Application of Double Sampling to Combine Measured and Imputed Genotype Data

in Genetic Association Studies

A Dissertation Presented

by

Qilong Yuan

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

December 2010

Stony Brook University

The Graduate School

Qilong Yuan

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Stephen J. Finch – Dissertation Advisor
Professor, Department of Applied Mathematics and Statistics

Nancy R. Mendell – Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Wei Zhu – Member
Professor, Department of Applied Mathematics and Statistics

Derek Gordon – Outside Member
Associate Professor, Department of Genetics
Rutgers, The State University of New Jersey

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Application of Double Sampling to Combine Measured and Imputed Genotype Data

in Genetic Association Studies

by

Qilong Yuan

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2010

Genotype imputation provides an essential technique for genome-wide association studies (GWAS) with hundreds of thousands of SNPs. Understanding the connection between imputation inconsistencies and the power to detect association at imputed markers or the disease genes close to them is important for the optimal design of imputation-based GWAS since genotype misclassification can significantly decrease statistical power to detect association. Double sampling of genotypes is a statistical procedure in which a portion of subjects receive a second and more precise genotyping. This paper applies the likelihood ratio test allowing for errors (LRT-AE), which incorporates double sample information for genotypes on a sub-sample of cases/controls, to correct for imputation inconsistencies. Parameters used to determine the log likelihoods are determined using the Expectation-Maximization (EM) algorithm. To compare the performance of the LRT-AE with the performance of the likelihood ratio test (LRT), which makes no adjustment for imputation inconsistencies, I perform simulation studies using a factorial design with high and low settings of: disease minor allele frequency (MAF), heterozygote relative risk, mode of inheritance (MOI), disease prevalence, and proportion of double sampled subjects. The LRT-AE method maintains correct type I error rates for all null simulations and all significance level thresholds (5%, 1%). Power improvement, however, is not significant unless more than 50% of subjects are in the double sampled group. Unbiased estimates of imputation inconsistency rates are also obtained from the LRT-AE method.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

Even though the human genome has approximately 3,000,000,000 base pairs (i.e., pairs of nucleotides on opposite complementary DNA or RNA strands that are connected via hydrogen bonds), the genetic sequences of different people are remarkably similar. According to the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov), when the chromosomes of two humans are compared, their DNA sequences can be identical for hundreds of bases. But at about one in every 1,000 to 2,000 bases, on average, sequences differ (Sachidanandam, et al. 2001, Olivier 2003). For example, one person may have a C nucleotide at a given location, while another may have a T nucleotide. In addition, there can be copy number variation, where a person may have extra bases or miss a segment of DNA at a specific location resulting from insertions, duplications, deletions, or inversions. A change in the DNA sequences can have no effect or a beneficial effect, or it can prevent the gene from functioning properly or completely and cause a disease. Identifying these genetic variants, where they occur in our DNA, and how they are distributed among people within populations and among populations in different parts of the world may help to establish connections between particular genetic variants and diseases.

Differences in individual bases account for a large fraction of the human genetic diversity and are by far the most common type of genetic variation. These genetic differences are known as single nucleotide polymorphisms (SNPs). A SNP is a DNA sequence variation occurring when a single nucleotide (i.e., A, T, C, or G) in the genome differs between members of a species (or between paired chromosomes in an individual). The alternative DNA sequences at the same physical locus are known as *alleles*. Most SNPs have only two alleles. For geneticists, SNPs act as *markers* to locate disease genes in DNA sequences. If researchers want to know where a disease gene is located, they can compare the genotype distribution of a SNP in people who have the disease to the distribution in people who do not. If the genotype distributions of a particular SNP are different in the cases and the controls, then that SNP may be close to the disease gene.

Genome-wide association studies (GWAS) are a common approach to find genetic variations associated with the presence of a particular disease or a certain trait. To carry out a GWAS, researchers use two groups of participants: people with the disease being studied and similar people without the disease. Researchers obtain the complete set of DNA, or genome, from each participant. The genomes are then scanned by automated laboratory machines. The machines quickly survey each participant's genome for strategically selected SNPs. If certain genetic variations are found to be significantly more frequent in people with the disease compared to people without disease, the SNP may be associated with the disease. Whole genome information, when combined with clinical and other phenotype data, offers the potential for increased understanding of basic biological processes affecting human health, improvement in the prediction of disease and patient care, and ultimately the realization of the promise of personalized medicine. In addition, rapid advances in understanding the patterns of human genetic variation and maturing high-throughput, cost-effective methods for genotyping are providing powerful research tools for identifying genetic variants that contribute to health and disease (http://grants.nih.gov/grants/gwas).

GWAS have successfully identified susceptibility genes for many common diseases. For example, GWAS have identified genetic variations that contribute to risk of age-related macular degeneration (AMD), type II diabetes, Parkinson's disease, heart disorders, obesity, Crohn's disease and prostate cancer, as well as genetic variations that influence response to anti-depressant medications (http://www.genome.gov). Once a genetic association has been established, follow up re-sequencing is usually required to further identify the actual causal variants, since SNPs identified in GWAS are probably genetic markers that are close to the true disease genes rather than being the actual causal genes themselves. However, there are very few examples of the actual variants being identified from a GWAS. One explanation is that rare variants (less common than those routinely studied in GWAS) with large effects account for or contribute to many of the identified association signals reported in GWAS (Dickson, Wang, Krantz, Hakonarson and Goldstein 2010).

Dense marker information has become available to researchers in recent years, making GWAS more powerful. Additionally, various genotype imputation methods have been employed to impute additional genotypes to supplement the available marker sets. Missing data imputation has been used to infer genotypes for known, but non-genotyped, variants. These variants can then be tested for association with the disease ("Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls" 2007, Scott, et al. 2007). There are several reasons for the incorporation of genotype imputation in GWAS:

First, genotype imputation may potentially increase the power to detect disease associations with a given marker set. It is understood that sometimes GWAS are underpowered to detect causal genes because of insufficient marker data (i.e., the available marker set is not dense enough to detect possible association signals in some studies). As a result, there is a need for genotype imputation at non-genotyped markers to achieve greater power. For studies with limited sample size, denser marker panels than those currently available are especially needed. Even if the sample size is large, the incorporation of imputed genotypes can lead to a substantial power gain for both common disease variants (Becker, Flaquer, Brockschmidt, Herold and Steffens 2009) and rare disease variants with population frequency less than 5% (Browning and Browning 2008). Browning and Browning showed that combining imputed and measured genotype data in multi-locus association studies appears to facilitate the detection of rare causative variants that might have been overlooked if use the original genotypes alone (Browning, et al. 2008).

Second, using imputed genotypes can reduce the cost of GWAS by using smaller, thus less expensive, arrays (such as the Illumina 300K rather than the Illumina 550K array), since genotype imputation generally costs less (both with respect to time and money) than actual genotyping. However, a larger sample size is needed to achieve comparable power when using a smaller array, as imputation is typically less accurate than genotyping (Anderson, et al. 2008).

Third, imputation can be valuable in the fine mapping of known disease-associated regions. It may help to identify additional candidate SNPs worth including in more detailed follow-up studies, or it may help narrow the region being searched for a gene.

Genotype imputation has several applications in the context of GWAS:

First, one can impute missing genotypes for markers that fail to pass quality control inspection. Standard quality control procedures remove all individuals and markers that have a large proportion of missing data, resulting in smaller sample size and lower marker density. One solution is to impute the missing genotypes using the almost complete data from the individuals and markers that have passed the quality control as the reference. Accuracy of imputed genotypes greater than 98% has been achieved in studies with 3,000+ individuals genotyped at the density of Affymetrix 500K array (Browning and Browning 2007).

Second, imputation of SNP genotypes has been proposed as a powerful means to include genetic markers into large-scale disease association studies without actually genotyping the markers (Marchini, Howie, Myers, McVean and Donnelly 2007, Servin and Stephens 2007). One can impute genotypes for markers of interest which are not genotyped in the study using a reference panel. A reference panel is composed of a group of individuals in whom markers are genotyped at extremely high density so that the genotyped SNPs in the reference panel are likely to include most of the SNPs genotyped in any commercially available gene chip. Reference panels such as the HapMap from the International HapMap Project are publicly available. The strategy of imputing HapMap SNPs has been adopted in several GWAS ("Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls" 2007, Scott, et al. 2007, Chambers, et al. 2008, Willer, et al. 2008, Zeggini, et al. 2008). This strategy has been successful in finding associations that would not have been found using only the original genotype data. For example, Zeggini et al. imputed 2.2 million HapMap SNPs in three collections of type II diabetes cases and controls (Zeggini, et al. 2008). Two of the collections were genotyped on the Affymatrix 500K GeneChip, and the third was genotyped on the Illumina 317K chip. The imputation resulted in two significant results that would not have been found using only the original genotype data.

Third, one can combine data from GWAS scans based on different genotyping platforms. Sometimes several datasets from different studies may be available to a researcher. If used individually, even the most significant association from an individual study may be neglected. Increased power can be obtained by combining all available datasets to get a larger sample with markers genotyped at higher density. This allows the detection of associations that are not among the top hits in any individual study, but that show a trend in each component study. The problem with combining data from different studies is that different studies may use different genotyping platforms, and the genotyped markers do not necessarily overlap. Thus, markers which are genotyped for individuals in some studies may not be genotyped in other studies. A powerful approach to assess markers genotyped in some component studies but not in others is through imputation (Pe'er, et al. 2006). Several groups have recently taken this approach, combining data from studies that used different genotyping platforms by using a reference panel to impute genotypes at non-genotyped markers for each individual study, and have found novel associations (Barrett, et al. 2008, Lettre, et al. 2008, Willer, et al. 2008). Here the reference panel is essential, unless the degree of overlapping among the genotyping platforms is very high (Browning 2008).

Genotype imputation provides an essential technique for GWAS with millions of SNPs (Servin, et al. 2007). The connection between imputation inconsistencies and the power to detect association at imputed markers or the disease genes close to them is important for the optimal design and interpretation of imputation based GWAS. A common assumption when analyzing data from GWAS is that genotypes obtained at each locus are correct for each individual. This assumption is reasonable because of maturing genotyping technologies and improved quality control procedures. Imputed genotypes, on the other hand, are generally less accurate than their measured counterparts (Anderson, et al. 2008). Non-differential genotyping errors (i.e., genotyping errors that occur with similar frequencies in the cases and the controls) can significantly decrease the statistical power to detect genetic association and thus inflate the sample size required to obtain comparative power and significance level (Gordon, Finch, Nothnagel and Ott 2002, Ahn, et al. 2007, Borchers, Brown, McLellan, Bekmetjev and Tintle 2009, Tintle, et al. 2009). Not surprisingly, the impact of genotype imputation inconsistencies is similar to the impact of genotyping errors: a significant loss of power. Huang et al. (2009) demonstrated that each 1% increase in the imputation inconsistency rate leads to an increase of approximately 5%-13% in the sample size required to achieve power at an imputed marker equal to that obtained if the genotype distribution of the marker is known with certainty (Huang, Wang and Rosenberg 2009). Typical imputation inconsistency rates (approximately 2%-6%) can lead to a large increase (approximately 10%-60%) in the required sample size (Huang, et al. 2009). To address the issues raised by Huang and others, this paper applies the double sampling method, incorporated in the likelihood ratio test allowing for errors (LRT-AE) (Gordon, et al. 2004b), to merge imputed and measured genotype data in genetic association studies. I use simulation studies to compare the empirical power to detect association using the standard likelihood ratio test (LRT) approach and the LRT-AE approach at various combinations of genetic parameters, and calculate the analytical power to verify the findings. Estimated imputation inconsistency rates are also reported.

Chapter 2 Methodology

2.1 Genotype Imputation

2.1.1 Haplotypes and Genotype Imputation

The term haplotype is a contraction of the term “haploid genotype” (http://en.wikipedia.org/wiki/Haplotype). For each chromosome, an individual gets one strand of DNA from his mother and the other from his father. During prophase I of meiosis, the eight available chromatids are in tight formation with one another and chromosomal crossover events may occur. Chromosomal crossover refers to recombination between the paired chromosomes inherited from each of one’s parents. The recombination frequency between two locations along a chromosome depends on their distance. For SNPs sufficiently distant on the same chromosome, the amount of crossover is high enough to destroy the correlation between the alleles, while SNPs that are near each other tend to be inherited together by offspring.

The SNPs within chromosomal regions that are always transferred to offspring as a whole are known as haplotypes (http://hapmap.ncbi.nlm.nih.gov/whatishapmap.html). In other words, a haplotype is a set of SNPs on the same strand of a chromosome that is transmitted together to offspring. In reporting genetic sequences, the identical base pairs are not reported, leaving only SNPs to specify the haplotypes. A haplotype may refer to as few as one SNP or to sequences of thousands of nucleotides, depending on the recombination events that have occurred between a given set of loci.

Over multiple successive generations, recombination and novel mutation events lead to a rearrangement of ancestral haplotypes. As a consequence, a SNP allele remains on the same portion of the ancestral haplotypes only with other SNP alleles that are in its close physical proximity. This non-random arrangement of adjacent loci, i.e., the maintenance of a small segment of the ancestral haplotypes, is called *linkage disequilibrium* (LD), or allelic association.

One measure of LD is r^2, where r^2 is the correlation coefficient between pairs of loci. For haplotypes of two loci A and B with two alleles each, Table 2.1 contains the frequency of each combination.

Table 2.1

Frequencies of the Combinations of Loci A and B with Two Alleles Each

Haplotype	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

One can use the above frequencies to determine the frequency of each of the alleles:

Table 2.2

Frequency of Each of the Alleles

Allele	Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$P_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} + x_{22}$

If the two loci are independent of each other (i.e., there is no LD between them), then x_{11}, the frequency of A_1B_1, is p_1q_1. One measurement of LD is the deviation of the observed frequency of a haplotype from the expected and is commonly denoted by capital D:

$$D = x_{11} - p_1 q_1 .$$

In genetics literatures "two alleles are in LD" means that $D \neq 0$. The measure D is easy to calculate but has the disadvantage of depending on the frequencies of the alleles. Another measure of LD is r^2, which is the correlation coefficient between pairs of loci. It is defined as:

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2} .$$

A value of r^2 close to one indicates that the two SNPs are in strong LD, and a value of r^2 close to zero indicates that the two SNPs are in weak LD.

If SNPs are not in LD, the alleles of the SNPs occur in seemingly random combinations on individual chromosomes. Therefore, in the absence of LD among SNP alleles, the alleles can form a large number of different haplotypes. For example, there are 2^n sequences of indicator functions of base pair changes for n consecutive SNP positions. In contrast, for a region where neighboring SNPs are in significant LD, only a small number of resulting haplotypes will be observed.

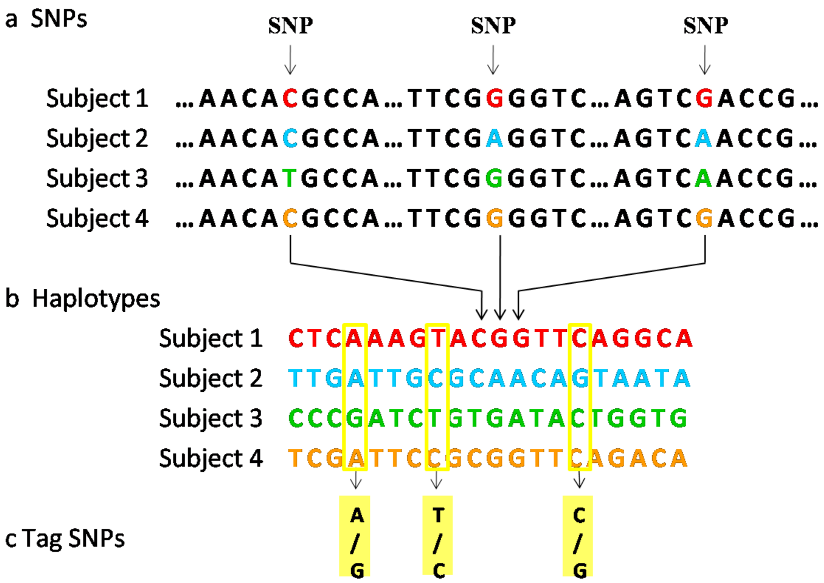
In fact, in many parts of our chromosomes, just a handful of haplotypes have been found. Patil et al. (2001) at Perlegen Sciences directly analyzed the haplotype patterns along the entire chromosome 21 (Patil, et al. 2001). They found that common haplotypes (with a frequency of greater than 10%) account for at least 80% of all haplotypes found in the entire sample. The resulting haplotypes contained between two and 114 SNPs, and the entire chromosome was covered by adjacent non-overlapping haplotypes.

Only a subset of SNPs is necessary to identify haplotypes uniquely. These SNPs are called tag SNPs. A set of tag SNPs contains non-redundant SNPs such that none of the SNPs included in this set would predict each other. The tag SNPs are representative of and capture the haplotype variation in the human genome. By genotyping an individual's tag SNPs, researchers are able to identify the collection of haplotypes in a person's DNA. The number of tag SNPs that contain most of the information about the patterns of genetic variations is estimated to be about 300,000 to 600,000, which is far fewer than the 10 million common SNPs (Kruglyak and Nickerson 2001) or the 3,000,000,000 base pairs. In Patil et al., only 2,793 SNPs needed to be genotyped to differentiate all common haplotypes that contain at least three SNPs, less than 12% of the total number of common SNPs on chromosome 21.

Figure 2.1 describes the relationship among SNPs, haplotypes, and tag SNPs. The same regions of a chromosome of four hypothetical individuals are presented as an example. Identical base pairs are printed in black, and SNPs are printed in color. The three adjacent SNPs in part *a* form a part of each individual's haplotype (in part *b*). Part *c* shows the three tag SNPs that are enough to identify uniquely the haplotype present in each individual, since only limited patterns of haplotypes exist in a given population.

Figure 2.1

SNPs, Haplotypes, and Tag SNPs


a SNPs
SNP
SNP
SNP
Subject 1 ...AACACGCCA...TTCGGGGTC...AGTCGACCG...
Subject 2 ...AACACGCCA...TTCGAGGTC...AGTCAACCG...
Subject 3 ...AACATGCCA...TTCGGGGTC...AGTCAACCG...
Subject 4 ...AACACGCCA...TTCGGGGTC...AGTCGACCG...
b Haplotypes
Subject 1 CTCAAAGTACGGTTCAGGCA
Subject 2 TTGATTGCGCAACAGTAATA
Subject 3 CCCGATCTGTGATACTGGTG
Subject 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A / G
T / C
C / G

The concepts of haplotypes and tag SNPs are used in genotype imputation. Genotype imputation is used to infer genotypes at non-genotyped markers based on known LD relationships. To do so, one needs to have a sample of individuals who are genotyped on a commercial array, which usually has 500,000 SNPs, as the basis of the imputation process. Then one needs another group of individuals who are genotyped at much higher density (i.e., the reference panel). The SNPs genotyped in the sampled individuals are only a small proportion of the SNPs genotyped in the individuals in the reference panel, but the genotyped SNPs in the reference panel do not have to include all of the SNPs genotyped in the sample. Genotype imputation programs impute genotypes for an individual by comparing the person's available genotypes to the genotypes of the individuals in the reference panel. First the SNPs genotyped in both the sample and the reference panel are used to identify the collection of haplotypes in each individual. Once the sets of haplotypes in a subject are identified, the missing genotypes could then be "filled in" simply by copying the genotypes at the corresponding loci from the reference panel. There is no genotype information on SNPs that are not genotyped in the reference panel, and those SNPs cannot be imputed.

Reference panels are publicly available. For example, the International HapMap Project ("A Haplotype Map of the Human Genome" 2005, Frazer, et al. 2007) has several million well-defined SNPs genotyped for 269 individuals: 30 trios of U.S. residents of northern and western European ancestry (CEU), 44 unrelated individuals from Tokyo (JPT), 45 unrelated Han Chinese from Beijing (CHB), and 30 trios from Ibadan, Nigeria (YRI). It is important to make sure that the reference panel selected is appropriate for the sample (i.e., the subjects in the reference panel should have similar haplotypes frequencies to the subjects in the sample), since haplotype frequencies differ widely between populations. When a reference panel from one ethnicity is used to impute variation in a sample taken from another ethnicity, the quality of imputation will be reduced somewhat. For samples where population stratification is likely to exist, using a pooled reference panel composed of all available ethnicities can give acceptable results (Chambers, et al. 2008).

2.1.2 Genotype Imputation

The imputation program used in this paper is MACH 1.0 (Li, Willer, Sanna and Abecasis 2009). MACH 1.0 was recommended by Pei et al. (2008) and Nothnagel et al. (2009) (Pei, Li, Zhang, Papasian and Deng 2008, Nothnagel, Ellinghaus, Schreiber, Krawczak and Franke 2009). They reported that MACH 1.0 has a relatively higher imputation consistency rate than other available imputation programs. It is also more user-friendly and generally requires less memory.

As to choosing the SNPs to impute, genetic researchers have imputed both genome-wide and selected regions of SNP data using different genotype imputation programs. Nothnagel et al. imputed genome-wide SNP data for 449 healthy blood donors of German descent (Nothnagel, et al. 2009). The SNPs were genotyped on three different commercial arrays, and they used one chip as the imputation basis and imputed the non-overlapping SNPs on the other two more densely genotyped arrays. They compared the imputed genotypes with the measured genotypes and reported a consistently high imputation consistency rate (>93%) for the four most widely used genotype imputation programs, namely BEAGLE, IMPUTE, MACH, and PLINK. Pei et al. imputed both simulated and real SNP data in selected regions on certain chromosomes (Pei, et al. 2008). Their findings about imputation consistency rates are consistent with Nothnagel et al.. They also pointed out that stronger LD, lower minor allele frequency (MAF) (i.e., the percentage of all living humans that have the rarer allele for this SNP, as opposed to the other more frequent nucleotide) for a non-genotyped marker, and higher marker genotyping density in the sample produce better imputation results.

MACH 1.0 expects the measured genotype data from the sample to be stored in a pedigree file and a matching data file. The two files are in Merlin format (Abecasis, Cherny, Cookson and Cardon 2002). For genotype imputation, MACH 1.0 requires a set of reference haplotypes as input in addition to the sample information. Phased haplotype information is encoded in two files: the SNP file and the haplotype file. A brief description of the input files is presented below:

1) The data file describes a variety of fields, including disease status information, quantitative traits and covariates, and marker genotypes. A simple MACH 1.0 data file simply lists names for a series of genetic markers in the sample. Each marker name appears on its own line prefaced by an “M” filed code. My data file looks like this:

   ```
   M Marker1
   M Marker2
   ......
   M Marker80
   ```

   In the MACH 1.0 command line, the data file is indicated with either the –d option (in short hand form) or the --datfile option (in long form).

2) The pedigree file stores the measured genotype data from the sample. It lists one individual per row. Each row starts with a family ID and an individual ID, followed by father and mother ID’s, and sex. Both father ID and mother ID should be zero since MACH 1.0 assumes individuals are unrelated. These initial columns are followed by a series of marker genotypes, each with two alleles. Alleles can be coded as either numbers

(1, 2, 3, and 4) or letters (A, G, C, and T). The pedigree file is indicated with either the –p option (in short hand form) or the --pedfile option (in long form).

3) The SNP file lists the markers in the phased haplotype file. It simply lists one marker name per line. The SNP file is indicated with either the –s option (in short hand form) or the --snps option (in long form).

4) The haplotype file lists the haplotypes from the reference panel. It lists one haplotype per line following the marker order indicated in the SNP file. The haplotypes can be prefaced by one or two optional labels followed by a series of single character alleles, one for each marker. Within each haplotype, spaces are ignored. The haplotype file is indicated with either the –h option (in short hand form) or the --haps option (in long form).

The data I used was obtained from the Framingham Heart Study (FHS) provided by the Genetics Analysis Workshop 16, problem two. The FHS research was supported by NHLBI Contract: 2 N01-HC-25195-06 and its contract with Affymetrix, Inc. for genotyping services (Contract No. N02-HL-6-4278). The FHS is a long-term, ongoing study of residents of the town of Framingham, Massachusetts. The study began in 1948 and is now on its third generation of participants. In the study, 6,476 individuals (in 942 pedigrees distributed among three generations and 188 singletons) from Framingham, Massachusetts were genotyped on both the GeneChip® Human Mapping 500K Array Set and the 50K Human Gene Focused Panel. MACH 1.0 assumes all sampled individuals are unrelated. As a consequence, 1,599 unrelated individuals (founders and singletons) were extracted from the original data to form my sample.

I selected three SNPs to impute. They are involved with nicotine dependence. Researchers have identified 51 SNPs as associated with nicotine dependence (Feng, et al. 2004, Bierut, et al. 2007, Saccone, et al. 2007, Voineskos, et al. 2007, Lou, et al. 2008, Schlaepfer, et al. 2008, Sherva, et al. 2008, Thorgeirsson, et al. 2008, Weiss, et al. 2008, Agrawal, et al. 2009, Hoft, et al. 2009). To obtain the imputation inconsistency rates for these SNPs, measured genotypes must be known. Out of the 51 nicotine dependence SNPs, only three have genotype data in the FHS dataset. The rs numbers of the three SNPs are: rs514743, rs2304297, and rs16969968. SNP rs514743 and SNP rs16969968 are on chromosome 15 and SNP rs2304297 is on chromosome 8. SNP rs514743 and SNP rs2304297 were located on the GeneChip® Human Mapping 500K Array Set, and SNP rs16969968 was located on the 50K Human Gene Focused Panel. It is obvious that the 500K array set is much more densely genotyped than the 50K panel. The reference panel used was obtained from the International HapMap Project, available at http://hapmap.ncbi.nlm.nih.gov.

To obtain the imputed genotypes for these three SNPs, I "masked" the measured genotypes of the SNPs and imputed their genotypes as if they were not available. Only SNPs in close physical proximity tend to be transferred together to offspring. In other words, SNPs that are far apart are more likely to have experienced recombination. As a consequence, the correlations between SNPs distant from the SNP to be imputed (i.e., the target SNP) are not large enough to justify including them in the imputation. Only a short region surrounding the target SNP needs to be considered in the imputation. For each individual nicotine dependence SNP, I selected 80 flanking markers as its imputation basis (40 on the left side and 40 on the right side). For the reference panel, I downloaded the phased haplotype data for each SNP only for the chromosomal region covered by its 80 flanking markers and the target SNP itself. SNP rs16969968 was originally located on the 50K panel, which was genotyped much less densely than the 500K array set. To obtain more accurate imputation results, I identified 80 flanking markers for SNP rs16969968 on the 500K array set using its physical position obtained from the 50K panel, and used them as the imputation basis instead. The population selected was CEU (U.S. residents with northern and western European ancestry, collected in 1980) to match our sample. The filter used was "Polymorphic in CEU" so that identical base pairs are excluded from the phased haplotype file. The physical distance (in kilo base pairs) of the chromosomal regions covered by the flanking markers of the three SNPs vary from 490,943 to 1,292,959, indicating different genotyping densities for different chromosomal regions. I ran MACH 1.0 with the default settings (mach1 –d sample.dat –p sample.ped –s ref.snps –h ref.haplos –r 50 --dosage --quality --greedy --geno --prefix out_file) and 50 iterations of the Markov sampler. The output file contains imputed genotypes for all SNPs genotyped only in the reference panel.

Next I introduce the likelihood ratio test allowing for errors (LRT-AE). My research area is to incorporate double sample information on the genotype data of a subset of the sampled individuals in genetic association studies. The loss of power due to inconsistently imputed genotypes may be reduced in imputation based GWAS using double sampling.

2.2 Likelihood Ratio Test Allowing for Errors

2.2.1 The Double Sampling Method

Tenenbein proposed the double sampling method (Tenenbein 1970). Suppose a researcher has two measuring devices to assign *N* sampling units to one of two mutually exclusive categories. One is a relatively inexpensive procedure but misclassifies units with non-zero error rate (fallible measuring device), and the other device is an expensive procedure but classifies units with perfect accuracy (infallible measuring device). Using only the fallible measuring device on all *N* sampling units results in classification errors and a biased estimate of the population parameter. A better estimate of *p*, the proportion of units which belong to one of the categories, can be obtained if the infallible device is used. However, the expense of using the infallible device on all *N* units in the sample may be high. Double sampling is presented as a compromise between the two extremes of using only fallible classification and using only infallible classification. To estimate *p*, first a random sample of *N* units is drawn from the population of interest. All *N* sampling units are classified using the fallible device. At the second stage, a subsample of *n* units is drawn from the main sample, and the true classifications for those *n* units are obtained using the infallible device. The unbiased maximum likelihood estimate (MLE) of *p* can be derived along with its asymptotic variance.

The double sampling method was then extended to the multinomial case (Tenenbein 1972), where more than two mutually exclusive categories that the sampling units are to be assigned to are present. Class frequencies for the units that are only classified by the fallible measuring device can be updated using the information provided by the double sampling units. Later in 1977, Dempster et al. introduced the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). Gordon et al. then applied the EM algorithm to the double sampling method and obtained more accurate class frequency estimates (Gordon, et al. 2004b).

In the application of the double sampling method to a genetics study, a SNP, which is a gene with two alleles, has three genotype categories; namely, the more common homozygote, the heterozygote, and the less common homozygote. I will treat a measured genotype as the "gold standard" measurement with a negligible classification error rate. I will treat an imputed genotype as the fallible measurement. While imputation requires little expense, genotyping all individuals in the sample on a dense marker set is expensive.

When the available marker set is considered not dense enough to be used in a GWAS, researchers usually choose one of two options: genotyping everyone for additional SNPs to be included in the study, which gives accurate results but with high expense; or imputing genotypes using a publicly available reference panel, which has higher classification error rates but at a much lower expense. I here suggest a method that at the first stage, impute genotypes for all *N* subjects in the sample. At the second stage, a subset of *n* subjects will be randomly chosen from the main sample and measured genotypes are obtained only for these *n* subjects. I merge the measured genotype data for the *n* subjects with the imputed genotype data for all *N* subjects in the main sample, and update the population genotype class frequencies. I hypothesize that tests incorporating double sample information have greater power to detect association, since the population genotype class frequency estimates are more accurate than the ones obtained using only imputation. The extra genotyping, however, does increase the cost. This extra cost should be taken into consideration and estimated when *n*, the number of individuals that need to be genotyped is specified.

The application of the double sampling method to imputation based genetic association studies can be particularly useful when a researcher has several studies in hand and the genotyped SNPs from different studies do not completely overlap. When a researcher wants to combine datasets generated in two different GWAS that use different genotyping platforms, there may be SNPs genotyped in one platform but not the other. A large amount of SNP data would be removed from the combined dataset if only SNPs genotyped in both platforms were used. To make the best use of the available data, researchers usually impute for each individual study the SNPs that are only genotyped in the other study. After imputing genotypes for non-overlapping SNPs for both studies, one can combine the datasets so that there is complete genotype data for the union of the SNPs from both studies. Since imputation inconsistency rates may be relatively high, double sampling is valuable because one can use the double sample information on SNPs genotyped in both studies to adjust genotype class frequencies.

Imputation inconsistencies, like genotyping errors, cause a decrease in statistical power to detect genetic association and produce biased estimates of population frequency parameters. Gordon and Ott considered the analysis of genetic data in the presence of genotyping errors (Gordon, et al. 2004a). They confirmed: (i) that there is no increase in type I error for certain tests of genetic association; (ii) that point estimates of SNP genotype frequencies are biased in the presence of genotyping errors; and (iii) that genotype errors lead to a loss of power to detect association between a disease allele and a marker. Recently, Gordon et al. produced a quantification of the loss in power for case/control studies of genetic association due to genotyping errors (Gordon, et al. 2002, Gordon, Haynes, Blumenfeld and Finch 2005). This quantification in different situations may be determined using the PAWE 3D web tool (available at http://linkage.rockefeller.edu/pawe3d/).

Imputation inconsistencies in imputation based GWAS have the same impact as genotyping errors on the power to detect association. A critical question is how one can use information about imputation inconsistencies to improve the power for genetic tests of association using case/control data. The analogous question for genotyping errors has been answered by Gordon et al. (Gordon, et al. 2004b). They used power simulations to show that in the presence of genotyping errors, the LRT-AE method improved the median power by 0.01 at the 5% significance level with equal costs for the LRT-AE method and the standard LRT method. The LRT-AE method incorporates double sample information on genotypes for a subsample of the case and control subjects and adjusts for the misclassification of the most commonly reported risk allele. It also produces unbiased estimates of the population frequency parameters and MLEs of the misclassification probabilities. The purpose of this work is to develop a statistical method that increases power to detect association (as compared to the standard method that considers only fallible data) in the presence of imputation inconsistencies. Our method assumes that double sample genotype data is available on a subsample of the case and control individuals. This method has three major advantages over the standard method that only considers the fallible data: its power can be equal to or greater than the standard method with little or small increase in the total cost; it provides unbiased estimates of population frequency parameters; and it provides MLEs of the imputation inconsistency probabilities.

I provide all notation and definitions for the mathematics presented in this work in Section 2.2.2. For all terms, the index i represents the phenotype category of a subject and is either 0 (control) or 1 (case). Throughout this work, I use prime superscripts to distinguish imputed genotype categories from measured genotype categories. For example, the indices j and j' represent the measured and the imputed genotype category of a subject, respectively, and range from one to three. Here I assume that I have SNP genotype data, but I note that our method can easily be applied to genes with three or more alleles and to haplotype data as well. I treat imputed genotype data as fallible and measured genotype data as infallible. I recognize that any classification technology has a non-zero error rate and that there is no such thing as a perfect classifier (Hochberg, 1977). Throughout this work, I use number superscripts (1, 2) to distinguish double sampled subjects (group I) from the subjects that only have imputed genotype data (group II).

2.2.2 Notation and Definitions for All Formulas Presented Throughout the Dissertation

Log-likelihood of the observed data and the LRT statistics

N: Sample size.

$n_{ijj'}^{(1)}$: Number of subjects with phenotype category *i*, measured genotype category *j* and imputed genotype category *j'*. These subjects are double sampled and are in group I.

$n_{ij'}^{(2)}$: Number of subjects with phenotype category *i* and imputed genotype category *j'*. These subjects have only imputed genotype data and are in group II.

Note: $\sum_i \sum_j \sum_{j'} n_{ijj'}^{(1)} + \sum_i \sum_{j'} n_{ij'}^{(2)} = N$.

$n_{ij'}$: Number of subjects with phenotype category *i* and imputed genotype category *j'* when double sampling is not available.

Note: $\sum_i \sum_{j'} n_{ij'} = N$.

Y_i: Event that a subject has phenotype category *i*, *i* = 0, 1.

X_j: Event that a subject has measured genotype category *j*, *j* = 1, 2, 3.

$X_{j'}$: Event that a subject has imputed genotype category *j'*, *j'*= 1, 2, 3.

$q_i = \Pr(Y_i)$: Sampling frequency of phenotype category *i*.

Note: $q_0 + q_1 = 1$.

$p_{j|i} = \Pr(X_j \mid Y_i)$: Measured population frequency of genotype category *j* for individuals with phenotype category *i*.

$p_{j'|i} = \Pr(X_{j'} \mid Y_i)$: Imputed population frequency of genotype category *j'* for individuals with phenotype category *i*.

Note: For each *i*, $\sum_j p_{j|i} = \sum_{j'} p_{j'|i} = 1$.

$p_j = \Pr(X_j)$: Measured population frequency of genotype category *j* under the null hypothesis that $p_{j|0} = p_{j|1} = p_j$.

$p_{j'} = \Pr(X_{j'})$: Imputed population frequency of genotype category *j'* under the null hypothesis that $p_{j'|0} = p_{j'|1} = p_{j'}$.

$\xi_{j'|j} = \Pr(X_{j'} \mid X_j)$.

Note: The parameter $\xi_{j'|j}$ is referred to as the misclassification parameter (Tenenbein 1972). Misclassification happens when *j'* ≠ *j*.

I make use of the double sampling data structure to determine the estimate of $\xi_{j'|j}$:

$$\hat{\xi}_{j'|j} = \frac{n_{jj'}^{(1)}}{n_{j}^{(1)}},$$

where

$n_{jj'}^{(1)} = \sum_{i} n_{ijj'}^{(1)}$: Number of group I subjects that have measured genotype category *j* and imputed genotype category *j'*.

$n_{j}^{(1)} = \sum_{j'} n_{jj'}^{(1)} = \sum_{i}\sum_{j'} n_{ijj'}^{(1)}$: Number of group I subjects that have measured genotype category *j*.

$\log\left(L_{0,\,ae}\right)$: Log-likelihood of the data under the null hypothesis, where genotype frequencies $p_{j|i}$ and $p_{j'|i}$ are constrained to be equal among different phenotype classes. That is, $p_{j|0} = p_{j|1} = p_{j}$ for all *j*'s and $p_{j'|0} = p_{j'|1} = p_{j'}$ for all *j'*'s.

$\log\left(L_{1,\,ae}\right)$: Log-likelihood of the data under the alternative hypothesis, where genotype frequencies $p_{j|i}$ and $p_{j'|i}$ are allowed to differ among different phenotype classes. That is, $p_{j|0}$ is not necessarily equal to $p_{j|1}$ for every *j* and $p_{j'|0}$ is not necessarily equal to $p_{j'|1}$ for every *j'*.

$\log\left(L_{0,\,std}\right)$: Log-likelihood of the data using only imputed genotypes, where genotype frequencies $p_{j'|i}$ are constrained to be equal among different phenotype classes. That is, $p_{j'|0} = p_{j'|1} = p_{j'}$ for all *j'*'s. Double sampling genotypes are not used.

$\log\left(L_{1,\,std}\right)$: Log-likelihood of the data using only imputed genotypes, where genotype frequencies $p_{j'|i}$ are allowed to differ among different phenotype classes. That is, $p_{j'|0}$ is not necessarily equal to $p_{j'|1}$ for every *j'*. Again, double sampling genotypes are not used.

EM algorithm estimates of true parameters

$p_{j|i}^{r}$: r^{th} step estimate of the parameter $p_{j|i}$.

p_{j}^{r}: r^{th} step estimate of the parameter p_{j} .

These two genotype frequency parameters are estimated using the EM algorithm developed by Dempster et al. (Dempster, Laird, and Rubin 1977).

$E[\ ]$: The expectation operator.

$I(\ )$: The indicator function.

2.2.3 Computation of the Log-likelihoods

I compute the log-likelihood of the observed data under both the null and the alternative hypothesis, allowing for imputation misclassifications. The null hypothesis is: $p_{j'|0} = p_{j'|1} = p_{j'}$ for all *j'*'s and $p_{j|0} = p_{j|1} = p_j$ for all *j*'s. The alternative hypothesis is $p_{j'|0} \neq p_{j'|1}$ for at least one *j'* and $p_{j|0} \neq p_{j|1}$ for at least one *j*. Under either hypothesis, by definition, the log-likelihood of the data is given by:

$$\log(L_{ae}) = \sum_i \sum_j \sum_{j'} n^{(1)}_{ijj'} \log(\Pr(Y_i, X_j, X_{j'})) + \sum_i \sum_{j'} n^{(2)}_{ij'} \log(\Pr(Y_i, X_{j'})), \tag{1a}$$

where the notation $\Pr(A, B, ...)$ is the probability of observing event *A* and event *B* and so forth. In equation (1a), the subscript *i* runs over all phenotype classifications and the subscripts *j* and *j'* run over all genotype classifications.

When there is no double sampling data or when one assumes that there is no genotype misclassification in the data, equation (1a) reduces to:

$$\begin{aligned} \log(L_{std}) &= \sum_i \sum_{j'} n_{ij'} \log(\Pr(Y_i, X_{j'})) \\ &= \sum_i \sum_{j'} n_{ij'} \log(\Pr(Y_i)\Pr(X_{j'} \mid Y_i)) \\ &= \sum_i \sum_{j'} n_{ij'} \log(q_i p_{j'|i}) && (1b) \\ &= \sum_i \sum_{j'} n_{ij'} (\log(q_i) + \log(p_{j'|i})). \end{aligned}$$

A key assumption in my work is that the imputation process (and hence imputation inconsistency rates) is independent of one's disease status (i.e., phenotype category). This assumption is reasonable because imputation programs (specifically, MACH 1.0) do not require phenotype status as input. I confirm the validity of this assumption using simulation studies (see the Results section).

From this assumption,

$$\Pr(X_{j'} \mid Y_i, X_j) = \Pr(X_{j'} \mid X_j).$$

It follows that:

$$\begin{aligned} \Pr(Y_i, X_j, X_{j'}) &= \Pr(X_{j'} \mid Y_i, X_j)\Pr(Y_i, X_j) \\ &= \Pr(X_{j'} \mid X_j)\Pr(X_j \mid Y_i)\Pr(Y_i) && (2) \\ &= \xi_{j'|j}\, p_{j|i}\, q_i . \end{aligned}$$

Using equation (2) and the fact that

$$\Pr(Y_i, X_{j'}) = \sum_j \Pr(Y_i, X_j, X_{j'}) \,, \tag{3}$$

I may rewrite the log-likelihood (1a) as:

$$\log(L_{ae}) = \sum_i \sum_j \sum_{j'} n_{ijj'}^{(1)} \log(\xi_{j'|j}\, p_{j|i} q_i) + \sum_i \sum_{j'} n_{ij'}^{(2)} \log(\sum_u \xi_{j'|u}\, p_{u|i} q_i) \,, \tag{4}$$

where I have replaced the index *j* in equation (3) by the index *u* in equation (4) for clarity. The index *u* ranges from one to three.

It follows from equation (4) that the log-likelihoods of the data under both the null and the alternative hypothesis are completely determined by the imputation misclassification parameters $\xi_{j'|j}$, the infallible genotype frequency parameters under the null and the alternative hypothesis p_j and $p_{j|i}$, the sampling frequency of the phenotype category *i* q_i, and the group counts $n_{ijj'}^{(1)}$ and $n_{ij'}^{(2)}$. The LRT-AE software uses the EM algorithm to determine the MLEs of the parameters. Parameter estimates are updated until the absolute difference between the sum of the n^{th} step and $n+1^{th}$ step estimates is no greater than 10^{-9} (summed over all parameters). That is, the stopping condition for parameter estimation is $\left| v^{r+1} - v^r \right| < 10^{-9}$. The log-likelihood of the data under each of the hypotheses is then computed.

The test of H_0 versus H_1 is a likelihood ratio test that is called LRT-AE in Gordon et al. (Gordon, et al. 2004b) and is given by:

$$LRT_{ae} = 2\left(\log\left(L_{1,\,ae}\right) - \log\left(L_{0,\,ae}\right)\right). \tag{5a}$$

Asymptotically, the null distribution of the LRT_{ae} statistic is a chi-square distribution with $K - 1$ degrees of freedom (DF), where the DF is $K - 1$ for a marker locus with *K* genotype categories ($K = 3$ for a SNP).

To compare the performance of the LRT_{ae} test statistic that corrects for imputation misclassifications with the standard LRT statistic, denoted by LRT_{std}, which does not make any correction, I compute log-likelihood solely from the observed data. That is,

$$LRT_{std} = 2\left(\log\left(L_{1,\,std}\right) - \log\left(L_{0,\,std}\right)\right), \tag{5b}$$

where the log-likelihoods under the null and the alternative hypothesis are computed using the estimates $\hat{p}_{j'|i} = \dfrac{n_{ij'}}{n_i}$, $\hat{p}_{j'} = \dfrac{n_{j'}}{N}$, $q_i = \dfrac{n_i}{N}$ that are then substituted into equation (1b).

For small samples or in situations where the number of observations in a particular genotype category is small, the asymptotic null distribution may not be valid. In such a situation, the *p*-values for the LRT-AE statistic should be computed using a permutation distribution.

Gordon et al. advised that permutation *p*-values should be computed and used even for relatively large samples (Barral, Haynes, Stone and Gordon 2006). When double sampling information is available for a subsample, the phenotype status for all individuals are randomly permuted, keeping the total number of cases and controls fixed in each replicate. The permutation *p*-value is then the proportion of replicates for which the LRT-AE statistic exceeds the observed LRT-AE statistic. Exact confidence intervals for the permutation *p*-values are computed using the method implemented in the BINOM software. The same procedures for the standard LRT method are performed in the LRT-AE software.

The LRT-AE software requires three input files:

1) The phenotype and genotype description file: This file contains the information identifying the phenotypes and genotypes in the fallible and infallible data files (items 2 and 3). The first column in this file indicates the nature of the categorical variable in the second column for the corresponding row (either phenotype or genotype). The second column is the name of the variable corresponding to the symbol in the first column. This variable name is also used in the first line of both the fallible and infallible data files to indicate the nature of the corresponding columns in those files. The third column, which is optional, indicates the symbol that is used to represent missing data in the fallible and infallible data files. The default values for missing phenotype and genotype data are "–1" and "0" respectively.

   My description file looks like this:

   ```
   P Disease
   A SNP2304297
   ```

2) Fallible data file: This file contains the genotype classifications for all individuals as measured with the fallible method. The format for this file is as follows:

   ```
   Ind_ID Order_of_genotype_data
   ```

   There are two key formatting issues regarding this file and the infallible data file (item 3). They are:

   a) The first line of the file always consists of the order of the phenotype and genotype data. The order is determined by using the variable names provided in the description file (item 1).

   b) The first column of each row is always the individual ID (Ind_ID), which must be an alphanumeric string of characters.

   My fallible data file looks like this:

   ```
   Disease SNP2304297
   A1 1 2 1
   ......

   A1599 0 2 1
   ```

3) Infallible data file: This file contains genotype classifications for individuals as measured with the infallible method. Note that the list of individuals in this file is a subset of the list of individuals in the fallible data file. An individual is listed in this file only if the individual has an infallible measurement for at least one of the genotype data variables in the description file.

I use simulation studies to examine how much the use of double sampling decreases the impact of inconsistently imputed genotypes on the power to detect association in imputation based GWAS. The simulation procedure is detailed in the next section.

2.3 Simulation Studies

In GWAS, the SNP genotype distributions in cases are compared to the corresponding SNP genotype distributions in controls to see whether there is a significant difference. I use simulation studies to compare the power to detect association using the LRT-AE to the power of the standard LRT, which does not correct for imputation inconsistencies.

I use a factorial design to assess the importance of the specified factors with regard to the relative effectiveness of the double sampling approach. The settings of the simulation parameters and inheritance models are presented in Table 2.3. There are a total of 12 simulation settings for my study of the null hypothesis. There are 24 settings for the study of the alternative hypothesis. I set the imputed SNPs as the disease susceptibility SNPs.

Table 2.3

Values of Parameters Used in the Creation of the Simulated Data

Parameter	Value
R_1 - Under the Null Hypothesis	1
R_1 - Under the Alternative Hypothesis	1.2
Mode of Inheritance	Dominant, Multiplicative
P_d (P_m)	0.25, 0.36
K	0.33, 0.50
Proportion that is double sampled	0.25, 0.50, 0.75

R_1 = Heterozygote relative risk; R_2 = Disease allele homozygote relative risk; Mode of Inheritance = Dominant ($R_2 = R_1$) and multiplicative ($R_2 = R_1^2$); $P_d = P_m$ = Disease (marker minor) allele frequency and K = Disease prevalence.

The disease prevalence K is selected to be 0.33 and 0.50 so that there are a sufficient number of subjects assigned to the case group. These fractions are common in case/control studies (Glasser, et al. 1998, Dye, Scheele, Dolin, Pathania and Raviglione 1999, Mukadi, Maher and Harries 2001). Assuming the common allele and the disease susceptibility allele are in Hardy-Weinberg equilibrium (HWE), the disease prevalence (K) in the general population can be expressed in terms of the genotype penetrances f_i's and the disease susceptibility allele frequency P_d as:

$$K = (1-P_d)^2 f_0 + 2P_d(1-P_d)f_1 + P_d^{\ 2} f_2,$$

where f_i = Pr (affection | i copies of the disease allele) for i = 0, 1, 2.

The disease prevalence K may also be written in terms of the genotype relative risks R_i's (i = 1, 2), where $R_i = \frac{f_i}{f_0}$ is obtained by dividing the genotype penetrances by the penetrance value in homozygotes for the common allele (the baseline disease penetrance):

$$K = (1-P_d)^2 f_0 + 2P_d(1-P_d) f_0 R_1 + P_d{}^2 f_0 R_2,$$

where R_1 = heterozygote relative risk and R_2 = disease allele homozygote relative risk.

Under the null hypothesis, both R_1 and R_2 are equal to one (i.e., all penetrances are equal). Under the alternative hypothesis, R_1 can be determined as a function of R_2 and the mode of inheritance (MOI):

1) For a dominant model, $R_2 = R_1$ and $R_1 > 1$.

2) For a multiplicative model, $R_2 = R_1{}^2$ and $R_1 > 1$.

The simulation program used in this research is FASTSLINK (Ott 1989, Cottingham, Idury and Schaffer 1993), which is similar to SLINK except that it can handle a larger number of pedigrees.

FASTSLINK requires three input files:

1) Simdata.dat: This is a standard LINKAGE data file (Lathrop and Lalouel 1984), which defines the locus systems. Values of the f_i's are estimated for different combinations of K, R_i, and P_d values:

$$f_0 = \frac{K}{(1-P_d)^2 + 2P_d(1-P_d)R_1 + P_d{}^2 R_2}$$

$$f_1 = f_0 \times R_1$$

$$f_2 = f_0 \times R_2 .$$

Since the imputed SNPs are assumed to be the disease alleles, disease marker haplotype frequencies are also determined. In this file we provide the haplotype frequencies rather than disease and marker allele frequencies. A sample data file for study setting "dominant MOI, R_1 = 1.2, P_d = 0.25, and K = 0.33" is presented in Table 2.4.

Table 2.4

FASTSLINK Code for Simdata.dat

```
2    0    0    5   << NO. OF LOCI, RISK LOCUS, SEXLINKED (IF 1),
PROGRAM
0    0.0  0.0  1    << MUT LOCUS, MUT MALE, MUT FEM, HAP FREQ
(IF 1)
1    2
1    2    << AFFECTION, NO. OF ALLELES
1    <<   N0. OF LIABILITY CLASSES
0.364137931    0.364137931    0.303448276
3    2     << ALLELE NUMBERS, NO. OF ALLELES
0.25 0    0    0.75 << HAP FREQ
0    0    << SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
0    << RECOMBINATION FRACTION
1    0.1  0.45 << REC VARIED, INCREMENT, FINISHING VALUE
```

2) Slinkin.dat: This is a file identifying various parameter values required in the simulation, such as the number of replicates (*rep*) specified (*rep* = 1,000 in our case), and a seed for the random number generator which must be updated each time FASTSLINK runs. The seed should be an integer between 1 and 30,000, and larger numbers (>25,000) are recommended to produce better results. I updated the seed for each setting of the parameter values by selecting a random number between 25,000 and 30,000.

3) Simped.dat: This is a standard LINKAGE pedigree file with an additional column inserted after the last phenotype column. This additional column contains the availability code, which controls what types of phenotypes are written to the output file.

   Since FASTSLINK requires pedigree data, dummy parents are created for each individual and are used in the simulations. That is, the pedigree structure is a trio where only the children have genotypes (availability code is 1, meaning that marker genotypes are available and phenotypes at each marker need to be assigned). Phenotypes for each child are generated conditional on their respective genotypes and the inheritance models provided in simdata.dat. The availability code for dummy parents is 0, meaning that marker genotypes are not available for parents, and phenotype “unknown” is assigned at each marker.

A total of 1,599,000 ($N \times rep$) trios are simulated for the 1,599 unrelated individuals for each study setting. FASTSLINK stores the simulated trios in the file pedfile.dat. Dummy parents are then removed from this file and only four columns of information are kept: individual ID, the simulated phenotype (disease status), and the measured genotypes (written in two columns). The simulated phenotypes in these modified pedfile.dat files are then used in the LRT-AE software.

In this research, each of the three nicotine dependence SNPs is treated as the causal disease SNP in turn. I calculate both the analytical power and the empirical power for both the LRT-AE method and the standard LRT method.

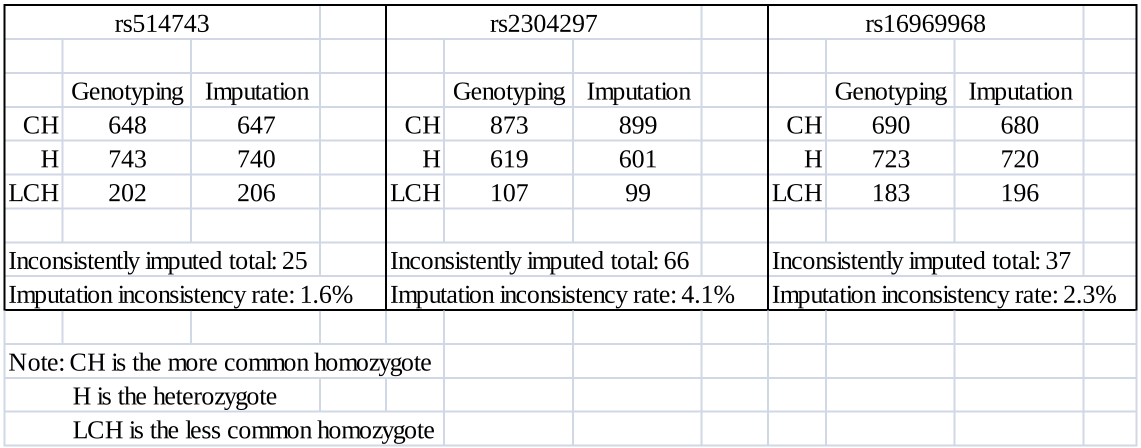
Chapter 3 Results

3.1 Imputation Results

The imputed genotypes for the three SNPs were extracted from the output files of FASTSLINK and compared to the measured genotypes in the FHS dataset. The imputation inconsistency rates for all three SNPs are reported in Table 3.1.

Table 3.1

Imputation Inconsistency Rates for All Three SNPS



rs514743			rs2304297			rs16969968		
	Genotyping	Imputation		Genotyping	Imputation		Genotyping	Imputation
CH	648	647	CH	873	899	CH	690	680
H	743	740	H	619	601	H	723	720
LCH	202	206	LCH	107	99	LCH	183	196
Inconsistently imputed total: 25			Inconsistently imputed total: 66			Inconsistently imputed total: 37		
Imputation inconsistency rate: 1.6%			Imputation inconsistency rate: 4.1%			Imputation inconsistency rate: 2.3%		

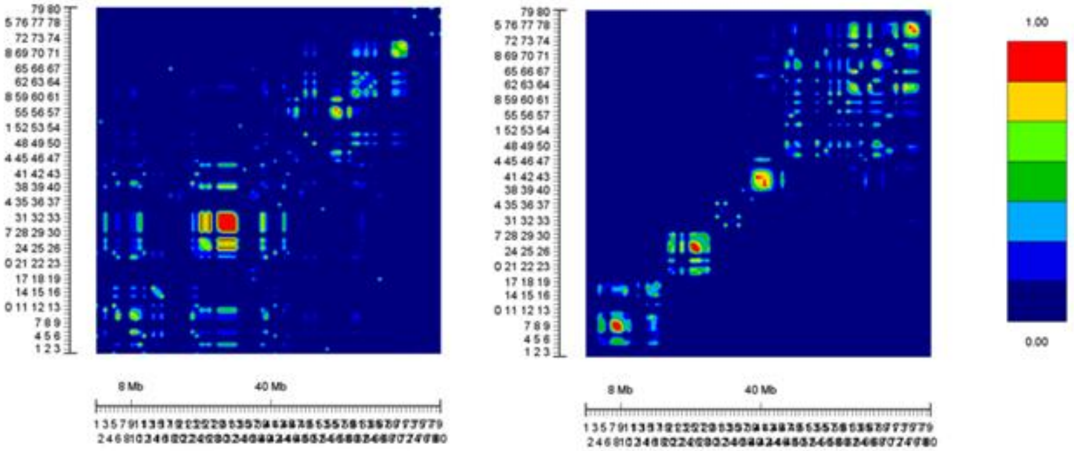
Note: CH is the more common homozygote
H is the heterozygote
LCH is the less common homozygote

The imputed genotypes were relatively consistent with the measured genotypes, with imputation inconsistency rates for the three SNPs ranging from 1.6% to 4.1%. SNP rs2304297 had a larger imputation inconsistency rate than the other two SNPs. To further investigate this, I compared the LD patterns for the chromosomal regions covered by the 80 flanking makers of SNP rs514743 and SNP rs2304297 and the two SNPs themselves using the program Graphical Overview of Linkage Disequilibrium (GOLD) (Abecasis and Cookson 2000), available at http://www.sph.umich.edu/csg/abecasis/GOLD. The plots shown in Figure 3.1 came from that program.

Figure 3.1

LD Plots for SNP rs514743 and rs2304297

	rs514743	rs2304297
Leftmost	76,498,858 (0)	42,235,549 (0)
SNP	76,671,282 (172,424)	42,727,356 (491,807)
Rightmost	76,990,629 (491, 771)	43,528,508 (1,292,959)

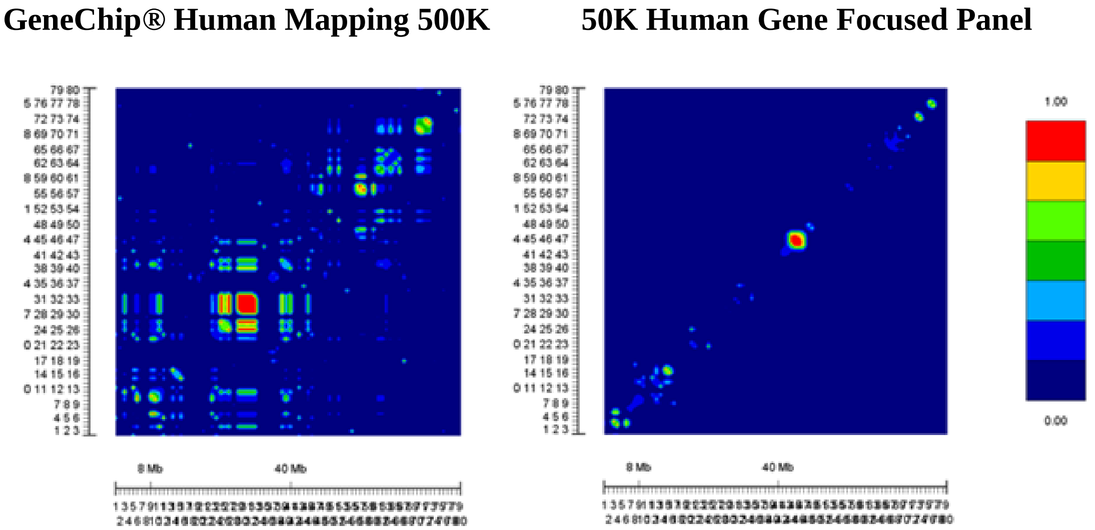


A value of r^2 close to one indicates strong LD among the SNPs within the region, and is printed in red. A value of r^2 equal to zero means that there is no LD among the SNPs, and is printed in blue. The left plot is for SNP rs514743 and the right plot is for SNP rs2304297. The physical positions of the leftmost flanking marker, the target SNP, and the rightmost flanking marker are listed above each plot, with their relative physical positions as in my sample enclosed in parentheses. For both SNPs, their physical positions were at about 1/3 of the total length of the imputation basis regions. When locating the target SNPs in the LD plots, SNP rs514743 is in a region marked in red, indicating that it is in very strong LD with the nearby flanking markers. SNP rs2304297, however, is in relatively weaker LD with the surrounding SNPs. This may explain why the imputation inconsistency rate of SNP rs2304297 was higher than that of SNP rs514743. Consistent with Pei et al., stronger LD results in higher imputation consistency rates (Pei, et al. 2008).

Another question in genotype imputation is what conditions are associated with accurate imputation. Originally I located SNP rs16969968 on the 50K Human Gene Focused Panel. I extracted 80 flanking markers from the panel and got a 26% imputation inconsistency rate. The LD plot for SNP rs16969968 in the 50K panel is presented in Figure 3.2 along with the plot in the GeneChip® Human Mapping 500K Array Set. The physical distance covered by the same number of flanking markers (i.e., 80) in the 50K panel is almost 8 times the length of the region covered in the 500K array set, and there is almost no LD among SNPs in the large region in the 50K panel as shown in the right plot in Figure 3.2. Consistent with Pei et al., higher marker density results in higher imputation consistency rates. The 2.3% imputation inconsistency rate using the 500K array set, compared to 26% using the 50K panel for the same SNP, suggested that if one's sample is genotyped at low density, no satisfactory imputation results may be obtained.

Figure 3.2

LD Plots for SNP rs16969968 in the 500K Array and 50K Panel



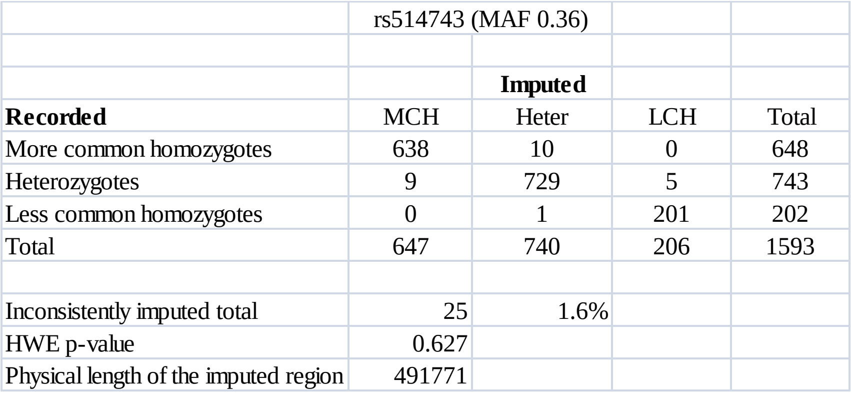
GeneChip® Human Mapping 500K

50K Human Gene Focused Panel

Not only should one look at the overall imputation inconsistency rates, but also one should study the pattern of imputation inconsistency rates since some types of imputation inconsistencies are more costly than others. When the imputation misclassification matrices for individual SNPs are considered, there are six different types of imputation inconsistencies: classifying the more common homozygote as the heterozygote or the less common homozygote, classifying the heterozygote as the more common homozygote or the less common homozygote, and classifying the less common homozygote as the more common homozygote or the heterozygote. According to Kang et al., the increase in the sample size required resulting from either recording the more common homozygote as the less common homozygote or recording the more common homozygote as the heterozygote becomes indefinitely large as the MAF goes to zero (Kang, Gordon and Finch 2004). That makes these two types of imputation inconsistencies more costly than other types of inconsistencies. In order to have a specific picture of what the imputation inconsistency matrices for the three SNPs look like, and what are the percentages of the more costly genotype imputation inconsistencies, I present the detailed imputation inconsistency matrices below in Table 3.2 to Table 3.4. The right two cells in the first row of each matrix are the more costly imputation inconsistencies.

Table 3.2

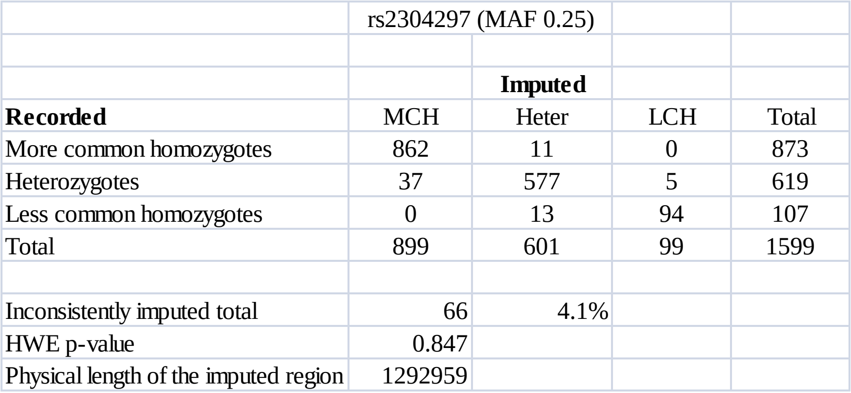
Imputation Misclassification Matrix for SNP rs514743



	rs514743 (MAF 0.36)			
		Imputed		
Recorded	MCH	Heter	LCH	Total
More common homozygotes	638	10	0	648
Heterozygotes	9	729	5	743
Less common homozygotes	0	1	201	202
Total	647	740	206	1593
Inconsistently imputed total	25	1.6%		
HWE p-value	0.627			
Physical length of the imputed region	491771			

Table 3.3

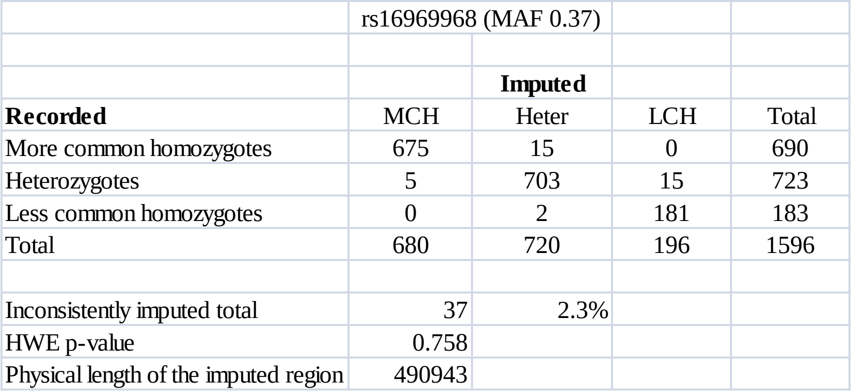
Imputation Misclassification Matrix for SNP rs2304297



	rs2304297 (MAF 0.25)			
		Imputed		
Recorded	MCH	Heter	LCH	Total
More common homozygotes	862	11	0	873
Heterozygotes	37	577	5	619
Less common homozygotes	0	13	94	107
Total	899	601	99	1599
Inconsistently imputed total	66	4.1%		
HWE p-value	0.847			
Physical length of the imputed region	1292959			

Table 3.4

Imputation Misclassification Matrix for SNP rs16969968



	rs16969968 (MAF 0.37)			
		Imputed		
Recorded	MCH	Heter	LCH	Total
More common homozygotes	675	15	0	690
Heterozygotes	5	703	15	723
Less common homozygotes	0	2	181	183
Total	680	720	196	1596
Inconsistently imputed total	37	2.3%		
HWE p-value	0.758			
Physical length of the imputed region	490943			

Note 1: The rows report the measured genotypes (the "infallible" measure), and the columns report the imputed genotypes (the "fallible" measure). MCH stands for the more common homozygote; that is, they are genotypes composed of two common alleles (population frequency greater than 50%). Heter stands for the heterozygote. LCH stands for the less common homozygote. HWE stands for the Hardy-Weinberg equilibrium test, and a *p*-value less than 5% rejects the null hypothesis that the SNP considered is in Hardy-Weinberg equilibrium.

Note 2: There are six people missing genotype data for SNP rs514743, and three for SNP rs16969968, making the numbers of total individuals in the comparison tables smaller than 1,599 in Table 3.2 and Table 3.4.

Although there were no imputation inconsistencies such that the measured more common homozygote was imputed as the less common homozygote for the three SNPs, imputing the more common homozygote as the heterozygote accounted for a considerable percentage of the imputation inconsistencies. This would inflate the sample size required to detect association.

In addition, even though small overall imputation inconsistency rates were observed, as I pointed out in the first chapter of this dissertation, a 2% imputation inconsistency rate could increase the sample size required by 10% to 26% to obtain comparative power and significance level. If researchers treat imputed genotypes as if they were true genotypes in imputation based GWAS, they would necessarily lose power.

Next I compare the power of the standard LRT method with the power of the LRT-AE method using both empirical power and asymptotic analytic power.

3.2 Simulation Results

3.2.1 Level of Significance

Non-differential genotyping errors/imputation inconsistencies do not have significant effect on type I error rates. I simulated 1,000 replicates for phenotypes for each combination of parameter settings under the null hypothesis that the SNP is not associated with the disease to check the empirical significance level in each situation. For each replicate, phenotypes were permuted 1,000 times. The permutation *p*-values of both the standard LRT and the LRT-AE were compared to 0.05 and 0.01. When the permutation *p*-value was less than 0.05/0.01, the hypothesis was rejected. The empirical type I error rates of both tests were estimated by the ratio of the number of significant replicates to the total number of replicates (i.e., 1,000).

I present empirical type I error rates of both the standard LRT and the LRT-AE for all 12 study settings under the null hypothesis and the 95% confidence intervals for type I error rates of the LRT-AE in Table 3.5. All 95% confidence intervals contain the nominal values of the type I error rates for each situation studied.

Table 3.5

Empirical Type I Error Rates

Type I Error Rates										
			α=0.05				α=0.01			
P_d (P_m)	*K*	*DS*	LRTstd	LRTae	LLae	ULae	LRTstd	LRTae	LLae	ULae
0.25	0.33	0.25	0.06	0.05	0.04	0.06	0.01	0.01	0.01	0.02
0.36	0.33	0.25	0.05	0.05	0.04	0.07	0.02	0.02	0.01	0.02
0.25	0.50	0.25	0.04	0.03	0.02	0.05	0.01	0.01	0.00	0.01
0.36	0.50	0.25	0.06	0.05	0.03	0.06	0.01	0.01	0.01	0.02
0.25	0.33	0.50	0.06	0.05	0.04	0.06	0.01	0.01	0.00	0.01
0.36	0.33	0.50	0.05	0.06	0.04	0.07	0.02	0.02	0.01	0.02
0.25	0.50	0.50	0.04	0.04	0.03	0.06	0.01	0.01	0.00	0.01
0.36	0.50	0.50	0.06	0.05	0.04	0.07	0.01	0.02	0.01	0.02
0.25	0.33	0.75	0.06	0.05	0.04	0.06	0.01	0.01	0.01	0.02
0.36	0.33	0.75	0.05	0.05	0.03	0.06	0.01	0.02	0.01	0.02
0.25	0.50	0.75	0.04	0.04	0.03	0.05	0.01	0.01	0.00	0.01
0.36	0.50	0.75	0.06	0.05	0.04	0.07	0.01	0.02	0.01	0.02

Note: Heterozygote relative risk is one for all study settings under the null hypothesis that the SNP is not associated with the disease.

P_d = P_m = Disease (marker minor) allele frequency; *K* = Disease prevalence; *DS* = Double sample proportion; α = Significance level; LRT_{std} = Standard LRT test where *p*-values are computed using permutation; LRT_{ae} = LRT-AE test where *p*-values are computed using permutation; LL_{ae} = Lower limit of the 95% confidence interval for the type I error rate of the LRT-AE where *p*-values are computed using permutation; UL_{ae} = Upper limit of the 95% confidence interval for the type I error rate of the LRT-AE where *p*-values are computed using permutation.

3.2.2 Empirical Power Calculation

I simulated 1,000 replicates for the disease status for each combination of the parameter settings. For each replicate, phenotypes were permuted 1,000 times. The permutation *p*-values were compared to 0.05 and 0.01. When the permutation *p*-value was less than 0.05/0.01, the hypothesis was rejected. The power of both the standard LRT and the LRT-AE was estimated by the ratio of the number of significant replicates to the total number of replicates (i.e., 1,000).

I report empirical power of the standard LRT and the LRT-AE for all 24 study settings under the alternative hypothesis and the 95% confidence intervals for the power of the LRT-AE in Table 3.6.

Table 3.6

Empirical Power of the Standard LRT and the LRT-AE

Power and 95% Confidence Intervals											
				α=0.05				α=0.01			
P_d (P_m)	*K*	MOI	*DS*	LRTstd	LRTae	LLae	ULae	LRTstd	LRTae	LLae	ULae
0.25	0.33	D	0.25	0.55	0.45	0.42	0.48	0.30	0.22	0.19	0.25
0.36	0.33	D	0.25	0.59	0.54	0.51	0.57	0.35	0.30	0.27	0.33
0.25	0.50	D	0.25	0.85	0.74	0.71	0.77	0.67	0.49	0.46	0.52
0.36	0.50	D	0.25	0.88	0.84	0.82	0.86	0.71	0.65	0.62	0.68
0.25	0.33	M	0.25	0.80	0.67	0.64	0.70	0.56	0.39	0.36	0.42
0.36	0.33	M	0.25	0.87	0.84	0.82	0.86	0.70	0.64	0.61	0.67
0.25	0.50	M	0.25	0.99	0.95	0.94	0.96	0.94	0.85	0.83	0.87
0.36	0.50	M	0.25	0.99	0.99	0.98	1.00	0.97	0.95	0.94	0.96
0.25	0.33	D	0.50	0.55	0.56	0.53	0.59	0.30	0.30	0.27	0.33
0.36	0.33	D	0.50	0.59	0.58	0.55	0.61	0.36	0.35	0.32	0.38
0.25	0.50	D	0.50	0.85	0.84	0.82	0.86	0.67	0.68	0.65	0.71
0.36	0.50	D	0.50	0.88	0.89	0.87	0.91	0.71	0.70	0.67	0.73
0.25	0.33	M	0.50	0.80	0.79	0.76	0.82	0.57	0.57	0.54	0.60
0.36	0.33	M	0.50	0.87	0.88	0.86	0.90	0.71	0.70	0.67	0.73
0.25	0.50	M	0.50	0.99	0.99	0.98	1.00	0.94	0.94	0.93	0.95
0.36	0.50	M	0.50	0.99	1.00	1.00	1.00	0.97	0.97	0.96	0.98
0.25	0.33	D	0.75	0.55	0.61	0.58	0.64	0.31	0.36	0.33	0.39
0.36	0.33	D	0.75	0.58	0.60	0.57	0.63	0.35	0.36	0.33	0.39
0.25	0.50	D	0.75	0.85	0.88	0.86	0.90	0.67	0.72	0.69	0.75
0.36	0.50	D	0.75	0.88	0.90	0.88	0.92	0.71	0.73	0.70	0.76
0.25	0.33	M	0.75	0.80	0.84	0.82	0.86	0.58	0.61	0.58	0.64
0.36	0.33	M	0.75	0.87	0.89	0.87	0.91	0.71	0.72	0.69	0.75
0.25	0.50	M	0.75	0.99	0.99	0.98	1.00	0.94	0.97	0.96	0.98
0.36	0.50	M	0.75	0.99	0.99	0.98	1.00	0.96	0.97	0.96	0.98

Note: Heterozygote relative risk is 1.2 for all study settings.

$P_d = P_m$ = Disease (marker minor) allele frequency; K = Disease prevalence; MOI = Mode of inheritance, D stands for dominant MOI and M stands for multiplicative MOI; DS = Double sample proportion; α = Significance level; LRT_{std} = Standard LRT test where p-values are computed using permutation; LRT_{ae} = LRT-AE test where p-values are computed using permutation; LL_{ae} = Lower limit of the 95% confidence interval for the power of LRT-AE where p-values are computed using permutation; UL_{ae} = Upper limit of the 95% confidence interval for the power of LRT-AE where p-values are computed using permutation.

I compare the power of the standard LRT with the power of the LRT-AE for all study settings under the alternative hypothesis in Figure 3.3 to Figure 3.5. The power of the LRT-AE with 25% double sampling is less than the power of the standard LRT. The power of the LRT-AE with 50% double sampling is roughly equal to the power of the standard LRT. Finally, the power of the LRT-AE with 75% double sampling is greater than the power of the standard LRT.

Figure 3.3

Power Comparison for 25% Double Sampling

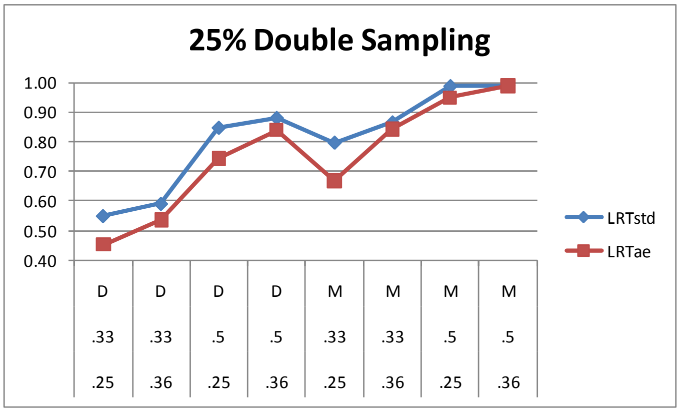


Figure 3.4

Power Comparison for 50% Double Sampling

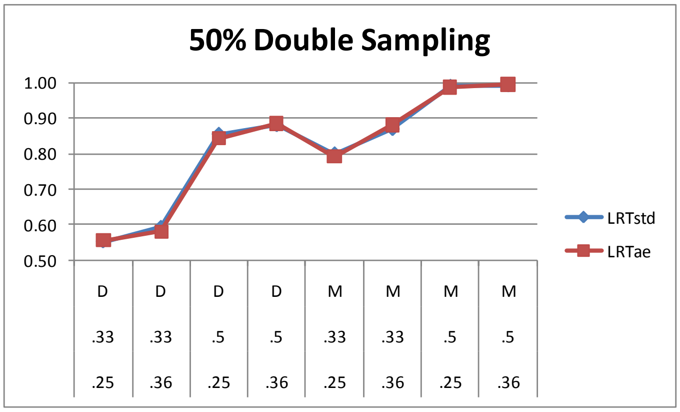
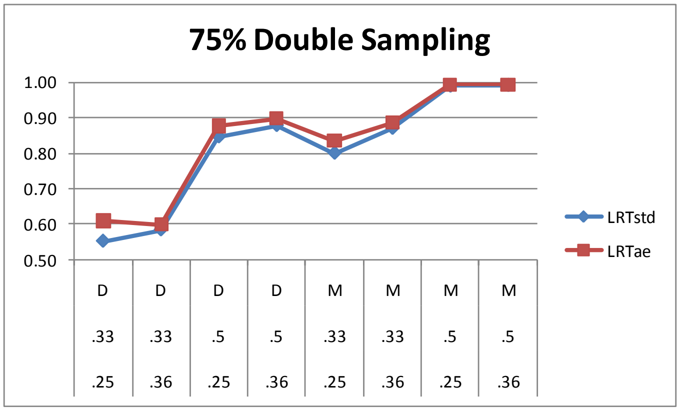


Figure 3.5

Power Comparison for 75% Double Sampling



3.2.3 Check for Differential Imputation Inconsistency Rates

Next I verify that there are non-differential genotype imputation inconsistency rates for cases and controls. Differential genotyping errors (i.e., genotyping errors that occur with different frequencies in the cases and the controls) can significantly increase type I error (Moskvina, Craddock, Holmans, Owen and O'Donovan 2006, Ahn, Gordon and Finch 2009). Since genotype imputation inconsistencies have the same effect as genotyping errors, problems may occur if we ignore differential misclassification if it does exist. I hypothesize that the imputation inconsistency rates for cases and controls are non-differential because major imputation programs do not take one's disease status into consideration when imputing missing genotypes. I calculated the imputation inconsistency rates for cases and controls separately under both the null and the alternative hypothesis, and tested whether there were differential imputation inconsistency rates.

Under the null hypothesis that the SNP is not associated with the disease, cases and controls have similar genotype frequencies. I analyzed one replicate of one study setting (disease allele frequency 0.36, disease prevalence 0.33) under the null hypothesis and imputed genotypes separately within the case and control groups. The overall imputation inconsistency rate for the 1,599 unrelated subjects was 1.6%. The imputation inconsistency rate was 1% for cases and 1.7% for controls. The difference in the imputation inconsistency rates is not significant at the 5% significance level ($\chi^2 = 1.2$, $p = 0.27$). I performed the same procedure for the other five study settings under the null hypothesis. The differences in the imputation inconsistency rates between case and control groups were not significant at the 5% significance level (data not shown).

Under the alternative hypothesis that the SNP is associated with the disease, cases and controls have different genotype frequencies. I analyzed one replicate of each study setting under the alternative hypothesis for non-differential imputation inconsistency rates. None of the study settings was significant at the 5% significance level (data not shown).

3.3 Estimation of Imputation Inconsistency Rates

An important advantage of the LRT-AE procedure is that it also gives the MLEs of the imputation inconsistency rates. Since Kang et al. found that misclassification of the more common homozygote to the less common homozygote and misclassification of the more common homozygote to the heterozygote were the two errors that were the most disadvantageous, I focused on these imputation inconsistency rates. The estimate of the rate of misclassifying the more common homozygote to the less common homozygote is always zero. Table 3.8 reports the mean and standard deviation of the rate of misclassifying the more common homozygote to the heterozygote for each situation.

Table 3.7

Summary Statistics for the Rate of Misclassifying the More Common Homozygote to the Heterozygote

				Case		Control	
P_d (P_m)	*K*	MOI	*DS*	Mean	Std	Mean	Std
0.25	0.33	D	0.25	1.7%	1.4%	1.9%	0.6%
0.36	0.33	D	0.25	2.1%	1.8%	2.0%	0.7%
0.25	0.50	D	0.25	1.8%	1.0%	1.8%	0.8%
0.36	0.50	D	0.25	2.0%	1.3%	2.0%	1.1%
0.25	0.33	M	0.25	1.9%	1.4%	1.8%	0.6%
0.36	0.33	M	0.25	2.0%	1.8%	2.0%	0.7%
0.25	0.50	M	0.25	1.8%	1.0%	1.9%	0.9%
0.36	0.50	M	0.25	2.0%	1.3%	2.1%	1.0%
0.25	0.33	D	0.50	1.4%	0.9%	1.4%	0.4%
0.36	0.33	D	0.50	2.0%	1.2%	1.9%	0.5%
0.25	0.50	D	0.50	1.4%	0.6%	1.4%	0.5%
0.36	0.50	D	0.50	1.9%	0.9%	1.9%	0.7%
0.25	0.33	M	0.50	1.5%	0.9%	1.4%	0.4%
0.36	0.33	M	0.50	2.0%	1.2%	1.9%	0.5%
0.25	0.50	M	0.50	1.4%	0.7%	1.5%	0.5%
0.36	0.50	M	0.50	1.9%	0.9%	1.9%	0.7%
0.25	0.33	D	0.75	1.2%	0.7%	1.2%	0.3%
0.36	0.33	D	0.75	1.5%	0.9%	1.5%	0.4%
0.25	0.50	D	0.75	1.2%	0.5%	1.2%	0.4%
0.36	0.50	D	0.75	1.5%	0.6%	1.5%	0.5%
0.25	0.33	M	0.75	1.3%	0.7%	1.2%	0.3%
0.36	0.33	M	0.75	1.5%	0.9%	1.4%	0.4%
0.25	0.50	M	0.75	1.2%	0.5%	1.2%	0.4%
0.36	0.50	M	0.75	1.5%	0.6%	1.5%	0.5%

These imputation inconsistency rates correspond to the true imputation inconsistency rates (within the 95% confidence intervals of the corresponding true values). Better estimates are associated with higher double sampling proportion.

Chapter 4 Discussions and Future Directions

My research questions were:

1. What situations produce more accurate imputation results?
2. Are there any indications of differential imputation inconsistency rates?
3. Are type I error rates inflated by imputation inconsistencies?

4a. Does the LRT-AE approach have greater power than the standard LRT approach that considers only fallible data?

4b. If the answer to question 4a is yes, what situations are associated with better LRT-AE performance?

4c. How big is the improvement?

5. Can researchers get unbiased estimates of imputation inconsistency rates using the LRT-AE approach?

My answers to these questions are:

1. My results are consistent with the findings of Pei et al. that stronger LD, lower MAF for the non-genotyped marker, and higher marker genotyping density in the sample produce better imputation results.
2. The imputation inconsistency rates appear to be non-differential under both the null and the alternative hypothesis.
3. Non-differential imputation inconsistencies affect only the power to detect association.

4a. There are circumstances in which the LRT-AE has greater power than the standard LRT.

4b. Better power of the LRT-AE occurs when the MAF for the disease marker is lower and imputation inconsistency rates are higher. For example, the power of the LRT-AE when the MAF is 0.25 and the imputation inconsistency rate is 4% is an average of 0.04 greater than the standard LRT when there is 75% double sampling.

4c. Both the average and the median power improvement is 0.02 for the 5% significance level and 0.03 for the 1% significance level using 75% double sampling.

5. Unbiased estimates of the imputation inconsistency rates are reported in the LRT-AE result files. They correspond to the true imputation inconsistency rates.

In summary, my dissertation compared the performance of the LRT-AE approach that incorporates double sample information for genotypes on a subsample to the standard LRT approach that only considers fallible data. I ran simulation studies under different combinations of various genetic parameters and calculated the empirical power. I also calculated the asymptotic power to verify the results.

Since my findings showed that a low MAF is associated with better performance of the LRT-AE approach, and Dickson et al. showed that rare variants create synthetic genome-wide associations, I plan to impute rare variants (MAF<0.05) and compare the performance of the LRT-AE approach with the performance of the standard LRT approach in the future.

References

Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002), "Merlin--Rapid Analysis of Dense Genetic Maps Using Sparse Gene Flow Trees," *Nat Genet*, 30, 97-101.

Abecasis, G. R., and Cookson, W. O. (2000), "Gold--Graphical Overview of Linkage Disequilibrium," *Bioinformatics*, 16, 182-183.

Agrawal, A., et al. (2009), "Further Evidence for an Association between the Gamma-Aminobutyric Acid Receptor a, Subunit 4 Genes on Chromosome 4 and Fagerstrom Test for Nicotine Dependence," *Addiction*, 104, 471-477.

Ahn, K., Gordon, D., and Finch, S. J. (2009), "Increase of Rejection Rate in Case-Control Studies with the Differential Genotyping Error Rates," *Stat Appl Genet Mol Biol*, 8, Article25.

Ahn, K., et al. (2007), "The Effects of Snp Genotyping Errors on the Power of the Cochran-Armitage Linear Trend Test for Case/Control Association Studies," *Ann Hum Genet*, 71, 249-261.

Anderson, C. A., et al. (2008), "Evaluating the Effects of Imputation on the Power, Coverage, and Cost Efficiency of Genome-Wide Snp Platforms," *Am J Hum Genet*, 83, 112-119.

Barral, S., Haynes, C., Stone, M., and Gordon, D. (2006), "Lrtae: Improving Statistical Power for Genetic Association with Case/Control Data When Phenotype and/or Genotype Misclassification Errors Are Present," *BMC Genet*, 7, 24.

Barrett, J. C., et al. (2008), "Genome-Wide Association Defines More Than 30 Distinct Susceptibility Loci for Crohn's Disease," *Nat Genet*, 40, 955-962.

Becker, T., Flaquer, A., Brockschmidt, F. F., Herold, C., and Steffens, M. (2009), "Evaluation of Potential Power Gain with Imputed Genotypes in Genome-Wide Association Studies," *Hum Hered*, 68, 23-34.

Bierut, L. J., et al. (2007), "Novel Genes Identified in a High-Density Genome Wide Association Study for Nicotine Dependence," *Hum Mol Genet*, 16, 24-35.

Borchers, B., Brown, M., McLellan, B., Bekmetjev, A., and Tintle, N. L. (2009), "Incorporating Duplicate Genotype Data into Linear Trend Tests of Genetic Association: Methods and Cost-Effectiveness," *Stat Appl Genet Mol Biol*, 8, Article24.

Browning, B. L., and Browning, S. R. (2008), "Haplotypic Analysis of Wellcome Trust Case Control Consortium Data," *Hum Genet*, 123, 273-280.

Browning, S. R. (2008), "Missing Data Imputation and Haplotype Phase Inference for Genome-Wide Association Studies," *Hum Genet*, 124, 439-450.

Browning, S. R., and Browning, B. L. (2007), "Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering," *Am J Hum Genet*, 81, 1084-1097.

Chambers, J. C., et al. (2008), "Common Genetic Variation near Mc4r Is Associated with Waist Circumference and Insulin Resistance," *Nat Genet*, 40, 716-718.

Cottingham, R. W., Jr., Idury, R. M., and Schaffer, A. A. (1993), "Faster Sequential Genetic Linkage Computations," *Am J Hum Genet*, 53, 252-263.

Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D. B. (2010), "Rare Variants Create Synthetic Genome-Wide Associations," *PLoS Biol*, 8, e1000294.

Dye, C., Scheele, S., Dolin, P., Pathania, V., and Raviglione, M. C. (1999), "Consensus Statement. Global Burden of Tuberculosis: Estimated Incidence, Prevalence, and Mortality by Country. Who Global Surveillance and Monitoring Project," *JAMA*, 282, 677-686.

Feng, Y., et al. (2004), "A Common Haplotype of the Nicotine Acetylcholine Receptor Alpha 4 Subunit Gene Is Associated with Vulnerability to Nicotine Addiction in Men," *Am J Hum Genet*, 75, 112-121.

Frazer, K. A., et al. (2007), "A Second Generation Human Haplotype Map of over 3.1 Million Snps," *Nature*, 449, 851-861.

, "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls." (2007)*Nature*, 447, 661-678.

Glasser, S., et al. (1998), "Prospective Study of Postpartum Depression in an Israeli Cohort: Prevalence, Incidence and Demographic Risk Factors," *J Psychosom Obstet Gynaecol*, 19, 155-164.

Gordon, D., Finch, S. J., Nothnagel, M., and Ott, J. (2002), "Power and Sample Size Calculations for Case-Control Genetic Association Tests When Errors Are Present: Application to Single Nucleotide Polymorphisms," *Hum Hered*, 54, 22-33.

Gordon, D., Haynes, C., Blumenfeld, J., and Finch, S. J. (2005), "Pawe-3d: Visualizing Power for Association with Error in Case-Control Genetic Studies of Complex Traits," *Bioinformatics*, 21, 3935-3937.

Gordon, D., et al. (2004a), "A Transmission Disequilibrium Test for General Pedigrees That Is Robust to the Presence of Random Genotyping Errors and Any Number of Untyped Parents," *Eur J Hum Genet*, 12, 752-761.

Gordon, D., et al. (2004b), "Increasing Power for Tests of Genetic Association in the Presence of Phenotype and/or Genotype Error by Use of Double-Sampling," *Stat Appl Genet Mol Biol*, 3, Article26.

, "A Haplotype Map of the Human Genome." (2005)*Nature*, 437, 1299-1320.

Hoft, N. R., et al. (2009), "Genetic Association of the Chrna6 and Chrnb3 Genes with Tobacco Dependence in a Nationally Representative Sample," *Neuropsychopharmacology*, 34, 698-706.

Huang, L., Wang, C., and Rosenberg, N. A. (2009), "The Relationship between Imputation Error and Statistical Power in Genetic Association Studies in Diverse Populations," *Am J Hum Genet*, 85, 692-698.

Kang, S. J., Gordon, D., and Finch, S. J. (2004), "What Snp Genotyping Errors Are Most Costly for Genetic Association Studies?," *Genet Epidemiol*, 26, 132-141.

Kruglyak, L., and Nickerson, D. A. (2001), "Variation Is the Spice of Life," *Nat Genet*, 27, 234-236.

Lathrop, G. M., and Lalouel, J. M. (1984), "Easy Calculations of Lod Scores and Genetic Risks on Small Computers," *Am J Hum Genet*, 36, 460-465.

Lettre, G., et al. (2008), "Identification of Ten Loci Associated with Height Highlights New Biological Pathways in Human Growth," *Nat Genet*, 40, 584-591.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009), "Genotype Imputation," *Annu Rev Genomics Hum Genet*, 10, 387-406.

Lou, X. Y., et al. (2008), "A Combinatorial Approach to Detecting Gene-Gene and Gene-Environment Interactions in Family Studies," *Am J Hum Genet*, 83, 457-467.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007), "A New Multipoint Method for Genome-Wide Association Studies by Imputation of Genotypes," *Nat Genet*, 39, 906-913.

Moskvina, V., Craddock, N., Holmans, P., Owen, M. J., and O'Donovan, M. C. (2006), "Effects of Differential Genotyping Error Rate on the Type I Error Probability of Case-Control Studies," *Hum Hered*, 61, 55-64.

Mukadi, Y. D., Maher, D., and Harries, A. (2001), "Tuberculosis Case Fatality Rates in High Hiv Prevalence Populations in Sub-Saharan Africa," *AIDS*, 15, 143-152.

Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M., and Franke, A. (2009), "A Comprehensive Evaluation of Snp Genotype Imputation," *Hum Genet*, 125, 163-171.

Olivier, M. (2003), "A Haplotype Map of the Human Genome," *Physiol Genomics*, 13, 3-9.

Ott, J. (1989), "Computer-Simulation Methods in Human Linkage Analysis," *Proc Natl Acad Sci U S A*, 86, 4175-4178.

Patil, N., et al. (2001), "Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21," *Science*, 294, 1719-1723.

Pe'er, I., et al. (2006), "Evaluating and Improving Power in Whole-Genome Association Studies Using Fixed Marker Sets," *Nat Genet*, 38, 663-667.

Pei, Y. F., Li, J., Zhang, L., Papasian, C. J., and Deng, H. W. (2008), "Analyses and Comparison of Accuracy of Different Genotype Imputation Methods," *PLoS One*, 3, e3551.

Saccone, S. F., et al. (2007), "Cholinergic Nicotinic Receptor Genes Implicated in a Nicotine Dependence Association Study Targeting 348 Candidate Genes with 3713 Snps," *Hum Mol Genet*, 16, 36-49.

Sachidanandam, R., et al. (2001), "A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms," *Nature*, 409, 928-933.

Schlaepfer, I. R., et al. (2008), "The Chrna5/A3/B4 Gene Cluster Variability as an Important Determinant of Early Alcohol and Tobacco Initiation in Young Adults," *Biol Psychiatry*, 63, 1039-1046.

Scott, L. J., et al. (2007), "A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants," *Science*, 316, 1341-1345.

Servin, B., and Stephens, M. (2007), "Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits," *PLoS Genet*, 3, e114.

Sherva, R., et al. (2008), "Association of a Single Nucleotide Polymorphism in Neuronal Acetylcholine Receptor Subunit Alpha 5 (Chrna5) with Smoking Status and with 'Pleasurable Buzz' During Early Experimentation with Smoking," *Addiction*, 103, 1544-1552.

Tenenbein, A. (1970), "A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications" *Journal of the American Statistical Association*, 65, 1350-1361.

Tenenbein, A. (1972), "A Double Sampling Scheme for Estimating from Misclassified Multinomial Data with Applications to Sampling Inspection" *Technometrics*, 14, 187-202.

Thorgeirsson, T. E., et al. (2008), "A Variant Associated with Nicotine Dependence, Lung Cancer and Peripheral Arterial Disease," *Nature*, 452, 638-642.

Tintle, N., et al. (2009), "Inclusion of a Priori Information in Genome-Wide Association Analysis," *Genet Epidemiol*, 33 Suppl 1, S74-80.

Voineskos, S., et al. (2007), "Association of Alpha4beta2 Nicotinic Receptor and Heavy Smoking in Schizophrenia," *J Psychiatry Neurosci*, 32, 412-416.

Weiss, R. B., et al. (2008), "A Candidate Gene Approach Identifies the Chrna5-A3-B4 Region as a Risk Factor for Age-Dependent Nicotine Addiction," *PLoS Genet*, 4, e1000125.

Willer, C. J., et al. (2008), "Newly Identified Loci That Influence Lipid Concentrations and Risk of Coronary Artery Disease," *Nat Genet*, 40, 161-169.

Zeggini, E., et al. (2008), "Meta-Analysis of Genome-Wide Association Data and Large-Scale Replication Identifies Additional Susceptibility Loci for Type 2 Diabetes," *Nat Genet*, 40, 638-645.