SAMPLE SIZE ESTIMATION AND TYPE I ERROR CORRECTION IN GENETIC ASSOCIATION STUDIES

A Dissertation

Presented in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

with a

Major in Bioinformatics and Computational Biology

in the

College of Graduate Studies

University of Idaho

by

Hua Feng

May 2016

Major Professor: Celeste Brown, Ph.D.

Committee Members: Larry J. Forney, Ph.D., Paul Joyce, Ph.D., Christopher Williams, Ph.D.

Department Administrator: Eva Top, Ph.D.

AUTHORIZATION TO SUBMIT DISSERTATION

The dissertation of Hua Feng, submitted for the degree of Doctor of Philosophy with a major in Bioinformatics and Computational Biology and titled, "SAMPLE SIZE ESTIMATION AND TYPE I ERROR CORRECTION IN GENETIC ASSOCIATION STUDIES," has been reviewed in final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: ___________________________________ Date: ______________
Celeste Brown, Ph.D.

Committee Members: ___________________________________ Date: ______________
Larry J. Forney, Ph.D.

___________________________________ Date: ______________
Paul Joyce, Ph.D.

___________________________________ Date: ______________
Christopher Williams, Ph.D.

Department
Administrator: ___________________________________ Date: ______________
Eva Top, Ph.D.

ABSTRACT

Background: Statistics is a key component of bioinformatics, which provides crucial insight into biological processes, such as testing genetic association with the risk of complex human diseases and variation of drug response. A lack of statistical power due to small sample size in genetic association studies increases the probability of type II error, and the determination of the correct sample size for these studies is influenced by various biological parameters. Additionally, multiple hypothesis testing, which is common in genetic association studies, leads to type I error inflation.

Objective and Methods: This study focused on statistical properties that are important in genetic association studies: 1) testing effects of biological factors on sample size estimation by regression analysis; 2) developing a two-stage Bonferroni type I error correction procedure using linkage disequilibrium (LD) to define independent haplotype blocks; and 3) adjusting alpha levels in sample size estimation based on LD structure among genetic markers in different racial groups.

Results: The first study showed that a recessive genetic model requires the largest sample size; the most significant factors for sample size estimation were minor allele frequency under the recessive genetic model, and genetic effect size under dominant and additive genetic models. The two-stage adjusted Bonferroni correction was less conservative than the standard Bonferroni correction, but less liberal than FDR. Sample sizes estimated using an adjusted alpha level based on LD structure could be reduced by 14% to 24% depending upon racial group, compared with the standard Bonferroni adjustment for alpha level.

Conclusion and implication: Genetic inheritance model, effect size, and allele frequency significantly impact sample size estimation. The results can be applied to genetic marker selection, sample size estimation, and statistical power prediction. The two-stage adjusted Bonferroni type I error correction procedure improves statistical power, and introduces a simple way to control for type I error in genetic association studies. Using LD structure across the tested DNA region to adjust the alpha value for sample size estimation by race can reduce the required total sample sizes, improve statistical power, and lead to cost-effective outcomes.

Keywords: Genetic association study; Sample size estimation; Statistical power; Genetic effect; Genetic inheritance model; Linkage disequilibrium; Type I error inflation; Bonferroni type I error correction; Haplotype block; FDR.

TABLE OF CONTENTS

AUTHORIZATION TO SUBMIT DISSERTATION ii

ABSTRACT iii

TABLE OF CONTENTS v

LIST OF TABLES vii

LIST OF FIGURES viii

ACKNOWLEDGEMENTS ix

DEDICATION x

CHAPTER 1: INTRODUCTION 1

GENETIC ASSOCIATION STUDY 2
THE OBJECTIVES OF THIS DISSERTATION 5

CHAPTER 2: INFLUENCE OF BIOLOGICAL PARAMETERS ON SAMPLE SIZE ESTIMATION IN GENETIC ASSOCIATION STUDIES 7

ABSTRACT 7
INTRODUCTION 8
STATISTICAL MODELS FOR SAMPLE SIZE ESTIMATION 10
Disease trait 10
Quantitative trait 13
FACTORS THAT INFLUENCE SAMPLE SIZE ESTIMATION 14
Type I error and power 14
Linkage disequilibrium 15
Marker allele frequency 17
Genetic inheritance model, disease prevalence, effect size, and disease prevalence 17
Population stratification by race 18
SIMULATING SAMPLE SIZES AND TESTING FACTOR EFFECTS ON SAMPLE SIZE ESTIMATION 19
Study design and outcome variables 20
Statistical models 20
Parameters for sample size estimation 20
Poisson regression analysis to examine factor effects on sample size estimation 21
Visualizing interaction effects by smoother surface plot 22
RESULTS 22
Sample sizes under different parameter combinations 22
Poisson regression analysis to examine factor effects on sample size estimation 25
Interaction effects illustrated by smoother surface plots 28
DISCUSSION AND CONCLUSIONS 30

CHAPTER 3: TWO-STAGE ADJUSTED BONFERRONI PROCEDURE FOR TYPE I ERROR CORRECTION IN MULTI-LOCUS GENETIC ASSOCIATION STUDIES 36

ABSTRACT 36

INTRODUCTION..37
CURRENT METHODOLOGY FOR TYPE I ERROR CORRECTION..39
Standard and modified Bonferroni correction methods..39
Type 1 error correction by permutation procedure..40
False Discovery Rate..41
MODELING AND PARAMETER TESTING..42
Data sets..42
P-value simulation and empirical distributions..42
Deriving the effective number of independent tests based on LD structure among multiple loci..45
Adjusted point-wise significance level..48
Adjusted point-wise error rate among and within haplotype blocks by the step-down Bonferroni method..49
Declaring significant SNP associations..49
RESULTS..50
P-value simulations based upon LD..50
Effective number of independent tests..53
Adjusted point-wise significance level..53
Number of tests declared significant..54
APPLICATION OF THE TWO-STAGE CORRECTION IN REAL CLINICAL STUDIES..56
DISCUSSION AND CONCLUSION..59

CHAPTER 4: APPLYING THE ADJUSTED TYPE I ERROR RATE IN SAMPLE SIZE ESTIMATION FOR GENETIC ASSOCIATION STUDIES..62

ABSTRACT..62
INTRODUCTION..63
METHODS..64
Find the effective number of independent tests..64
Calculate the point-wise error rate..65
Estimate the sample size..65
CASE STUDY..67
Introduction..67
Methods..67
RESULTS..68
Structure of haplotype block and pointwise error rate..68
Haplotype Structure in ASW population..69
Haplotype Structure in CEU population..70
Haplotype Structure in CHB population..70
Sample size estimation..71
DISCUSSION..73

REFERENCES..76

LIST OF TABLES

TABLE 2-1. THE RELATIONSHIP BETWEEN RELATIVE RISK AND GENETIC INHERITANCE MODELS 18

TABLE 2-2. THE NUMBER OF CASE-CONTROL PAIRS NEEDED TO DETECT GENETIC ASSOCIATION WITH DICHOTOMOUS OUTCOMES UNDER DIFFERENT PARAMETER ASSUMPTIONS AT A=0.05 AND B=0.80 23

TABLE 2-3. THE NUMBER OF INDIVIDUALS NEEDED TO DETECT GENETIC ASSOCIATION WITH QUANTITATIVE UNDER DIFFERENT PARAMETER ASSUMPTIONS WHEN POPULATION MEAN=100, STANDARD DEVIATION=10, A=0.05 AND B=0.80 24

TABLE 2-4. THE TESTS OF MAIN EFFECTS FOR THREE PARAMETERS IN POISSON REGRESSION ANALYSES 26

TABLE 2-5. THE TESTS OF MAIN AND INTERACTION EFFECTS FOR THREE PARAMETERS IN POISSON REGRESSION ANALYSES 27

TABLE 3-1. DESCRIPTIVE STATISTICS FOR THE THREE LEVELS OF P VALUES 51

TABLE 3-2. THE NUMBER OF INDEPENDENT TESTS (HAPLOTYPE BLOCKS) ACROSS HAPLOTYPE BLOCKING ALGORITHMS..53

TABLE 3-3. POINT-WISE ERROR RATES UNDER FAMILY-WISE ERROR RATE OF 0.05 AMONG 267 SNPS IN THREE HAPLOTYPE INFERRING ALGORITHMS...54

TABLE 3-4. NUMBERS OF SNPS WITH SIGNIFICANT ASSOCIATIONS WITH DISEASE ACROSS TYPE I ERROR ADJUSTMENT METHODS AMONG 267 SNPS IN THREE HAPLOTYPE INFERRING ALGORITHMS 55

TABLE 3-5. INFLUENCE OF *NPPA* SNPS ON THE RISK OF ASTHMA RISK FOR WHITE PARTICIPANTS IN THE SCREENING COHORT[55] 57

TABLE 3-6. INFLUENCE OF *NPPA* SNPS ON THE RISK OF ASTHMA RISK FOR WHITE PARTICIPANTS IN THE REPLICATE COHORT[55]..58

TABLE 4-1. SAMPLE SIZE REQUIRED TO TEST GENETIC ASSOCIATION WITH HYPOVITAMINOSIS D...........................72

TABLE 4-2. RESULTS FROM HONG'S STUDY[28] IN SAMPLE SIZES WITH 80% POWER BY INCREASING NUMBER OF SNP MARKERS IN CASE-CONTROL AND CASE-PARENT STUDIES..73

List of Figures

Figure 2-1. Surface plot showing effect of minor allele frequency and genetic effect size on sample size for a quantitative outcome under a A) recessive, B) dominant or C) additive genetic model 29

Figure 3-1. Procedure for p value simulation 44

Figure 3-2. Haplotype blocks from SNP sequence data at chr10:67587369..67787368 in 174 Utah residents. 47

Figure 3-3. Distribution of 267 pair-wise r^2 among 267 SNPs calculated by Haploview[74] from 200 kbp at chr10:67587369..67787368 in 174 Utah residents 51

Figure 3-4. P value distributions for large, medium and small p value generated by multiplying uniform random numbers by 0.1, 0.05, and 0.01, respectively. 52

Figure 4-1. Linkage disequilibrium plot for SNPs in vitamin D binding protein gene (*GC*) region at chr4:72824381..72862352 among African ancestry in Southwest USA 69

Figure 4-2. Linkage disequilibrium plot for SNPs in vitamin D binding protein gene (*GC*) DNA region at chr4:72824381..72862352 among Utah residents with Northern and Western European ancestry 70

Figure 4-3. Linkage disequilibrium plot for SNPs in vitamin D binding protein gene (*GC*) gene DNA region at chr4:72824381..72862352 among Han Chinese in Beijing 71

ACKNOWLEDGEMENTS

It has been a long journey to pursue my degrees at the University of Idaho. 16 years ago, I started my first class in the university. Day by day, year by year, I completed my two master degrees, and now I am going to reach my final destination of the PhD.

During the journey, I have experienced so many things that I never experienced before, have had tough times and good times mixed with hope, desperation and joy, and have been surrounded and blessed by incredible people. Without their generous help and support, I would not be able to go through these trials. All of these people are such precious and unexpected gifts for me. I feel wordless in expressing my sincere great appreciation to you!

I would like to thank you, Celeste, for your mentoring, inspiration, and encouragement of my dissertation work;

I would like to thank you, Paul, for providing me the great opportunity enrolled in the BCB PhD program and for your consistent support;

I would like to thank you, Chris and Larry, for serving on my PhD committee and your great advice.

DEDICATION

My dissertation is gratefully dedicated to all my loved ones, who are my dear family members, supportive academic mentors, and wonderful friends.

CHAPTER 1: INTRODUCTION

Bioinformatics is a recently emerged interdisciplinary field that draws scientists from statistics, mathematics, computer science and engineering to study, understand and process molecular and genetic data[1]. The development of high-throughput sequencing and genotyping technology has led to the production of huge amounts of genetic data. This flood of data introduces significant challenges in bioinformatics, such as data storage and retrieval, data manipulation and analysis. Statistics is the key component of bioinformatics that provides crucial insight into biological processes, to understand molecular level biological systems and to allow the simultaneous analysis of millions of data points. Without statistics and statistical techniques required to analyze, summarize and interpret these data, we are very limited to learn from our observations, which will in turn inhibit our ability to move forward in genetic research. Although basic statistical concepts help biologists to prepare experiments, verify conclusions and interpret results, these traditional statistical tools often fail in the face of an immense challenge due to high volume and large heterogeneity of biological data.

Genetic variants, a type of biological data, are often investigated to detect genetic association with complex human diseases and variation in drug response. Various biological factors affect statistical power in a genetic association study and as more genetic variants are tested simultaneously, larger sample sizes are required to achieve adequate statistical power to detect the association. Also, when a large number of single nucleotide polymorphisms and/or genes are tested, multiple hypothesis tests can lead to type I error inflation, resulting in false positive results. To confront these challenges, we applied basic statistical concepts to develop new methods for finding genetic association with targeted phenotypes.

This dissertation used a bioinformatics approach to understand parameter influence on sample size estimation in study design, to address and solve type I error inflation problems in multiple tests, and to apply type I error adjustment methods in sample size estimation for multi-locus genetic association studies. The dissertation involved a broad range of bioinformatics across statistics, genetics, and computational biology, including single nucleotide polymorphism genotype data query, computational simulation, linkage disequilibrium test, haplotype block inference, genetic inheritance model assumption, statistical analysis, etc.

GENETIC ASSOCIATION STUDY

Genetic association studies are a major tool for identifying genes conferring susceptibility to complex disorders. Genetic association studies are performed to determine the associations between a phenotype/phenotypes and a genetic variant or multiple genetic variants, such as to test whether a genetic variant is associated with a disease or a trait (e.g. variation in drug dose response). Despite the potential problem of low power to definitively identify the associations, the genome-wide studies are applied to test the large number of polymorphisms simultaneously. If an association is present, a particular allele, genotype or haplotype of a polymorphism or polymorphisms will be seen more often than expected by chance in an individual carrying the trait. Thus, a person carrying one or two copies of a high-risk variant is at increased risk of developing the associated disease or having the associated trait[2].

Genetic association can be between phenotypes, between a phenotype and a genetic polymorphism, or between two genetic polymorphisms[3-5]. Single nucleotide polymorphisms (SNP) are the most common source of genetic polymorphism in the human genome[6]. A SNP is the result of point mutations that produce single base-pair differences (substitutions or deletions) among chromosome sequences[7]. Traits and diseases are termed “complex” when both genetic and environmental factors contribute to the susceptibility risk. Extensive experience in genetic studies for many complex disorders (such as diabetes, heart disease, autoimmune diseases, and psychiatric traits) confirms that many different genetic variants control disease risk, with each variant having only a subtle effect[2].

Advanced genotyping technology has made genome-wide association studies possible to test if any variants among hundreds of thousands or even millions of polymorphisms are associated with the phenotypes of interest. However, when a large number of polymorphisms are tested in a genome-wide study, the power to definitively identify associations is low. Therefore, most recent genetic association studies still examine a single polymorphism or a set of polymorphisms near a single gene, or focus on a candidate region that we have prior reason to believe might be associated with the phenotype of interest. Bioinformatics is commonly used to identify the candidate genes and nucleotides (SNPs) to help researchers to better understand the genetic basis of disease.

A significant genetic association between a SNP and a disease could be interpreted as either (1) direct association, in which the genotyped SNP is the true causal variant conferring disease susceptibility; (2) indirect association, in which a SNP is in linkage disequilibrium (LD) with the true causal variant; or (3) a false-positive result, in which there is either chance or systematic confounding, such as population stratification or type I error inflation. Distinguishing between direct and indirect association is challenging and may require resequencing of the candidate region, dense genotyping of all available SNPs, or functional studies to confirm the role of a putative mutation in disease[2].

The simplest study design used to test for association is the case–control study[8], in which a number of cases affected with the disease of interest are collected together with a number of control individuals without the presence of the phenotype of interest. The specific choice of phenotype for the cases may define the exact hypothesis to be tested, and applying strict clinical criteria for ascertainment is necessary to ensure a homogeneous set of cases. Controls are collected from individuals who have been screened as negative for presence of the phenotype of interest, or randomly ascertained from the population, whose disease status is unknown. A study with screened unaffected controls will have higher power to detect an association compared with a study using population-based controls. However, for some diseases, screening controls for the presence or absence of the disease may be difficult, and using a larger sample of unscreened controls may be more efficient[2].

A sample size with sufficient statistical power is critical to the success of genetic association studies in detecting causal variants of complex human diseases and drug responses. Sample size is typically estimated according to an appropriate statistical model that is derived under the hypothesis, the study design, the primary study endpoint, and the clinically meaningful difference of the primary study endpoint between case and control.

To test for association between polymorphisms and an outcome (e.g. disease), a genetic model needs to be assumed[8]. If we assume a dominant or recessive genetic model, the SNP genotypes are dichotomized to force heterozygotes to have the same risk or mean phenotype as one of the homozygotes. Additive models impose a structure in which each additional copy of the variant allele increases the response, whether log odds ratio, log hazard ratio, or mean phenotype, by the same amount. For categorical outcomes, the simplest association test is a Chi-Square test of independence computed on a cross-classification table of outcome versus alleles or genotypes for each variant. For quantitative phenotypes, ANOVA, a type of linear regression models for quantitative outcomes and categorical predictors, can be used to test for association between a genotype and a phenotype. Each of the hypothesis tests yields a p value.

In statistics, the p value is the probability of obtaining the observed sample results, when the null hypothesis is actually true[9]. In inferential statistics, the null hypothesis usually refers to a general statement or default position that there is no relationship between two measured phenomena (e.g. genetic variants and disease state), or no difference among groups[10]. The p values are used to help scientists determine whether or not their hypotheses are correct. A small p value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so we reject the null hypothesis. A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so we fail to reject the null hypothesis. The probability of achieving statistical significance is called statistical power, when in fact the alternative hypothesis (contrary to the null hypothesis) is true. Such as a p value of 0.01 may imply a significant association between a genetic variant and a disease. Thus, a person carrying one or two copies of the genetic variant is at increased risk of developing the associated disease or having the associated trait. To declare the statistical significance, we need a criterion/ measure to compare with the p value. Therefore, before each of a hypothesis tests is performed, a threshold value is chosen, called the significance level of the test, traditionally 5% or 1% and denoted as α[10, 11]. If the p-value is equal to or smaller than the significance level (α), it suggests that the observed data are inconsistent with the assumption that the null hypothesis is true and thus that hypothesis must be rejected. When a true null hypothesis is incorrectly rejected, a type I error (false positive) occurs. The significance level (α) is also the probability of Type I error[10]. However, when a false null hypothesis is accepted, a type II error (false negative) occurs. As the statistical power increases, the chance of type II error decreases. Most genetic association studies involve multiple SNPs and/or genes, to test a large number of hypotheses for direct or indirect association with phenotypes of interested. When multiple hypothesis tests are performed simultaneously in a study, the risk of inflation of the type I error rate increases, resulting in false positive results[10, 11].

The Objectives of This Dissertation

The dissertation consists of three main parts: 1) testing parameter effects on sample size estimation for single locus genetic association studies; 2) developing a two-stage adjusted Bonferroni correction procedure for multiple hypothesis tests of association between multiple genetic markers and disease; and 3) developing a method to calculate the alpha level in sample size estimation in multi-locus genetic studies by taking linkage disequilibrium between genetic markers into account.

Sample size determination is among the most commonly encountered tasks in statistical practice. The sample size required to detect genetic association is influenced by various parameters specific to these types of studies. Our first objective was to test the effects of minor allele frequency, genetic inheritance model, and genetic effect size on sample size estimation to achieve adequate power in genetic association studies. Based on the different combinations of these three parameters, two statistical models (logistic regression and linear regression) were used to estimate sample sizes in accordance with two phenotypes, disease state and quantitative trait, respectively. A case-control study design was used for disease outcome while a study cohort with independent individuals was used for a quantitative trait. Minor allele frequencies were from 0.01 to 0.30, effect sizes were from small to medium, and genetic inheritance models were dominant, recessive, and additive. Poisson regression models were applied to examine the main and interaction effects among the three factors on sample size estimation. The results of the effects were represented by regression coefficients and surface plots. Finally, real clinical studies were given as examples to illustrate the importance and effects of these parameters on sample size estimation.

The multiple-comparison problem with type I error inflation arises in genetic association studies when the association between phenotypes and multi-locus genotypes is examined. The Bonferroni correction directly targets the type I error problem but it may yield conclusions that are too conservative due to correlation among genetic markers. The correlation among these markers violates the independence assumption of the Bonferroni procedure, resulting in type II error inflation. Our second objective was to develop a two-stage adjusted Bonferroni correction procedure, which corrects for the problem of correlation among genetic markers. The procedure was to 1) simulate p values based on the LD structure of 267 SNPs taken from

HapMap[12] to mimic the p values from the statistical association tests between the 267 SNPs and disease state; 2) derive the effective number of independent tests based on linkage disequilibrium (LD) structure among the 267 SNPs; 3) calculate the point-wise error rates based on the effective number of independent tests as the threshold values to determine whether or not these SNPs were significantly associated with the disease; 4) compare the p values with the point-wise error rates across the blocks and singletons; and 5) apply the Holm–Bonferroni method and dependent false discovery rate (FDR) to highly-correlated test statistics within each LD block when there are SNPs in the LD block that have smaller p values compared to the point-wise error rate. Our third aim was to apply the type I error adjustment method to calculate the alpha level in sample size estimation in accordance with the multi-locus genetic association study design.

CHAPTER 2: INFLUENCE OF BIOLOGICAL PARAMETERS ON SAMPLE SIZE ESTIMATION IN GENETIC ASSOCIATION STUDIES

ABSTRACT

Background: Genetic association (GA) studies assess the association between genetic polymorphisms and phenotypes of interest, such as disease states and drug responses. A sample size with sufficient statistical power is critical to the success of genetic association studies. Various biological factors in GA studies could affect statistical power, including frequency of the risk allele, genetic effect size and mode of inheritance.

Objective and Methods: To optimize statistical power, we studied the effects of three parameters on sample size estimation: minor allele frequency, genetic inheritance model, and genetic effect size. Based on different combinations of the three parameters, two statistical models (logistic regression model and linear regression model) were used to estimate the sample sizes in accordance with two phenotypes, disease state and quantitative trait, respectively. A case-control study design was used for disease outcome while a study cohort with independent individuals was used for a quantitative trait. Minor allele frequencies ranged from 0.01 to 0.30, effect sizes were small or medium, and genetic inheritance models were dominant, recessive, and additive. Poisson regression models were used to test for main and interaction effects among the three factors on sample size estimation.

Results: The main and the interaction effects among the three factors were statistically significant (p<0.001). Among the three genetic models, the largest sample size is required to detect the genetic effect of variants under the recessive model in both dichotomous and quantitative outcomes. To reach a specific statistical power to discover causal variants, a larger sample size was needed for small minor allele frequency and genetic effect size. As such, we need to select genetic markers with allele frequencies at least 1% and genetic effect size being larger than small. The significant interaction indicates that the effect of one factor on the sample size is different at different values of the other two factors. Minor allele frequency is the most important factor under the recessive genetic model, and genetic effect size has the most influence under dominant and additive genetic models.

Conclusion and Implication: Required sample sizes in genetic association studies were significantly associated with genetic inheritance model, effect size, and allele frequency. Our simulation results could be applied to real genetic association studies to help researchers in genetic marker selection, sample size estimation, and statistical power prediction.

Keywords: Genetic association study; Sample size estimation; Statistical power; Genetic effect; Genetic inheritance model.

INTRODUCTION

Genetic makeup plays an important role in the development of human diseases such as asthma, hypertension, and diabetes. Interacting with lifestyle and environmental factors, single gene or polygenic disorders cause at least six thousand human diseases[13]. A person's chance of developing or passing on a genetic disorder can be influenced by many factors. Therefore, only part of these genetic diseases can be passed on from generation to generation. For instance, asthma is a heritable disease and polygenic disorder; the onset of asthma and related traits is associated with at least 200 genes[14]. Four of these genes (*NPPA, NPRA, CLCN6*, and *ILF2*) were found to be involved in lung vasodilatation, bronchorelaxation, pulmonary permeability, and surfactant production and action[15, 16]. Sickle cell disease is also one of the most common genetic diseases with prevalence of 0.16% in African Americans. Acute chest syndrome is the leading cause of mortality[17-19] in patients with sickle cell disease. The etiology of acute chest syndrome in sickle cell disease is not fully understood. However, both the *NOS1* AAT-repeat polymorphism and *NOS3* T-786C polymorphism were reported to be associated with acute chest syndrome in sickle cell disease[20].

Genetic variants also affect individual variability in drug response. For example, warfarin is the most widely prescribed anticoagulant drug worldwide to prevent heart attacks, strokes, and blood clots[21]. The optimal warfarin dose is difficult to establish because it can vary 10-fold among patients. An incorrect dose could lead to life threatening side effects, such as severe bleeding. Previous studies indicated that the single nucleotide polymorphisms in vitamin K epoxide reductase complex 1 (*VKORC1)* and cytochrome P450 2C9 (*CYP2C9)* genes have prominent effects on warfarin dose requirements[22]. The new disciplines of pharmacogenetics and pharmacogenomics study genetic effects on drug response, challenge the traditional one-size-fits-all dosage, and develop individualized medicine based on the patient’s genetic background. The ultimate goal of pharmacogenetics and pharmacogenomics is to maximize efficacy and minimize toxicity by individualized drug therapy.

A genetic association study identifies genetic polymorphisms or variants for the trait of interest in a population. There are many types of genetic polymorphisms, but this study focuses on single-nucleotide polymorphisms (SNP). A SNP is defined as a genetic variant that occurs in coding, non-coding, or untranslated regions of the genome in at least 1% of the population. Some of these variants in human DNA sequences can affect how humans develop diseases and respond to pathogens, drugs and other agents.

A fundamental notion in genetic association studies is the linkage disequilibrium (LD) between a genetic marker and the locus/loci that affect the trait (e.g. a specific disease) under study. LD is a measure of deviation from random association among alleles at two or more loci from their close proximity on a chromosome in the population. The genetic marker is expected to be in strong LD with a nearby disease locus. Therefore, the association between an unobserved genetic variant and the disease is indirectly measured between a candidate marker and the disease. In the indirect genetic association study, the strength of the association between the genetic markers and the phenotype of interest is attenuated along with the decrease in LD strength between the genetic markers and the causal variant. Many polymorphisms in the human genome carry redundant information since they are in LD with each other. As such, representative SNPs are commonly used in candidate gene/variant association studies, which specifically test one or a few genetic regions with prior hypotheses on sets of genes or polymorphisms associated with the phenotype of interest. Furthermore, genome-wide association studies investigate hundreds of thousands or even millions of polymorphisms in the entire genome.

Drawbacks in genetic association studies include low statistical power and type I error inflation due to insufficient sample size[23]. Sample size estimation is crucial to assure validity, accuracy, reliability, and integrity in a genetic association study. An insufficient sample size lacks power, leading to a false negative conclusion (type II error), whereas an over-powered study wastes time and resources. Gauderman[24-26] implemented regression models in sample size or power estimation for association studies of genes, gene-environment interaction, or gene-gene interaction for study designs with matched case-control, case-sibling, case-parent, and case-only designs. Pfeiffer[27] calculated sample sizes for unmatched case-control and sibling case-control studies to detect genetic association, and discovered factors affecting required sample size. Hong[28] estimated statistical power with increasing numbers of markers and compared sample sizes that were required in case-control/case-parent studies under various assumptions.

This study investigates the influence of several biological factors that are important for sample size estimation in genetic association studies, and determines how adequate statistical power can be achieved for candidate gene/variant association studies. The factors include linkage disequilibrium (LD), minor allele frequency, genetic inheritance model, genetic effect, and disease prevalence. Clinical studies are given as examples to illustrate the importance of these parameters and interaction effects in sample size estimation.

Statistical Models for Sample Size Estimation

Sample size is typically estimated according to an appropriate statistical model that is derived under the hypothesis, the study design, the primary study endpoint, and the clinically meaningful difference of the primary study endpoint between cases and controls[29]. In our study, different statistical models were used to estimate sample size for two phenotypes, disease state and quantitative trait. Logistic regression models were used for genetic association with diseases; and linear regression models were used for genetic association with quantitative outcomes. The estimation models are introduced as follows.

Disease trait

Population-based genetic association studies[24] are commonly used to investigate genetic effects on phenotype by randomly sampling diseased (case) and non-diseased (control) individuals from targeted populations. The mean proportions of genetic risk variants are compared between cases and controls to determine how these genetic factors are related to disease risk.

A case-control genetic association study determines whether the presence of a genetic variant increases the risk of a disease in a large population of unrelated individuals. In the study, there are two well-designed groups: the case group includes individuals with the disease under study, and the control group includes individuals without the targeted disease. Causal genetic variants are sought based on the assumption that an individual carrying one or two copies of a disease risk variant is more likely to develop the disease, and therefore an increased frequency of a variant or genotype in cases compared with controls implies that the variant may be associated with the disease.

A logistic regression model is used because a subject either has the disease or does not. Suppose that *N* denotes the number of disease affected subjects (cases) and *NK* is the number of disease unaffected individuals (controls), where K is the ratio of controls to cases. We wish to determine the smallest N that will give us sufficient power to reject the null hypothesis when it is false. Sample sizes can be estimated for a single marker or multi-locus markers.

To test the effect of a single genetic marker, we assume that it is the causal mutation or that it has complete or strong LD with the causal locus. The biallelic genetic marker has two possible alleles "*A*" or "*a*", and therefore there are three possible genotypes, "*AA*", "*Aa*" and "*aa*". The subject carrying the allele "*A*" is more prone to develop the disease than one with allele "*a*" only and this is indicated by the genetic factor, *G*. *G* is determined by genotype *g* and the inheritance model as follows:

Dominant: *G*=1 for *g*=*AA*, *Aa*; *G*=0 for *g*=*aa*

Recessive: *G*=1 for *g*=*AA*; *G*=0 for g=*Aa*, *aa*

Log-additive: *G*=2 for *g*=*AA*; *G*=1 for *g*=*Aa*; *G*=0 for *g*=*aa*

Disease occurrence in the population is given by the logistic regression model[24, 25, 30, 31].

$$P_r(S = 1|G) = \frac{e^{\beta_0+\beta_1 G}}{1 + e^{\beta_0+\beta_1 G}} \quad ,$$

where *S* is an indicator of disease status (1=diseased; 0=not diseased). The baseline probability of disease is $P_0 = e^{\alpha}/(1 + e^{\alpha})$, which is the disease risk in genetically normal (*G*=0) subjects. The disease odds ratio for carriers (*G*=1) compared to non-carriers (*G*=0) is the quantity $R_g = e^{\beta_1}$, which is the population-average genetic relative risk in an epidemiological study of the genetic factor alone. The mean population genetic odds ratio with genetic factor alone also can be estimated as

$$\overline{R_g} = \frac{P_r(S = 1|G = 1)}{P_r(S = 1|G = 0)}$$

To obtain a maximum likelihood estimate for the logistic regression model, the following likelihood for the logistic model with N cases and N$\times$ K $(K = 1{,}2, .. k)$ controls is maximized:

$$L(\alpha, \beta_g) = \prod_{i=1}^{N} \frac{e^{\beta_0+\beta_1 G}}{1 + e^{\beta_0+\beta_1 G}} \prod_{j=1}^{N\times K} \frac{e^{\beta_0+\beta_1 G}}{1 + e^{\beta_0+\beta_1 G}}$$

To demonstrate algorithms for testing a genetic factor, we supposed a null hypothesis H_0 of interest is that $\beta_1 = 0$ (or $R_g = 1$), which means no significant genetic effect on the disease with the alternative hypothesis H_1

$$H_0{:}\, \beta_1 = 0 \ versus\ H_1{:}\, \beta_1 \neq 0\ .$$

The likelihood ratio statistic (deviance) for testing the lack of fit for a logistic regression model is given by

$$\lambda(\beta) = -2\ln\left[\frac{\mathrm{L}\hat{\beta}_0}{\mathrm{L}\hat{\beta}}\right] = -2\ln\left[\frac{\prod_{i=1}^{N} \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} \prod_{j=1}^{N\times K} \frac{1}{1+e^{\hat{\beta}_0}}}{\prod_{i=1}^{N} \frac{e^{\hat{\beta}_0+\hat{\beta}_1 Gi}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 Gi}} \prod_{j=1}^{N\times K} \frac{1}{1+e^{\hat{\beta}_0+\hat{\beta}_1 Gj}}}\right]$$

$\lambda(\beta)$ has a chi-square distribution with degrees of freedom equal to the difference in the number of degrees of freedom (*df*) between the two models (i.e., the number of variables added to the model). If $\lambda(\beta) > x^2_{\alpha,df}$, the lack of fit is significant at significance level α.

A sufficiently large deviance implies the logistic model is inappropriate. Inference concerning any regressor or subset can be computed by determining how much the presence of each regressor contributes to the reduction in deviance. Therefore, to test the hypothesis that the genetic factor is significantly associated with the disease, we are going to test the difference in variation explained between the full and reduced models by

$$\Lambda = \lambda(\beta_{0,}\, \beta_1) - \lambda(\beta_0) = 2(lnL(\hat{\beta}_0, \hat{\beta}_1) - \ln\ L(\hat{\beta}_0))$$

Under the null hypothesis ($\beta_1 = 0$), Λ is asymptotically distributed as a chi-square random variable with one degree of freedom. When H_0 is rejected, NΛ is the non-centrality parameter of the chi-squared distribution for a given sample size *N*. Sample size can be computed by $N = \frac{(z_{\alpha/2}+z_{\beta})^2}{\Lambda}$, where the parameters are type I error rate (α-0.05), type II error rate (β=0.80) and *K* = controls/cases. Since Λ is a function of the odds ratio $R_g = e^{\beta_1}$, a smaller odds ratio requires a larger sample size. Therefore, the total sample size is equal to 2N for the matched case-control study, and *N*+*KN* for unmatched case-control study.

Quantitative trait

A quantitative trait genetic association study can be used to determine genetic impact on quantitative outcomes. For instance, fluticasone propionate is used to treat asthma. The genetic variant Gly16Arg of *ADRB2* was found to be implicated with responsiveness to this drug[32]. A study was conducted to compare the measurements of lung function between variant carriers and non-variant carriers in patients who take the medication. Sample size estimation at the beginning of this study was performed based on a two-sample parallel design with a dominant genetic model (Arg/Arg+Arg/Gly vs. Gly/Gly in Gly16Arg of *ADRB2*).

For the quantitative outcome trait, we assume a linear model relating the phenotype to the genetic factor, as $Y=\beta_0+\beta_1 G+e$. The residual e is assumed to be normally distributed, with mean zero and variance σ^2. The parameter β_1 is the main effect of the genotype. It is measured by the change in mean Y (outcome) per unit increase in genotype, such as for the dominant genetic model: G=1 for genotypes *AA*, and *Aa*; G=0 for genotype aa.

Rg^2 denotes the proportion of variation in Y explained by the marginal genetic effect. The marginal genetic effect is the effects of a gene risk factor in the population as a whole, averaged over all other variables. Rg^2 can be computed as

$$R_g^2 = \left(\overline{\beta_1}\right)^2 Var(G)/Var(Y)$$

$Var(G)$ can be computed based on the assumed genetic model and allele frequency. For example, if a dominant model is assumed,

$$Var(G) = [q_A^2 + 2q_A(1-q_A)] \times [(1-q_A)^2]$$

Assume that the null hypothesis H_0 of interest is that $\beta_1 = 0$, which means no significant genetic effect on the outcome *Y* with the alternative hypothesis H_1

$$H_0\colon \beta_1 = 0 \ versus \ H_1\colon \beta_1 \neq 0$$

The maximized likelihood under the null hypothesis is given as

$$\hat{L}_0 = L_0(\hat{\beta}_0)$$

The maximized likelihood under the alternative hypothesis is given as

$$\hat{L}_1 = L_1(\hat{\beta}_0, \hat{\beta}_1)$$

The likelihood-ratio test statistics is

$$\Lambda=2*[Ln(L_1)-Ln(L_0)]$$

Assuming a two-sided alternative hypothesis, N is then computed as

$$N = \frac{(z_{\alpha/2}+z_{\beta})^2}{\Lambda},$$

where a is the significance level (set to 0.05), $1 - \beta$ is the desired power (set to 0.8). For a given number N of sampling units, the quantity $N\Lambda$ is the non-centrality parameter of the chi-square distribution under H_1.

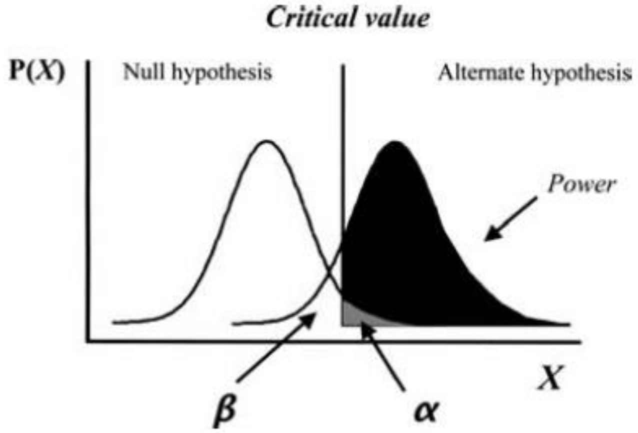
Factors that Influence Sample Size Estimation

Sample size estimation in genetic association studies requires making assumptions about biological parameters. Some of these parameters are under the control of the investigator or are known, such as power, type I error rate, or disease prevalence. However, we need to make assumptions for some of these parameters, i.e. genetic effect size and model of inheritance (additive, dominant, or recessive), degree of linkage disequilibrium between markers and trait loci, and allele frequencies at these loci, based on expert knowledge and results from previous studies. These parameters are described in the following sections.

Type I error and power

In a statistical hypothesis test, a type I error is the incorrect rejection of a true null hypothesis, while a type II error is the failure to reject a false null hypothesis. In the following graph, α refers to the type I error while β denotes type II error. The distribution of the test statistic is shown for both the null (left) and alternative (right) hypotheses. The vertical line is the critical value of the test. The black and gray areas under the alternative hypothesis represent the power of the test. The shaded gray area refers to the type I error rate of the test.

The unshaded area under the alternative hypothesis is the probability of committing a type II error.



If the null hypothesis is true, we may correctly accept the null hypothesis, or incorrectly reject the null hypothesis at rate α, which is determined by the experimenter to decide at what threshold the test will be declared significant. In this way, the sensitivity and specificity of a statistical test can be described and controlled. There is a trade-off between these two properties: Lowering the threshold increases sensitivity (i.e., increases power, reduces type II errors) but also decreases specificity (i.e., increases type I errors). Traditionally, the values of α=0.05 and β=0.2 are adopted to represent a realistic and adequate tradeoff between type I and II errors.

Linkage disequilibrium

Linkage disequilibrium (LD) is a measure of deviation from random association among alleles at two (or more) loci resulting from their close proximity on a chromosome. A genetic marker may not be the functional locus responsible for a disease but may be in linkage disequilibrium with the true functional locus.

Suppose that a sample of n_T haploid individuals is drawn from a large population, and there are only two possible alleles at each locus A and B. The four possible haplotypes are $A_0B_0, A_0B_1, A_1B_0, and\, A_1B_0$, and the observed numbers of these haplotypes in the sample will be denoted by $n_{00}, n_{01}, n_{10}, and\, n_{11}$ respectively, with $n_{00}+n_{01}+n_{10}+n_{11}=n_T$. The random variable n follows a multinomial distribution as[33]

$$P(n) = \frac{n_T!}{\prod_{i,j} n_{ij}!} \prod_{i,j} p_{ij}^{\;n_{ij}}$$

The maximum likelihood estimate of p_{ij} is n_{ij} / n_T . The frequency of the A_i allele is denoted $p_{i.}$, and its maximum likelihood estimate is $n_{i.} / n_T$, where $n_{i.}$ is the number of A_i alleles in the sample. The maximum likelihood estimate of the frequency of the B_j allele is obtained similarly. Measures of linkage disequilibrium between allele A_i and B_j include:

$$D_{ij} = p_{ij} - p_{i.}p_{.j}\text{, } D^{'} = D / D_{\max}$$

$$\text{and } \; r^2 = D^2 / P_{00}P_{01}P_{10}P_{11}.$$

The closer to zero the D' value is, the greater the amount of historical recombination between the two loci. Therefore, the value of 0 for D' indicates that the examined loci are independent of one another, while a value of 1 demonstrates complete dependency. The measure of r^2, however, has a stricter interpretation than that of D'; $r^2 = 1.0$ only when the marker loci also have identical allele frequencies. The allele at the one locus can always be predicted by the allele at the second locus. D' is affected solely by recombination and not by differences in allele frequencies between sites. r^2 is also affected by differences in allele frequencies at the two sites, and is therefore a better measure of potential allele-trait associations than D'.

Currently, empirical power estimation is usually based on the assumption that the genetic marker is a causal locus or both the marker and disease loci are in complete linkage disequilibrium. However, if the genetic marker is linked to a trait-influencing locus with a certain degree of LD, the sample size should be the total sample size divided by LD strength. As mentioned above, the degree of LD can be estimated by the correlation coefficient (r) between the genetic marker and the potential causal variant. The correlation determination r^2 is 1 if two SNPs arose from the same branch of the genealogy with no recombination between them; but it tends to be less than 1 if these SNPs came from different branches or if a cross-over event occurred between them[34-37]. Larger r^2 between genetic marker and causal locus allows smaller sample size. If N subjects are needed to reach a certain power in a direct genetic association test, the minimum sample size of N/r^2 will be required for an indirect test[38].

Marker allele frequency

When trait and marker loci with high LD have similar allele frequencies, the power to detect association is optimized. Genetic variants with low frequency require a large sample size. Therefore, a rare variant is difficult to detect in association studies.

Genetic inheritance model, disease prevalence, effect size, and disease prevalence

A genetic inheritance model (dominant, recessive, additive) describes the relationship between an individual's genotype and their phenotype or trait. The term model of inheritance refers to exactly how phenotypic values depend on the number of risk alleles. When only one copy of the disease allele is required to induce an effect on the disease phenotype, the mode of inheritance is called dominant. However, if two copies of the disease allele are required to elevate the disease risk, we speak of a recessive model of inheritance. Depending on the 'scale' with an additive mode of inheritance, the penetrance probability of heterozygous genotype is mid-way between the penetrance probabilities of both homozygous genotypes. For example, a single gene with two alleles (*A* and *a*) is related with a disease, and the two alleles contribute to phenotypic variability. An individual can either be affected or unaffected based on the genotype that the individual carries. Let q_A=frequency of the allele *A* increasing risk of disease, where $q_A+q_a=1$. The penetrance parameters are f_{AA}=probability of being affected given AA genotype, f_{Aa} probability of being affected given *Aa* genotype, and faa=probability of being affected given aa genotype, and K_p=population prevalence of the disease. K_p is the overall disease risk in the general population as follows:

$$K_p = P_r(S=1) = \sum_G P_r(S=1|G)P_r(G|q_A) = {q_A}^2 fAA + 2\, q_A q_a fAa + {q_a}^2 faa\,,$$

where S is an indicator of disease status (1=diseased; 0=not diseased). The disease risk of a genotype (*AA*, *Aa*, and *aa*) relative to the average population can be calculated as $f_{AA}/K_{p},$ $f_{Aa}/K_{p},$ *and* f_{aa}/k_p, respectively. As the penetrance of the disease variant decreases, the power to detect the genetic association also attenuates.

The term effect size can refer to a standardized measure of effect (such as *r***,** Cohen's ***d*,** and odds ratio), or to an unstandardized measure (e.g., the raw difference between group means and unstandardized regression coefficients). For instance, the relative risk (or odds ratio) parameters R_{AA} and R_{Aa} quantify the effect size of a genetic variant, and they represent that individual with genotype *AA* or *Aa* are R_{AA} or R_{Aa} times more likely to have the disease than individuals with the *aa* genotype (arbitrarily assuming that *aa* is the low-risk, reference genotype). Genotypic relative risks are likely to be in the range of 1.1–1.5 for a typical genotype[39, 40].

The relationship between the relative risk parameters R_{AA} and R_{Aa} can be specified by the underlying genetic model of inheritance as follows:

Table 2-1. The relationship between relative risk and genetic inheritance models

Genetic Model of Inheritance	Relative Risk (Genotype *aa* as the reference)	
	R_{AA}	R_{Aa}
Dominant	RR	RR
Recessive	RR	1
Additive	2RR	RR
Multiplicative genetic model	RR^2	RR

For example, relative risks of $R_{AA} = 4$ and $R_{Aa} = 1$ indicate a recessive mode of inheritance for the *A* allele, because the *Aa* heterozygote contributes no more risk than the recessive *aa* homozygote. In contrast, relative risks of $R_{AA} = 16$ and $R_{Aa} = 4$ imply a multiplicative model of inheritance, because the risk of the *AA* genotype is the square of the risk associated with the heterozygote.

Population stratification by race

Epidemiologic studies of genetic factors and disease are sensitive to population stratification into racial or ethnic groups. Numerous studies have reported that unrecognized population structure could induce bias, false-positive associations, and lack of replication in association studies[41-46]. Population stratification (i.e. confounding by ethnicity) can occur if both baseline disease risks and risk-conferring allele frequencies differ across the groups being studied (e.g., races or ethnicities).

Genome-wide association or linkage methods are dependent on the linkage disequilibrium among genetic variants on a chromosome. Differences in the pattern of linkage disequilibrium by race have been reported[3], which could affect the success of gene discovery efforts. Previous work[45] also reported that ignoring ethnicity in molecular epidemiologic studies can lead to some distortion of association estimates. Therefore, all studies should carefully consider the potential for confounding by ethnicity, ancestry, or race, and respond with an appropriate study design or analytic methods. As such, the statistical analyses in genetic association studies are often performed by race to correct for different population structures. The number of subjects required for one racial group could also be different from another group. Therefore, in genetic research design, sample size estimation should also be performed by race to optimize statistical power.

Simulating Sample Sizes and Testing Factor Effects on Sample Size Estimation

An appropriate sample size for genetic association studies requires assumptions about potential impact factors. Some factors mentioned above may have greater influence on the determination of sample size than others. For instance, in a power analysis, a non-centrality parameter is estimated, which is the effect size multiplied by a sample size factor. Therefore, a large effect size can result in a large non-centrality coefficient lambda, which produces a high power to detect an association. Under the same power level, only a small sample size is required to detect a large effect size, while a large sample size is needed to detect a small effect size.

In this study, a systematic evaluation of factor influence on sample size estimation was conducted. Instead of time-consuming data simulations, a closed-form expression of statistic formulae was applied to calculate sample sizes under a variety of different scenarios, and to test the effects under various parameter settings using SAS 9.4. Cary, NC: SAS Institute Inc., and QUANTO software. QUANTO is a computational program developed by WJ Gauderman et al. to calculate sample sizes in genetic association studies with dichotomous and quantitative outcomes under various assumptions[24, 25, 47, 48]. Using the program, we calculated sample sizes under different parameters to investigate main and interaction effects of several factors on sample size estimation, and determined how adequate statistical power can be achieved in a genetic association study.

Study design and outcome variables

Sample size estimation was performed based on genetic association studies, and single nucleotide polymorphisms (SNP) were the genetic markers for the studies. The case-control study was used to identify candidate genes that contribute to a specific disease. The outcome variable was disease status (case-control). In the case-control study, a higher frequency of the SNP allele or genotype in cases (affected individuals with a disease) compared with controls (non-affected individuals with a disease), implies that the tested variant increases the risk of the specific disease. A cohort study was used by drawing an independent sample from unrelated individuals to test genetic effect on drug/treatment response. A significant difference in mean responses between variant and non-variant carriers suggests a genetic association with the outcomes.

Statistical models

Two statistical models were applied to estimate the sample size in accordance with two different phenotypes (disease state and quantitative trait). Logistic regression was used for genetic association with diseases. Linear regression was applied for genetic association with quantitative outcomes. In the logistic regression model, an odds ratio (OR) was calculated to measure the associations between genetic variants and a dichotomous outcome (e.g. disease status). In the linear regression model, β_g was used to measure the difference in the predicted value of outcomes (e.g. drug dose) between genotypes of a genetic variant, holding other variables constant. Sample sizes were calculated by the likelihood ratio test statistic Λ for a single sampling unit based on the expected maximum log-likelihoods under the research hypothesis and the null hypothesis. For a given number N of sampling units, the quantity $N\Lambda$ was the non-centrality parameter of the chi-squared distribution under the alternative hypothesis.

Parameters for sample size estimation

For each analysis, a variety of parameter settings was tested. We assumed 1) the power of the test is 80%; 2) a significance threshold α=0.05 for a single genetic variant; 3) R^2 for pair-wise Linkage Disequilibrium (LD) was 0.5 or 1; 4) case-control ratio was 1:1for disease outcomes; 5) minor allele frequencies were 1%, 5%, 20%, or 30%; 6) genetic inheritance was additive, dominant, or recessive; and 7) genetic effect sizes varied according to trait. It is generally accepted that genetic effects on phenotypes are likely to be small, therefore, small to medium effect sizes were used. Effect sizes are defined by H. Chen[49] as small - OR= 1.68, medium - OR=3.47, and large - OR= 6.71. Cohen's f^2 method define the effect sizes as small (f^2=0.02), medium (f^2= 0.15), and large (f^2=0.35) for the difference of more than two means in ANOVA or multiple regression[50]. The f^2 effect size measure for multiple regression is defined as $f^2=R^2/(1-R^2)$, where R^2 is the squared multiple correlation. Therefore, translated the effect sizes of f^2 =0.02 (small), 0.15 (medium), or 0.35(large) into R^2, this gives 0.02, 0.13, or 0.26, respectively[50]. In a genetic association test, Rg^2 is the marginal proportion of variance in Y (outcome) explained by genetic effect of the tested genetic markers.

Since genetic effect size is generally small[40], we set up the small effect sizes of odds ratios at 1.2, 1.5, 2.0, and 2.5 to calculate sample sizes for disease outcome, and we applied Cohen's f^2 values 0.02, 0.15, and 0.35 to set up R^2 at a range from small to medium of 0.01, 0.05, 0.10, and 0.25 for sample size estimation with a quantitative outcome.

Poisson regression analysis to examine factor effects on sample size estimation

Poisson regression is used to model count variables, such as the number of subjects. To test the effects of minor allele frequency, genetic inheritance model, and genetic effect size on sample size estimation, we applied Poisson regression analysis at α=0.05, β=0.80 and 1% disease prevalence with the estimated sample size as outcome variable and the three factors as predict variables. Here, the genetic markers were assumed to be causal variants in the regression model and LD was not used in the analysis. For the quantitative outcome, the genetic effect size of β_g was used for the expected outcome difference between genotypes. For the dichotomous outcome, the genetic effect size of OR was used to measure the genetic association with disease.

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in specified intervals such as time, distance, experiment, area or volume. Assume that the number of subjects in a study follows a Poisson probability distribution. Log-linear models were applied to investigate main and interaction effects of three factors on the number of required subjects. Given that N=sample size, q_A=minor allele frequency, E=genetic effect, and M=genetic inheritance mode with the additive genetic model as the reference group, the log-linear model to assess factor main effects only was:

$$\log(\mathrm{N}) = \beta_0 + \beta_1 q_A + \beta_2 \mathrm{E} + \beta_3 \mathrm{Recessive} + \beta_4 \mathrm{Dominant}$$

The full log-linear model to test interaction effects among the three predictors was:

$$\begin{aligned}\log(\mathrm{N}) = {} & \beta_0 + \beta_1 q_A + \beta_2 \mathrm{E} + \beta_3 \mathrm{Recessive} + \beta_4 \mathrm{Dominant} + \beta_5 q_A \mathrm{E} + \beta_6 q_A \mathrm{Recessive} \\ & + \beta_7 q_A \mathrm{Dominant} + \beta_8 \mathrm{ERecessive} + \beta_9 \mathrm{ERecessive} + \beta_{10} q_A \mathrm{ERecessive} \\ & + + \beta_{10} q_A \mathrm{EDominant}\end{aligned}$$

To assess the interrelationship of the three factors, the model fit was assessed, and tests were done to check the significance of the relationships. The criteria for assessing goodness of fit include Akaike Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC).

Visualizing interaction effects by smoother surface plot

To visualize the impact of the three parameters on sample size estimation, smoother surface plots were generated using a bivariate spline interpolation method[51-53] with smoothing. The interpolation method trades closeness to the original data points for smoothness by generating the z values from x, y points since the original data does not contain enough combinations of x, y, and z values to generate a surface plot. The surface formed by the interpolated data passes precisely through the data points in the raw data set.

RESULTS

Sample sizes under different parameter combinations

Two statistical models were applied to estimate the sample size in accordance with two different phenotypes (disease state and quantitative trait). Logistic regression was used for genetic association with diseases, and linear regression was used for genetic association with quantitative outcomes. For a dichotomous outcome, the genetic effect size of odds ratio was used for the measure of association between an exposure and an outcome. For a quantitative outcome, the genetic effect size of β_g was used for the expected outcome difference between genotypes.

The subject numbers required for dichotomous and quantitative outcomes with genotype as an independent variable were estimated under different parameter combinations at α=0.05 and β=0.80. For each parameter, the results were generated holding other factors constant; for instance, the minor allele frequencies were changed from 0.01 to 0.30, while the odds ratio, genetic model, linkage disequilibrium and disease prevalence were held constant.

Table 2-2 lists the sample sizes required for a dichotomous outcome (disease state) with genotype as the independent variable under different parameter combinations at α=0.05, β=0.80, and 1% of disease prevalence.

Table 2-2. The number of case-control pairs needed to detect genetic association with dichotomous outcomes under different parameter assumptions at α=0.05 and β=0.80

		Mode of Genetic Inheritance and Strength of LD[a]					
OR[b]	q_A[c]	Additive		Dominant		Recessive	
		R^2=1	R^2=0.5	R^2=1	R^2=0.5	R^2=1	R^2=0.5
1.2	0.01	21861	43722	22233	44466		>1,000,000
	0.05	4593	9186	5002	10004	173299	346598
	0.20	1405	2810	2007	4014	11337	22674
	0.30	1091	2182	1901	3802	5367	10734
1.5	0.01	3978	7956	4056	8112	783620	1567240
	0.05	847	1694	933	1866	31457	62914
	0.20	271	542	400	800	2080	4160
	0.30	216	432	391	982	998	1996
2.0	0.01	1192	2384	1219	2438	233342	466684
	0.05	259	518	290	580	9378	18756
	0.20	89	178	136	272	631	1262
	0.30	74	148	139	278	310	620
2.5	0.01	617	1234	632	1264	119982	239964
	0.05	137	274	155	310	4827	9654
	0.20	50	100	79	158	330	660
	0.30	43	86	83	166	165	330

[a] Linkage Disequilibrium (LD) between a genetic marker and the locus/loci that affect the trait is measured by correlation determination R^2.

[b] OR ($R_g = e^{\beta_1}$) = Odds Ratio of disease is calculated by the ratio of allele carriers (exposed to genetic risk factor) to non-carriers (unexposed to genetic risk factor) in cases compared with controls. Small effect size OR= 1.68, medium effect size OR=3.47, and large effect size OR= 6.71 by Chen[49].

[c] 1% disease prevalence is assumed in disease status outcome.

Among three genetic inheritance models, the largest sample sizes were required for a recessive trait while the smallest sample sizes were needed in the additive model. As expected, the higher LD between a genetic marker and the locus/loci that affect the trait (e.g. a specific disease) under study, the fewer subjects are needed to detect the genetic association with the disease. Additionally, larger effect size and larger minor allele frequency both lead to smaller sample sizes.

Table 2-3 lists the sample sizes required for quantitative outcomes (i.e. drug dose) with genotype as an independent variable under the different parameter combinations at α=0.05 and β=0.80 from general linear regression modeling. Rg^2 is the marginal proportion of variance in Y (outcome) explained by genetic effect of the tested genetic markers. For a quantitative outcome, the β_g was used for the expected outcome difference between genotypes.

Table 2-3. The number of individuals needed to detect genetic association with quantitative under different parameter assumptions when population mean=100, standard deviation=10, α=0.05 and β=0.80

		Mode of Genetic Inheritance and Strength of LD[a]					
R_g^{2b}	q_A^c	Additive		Dominant		Recessive	
		R^2=1	β_g^d	R^2=1	β_g	R^2=1	β_g
0.01	0.01	781	7.11	781	7.16	781	100.01
	0.05	781	3.24	781	3.37	781	20.03
	0.20	781	1.77	781	2.08	781	5.10
	0.30	781	1.54	781	2.00	781	3.49
0.05	0.01	153	15.89	153	16.01	153	223.62
	0.05	153	7.25	153	7.54	153	44.78
	0.20	153	3.95	153	4.66	153	11.41
	0.30	153	3.45	153	4.47	153	7.81
0.10	0.01	74	22.47	74	22.64	74	316.24
	0.05	74	10.26	74	10.66	74	63.32
	0.20	74	5.59	74	6.59	74	16.14
	0.30	74	4.88	74	6.33	74	11.05
0.25	0.01	27	35.53	27	35.80	27	500.03
	0.05	27	16.22	27	16.86	27	100.13
	0.20	27	8.84	27	10.42	27	25.52
	0.30	27	7.72	27	10.00	27	17.47

[a] Linkage Disequilibrium (LD) between a genetic marker and the locus/loci that affect the trait is measured by correlation determination R^2.

[b] R_g^2 is the marginal proportion of variance in Y (outcome) explained by genetic effect. It is also a measure of marginal genetic effect. A small effect size R_g^2= 0.02, a medium effect size R_g^2=0.13, and a large effect size R_g^2 = 0.26 Based on Cohen' f^2 index[50].

[c] q_A=minor allele frequency.

[d] β_g=Genetic effect (Difference in expected outcome between genotypes). Measured by the change in mean Y (outcome) per unit increase in genotype (e.g. dominant genetic model: *G*=1 for genotypes *AA*, and *Aa*; *G*=0 for genotype *aa*.

To detect the genetic association in the same number of subjects, the largest genetic effect was needed for a recessive trait while the smallest genetic effect was required in the additive genetic model. For example, If there are 781 subjects in a study, we expect to detect a genetic variant associated with a quantitative outcome if the genetic marker's minor allele frequency is at least 1% and the mean difference between genotypes (main effects) is at least 7.11 (additive), 7.16 (dominant), or 100.01 (recessive). Thus, a variant with recessive inheritance model has the lowest power to detect statistical significance. A negative association was also observed between minor allele frequencies and genetic effects on sample size estimation. For example, among 781 subjects, we expect to detect an additive variant associated with a quantitative outcome if the genetic marker has a small minor allele frequency 0.01 and large main effect size 7.11, or a large minor allele frequency 0.30 and a small main effect size 1.54. As we expected, higher marginal genetic effects R_g^2 require smaller sample sizes, such as when R_g^2 was 0.01, 0.05, 0.10, and 0.25, the required sample sizes were 781, 153, 74, and 27, respectively, when other parameters were held constant. From the results in Tables 2-1 and 2-2, the sample size estimation in genetic association study is significantly impacted by genetic effect size, minor allele frequency, and genetic inheritance model. Other factors such as LD between causal locus and genetic marker and disease prevalence also play an important role in the sample size estimation.

Poisson regression analysis to examine factor effects on sample size estimation

To test the effects of the three factors on sample size estimation, these parameters were used in a Poisson regression analysis at α=0.05, β=0.80 and 1% disease prevalence in disease status outcome. The estimated regression models for main effects were:

- Estimated sample size (N) for dichotomous outcome

$$\log(N) = 14.32 - 25.88q_A - 3.19E + 3.50Recessive + 0.16Dominant$$

- Estimated sample size(N) for quantitative outcome

$$\log(N) = 5.27 - 2.22q_A - 0.02E + 0.51Recessive + 0.02Dominant$$

The regression coefficients, standard deviation of the coefficients, and p values from two regression models for both dichotomous and quantitative outcomes were summarized in (Table 2-4), minor allele frequency and genetic effect size had significant inverse effects on sample size estimation (P<0.001). That is, the required sample size decreases as the allele frequency and genetic effect size increase. Holding other factors constant, the additive genetic inheritance model required significantly smaller sample sizes compared with the dominant and recessive genetic models (P<0.001).

Table 2-4. The tests of main effects for three parameters in Poisson regression analyses

Parameter[a]	Dichotomous Outcome			Quantitative Outcome		
	Maximum Likelihood Parameter Estimates			Maximum Likelihood Parameter Estimates		
	Estimate	Standard Error	P value	Estimate	Standard Error	P value
β_0	14.32	0.002	<0.001	5.27	0.006	<0.001
q_A	-25.88	0.006	<0.001	-2.22	0.028	<0.001
M						
Recessive	3.50	0.001	<0.001	0.51	0.006	<0.001
Dominant	0.16	0.002	<0.001	0.02	0.005	<0.001
E	-3.19	0.001	<0.001	-0.02	0.000	<0.001
Assessing goodness of fit						
AIC	11984452.06			281428.64		
BIC	11984478.05			281457.24		

[a] q_A=minor allele frequency; M=genetic inheritance mode, and additive genetic model as the reference group; and E=genetic effect size.

We have reported the main effects of the three parameters on the sample size estimation. The following included the results of testing interaction effects among the three parameters on sample size estimation.

The estimated full regression models including interaction terms were:

- Estimated sample size (N) for dichotomous outcome

$$\begin{aligned}\log(N) = 16.64 - 14.81q_A - 5.69E - 1.07\text{Recessive} - 0.07\text{Dominant} - 2.98q_A E \\ + 58.39q_A\text{Recessive} + 1.43q_A\text{Dominant} + 4.31E\text{Recessive} \\ - 0.01E\text{Recessive} - 59.96q_A E\text{Recessive} + 0.97q_A E\text{Dominant}\end{aligned}$$

- Estimated sample size(N) for quantitative outcome

$$\begin{aligned}\log(N) = 6.71 + 3.38q_A - 0.11E + 0.54\text{Recessive} - 0.13\text{Dominant} - 2.11q_A E \\ - 3.05q_A\text{Recessive} + 0.60q_A\text{Dominant} + 0.11E\text{Recessive} \\ - 0.01E\text{Recessive} + 1.22q_A E\text{Recessive} + 0.48q_A E\text{Dominant}\end{aligned}$$

Table 2-5 shows the results from the full models of the Poisson regression analysis. The full models fit the data better than the main effects only models based on AIC and BIC model selection criteria. The main, two-way, and three-way interaction effects among minor allele frequency, the type of genetic inheritance model, and genetic effect size were significantly associated with the estimated sample sizes in both dichotomous and quantitative outcomes

(P<0.0001). However, the genetic effects were not significant different between dominant and additive genetic inheritance models (p=0.59), while additive genetic model is reference model. In the three-way interactions in dichotomous outcome, the required sample size in the recessive inheritance model significantly decreased more (P<0.001) when the minor allele frequency and effect size increased, compared with that in the additive genetic model.

Table 2-5. The tests of main and interaction effects for three parameters in Poisson regression analyses

Parameter	Dichotomous Outcome				Quantitative Outcome			
	P value[b] by LR statistics for Type 3 analysis	Maximum Likelihood Parameter Estimates			P value[b] by LR statistics for Type 3 analysis	Maximum Likelihood Parameter Estimates		
		Estimate	Standard Error	P value		Estimate	Standard Error	P value
β_0		16.64	0.01	<0.001		6.71	0.01	<0.001
q_A	<0.001	-14.81	0.13	<0.001	<0.001	3.38	0.07	<0.001
M	<0.001				<0.001			
Recessive		-1.07	0.01	<0.001		0.54	0.02	<0.001
Dominant		-0.07	0.02	<0.001		-0.13	0.02	<0.001
E	<0.001	-5.69	0.01	<0.001	<0.001	-0.11	0.001	<0.001
q_A×E	<0.001	2.98	0.10	<0.001	<0.001	-2.11	0.01	<0.001
q_A× M[a]	<0.001				<0.001			
q_A× Recessive		58.39	0.13	<0.001		-3.05	0.09	<0.001
q_A× Dominant		1.43	0.16	<0.001		0.60	0.11	0.002
E× M[a]	<0.001				<0.001			
E× Recessive		4.31	0.01	<0.001		0.11	0.001	<0.001
E× Dominant		-0.01	0.02	0.59		-0.01	0.002	<0.001
q_A× E× M[a]	<0.001				<0.001			
Recessive		-59.96	0.11	<0.001		1.22	0.01	<0.001
Dominant		0.97	0.13	<0.001		0.48	0.02	<0.001
Assessing goodness of fit								
AIC	6748165.45				34880.50			
BIC	6748227.95				34949.13			

[a] q_A=minor allele frequency; M=genetic inheritance model, and additive genetic model as the reference group; and E=genetic effect size (In quantitative outcome, β_G was used for the expected outcome difference between genotypes. In dichotomous outcome, OR was used for odds ratio as a measure of association between an exposure and an outcome).

[b] P value by likelihood ratio statistics for type 3 overall test of significance.

Interaction effects illustrated by smoother surface plots

The surface plots revealed the interaction effects of the three genetic inheritance models in dichotomous and quantitative outcome variables. The axes are scaled to include all positive values representing the sample size. Each axis is labeled with the corresponding variable. The tick marks on the axes are divided into five even intervals in X (minor allele frequency from 0.01 to 0.30) and Y (genetic effect size from 1.5 to 25 for the expected mean difference between genotypes) axes, and eight even intervals in Z (sample size) axis. The reference lines at the major tick marks on all axes were drawn.

Surface plots graphically represent the relationship between independent and dichotomous dependent variables, as well as the interaction between minor allele frequency and genetic effect size of odds ratios on the estimated sample sizes in the three genetic inheritance models (data not shown). The plots revealed that when minor allele frequency approaches 0.01 and odds ratio approaches 1, the recessive genetic inheritance model required much larger sample size compared with dominant and additive genetic models. Furthermore, when minor allele frequency and odds ratio increase, the required sample size in the recessive genetic inheritance model decreased faster compared with other two models. The results were consistent with the results in Table 2-5 from Poisson regression analyses.

In Figure 2-1, three surface plots graphically represent the relationship between the independent and quantitative dependent variables, and illustrate the interaction between minor allele frequency and genetic effect size of the expected mean differences between genotypes on the estimated sample sizes in the three genetic inheritance models. Figure 2-1 A reveals that the surface plane declines along with the increase in minor allele frequency. However, Figure 2-1 B and 2-1 C shows that the surface plane declines along with the increase in effect size. The results indicate that minor allele frequency plays an important role in estimation of sample size in the recessive genetic model, while genetic effect size has a more significant influence on the desired sample size in dominant and additive genetic models.

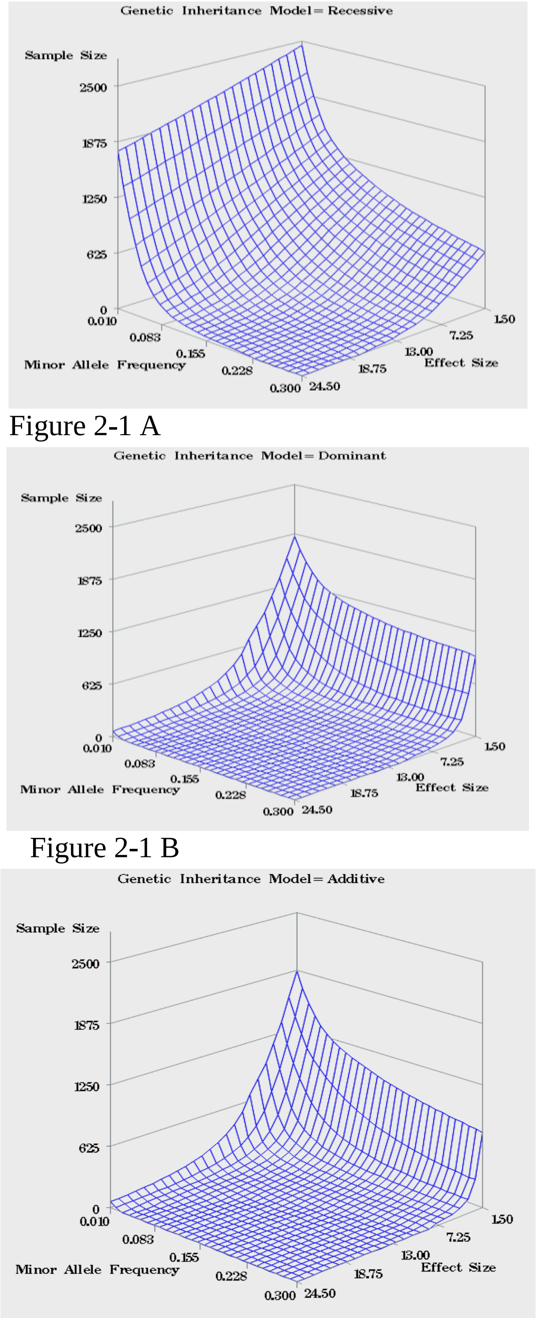


Figure 2-1 A

Figure 2-1 B

Figure 2-1 C

Figure 2-1. Surface plot showing effect of minor allele frequency and genetic effect size on sample size for a quantitative outcome under a A) recessive, B) dominant or C) additive genetic model

DISCUSSION AND CONCLUSIONS

Genetic association studies assess the association between genetic polymorphisms and phenotypes, such as disease states and drug responses. Low power and type I error inflation are major problems in these studies. Adequate sample size is crucial to assure study validity, accuracy, reliability, and integrity. An insufficient sample size lacks precision, leading to a false negative conclusion (type II error), whereas an over-powered study wastes time and resources.

Study design and choice of outcomes for genetic association studies follows the same concepts and principles as any epidemiological study. A case-control study is a typical observational study. For dichotomous outcomes, we applied the case-control design to compare the frequency of SNP alleles in two well-designed groups of individuals: cases, who have been diagnosed with the disease under study, and controls, who are either known to be unaffected, or who have been randomly selected from the population. The power in genetic association studies is a function of three things: the strength of association or outcome difference between the case and control subjects, the allele frequency, and the sample size[54]. Although there is a concern about linkage disequilibrium between genetic markers and the true causal allele, it may be true that the allele itself is functional and directly affects the expression of the phenotype or the genetic marker may have complete linkage disequilibrium with the true causal allele. Therefore, among the five parameters, we mainly focused on testing minor allele frequency, genetic effect size and the genetic inheritance model by holding other parameters constant.

Sample size estimation is influenced by multiple factors. We report the minimum sample sizes required under different parameter combinations in Tables 2-2 and 2-3. The Tables present reference information for required sample sizes for readers to make practical assumptions for their studies. A reasonable sample size should depend on research objective, study design, and technique/financial limitations.

Minor allele frequency has a significant impact on the desired sample size. From Table 2-2, we see that the required sample size is very large (more than one million) if we want to detect a small genetic effect size (OR=1.2) for a variant with a minor allele frequency 0.01 under a recessive genetic effect model. To reach a statistical power of 80% when minor allele frequency is lower than 20%, the required sample sizes were at least two times more than the sample sizes for minor allele frequencies greater than 20%. This relationship between minor allele frequency and statistical power was also reported by Scott[54], who indicated that when allele frequencies are less than 20%, the power falls off dramatically in a case-control genetic association study.

Using our previously published study to illustrate how statistical power was impacted by the factors of minor allele frequency, genetic inheritance model, and the strength of LD between the genetic markers and causal variant. The case study used a case–control design to determine associations between asthma and four common SNPs of the *NPPA* gene in white participants[55]. Atrial natriuretic peptide (*ANP*) plays an important role in the lung and in augmenting allergic inflammation in asthma. The gene encoding *ANP, NPPA,* is located on chromosome 1p36, a region that has been linked to asthma. There were two cohorts in the study. A screening cohort was used to find significant genetic variants with asthma, and a replicate cohort was used to confirm the findings from the screening cohort. The screening cohort consisted of 336 asthmatic and 154 non-asthmatic controls. The replicate cohort consisted of 172 asthmatic cases and 115 healthy controls. The frequency of minor allele C for SNP rs5067 was 0.10. A dominant genetic model was assumed with the risk range of asthma in C carriers having odds ratios from 0.20 to 0.50.

In the post hoc power analysis, the power to detect associations between the C allele and asthma could be 75% at odds ratio 0.50 in the screening cohort, and 98% at odds ratio 0.24 in the replicate cohort. To confirm the post hoc power prediction, we compared the calculated power with the study results. The results showed that the C allele of SNP rs5067 was associated with asthma in the screening and replicate cohorts: ORs (95% confidence intervals) were 0.50 (0.29–0.84; P = 0.009) and 0.24 (0.11–0.53; P=0.0001) adjusted by age, gender and body mass index, respectively. These results confirmed our prediction for statistical power at 75% for an odds ratio 0.5 in screening cohort, and at 98% of the power for an odds ratio 0.24 in the replicate cohort. The significant association results of ORs (95% confidence intervals) 0.50 (0.29–0.84; P = 0.009) and 0.24 (0.11–0.53; P=0.0001) provided the evidence of that the genetic marker SNP rs5067 was statistically significantly associated with asthma.

To test whether our prediction is reliable and valid in the dichotomous outcome study, we also compared our results with others. E. P. Hong et al[28] computed the effective sample size and statistical power based on case-control study design using Genetic Power Calculator developed by Purrcell *et al*.[56], while we applied QUANTO program[24, 25, 47, 48]. They reported that 248 cases were needed to test a SNP marker under the assumption of an odds ratio of 2, 5% disease prevalence, 5% minor allele frequency, complete linkage disequilibrium, 1:1 case/control, and a 5% error rate in an allelic test. They also found that a smaller sample size was required under a dominant model than other genetic models, and a much lower sample size was required under a dominant model with a strong effect size, common SNP, and increased LD. Our study results confirmed their findings, and additionally we applied Poisson regression models to test the main and interaction effects on sample sizes among the three factors of genetic inheritance model, minor allele frequency, and genetic effect size (OR). The results from the Poisson regression analyses help us to better understand how these factors influence each other in sample size estimation.

We have discussed how the factors affect the sample size estimation in a dichotomous outcome study. Here we are discussing the parameter influence on sample size estimation in a quantitative outcome study. From Table 2-3, we could find that the allele frequency (q_A) has a significant impact on sample size estimation. At R_g^2 =0.25, when the q_A changes from 0.01 to 0.30, the genetic effect sizes must be at least as a range of 35.53 to 7.72, respectively, to be able to detect the effects of an additive genetic inheritance variant in 27 subjects. In other words, among the same number of subjects, a lower allele frequency q_A requires a larger genetic effect size. The genetic association study with quantitative outcome has the lowest power to detect the effects of a variant under a recessive genetic model. For instance, if we have 27 subjects, it is unlikely we will find a genetic association in a variant with minor allele frequency 1% under a recessive genetic model because it is unlikely the outcome difference can reach 500 units between genotypes.

To verify our research findings in the quantitative outcome study, we used our previously published study to perform a post hoc power prediction. The study was to determine the quantitative influence of vitamin K epoxide reductase complex subunit 1 (*VKORC1*) polymorphism G3673A on warfarin dose requirements in 205 Turkish patients[57] . Warfarin is the most widely prescribed anticoagulant for the prophylaxis and treatment of venous and arterial thromboembolic disorders. Warfarin has a narrow therapeutic range and shows large inter-individual variation in dose requirements. The genetic variability is in part responsible for large differences in required warfarin dose. Based on an additive genetic model, the genotype of *VKORC1* 3673 was coded as *AA* as 2, GA as 1 and GG as 0. The frequencies of the minor allele (*A*) in Turkish, Asian, African-American, and Caucasian populations were 0.50, 0.12, 0.12, and 0.43, respectively[57]. The estimated mean Warfarin dose and standard deviation in the Turkish population were 34.2±16.8, and the estimated partial R^2=0.17 for the SNP effect, which is a medium-to-large effect size by Cohen's f^2 (R^2/ (1- R^2)) was 0.20. Based on these parameters of sample size 205, additive genetic model, and a medium-to-large genetic effect size, the power to detect the association between *VKORC1* G3673A and warfarin dose requirements was 99.99%. To confirm the post hoc power prediction, we compared the calculated power with the published results from the study. The results showed that the *VKORC1* G3673A promoter polymorphism was associated with variation of weekly mean warfarin dose: for GG genotype the dose was 43.18 mg/week, for GA genotype 33.78 mg/week and for AA genotype 25.83 mg/week (P<0.0001). Patients who carried *VKORC1* variants needed a 40% lower mean weekly warfarin dose compared to wild type (P<0.0001). The results confirmed our prediction of 99.99% statistical power to discover the genetic association with warfarin dose requirements. The results indicated that effect size plays an important role in sample size estimation. A large genetic effect size predicts that not only a small sample size is required but also the genetic markers are possible causal variants or they are in a high LD with causal variants.

The main contribution of the study was to understand the main and interaction effects on sample size estimation, and the study results could be applied in other genetic association studies for genetic marker selection, sample size estimation, and statistical power prediction. Our regression models provided deep insight into how the three factors interact with each other in sample size estimation for both quantitative and dichotomous outcome studies. We tested the main and interaction effects on sample size estimation among the three parameters, including effect size, allele frequency, and genetic inheritance model by Poisson regression analysis (Table 2-4 and 2-5). The negative coefficients in minor allele frequency and genetic effect size variables indicated that the lower allele frequency and smaller genetic effect size were related with larger sample size, and the positive coefficients for recessive and dominant model variables indicated that the two genetic models were associated with larger sample sizes compared with additive genetic effect models. The results confirmed the findings in Tables 2-2 and 2-3. In the Poisson regression models for testing main and interaction effects with dichotomous and quantitative outcomes (Table 2-5), all the two-way and three-way interaction effects were statistically significant with the sample sizes among the three parameters. Therefore, when we select the genetic markers in a genetic association study, we should consider the significant interaction effects among the three factors. Such as, we need to set up the different minimum accepted minor allele frequencies for different genetic inheritance models to reach a specific statistical power.

To visualize the evidence of interaction effects among the three factors, surface plots were presented to show the three factors influence on sample sizes. Under the recessive model, a higher peak in Figure 2-1A was observed at the back corner than that in Figure 2-1B and 2-1C, which implies that the highest sample size required when the minor allele frequency approaches 0.01 and OR tends to 1. The Figure 2-1A shows that as allele frequency increases, the required sample size significantly decreases. From Figure 2-1B and 2-1C, we also observe that as effect size increases, the required sample size significantly decreases. These plots provide visual evidence that a recessive genetic model has the lowest power in genetic association studies, and therefore, the largest sample size is required; genetic effect size is an important factor on sample size under dominant and additive genetic model; and minor allele frequency significantly impacts sample size under the recessive genetic model.

In summary, the study provided evidence that required sample sizes in genetic association studies were significantly associated with genetic inheritance model, effect size, and allele frequency at a specific type I error rate and power level. Among the three genetic models the recessive model required the largest sample size to detect the effect of a variant. Since a very large sample size needed to discover the causal variants if the minor allele frequency and genetic effect size are small, we have to select genetic markers with allele frequencies at least 1% and genetic effect size at least 2. The interaction effects among the three factors were statistically significant, which implies that the three factors affect the required sample sizes, and the effect of one factor on the sample size estimation is different at the different values of the other two factors. Minor allele frequency is the most important factor under the recessive genetic model, and the genetic effect size has the most influence under dominant and additive genetic models on the sample size estimation.

We tested parameter influence on the sample size estimation under various assumptions. Our simulation results could be applied in real genetic association studies to help researchers in genetic marker selection, sample size estimation, and statistical power prediction.

Chapter 3: Two-stage Adjusted Bonferroni Procedure for Type I Error Correction in Multi-Locus Genetic Association Studies

Abstract

Background: Genetic association studies assess the association between phenotypes of interest and multiple genetic polymorphisms or markers. Because there are multiple genetic markers, these association tests involve multiple statistical comparisons, leading to type I error inflation. The Bonferroni correction directly targets the type I error problem but it may yield too conservative conclusions due to correlation among genetic markers. The correlation among these markers violates the independence assumption of the Bonferroni procedure, resulting in type II error inflation.

Objectives and Methods: We proposed a two-stage adjusted Bonferroni correction procedure, which corrects for the multiple-comparison problem with type I error inflation. We first simulated p-values that indicate the significance of association between single nucleotide polymorphisms (SNPs) and a disease state. The simulation was based on the linkage disequilibrium (LD) structure of 267 SNPs taken from HapMap[12] and reflected small, medium and large associations. The second step was to derive the effective number of independent tests based on LD structure among the 267 SNPs. Three algorithms were used to separately estimate the effective number of independent tests, resulting in numbers of Haplotype blocks and singleton SNPs. The point-wise error rates were calculated using the family-wise error rate divided by the number of blocks and singleton SNPs. The point-wise rates provided the threshold values to determine whether or not SNPs were significantly associated with the disease. Finally, we compared the p values with the point-wise error rates across blocks and singletons. If some SNPs in an LD block had smaller p values compared with the point-wise error rate, then the Holm–Bonferroni method and dependent false discovery rate to highly-correlated test statistics (FDR) were applied within each significant LD block. If a p value is equal to or smaller than the threshold value, it suggests the SNP is statistically significantly associated with the disease.

Results: Among three haplotype blocking methods, the Gabriel algorithm generated the largest number of independent tests, resulting in the smallest number of significant SNPs. The numbers of significant SNPs were 4-9 by the standard Bonferroni correction and Holm-Bonferroni method, 9 from the empirical experiment-wise critical value method, and 2-267 by FDR adjustment across three levels of p values at the family-wise error rate 0.05. Compared with the five correction methods, the two-stage adjusted Bonferroni correction generated numbers of significant SNPs falling between the conservative standard Bonferroni correction and the liberal dependent FDR.

Conclusion and Implication: Our two-stage adjusted Bonferroni type I error correction procedure applied the statistics technique for understanding biological/genetic data, provided a new, simple, easy way to control for type I error to increase the specificity in hypothesis testing accounting for LD variation for both within- and across-blocks, improved the statistical power by increasing the testing sensitivity, and introduced a better way than the traditional Bonferroni correction to control for type I error in genetic association studies.

Keywords: Genetic association study; Type I error; Linkage disequilibrium; Bonferroni type I error correction; Haplotype block; Point-wise error rate; Family-wise error rate, False discovery rate.

INTRODUCTION

Genetic makeup plays an important role in the development of human diseases such as asthma, hypertension, and diabetes. Interacting with lifestyle and environmental factors, single gene or polygenic disorders cause at least six thousand human diseases[13]. A genetic association study identifies the underlying genetic basis of a particular disease trait by finding genetic polymorphisms or variants associated with the trait. There are many types of genetic polymorphisms, but this study focused on single-nucleotide polymorphisms (SNP). A SNP is defined as a genetic variant that occurs in coding, non-coding, or untranslated regions of the genome in at least 1% of the population. These variants in human DNA sequences can affect how humans develop diseases and respond to pathogens, drugs and other agents.

In genetic research, the case-control design has been used widely to evaluate genetic susceptibilities to complex human diseases and markers, i.e. SNPs, to localize disease gene variants. A case-control genetic association study determines whether the presence of a genetic variant increases the risk of a disease in a large population of unrelated individuals. In the study, there are two well-designed groups: the case group includes individuals with the disease under study, and the control group includes individuals without the targeted disease. Causal genetic variants are sought based on the assumption that an individual carrying one or two copies of a disease risk variant is more likely to develop the disease, and therefore an increased frequency of a variant or genotype in cases compared with controls implies that the variant may be associated with the disease.

The development of high-throughput genotyping technology has led to the simultaneous analysis of millions of single nucleotide polymorphism markers. Therefore, modern genetic association studies involve multiple SNPs and/or genes, to test a large number of hypotheses for direct or indirect association with disease phenotypes. When multiple hypothesis tests are performed in a study, there is a risk of inflation of the type I error rate. In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (a "false positive")[58]. The Bonferroni correction is commonly used to control for the Family Wise Error Rate (FWER), which is the probability of making one or more false discoveries, or type I errors, among all the hypotheses when performing multiple hypotheses test. The correction method reduces the chances of obtaining false-positive results when multiple pair-wise tests are performed on a single set of data. The procedure is simple, but conservative and lacks power if several highly correlated tests are undertaken.

In multi-locus genetic association studies, closely spaced genetic markers often yield high correlations because of extensive linkage disequilibrium (LD) among them. Therefore, tests performed on each genetic marker are usually not independent of each other. This violation of the independence assumption makes the Bonferroni procedure less effective, and the point-wise error rate for each test should be adjusted to achieve the experiment error rate at a nominal level. However, if the number of independent tests can be inferred correctly, the standard Bonferroni correction can still be applied to rapidly adjust for multiple testing. Based on this idea, previous studies have reported using the number of independent tests by haplotype blocks and principle component method[59-61]. Since the strength of LD varies from moderate to high in the haplotype blocks, the same point-wise error rate applied to all SNPs within a haplotype block could still inflate the experiment-wise type I error. For instance, when a point-wise error rate α_1 is applied to a haplotype block with *n* SNPs, the real error rate for the block may be over α_1 due to recombination that might have occurred among these SNPs, and therefore these SNPs are not in complete linkage disequilibrium.

Considering these issues, we developed a two-stage adjusted Bonferroni correction procedure. The correction is to 1) derive the effective number of independent tests based on LD structure among multiple loci; 2) use standard Bonferroni to calculate the point-wise error rate for each derived independent test; and 3) apply the Holm–Bonferroni method and dependent FDR to highly-correlated test statistics within each significant LD block, which is defined as a block in which any p values for SNPs are smaller than the point-wise error rate. We validated this new procedure using standard Bonferroni, Bonferroni adjusted by number of independent tests, empirical experiment-wise critical value method, Holm–Bonferroni method, dependent FDR[62], and finally a real clinical data study.

Current Methodology for Type I Error Correction

Standard and modified Bonferroni correction methods

The Bonferroni correction is applied when multiple independent or dependent hypotheses are tested[63]. In our previously published study, we applied the Bonferroni correction to establish a P value cutoff of 10^{-7} for a significant association for type I error of 0.05 assuming 550,000 tests for a GWAS with warfarin maintenance dose[64]. Although the Bonferroni correction can be applied to independent and dependent hypotheses, RC Johnson[65] indicated that one of the key assumptions of a Bonferroni adjustment is that all comparisons are independent, and for non-independent tests, Bonferroni adjustment may lead to over-correction. In a genetic association study, neighboring SNPs on a chromosome tend to be inherited together in blocks and are non-independent[66], making a strict Bonferroni adjustment overly conservative. However, modified Bonferroni correction methods were developed to adjust the type I error when dependent hypotheses are tested. The following includes a brief introduction to standard and modified Bonferroni correction methods that were applied in our study.

In hypothesis testing, a test statistic *T* from the observed data is calculated to decide whether the null hypothesis (H_0) should be rejected. The p value is defined as the probability of obtaining a test statistic at least as extreme as the *T* value under the condition that H_0 is true: p=P ($T \geq t \mid H_0$). If the *p*-value is smaller than a threshold α (traditionally set at 0.05), then the null hypothesis is incorrectly rejected. If *m* multiple hypotheses are tested at α level = 0.05 for each test in a study, it is expected that 5% of the *n* tests will be declared statistically significant, when H_0 is in fact true for all of these tests. The standard Bonferroni correction[67] simply sets the point-wise significance cut-off at α/*m* as in the following inequality,

$$p_r\left\{\bigcup_{i=1}^{m}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \alpha \;\; (0 \leq \alpha \leq 1)$$

to control *m* independent tests with corresponding p values P_i under the family-wise error rate α. Therefore, standard Bonferroni correction simply sets the point-wise significance cut-off at α/m SNPs. The step-down Bonferroni method is more powerful (smaller adjusted *p*-values) while in most cases maintaining strong control of the family-wise error rate. The step-down method was pioneered by Holm[68] (1979) . Bonferroni-Holm correction for multiple comparisons is a sequential rejection version of the simple Bonferroni correction for multiple comparisons and strongly controls the family-wise error rate at level alpha. The Holm Bonferroni method controls the family wise error rate without assuming independence.
The Bonferroni step-down (Holm) *p*-values are obtained from

$$\tilde{p}_{(i)} = \begin{cases} mp_{(1)} & for\; i = 1 \\ \max(\tilde{p}_{(i+1)}, (m - i + 1)p_i & for\; i - 2 \ldots\ldots m \end{cases}$$

As always, if any adjusted p value exceeds 1, it is set to 1.

Type 1 error correction by permutation procedure

The robust but computationally intensive permutation test[69] is widely used in genetic association studies as an alternative to the Bonferroni correction for multiple-testing correction among dependent tests. To find the permutation-based empirical experiment-wise critical value for the overall 0.01 or 0.05 type I error rate, a permutation is performed by random reassignment of case-control status to the data. A p value is calculated for each of *m* multiple tests (e.g. the number of *m* tested genetic markers) on the reassigned data set, and the smallest *p*-value in the *m* multiple tests is recorded. The procedure is then repeated for a large number of *x* times, such as 1000 times. Therefore, there are in total *x* smallest p-values, which are used to construct an empirical null frequency distribution of the smallest p-values under the null hypothesis of no true associations in the study. The smallest p values are arranged in ascending order and the first and fifth percentiles are the permutation-based empirical experiment-wise critical values for the overall 0.01 and 0.05 type I error rates, respectively. The p value calculated from the real data (such as a p value resulting from an association test between a genetic variant and a disease) can be compared with the permutation-based empirical experiment-wise critical value. If the p value is smaller than the critical value, the null hypothesis is rejected, and statistical significance is declared.

False Discovery Rate

The false discovery rate (FDR) was proposed by Benjamini and Hochberg[62] for multiple-testing inference in behavior genetics research. FDR has since been applied to adjust for statistical assessment of microarray studies[70] and for genetic association in autism[66]. The method controls the expected number of false discoveries in *n* tests by setting α at an appropriate level (i.e. 0.05) and making no assumptions about the relationship between the number of tests and the prior probability that H_0 is true. These adjustments do not necessarily control the family-wise error rate. However, FDR-controlling methods are more powerful and more liberal, and hence reject more null hypotheses, than adjustments protecting the family-wise error rate.

Suppose we test null hypotheses $H_{01} \dots \dots H_{0m}$, and obtain the p values $p_1 \dots \dots p_m$. Denote that ordered p values as $p_{(1)} \leq \cdots \leq p_{(m)}$, and order the tests appropriately: $H_{01} \dots \dots H_{0m}$. Suppose we know m_0 of the null hypotheses are true and $m_1 = m - m_0$ are false. Let R indicate the number of null hypotheses rejected by the test, where V of these are incorrectly rejected (that is, V tests are type I errors) and R-V are correctly rejected (so $m_1 - R + V$ tests are Type II errors). The FDR is defined as

$$FDR = E\left(\frac{V}{R}\right) = E\left(\frac{V}{R}\middle|R > 0\right)\Pr(R > 0) \quad where\ \frac{V}{R} = 0\ when\ V = R = 0$$

A dependent false discovery rate control method can be applied to control the false discover rate for p-values under any kind of dependence[71-73]. Let $\gamma = \sum_{i=1}^{n} \frac{1}{i}$. The dependent FDR can always control the false discovery rate at level

$$\le \frac{m_0}{m}\alpha\gamma, \quad \text{where } \alpha \text{ is the significance level.}$$

The adjusted *p*-values are computed as

$$\tilde{p}_i = \begin{cases} \gamma p_{(m)} & for\ i = m \\ \min(\tilde{p}_{(i+1)},\ \gamma \frac{m}{i} p_i) & for\ i = m-1 \ldots\ldots 1 \end{cases}$$

This formula is introduced in the SAS MULTTEST procedure.

Modeling and Parameter Testing

Data sets

SNP genotype data was downloaded from the international HapMap project[12] (HapMap Data Rel 28 PhaseII+III, August 2010, on NCBI B36 assembly, dbSNP b126). The genotype data was extracted from 200 kbp at chr10:67587369..67787368 in 174 Utah residents with Northern and Western European ancestry from the CEPH collection in 27 trios from 20 families. DNA regions including at least 10 haplotype blocks were selected. Since there is racial variation in genetic effects, only data for white individuals was extracted from the database. Individuals with >50% missing genotypes were excluded from the study cohort. The 267 SNPs in this sample with a minor allele frequency >0.1 were used. Among the 267 SNPs, mean pair-wise linkage disequilibrium, r^2, and standard deviation were 0.36 ± 0.36.

P-value simulation and empirical distributions

We performed p-value simulations for two purposes: 1) to determine empirical experiment-wise critical values, and 2) to simulate p values for the 267 SNPs. The p value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H_0) of a study question is true. The null hypothesis is usually that there is "no association" e.g. no association between a tested genetic variant and a disease. These p values are used to determine statistical significance in a hypothesis test with significance level, alpha, which is a pre-chosen probability from the study design. If the p value is less than the chosen significance level, then the null hypothesis is rejected. A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis i.e. the evidence supports an association between the tested genetic variant and the disease.

In a genetic association study, since there is linkage disequilibrium among some of the tested SNPs in a region of the genome, the p values from the association tests among these SNPs will be correlated with each other. Therefore, we simulated the p values for the null distribution based on the LD structure among these genetic markers.

We simulated 267 uniform random numbers 1000 times and then transformed these independent random numbers into p values based on the LD structure among the 267 SNPs. The details of the procedure are shown in Figure 3-1.

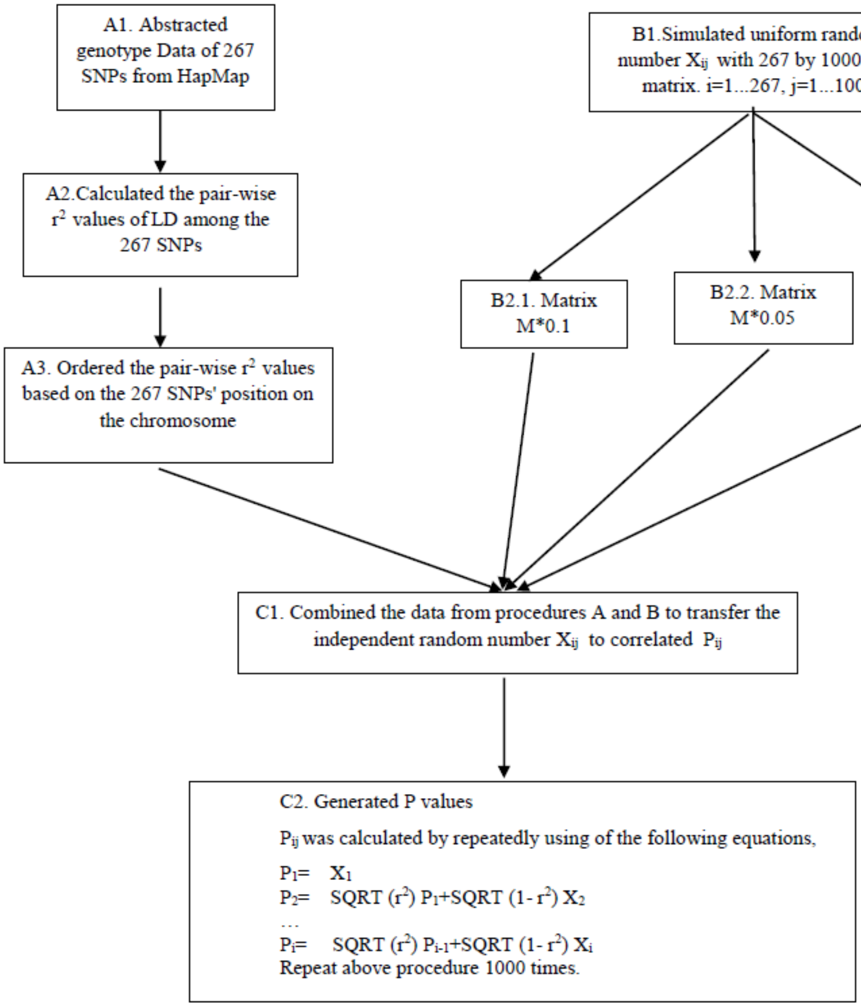

A1. Abstracted genotype Data of 267 SNPs from HapMap
A2.Calculated the pair-wise r^2 values of LD among the 267 SNPs
A3. Ordered the pair-wise r^2 values based on the 267 SNPs' position on the chromosome
B1.Simulated uniform random number X_{ij} with 267 by 1000 (M) matrix. i=1...267, j=1...1000
B2.1. Matrix M*0.1
B2.2. Matrix M*0.05
B2.3. Matrix M*0.01
C1. Combined the data from procedures A and B to transfer the independent random number X_{ij} to correlated P_{ij}
C2. Generated P values
P_{ij} was calculated by repeatedly using of the following equations,
$P_1 = X_1$
$P_2 = \text{SQRT}(r^2) P_1 + \text{SQRT}(1 - r^2) X_2$
...
$P_i = \text{SQRT}(r^2) P_{i-1} + \text{SQRT}(1 - r^2) X_i$
Repeat above procedure 1000 times.

Figure 3-1. Procedure for p value simulation

SNP genotype data from the HapMap[12] project were used to calculate pair-wise r^2 values of LD among 267 SNPs using the program Haploview[74] (A1 and A2 in Figure 3-1). These r^2 values were ordered based upon the positions of the SNPs on the chromosome (A3 in Figure 3-1). Meanwhile, 1000 repeats of 267 independent random numbers (0-1) from a uniform distribution were generated (B1 in Figure 3-1) and the 1000 x 267 matrix of these random numbers was multiplied by 0.1, 0.05, or 0.01 (B2 in Figure 3-1), to produce three matrices corresponding to large, medium and small levels of p values, respectively. The level of p values indicates how strong the evidence against the null hypothesis. A small p value (typically ≤ 0.05) indicates strong evidence against the null hypothesis (i.e. no association between a genetic marker and a disease), therefore, reject the null hypothesis to declare that the tested genetic marker is significantly associated with the disease. Finally, the numbers in the three matrices were transformed into correlated p values using the formulas shown in C1 and C2 in Figure 3-1.

To determine the empirical experiment-wise critical values, the smallest p value among the 267 SNPs was recorded for each of the three levels of correlated p values. There were 1000 smallest p values from 1000 repeated simulations for each level of p values. This is equivalent to the 1000 smallest p values derived by the permutation test that are used to construct an empirical null frequency distribution of the smallest p values under the null hypothesis of no true associations in the study. The smallest *p*-values are arranged in ascending order and the first and fifth percentiles are the permutation-based empirical experiment-wise critical value for the overall 0.01 and 0.05 type I error rates, respectively.

We then used the p-values simulated above to assign p values for the 267 SNPs. For each of the 267 SNPs, we found the smallest p-value among 1000 repeats for each of 267 SNPs at each of the three levels of p values. The p value for each SNP denotes the probability of obtaining an effect at least as extreme as the one in the sample data, assuming the SNP is not associated with the targeted phenotype (e.g. disease), that is the truth of the null hypothesis.

Deriving the effective number of independent tests based on LD structure among multiple loci

Linkage disequilibrium (LD) provides insights into how two or more loci in gametes or on chromosomes are correlated with each other. Haplotype blocks were inferred based on the linkage disequilibrium structure among markers. Suppose that a sample of n_T haploid individuals is drawn from a large population, and there are only two possible alleles at each locus A and B. The four possible haplotypes are A_0B_0, A_0B_1, A_1B_0, A_1B_1, and the observed numbers of these haplotypes in the sample will be denoted by n_{00}, n_{01}, n_{10}, and n_{11}, respectively, with $n_{00}+n_{01}+n_{10}+n_{11}=n_T$.

The random variable *n* follows a multinomial distribution as[18]

$$P(n)=\frac{n_T!}{\prod_{i,j} n_{ij}!}\prod_{i,j} p_{ij}{}^{n_{ij}}$$
.

The maximum likelihood estimate of p_{ij} $is\, n_{ij}/n_T$. The frequency of the A_i allele is denoted P_i.

The maximum likelihood estimate of this quantity is $n_{i.}/n_T$. n_i is the number of A_i alleles in the sample. The maximum likelihood estimate of the frequency of the B_j allele is obtained similarly. Measures of linkage disequilibrium between allele A_i and B_j includes $D_{ij} = P_{ij} - P_iP_j$, $D' = D/D_{max}$, and $r^2 = D^2 / P_{00}P_{01}P_{10}P_{11}$.

The closer the *D′* value is to zero, the greater the amount of historical recombination between the two loci, therefore the value of 0 for *D′* indicates that the examined loci are in fact independent of one another, while a value of 1 demonstrates complete dependency. However, the measure of r^2 has a stricter interpretation than that of *D′*; r^2 = 1.0 only when the marker loci also have identical allele frequencies. The allele at the one locus can always be predicted by the allele at the second locus. *D*' is affected solely by recombination and not by differences in allele frequencies between sites. r^2 is affected by differences in allele frequencies at the two sites, and is therefore a better measure of potential allele-trait associations than *D*'.

To obtain the number of independent tests, we constructed haplotype blocks to define the underlying pair-wise linkage disequilibrium structure among genetic markers. We applied three haplotype blocking algorithms: Gabriel's algorithm, the 4-gamete test, and the solid spine of LD measure using the Haploview software[74]. Gabriel defined haplotype blocks based on the confidence interval of the LD measure D' in 2002. In Gabriel's algorithm[75], a haplotype block is defined as a region over which a very small proportion (<5%) of comparisons among informative SNP pairs show strong evidence of historical recombination. By the algorithm, 95% confidence intervals (CI) on D' are generated and each comparison between informative SNP pairs is called 'Strong LD', 'inconclusive' or 'strong recombination'. A strong LD is defined for the pairwise D' when CI minima for upper CI bound = 0.98 and CI minima for lower CI bound = 0.70. A block is created if 95% of informative comparisons are ‘strong LD'. Wang et al[76] introduced the four Gamete rule in 2002. The population frequencies of the 4 possible two-marker haplotypes are computed for each marker pair. If all 4 are observed with at least frequency 0.01, a recombination is deemed to have taken place. Blocks are formed by consecutive markers where only 3 gametes are observed. Barrett et al[77] proposed the solid spine of the LD algorithm (SSLD). The SSLD method creates blocks of SNPs that have contiguous pairwise D' values of greater than 0.80. The method searches for a 'spine' of strong LD running from one marker to another along the legs of the triangle in an LD chart (e.g.

Figure 3-2). This search results in the first and last markers in a block that are in strong LD with all intermediate markers while intermediate markers are not necessarily in LD with each other. For all three algorithms, the effective number of independent tests (N_e) is equal to the number of LD blocks across the DNA region and the number of inter-block SNPs (i.e. singleton SNPs). For instance, if there are 52 singleton SNPs and 17 haplotype blocks among 267 SNPs, then the effective number of independent tests is 69. The effective number of independent tests is used in a Bonferroni adjustment to estimate study-wide significance thresholds for multiple-test correction in a study.

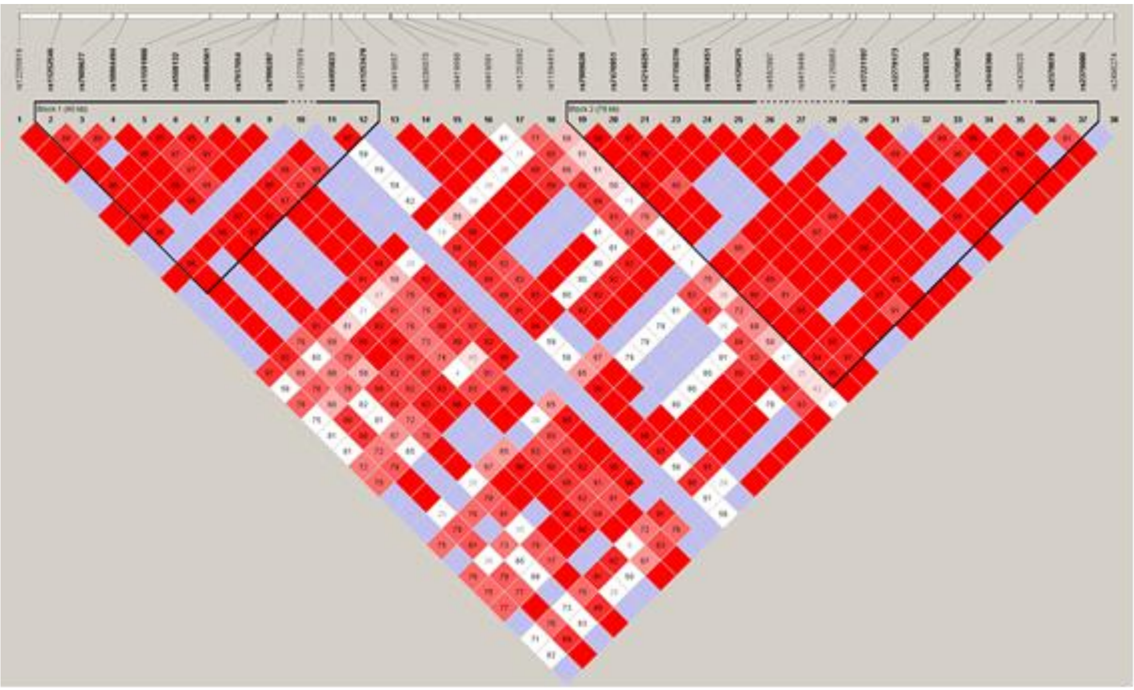


Figure 3-2. Haplotype blocks from SNP sequence data at chr10:67587369..67787368 in 174 Utah residents.

The genotype data for 267 SNPs were extracted from HAPMAP[12], and the LD structure among the 267 SNPs was determined using each of three algorithms. Figure 3-2 illustrates the partial LD structure for these data. The plot includes eight singleton SNPs and two haplotype blocks generated by the Gabriel algorithm. The color schemes in the plot are in accordance with the strength of LD between SNPs. The log of the odds (LOD) was used to measure LD between loci. LOD>2 indicated significant LD. Therefore, bright red indicates a strong LD with the log of the odds (LOD) ≥2 and D'=1, while white indicates a weak LD with

LOD<2 and D'<1. Between the strong and weak LDs, blue represents LOD<2 and D'=1 while shades of pink/red shows LOD≥2 and D'<1. The plot includes eight singleton SNPs and two haplotype blocks generated by the Gabriel algorithm.

Adjusted point-wise significance level

The point-wise error rate (PWER) is the type I error rate for an individual test or the probability of incorrectly rejecting the null hypothesis. Experiment-wise error rate, also called family-wise error rate (FWER), is the probability of making at least one type I error when performing a large number of related tests. Keeping FWER (a_f) at a nominal significance level, the adjusted PWER is denoted as a_p. Since current genetic association analyses involve a large number of markers, multiple statistical tests are commonly performed. The multiple testing could result in a large FWER. However, if we set the PWER, α_p to a low level, the family-wise error rate can be controlled. In our study, we calculated the adjusted PWER thresholds for standard Bonferroni, adjusted Bonferroni by the number of independent tests, and percentiles from the empirical distribution of p values.

Under the standard Bonferroni correction, we assumed that the *m* hypothesis tests are independent, and the point-wise error rates are obtained as

$$\alpha_B = {}^{\alpha_F}/_{m} \; where \; \alpha_F = 0.05 \; and \; (1-\alpha_B)^m \approx 1 - m\alpha_B \; for \; small \; \alpha_B$$

Under the adjusted standard Bonferroni correction, we used the effective number of independent tests (N_e) to estimate the point-wise error rate

$$a_j = {}^{a_F}/_{N_e} \; where \; a_F = 0.05 \; and \; (1-a_j)^{N_e} \approx 1 - N_e a_j \text{ for small } a_j.$$

Finally we obtained three point-wise error rates for the large, medium, and small simulated p values. These simulated p values were arranged in ascending order, and the first and the fifth percentiles became the simulation-based empirical experiment-wise critical values for the overall 0.01 and 0.05 type I error rates, respectively.

Adjusted point-wise error rate among and within haplotype blocks by the step-down Bonferroni method

To calculate the point-wise error rate among and within haplotype blocks, we proposed a two-stage adjusted Bonferroni correction procedure to 1) derive the effective number of independent tests based on linkage disequilibrium (LD) structure among multiple loci; 2) calculate the point-wise error rate among haplotype blocks based on the effective number; and 3) apply the Holm–Bonferroni method and dependent false discovery rate (FDR) to highly-correlated test statistics within each LD block.

The three haplotype blocking algorithms group correlated SNPs across the DNA region into haplotype blocks. However, the strength of LD varies from moderate to high within the haplotype blocks. For instance, 267 SNPs could be grouped into 17 haplotype blocks and 52 singleton SNPs by the Gabriel method, resulting in 69 independent tests. The point-wise error rate is calculated as $0.05/69=7.25\times10^{-4}$. In this situation, the same point-wise error rate is applied to all SNPs within a haplotype block, such as there are z SNPs in a haplotype block, and then the significant level 7.25×10^{-4} are applied into the significant test in the z SNPs. This adjustment may still inflate the family-wise type I error. Therefore, if there are significant SNPs within a haplotype block, we further apply the Holm–Bonferroni method and dependent FDR to highly-correlated test statistics within each LD block to get the adjusted the p values for SNPs within each of haplotype blocks.

Declaring significant SNP associations

For each of the type I error adjustment methods, the 267 simulated p-values were compared to the three empirical experiment-wise critical values (generated from three-level p values) and the adjusted point-wise significance levels to control the type I error rate at 0.05. If the p-value is smaller than the critical value or the adjusted point-wise significance level, the null hypothesis is rejected, indicating that the SNP has a statistically significant association with the targeted phenotype (e.g. a disease).

RESULTS

P-value simulations based upon LD

The genotype data for 267 SNPs were extracted from HAPMAP[12], and the pair-wise r^2 as a measure of LD among the 267 SNPs was calculated using Haploview[74]. There were 35778 total pair-wise r^2 values among the 267 SNPs, however, we only used 267 r^2 values. For example, SNPs A, B, C, D, and E are next to each other on the DNA sequence, and the pair-wise r^2 values between AB, BC, CD, and DE were taken for the p value simulation. The mean and standard deviation of the 267 pair-wise r^2values was 0.36 ± 0.36. Most of the r^2 values were smaller than 0.60 (77.15%), which implies that a large number of independent tests could be inferred. The LD structures among the SNPs on the tested DNA region resulted in certain correlation patterns of p-values if we replicate the hypotheses a number of times. Also the LD structure can derive different numbers of independent tests based on different haplotype algorithms. Therefore, it is important to understand the distribution of these 267 pair-wise r^2 values as shown in Figure 3-3.

Using the 267 pair-wise r^2 values between two adjacent SNPs, we produced correlated p values based upon independent uniform random numbers at three levels to understand if the performance of the two-stage type I error correction is different at the different strengths of association between SNPs and disease. Figure 3-4 shows the p value distributions for large, medium and small p values generated by multiplying uniform random numbers by 0.1, 0.05, and 0.01, respectively, and then transferred the independent numbers into correlated p values based on the LD structure among these SNPs. The three distributions are approximately symmetric. The ranges for large, medium and small p value levels were 0.000-0.021, 0.0000-0.0096, and 0.0000-0.0021, respectively. Table 3-1 summarizes the descriptive statistics for the three levels.

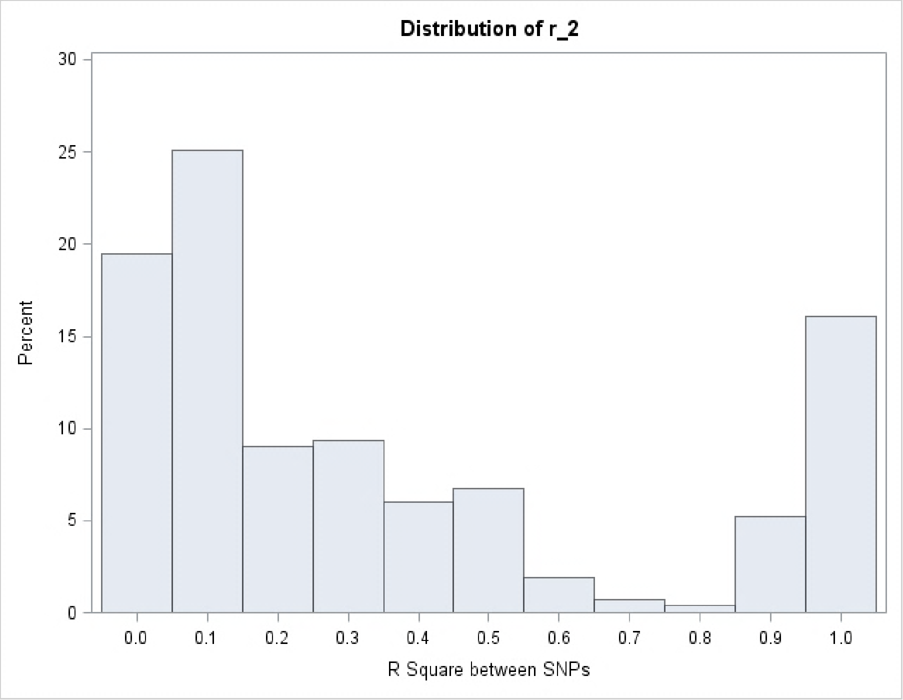


Figure 3-3. Distribution of 267 pair-wise r^2 among 267 SNPs calculated by Haploview[74] from 200 kbp at chr10:67587369..67787368 in 174 Utah residents

Table 3-1. Descriptive statistics for the three levels of p values

Level	Mean ± Std.	1st percentile	5th percentile
Small	$7.5\times10^{-4} \pm 3.8\times10^{-4}$	4.0×10^{-5}	1.6×10^{-4}
Medium	$3.8\times10^{-3} \pm 1.9\times10^{-3}$	2.0×10^{-4}	8.0×10^{-4}
Large	$7.5\times10^{-3} \pm 3.8\times10^{-3}$	4.0×10^{-4}	1.6×10^{-3}

The first and the fifth percentiles are the simulation-based empirical experiment-wise critical values for the overall 0.01 and 0.05 type I error rates, respectively. In the following analyses, we set up the significance level at 0.05. Therefore, we will use only the 5th percentile values (αe) to control the type I error rate under 5%.

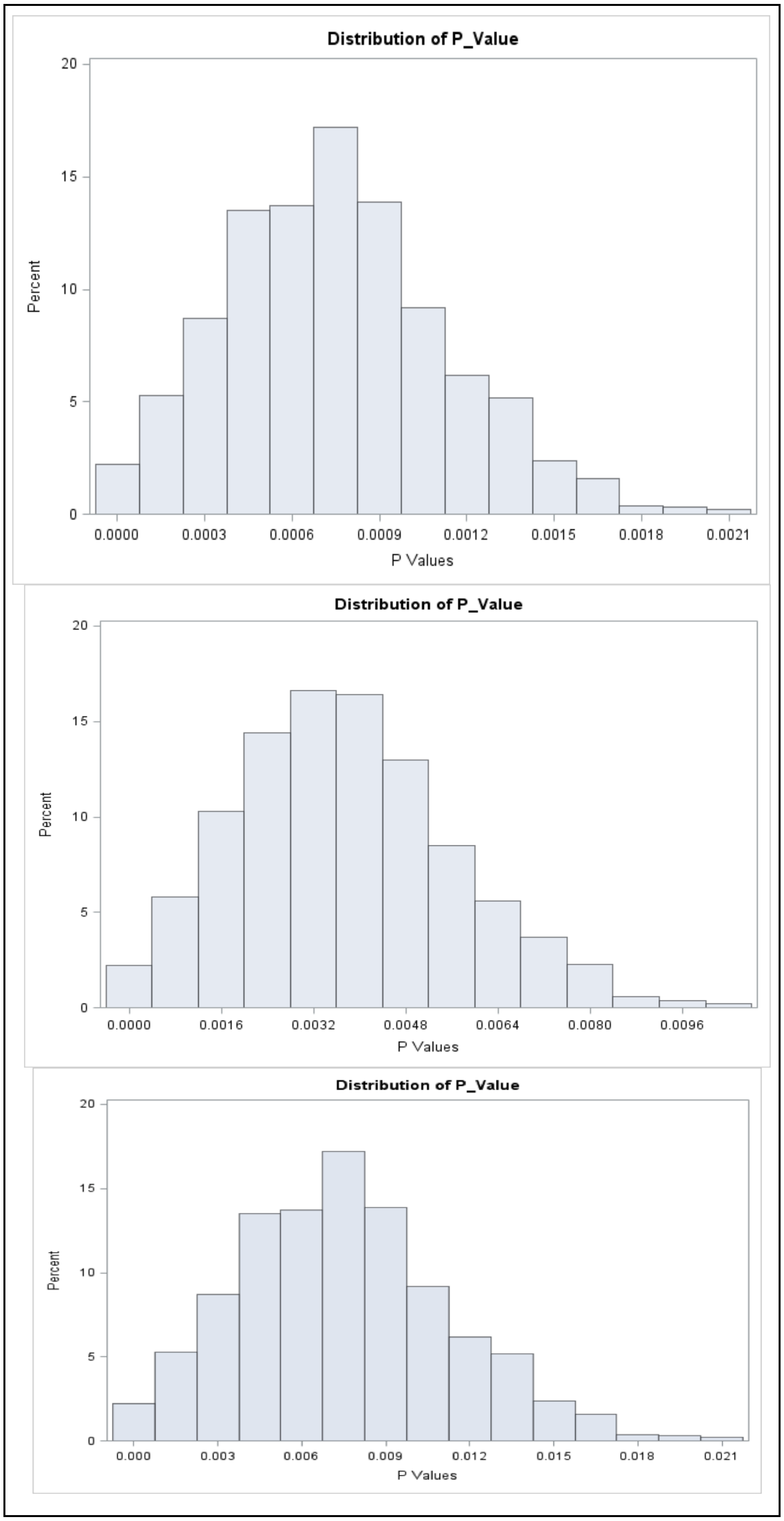


Figure 3-4. P value distributions for large, medium and small p value generated by multiplying uniform random numbers by 0.1, 0.05, and 0.01, respectively.

Effective number of independent tests

To estimate the effective number of independent tests, we used three haplotype blocking algorithms: Gabriel's algorithm, the 4-gamete test, and the solid spine of LD measure, to find the number of haplotype blocks among the 267 SNPs. The effective number of independent tests is the sum of the number of haplotype blocks and the number of singleton SNPs. The effective numbers of tests across the three haplotype block algorithms are shown in Table 3-2.

Table 3-2. The number of independent tests (haplotype blocks) across haplotype blocking algorithms

Algorithm for haplotype blocking	Singleton SNP	Haplotype Block	Number of Independent Tests
Gabriel method	52	17	69
4-Gamete test	18	42	60
Solid spine of LD measure	1	10	11

Among the three LD-based methods, the Gabriel algorithm generated the largest number of independent tests, therefore, it is the most conservative. However, compared with the standard Bonferroni correction, the three LD-based methods are anti-conservative when constructing Bonferroni significance thresholds.

Adjusted point-wise significance level

Applying the effective number of independent tests, we calculated point-wise error rates using standard and adjusted Bonferroni procedures across the three haplotype blocking algorithms. We also found the empirical experiment-wise critical values for the p values at three levels. Setting the nominal significance level to be 0.05, we calculated the point-wise error rates by dividing the nominal significance level of 0.05 by the number of independent tests for the standard Bonferroni and adjusted Bonferroni corrections by the Gabriel algorithm, 4-Gamete test and Solid spine of LD measure. The empirical experiment-wise critical values for the fifth percentile p-values at large, medium, and small levels are also shown in Table 3-3.

Table 3-3. Point-wise error rates under family-wise error rate of 0.05 among 267 SNPs in three haplotype inferring algorithms

Adjustment Method	Number of Independent Tests	Point-wise error rates
Standard Bonferroni α_β[a]	267	1.9×10^{-4}
Adjusted Bonferroni α_j[b]		
Gabriel method	69	7.2×10^{-4}
4-Gamete test	60	8.3×10^{-4}
Solid spine of LD measure	11	4.6×10^{-3}

[a] α_β is calculated by α/N of unadjusted Bonferroni correction for comparison purpose, where N is the total number of unadjusted tests such as the total number of SNPs.

[b] α_j is point error rate of adjusted Bonferroni correction by the nominal level α divided by the number of independent tests (sum of haplotype blocks plus the singleton SNPs)

Among the three haplotype algorithms in the adjusted Bonferroni methods, the solid spine of LD measure is the most liberal and yields the largest point-wise error rate because it has the fewest number of independent tests. In the point-wise error rate (1.61×10^{-4}) from the empirical p value distribution, the small-level p values is close to the point-wise error rate (1.87×10^{-4}) for the standard Bonferroni correction.

Number of tests declared significant

After determining the point-wise error rate by LD structure, we then applied the step-down Bonferroni method to get the adjusted point-wise error rate within haplotype blocks and to declare whether the tests were statistically significant. Table 3-4 lists the number of SNPs declared as significant across these methods to validate the new two-stage Bonferroni procedure and to test the procedure's performance.

Table 3-4. Numbers of SNPs with significant associations with disease across type I error adjustment methods among 267 SNPs in three haplotype inferring algorithms

Adjustment Method	Haplotype Algorithms[a]								
	Gabriel method			4-Gamete Test			Solid Spine of LD measure		
	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
Standard Bonferroni[b]	4	4	9	4	4	9	4	4	9
Adjusted Bonferroni by number of independent tests[c]	4	7	117	4	9	138	49	145	267
Empirical experiment-wise critical value method[d]	9	9	9	9	9	9	9	9	9
Holm–Bonferroni method[e]	4	4	9	4	4	9	4	4	9
Dependent FDR[f]	2	145	267	2	145	267	2	145	267
Two-stage adjusted Bonferroni correction[g]									
Holm–Bonferroni method	4	5	28	3	4	21	2	4	26
Dependent FDR	4	5	28	3	3	18	2	2	75

[a] Values of 0.1, 0.05 or 0.01 indicate that p values were generated by uniform random numbers×0.1, 0.05 or 0.01.

[b] Standard Bonferroni correction simply sets point-wise significance cut-off at α/Ns, where Ns is the total number of SNPs.

[c] Adjusted Bonferroni correction uses the nominal level α divided by the number of independent tests (sum of haplotype blocks plus singleton SNPs).

[d] Simulation-based empirical experiment-wise critical value for the overall 0.05 type I error rate served as true cutoff.

[e] Holm–Bonferroni method without assumptions required to control family-wise error rate.

[f] Dependent FDR procedure controls the expected proportion of incorrectly rejected null hypotheses ("false discoveries").

[g] Two-stage adjusted Bonferroni correction applies two procedures to control for family-wise error rate. The first stage applies the adjusted Bonferroni correction, then the second step applies Holm–Bonferroni method or the dependent FDR within haplotype blocks.

Among the 267 SNPs, the numbers of significant SNPs were 4-9 by both the standard Bonferroni correction and the Holm-Bonferroni method, 9 from the empirical experiment-wise critical value method, and 2-267 by FDR adjustment across the three p-value simulation models (large, medium, and small) in three different haplotype algorithms at the family-wise error rate 0.05. It was not surprising that the empirical experiment-wise critical value method resulted in the same number of nine significant tests among three levels of p values because the p value distributions from the random numbers were the same since the random numbers were multiplied by constants.

After adjustment by Bonferroni correction based on the number of independent tests generated from three haplotype algorithms, the Gabriel method resulted in the smallest number of significant SNPs. Among above five correction method, the standard Bonferroni correction was the most conservative while the dependent FDR was liberal. Compared with the five correction methods, two-stage adjusted Bonferroni correction generated numbers of significant SNPs that were between the conservative standard Bonferroni correction and the liberal dependent FDR.

Application of the Two-stage Correction in Real Clinical Studies

To evaluate the two-stage Bonferroni correction, we applied the method to our previously published study that tested the association between polymorphisms in the *NPPA* gene and asthma[55]. In the published study, we applied a Bonferroni's factor to define the threshold of significance in a study of the association of *NPPA* gene polymorphisms with asthma. We compared our results from the two-stage Bonferroni correction with the published results which used the unadjusted Bonferroni type I error correction. In the genetic association study, three common SNPs (rs5063, rs5065, and rs5067) of the *NPPA* gene were genotyped among white participants in a case-control study. Among these SNPs, haplotypes were constructed using imperfect phylogeny by the software package HAP[78]. Additive and dominant genetic models were applied to assess genetic association. The following includes study background, objective, statistical method, results and conclusion, and shows how the two-stage Bonferroni correction method can be used to adjust for the type 1 error in the study.

Asthma is a complex chronic disease characterized by inflammation, constriction of the airways and bronchial hyper-responsiveness to external stimuli. Asthma affects an estimated 20 million people in the US and 300 million people world-wide. It is the most common disease of childhood and is the third leading cause of hospitalization among individuals<18 years old. Numerous candidate genes have been linked to asthma, and susceptibility loci for asthma have been mapped to regions in most chromosomes[79-85]. Moreover, a region on chromosome 1p36 has been identified as an asthma candidate locus[81]. Atrial natriuretic peptide (ANP) plays an important role in the lung and in augmenting allergic inflammation in asthma. The gene encoding *ANP*, *NPPA*, is located on chromosome 1p36, a region that has been linked to asthma. The objective of this study was to determine associations between asthma and four common SNPs on the *NPPA* gene: C/G (rs13305986) in the promoter; G/A (rs5063) in Exon 1 resulting in *NPPA* Met32→Val substitution; T/C (rs5065) in Exon 3 resulting in an Arg152→Ter substitution; and T/C in the 3'UT region (rs5067).

A case–control design was used in this study of asthma in Caucasians[55]. The screening cohort consisted of 336 asthmatic cases who participated in a large clinical trial and 154, non-asthmatic controls. The replicate cohort consisted of 172 asthmatic cases from a second clinical trial and 115 healthy controls. Demographic characteristics were well matched for cases and controls in the screening cohort. Adjusted (age, gender, body mass index) odds ratios (OR) were calculated by χ^2 and logistic regression. A dominant genetic inheritance model was assumed for all of the three SNPs. The Bonferroni method was used in the publication to define the threshold of significance, which was 0.0167 for the three SNPs, rs5063, rs5067, and rs5065, (0.05/3 SNPs).

Table 3-5. Influence of *NPPA* SNPs on the risk of asthma risk for white participants in the screening cohort[55]

SNP	Genotype	Participants		Adjusted odds ratio (95% CI)[a]	P-value
		Cases N(%)	Controls N(%)		
rs5063	AG	18(6.1)	19(12.8)	0.43(0.21-0.88)	0.02
	GG	275(93.9)	129(87.2)	1	
rs5065	CC+TC	70(24.1)	60(41.4)	0.45(0.29-0.70)	<0.0001
	TT	220(75.9)	85(58.6)	1	
rs5067	CC+TC	40(13.6)	37(25.0)	0.50(0.29-0.84)	0.009
	TT	254(86.4)	111(75.0)	1	

[a] Adjusted for age, gender and body mass index. CI=confidence interval.

Tables 3-5 and 3-6 include the p values from the screening and replicate cohorts[55]. Using the standard Bonferroni correction for the screening cohort, rs5065 and rs5067 had p-values that were less than the threshold of 0.0167, and thus were reported to be significantly associated with asthma (Table 3-4). Since p value of rs5063 (P=0.02) was larger than 0.0167, no significant association was found between it and asthma. In the replicate cohort, only the p value of rs5067 (P<0.0001) was smaller than 0.0167 based upon the standard Bonferroni correction.

Table 3-6. Influence of *NPPA* SNPs on the risk of asthma risk for white participants in the replicate cohort[55]

SNP	Genotype	Participants		Adjusted odds ratio (95% CI)[a]	P-value
		Cases N (%)	Controls N (%)		
rs5063	AG	17(9.9)	9(8.1)	1.02(0.36-2.83)	0.98
	GG	155(90.1)	102(91.9)	1	
rs5065	CC+TC	53(30.8)	24(22.4)	1.16(0.60-2.24)	0.66
	TT	119(69.2)	83(77.6)	1	
rs5067	CC+TC	23(13.4)	28(25.0)	0.24(0.11-0.53)	<0.0001
	TT	149(86.6)	84(75.0)	1	

[a] Adjusted for age, gender and body mass index. CI=confidence interval.

To test our two-stage Bonferroni correction using data from the screening cohort, the number of independent tests was first determined using the three haplotype block algorithms (Gabriel confidence intervals, four Gamete rule, and solid spine of LD). The number of independent tests was based on the number of haplotype blocks and singleton SNPs. Using the confidence intervals for the Gabriel algorithm on the screening cohort data, one haplotype block of rs5067 and rs5065 and a singleton SNP (rs5063) were found. Therefore, the point-wise error rate (the threshold of significance) at the first stage was 0.05/2=0.025 for rs5063 and the haplotype block of rs5067 and rs5065. To test the significance for the singleton SNP rs5063, the p value of rs5063 (p=0.02) was compared with the point-wise error rate 0.025. Since this p value is less than the threshold, SNP rs5063 is declared to be significantly associated with asthma. For SNPs (rs5067 and rs5065) in the haplotype block, the second stage type I error adjustment using the Holm-Bonferroni step-down method within the haplotype block was performed. First the two SNPs were ordered from smallest to largest p-value: rs5065 (p<0.001) < rs5067 (p=0.009). Next the p-value for rs5065 was compared to 0.025/2. Since 0.001 is less than 0.025/2 the block is considered significant and the next p-value is tested. The p-value for rs5067 was compared to 0.025/1, and both SNPs within this block are also significantly associated with asthma. Thus rs5063 is significantly associated with asthma using our method but is not according to the standard Bonferroni correction.

Using the algorithm of four Gamete rule and solid spine of LD methods, we only discovered one haplotype block among the three SNPs. The Holm-Bonferroni step-down correction was applied therefore using 0.05 as the threshold. In this case, the three p values are ordered smallest to largest; the smallest p-value rs5065 (p<0.001) is compared to 0.05/3.

Since this value is less than 0.05/3, rs5065 is significantly associated with asthma, and we continue to test the other two SNPs. The p-value for rs5067 (p=0.009) is compared to 0.05/2, which again indicates a significant association. The p-value for rs5063 is then compared to 0.05/1 and again an association is found. Thus when all of the values are within a single block, the results are the same as the Holm-Bonferroni step-down method, and again, the p-value for rs5063 is significant when it was not for the standard Bonferroni correction.

DISCUSSION AND CONCLUSION

Our two-stage adjusted Bonferroni type I error correction procedure is a simple and easy way to reduce the chances of obtaining false-positive genetic markers in candidate gene and genome-wide SNP studies. In multiple-locus genetic association studies, the traditional Bonferroni correction is commonly applied to control for type I error. However, the correlation among genetic markers can result in non-independent tests, which violates the independence assumption of the Bonferroni procedure to control type I error effectively and inflates the type I error rate, resulting in more false positive genetic markers. With a small sample size, the type II error rate may be detrimentally inflated by a Bonferroni correction[65], resulting in low power to detect positive genetic markers.

Our two-stage adjusted Bonferroni type I error correction procedure introduces a better way than the traditional Bonferroni correction to control for type I error accounting for the correlation among genetic markers. Although the gold standard for multiple testing adjustment in genetic association studies is the permutation test, this test is computationally intensive, and therefore, it is not commonly used. Instead, the blocking method is a biologically meaningful, easy and fast way to achieve type I error rates closer to the desired value over a range of LD levels. Johnson et al[65] determined whether the number of informative SNPs inferred by principal components analysis (PCA) and haplotype blocking based on the LD structure could increase statistical power, i.e. the ability to detect true associations. The inferred number of informative SNPs was used in a Bonferroni adjustment to obtain multiple candidate SNPs or GWAS-wide significance threshold. Their study assumed that the SNPs within haplotype blocks and principle component groups were in complete LD or strongly correlated. They reported a moderate increase in power using this method compared to the traditional Bonferroni method. Since the strength of LD varies from moderate to high within haplotype blocks, applying the same point-wise error rate to all SNPs within a haploype block could still inflate the experiment-wise type I error. For instance, when a point-wise error rate α_1 is applied to a haplotype block with *n* SNPs, the real error rate for the block may be over α_1 due to recombination that might have occurred among these SNPs. As a result, these SNPs are not in complete linkage disequilibrium as Johnson et al[65] assumed.

To solve these problems, our two-stage adjusted Bonferroni type I error correction procedure applies linkage disequilibrium structure into statistics to account for the variation in linkage disequilibrium across-blocks and within blocks. In the first step, we adopted three haplotype block algorithms[3, 4] to find the number of independent tests based on LD structure, and then to calculate the pointwise error rate using the number of independent tests among multiple loci. Among the three algorithms, the solid spine of LD measure generated the smallest number of informative SNPs (Table 3-2) and the largest point-wise error rate (Table 3-3), and therefore, it is the least conservative type I error correction. The Gabriel method generated the largest number of independent tests, resulting in the smallest pointwise error rate compared with the other two algorithms. A small threshold value is less likely to reject a null hypothesis, resulting in a significant test. Therefore, a smaller pointwise error rate among multiple SNPs could declare less SNPs to be significantly associated with a disease. The results across the three haplotype blocking methods in our study agreed with the results reported by K. K Nicodemus[4]. Since the three blocking methods generated different pointwise errors, we would consider the LD structure among genetic markers to decide which blocking algorithm should be used. Based on the report from K. K Nicodemus[4], Gabriel blocking algorithm consistently gave a ~3.4% type I error rate across moderate and high LD conditions, which is close to the desired 5% level. Therefore, we would recommend that the Gabriel blocking algorithm is the first choice for inferring the number of independent tests in our two-step method.

In the second step, we further adjust the type I error within haplotype block accounting for the LD within each block. Our adjustment procedure was more liberal than standard Bonferroni and often more liberal than the Holm–Bonferroni correction method, but stricter than the type I error adjustment by the number of informative SNPs. Since the number of independent tests is smaller than the number of genetic markers, the pointwise error rate (the threshold of significance) is larger than that from a standard Bonferroni correction. The pointwise error rate is calculated by the familywise error rate/the number of independent tests. If there are m genetic markers and M independent tests (M<m), the pointwise error rates for standard Bonferroni correction and two stage standard Bonferroni correction are 0.05/m and 0.05/M, respectively, while 0.05/m<0.05/M. As a result, the declared number of significant tests after type I error adjustment from our two-stage adjusted Bonferroni correction procedure is more than that from standard Bonferroni method. Table 3-4 shows that the number of tests declared as significant from our two-stage Bonferroni adjustment was larger than the numbers from the standard Bonferroni correction, but smaller than the numbers from the Holm-Bonferroni adjustment by number of independent tests. Compared with the dependent FDR method, our method appears to be more conservative. FDR is commonly used in studies involving large amounts of true alternatives, such as in micro-array data analysis of differentially expressed genes. In the simulated small level p values, all tests were declared significant by FDR. Furthermore, compared with the PCA and haplotype blocking to infer number of informative SNPs used by Johnson et al[65], our two-stage type 1 error adjustment would increase the specificity in hypothesis testing to avoid false positive variants.

In summary, our two-stage adjusted Bonferroni correction procedure provides a new, simple, easy way to control for type I error by incorporating informative SNPs. The correction method accounts for LD variation both within- and across-blocks, and therefore, it could increase the specificity in hypothesis testing of genetic association. Additionally, the two-stage adjustment method would also increase the sensitivity and power to discover causal genetic variants, compared with commonly used standard Bonferroni correction.

CHAPTER 4: APPLYING THE ADJUSTED TYPE I ERROR RATE IN SAMPLE SIZE ESTIMATION FOR GENETIC ASSOCIATION STUDIES

ABSTRACT

Background: In multi-locus genetic association studies, as more SNP markers are tested simultaneously, a larger sample size is needed to reduce false positive association and to increase the reliability of a study. However, a study with large sample size may not be cost-effective, leading to wasted time and resources. It is crucial, therefore, to find an effective sample size, which is the minimum number of samples that achieves adequate statistical power to detect the genetic association with a targeted phenotype. Because of the complicated pattern of linkage disequilibrium (LD) in humans, statistical power is influenced by the LD structure of the tested DNA region.

Objective and Methods: We incorporated linkage disequilibrium structure into estimating sample size for multi-locus genetic association studies. Haplotype blocking was first used to find the effective number of independent tests among multiple loci by race. We then applied the effective number of independent tests in the calculation of point-wise error rates using standard Bonferroni procedure and adjusted Bonferroni correction. We then used the point-wise error rates to estimate sample sizes. Finally, we applied our procedure to a case study to illustrate the procedure of using adjusted significant levels in sample size estimation for three racial groups.

Results: The case study was to estimate sample size in a multi-locus genetic association study to determine the associations between SNPs in the vitamin D binding protein gene, *GC*, and hypovitaminosis D. Compared with sample sizes estimated using a significance level from a standard Bonferroni type I error correction, the sample sizes required using the significance level from our haplotype blocking method were 14.94% and 28.50% lower in Han Chinese in Beijing, China and Utah residents with Northern and Western European ancestry populations, respectively. Since no haplotype blocks were inferred in the *GC* gene region in ASW, the haplotype block method resulted in the same sample size as in the standard Bonferroni method.

Conclusion and Implication: Population structure impacts the statistical power in an association test. Applying LD structure across the tested DNA region to adjust the alpha value for sample size estimation by race could reduce the sample size to reach enough statistical power to find the genetic association with the targeted phenotype, which would have cost-effective outcomes in these studies.

Keywords: Type I error; LD; Bonferroni type I error correction; Haplotype block; Point-wise error rate; Family-wise error rate.

INTRODUCTION

A common problem in bioinformatics is multi-locus or genome-wide association studies, which are powerful and widely-used studies to find genetic variants that impact a drug response or increase the risk of developing a particular disease. These studies are complex and must be planned carefully in order to maximize the probability of finding causal genetic variants that are associated with phenotypes. For these studies, design choices must balance sample size with budget constraints to optimize the power in detecting associations.

In Chapter 2, we addressed the problem of sample size estimation to achieve adequate power. We tested various combinations of parameters that are specific to genetic association studies and the influence of interactions among these parameters on sample size calculation. Type I error affects the number of individuals needed in a genetic association study, and we held the type I error at 0.05 because we were studying a single locus. For a single variant association test, the family-wise error rate is as the same as the point-wise error rate at a nominal level of α. However, for the multi-locus or genome-wide association study, the family-wise error rate must be controlled by adjusting the point-wise error rate for each genetic marker tested.

In Chapter 3, we proposed a two-stage adjusted Bonferroni correction procedure to control for inflated type I error rates in multiple testing among genetic markers. This procedure not only takes into account the linkage pattern among haplotype blocks to determine the number of independent tests that arise in the data but also within haplotype blocks to adjust the type I error among correlated genetic markers within blocks. Our two-stage adjustment method accounts for Linkage Disequilibrium (LD) variation within the haplotype blocks to control for block-wise type I error and overcomes the too conservative limitation of traditional Bonferroni correction to increase power. On the other hand, the two-stage correction method is less liberal to control for the type I error inflation compared with the haplotype blocking method only.

In this chapter, we focused on testing how sample size estimation can take advantage of this adjusted Bonferroni correction for type I error rate in multi-locus association studies. More subjects in a study leads to higher statistical power to detect the association of genetic variants with a targeted phenotype. However, a too large sample size could result in an economic burden and a great effort to collect sample. Therefore, finding an effective sample size, which is the minimum number of subjects to achieve enough statistical power, is an important step in study design. Here, we applied LD structure in sample size estimation by adjusting the significance level using the number of independent tests.

Linkage disequilibrium, the non-random association between alleles of different loci, is extremely important for the dissection of complex traits. Linkage disequilibrium patterns in the human genome are different across populations. The level and pattern of LD is influenced by many factors, e.g. genetic drift, admixture and inbreeding, which are population specific factors[86]. Additionally, statistical power is dependent on the frequency of the genotype or exposure being studied[87]. Genotype frequencies tend to vary by ethnicity, ancestry, or race. If race/ethnicity-specific estimates of genotype or exposure frequencies are not considered in sample size estimation, studies may have inadequate power or may be inefficient (i.e., using larger sample sizes than may be necessary). Therefore, to incorporate the different patterns of LD structure and genotype frequencies among populations into sample size estimation, we used a blocking method discussed in Chapter 3 to find the adjusted significance level for different populations, and then used the adjusted significance level to calculate sample sizes.

METHODS

Find the effective number of independent tests

To obtain the number of independent tests, we constructed haplotype blocks to define the underlying pair-wise linkage disequilibrium structure among genetic markers. The effective number of independent tests is the sum of the number of haplotype blocks and the number of singleton single nucleotide polymorphisms (SNP). There are three haplotype blocking algorithms for inferring haplotype blocks, including Gabriel's algorithm, the 4-gamete test, and the solid spine of LD measure. The three haplotype blocking algorithms group correlated SNPs across a DNA region into haplotype blocks. Our study results in Chapter 3 showed that the Gabriel algorithm generated the largest number of independent tests among the three haplotype blocking methods. Since the Gabriel algorithm is the most conservative, we used the method to find the effective number of independent tests for point-wise error calculation.

Calculate the point-wise error rate

We calculated point-wise error rates using the standard Bonferroni correction and the Bonferroni correction adjusted by the effective number of independent tests. Setting the nominal significance level at 0.05, we calculated the point-wise error rates by dividing the nominal significance level of 0.05 by the number of genetic markers for the standard Bonferroni and by the number of independent tests generated by Gabriel algorithm for the adjusted Bonferroni correction. For instance, if there are n SNPs and m independent tests, then the point-wise error rates are 0.05/n for the standard Bonferroni correction and 0.05/m for the adjusted Bonferroni correction.

Estimate the sample size

We assumed the simplest disease model in which a genetic marker is correlated with the causal variant. Logistic regression models were used to estimate sample sizes in accordance with a dichotomous outcome variable with genotype as the predictor variable. The genetic inheritance model is dominant, recessive, or additive. Assuming a null hypothesis of no significant genetic effect on the outcome variable, we performed a likelihood ratio test to compare the goodness of fit of two models, the null model (reduced model) and the alternative model (full model). The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. The likelihood ratio statistic (deviance) has a chi-square distribution with degrees of freedom, which is equal to the difference in the number of degrees of freedom between the full and reduced models. A sufficiently large deviance implies the logistic model is inappropriate. Inference concerning any regressor or subset can be computed by determining how much the presence of each regressor contributes to the reduction in deviance. Therefore, to test the hypothesis that the genetic factor is significantly associated with the disease, we are going to test the difference in variation explained between the full and reduced models by

$$\Lambda = \lambda(\beta_{0,} \beta_1) - \lambda(\beta_0) = 2(lnL(\hat{\beta}_0, \hat{\beta}_1) - \ln\ L(\hat{\beta}_0)),$$

where β_1 is coefficient of genetic factor.

Under the null hypothesis ($\beta_1 = 0$), Λ is asymptotically distributed as a chi-square random variable with one degree of freedom. When the null hypothesis is rejected, NΛ is the non-centrality parameter of the chi-squared distribution for a given sample size *N*. Therefore, sample size can be computed by

$$N = \frac{(z_{\alpha/2}+z_{\beta})^2}{\Lambda} \quad ,$$

where a two-sided alternative hypothesis is tested, α is the significance level, $1 - \beta$ is the desired power (set to 0.). To test a single locus, the significance level α is equal to the family-wise error rate or point-wise error rate. To test multiple loci, we should control the family-wise error rate under a specific level (e.g. 0.05), therefore, we need to replace the significance level α by the point-wise error rates calculated by standard and adjusted Bonferroni corrections. Under the null hypothesis (e.g. the genetic marker is not associated with the disease), Λ is asymptotically distributed as a chi-square random variable with one degree of freedom. When the null hypothesis is rejected (e.g. the genetic marker is associated with the disease), NΛ is the non-centrality parameter of the chi-squared distribution for a given sample size *N*. Since Λ is a function of the odds ratio, a smaller odds ratio requires a larger sample size. The total sample size is equal to 2N for the matched case-control study, and *N*+*KN* for an unmatched case-control study, where *K* = controls/cases.

The following includes a case study to illustrate how we apply the adjusted pointwise error rates to calculate sample size in multi-locus genetic association studies.

Case Study

The case study introduces the sample size estimation procedure in a multi-locus genetic association study to determine the associations between SNPs in vitamin D binding protein gene (*GC*) and hypovitaminosis D.

Introduction

One-half of healthy adults in developed countries might suffer from vitamin D insufficiency[88]. Vitamin D is crucial to maintain human health. Recent studies reported that vitamin D insufficiency has been linked to diabetes, cancer, and cardiovascular disease. The level of 25-OH D is the widely-accepted biomarker of vitamin D status. Determinants of circulating 25-hydroxyvitamin D (25-OH D) include sun exposure and dietary intake. However, only about a quarter of the inter-individual variability in 25-OH D is associated with the factors of season, geographic latitude, and vitamin D intake[89, 90]. Previous studies suggest that genetic determinants may also play a role in the variability of 25-OH D[91, 92] with estimates of heritability as high as 53%. Candidate gene studies with modest sample sizes and small numbers of variants and genome-wide association studies have been performed to examine the effect of specific vitamin D-pathway genes on the vitamin D plasma level[91-95].

The prevalence of vitamin D deficiency varies by race/ethnic group. Previous studies showed that vitamin D deficiency is more common among African Americans than among European Americans[93]. Previous genome-wide association studies in European populations identified vitamin D pathway gene single-nucleotide polymorphisms (SNPs) associated with serum vitamin D [25(OH)D] levels, but a few of these SNPs have been replicated in African Americans[93]. We calculated sample size for investigating the associations of vitamin D binding protein gene, *GC*, with hypovitaminosis D in asthmatic populations of Caucasians, African Americans and Asians, separately.

Methods

We calculated sample sizes based on a logistic regression model, in which the dichotomous outcome variable is the disease status if a patient had hypovitaminosis D or not. The sample sizes were estimated by three different racial groups: Caucasian, African

Americans and Asian. SNP genotype data in the vitamin D binding protein gene were downloaded for each racial group from HapMap[12] Genome Browser release #27 (Phase1, 2, &3-merged genotypes &frequencies). Haplotype block structure from these SNP genotype data for each of the racial groups were inferred by Haploview[12] software. Based on the haplotype structure in each of the three race groups, the number of independent tests was calculated. And then the point-wise error rate (alpha value) was calculated using the number of independent tests and family-wise error rate. Finally the sample sizes were estimated by genetic program Quanto[24, 25, 47, 48] with parameters as follows:

- The prevalence of the hypovitaminosis D in asthmatic population is 80%.
- A matched case-control design.
- Dominant inheritance model was assumed.
- Odds ratios for risk allele carriers compared to normal from 1.5 to 3.0.
- Desired power is 80%.
- The type I error rates are set at 0.05/ *m*, *m*=the number of candidate SNPs (standard Bonferroni type I error correction), 0.05/*x*, x=total number of haplotype blocks and singletons among the candidate SNPs, and 0.05 (assuming a single SNP) with a 2-sided alternate hypothesis.

RESULTS

Structure of haplotype block and pointwise error rate

Haplotype blocks were inferred using SNP genotype data from HapMap[12, 96], and the pointwise error rate was calculated by the number of haplotype blocks and singletons from the Gabriel method. SNP genotype data for the *GC* gene DNA region at chr4:72824381.. 72862352 were downloaded for each of the three racial groups. The racial groups were: African ancestry in Southwest USA (ASW) including 12 singletons and 11 trios from 40 families; Utah residents with Northern and Western European ancestry (CEU) from the CEPH collection including 27 trios from 20 families; Han Chinese in Beijing, China (CHB) including 45 singletons from 45 families. SNPs were selected if the p value for a Hardy Weinberg Equilibrium test was larger than 0.001, non-missing genotype percentage was at least 75%, maximum number of Mendelian errors was 1, and the minimum minor allele frequency was at least 0.001. The haplotype block structures for the three racial groups are shown below. The color schemes in LD plot are in accordance with the strength of LD between SNPs. The *D'* value and the log odds (LOD) were used to measure LD between loci. The closer the *D'* value is to zero, the greater the amount of historical recombination between the two loci. LOD>2 indicates significant LD, therefore, bright red indicates a strong LD with LOD ≥2 and *D'*=1, while white indicates a weak LD with LOD<2 and *D'*<1. Between strong and weak LDs, blue represents LOD<2 and D'=1 while shades of pink/red shows LOD≥2 and D'<1.

Haplotype Structure in ASW population

Haplotype blocks were inferred for the *GC* gene region for the ASW population. There were a total of 36 SNPs in the *GC* gene region. The minor allele frequencies among the 36 SNPs ranged from 0.009-0.462. The linkage disequilibrium plot in Figure 4-1 showed that no haplotype blocks were inferred among these SNPs. Therefore, the number of independent tests was equal to the number of SNPs, which was 36. The point-wise error rate under the family-wise error rate 0.05 was calculated by 0.05/36=0.001, which was equal to the error rate calculated by the standard Bonferroni correction.

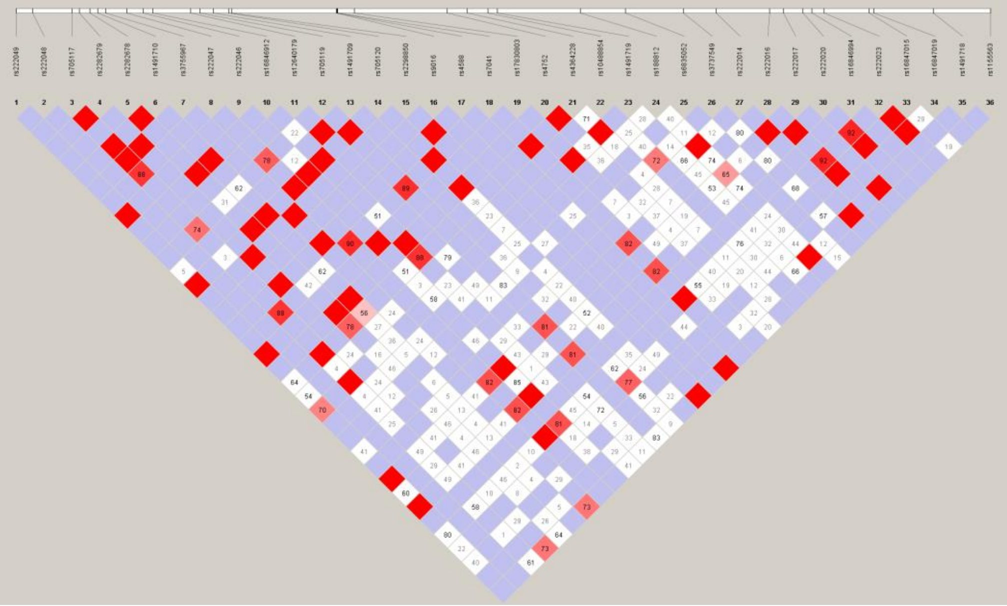


Figure 4-1. Linkage disequilibrium plot for SNPs in vitamin D binding protein gene (*GC*) region at chr4:72824381..72862352 among African ancestry in Southwest USA

Haplotype Structure in CEU population

Haplotype blocks were inferred in the *GC* gene region for the CEU population. There were a total of 36 SNPs selected to infer the haplotype block structure. The minor allele frequencies among the 36 SNPs ranged from 0.009-0.462. The following LD plot (Figure 4-2) showed that two haplotype blocks and four singletons were inferred for these SNPs. Therefore, the number of independent tests was equal to the sum of the number of haplotype blocks (2) and the number of singletons (4), which was 6. The point-wise error rate under the family-wise error rate 0.05 was calculated by 0.05/6=0.008. The error rate calculated by standard Bonferroni correction was equal to 0.05/36=0.001.

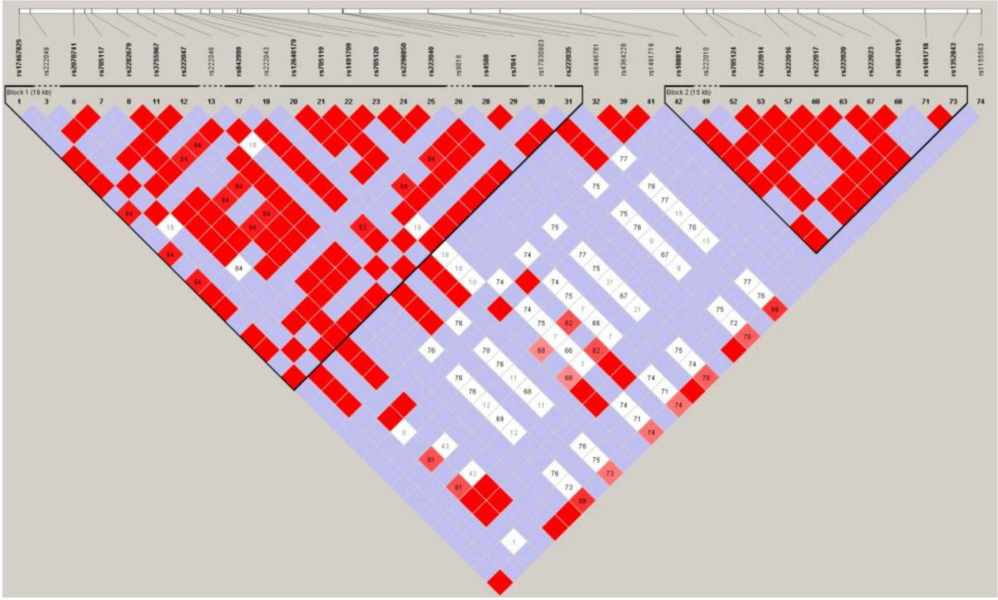


Figure 4-2. Linkage disequilibrium plot for SNPs in vitamin D binding protein gene (*GC*) DNA region at chr4:72824381..72862352 among Utah residents with Northern and Western European ancestry

Haplotype Structure in CHB population

Haplotype blocks were inferred in the *GC* gene region of the CHB population. A total of 45 SNPs at the *GC* gene region were selected to infer the haplotype block structure. The minor allele frequencies among the 45 SNPs ranged from 0.011-0.465. The following LD plot (Figure 4-3) showed that two haplotype blocks and four singletons were inferred among these SNPs. Therefore, the number of independent tests was equal to the sum of the number of haplotype blocks (5) and the number of singletons (14), which was 19. The point-wise error rate under the family-wise error rate of 0.05 was calculated by 0.05/19=0.003. The error rate calculated by standard Bonferroni correction was equal to 0.05/45=0.001.

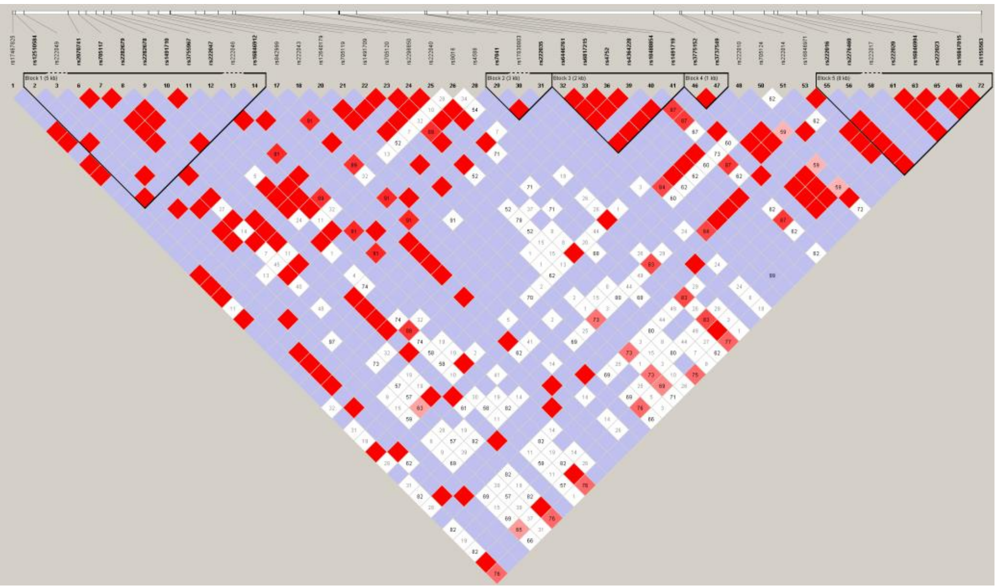


Figure 4-3. Linkage disequilibrium plot for SNPs in vitamin D binding protein gene (*GC*) gene DNA region at chr4:72824381..72862352 among Han Chinese in Beijing

Sample size estimation

Table 4-1 shows the estimated sample sizes for testing the main effects of SNPs in the *GC* gene region with hypovitaminosis D. Because the pattern of linkage disequilibrium in a genome is impacted by population structure, there are differences in significance (α) levels among the three racial groups and hence different sample sizes. When the haplotype blocking method for multi-locus tests was used to calculate the alpha level, the estimated sample sizes from the adjusted alpha value were between the sample sizes from the standard Bonferroni correction and the unadjusted single locus calculation. The sample sizes required using the significant level from our haplotype blocking method were 14.94% and 28.50% lower than for the standard Bonferroni correction in CHB and CEU populations, respectively. Since no haplotype blocks were inferred at the *GC* gene region in ASW, the haplotype block method resulted in the same sample size as in the standard Bonferroni method. The unadjusted one locus method in sample size estimation resulted in the smallest sample sizes among the three methods and required on average 54% smaller sample sizes compared to the standard

Bonferroni correction, however, we would lack of power to test the multiple SNPs with the targeted phenotype, simultaneously.

Table 4-1. Sample size required to test genetic association with hypovitaminosis D

Minor Allele Frequency	Odds Ratio	Number of Case-Control Pairs (percentage decrease[a])			
		Standard Bonferroni (ASW, CEU, CHB)[b]	Haplotype Block Adjustment		One Locus Method (ASW, CEU, CHB)
		α=0.001	α=0.003 (CHB)	α=0.008 (CEU)	α=0.05
0.10	1.5	1475	1253 (15.05)	1054(28.54)	678(54.03)
0.10	2.0	546	464 (15.02)	390(28.57)	251(54.03)
0.10	2.5	334	284 (14.97)	239(28.44)	154(53.89)
0.10	3.0	247	210 (15.00)	177(28.34)	114(53.85)
0.15	1.5	1109	942 (15.06)	793(28.49)	510(54.01)
0.15	2.0	404	344 (14.85)	289(28.47)	186(53.96)
0.15	2.5	245	208 (15.10)	175(28.57)	113(53.88)
0.15	3.0	180	153 (15.00)	129(28.33)	83(53.89)
0.20	1.5	946	804 (15.01)	676(28.54)	435(54.02)
0.20	2.0	340	289 (15.00)	243(28.53)	156(54.12)
0.20	2.5	204	174 (14.71)	146(28.43)	94(53.92)
0.20	3.0	149	126 (15.44)	106(28.86)	68(54.36)
0.25	1.5	869	739 (14.96)	621(28.54)	399(54.09)
0.25	2.0	309	262 (15.21)	221(28.48)	142(54.06)
0.25	2.5	183	156 (14.75)	131 (28.41)	84(54.10)
0.25	3.0	132	112 (15.15)	95 (28.03)	61(53.79)
0.30	1.5	841	715 (14.98)	601 (28.54)	387(53.98)
0.30	2.0	295	251 (14.92)	211 (28.47)	136(53.90)
0.30	2.5	174	148 (14.94)	124 (28.74)	80(54.02)
0.30	3.0	124	106 (14.52)	89 (28.23)	57(54.03)
0.35	1.5	847	720 (14.99)	606 (28.45)	389(54.07)
0.35	2.0	294	250 (14.97)	210 (28.57)	135(54.08)
0.35	2.5	171	146 (14.62)	122 (28.66)	79(53.80)
0.35	3.0	122	104 (14.75)	87 (28.69)	56(54.10)
0.40	1.5	884	751 (15.05)	632 (28.51)	406(54.07)
0.40	2.0	303	258 (14.85)	217 (28.38)	139(54.13)
0.40	2.5	175	149 (14.86)	125 (28.57)	81(53.71)
0.40	3.0	124	105 (15.32)	88 (29.03)	57(54.03)
0.45	1.5	951	809 (14.93)	680 (28.50)	437(54.05)
0.45	2.0	323	275 (14.86)	231 (28.48)	149(53.87)
0.45	2.5	185	158 (14.59)	133 (28.11)	85(54.05)
0.45	3.0	130	111 (14.62)	93 (28.46)	60(53.85)

[a] Percentage decrease = (the number from haplotype block method or one locus method - the number from standard Bonferroni method)/the number from standard Bonferroni method) ×100

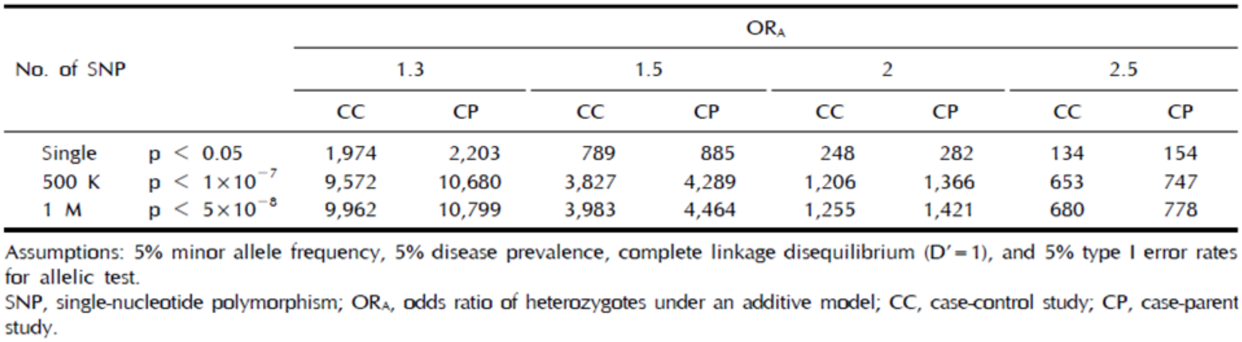
[b] ASW: African ancestry in Southwest USA including 12 singletons and 11 trios from 40 families, CEU: Utah residents with Northern and Western European ancestry from the CEPH collection including 0 singletons and 27 trios from 20 families, CHB: Han Chinese in Beijing, China,2.The prevalence of the hypovitaminosis D in asthmatic population is 80%; dominant inheritance model was assumed; desired power is 80%; The type I error rates were set at 0.05/ *m* the number of candidate SNPs (standard Bonferroni type I error correction), 0.05/*x*, *x*=total number of haplotype blocks and singletons among the candidate SNPs, and 0.05 (assuming a single SNP) with a 2-sided alternate hypothesis; A matched case-control design.

DISCUSSION

Genetic association studies commonly test multiple genetic variants. If we estimate the sample size based on a single variant for a multi-locus genetic association study, the estimated sample size might result in insufficient statistical power to detect true evidence for an association, leading to high false negative rates and reducing the reliability of the study. Therefore, a large number of SNP markers requires a large sample size to reduce false positive association due to testing multiple hypotheses.

For instance, E.P. Hong et al.[28] reported that as the number of genetic markers tested increased, the Bonferroni p-value that was specific to the number of SNP makers tested decreased, and the required sample sizes increased. In their study, they set up the family-wise error rate at 0.05. Therefore, for a single SNP marker p=0.05, for 500k SNP markers $p=1\times10^{\sim7}$, and for 1M SNP markers $p=5\times10^{-8}$. They calculated the sample sizes with 80% power for increasing number of SNP markers in case-control and case-parent studies as shown in the following table[28].

Table 4-2. Results from Hong's study[28] in sample sizes with 80% power by increasing number of SNP markers in case-control and case-parent studies



No. of SNP		OR_A							
		1.3		1.5		2		2.5	
		CC	CP	CC	CP	CC	CP	CC	CP
Single	$p < 0.05$	1,974	2,203	789	885	248	282	134	154
500 K	$p < 1\times10^{-7}$	9,572	10,680	3,827	4,289	1,206	1,366	653	747
1 M	$p < 5\times10^{-8}$	9,962	10,799	3,983	4,464	1,255	1,421	680	778

Assumptions: 5% minor allele frequency, 5% disease prevalence, complete linkage disequilibrium (D′ = 1), and 5% type I error rates for allelic test.

SNP, single-nucleotide polymorphism; OR_A, odds ratio of heterozygotes under an additive model; CC, case-control study; CP, case-parent study.

From Table 4-2, we can see that as the point-wise error decreased from 0.05 to 5×10^{-8}, the required sample sizes increased. They used standard Bonferroni correction to get the point-wise error rates to account for the effects for multiple genetic markers in sample size estimation and to control for the family-wise error rate at the specific level.

However, the adjusted alpha level from the standard Bonferroni correction can generate much larger sample sizes compared with an un-adjusted alpha value. Though a large sample size improves the ability to detect genetic association with a disease, it may not be cost-effective. To find the minimum number of samples to achieve adequate statistical power, we investigated the pattern of LD structure across the tested DNA region among the racial groups by the haplotype blocking method (BCM). BCM applied probability theory to identify statistical associations among a set of SNPs across a specific DNA region, and grouped these SNPs into related subsets of SNPs. The number of subsets of SNPs and singletons of SNPs formed the number of independent tests, which accounts for historical biological relationships among these genetic markers in the point-wise error calculation. Compared with the standard Bonferroni type I error correction calculation for the alpha level, our method generated a more precise pointwise error rate to get a more appropriate and effective sample size with enough power and reasonable cost to realistically collect samples.

Genome-wide association or linkage methods are dependent on the linkage disequilibrium among genetic variants on a chromosome. Differences in the pattern of linkage disequilibrium by race have been reported[3], this could affect the success of gene discovery efforts. A previous study[45] also reported that ignoring ethnicity in molecular epidemiologic studies can lead to some distortion of estimates of association. Therefore, all studies should carefully consider the potential for confounding by ethnicity, ancestry, or race, and respond with appropriate study design or analytic methods. As such, the statistical analyses in genetic association studies are often performed by race due to the different population structure. Therefore, the number of subjects required for one race group could be different from another group. In our case study, ASW required a larger sample size than the other two race groups to detect genetic association in the same DNA region with hypovitaminosis D since there was weaker LD among the genetic variants in the ASW population compared with that in CEU and CHB populations. Using one standard method for the different racial groups to calculate sample sizes, might result in a lack of power to find the causal genetic variants in the ASW population, but overestimate sample sizes needed to test the genetic association in other two populations.

The haplotype block by race accounts for the historical biological relationships among genetic makers at a specific DNA region to group these genetic variants with high linkage disequilibrium into subsets. The subset of SNPs not only can be used to find a tag SNP, but also the number of independent tests. A tag SNP is a representative single nucleotide polymorphism in a region of the genome with high linkage disequilibrium that represents a group of SNP**s** called a haplotype. It is possible to identify genetic variation and association to phenotypes without genotyping every SNP in a chromosomal region. We used the number of independent tests or tag SNPs across the tested DNA region to adjust the alpha value for sample size estimation by race. This could reduce the number of subjects required, but have enough statistical power, to find evidence of genetic association with the targeted phenotype, leading to cost-effective outcomes in these studies.

REFERENCES

1. Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med*. 2001;40:346-358
2. Lewis CM, Knight J. Introduction to genetic association studies. *Cold Spring Harb Protoc*. 2012:297-306
3. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med*. 2002;4:45-61
4. Srivastava K, Srivastava A. Comprehensive review of genetic association studies and meta-analyses on mirna polymorphisms and cancer risk. *PLoS One*. 2012;7:e50966
5. Fu L, Jin L, Yan L, Shi J, Wang H, Zhou B, et al. Comprehensive review of genetic association studies and meta-analysis on mirna polymorphisms and rheumatoid arthritis and systemic lupus erythematosus susceptibility. *Hum Immunol*. 2016;77:1-6
6. Ramenskii VE, Siuniaev Sh R. [computational analysis of human genome polymorphism]. *Mol Biol (Mosk)*. 2009;43:286-294
7. Muehlenbein MP. *Human evolutionary biology*. Cambridge University Press; 2010.
8. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. *Nat Protoc*. 2011;6:121-133
9. Gelman A. P values and statistical practice. *Epidemiology*. 2013;24:69-72
10. Cowan G. *Statistical data analysis*. Oxford University Press Inc.; 1998.
11. Jerome L. Myers ADW, Robert F. Lorch Jr. Research design and statistical analysis. 2010
12. Nannipieri M, Posadas R, Williams K, Politi E, Gonzales-Villalpando C, Stern MP, et al. Association between polymorphisms of the atrial natriuretic peptide gene and proteinuria: A population-based study. *Diabetologia*. 2003;46:429-432
13. Nguyen D, Xu T. The expanding role of mouse genetics for understanding human biology and disease. *Dis Model Mech*. 2008;1:56-66
14. Halapi E, Bjornsdottir US. Overview on the current status of asthma genetics. *Clin Respir J*. 2009;3:2-7
15. Perreault T GJ. Role of atrial natriuretic factor in lung physiology and pathology. *Am J Respir Crit Care Med*. 1995;151:226-242
16. Sakamoto M NK, Morii N, Sugawara A, Yamada T, Itoh H, et al. . The lung as a possible target organ for atrial natriuretic polypeptide secreted from the heart. *Biochem Biophys Res Commun* 1986;135 515-520

17. Buchanan GR, DeBaun MR, Quinn CT, Steinberg MH. Sickle cell disease. *Hematology Am Soc Hematol Educ Program*. 2004:35-47

18. Platt OS, Brambilla DJ, Rosse WF, Milner PF, Castro O, Steinberg MH, et al. Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med*. 1994;330:1639-1644

19. Stuart MJ, Setty BN. Sickle cell acute chest syndrome: Pathogenesis and rationale for treatment. *Blood*. 1999;94:1555-1560

20. Sharan K, Surrey S, Ballas S, Borowski M, Devoto M, Wang KF, et al. Association of t-786c enos gene polymorphism with increased susceptibility to acute chest syndrome in females with sickle cell disease. *Br J Haematol*. 2004;124:240-243

21. Consortium IWP. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med*. 2009;360:753-764

22. Cooper GMJ, J. A. Langaee, T. Y. Feng, H. Stanaway, I. B. Schwarz, U. I. Ritchie, M. D. Stein, C. M. Roden, D. M. Smith, J. D. Veenstra, D. L. Rettie, A. E. Rieder, M. J. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood*. 2008;112:1022-1027

23. Tantisira K, Weiss S. The pharmacogenetics of asthma treatment. *Curr Allergy Asthma Rep*. 2009;9:10-17

24. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med*. 2002;21:35-50

25. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol*. 2002;155:478-484

26. Gauderman WJ. Candidate gene association analysis for a quantitative trait, using parent-offspring trios. *Genet Epidemiol*. 2003;25:327-338

27. Pfeiffer RM, Gail MH. Sample size calculations for population- and family-based case-control association studies on marker genotypes. *Genet Epidemiol*. 2003;25:136-148

28. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. *Genomics Inform*. 2012;10:117-122

29. S.C.Chow JS, H. Wang. Sample size calculations in clinical research. 2008

30. Goodman M, Dana Flanders W. Study design options in evaluating gene-environment interactions: Practical considerations for a planned case-control study of pediatric leukemia. *Pediatr Blood Cancer*. 2007;48:375-379

31. Myers RH. *Classical and modern regression with application*. Duxbury Press, An Imprint of Wadsworth Publishing Company Adivision Wadsworth, Inc.; 1990.

32. Hizawa N. Pharmacogenetics of beta2-agonists. *Allergol Int*. 2011;60:239-246

33. Balding DJ. Likelihood-based inference for genetic correlation coefficients. *Theor Popul Biol*. 2003;63:221-230

34. Sultan SN, C. Shah, P. Feng, H**.** Provenzale, D. Beyth, R. . Marital status is an important predictor of adherence to colonoscopy among high risk veterans. . *The 27th VA Health Services Research and Development Service (HSR&D) National Meeting, February 11-13, 2009 in Baltimore, MD; Society of General Internal Medicine 2009.*

35. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nat Genet*. 2001;29:229-232

36. Documentation R. Http://www.R-project.Org/other-docs.Html.

37. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: Models and data. *Am J Hum Genet*. 2001;69:1-14

38. Moskvina V, O'Donovan MC. Detailed analysis of the relative power of direct and indirect association studies and the implications for their interpretation. *Hum Hered*. 2007;64:63-73

39. Koppelman GH, te Meerman GJ, Postma DS. Genetic testing for asthma. *Eur Respir J*. 2008;32:775-782

40. Ioannidis JP, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am J Epidemiol*. 2006;164:609-614

41. Khlat M, Cazes MH, Genin E, Guiguet M. Robustness of case-control studies of genetic factors to population stratification: Magnitude of bias and type i error. *Cancer Epidemiol Biomarkers Prev*. 2004;13:1660-1664

42. Kittles RA, Chen W, Panguluri RK, Ahaghotu C, Jackson A, Adebamowo CA, et al. Cyp3a4-v and prostate cancer in african americans: Causal or confounding association because of population stratification? *Hum Genet*. 2002;110:553-560

43. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet*. 2004;36:512-517

44. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias. *J Natl Cancer Inst*. 2000;92:1151-1158

45. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*. 1999;65:220-228

46. Millikan RC. Re: Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias. *J Natl Cancer Inst*. 2001;93:156-158

47. Gauderman WJ. Air pollution and children--an unhealthy mix. *N Engl J Med*. 2006;355:78-79

48. Gauderman WJ, Gilliland GF, Vora H, Avol E, Stram D, McConnell R, et al. Association between air pollution and lung function growth in southern california children: Results from a second cohort. *Am J Respir Crit Care Med*. 2002;166:76-84

49. Henian Chen PCSC. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics - Simulation and Computation* 06 Apr 2010;39:860-864

50. Cohen J. *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates; 1988.

51. Green PJaS, B.W. Nonparametric regression and generalized linear models. . *London, : Chapman & Hall*. 1994

52. Harder RLaD, R.N. Interpolation using surface splines. *Journal of Aircraft* 1972;9, 189–191

53. Meinguet J. Multivariate interpolation at arbitrary points made simple. *Journal of Applied Mathematics and Physics* 1979;30, 292–304

54. Weiss ST. Association studies in asthma genetics. *American Journal of Respiratory and Critical Care Medicine*. 2001;164

55. Lima JJM, S. Feng, H. Lockey, R. Jena, P. K. Castro, M. Irvin, C. Johnson, J. A. Wang, J. Sylvester, J. E. A polymorphism in the nppa gene associates with asthma. *Clin Exp Allergy*. 2008;38:1117-1123

56. Purcell S, Cherny SS, Sham PC. Genetic power calculator: Design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*. 2003;19:149-150

57. Oner Ozgon GL, T. Y. Feng, H. Buyru, N. Ulutin, T. Hatemi, A. C. Siva, A. Saip, S. Johnson, J. A. Vkorc1 and cyp2c9 polymorphisms are associated with warfarin dose requirements in turkish patients. *Eur J Clin Pharmacol*. 2008;64:889-894

58. http://en.wikipedia.org/wiki/Type_I_and_type_II_errors.

59. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol*. 2008;32:361-369

60. Nicodemus KK, Liu W, Chase GA, Tsai YY, Fallin MD. Comparison of type i error for multiple test corrections in large single-nucleotide polymorphism studies using principal components versus haplotype blocking algorithms. *BMC Genet*. 2005;6 Suppl 1:S78

61. Duggal P, Gillanders EM, Holmes TN, Bailey-Wilson JE. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics*. 2008;9:516

62. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res*. 2001;125:279-284

63. SIMES RJ. An improved bonferroni procedure for multiple tests of significance. *Biometrika*. 1986;73:751-754

64. Cooper GM, Johnson JA, Langaee TY, Feng H, Stanaway IB, Schwarz UI, et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood*. 2008;112:1022-1027

65. Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, et al. Accounting for multiple comparisons in a genome-wide association study (gwas). *BMC Genomics*. 2010;11:724

66. A haplotype map of the human genome. *Nature*. 2005;437:1299-1320

67. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*. 1936

68. Holm S. A simple sequentially rejective bonferroni test procedure. *Scandinavian Journal of Statistics*. 1979;6, 65–70

69. So HC, Sham PC. Multiple testing and power calculations in genetic association studies. *Cold Spring Harb Protoc*. 2011;2011:pdb top95

70. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*. 2005;21:3017-3024

71. Benjamini YaH. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. 1995;B:289–300

72. Benjamini YaY, D. . The control of the false discovery rate in multiple testing under dependency,. *Annals of Statistics*. 2001;29, 1165–1188

73. Benjamini YaY, D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*. 2001;29:1165-1188

74. Dbsnp- the single nucleotide polymorphism dtabase.

75. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. *Science*. 2002;296:2225-2229

76. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. *Am J Hum Genet*. 2002;71:1227-1234

77. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: Analysis and visualization of ld and haplotype maps. *Bioinformatics*. 2005;21:263-265

78. Halperin E, Eskin E. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*. 2004;20:1842-1849

79. Cookson WO. Asthma genetics. *Chest*. 2002;121:7S-13S

80. Palmer LJ, Cookson WO. Using single nucleotide polymorphisms as a means to understanding the pathophysiology of asthma. *Respir Res*. 2001;2:102-112

81. Haagerup A, Bjerke T, Schiotz PO, Binderup HG, Dahl R, Kruse TA. Asthma and atopy - a total genome scan for susceptibility genes. *Allergy*. 2002;57:680-686

82. Malerba G, Pignatti PF. A review of asthma genetics: Gene expression studies and recent candidates. *J Appl Genet*. 2005;46:93-104

83. Park HS, Kim SH, Park CS. The role of novel genes in modifying airway responses in asthma. *Curr Allergy Asthma Rep*. 2006;6:112-116

84. Bosse Y, Hudson TJ. Toward a comprehensive set of asthma susceptibility genes. *Annu Rev Med*. 2007;58:171-184

85. Holloway JW, Koppelman GH. Identifying novel genes contributing to asthma pathogenesis. *Curr Opin Allergy Clin Immunol*. 2007;7:69-74

86. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. *Nature*. 2001;411:199-204

87. Rebbeck TR, Sankar P. Ethnicity, ancestry, and race in molecular epidemiologic research. *Cancer Epidemiol Biomarkers Prev*. 2005;14:2467-2471

88. Holick MF. Vitamin d deficiency. *N Engl J Med*. 2007;357:266-281

89. Livshits G, Karasik D, Seibel MJ. Statistical genetic analysis of plasma levels of vitamin d: Familial study. *Ann Hum Genet*. 1999;63:429-439

90. Shea MK, Benjamin EJ, Dupuis J, Massaro JM, Jacques PF, D'Agostino RB, Sr., et al. Genetic and non-genetic correlates of vitamins k and d. *Eur J Clin Nutr*. 2009;63:458-464

91. Sinotte M, Diorio C, Berube S, Pollak M, Brisson J. Genetic polymorphisms of the vitamin d binding protein and plasma concentrations of 25-hydroxyvitamin d in premenopausal women. *Am J Clin Nutr*. 2009;89:634-640

92. Lauridsen AL, Vestergaard P, Hermann AP, Brot C, Heickendorff L, Mosekilde L, et al. Plasma concentrations of 25-hydroxy-vitamin d and 1,25-dihydroxy-vitamin d are related to the phenotype of gc (vitamin d-binding protein): A cross-sectional study on 595 early postmenopausal women. *Calcif Tissue Int*. 2005;77:15-22

93. Wang TJ, Zhang F, Richards JB, Kestenbaum B, van Meurs JB, Berry D, et al. Common genetic determinants of vitamin d insufficiency: A genome-wide association study. *Lancet*. 2010;376:180-188

94. Engelman CD, Fingerlin TE, Langefeld CD, Hicks PJ, Rich SS, Wagenknecht LE, et al. Genetic and environmental determinants of 25-hydroxyvitamin d and 1,25-dihydroxyvitamin d levels in hispanic and african americans. *J Clin Endocrinol Metab*. 2008;93:3381-3388

95. Ramos-Lopez E, Bruck P, Jansen T, Herwig J, Badenhoop K. Cyp2r1 (vitamin d 25-hydroxylase) gene is associated with susceptibility to type 1 diabetes and vitamin d levels in germans. *Diabetes Metab Res Rev*. 2007;23:631-636

96. Hapmap-international hapmap project.