Abstract

MENG, ZHAOLING. Statistical Topics in Disease Gene Mapping (Under the direction of DRS. BRUCE S. WEIR AND MARGARET G. EHM)

Efforts in disease gene mapping have achieved a great deal of success in mendelain diseases, but made slower progress in common disease studies because of their complexity. The rapid development of genetics and molecular technologies provides an immense amount of DNA data; developing powerful and efficient statistical methodologies is under high demand. This dissertation explored some aspects of the problem. The power of two genome-wide disease gene mapping strategies is investigated. One applies linkage analysis and then linkage disequilibrium (LD) tests to markers within linked regions. The other looks for LD with disease using all markers. The results showed that the genome-wide association based tests are much more likely to identify genes. Genotyping closely spaced Single Nucleotide Polymorphisms (SNPs) frequently yields highly correlated data due to extensive LD, and gives association studies unnecessary and unaffordable burden when these markers don't yield significantly different information. Two procedures are developed to select an optimum subset of SNPs that could be efficiently genotyped on larger numbers of samples while retaining most of the information based on genotypes of a large initial set of SNPs on a small number of samples. One utilizes a spectral decomposition method based on matrices of pair-wise LD, and the other extends David Clayton's htSNP selection method. Properties of the procedures are studied; minimum sample sizes needed for achieving consistent results are

recommended; the procedures are evaluated using experimental data. Studying gene-treatment interaction is a long desired problem. When the genetic variation that is being tested is not specific functional sites but randomly selected polymorphisms, a source of randomness is introduced. A mixed effect model is developed to fit fixed treatment effects, random haplotypic effects, and random gene-treatment interactions in this scenario; likelihood ratio tests are applied for testing the random effects. Our simulation results showed that the mixed effect model is valid and generally behaves better than the fixed haplotypic effects model in the exploratory phase of a study.

Statistical Topics in Disease Gene Mapping

By
Zhaoling Meng

A dissertation submitted to the Graduate Faculty of
North Carolina State University
In partial fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Bioinformatics

Raleigh

2003

APROVED BY:

_____	_____
_____	_____
_____	_____
Co-chair of Advisory Committee	Co-chair of Advisory Committee

To my husband, my parents and my parents-in-law

Biography

Zhaoling Meng was born in Hefei, Anhui Province, China on June 26, 1975. She finished her secondary education at Hefei No. 1 Middle School in Hefei, Anhui, China in 1993.

From August 1993 to July 1998, Zhaoling attended the University of Science and Technology of China (USTC), Hefei, Anhui Province, China, where she received a Bachelor of Science degree in Biological Sciences.

Zhaoling entered the University of Toledo at Toledo, Ohio in August 1998 and got her Master of Science degree in Statistics in May 2000. During her stay at Toledo, she worked as a teaching assistant and later an instructor in the Department of Mathematics.

Since May 2000, Zhaoling has been studying for her Ph.D. in the Bioinformatics Research Center (BRC), North Carolina State University, and working as a graduate industrial trainee (GIT) at GlaxoSmithKline (GSK, the then GlaxoWellcome before the merge) in the program of Population Genetics, providing support in statistical genetics and programming.

Acknowledgements

First, I would like to humbly acknowledge my dissertation advisors, Drs. Bruce S. Weir and Margaret G. Ehm, and thank them for their advice, support and patience. Without their help and guidance, this dissertation would never be completed. In addition, I would like to extend my profound gratitude to Drs. Zhao-Bang Zeng, Greg Gibson, and Russ Wolfinger for their valuable advice on the dissertation. Also, I am grateful to Dr. Jonathan Allan for his service on the dissertation committee. Special thanks go to Dr. Dmitri Zaykin for his mentoring and collaboration in the research work at GSK. I also thank Drs. Dennis Boos, John Monahan and Daowen Zhang at the statistics department for their lectures and help. Thanks GlaxoSmithKline for supporting my internship. I am particularly grateful to the entire population genetics group at GSK for their help and friendship in the past three years. I would like to thank the entire Bioinformatics Research Center for support, with special thanks to Amy, Debbie, Juliebeth and Lisa for their smiles and all the paperwork. Thank my parents and parents-in-law for their understanding and encouraging support in the past years. Finally my most heartfelt thanks go to my husband, Xiaohui Luo, for his love and support throughout the years.

Contents

List of Tables **Page**

List of Figures **Page**

Chapter One

Introduction

1.1 Complex trait gene mapping

Studying genetic variation, especially mapping human drug response and disease susceptibility genes, has drawn increasing attention since the near-completion of human genome sequencing (Venter et al. 2001; Lander et al. 2001). With a lot of fruitful results accomplished since 1913 (Sturtevant 1913), mapping genetic traits remains a hard task. One main reason is that mapping genes is always like fishing in a sea of large size genomes, such as the human genome consisting of 3×10^9 base pairs. Another reason is the complexity of genetic traits such as incomplete penetrance, genetic heterogeneity, and polygenic inheritance. Due to the development of technologies and methodologies, the detection can be conducted based on the relations between inheritance patterns of a trait and chromosome components located by genetic markers, instead of the knowledge of the gene functions (Nakamura et al. 1987; Lander and Schork 1994). Most frequently used methods include linkage analysis (including model based and allele sharing methods), association studies, and experimental crosses.

Linkage analysis methods rely on the assumption that the chromosome region closely linked to the disease mutation allele tends to be conserved in pedigrees and leads to a certain inheritance pattern of this chromosome piece among affected individuals. With current genotyping techniques, linkage analysis is widely applied to genome-wide gene mapping based on several hundred markers spread across the genome. It has been utilized for a relatively long time and quite successful in mapping Mendelian diseases (Kerem et al. 1989) and some complex disease such as Alzheimer's disease and psoriasis (Pericak-Vance et al. 1991; Tomfohrde et al. 1994). Model-based linkage analysis methods are believed to be more powerful if the model parameters, such as penetrance functions or the underlying genetic models, are correctly specified. But the

results can be misleading if specified parameters do not mimic the reality correctly. On the other hand, allele-sharing methods are usually non-parametric, more robust, but presumably less powerful. The success of the linkage analysis in mapping complex human traits is still limited. Multiple genes with intermediate or small effects are believed to be the genetic basis of the complex traits. Based on their calculations, Risch and Merikangas (1996) predicted that linkage analysis would require unrealistic large sample sizes to obtain the statistical power required to detect relatively medium or small genetic effects comparing to the effects of some Mendelian diseases, and therefore might not be suitable for mapping complex traits. Several multi-point linkage analysis methods (Kruglyak et al. 1996; O'Connell 2001) were proposed and presumed to be more powerful than single-point methods (Penrose 1953; Elston and Stewart 1971). Obtaining large pedigrees is also considered crucial. However, their further achievements are still under inspection. Another limitation of linkage analysis is that the size of the detected region usually extends over more than 1cM (approximately 1000kb in human genome), which might be too large to pinpoint the targeted trait gene. Narrowing down a linkage region depends on the number of recombinations in the pedigree, which, in turn, depends on the number of meioses, pedigree structure and the sample sizes.

Association analysis methods are more "population based' in a sense that they try to locate the trait locus of interest by detecting the differences in marker allele frequencies between cases and matched controls from a population. The marker (markers) showing a significant difference is assumed to be either the trait locus or in linkage disequilibrium (LD) with it. The length of the detected region in an association study depends on a sample representing the recombination patterns and the evolution history of the population under study as whole, and therefore usually

much smaller than that from linkage analyses, which depends the patterns within families. It might vary in different chromosome regions from 1 kb to several hundred kb and from population to population. Relying on the inheritance pattern of a whole population, association studies are believed to have more power in mapping complex trait genes with small or intermediate effects (Jorde 1995). Therefore, they play more and more important roles in complex trait mapping (Ryder et al. 1979; Martin et al. 2000). Currently, most association studies are conducted on either candidate genes or pre-identified linkage regions. On the other hand, genome-wide association studies rely on relatively short distances of LD, and hence are constrained by the required high marker density, corresponding high genotyping costs and lack proper analysis methods. However, the fast development of high throughput single nucleotide polymorphism (SNP) genotyping techniques (Prince and Brookes 2001; Cutler et al. 2001) and statistical analysis methods are expected to be able to address these problems in the near future.

Utilizing unrelated population case-control samples is a major advantage of the association studies since the samples are relatively easy to obtain, but also a major limitation. Spurious associations could be caused by population stratification or recent population admixture (Weiss 1993), and by any markers confounding with trait locus of interest. Furthermore, the effects might be presented even after a careful matching of cases and controls. Two approaches were proposed to solve this problem. One relies on within family controls (Spielman et al 1993; Knapp et al 1993). Within family controls also provide good matches for the environments, although the concern of losing power due to over-matches of siblings was also raised (Risch and Teng 1998). The other approach is to develop methods either correcting or detecting stratification (Devlin and Roeder 1999; Pritchard and Rosenberg 1999; Pritchard et al. 2000).

Although the possible spurious association is the most frequently raised problem of the association studies, its effect on false positive findings is still unclear.

Experimental crosses have been also widely applied to genetic trait mapping and have already made great contributions to the study of human complex diseases like diabetes and obesity (Forsell et al. 2000; Hamilton-Williams et al. 2001). Although their applicability is limited in studying certain human genetic traits, experimental crosses are viewed to be powerful and even the "limit-breaking" tool in future genetic variation studies. These methods have the advantage of relying on the relative genetic homogeneity of animal or plant models and the ability to study a relatively large number of animal or plant generations at a time. Therefore, they can be employed to solve the genetic heterogeneity problem in the complex traits.

Replicating positive findings is essential in proving that the findings are true positives in any study. Contradicting reports about the same regions or genes are often seen in genetic trait mapping. One possible reason is false positives in some studies. Another possibility is lack of an appropriate design in the replicated studies. Therefore, replications need to be carefully selected to avoid the heterogeneity from the data showing the original positive signal, and large enough to possess the statistical power of replicating the original signal (Lernmark and Ott J1998). The biological and experimental proof should be considered critical and might be considered as the only "real" proof.

1.2 Power comparison of genome wide disease gene mapping strategies.

As mentioned above, linkage studies have been widely applied to genome wide gene mapping,

but usually give resolutions larger than 1cM subject to pedigree structure and sample sizes. Therefore, association studies are usually conducted to do fine mapping under linkage peak showing significant signals. Furthermore, linkage analysis might lack statistical power to detect moderate genetic effects of complex diseases. On the other hand, association tests have the ability to narrow disease susceptibility regions down to 1-100kb depending on the extent of LD, and are believed to be more powerful than traditional linkage studies. However, conducting a genome-wide association study requires a large genotyping effort due to the required high marker density. These pros and cons of both approaches lead to the question of which of the two genome-wide gene mapping strategies, applying association tests as a primary approach vs. as a follow-on to family-based linkage studies, has more power in genome wide studies.

In chapter 2, these two approaches were investigated utilizing GAW12 simulated data and methodologies suitable for the pedigree structure under study. Furthermore, a strategy of using Simes test (Simes 1986) as a method for combining results from LD tests of adjacent markers was investigated in order to control false positives. The results showed that the genome-wide association based tests are much more likely to identify genes, although a denser map of markers is required. The results were published. (**Meng Z**, Zaykin DV, Karnoub MC, Sreekumar GP, St Jean PL, Ehm MG. "Identifying susceptibility genes using linkage and linkage disequilibrium analysis in large pedigrees." Genet Epidemiol. 2001; 21 Suppl 1:S453-8.)

1.3 LD structure study and marker selection for association studies

One limitation of the association studies is the high genotyping cost due to the required high marker densities and large sample sizes in complex trait gene mapping. The intention of high

marker densities is to increase the chance of either including disease susceptibility loci or markers in LD with them and to provide enough statistical power to detect them. The current development of high-density maps of Single Nucleotide Polymorphisms (SNPs) provides a great source for such markers. However, genotyping of closely spaced SNPs frequently yields highly correlated data due to extensive LD between markers, it might be considered as "wasting resources" when these markers don't yield significantly different information in association studies.

Several recent studies investigated the empirical LD structure on different human chromosomal regions, and discovered that LD appears to be organized in block-like structures. Within these "blocks", limited genetic variation was observed. Daly et al. (2001) analyzed 516 chromosomes from a European-derived population typed for 103 SNPs in a 500 kb region on chromosome 5q31, and found the region could be decomposed into discrete haplotype blocks, which spanned up to 100 kb and contained 5 or more common SNPs. Johnson et al. (2001) scanned 135 kb of DNA, genotyped 122 markers in 9 genes, and determined haplotypes in a minimum of 384 European individuals. They advocated determining haplotypes as an approach that would provide the relationships between all alleles in the region. Based on all these observations, researchers raised a possibility of developing the "haplotype map" for human, which is a map consisting of haplotype blocks and SNPs in relatively low LD linking between blocks. Because of the low genetic diversities within blocks, only a relatively small number of SNPs is required to retain most of the information. Therefore, the association studies can provide a much clearer picture by conducting analyses based on haplotype blocks and much lower cost for genotyping. However, there were also doubts about the existence of haplotype blocks (Couzin 2002);

simulation results showing that the required SNP number might be much larger (Kruglyak 1999). The future of the haplotype map is still unknown, and its impact on association studies still requires further investigation.

Regardless of these contradictory views, using LD or haplotype information to select a subset of SNPs that optimizes the information retained in a genomic region while reducing the genotyping cost and simplifying the analysis without relying on "haplotype blocks" is still possible and can be quite helpful in association studies. In chapter 3, two procedures are developed to achieve this goal. One utilizes a spectral decomposition method based on matrices of pair-wise LD between markers, and the other extends David Clayton's htSNP selection method. The procedures require genotype information for a large initial set of SNPs on a small number of samples to select an optimum subset of SNPs that could be efficiently genotyped on larger numbers of samples while retaining most of the information on genetic variation. The properties of these procedures were studied using simulated data sets; minimum sample sizes needed for achieving consistent results were recommended; the procedure performances were evaluated using experimental data sets with measures of haplotype information; the possible impact of the marker selection on the association study was illustrated using a real example.

1.4 Mixed model for association study considering drug and gene-drug interaction

Investigating genetic effects and gene×treatment interactions is essential in determining individual differences in drug responses, a long desired problem in clinical trial studies. With this knowledge, we will have more confidence in reducing drug adverse events and prescribing the

right treatment to the right person. In order to study how genes are related to efficacy and safety of a medicine, multiple markers are genotyped in candidate genes on samples collected in clinical trials. Challenges for analysis include lack of validated statistical genetic methods, multiple correlated genetic markers, and low sample sizes. Quantitative response traits are of interest because they better reflect drug pharmacokinetics and pharmacodynamics, therefore provide more information on drug responses. Models taking into account the nature of genetic data including allelic, genotypic and haplotype effects are critical to extracting maximum information.

A linear regression approach considering genetic, treatment effects and their interactions directly should address this problem well. However, further studies are needed to understand the effects of the multiple factors involved, such as the nature of genetic effects, gene-drug interactions, and marker spacing in candidate genes, since the testing powers of different models, and "the best" approach vary under those effects. Therefore, in chapter 4, a model is proposed to separate the genetic effect into haplotype allelic additive effects and dominant effects of multiple markers in the candidate genes (Weir and Cockerham 1977; Zaykin et al. 2002), model the treatment, haplotypes additive effects, and treatment-haplotype interactions, and relate these effects to a quantitative trait. Furthermore, the variance components of a marker locus were related to those of a functional trait locus through genetic effects of trait alleles, LD between trait locus and markers, as well as their allele frequencies. Therefore, randomly genotyping markers in the candidate genes or regions of interests will introduce uncertainty into the study when the typed markers are not functional sites in question. Hence, more appropriate statistical models are required. A random effect model is proposed to treat haplotype effects and treatment-haplotype

interactions as random to account for this uncertainty. The model is compared to an analysis of variance approach with single marker genotypic classes and haplotype effects treated as fixed effects, and the power of utilizing multiple markers with a haplotype approach versus single markers is illustrated using simulated data.

Reference

Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA, Chakravarti A (2001) High-throughput variation detection and genotyping using microarrays. Genome Res 11(11):1913-25

Devlin B and Roeder K (1999) Genomic Control for Association Studies. Biometrics 55:997-1004

Elston RC and Stewart J (1971) A general model for the analysis of pedigree data. Hum Hered 21:523-542.

Forsell PA, Boie Y, Montalibet J, Collins S, Kennedy BP (2000) Genomic characterization of the human and mouse protein tyrosine phosphatase-1B genes. Gene 260:145-53

Hamilton-Williams EE, Serreze DV, Charlton B, Johnson EA, Marron MP, Mullbacher A, Slattery RM (2001) Transgenic rescue implicates beta2-microglobulin as a diabetes susceptibility gene in nonobese diabetic (NOD) mice. PNAS 98(20):11533-8

Jorde LB (1995) Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet 56:11-14

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073-80

Knapp M, Seuchter SA, Baur MP (1993) The haplotype-relative risk (HRR) method for analysis of association in nuclear families. Am J Hum Genet 52:1085-1093

Kruglyak L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139-44

Lander ES and Schork NJ. (1994) Genetic dissection of complex traits. Science 265:2037-48

Lander ES, Linton LM (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Lernmark A and Ott J. (1998) Sometimes it's hot, sometimes it's not. Nat Genet 19:213-214

Kruglyak L, Daly MJ, Reeve-Daly M, Lander ES (1996) Parametric and non-parametric linkage analysis: a unified multipoint approach. Am J of Hum Genet 58:1347-1363

Martin ER, Gilbert JR, Lai EH, Riley J, Rogala AR, Slotterbeck BD, Sipe CA, Grubber JM, Warren LL, Conneally PM, Saunders AM, Schmechel DE, Purvis I, Pericak-Vance MA, Roses AD, Vance JM. (2000) Analysis of association at single nucleotide polymorphisms in the APOE region. Genomics 63:7-12

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, et al. (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616-22

O'Connell JR. (2001) Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. Hum Hered 51:226-240.

Pericak-Vance MA, Bebout JL, Gaskell PC Jr, Yamaoka LH, Hung WY, Alberts MJ, Walker AP, Bartlett RJ, Haynes CA, Welsh KA, et al. (1991) Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. Am J Hum Genet. 48:1034-50

Penrose LS. (1953) The general purpose sib-pair linkage test. Annals of Eugenics (London). 18, 120-124

Pritchard JK and Rosenberg NA. (1999) Use of Unlinked Genetic Markers to Detected Population Stratification in Association Studies. Am. J. Hum. Genet. 65:220-228

Pritchard JK, et al. (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155: 945-959

Prince JA, Brookes AJ Towards high-throughput genotyping of SNPs by dynamic allele-specific hybridization Expert Rev Mol Diagn (2001) 1:352-358

Risch N, Teng J (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res 8:1273-88

Ryder LP, Christy M, Nerup J, Platz P, Svejgaard A, Thomsen M (1979) HLA studies in diabetics. Adv Exp Med Biol 119:41-8

Simes RJ (1986): An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751-754.

Spielman RS, McGinnis RE, Ewens WJ (1993) The transmission test for linkage disequilibrium: the insulin gene and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-516

Sturtevant AH (1913) J.EXP. Zool. 14, 43

Tomfohrde J, Silverman A, Barnes R, Fernandez-Vina MA, Young M, Lory D, Morris L, Wuepper KD, Stastny P, Menter A, et al. (1994) Gene for familial psoriasis susceptibility mapped to the distal end of human chromosome 17q. Science 264:1141-5

Venter JC, et. al. (2001) The sequence of the human genome. Science 291:1304-1351

Weir BS and Cockerham CC (1977): Two-locus theory in quantitative genetics. In: Proceedings of the international Conference on Quantitative Genetics, (eds) Pollak E, Kempthorne O, Bailey TB, Iowa State University Press, Ames, Iowa, pp 247-269

Weiss KW. (1993) Genetic variation and human Disease. Cambridge Univ. Press, Cambridge

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. (2002) Testing Association of Statistically Inferred Haplotypes with Discrete and Continuous Traits in Samples of Unrelated Individuals. Hum Hered 53:79-91

Chapter 2

Identifying Susceptibility Genes Using Linkage and Linkage Disequilibrium Analysis in Large Pedigrees

SUMMARY

Linkage and linkage disequilibrium tests are powerful tools for mapping complex disease genes. We investigated two approaches to identifying markers associated with disease. One method applied linkage analysis and then linkage disequilibrium tests to markers within linked regions. The other method looked for linkage disequilibrium with disease using all markers. Additionally, we investigated using Simes test to combine p-values from linkage disequilibrium tests for nearby markers. We applied both approaches to all replicates of the GAW12 problem 2 isolated population data set. We reported results from the 25th replicate as if it were a real problem and assessed the power of our methods using all replicates. Using all replicates, we found that testing all markers for linkage disequilibrium with disease was more powerful than identifying markers that were in linkage with disease and then testing markers within those regions for linkage disequilibrium with the implementations that we chose. Using Simes test to combine p-values for linkage disequilibrium tests on correlated markers seemed to be of marginal value.

INTRODUCTION

In this paper, we aim to compare the strategies of identifying markers in linkage disequilibrium within regions linked to disease versus identifying markers in linkage disequilibrium across the genome. Furthermore, we investigate the strategy of using Simes test as a method for combining results from linkage disequilibrium tests for nearby markers. We analyzed GAW12 simulated data using one linkage analysis method (SimIBD) applied to a 10cM map of markers to identify broad regions likely to contain disease genes and then one linkage disequilibrium analysis method (PDT) applied to all markers within those regions. We compared the results to applying another LD analysis method (Transmit) on all markers. We investigated use of the Simes test to combine results from linkage disequilibrium tests for adjacent markers. We feel that these are reasonable approaches for identifying markers nearest to disease genes without an abundance of false positives.

The present study was completed on all the replicates of the isolated population for GAW12 problem 2. We investigated the affection status trait provided to search for disease susceptibility genes. To illustrate our approaches, we report the results for the 25^{th} replicate. We estimated the power to find each of the 7 genes for a preset equal false positive rate for both approaches. The GAW12 problem 2 data set included 2855 STR markers with an average inter-marker distance of 1 cM in each replicate.

METHODS

We selected a single-point linkage analysis method, SimIBD [Davis, et al., 1996], to investigate linkage of the affection status trait with genetic makers. This method was selected for its speed and ability to analyze large pedigrees. SimIBD presents a non-parametric simulation-based statistic, which measures identity by descent sharing of alleles between affected relative pairs, reports a normalized Z statistic, weighted using population allele frequencies and gives a conditional empirical p-value. An empirical null distribution is determined by simulating marker genotypes in the affected subjects conditional on the marker genotypes in the unaffected subjects. The p-value reported is determined by the proportion of points in the null distribution that has a Z value greater than the observed Z statistic. To imitate a genome scan approach, we selected one marker every 10cM for analysis. We thought that it was a reasonable approach since, for micro-satellite makers, adding markers at a finer density isn't likely to increase the information available for linkage: linkage usually extends more than 10cM. We defined a linkage region as significant if the p value at a peak was less than 0.05. The region started from the first marker with a p value greater than 0.17 (Lod score = 0.2) to the left of the peak to the first marker with a p value greater than 0.17 to the right of the peak.

To identify markers in linkage disequilibrium with disease within linkage regions, we applied the transmission disequilibrium test [Spielman, et al., 1993] using the PDT program [Martin et al., 2000]. PDT uses data from affected and unaffected individuals in related nuclear families in extended pedigrees. It calculates a statistic, T, which sums weighted transmission information from discordant sibling pairs and trios within the pedigrees. Under the null

hypothesis of no association in the presence of linkage, this statistic is asymptotically normal with mean 0 and variance 1. When investigating age of onset and age of exam, we noted that many of the individuals identified as unaffected were examined at ages that were considerably less than the mean age of onset. We suspected that some people classified as unaffected should have been called unknown, because they were too young to develop the disease. By investigating the distributions for age of onset in affected people and age of exam in unaffected people, we decided that individuals needed to be at least 45 years at exam to be identified as unaffected with disease. Individuals less than 45 years at exam and not affected, were assigned an unknown affection status. Then we calculated the above statistic using the new affection status. The p-values reported were based on the normal distribution.

We applied the transmission disequilibrium test using Transmit [Clayton D, 1999] to test for linkage and association using all the markers, not just ones within linkage regions. This method calculates a score test based on a partial score function that omits the terms most influenced by hidden population stratification. The test is proposed for the situation in which transmission is uncertain and is applied to all nuclear families within the extended families. Under the null hypothesis of no linkage or association, the score vector asymptotically follows a χ^2 distribution with degrees of freedom (df) equal to $H-1$ where H is the number of distinct alleles. The reported statistic uses transmission information only to affected individuals. It has been shown [Clayton D, 1999] that the Transmit statistic is still valid while using all the affected siblings, even when you are assuming linkage. However, linkage induces a correlation structure within distant pedigree members that is not taken into account in the implementation of Transmit that was available to us. Since we were using the entire extended pedigrees when calculating the

Transmit statistic, we tested the null hypothesis of no linkage and no association. This is distinct from the PDTest, which is a valid test for association in the presence of linkage. Because the Transmit test requires the assumption of Hardy-Weinberg equilibrium, we tested each marker to see if the genotype frequencies were in HWE.

To identify significant regions of linkage disequilibrium based on statistics calculated with correlated markers, we applied Simes test [Simes, RJ 1986], which is a method for testing the intersection of hypotheses and controlling type 1 error for the whole set. Sarkar and Chang [1997] proved that it was applicable to positively correlated dependent hypotheses. Simes test reports a combined p-value for a set of p-values based on the following formula, where $p_{[i]}$ is the i^{th} order statistic and n is the number of p-values in the set.

$$Min[np_{[1]}, \frac{n}{2}p_{[2]}, \frac{n}{3}p_{[3]}, \ldots, \frac{n}{n-1}p_{[n-1]}, p_{[n]}]$$

We applied Simes test replacing the p-value of the middle marker by the combined p-value from the first n markers starting from the p-ter of the chromosome. Note that this would control the type 1 error rate of only the n markers, if we had one fixed. However, we are using this method as a smoothing technique for highly variable p-values. We continued this for each set of n markers in a sliding window across the chromosome. To determine the window size, we calculated the correlation of the marker p-values for the PDT and the Transmit results. The size of the window remained constant for each method. We assumed that the markers were equally spaced and the autocorrelation depended only on the number of markers between the two markers rather than either of the markers themselves.

We analyzed all 50 replicates using these 4 methods (1-SimIBD & PDT, 2-Transmit, 3-SimIBD, PDT with Simes, and 4-Transmit with Simes). When comparing the markers with significant linkage disequilibrium tests to the "Answers", any marker was a true positive if it was 3 cM either to the left or to the right of the disease gene. Applying SimIBD and PDT tests sequentially to the data made it difficult to control the type I error for each approach. Applying Simes test complicated this further. To make the comparison of these 4 methods fair, we adjusted the α level used for each test within each approach so that the observed false positive rate, for the approach, was close to 0.05.

The observed false positive rate was calculated as the proportion of all markers tested identified as significant that were not within 3cM of a gene. We summarized the number of times we identified each major gene in 50 replicates and computed the true discovery rate (the number of true positives over the total number of positives) for each approach.

RESULTS

We didn't find any evidence against the HWE assumption. The p-values of linkage disequilibrium analysis results (PDT and Transmit) were positively correlated and the average correlation extended to 3 markers. We smoothed the PDT and the Transmit results using Simes test with window sizes of 3.

Replicate 25

We controlled the false positive rate to be 0.05 for all approaches by choosing a different α-level for each test as follows: when applying Simes test, the α–level was 0.05 for SimIBD, 0.09 for PDT and 0.03 for Transmit. Without Simes test, the α-level was 0.05 for SimIBD, 0.07 for PDT, and 0.05 for Transmit. The results are summarized in Table 1. For each linkage region, we listed the boundaries of the region and the peak marker name along with the corresponding p-values. For each set of markers showing significant linkage disequilibrium, we listed p-values only. Analysis using Transmit resulted in so many significant results that, we listed only the first 12 most significant ones.

We identified 12 linkage regions and 4 LD regions under the linkage regions (SimIBD & PDT). Among them, the region D09G120-122 contains MG3. The true discovery rate for regions was 25.0% (1/4) considering 4 regions identified by PDT. Transmit identified 41 LD regions looking for linkage and association (not all listed). Among them, the region D06G032-043

contains MG6 & MG7 and the region D19G026-032 contains MG1. The true discovery rate was 4.9% (2/41).

Power Study

We counted the number of times each major gene was found in the 50 replicates, and summarized the results in Table 2. We also listed the average number of tests applied, the true discovery rate and the false positive rate for each approach.

Comparing the SimIBD & PDT approach to the Transmit approach, we can see that a genome-wide association test did increase our chances of finding the disease genes. Applying Simes test reduced our power to locate disease genes, but a greater percentage of the markers we identified as significant were near genes.

DISCUSSION

We analyzed GAW12 simulated data to compare the two commonly used strategies of identifying disease genes (linkage disequilibrium analyses within regions linked to disease and linkage disequilibrium analyses across the genome). Neither of these two approaches was extremely powerful, but our results show that genome wide linkage disequilibrium analyses increase the chance of finding genes compared to the linkage and association strategy. The above conclusion is only based on the implementations we chose in this paper. A possible shortcoming is that the number of markers to be typed will be significantly higher, but the probability of finding the genes using any replicate is also much higher with in this approach. Our approach of using Simes' test for smoothing p-values is similar in spirit to the method of Goldin, et al., [1999] who suggested that p-values could be averaged across certain genetic distances. The underlying idea of such approaches is that multiple significant results in a chromosomal region provide more support for the presence of a gene than a single significant result. Thus, the false discovery rate might be decreased. For similar reason, Juo, et al., [1997] required that the p-values for flanking markers around significant tests must also show the tendency to be "small", if not significant. Terwilliger, et al., [1997] and Knapp [1998] presented theoretical considerations pointing out that the positive correlation between test statistic values extends further around true positives. We have chosen Simes' test based on preliminary simulation study (data not shown) that revealed its better performance. Both approaches (averaging and applying Simes' test) however suffer from the fact that the resulting overall p-value cannot possibly be smaller than the minimum p-value in the window. As a consequence, the power can be reduced, and Table 2 indicates this. On the other hand, Table 3 shows that the true discovery rate is somewhat

increased. During analysis of GAW12 simulated data sets, Hardy et al. (personal communication) used another method proposed by Zaykin, et al., [2001] that may not suffer from the mentioned potential loss of power, allowing the overall p-value to be smaller than the minimum in the combined set. Further study is needed to compare suggested techniques. Both, original and smoothed sets of significant results show that the genome-wide association based tests (i.e. Transmit) are much more likely to identify genes, although a denser map of markers is required.

ACKNOWLEDGMENTS

Thanks to Xiaobin Li, Xiaohui Luo, Brad Freeman, Santhi Sampath, and Rusty Czerwinski for technical assistance and thanks to Michael Wagner and Bruce Weir for valuable discussions.

REFERENCES

Clayton D (1999): A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am. J. Hum. Genet 65: 1170-1177.

Cottingham RW Jr, Idury RM, Schaffer AA (1993): Faster sequential genetic linkage computations. Am. J. Hum. Gent. 53: 252-263.

Davis S, Schroder M, Goldin LR, Weeks DE (1996): Non-parametric simulation-based statistics for detecting linkage in general pedigrees. Am. J. Hum. Genet. 58: 867-880.

Goldin LR, Chase GA, Wilson AF (1999): Regional inference with averaged p-values increases the power to detect linkage. Genetic Epidemiology 17:157-164.

Juo SH, Beaty TH, Duffy DL, Maestri NE, Prenger VL, Zeiger J, Lei HH, Coresh J (1997): A comprehensive analysis of complex traits in problem 2A. Genetic Epidemiology 14(S):815-820.

Knapp M (1998): Discriminating between true and false-positive peaks in a genomewide linkage scan, by use of the peak length. Am. J. Hum. Genetics. 62:1561-1562.

Martin ER, Monks SA, Warren LL, Kaplan NL (2000): A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am. J. Hum. Genet. 67: 146-154.

Ott J (1989): Computer-simulation methods in human linkage analysis. Proc. Natl. Acad. Sci. USA 86(11):4175-8

Sarkar SK, Chang CK (1997): The Simes method for multiple hypothesis testing with positively dependent test statistics. J. Am. Stat. Assoc. 92, 1601-1608.

Simes RJ (1986): An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751-754.

Spielman RS, McGinnis RE, Ewens WJ (1993): Transmission test for linkage disequilibrium: the insulin-dependent diabetes mellitus. Am. J. Hum. Genet. 52: 506-516.

Terwilliger JD, Shannon WD, Lathrop GM, Nolan JP, Goldin LR, Chase GA, Weeks DE (1997): True and false positive peaks in genomewide scans: applications of length-biased sampling to genome mapping. Am. J. Hum. Genet. 61:430-438.

Weeks DE, Ott J, Lathrop GM (1990): SLINK: a general simulation program for linkage analysis. Am. J. Hum. Genet. 47: A204

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS (2001): Truncated product method for combining p-values. (Genetic Epidemiology, *in press*)

Table 2.1 Analysis result for isolate replicate 25

SimIBD		SimIBD & PDT with Simes		Transmit with Simes	
Linkage regions	**Marker & p-value at peak**	**LD regions**	**p value at peak**	**LD regions**	**p-value at peak**
D01G115-143	D01G120 & 0.0061			D01G050-052	0.0009
D02G167-186	D02G177 & 0.0241			D01G126-128	0.0045
D03G010-057	D03G023 & 0.0037	D03G024-026	0.0858	D03G043-045	0.0001
D03G057-073	D03G065 & 0.0208			D04G153-155	0.0015
D04G010-032	D04G026 & 0.0214			D05G045-050	0.0012
D05G112-142	D05G132 & 0.0089			D06G032-043*	0.0003
D06G085-119	D06G097 & 0.0472			D06G055-057	0.0024
D06G129-152	D06G146 & 0.0257	D06G142-144	0.0303	D08G055-071	0.0018
D09G102-129	D09G110 & 0.0197	D09G120-122*	0.0432	D12G050-052	0.0021
D18G086-106	D18G103 & 0.0091			D13G073-075	0.0001
D19G102-105	D19G105 & 0.0130			D16G107	0.0030
D22G010-026	D22G016 & 0.0169	D22G024-026	0.0678	D19G026-032*	0.0015

Table 2.2 Number (percentage) of replicates correctly identifying each of the 7 major genes

Major Gene	With Simes		Without Simes	
	Linkage & PDT	Transmit	Linkage & PDT	Transmit
MG1	3 (6%)	4 (8%)	1 (2%)	8 (16%)
MG2	0 (0%)	5 (10%)	1 (2%)	14 (28%)
MG3	1 (2%)	4 (8%)	1 (2%)	14 (28%)
MG4	0 (0%)	4 (8%)	0 (0%)	10 (20%)
MG5	1 (2%)	7 (14%)	2 (4%)	12 (24%)
MG6	8 (16%)	44 (88%)	10 (20%)	49 (98%)
MG7	7 (14%)	41 (82%)	10 (20%)	47 (94%)

Table 2.3 False positive and true discovery rates over all 50 replicates

	With Simes		Without Simes	
	Linkage & PDT	**Transmit**	**Linkage & PDT**	**Transmit**
Total no. of markers tested	15458	142750	15643	142750
Total no. false positives	765	6947	762	7291
Total no. true positives	38	294	24	233
False positive rate	4.95%	4.87%	4.87%	5.10%
True discovery rate	4.73%	4.06%	3.05%	3.10%

Chapter 3

Selecting Genetic Markers for Association Analyses Using LD and Haplotypes

Abstract

Genotyping closely spaced SNP markers frequently yields highly correlated data due to extensive linkage disequilibrium (LD) between markers. The extent of LD varies widely across the genome, and drives the number of frequent haplotypes observed in small regions. Several studies have illustrated that it may be possible to use LD or haplotype data to select a subset of SNPs that optimizes the information retained in a genomic region while reducing the genotyping cost and simplifying the analysis. Generally applicable methods are needed to select a minimum subset of SNPs sufficiently retaining most information provided by haplotypes observed in a region. We proposed a spectral decomposition method based on the matrices of pairwise LD between markers, modified David Clayton's htSNP selection method that utilizes haplotype information, and proposed algorithms allowing the methods to be applied to large chromosomal regions. Our procedures require genotype information for a large initial set of SNPs on a small number of individuals, and select an optimum subset of SNPs that could be efficiently genotyped on larger numbers of samples while retaining most of the genetic variation in samples. We studied the properties of procedures using simulated data sets in linkage equilibrium and disequilibrium, and reported minimum sample sizes needed for consistent results. Procedures were applied to experimental data sets with SNPs at average densities of one SNP every 20 or 30 kb and evaluated using haplotype information measures. Both procedures were similarly effective at reducing the genotyping requirement while maintaining the genetic information content throughout the regions. We also illustrated the procedure impact on an association study result in a region around the CYP2D6 gene (Hosking et al. 2002).

Introduction

Efforts to positionally clone susceptibility genes for common, oligogenic diseases have led to the development of high-density maps of Single Nucleotide Polymorphisms (SNPs) distributed across the human genome (Sachidanandam et al. 2001). Theoretical studies have suggested that association tests employing such high-density SNP maps, either as a primary approach or as a follow-on to family-based linkage studies, should be more powerful in detecting disease susceptibility genes than traditional linkage approaches (Risch and Merikangas 1996). However, the precise numerical meaning of "high-density" is a matter of debate, and has significant implications on the cost and practicality of conducting SNP association studies. An optimum strategy would be to genotype enough SNPs to capture the large majority of information on genetic variation within a defined chromosomal region while avoiding typing SNPs that yield redundant information due to extensive linkage disequilibrium (LD) between nearby SNPs. Defining the optimum set of SNPs will require knowledge of the patterns of linkage disequilibrium across the human genome.

Recently, several studies investigated the empirical LD structure on different human chromosomal regions, and discovered a common pattern that LD appears to be organized in block-like structures, where a contiguous group of SNPs comprising a block show high levels of pair-wise LD between SNPs and where there is little LD between SNPs in different blocks. Authors (Subrahmanyan et al. 2001; Daly et al. 2001; Johnson et al. 2001; Dawson et al. 2002; Gabriel et al. 2002; Patil et al. 2001) have reported block-like LD structures showing considerable spatial variation across different genomic regions, extending up to several hundred kb, and exhibiting differing boundaries in samples from different ethnic groups. Reduced

haplotype diversities within blocks, given the number of SNPs involved, are observed not only in genotype data with numerically inferred haplotypes, but also in experimentally determined haplotype data (Patil et al. 2001). The reduction of haplotype diversities suggests the possibility of identifying SNPs to define the common haplotypes thereby reducing the number of markers needed to capture the majority of the genetic information about the region. A procedure that utilizes genotype information on a small number of samples to prioritize SNPs for typing on a large number of samples could be useful in increasing the experimental efficiency in any project involving a high-density map of SNPs. Examples include testing multiple SNPs within a candidate gene for association, fine-mapping a region identified using linkage analysis, and testing thousands of SNPs as part of a genome-wide association study. Furthermore using a technique to identify the most independent and informative SNPs could be helpful in interpreting analyses across a region where a large number of highly correlated SNPs have been typed. Such a procedure could helping an analyst see real support for the association in a region without the redundant information provided by highly correlated SNPs. On the other hand, any "marker selection procedure" relies on an arguable assumption that common SNP variation can provide high predictive values for risks associated with complex diseases (Couzin 2002). Thus, these procedures are only valuable to the extent that the original set of SNPs is useful for association mapping purposes. Nevertheless, marker selection can be viewed as a procedure for identifying polymorphisms most characteristic of underlying populations.

Several algorithms have been proposed to detect haplotype blocks and (or) select markers. Patil et al (2001) utilized a greedy algorithm to partition the entire chromosome into a set of contiguous haplotype blocks while minimizing the total number of representative SNPs that

distinguish at least α percent of the unambiguous haplotypes in each block. Zhang et al (2002) extend Patil et al's greedy algorithm to a dynamic programming algorithm, which can guarantee an optimal solution for haplotype partitioning. Therefore, both Patil et al and Zhang et al selected markers by identifying the minimum number of SNPs distinguishing at least α percent of the unambiguous haplotype in the blocks. These algorithms require haplotype phase known data. In this case, haplotypes were determined experimentally. These procedures are not applicable to unphased genotype data, and rely on their definitions of block boundaries to select markers. Other "block defining" algorithms such as those described in Daly et al (2002) and Gabriel et al (2002) do not require phase known haplotype data. Daly et al used a combination of methods including familial data and the EM algorithm to estimate haplotype frequencies. Then they initially defined blocks by comparing the observed haplotype heterozygosity with that expected assuming Hardy Weinberg Equilibrium within consecutive five markers windows, identifying "lower diversity haplotypes cores", and constructing "blocks" by extending or shrinking two ends of the cores until reaching the longest local minimum "blocks". Next, a hidden Markov model was also used to formally define the blocks by assigning observed chromosomes to one of the four ancestral haplotypes, accessing the significance of the estimations of the historical recombination rate (θ) between each pair of markers. Gabriel et al used *D'* and associated confidence intervals as a measure of the historical recombination and defined "blocks". Both methods appear to be specific for the particular data sets used and their general applicability is not known. Johnson et al (2001) proposed two methods to select markers within genes based on gene haplotypes constructed either using the family data or the expectation-maximization (EM) algorithm on unrelated individuals. One method orders the haplotypes by their similarities and requires selecting SNPs by eye. The other suggests a htSNP diversity method proposed by

Clayton (2001) to select haplotype tagging SNPs (htSNPs) to best extract the haplotype information in a gene. The first method is difficult to automate, and the second can become quite computationally intensive when a large number of markers are considered in a region. (Detailed reason is shown later.) Both methods require the predetermination of haplotypes for the region considered, which is difficult when the region contains a large number of markers.

It is worth noticing that all the "block detecting" methods mentioned may result in differing block boundaries. Given the diversity of methods used to define blocks and conflicting assertions as to whether they exist at all (Couzin 2002), we choose to develope marker selection procedures that do not rely on defining "blocks". Instead, we select a set of SNPs that retain haplotype information similar to an original (and presumably) larger set of SNPs. Furthermore, the procedure should be applicable to regions with a large number of SNPs, and data sets without haplotype phase information or family information. We propose a method based on the spectral decomposition of the matrix of the pair-wise linkage disequilibrium coefficients of the markers, and compare it to the htSNP diversity method proposed by Clayton (2001). Both methods can be utilized to select a subset of markers that maintain haplotype information available from the set of markers before selection. The spectral decomposition of the LD matrix is a reductionist approach, which considers many pair-wise LD coefficients at one time. Clayton's htSNP method relies on haplotype information rather than pairwise LD coefficients making it a complementary approach. The spectral decomposition method has a population-genetics justification and advantages over considering a single pair-wise LD coefficient at a time. In addition, we propose a procedure summarizing the information obtained from a sliding window approach to allow both methods to be applied in large chromosomal regions. Our procedures are

local in that they are applied to genetically proximal sets of markers by considering relatively short windows of markers covering genetic distances that are generally less than 500 kb. Furthermore, none of the existing marker selection methods have been evaluated using quantitative criteria describing the proportion of the information retained in the selected marker sets. We compare two local haplotype diversity measures: the haplotype heterozygosity and the number of frequent haplotypes before and after application of the procedures to access the information retained and measure the success of these procedures. We summarize the results from two simulation studies to evaluate the performance of these procedures, apply them to two experimental data sets as examples, and also apply them to markers typed around CYP2D6 where an association has been identified to show how the marker selection procedures impact these association study results. (Hosking et al. 2002)

Materials and Methods

We describe two marker selection methods, the procedure extending them to a large chromosomal region, two simulation studies to evaluate the procedures, and criteria by which we evaluate the performance of the procedures when applied to two experimental data sets. Our selection procedures study relatively polymorphic SNPs with minor allele frequency (MAF) of at least 0.05.

Spectral decomposition (spD)

Population genetics theory predicts that the linkage disequilibrium associated with alleles from three or more markers decays more rapidly than LD associated with alleles from two markers (Bennet 1954). Therefore, it is reasonable to describe dependencies between markers by considering only pairwise correlations. Moreover, the precision of estimates and power to detect LD associated with alleles from three or more markers quickly diminishes with their order. The essence of spectral decomposition is to represent an entire variance-covariance matrix (LD matrix in our case) in terms of its eigenvalues and eigenvectors. Since the spectral decomposition-based method (spD) that we propose takes into account all pairwise disequilibria for a set of markers, it assumes that the LD associated with alleles from three or more markers is negligible. Therefore, the method assumes that most of the practically available haplotype information can be recovered from pairwise LD and single marker characteristics. Spectral decomposition is also the basis for the principal component analysis (PCA). In PCA, the sample variation is represented by a few linear combinations (the eigenvectors) of all original variables (i.e. SNPs), taken with different weights (the eigenvalues) to reflect their importance. In contrast, we examine all eigenvectors (linear combinations of the marker contributions) and eigenvalues

(the importance of the corresponding combinations), and retain only a subset of the original variables that contribute more to the more important weights**.** Note that this procedure allows us to consider the pairwise LD coefficients of all markers at once instead of only considering the LD measure for a pair of markers at a time. Let L be the number of markers evaluated. For a set of markers, $m_1, \ldots, m_L$, the LD matrix is $\boldsymbol{R}_{LxL}$ with the pair-wise correlation r_{ij} as components, where Δ_{ij} is the composite LD (Weir 1996) between markers i and j. (See Appendix 1)

$$r_{ij} = \hat{\Delta}_{ij} / \sqrt{\mathrm{var}(\hat{\Delta}_{ij})}$$

Applying the spectral decomposition technique, $\boldsymbol{R}$ can be written as

$$\mathbf{R} = \sum_{i=1}^{L} \lambda_i \mathbf{e}_i \mathbf{e}_i^T$$

where e_i and λ_i are eigenvectors and eigenvalues of $\boldsymbol{R}$, $i = 1, \ldots, L$, and $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_L$. Note that the variables (markers) that contribute more to the eigenvectors associated with the first several large eigenvalues, are considered the more influential variables (markers) for that LD matrix, $\boldsymbol{R}$**.** Variables that contribute more to the eigenvectors associated with subsequent eigenvalues, are considered less influential.

To determine if there are variables or markers, which have little or no influence on the LD matrix, the following L_r index is calculated (See Appendix 2):

$$L_r = L \frac{\sum \lambda_i^2}{(\sum \lambda_i)^2} - 1$$

L_r equal to 0 indicates that all the markers in the set provide important information and the whole set should be kept. This measure is derived by examining the conditions of extreme disequilibrium and complete independence. We find it useful in identifying when no SNPs should be eliminated from the set. If $L_r > 0$, the actual number of markers to be retained, x, is most precisely determined from the inequality:

$$\frac{\sum_{i=1}^{x} \lambda_i}{\sum_{i=1}^{L} \lambda_i} \geq \alpha$$

where α is the proportion of information retained (proportion of variation explained). Therefore, we retain markers while the sum of the eigenvalues corresponding to the eigenvectors they contribute more is a high proportion of the sum of all eigenvalues. Appropriate levels for α will be investigated in the simulation studies.

It is not always clear which marker contributes more to which eigenvalue or eigenvector. To sharpen marker loadings to particular eigenvectors, we apply the varimax-rotation procedure to the original set of eigenvectors, $\boldsymbol{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_L\}$. This procedure finds an orthogonal transformation $\boldsymbol{T}$, $\boldsymbol{E}^* = \boldsymbol{ET}$ that will confine influence of each marker to a particular eigenvector. For each marker, m_j, compute the following:

$$\Gamma_j = \frac{1}{x} \sum_{v=1}^{x} | e^*_{jv} |$$

$$\gamma_j = \frac{1}{L - x} \sum_{v=x+1}^{L} | e^*_{jv} |$$

where e_{jv}* is the j^{th} element of v^{th} eigenvector of E*. Marker, m_j, is selected if $\Gamma_j > \Upsilon_j$. That is, this marker contributes more to the more independent variations than the redundancy.

Haplotype diversity (div)

Clayton (2001) proposed the following method to select a subset of SNPs using haplotype information. Let N be the total number of haplotypes in the sample, which is two times the number of individuals for a diploid population. For L-diallelic-marker haplotypes, each haplotype can be written as a vector $z_i = \{z_{ij}, j = 1\ldots L, i = 1\ldots N\}$, where z_{ij} is either 0 or 1 representing one of the two alleles. Haplotype diversity can be defined as the total number of differences in all N^2 pair-wise comparisons between a pair of haplotypes. $(z_{ij}\text{-}z_{kj}) = 0$ if haplotype i and k are the same at locus j. $(z_{ij}\text{-}z_{kj}) = \pm 1$ if they differ. Haplotype diversity at locus j is calculated as

$$D_j = \sum_{i=1}^{N}\sum_{k=1}^{N}(z_{ij} - z_{kj})^2 = 2\left\{N\sum_{i=1}^{N} z_{ij}^2 - (\sum_{i=1}^{N} z_{ij})^2\right\}$$

Clayton proposed that the total haplotype diversity, given below, is calculated as the summation over all loci, which is analogous to the total sum of squares in an ANOVA setting.

$$D = \sum_{i=1}^{N}\sum_{k=1}^{N}(\mathbf{Z_i} - \mathbf{Z_k})^T(\mathbf{Z_i} - \mathbf{Z_k}) = \sum_{j=1}^{L} D_j$$

where L is the number of loci.

Haplotype tagging SNPs (htSNPs) are a set of SNPs that retain most of the information available in the full haplotype. After selecting a set of htSNPs, N haplotypes are collapsed into groups

according to different htSNP allele combinations. That is, if haplotypes consisting of L SNPs are under study, and H out of L SNPs are selected to be candidate htSNPs, any haplotypes will belong to the same group as long as they have the same alleles at these H loci. Then the N full haplotypes are divided into $G=2^H$ (at most) groups. Within each group, a similar diversity measure to that above is computed. Within group haplotype diversity is then summed over all groups, which is analogous to the residual sums of squares.

$$R = \sum_{g=1}^{G} \left\{ \sum_{i \in G_g} \sum_{k \in G_g} (\mathbf{Z}_i - \mathbf{Z}_k)^T (\mathbf{Z}_i - \mathbf{Z}_k) \right\}$$

Then Clayton (2001) calculated the proportion of diversity explained by a set of htSNPs as $p = 1 - R/D$. R/D is preferred to be as close to 0 as possible indicating that there is little diversity left when the haplotype is represented by the subset of htSNPs. The optimal htSNP set is obtained by an exhaustive search from the possible 2^L-1 candidate sets. This is very computationally intensive as the number of markers, L, increases. Since Clayton (2001) does not provide guidance on obtaining a good set of htSNPs without searching the entire set, we propose selecting a set of htSNPs by minimizing the number of SNPs selected and maintaining the proportion of diversity explained by htSNPs, p, greater than a desired value, say α.

We have simplified the expressions for both D and R in Clayton's formula to the following.

$$D = \sum_{j=1}^{L} D_j = \sum_{j=1}^{L} 2(Nn_{0j} - n_{1j}^{\ 2}) = \sum_{j=1}^{L} 2n_{0j}n_{1j} = N^2 \sum_{j=1}^{L} 2p_{0j}p_{1j}$$

$$R = \sum_{j=1}^{L} \sum_{g=1}^{G} 2n_{0jg}n_{1jg} = N^2 \sum_{j=1}^{L} \sum_{g=1}^{G} 2p_{0jg}p_{1jg}$$

where n_{0j} and n_{1j} are the number of 0's and 1's at locus j, p_{0j} and p_{1j} are the frequencies of "0" and "1" alleles at locus j. Here $2p_{0j}p_{1j}$ is the expected heterozygosity measure for the j^{th} locus. Correspondingly, the proportion of diversity explained by htSNPs can be written as

$$p = 1 - \frac{\sum_{j=1}^{L}\sum_{g=1}^{G} 2 p_{0jg} p_{1jg}}{\sum_{j=1}^{L} 2 p_{0j} p_{1j}} > \alpha$$

Therefore, htSNPs are selected by trying to minimize the within group loci heterozygosity. After the simplification, the above measure can be extended to analyze multi-allelic markers by extending p_0, p_1 to p_i, where $i = 0 \ldots T$ and T is the total number of alleles at this marker. Clayton (2001) suggested a Kappa correction, which corrects for the fact that selecting a set of ht SNPs will always reduce the residual diversity. Also note that when haplotypes are not known with certainty, the EM algorithm is used to infer haplotype frequencies.

Applying spD or div to a large chromosome region

In selecting markers that maintain haplotype information, it is important for us to consider the haplotype information they provide in the context of nearby markers rather than for any marker – regardless of its position. That is, we decide not to include markers not only if they provide similar information, but also if they are fairly close to each other. Therefore, we propose the following procedure to apply either spD or div to a large chromosome region with a large number of SNPs. First, we assume that markers are arranged according to an ordered map. Next, a sliding window with a relatively small window size is moved along the map. Either spD or div method is used to select informative SNPs in each window. The event of selecting or failing to select a SNP is recorded in a vector $Wi = \{w_{ij}, j = 1 \ldots L\}$, where L is the number of SNPs in a

window (or the window size). $w_{ij} = 1$ indicates that j^{th} SNP is not selected in i^{th} window, 0 otherwise. Most of SNPs appear in multiple windows with the maximum number of windows equal to the sliding window size. Each marker's relative redundancy is computed by averaging its corresponding w_{ij} over all the windows in which it appears. The relative redundancies of all markers are recorded in another vector $RR = \{rr_m, m = 1, \ldots, M\}$, where M is the total number of markers in the chromosome region, and rr_m is called the relative redundancy of m^{th} SNP. A SNP is dropped from the final list when its relative redundancy is above a predetermined threshold. Note that the window size (L) and relative redundancy threshold are adjustable parameters.

Ideally, the sliding window size should be changed to reflect differing amounts of LD in the data. More SNPs should be included in a window and examined together when they are in high LD and fewer SNPs examined together when there is less LD between them. Practically, it is difficult to identify regions of high and low LD and choose the window sizes accordingly. Therefore, we propose applying the procedure multiple times so that we drop SNPs that provide redundant information in the first run and then rerun the procedure on the resulting data set. Then if there are some regions with highly correlated markers interspersed with regions having fewer correlated markers, the initial runs will drop markers within the highly correlated regions and retain markers in the regions with less correlation. Subsequent runs will then examine the resulting markers using small window sizes. Note that they will have more similar levels of correlation and will be appropriate for a fixed window size. We will refer to this set up as "repeated runs". Additional runs can be repeated until the procedure "converges." Convergence is achieved when the difference between the number of markers before and after selection

represents no more than 5% of the markers before selection. We will refer to the procedures based on the spectral decomposition and the diversity methods as upper case SD and DIV, respectively.

Simulation Studies

We have designed two simulation studies to study the performance of SD and DIV. The first study investigates how the procedures behave when applied to SNPs in linkage equilibrium (LE). The second study investigates what sample sizes provide consistent selection results. If we put these procedures into a hypothesis-testing framework, the first study is similar to controlling the false positive rate under the null hypothesis– how often do we drop important markers that should be included in our set. Also note that we are interested in the true positive rate, or how often we drop markers and maintain the desired information when there are redundancies among them. This will be addressed by using the experimental data rather than a simulation approach.

Simulation study I: will the SNP selection procedure drop "important" SNPs?

When the SNPs are in LE, all SNPs should be selected if both SD and DIV identify informative SNPs and only drop SNPs that are redundant. Note that the performance of both methodologies, SD and DIV, will be affected by the set of parameters used, such as the sliding window size, the percentage of the variation explained, the relative redundancy threshold for each marker, and the sample size. We identify suitable parameter combinations by studying the procedures' behaviors for SNPs in LE and controlling the SNPs' drop rate (the false positive rate) to be less than 5%. First, we simulate genotype data for 50 SNPs with MAF greater than 5%, randomly drawn from the uniform distribution and in linkage equilibrium with each other. We apply either the SD or

DIV method, record the percentage of markers, out of 50, dropped, and average this percentage over 100 simulation runs. The "repeated runs" setup is not implemented here, since the purpose is to obtain the proportion of SNPs dropped at each single run and find parameter combinations than ensure this proportion is less than 5%. Note this threshold is the convergence criterion for "repeated runs". In practice, we recommend applying the procedures with the parameter combinations determined here. If the drop percentage in one run is below 5%, we stop the procedure and declare "convergence" to prevent dropping informative SNPs. For SD, we investigate the following parameter combinations: sliding window sizes of 2, 5, 10, 15 and 30, and percentages of the variation explained of 85% and 90%. For DIV, we investigate sliding window sizes of 2, 3, 5, and 7, and percentages of the variation explained of 92% and 96%. Note that values of the percentage of the variation explained are calculated using different methods for SD and DIV, have different interpretations, and cannot be directly related to each other. For both SD and DIV, we investigate relative redundancy thresholds of 50%, 70% and 90% and sample sizes of 10, 50, 100 and 200 individuals. In addition, we look at the effect of availability of haplotype phase information by providing the same data in both haplotype phase-known and phase-unknown forms.

The percentage of variation explained, as a parameter, significantly determines the proportion of the LD information conserved in the selected marker set. We like to set this parameter high enough to conserve the required amount of LD to map susceptibility genes successfully, but low enough so that we can afford to genotype all the required markers. However, the optimal amount of information is affected by many factors including the effects of the gene, the degree of LD between markers, marker allele frequencies, and distribution of markers. Therefore an optimal

value of the percentage of variation explained for all cases will not exist. To provide a reasonable value, we propose to identify a variation explained percentage parameter that results in less than 5% of markers in LE being dropped. We reason that it is important to control the percentage of markers dropped when all of them should be retained. Note that when there is some LD between markers, this variation explained parameter will result in dropping more than 5% of the markers since there is LD between them. In this situation, it is difficult to determine if the percentage of markers being dropped is correct. To assess this, we use the real data to investigate the haplotype information in the region before and after dropping markers. Note that the variation explained percentages are calculated differently for SD and DIV, and can't be compared to each other directly. Despite this, we find a rough correspondence between them so that we can compare their behavior fairly. This allows us to calibrate the behavior of the methods under "the null hypothesis" of high importance of all markers, and is analogous to setting a common rejection region for power comparisons of statistical tests. Note the SD procedure calculates a L_r measure that essentially prevents dropping SNPs under LE, thus preventing us from studying its behavior under LE and finding the corresponding DIV variation explained percentages. We will not use this measure for this part of the simulation study. We simulated data as described above, applied SD or DIV with a range of the variation explained percentages, and computed the average percentage of SNPs dropped. The sample sizes considered were 50 and 100. The rest of parameters were chosen based on the results from the first part of the simulation I.

Simulation Study II: what sample size is required to ensure consistent results across sample?

Ideally, we would like to obtain information on the most informative markers representing the population we are interested in, but with as few samples as possible to keep genotyping costs

low. Therefore we design a simulation study that investigates the consistency of our procedures as a function of the sample size used for marker selection. First, we simulate a large diploid data set, referred to as "the population" later, containing a chromosomal region with 50 SNPs and 20,000 individuals using a forward simulation model that assumes constant population size, non-overlapping generations, random mating and no other disturbing forces except recombination. Initial LD in the data is created by mixing two populations with discrepant allele frequencies. The number of generations, the recombination rate and the initial LD determine the degree of LD in the final generation. We also investigate whether the degree of LD has an effect on the sample size required to achieve consistent results. To vary the data sets' degree of LD, we studied the LD patterns in several regions for which we have experimental data. We selected "high" and "low" LD regions, with criteria defined later in the results section, and adjust the parameters in our simulation program to mimic these patterns in our simulated data sets. Then we sample a certain number of individuals without replacement from "the population", and apply either SD or DIV to each sample. For both methods, we fix the relative redundancy threshold of each marker at 75% and sliding window size at 5, and test sample sizes of 10, 50, 100 and 200 individuals using the variation explained percentages 85% and 90% for SD, and 92% and 96% for DIV. We also apply "repeated runs" for each procedure until it meets the convergence criteria. For each parameter combination, we record, in a vector $v_1, \ldots v_{50}$, the percentage of times that each SNP, 1, 2, …50, is dropped in 100 non-overlapping samples from "the population". Note that v_i's close to 0 or 1 are preferred since this indicates that the SNP either gets dropped or kept each time in the simulations. The consistency for each marker is evaluated using the mean square error of its dropping

$$MSE_i = \frac{1}{100}\sum_{k=1}^{100}(y_{ik} - v_i)^2 = v_i(1 - v_i)$$

where y_{ik} is an indicator variable indicating whether i^{th} marker gets dropped in k^{th} sample, and v_i is the drop percentage for the i^{th} marker over 100 simulations. Then the overall consistency, the average MSE of the drop percentage for all markers, is calculated as

$$MSE = \frac{1}{50}\sum_{i=1}^{50} v_i(1 - v_i)$$

We also provide the same data in both haplotype phase-known and phase-unknown formats to study the effect of haplotype information on our results.

Validation Criteria

Our goal for marker selection is to identify a set of SNPs in a region that can retain a majority of the haplotype information available. We assume that the haplotype information is summarized by the number of frequent haplotypes and their haplotype frequencies. Therefore, we propose that we can evaluate the information retained about a region using these metrics. Unfortunately, haplotype phase is unobservable in most cases. We suggest using the EM algorithm with a sliding window procedure to infer the haplotype frequencies. Our validation procedure is the following: apply a sliding window with window size equal to 5 and estimate the haplotypes for 5 SNPs using the EM algorithm in each window. We chose a window size of 5 because calculating EM frequencies for 5 SNPs is computationally feasible and sample sizes of 50 and 100 individuals provide enough genotype information to get reasonable estimates. It is possible that unique situations may require other window sizes. Compute two measures to evaluate the information in the data: count the number of the frequent haplotypes and calculate the

heterozygosity for these haplotypes. Frequent haplotypes are defined as the haplotypes with frequencies greater than 5% and heterozygosity is calculated as $1-\Sigma p_i^2$. Then the selection procedure is applied with either SD or DIV to select "informative" SNPs. If the subset of SNPs selected can represent most of the information in the data, then in each window, we expect to observe the same number of frequent haplotypes and the same frequencies when we only use the selected SNPs to infer these. Therefore we only use the selected SNPs to estimate the haplotype frequencies within the previously defined windows, compute the above two measures again, and compute the differences of the two measures before and after selection. We define an acceptable difference before and after selection for haplotype heterozygosity as more than 90% of the windows having a heterozygosity difference of 0.1 or less. When repeated runs are used, we evaluate the final selected marker set against the initial full data set. (See discussion for details) We judge the performance of the procedures by looking at the differences along the chromosomal region, as well as their overall distributions.

Data Sets

Using linkage analysis, we identified a 12 cM region on chromosome 12 centered at D12S853 as likely to contain a susceptibility gene for type 2 diabetes (Ehm et al. 2000). 649 SNPs distributed across this region were genotyped on 138 unrelated Caucasian individuals. The SNPs have been placed on a 12 Mb composite map using a combination of STS content mapping and sequence analysis. 604 out of 649 SNPs have MAF greater than 5%.

To study linkage disequilibrium in the region surrounding the CYP2D6 gene on chromosome 22, 32 markers were typed with 27 having MAF greater than 5% on 1018 Caucasians. The markers

were identified from The SNP Consortium (TSC) (http://snp.cshl.org/) release 5 (Sept2000) and dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) public databases. The samples consisted of 230 samples originating from CEPH or Coriell cell repositories (Camden, New Jersey, US) and 788 originating from patients in GlaxoSmithKline (GSK) Clinical Pharmacology studies with consent for non-identified genotyping (428 from North America and 360 from the United Kingdom) (Hosking et al. 2002). All SNPs map to an 879 kb contig flanking the CYP2D6 locus. Hosking et al. (2002) reported the association study results of the above 27 SNPs with poor drug metabolizing phenotype using Fisher's exact test for single marker genotypic tests and a regression based model for haplotype tests using a sliding window approach with window size equal to 5. To illustrate the impact of the marker selection on the results of an association study, first, we reproduced Hosking et al's association study results using their data and the same association tests they used. We, then, applied the SD procedure to select markers using 100 randomly selected controls. Using only the selected markers and all 1018 individuals, we conducted Fisher's exact test for single marker genotypic tests, used only selected SNPs to estimate haplotype frequencies within the previously defined windows, and applied the regression based haplotype tests mentioned above with the estimated haplotype frequencies. We plot the test p-values versus the marker positions of the full data set and the selected set, and compare their patterns.

Results

We report the results for the two simulation studies described and study the behavior of the marker selection algorithms on several experimental data sets.

Simulation Study I

Table 1 shows the percentage of SNPs dropped when markers are in LE using SD with the percentage of variation explained set to 85% when the haplotype phase is unknown for sample sizes of 10 and 50. No markers were dropped in the simulations for sample sizes of 100 and 200 and therefore these percentages are not shown**.** We only show the table for a variation explained value of 85% because the pattern of dropping markers was similar for a variation explained parameter of 90%, only slightly lower. The results show that SD won't drop SNPs in LE under most parameter combinations unless the sample size is small relative to the window size. With a small sample size and relatively large window size, some independencies of SNPs will not be represented in the sample and SNPs will be dropped as a result. Note that the small percentages of SNP dropped are probably the result of using the L_r measure to determine if there is redundant information in the sample and dropping SNPs only if redundancy appears to exist. In conclusion, for sample sizes of 50 individuals or more, this procedure will retain important SNPs up to window sizes of 15.

Table 2 shows results for the DIV method when the percentage of variation explained is 92%. The results are similar when the percentage of variation explained is 96% only fewer markers are dropped. DIV has more of a tendency to drop SNPs in LE compared to SD. A similar pattern is observed that more SNPs are dropped when window size is large and sample size is relatively

small. The percentage of SNPs dropped is relatively stable when the sample size is greater than 50 and the relative redundancy threshold is greater than 70%. In choosing a sliding window size, we would like it to be large enough to include certain variation, but small enough that we don't require large sample sizes to capture the important variation. If we choose a window size equal to 5 and a relative redundancy threshold greater than 70%, then the percentage of SNP dropped is close to 5% even with a small sample size like 50. Based on these simulations, we have selected a sliding window size of 5 and the relative redundancy threshold of 75% for many of our further analyses. To make SD and DIV easier to compare, we choose to use the sliding window size equal to 5 and relative redundancy threshold to be 75% for SD also.

To explore these two procedures when haplotypes are known rather than inferred, we generated the same data with the haplotype phase information, repeated the same simulation study, and computed the average percentage of SNP dropped. We present the results for DIV since it relies more on the haplotype information. Table 3 shows the results for DIV with the percentage of variation explained 92%. The percentages of SNPs dropped are smaller when the haplotype phase is known because inferring this information using a mathematical algorithm such as EM results in an information loss. Nonetheless, the percentages show a similar pattern that, to control the percentages of SNPs dropped, relatively small window sizes are needed when the sample size is small. When comparing the results in tables 2 and 3, we observed that the differences in the percentages of SNPs dropped between inferred and known haplotype phase data are smaller with increases in the sample size. It suggests that, when the haplotype phase is known, it may be possible to apply the algorithms using a smaller sample size.

Table 4 shows the percentages of SNPs dropped for SNPs in LE for a range of variation explained percentages using either SD or DIV. The variation explained percentages should be at least 94% for DIV and 75% for SD when the sample size is 50 and 92% for DIV and 70% for SD when sample size is 100, if the false positive rate for the markers in LE is controlled to be less than 5%. We only propose these values as the procedure's "safe" starting point to avoid dropping important markers, when the degree of LD in the data is either hard to measure or complex. The variation explained percentage could be determined by testing a large range of values on the experimental data, and a value chosen when the haplotype information before and after selection is similar. However, the results might become too complicated to interpret when the real data consists of regions with differing degrees of LD, and "repeated runs" is applied. Therefore, using a variation explained parameter for which we know does not result in dropping too many SNPs under LE is simpler.

Simulation Study II

To study the sample sizes needed to obtain consistent selection results, we simulated data sets containing differing degrees of LD with similar patterns as we observed in our experimental data. We treated chromosome 22 data as one single region, and divided chromosome12 data roughly into 6 regions each containing about 110 markers to study their LD patterns. D' was calculated for each marker pair within each region, and averaged according to the number of intervening markers. The averaged-D' was plotted versus the number of intervening markers. As expected, for all regions, the average-D' decreases as the number of intervening markers increases (graphs are not shown.). Two regions, one with fastest and one with slowest decreasing LD were selected as "high" and "low" LD regions, respectively. The average-D' for

30 markers' apart decreases approximately to the half of its maximum, 1, for the high LD region. For the low LD region, the average-D' drops under 0.5 after 5 markers apart. Two data sets, each with 50 SNPs, were generated to produce similar LD patterns by using the simulation procedure we mentioned above adjusting the number of generations evolved and the recombination rate.

Figure 1 shows each SNP's average drop percentage across 100 simulations when SD is applied with the variation explained percentage 90% and the window size equal to 5 on the "high LD" data. Figure 2 shows each SNP's average drop percentage across 100 simulations when DIV is used with the variation explained percentage 92% and the window size equal to 5 on the "high LD" data. Note, SD or DIV could select different markers across different samples from the population because these markers are highly correlated and provide somewhat equivalent information, and which marker (markers) provides more information in each sample is affected by the statistical sampling variation. Therefore, we are expecting relative stable patterns, but not absolute 0% or 100% drop for each marker. In addition, the observation that SD and DIV drop different markers for the same data is expected for the same reason. Both Figures 1 and 2 suggest that sample sizes greater than or equal to 50 are needed to achieve consistent results. When the sample size is increased from 50 to 100 or from 100 to 200, the consistency improves a little. With higher variation explained values or a low degree of LD in the data, a similar pattern was obtain with even more consistent results (data not shown). Knowing haplotype phase information does improve the consistency, but its effect is small and can be compensated for with a slight increase of the variation explained percentage or the sample size. For both SD and DIV, almost the same number of markers and the same markers are dropped when the same parameters are used and the sample size is large enough (the sample size >= 50), regardless whether the

haplotype information is observable. According to simulation study II results, we recommend a sample size of 50-100 individuals depending on the expected amount of missing data.

Experimental data results

In applying the procedure to the experimental data, we also used a sliding window size equal to 5 and retained a SNP when its relative redundancy was less than 75%. We used "repeated runs" and utilized the convergence criterion as we mentioned above. Based on the results from table 4, we chose variation explained starting values of 70% for SD and 92% for DIV, respectively, since we use sample sizes close to 100 individuals for all experimental data. For validating the information retained, we used the sliding window size equal to 5 when calculating the number of frequent haplotypes and the haplotype heterozygosity. To obtain the haplotype structure information in the data before selecting markers, we included all the available SNPs even when their allele frequencies were less than 5% to achieve a relatively comprehensive picture of LD regardless of the SNP allele frequencies. We adjusted the variation-explained for each method until more than 90% of the windows had haplotype heterozygosity differences less than 0.1. We chose the haplotype heterozygosity because we found the number of frequent haplotypes to be unreliable in certain situations. For example, a haplotype can change its status from an infrequent haplotype to a frequent haplotype by a very small frequency change, such as from 4% to 5.5%, which results in a difference in the number of frequent haplotypes. These haplotype frequency changes contribute little to the difference in heterozygosity. For all experimental data sets, we presented results as long as one procedure (SD or DIV) selected markers to achieve this goal, and adjusted the variation-explained percentage for the other so that it retained a similar number of markers in order to make our comparisons fair.

For chromosome 12 data, the above procedure resulted in a variation-explained value of 90% for SD and 96% for DIV. Using these settings, we selected 415 (63.9%) using SD and 412 (63.5%) markers using DIV out of 649 markers, respectively. Histograms of the differences in the haplotype heterozygosity and number of frequent haplotypes for each window are shown in Figure 3. The proportion of windows with differences in the heterozygosity of less than 0.1 is 85.2% for the SD method, and 92.2% for the DIV method. The proportion of windows with differences in the number of frequent haplotypes of 1 or less was 93.4% for SD, and 95.3% for DIV.

For the chromosome 22 region, we have a much larger sample size than that required for the marker selection. We randomly selected 100 controls and applied our procedure as if this was the sample size collected for the marker selection purposes. Variation explained percentages 75% and 92% were used for SD and DIV, and 20 (62.5%) markers were selected in both instances. The proportion of windows with differences in heterozygosity of less than 0.1 was 92.9% and 88.9% for each method. We used relatively low variation explained percentages for each method than those for the chromosome 12 data, and achieved smaller differences before and after selection. The homogenous of the LD pattern in this data set might be one explanation.

We analyzed the selected markers (SD method) for association with the poor metabolizer phenotype. Note that because almost a half of the markers show strong association with the phenotype, it does not make sense to evaluate the procedure based on whether the selected marker set includes the significant signals. Therefore, we chose to compare the test p-values'

patterns using the full data and the selected data. Note that 20 of 27 markers were selected and the markers with the most significant genotypic tests were among those selected (data not shown). Figure 4 contains the haplotype tests results for the full marker set with a sliding window size of 5, and the selected marker set within the context of the windows defined by the full data. The two curves almost overlap, which is not surprising, since the selected markers preserved the information content of the full data well. Therefore marker selection had a negligible impact on the results of this association study.

Discussion

There are several fundamental differences in two methods, spD and div, we considered. Spectral Decomposition (spD) is less computational intensive and can be applied to analyze a larger number of SNPs, such as candidate genes typed for several dozen markers, without using sliding windows since it is based on a summary of pair-wise LD and does not require haplotype phase information. The haplotype diversity (div) is constrained by its computational limitations because it relies on the haplotype information that has to be estimated if unknown. The required computational time on estimating haplotype frequencies in unrelated individuals using a numerical algorithm such as EM increases dramatically as the number of markers increases. Furthermore, as we mentioned in the method section, the optimal htSNP set is found by an exhaustive search from all the possibilities. Therefore, div, by itself, is quite time-consuming when more than 7 SNPs are considered.

In addition to having different computational requirements, SD and DIV represent two different approaches to the problem of marker selection. SD is a procedure that relies on two locus LD (i.e. pair-wise correlation) and single marker characteristics whereas DIV relies on haplotype frequencies. The pair-wise correlation relies on two-locus LD measures and marker allele frequencies. Haplotype frequencies involve not only two-locus LD coefficients and allele frequencies, but also LD coefficients involving alleles at three or more markers (Bennett 1954). Haplotypes will provide more information than pair-wise LD measures if the LD measures involving three or more markers makes a significant contribution. Otherwise, haplotype frequencies are just linear combinations or summaries of pair-wise LD and allele frequencies. Ehm et al (unpublished data) summarized the decay and extent of two and three locus LD in

several genomic regions including the chromosome 12 and 22 regions described here. They found that LD based on alleles at three loci decays more quickly than two locus LD. The extent of three locus LD is relatively small across the chromosome 12 locus except for the central region, near 5000 or 6000 kb, where there is a large amount of three locus LD. Therefore it is reasonable to investigate marker selection procedures based on two locus LD and single marker characteristics.

We did not find major differences in the overall performance of the two procedures. A minor observation is that SD tends to drop markers with more equal or higher allele frequencies because it relies on *r*, which achieves higher values when marker allele frequencies are equal or relatively high. DIV tends to drop markers with disparate allele frequencies because it is based on the heterozygosity and markers with less frequent alleles contribute less to this measure. While we did not see significant differences in the performance of the methods when the results are evaluated using our validation criteria, this difference may explain why the methods do select different markers. Note that for both methods, we pre-select markers according to their allele frequencies, and retain SNPs only when their minor allele frequencies are greater than 5%. This may make the two procedures more comparable since SNPs with very disparate allele frequencies will not be retained. Furthermore, typing markers with frequencies less than 5% might be considered less efficient in the association study design.

Zhang et al (2002), proposed a dynamic programming method to perform haplotype block partitioning to minimize the SNPs needed to represent common haplotypes. This method can utilize different measures of the block quality including the ratio of the number of SNPs in the

block to the minimum number of SNPs required to define haplotypes or the proportion of haplotype diversity explained by a subset of the SNPs in a block (Clayton 2001). Similarly, another possible measure of the block quality could be the number of SNPs needed to achieve a pre-specified measure of the variation explained using the spD method. Note that Zhang et al 's goal is to define blocks with a minimum number of SNPs using any of a number of measures of block quality. Our goal is to select a subset of SNPs, which capture the common haplotypes within a region and preserve local haplotype frequencies within relatively short genetic distance. Thus, our objective is to maintain similar measures of haplotype information before and after marker selection. This leads us to the sliding window approach, which does not rely on the block definitions. Furthermore, our procedure is designed to be applicable to data when haplotype phase information is not available.

There are several approaches that could be considered in defining the convergence for "repeated runs". One choice is to run the procedure until no more markers are dropped. We found that this worked well for a homogeneous LD region, but when a region exhibits a mixed amount of LD, too many repeated runs can lead to dropping informative markers and information loss. One possible reason is that markers in LE get dropped if the procedure is applied to data with less and less correlation. Therefore, for a complex chromosome region, we suggested the 5% cutoff as described above. One might consider whether it makes sense to use repeated runs at all. We found that using a higher variation explained percentage value combined with "repeated runs" was preferable (less information loss) to one with a lower variation explained value with no repeated runs.

In evaluating our procedures, we used the heterozygosity and the number of the frequent haplotypes to measure the haplotype information content. We felt that these measures captured aspects of haplotypes important in association studies. Any other suitable measures, such as matching haplotype frequencies before and after selection, could be used. Furthermore, the validation procedure can vary according to the different requirements of the studies. We propose to use a fixed variation explained percentage, run the procedures repeatedly until they converge, and evaluate the information content of the selected marker set against that of its initial full data within the initially defined sliding windows. One advantageous of this approach is that it provides an overall evaluation for all repeated runs. However, it might be too conservative since markers are selected based on their LD structure in the sliding windows and some of the windows may loss their initial meanings after the repeated marker reductions, which result in some large differences in the haplotype information measures between before and after selection that don't necessarily indicate serious information loss. An alternative approach would be to evaluate the information content of the selected marker set against that of its immediate input data for each repeated run. There are two advantages for this approach: first, all selected markers are evaluated in the windows they are selected. Second, the variation explained value could be adjusted to control the information loss for each repeated run. However, it is not clear how we summarize the overall performance of such a procedure.

In this paper, the selection procedures are applied to the markers discovered and typed on population controls: samples chosen regardless of their phenotype. We have found it useful to ensure that SNPs are polymorphic in the ethnic group of interest before typing them on precious disease samples. Currently we are typing a large number of SNPs on a panel of 100 population

samples and using these data in the marker selection. Since we are studying several diseases, using a population sample allows us to use the marker selected in further studies regardless of the disease. However, this approach does assume the common disease, common allele hypothesis. If a rare allele at a marker is responsible for the disease, then it is unlikely to be selected in such an approach. Another choice would be to apply marker selection to markers discovered in case-control samples and typed on cases only. This way, the disease alleles, although rare in the population, will have increased frequency in the sample and the selection favoring polymorphic markers would have a good justification. When case-control samples are available, procedures similar to the ones described here can be useful in selecting most discriminative subsets of markers among those that show frequency differences between cases and controls. Analogously, the marker selection procedure might need to be conducted separately using different samples from different populations if we wish to study different populations possibly having somewhat different haplotype structures.

We do realize that procedures such of these have consequences. As with any statistical procedure, the marker selection is always a gamble, since markers are selected mainly based on LD structure regardless of any phenotypes. Therefore the impact of selecting markers on the results of an association study, in general, is not known. Although we have showed one example where marker selection had a negligible effect on the association results, the impact can vary greatly from case to case. For instance, the required percentage of information retained in the association study might depend on effects of disease susceptibility genes, which are hard to access before actually conducting the association test. Furthermore, it is possible that a causal marker is not selected because it is highly correlated with nearby markers. Fortunately, it appears that analyzing the data using haplotype analysis would reduce the impact of such a

selection. In summary, the marker selection should be viewed as providing a way to prioritize markers for a first genotyping screen and more markers can always be typed in the regions of interest later.

Acknowledgements

We like to thank Bruce Weir and Mike Boehnke for their helpful comments on the manuscripts, Eric Lai, Clive Bowman and Mike Mosteller for their valuable advice, and Achamma Philip for her technique support.

Appendix

Appendix 1. Obtaining the matrix of pairwise LD for bi-allelic markers

Calculation of matrices of pairwise LD is most straightforward for markers with two alleles and can be handled with a simple command using standard statistical software (e.g. by invoking cor() function in R/Splus). The following method requires that marker genotypes are recoded as follows:

$$\text{New value} = \begin{cases} -1 & \text{if genotype is "11"} \\ 0 & \text{if genotype is "12"} \\ 1 & \text{if genotype is "22"} \end{cases}$$

A pair of SNPs will be represented by two vectors x and y with entries as just indicated. It can be easily shown that the usual sample covariance

$$C_{AB}(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum x_i y_i - \frac{1}{n^2}\sum x_i \sum y_i$$

is twice the composite linkage disequilibrium, Δ_{AB}, of Weir, 1996. To see this, the terms of the covariance can be written in terms of di-locus counts as,

$$\sum x_i y_i = n_{AABB} - n_{AAbb} - n_{aaBB} + n_{aabb}$$

$$\sum x_i \sum y_i = (n_{aa} - n_{AA})(n_{bb} - n_{BB})$$

Then $C(\mathbf{x},\mathbf{y}) = 2\Delta_{AB}$ follows from the relation

$$\Delta_{AB} = \frac{1}{4}(\Delta_{AB} + \Delta_{ab} + \Delta_{Ab} + \Delta_{aB})$$
$$= \frac{1}{2n}(n_{AABB} - n_{AAbb} - n_{aaBB} + n_{aabb}) - \frac{1}{2n^2}(n_{aa} - n_{AA})(n_{bb} - n_{BB})$$

These composite coefficients are unbiased estimates of the population LD under HWE. When HWE does not hold, they include an additional component measuring covariance between alleles

between different haplotypes in an individual. The diagonal elements $\mathbf{C}_A(\mathbf{x},\mathbf{x})$ of the variance-covariance matrix $\mathbf{C}(\mathbf{x},\mathbf{x})$, are variances of allele frequencies, $\mathrm{Var}(p_A) = (p_A(1-p_A) + D_A)/2n$ where p_A is the allele frequency of allele A, $D_A = P_{AA} - p_A$ is the deviation from HWE, and P_{AA} is the frequency of genotype AA. In terms of recoded values the allele frequencies are

$$\tilde{p}_A = \frac{1}{2} - \frac{\sum x_i}{2n}$$
$$\tilde{p}_B = \frac{1}{2} - \frac{\sum y_i}{2n}$$

Finally, the correlation of Weir (1996) defined as

$$r_{AB} = \frac{\hat{\Delta}_{AB}}{\sqrt{(\tilde{p}_A(1-\tilde{p}_A) + \hat{D}_A)(\tilde{p}_B(1-\tilde{p}_B) + \hat{D}_B)}}$$

can be computed from recoded values as

$$r_{AB} = \frac{C(\mathbf{x},\mathbf{y})}{\sqrt{C(\mathbf{x},\mathbf{x})C(\mathbf{y},\mathbf{y})}}$$

or by an R/Splus function call, cor().

Appendix 2. Determining effective/redundant numbers of markers, L_e, L_r.

Let L be the actual number of markers and $\{\lambda_i\}$ is a set of eigenvalues associated with the matrix of pairwise LD coefficients. From Cauchy-Schwarz inequality and noting that $\{\lambda_i\}$ are non-negative we have $\sum \lambda_i^2 \geq \left(\sum \lambda_i\right)^2 / L$. This bound corresponds to the zero LD situation, when $\lambda_i = \lambda_j, \forall i, j$. In this case $\sum \lambda_i^2 = \sum \lambda_i \frac{\sum \lambda_i}{L} = \left(\sum \lambda_i\right)^2 / L$. From expanding $\left(\sum \lambda_i\right)^2$ we also see that $\sum \lambda_i^2 \leq \left(\sum \lambda_i\right)^2$. This bound corresponds to the maximum possible LD with a

single non-zero eigenvalue. In this case $\sum \lambda_i^2 = \left(\sum \lambda_i\right)^2$. Putting together, $\left(\sum \lambda_i\right)^2 / L \le \sum \lambda_i^2 \le \left(\sum \lambda_i\right)^2$. Then we have

$$0 \le L\frac{\sum \lambda_I^2}{\left(\sum \lambda_i\right)^2} - 1 \le L - 1$$

So that the number of redundant markers can be defined as

$$L_r = L\frac{\sum \lambda_i^2}{\left(\sum \lambda_i\right)^2} - 1$$

and the effective number of markers, $1 \le L_e \le L$, reduced due to LD, is

$$L_e = 1 + L\left(1 - \frac{\sum \lambda_i^2}{\left(\sum \lambda_i\right)^2}\right)$$

Reference

Bennett J (1954) On the theory of random mating. Annals of Eugenics 18(4): 331-317.

Clayton D (2001) Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. ftp-gene.cimr.cam.ac.uk/software.

Couzin J (2002) New mapping project splits the community. Science 296: 1391-1393.

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

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J et al. (2002). A first-generation linkage disequilibrium map of human chromosome 22. Nature 418: 544 - 548.

Ehm MG, Karnoub MC, Sakul H, Gottschalk K, Holt DC, Weber JL, Vaske D, Briley D, Briley L, Kopf J, McMillen P, Nguyen Q, Reisman M, Lai EH, Joslyn G, Shepherd NS, Bell C, Wagner MJ, Burns DK (2000). Genome-wide Search for Type 2 Diabetes Susceptibility Genes in Four American Populations. Am J Hum Genet 66: 1871-1881.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002). The structure of haplotype blocks in the human genome. Science 296: 2225-2229.

Hosking LK, Boyd PR, Xu CF, Nissum M, Cantone K, Purvis IJ, Khakhar R, Barnes MR, Liberwirth U, Hagen-Mann K, Ehm MG, Riley JH (2002). Linkage disequilibrium mapping identifies a 390kb region flanking CYP2D6 associated with CYP2D6 poor drug metabolising activity. Pharmacogenetics Journal 2: 165-175.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001). Haplotype tagging for the identification of common disease genes. Nat Genet 29: 233-237.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001). Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. Science 294: 1719-1723.

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516-1517

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, et. al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933

Subrahmanyan L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA (2001) Sequence variation and linkage disequilibrium in the human T-cell receptor (TCRB) locus. Am J Hum Genet 69: 381-395.

Weir BS (1996). Genetic Data Analysis II, Sinauer Associates, Inc

Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. PNAS 99(11): 7335-7339.

Table 3.1 Percentage of SNPs dropped in LE using SD with variation explained 85% haplotype phase-unknown

Percentage of SNPs Dropped		Sample Size					
		10			50		
		Relative redundancy			Relative redundancy		
		50%	70%	90%	50%	70%	90%
Sliding Window Size	2	0	0	0	0	0	0
	5	7	2	1	0	0	0
	10	45	21	6	0	0	0
	15	70	41	12	0	0	0
	30	91	74	28	36	15	5

Note: Results averaged over 100 simulations.

Table 3.2 Percentage of SNPs dropped in LE using DIV with variation explained 92% haplotype phase-unknown

Percentage of SNPs Dropped		Sample Size					
		10			50		
		Relative redundancy			Relative redundancy		
		50%	70%	90%	50%	70%	90%
Sliding Window Size	2	3	<1	<1	<1	0	0
	3	6	2	2	<1	0	0
	5	24	14	9	10	6	4
	7	39	28	15	21	16	8
Percentage of Drop (100 simulations)		Sample Size					
		100			200		
		Relative redundancy			Relative redundancy		
		50%	70%	90%	50%	70%	90%
Sliding Window Size	2	0	0	0	0	0	0
	3	<1	0	0	0	0	0
	5	7	5	3	5	4	2
	7	15	11	6	13	10	6

Note: Results averaged over 100 simulations.

Table 3.3 Percentage of SNPs dropped in LE using DIV with variation explained 92% haplotype phase-known

Percentage of SNPs Dropped		Sample Size					
		10			50		
		Relative redundancy			Relative redundancy		
		50%	70%	90%	50%	70%	90%
Sliding Window Size	2	1	<1	<1	<1	0	0
	3	2	<1	<1	0	0	0
	5	8	5	2	4	2	2
	7	13	8	3	10	7	4
Percentage of Drop (100 simulations)		Sample Size					
		100			200		
		Relative redundancy			Relative redundancy		
		50%	70%	90%	50%	70%	90%
Sliding Window Size	2	<1	0	0	0	0	0
	3	0	0	0	0	0	0
	5	4	3	2	4	2	2
	7	9	7	5	9	7	5

Note: Results averaged over 100 simulations.

Table 3.4 Percentage of SNPs dropped in LE using different variation explained values

Percentage of SNPs Drop		Sample Size			
		50		100	
		DIV	SD	DIV	SD
Percentage of the Variation Explained	65%	/	12.6	/	11.8
	70%	/	8.9	/	4.2
	75%	/	3.9	/	3.4
	80%	19.7	3.4	17.0	3.4
	85%	14.7	3.2	13.3	1.7
	90%	10.0	0.1	7.9	0
	92%	6.5	/	4.8	/
	94%	3.4	/	1.7	/
	95%	/	0	/	0
	96%	1.0	/	0.3	/
	98%	0	/	0	/

Note: "/" means the simulation is not done for this value. Results averaged over 100 simulations.

Figure Legends

Figure 3.1: 50 SNPs' average drop percentage across 100 simulations on the "high LD" data when haplotype phase is unknown. SD is used with variation explained 90% and window size 5. The graphs from top to bottom are for the sample sizes equal to 10, 50, 100 and 200. The average number of SNPs dropped is 25.5, 19.7, 20.1 and 20.2 and the average MSE of dropping is 0.17, 0.11, 0.09 and 0.08 for each graph.

Figure 3.2: 50 SNPs' average drop percentage across 100 simulations on the "high LD" data when haplotype phase is unknown. DIV is used with variation explained 92% and window size 5. The graphs from top to bottom are for the sample sizes equal to 10, 50, 100 and 200. The average number of SNPs dropped is 26.2, 20.4, 20.0 and 19.4 and the average MSE of dropping is 0.16, 0.09, 0.07 and 0.06 for each graph.

Figure 3.3: Apply SD and DIV on on the chromosome 12 region with 649 SNPs. The graphs on the top and bottom are the overall evaluation for the procedure with SD method using variation explained 90% and DIV using variation explained 96%, respectively. The histograms on the left and right are for differences in the number of frequent haplotypes and differences in heterozygosity before and after marker selection. For both methods, sliding window size of 5 was used and the procedure converged at the third run.

Figure 3.4: Association between CYP2D6 PM phenotype and haplotypes. Haplotypes were derived by EM algorithm from windows of consecutive SNPs. Characters indicate the first

markers in window. "◊" represent haplotype test results of 27 markers with window size 5. "▲" represent haplotype test results of 20 selected markers with placing markers in previously defined windows.

Figure 1

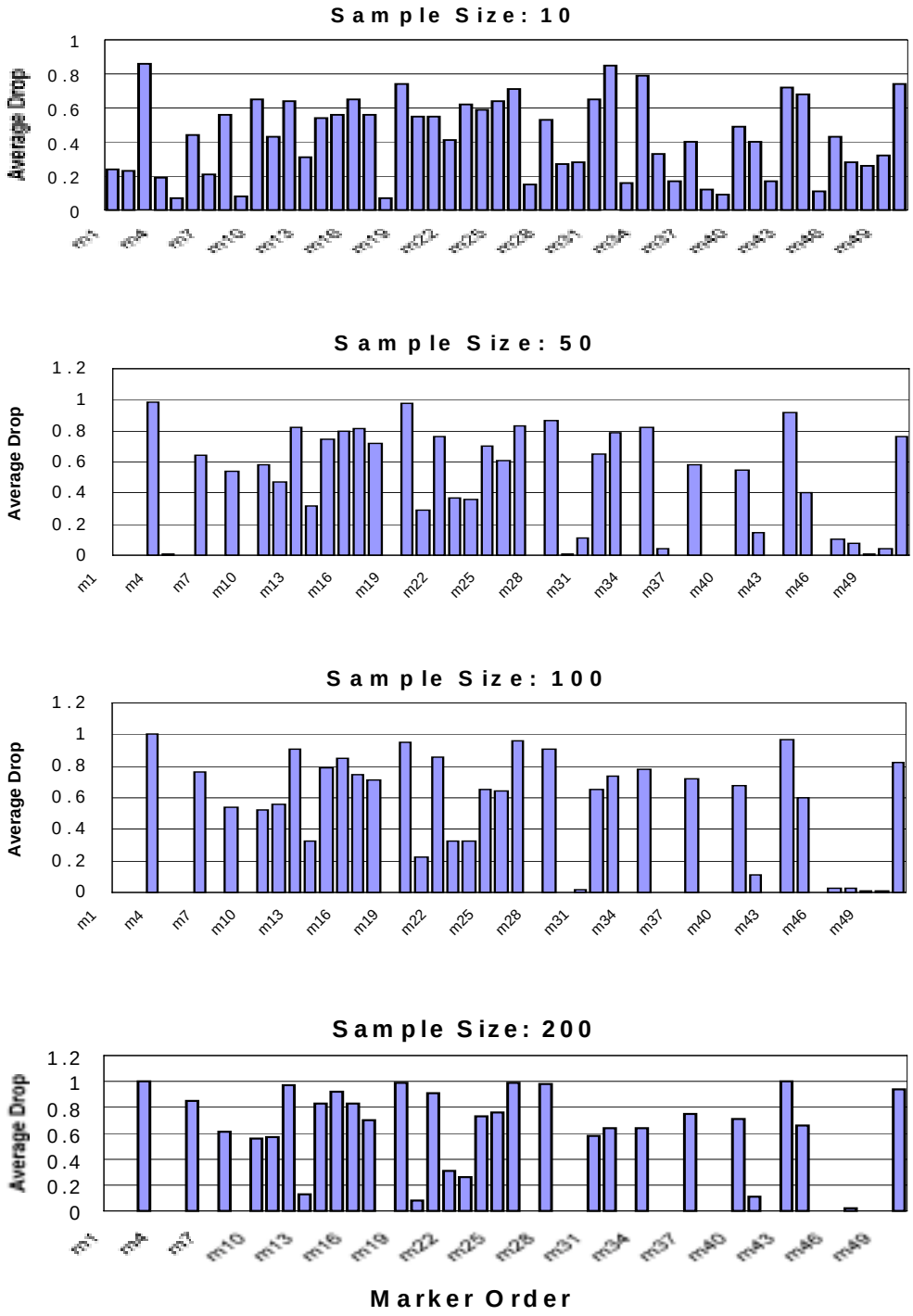

Sample Size: 10
Average Drop
1
0.8
0.6
0.4
0.2
0
m1 m4 m7 m10 m13 m16 m19 m22 m25 m28 m31 m34 m37 m40 m43 m46 m49
Sample Size: 50
Average Drop
1.2
1
0.8
0.6
0.4
0.2
0
m1 m4 m7 m10 m13 m16 m19 m22 m25 m28 m31 m34 m37 m40 m43 m46 m49
Sample Size: 100
Average Drop
1.2
1
0.8
0.6
0.4
0.2
0
m1 m4 m7 m10 m13 m16 m19 m22 m25 m28 m31 m34 m37 m40 m43 m46 m49
Sample Size: 200
Average Drop
1.2
1
0.8
0.6
0.4
0.2
0
m1 m4 m7 m10 m13 m16 m19 m22 m25 m28 m31 m34 m37 m40 m43 m46 m49
Marker Order

Figure 2

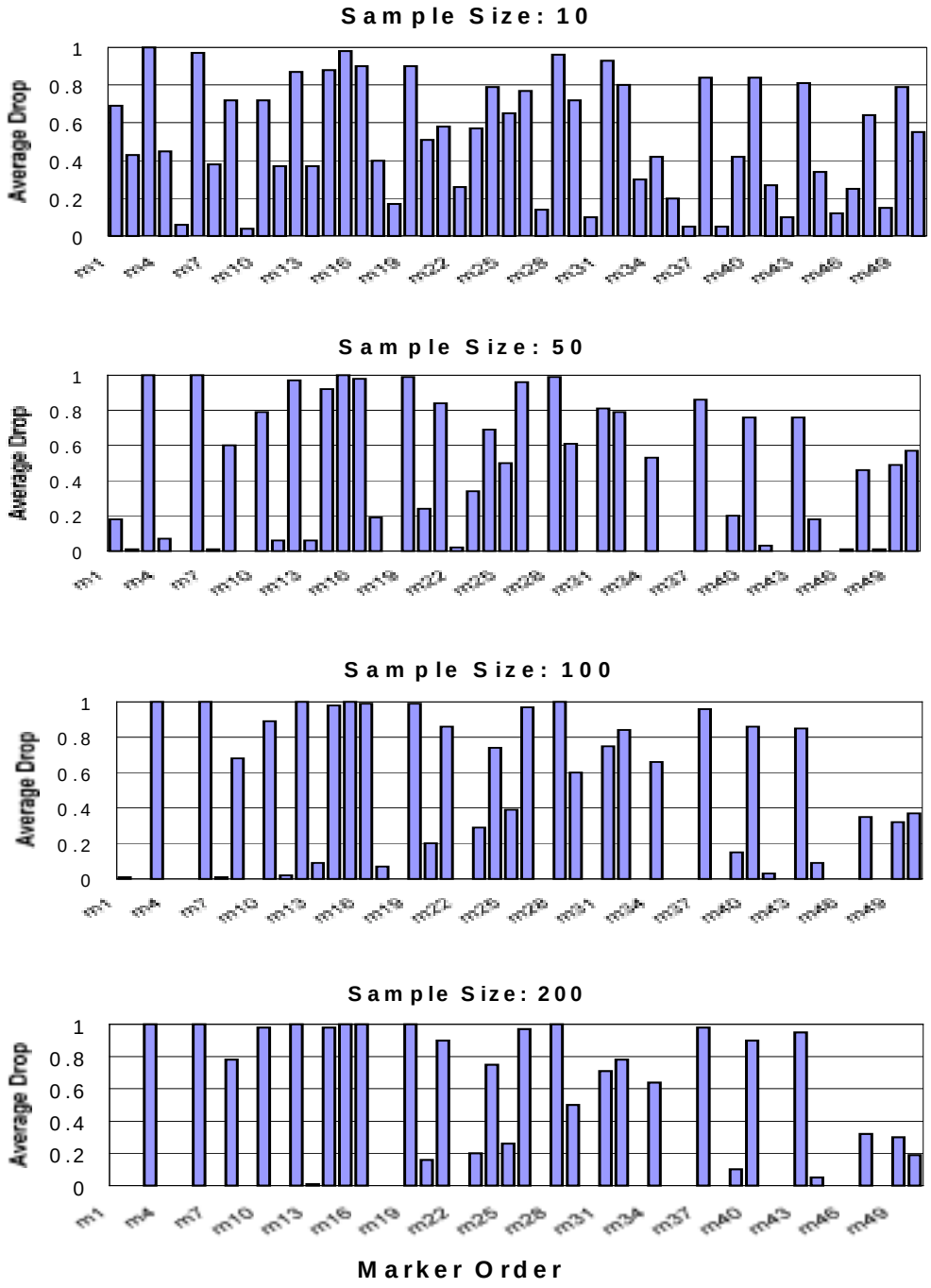

Sample Size: 10
Sample Size: 50
Sample Size: 100
Sample Size: 200
Average Drop
0
0.2
0.4
0.6
0.8
1
m1
m4
m7
m10
m13
m16
m19
m22
m25
m28
m31
m34
m37
m40
m43
m46
m49
Marker Order

Figure 3

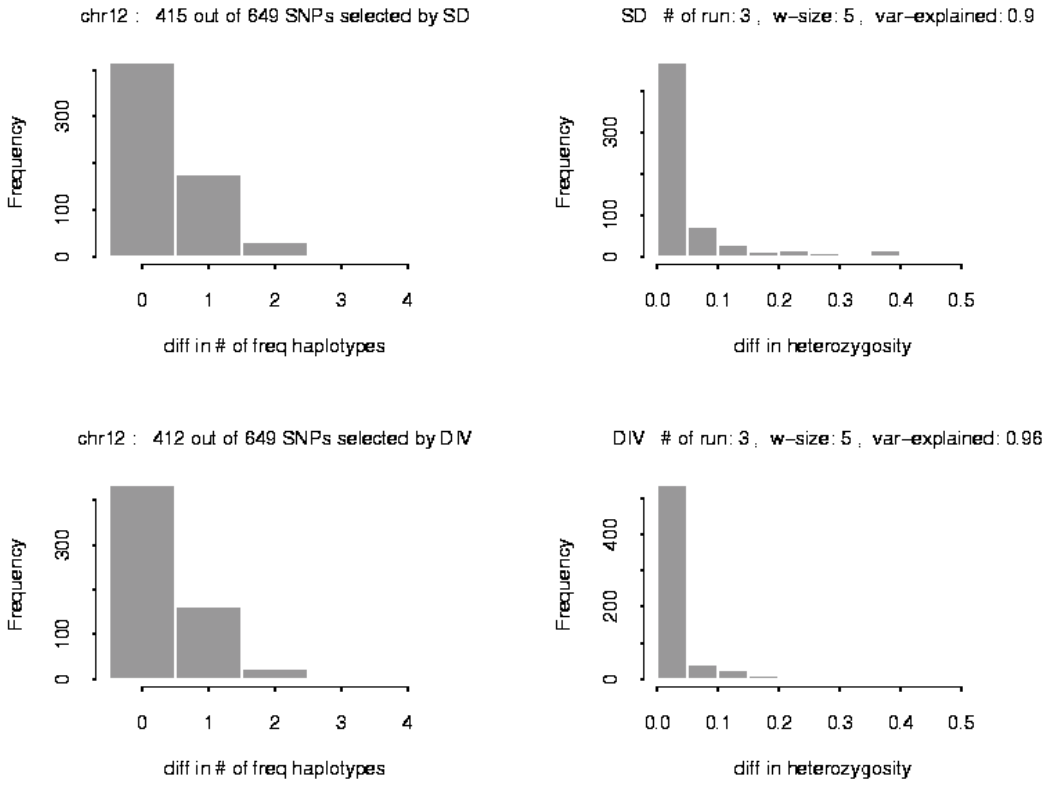
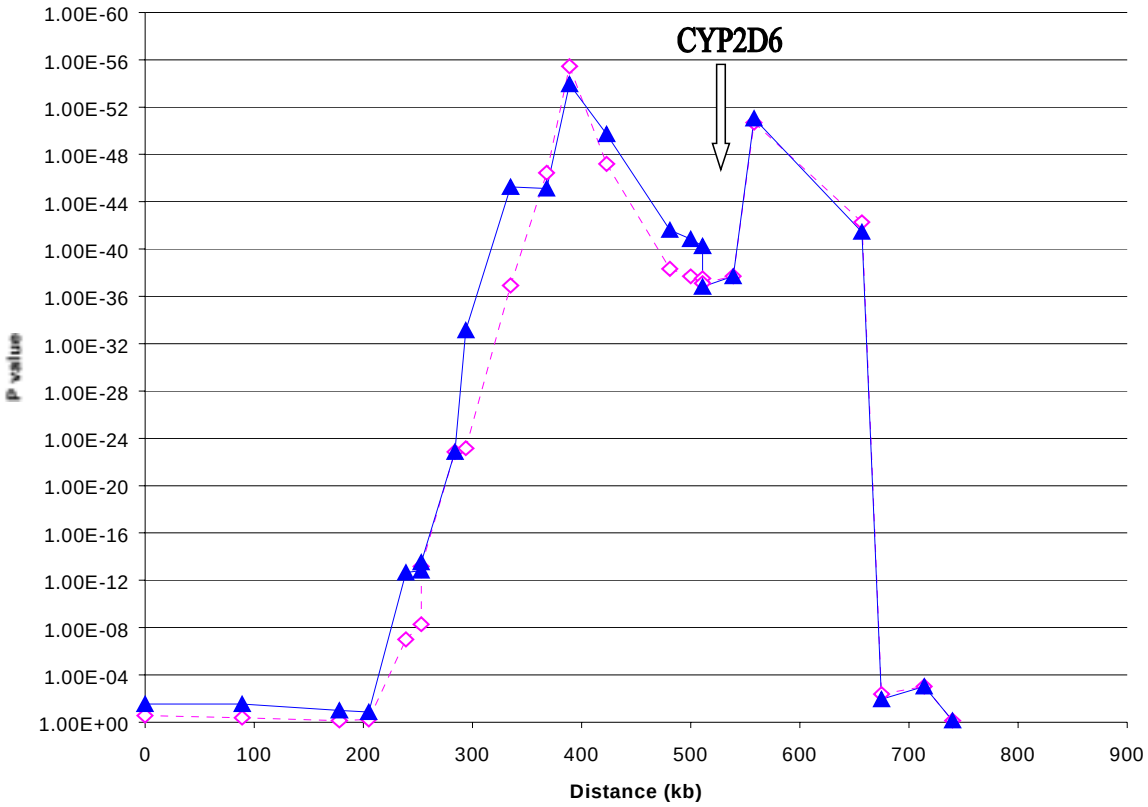

chr12 : 415 out of 649 SNPs selected by SD
Frequency
0
100
300
0
1
2
3
4
diff in # of freq haplotypes
SD # of run: 3 , w-size: 5 , var-explained: 0.9
Frequency
0
100
300
0.0
0.1
0.2
0.3
0.4
0.5
diff in heterozygosity
chr12 : 412 out of 649 SNPs selected by DIV
Frequency
0
100
300
0
1
2
3
4
diff in # of freq haplotypes
DIV # of run: 3 , w-size: 5 , var-explained: 0.96
Frequency
0
200
400
0.0
0.1
0.2
0.3
0.4
0.5
diff in heterozygosity

Figure 4.


CYP2D6
P value
1.00E-60
1.00E-56
1.00E-52
1.00E-48
1.00E-44
1.00E-40
1.00E-36
1.00E-32
1.00E-28
1.00E-24
1.00E-20
1.00E-16
1.00E-12
1.00E-08
1.00E-04
1.00E+00
0
100
200
300
400
500
600
700
800
900
Distance (kb)

Chapter 4

A Random Effect Model for Quantitative Trait and Haplotypes Association Test Considering Treatments and Gene×treatment Interactions

Abstract

To study how genes are related to the efficacy and safety of a medicine, multiple markers are genotyped in candidate genes on samples collected in clinical trials. The development of analysis methods that take into account the nature of genetic data is critical to extracting the maximum information. Currently, a linear regression approach, where genetic effects are treated as fixed effects, is widely used to model the association between a quantitative trait and genetic markers. However, in an exploratory phase of a study, the genetic variations under test are usually not specific functional sites but randomly selected polymorphisms within a gene. It is more appropriate statistically to treat these effects as random, especially when the inference scope is regard to the whole candidate gene but not to the sites under test. To illustrate the randomness introduced by these non-specific polymorphisms, we derive the relation between the variance components of a causal quantitative trait locus (QTL) and a randomly selected locus in a nearby region. Furthermore, a mixed effect model is developed to fit fixed treatment effects, random haplotypic effects, and random gene×treatment interactions in this scenario; likelihood ratio tests are applied for testing the random effects. We illustrate and compare our random genetic effect model (Model III) to a fixed genetic effect model with single SNP genotypic effects (Model I) or haplotypic effects (Model II) with simulated data under a variety of conditions. Our simulation results showed that Model III is more powerful than or equivalent to Model II under most of the circumstances. Model I is always more powerful than or equivalent to the other two models under the recessive genetic model. When the genetic model is dominance or additive, in testing genetic effects, Model I is more powerful than or equivalent to Model II and III when the QTL is caused by a single mutation and included in the data. Model II and III are more powerful when the opposite is true. When the genetic model is dominance or additive, in testing gene-drug

interactions, Model III is more powerful than or equivalent to Model I and II regardless the type of the QTL and whether the QTL is included in the data. Therefore, our mixed effect model generally behaves equivalently or better than the fixed haplotypic effects model, and is more appropriate in the exploratory phase of a study.

Introduction

Individuals respond to medicines in unique ways. Despite extensive research efforts into developing safer and more effective medicines and strict FDA guidelines, adverse drug events are still estimated to cost US $100 billion and over 100,000 deaths per year (Marshall, 1997). Furthermore, response rates for medicines vary widely. Variation in response could be due to many factors including drug-drug interactions, mis-dose, drug allergies, and medication error (Spear et al. 2001). However, a large proportion of the variation is known to be due to genetic variation among individuals. Variation in several genes has been associated with response to medicines including CYP2C9 (Aithal et al. 1999), CYP2D6 (Raimundo et al. 2000) and HLA-B region (Hetherington 2002). Some of the findings have been utilized in increasing drug discovery efficiencies (Gao 2002) and will eventually help in personalizing medicines and decreasing drug adverse events with the development of high throughput genotyping techniques. Methods to investigate the relationship between drug efficacy and adverse events and genetic variation are in their infancy. Therefore, identifying methods for investigating genetic effects and gene×treatment interactions that may be associated with drug efficacy or safety is still a task of high priority for clinical trial studies.

The rapid discovery of single nucleotide polymorphisms (SNPs) and the dense SNP map are terrific tools necessary for identifying variation associated with drug efficacy and safety (Sachidanandam et al 2001). Multiple markers are genotyped in candidate genes or regions on unrelated samples collected in clinical trials. Association studies are widely applied because they utilize dense marker information efficiently, can be applied to unrelated individuals, provide more power in detecting susceptibly genes with small effects than do traditional linkage studies

(Risch and Merikangas 1996), and have the ability to narrow down the candidate regions to more experimentally manageable sizes. Association study designs that relate single SNP markers and haplotypes formed by multiple SNPs in regions of interest are the two main investigative approaches. While a single SNP could be made into a diagnostic product more easily than several SNPs, recent work has shown that the grouping and interactions of markers in haplotypes might play an important role as well. (Drysdale et al. 2000; Hoehe et al. 2000; Davidson 2000; Morris and Kaplan 2002) One rationale is that some complex diseases are not caused by a single mutation but the combination of several mutations (detected by haplotypes). Therefore the haplotype approach will detect more significant evidence for these kinds of epistatic effects. Also, the haplotype of the original carrier provides a good surrogate for the disease susceptibility gene since it tends to remain intact if recombination around the disease-causing mutation is rare. Finally, haplotype analysis has an advantage as it uses linkage disequilibrium (LD) between markers to capture more information in the regions than do single SNPs. Even with current typing techniques, it is not possible to type every SNP in candidate genes, and therefore need to rely on LD between markers and disease locus to detect the disease locus effects. Even though actual haplotypes can be determined experimentally using techniques such as atomic imaging microscopy (Woolley et al. 2000) or somatic cell hybrid (Douglas et al. 2001) approaches, these type of data is not routinely available. Therefore, the loss of the haplotype phase information might result in a loss of the statistical testing power. However, haplotype phases can be reasonable inferred from genotypes computationally (Clark 1990; Excoffier and Slatkin 1995; Long et al. 1995) in many cases.

Continuous human traits such as blood pressure or drug efficacy and binary traits such as the presence or absence of disease have been investigated for relationships with genetic variation in association studies. Quantitative response traits are likely to provide more power for identifying genetic variation associated with drug response. A linear regression setting is preferred because it utilizes quantitative trait information and other covariates, treatment effects in our case, more easily. For binary response traits, non-identity links (e.g., logistic) can be applied. In studying the association of a quantitative trait and genetic variation in a regression setting, the following two models are frequently used (Zaykin et al. 2002). The first is a linear model that associates a quantitative trait with genotypic effects, where the genotypic effects are treated as fixed factors with discrete levels. If multiple SNPs (*S* SNPs) at a locus are tested simultaneously, the number of possible genotypes, $2^S(2^S+1)/2$, increases dramatically with the number of alleles, 2^S. The power of the genotypic test will decrease with increasing degrees of freedom. Zaykin et al. (2002) proposed a linear model that split genotypic effects into allelic additive effects (haplotypic additive effects) and dominance effects, omitted the dominance effects from the model, associated the response trait with the haplotypic additive effects, and utilized an F-test in testing the significance of the haplotypic effects. Again, all genetic effects are treated as fixed factors with discrete levels. This second model might increase the testing power by decreasing the test degrees of freedom from $2^S(2^S+1)/2$ to 2^S, but it might also display a lack of fit because the dominance effects are omitted. Which model is more powerful will vary case by case. We extend the above two models to include treatments and gene×treatment interactions, and call them Model I and Model II, respectively. Model I includes a quantitative trait, treatment effects, genotypic effects, and genotype treatment interactions, where all effects are treated as fixed factors with discrete levels. Model II includes a quantitative trait, treatment effects, haplotype

additive effects, and haplotype treatment interactions, where all effects are treated as fixed factors with discrete levels. In this paper, we are mainly interested in comparing two types of the genotypic effects: the effects of a single SNP and the additive haplotype effects, where the haplotypes are formed by multiple SNPs within a gene.

When the genetic variation under testing is not from specific functional sites but rather from randomly selected markers within the genes, it might not be appropriate to treat the genetic effects of the markers as fixed effects in the model. When the functional sites of candidate genes are genotyped and tested for association, the question is whether these specific sites have significant effects on the quantitative trait under study. Therefore, it is reasonable to model the effects of these functional sites as fixed effects. However, when markers with no prior knowledge of function are selected randomly, as a result of either the availability of SNPs in databases or the SNP discovery process based on a limited number of individuals, uncertainty is introduced in the study. First, most of these markers don't directly cause the quantitative trait variation, but are associated with it through the functional sites in linkage disequilibrium with these SNPs under study. We show that the variance components (both additive and dominance) of a marker locus can be related to the variance components of the functional trait locus through genetic effects of trait alleles, LD between trait locus and markers, and their allele frequencies. Therefore, a statistical model that takes this uncertainty and randomness into account is preferred. Secondly, because of the limitations of both statistical and genetic sampling, the markers genotyped in the data represent only a portion of the total variation in the population and the specific population under study. Thus, based on our understanding of population genetics, it is desirable to model genetic effects as random, instead of fixed effects. Therefore, we adopt the

linear regression model setting of Model II to treat the haplotype additive effects and the haplotype gene×treatment interactions as random, and utilize likelihood ratio statistics to test the significance of the random effects. Furthermore, the scope of statistical inference based on our proposed model is extended to all markers in LD with the tested markers in the candidate gene including the "hidden" functional sites rather than for only the tested markers. Then if markers within a candidate gene show association, additional markers can be tested to identify the functional polymorphism.

Since the scope of our hypotheses is wider than that of the fixed effect models, the power comparison of the two types of models is in favor of the fixed effect model if the two methods are equally powerful. We wish to investigate the differences of the above three models under several different circumstances, such as different genetic models, different allele frequencies of the quantitative trait locus (QTL), different types of mutations causing the QTL, and different magnitudes of the gene×treatment interactions, etc. A random effect model is more appropriate if it won't loss much power to a fixed effect model in an exploratory stage of a study. In this paper, we concentrate on the haplotype phase-known situation, but illustrate that our proposed model can be easily extended to the haplotype phase-unknown cases by inferring haplotype frequencies using the Expectation-Maximization algorithm and fitting the inferred haplotype frequencies as "weights" in the model. .

Method

We briefly introduce Models I and II, derive the relationship between the variance components of the quantitative trait locus and a random SNP in LD with it, and propose our random effect model. Furthermore, the test statistics used in testing treatment effects, genetic effects and gene×treatment interactions in each model are described. Simulation studies are conducted to study the powers of detecting treatment effects, genetic effects, and gene×treatment interactions using different models under several different scenarios.

A brief introduction to Model I and II

Model I:

Let Y_{tgk} is the quantitative trait value for the k^{th} individual with g^{th} genotype and taking t^{th} treatment. We model

$$Y_{tgk} = \mu + \tau_t + G_g + (\tau G)_{tg} + \varepsilon_{tgk} \quad (1)$$

where, μ is the overall mean, τ_t is the t^{th} treatment effect, G_g is the genetic effect of g^{th} genotype at this locus, $(\tau G)_{tg}$ is the genotype gene×treatment interaction of the g^{th} genotype with the t^{th} treatment, and ε_{tgk} is the random error. All effects, τ_t $(t = 1..T,$ for T treatments), G_g $(g = 1..G,$ for G genotypes), and $(\tau G)_{tg}$, are regarded as fixed effects with discrete levels except ε_{tgk} iid $\sim N(0, \sigma_e^2)$. g can be viewed as a mapping function, $g()$, that an dependent variable, any genotype g, is uniquely determined by a independent variable, an individual k.

Weir and Cockerham (1977) and Nielsen and Weir (1999) decomposed the genotypic effects into average effects of individual alleles and interaction effects between alleles. Then, the response variables can be explained by the allelic additive "main effects" and their interaction, the dominance effects. The alleles inherited from two parents are assumed to be independent if the parents are not related. Therefore, each allele is an independent observation and the allele-based test is a valid test and maintains the proper distribution under the null hypothesis. To apply to Model I, we split the genotypic effect G_g into two allelic additive effects and the dominance effect, and the genotype×treatment interaction $(\tau G)_{tg}$ into the interactions between the treatments and two alleles and the interactions between the treatments and the dominance effects. If Y_{tijk} is the quantitative trait value for the k^{th} individual with genotype containing alleles i and j and taking t^{th} treatment, then

$$Y_{tijk} = \mu + \tau_t + \alpha_i^m + \alpha_j^p + d_{ij} + (\tau\alpha)_{ti}^m + (\tau\alpha)_{tj}^p + (\tau d)_{tij} + \varepsilon_{tijk} \quad (2)$$

where μ is the overall mean, τ_t is the t^{th} treatment effect, α_i^m and α_j^p are the allelic effects of the maternal and paternal alleles i and j, respectively, d_{ij} is the dominance effect of alleles i and j, the deviation from their additivity, $(\tau\alpha)_{ti}^m$ and $(\tau\alpha)_{tj}^p$ are the treatment by allelic interactions, $(\tau d)_{tij}$ is the treatment by dominance interaction, and ε_{tijk} is a random error. Again, all effects are regarded as fixed effects with discrete levels except ε_{tijk} iid ~ $N(0, \sigma_e^2)$. Note that α_i^m and α_j^p take the same number of levels, H the total number of alleles (haplotypes). For example, there are two markers: one has alleles A and a; the other has alleles B and b. The possible formed haplotypes are *"AB", "Ab", "aB", "ab*". Then, α_i^m and α_j^p can both be α_{AB}, and α_{Ab}, α_{aB}, and α_{ab}. It is neither possible nor necessary to distinguish which allele is the actual

maternal or paternal allele in modeling since both maternal and paternal alleles are independent samples from the population and are assumed to contribute the same to the trait variations. Similarly, ij can also be viewed as a mapping function, *ij()*, that a dependent variable, any combination of allele *i* and *j*, is uniquely determined by an individual *k*. The above model is, in fact, an extension of the Zaykin et al. (2002) model with the inclusion of treatments and gene×treatment interactions. Zaykin et al. (2002) proposed to drop the dominance effects to balance between the test degrees of freedom and the testing power constrained by the sample size. We omit the dominance effects for the same reason, and have the following:

Model II

$$Y_{tijk} = \mu + \tau_t + \alpha_i^m + \alpha_j^p + (\tau\alpha)_{ti}^m + (\tau\alpha)_{tj}^p + \varepsilon_{tijk} \quad (3)$$

Zaykin et al. (2002) also showed that the expectation-maximization (E-M) algorithm could be applied to infer haplotype frequencies when the haplotype phases are not observable. In this case, each individual's allelic effects were expanded into all possible haplotype effects with their corresponding frequencies as weights. We are mainly interested in the haplotype phase known cases here.

Relating the variance components of a QTL and a marker

Suppose the QTL we are interested in has alleles A_r and A_s and its genotypic value G_{rs} can be written as

$$G_{rs} = \mu + \alpha_r + \alpha_s + d_{rs} \quad (4)$$

where μ is the overall mean of the genotypic effects, α_r and α_s are the allelic additive effects for allele A_r and A_s, and d_{rs} are their dominance effects. Under the random mating assumption, constraints $\Sigma_r p_r \alpha_r = \Sigma_r p_r d_{rs} = 0$ hold (Weir and Cockerham, 1977), where p_r is the allele frequency of A_r. Therefore, the additive and dominance variance components of the trait can be written as

$$\sigma_A^2 = 2\sum_r p_r (\alpha_r)^2 \quad (5)$$

$$\sigma_D^2 = 2\sum_{r,s} p_r p_s (d_{rs})^2 \quad (6)$$

where p_s is the allele frequency of A_s. Similarly, the genotypic value of a marker locus with alleles M_i and M_j can be written as

$$G_{ij}^{(m)} = \mu + \alpha_i^{(m)} + \alpha_j^{(m)} + d_{ij}^{(m)} \quad (7)$$

Constraints $\Sigma_i q_i \alpha_i = \Sigma_i q_i d_{ij} = 0$ are imposed. The additive and dominance variance components of the marker can be written as $\sigma_A^2(m) = 2\Sigma_i q_i (\alpha_i^{(m)})^2$ and $\sigma_D^2(m) = \Sigma_{i,j} q_i q_j (d_{ij}^{(m)})^2$,), where q_i and q_j are the allele frequencies of M_i and M_j. When the QTL and the marker locus are in LD, Nielsen and Weir (1999) showed

$$\alpha_i^{(m)} = \frac{1}{q_i}\sum_r \alpha_r D_{ri} \quad (8)$$

$$d_{ij}^{(m)} = \frac{1}{q_i q_j}\sum_{r,s} d_{rs} D_{ri} D_{sj} \quad (9)$$

where D_{ri} is the LD coefficient of the trait allele A_r and the marker allele M_i. Allelic additive effects and dominance effects showed at the markers are functions of the causal effects at the trait, marker allele frequencies and the LD coefficients between alleles of two loci. Furthermore,

$$\sigma_A^2(m) = 2\sum_i \frac{1}{q_i}(\sum_r \alpha_r D_{ri})^2 \tag{10}$$

$$\sigma_D^2(m) = \sum_{i,j} \frac{1}{q_i q_j}(\sum_{r,s} d_{rs} D_{ri} D_{sj})^2 \tag{11}$$

If both loci have only two alleles, we can simplify the expressions (10) and (11) by denoting the single LD coefficient $D_{11} = -D_{12} = -D_{21} = D_{22} = D$ and using constraints $\Sigma_r p_r \alpha_r = 0$ *and* $\Sigma_i p_i d_{ij} = 0$. (Details see Appendix I)

$$\sigma_A^2(m) = \frac{D^2 \sigma_A^2}{q_1 q_2 p_1 p_2} = \mathrm{r}^2 \sigma_A^2 \tag{12}$$

$$\sigma_D^2(m) = \frac{D^4 \sigma_D^2}{q_1^2 q_2^2 p_1^2 p_2^2} = \mathrm{r}^4 \sigma_D^2 \tag{13}$$

Here r is the correlation coefficient between allele frequencies at the trait and the marker loci. The above derivations show that choosing different markers will result in different ratios between the variance components at the QTL and marker locus. In another ward, there is a considerable amount of randomness introduced by randomly choosing SNPs to test for the effects of the candidate genes. To take this randomness into account, we argue that the effects

of the QTL and the effects of all the possible combinations of the markers in LD with the QTL follow a certain distribution, and the effects of the markers are drawn from that distribution. Therefore, we propose to model marker genetic effects, haplotype additive effects and haplotype gene×treatment interaction, as random effects under this scenario. In this paper, we strict our attention to the additive effects of the haplotypes formed by several consecutive SNPs since it is reasonable to assume the effects of multiple alleles are drawn randomly from a distribution. Similarly, we can model the genotypic effects of a locus with multiple alleles as effects randomly drawn from a distribution as well, as long as there are multiple possible genotypes at the locus. However, if a single SNP is tested, the number of effects will be only two or three depending allelic or genotypic tests were conducted, and random effect approach will not be appropriate.

The random effect model

We adapt the regression setting in Model II and propose to treat all the genetic effects, the haplotype additive effects and haplotype gene×treatment interactions as random.

Model III

$$Y_{tijk} = \mu + \tau_t + \alpha_i^m + \alpha_j^p + (\tau\alpha)_{ti}^m + (\tau\alpha)_{tj}^p + \varepsilon_{tijk} \quad (14)$$

The differences between Model II and III are that the haplotype additive effects α_i^m and α_j^p iid $\sim N(0, \sigma_\alpha^2)$, $(\tau\alpha)_{ti}^m$ and $(\tau\alpha)_{tj}^p$ iid $\sim N(0, \sigma_{\tau\alpha}^2)$. In the rest of the paper, we use the vector $\boldsymbol{\theta} = \{\sigma_\alpha^2, \sigma_{\tau\alpha}^2, \sigma_e^2\}$ denote the variance components for this model. α_i^m and α_j^p and $(\tau\alpha)_{ti}^m$ and

$(\tau\alpha)_{tj}^{p}$ are independent since they refer to the effects of alleles from unrelated parents. We then have a mixed model with fixed treatment effects, random haplotype additive effects, and random gene×treatment interactions. Because individuals share the same allele (or alleles), their trait values are correlated. The variance-covariance coefficients between individuals are given as following:

$$\mathrm{cov}(y_{tijk}, y_{t'i'j'k'}) = \begin{cases} 2\sigma_\alpha^2 + 2\sigma_{\tau\alpha}^2 + \sigma_e^2 & \text{for any person} \\ 2\sigma_\alpha^2 + 2\sigma_{\tau\alpha}^2 & \text{for two persons sharing the same drug and both alleles} \\ \sigma_\alpha^2 + \sigma_{\tau\alpha}^2 & \text{for two persons sharing the same drug and one allele} \\ 2\sigma_\alpha^2 & \text{for two persons sharing both alleles} \\ \sigma_\alpha^2 & \text{for two persons sharing one allele} \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

where $2\sigma_\alpha^2 = \sigma_A^2$, the additive genetic variation in a classical quantitative genetics setting (Lynch and Walsh 1998).

Under the haplotype phase unknown cases, the EM algorithm (Weir and Cockerham. 1979; Excoffier and Slatkin 1995; Long et al. 1995) can be used to infer the population haplotype frequencies. Each person's possible haplotypes with the corresponding probabilities of having these haplotypes are determined by jointly considering the compatibility of haplotypes to his (her) genotype and haplotype population frequencies. Then, for each person, his (her)

probabilities of having certain haplotypes are fitted as weights in the design matrix of the model.

$$Y_{tgk} = \mu + \tau_t + \Sigma_i p_{i(g)}{}^m \alpha_i{}^m + \Sigma_j p_{j(g)}{}^p \alpha_j{}^p + \Sigma_i p_{i(g)}{}^m (\tau\alpha)_{ti}{}^m + \Sigma_j p_{j(g)}{}^p (\tau\alpha)_{tj}{}^p + \varepsilon_{tgk} \qquad (16)$$

where $p_{i(g)}{}^m$ and $p_{j(g)}{}^p$ are this person's probabilities of having *i* maternal and *j* paternal haplotypes given his (her) genotype g and haplotype population frequencies. $p_{i(g)}{}^m$ and $p_{j(g)}{}^p$ are either 0 or 1 if haplotype phase is known. Again, g can be viewed as a link function, *g()*, that an dependent variable, any genotype *g*, is uniquely determined by a independent variable, an individual *k*.

Then variance-covariance coefficients between individuals are

$$\mathrm{cov}(y_{tgk}, y_{t'g'k'}) = \begin{cases} (\sum (p_{i(g)}^m)^2 + \sum (p_{j(g)}^p)^2)\sigma_\alpha^2 + (\sum (p_{i(g)}^m)^2 + \sum (p_{j(g)}^p)^2)\sigma_{\alpha\tau}^2 + \sigma_e^2 \\ \qquad\qquad \text{for any person} \\ (\sum p_{i(g)}^m \times p_{i'(g')}^m + \sum p_{j(g)}^p \times p_{j'(g')}^p)\sigma_\alpha^2 + (\sum p_{i(g)}^m \times p_{i'(g')}^m + \sum p_{j(g)}^p \times p_{j'(g')}^p)\sigma_{\alpha\tau}^2 \\ \qquad\qquad \text{for any two persons taking the same drug} \\ (\sum p_i^m \times p_{i'}^m + \sum p_j^p \times p_{j'}^p)\sigma_\alpha^2 \\ \qquad\qquad \text{for any two persons taking different drugs} \end{cases} \qquad (17)$$

Note that the variance of an individual's trait value, $((\Sigma(p_{i(g)}{}^m)^2+\Sigma(p_{j(g)}{}^p)^2)\sigma_\alpha{}^2 + (\Sigma(p_{i(g)}{}^m)^2+\Sigma(p_{j(g)}{}^p)^2)\sigma_{\tau\alpha}{}^2 + \sigma_e{}^2)$, when haplotype phase is unknown, is less than the variance of an individual's trait value $(2\sigma_\alpha{}^2 + 2\sigma_{\tau\alpha}{}^2 + \sigma_e{}^2)$, when haplotype phase is known. Unknown

phase represents an information loss and reduces the power for detecting the haplotype effects comparing to haplotype phase known cases, since we have to use the combination of all the possible phases. We focus on the haplotype phase-known case only in this paper.

Using matrix notation, equations (13) or (15) can be written as:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zb} + \boldsymbol{\varepsilon} \tag{18}$$

where $\boldsymbol{X}$ is the design matrix for all the fixed effects (treatments) including the overall mean, $\boldsymbol{\beta}$ is the vector of the fixed effects including the overall mean, $\boldsymbol{Z}$ is the design matrix for the random effects including the haplotype additive effects (denoted as $\boldsymbol{Z_H}$) and interactions (denoted as $\boldsymbol{Z_{INT}}$), and $\boldsymbol{b}$ is the vector of the random effects for the haplotypes and interactions. If there are H possible haplotypes at this locus and T possible treatments, the random effect $\boldsymbol{b}$ can be written as $(\alpha_1^m, \ldots, \alpha_H^m, \ldots, \alpha_1^p, \ldots, \alpha_H^p, (\tau\alpha)_{11}^m, \ldots, (\tau\alpha)_{TH}^m, \ldots, (\tau\alpha)_{11}^p, \ldots, (\tau\alpha)_{TH}^p)$. And

$$\mathbf{b} \sim N(0, \begin{matrix} \sigma_\alpha^2\mathbf{I}_{H\times H} & 0 \\ 0 & \sigma_{\alpha\tau}^2\mathbf{I}_{TH\times TH} \end{matrix}) = \mathrm{N}(0, \mathbf{G}) \tag{19}$$

Then $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \boldsymbol{ZG(\theta)Z}^T + \boldsymbol{I}\sigma_e^2)$, and $\boldsymbol{Y}|\boldsymbol{b} \sim N(\boldsymbol{X\beta}+\boldsymbol{Zb}, \boldsymbol{I}\sigma_e^2)$. After estimating the variance components of the random effects, the fixed effects can be estimated using $\boldsymbol{\beta}^\wedge = (\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\theta}^\wedge)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\theta}^\wedge)^{-1}\boldsymbol{Y}$, where $\boldsymbol{V}(\boldsymbol{\theta}^\wedge) = \boldsymbol{ZG}(\boldsymbol{\theta}^\wedge)\boldsymbol{Z}^T + \boldsymbol{I}\sigma_e^{2\wedge}$. Each level of the random effects can be predicted as $\boldsymbol{b}^\wedge = \boldsymbol{G}(\boldsymbol{\theta}^\wedge)\boldsymbol{Z}^T\boldsymbol{V}(\boldsymbol{\theta}^\wedge)^{-1\wedge}(\boldsymbol{Y}-\boldsymbol{X\beta}^\wedge)$. We are not interested in estimating $\boldsymbol{\beta}^\wedge$ or predicting $\boldsymbol{b}^\wedge$ in the paper.

Hypothesis Testing

To compare the above three models, we are interested in testing the following hypotheses:

1. Whether treatments have significant effects

For all models: H_0: $\tau_1 = \tau_2 = \ldots = \tau_T = 0$ vs. H_a: at least one of them $\neq 0$.

2. Whether the gene (or the region) has significant effects

Model I: H_0: $G_1 = G_2 = \ldots = G_G = 0$ vs. H_a: at least one of them $\neq 0$

Model II: H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_H = 0$ vs. H_a: at least one of them $\neq 0$

Model III: H_0: $\sigma_\alpha^2 = 0$ vs. H_a: $\sigma_\alpha^2 > 0$

3. Whether gene×treatment interactions have significant effects

Model I: H_0: $(\tau G)_{11} = (\tau G)_{21} = \ldots = (\tau G)_{TG} = 0$ vs. H_a: at least one of them $\neq 0$

Model II: H_0: $(\tau\alpha)_{11} = (\tau\alpha)_{21} = \ldots = (\tau\alpha)_{TH} = 0$ vs. H_a: at least one of them $\neq 0$

Model III: H_0: $\sigma_{\tau\alpha}^2 = 0$ vs. H_a: $\sigma_{\tau\alpha}^2 > 0$

For all three models, we fit and test all effects in a similar order. We test the treatment effects when the genetic effects are not included (so called Type I tests) or included (so called Type III tests) in the model, test the genetic effects when the treatments are not included (Type I) or included (Type III) in the model, and test the interactions when both treatments and genetic main effects are included in the model. Since the major sources of genetic variation that we are

interested in are the genetic effects and the treatment by gene interactions, we test whether both are significantly different from zero when the treatment effects are included in the model.

To test any effects in a fixed effects model, we utilize the most frequently used F test:

$$F = \frac{[ESS(reduced) - ESS(full)]/[df(reduced) - df(full)]}{ESS(full)/df(full)} \tag{20}$$

$$\sim F_{df(reduced)-df(full),df(full)} \qquad \text{under } H_0$$

where ESS is the error sum of square for each model, the reduced model is the model without the effects under testing, the full model is the model with the effects under testing, and *df* denotes the degrees of freedom for each model. For Model I, the genotypic test for each SNP is conducted individually, the smallest *p* value of the all tests is recorded as the *p* value for this set of SNPs, and a Bonferroni correction is then applied to this *p* value.

To test Type III treatment effects in a random effect model, we use an F test with a Satterthwaite approximation for the denominator degrees of freedom.

$$F = \frac{\hat{\beta}'\mathbf{L}'(\mathbf{L}\hat{C}\mathbf{L}')^{-1}\mathbf{L}\hat{\beta}}{q} \tag{21}$$

Where $\boldsymbol{\beta}$ is a vector of fixed effects under testing, $\boldsymbol{L}$ is an estimable contrast matrix of rank $q > 1$, and $C^{\wedge}$ is $(\boldsymbol{X'V^{\wedge -1}X})^{-}$.

The variance components of the random effects are estimated and tested using the Restricted Maximum Likelihood (REML) approach. The likelihood ratio statistic is constructed as *–2 * [the likelihood value of the reduced model – the likelihood values of the full model]*, where the likelihood values are obtained from REML. When testing whether a single variance component (σ_α^2 or $\sigma_{\tau\alpha}^2$) is significantly greater than zero, the likelihood ratio statistic follows a $50\%\ \chi_1^2 + 50\%\ \chi_0^2$ distribution if asymptotic theory holds (Self SG and Liang KY, 1987). In our simulation study, the validity of the tests is verified and showed to be slightly conservative. Therefore, a permutation test is implemented for each test. When testing whether σ_α^2 is significantly greater than zero, the rows of the design matrix **Z** are shuffled with respect to **Y** and **X**, and likelihood ratio statistic is recalculated for each shuffling. When testing whether $\sigma_{\tau\alpha}^2$ is significantly greater than zero, the rows of the design matrix for the interactions $\mathbf{Z}_{INT}$ are shuffled with respect to **Y**, **X** and **ZH** to separate the interactions from the main effects and likelihood ratio statistic is recalculated for each shuffling. The permutation p value is calculated as the proportion of the times that the permuted likelihood ratio statistic is great than or equal to the likelihood ratio from the non-permuted data. Whether at least one of the two variance components (σ_α^2 and $\sigma_{\tau\alpha}^2$) is significantly from zero can also be tested. The likelihood ratio statistic follows a $25\%\ \chi_2^2 + 50\%\ \chi_1^2 + 25\%\ \chi_0^2$ distribution if asymptotic theory holds (Self SG and Liang KY, 1987). This type of tests is not studied here.

Simulation study

To illustrate the performances of the above three models under the influences of different factors such as genetic models and mutation origins of QTL, we conduct a simulation study. One thousand unrelated individuals were collected in a case-control clinical trial study, where

half of the patients were randomly selected and given one drug and the other half were given placebo. To study the possible genetic effects in the drug efficacy or adverse event, five SNPs were genotyped in a candidate gene at a density of one SNP every 30kb in all 1000 individuals. Several assumptions are made to keep our simulation study relatively simple and representative. First, the response variable is a quantitative trait representing patients' responses to the drug. Second, there are no missing data. As we mentioned above, the haplotype phases are usually unobservable but can be inferred reasonably well using EM algorithm. We assume that haplotype phases are known. Last, the QTL is within the gene under study, but can be either included in our data or not depending on the selection of the SNPs within the gene. We simulate three types of genetic causation for the trait variation. First, a single site mutation (one SNP) might cause the change of the quantitative value. The QTL minor allele frequency (MAF) is relatively uncommon, 5%-15%, and presumably a relatively young mutation. Second, a single site mutation (one SNP) with MAF from 15% to 50% presumably a relatively old mutation might be the QTL. Last, the quantitative value change might be caused jointly by three single mutations (3 SNPs) in the gene with non-additive effects of three mutations. Hudson (2002) simulation program based on a coalescent approximation to the Wright-Fisher neutral model is used to generate gametes (haplotypes). The program assumes a diploid panmictic population of constant size N, an infinite-sites model of mutations, and no selection. A random genealogy tree for a piece of chromosome is generated for each sample. Mutations are created according to a Poisson process with the mean equal to the product of the mutation rate and the branch length. Only mutation and recombination processes are assumed in our simulation. In each simulation run, 2000 gametes were generated and randomly paired into 1000 individuals. When QTL is a single SNP

mutation, six linked loci with 10 SNPs in each locus are simulated with equal possible recombination occurring between loci and no recombination within the locus. One out of 10 SNPs in each locus is selected if its MAF is greater than 5%, which results a region with six SNPs numbered as 1-6. The QTL is always the third SNP, which can be typed into the data or not. If the QTL is typed, SNP 1, 2, 3, 5, 6 are included in the data. If the QTL is not typed, SNP 1, 2, 4, 5, 6 are included in the data. To mimic a young QTL, SNP 3's MAF is controlled between 5%-15%. To mimic an old QTL, SNP 3's MAF is controlled between 15-50%. When QTL is a combination of multiple SNP mutations, the above simulation steps are used similarly except 7 SNPs generated and numbered as 1-7. When a QTL is typed, SNPs 1, 3, 4, 5, 7 are included in the data. Otherwise, SNPs 2, 3, 4, 6, 7 are included. For both types of QTL mutation origins, either the mutation site (sites) or a nearby SNP (SNPs) are included in the data, which all result in 5 SNPs in the gene under study. One parameter required by the simulation program is the scaled recombination size of the region under study, *4Nc*, where *N* is the effective population size and c is the recombination rate per gamete per generation between the two ends of the region. If we assume *4N* equal to 10,000 and set *4Nc* equal to15, the simulated region corresponds to a region of 0.15Mb. Since five SNPs are finally selected and "typed" in the region, a map density of one SNP every 30kb is then resolved. Our study is based on 500 simulation runs for each model. Then the quantitative trait values are generated by jointly considering the QTL and effects in table 1 or 2. Each individual's quantitative trait value is generated using the following formula.

$$Y_{tik} = \mu + \tau_t + G_i + (G\tau)_{ti} + \varepsilon_{tik}$$

Where μ is always set to zero, ε_{tik} is randomly generated from *N(0,1),* and the drug *(τ_t),* genetic *(G_i)* and interactions *($(G\tau)_{ti}$)* are determined by values in the tables 1 or 2 and the types of the QTL causation, the genetic model, and whether an individual took the drug or not. Proportions of the drug effects, gene effects and gene×treatment interactions for 6 different conditions and three types of genetic models, additive, dominance and recessive, are considered and shown in tables 1 and 2. The percentages of genetic effects and gene×treatment interactions out of the total variation are calculated using the averaged percentages of the sum of squares of the effects over the sum of squares of the total corrected by mean and shown in corresponding graphs. For the permutation tests of the random effects, 99 permutations are conducted for each simulation and a regression-based method from Boos and Zhang (2002) is utilized to reduce the bias of power for the study. Estimated type I error rates and the power for each model under different circumstances are calculated in testing the genetic effects and gene×treatment interactions at $\alpha = 0.05$ level. In generating data, we try to keep all the effects the same for all genetic models and different types of single mutation based QTLs in order to show the impact of the genetic models and QTL compositions in the study. However, some effects were increased in data generation for either additive or recessive models because the original effects were too small to be detected using the sample size we studied. Increasing the sample sizes might increase the testing power. However, it is not an issue we are interested in this project.

Results

Estimated type I error rate

Table 3 summarizes the estimated type I error rates in testing the type I and type III genetic effects and gene-drug interactions when the QTL is a single young mutation, a single old mutation or a combination of multiple mutations, and when the QTL is included in the data or not for Model I, Model II, Model III with the asymptotic test and Model III with the permutation test. All the tests for Model I and Model II are valid. Model III with asymptotic tests are slightly conservative. When the permutation tests are utilized for Model III, the test sizes for genetic effects are corrected, but the test sizes for gene-drug interactions are still slightly conservative. The test sizes are similar regardless the structure of the genetic components.

Estimated Power

Drug effects

Figure 4.1 shows the power of the drug type III effect tests when the QTL is a single young mutation, a single old mutation or multiple mutations, when the genetic model is dominance, additive or recessive, and when there is drug effects only, drug and gene effects, or drug effects and different magnitude gene-drug interactions. The power of the drug type I effect tests is the same as that for type III tests since the experimental design is balanced here. The powers of the drug effect tests are the same for all three models. When the genetic model is recessive, the effects of the gene-drug interactions on the drug effects are hard to detect except when the effective allele is quite frequent.

Genetic effects

When the QTL is a single young mutation and included in the data, the fixed genotypic test (FG) is the most powerful test. The fixed haplotype effect test (FH) and the random haplotype effect with permutation test (RHA) have similar power. The random haplotype effect with the asymptotic test (RHP) is the least powerful. When the QTL is not included in the data and the genetic model is either dominance or additive, RHP is the most powerful test; FG is the least powerful test; RHA is as same as or slightly more powerful than FH. When the QTL is single old mutation and included in the data, FG is the most powerful test; FH is the least powerful test; RHP is more powerful than PHA. When QTL is not included and the genetic model is either dominance or additive, RHP is the most powerful test; FG is the least powerful test; RHA is as same as or slightly more powerful than FH. When the QTL is a combination of multiple mutations and the genetic model is either dominance or additive, FG, FH, RHA and RHP are almost equivalent except FH is slight powerful under some circumstances. When QTL is not included and the genetic model is either dominance or additive, RHP is most powerful; FG is least powerful; RHA and FH are almost equivalent. Under the recessive model, FG is always the most powerful test.

Gene-drug interactions

When there is no gene-drug interaction, the power is the type I error rate under the null hypotheses. When the QTL is a single young mutation and the genetic model is either dominance or additive, RHA and RHP are the most powerful tests. RHP is slightly more powerful than PHA. When QTL is included in the data, FH is least powerful. When the QTL is not included, FG is least powerful. When the genetic model is recessive, FG is always the most

powerful test. When the QTL is single old mutation and included in the data, RHA and RHP are the most powerful tests. RHP is slightly more powerful that PHA. When the QTL is included in the data, FH is least powerful. When the QTL is not included, FG and FH are equivalently the least powerful test. When the genetic model is recessive and the QTL is included, FG is most powerful; FH is least powerful. . When the genetic model is recessive and the QTL is not included, RHA and RHP are as same as or slightly more powerful than FG and FH. When the QTL is a combination of multiple mutations and the genetic model is either dominance or additive, RHA and RHP are the most powerful tests. RHP is as same as or slightly more powerful that PHA. FG is least powerful. Under the recessive genetic model, RHA and RHP are as same as or slightly more powerful than FG.

Discussion

In detecting genotypic effects, our study shows that the Model I based F-test is most powerful when the QTL is a single mutation and included in the data or the genetic model is recessive. Note, when the QTL is a single mutation and included in the data, Model I is the statistically appropriate model. The haplotype additive model (Model II and Model III) based tests are more powerful when the single mutation based QTL is not included in the data and the genetic model is either dominance or additive, or the QTL is a combination of several mutations. One observation is that haplotypes based tests start to have advantages over single SNP genotypic test when the number of mutations increases in the QTL. Although, this study does not serve as a proof that haplotypes based tests are more powerful than single SNP based tests since only limited circumstances are explored and we assume that the haplotype phases are observed, which is seldom the case because of the difficulties in obtaining haplotype data. Losing haplotype phase information might result in losing statistical power of detecting the effects. However, our results coincide with some recent published work. Morris and Kaplan (2002) showed that the haplotype analysis has advantages when multiple disease susceptibility alleles present under the additive genetic model. It further shows the importance of developing statistical appropriate and more powerful haplotype analysis methods. The fixed effect models based F-test is asymptotically equivalent to the likelihood ratio test and equivalent when the normality holds. Our simulation study shows that Model I based tests have correct size when the Bonferroni correction is applied. One reason might be that the LD between SNPs is relatively weak.

In testing gene×treatment interactions, our Model III based tests are the most powerful tests under most circumstances we explored except when the disease gene model is recessive and the QTL is included it the data. It shows that the random effect model is not only statistically rigorous but also having comparable power to the other tests. For Model III, Permutation tests increase the power of likelihood ratio tests based on asymptote, but not dramatically. The asymptote based likelihood ratio tests can be applied directly although they are slightly conservative under the null hypotheses.

When the dominance effects (d_{ij}) are fitted, both Model II and III are almost equivalent to Model I with genotype classes of multiple loci, but providing more knowledge on the additive and dominance proportions of the genetic effects. However, for any particular analysis, there is a consideration of the sample size versus the test degrees of freedom. In this study, we choose not to include the dominance terms in the haplotype analysis. Extending Model II and III to include the dominance term is certainly if interest.

Another observation is that the type I and type III tests of the genetic and drug effects are almost similar in our study results. The reason is that we have a balanced design in our simulation study and the design matrices of these two independent effects are orthogonal to each other. Therefore, the power of detecting one type of effects is almost independent from the presents of the other factors. We present the type III test results. However, type III tests should always be used in case there is unbalance in the experimental design and the type I test results might be misleading.

There are several issues worth considering, such as the sample sizes needed or the appropriate SNP densities. Similar to causes of the complex diseases, multiple genes with moderate effects are likely to play a role in complex drug treatment interactions as well. Therefore, a relatively large sample size may need in detecting the effects. Fortunately, phase III or IV clinical trials might provide enough sample sizes. Further study is certainly of interest.

Acknowledgment

We thank John Monahan, Xiaohui Luo, Russ Wolfinger, Daowhen Zhang, and Zhao-Bang Zeng for helpful discussions and advice. This work was supported in part by NIH grant GM45344.

Appendix

Nielsen and Weir (1999) showed

$$G_{ij}^{(m)} = \sum_{r,s} \Pr(A_r A_s \mid M_i M_j) G_{rs}$$
$$= \mu + \frac{1}{q_i}\sum_r \alpha_r D_{ri} + \frac{1}{q_j}\sum_r \alpha_r D_{rj} + \frac{1}{q_i q_j}\sum_{r,s} d_{rs} D_{ri} D_{sj} \quad \text{(A1)}$$

Combining equation (7) and (A1), equations (8) and (9) held. If both the trait and the marker loci have only two alleles, equation (10) can be further simplified.

$$\sigma_A^2(m) = 2\sum_i q_i (\alpha_i^{(m)})^2$$
$$= 2\sum_i \frac{1}{q_i}(\sum_r \alpha_r D_{ri})^2$$
$$= 2(\frac{1}{q_1} + \frac{1}{q_2})(\alpha_1^2 + \alpha_2^2 - 2\alpha_1\alpha_2) D^2 \quad \text{(A2)}$$

From the constrain $\Sigma_r p_r \alpha_r = 0$, we have

$$(p_1\alpha_1 + p_2\alpha_2)^2 = p_1^2\alpha_1^2 + p_2^2\alpha_2^2 + 2p_1p_2\alpha_1\alpha_2 = 0 \quad \text{(A3)}$$

It follows

$$-2\alpha_1\alpha_2 = \frac{p_1^2\alpha_1^2 + p_2^2\alpha_2^2}{p_1 p_2} \quad \text{(A4)}$$

Substitute equation (A4) into equation (A2), we have

$$\sigma_A^2(m) = 2(\frac{1}{q_1} + \frac{1}{q_2})(\alpha_1^2 + \alpha_2^2 + \frac{p_1^2\alpha_1^2 + p_2^2\alpha_2^2}{p_1 p_2})D^2$$
$$= \frac{2\mathrm{D}^2}{\mathrm{q_1 q_2 p_1 p_2}}(\mathrm{p_1}\alpha_1^2 + \mathrm{p_2}\alpha_2^2)$$
$$= \frac{2\mathrm{D}^2}{\mathrm{q_1 q_2 p_1 p_2}}\sigma_A^2 \qquad \text{(A5)}$$

Similarly, the marker dominance variation (11) can be written as

$$\sigma_D^2(m) = \sum_{i,j}\frac{1}{q_i q_j}(\sum_{r,s} d_{rs} D_{ri} D_{sj})^2$$
$$= (\frac{1}{\mathrm{q_1^2}} + \frac{1}{\mathrm{q_2^2}} + \frac{2}{\mathrm{q_1 q_2}})\mathrm{D}^4(d_{11} + d_{22} - 2d_{12})^2$$
$$= \frac{D^4}{q_1^2 q_2^2}(d_{11}^2 + d_{22}^2 + 4d_{12}^2 - 4d_{11}d_{12} - 4d_{12}d_{22} + 2d_{11}d_{22}) \qquad \text{(A6)}$$

From the constrain $\Sigma_r q_r d_{rs} = 0$, we have

$$(p_1 d_{11} + p_2 d_{12})^2 = p_1^2 d_{11}^2 + p_2^2 d_{12}^2 + 2p_1 p_2 d_{11} d_{12} = 0 \qquad \text{(A7)}$$

Then,

$$-2d_{11}d_{12} = \frac{p_1^2 d_{11}^2 + p_2^2 d_{12}^2}{p_1 p_2} \qquad \text{(A8)}$$

Similarly,

$$-2d_{12}d_{212} = \frac{p_1^2 d_{12}^2 + p_2^2 d_{22}^2}{p_1 p_2} \qquad \text{(A9)}$$

Also,

$$d_{11} = \frac{p_2^2}{p_1^2}\mathrm{d}_{22} \qquad \text{(A10)}$$

Substitute equation (A8), (A9) and (A10) into (A6), we have

$$\sigma_D^2(m) = \frac{D^4}{q_1^2 q_2^2}(d_{11}^2 + d_{22}^2 + 4d_{12}^2 - 4d_{11}d_{12} - 4d_{12}d_{22} + 2d_{11}d_{22})$$
$$= \frac{\mathrm{D}^4}{q_1^2 q_2^2 p_1^2 p_2^2}(\mathrm{p}_1^2\mathrm{d}_{11}^2 + 2\mathrm{p}_1\mathrm{p}_2\mathrm{d}_{12}^2 + \mathrm{p}_2^2\mathrm{d}_{22}^2)$$
$$= \frac{\mathrm{D}^4}{q_1^2 q_2^2 p_1^2 p_2^2}\sigma_{\mathrm{D}}^2 \qquad \text{(A11)}$$

If either the trait or the marker locus has more than two alleles, there are no simple relationships of the variance components of the trait and the marker loci as the above held. Instead, the effects of the trait is related to the effects of the marker though a rather complicated equation.

Reference

Aithal GP, Day CP, Kesteven PJ, Daly AK. (1999) Association of polymorphisms in the cytochrome p450 CYP2C9 with warfarin dose requirement and risk of bleeding complications. Lancet 353: 717-719

Boos D. and Zhang J (2000) Monte Carlo evaluation of resampling-based hypothesis tests. JASA Vol 95 No. 450 486-492

Clark AG. Inference of haplotype from PCR-amplified samples of a diploid population. (1990) Mol Biol Evol 7:111-122

Davidson S (2000) Research suggests importance of haplotypes over SNPs. Natures Biotechnology 18: 1134-1135

Douglas J, Boehnke M, et al. (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. Nat Genet 28:361-364

Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, and Liggett SB (2000) Complex promoter and coding region $_2$-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness**.** PNAS 97: 10483-10488

Excoffier L and Slatkin M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921-927

Gao F, Johnson DL, Ekins S, Janiszewski J, Kelly KG, Meyer RD, West M. Optimizing higher throughput methods to assess drug-drug interactions for CYP1A2, CYP2C9, CYP2C19, CYP2D6, rCYP2D6, and CYP3A4 in vitro using a single point IC(50). (2002) J Biomol Screen 7:373-382

Hetherington S, Hughes AR, Mosteller M, Shortino D, Baker KL, Spreen W, Lai E, Davies K, Handley A, Dow DJ, Fling ME, Stocum M, Bowman C, Thurmond LM, Roses AD. (2002) Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. Lancet. 359: 1121-1122

Hudson RR. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics Vol. 18: 337-338

Hoehe MR, Kopke K, Wendel B, Rohde K, Flachmeier C, Kidd KK, Berrettini WH, Church GM. (2000) Sequence variability and candidate gene analysis in complex disease: association of mu opioid receptor gene variation with substance dependence. Hum Mol Genet 9:2895-2908

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56: 799 –810

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sunderland, Mass: Sinauer

Marshall A. (1997) Getting the right drug into the right patient. Nature Biotechnol 15:1249-1251

Morris RW, Kaplan NL. (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221-233

Nielsen DM, Weir BS (1999): A classical setting for association between markers and loci affecting quantitative traits. Genet Res 74:271-277

Raimundo S, Fischer J, Eichelbaum M, Griese E-U, Schwab M, Zanger UM. (2000) Eluciatdion of the genetic basis of the common “intermediate metabolizer” phenotype for drug oxidation by CYP2D6. Pharmacogenetics. 10: 1-5

Risch, N, Merikangas, K (1996) The future of genetic studies of complex human diseases. Science 273: 1516-1517

Sachidanandam, R., Weissman, D et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928-933

Self SG and Liang KY (1987) Asymptotic properties of maximum likelihood estimators and like likelihood ratio tests under nonstandard conditions. JASA Vol 82, No 398: 605-610

Spear BB, Heath-Chiozzi M. Huff J. (2001) Clinical application of pharmacogenetics TRENDS in Molecular Medicine. Vol. 7 No.5

Weir BS, Cockerham CC (1977) Two-locus theory in quantitative genetics. In: Proceedings of the international Conference on Quantitative Genetics. (eds) Pollak E, Kempthorne O, Bailey TB, Iowa State University Press, Ames, Iowa 247-269

Weir, B.S. and C.C. Cockerham. (1979) Estimation of linkage disequilibrium in randomly mating populations. Heredity 42:105-111

Woolley AT, Guillemette C, Li Cheung C, Housman DE, Lieber CM. (2000) Direct Haplotyping of Kilobase-Size DNA Using Carbon Nanotube Probes. Nature Biotechnology 18:760-763.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. (2002) Testing Association of Statistically Inferred Haplotypes with Discrete and Continuous Traits in Samples of Unrelated Individuals. Hum Hered 53:79-91

Figure legends

Figure 4.1: Power of drug effects type III tests. Three types of vertical bars stand for three genetic models: the light solid bar: the dominance model; the dark solid bar: the additive model; the empty bar: the recessive model. Five groups of bars are results for different types of effects in the model. From left to right, drug effects only, drug and genetic effects, drug and small gene-drug interactions, drug and medium gene-drug interactions, and drug and large gene-drug interactions. Figures from top to bottom: A: the QTL is a single young mutation. B: the QTL is a single old mutation. C: the QTL is a combination of multiple mutations.

Figure 4.2: Power of genetic effects type III tests when the QTL is a single young mutation. Four types of vertical bars stand for different models and tests: the dark gray bar: Model I; the light bar: Model II; the light gray bar: Model III with the asymptotic test; the dark bar: Model III with the permutation test. For each picture, four groups of bars on the left are the results when QTL is included in the data; four groups of bars on the right are the results when QTL is not included in the data. Within each four groups of bars, from left to right, the groups are results when there are different types of effects in the model: drug and genetic effects, drug and small gene-drug interactions, drug and medium gene-drug interactions, and drug and large gene-drug interactions. Proportions of effects out of total variation are shown. Figures from top to bottom, A: Dominance genetic model; B: Additive genetic model with increased effects; C: Recessive genetic model with increased effects.

Figure 4.3: Power of genetic effects type III tests when the QTL is a single old mutation. All the notations are as the same as in Figure 2. Figures from top to bottom, A: Dominance genetic model; B: Additive genetic model; C: Recessive genetic model.

Figure 4.4: Power of genetic effects type III tests when the QTL is a combination of multiple mutations. All the notations are as the same as in Figure 2. Figures from top to bottom, A: Dominance genetic model; B: Additive genetic model; C: Recessive genetic model.

Figure 4.5: Power of gene-drug interaction tests when the QTL is a single young mutation. All the notations are as the same as in Figure 2. Figures from top to bottom, A: Dominance genetic model; B: Additive genetic model with increased effects; C: Recessive genetic model with increased effects.

Figure 4.6: Power of gene-drug interaction tests when the QTL is a single old mutation. All the notations are as the same as in Figure 2. Figures from top to bottom, A: Dominance genetic model; B: Additive genetic model; C: Recessive genetic model.

Figure 4.7: Power of gene-drug interaction tests when the QTL is a combination of multiple mutations. All the notations are as the same as in Figure 2. Figures from top to bottom, A: Dominance genetic model; B: Additive genetic model; C: Recessive genetic model.

Table 4.1 Simulation effects when quantitative trait locus (QTL) is a single SNP mutation.

MODEL	Genetic Model	Additive			Dominance			Recessive		
	Genetic Components At QTL	1/1	1/0	0/0	1/1	1/0	0/0	1/1	1/0	0/0
1	τ_t	0	0	0	/	/	/	/	/	/
	G_i	0	0	0	/	/	/	/	/	/
	$(G\tau)_{ti}$	0	0	0	/	/	/	/	/	/
2	τ_t	0.08	0.08	0.08	/	/	/	/	/	/
	G_i	0	0	0	/	/	/	/	/	/
	$(G\tau)_{ti}$	0	0	0	/	/	/	/	/	/
3	τ_t	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
	G_i	0.28	0.14	0	0.28	0.28	0	0.28	0	0
	$(G\tau)_{ti}$	0	0	0	0	0	0	0	0	0
4	τ_t	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
	G_i	0	0	0	0	0	0	0	0	0
	$(G\tau)_{ti}$	0.32	0.16	0	0.32	0.32	0	0.32	0	0
5	τ_t	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
	G_i	0	0	0	0	0	0	0	0	0
	$(G\tau)_{ti}$	0.48	0.24	0	0.48	0.48	0	0.48	0	0
6	τ_t	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
	G_i	0	0	0	0	0	0	0	0	0
	$(G\tau)_{ti}$	0.64	0.32	0	0.64	0.64	0	0.64	0	0

Note: 1. “1” is always the minor allele.

2. Model 1-6 represent 6 different proportions of the drug effect, gene effects and interactions.

Table 4.2 Simulation effects when quantitative trait locus (QTL) is from multiple SNP mutations.

Model	Drug Effect	Genetic Allelic Effect			Gene×treatment Allelic Effect		
		11*	100	00*	11* with drug	100 with drug	00* with drug
1	0	0	0	0	0	0	0
2	0.08	0	0	0	0	0	0
3	0.08	0.32	0.16	-0.06	0	0	0
4	0.08	0	0	0	0.60	0.30	-0.06
5	0.08	0	0	0	1.20	0.30	-0.06
6	0.08	0	0	0	1.80	0.30	-0.06

Note: 1. "1" is always the minor allele.

2. '*' means any allele.

3. Model 1-6 represent 6 different proportions of the drug effect, gene effects and interactions.

4. Additive, dominance and recessive models are also applied to Model 3-6. The effects of alleles are dominated by alleles from left to right.

Table 4.3 Type I error rates with 95% confidence interval.

Effects	Model	Single Young SNP QTL		Single Old SNP QTL		Multiple SNPs QTL	
		W/ QTL	W/O QTL	W/ QTL	W/O QTL	W/ QTL	W/O QTL
Genetic Effect Type I	I	0.06 ± 0.02	0.07 ± 0.02	0.04 ± 0.02	0.04 ± 0.02	0.04 ± 0.02	0.05 ± 0.02
	II	0.06 ± 0.02	0.05 ± 0.02	0.07 ± 0.02	0.05 ± 0.02	0.06 ± 0.02	0.05 ± 0.02
	III	0.04 ± 0.02	0.04 ± 0.02	0.03 ± 0.02	0.03 ± 0.02	0.03 ± 0.02	0.04 ± 0.02
	III w/ perm	0.06 ± 0.02	0.07 ± 0.02	0.05 ± 0.02	0.05 ± 0.02	0.05 ± 0.02	0.07 ± 0.02
Genetic Effect Type III	I	0.06 ± 0.02	0.07 ± 0.02	0.04 ± 0.02	0.04 ± 0.02	0.04 ± 0.02	0.05 ± 0.02
	II	0.06 ± 0.02	0.05 ± 0.02	0.07 ± 0.02	0.05 ± 0.02	0.06 ± 0.02	0.06 ± 0.02
	III	0.05 ± 0.02	0.04 ± 0.02	0.03 ± 0.02	0.03 ± 0.02	0.03 ± 0.02	0.04 ± 0.02
	III w/ perm	0.06 ± 0.02	0.07 ± 0.02	0.05 ± 0.02	0.05 ± 0.02	0.04 ± 0.02	0.07 ± 0.02
Gene-drug interaction	I	0.05 ± 0.02	0.05 ± 0.02	0.03 ± 0.02	0.04 ± 0.02	0.04 ± 0.02	0.05 ± 0.02
	II	0.05 ± 0.02	0.05 ± 0.02	0.04 ± 0.02	0.05 ± 0.02	0.05 ± 0.02	0.05 ± 0.02
	III	0.03 ± 0.02	0.03 ± 0.02	0.04 ± 0.02	0.03 ± 0.02	0.03 ± 0.02	0.03 ± 0.02
	III w/ perm	0.05 ± 0.02	0.04 ± 0.02	0.05 ± 0.02	0.04 ± 0.02	0.04 ± 0.02	0.03 ± 0.02

Figure 4.1

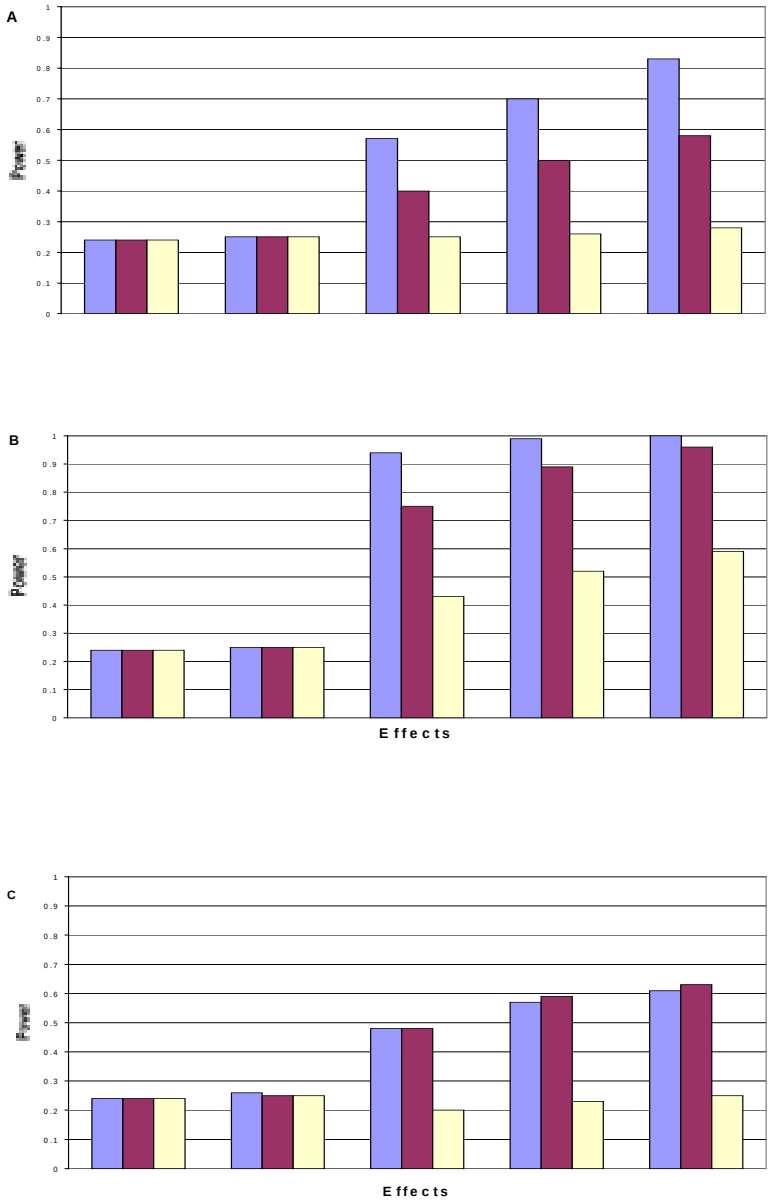
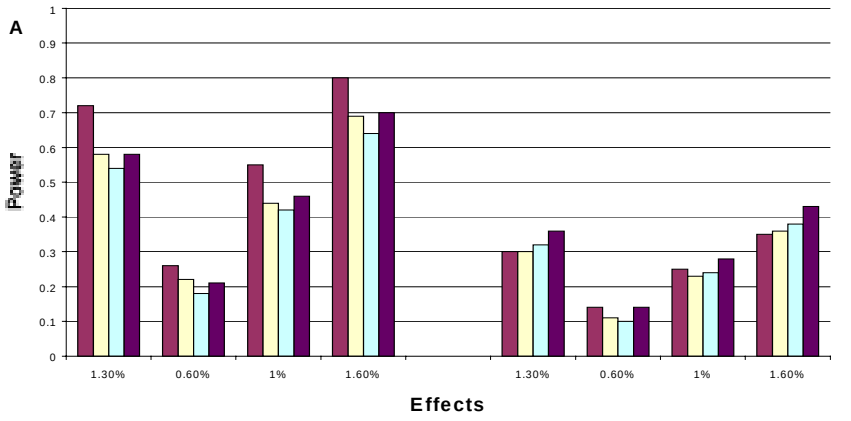
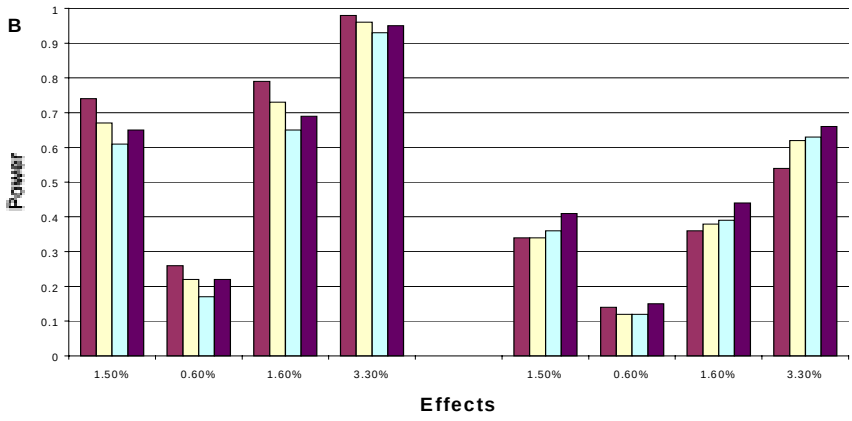


Figure 4.2


A
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1.30%
0.60%
1%
1.60%
1.30%
0.60%
1%
1.60%
Effects


B
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1.50%
0.60%
1.60%
3.30%
1.50%
0.60%
1.60%
3.30%
Effects

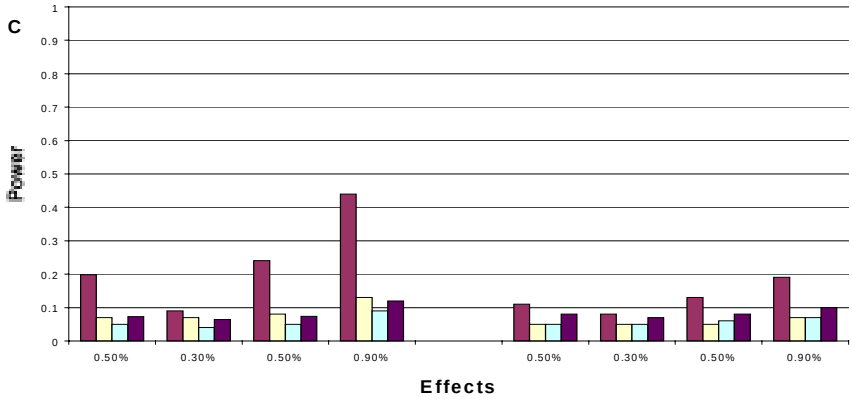
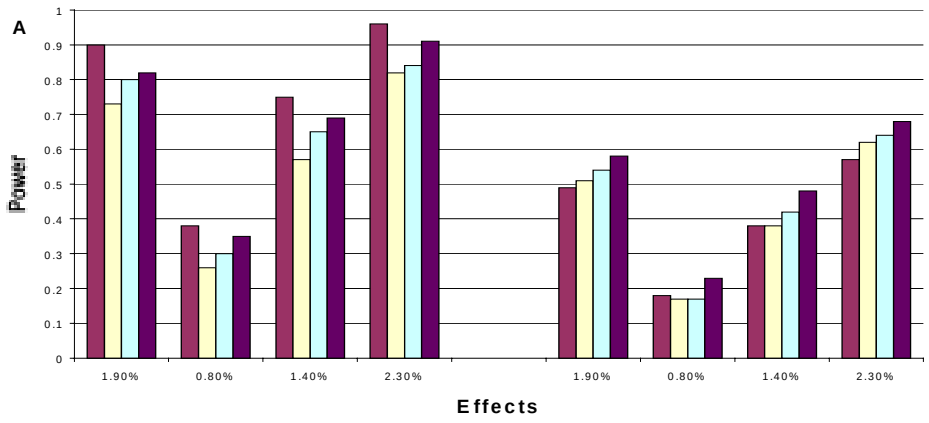
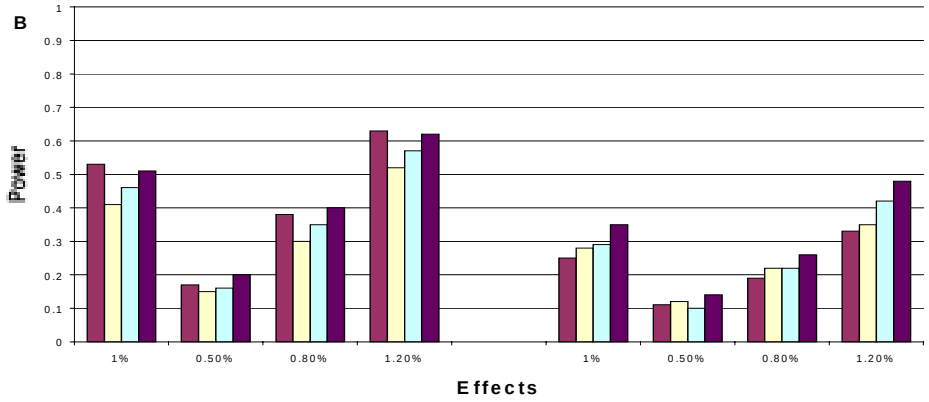

C
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.50%
0.30%
0.50%
0.90%
0.50%
0.30%
0.50%
0.90%
Effects

Figure 4.3


A
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1.90%
0.80%
1.40%
2.30%
1.90%
0.80%
1.40%
2.30%
Effects


B
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1%
0.50%
0.80%
1.20%
1%
0.50%
0.80%
1.20%
Effects

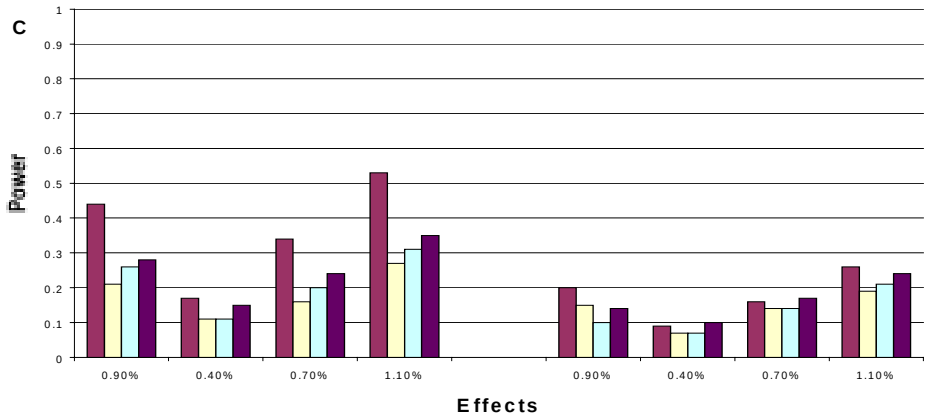

C
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.90%
0.40%
0.70%
1.10%
0.90%
0.40%
0.70%
1.10%
Effects

Figure 4.4

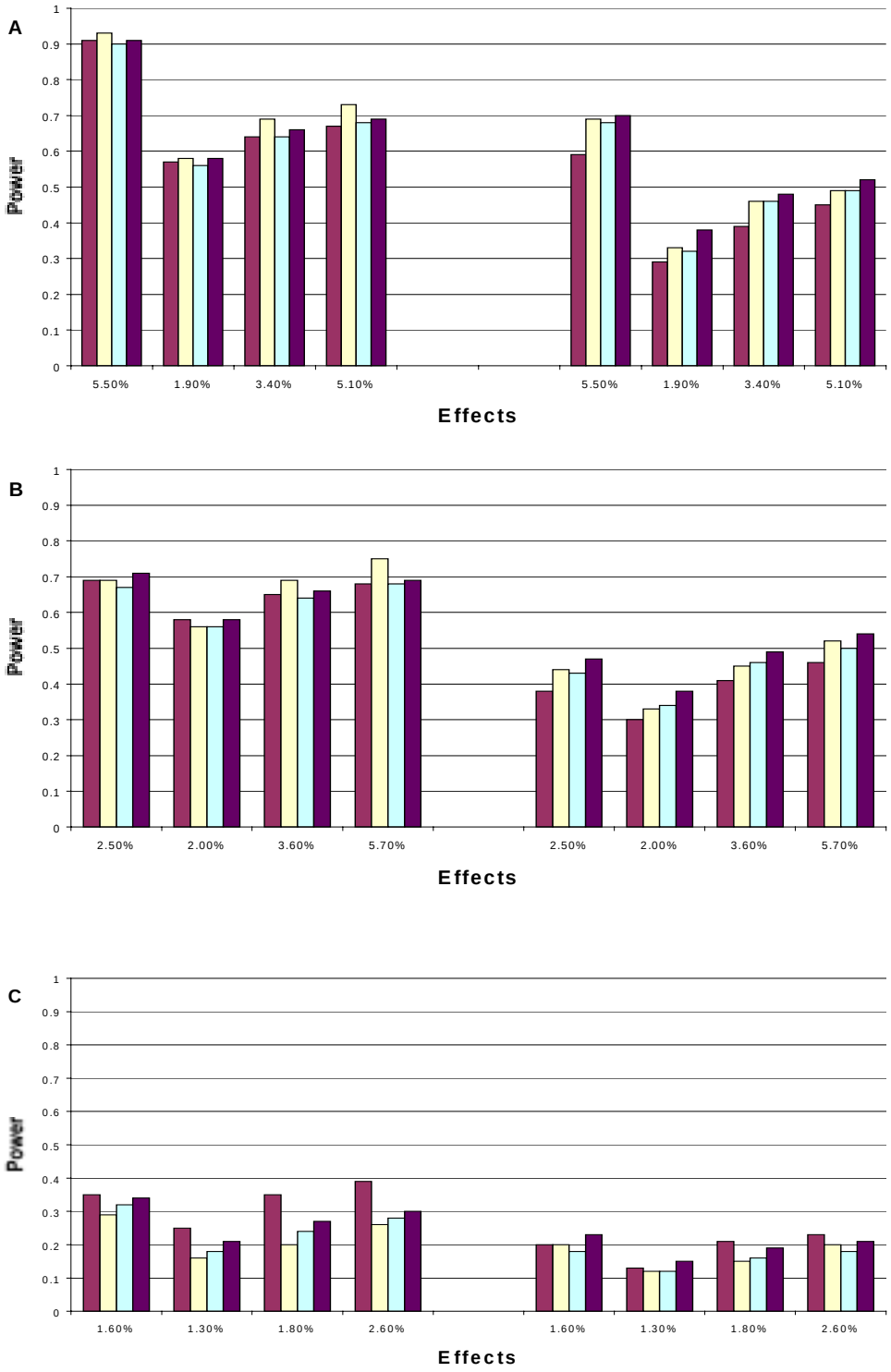

A
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5.50%
1.90%
3.40%
5.10%
5.50%
1.90%
3.40%
5.10%
Effects
B
Power
2.50%
2.00%
3.60%
5.70%
2.50%
2.00%
3.60%
5.70%
Effects
C
Power
1.60%
1.30%
1.80%
2.60%
1.60%
1.30%
1.80%
2.60%
Effects

Figure 4.5

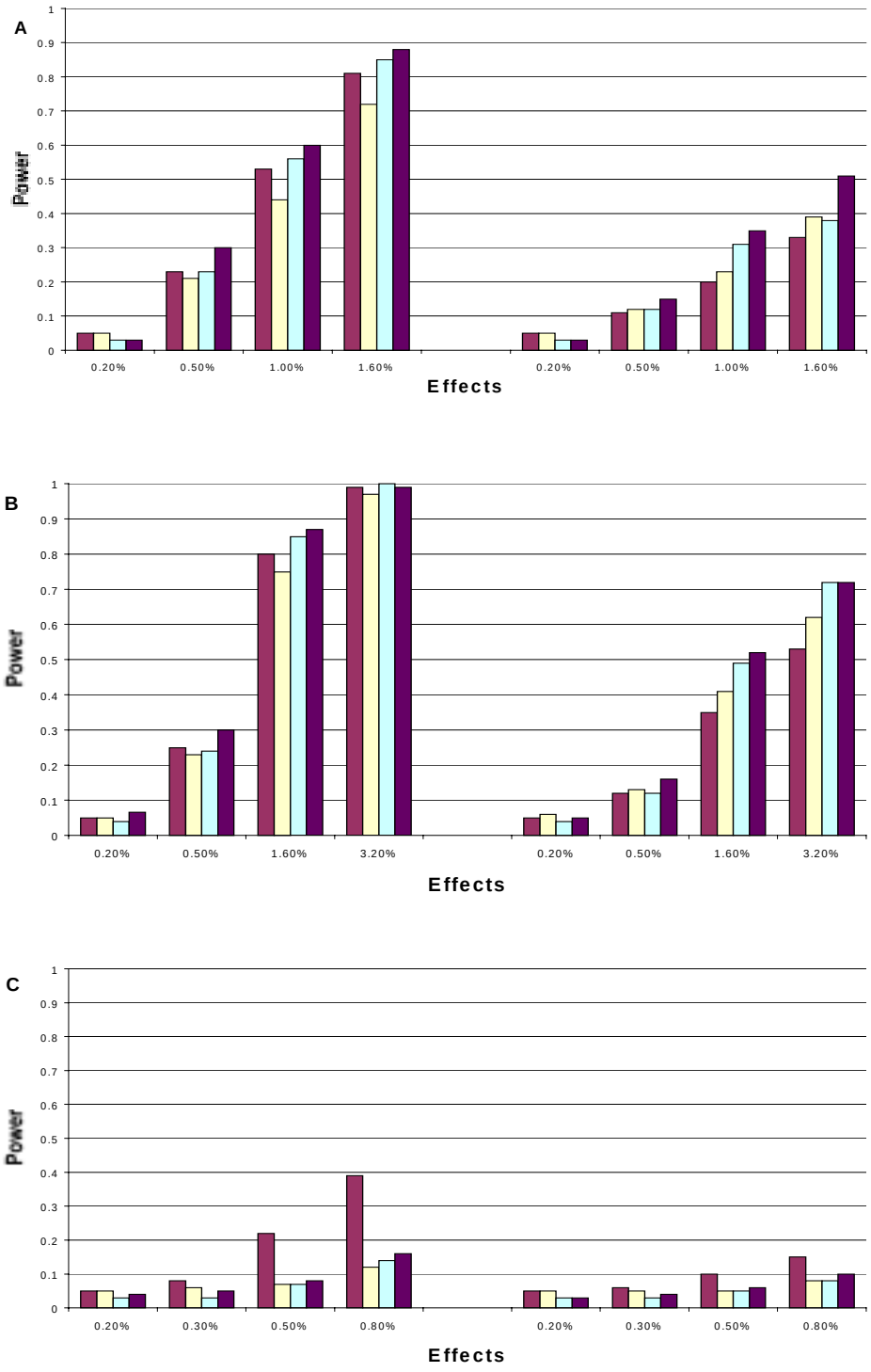
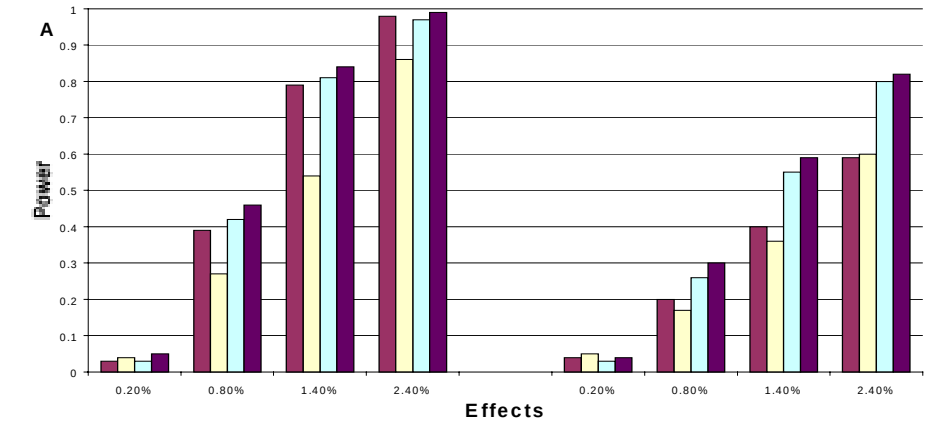
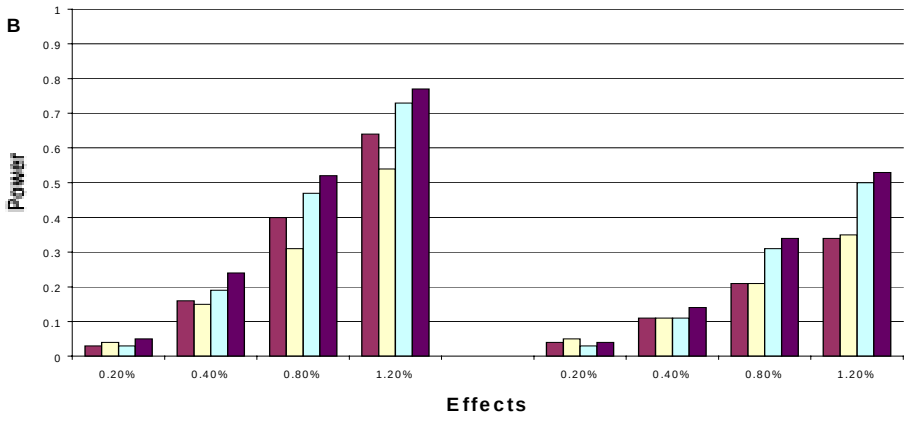

A
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.20%
0.50%
1.00%
1.60%
0.20%
0.50%
1.00%
1.60%
Effects
B
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.20%
0.50%
1.60%
3.20%
0.20%
0.50%
1.60%
3.20%
Effects
C
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.20%
0.30%
0.50%
0.80%
0.20%
0.30%
0.50%
0.80%
Effects

Figure 4.6


A
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.20%
0.80%
1.40%
2.40%
0.20%
0.80%
1.40%
2.40%
Effects


B
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.20%
0.40%
0.80%
1.20%
0.20%
0.40%
0.80%
1.20%
Effects

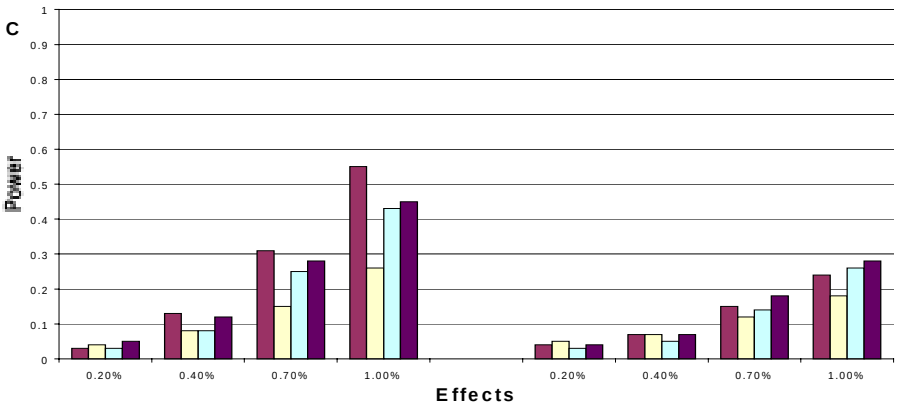
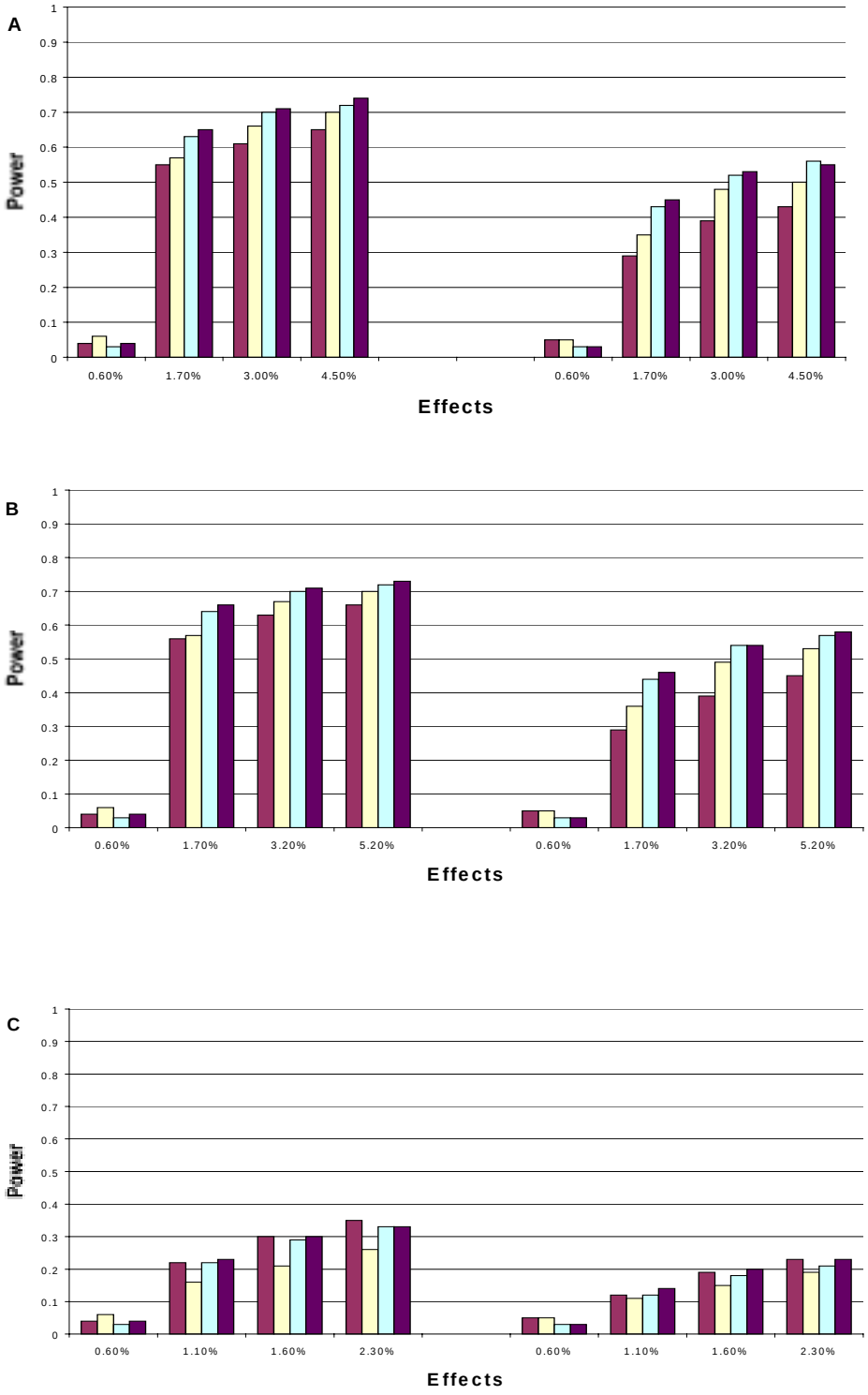

C
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.20%
0.40%
0.70%
1.00%
0.20%
0.40%
0.70%
1.00%
Effects

Figure 4.7


A
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.60%
1.70%
3.00%
4.50%
0.60%
1.70%
3.00%
4.50%
Effects
B
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.60%
1.70%
3.20%
5.20%
0.60%
1.70%
3.20%
5.20%
Effects
C
Power
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.60%
1.10%
1.60%
2.30%
0.60%
1.10%
1.60%
2.30%
Effects