ABSTRACT

NIELSEN, DAHLIA MICHELLE. An Examination of Association Based Tests for Localizing Genes in Outbred Populations. (Advisor: Bruce S. Weir)

Association based tests are designed to capitalize on evolutionary forces and population history in order to localize genes affecting the traits of interest to within very small regions. In the case-control test, a sample of affected individuals (the cases) and a matched set of unaffected individuals (the controls) are collected, and marker allele frequency differences between the two groups are compared. If a significant difference between allele frequencies is found, it is determined that there is an association between the marker and a disease susceptibility locus. One shortcoming of this test is that if the cases and controls are not well matched, or if the controls are chosen from different subpopulations than the cases, spurious associations may be detected within the samples which do not reflect actual population values. Additionally, it is possible that genotype information on a set of controls is simply not available. We explore the relationship between Hardy-Weinberg disequilibrium among affected individuals at a marker locus and linkage disequilibrium between the marker and a disease susceptibility locus and show that there is a connection between these disequilibrium measures which may be useful for detecting association using affected individuals only. As part of this work, we introduce two summary disequilibrium terms, one allelic and one genotypic, which appear as factors in various association-based measures.

Following up on several suggestive equations which led to the summary disequilibrium terms, we examine the relationship between phenotype and marker genotypes through the perspective of classical quantitative genetics. Within this

framework, we show that in a randomly mating population there is a simple connection between the additive effects of a marker locus and the additive effects of an associated trait locus. An equivalent relationship holds between the dominance deviations at the marker and the dominance deviations at the trait locus. These relationships are captured by the summary disequilibrium terms introduced earlier.

Using these results, we characterize the genetic properties that loci affecting a quantitative trait must express in order for common tests of association to be able to detect them. We examine the case-control test and the basic form of the transmission/disequilibrium test (TDT), and show that by focusing on alleles rather than on genotypes, these tests are sensitive mainly to additive genetic effects at the susceptibility loci. We offer several illustrations of the effectiveness of these tests in detecting association under various genetic models.

AN EXAMINATION OF ASSOCIATION BASED TESTS FOR LOCALIZING GENES IN OUTBRED POPULATIONS

by

Dahlia M. Nielsen

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

DEPARTMENT OF GENETICS

Raleigh
1999

APPROVED BY:

_____	_____
Chair of Advisory Committee	Co-Chair of Advisory Committee

BIOGRAPHY

Personal History

- Born November 15, 1966 in Salt Lake City, UT.

Education

- B.A. in Political Science, University of Utah, June 1990.
- B.S. in Computer Science, University of Utah, June 1992.
- M. Stat., North Carolina State University, August 1997.
- Ph.D. in Genetics, North Carolina State University, August 1999
  Dissertation topic: An Examination of Association Based Tests for Localizing Genes in Outbred Populations.

Professional Experience

- Programmer, Genmark, Inc., Salt Lake City, UT
  May 1991 – July 1993
- Programmer, University of Utah, Bioinformatics, Department of Human Genetics, Salt Lake City, UT
  January 1994 – March 1995
- Research Associate, Utah State University, Department of Animal, Dairy and Veterinary Sciences, Logan, UT
  December 1993 – March 1995
- Teaching Assistant at North Carolina State University, Raleigh, NC
  Spring 1998 and Fall 1998
- Graduate Research Internship at Glaxo Wellcome, Bioinformatics, Reasearch Triangle Park, NC
  May 1997 – February 1999
- Statistical Geneticist at Glaxo Wellcome, Bioinformatics, Reasearch Triangle Park, NC
  February 1999 – Current

ACKNOWLEDGEMENTS

For this work and all that it took to lead up to it, I am very grateful for the support and advice of my advisor, Bruce Weir. Thanks also to my committee, Trudy Mackay, Ted Emigh, and Sujit Ghosh, my substitute member, B.B. Bhattacharyya, and to my "honorary" member, Meg Ehm. None of this would have been possible had it not been for the influence and inspiration of Michel Georges. I especially thank my parents, Ken and Carmela Nielsen, and my brother, Erik, for always being there, in more ways than I can imagine.

I am indebted to all the members of the Program in Statistical Genetics for their support and friendship, particularly Dmitri Zaykin for numerous helpful conversations and Debbie Hibbard for endless encouragement and support.

I am grateful for the financial support from the NIH Training grant, and the GIT program at Glaxo Wellcome, Inc. This work was supported in part by NIH Grant GM45344 to NC State University.

Contents

LIST OF TABLES **vi**

LIST OF FIGURES **vii**

1 REVIEW **1**

1.1 Introduction 2

1.2 Genetic Mapping 3

1.2.1 Linkage Analysis 4

1.2.2 Association Studies 5

1.3 References 12

2 DETECTING MARKER-DISEASE ASSOCIATION BY TESTING FOR HARDY-WEINBERG DISEQUILIBRIUM AT A MARKER LOCUS **17**

2.1 Abstract 18

2.2 Introduction 19

2.3 Methods 22

2.3.1 Recessive Disease Model 22

2.3.2 General Disease Model 26
2.3.3 Test Statistics 29
2.3.4 Simulations 31
2.4 Results 33
2.4.1 Power 33
2.4.2 Size 34
2.5 Discussion 35
2.6 Acknowledgements 38
2.7 Appendix 1 39
2.8 Appendix 2 – Special Cases 41
2.8.1 Heterogeneous Recessive Model 41
2.8.2 General Diallelic Model 41
2.8.3 Additive Susceptibilities 41
2.8.4 Multiplicative Susceptibilities 42
2.9 References 43
2.10 Figures 46
2.11 Tables 49

3 A CLASSICAL SETTING FOR ASSOCIATIONS BETWEEN MARKERS AND LOCI AFFECTING QUANTITATIVE TRAITS 50

3.1 Abstract 51
3.2 Introduction 52
3.3 Methods 54

3.4 Tests for Association 57
3.4.1 Dichotomous Trait Case-Control Tests 57
3.4.2 Quantitative Trait Case-Control Tests 59
3.4.3 Dichotomous Trait Transmission/Disequilibrium Tests . . . 60
3.4.4 Quantitative Trait Transmission/Disequilibrium Tests 63
3.5 Discussion 67
3.6 Acknowledgments 69
3.7 References 70

4 EXAMINING THE GENETIC EFFECTS WHICH ALLELIC ASSOCIATION-BASED TESTS ARE DESIGNED TO DETECT 73

4.1 Abstract 74
4.2 Introduction 74
4.3 Methods 76
4.3.1 Case-Control Test 78
4.3.2 Transmission/Disequilibrium Test 81
4.3.3 Simulations 85
4.4 Results 90
4.5 Discussion 93
4.6 References 95

5 DISCUSSION 106

5.1 References 115

List of Tables

2.1 Parameters of the Simulated Disease Models 49

4.1 Parameters for the APOE Model 97

4.2 Parameters of Disease Models I 98

4.3 Parameters of Disease Models II 99

4.4 Marker Association Parameters I 100

4.5 Marker Association Parameters II 101

4.6 Results for Genetic Models I 102

4.7 Results for Genetic Models II 103

4.8 Marker Association Results I 104

4.9 Marker Association Results II 105

List of Figures

2.1 Power results for the chi-square tests for linkage disequilibrium. . . 46

2.2 Size of the the chi-square tests for linkage disequilibrium. 46

Chapter 1

REVIEW

1.1 Introduction

The search for genes affecting traits in humans has a long history, with its origins predating the discovery of DNA as the genetic material (Fisher 1934; Haldane 1934; Fisher 1935). Since this time, much progress has been made. Genes involved in various human diseases have been located (Riordan et al. 1989; Saunders et al. 1993; Miki et al. 1994), and their sequences determined. From this information, protein sequences of normal functioning and also incorrectly formed gene products have been analyzed, and their roles in biological processes have begun to be unraveled. Researchers now hope to be able to correct the conditions brought on by genetic disorders, either through direct genetic manipulation or with the use of individually tailored medications.

Much of the success to date has been in locating genes involved in simple mendelian traits, where single genes are responsible for the alternate phenotypes. Genes involved in complex traits – traits affected by multiple genes and also by environmental variations – have been much more difficult to isolate. However, the vast majority of the traits of interest fit this category (Moises 1995; Allison et al. 1998). More powerful methods of detection for genes involved in the complex traits, as well as a better understanding of the currently available techniques are necessary if these genes are to be effectively detected and isolated.

1.2 Genetic Mapping

Regions of DNA involved in regulation and coding for functional gene products lie along chromosomes at non-regular intervals, often with large amounts of non-coding DNA among and between them. The total number of genes in the human genome has been estimated to be between 50,000 and 100,000, with a total genome size of approximately 3,000,000,000 base pairs (Centre and Center 1998). Given these large values, the ability to isolate and narrow individual regions containing specific genes of interest is paramount if the actual DNA sequence of these genes is to be determined.

One of the major advances in the effort to localize genes was the discovery of widely dispersed genetic markers (Wyman et al. 1980; Nakamura et al. 1987; Litt and Luty 1989). These generally non-coding stretches of DNA are easily quantified, and are highly variable between individuals. They can be detected and their postions determined through molecular means, and thus provide ideal landmarks in determining the positions of the unobserved trait loci.

For genetic mapping studies, distances between loci are measured in terms of the rate of recombination between them. Morgan (1911) proposed this measure, noting that loci which were further apart on a chromosome were more likely to experience a crossover event between them. Loci which are unlinked, either on different chromosomes or far apart on the same chromosome, experience the maximum 50% recombination. For smaller distances between loci (less than 10% recombination), these recombination rates are roughly equivalent to physical distances, measured in units of centiMorgans. One centiMorgan is approximately one

megabase (1,000,000 bases) along a chromosome.

1.2.1 Linkage Analysis

A very successful area of analysis used for estimation of recombination rates between loci is linkage analysis. These methods utilize extended pedigrees which display segregation of the trait of interest, and test whether this segregation pattern of the trait is consistant with the segregation patterns of the markers (Lathrop et al. 1984; Kruglyak et al. 1996). Parametric tests assume a genetic model for the trait locus, such as autosomal dominant or X-linked recessive. Non-parametric approaches test whether the amount of allele sharing among individuals expressing the trait in the pedigree is consistant with a putative trait locus at the map location under observation. Recombination rate between markers and the putative trait locus are estimated by calculating the number of recombination events necessary to explain the pedigree given the total number of meioses which occurred within the pedigree.

These types of approaches are useful in estimating map positions for a gene of interest to within a few centiMorgans, but the accuracy of the estimates are limited to the number of meioses in the pedigree. Because of this limitation, very small distances between loci cannot be accurately estimated, so that fine-scale mapping is not feasible.

1.2.2 Association Studies

In order to define more localized regions containing the gene of interest, many association based methods have been developed. These methods rely on population history and evolution as a means to capture information from numerous generations of meioses; potentially orders of magnitudes more than the number of generations available in pedigree analysis.

In terms of genetic mapping, association is the tendency for marker alleles to be found more or less frequently than expected by chance among individuals in a population who share a common phenotype. The hope is that in sharing a phenotype, these individuals also share a common history, and it is this history which potentially provides information useful for mapping. As with linkage mapping, the goal of association mapping is to detect markers which are tightly linked to the locus affecting the trait of interest.

The statistical quantities of interest in these studies are the linkage disequilibria (often referred to simply as disequilibria) between alleles at the two loci. If one locus has alleles A_r at population frequencies p_r and the other locus has alleles M_i at frequencies q_i, disequilibrium between alleles A_r and M_i is defined as $D_{ri} = P_{ri} - p_r q_i$, where P_{ri} is the frequency of haplotype $A_r M_i$. This is of interest in genetic mapping, as in the absence of propagating forces, linkage disequilibrium decays over time (by means of recombination) according to the equation $D_{ri}^{(g)} = D_{ri}^{(0)}(1-c)^g$, where g is the number of generations since disequilibrium was created and c is the recombination rate between loci. By this equation, it is evident that linkage disequilibrium can provide a measure of recombination rate, hence distance,

between loci, as loci which are further apart should experience weaker degrees of disequilibrium than loci which are closer together.

Linkage disequilibrium can initially be created between loci by many evolutionary forces. In a small population, the random process of passing haplotypes from one generation to the next is variable enough that some haplotypes may, by chance, be inherited at proportions deviating from the product of their allele frequencies. If the population expands, this initial disequilibrium may become substantial in the larger population. A mutation which creates a new allele at one of the loci will create the same initial effect by creating a new haplotype in the population. If this occurs within a small but expanding population, this new haplotype can, by drift, become more prevalent in the population than expected by chance. Other forces which can create disequilibrium between loci include selection of certain haplotypes and population admixture. Selection for certain haplotypes creates disequilibrium which does not necessarily decay over time, provided the selecting circumstances remain. This maintenance of disequilibrium from generation to generation can interfere with association-based mapping principles, which rely on decay of disequilibrium over time. Events such as population admixture can create disequilibrium in an initial combined population which only begins to decay after several generations of random mating within the whole population (Ewens and Spielman 1995). This, too, can create spurious associations between alleles which may cause tests of association to give false positive results.

If the alleles at both loci are observable, linkage disequilibrium can be estimated directly (Hill 1974; Weir and Cockerham 1979). When the alleles at one of the loci cannot be observed, such as those of the unknown loci which are to be mapped,

these quantities cannot be directly estimated. Instead, these measures must be determined indirectly via marker-phenotype associations.

Many measures and tests have been proposed for detection of marker-disease assocations. The case-control test compares the marker allele frequencies found among a sample of affected individuals (the cases) with the frequencies found amoung a sample of unaffected individuals (the controls). If significant differences are found between these allele frequencies, it is viewed as evidence that the marker is in disequilibrium with a disease susceptibility locus. The measure P_{excess} (Lehesjoki, et al. 1993) was proposed as a means of comparing the degrees of associations of different markers.

In performing the case-control test, it is important that the cases and controls are drawn from the same underlying population. If this does not occur, differences between the two samples may reflect differences in allele frequencies between populations rather than associations between the marker and the disease. To address this issue, several alternative strategies have been proposed which utilize family-based information as the "control" information. These tests include the haplotype relative risk (HHR) (Rubinstein et al. 1981; Falk and Rubinstein 1987; Knapp 1993), the haplotype-based haplotype relative risk test (HHRR) (Terwilliger and Ott 1992), and the affected family-based control test (AFBAC) (Thomson 1995). By collecting parents along with affected individuals, two sets of genotypes can be created for each parent-offspring trio. The "affected" genotypes (or cases) are the genotypes of the affected offspring. The "unaffected" genotypes (or the controls) are composed of the alleles which were not transmitted from the parents to the offspring. This produces a set of case and control genotypes which are not

susceptible to events such population stratification or admixture.

In Chapter 2 we propose a method when only unrelated affected individuals are available. We demonstrate a relationship between linkage disequilibrium between a marker and a disease susceptibility locus in the whole population and Hardy-Weinberg disequilibrium at the marker locus among affected individuals. Drawing a sample of affected individuals causes selection of individuals based on genotypes, not individual alleles. Thus, when the overall population is in Hardy-Weinberg equilibrium, sampling individuals in this manner can create Hardy-Weinberg disequilibrium within the sample of affected individuals. This effect will be seen not only at the disease susceptibility locus itself, but also at marker loci which are in linkage disequilibrium with the disease susceptibility locus. The amount of deviation which can be expected depends on the amount of linkage disequilibrium in the whole population and on the manner in which the disease susceptibility locus acts. If the overall population is in Hardy-Weinberg equilibrium, a significant finding of Hardy-Weinberg disequilibrium within the affected individuals may be considered evidence for non-zero linkage disequilibrium between the marker and a disease susceptibility locus.

Since linkage disequilibrium can be created between loci which are not linked, conclusions regarding the position of a disease susceptibility locus cannot be made based on case-control or case-only evidence alone. To address this issue, the transmission/disequilibrium test (TDT) was introduced to test for linkage and linkage disequilibrium between a diallelic marker and a disease susceptibility locus (Spielman et al. 1993; Ewens and Spielman 1995). In the basic form of this test, trios containing affected offspring and their parents are collected and genotyped for the

marker loci. Each heterozygous parent is examined to determine which marker allele it transmitted to the affected offspring. Affected offspring are more likely to have received alleles at the trait locus which increase susceptibility to disease. For a linked marker locus, any marker alleles that are in positive disequilibrium with the susceptibility alleles will also be found more often among the affected offspring. Therefore, evidence for unequal transmission of alleles at the marker locus indicates linkage disequilibrium and linkage between the marker and the disease susceptibility locus. This result potentially offers a finer resolution than linkage analysis alone, as it is assumed that only markers which are very close to a susceptibility locus will maintain non-zero linkage disequilibrium with that locus in a randomly mating population.

Many extentions to the TDT have been proposed. Bickeböller et al. (1995) extended the test to consider markers with multiple alleles. Trégouët et al. (1997) proposed using estimating equations to estimate association parameters in samples of nuclear families of varying sizes and mixtures of related and unrelated individuals. Martin et al. (1997) proposed two test statistics for association that use data from all affected children (and their parents) in a nuclear family. Spielman and Ewens (1998) proposed a test statistic that tests for linkage disequilibrium using affected and unaffected siblings when parent data are not available. These tests have also been extended to the analysis of quantitative traits (Allison 1997; Rabinowitz 1997; Martin 1997).

All of these tests examine alleles individually, rather than as genotypes, though it is genotypes, rather than individual alleles, which generally affect phenotype. In Chapter 3, we propose a classical genetics framework whereby the consequences

of examining genetic data based on observations at the alleles alone can be determined. We considering a linear model defining genetic effects in terms of the additive effects of alleles and dominance deviations. The genetic effects are considered in terms of the marker genotypes, which can be observed. Using the least squares solutions, we show that in a randomly mating population, the additive effects of a marker locus are a function of the additive effects of an associated trait locus and the degree of linkage disequilibrium between the loci. Similarly, the dominance deviations at the marker locus are a function of the dominance deviations at the trait locus and the linkage disequilibrium between loci.

By examining the properties of the statistics involved in allele-based association tests, we can determine to which genetic effects these tests are sensitive. In Chapter 3 we examine the case-control test, the basic form of the TDT for dichotomous traits, and a basic form of the TDT for quantitive traits. We find that in a randomly mating population, the expected values of the statistics are sensitive only to the additive effects of the trait loci. However, the variances of these statistics are sensitive to both additive and dominance effects. This implies that while the factor that has the largest effect on the power of a test to detect a trait locus is the additive genetic effect of that locus, power may be increased (or decreased) beyond the nominal α level by dominance deviations. Strength of linkage disequilibrium alone is not sufficient in determining the strength of association between a marker and a trait phenotype; the manner in which the trait locus acts to affect phenotype is also important. Even markers in strong disequilibrium with trait loci with large additive components may not exhibit strong association with the trait phenotype, as the overall level of assocation must be captured by the marker through summary

terms which confound disequilibria and the genetic effects.

In Chapter 4 we extend our results for tests of dichotomous traits, examining more closely the effects of the additive and dominance components of the trait locus on the expected values and the variances of the statistics involved in tests of assocation. We provide some simulations under various genetic models to illustrate our results.

1.3 References

Allison DB (1997) Transmission-Disequilibrium Tests for quantitative traits. Am J Hum Genet 60(3):676–90.

Allison DB, Thiel B, St Jean P, Elston RC, Infante MC, Schork NJ (1998) Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. Am J Hum Gen 63(4):1190–201

Bickeböller H, Clerget-Darpoux F (1995) Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. Genet Epidem 12:865–870

Centre TS, Center TWUGS (1998) Toward a Complete Human Genome Sequence. Genome Res 8:1097–1108

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision and admixture. Am J Hum Genet 57: 455–464

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a control sample for risk calculations. Am J Hum Genet 51: 227-233

Fisher RA (1934) The amount of information supplied by records of families as a function of the linkage in the population sampled. Ann Eugen 6:66–70

Fisher RA (1935) The detection of linkage with 'dominant' abnormalities. Ann Eugen 6:187–201

Haldane JBS (1934) Methods for the detection of autosomal linkage in man. Ann Eugen 6:26–65

Hästbacka J, de la Chappelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nature Genet 2:204–211

Hill WG (1974) Estimation of linkage disequilibrium in randomly mating population. Heredity 33:229–239

Kaplan NL, Martin ER, Weir BS (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 60:691-702

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. Daly MJ. Reeve-Daly MP. Lander ES. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Gen 58(6):1347-63

Lathrop GM, Lalouel JM, Julier C, Ott J (1984) Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci 81:3443-3446

Lehesjoki A-E, Koskiniemi M, Norio R, Tirrito S, Sistonen P, Lander E, de la Chapelle A (1993) Localization of the EPM1 gene for progressive myoclonus

epilepsy on chromosome 21: Linkage disequilibrium allows high resolution mapping. Hum Mol Genet 2:1229-1234

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet. 44(3):397–401

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61:439–448

Martin ER (1997) Extensions of the transmission/disequilibrium test for identifying human disease genes. Ph.D. Dissertation, North Carolina State University, Chapter 3 (pp 48–86).

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W, Bell R, Rosenthal J, et al. (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266:66–71

Moises HW, Yang L, Kristbjarnarson H, Wiese C, Byerley W, Macciardi F, Arolt V, Blackwood D, Liu X, Sjogren B, et al. (1995) An international two-stage genome-wide search for schizophrenia susceptibility genes. Nat Gen 11(3):321–4

Morgan TH (1911) Random segregation versus coupling in mendelian inheritance. Science 34:384

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

Ott J (1999) Analysis of Human Genetic Linkage. The Johns Hopkins University Press, Baltimore, MD

Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. Human Hered 47(6):342–50.

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, et al. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245:1066–73

Rubinstein P, Walker M, Carpenter C, Carrier C, Krassner J, Falk C Ginsberg F (1981) Genetics of HLA associations: the use of haplotype relative risk (HRR) and the "haplo-delta" (Dh) estimates in juvenile diabetes from three racial groups. Hum Immunol 3:384

Saunders AM, Schmader K, Breitner JC, Benson MD, Brown WT, Goldfarb L, Goldgaber D, Manwaring MG, Szymanski MH, McCown N, et al. (1993) Apolipoprotein E epsilon 4 allele distributions in late-onset Alzheimer's disease and in other amyloid-forming diseases. Lancet 342:710–1

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus

(IDDM). Am J Hum Genet 52:506–516

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the S-TDT. Am J Hum Genet 62:450–458

Thomson G (1995) Mapping disease genes: family-based association studies. Am J Hum Genet 57:487-498

Trégouët D-A, Ducimetire P, Tiret L (1997) Testing association between candidate-gene markers and phenotype in related individuals, by use of estimating equations. Am J Hum Genet 61:189–199

Wyman AR, White R (1980) A highly polymorphic locus in human DNA. Proc Nat Ac Sci 77(11):6754–8

Weir BS, Cockerham CC (1979) Estimation of linkage disequilibrium in randomly mating populations. Heredity 42:105–111

Weir BS (1996) Genetic Data Analysis II. Sinauer, Sunderland, MA

Chapter 2

DETECTING MARKER-DISEASE ASSOCIATION BY TESTING FOR HARDY-WEINBERG DISEQUILIBRIUM AT A MARKER LOCUS

Nielsen DM, Ehm MG and Weir BS (1998)

Am J Hum Genet 63:1531–1540

2.1 Abstract

We review and extend a recent suggestion that fine-scale localization of a susceptibility locus for a complex disease be done on the basis of deviations from Hardy-Weinberg equilibrium among affected individuals. This deviation is driven by linkage disequilibrium between disease and marker loci in the whole population, and requires a heterogeneous genetic basis for the disease. A finding of marker-locus Hardy-Weinberg disequilibrium therefore implies disease heterogeneity and marker-disease linkage disequilibrium. Although a lack of departure of Hardy-Weinberg disequilibrium at marker loci implies that susceptibility-weighted linkage disequilibria are zero, given disease heterogeneity, it does not follow that the usual measures of linkage disequilibrium are zero. For susceptibility loci with more than two alleles, therefore, care is needed with drawing inferences from marker Hardy-Weinberg disequilibria.

2.2 Introduction

We will refer to fine mapping as attempting to narrow what may be a 10 cM region indicated by linkage analysis to approximately a 1 cM or less region containing the susceptibility locus. Fine mapping methods for qualitative and quantitative phenotypic traits have been under constant development in recent years. Simple Mendelian traits with high penetrance are often fine mapped by recombinant mapping: typing markers every 1-2 cM, determining haplotypes using extended family information, and identifying recombination events on either side of the supposed susceptibility locus (Boehnke 1994). Glaser et al. (1995) illustrate this approach in a search for the gene responsible for familiar hyperinsulinism. In the absence of high penetrance or sufficient numbers of patients, linkage disequilibrium methods in isolated populations have been used. Hästbacka et al. (1992) used linkage disequilibrium to map diastrophic dysplasia (DTD) in Finland, and indicated that the DTD gene should lie within 0.06 cM of the CSF1R gene which was later confirmed (Hästbacka et al. 1994). While the identification of susceptibility loci for complex traits has been slow, several fine mapping methods have been employed. Identification of IDDM2 (insulin gene) was accomplished using linkage disequilibrium (Bennett et al. 1995). Association studies played a key role in implicating the apolipoprotein E gene in late-onset Alzheimer's disease and heart disease (Corder et al. 1993).

While methods for mapping simple Mendelian diseases utilize extended families collected for the genomic scan to refine disease gene locations, fine mapping techniques for complex diseases utilize samples with varying characteristics. Several

methods have been proposed. Association methods utilize unrelated case and controls. Traditional transmission disequilibrium tests require affected children and their parents (Spielman et al. 1993, Kaplan et al. 1997). Several tests utilizing data sets that are simpler to collect and are more similar to those used a genomic scan have been proposed. Trégouët et al. (1997) proposed using estimating equations to estimate association parameters in samples of nuclear families of varying sizes and mixtures of related and unrelated individuals. Martin et al. (1997) proposed two test statistics for association that use data from all affected children (and their parents) in a nuclear family. Spielman and Ewens (1998) proposed a test statistic that tests for linkage disequilibrium using affected and unaffected siblings. Feder et al. (1996) suggested that fine localization of a susceptibility locus could be accomplished using deviation from Hardy-Weinberg equilibrium among affected individuals.

Feder et al. (1996) studied hereditary haemochromatosis (HH), a common autosomal recessive disorder of iron metabolism. As described in their paper, previous localization of the HH gene placed it near the major histocompatibility complex on chromosome 6p and within 1-2 cM of the HLA-A gene, although many of the reports have been contradictory. Linkage disequilibrium studies confirmed the existence of a founder effect. To proceed with the localization of a gene involved in this disease, Feder et al. (1996) developed 45 STRP and SNP markers lying within an 8 cM region suspected of containing the gene. All 45 markers were typed on 101 HH patients and 64 controls.

To estimate the position of the gene relative to these closely spaced markers, Feder et al. (1996) used the measure p_{excess} (Lehesjoki, et al. (1993); Bengtsson and

Thomson 1981). This is a measure of linkage disequilibrium that in the presence of linkage is expected to be maximized at the marker nearest the gene. Feder et al. (1996) plotted p_{excess} for each marker along the marker map. This plot had a peak representing the maximum p_{excess} in the region; however, the peak was not very sharp, causing concern regarding the accuracy of the results using this measure alone.

In examining the data used in this study, Feder et al. (1996) noted that among the affected individuals, there appeared to be an excess of homozygosity at the marker loci. They considered several explanations for this; the most likely of which proved to provide the basis for a new measure for linkage disequilibrium. They noted that for heterogeneous recessive traits such as theirs, not only will an excess of homozygosity exist among affected individuals, but also this excess homozygosity should decrease with decreased linkage disequilibrium between the marker loci and the susceptibility locus. A disequilibrium measure based on this observation has the advantage that only affected individuals need to be collected and genotyped, as opposed to the case-control type studies necessary for most measures of association, such as p_{excess}. In their paper, Feder et al. (1996) plotted a measure of Hardy-Weinberg disequilibrium within their HH affected individuals for each marker along the marker map. This plot had a maximum at approximately the same point in the map as p_{excess}, but the peak was much sharper. From this, Feder et al. (1996) concluded that their initial results from the p_{excess} measures had been confirmed, and the region in which the gene lies had been more accurately defined.

We were impressed with these results and were interested in exploring the

properties of this measure. We have extended the model and have examined some general results for this and other measures. We also compare a test for Hardy-Weinberg disequilibrium to a direct test for linkage disequilibrium.

2.3 Methods

2.3.1 Recessive Disease Model

Feder et al. (1996) examined a heterogeneous recessive model in which a subset of the disease cases are due to a mutation in the region of interest, while other disease cases are due to unrelated genetic loci or to non-genetic factors. Using A to denote the disease allele and $\bar{A}$ to denote all other alleles at the susceptibility locus, this model can be summarized as:

$$\begin{aligned} \Pr(\text{Affected}|AA) &= 1 \\ \Pr(\text{Affected}|A\bar{A}) &= \psi \\ \Pr(\text{Affected}|\bar{A}\bar{A}) &= \psi \end{aligned}$$

where ψ is the probability that an individual will exhibit the disease due to causes other than this locus. Assuming random mating in the population, genotype and allele probabilities at the susceptibility locus among affected individuals can be calculated and include

$$\begin{aligned} \Pr(AA|\text{Affected}) &= P_{AA|\text{Aff.}} = p_A^2/\phi \\ \Pr(A|\text{Affected}) &= p_{A|\text{Aff.}} = p_A(p_A + \psi p_{\bar{A}})/\phi, \end{aligned}$$

where ϕ is the prevalence of the disease in the population. We have used $p_{A|\text{Aff.}}$ and $P_{AA|\text{Aff.}}$ to distinguish frequencies among affected individuals from p_A and P_{AA}, the whole-population frequencies. For this model, $\phi = p_A^2 + \psi(1 - p_A^2)$.

Departure from Hardy-Weinberg equilibrium at the susceptibility locus can be measured by the disequilibrium coefficient $\mathcal{D}_{AA} = P_{AA} - p_A^2$ (Weir 1996). Among affected individuals, this coefficient becomes

$$\begin{aligned} \mathcal{D}_{AA|\text{Aff.}} &= P_{AA|\text{Aff.}} - p_{A|\text{Aff.}}^2 \\ &= \psi(1-\psi)p_A^2(1-p_A)^2/\phi^2. \end{aligned}$$

Feder et al. (1996) quantified departure from Hardy-Weinberg equilibrium at the susceptibility locus with a measure F_A, defined as $(H_o - H_e)/(1 - H_e)$, where H_o and H_e are observed and expected homozygosities. Although they did not give an explicit expression for this quantity, it appears to us that they used the formulation

$$\begin{aligned} F_A &= \frac{P_{AA|\text{Aff.}} + P_{\bar{A}\bar{A}|\text{Aff.}} - p_{A|\text{Aff.}}^2 - p_{\bar{A}|\text{Aff.}}^2}{1 - p_A^2 - p_{\bar{A}}^2} \\ &= 2\mathcal{D}_{AA|\text{Aff.}}/[2p_A p_{\bar{A}}] \\ &= \psi(1-\psi)p_A p_{\bar{A}}/\phi^2. \end{aligned}$$

Association between the susceptibility allele A and a marker allele M can be expressed using the linkage disequilibrium measure $D_{AM} = P_{AM} - p_A q_M$, where q_M is the frequency of marker allele M. This quantity compares the frequency (P_{AM}) of haplotypes carrying both alleles A and M with the product of the separate frequencies of the two alleles. D_{AM} is positive when marker allele M is more likely to be associated with susceptibility allele A than would be expected by chance.

Feder et al. (1996) also discussed Hardy-Weinberg disequilibrium at a diallelic marker locus. Assuming random mating in the whole population, probabilities for the marker alleles and marker genotypes conditioned on having the disease include

$$\begin{aligned} P_{MM|\text{Aff.}} &= [(1-\psi)(p_A q_M + D_{AM})^2 + \psi q_M^2]/\phi \\ q_{M|\text{Aff.}} &= [\psi q_M + (1-\psi)p_A(p_A q_M + D_{AM})]/\phi. \end{aligned}$$

The Hardy-Weinberg disequilibrium coefficient at the marker locus among affected individuals is

$$\mathcal{D}_{MM|\text{Aff.}} = \psi(1-\psi)D_{AM}^2/\phi^2. \tag{2.1}$$

This is non-zero only if ψ is neither one nor zero, implying the disease must be heterogeneous, and if there is linkage disequilibrium, $D_{AM} \neq 0$. The Hardy-Weinberg departure measure of Feder et al. (1996) for the marker locus is

$$\begin{aligned} F_M &= \frac{P_{MM|\text{Aff.}} + P_{\bar{M}\bar{M}|\text{Aff.}} - q_{M|\text{Aff.}}^2 - q_{\bar{M}|\text{Aff.}}^2}{1 - q_M^2 - q_{\bar{M}}^2} \\ &= \mathcal{D}_{MM|\text{Aff.}}/[q_M q_{\bar{M}}] \\ &= \psi(1-\psi)D_{AM}^2/[\phi^2 q_M q_{\bar{M}}]. \end{aligned}$$

As stated by Feder et al. (1996),

$$F_M = \Delta_{AM}^2 F_A \tag{2.2}$$

where

$$\Delta_{AM}^2 = \frac{D_{AM}^2}{p_A p_{\bar{A}} q_M q_{\bar{M}}}.$$

Equations 2.1 and 2.2 capture the essential point that Hardy-Weinberg disequilibrium at a marker locus among affected individuals depends on the whole-population linkage disequilibrium between the marker and the disease locus. Although it is the latter quantity that is of interest, it is easier to test for the former. A test for Hardy-Weinberg disequilibrium at the marker locus can serve as a test for linkage disequilibrium. It should be noted, however, that the measure F_M proposed by Feder et al. (1996) depends on the values q_M and $q_{\bar{M}}$ which are whole-population parameters and cannot be estimated using affected individuals alone.

A common direct measure of linkage disequilibrium is the quantity p_{excess} (Lehesjoki, et al. (1993); Bengtsson and Thomson 1981). This measure compares the frequency of a marker allele M among affected individuals ($q_{M|\text{Aff.}}$) to the frequency among unaffected individuals ($q_{M|\text{Unaff.}}$). It is defined as

$$p_{\text{excess}} = \frac{q_{M|\text{Aff.}} - q_{M|\text{Unaff.}}}{1 - q_{M|\text{Unaff.}}}.$$

For the model of Feder et al. (1996),

$$\begin{aligned} q_{M|\text{Aff.}} &= [\psi q_M + (1-\psi)p_A(p_A q_M + D_{AM})]/\phi \\ q_{M|\text{Unaff.}} &= (1-\psi)[q_M - p_A(p_A q_M + D_{AM})]/(1-\phi) \end{aligned}$$

so that

$$p_{\text{excess}} = \frac{(1-\psi)p_A D_{AM}}{\phi(1-\phi)\left(q_{\bar{M}} + \dfrac{(1-\psi)p_A D_{AM}}{(1-\phi)}\right)}. \tag{2.3}$$

Therefore p_{excess} is proportional to D_{AM} and it reaches its maximum at the marker with the greatest disequilibrium with the disease. Note that ψ must be less than

one. Hardy-Weinberg disequilibrium is proportional to the square of disequilibrium so that F_M is expected to be a more sensitive indicator of linkage in the presence of linkage disequilibrium (Equation 2.2). This appears to have been the case in the analyses reported by Feder et al. (1996).

2.3.2 General Disease Model

We wished to know if Equation 2.1 might be generalized to other disease models, and so we considered a more general model with susceptibility affected by a locus with an arbitrary number of alleles, denoted by A_r. Under this model, the conditional probability of an individual having the disease, given that the individual has genotype A_rA_s at the susceptibility locus, is ϕ_{rs}. We will refer to these values as penetrances, though we recognize that for some of the A_rA_s genotypes, the ϕ_{rs} values should properly be called phenocopy rates. These values could equivalently be called prevalences: they represent the prevalence of the disease within a genotypic class. This relates the notation ϕ_{rs} to the use of ϕ, the whole-population prevalence (the unconditional probability of an individual having the disease). This value is

$$\phi = \sum_r \sum_s \phi_{rs} p_r p_s$$

where p_r is the population frequency of allele A_r at the susceptibility locus, and Hardy-Weinberg equilibrium is assumed.

In addition to the genotypic penetrances ϕ_{rs}, we find it convenient to define an allelic penetrance, ϕ_r:

$$\phi_r = \sum_s p_s \phi_{rs}$$

which is the conditional probability an individual will have the disease, given that the individual has allele A_r (the other allele being a random allele from the population). Note that $\phi = \sum_r p_r \phi_r$.

We consider a marker locus with alleles M_i occuring at frequencies q_i. For such a marker, we are likely to concentrate on those alleles that show a positive association with the disease, meaning that they have a higher frequency among affected than among unaffected individuals. If P_{ri} is the population frequency of haplotypes carrying susceptibility allele A_r and marker allele M_i, then the population linkage disequilibrium D_{ri} between these alleles is defined by

$$D_{ri} \quad = \quad P_{ri} - p_r q_i.$$

These coefficients sum to zero over all the alleles at either locus, so that $\sum_r D_{ri} = \sum_i D_{ri} = 0$. We also wish to describe the linkage disequilibrium between marker allele M_i and the disease locus as a whole, and do so by weighting the D_{ri} terms by the allelic penetrances. This measure is written as δ_i:

$$\delta_i \quad = \quad \sum_r \phi_r D_{ri} \quad = \quad \sum_r \sum_s p_s \phi_{rs} D_{ri}$$

and sums to zero over i. The quantity δ_i is zero if all susceptibility-locus alleles have the same penetrances.

For a susceptibility locus with two alleles, A_1 and A_2, δ_i is a multiple of D_{1i} and so is proportional to the usual linkage disequilibrium coefficient, and will maximize at the same point as does linkage disequilibrium. In this two-allele case, if the penetrances are not the same, a zero value of δ_i implies that there is no linkage disequilibrium between susceptibility and marker loci. For a susceptibility

locus with more than two alleles, however, it is possible for δ_i to be near or equal zero even when there is linkage disequilibrium since the D_{ri} do not all have the same sign and may have a (penetrance-weighted) sum close to zero.

The penetrance-weighted linkage disequilibrium coefficient allows simple expressions for marker allele frequencies among affecteds:

$$q_{i|\text{Aff.}} = q_i + \frac{\delta_i}{\phi}$$

as is shown in Appendix 1. This equation shows that marker allele frequencies among affected individuals deviate from the overall population frequencies by an amount which depends on the strength of association between the marker allele and the susceptibility alleles, weighted by the penetrances of those alleles. A similar expression holds for the marker allele frequency among unaffected individuals:

$$q_{i|\text{Unaff.}} = q_i - \frac{\delta_i}{1-\phi}$$

so that in the whole population, $q_i = \phi q_{i|\text{Aff.}} + (1-\phi) q_{i|\text{Unaff.}}$.

As a generalization of Equation 2.3, the quantity p_{excess} for marker allele M_i becomes

$$p_{\text{excess}_i} = \frac{\delta_i}{\phi(1-\phi)\left((1-q_i) + \frac{\delta_i}{1-\phi}\right)}. \tag{2.4}$$

If M_i is a marker allele showing a positive association with the disease $p_{\text{excess}_i} \geq 0$, so that $\delta_i \geq 0$ and these two quantities are maximized together. However, it is not necessary that each individual linkage disequilibrium coefficient D_{ri} is positive.

For the general disease model, discussion of marker-locus Hardy-Weinberg disequilibrium requires an additional summary measure of linkage disequilibrium. This

quantity, δ_{ij}, is defined for pairs of marker alleles, M_i, M_j, instead of single marker alleles

$$\delta_{ij} \quad = \quad \sum_r \sum_s \phi_{rs} D_{ri} D_{sj}.$$

We term it "genotypic disequilibrium" as opposed to the "allelic disequilibrium" δ_i. Note that $\sum_i \delta_{ij} = \sum_j \delta_{ij} = 0$. Among affected individuals, the marker locus homozygote Hardy-Weinberg disequilibrium coefficients can now be written as

$$\mathcal{D}_{ii|\text{Aff.}} \quad = \quad P_{ii|\text{Aff.}} - q^2_{i|\text{Aff.}} = \frac{\phi\delta_{ii} - \delta_i^2}{\phi^2} \tag{2.5}$$

and heterozygote disequilibria (Weir 1996) are

$$\mathcal{D}_{ij|\text{Aff.}} \quad = \quad P_{ij|\text{Aff.}} - 2q_{i|\text{Aff.}} q_{s|\text{Aff.}} = \frac{2(\phi\delta_{ij} - \delta_i\delta_j)}{\phi^2}.$$

For this more general model, it is not clear that Hardy-Weinberg disequilibria, $\mathcal{D}_{ij|\text{Aff.}}$, $\mathcal{D}_{ii|\text{Aff.}}$, are maximized when linkage disequilibria, δ_i, are maximized. It is clear, however, that some patterns of non-zero linkage disequilibrium will result in zero departure from Hardy-Weinberg at a marker locus. Conversely, a departure from Hardy-Weinberg at a marker locus provides evidence both for linkage disequilibrium between marker and susceptiblity loci and heterogeneity of susceptibility.

2.3.3 Test Statistics

We have discussed two measures that can be used to characterize marker disease associations. One is p_{excess} which is directly proportional to linkage disequilibrium measured on unrelated affected and unaffected individuals and the other is the Hardy-Weinberg disequilibrium coefficient measured among affected individuals.

To compare these two approaches we consider the statistical power of corresponding test statistics.

A widely used statistical test for association using unrelated affected and unaffected individuals, i.e., a case-control design, is the $(m-1)$ df chi-square test based on the statistic X^2_{CC} when the marker locus has m alleles. When the marker alleles have sample frequencies $\tilde{p}_{i|\text{Aff.}}$ and $\tilde{p}_{i|\text{Unaff.}}$ among n affecteds and n unaffecteds:

$$X^2_{CC} = 2n \sum_i \frac{(\tilde{p}_{i|\text{Aff.}} - \tilde{p}_{i|\text{Unaff.}})^2}{\tilde{p}_{i|\text{Aff.}} + \tilde{p}_{i|\text{Unaff.}}}. \tag{2.6}$$

When alternatives to the null hypothesis of no disequilibrium are of the Pitman type (i.e., departures tend to zero with sample size), the non-centrality parameter of this statistic is (Meng and Chapman, 1966)

$$\begin{aligned} \lambda_{CC} &= 2n \sum_i \frac{(q_{i|\text{Aff.}} - q_{i|\text{Unaff.}})^2}{q_{i|\text{Aff.}} + q_{i|\text{Unaff.}}} \\ &= 2n \sum_i \frac{\delta_i^2}{\phi^2(1-\phi)^2 \left(2q_i + \frac{(1-2\phi)\delta_i}{\phi(1-\phi)}\right)}. \end{aligned}$$

We have previously (Kaplan et al. 1997) written the sum in this expression as I^*.

To test for Hardy-Weinberg at the marker locus among the same total number of individuals, $2n$ affecteds, the test statistic X^2_{HW} is (Weir 1996)

$$X^2_{HW} = n \sum_i \frac{(\tilde{P}_{ii|\text{Aff.}} - \tilde{q}^2_{i|\text{Aff.}})^2}{\tilde{q}^2_{i|\text{Aff.}}} + 2n \sum_{i<j} \frac{(\tilde{P}_{ij|\text{Aff.}} - 2\tilde{q}_{i|\text{Aff.}}\tilde{q}_{j|\text{Aff.}})^2}{2\tilde{q}_{i|\text{Aff.}}\tilde{q}_{j|\text{Aff.}}}. \tag{2.7}$$

This has $m(m-1)/2$ df, and a non-centrality parameter of

$$\lambda_{HW} = 2n \sum_i \sum_j \frac{(\phi\delta_{ij} - \delta_i\delta_j)^2}{\phi^2(\phi q_i + \delta_i)(\phi q_j + \delta_j)}.$$

When there are just two marker alleles, $m = 2$, the power of the two chi-square tests can be compared directly by comparing λ_{CC} with λ_{HW}. For this

case, Hardy-Weinberg disequilibrium decays at a rate proportional to the square of linkage disequilibrium (Appendix 2). This indicates that the measure of Hardy-Weinberg disequilibrium should be a more sensitive indicator of position, decaying more quickly than the measure of linkage disequilibrium as the distance between the marker and susceptibility locus increases.

2.3.4 Simulations

To illustrate our theoretical results, we performed simulations of evolving populations segregating for a diallelic disease susceptibility locus and several diallelic markers. We performed these simulations under four different disease models, representing special cases of the general model. Analytical results for these special cases can be found in Appendix 2. For the four models we performed chi-square tests for linkage disequilibrium and for Hardy-Weinberg disequilibrium and compared the estimated power of the results.

For all four models, we considered a marker allele M at frequency $q_M = 0.20$ which had a positive association with the disease allele. Our first simulated model was the heterogeneous recessive model of Feder et al. (1996) with $p_A = 0.10$ and $\psi = 0.05$. For these parameters, the maximum linkage disequilibrium expected is 0.08. The second model was also of the type of Feder et al. (1996), but with different parameter values. For this model we chose the parameters $p_A = 0.05$ and $\psi = 0.05$. Since p_A is smaller in the second model, less linkage disequilibrium is possible; reaching a maximum expected value of 0.04, half of what was expected in the first model.

The third model was an additive model for penetrance. We set the effect of the disease-causing allele (A) to be 0.50, and the effect of the non-disease allele ($\bar{A}$) to be 0. This yields $\phi_{AA} = 1.0$, $\phi_{A\bar{A}} = 0.5$ and $\phi_{\bar{A}\bar{A}} = 0.0$. The frequency of the disease allele in the population, p_A, was 0.10. For the additive model, Hardy-Weinberg disequilibrium is expected to be negative (Appendix 2), and will increase in absolute value with increasing linkage disequilibrium.

A multiplicative model for penetrance was assumed for the fourth set of simulations. We set the effect of the disease-causing allele (A) to be 0.9 and the effect of the non-disease allele ($\bar{A}$) to be 0.05. This leads to $\phi_{AA} = 0.8100$, $\phi_{A\bar{A}} = 0.0450$ and $\phi_{\bar{A}\bar{A}} = 0.0025$. The frequency of the disease allele in the population, p_A, was 0.10. We did not expect to see any Hardy-Weinberg disequilibrium among the affected individuals (Appendix 2).

A summary of the parameter values used in these four models can be found in Table 2.1.

For our simulated populations we considered marker loci positioned at distances of 0 to 2 centiMorgans from the susceptibility locus, considering one marker every 0.25 centiMorgans. The populations started at generation G_0 with complete association between the disease allele and one allele at each marker locus, then evolved for 50 generations of random mating. For each model we retained the first 100 populations which after 50 generations had not experienced substantial genetic drift at the disease locus. For a population to be accepted, the frequency of the disease allele at the end of the evolution could not deviate from the original frequency by more than 0.05. We made no adjustments for genetic drift at the marker locus.

2.4 Results

2.4.1 Power

To determine the power to detect Hardy-Weinberg and linkage disequilibria, we performed the chi-square tests X^2_{CC} and X^2_{HW} (Equations 2.6 and 2.7) on samples taken from each population. For the case-control test, we sampled 50 affected and 50 unaffected individuals from each population. For the test for Hardy-Weinberg disequilibrium, we sampled 100 affected individuals. We repeated both tests 5000 times for each population, recording the percentage of times we rejected the hypothesis of no disequilibrium. This rejection percentage gave us an estimate of the power of the respective tests. The comparisons of these results can be seen in Figure 2.1. The symbols in this figure are box plots of the results; the bottom and top edges of the box are located at the sample 25th and 75th percentiles, the point joined by the connecting line is the median, and the whiskers extend the range of the results. This figure shows that for the Feder et al. type models (panels A and B), the power to detect Hardy-Weinberg disequilibrium is greater in general than the power to detect linkage disequilibrium. This is particularly noteworthy in the case of the second Feder et al. type model (panel B), where the power to detect linkage disequilibrium using X^2_{CC} is not very different from the $\alpha = 0.05$ nominal level. The additive model (panel C) shows high power for both tests. For the multiplicative model (panel D), we expected to find no Hardy-Weinberg disequilibrium among affected individuals. These experiments showed that while power to detect Hardy-Weinberg disequilibrium was very low for this model, it was very frequently above the $\alpha = 0.05$ nominal level. This appears to be due to

increased variance of Hardy-Weinberg disequilibrium values created by sampling affected individuals.

Figure 2.1 reveals the variability of the power of the two chi-square tests. For several of these experiments, the power of X^2_{HW} varied from the nominal 0.05 level to values close to 1.0. X^2_{CC} was less varied in its power.

In these experiments, we generated linkage disequilibrium in the presence of physical linkage. Thus, both tests showed reduced power at greater distances between the loci. As expected, the power to detect Hardy-Weinberg disequilibrium decayed more quickly than that of linkage disequilibrium in the three models where we expected to find Hardy-Weinberg disequilibrium.

2.4.2 Size

In order to determine the size of our tests, we simulated a second set of populations under the same four disease models, but segregating for a diallelic marker located at 50% recombination from the disease locus. We performed the same sampling and testing experiments as before. The results from this experiments are displayed in Figure 2.2. These results showed that the case-control test, X^2_{CC}, was conservative; in most cases examined it rejected the true null hypothesis at a rate less than the $\alpha = 0.05$ nominal level. The rejection rate of test for Hardy-Weinberg disequilibrium, X^2_{HW}, appeared to be centered around the nominal $\alpha = 0.05$ level for the first three models, but was higher for model 4, the multiplicative model. For model 4, the variance of the rejection rate appeared quite large. It was, however, very similar to the values seen for the populations generated with linked markers

as shown in Figure 2.1.

2.5 Discussion

We have examined departures from Hardy-Weinberg equilibrium created by sampling individuals based on presence of a disease phenotype. These departures from equilibrium are created because the selection criteria is based on susceptibility genotypes, rather than independently selected alleles. Alleles within genotypes which confer greater susceptibilities are represented in the sample at disproportionally high rates. Disequilibrium is expected to be the greatest at the susceptibility locus itself, since this is the factor which determines the selection criteria. Loci which are phenotypically neutral, but are somehow associated with the susceptibility locus, such as genetic markers in linkage disequilibrium with the susceptibility locus, also experience disproportionate genotype selection. As the degree of association between susceptibility and marker loci decreases, Hardy-Weinberg disequilibrium at the marker loci is also expected to decrease. We have examined measures which capture this relationship, potentially offering fine-mapping techniques which can be performed on samples of affected individuals when an appropriate control sample is not available. For a general disease model which considers an arbitrary number of alleles at the susceptibility locus, we have proposed the measures δ_i and δ_{ij}. These are summary measures, useful in quantifying linkage disequilibria between marker allele M_i and the susceptibility alleles. Since these measures allow for a simple expression of the marker allele frequencies within affected and unaffected individuals, they can be readily incorporated into many

established measures of association. This allows for simpler interpretation of these measures.

Under certain disease models, and when physical linkage and linkage disequilibrium exist between the markers and a susceptibility locus, conventional tests for Hardy-Weinberg disequilibrium at marker loci can be used to fine map disease susceptibility loci. For diallelic locus models (in which both the susceptibility locus and the marker loci have only two alleles per locus), Hardy-Weinberg disequilibrium is proportional to the square of linkage disequilibrium (Appendix 2). This indicates that measures of Hardy-Weinberg disequilibrium are expected to decay more rapidly than direct tests for linkage disequilibrium as linkage disequilibrium diminishes. The results of Feder et al (1996) illustrated this: their curve plotting the marker map versus Hardy-Weinberg equilibrium was sharper than their curve plotting marker map versus p_{excess}, a measure of linkage disequilibrium. For a general disease model, allowing for two or more alleles at the marker and susceptibility loci, the relationship between linkage disequilibrium and Hardy-Weinberg disequilibrium becomes less clear. However, departure from Hardy-Weinberg at a marker locus provides evidence both for linkage disequilibrium between marker and susceptiblity loci and heterogeneity of susceptibility, so tests for Hardy-Weinberg disequilibrium could still be useful.

There are some caveats which should be considered when examining a general disease model. For simple diallelic locus models, the interpretation of disequilibrium measures is straight-forward. However, when more than two alleles exist at the marker and/or susceptibility locus, complications arise. For these models, summary measures may be used to quantify association between loci; however, while

disequilibria between specific alleles may exist, these disequilibria may cancel out when combined into the summary measure. This poses a challange in mapping loci involved in complex traits, as it is doubtful that many of the traits of interest are diallelic. With the use of single nucleotide polymorphisms (SNPs) as genetic markers, some of the problems regarding multiple alleles disappear. In this case $\delta_2 = -\delta_1$. However, if there are more than two alleles at the susceptibility locus, the problem of the disequilibria between the susceptibility alleles and the marker allele cancelling within δ_1 is still a concern.

Tests for Hardy-Weinberg disequilibrium will be the most powerful when large amounts of disequilibrium within a sample of affected individuals are expected. The amount of Hardy-Weinberg disequilibrium expected depends on both the degree to which the susceptibility locus affects disease status and on the manner in which the alleles within a genotype interact. In sampling affected individuals, genotypes will be sampled proportionally to the rate of susceptibility they confer. By definition, Hardy-Weinberg disequilibrium is the difference between genotype proportions and the product of the proportions of the composite alleles. If the alleles within a genotype act in a multiplicative manner to cause increased levels of susceptibility, the genotypes are expected to be selected proportionally to the product of the allele frequecies. Thus, with disease models where the alleles act in a multiplicative manner, Hardy-Weinberg disequilibrium is not expected to be created in the sample. The more the effects of alleles deviate from multiplicative interactions, the greater the amount of Hardy-Weinberg disequilibrium expected. This is seen in the theoretical results of Appendix 2 and was illustrated in the results of the simulations we performed.

We note that if the penetrances ϕ_{rs} are regarded as genotypic values, much of the theory in this paper can be applied to the study of quantitative traits.

2.6 Acknowledgements

This work was supported in part by NIH Grant GM45344 to NC State University.

2.7 Appendix 1

The two-locus genotypes and their frequencies, for susceptibility-locus homozygotes, are

$$\begin{array}{ll} A_rA_rM_iM_i & (p_rq_i + D_{ri})^2 \\ A_rA_rM_iM_j & 2(p_rq_i + D_{ri})(p_rq_j + D_{rj}),\ r \neq s \end{array}$$

and for susceptibility-locus heterozygotes, $r \neq s$,

$$\begin{array}{ll} A_rA_sM_iM_i & 2(p_rq_i + D_{ri})(p_sq_i + D_{si}) \\ A_rA_sM_iM_j & 2(p_rq_i + D_{ri})(p_sq_j + D_{sj}) + 2(p_rq_j + D_{rj})(p_sq_i + D_{si}),\ r \neq s \end{array}$$

Among affected people, therefore, the marker genotype frequencies are

$$\begin{aligned} P_{ii|\text{Aff.}} &= \frac{1}{\phi}\sum_r\sum_s \phi_{rs}(p_rq_i + D_{ri})(p_sq_i + D_{si}) \\ &= q_i^2 + \frac{2q_i\delta_i}{\phi} + \frac{\delta_{ii}}{\phi} \end{aligned}$$

$$\begin{aligned} P_{ij|\text{Aff.}} &= \frac{1}{\phi}\sum_r\sum_s \phi_{rs}[(p_rq_i + D_{ri})(p_sq_j + D_{sj}) \\ &\quad + (p_rq_j + D_{rj})(p_sq_i + D_{si})] \\ &= 2q_iq_j + \frac{2(q_i\delta_j + q_j\delta_i)}{\phi} + \frac{2\delta_{ij}}{\phi},\ i \neq j \end{aligned}$$

where

$$\begin{aligned} \delta_i &= \sum_r\sum_s p_s\phi_{rs}D_{ri} = \sum_r \phi_rD_{ri} \\ \delta_{ij} &= \sum_r\sum_s \phi_{rs}D_{ri}D_{sj} \end{aligned}$$

Adding over genotypes provides the marker allele frequencies:

$$\begin{aligned} q_{i|\text{Aff.}} &= P_{ii|\text{Aff.}} + \frac{1}{2}\sum_{j\neq i} P_{ij|\text{Aff.}} \\ &= \sum_j \left[q_i q_j + \frac{(q_i\delta_j + q_j\delta_i)}{\phi} + \frac{\delta_{ij}}{\phi} \right] \\ &= q_i + \frac{\delta_i}{\phi}. \end{aligned}$$

2.8 Appendix 2 – Special Cases

2.8.1 Heterogeneous Recessive Model

Susceptibility locus alleles $A, \bar{A}$; marker alleles $M, \bar{M}$;

$\phi_{AA} = 1$, $\phi_{A\bar{A}} = \phi_{\bar{A}\bar{A}} = \psi$:

$$\begin{aligned} \phi &= p_A^2 + \psi(1 - p_A^2) \\ \delta_M &= (1 - \psi) p_A D_{AM} \\ \delta_{MM} &= (1 - \psi) D_{AM}^2 \\ \mathcal{D}_{MM|\text{Aff.}} &= \frac{\psi(1 - \psi) D_{AM}^2}{\phi^2} \geq 0 \end{aligned}$$

2.8.2 General Diallelic Model

Susceptibility locus alleles $A, \bar{A}$; marker alleles $M, \bar{M}$:

$$\begin{aligned} \delta_M &= [p_A(\phi_{AA} - \phi_{A\bar{A}}) + (1 - p_A)(\phi_{A\bar{A}} - \phi_{\bar{A}\bar{A}})] D_{AM} \\ \delta_{MM} &= (\phi_{AA} - 2\phi_{A\bar{A}} + \phi_{\bar{A}\bar{A}}) D_{AM}^2 \\ \mathcal{D}_{MM|\text{Aff.}} &= \frac{(\phi_{AA}\phi_{\bar{A}\bar{A}} - \phi_{A\bar{A}}^2) D_{AM}^2}{\phi^2} \end{aligned}$$

2.8.3 Additive Susceptibilities

If $\phi_{rs} = \alpha_r + \alpha_s$:

$$\phi = 2 \sum_r \alpha_r p_r$$

$$
\begin{aligned}
\delta_i &= \sum_r \alpha_r D_{ri} \\
\delta_{ij} &= 0 \\
\mathcal{D}_{ij|\text{Aff.}} &= -\left(\frac{\sum_r \alpha_r D_{ri}}{2\sum_r \alpha_r p_r}\right)\left(\frac{\sum_s \alpha_s D_{sj}}{2\sum_s \alpha_s p_s}\right) \\
&\leq 0, \text{ if } r = s
\end{aligned}
$$

2.8.4 Multiplicative Susceptibilities

If $\phi_{rs} = \alpha_r \alpha_s$:

$$
\begin{aligned}
\phi &= (\sum_r \alpha_r p_r)^2 \\
\delta_i &= (\sum_r \alpha_r p_r)(\sum_r \alpha_r D_{ri}) \\
\delta_{ij} &= (\sum_r \alpha_r D_{ri})(\sum_s \alpha_s D_{sj}) \\
\mathcal{D}_{ij|\text{Aff.}} &= 0
\end{aligned}
$$

2.9 References

Bengtsson BO, Thomson G (1981) Measuring the strength of associations between HLA antigens and diseases. Tissue Antigens 18:356–363.

Bennett ST, Lucassen AM, Gough SCL, Powell EE, Undlien DE, Pritchard LE, Merriman ME, et al (1995) Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene mini satellite locus. Nat Genet 9:284– 292

Boehnke, M (1994) Limits of resolution of genetic linkage studies: implications for the positional closing of human disease genes. Am J Hum Genet 55:379–390

Corder EH, Saunders AM, Strittmatter, Schmechel DE, Gaskell PC, Small GW, Roses AD, et al (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261:921–923

Feder JN, Gnirke A, Thomas W, Tsuchihasi Z (1996) A novel MHC class I- like gene is mutated in patients with hereditary haemochromatosis. Nat Genet 13:399–408

Glaser B, Chiu KC, Liu L, Anker A, Nestorowicz A, Cox NJ, Landau N, et al (1995) Recombinant mapping of the familial hyperinsulinsim gene to an 0.8 cM region on chromosome 11p15.1 and demonstration of a founder effect in Ashkenazi Jews. Hum Mol Genet 4:879–886

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander ES (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211

Hästbacka J, de la Chapelle A, Mahtani MM, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, et al. (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: position cloning by fine-structure linkage disequilibrium mapping. Cell 78:1073–1087

Kaplan NL, Martin ER, Weir BS (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 60:691–702

Lehesjoki A-E, Koskiniemi M, Norio R, Tirrito S, Sistonen P, Lander E, de la Chapelle A (1993) Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: Linkage disequilibrium allows high resolution mapping. Hum Mol Genet 2:1229-1234

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61:439–448

Meng RC, Chapman DG (1966) The power of chi-square tests for contingency tables. J Am Stat Assoc 61:965–975

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

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the S-TDT. Am J Hum Genet 62:450–458

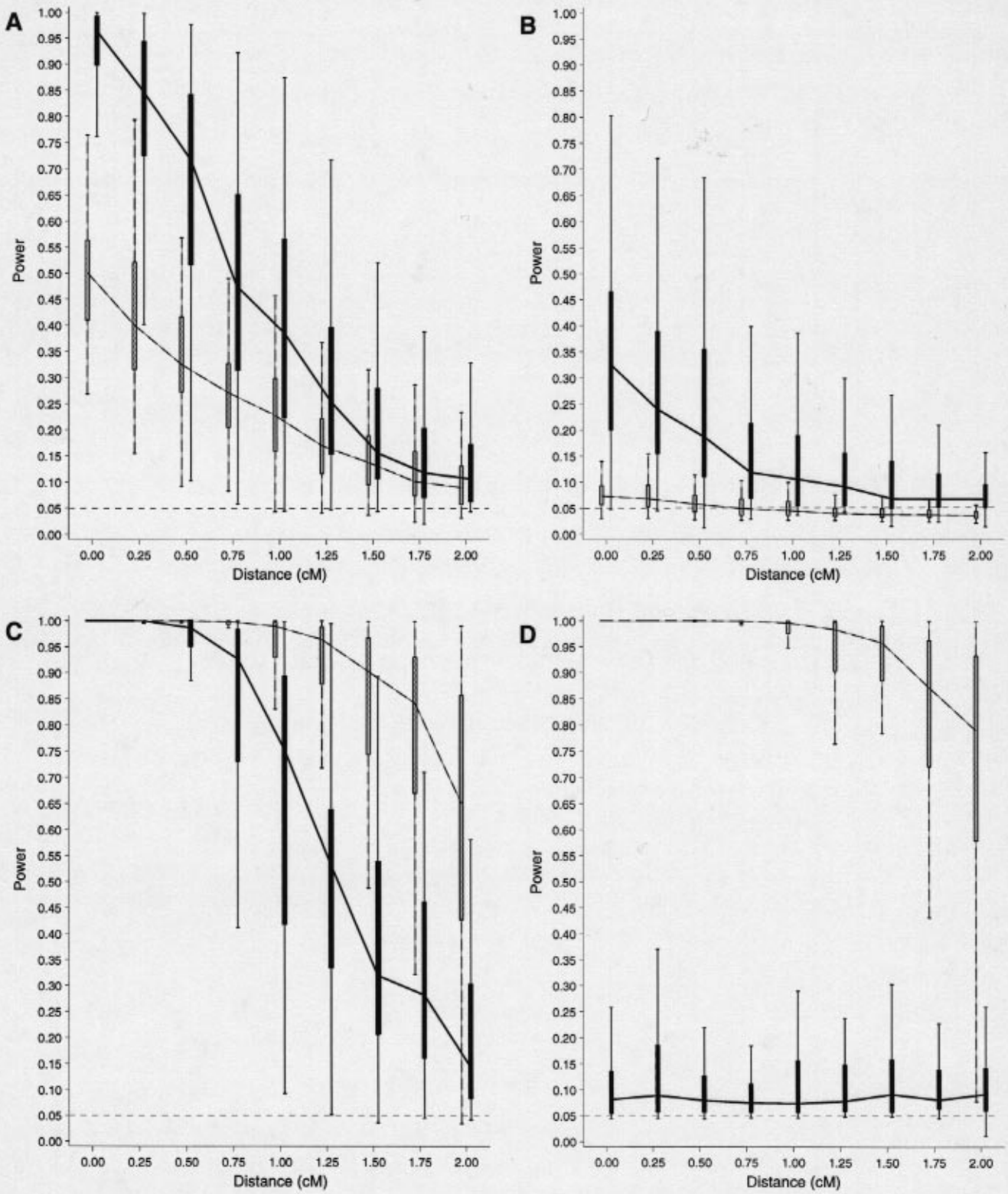
Trégouët D-A, Ducimetire P, Tiret L (1997) Testing association between candidate-gene markers and phenotype in related individuals, by use of estimating equations. Am J Hum Genet 61:189–199

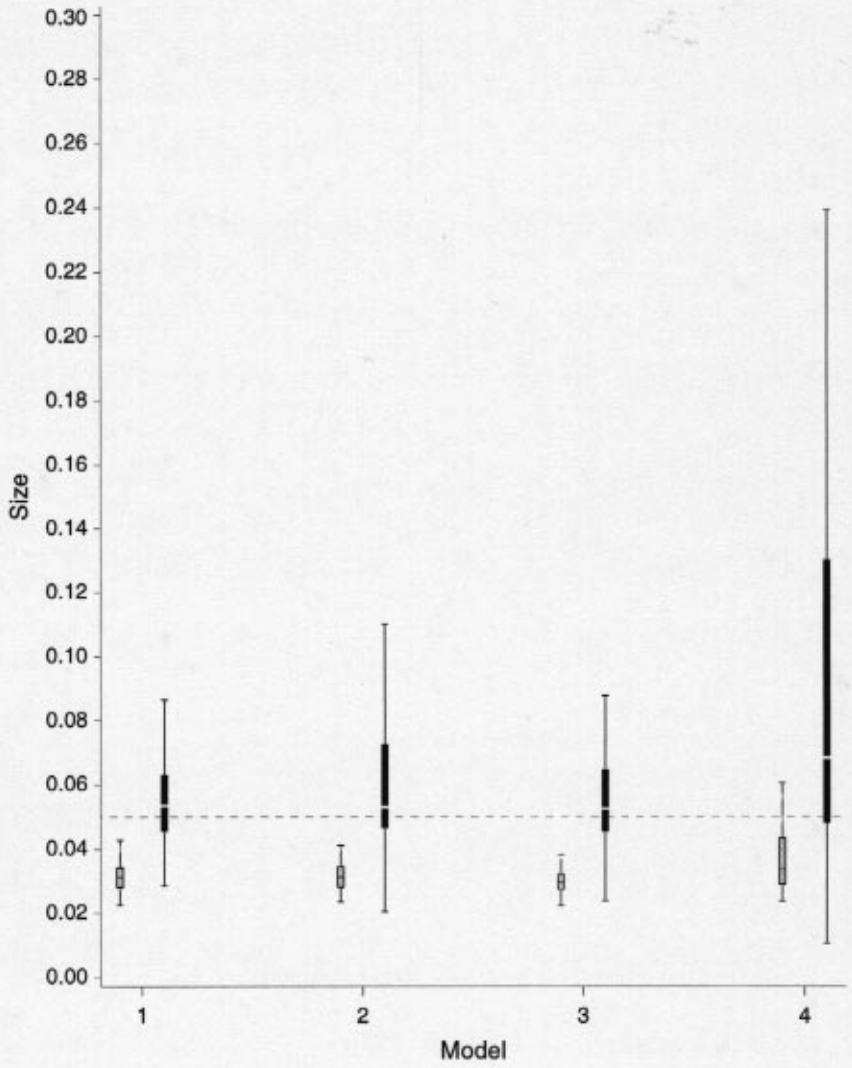
Weir BS (1996) Genetic Data Analysis II. Sinauer, Sunderland, MA.

2.10 Figures

Figure 2.1: Power results for the chi-square tests for linkage disequilibrium (grey-filled boxes) and Hardy-Weinberg disequilibrium (black boxes). Panels A and B are the first and second heterogeneous recessive models of Feder et al, panel C is the additive model, and panel D is the multiplicative model. The symbols represent the range of the proportions of times the hypothesis of no disequilibrium was rejected for the 100 populations. The bottom and top edges of the box represent the sample 25th and 75th percentiles, the point joined by the connecting line is the median, and the whiskers extend the range of the results.

Figure 2.2: Size of the the chi-square tests for linkage disequilibrium (grey-filled boxes) and Hardy-Weinberg disequilibrium (black boxes). Models 1 and 2 were the first and second heterogeneous recessive models of Feder et al., model 3 was the additive model, and the multiplicative model was model 4. These symbols represent the proportion of times a true null hypothesis was rejected.


A
B
C
D
Power
Distance (cM)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


0.30
0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Size
1
2
3
4
Model

2.11 Tables

Table 2.1: Parameters of the Simulated Disease Models

Model	Type	ϕ_{AA}	$\phi_{A\bar{A}}$	$\phi_{\bar{A}\bar{A}}$	p_A	p_B	D_{max} [a]
1	Heterogeneous Recessive	1.00	0.05	0.05	0.10	0.20	0.08
2	Heterogeneous Recessive	1.00	0.10	0.10	0.05	0.20	0.04
3	Additive	1.00	0.50	0	0.10	0.20	0.08
4	Multiplicative	0.81	0.045	0.0025	0.10	0.20	0.08

[a] Maximum expected disequilibrium for these parameters

Chapter 3

A CLASSICAL SETTING FOR ASSOCIATIONS BETWEEN MARKERS AND LOCI AFFECTING QUANTITATIVE TRAITS

Nielsen DM and Weir BS (1999)

Accepted: Genetical Research

3.1 Abstract

We examine the relationships between a genetic marker and a locus affecting a quantitative trait by decomposing the genetic effects of the marker locus into additive and dominance effects under a classical genetic model. We discuss the structure of the associations between the marker and the trait locus, paying attention to non-random union of gametes, multiple alleles at the marker and trait loci, and non-additivity of allelic effects at the trait locus. We consider that this greater-than-usual level of generality leads to additional insights, in a way reminiscent of Cockerham's decomposition of genetic variance into five terms: three terms in addition to the usual additive and dominance terms. Using our framework, we examine several common tests of association between a marker and a trait.

3.2 Introduction

Since the discovery of numerous polymorphic markers spread across the genomes of many species, linkage analysis has been highly successful in localizing regions of chromosomes containing genes affecting many traits of interest (Paterson et al., 1988; Georges et al., 1995; Comuzzie et al., 1997). Often, however, these chromosomal regions are very large, spanning many millions of bases and containing many genes. In order to narrow these regions to areas which are more amenable to molecular characterization, there has been an increasing amount of interest in fine-mapping techniques. These methods capitalize on evolutionary history and population genetics to capture the relationships between markers in very close proximity to genes affecting the trait of interest. The parameters of interest in these studies include the linkage disequilibria between the gene and a marker, which it is hoped can give an indication of degree of proximity between the two loci. The general theory behind this method is that markers which are very tightly linked to a gene should show high association with the trait, reflecting linkage disequilibria between the alleles at the marker and those at the gene affecting the trait. Markers which are less closely linked to the gene will have lost much of their association with the trait due to recombination over time.

In the analysis of dichotomous traits, two basic types of study designs are often used in fine-mapping experiments to detect association between a marker and a gene. The first is a case-control design, in which individuals are collected for both categories of the trait (i.e. "affected" and "unaffected") and then genotyped for the genetic markers. If marker allele frequencies among the two groups differ

significantly, then it is concluded that the marker is associated with a gene affecting the trait; there is non-zero linkage disequilibrium between the gene and the marker. This type of analysis does not control for population dynamics such as admixture or selection, so association does not necessarily imply linkage.

The second study design which has been proposed for dichotomous traits is the transmission/disequilibrium design (Spielman et al., 1993; Kaplan et al., 1997). These methods utilize random population samples of small nuclear families, and test for non-equal segregation of marker alleles from heterozygous parents to affected offspring. If a marker is linked to a gene affecting the trait, then marker alleles which are in association with the alleles of the gene should be preferentially transmitted to affected offspring. A significant result for this test is evidence of both linkage and association, suggesting a more precise indication of location of the gene.

Several extensions to the transmission/disequilibrium design have been proposed for fine-mapping of quantitative trait loci (QTL). Allison (1997) suggested five designs, each considering a different scheme for sampling based on phenotype data. Each of these designs utilizes data collected for trios of two parents with one offspring, where at least one of the parents is heterozygous for the marker locus being examined. Markers are assumed to be biallelic. Under the null hypothesis, the mean values of offspring within the three marker genotype classes are equal. If, however, the marker is associated with a gene affecting the trait, and the recombination rate between the two loci is less than 0.5, then the null hypothesis will not be true and the three means will not be equal. If a sampling scheme based on offspring phenotype is chosen, then under the alternative hypothesis, unequal transmission

of marker alleles to offsprings can also be used as an indication of association in the presence of linkage. Rabinowitz (1997) proposed a similar design to test for linkage in the presence of association, but allowed for larger nuclear families to be collected. In the case of families with a single offspring, his design reduces to the randomly sampled design of Allison (1997). Martin (1997) extended the work of Allison (1997) by allowing for an arbitrary number of marker alleles, all considered simultaneously. This test is based on the difference between the mean phenotypic value among offspring of parents who transmit a particular marker allele and the mean phenotypic value among offspring of parents who do not transmit that allele.

All of these tests examine alleles individually, rather than as genotypes, though it is genotypes, rather than individual alleles, which generally affect phenotype. Here, we propose a classical genetics framework whereby the consequences of examining genetic data based on observations at the alleles alone can be determined.

3.3 Methods

We consider a gene with an arbitrary number of alleles which contributes to the genetic component of the quantitative trait of interest. Alleles at this gene are designated A_r, with population frequencies p_r. The genetic effect of genotype A_rA_s on the trait, G_{rs}, can be described by the classical linear model

$$G_{rs} = \mu + \alpha_r + \alpha_s + d_{rs} \tag{3.1}$$

where μ is the genotypic mean, α_r and α_s are the additive effects of the alleles, and d_{rs} is the deviation from additivity. For a random mating reference population, the

least squares solutions for the parameters are described in Weir and Cockerham (1977):

$$\mu = \sum_{r,s} p_r p_s G_{rs} = G_{..} \tag{3.2}$$

$$\alpha_r = \sum_{s} p_s G_{rs} - \mu = G_{r.} - G_{..} \tag{3.3}$$

$$\alpha_s = \sum_{r} p_r G_{rs} - \mu = G_{.s} - G_{..} \tag{3.4}$$

$$d_{rs} = G_{rs} - \alpha_r - \alpha_s - \mu = G_{rs} - G_{r.} - G_{.s} + G_{..} \tag{3.5}$$

These solutions embody the constraints $\sum_r p_r \alpha_r = \sum_r p_r d_{rs} = 0$.

The additive and dominance effects are generally regarded as being random, and the genetic variance contributed by the trait locus in a random mating population is

$$\sigma^2_G = \sum_{r,s} p_r p_s (G_{rs} - \mu)^2 = \sigma^2_A + \sigma^2_D$$

where the additive and dominance variance components are

$$\sigma^2_A = 2 \sum_{r} p_r (\alpha_r)^2 \tag{3.6}$$

$$\sigma^2_D = \sum_{r,s} p_r p_s (d_{rs})^2 \tag{3.7}$$

Since the genotypes at the trait locus are generally not observable, we also consider a marker locus with an arbitrary number of alleles, designated M_i and having population frequencies q_i. We are interested in determining the relationship between the genotypes at the marker locus and the gene affecting the trait. For this, we consider the same type of linear model as in Equation 3.1, but now defined in terms of the marker genotypic classes, $M_i M_j$, and their genetic effects $G^{(m)}_{ij}$:

$$G^{(m)}_{ij} = \mu^{(m)} + \alpha^{(m)}_i + \alpha^{(m)}_j + d^{(m)}_{ij} \tag{3.8}$$

Here $\alpha_i^{(m)}$ and $\alpha_j^{(m)}$ are the additive effects of marker alleles M_i and M_j on the trait, and $d_{ij}^{(m)}$ is the dominance deviation. Using the solutions of Equations 3.2-3.5, we see that

$$\mu^{(m)} = \sum_{i,j} q_i q_j G_{ij}^{(m)} = \mu \tag{3.9}$$

$$\alpha_i^{(m)} = \sum_j q_j G_{ij}^{(m)} - \mu^{(m)} \tag{3.10}$$

$$d_{ij}^{(m)} = G_{ij}^{(m)} - \mu^{(m)} - \alpha_i^{(m)} - \alpha_j^{(m)} \tag{3.11}$$

Each marker genotypic class M_iM_j is composed of a mixture of elements from all of the trait classes, A_rA_s, and the proportion of class A_rA_s contained within class M_iM_j is $\Pr(A_rA_s|M_iM_j)$:

$$\begin{aligned} G_{ij}^{(m)} &= \sum_{r,s} \Pr(A_rA_s|M_iM_j) G_{rs} \\ &= \sum_{r,s} \frac{P_{sj}^{ri}}{\Pr(M_iM_j)} G_{rs} \end{aligned}$$

Here P_{sj}^{ri} is the frequency of A_rM_i/A_sM_j genotypes. With random union of gametes, and writing D_{ri} for the linkage disequilibrium between marker allele M_i and allele A_r at the gene, this becomes

$$\begin{aligned} G_{ij}^{(m)} &= \sum_{r,s} \frac{(p_r q_i + D_{ri})(p_s q_j + D_{sj})}{q_i q_j} G_{rs} \\ &= \mu + \frac{1}{q_j}\delta_j + \frac{1}{q_i}\delta_i + \frac{1}{q_i q_j}\delta_{ij} \end{aligned}$$

where $\delta_i = \sum_r \sum_s p_s G_{rs} D_{ri} = \sum_r \alpha_r D_{ri}$, and $\delta_{ij} = \sum_{r,s} D_{ri} D_{si} G_{rs} = \sum_{r,s} D_{ri} D_{si} d_{rs}$. Using this result and applying Equations 3.10 and 3.11, and noting that $\sum_i D_{ri} = 0$, it follows that

$$\alpha_i^{(m)} = \frac{1}{q_i}\delta_i = \frac{1}{q_i}\sum_r \alpha_r D_{ri} \tag{3.12}$$

$$d_{ij}^{(m)} = \frac{1}{q_i q_j}\delta_{ij} = \frac{1}{q_i q_j}\sum_{r,s} d_{rs} D_{ri} D_{si} \tag{3.13}$$

This shows that the additive effect of a marker allele M_i is the weighted sum of the additive effects of the alleles at the trait locus, where the weights are the linkage disequilibria between that marker allele and the alleles at the trait locus. Similarly, the dominance deviation of a marker genotype is the weighted sum of the dominance deviations of the genotypes at the trait locus.

It is important to distinguish the association measure δ_i between marker allele M_i and a trait from the linkage disequilibrium measure D_{ri} between marker allele M_i and trait allele A_r. In the special case of the trait locus having only two alleles, A_1, A_2, we have $D_{1i}+D_{2i} = 0$ and the association parameter is proportional to linkage disequilibrium. More generally, non-zero association implies non-zero linkage disequilibrium but zero association does not imply zero linkage disequilibrium.

3.4 Tests for Association

These relationships between marker and trait locus effects suggest several approaches to testing for association between the marker and a gene affecting the trait.

3.4.1 Dichotomous Trait Case-Control Tests

For a dichotomous trait, the genetic values G_{rs} may be regarded as susceptibilities, or the probabilities that $A_r A_s$ individuals are affected (Nielsen et al., 1998). These values are quantitative in nature; having a continuous distribution and relying on

both genetic and environmental influences. The mean value μ is the probability of a random individual being affected, i.e. the population prevalence ϕ of the disease. The frequency of marker allele M_i among affected individuals is found by taking the sum over all (unobserved) trait genotypes. Under the assumption of random union of gametes

$$\begin{aligned} q_{i|\text{Aff.}} &= \sum_j \sum_{r,s} \Pr(A_r M_i / A_s M_j | \text{Aff.}) \\ &= \sum_j \sum_{r,s} (p_r q_i + D_{ri})(p_s q_j + D_{si}) G_{rs} / \phi \\ &= q_i + \delta_i / \phi \end{aligned}$$

Among unaffected individuals the marker allele frequency is

$$q_{i|\text{Unaff.}} = q_i - \delta_i / (1 - \phi)$$

suggesting that association, δ_i, can be detected by comparing these two frequencies:

$$q_{i|\text{Aff.}} - q_{i|\text{Unaff.}} = \frac{\delta_i}{\phi(1-\phi)}$$

A goodness-of-fit test statistic (Kaplan et al., 1998) uses sample marker allele frequencies $\tilde{q}_i$ among affecteds and unaffecteds. For samples of n affected and n unaffected individuals:

$$X_c^2 = n \sum_i \frac{(\tilde{q}_{i|\text{Aff.}} - \tilde{q}_{i|\text{Unaff.}})^2}{(\tilde{q}_{i|\text{Aff.}} + \tilde{q}_{i|\text{Unaff.}})}$$

The allelic case-control test is therefore a test for additive effects at the trait locus, mediated by linkage disequilibrium between the trait and marker loci. Power to detect association, therefore, depends on both non-zero additive effects and non-zero linkage disequilibria, and that these terms do not cancel in the summary measure, δ_i.

An alternative would be to compare marker genotype frequencies among affecteds and unaffecteds. With random union of gametes

$$\begin{aligned}
\Pr(M_iM_j|\text{Aff.}) &= \sum_{r,s} \Pr(A_rM_i/A_sM_j|\text{Aff.}) \\
&= \sum_{r,s} (p_rq_i + D_{ri})(p_sq_j + D_{si})G_{rs}/\phi \\
&= q_iq_j + (\delta_i + \delta_j)/\phi + \delta_{ij}/\phi \\
\Pr(M_iM_j|\text{Unaff.}) &= q_iq_j - (\delta_i + \delta_j)/(1-\phi) - \delta_{ij}/(1-\phi)
\end{aligned}$$

The difference is

$$\Pr(M_iM_j|\text{Aff.}) - \Pr(M_iM_j|\text{Unaff.}) = \frac{\delta_i + \delta_j + \delta_{ij}}{\phi(1-\phi)}$$

and this contrast leads to a joint test of both additive and dominance components for the trait locus, mediated by linkage disequilibria.

3.4.2 Quantitative Trait Case-Control Tests

For the continuous trait values of primary interest here, the viewpoint is reversed: conditioning is on marker type instead of on disease status (trait value). The simplest procedure is to compare trait means among individuals distinguished by their marker types. Suppose there are n_{ij} individuals with marker genotype M_iM_j. The trait value for the kth of these individuals is

$$Y_{ijk} = G_{ij}^{(m)} + \epsilon_{ijk}, \quad k = 1, 2, \ldots, n_{ij} \tag{3.14}$$

where ϵ_{ijk} is an error term. We assume the errors are independent of both marker and trait genotypes, and are distributed with a mean of zero and variance σ_e^2.

The mean squares between and within marker genotype classes have expected values of

$$\begin{aligned}\mathcal{E}(\text{MSB}) &= \sigma_e^2 + \frac{1}{m-1}\sum_{i,j} n_{ij}(G_{ij}^{(m)} - \overline{G_{\cdot\cdot}^{(m)}})^2 \\ \mathcal{E}(\text{MSW}) &= \sigma_e^2\end{aligned}$$

where m is the number of distinct marker genotypes in the data, and

$$\overline{G_{\cdot\cdot}^{(m)}} = \sum_{i,j} n_{ij} G_{ij}^{(m)} / \sum_{i,j} n_{ij}$$

For normally-distributed trait values, the F-test will provide a test for association between trait and marker loci, but will not distinguish between additive and dominance effects at the trait locus.

Of course, it would be possible to find least-squares estimates of the effects $\alpha_i^{(m)}$ and $d_{ij}^{(m)}$ and form the summary statistics

$$\begin{aligned}\hat{\sigma}^2_{A^{(m)}} &= 2\sum_i \hat{q}_i(\widehat{\alpha_i^{(m)}})^2 \\ \hat{\sigma}^2_{D^{(m)}} &= \sum_{i,j} \hat{q}_i\hat{q}_j(\widehat{d_{ij}^{(m)}})^2\end{aligned}$$

and construct tests that the corresponding parameters were zero. Non-zero values of $\sigma^2_{A^{(m)}}$ and $\sigma^2_{D^{(m)}}$ imply non-zero additive and dominance variances for the trait, although the converse does not apply: zero values of $\sigma^2_{A^{(m)}}$ and $\sigma^2_{D^{(m)}}$ do not imply zero additive and dominance variances for the trait.

3.4.3 Dichotomous Trait Transmission/Disequilibrium Tests

A finding of association between a marker and the trait does not imply genetic linkage between marker and trait loci. In order to test for linkage, Spielman et al.

(1993) introduced the TDT based on the marker allele transmitted from parent to child. For the basic design of this test, trios of parents and an affected offspring are collected, and the number of times an M_i allele and not an M_j allele are transmitted from a parent to an affected offspring is calculated. If no assumptions regarding the mating structure of the population are made, transmissions from parents to offspring must be considered jointly, as parents are not necessarily independent. The probability of a parent transmitting an M_i and not an M_j allele to an affected offspring, averaging over all trait alleles A_t transmitted by the other parent is T_{ij}, where

$$\begin{aligned} T_{ij} &= \frac{1}{\phi}\sum_{r,t}\left[(1-c)\Pr(A_rM_i/M_j,\ A_t) + c\Pr(A_rM_j/M_i,\ A_t)\right]G_{rt} \\ &= \frac{1}{\phi}\sum_{r,t}\left[(1-c)P^{ir}_{j|t} + cP^{jr}_{i|t}\right]G_{rt} \end{aligned}$$

where c is the recombination rate between the gene and the marker and $\Pr(A_rM_i/M_j,\ A_t) = P^{ir}_{j|t}$ is the joint probability that one parent has haplotype A_rM_i and has M_j as the other marker allele and the other parent carries the A_t allele (Weir et al., 1990). In this design, conditioning is based on the affection status of the offspring (only trios with affected offspring are chosen) instead of on the marker genotype of the parents, which are not known in advance. The probability of a parent transmitting an M_i and not an M_j allele is equivalent to the probability of an M_iM_j parent transmitting an M_i allele.

The difference between the transmission probabilities for M_i and M_j from M_iM_j parents to affected offspring is

$$T_{ij} - T_{ji} = \frac{(1-2c)}{\phi}\sum_{r,t}(P^{ir}_{j|t} - P^{jr}_{i|t})G_{rt}$$

which suggests a procedure for testing for linkage ($c \neq 0.5$) in the presence of association, i.e. $P^{ir}_{j|t} \neq P^{jr}_{i|t}$. Ewens and Spielman (1995) noted that the TDT does not detect association in a structured population for which there is no linkage disequilibrium within each subpopulation and there has not been more than one generation of mating between subpopulations. This is an example where $P^{ir}_{j|t} = P^{jr}_{i|t}$ among parents in the whole population.

When there is random union of gametes and no higher-order disequilibria

$$T_{ij} = q_i q_j + [(1-c)q_j\delta_i + cq_i\delta_j]/\phi$$

and

$$T_{ij} - T_{ji} \quad = \quad \frac{(1-2c)}{\phi}(q_j\delta_i - q_i\delta_j),$$

suggesting a test for association. In general, however, the difference depends also on the associations among all subsets of the four alleles M_i, M_j, A_r, A_t (Cockerham and Weir, 1973; Weir, 1996). Disequilibria among non-gametic allele pairs such as M_j, A_r and M_j, A_t are likely to be small when there is Hardy-Weinberg equilibrium.

It is customary to concentrate only on the transmitted marker allele. The transmission probabilities T_{ij} are summed over the non-transmitted allele and the contrast between transmitted and non transmitted probabilities for a particular allele is

$$T_{i.} - T_{.i} \quad = \quad \frac{(1-2c)}{\phi}(\delta^*_i - \delta^{*c}_i)$$

where $\delta^*_i = \sum_{r,t}(D_{ri|t} + p_t D_{ri})G_{rt}$ and $\delta^{*c}_i = \sum_{r,t}(D_{r/i|t} + p_t D_{r/i})G_{rt}$ are necessary when there is not random union of gametes.

Although rejection of the hypothesis $H_0 : T_{i.} = T_{.i}$ implies that $(\delta_i^* - \delta_i^{*c}) \neq 0$, and therefore that there is linkage disequilibrium between trait and marker loci and/or Hardy-Weinberg disequilibrium in the population from which the M_iM_j parent is drawn, the converse does not hold. There may be linkage and/or Hardy-Weinberg disequilibrium but little association. The various disequilibrium terms D add to zero over any subscript, and weighting them by terms such as α_r or d_{rt} can also give sums close to zero, especially for loci with little effect on the trait. With random union of gametes and no non-gametic disequilibria, this TDT is addressing the additive components of the trait-locus effects. Otherwise dominance at the trait locus does contribute to the term $(T_{i.} - T_{.i})$.

If there are n_{ij} parent-affected offspring pairs where M_i is transmitted and M_j is not transmitted, then a test for equality of marginal totals for the $\{n_{ij}\}$ contingency table has test statistic

$$X^2_{\mathrm{m}} = \sum_i \frac{(n_{i.} - n_{.i})^2}{n_{i.} + n_{.i}}$$

where $n_{i.} = \sum_j n_{ij}, n_{.i} = \sum_j n_{ji}$. It is more usual not to include transmissions from M_iM_i homozygous parents in this calculation (Spielman and Ewens, 1996). The test statistic is modified to

$$X^2_{\mathrm{mhet}} = \frac{m-1}{m} \sum_{i=1}^{m} \frac{(n_{i.} - n_{.i})^2}{n_{i.} + n_{.i} - 2n_{ii}}$$

3.4.4 Quantitative Trait Transmission/Disequilibrium Tests

For a quantitative trait, in place of the marker allele transmission probabilities T_{ij} to affected offspring, we consider the trait value $H_{ij}^{(m)}$ for an offspring that receives

M_i and not M_j from parent M_iM_j. From the same argument as above, when there is random union of gametes,

$$\begin{aligned} H_{ij}^{(m)} &= \frac{1}{q_i q_j} \sum_{r,t} \left[(1-c) \Pr(A_r M_i / M_j,\ A_t) + c \Pr(A_r M_j / M_i,\ A_t) \right] G_{rt} \\ &= \mu + (1-c)\frac{\delta_i}{q_i} + c\frac{\delta_j}{q_j} \end{aligned}$$

The contrast between expected trait values of individuals receiving M_i versus M_j from an M_iM_j parent is

$$H_{ij}^{(m)} - H_{ji}^{(m)} = (1-2c)\left(\frac{\delta_i}{q_i} - \frac{\delta_j}{q_j}\right)$$

suggesting a test statistic for the null hypothesis of no linkage or no association.

Allison (1997) and Martin (1997) worked with marginal expected trait values $H_{i.}^{(m)}$ for offspring that received marker allele M_i from a parent that carried that allele. The other parental marker allele does not need to be specified. Summing over j:

$$\begin{aligned} H_{i.}^{(m)} &= \sum_j q_j H_{ij}^{(m)} \\ &= \mu + (1-c)\frac{\delta_i}{q_i} \end{aligned}$$

so that the expected difference in trait values for individuals that either do or do not receive marker allele M_i from a parent is

$$H_{i.}^{(m)} - H_{.i}^{(m)} = (1-2c)\frac{\delta_i}{q_i} \tag{3.15}$$

The trait difference between marginals is expected to be zero if the recombination rate is 0.5 or if all the δ_i are zero. Since δ_i depends only on the additive effects of the trait, dominance at the trait locus does not affect these measures.

A statistical test for differences of expected trait values, $H_{i.}^{(m)} - H_{.i}^{(m)}$, is based on the set of observed phenotypes of offspring, $\mathbf{Y}$. A possible statistic to measure differences in marginal values is $\overline{Y}_{i.} - \overline{Y}_{.i}$, where $\overline{Y}_{i.}$ is the mean phenotypic value of all offspring who have received an M_i allele. When there is random mating in the population, the expected value of this contrast is the expression in Equation 3.15.

To calculate the variance of the contrast, both parents' transmissions to an offspring must be considered jointly. This is necessary, as each individual's phenotype is entered into the table twice (once for each parent-offspring transmission), which creates dependencies within the table. We use $Y_{ik,jl(a)}$ to denote the phenotype of the a^{th} offspring with genotype M_iM_k and parents M_iM_j and M_kM_l. We assume that these phenotypes conditional on genotypes are independent, with expected values $H_{ik,jl}^{(m)}$ and variances σ_ϵ^2, the environmental variance. The sum of the phenotypes of all offspring with the same transmitted/non-transmitted pattern is

$$Y_{ik,jl(.)} = \sum_{a=1}^{n_{ik,jl}} Y_{ik,jl(a)}$$

where $n_{ik,jl}$ is the number of offspring with genotype M_iM_k and parents M_iM_j and M_kM_l. The terms $n_{ik,jl}$ are random variables, following a multinomial distribution with probabilities $P_{ik,jl}$.

We can calculate $Y_{ij(.)}$, the sum of the phenotypes of all individuals receiving an M_i allele from an M_iM_j parent (the typical measure defined for the quantitative TDT) as:

$$Y_{ij(.)} = 2Y_{ii,jj(.)} + \sum_{k \neq i} Y_{ik,jj(.)} + \sum_{l \neq j} Y_{ii,jl(.)} + \sum_{k \neq i} \sum_{l \neq j} Y_{ik,jl(.)}.$$

For the marginal sums, $Y_{i.}$ and $Y_{.i}$, we need additional terms:

$$Y_{i.} = 2\sum_j Y_{ii,jj(.)} + \sum_j \sum_{k\neq i} Y_{ik,jj(.)} + \sum_j \sum_{l>j} Y_{ii,jl(.)} + \sum_j \sum_{k\neq i} \sum_{l>j} Y_{ik,jl(.)}$$

$$Y_{.i} = 2\sum_j Y_{jj,ii(.)} + \sum_j \sum_{k>j} Y_{jk,ii(.)} + \sum_j \sum_{l\neq i} Y_{jj,il(.)} + \sum_j \sum_{k>j} \sum_{l\neq i} Y_{jk,il(.)}.$$

Since the phenotypes conditional on genotypes are independent, we only need to consider the sum of the variances of the individual terms to calculate the variance of each marginal sum.

The statistic of interest for the marginal test is $\overline{Y}_{i.} - \overline{Y}_{.i} = Y_{i.}/n_{i.} - Y_{.i}/n_{.i}$, where

$$\begin{aligned} n_{i.} &= 2\sum_j n_{ii,jj} + \sum_j \sum_{k\neq i} n_{ik,jj} + \sum_j \sum_{l>j} n_{ii,jl} + \sum_j \sum_{k\neq i} \sum_{l>j} n_{ik,jl} \\ n_{.i} &= 2\sum_j n_{jj,ii} + \sum_j \sum_{k>j} n_{jk,ii} + \sum_j \sum_{l\neq i} n_{jj,il} + \sum_j \sum_{k>j} \sum_{l\neq i} n_{jk,il} \end{aligned}$$

The variance of the difference is the sum of the variances of the individual ratios minus two times the sum of the covariances. Covariances are non-zero, as there are terms common to both $Y_{.i}$ and $Y_{.i}$. These covariances will take on the form of weighted variances; $\mathrm{Cov}(Y_{ii,ik(.)}/n_{i.}, Y_{ii,ik(.)}/n_{.i})$.

While the variances of the sums $Y_{i.}$ and $Y_{.i}$ are not difficult to derive, the variances of the ratios $Y_{i.}/n_{i.}$ and $Y_{i.}/n_{.i}$ involve lengthy expressions, as both numerators and denominators are random variables. Simpler approximate solutions can be found using Taylor's series expansions, or by using Central Limit Theorem approximations.

While the expected values of the contrasts $\overline{Y}_{i.} - \overline{Y}_{.i}$ involve only the additive terms, the variances are functions of both the additive and dominance terms. This

means that an expected value of zero for the contrasts does not imply that the test will behave as expected under the null hypothesis that all linkage disequilibria, D_{ri}, are equal to zero.

3.5 Discussion

We have proposed a classical linear model for a quantitative trait in terms of observable marker genotypes and we have shown that there is a simple relationship between the marker being examined and an associated locus which affects the trait of interest. Additive effects for the marker alleles are functions of the additive effects of the trait locus and the disequilibria existing between the marker and the trait locus. Dominance effects of the marker genotypes are functions of the dominance effects of the trait locus and the disequilibria between the loci. For the simplified case in which random mating within the population is assumed, the relationships between the marker and the trait locus involve only linkage disequilibria. For the general model, making no assumptions about random mating, higher order disequilibria are involved.

Since this genetic model can offer insight into the degree of association between the marker and the trait, a straight-forward test of association could be performed using an analysis of variance. One method is to compare the expected mean squares between and within genotype classes. A significant F- test indicates the presence of association between the marker and the trait locus, but does not distinguish between additive and dominance effects.

Another possibility is to use the anova estimates of additive terms of the marker.

Estimates of additive effects which are significantly different from zero imply both non-zero additive effects at the trait locus and an association between the loci. Significantly non-zero dominance estimates at the marker imply non-zero dominance terms at the trait locus and associations between loci. This test, in a sense, provides a case-control type test for quantitative traits; indicating possible allele or genotype associations between the marker and the trait locus. The principle behind this method is similar to that of Luo (1998) and Luo and Suhai (1999); however our model is the classical genetic model defined on marker genotypes rather than a combination of marker and trait genotypes.

We examined several common tests of association in light of our results for additive and dominance estimators at the marker locus, including the case-control test and the TDT for dichotomous and quantitative traits. For the tests based on allelic rather than genotypic associations, we find that the primary focus is additive genetic components of the trait locus, as expected. While the expected values of the statistics used in these tests are zero when the additive effects at the trait locus are zero or when linkage disequilibria are zero, the variances of the statistics are functions of both additive and dominance terms in conjunction with non-zero disequilibria. This implies that tests based on these statistics may have an increased or decreased variance due to dominance, even when the expected values of the statistics are zero. Thus, the null hypothesis being tested must be formulated correctly if the test is to have the proper size under the null hypothesis. Hypothesizing that all $D_{ri} = 0$ provides the proper null distribution. Conditions such as small additive effects at the trait locus may substantially reduce the power of these tests in spite of the existence of strong disequilibria and overall large genetic effects.

3.6 Acknowledgments

This work was supported in part by NIH grant GM45344.

3.7 References

Allison DB. (1997) Transmission-Disequilibrium Tests for quantitative traits. Am J Hum Genet 60(3):676-90.

Comuzzie AG, Hixson JE, Almasy L, Mitchell BD, Mahaney MC, Dyer TD, Stern MP, MacCluer JW, Blangero J (1997) A major quantitative trait locus determining serum leptin levels and fat mass is located on human chromosome 2. Nature Genet 15(3):273-6.

Ewens, W.J. and R.S. Spielman. 1995. The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455-464.

Falconer DS and MacKay TFC (1996) Introduction to Quantitative Genetics (fourth edition). Longman, Essex, England.

Georges M, Nielsen D, Mackinnon M, Mishra A, Okimoto R, Pasquino AT, Sargeant LS, Sorensen A, Steele MR, Zhao X, Womack JE, Hoeschele I (1994) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics 139(2):907-20.

Kaplan NL, Martin ER, Weir BS (1997) Power studies for the transmission / disequilibrium tests with multiple alleles. Am J Hum Genet 60:691-702

Luo ZW (1998) Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Heredity 80:198-208.

Luo, Z.W. and S. Suhai. 1999. Estimating linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Genetics 151:359-371.

Martin ER (1997) Extensions of the transmission/disequilibrium test for identifying human disease genes. Ph.D. Dissertation, North Carolina State University, Chapter 3 (pp 48-86).

Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335:721-726.

Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. Human Hered 47(6):342-50.

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

Weir BS (1996) Genetic Data Analysis II. Sinauer, Sunderland, MA.

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

Weir, B.S., J. Reynolds and K.G. Dodds. 1990. The variance of sample heterozygosity. Theor Pop Biol 37:235-253.

Chapter 4

EXAMINING THE GENETIC EFFECTS WHICH ALLELIC ASSOCIATION-BASED TESTS ARE DESIGNED TO DETECT

4.1 Abstract

It is generally accepted that the degree to which a disease susceptibility locus affects phenotype influences the power to detect and localize that locus. What is equally important is the manner in which these loci act. Since most association based measures examine alleles individually rather than as genotypes, it is reasonable to assume that tests based on these measures are effective at detecting genes with large additive effects. Loci which may have a substantial biological effect on the disease of interest, but which may not have a large additive component might not be readily detected using these approaches. This idea can be quantified for association based studies by examining the relationships between phenotype and the disease susceptibility locus and those between the susceptibility locus and the marker. By applying insights from classical quantitative genetics, it is possible to formalize these relationships in a manner which allows us to determine which genetic properties are necessary for genes to display in order for common association based tests to have power to detect them. We examine these principles here, illustrating how the properties of the statistics used in these tests influence their power to detect association under different genetic models.

4.2 Introduction

There has been a surge of effort in recent years in the development of association based methods for fine mapping genes. These methods rely on population dynamics and population history to establish connections between disease susceptibility

loci and the genetic markers closest to them. In this paper, we refer to "association" as the tendency for certain marker alleles to appear along with the disease at frequencies differing from those expected by chance, and "linkage disequilibrium," (often referred to simply as "disequilibrium") as a statistical measure of this association.

In principle, after an initial amount of disequilibrium between markers and a disease susceptibility locus has been created (either due to a new mutation or to an event such as population admixture), recombination between the loci from generation to generation causes a decay in the magnitude of the disequilibrium between loci. Markers further away from the disease susceptibility locus should experience this decay more quickly than those more tightly linked. This principle serves as the basis for association based fine mapping methods, as markers showing stronger evidence for association are concluded to be closer to the gene of interest.

Two common designs used to detect association between markers and disease susceptibility are the case-control (CC) and transmission/disequilibrium tests (TDT). In the CC design, a group of affected individuals (the cases) and a matched group of unaffected individuals (the controls) are collected and genotyped. If a statistically significant difference in marker allele frequencies between the two groups is found, it is concluded that there is an association between the marker and a disease susceptibility locus. This association can exist between loci which are not linked, so that conclusions regarding localization of a disease susceptibility locus cannot be made based on this evidence alone.

The TDT, introduced by Spielman et al. (1993), provides a method of determining if there is linkage and association between a marker and a disease suscepti-

bility locus. In this test, transmission frequencies of marker alleles from heterozygous parents to their affected offspring are examined. In the case of no linkage or no disequilibrium, the fact that only affected offspring are being examined should not affect the independent segregation probabilities of marker alleles from parents to offspring. Significant deviations from equal segregation across families is evidence that the marker is linked to and in disequilibrium with a disease susceptibility locus.

In Nielsen and Weir (1999) we showed that while, in general, evidence for association between disease susceptibility and a marker implies both linkage and non-zero disequilibrium between a disease susceptibility locus and the marker, the converse is not necessarily true. It is possible that disequilibria between the loci exist, but that association will not be detected. We investigate this tendency more closely here, examining the properties of the statistical measures used in association based tests, and how these properties are influenced by the genetic effects of the susceptibility locus itself. We illustrate our findings using examples of genetic models and simulation.

4.3 Methods

We consider a general model in which disease susceptibility is affected by a locus with an arbitrary number of alleles, denoted by A_r, with population frequencies p_r. The penetrances of each genotype A_rA_s at this locus are given by ϕ_{rs}, indicating the probability of getting the disease, given that an individual has the A_rA_s genotype. For randomly mating populations, the overall disease prevalence

is $\phi = \sum_r \sum_s p_r p_s \phi_{rs}$.

We also consider a marker locus with an arbitrary number of alleles, denoted by M_i, with population frequencies q_i. Linkage disequilibrium between allele A_r at the disease susceptibility locus and marker allele M_i is $D_{ri} = P_{ri} - p_r q_i$, where P_{ri} is the frequency of the $A_r M_i$ haplotype. We denote the genotype frequency of $M_i M_j$ among affecteds as $P_{ij|\text{affected}}$ and define the summary disequilibria terms $\delta_i = \sum_r \sum_s p_s \phi_{rs} D_{ri}$ and $\delta_{ij} = \sum_r \sum_s \phi_{rs} D_{ri} D_{sj}$. Using this notation, we showed (Nielsen et al. 1998) that

$$P_{ii|\text{affected}} = q_i^2 + \frac{2q_i\delta_i}{\phi} + \frac{\delta_{ii}}{\phi} \quad (4.1)$$

$$P_{ij|\text{affected}} = 2q_i q_j + \frac{2(q_i\delta_j + q_j\delta_i)}{\phi} + \frac{2\delta_{ij}}{\phi} \quad (i \neq j), \quad (4.2)$$

By an application of Bayes' rule,

$$P_{\text{affected}|ij} = \phi + \frac{1}{q_i}\delta_i + \frac{1}{q_j}\delta_j + \frac{1}{q_i q_j}\delta_{ij}.$$

This equation is a specific instance of the linear model described in Nielsen and Weir (1999), where we modeled the genotypic effects of a marker in terms of additive terms and dominance deviations:

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

where $G_{ij}^{(m)}$ is the genotypic value of marker genotype $M_i M_j$, $\mu^{(m)}$ is the overall mean value, $\alpha_i^{(m)}$ is the additive effect of marker allele M_i, and $d_{ij}^{(m)}$ is the dominance deviation. Using the classical least squares solutions (Weir and Cockerham 1977), we showed (Nielsen and Weir 1999) that

$$\mu^{(m)} = \phi$$

$$\alpha_i^{(m)} = \delta_i/q_i = (1/q_i)\sum_r \alpha_r D_{ri} \quad (4.3)$$

$$d_{ij}^{(m)} = \delta_{ij}/(q_i q_j) = (1/q_i q_j)\sum_{r,s} d_{rs} D_{ri} D_{sj}, \quad (4.4)$$

where α_r is the additive effect of allele A_r at the trait locus (we referred to α_r as ϕ_r in Nielsen et al. 1998), and d_{rs} is the dominance deviation for trait locus genotype A_rA_s. This shows that the additive effect of marker allele M_i is a function of the additive effects of the A_r alleles at the trait locus, weighted by the linkage disequilibria between the marker and the trait locus. The dominance deviation for marker genotype M_iM_j is a function of the dominance deviations of the trait locus genotypes A_rA_s, weighted by the linkage disequilibria. In this case, the genetic trait of interest is disease susceptibility. This parameterization, using the classical quantitatitive model, allows us to examine various association-based tests to determine the genetic properties to which they are sensitive.

4.3.1 Case-Control Test

In the case-control design, significant differences in marker allele frequencies between cases and controls is evidence for rejecting a null hypothesis of no association between the marker and a locus affecting the trait. For n cases and n controls, these differences can be measured by

$$\tilde{Q}_i = \tilde{q}_{i|\text{affected}} - \tilde{q}_{i|\text{unaffected}},$$

where $\tilde{q}_{i|\text{affected}} = n_{i|\text{affected}}/2n$ and $\tilde{q}_{i|\text{unaffected}} = n_{i|\text{unaffected}}/2n$. Using Equations 4.1 and 4.2, the expected value of $\tilde{Q}_i$ is

$$\text{E}[\tilde{Q}_i] = \text{E}[\tilde{q}_{i|\text{affected}}] - \text{E}[\tilde{q}_{i|\text{unaffected}}]$$

$$
\begin{aligned}
&= q_{i|\text{affected}} - q_{i|\text{unaffected}} \\
&= q_i + \frac{\delta_i}{\phi} - \left(q_i - \frac{\delta_i}{1-\phi} \right) \\
&= \frac{\delta_i}{\phi(1-\phi)}.
\end{aligned}
$$

If all the δ_is are zero, $\mathrm{E}[\tilde{Q}_i] = 0$. From Equation 4.3, this is the case when the linkage disequilibria, D_{ri} between the marker and the disease susceptibility locus are all zero. It is also be the case, however, if the additive effects of the trait locus are zero, or if the disequilibrium weighted additive effects cancel out in the summary measures δ_i.

In order to characterize the properties of a test statistic which is based on $\tilde{Q}_i$, it is necessary to calculate its variance. Assuming the cases are drawn independently of the controls, for n cases and n controls,

$$
\mathrm{Var}(\tilde{Q}_i) = \mathrm{Var}(\tilde{q}_{i|\text{affected}}) + \mathrm{Var}(\tilde{q}_{i|\text{unaffected}}).
$$

The variance of the allele frequency estimator $\tilde{q}_i = n_i/2n$ is

$$
\mathrm{Var}(\tilde{q}_i) = \frac{1}{2n}(q_i + P_{ii} - 2q_i^2)
$$

(Weir 1996), so that

$$
\begin{aligned}
\mathrm{Var}(\tilde{Q}_i) &= \frac{1}{2n}\Big(q_{i|\text{affected}} + P_{ii|\text{affected}} - 2q^2_{i|\text{affected}} \\
&\qquad + q_{i|\text{unaffected}} + P_{ii|\text{unaffected}} - 2q^2_{i|\text{unaffected}} \Big) \\
&= \frac{1}{2n}\left[2q_i(1-q_i) + \delta'_i(1-2\phi)(1-2q_i) \right. \\
&\qquad \left. -2(\delta'_i)^2(1-2\phi+2\phi^2) + \delta'_{ii}(1-2\phi) \right],
\end{aligned}
$$

where $\delta'_i = \delta_i/\phi(1-\phi)$ and $\delta'_{ii} = \delta_{ii}/\phi(1-\phi)$. Whereas the expected value of $\tilde{Q}_i$ depends only on the additive components of the trait at the disease susceptibility

locus, the variance of $\tilde{Q}_i$ involves both the additive components and the dominance deviations at this locus. Therefore, although a test based on $\tilde{Q}_i$ may have a zero expectation even if there are non-zero disequilibria (D_{ri}), the potentially increased or decreased variance may increase or decrease the rejection region of the test.

If the summary disequilibria terms δ_i and δ_{ii} are zero, the variance of $\tilde{Q}_i$ reduces to

$$\mathrm{Var}(\tilde{Q}_i) \quad = \quad \frac{q_i(1-q_i)}{n}.$$

The usual statistic for the case-control test when n cases and n controls are sampled is

$$X^2_{CC} = 2n \sum_i \frac{\left[\tilde{q}_{i|\text{affected}} - \tilde{q}_{i|\text{unaffected}}\right]^2}{\tilde{q}_{i|\text{affected}} + \tilde{q}_{i|\text{unaffected}}},$$

which is a sum over the squared differences, $\tilde{Q}_i{}^2$, divided by an estimator of the variance of those differences under the null hypothesis of all $D_{ri} = 0$. As stated above, under this null hypothesis, the expected values of all the $\tilde{Q}_i$ are equal to zero, so this statistic is asymptotically chi-squared in distribution with degrees of freedom equal to the number of marker alleles minus one.

Since the variance of $\tilde{Q}_i$ depends on δ_{ii}, X^2_{CC} may be larger or smaller than would be expected under the hypothesis that all $D_{ri} = 0$, even if the expected value of all the $\tilde{Q}_i$s are zero. If the variance is increased, there may be low power to detect a disease susceptibility locus with a large genetic effect but with small additive effects, but the rejection region will be larger than the size under the hypothesis that all the D_{ri} are zero. If the variance is reduced due to the δ_{ii} terms, X^2_{CC} is reduced. This decreases the power of the test to smaller than the size expected under the null hypothesis that all $D_{ri} = 0$.

A measure of the non-centrality parameter of X^2_{CC} is λ_{CC}:

$$\lambda_{CC} = 2n \sum_i \frac{(\delta_i')^2}{2q_i + \delta_i'(1-2\phi)}$$

The sum in this expression was defined by Kaplan et al. (1997) as I^*. From this, it is evident that λ_{CC} increases as the δ_is increase.

4.3.2 Transmission/Disequilibrium Test

In the basic form of the TDT, trios which include a single affected offspring and its parents are collected (Spielman et al. 1993). Under the null hypothesis of no linkage or no disequilibrium, M_iM_j parents are expected to transmit M_i alleles and M_j alleles in equal frequencies to an affected offspring. A significant deviation from this equal transmission frequency is evidence for linkage and association between the marker and a locus affecting the disease.

Differences of transmission rates are estimated by genotyping n parent-offspring trios and filling out the transmitted/non-transmitted allele table, where n_{ij} is the number of offspring who received an M_i allele from a M_iM_j parent. If the difference $n_{ij} - n_{ji}$ deviates significantly from zero, this is considered to be evidence for unequal transmission frequencies of alleles to affected offspring. The probability that an affected offspring receives M_i from an M_iM_j parent is

$$P_{ij} = q_iq_j + [(1-c)q_j\delta_i + cq_i\delta_j]/\phi$$

(Nielsen and Weir 1999). For n parent-offspring trios, or $2n$ transmissions, the expected value of the difference $n_{ij} - n_{ji}$ is $2n(P_{ij} - P_{ji})$, where

$$P_{ij} - P_{ji} = (1-2c)(q_j\delta_i - q_i\delta_j)/\phi$$

$$= \quad (1-2c)(\delta_i/q_i - \delta_j/q_j)q_i q_j/\phi.$$

Since δ_i/q_i is the additive effect of marker allele M_i, $\delta_i/q_i - \delta_j/q_j$ is the average effect of allele substitution (Falconer and MacKay 1996). This shows that $P_{ij} - P_{ji}$ is composed of a factor which is zero if $c = 1/2$ and another factor which is zero if the average effect of allele substitution is zero. The latter is true if $\delta_i = \delta_j = 0$, but is also true if each allele has a similar additive effect.

In calculating the variance of n_{ij}, covariances among the elements of n_{ij} must be considered. We can define these relationships by noting that

$$n_{ij} = 2n_{ii,jj} + \sum_{k \neq i} n_{ik,jj} + \sum_{l \neq j} n_{ii,jl} + \sum_{k \neq i} \sum_{l \neq j} n_{ik,jl}, \tag{4.5}$$

where $n_{ik,jl}$ is the number of trios for which the parents have genotypes M_iM_j and M_kM_l and the affected offspring has genotype M_iM_k. Assuming trios are chosen independently of one another, the $n_{ik,jl}$ terms follow a multinomial distribution with probabilities $P_{ik,jl}$. A trio must be considered as a unit, rather than as two sets of parent-offspring pairs, since the offspring's affection status relies on its entire genotype, and trios are chosen based on the affection status of the offspring. By taking the variance of the sum in Equation 4.5 as the sum of the variances plus two times the sum of the covariances,

$$\text{Var}(n_{ij}) \quad = \quad 2n(P_{ii,jj} + P_{ij} - 2P_{ij}^2).$$

This variance involves the additive terms δ_i and the dominance terms δ_{ij}.

It is of interest to note that if $P_{ii,jj}^2 = P_{ij}^2$, the variance reduces to

$$\text{Var}(n_{ij}) \quad = \quad 2nP_{ij}(1 - P_{ij}),$$

which does not rely on the dominance deviations. This is true if there is Hardy-Weinberg equilibrium among the affected individuals at the marker locus. In Nielsen et al. (1998), we showed that by sampling affected individuals from a population in Hardy-Weinberg equilibrium, markers which are in linkage disequilibrium with a disease susceptibility locus may display Hardy-Weinberg disequilibrium within the affected sample. The degree of Hardy-Weinberg disequilibrium created within the sample depends on the genetic model of the disease susceptibility locus, and on the amount of linkage disequilibrium present in the whole population. If the disease susceptibility locus acts in a multiplicative manner, whereby the probability of becoming affected for a given genotype is a product of the effects of the individual alleles, Hardy-Weinberg disequilibrium is not created in a sample of affected individuals, even if the marker is in complete linkage disequilibrium with the disease susceptibility locus. In this case, the expected value and the variance of n_{ij} depend only on the additive terms, δ_i.

For the contrast $n_{ij} - n_{ji}$, additional covariances terms must be considered:

$$\begin{aligned} \mathrm{Var}(n_{ij} - n_{ji}) &= 2n(P_{ii,jj} + P_{ij} - 2P_{ij}^2 + P_{jj,ii} + P_{ji} - 2P_{ji}^2) \\ &\quad -4n\mathrm{Cov}(n_{ij}, n_{ji}), \end{aligned}$$

which is a lengthy expression involving the additive and dominance terms. The dependence on the dominance deviations causes the same phenomenon as seen with X^2_{CC}, where power may be increased or decreased even if the expectations of the $n_{ij} - n_{ji}$ contrasts are zero. A null hypothesis of $c = 1/2$ or all $D_{ri} = 0$ provides a proper size test. In the specific case in which the genetic model of the susceptibility locus is multiplicative, the hypothesis that all $\delta_i = 0$ yields a valid

test.

Bickeböller and Clerget-Darpoux (1995) suggested the test for differences of transmission rates:

$$T_c = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}.$$

This statistic is similar in form to X^2_{CC}, in that it is composed of the squares of the differences in transmission counts, divided by an estimator of the variances of the differences under the null hypothesis that all the $D_{ri} = 0$ are zero or $c = 1/2$. Under this null hypothesis, the expected values of the difference in counts, $n_{ij} - n_{ji}$, are zero for all i and j, and this statistic is asymptotically chi-squared in distribution, with $m(m-1)/2$ degrees of freedom for m marker alleles.

A similar test with fewer degrees of freedom was proposed by Bickeböller and Clerget-Darpoux (1995) using contrasts of marginal allele counts, $n_{i.} - n_{.i}$, where $n_{i.} = \sum_j n_{ij}$:

$$T_m = \sum_i \frac{(n_{i.} - n_{.i})^2}{n_{i.} + n_{.i}}.$$

This statistic is asymptotically chi-squared with $(m-1)$ degrees of freedom for m marker alleles under the null hypothesis that all $D_{ri} = 0$ or $c = 1/2$.

The expected value of the contrast $n_{i.} - n_{.i}$ is

$$\begin{aligned} \mathrm{E}(n_{i.} - n_{.i}) &= \mathrm{E}\left(\sum_j n_{ij} - \sum_j n_{ji}\right) \\ &= \sum_j \mathrm{E}(n_{ij} - n_{ji}) \\ &= 2n(1-2c)\delta_i/\phi. \end{aligned}$$

This is zero when the additive effect of allele M_i is zero or when $c = 1/2$. The variances also rely on the dominance terms.

Homozygous M_iM_i parents provide no information regarding frequencies of transmission, and can be excluded from the marginal allele counts when formulating a test statistic. Spielman and Ewens (1996) proposed the test

$$T_{mhet} = \frac{m-1}{m} \sum_i \frac{(n_{i.} - n_{.i})^2}{n_{i.} + n_{.i} - 2n_{ii}}.$$

Excluding the n_{ii} values does not change the statistical properties of the contrasts $n_{i.} - n_{.i}$, however, since the n_{ii} term is common to both marginal counts and cancels in the difference.

4.3.3 Simulations

To illustrate our theoretical results, we created several genetic models designed to test different principles. As a general basis for our genetic models, we used the effect of the gene which encodes the protein Apolipoprotein E (apoE) on late-onset Alzheimer disease, for which much is understood. This gene, denoted APOE, has three common allelic variants; alleles $\epsilon 2$, $\epsilon 3$, and $\epsilon 4$. The $\epsilon 3$ allele is the most common in most populations, and is considered the "ancestral" allele (Brouwer et al. 1996). The $\epsilon 2$ and $\epsilon 4$ alleles each differ from $\epsilon 3$ by single nucleotide substitutions, in codons 112 and 158 (Brouwer et al. 1996). The different APOE genotypes, rather than affecting whether or not a person is susceptibile to disease, appear to affect the age at which disease onset is expected to occur (Strittmatter et al. 1996). The APOE alleles $\epsilon 2$ and $\epsilon 4$ appear to have quite different effects on the Alzheimer disease etiology; allele $\epsilon 2$ appears to be protective, reducing the probability of disease onset by age 85, while the $\epsilon 4$ allele appears to increase risk substantially by age 85 (Corder et al. 1994). When we consider penetrances of the

different genotypes, we consider the probability of getting the disease by age 85, which allows us to use genotypic penetrances in the usual way.

For the actual APOE population parameters, we used the estimates derived in Farrer et al. (1997) for a Caucasian population (the population with information from the largest number of individuals). In this meta-analysis, Farrer et al. (1997) estimated population allele frequencies for the APOE alleles at 8.4%, 77.9% and 13.7% for alleles $\epsilon 2$, $\epsilon 3$ and $\epsilon 4$, respectively. They also reported finding no evidence for deviation from Hardy-Weinberg equilibrium at the APOE locus.

The penetrances of the APOE genotypes we used were derived using the estimates of the odds ratios for the Caucasion population from Farrer et al. (1997) and the approximation that 60% of $\epsilon 4\epsilon 4$ homozygous individuals may be expected to display symptoms of Alzheimer disease by age 85. This yielded an overall prevalence (probability of the disease by age 85), of about 7.3%. This appears to be reasonable compared with published population estimates of prevalences for individuals by age 85 (Rocca et al. 1991). Table 4.1 summarizes the parameters of the APOE model.

For the first set of simulations, we were interested in examining the consequences of additive and dominance effects on the CC test and the TDT. As it is intuitively the case that loci with different penetrances should yield different test statistic values, we chose to examine three models which had quite different penetrance values but for which the additive effects of the alleles were the same. We considered a genetic basis similar to the APOE model, with susceptibility alleles $\epsilon 2$, $\epsilon 3$ and $\epsilon 4$ (denoted by subscripts 2, 3 and 4) with population frequencies of the APOE model (Table 4.1). Of these APOE-type models, described in Table 4.2,

Model 1 was additive only; the effects of the genotypes were the sum of the individual allele effects. The APOE model itself could not be reduced to its additive components alone, as penetrance values are restricted to be between zero and one. Instead, we chose a set of additive penetrance values which could be modified with appropriate dominance terms to be similar to the penetrance values of the APOE model. This became Model 2. Model 3 was chosen to have the same additive components as Models 1 and 2, but to have quite different overall penetrances. Model 4 was a null model, for which disease susceptibility was independent of genotype. This model was included to ensure the tests had the appropriate size under the null hypothesis.

The penetrances of Models 5 and 6, summarized in Table 4.3, included dominance components only; additive effects of the alleles were all zero. These are models for which there is a large genetic effect on disease phenotype from this locus, though the effects can only be partitioned by examining whole genotypes. If the association-based tests used to test these models are sensitive only to additive effects, then we would expect to see little power to detect these loci. Allele frequencies for these models were all equal to 1/3. Model 7 was another null model.

To perform the TDT and the CC test, we considered the optimal case, in which the marker locus was the disease susceptibility locus. This provided a marker which was completely linked and in complete disequilibrium with the susceptibility locus. For the TDT, we also considered a marker with three alleles in complete disequilibrium with the susceptibility locus alleles, but which was unlinked. This was to verify the size of the TDT when $c = 1/2$.

Since we were interested in testing our results under the appropriate theorized

conditions rather than determining effects of variation due to evolutionary forces, we created populations with parameters taking on their expected values under the genetic model being studied. This included haplotype and allele frequencies for the first set of simulations. For each model, we drew population samples of 50 cases and 50 controls and 50 parent-affected offspring trios. We replicated each simulation 10,000 times, and for each sample, calculated X^2_{CC} and T_{mhet}. We recorded both the value of the test statistics and whether these values exceeded an appropriate significance threshhold. This threshhold was determined based on a size $\alpha = 0.05$ test following a chi-squared with two degrees of freedom under the null hypothesis that all $D_{ri} = 0$ for the CC test, and that all $D_{ri} = 0$ or $c = 1/2$ for the TDT.

For the second set of simulations, we were interested in examining the properties of these tests using markers distinct from the disease susceptibility locus. For these simulations, we used the APOE model (Table 4.1) for the disease susceptibility locus and assumed a marker with two alleles at frequencies $q_1 = 0.15$ and $q_2 = 0.85$.

In the first of these studies we examined a marker which was in linkage disequilibrium with and completely linked to the susceptibility locus ($c = 0$). However, the magnitudes and signs of the disequilibria and the additive effects of the susceptibility alleles were such that they cancelled in the summary term δ_i. If the tests of association are sensitive to linkage disequilibrium alone, rather than to the confounded effect including disequilibria and allele effects, then there should still be some evidence to detect this gene. Otherwise, this is an example where there is non-zero linkage disequilibrium between the marker and the susceptibility locus, but for which little association may be detected. As before, we sampled 50 cases and 50 controls for the CC test and 50 trios for the TDT then recorded the

resulting test statistics and whether they exceeded an $\alpha = 0.05$ significance level for a chi-squared distribution with 1 degree of freedom.

Finally for these studies, we examined two markers at different distances from the disease susceptibility locus. Here we were interested in testing the principle that markers closer to the susceptibility locus are expected, on average, to show a stronger association with the disease. We created two markers, both with the same initial level of disequilibrium with the susceptibility locus, but at different distances and with different haplotype patterns. We eliminated the effect of random evolutionary variation by creating a population which contained haplotype frequencies equal to their expected values at each generation. The first marker was very tightly linked with the susceptibility locus, with $c = 0.0005$ (0.05% recombination). For this marker, the rarer marker allele was in complete linkage disequilibrium with disease susceptibility allele $\epsilon 2$ and the common marker allele was in complete disequilibrium with allele $\epsilon 3$, the common susceptibility allele. The second marker was linked to the susceptibility locus at $c = 0.005$ (0.5% recombination). The rare allele for this marker was in disequilibrium with susceptibility allele $\epsilon 4$, at an initial level equal to the initial disequilibrium between the first marker and allele $\epsilon 2$ (Table 4.5). This provided two markers with identical magnitudes of disequilibrium with the disease susceptibility locus, but at different distances and with different patterns of association. We sampled individuals from the population at generations $g = 0$, $g = 50$ and $g = 100$ and performed the CC test and the TDT, recording the results as before.

4.4 Results

The first set of simulations were based on Models 1 through 3 (Table 4.2). The results are given in Table 5. These were models with quite different penetrances, but for which the additive effects of the alleles were the same. If power to detect the susceptibility locus depended on overall penetrances, we would expect the power of these tests to be quite different. In fact, we found that the values of both X^2_{CC} and T_{mhet} averaged across the 10,000 simulations were very similar for Models 1 through 3. The variance of the statistics did differ somewhat between models, with the variance for Model 3 being higher for both X^2_{CC} and T_{mhet}. This model also displayed a slightly elevated power. Model 4 was a null model, and showed an average rejection rate very close to the $\alpha = 0.05$ size expected under the null hypothesis. For the TDT, we also tested these models for the case when $c = 1/2$. The results for these tests are also consistant with a size $\alpha = 0.05$ test.

In Models 5 and 6, the simulated gene had a large effect on disease susceptibility, but the effect was soley due to the dominance components. Under the premise that these tests are sensitive only to additive effects of the susceptibility alleles, we expected to find rejection regions close to 0.05, depending on the effect of the dominance terms on the variance of the statistics. These experiments provided quite interesting results (Table 4.7). In Model 5, the CC test rejected very close to 5% of the time. However, in Model 6, which was a very similar model, the CC test rejected 9.4% of the time – almost twice as often. This increase in the rejection region was caused by an increase in variance due to the dominance terms of the susceptibility locus. These dominance terms differ from those in Model 5

by a change in sign only. However, since the disequilibria do not change between models, new combinations of terms are created in the summary measures δ_{ij}, which are composed of both the dominance terms and the linkage disequilibria. The subsequent changes to the variances caused an overall increase in the chi-square statistic.

The results of the TDT were equally interesting. For this test, Model 5 produced a smaller size than expected had the null hypothesis been true. This was due to a reduction in the variances of the statistics from the δ_{ij} terms. For this genetic model, the false positive rate which would have occured if the null hypothesis had been true was higher than the power to detect the true effect. With Model 6, as seen for the CC test, a larger variance increased power to detect this gene to about 12%, a substantial improvement over the power under Model 5.

All results for Model 7, the null model, were consistant with expectations under the null hypothesis, as were the results for the TDT when there was free recombination between the marker and the susceptibility locus.

For the second set of simulations, we considered a diallelic marker which was in disequilibrium with and completely linked to the susceptibility locus, but for which the individual terms ($\alpha_r D_{ri}$) cancelled out in the summary measure, δ_i. The results are summarized in Table 4.8. For this model, the average value of the chi-squared test took on its expected value under the null hypothesis almost exactly, with a mean value of 1.003. The variance of the X^2_{CC} over all replicate samples was also very close to the null variance of 2, and the rejection percentage near $\alpha = 0.05$. This implies that non-zero disequilibria alone is not sufficient in detecting associations between markers and affection status.

In the final set of simulations, we considered a genetic map containing a susceptibility locus and two adjacent markers, linked at $c = 0.0005$ and $c = 0.005$. Both markers had equivalent magnitudes of disequilibrium with the susceptibility locus, but with different patterns of association. Population samples were drawn and tested for both markers using both methods at several points in the population evolution. The results for the different tests are listed in Table 4.9. For these two markers, despite the fact that the amount of disequilibrium with the susceptibility locus was initially the same, the power to detect association with the disease was quite different. At generation $g = 0$, the approximate power of the CC test was about 22% for the closer marker and over 56% for the marker further away. The results were slightly lower for the TDT. The difference in power for these two markers is caused by the fact that the susceptibility alleles most strongly associated with the alleles of the closer marker have less of an effect on susceptibility than the alleles more strongly associated with the further marker. Specifically, the effect of the model APOE $\epsilon 4$ allele (which increased susceptibility) was stronger than the effect of allele $\epsilon 2$, which decreased susceptibility. The closer marker had an overall greater degree of disequilibrium with the $\epsilon 2$ allele, the weaker allele. This difference in strength of allelic effect was substantial enough to cause a difference in the power to detect association.

As the population evolved, the disequilibria decayed over time as expected (the population was designed in such a way that this decay occurred according to expectation, with no fluctuation due to evolutionary forces). However, after 100 generations of random mating, the degree of decay was insufficient to overcome the effect of the genetic contributions of the susceptibility alleles, and the marker

showing the greater degree of association with the disease was still the further marker.

4.5 Discussion

Using principles of classical quantitative genetics, we have examined the relationships between the genetic components of a disease susceptibility locus and a marker locus in linkage disequilibrium with it. For randomly mating populations, a simple connection exists between the genetic components of the two loci. The additive effects of the marker alleles are functions of the additive effects of the susceptibility alleles and the linkage disequilibria between loci. The dominance deviations of the marker genotypes are functions of the dominance components of susceptibility locus and the disequilibria. These additive and dominance components for the marker locus are the weighted summary terms δ_i/q_i and δ_{ij}/q_iq_j.

The principle behind association based mapping relies on the decay of linkage disequilibrium between markers and disease susceptibility loci as a function of distance between loci. Commonly used statistical measures of association, however, are able only to capture the δ_i and δ_{ij} summary terms. Because of this, estimates of distances between loci based on degree of association between the marker and the disease are confounded with genetic effects.

The CC test and the TDT are both based on statistics for which the expected values depend only on the allelic δ_i terms, though the variances also depend on the genotypic δ_{ij} terms. Since these statistics increase in expectation as the magnitudes of the δ_is increase, the factors which most directly affect the power to detect

association are the strengths of the additive components of the susceptibility locus and the degree to which these alleles are in linkage disequilibrium with the marker. However, these conditions do not guarantee high power, as both additive effects and linkage disequilibrium can be positive or negative, and the weighted disequilibrium terms $\alpha_r D_{ri}$ can cancel out in the summary term δ_i. This implies that a marker locus may be in stong disequilibrium with a susceptibility locus which has a large additive effect on phenotype, but still might not exhibit strong association with the disease. Susceptibility loci which have a large effect on phenotype, but do not act in an additive fashion will be difficult to detect using these types of association based tests.

In Nielsen and Weir (1999), we examined a few of the properties of the TDT in a structured population where the assumption of random mating cannot be made. In these populations, higher order disequilibria exist which cannot be ignored. These terms cause dependencies between alleles which may allow association based test to become more sensitive to non-additive genetic effects.

4.6 References

Bickeböller, Clerget-Darpoux (1995) Statistical properties of the allelic and genotypic transmission/disequilibrium tests for multiallelic markers. Genet Epidemiol 12:856–870.

Brouwer DA, van Doormaal JJ, Muskiet FA (1996) Clinical chemistry of common apolipoprotein E isoforms. Journal of Chromatography B: Biomedical Applications 678(1):23–41.

Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Gaskell PC Jr, Rimmler JB, et al. (1994) Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. Nature Genetics 7(2):180–4.

Falconer DS, MacKay TFC (1996) Introduction to Quantitative Genetics (fourth edition). Longman, Essex, England.

Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, et al. (1997) Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: A meta-analysis. JAMA 278(16):1349–56.

Kaplan NL, Martin ER, Weir BS (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 60:691–702

Nielsen DM, Ehm MG, Weir BS (1998) Detecting Marker-Disease Association by

Testing for Hardy-Weinberg Disequilibrium at a Marker Locus. Am J Hum Genet 63:1531–1540.

Nielsen DM, Weir BS (1999) A classical setting for associations between markers and loci affecting quantitative traits. Submitted, Genet Res.

Rocca WA, Hofman A, Brayne C, Breteler MM, Clarke M, Copeland JR, Dartigues JF, et al. (1991) Frequency and distribution of Alzheimer's disease in Europe: a collaborative study of 1980-1990 prevalence findings. Annals of Neurology 30(3):381–90.

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983-989.

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

Strittmatter WJ, Roses AD (1996) Apolipoprotein E and Alzheimer's disease. Annual Review of Neuroscience 19:53-77.

Weir BS (1996) Genetic Data Analysis II. Sinauer, Sunderland, MA.

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

Table 4.1: Parameters for the APOE Model

Disease prevalence	ϕ	0.07345
Allele frequencies	p_2	0.084
	p_3	0.779
	p_4	0.137
Penetrances	ϕ_{22}	0.0432
	ϕ_{23}	0.0288
	ϕ_{24}	0.0576
	ϕ_{33}	0.048
	ϕ_{34}	0.1296
	ϕ_{44}	0.6
Additive effects	α_2	-0.0395
	α_3	-0.01589
	α_4	0.1145
Dominance deviations	d_{22}	0.04874
	d_{23}	0.01073
	d_{24}	-0.0909
	d_{33}	0.006319
	d_{34}	-0.04251
	d_{44}	0.2975

Table 4.2: Parameters of Disease Models I

Allele frequencies: $p_2 = 0.084$, $p_3 = 0.779$, $p_4 = 0.137$

	Model	1	2	3	4
Disease prevalence	ϕ	0.08371	0.08371	0.08371	0.08371
Penetrances	ϕ_{22}	0.0029	0.04500	0.5	0.08371
	ϕ_{23}	0.01545	0.04018	0.001672	0.08371
	ϕ_{24}	0.2265	0.06	0	0.08371
	ϕ_{33}	0.028	0.02682	0.0422	0.08371
	ϕ_{34}	0.239	0.2306	0.1667	0.08371
	ϕ_{44}	0.45	0.6	1	0.08371
Additive effects	α_2	-0.0404	-0.0404	-0.0404	0
	α_3	-0.02785	-0.02785	-0.02785	0
	α_4	0.1831	0.1831	0.1831	0
Dominance deviations	d_{22}	0	4.210e-02	4.971e-01	0
	d_{23}	0	2.473e-02	-1.378e-02	0
	d_{24}	0	-1.664e-01	-2.265e-01	0
	d_{33}	0	-1.184e-03	1.420e-02	0
	d_{34}	0	-8.432e-03	-7.231e-02	0
	d_{44}	0	1.500e-01	5.500e-01	0
Allelic disequilibria[a]	δ_1	-0.04425	-0.04425	-0.04425	0
	δ_2	-0.2829	-0.2829	-0.2829	0
	δ_3	0.3271	0.3271	0.3271	0
Genotypic disequilibria[a]	δ_{11}	0	0.003873	0.04573	0
	δ_{12}	0	0.0211	-0.01175	0
	δ_{13}	0	-0.02497	-0.03398	0
	δ_{22}	0	-0.009369	0.1124	0
	δ_{23}	0	-0.01173	-0.1006	0
	δ_{33}	0	0.03671	0.1346	0

[a]Normalized by $\phi(1-\phi)$

Table 4.3: Parameters of Disease Models II

Allele frequencies: $p_2 = 1/3$, $p_3 = 1/3$, $p_4 = 1/3$

	Model	5	6	7
Penetrances	ϕ_{22}	0.05	0.75	0.05
	ϕ_{23}	0.75	0.05	0.05
	ϕ_{24}	0.75	0.05	0.05
	ϕ_{33}	0.05	0.75	0.05
	ϕ_{34}	0.75	0.05	0.05
	ϕ_{44}	0.05	0.75	0.05
Additive effects	α_2	0	0	0
	α_3	0	0	0
	α_4	0	0	0
Dominance deviations	d_{22}	-0.4667	0.4667	0
	d_{23}	0.2333	-0.2333	0
	d_{24}	0.2333	-0.2333	0
	d_{33}	-0.4667	0.4667	0
	d_{34}	0.2333	-0.2333	0
	d_{44}	-0.4667	0.4667	0
Allelic disequilibria[a]	δ_1	0	0	0
	δ_2	0	0	0
	δ_3	0	0	0
Genotypic disequilibria[a]	δ_{11}	-0.2076	0.2554	0
	δ_{12}	0.1038	-0.1277	0
	δ_{13}	0.1038	-0.1277	0
	δ_{22}	-0.2076	0.2554	0
	δ_{23}	0.1038	-0.1277	0
	δ_{33}	-0.2076	0.2554	0

[a]Normalized by $\phi(1-\phi)$

Table 4.4: Marker Association Parameters I

Allele frequencies	q_1	0.15
	q_2	0.85
Disequilibria [a]	D_{21}	0.7207
	D_{31}	-0.5692
	D_{41}	0.1053
Summary Disequilibrium	δ_1	0

[a]Normalized by $[p_r(1-p_r)q_1(1-q_1)]^{1/2}$

Table 4.5: Marker Association Parameters II

		Marker 1	Marker 2
Recombination rate	c	0.0005	0.005
Allele frequencies	q_1	0.15	0.15
	q_2	0.85	0.85
Haplotype frequencies	P_{21}	0.084	0.05805
	P_{31}	0	0
	P_{41}	0.066	0.09195
	P_{22}	0	0.02595
	P_{32}	0.779	0.779
	P_{42}	0.071	0.04505
Initial Disequilibria[a]	D_{21}	0.0714	0.04545
	D_{31}	-0.1169	-0.1169
	D_{41}	0.04545	0.0714
Summary Disequilibrium[b]	δ_1	0.06233	0.1211

[a]Not Normalized

[b]Normalized by $\phi(1-\phi)$

Table 4.6: Results for Genetic Models I

	Model	1	2	3	4
CC Results	Average χ^2_{CC}	28.69	28.79	29.33	2.015
	Variance (χ^2_{CC})	65.07	71.36	87.18	3.906
	Percent rejected	0.9998	0.9998	0.9992	0.0498
TDT Results ($c = 0$)	Average T_{mhet}	24.55	24.61	25.18	2.014
	Variance T_{mhet}	57.32	63.66	77.4	4.205
	Percent rejected	0.9984	0.9976	0.9958	0.0511
TDT Results ($c = 1/2$)	Average T_{mhet}	2.025	2.011	2.005	1.986
	Variance T_{mhet}	4.262	4.237	4.168	4.041
	Percent rejected	0.053	0.0528	0.0503	0.0492

Table 4.7: Results for Genetic Models II

	Model	5	6	7
CC Results	Average χ^2_{CC}	2.057	2.549	1.999
	Variance (χ^2_{CC})	4.194	6.328	3.987
	Percent rejected	0.0523	0.094	0.0504
TDT Results ($c = 0$)	Average T_{mhet}	1.55	2.841	2.005
	Variance T_{mhet}	2.391	7.766	3.844
	Percent rejected	0.0211	0.1199	0.049
TDT Results ($c = 1/2$)	Average T_{mhet}	2.016	1.995	1.994
	Variance T_{mhet}	3.974	3.825	3.916
	Percent rejected	0.0508	0.0475	0.0485

Table 4.8: Marker Association Results I

Summary Disequilibrium	δ_1	0
CC Results	Average χ^2_{CC}	1.003
	Variance (χ^2_{CC})	1.97
	Percent rejected	0.0515
TDT Results ($c = 0$)	Average T_{mhet}	1.008
	Variance T_{mhet}	1.972
	Percent rejected	0.054

Table 4.9: Marker Association Results II

Marker		1			2	
Generation	0	50	100	0	50	100
$D_{21}{}^a$	7.140e-02	6.964e-02	6.792e-02	4.545e-02	3.537e-02	2.753e-02
$D_{31}{}^a$	-1.168e-01	-1.140e-01	-1.111e-01	-1.168e-01	-9.095e-02	-7.078e-02
$D_{41}{}^a$	4.545e-02	4.433e-02	4.323e-02	7.140e-02	5.557e-02	4.325e-02
$\delta_1{}^b$	0.06233	0.06079	0.0593	0.1211	0.09423	0.07334
Average χ^2_{CC}	2.385	2.253	2.169	5.45	3.876	2.801
Variance (χ^2_{CC})	7.279	6.772	6.326	18.61	12.96	8.982
Average Power	0.2227	0.2052	0.194	0.5687	0.4056	0.275
Average T_{mhet}	2.093	2.086	1.994	4.658	3.281	2.459
Variance T_{mhet}	5.887	6.073	5.582	14.61	10.17	7.352
Average Power	0.1878	0.1846	0.1787	0.4977	0.3403	0.2322

[a]Not normalized

[b]Normalized by $\phi(1-\phi)$

Chapter 5

DISCUSSION

Within the context of association-based fine-mapping, there are many factors which affect the ability to detect and localize genes underlying traits of interest. Association-based methods rely on linkage disequilibrium between marker and trait loci as the basis for localizing genes, so that understanding the behaviour of this measure is fundamental in determining the properties of association-based fine-mapping techniques. Forces creating initial levels of linkage disequilibrium and those maintaining or diminishing it, amount of time since disequilibrium was created, and the properties of the techniques used to measure disequilibrium are all factors which play a role in how useful and reliable association-based methods may be for the purpose of fine-mapping genes.

Linkage disequilibrium between loci can be created by a number of evolutionary forces, including mutation, selection, population admixture, and bottlenecks followed by population expansion. After an initial level of disequilibrium has been created in a population, it is expected to decay over time at a rate which depends on the recombination rate between loci. As there is generally a higher rate of recombination between loci which are further apart than between those closer together, it is assumed that the markers which show the highest levels of association with the phenotype are the ones which are closer to genes affecting phenotype, as recombination over time will have disassociated the further markers.

There are several problems with this generalization that arise in practice. For any given marker-trait locus pair, there is a unique history associating the two. Even if the time of initial association with the gene of interest is the same for all markers being examined, linkage disequilibrium is bounded in value by allele frequencies, which differ from marker to marker. For association mapping, there is

the added problem of not directly observing the genotypes at the trait locus, but instead observing the phenotypes which are presumed to be affected by the trait locus. This is further complicated when the manner by which genotypes affect phenotype is unknown. Additionally, linkage disequilibrium is a measure of alleles and haplotypes, while phenotypes are generally affected by whole genotypes, so that phenotypic associations may not clearly indicate allelic relationships.

Associations between loci which are not linked can arise and be maintained in populations where propagating forces, such as selection or recurrent mutation, exist. Admixture and population stratification create linkage disequilibrium in the overall population between unlinked loci which only begins to decay after several generations of random mating (Ewens and Spielman 1995). Association studies which do not simultaneously test for linkage (such as the case-control test) do not distinguish between linked and unlinked markers associated with the phenotype. Tests for Hardy-Weinberg equilibrium at unlinked markers can be used to test whether spurious association (association in the absence of linkage) resulting from events such as population stratification and admixture exits. Pritchard and Rosenberg (1999) advocate performing a case-control test on multiple unrelated markers along with testing the markers in the region of interest. If the markers unlinked to the region of interest show no significant results of association, the assumption that linkage exists between the associated markers and a locus affecting the trait can be made. This does not, however, provide any true information regarding linkage, as the tests are not sensitive to increased or decreased recombination rates, as are combined tests.

Another problem which arises in practice when performing fine-mapping studies

is that the amount of association between the phenotype and the markers must generally be estimated by examining a subset of individuals from a population, or perhaps by sampling individuals from several populations. There is a very large variance associated with statistical estimates of linkage disequilibrium caused by both statistical sampling (selecting a subset of individuals from a population) and genetic sampling (sampling inherent in the transmission of genetic material from parents to offspring) (Weir, 1996). While the problem of variance due to statistical sampling might be addressed by collecting large samples or by taking replicate samples, variance due to genetic sampling is more difficult, if not impossible, to control experimentally.

These factors greatly influence the effectiveness of association-based tests for fine-scale localization. One question which arises is whether it is reasonable to expect to position genes based on determining the individual markers which display the highest levels of association with the phenotype, or whether, instead, we simply hope to be able to identify regions containing the genes of interest based on the results for groups of linked markers. If it is the latter which is the more realistic goal, then the question becomes determining how large (or narrow) we expect the regions we isolate to be. These are questions which have been addressed by various studies, both theoretically and empirically.

Jorde et al. (1994) performed a series of experiments designed to test the correlation between statistical measures of disequilibrium and physical distances in human populations for polymophic sites within and between coding regions of the APC and MCC genes (within the APC region of chromosome 5). They summarized their results together with a review of the published results from a

number of other similar studies. From their analyses and the results of the other published studies, Jorde et al. (1994) concluded that there is significant negative correlation between physical distance and linkage disequilibrium between loci in genomic regions of approximately 50-500 kb, but this correlation does not exist when the loci lie within regions smaller than about 50 kb. As noted by Jorde et al. (1994), this appears to be consistant with the idea that at very small intervals, for which recombination events are very rare, the effects of evolutionary forces such as mutation are expected to contribute sufficiently to linkage disequilibrium as to balance out the effect of recombination. Jorde et al. (1994) also note that the degree of correlation between physical distance and disequilibrium is dependent on chromosomal location. Inter-gene regions appear to experience higher levels of disequilibrium, possibly implying reduced recombination. Higher levels of disequilibrium also appear to exist in centromeric regions than telomeric regions, which appears consistant with the evidence that recombination rates are higher in telomeric regions. From the experments of Jorde et al. (1994, Figure 2), it appears that substantial levels of disequilibrium were found within regions of up to about 250-300 kb, but not in larger regions.

In an outbred population of *Drosophila melanogaster*, Long et al. (1998) performed experiments whereby physical distance and estimates of linkage disequilibrium within the *Delta* gene region could be compared. In the sample examined, linkage disequilibrium decayed very rapidly with distance, with substantial levels of disequilibrium occuring only between loci separated by 5 kb or less. While the overall pattern of disequilibrium appears to be consistant with the principle that linkage disequilibrium is inversely proportional to physical distance, there

appeared to be very large variation in the smaller regions, with values of disequilibrium ranging the entire scale. These results were consistant with the findings of many classic studies examining linkage disequilibrium within natural populations of *D. melanogaster* (Langley et al. 1987, Aquadro et al. 1992, Miyashita et al. 1993).

In a proof-of-principle study, Martin et al. (1999) examined sib-pair and case-control data for randomly selected American Caucausians for ten SNPs in the region of APOE (chromosome 19q13.2). Variants at this gene are responsible for altered levels of susceptibility to Alzheimer Disease (Corder et al. 1993). The gene was originally detected and isolated using linkage studies (Pericak-Vance et al. 1991). The proof-of-principle study was designed to determine if association-based techniques would have been effective at isolating this gene, and how close markers would need to be in order to be able to detect an association with the disease. For their sample, Martin et al. (1999) found the most significant association between the SNP marker closest to APOE and Alzheimer disease. This SNP, which was part of a different gene (APOC), showed a higher statistically significance association than did the most common APOE variant with the disease (which was also highly significant), though there is no biological evidence that this gene plays any role in Alzheimer disease etiology. Markers further from the gene showed little or no significant association with the disease. Based on their overall results, Martin et al. (1999) concluded that that loci must be within 20-40 kb for linkage disequilibrium to be detected between them for this population.

Kruglyak (1999) used coalescent simulations under various model assumptions in order to evaluate the distance for which linkage disequilibrium is expected to

extend in general populations and in population isolates. Based on his simulations, Kruglyak (1999) concludes that that linkage disequilibrium is unlikely to extend beyond an average distance of approximately 3 kb in the general population. He found similar results for isolated populations for which the founding population size is not very small. His studies, however, do not consider the effects of dynamics such as selection, admixture, and non-random mating, which are aspects of many current human populations. His results indicated that while linkage disequilibrium was inversely proportional to distance, the variation seen between disequilibrium measures in the smaller regions was also larger, with values extending the entire range of possibilities. As the distance between loci increases, average linkage disequilibrium decreases, as does its variability. This result was consistant with the empirical evidence of Long et al. (1998). It also appears to be consistant with the results of Jorde et al. (1994) presented in their Figure 2.

Jorde et al. (1994), based on their own evidence and their review of similar experiments, concluded that linkage disequilibrium correlates well with physical distance within certain size regions. This conclusion implies that linkage disequilibrium is, on average, an indirect measure of distance, and therefore that genes can be positioned based on detecting the markers showing the highest levels of disequilibrium with these genes. However, these studies are based on direct estimates of disequilibrium, for which genotypes at each loci must be observed. For smaller regions (less than 50 kb), the correlation between disequilibrium and distance is lost, though significant levels of disequilibrium do exist. The results of Long et al. (1998) and Kruglyak (1999) indicate that while disequilibrium appears or is expected to be greater within smaller regions, the variation of the measure is also

greater. This implies that while regions containing genes may be detected using linkage disequilibrium as a tool, specific positioning of genes based on this measure is not practical. In Chapters 2 and 3, we show that common measures of association designed to detect linkage disequilibrium between markers and trait loci confound disequilibrium with genetic effects. This confounding of effects, along with the problem of variability of actual disequilibrium and the variance inherent in the statistical measures of it, implies that the more realistic goal of fine-mapping may be to identify regions containing genes, rather than positioning loci along a map of markers. The size of the regions which can be isolated depends on the population history and the recombination rate at the specific location, but appear to be quite small for outbred populations. Long et al. (1998) and Kruglyak (1999) conclude that this size is not larger than approximately 3 to 5 kb, Martin et al. (1999) determined the size to be 20-40 kb, and Jorde showed evidence that these regions may be up to about 250 kb.

Once a set of markers has been examined, and significant associations detected for some subset of markers, another problem may arise. For many studies, thousands or even hundreds of thousands of markers may have been tested across the genome, and the problem of multiple testing must be addressed. With very large numbers of tests, a Bonferroni-type adjustment is generally infeasible, as the resulting significance threshhold becomes infinitesimal. In this case, alternative methods must be utilized to protect the genome-wide significance level. Zaykin et al. (1999) proposed a method whereby multiple markers can be tested and the resulting p-values combined to form a joint test, taking into consideration the correlation structure that exists between markers. The hypothesis examined in this

method is whether there is evidence of association within a region containing multiple markers, considering all markers simulataneously rather than independently. Another possibility for determining an appropriate significance threshhold when performing multiple tests is an empirical approach. A general strategy using permutation was proposed by Westfall and Young (1993) and applied to genetic data by Doerge and Churchill (1996). This method also maintains the correlation structure implicit in the genetic map. As a possible alternative strategy in addressing the multiple testing issue, approaches in data-mining, such as decision trees and neural networks, can be used initially to identify regions which appear promising, thus limiting the total number of tests to be performed.

Ultimately, any conclusions regarding significant results for association-based tests should be made with caution. Whether the most significant results actually represent the markers closest to the trait locus is not clear, and perhaps instead, the region surrounding the markers showing significant association should be examined. How large this region can be expected to be must be determined based on what is understood of the history of the population being examined and the biology of the region itself. Estimates for the sizes of these regions in human populations vary substantially. Significant associations with polymorphisms within known genes might not indicate that the site being tested is actually the functional polymorphism. Other genes in the region and other polymorphisms within the same gene should also be considered. However, in spite of these cautions, the ability for association-based mapping methods to narrow significantly the regions containing genes of interest is substantial. The degree of resolution potentially provided by these methods should allow a much quicker transition from p-values to DNA sequences.

5.1 References

Aquadro CF, Jennings RM Jr, Bland MM, Laurie CC, Langley CH (1992) Patterns of naturally occurring restriction map variation, dopa decarboxylase activity variation and linkage disequilibrium in the Ddc gene region of Drosophila melanogaster. Genetics 132(2):443–52

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261:921-3

Doerge RW and Churchill GA (1996) Permutation tests for multiple loci affecting quantitative character. Genetics 142:285–294

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision and admixture. Am J Hum Genet 57:455–464

Jorde LB, Watkins WS, Carlson M, Groden J, Albertsen H, Thliveris A, Leppert M (1999) Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. Am J Hum Genet 54:884–898

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144

Langley CH and Aquadro CF (1987) Restriction-map variation in natural popula-

tions of Drosophila melanogaster: white-locus region. Molecular Biology and Evolution 4(6):651–63

Long AD, Lyman RF, Langley CH, and Mackay TFC (1998) Two Sites in the *Delta* gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. Genetics 149:999–1017

Martin ER, Gilbert JR, Lai E, Riley J, Slotterbeck BD, Rogala AR, Sipe CA, Grubber J, Purvis I, Pericak-Vance MA, Roses AD, Vance JM (1999) Association mapping of APOE with single single nucleotide polymorphisms (in preparation)

Miyashita NT, Aguade M, Langley CH (1993) Linkage disequilibrium in the white locus region of Drosophila melanogaster. Genetical Research 62(2):101–9

Pericak-Vance MA, Bebout JL, Gaskell PC, Yamaoka LH, Hung W-Y, Alberts MJ, Walker AP, et al. (1991) Linkage studies in familial Alzheimer's disease: evidence for chromosome 19 linkage. Am J Hum Gen 48:1034–1050

Pritchard JK and Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

Westfall PH and Young SS (1993) Resampling-Based Multiple Testing. Wiley, New York

Zaykin D, Zhivotovsky LA and Weir BS (1999) Combining p-values obtained from multiple tests of significance (submitted: Biometrics)