0 \mid T = 0)$$

Each CNV calling program typically reports a value for calling confidence. Let *Z* denote a set of CNV characteristics, including percentile distribution of confidence scores from each program and the number of programs that report a CNV at a particular CNV segment. Let *k* denote the true base rate at this region. Therefore, we have

$$k = \Pr(T = 1)$$

Then the probability of observing a CNV (duplication or deletion) at evaluation *i* is given by

$$\text{(1)} \qquad P_+ = \Pr(O_i = 1 \mid Z) = pk + (1-q)(1-k)$$

Similarly, the probability of observing a case at the second time, conditional on observing a case at the first time at condition *Z* is

$$\text{(2)} \quad \Pr(O_2 = 1 \mid O_1 = 1, Z) = p\Pr(T = 1 \mid O_1 = 1, Z) + (1-q)(1 - \Pr(T = 1 \mid O_1 = 1, Z))$$

In order to estimate theoretical sensitivity for this model, we used the following methodology. At time 2, the probability of confirming a true positive which is discovered at time 1 (i.e. identifying a CNV in the second replicate, given that it is true and it was seen in the first replicate) is the theoretical sensitivity of this model. Let Z_{max} denote the theoretical condition for which all three CNV estimation programs asymptotically reach the maximum value for prediction of a CNV. Then, we make the assumption that if a CNV is identified (O_1=1) with the theoretical maximum values of the three programs (Z_{max}), then the CNV is a true positive (T=1):

$$\Pr(T=1 \mid O_1=1, Z_{\max})=1$$

This is a theoretical situation. Because we used percentile distributions of confidence scores, Z_{max} should be 100 for each confidence score, and 3 for the consistence rate, which are the theoretical maximum values that Z_{max} can reach. Combining $\Pr(T=1 \mid O_1=1, Z_{\max})=1$ with Function (2), we have:

$$p=\Pr(O_2=1 \mid O_1=1, Z_{\max})$$

where p is the theoretical sensitivity of our mathematical model. The value of $\Pr(O_2=1 \mid O_1=1, Z)$ can be modeled by a logistic regression model because O_2, O_1, and Z are all observed values:

$$\Pr(O_2=1 \mid O_1=1, Z)=1/[1+\exp(-\alpha-\beta Z)]$$

The percentile of confidence scores from CNVpartition, PennCNV and QuantiSNP, as well as the consistency rate, were all significant for duplications or deletions. Based on the logistic regression and the formula $p=\Pr(O_2=1 \mid O_1=1, Z_{\max})$, we estimated that the theoretical sensitivity is 0.91 for duplications, and 0 .97 for deletions. This is the theoretical sensitivity for the mathematical model, and should be distinguished from the sensitivity *p'* for each subcategory. We also wanted to point out that these regions are restricted to 851 potential regions for duplications and 2,348 potential regions for deletions.

We have duplicate samples. The probability of observing a CNV (duplication or deletion) at both times is given by

$$(3) \qquad P_{++} = \Pr(O_2 = 1, O_1 = 1 \mid Z_1, Z_2) = p^2 k + (1-q)^2 (1-k)$$

Now, *q* can be solved by combining function (1) and (3),

$$q = \frac{p + P_{++} - P_{+} - pP_{+}}{p - P_{+}}$$

From this formula, we also can estimate that the theoretical specificity, which is 0.986 for duplications and 0.989 for deletions.

For each CNV, by solving function (2), we can obtain the positive predicted rate R_+, which is the probability of being a CNV, conditioned on being positive at time 1.

$$R_+ = \Pr(T = 1 \mid O_1 = 1, Z) = [\Pr(O_2 = 1 \mid O_1 = 1, Z) + q - 1] / (p + q - 1)$$

R_+ was calculated for all CNVs, and we took the mean value of R_+ of CNVs within each category. This is the positive predicted rate for each subcategory reported in Tables 3.2-3.5. And this value is later used to calculate the sensitivity for each subcategory as described below. We assume *p'*, *q'* are the sensitivity and specificity within each subcategory, respectively, and *k* is the true base rate. We have

(4) $$p' = \frac{P_+}{k} R_+$$

Functions (1), (3) and (4) together only have 3 unknown variables: k, p' and q. P_+ and P_{++} can be obtained from the data directly. By solving functions (1), (3) and (4), we can obtain the following formula for p'.

$$ap'^2 + bp' + c = 0$$

Among them,

$$a = P_+ R_+,\quad b = (P_+ - P_+ R_+)^2 - P_{++} - P_+^2 R_+^2,\quad c = P_+ R_+ P_{++}$$

Finally,

$$p' = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

The negative root is ignored because it is out of boundary.

Given that p' can be calculated from the function above, and both P_+ and R_+ are available, we can also estimate the base rate k by solving function (4). Therefore, we have

$$k = \frac{P_+}{p'} R_+$$

When all CNV callings are considered, the base rate k is 0.016 for deletions and 0.012 for duplications.

Based on this model, we estimated the probability that an observed CNV is a true positive, and further the sensitivity for different methods. Duplications and deletions were modeled separately. The percentile of confidence scores from CNVpartition, PennCNV and QuantiSNP, as well as the consistency rate, were all significant for duplications or deletions, and thus were included in our model (See supplementary materials for more details).

Model validation

In calculating the positive predicted rate R_+, we assumed that $\Pr(T=1 \mid O_1=1, Z_{\max})=1$, i.e., a theoretical CNV is true if three programs simultaneously report the highest confidence scores. This assumption is only used to estimate the theoretical sensitivity of our mathematical model. Z_{max} is the characteristic from a theoretical CNV that has the highest confidence score and is confirmed by all three programs. Because we used percentile distributions, the vector of Z_{max} includes 100 for each confidence score and 3 for the consistence rate. For a theoretical CNV like this, our mathematical model can only capture part of it; therefore probability is the theoretical sensitivity of our mathematical model. This theoretical CNV does not exist in reality. We can only test some CNVs with weaker characteristics. Our rationale is that if a CNV with slightly weaker characteristics can meet the requirement, a theoretical CNV with perfect characteristics can meet the requirement as well. We tested CNVs with weaker characteristics on Chr 14 by qPCR. Among them, 48 subjects have CNVs with R_+ more than 0.91, with the average as high as 0.98. The qPCR experiment confirmed that all 48 subjects were reported as having the CNVs and 0 were false positive.

Based on our model, we were able to calculate the positive predicted rate for each CNV. We grouped CNVs with similar positive predicted rates together and compared the positive predicted rate of each group against the proportion of CNVs from that group that can be reproduced. We reported a CNV as reproduced in the duplicate if the CNV detected by the two independent genotyping shares more than 30% of the total coverage. We were able to obtain agreement between theoretical positive predicted rate and experimental reproducibility in duplicates (Figure 3.1).

We also randomly selected 90% of replicate pairs, and randomly assigned status as discovery or replication, and then we calculated the positive predicted rate for “any 1 of 3 programs” method. We repeated the process 100 times. The positive predicted rate was stable across many repeats (Figure 3.2), indicating that our result is not subject to serious random fluctuations.

RESULTS

We tested the concordance rate of CNV calls from each program in duplicate samples. The concordance rates for the three programs range from 98.0% to 99.3% (Table 3.1). However, concordance rate is not a good indicator of CNV calling reliability, because the concordance rate also includes the agreement of the absence of CNVs. Similar to SNPs with very low minor allele frequencies [71], a large portion of agreement is due to the chance agreement of negatives. Because of this, we believe that the reproducibility rate is a more appropriate measure for CNV calling reliability. We reported a CNV as reproduced in the duplicate if the CNV detected by the two independent genotyping shares more than 30% of the total coverage. The reproducibility among deletions ranged from 59% to 62%, and the reproducibility among duplications ranged from 43% to 57% (Table 3.1). This highlights the variation between methods and the low reliability of all three methods.

We then estimated the reproducibility rate, the positive predicted rate and the sensitivity for each CNV calling method (Table 3.2). As expected, deletions have higher reproducibility rates, higher positive predictive rates and better sensitivity. For both duplications and deletions, the method that requires CNVs to be reported by all three programs has the highest reproducibility rate and the highest positive predicted rate.

False CNV calling may be caused by intensity variation (noise) from the microarray. A short CNV segment with few probes is particularly vulnerable to noise. Because of this, we estimated both the reproducibility rate and the positive predicted rate R_+ within four subcategories for each method based upon number of probes contained within the CNV (Table 3.3). Some of these subcategories are often used in the literature as thresholds for quality controls [28]. Not surprisingly, a higher positive predicted rate R_+ was seen when there were more probes in a single CNV. We also tested the relationship between the size of CNV segments and positive predicted rate R_+ (Table 3.5). As expected, the result was similar to Table 3.3, because a larger CNV segment typically contains more probes.

The primary purpose of this study was to determine the reliability of CNVs found in microarrays, such as in GWAS. We found that if a CNV is reported by at least 3 of 3 programs, it has the highest positive predicted rate. Moreover, in a microarray, probes are not always evenly spaced. We hypothesized that the combination of the number of probes and the size would boost the positive predicted rate. We tested this hypothesis by using both the number of probes and the size as filters. The results suggest that a minimum of 10 probes and 10K base pairs are necessary to reach > 80% positive predicted rate (Table 3.4).

DISCUSSION

Data from genome wide association studies can be used to estimate locations of CNVs and their potential effects on disease. There is disturbing evidence that calling CNVs from SNP microarray data is not reliable [56]. For this reason, investigators are interested in quantifying the reliability. To our knowledge, this is the first study that compares CNV calls from a considerable number of duplicate samples.

Although experimental validation is necessary for CNV association studies, it is both demanding and costly and should be limited to regions most likely to contain true CNVs associated with disease. In this study, we introduced a convenient way to identify potential false positive CNVs on a genome wide scale, using an estimated positive predicted rate for CNV callings. Our results confirmed that combining CNVs from different programs is one way to improve the positive predicted rate.

In this study, we found that 10 probes and 10 kb in size maximize CNV calling quality. We also discovered that deletions are much easier to detect than duplications. The reason is that when calling genotypes from the microarray, one deletion represents a 50% decrease in signal intensity, rather than the 33% increase caused by one duplication. In addition, the B allele frequencies—a reported measure from microarray—of those SNPs at a particular deletion region usually take the value of 0% or 100%, leading to a distinctive pattern that is relatively easy to spot.

Different methods for estimating the locations of CNVs use different mathematical models. Both PennCNV and QuantiSNP use hidden Markov models [28, 61], while CNVPartition estimates model parameters using bivariate Gaussian distributions. Each method has its own strengths, but all also have relatively high frequencies of false-positive CNVs. The “3 of 3” method, however, minimizes false positives.

When 3 different programs call the same CNV, different boundaries may be reported, leading to a quandary on how to categorize this particular CNV. To resolve this, we included all CNVs for one category if a CNV reported by any program satisfies the category. Therefore, the total number of CNVs for “3 of 3 programs” may be higher than the total number of CNVs reported by each program alone.

Moreover, the reproducibility in our manuscript is defined either as being reproduced by itself or being reproduced by any of the three algorithms. The exact definition is indicated below in Tables 3.1-3. 2. The reason for this is to address both self-reproducibility and across-the-spectrum reproducibility. In Table 3.1, we adopted “being reproduced by itself” as the criterion in order to show self reproducibility. That is because self reproducibility is a good indicator of reliability when the truth is not known, and also a good point to start with. The fact that a program cannot even reproduce its result is surely a good sign of poor reliability. In Table 3.2, we want to compare the reproducibility among the three algorithms and the three combinational methods, therefore, a consistent criterion, which is across-the-spectrum reproducibility for this table, is needed in order to make the comparison fair and meaningful.

The sensitivity here is restricted to CNVs that can be detected by a microarray. In our data from 96 replication subjects, we identified 2348 potential regions across the genome for deletions and 851 potential regions for duplications. For any particular potential region, at least one of these 96 subjects had a duplication or deletion in this region. Among these regions, the true base rate *k* is 0.016 for deletions and 0.012 for duplications (See the supplementary materials). We restricted our study only to these potential regions. Some CNVs in the genome may be located at particular regions where no probes or very few probes exist. Those CNVs can never be detected by microarray technology, and therefore are excluded from the estimation of sensitivity. The sensitivity here may be better understood as the sensitivity adjusted by the total number of those potential CNV regions. Therefore, the sensitivity reported by our study should not be directly compared to other studies [56, 72].

Based on our model parameters, investigators can estimate the probability that an estimated CNV is true. Interested researchers can estimate the positive predicted rate for their own data if confidence scores and some other information can be provided. Finally, it is important to emphasize that there are benefits to be gained from utilizing multiple CNV calling approaches and then comparing the results between them. This can maximize the sensitivity for discovery, maximize the positive predicted rate for verification, or balance the sensitivity and the positive predicted rate to a desired point. As genome wide association studies move forward from SNPs to CNVs, investigators can better identify CNVs associated with human disease by using multiple estimation programs and calculating the positive predictive rates of observed CNVs.

Table 3.1 - Concordance rate and reproducibility rate for CNVPartition, PennCNV and QuantiSNP

	CNVPartition		PennCNV		QuantiSNP	
	Concordance rate	Reproducibility rate[a]	Concordance rate	Reproducibility rate[b]	Concordance rate	Reproducibility rate[c]
Duplication	99.2%	54%	98.0%	41%	98.9%	48%
Deletion	99.3%	61%	98.6%	62%	98.9%	63%

[a] Reproduced by CNVPartition
[b] Reproduced by PennCNV
[c] Reproduced by QuantiSNP

Table 3.2 - Positive predicted rate R_+ , sensitivity p' and total number of CNVs for different CNV calling methods.

Method	Duplication				Deletion			
	Sensitivity p'	Positive predicted rate R_+	Reproducibility rate*	Total No. of CNVs	Sensitivity p'	Positive predicted rate R_+	Reproducibility rate*	Total No. of CNVs
CNVPartition	0.77	0.69	0.63	849	0.75	0.78	0.77	2227
PennCNV	0.92	0.46	0.41	2001	0.94	0.65	0.64	2348
QuantiSNP	0.83	0.58	0.55	1177	0.91	0.69	0.68	4171
Any 1 of 3 Programs	0.94	0.43	0.40	2199	0.96	0.59	0.59	5767
Any 2 of 3 Programs	0.82	0.61	0.56	1169	0.88	0.76	0.75	3565
3 of 3 Programs	0.75	0.79	0.73	642	0.72	0.89	0.85	1816

*Reproduced by any one of the three programs: CNVPartition, PennCNV and QuantiSNP

Table 3.3 - The positive predicted rate R_+ within subcategories defined by the number of probes

No. of Probes	**Duplication**				**Deletion**			
	<10	**10-50**	**50-100**	**⩾100**	**<10**	**10-50**	**50-100**	**⩾100**
CNVPartition	0.54 (150)	0.69 (486)	0.77 (143)	0.88 (84)	0.70 (1176)	0.85 (974)	0.87 (82)	0.88(54)
PennCNV	0.29 (1009)	0.56 (930)	0.88 (117)	0.95 (39)	0.58 (3433)	0.81 (1334)	0.95 (74)	0.99 (45)
QuantiSNP	0.42 (228)	0.57 (757)	0.70 (163)	0.84 (77)	0.63 (2411)	0.74 (1644)	0.74 (178)	0.89 (81)
3 of 3 programs	0.65 (122)	0.77 (461)	0.87 (157)	0.91 (83)	0.84 (993)	0.92 (1013)	0.96 (101)	0.98 (67)

The total number of CNVs is indicated in parentheses. In calculating the total number of CNVs, we included a CNV for a certain category if the report from any one of the specified programs' satisfied this category.

Table 3.4 - The positive predicted rate R_+ for the “3 of 3 method”

	Duplication size					Deletion size				
No. of Probes	**<1kb**	**1-10kb**	**10-100kb**	**100-500kb**	**⩾500kb**	**<1kb**	**1-10kb**	**10-100kb**	**100-500kb**	**⩾500kb**
<10	0.58 (3)	0.68 (39)	0.64 (81)	0.62 (6)	---	0.80 (47)	0.83 (595)	0.84 (426)	0.76 (9)	---
10-50	---	0.72 (20)	**0.79 (250)**	**0.74 (239)**	**0.72 (20)**	---	0.91 (144)	**0.93 (731)**	**0.91 (247)**	**0.86 (24)**
50-100	---	---	**0.89 (19)**	**0.91 (73)**	**0.82 (74)**	---	---	**0.95 (39)**	**0.98 (45)**	**0.96 (24)**
>100	---	---	**0.79 (2)**	**0.96 (29)**	**0.90 (56)**	---	---	**0.99 (3)**	**0.99 (32)**	**0.97 (37)**

The total number of CNVs is indicated in parentheses. In calculating the total number of CNVs, we included a CNV for a certain category if the report from any one of the specified programs’ satisfied this category.

Table 3.5 - The positive predicted rate R_+ within subcategories defined by size

	Duplication					Deletion				
Size	**<1kb**	**1-10kb**	**10-100kb**	**100-500kb**	**≥500kb**	**<1kb**	**1-10kb**	**10-100kb**	**100-500kb**	**≥500kb**
CNVPartition	0.46 (4)	0.54 (75)	0.68 (360)	0.72 (300)	0.77 (120)	0.63 (57)	0.72 (784)	0.80 (1115)	0.85 (255)	0.88 (64)
PennCNV	0.26 (39)	0.31 (428)	0.43 (1076)	0.60 (478)	0.78 (79)	0.51 (191)	0.60 (2229)	0.69 (2091)	0.86 (342)	0.97 (29)
QuantiSNP	0.30 (8)	0.47 (91)	0.57 (515)	0.60 (442)	0.64 (161)	0.41 (305)	0.68 (1505)	0.74 (1745)	0.67 (623)	0.69 (141)
3 of 3 programs	0.58 (3)	0.69 (56)	0.77 (322)	0.78 (311)	0.83 (114)	0.80 (47)	0.84 (690)	0.90 (1075)	0.93 (304)	0.94 (79)

The total number of CNVs is indicated in parentheses. In calculating the total number of CNVs, we included a CNV for a certain category if the report from any one of the specified programs' satisfied this category.

Figure 3.1 - The relationship between positive predicted rate and reproducibility rate in duplicate samples

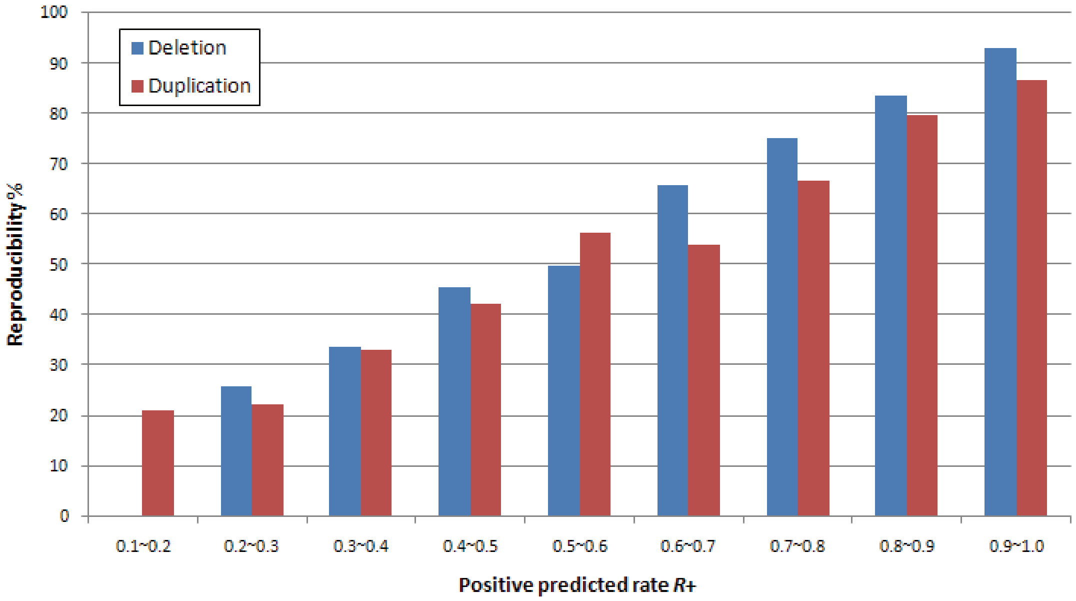
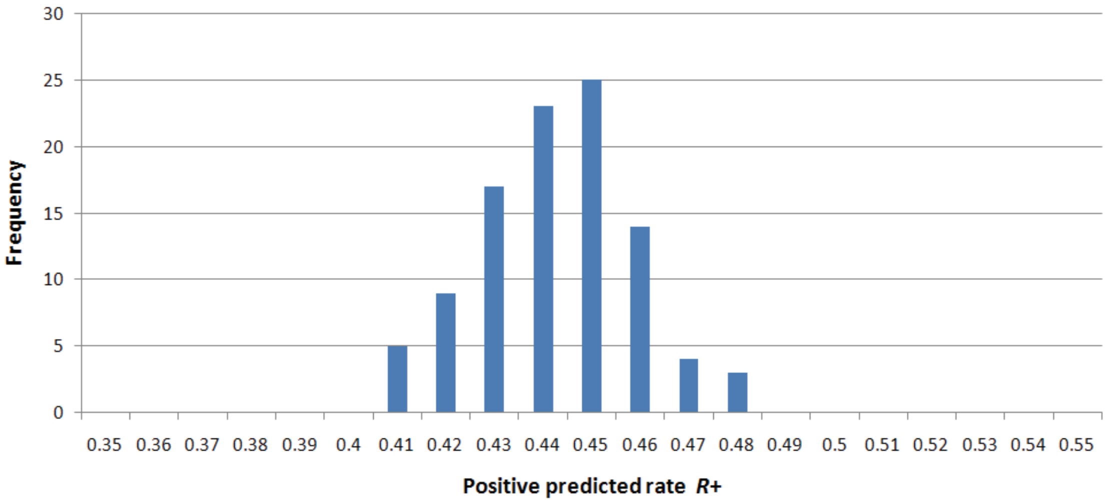
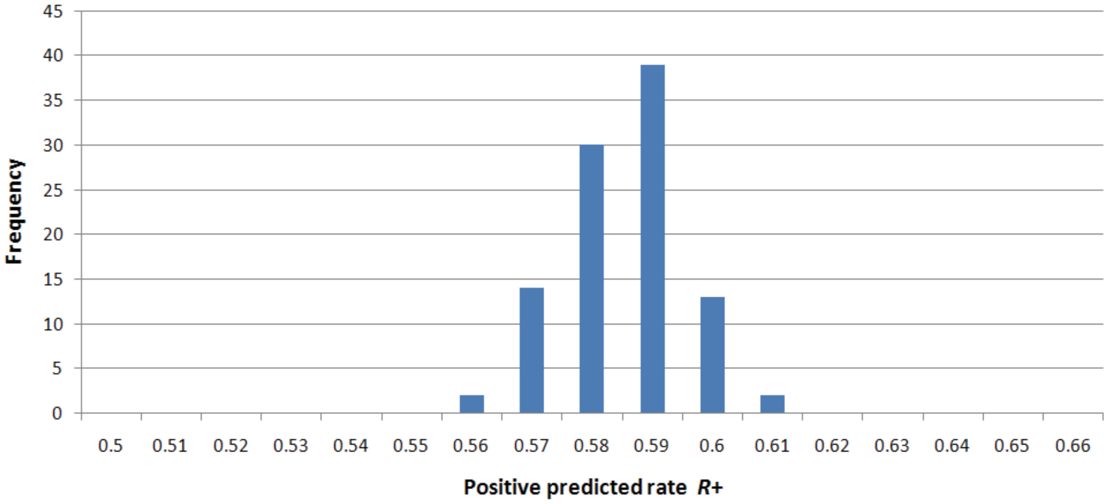


Figure 3.2 - The positive predicted rate is stable across many replications

A



B



We randomly selected 90% of replicate pairs, and randomly assigned status as discovery or replication, and then we calculated the positive predicted rate for "any 1 of 3 programs" method. We repeated the process 100 times. The positive predicted rate fluctuates in a narrow range for duplications (A) and deletions (B).

Chapter 4: CNVs and Alcohol Dependence

* This chapter is adapted from: Lin P et al. Copy number variations in 6q14.1 and 5q13.2 are associated with alcohol dependence. In Preparation.

ABSTRACT

Excessive alcohol use is the third leading cause of preventable death and is highly correlated with alcohol dependence, a heritable phenotype. Many genetic factors for alcohol dependence have been found, but several remain unknown. In search of additional genetic factors, we examined the association between DSM-IV alcohol dependence and all common copy number variations (CNV) with good reliability in the Study of Addiction: Genetics and Environment (SAGE).

All participants in SAGE were interviewed using the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA), as a part of three contributing studies. 2,610 non-Hispanic European American samples were genotyped on the Illumina Human 1M array. We performed CNV calling by CNVpartition, PennCNV and QuantiSNP and only CNVs identified by all three software programs were examined. Association was conducted with the CNV (as a deletion/duplication) as well as with probes in the CNV region. Quantitative polymerase chain reaction (qPCR) was used to validate the CNVs in the laboratory.

CNVs in 6q14.1 (P= 1.04 x10^{-6}) and 5q13.2 (P= 3.37 x10^{-4}) are significantly associated with alcohol dependence after adjusting multiple tests. On chromosome 5q13.2 there were multiple candidate genes previously associated with various neurological disorders. The region on chromosome 6q14.1 is a gene desert that has been associated with mental retardation, and language delay. The CNV in 5q13.2 was validated whereas only a component of the CNV on 6q14.1 was validated by qPCR.

This is the first study to show an association between DSM-IV alcohol dependence and common CNVs. CNVs in regions previously associated with neurological disorders may be associated with alcohol dependence.

INTRODUCTION

During 2001-2005, excessive alcohol use contributed to about 79,000 deaths and 2.3 million years of potential life lost in the United States [73] . Excessive alcohol consumption, the third leading cause of preventable death in the United States, can cause damage to the central and peripheral nervous system, and to nearly every organ system in the body [74, 75]. It is also strongly correlated with alcohol dependence, a serious psychiatric disorder that affects about 12% of American adults across their lifetime [76]. Alcohol dependence is a common, complex disease characterized by compulsive and uncontrolled alcohol consumption despite its negative effects on the drinker's health, relationships, and social standing.

There is robust evidence for heritable influences on the liability to alcohol dependence [77]. Siblings of alcohol dependent individuals have a 3-8 fold increased risk of developing alcoholism [78] with twin studies revealing the heritability of alcohol dependence to be ≈50% [79-81]. Given its serious public health impact [82] and the strong evidence for its biological underpinnings, numerous linkage and association studies have been targeted at gene identification for alcohol dependence [78, 83-86]. Recently, several genome wide association studies (GWAS) queried the genome for association [83, 87-89]. Results surpassed genome-wide significance in one study of early-onset male alcoholics [87], but across the multiple efforts, effect sizes were small and did not replicate. This has generated considerable interest in the examination of other possible contributors to the "missing heritability" for alcohol dependence. One such contributor is copy number variations (CNVs).

CNVs are duplications or deletions of a particular segment of an individual's genome and reflect inherent structural instability in the architecture of the genome. They are prevalent forms of common genetic variation and can have a substantial influence on gene expression levels [12]. For instance, Mendelian disorders such as Williams-Beuren Syndrome (due to a deletion at chromosome region 7q11.23) and Charcot-Marie-Tooth neuropathy Type 1A (caused by duplication at chromosome region 17p11.2 [90, 91] are attributable to CNVs. Despite the deleterious effects of CNVs and their links to disease, few studies have examined CNVs in the context of psychiatric illness, particularly alcohol dependence. This is primarily due to the inherent challenges involved in identification of what constitutes a CNV. While traditional methods of CNV identification involve laboratory-based experiments, they can also be identified (or "called") using GWAS data where a series of single nucleotide polymorphisms (SNPs) or "intensity" probes are interrogated for their occurrence in a state other than the expected disomic (i.e. 2 copy) state. Typically, the intensity of the probe signal that is expected when two copies of the probe are present is compared with the observed intensity, which is expected to be enhanced for duplications, or suppressed for deletions. These probes are routinely included in GWAS chips and thus, as GWAS technology became more accessible, there was an up-swell in CNV identification efforts. However, this method of CNV calling from GWAS microarrays can be associated with relatively high error rates. For instance, in a previous study, we demonstrated the relatively modest concordance in CNV detection using three widely utilized software packages with varying algorithms. In that study, we implemented statistical measures that enhance the reliability of the detected CNVs using multiple algorithms and further, validated the CNVs identified using statistical programs by quantitative Polymerase Chain Reaction (qPCR) in the laboratory [29].

Other challenges of CNV detection include (a) size of the CNV, with smaller CNVs (<10 kb) being harder to detect, (b) number of CNV probes in the region of the CNV, with fewer probes resulting in greater noise, (c) the general quality of the data (including artifacts in the SNP data) and genomic waves (intensity variations in normalized GWAS data), (d) ethnic variations and (e) source of the sample that was genotyped – for instance, it now well known that deletions and duplications can arise in DNA drawn from cell lines (i.e. extracting cells from a DNA source and maintaining them in laboratory cultures to enhance longevity) and, thus CNV detection using cell cultures requires caution. Yet, if attention is paid to these challenges, CNVs represent a unique route for enquiry into the genetic architecture of alcohol dependence.

There continues to be a great deal of progress in statistical methods for CNV detection. In tandem, there is growing excitement about the association between these CNVs and human behavior and the extent to which these intriguing variations in the human genome may contribute to that elusive “missing heritability” in complex behavioral phenotypes and psychiatric illness. While there has been some promise in studies of autism, and intellectual disabilities [92, 93], as well as schizophrenia and bipolar disorder [94, 95], research on CNVs in studies of addiction, particularly alcohol dependence, is lacking. In this study, we examine the CNVs for DSM-IV alcohol dependence in a large sample of European-American subjects.

MATERIALS AND METHODS

Samples

Data were drawn from the Study of Addiction: Genetics and Environment (SAGE) [68]. SAGE is one study of the Gene Environment Association Studies (GENEVA) project [58]. Cases and controls for the SAGE sample were drawn from 3 contributing projects: the Collaborative Study on the Genetics of Alcoholism (COGA), the Collaborative Study on the Genetics of Nicotine Dependence (COGEND) and the Family Study of Cocaine Dependence (FSCD). While the contributing studies originally ascertained subjects for alcohol dependence (COGA), nicotine dependence (COGEND: based on an FTND score of 4 or greater in current smokers, controls being smokers) and for cocaine dependence (FSCD), the subset of cases selected for genotyping in SAGE were uniformly defined as those meeting criteria for DSM-IV alcohol dependence (N=1899) while controls (N=1946) were individuals who reported drinking alcohol but not meeting criteria, during their lifetime, for alcohol dependence. Of these, 1,186 cases and 1,397 controls are of self-reported non-Hispanic European-American descent. All participants agreed to share their DNA and phenotypic information for research purposes and provided written informed consent following instructions from institutional review boards at all data collection sites.

Measures

A lifetime diagnosis of DSM-IV alcohol dependence was determined via self-reported interview information collected using the Semi-Structured Assessment for the Genetics of

Alcoholism (SSAGA). Controls were individuals who had drunk alcohol atleast once in their lifetime but did not meet criteria for alcohol dependence.

Genotyping and quality control

The Center for Inherited Disease Research (CIDR) at Johns Hopkins University genotyped all samples on the Illumina Human 1M array. An extensive data cleaning effort had been made to ensure data quality. These procedures included, but not limited to, using HapMap controls, detection of gender mis-annotation and chromosomal anomalies, cryptic relatedness, population structure, batch effects, Mendelian error detection, and duplication error detection. A detailed description of data cleaning effort is described elsewhere [23, 83].

CNV calling

The Illumina 1M array has a total of 1,072,820 probes, predominantly indexed by polymorphic SNPs. 23,812 of these probes are non-SNP "intensity-only" markers for CNV detection. All of the 1,072,820 probes were used for the CNV analyses. Three common programs were used to call CNVs: CNVPartition, PennCNV [28], and QuantiSNP[61]. Genomic waves were also adjusted when we called CNVs by PennCNV and QuantiSNP [60]. Both PennCNV and QuantiSNP report a metric score for quality control purposes. As recommended by QuantiSNP documentation, we removed all CNV calls that had Log Bayes Factor (LBF) less than 10, as well as poor quality samples based on quality control measures for CNV analysis, following the approaches described in our previous work [29]. In total, we genotyped 2,583 non-Hispanic European American samples in SAGE and among them 95 samples failed to pass quality controls for CNV analysis.

Comparative statistics

The CNV calls from different programs were compared against each other. In our previous work, we have demonstrated that a CNV that is confirmed by all three CNV calling programs has a higher reproducibility rate, and thus, a higher reliability [29]. Therefore, we required that in this study, only CNVs detected by all three programs would be studied.

Association analysis

Logistic regressions were performed on all CNV regions. After identifying potential regions, individual dummy variables for duplications and deletions were created to dissect the association signal with DSM-IV alcohol dependence. Several covariates were included in the model based on the previous GWAS of these data [83], including sex, age, and two principal components indexing continuous ancestral genetic variation. We also included a dummy variable to indicate the source of DNA (cell line versus whole blood). In addition, we ensured that these potentially confounding variables were not directly associated with the identified CNVs.

CNVs with different starting and ending point

Even when all three programs detect a CNV, they often report different starting and ending points for the same CNV segment, which leads to computational challenges in combining CNV reports. There is a lack of consensus in the research community regarding this issue and therefore, in addition to studying the CNV as a deletion or duplication, we adopted an additional straightforward approach for association. First, SNP probes and intensity-only probes were used to detect CNVs by multiple programs. Second, a change of copy number at a particular probe was considered detected when all CNV programs reported CNV segments that cover the probe. Third, association between alcohol dependence and each probe (assigned the same copy number as the CNV) was examined. For instance, if a CNV (duplication or deletion) was detected in region X, using probes (SNP or intensity probes) A, B C, D, E, F and G by three different programs, then the results from the three programs for each probe were compared against each other (Figure 4.1). If agreement was reached among three programs, then the CNV for these probes (Probe D and E) were confirmed and would be used in the following analysis. If there was disagreement among the three programs, then a missing value was assigned to these probes (Probe B, C, and F).

Validation

CNVs identified by 3 independent programs were validated in subjects carrying the variant with quantitative PCR (qPCR). We selected a TaqMan CNV probe in the target region. The probe was predesigned by Applied Biosystems (Applied Biosytems, Foster City, CA, USA). Genomic DNA was analyzed with real-time PCR using an ABI-7900 Fast PCR system. Each real-time PCR run included within-plate duplicates. Correction for sample-to-sample variation was done by simultaneously amplifying a standard CNV reference assay, RNAse P. Real-time data were analyzed using the comparative C_t

method [96]. The C_t values of each sample were normalized with the C_t value for the RNAse P assay. Only the samples with a standard error <0.15 were analyzed. Copy numbers were calculated using ABI CopyCaller™ Software v1.0.

RESULTS

Socio-demographic characteristics

Of the 2,583 non-Hispanic European American samples from SAGE, 95 failed to pass quality control for CNV analysis, leaving 1,140 cases and 1,348 controls. The mean ages among subjects with alcohol dependence was 38.2 [SD=10.0], and for controls was 39.0 [SD=9.5]. Sixty percent of the cases and 29.2% of the controls were male. As shown in Table 4.1, cases were more likely to be dependent on nicotine and illicit drugs, including nicotine, cocaine, and marijuana. They were also more likely to meet criteria for a lifetime history of conduct disorder and major depression.

Alcohol history

Cases also reported an earlier age of heavy and regular alcohol use, and, by definition, reported more alcohol symptoms (Table 4.2).

CNV detection

Of the samples that passed quality control, we identified 1,139 CNV regions with length greater than 50 kb and number of probes not less than 10 [29]. Among them, only 141 CNV regions have frequency higher than 1%. All of these CNV regions had previously been documented in the database of genetic variants [97]. Thus, after adjusting for multiple tests, our significance threshold for association analyses is $0.05/141=3.54 \text{ x}10^{-4}$.

Association between CNVs and Alcohol Dependence

Two CNV regions were significantly associated with alcohol dependence (Table 4.3): chromosome 6q14.1 (OR=2.86, *P*= 1.04 x10^{-6}, n=121 subjects with the duplication) and chromosome 5q13.2 (OR=1.99, *P*= 3.37 x10 $^{-4}$, n=59 subjects with the duplication, and n=58 subjects with deletions). The P values for each probe in these two regions are listed in Tables 4.4 & 4.5.

Validation using qPCR.

For the CNV at 5q13.2, over 97% of these CNVs were confirmed as true CNVs using qPCR. However, for 6q14.1, while all deletions were confirmed, none of the duplications were reproduced via qPCR. This suggests that the result for 6q14.1 should be viewed with caution.

Relationship among Personality, Alcohol Dependence, and CNVs

We noticed that alcohol dependence is significantly associated with agreeableness (*P*=1.04 x10^{-20}), conscientiousness (*P*=3.93 x10^{-22}), extraversion (*P*=1.15 x10^{-12}) and neuroticism (*P*=6.95 x10^{-40}). . We also found an exceptional P value (*P*=4.8 x10^{-5}) for conscientiousness in Chr5: 68,921,426 - 70,412,247, but not for the other four factors. This observation drives us to hypothesize that Chr5: 68,921,426 - 70,412,247 increase the risk of alcohol dependence by lowering conscientiousness, or more specifically self-discipline (Table 4.6).

DISCUSSION

These analyses evaluated the association between CNVs and alcohol dependence among a relatively large sample of alcohol-dependent cases and non-dependent alcohol exposed controls. We found two regions significantly associated with alcohol dependence: Chr5: 69,916,523- 70,373,564 and chr6:79,034,386-79,090,197. To our knowledge, this is the first study to connect common CNVs and alcohol dependence.

The identified chromosomal regions have been previously associated with several neurological and other disorders. Chr5: 69,916,523- 70,373,564 covers several genes, including *SMA4, SERF1, SERF1B, SMN2, SMA3, NAIP, GTF2H2, GTF2H2D* and the downstream *OCLN*. Among them, *SMA3, SMA4* and *SMN2* are known to be associated spinal muscular atrophy [98, 99]. Recent research shows that the genes in this region have a function in the nervous system [100], including *OCLN,* another candidate in this region, which is an integral membrane protein that is required for cytokine-induced regulation of the tight junction paracellular permeability barrier. Mutations in this gene are thought to be a cause of pseudo-*TORCH* syndrome, an autosomal recessive neurologic disorder that mimics the clinical characteristics (e.g. microcephaly, seizures, spasticity) attributable to congenital infections due to **T**oxoplasmosis, **O**ther Agents, **R**ubella, **C**ytomegalovirus or **H**erpes Simplex [101]. While the CNV in Chr6:79,034,386-79,090,197 is located in a gene desert, there is evidence that suggests a link between chromosome region 6q14.1 and mild mental retardation, language delay, and minor dysmorphisms [102, 103]. Also, it is hypothesized that non-coding intergenic regions, such as this, may contain regulatory elements, such as enhancers and chromosome scaffold components that are capable of changing gene expression.

The observation that conscientiousness is associated with Chr5: 68,921,426 - 70,412,247 drives us to hypothesize that CNVs at 5q13.2 increase the risk of alcohol dependence by lowering conscientiousness. But it is possible that this link can be contributed to confounding effects. We believed it is not likely the case, because agreeableness, extraversion and neuroticism, all of them are associated with the risk of alcohol dependence but none of them are linked to this CNV.

We restricted our association tests to non-rare (>1%) CNVs for two reasons. First, the traditional genome wide association study design has little power to detect rare genetic variants (<1%), and the case control study design of this project cannot provide enough power to detect rare CNVs. Second, accuracy of CNV detection diminishes with decreasing frequency.

In addition we required that all CNVs be reproduced by 3 independent programs, a step that increases confidence in the results but that raises the potential problem of the same CNV region being detected with different starting and ending points, which results in uncertainty on how to combine these different CNV calls. In order to avoid this controversy, we adopted an intuitive method where we tested each genetic marker instead of a particular CNV segment. A CNV status is assigned to a particular genetic marker when all programs report a CNV that covers this probe (Tables 4.4 & 4.5). We validated our findings by qPCR – while the region on chromosome 5q13.2 replicated, the duplication on 6q14.1 did not. This underscores the considerable importance of experimental validation of CNVs identified using software algorithms.

Our finding of the association between these CNVs and alcohol dependence is encouraging because it identifies regions previously associated with neurological disorders, however these findings will require replication. Nonetheless, our study is amongst the first to examine the role of CNVs in the etiology of alcohol dependence. This reflects the exciting phase of the post-GWAS genomics era where the quest to articulate the genetic architecture of serious psychiatric problems like alcohol dependence moves beyond single SNP association to new frontiers, such as CNVs and rare variants.

Table 4.1 - Socio-demographic characteristics

	Total No.	Gender (male %)	Ave. Age	Comorbidity with other addictions*			
				Nicotine dependence	Cocaine dependence	Marijuana dependence	Opiates dependence
Alcohol dependence	1,140	60.0%	38.2	70.4%	38.3%	34.2%	14.7%
No dependence	1,348	29.8%	39.0	22.5%	0.0 %	0.0 %	0.0 %
Total	2,488	43.7%	38.7	44.5%	17.5%	15.6%	6.7 %

*All Psychiatric diagnoses were categorized by Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition.

Table 4.2 - Alcohol history

	Age of onset of regular drinking [a]	Age when first got drunk	Maximum drinks per day [b]	Maximum drinks per week	Number of alcohol symptoms endorsed
Alcohol dependence	17.8	15.2	30.1	54.7	5.3
No dependence	20.8	18.2	9.9	10.7	0.7
Total	19.3	16.7	19.1	27.5	2.8

[a] Regular drinking is defined as drinking once a month for 6 months or more

[b] Largest number of alcoholic drinks consumed in 24 hours

Table 4.3 - Associations of CNVs and alcohol dependence

Locus [a]	CNV type	CNV frequency			Beta [b]	P value [c]	Genes
			Cases	Controls			
Chr5: 68,921,426- 70,412,247	Copy number	Duplication	43 (4.97%)	16 (1.59%)	0.69	3.37 x10^{-4}	*SMA4, SERF1, SERF1B,* *SMN2, SMA3, NAIP,* *GTF2H2, GTF2H2D, OCLN*[d]
		Deletion	22 (2.54%)	36 (3.57%)			
		Non-missing	865 (100%)	1008 (100%)			
chr6:79,034,386-79,090,197	Duplication	Duplication	83 (8.99%)	38 (3.49%)	1.05	1.04 x10^{-6}	Gene desert
		Deletion	448 (48.54%)	542 (49.77%)			
		Non-missing	923 (100%)	1089 (100%)			

[a.] The starting and the ending point are defined by the probes whose P values have a clear deviation from the rest.

[b.c.] The P value and beta of this region is annotated by the most significant probe in this CNV.

[d.] *OCLN* is located less than 30 kb downstream of this region.

Table 4.4 - Associations of each probe and alcohol dependence in Chr5: 68,921,426-70,412,247

Chr	Position	RS ID or probe ID	Beta	Std Error	P value
5	68921426	cnvGap_CNV_9695.2p30	1.199816	0.384087	0.001785
5	68930217	cnvGap_CNV_9695.2p34	1.201797	0.384229	0.001761
5	68940434	cnvGap_CNV_9695.2p36	1.022875	0.354005	0.003859
5	68950807	cnvGap_CNV_9695.2p40	0.752159	0.317977	0.018008
5	68956063	rs2872744	0.791176	0.326499	0.015384
5	68977500	cnvGap_CNV_9695.2p52	0.731418	0.311746	0.018966
5	68989477	cnvGap_CNV_9695.2p56	0.708313	0.313653	0.023929
5	68999018	cnvGap_CNV_9695.2p60	0.639949	0.308121	0.037807
5	69011885	cnvGap_CNV_9695.2PP1	0.768489	0.300768	0.010616
5	69027097	cnvGap_CNV_9695.2p74	0.631037	0.291654	0.030491
5	69031745	cnvGap_CNV_9695.2p75	0.584503	0.269701	0.030218
5	69041250	cnvGap_CNV_9695.2p77	0.610001	0.258246	0.018172
5	69049740	cnvGap_CNV_9695.2p82	0.558177	0.250013	0.025576
5	69061359	cnvGap_CNV_9695.2p88	0.454047	0.242656	0.061324
5	69096757	cnvGap_CNV_9695.2p10	0.432948	0.23653	0.067188
5	69099351	cnvGap_CNV_9695.2p10	0.432948	0.23653	0.067188
5	69111519	cnvGap_CNV_9695.2p11	0.41567	0.2376	0.080214
5	69118987	cnvGap_CNV_9695.2p11	0.415922	0.237575	0.079999
5	69130746	cnvGap_CNV_9695.2PP2	0.368784	0.233744	0.114628
5	69145115	cnvGap_CNV_9695.2PP2	0.334341	0.23186	0.149303
5	69150402	cnvGap_CNV_9695.2p13	0.353716	0.230783	0.125355
5	69159243	cnvGap_CNV_9695.2p13	0.367197	0.229883	0.110194
5	69172823	cnvGap_CNV_9695.2p14	0.32113	0.231953	0.166217
5	69183643	cnvGap_CNV_9695.2p14	0.321074	0.231953	0.166291
5	69190743	cnvGap_CNV_9695.2p14	0.321015	0.231974	0.166407
5	69199857	cnvGap_CNV_9695.2p14	0.312488	0.226807	0.168275
5	69210590	cnvGap_CNV_9695.2p15	0.321634	0.228144	0.158604
5	69221673	cnvGap_CNV_9695.2p16	0.451626	0.238605	0.058388
5	69233077	cnvGap_CNV_9695.2p16	0.452504	0.238721	0.058022
5	69247965	cnvGap_CNV_9695.2p17	0.464867	0.23734	0.050154
5	69251715	cnvGap_CNV_9695.2p17	0.503588	0.235621	0.032575
5	69262959	cnvGap_CNV_9695.2p18	0.503588	0.235621	0.032575
5	69280018	cnvGap_CNV_9695.3p28	0.509937	0.237742	0.031959
5	69289210	cnvGap_CNV_9695.3PP1	0.48309	0.239443	0.043637
5	69304294	cnvGap_CNV_9695.3p11	0.534143	0.248472	0.031578
5	69315096	cnvGap_CNV_9695.3p18	0.520903	0.249389	0.036733
5	69326897	cnvGap_CNV_9695.3p26	0.674162	0.253549	0.00784
5	69331520	cnvGap_CNV_9695.3p30	0.674162	0.253549	0.00784
5	69345138	cnvGap_CNV_9695.3p39	0.722887	0.249519	0.003766
5	69359352	cnvGap_CNV_9695.3p46	0.723277	0.249613	0.00376
5	69369617	cnvGap_CNV_9695.3p50	0.723277	0.249613	0.00376
5	69390807	cnvGap_CNV_9695.3p54	0.756269	0.249072	0.002395
5	69400684	cnvGap_CNV_9695.3p58	0.756269	0.249072	0.002395

5	69409672	cnvGap_CNV_9695.3p61	0.741535	0.249818	0.002994
5	69428601	cnvGap_CNV_9695.3p65	0.7834	0.260774	0.002663
5	69430212	cnvGap_CNV_9695.3p66	0.7834	0.260774	0.002663
5	69439738	cnvGap_CNV_9695.3p69	0.809719	0.26522	0.002266
5	69455867	cnvGap_CNV_9695.3p74	0.667099	0.280284	0.017309
5	69459806	cnvGap_CNV_9695.3p76	0.667099	0.280284	0.017309
5	69476562	cnvGap_CNV_9695.3p87	0.66676	0.280188	0.017327
5	69510771	cnvGap_CNV_9695.4p77	0.714841	0.286388	0.012558
5	69522322	cnvGap_CNV_9695.4p11	0.713733	0.286203	0.012638
5	69551174	cnvGap_CNV_9695.4p24	0.666819	0.270958	0.013856
5	69566693	cnvGap_CNV_9695.4p27	0.683789	0.269716	0.011238
5	69570933	cnvGap_CNV_9695.4p27	0.701269	0.268428	0.008988
5	69586554	cnvGap_CNV_9695.4p35	0.538879	0.235836	0.022314
5	69590946	cnvGap_CNV_9695.4p39	0.538879	0.235836	0.022314
5	69606832	cnvGap_CNV_9695.4p44	0.563723	0.2313	0.014802
5	69611483	cnvGap_CNV_9695.4p46	0.524986	0.227573	0.021061
5	69629469	cnvGap_CNV_9695.4p53	0.485993	0.218098	0.025859
5	69636067	cnvGap_CNV_9695.4p56	0.485993	0.218098	0.025859
5	69641101	cnvGap_CNV_9695.4p59	0.485993	0.218098	0.025859
5	69651103	cnvGap_CNV_9695.4p63	0.486	0.218117	0.025869
5	69662828	cnvGap_CNV_9695.4p67	0.509414	0.21455	0.017581
5	69670671	cnvGap_CNV_9695.4PP1	0.460186	0.212049	0.029992
5	69685334	cnvGap_CNV_9695.4p79	0.497897	0.210767	0.018161
5	69692210	cnvGap_CNV_9695.4p83	0.448661	0.208357	0.031293
5	69706422	cnvGap_CNV_9695.4p92	0.4288	0.202809	0.034489
5	69713288	cnvGap_CNV_9695.4p95	0.42941	0.202791	0.034217
5	69724106	cnvGap_CNV_9695.4p10	0.437324	0.20042	0.029106
5	69733571	cnvGap_CNV_9695.4PP2	0.371912	0.191615	0.052267
5	69741366	cnvGap_CNV_9695.4p10	0.386067	0.194716	0.047399
5	69751404	cnvGap_CNV_9695.4p11	0.450406	0.20269	0.026274
5	69768625	cnvGap_CNV_9695.4p11	0.528363	0.213867	0.013492
5	69773055	cnvGap_CNV_9695.4p11	0.559303	0.215967	0.009604
5	69788867	cnvGap_CNV_9695.4p12	0.564798	0.233271	0.015469
5	69791981	cnvGap_CNV_9695.4p12	0.58497	0.238321	0.014106
5	69809114	cnvGap_CNV_9695.4p13	0.801829	0.277339	0.003838
5	69810251	cnvGap_CNV_9695.4p13	0.801829	0.277339	0.003838
5	69823637	cnvGap_CNV_9695.4p13	0.837268	0.275872	0.002405
5	69831300	cnvGap_CNV_9695.4p14	0.837268	0.275872	0.002405
5	69841867	cnvGap_CNV_9695.4p14	0.87538	0.279896	0.001763
5	69851260	cnvGap_CNV_9695.4PP3	0.875937	0.27989	0.001751
5	69865823	cnvGap_CNV_9695.4p15	0.862827	0.280838	0.002124
5	69876032	cnvGap_CNV_9695.4p15	0.895296	0.279328	0.00135
5	69882833	cnvGap_CNV_9695.4PP3	0.895296	0.279328	0.00135
5	69893882	cnvGap_CNV_9695.4PP3	0.896231	0.279376	0.001337
5	69902168	cnvGap_CNV_9695.4p16	0.839408	0.275278	0.002294
5	69916523	cnvGap_CNV_9695.4p16	0.929826	0.270602	0.00059
5	69922377	cnvGap_CNV_9695.4p17	0.929826	0.270602	0.00059
5	69936538	cnvGap_CNV_9695.4PP3	0.844825	0.267901	0.001613

5	69943485	cnvGap_CNV_9695.4p18	0.858466	0.266437	0.001273
5	69954494	cnvGap_CNV_9695.4p18	0.858466	0.266437	0.001273
5	69973658	cnvGap_CNV_9695.4p19	0.756466	0.265982	0.004454
5	69980978	cnvGap_CNV_9695.4p19	0.756384	0.265997	0.004461
5	69997937	cnvGap_CNV_9695.4p20	0.724284	0.262874	0.005865
5	70002621	cnvGap_CNV_9695.4p20	0.731238	0.267667	0.006297
5	70022132	cnvGap_CNV_9695.4p21	0.759332	0.272195	0.005276
5	70033323	cnvGap_CNV_9695.4p22	0.759684	0.272224	0.00526
5	70054944	cnvGap_CNV_9695.4PP5	0.764576	0.279212	0.006175
5	70065114	cnvGap_CNV_9695.4PP5	0.765482	0.279118	0.006097
5	70078585	cnvGap_CNV_9695.4p23	0.634303	0.286028	0.02658
5	70093265	cnvGap_CNV_9695.4p24	0.601258	0.293154	0.040267
5	70104506	cnvGap_CNV_9695.4p24	0.416646	0.324074	0.198565
5	70132660	cnvGap_CNV_9695.4p26	0.742946	0.34966	0.033606
5	70145118	cnvGap_CNV_9695.4p26	0.550989	0.321841	0.086898
5	70160109	cnvGap_CNV_9695.4p27	0.496704	0.274422	0.070295
5	70161507	cnvGap_CNV_9695.4p27	0.496537	0.274459	0.070428
5	70179802	cnvGap_CNV_9695.4p28	0.515022	0.272403	0.058669
5	70181307	cnvGap_CNV_9695.4p28	0.515022	0.272403	0.058669
5	70202676	cnvGap_CNV_9695.4p29	0.484197	0.260942	0.063515
5	70212521	cnvGap_CNV_9695.4p30	0.495478	0.260042	0.056731
5	70222045	cnvGap_CNV_9695.4p30	0.494999	0.260111	0.057037
5	70232726	cnvGap_CNV_9695.4p31	0.493616	0.260006	0.057633
5	70245258	cnvGap_CNV_9695.4p32	0.512855	0.262521	0.050751
5	70256087	cnvGap_CNV_9695.4p32	0.513487	0.262557	0.050498
5	70261872	cnvGap_CNV_9695.4p32	0.513487	0.262557	0.050498
5	70274080	cnvGap_CNV_9695.4p32	0.603484	0.249562	0.015599
5	70300876	cnvGap_CNV_9695.5p24	0.521426	0.217213	0.016372
5	70304036	cnvGap_CNV_9695.5p42	0.497972	0.213587	0.019729
5	70309633	rs575909	0.472588	0.197371	0.016647
5	70311476	cnvGap_CNV_9695.5p73	0.46997	0.196627	0.016841
5	70327006	cnvGap_CNV_9695.5p12	0.512376	0.169714	0.002536
5	70335524	cnvGap_CNV_9695.5p15	0.553221	0.172349	0.001328
5	70341309	rs28751879	0.5665	0.168459	0.000771
5	70341452	rs28538463	0.5665	0.168459	0.000771
5	70342434	rs28447466	0.566707	0.168305	0.000759
5	70343142	rs36065930	0.567762	0.168294	0.000742
5	70343220	rs4976210	0.566641	0.171011	0.000921
5	70351828	cnvGap_CNV_9695.6p46	0.620636	0.18749	0.000932
5	70368925	cnvGap_CNV_9695.6p10	0.650221	0.189743	0.000611
5*	70373564	cnvGap_CNV_9695.6p11	0.693727	0.193509	0.000337
5	70391854	cnvGap_CNV_9695.6p16	0.55771	0.194912	0.004218
5	70405153	cnvGap_CNV_9695.6p22	0.576782	0.202344	0.004365

* The probe with the most significant P value in this CNV region.

Table 4.5 - Associations of each probe and alcohol dependence in Chr6:79,034,386-79,090,197

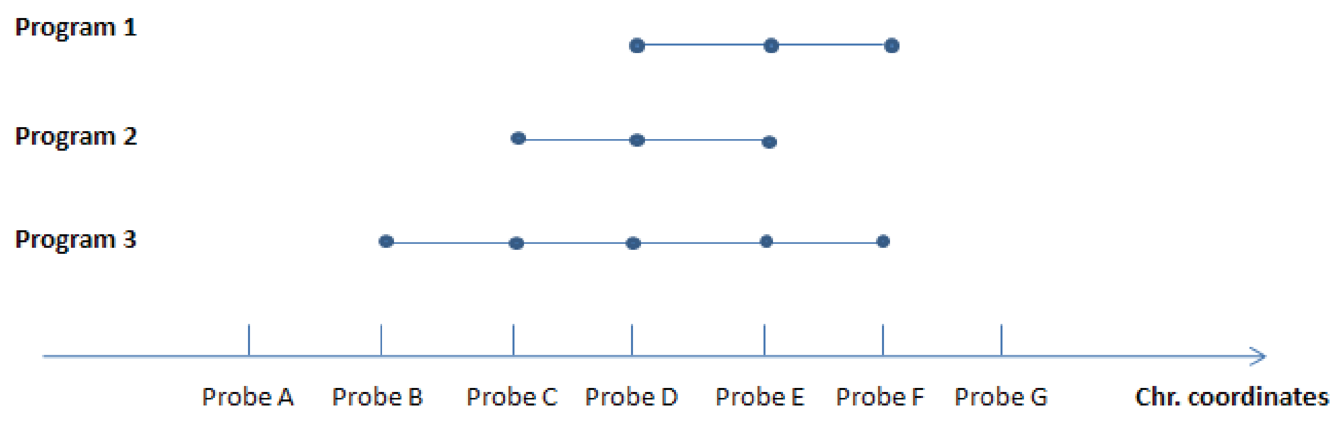
Chr	Position	RS ID or probe ID	Beta	Std Error	P value
6	79034386	rs818258	1.208651	0.25355	1.87065E-06
6	79036117	rs818262	1.189742	0.249361	1.83165E-06
6	79039487	rs818313	1.095437	0.232592	2.48088E-06
6	79042356	rs818310	1.07382	0.225953	2.01012E-06
6	79052979	rs11964123	1.017069	0.219191	3.48217E-06
6	79056617	rs1093580	1.053288	0.217916	1.34181E-06
6	79056822	rs818301	1.053288	0.217916	1.34181E-06
6	79059458	rs6932920	1.053288	0.217916	1.34181E-06
6	79063712	rs6918807	1.004236	0.212568	2.30915E-06
6	79065940	rs6911209	1.003881	0.212539	2.32078E-06
6	79065999	rs6931912	1.003881	0.212539	2.32078E-06
6	79067895	rs9361392	1.001207	0.212502	2.4589E-06
6	79069278	rs818295	1.005926	0.21264	2.23806E-06
6	79069674	rs9448350	1.005926	0.21264	2.23806E-06
6	79070425	rs964927	1.005926	0.21264	2.23806E-06
6	79075016	rs7749022	1.01238	0.212543	1.90567E-06
6	79076024	rs9448356	1.046794	0.214308	1.03678E-06
6	79076473	rs9448357	1.046794	0.214308	1.03678E-06
6*	79077158	rs818290	1.046827	0.214304	1.03542E-06
6*	79077999	rs7774454	1.046827	0.214304	1.03542E-06
6	79078423	rs818288	1.042325	0.214243	1.14368E-06
6	79081009	rs16889854	1.00384	0.224986	8.12803E-06
6	79082584	rs16889859	0.968943	0.226665	1.91323E-05
6	79083049	rs818285	0.969383	0.226686	1.90001E-05
6	79083083	rs818284	0.965082	0.226739	2.07794E-05
6	79083326	rs9443550	0.974056	0.226628	1.72314E-05
6	79086086	rs9448361	0.976535	0.226689	1.64874E-05
6	79088461	rs818280	0.964914	0.227086	2.1462E-05
6	79090197	rs7773124	0.951371	0.228169	3.05157E-05

* The probes with the most significant P value in this CNV region.

Table 4.6 - The relationship among the FFM dimensions, alcohol dependence and CNVs

	Alcohol dependence		Chr5: 68,921,426 - 70,412,247		Chr6:79,034,386 -79,090,197	
	P value	Beta	P value	Beta	P value	Beta
Agreeableness	1.04 x10^{-20}	-0.0975	0.1818	-0.94	**0.0020**	-2.38
Conscientiousness	3.93 x10^{-22}	-0.0913	**4.8 x10^{-5}**	-3.19	0.0226	-1.85
Extraversion	1.15 x10^{-12}	-0.0622	0.2672	-0.88	0.1968	-1.06
Openness	0.7075	-0.0034	0.0707	1.37	0.5652	-0.45
Neuroticism	6.95 x10^{-40}	0.0997	0.1325	1.59	**0.0010**	3.55

Figure 4.1 - Comparing CNV calls from different programs



A straightforward approach was adopted to treat inconsistent CNV calls from different programs. Three different programs were used to call CNVs. Figure 4.1 shows that for the same subject, three programs detected a CNV (duplication or deletion) in region X by using probes (SNP or intensity probes) A, B C, D, E, F and G. The results from the three programs then were compared against each other. If agreement was reached among three programs, then the CNV calls for these probes (Probe D and E) were confirmed and used in the following analysis. If there was a disagreement among the three programs, a missing value was assigned (Probe B, C, and F). Other probes (Probe A and G) did not have copy number changes.

Chapter 5: Challenges and directions: an analysis of Genetic Analysis Workshop 17 data by collapsing rare variants within family data

* This chapter is adapted from: Lin P et al. (2011) Challenges and directions: an analysis of Genetic Analysis Workshop 17 data by collapsing rare variants within family data. BMC proceedings. In Press.

ABSTRACT

Recent studies suggest that the traditional case-control study design does not have sufficient power to discover rare risk variants. Two different methods—collapsing and family data—are suggested as alternatives for discovering these rare variants. Compared with common variants, rare variants have unique characteristics. In this paper, we assess the distribution of rare variants in family data. We notice that a large number of rare variants exist only in one or two families and that the association result is largely shaped by those families. Therefore we explore the possibility of integrating both the collapsing method and the family data method. This combinational approach offers a potential power boost for certain causal genes, including *VEGFA*, *VEGFC*, *SIRT1*, *SREBF1*, *PIK3R3*, *VLDLR*, *PLAT*, and *FLT4*, and thus deserves further investigation.

BACKGROUND

Genome-wide association studies have accelerated the discovery of genetic variants that cause disease. Thus far, nearly 600 genome-wide association studies have examined about 150 distinct diseases or traits, and more than 800 SNPs associated with these diseases or traits have been identified[104]. Recent studies have suggested that rare variants contribute to common diseases, but the case-control study design does not have sufficient power to discover rare causal variants.

Two common approaches are used to increase the power to detect rare variants. One method is to collapse rare variants on the basis of predetermined criteria. By grouping risk variants together, the frequency of rare risk variants can be increased in the data set. Extensive research on collapsing has been done for population-based data [105]. Another approach is to examine family data. The potential advantage of family data is that a particular rare variant found in an affected individual is more common in that individual's family than in subjects randomly sampled in the population.

The Genetic Analysis Workshop 17 (GAW17) is a collaborative effort among researchers to improve our current understanding of genetic architecture. It provides simulated data based on real exon sequence data, and thus offers a unique and relatively realistic opportunity to evaluate statistical genetic methods that are relevant to current analytical problems. In For this workshop, we designed this a study to (1) test both the collapsing methodology and the family design in data sets generated with the same biological model, and (2) assess the power of combining these two approaches: (collapsing rare variants within family data). This study will help guide researchers to design and analyze future studies for the detection of rare genetic variants.

METHODS

Family-based association testing

To test genetic associations in family data, investigators need to address the correlation among family members. Several methods are available [31, 106, 107]. We accounted for correlated genotypes by using the modified quasi-likelihood score test (M_{QLS}) developed by Thornton and McPeek [106]. This method is implemented in the computer program M_{QLS}.

M_{QLS} is an improvement on the previous quasi-likelihood score test, W_{QLS}, developed by Bourgain et al. [108]. It accounts for the correlations among related individuals by using a defined kinship matrix and assigns optimal weights depending on the pedigree information, thus providing an efficient estimator of allele frequency under the null hypotheses. Interested researchers should refer to Thornton and McPeek's paper for more details [106].

Collapsing rare variants within family-based association testing

A causal gene can be shared by more than one or two families, although this gene can have different rare risk variants in those families. Traditional family-based association tests fail to combine signals from different rare variants. To address this issue, we proposed to collapse these rare variants. Many collapsing methods are available. Some methods simply account for the presence or absence of rare variants, whereas others assign an adjustable weight to different types of rare variants, based on biological function or minor allele frequency, and then calculate a final score for each gene [105].

Currently, there is no conclusive evidence to argue for or against a particular collapsing method. To generate data that can be analyzed by M_{QLS}, we created a gene indicator that collapses rare variants within the same gene. Similar to SNPs, the gene indicator is a dichotomous variable that indicates presence or absence of any rare variant within the region of interest, so it can be processed by the M_{QLS} program. A gene indicator variable *G* for the *n*th subject is defined as

$$G_n = \begin{cases} AB & \text{if any predefined rare variants exist in a particular gene,} \\ AA & \text{if no predefined rare variants exist in a particular gene.} \end{cases} \tag{1}$$

Although genotype *BB* can be defined when both alleles of a particular SNP are the rare alleles, the likelihood of this situation is small, because we are dealing with rare variants. We have developed a SAS macro to implement our method with the M_{QLS} program. The SAS macro is available to interested investigators.

Power analysis

A subset of genes that had sequence data available in the 1000 Genomes Project was included in this GAW17 project. GAW17 simulated the phenotype based on a predefined simulation model and generated 200 different phenotype files under the same model. Thus the 200 replicate phenotype files provide a unique opportunity to estimate power. We tested associations under different conditions and calculated the power of different approaches. Power is defined as the proportion of times that a particular test reaches the significance threshold.

RESULTS

Distribution of rare variants within family data

The GAW17 data set has 697 subjects (209 case subjects and 488 control subjects) from 8 families. A total of 24,487 SNPs were simulated for 3,205 genes. Fully informative identical-by-descent (IBD) scores were also provided for each gene.

We defined a SNP as rare if its minor allele frequency (MAF) in the population was less than 0.01. By this definition, in the GAW17 data there are 18,131 rare SNPs, 56.4% of which do not exist in the family data. According to the simulation model, 162 SNPs underlie the disease status. Among them, 145 are rare SNPs. Unfortunately, more than 70% of these rare SNPs do not exist in the family data. In addition, a large proportion (85%) of the remaining SNPs exist in only one or two families (Figure 5.1).

Moreover, many existing rare variants are not passed on in the family. Analysis of the family data shows that 30 of the 42 rare variants that exist in founders are not passed on to offspring. In fact, only 10 of the 42 rare SNPs (7% of all the causative rare SNPs) have an allele frequency (frequency in family data) greater than 0.01.

Family-based association test

Because 85% of the 36 rare SNPs found in families exist in only one or two families, it is expected that only one or two families can contribute to the final association result. Among the 145 rare SNPs that underlie the disease status, most signals exist in only one or two families. The distribution of signals is shown in Figure 5.2, and it matches the distribution of rare SNPs within families well.

In addition, combining families that have a particular risk allele with families that do not have the particular risk allele unintentionally diminishes the power. We compared the association result from all families and the association result from each family. Seventy-seven percent of rare causal SNPs have more significant *P*-values from one family than from all data analyzed together.

Collapsing rare variants within family-based association test

As we have shown, for a particular rare risk variant, only one or two families contribute to the signal, but one gene may have multiple risk variants, each of which may be possessed by different families. Cystic fibrosis transmembrane conductance regulator (*CFTR*) is a good example. Since *CFTR* was identified, more than 1,000 mutations have been found for cystic fibrosis [109]. And similar to *CFTR*, a causal gene may have multiple mutations, and different families may have different risk mutations within the same gene. Because these different mutations can be designated by a risk gene indicator, we believe that collapsing those different mutations to a gene indicator may provide an additional boost on power.

We tested collapsing within family data using the method described in the Methods section. One particular question we want to address here is whether there is any benefit to collapsing within families compared to collapsing in population-based data, which has been extensively researched.

We set our significance level to a loose level of $P < 0.05$ for power calculation and repeated our analysis in the 200 phenotype data sets. We collapsed all rare SNPs (MAF < 0.01) within genes. The SNP for *GCKR* has a MAF greater than 0.01 and thus was excluded from analysis. Among 35 available genes, 17 reached the significance threshold. The power for these genes is shown in Table 5.1. For comparison, we did similar analyses in the population data with two dummy variables to adjust for ancestry. From the table, we notice that family-based collapsing is more useful for certain genes.

Among those genes for which the family-based collapsing has power, we set our significance threshold to the stringent level of $0.05/3{,}205 = 1.56 \times 10^{-5}$. The power for *VEGFC* and *VEGFA* is 99% and 94.5%, respectively. Population-based collapsing, however, has no power to detect these two genes. Among the 200 phenotypes, the population-based collapsing reported a median *P*-value of 0.98 for *VEGFC* and 0.54 for *VEGFA*.

Another issue we want to address is whether there is any gain in power for collapsing within families compared to the family approach without collapsing. We tested each SNP using M_{QLS} within family data. The result is shown in Table 5.2. The comparison shows that collapsing may be useful for some variants and may be detrimental for some other variants. In fact, collapsing a causal variant with a noncausal variant will diminish power.

We found that *SIRT1* and *VLDLR* have a power drop, but for some other genes, such as *SREBF1*, *PIK3R3*, *PLAT*, and *FLT4*, there are considerable power gains. Further analysis shows that among those genes that have power gains by the family-based collapsing, many families that possess different risk variants have contributed to the signal.

DISCUSSION

Recent advances in genome-wide association studies have identified hundreds of common SNPs that are associated with different diseases, but collectively they can explain only a small fraction of variation. Many investigators believe that the missing heritability may be partly explained by the rare variants, which are difficult to discover in the common case-control study design. One reason that the existing study design does not have sufficient power is simply because these rare variants are rare. In general, for any statistical test, a certain number of subjects who possess this particular rare variant are required in order to obtain enough power. From this perspective, the family design and the collapsing approach, both of which are potential methods for discovering rare variants, aim to increase the presence or the frequency of the risk variant or haplotype in the data set. However, some challenges are associated with these two methods.

It is generally thought that because a rare mutation can be transmitted to offspring, family data may have more copies of rare mutations than can be found in population-based data. However, a large number of rare mutations that are possessed by founders are not passed on in the family data. Among 145 rare SNPs, only 10 have an allele frequency (frequency in family data) greater than 0.01. This may partly explain the general conclusion reached in the GAW17 meeting that family data are not particularly helpful for discovering rare risk variants.

In addition, collapsing should be used with caution. The assumption behind collapsing is that risk alleles tend to be rare. This assumption may be supported by evolution theory. If one new variant is generated by mutation and is beneficial, then this new variant will be favored by selection and therefore its frequency will increase over time. Similarly, malicious alleles are selected against, and therefore their frequency will decrease over time. Moreover, if a nonsynonymous mutation occurs at a conservative gene coding region, it is likely that the mutation will be malicious, because that is why the sequence is otherwise conservative. However, some neutral rare variants can exist in the population as a result of random mutation. Grouping a risk variant with a neutral variant may decrease the power, as we have shown in Table 5.1.

In GAW17, all risk variants are nonsynonymous SNPs. In Table 5.1, the power is lower when collapsing all rare variants than when collapsing only nonsynonymous SNPs. It is tempting to argue that we should collapse only nonsynonymous SNPs. In reality, however, synonymous SNPs may play a significant role in biological function, for example, alternative splice site, transcription factor binding site, or even chromatin structure protein binding site. Meanwhile, nonsynonymous SNPs may have no function at all. At the protein level, an amino acid change, which is usually the result of nonsynonymous SNPs, often fails to change the secondary structure and tertiary structure of a protein and therefore may have no impact on protein function. Although it is generally difficult to predict whether a synonymous SNP or a nonsynonymous SNP is biologically functional or not, we believe that the use of prediction algorithms for function will be helpful. Several function prediction algorithms are available, for example, SIFT and PolyPhen-2 [110, 111]. Unfortunately, all causal variants in the GAW17 simulation data were chosen based on PolyPhen and SIFT predictions of the likelihood that the variant would be deleterious. Thus the application of the function prediction algorithm to the GAW17 simulation data, which were generated using the same function prediction algorithm, may not be illuminating.

One purpose of this study is to cast new light on future study designs. We noticed that in family data, the association signals exist in only one or two families. We also noticed that combining these families with families that do not possess these risk variants unintentionally diminishes power. Therefore we argue that, given a limited sample size, a large pedigree may be more useful for discovering rare risk variants. Although many rare variants cannot be discovered, a large pedigree is still useful because at least some causal rare variants are more likely to be found.

In conjunction with association testing, linkage can identify regions of interest. Therefore regional sequencing can be done instead of whole genome sequencing. In addition, the selection of the most informative families or family members may further reduce the total genotyping cost. In addition, the use of extremes of a phenotypic distribution may provide tremendous information and reduce the required sample size [112].

In this study, we tested collapsing within family data, which combines the two widely proposed methods: the family design and the collapsing approach. The new combinational method provides considerable power gain for some genes. Although we noticed that the power gain is obtained at the cost of power for some other genes, this is still useful, especially if the alternative is that nothing can be found. As we have shown in this paper, this method can be useful for discovering novel variants associated with disease, and thus it merits further study.

CONCLUSIONS

Family data are believed to be one way to increase the presence of rare variants in the data set. But a large number of rare risk variants cannot be sampled in the family data. Even for existing rare risk variants, a large proportion of them are not passed on in the family. Many existing rare risk variants are seen in only one or two families, and the result from association is largely shaped by those families. To aggregate signals from different rare variants in different families, we integrated the collapsing method within the family data method. To our knowledge, this is the first attempt in the literature to do collapsing within family data. This combinational approach offers a promising power boost for certain causal genes and thus deserves further investigation.

Table 5.1 - Comparison of collapsing within family data and collapsing within population-based data

Power is calculated based on the threshold $P < 0.05$. Because of limited space, only those genes whose power is greater than 10% are shown.

						Family		Population	
Chromosome	Gene	Number of synonymous SNPs	Number of nonsynonymous SNPs	Total number of SNPs	Number of risk SNPs	Power (All) (%)[a]	Power (nonsynonymous) (%)[b]	Power (All) (%)[a]	Power (nonsynonymous) (%)[b]
6	*VEGFA*	2	2	4	1	100	100	13	10
4	*VEGFC*	0	1	1	1	100	100	0	0
10	*SIRT1*	9	14	23	9	19	47	7	8
17	*SREBF1*	5	16	21	10	19	36	17	18
1	*PIK3R3*	3	2	5	1	11	1	2	2
9	*VLDLR*	8	15	23	8	10	4	10	9
8	*PLAT*	14	11	25	8	8	34	6	7
5	*FLT4*	3	5	8	2	6	13	15	15
4	*KDR*	5	9	14	8	0	0	45	35
18	*PIK3C3*	5	1	6	1	0	0	38	0
8	*PTK2B*	6	3	9	3	0	0	31	4
14	*SOS2*	1	6	7	2	0	0	22	25
13	*FLT1*	8	17	25	8	2	1	21	17
3	*BCHE*	3	25	26	13	0	1	19	19
11	*PDGFD*	0	6	6	4	6	6	19	19
14	*HIF1A*	2	5	7	3	0	0	17	17
8	*PTK2*	4	5	9	2	0	0	14	6
1	*PIK3C2B*	22	38	60	23	2	0	14	16
1	*SHC1*	3	3	6	1	0	0	13	5
6	*VNN1*	4	2	6	1	1	2	7	13

[a] All rare SNPs were collapsed.

[b] Only nonsynonymous rare SNPs were collapsed.

Table 5.2 - Comparison of family data with collapsing and family data without collapsing

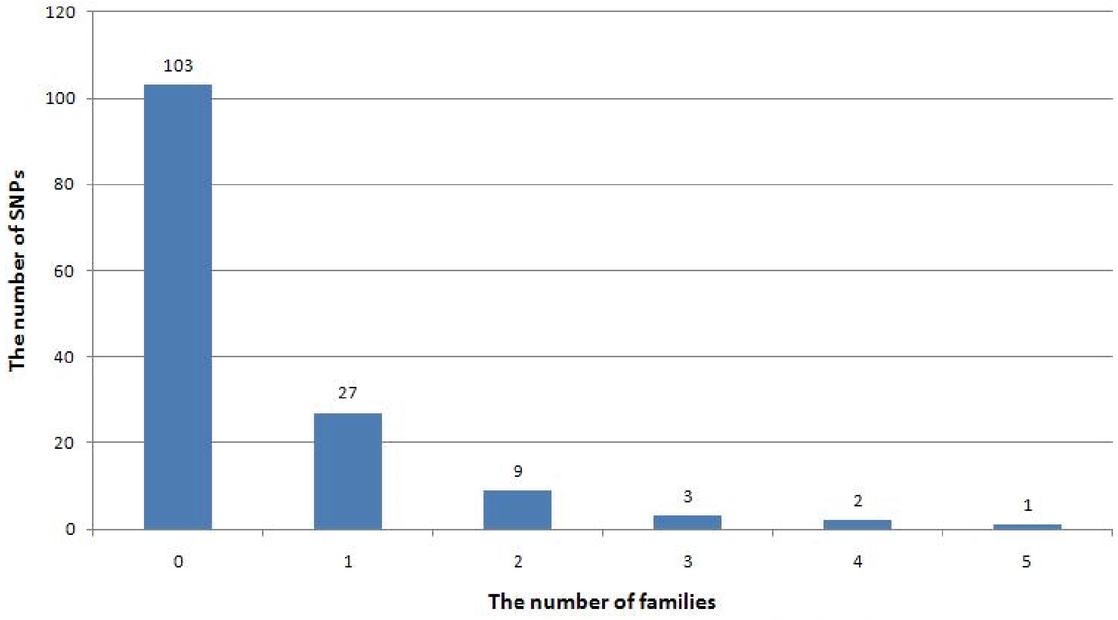
Power is calculated based on the threshold $P < 0.05$. Because of limited space, only those genes whose power is greater than 10% are shown.

						Collapsing		
Chromosome	Gene	Number of synonymous SNPs	Number of nonsynonymous SNPs	Total number of SNPs	Number of risk SNPs	Power (All) (%)[a]	Power (nonsynonymous) (%)[b]	Noncollapsing, power (%)
6	*VEGFA*	2	2	4	1	100.0	100.0	100.0
4	*VEGFC*	0	1	1	1	100.0	100.0	100.0
10	*SIRT1*	9	14	23	9	19.0	47	99
9	*VLDLR*	8	15	23	8	10.0	3.50	58.50
17	*SREBF1*	5	16	21	10	18.5	36	22.50
1	*PIK3R3*	3	2	5	1	11.0	1	0
8	*PLAT*	14	11	25	8	8.0	34	0
5	*FLT4*	3	5	8	2	6.0	12.50	0

[a] All rare SNPs were collapsed.

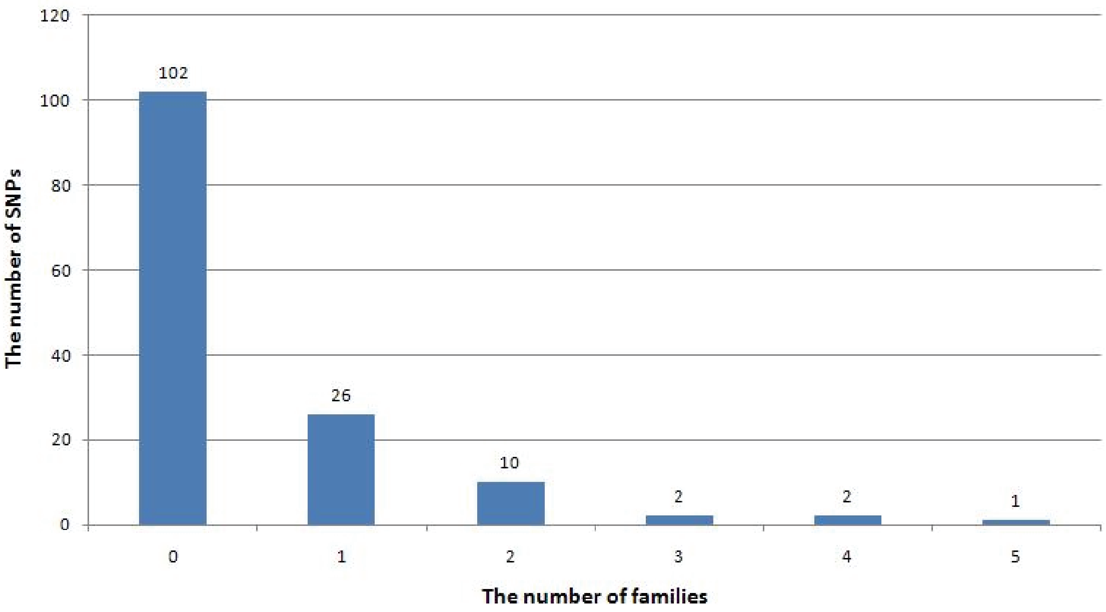
[b] Only nonsynonymous rare SNPs were collapsed.

Figure 5.1 - Distribution of rare causal SNPs within families



In the GAW17 data set, 145 of 162 casual SNPs are rare variants. Of these 145 rare variants, 103 do not exist in the family data. Eighty-five percent of the existing rare variants exist in only one or two families. The number above each bar indicates the exact number of rare SNPs in this category. It partly explains why many rare variants cannot be discovered using family data.

Figure 5.2 - Distribution of association signals within families



Each category indicates the number of families that report an association signal for each SNP. The number above each bar indicates the total number of rare causal SNPs in this category. The distribution of association signals matches well to the distribution of rare SNPs within families. It shows that when all families are analyzed together, the final result is largely shaped by only a few families.

Chapter 6: Conclusions and Future Directions

CONCLUSIONS

The human genome has around 3 billion base pairs of DNA and stores most, if not all, the information needed to build up a human being from scratch. Our very existence − both physical existence and mental existence − is determined by the combinational effects of gene and environment. Human diseases, ranging from Alzheimer’s disease to Zadik Barak Levin syndrome, are no exceptions. Since the discovery of Mendel’s laws, one of the most challenging problems in genetic research is to find and characterize different genetic variants that contribute to various human diseases. In the past decade, the research community has made impressive progress. New technologies were introduced, numerous methods were proposed and a large number of disease-associated genes were found [6]. The work presented here has largely reflected the recent development of this field, and as a result this work presented several new methods to address current research challenges.

One of the most important developments in the past 5 years is genotyping imputation [8]. Genotyping imputation allows researchers to evaluate the evidence for association at the genetic markers that are not directly-genotyped. It can improve power of individual scans and is particularly useful for combining results from different studies. However, there are two situations for which imputation has been problematic: (1) polymorphisms with low minor allele frequency (MAF), and (2) datasets where subjects are genotyped on different platforms. The imputation quality score that we introduced in Chapter 2 is sufficient to address these two issues. After filtering out poorly-imputed SNPs, we were able to remove thousands of false positives and obtain an acceptable Q-Q plot. We concluded that IQS is particularly useful for SNPs with low minor allele frequency and when datasets are genotyped on different platforms.

By using the same approach described in chapter 2, we discovered that CNVs with low minor allele frequency also have similar problems in the dataset of the CNV Discovery Project [113]. Our intent was to evaluate the relative performance of CNV calling in a genome wide scale despite the lack of experimental validation at individual CNV loci. The underlying rational is that whether a CNV that is called the first time can be confirmed the second time is restricted by both sensitivity and specificity. This kind of information in turn can give us some clues on sensitivity and specificity. By using the proposed method, we found that the positive predictive rate increases with the number of probes in the CNV and the size of the CNV. We also noticed that CNVs reported by multiple programs have a higher reproducibility rate and positive predicted rate. Our method was intended to find an efficient way to evaluate CNV calling in a genome wide scale. The fact that CNVs that are reported by multiple programs have a higher reliability is not part of our method, but rather our observation. The exact reasons for our observation may need further investigation, but it may be due to the fact that real CNVs have a more distinguished pattern that is easier to be spotted by different programs.

We applied our CNV methods to the Study of Addiction: Genetics and Environment [68]. Our analysis revealed that CNVs in 6q14.1 (P= 1.04 x10^{-6}) and 5q13.2 (P= 3.37 x10^{-4}) are significantly associated with alcohol dependence after adjusting multiple tests. The following qPCR experiments on these two CNV loci showed over a 97% agreement rate for our CNV calls. The experimental validation not only confirmed the association signal for alcohol dependence, but also demonstrated the power and legitimacy of our methods.

Interestingly, we also noticed the connection between alcohol dependence and personality. Among the five factors of the FFM dimensions, agreeableness (*P*=1.04 x10^{-20}), conscientiousness (*P*=3.93 x10^{-22}), extraversion (*P*=1.15 x10^{-12}) and neuroticism (*P*=6.95 x10^{-40}) are all significantly associated with alcohol dependence. We also found an exceptional P value (*P*=4.8 x10^{-5}) for conscientiousness in Chr5: 68,921,426 - 70,412,247, but not for the other four factors. This observation drives us to hypothesize that Chr5: 68,921,426 - 70,412,247 increases the risk of alcohol dependence by lowering conscientiousness, or more specifically self-discipline. Because conscientiousness and alcohol dependence are associated, it is possible that this link can be contributed to confounding effects. We believed that this is not likely the case, because agreeableness, extraversion and neuroticism are all associated with the risk of alcohol dependence but none of them are linked to Chr5: 68,921,426 - 70,412,247.

As the whole research community shifted its focus from common variants to rare variants [30], we also explored the possibility of applying our methods described in Chapter 2 & 3 to rare variants. Particularly, we noticed that some of the issues that we intended to address in Chapter 2 & 3 arise from rare variants. We had evidence that our methods are applicable to rare variants as well. However, we also noticed that our methods cannot address one major problem of detecting rare variants – the lack of power. By taking advantage of the simulation data from the Genetic Analysis Workshop, we integrated both the collapsing method and the family data method in an attempt to increase power. We concluded that our combinational approach offers a substantial power boost for certain causal genes that cannot be discovered otherwise.

There are several highlights of the work presented here. The motivation for Chapter 2 and Chapter 3 was to address data quality of imputation and CNV calling. The data quality issues are more noticeable when the allele frequency of SNPs and CNVs becomes rarer. Chapter 5 confirmed the commonly-held belief that traditional study design does not have sufficient power to discover rare variants, and hence proposed a new method to increase power. As a result, the work presented here aimed to provide solutions to the issues regarding the currently-ongoing transition from common variant research to rare variant research.

Many new methods are justifiable in mathematical theory, but may not be applicable in real data. The work presented here was also fortunate because it reaches beyond theory. The method and the Imputation Quality Score proposed in Chapter 2 were used in real SAGE data to assess imputation quality of some interesting SNPs, especially those at some target loci. The methods proposed in Chapter 3 were used to improve CNV calling in the alcohol dependence study and the BMI study. The results of connecting CNV calls and alcohol dependence by the method described in Chapter 3 were reported in detail in Chapter 4. Though at the current stage, we do not have enough sequencing data to test the method described in Chapter 5. But it is expected that the method we propose here will be applied to real sequencing data when it soon becomes available. In fact, several groups in the Genetic Analysis Workshop showed great interest in this combinational approach or some similar methods.

Another highlight of this work is that we had *in silico* and experimental validation for the methods described here. In Chapter 2, we used IQS to filter out poorly-imputed SNPs, and successfully removed many false positives. In Chapter 4, we did qPCR to validate CNVs that were called by the methods described in Chapter 3. We were able to confirm over 97% of CNV calls. The *in silico* and experimental validation showed the methods presented here are valid and can be very useful for future research.

FUTURE DIRECTIONS

The work presented here has provided several solutions to the current research challenges, but it also has prompted a need to investigate further.

As more and more genome wide association studies are completed, imputation will become more popular, and meta-analysis based on imputation may become a routine procedure. As a result, over-estimation of the quality of imputation due to chance agreement will be more common. Based on our method, a database can be constructed to document poorly-imputed SNPs and used to remove false positive associations. We envision this as a dynamic database to be updated when new datasets include subjects genotyped on multiple platforms. The future database will include, but will not be limited to, IQS scores for the following imputations: (1) from Affymetrix 6.0 to Illumina 1M, (2) from Illumina 1M to Affymetrix 6.0, (3) from Illumina 300K to Affymetrix 6.0 plus Illumina 1M, (4) from Illumina 550K to Affymetrix 6.0 plus Illumina 1M, and (5) from Affymetrix 5.0 to Affymetrix 6.0 plus Illumina 1M. We expected that this database can greatly help decrease the amount of false positive findings, making follow up of positive associations practical. In order to successfully build up and maintain a database proposed here, we anticipate a server with great computation capability and a team with good programming skills.

The methods that we introduced in Chapter 3 are a short-cut to evaluate CNV calling accuracy in a large scale. We concluded that the positive predictive rate increases with number of probes and the size of CNVs and that CNVs reported by multiple programs have a higher reliability. We believed that our general conclusion can hold true concerning other studies, but we also noticed some variations among different platforms. We expect that it will be more useful if we can provide a user-friendly program to different researchers so that they can evaluate the performance of CNV calling in the platform they have chosen.

Our finding of genome-wide association between alcohol dependence and the CNVs at 6q14.1 and 5q13.2 is encouraging. These two regions were previously shown to be associated with neurological disorders. But like most genome wide association studies, replication in an independent dataset is necessary. Unfortunately, no datasets with both appropriate genotyping and alcohol dependence measures are currently available. We are looking forward to new studies that can be used to validate our findings.

SAGE also includes a large number of addiction-related covariates. These covariates may be interesting to many researchers, for example, height, weight, BMI, the number of cigarettes and drinks per day, pre-term birth and age at menarche etc. It can be extremely rewarding to study the connection between these covariates and genetic variants. Particularly, we can further investigate the relationship between CNVs and these different traits by using the methods described in Chapter 3 & 4.   SAGE also has genotypes for African Americans. We expect that it will be equally rewarding to compare and contrast our findings between European Americans and African Americans [114].

We discussed our combinational approach to increase power for rare variants in Chapter 5. But we also noticed that the power gain is obtained at the cost of power for some other genes. Further study is required. Particularly, we expected an improved method that gives different weight to SNPs in the collapsing process based on SNP function. Several function prediction algorithms are available, for example, SIFT and PolyPhen-2 [110, 111]. We believed that the use of prediction algorithms for function will be helpful. In conjunction with association testing, linkage can identify regions of interest. Sequencing may only need to target regions that are discovered by linkage. In addition, the selection of the most informative families may further reduce the total genotyping cost. The use of extremes of a phenotypic distribution may further provide additional information [112]. Moreover, with the help of imputation, we may only need to do sequencing on a limited number of family members to obtain the whole genome sequencing data for all members of a family. All of these directions are promising and have the potential to make an impact in this field, but it takes time and effort to complete.

The work presented here is in parallel with the current development of genetic research and provides a blueprint for the future. At the present stage, we still know little about human disease and human gene. Even though hundreds of genome-wide association studies turned up thousands of genetic variants, they did very little to predict disease risk [115]. More research still needs to be done in order to better appreciate the relationship between gene and disease. But since the human genome only has 3 billion base pairs of DNA, compared to other disciplines without known boundaries, we should feel fortunate, because in the end we are searching in a finite universe.

References

1. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N: **Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia**. *Science (New York, NY* 1985, **230**(4732):1350-1354.
2. Queller DC, Strassmann JE, Hughes CR: **Microsatellites and kinship**. *Trends in ecology & evolution* 1993, **8**(8):285-288.
3. Feingold E: **Methods for linkage analysis of quantitative trait loci in humans**. *Theoretical population biology* 2001, **60**(3):167-180.
4. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature* 2007, **447**(7145):661-678.
5. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST *et al*: **Complement factor H polymorphism in age-related macular degeneration**. *Science (New York, NY* 2005, **308**(5720):385-389.
6. Johnson AD, O'Donnell CJ: **An open access database of genome-wide association results**. *BMC medical genetics* 2009, **10**:6.
7. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM *et al*: **A second generation human haplotype map of over 3.1 million SNPs**. *Nature* 2007, **449**(7164):851-861.
8. Li Y, Willer C, Sanna S, Abecasis G: **Genotype imputation**. *Annual review of genomics and human genetics* 2009, **10**:387-406.
9. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU *et al*: **A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants**. *Science (New York, NY* 2007, **316**(5829):1341-1345.
10. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A *et al*: **Finding the missing heritability of complex diseases**. *Nature* 2009, **461**(7265):747-753.
11. Visscher PM: **Sizing up human height variation**. *Nature genetics* 2008, **40**(5):489-490.
12. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R *et al*: **Diet and the evolution of human amylase gene copy number variation**. *Nature genetics* 2007, **39**(10):1256-1260.
13. Le Marechal C, Masson E, Chen JM, Morel F, Ruszniewski P, Levy P, Ferec C: **Hereditary pancreatitis caused by triplication of the trypsinogen locus**. *Nature genetics* 2006, **38**(12):1372-1374.
14. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ *et al*: **The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility**. *Science (New York, NY* 2005, **307**(5714):1434-1440.
15. Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E *et al*: **Copy number polymorphism**

in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* 2006, **439**(7078):851-855.

16. Padiath QS, Saigoh K, Schiffmann R, Asahara H, Yamada T, Koeppen A, Hogan K, Ptacek LJ, Fu YH: **Lamin B1 duplications cause autosomal dominant leukodystrophy**. *Nature genetics* 2006, **38**(10):1114-1123.
17. Lee JA, Lupski JR: **Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders**. *Neuron* 2006, **52**(1):103-121.
18. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J *et al*: **Strong association of de novo copy number mutations with autism**. *Science (New York, NY* 2007, **316**(5823):445-449.
19. Fan YS, Jayakar P, Zhu H, Barbouth D, Sacharow S, Morales A, Carver V, Benke P, Mundy P, Elsas LJ: **Detection of pathogenic gene copy number variations in patients with mental retardation by genomewide oligonucleotide array comparative genomic hybridization**. *Human mutation* 2007, **28**(11):1124-1132.
20. Bodmer W, Bonilla C: **Common and rare variants in multifactorial susceptibility to common diseases**. *Nature genetics* 2008, **40**(6):695-701.
21. Carletta J: **Assessing agreement on classification tasks: The kappa statistic**. *Computational Linguistics* 1996, **22**(2):249-254.
22. Abecasis GR, Cherny SS, Cookson WO, Cardon LR: **Merlin--rapid analysis of dense genetic maps using sparse gene flow trees**. *Nature genetics* 2002, **30**(1):97-101.
23. Laurie C, Bierut L, Bhangale T, Boehm F, Caporaso N, Doheny K, Gabriel S, Harris E, Hu F, Jacobs K *et al*: **Genotype data cleaning for whole-genome association studies**. *In preparation*.
24. Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies**. *PLoS genetics* 2009, **5**(6):e1000529.
25. Asimit J, Zeggini E: **Rare Variant Association Analysis Methods for Complex Traits**. *Annual Review of Genetics, Vol 44* 2010, **44**:293-308.
26. Lin P, Hartz SM, Zhang Z, Saccone SF, Wang J, Tischfield JA, Edenberg HJ, Kramer JR, A MG, Bierut LJ *et al*: **A new statistic to evaluate imputation reliability**. *PloS one* 2010, **5**(3):e9697.
27. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: **Detection of large-scale variation in the human genome**. *Nature genetics* 2004, **36**(9):949-951.
28. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M: **PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data**. *Genome research* 2007, **17**(11):1665-1674.
29. Lin P, Hartz SM, Wang J-C, Krueger RF, Foroud TM, Edenberg HJ, Jr. JIN, Brooks AI, Tischfield JA, Almasy L *et al*: **Copy Number Variation Accuracy in Genome Wide Studies**. *human heredity* 2011, **In Press**.
30. Bansal V, Libiger O, Torkamani A, Schork NJ: **Statistical analysis strategies for association studies involving rare variants**. *Nature reviews*, **11**(11):773-785.

31. Li B, Leal SM: **Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data**. *American journal of human genetics* 2008, **83**(3):311-321.
32. Niccols A: **Fetal alcohol syndrome and the developing socio-emotional brain**. *Brain and cognition* 2007, **65**(1):135-142.
33. Altshuler D, Daly M: **Guilt beyond a reasonable doubt**. *Nature genetics* 2007, **39**(7):813-815.
34. Marchini J, Howie B, Myers S, McVean G, Donnelly P: **A new multipoint method for genome-wide association studies by imputation of genotypes**. *Nature genetics* 2007, **39**(7):906-913.
35. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM *et al*: **Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease**. *Nature genetics* 2008, **40**(8):955-962.
36. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C *et al*: **Identification of ten loci associated with height highlights new biological pathways in human growth**. *Nature genetics* 2008, **40**(5):584-591.
37. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM *et al*: **Newly identified loci that influence lipid concentrations and risk of coronary artery disease**. *Nature genetics* 2008, **40**(2):161-169.
38. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM *et al*: **Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes**. *Science (New York, NY* 2007, **316**(5829):1336-1341.
39. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G *et al*: **Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes**. *Nature genetics* 2008, **40**(5):638-645.
40. Browning BL, Browning SR: **Haplotypic analysis of Wellcome Trust Case Control Consortium data**. *Human genetics* 2008, **123**(3):273-280.
41. Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A: **A comprehensive evaluation of SNP genotype imputation**. *Human genetics* 2009, **125**(2):163-171.
42. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering**. *American journal of human genetics* 2007, **81**(5):1084-1097.
43. Manolio TA, Rodriguez LL, Brooks L, Abecasis G, Ballinger D, Daly M, Donnelly P, Faraone SV, Frazer K, Gabriel S *et al*: **New models of collaboration in genome-wide association studies: the Genetic Association Information Network**. *Nature genetics* 2007, **39**(9):1045-1051.
44. Cohen J: **A coefficient of agreement for nominal scales**. *Educ psychol Meas* 1960(20):37-46.

45. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nature genetics* 2006, **38**(8):904-909.

46. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF: **Practical aspects of imputation-driven meta-analysis of genome-wide association studies**. *Human molecular genetics* 2008, **17**(R2):R122-128.

47. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE *et al*: **Population structure, differential bias and genomic control in a large-scale, case-control association study**. *Nature genetics* 2005, **37**(11):1243-1246.

48. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL *et al*: **New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk**. *Nature genetics* 2010.

49. Repapi E, Sayers I, Wain LV, Burton PR, Johnson T, Obeidat M, Zhao JH, Ramasamy A, Zhai G, Vitart V *et al*: **Genome-wide association study identifies five loci associated with lung function**. *Nature genetics* 2009, **42**(1):36-44.

50. Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, Lyssenko V, Bouatia-Naji N, Dupuis J, Jackson AU *et al*: **Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge**. *Nature genetics* 2010.

51. Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, Marciante KD, Franceschini N, van Durme YM, Chen TH, Barr RG *et al*: **Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function**. *Nature genetics* 2009, **42**(1):45-52.

52. McMahon FJ, Akula N, Schulze TG, Muglia P, Tozzi F, Detera-Wadleigh SD, Steele CJ, Breuer R, Strohmaier J, Wendland JR *et al*: **Meta-analysis of genome-wide association data identifies a risk locus for major mood disorders on 3p21.1**. *Nature genetics* 2010.

53. Pfeufer A, van Noord C, Marciante KD, Arking DE, Larson MG, Smith AV, Tarasov KV, Muller M, Sotoodehnia N, Sinner MF *et al*: **Genome-wide association study of PR interval**. *Nature genetics* 2010.

54. Beckmann JS, Estivill X, Antonarakis SE: **Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability**. *Nature reviews* 2007, **8**(8):639-646.

55. Rice JP, Rochberg N, Endicott J, Lavori PW, Miller C: **Stability of psychiatric diagnoses. An application to the affective disorders**. *Archives of general psychiatry* 1992, **49**(10):824-830.

56. Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ: **Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays**. *Nucleic acids research*, **38**(9):e105.

57. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P *et al*: **Origins and functional impact of copy number variation in the human genome**. *Nature*, **464**(7289):704-712.

58. Cornelis MC, Agrawal A, Cole JW, Hansel NN, Barnes KC, Beaty TH, Bennett SN, Bierut LJ, Boerwinkle E, Doheny KF *et al*: **The gene, environment**

association studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genetic epidemiology* 2010:364-372.

59. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI *et al*: **Population analysis of large copy number variants and hotspots of human genetic disease**. *American journal of human genetics* 2009, **84**(2):148-161.

60. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K: **Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms**. *Nucleic acids research* 2008, **36**(19):e126.

61. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J: **QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data**. *Nucleic acids research* 2007, **35**(6):2013-2025.

62. Wain LV, Pedroso I, Landers JE, Breen G, Shaw CE, Leigh PN, Brown RH, Tobin MD, Al-Chalabi A: **The role of copy number variation in susceptibility to amyotrophic lateral sclerosis: genome-wide association study and comparison with published loci**. *PloS one* 2009, **4**(12):e8175.

63. Breitling LP, Dahmen N, Mittelstrass K, Illig T, Rujescu D, Raum E, Winterer G, Brenner H: **Smoking cessation and variations in nicotinic acetylcholine receptor subunits alpha-5, alpha-3, and beta-4 genes**. *Biol Psychiatry* 2009, **65**(8):691-695.

64. Bucan M, Abrahams BS, Wang K, Glessner JT, Herman EI, Sonnenblick LI, Alvarez Retuerto AI, Imielinski M, Hadley D, Bradfield JP *et al*: **Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes**. *PLoS genetics* 2009, **5**(6):e1000536.

65. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE *et al*: **Copy number variation at 1q21.1 associated with neuroblastoma**. *Nature* 2009, **459**(7249):987-991.

66. Bowden W, Skorupski J, Kovanci E, Rajkovic A: **Detection of novel copy number variants in uterine leiomyomas using high-resolution SNP arrays**. *Molecular human reproduction* 2009, **15**(9):563-568.

67. Li J, Yang T, Wang L, Yan H, Zhang Y, Guo Y, Pan F, Zhang Z, Peng Y, Zhou Q *et al*: **Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations**. *PloS one* 2009, **4**(11):e7958.

68. Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, Fisher S, Fox L, Howells W, Bertelsen S *et al*: **A genome-wide association study of alcohol dependence**. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(11):5082-5087.

69. Rice JP, Endicott J, Knesevich MA, Rochberg N: **The estimation of diagnostic sensitivity using stability data: an application to major depressive disorder**. *Journal of psychiatric research* 1987, **21**(4):337-345.

70. Rice JP, McDonald-Scott P, Endicott J, Coryell W, Grove WM, Keller MB, Altis D: **The stability of diagnosis with an application to bipolar II disorder**. *Psychiatry research* 1986, **19**(4):285-296.

71. Lin P, Hartz SM, Zhang Z, Saccone SF, Wang J, Tischfield JA, Edenberg HJ, Kramer JR, A MG, Bierut LJ *et al*: **A new statistic to evaluate imputation reliability**. *PloS one*, **5**(3):e9697.

72. Winchester L, Yau C, Ragoussis J: **Comparing CNV detection methods for SNP arrays**. *Briefings in functional genomics & proteomics* 2009, **8**(5):353-366.

73. Kanny D, Liu Y, Brewer RD: **Binge drinking - United States, 2009**. *MMWR Surveill Summ* 2011, **60 Suppl**:101-104.

74. Testino G: **Alcoholic diseases in hepato-gastroenterology: a point of view**. *Hepato-gastroenterology* 2008, **55**(82-83):371-377.

75. Caan W, De Belleroche J: **Drink, drugs and dependence : from science to clinical practice**. London ; New York, NY: Routledge; 2002.

76. Hasin DS, Stinson FS, Ogburn E, Grant BF: **Prevalence, correlates, disability, and comorbidity of DSM-IV alcohol abuse and dependence in the United States: results from the National Epidemiologic Survey on Alcohol and Related Conditions**. *Archives of general psychiatry* 2007, **64**(7):830-842.

77. Bierut LJ, Dinwiddie SH, Begleiter H, Crowe RR, Hesselbrock V, Nurnberger JI, Jr., Porjesz B, Schuckit MA, Reich T: **Familial transmission of substance dependence: alcohol, marijuana, cocaine, and habitual smoking: a report from the Collaborative Study on the Genetics of Alcoholism**. *Archives of general psychiatry* 1998, **55**(11):982-988.

78. Reich T, Edenberg HJ, Goate A, Williams JT, Rice JP, Van Eerdewegh P, Foroud T, Hesselbrock V, Schuckit MA, Bucholz K *et al*: **Genome-wide search for genes affecting the risk for alcohol dependence**. *American journal of medical genetics* 1998, **81**(3):207-215.

79. Heath AC, Bucholz KK, Madden PA, Dinwiddie SH, Slutske WS, Bierut LJ, Statham DJ, Dunne MP, Whitfield JB, Martin NG: **Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men**. *Psychological medicine* 1997, **27**(6):1381-1396.

80. Knopik VS, Heath AC, Madden PA, Bucholz KK, Slutske WS, Nelson EC, Statham D, Whitfield JB, Martin NG: **Genetic effects on alcohol dependence risk: re-evaluating the importance of psychiatric and other heritable risk factors**. *Psychological medicine* 2004, **34**(8):1519-1530.

81. Kendler KS, Neale MC, Heath AC, Kessler RC, Eaves LJ: **A twin-family study of alcoholism in women**. *The American journal of psychiatry* 1994, **151**(5):707-715.

82. Braillon A, Dubois G: **Alcohol and public health**. *Lancet* 2005, **365**(9468):1387.

83. Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, Fisher S, Fox L, Howells W, Bertelsen S *et al*: **A genome-wide association study of alcohol dependence**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(11):5082-5087.

84. Long JC, Knowler WC, Hanson RL, Robin RW, Urbanek M, Moore E, Bennett PH, Goldman D: **Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 from an autosome-wide scan in an American Indian population**. *American journal of medical genetics* 1998, **81**(3):216-221.

85. Hill SY, Shen S, Zezza N, Hoffman EK, Perlin M, Allan W: **A genome wide search for alcoholism susceptibility genes**. *Am J Med Genet B Neuropsychiatr Genet* 2004, **128B**(1):102-113.

86. Edenberg HJ, Foroud T: **The genetics of alcoholism: identifying specific genes through family studies**. *Addiction biology* 2006, **11**(3-4):386-396.

87. Treutlein J, Cichon S, Ridinger M, Wodarz N, Soyka M, Zill P, Maier W, Moessner R, Gaebel W, Dahmen N *et al*: **Genome-wide association study of alcohol dependence**. *Archives of general psychiatry* 2009, **66**(7):773-784.

88. Heath AC, Whitfield JB, Martin NG, Pergadia ML, Goate AM, Lind PA, McEvoy BP, Schrage AJ, Grant JD, Chou YL *et al*: **A Quantitative-Trait Genome-Wide Association Study of Alcoholism Risk in the Community: Findings and Implications**. *Biological psychiatry* 2011.

89. Edenberg HJ, Koller DL, Xuei X, Wetherill L, McClintick JN, Almasy L, Bierut LJ, Bucholz KK, Goate A, Aliev F *et al*: **Genome-wide association study of alcohol dependence implicates a region on chromosome 11**. *Alcoholism, clinical and experimental research* 2010, **34**(5):840-852.

90. Martens MA, Wilson SJ, Reutens DC: **Research Review: Williams syndrome: a critical review of the cognitive, behavioral, and neuroanatomical phenotype**. *Journal of child psychology and psychiatry, and allied disciplines* 2008, **49**(6):576-608.

91. Krajewski KM, Lewis RA, Fuerst DR, Turansky C, Hinderer SR, Garbern J, Kamholz J, Shy ME: **Neurological dysfunction and axonal degeneration in Charcot-Marie-Tooth disease type 1A**. *Brain* 2000, **123 (Pt 7)**:1516-1527.

92. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP *et al*: **Autism genome-wide copy number variation reveals ubiquitin and neuronal genes**. *Nature* 2009, **459**(7246):569-573.

93. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS *et al*: **Functional impact of global rare copy number variation in autism spectrum disorders**. *Nature*, **466**(7304):368-372.

94. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE *et al*: **Large recurrent microdeletions associated with schizophrenia**. *Nature* 2008, **455**(7210):232-236.

95. Lachman HM, Pedrosa E, Petruolo OA, Cockerham M, Papolos A, Novak T, Papolos DF, Stopkova P: **Increase in GSK3beta gene copy number variation in bipolar disorder**. *Am J Med Genet B Neuropsychiatr Genet* 2007, **144B**(3):259-265.

96. Muller PY, Janovjak H, Miserez AR, Dobbie Z: **Processing of gene expression data generated by quantitative real-time RT-PCR**. *BioTechniques* 2002, **32**(6):1372-1374, 1376, 1378-1379.

97. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW: **Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome**. *Cytogenetic and genome research* 2006, **115**(3-4):205-214.

98. Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE *et al*: **A comprehensive analysis of common copy-number variations in the human genome**. *American journal of human genetics* 2007, **80**(1):91-104.

99. Also-Rallo E, Alias L, Martinez-Hernandez R, Caselles L, Barcelo MJ, Baiget M, Bernal S, Tizzano EF: **Treatment of spinal muscular atrophy cells with drugs that upregulate SMN expression reveals inter- and intra-patient variability**. *Eur J Hum Genet* 2011.

100. Dachs E, Hereu M, Piedrafita L, Casanovas A, Caldero J, Esquerda JE: **Defective neuromuscular junction organization and postnatal myogenesis in mice with severe spinal muscular atrophy**. *Journal of neuropathology and experimental neurology* 2011, **70**(6):444-461.

101. O'Driscoll MC, Daly SB, Urquhart JE, Black GC, Pilz DT, Brockmann K, McEntagart M, Abdel-Salam G, Zaki M, Wolf NI *et al*: **Recessive mutations in the gene encoding the tight junction protein occludin cause band-like calcification with simplified gyration and polymicrogyria**. *American journal of human genetics* 2010, **87**(3):354-364.

102. Puppala S, Coletta DK, Schneider J, Hu SL, Farook VS, Dyer TD, Arya R, Blangero J, Duggirala R, Defronzo RA *et al*: **Genome-Wide Linkage Screen for Systolic Blood Pressure in the Veterans Administration Genetic Epidemiology Study (VAGES) of Mexican-Americans and Confirmation of a Major Susceptibility Locus on Chromosome 6q14.1**. *Hum Hered* 2011, **71**(1):1-10.

103. Lespinasse J, Gimelli S, Bena F, Antonarakis SE, Ansermet F, Paoloni-Giacobino A: **Characterization of an interstitial deletion 6q13-q14.1 in a female with mild mental retardation, language delay and minor dysmorphisms**. *European journal of medical genetics* 2009, **52**(1):49-52.

104. Manolio TA: **Genomewide association studies and assessment of the risk of disease**. *The New England journal of medicine*, **363**(2):166-176.

105. Dering C, Pugh E, Ziegler A: **Statistical analysis of rare sequence variants: An overview of collapsing methods**. *BMC proceedings (ms in prep)*

106. Thornton T, McPeek MS: **Case-control association testing with related individuals: a more powerful quasi-likelihood score test**. *American journal of human genetics* 2007, **81**(2):321-337.

107. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB *et al*: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness**. *Nature genetics* 2006, **38**(2):203-208.

108. Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS: **Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus**. *American journal of human genetics* 2003, **73**(3):612-626.

109. Doull IJ: **Recent advances in cystic fibrosis**. *Archives of disease in childhood* 2001, **85**(1):62-66.

110. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function**. *Nucleic acids research* 2003, **31**(13):3812-3814.

111. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nature methods*, **7**(4):248-249.

112. Guey LT, Kravic J, Melander O, Burtt NP, Laramie JM, Lyssenko V, Jonsson A, Lindholm E, Tuomi T, Isomaa B *et al*: **Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants**. *Genetic epidemiology* 2011.

113. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W *et al*: **Global variation in copy number in the human genome**. *Nature* 2006, **444**(7118):444-454.

114. Saccone NL, Wang JC, Breslau N, Johnson EO, Hatsukami D, Saccone SF, Grucza RA, Sun L, Duan W, Budde J *et al*: **The CHRNA5-CHRNA3-CHRNB4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans**. *Cancer research* 2009, **69**(17):6848-6856.

115. Maher B: **Personal genomes: The case of the missing heritability**. *Nature* 2008, **456**(7218):18-21.