Epistasis in Complex Human Traits

Jordana Tzenova Bell

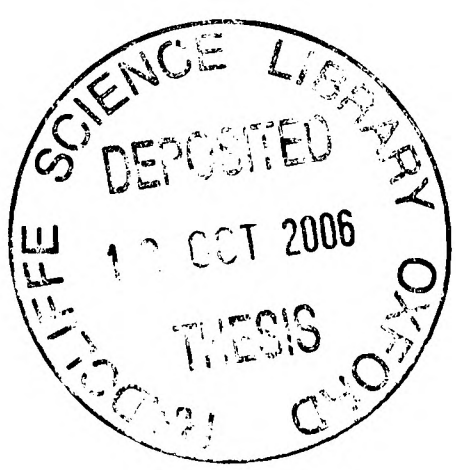
Keble College

Department of Cardiovascular Medicine

Wellcome Trust Centre for Human Genetics

University of Oxford

Thesis submitted for the degree of Doctor of Philosophy



Epistasis in Complex Human Traits

Jordana Tzenova Bell

Keble College
Department of Cardiovascular Medicine
Thesis submitted for the degree of Doctor of Philosophy, Hilary term 2006

ABSTRACT

Epistasis (gene-gene interaction) is a universal component of common complex genetic traits. The identification and characterization of epistatic interactions are crucial to a full understanding of the complex genetic mechanisms that underlie human disease. The aim of this thesis is to examine epistasis in non-parametric linkage analysis of human complex traits, with an emphasis on the affected sibling pair (ASP) study design.

Following an overview of approaches that model and detect epistasis in linkage analysis of human complex traits, I present an extension of a two-locus non-parametric linkage method in ASPs. The new two-locus approach, Merloc, jointly models pair-wise interactions between susceptibility loci in different types of affected relative pairs and estimates of the most likely underlying genetic model for a pair-wise interaction, implemented to genome-wide applications. To test the performance of the approach, Merloc was compared to two multilocus non-parametric conditional linkage approaches. Power and type I error rates under null, single-locus, and two-locus genetic models of epistasis and heterogeneity indicated that Merloc out-performed the other methods.

The method was applied to type 2 diabetes data to assess the evidence for epistasis between two susceptibility loci. Significant evidence for epistasis was obtained supporting previous findings from conditional interaction analysis. A search through the space of parametric two-locus models indicated that nine two-locus models best approximated the pair-wise interaction.

Genome-wide strategies to detect epistasis were also examined in this thesis and the simultaneous search for genome-wide interactions was explored in detail. Two-dimensional (2D) linkage scans were performed using Merloc in three complex traits, essential hypertension, autism, and type 2 diabetes. Several peaks were detected in the two-dimensional likelihood surfaces with genome-wide suggestive evidence for linkage. Extensive simulations were used to examine the distribution of the test statistic under the null hypothesis in the context of two-dimensional linkage scans.

Finally, two main extensions of this approach were considered - linkage approaches to examine more than two loci, and extending the method in this study to include a test of association.

ACKNOWLEGDMENTS

I would like to thank my supervisor, Martin Farrall, for his guidance, for providing a challenging environment, and for his motivation. Lon Cardon and Jonathan Flint were my transfer-report advisors and their helpful comments improved my thesis. I would like to thank Steven Wiltshire for many insightful discussions and collaborations in type 2 diabetes. The BRIGHT consortium[1] provided extensive and prolonged access to their data for analysis. I thank Chris Wallace for helping with data clean-up and map problems. Janine Lamb kindly provided access to the IMGSAC autism data for analysis. I would like to acknowledge the Warren 2 consortium[2] for providing access to the type 2 diabetes data. The PROCARDIS group, especially Fiona Green, John Peden, Helen Broadbent, and Elizabeth Taylor gave me helpful advice. I also thank the Bioinformatics group and in particular John Broxholme and Stacey Cherny, for computing and statistical help.

I would like to thank a number of people at the Wellcome Trust Centre in Oxford for their advice, discussions, and laughter, in particular Gbenga Kazeem, Blanca Herrera, William Valdar, Emma Banfield, Caroline Durrant, and Denise Brockleback. Finally, I want to especially thank Tom, and my parents and my sister.

I was funded by the PGS-B doctoral scholarship from the Natural Sciences and Engineering Research Council of Canada, the PhD scholarship from the Fonds quebecois de la recherche sur la nature et les technologies from Quebec, and the Clarendon Fund scholarship from the University of Oxford.

---

[1] BRIGHT consortium members: Mark Caulfield, Nigel Benjamin, Morris Brown, Deavid Clayton, John Connell, Anna Dominiczak, Martin Farrall, Mark Lathrop, Patricia Munroe, Nilesh Samani, John Webster.

[2] Warren 2 consortium members: Tim Frayling, Andrew Hattersley, Graham Hitman, Jonathan Levy, Mark McCarthy, Steve O'Rahilly, Mike Sampson, Mark Walker.

TABLE OF CONTENTS

ABSTRACT 2

ACKNOWLEGDMENTS 3

TABLE OF CONTENTS 4

LIST OF TABLES 6

LIST OF FIGURES 7

CHAPTER I. Introduction 8

1.1 Epistasis – a historical perspective 8

1.2 Epistasis in human genetics 12
1.2.1 Models 13
1.2.2 Methods 16
1.2.3 Power 18
1.2.4 Search strategies 19

1.3 Thesis summary 21

CHAPTER II. Merloc: a two-locus non-parametric linkage method for genome-wide applications 23

2.1 Introduction 23

2.2 Methods 23
2.2.1 Two-locus non-parametric linkage test 24
2.2.2 Extension to different types of affected relative pairs 26
2.2.3 Two-locus genetic models 30
2.2.4 Likelihood ratio statistic: null distribution and software implementation 34

2.3 Application to type 1 diabetes 41
2.3.1 Type 1 diabetes results 41
2.3.2 Simulations 45

2.4 Discussion 48

CHAPTER III. Comparison of multilocus linkage methods 50

3.1 Introduction 50

3.2 Methods 50
3.2.1 Genetic models 52
3.2.2 Multilocus Linkage Methods 56

3.3 Results 59
3.3.1 Significance thresholds and Type I error 59
3.3.2 Power 64
3.3.3 Parameter estimates 71

3.4 Discussion 74

CHAPTER IV. Interaction between type 2 diabetes susceptibility loci on chromosomes 1q21-q25 and 10q23-q26 84

4.1 Introduction 84

4.2 T2D data sets 84

4.3 Analysis 86
4.3.1 Two-locus analyses 86
4.3.2 Two-locus parametric model search 93
4.3.3 Localization support intervals 98

4.4 Discussion 99

CHAPTER V. Two-dimensional genome scan of hypertension 104

5.1 Introduction 104

5.2 BRIGHT dataset 105

5.3 Determining significance thresholds in a 2D scan 105

5.4 BRIGHT Analysis Results 110
5.4.1 Single-locus results 110
5.4.2 Two-locus results 111
5.4.3 Sensitivity to map misspecification 117
5.4.4 Locus-counting results 120

5.5 Discussion 121

CHAPTER VI. Two-dimensional linkage scans of complex traits 129

6.1 Introduction 129

6.2 Autism 129
6.2.1 2D scan 130
6.2.2 Chromosome 7 134

6.3 Type 2 Diabetes 137
6.3.1 2D scan 138

6.4 Comparison of 2D scans across complex traits 141

CHAPTER VII. Extensions and Conclusion 150

7.1 Linkage approaches to examine more than two genes 150

7.2 Association studies 155

7.3 Conclusion 158

APPENDIX A. Variance components in 13 two-locus models 160

APPENDIX B. Genome-scans for hypertension and variation in blood pressure. 163

APPENDIX C. Candidate genes at peak coordinates in the BRIGHT 2D scan 165

ASSOCIATED PUBLICATIONS 170

REFERENCES 171

LIST OF TABLES

Table 2.1. Covariances for ASP sharing *i* alleles at locus 1 and *j* alleles at locus 2....25
Table 2.2 Coefficients of relatedness and probabilities of sharing two alleles IBD....27
Table 2.3. Pointwise significance thresholds for unlinked pairs of loci. 36
Table 2.4. Maximum-likelihood estimates of IBD probabilities in type 1 diabetes....46
Table 3.1. Genetic models in the multilocus comparison simulations. 53
Table 3.2 Significance thresholds in Merloc, Genehunter-Plus, and IBD regression. 61
Table 3.3 Type I error and power in Merloc, GHP, and IBDreg for models m0-m3..63
Table 3.4. Comparison across Merloc, GHP, and IBDreg....71
Table 3.5 Parameter estimates in Merloc, GHP, and IBDreg....72
Table 3.6. Variance components from 13 two-locus models....81
Table 4.1. Maximum-likelihood estimates of IBD sharing in the British T2D data. ..90
Table 4.2. Merloc results in the T2D data sets on chromosomes 1 and 10. 91
Table 4.3. Variance components estimates in the British T2D data....101
Table 5.1 Most significant peaks from the BRIGHT 2D scan....113
Table 5.2 Genome-wide suggestive results from the hypertension 2D scan. 116
Table 6.1. Genome-wide suggestive results from the autism 2D scan. 133
Table 6.2. Genome-wide suggestive results from the T2D 2D scan. 140
Table 6.3 Comparison of unlinked pairwise suggestive results across 2D scans. 143
Table B.1 Significant and suggestive results from 25 genome scans of hypertension or variation in blood pressure....163
Table C.1. Summary of molecular interactions in BIND. 168

LIST OF FIGURES

Figure 2.1. The range of plausible epsilon values. 33
Figure 2.2 Distribution of the two-locus MLS under the null. 37
Figure 2.3 Genotype configurations reflecting IBD sharing states..... 39
Figure 2.4. T1D chromosome 6 results in the data of Mein et al. (1998)..... 44
Figure 2.5. T1D chromosome 6 simulation results..... 47
Figure 3.1. Parameters specifying the two-locus parametric models examined..... 51
Figure 3.2 Power estimates from the parametric two-locus Lod score in GHT. 64
Figure 3.3. Power estimates for MLS statistics in Merloc at $\alpha = 0.05$ compared to m0 as the null model. 66
Figure 3.4. Power estimates in GHP at $\alpha = 0.05$..... 69
Figure 3.5. Power estimate from IBDreg at $\alpha = 0.05$..... 70
Figure 3.6. Histograms of the estimates of ε..... 73
Figure 4.1. Two-locus results in the British data. 88
Figure 4.2 Histogram of the two-locus parametric Lod scores..... 96
Figure 4.3. Penetrance structures of the 113 parametric models with Lod > 4.45. 97
Figure 5.1. 2D significance thresholds for linked pairs of genes..... 107
Figure 5.2. Genome scans of essential hypertension. 111
Figure 5.3. Fine-scale examination 2D peaks. 114
Figure 5.4. BRIGHT genome analysis under the Marshfield genetic map..... 119
Figure 5.5. Two-locus counting results for unlinked loci. 121
Figure 6.1. Autism genome-wide linkage scans. 131
Figure 6.2. Autism chromosome 7 results. 136
Figure 6.3. T2D genome-wide linkage scans..... 139
Figure 6.4 Schematic representation of the 2D suggestive interactions. 147

CHAPTER I. Introduction

The identification of susceptibility genes contributing to human complex disease is of great medical and scientific interest. Current methods used to locate genes involved in human traits are based on linkage analysis and association tests, and aim to identify genes with major phenotypic effects. However, difficulties in mapping susceptibility loci have frustrated progress in this field, perhaps because of insufficient sample sizes or low power to detect loci of modest main effects. An alternative explanation is that the effects of individual genes are not independent in predicting complex phenotypes and that epistasis (gene-interaction) contributes substantially to many human diseases.

1.1 Epistasis – a historical perspective

At the beginning of the 20th century the re-discovery of Gregor Mendel's experiments (1865) sparked much interest in heredity and set in motion the progress in genetic research that persists into the present day. The relevance of Mendel's work to human traits was first outlined by Garrod (1902) in alkaptonuria, a rare metabolic condition. Following the success of the first gene-mapping experiments in the fruit-fly (Morgan 1910; Sturtevant 1913), interest focused on localizing genes involved in human traits. Subsequent empirical and theoretical studies by Haldane, Fisher, Wright, Penrose, Morton and many others (see Freire-Maia 1990; Lynch and Walsh 1998; Risch 2000; Hoh and Ott 2003) have made gene-mapping in humans feasible and have contributed to a broader understanding of genetics and the inherited basis of human disease.

William Bateson (1909) first coined the term 'epistastic' to describe the case in which alleles at one locus mask the effect of alleles at a different locus. This view

of gene-interaction is now described as physiological or mechanistic epistasis, while for gene-interaction to be identified in the population (as population or statistical epistasis) there must be allelic variation at all relevant loci (Brodie 2000). Fisher (1918) expanded the definition of epistasis to population genetics by describing epistasis as the deviation from additive effects of multiple loci on the trait. The phenotypic variance (V_P) may be partitioned into hereditary (V_G), environmental (V_E) and interaction (V_{GxE}) variances and covariance terms if these components are correlated.

$$V_P = V_G + V_E + V_{G\times E} \tag{1.1}$$

Fisher (1918) further partitioned the genetic component of the phenotypic variance into additive (V_A), dominance (V_D) and epistatic (V_I) components, using the least-squares approach.

$$V_G = V_A + V_D + V_I \tag{1.2}$$

To resolve the complexity of formulating epistasis for populations, Cockerham (1954) and Kempthorne (1954) partitioned the epistatic variance (V_I) into eight components to describe the contribution of the epistatic variance to the covariance between relatives. Kempthorne (1957) and Hayman and Mather (1955) adopted a similar epistatic model with eight parameters, which mathematically describes the phenotypes in the nine genotype partitions formed by two loci each with two alleles. These studies formulate a general quantitative genetics model of epistasis within the variance components framework for a two-locus trait.

To describe the general epistatic model in more detail, first consider a single biallelic locus, with alleles A and a, which contributes to a measured quantitative trait of interest, y. The contribution of the locus to the trait may be expressed in the following linear regression, assuming that alleles A and a are co-dominant,

$$y = \mu + \alpha x \tag{1.3}$$

where μ is the trait population mean and x is the number of A alleles present and ranges from 0 to 2. The assumption that the effects of A and a are completely additive (or that the effect of genotype Aa is exactly half-way between the effects of the two homozygotes) may be unrealistic. Another term can be incorporated in the regression to model the genetic effect as the additivity of allelic effects and the dominance deviation,

$$y = \mu + \alpha X_{ai} + \delta X_{di} \tag{1.4}$$

where μ is the trait population mean, α represents the additive main effect at the locus, δ is the dominance main effect, and X_{ai} and X_{di} are scales that express the genotype combinations for genotype i at the locus, for example, $X_{aAA} = -1$, $X_{aAa} = 0$, $X_{aaa} = 1$ and $X_{dAA} = 0$, $X_{dAa} = 1$, $X_{daa} = 0$.

Under the two-locus general epistatic model, if two biallelic genes contribute to the trait y, each with two alleles, locus 1 with alleles A and a, and locus 2 with alleles B and b, the effects of the alleles can be partitioned into main effects and interaction terms. The main effect of an allele is its additive contribution to the trait, calculated without considering contributions from other loci; and the interaction effects between the four alleles at the two loci constitute additive-by-additive, additive-by-dominance, and dominance-by-dominance effects. The general biometric model can be expressed in terms of the components of the genotypic means; alternatively, one can represent this general model as a linear regression model as shown above for the single-locus model,

$$y = \mu + \alpha_1 X_{ai} + \delta_1 X_{di} + \alpha_2 X_{aj} + \delta_2 X_{dj} + \alpha\alpha X_{ai} X_{aj} + \alpha\delta X_{ai} X_{dj} + \delta\alpha X_{di} X_{aj} + \delta\delta X_{di} X_{dj} \tag{1.5}$$

where y is the continuous trait of individuals with genotype i and locus 1 and j and locus 2, μ is the trait population mean, X_{ai} and X_{di}, and X_{aj} and X_{dj} are scales that

express the genotype combination at each locus with X_{ai} and X_{di} for locus 1, and X_{aj} and X_{dj} for locus 2; α_1 and α_2 are the additive main effects at loci 1 and 2, respectively, δ_1 and δ_2 are the dominance main effects at loci 1 and 2, respectively; the four epistatic effects for the two genes are $\alpha\alpha$ - additive-by-additive, $\alpha\delta$ - additive-by-dominance, $\delta\alpha$ - dominance-by-additive, and $\delta\delta$ – dominance-by-dominance. Similar models can be applied to a two-locus dichotomous (qualitative) trait by using the joint genotype penetrances instead of the quantitative trait outcome (see Dupuis et al. 1995; and section 1.2.1).

Hayman and Mather (1955), Van der Veen (1959), and Mather and Jinks (1982) studied the properties of the general epistatic model in more detail and proposed several tests for epistasis in animal crosses. Many studies in model organisms have attempted to map multiple interacting quantitative trait loci (QTL) by incorporating epistasis. Several studies have extended interval mapping to multiple loci either in a series of one-dimensional searches using marker subsets (Jansen 1993; Zeng 1993; Jansen and Stam 1994; Zeng 1994) or genetic background (Charcosset et al. 1994; Rebai et al. 1997; Jannink and Jansen 2001; Boer et al. 2002) as covariates, or in a multiple regression framework with two-dimensional searches (Haley and Knott 1992; Chase et al. 1997; Holland 1998; Wang et al. 1999), or by extending interval mapping directly to multiple QTL (Kao et al. 1999; Zeng et al. 1999), which is perhaps the most interesting method of the approaches listed above because it models multiple loci jointly in multiple interval mapping (MIM). The main difficulty for most methods is that the search across the space of epistatic models and marker subsets remains complex and computationally intensive. Other approaches have also been used to map multiple epistatic QTL, for example, genetic algorithms have been used to detect epistasis (Carlborg et al. 2000; Carlborg and Andersson 2002; Carlborg

et al. 2003) and Bayesian methods have been developed to estimate the number of QTL involved in the trait (Satagopan et al. 1996; Heath 1997; Sillanpaa and Arjas 1998; Stephens and Fisch 1998; Sillanpaa and Arjas 1999) and test for gene-interactions (Sen and Churchill 2001; Yi and Xu 2002).

To date, many methods that map multiple QTL with epistasis in model organisms have been applied and have successfully identified a large number of interactions. Morphological and genetic markers have been used to demonstrate the presence of epistasis in different organisms (Fasoulas and Allard 1962; Fijneman et al. 1996; Long et al. 1996; Li et al. 1997; Shook and Johnson 1999; Leips and Mackay 2000) and biological processes (Araujo and Bier 2000; Beaudoin et al. 2000; Khazanehdari and Borts 2000; Luschnig et al. 2000; Scanga et al. 2000). For example, epistatic interactions between alleles at different loci determine coat colour in mice (Wright 1977).

Epistasis describes departure from additivity, but the term is broadly defined – epistasis can be positive, negative, synergistic (Crow 1970) or reinforcing (Crow and Kimura 1970), antagonistic or diminishing returns (Crow and Kimura 1970), multiplicative (Clark et al. 1981; Hodge 1981; Risch 1990a), or additive (Clark et al. 1981; Puniyani and Feldman 2005). These terms are neither mutually exclusive nor synonymous, and may be used in different disciplines or in the same area of research, stressing the importance of a precise definition of epistasis in each study.

1.2 Epistasis in human genetics

Epistasis appears to be an important and ubiquitous component of complex traits and has generated considerable interest particularly in recent years (Phillips 1998; Templeton 2000; Moore 2003). Examples of epistasis in human complex traits exist for pair-wise interactions between susceptibility loci in type 1 diabetes (Cordell

et al. 1995), type 2 diabetes (Cox et al. 1999), asthma (Xu et al. 2001), and coronary artery disease (Nelson et al. 2001). Specific examples of epistasis between genes contributing to complex traits include interactions between *Apolipoprotein E* (ApoE4) and the *Low-Density Lipoprotein Receptor* (LDLR) genes in coronary artery disease (Pedersen and Berg 1989), and between *Angiotensin Converting Enzyme* (ACE) and *G Protein-coupled Receptor Kinase* (GRK4) in hypertension (Williams et al. 2004b). Analyses that incorporate multi-locus interactions in humans require careful consideration of the particular models, methods and resulting power, and search strategies used. I outline these below.

1.2.1 Models

There are many models of epistasis proposed and discussed in the literature. Two-locus genetic models can be specified in terms of the penetrance structure for the two-locus joint genotypes. If two unlinked loci contribute to the trait with $i = 1, \ldots, n$ possible genotypes at locus 1 and $j = 1, \ldots, m$ genotypes at locus 2, the penetrance of the ij joint genotype can be defined as f_{ij}, and the marginal locus penetrances as f_i for locus 1 and f_j for locus 2. Li and Reich (2000) list all possible such two-locus two-allele fully-penetrant disease models, where f_{ij} can be 0 or 1. In reality f_{ij} can take any value between 0 and 1, and so the space of realistic two-locus models is large, ranging across all possible penetrances and allele frequency values, and number of n and m genotypes at each locus, also expanding to more than two loci.

A subset of these two-locus models can be termed epistatic according to some clearly defined criteria. Hodge (1981) first proposed an epistatic model, later described as the multiplicative genetic model (Risch 1990a). In this model the joint genotype penetrances are specified as a function of the marginal single-locus penetrances. Under the multiplicative model,

$$f_{ij} = f_i \times f_j \tag{1.6}$$

with similar relationships for the population prevalence and sibling recurrence risk-ratio factors (more details provided in chapter II). Risch (1990a) examined the multiplicative genetic model, as well as models without epistasis (additive and heterogeneity models) and their properties in affected relatives. Under the additive model,

$$f_{ij} = f_i + f_j \tag{1.7}$$

and under a heterogeneity model,

$$f_{ij} = f_i + f_j - f_i \times f_j \tag{1.8}$$

For the latter two models the relationship holds for population prevalence factors, but the expression for the multilocus sibling risk ratios is more complex (Risch 1990a). A different definition of epistatic models may restrict such models to two-locus models with no main effects (Culverhouse et al. 2002). In such cases, each of the marginal penetrances for the three genotypes (at each locus), f_i and f_j, will be equal to the population prevalence of the trait, leading to perhaps more 'extreme' models of epistasis.

The epistatic models, described so far, are related to the general epistatic model in section 1.1. The general biometrical model of epistasis can be applied to a dichotomous trait by using the joint genotype penetrances at two loci, for example, the additive and multiplicative models are special cases of the general epistatic model (see Dupuis et al. 1995). The general epistasis model for a two-locus dichotomous trait can be expressed in terms of the additive and dominance variance components at each locus (V_{A1}, V_{D1}, V_{A2} and V_{D2}), and the four interaction terms between them (V_{A1A2}, V_{A1D2}, V_{A2D1} and V_{D1D2}) together with the population prevalence (K) of the disease, using the formulation of Kempthorne (1957) and James (1971). Tiwari and

Elston (1997; 1998), Sham (1998), and Culverhouse (2002) derived variance components for two-locus genetic models in dichotomous traits and examined properties and parameter restrictions for specific two-locus models. Dupuis et al. (1995) and Tang and Siegmund (2002) proposed variations of this general model specifically for qualitative traits. In this context, an epistatic model can be defined as a model for which the sizes of the epistatic variance components are significantly greater than zero.

Several studies have examined properties of multilocus models and identified specific parameter restrictions with respect to disease allele frequencies (Neuman and Rice 1992), joint genotype penetrances (Dupuis et al. 1995; Culverhouse et al. 2002) and heritability (Culverhouse et al. 2002), relative recurrence risk ranges (Neuman and Rice 1992; Craddock et al. 1995; Rybicki and Elston 2000), expected identity-by-descent (IBD) joint two-locus and single-locus sharing probabilities (Cordell et al. 1995; Li and Reich 2000; Tang and Siegmund 2002), and epistatic variance components (Tiwari and Elston 1997; Tiwari and Elston 1998), which are either general or specific to the approach considered in each study. It is important to take these restrictions into account when incorporating epistasis in gene-mapping studies for a mathematically valid and meaningful model of epistasis. Unfortunately, the biological relevance of each mathematical model of epistasis is often unclear. Interpreting epistatic interaction in a biological setting may relate to biochemical pathways, in which mutations at enzymes at different stages of the pathway hinder the production of the final product, or to the formation of protein complexes, in which all interacting partners must be present in the non-mutant state to form the necessary complex. Very few studies have attempted to address this issue theoretically (Li and Reich 2000; Cordell 2002; Elston et al. 2005) or experimentally (Cordell et al. 2001).

Lastly, it should be noted that definitions of epistasis may vary across studies. In some cases epistasis corresponds to a deviation from the additive or heterogeneity model (Risch 1990a; Cordell et al. 1995), while others define epistasis as a deviation from the multiplicative model (Cox et al. 1999). For example, Vieland and Huang (2003) have argued that it is not possible to distinguish between models of heterogeneity and epistasis in affected sib-pair samples. However, this finding appears to depend on how heterogeneity is defined in the model (Cordell 2003; Farrall 2003) because some of the epistatic models that Vieland and Huang (2003) consider fall under the additive definition of Risch (1990a). I use the term "joint-action" to denote any type of multilocus model (not necessarily epistatic), the term "additive" to specify an additive penetrance model (which I also use as an approximation of heterogeneity), and finally the term "epistasis" to define departure from additivity (unless otherwise specified).

1.2.2 Methods

Several approaches incorporate epistasis in gene-mapping linkage and association analyses in humans. Linkage analysis tests for co-segregation of disease and genetic variation at a chromosomal region in relatives. Linkage analysis can either assume an underlying genetic model for the trait locus (or loci), in parametric linkage analysis, where the logarithm-of-odds (Lod) score (Morton 1955) is used to assess evidence for linkage, or be 'model-free', in non-parametric linkage. Lathrop and Ott (1990), Neuman and Rice (1992), Schork et al (1993), and Strauch (2000) present or implement extensions of parametric linkage to two-loci by specifying the allele frequencies at the two loci and the joint genotype penetrances. In complex traits the genetic model at the disease locus is often unknown and in such cases single-locus Lod scores may be maximized over a range of genetic models (the mod score method)

for the purpose of both detecting linkage and determining the appropriate genetic model (Hodge and Elston 1994). Extending this approach to two loci would be computationally challenging because the space of possible two-locus models is large.

Non-parametric linkage analysis was first considered by Penrose (1935) in studying the association between allele sharing and sharing of affection status by pairs of relatives (Sham 1998), usually affected-sib-pairs (ASPs). Non-parametric linkage is typically more suited to complex traits because the mode of inheritance does not have to be specified (Kruglyak and Lander 1995). Two-locus extensions of non-parametric methods in ASPs have been considered and implemented originally in joint two-locus analysis (when the allele-sharing probabilities or effects at two loci are considered simultaneously) by Dizier and Clerget-Darpoux (1986). Several studies have extended the single-locus maximum-Lod score (MLS) method of Risch (1990a; 1990b; 1990c) to two loci in joint two-locus analysis. Cordell et al. (1995) and Farrall (1997) proposed a two-locus MLS test for affected sib-pairs, by specifying the joint two-locus probability of sharing alleles identical by descent (IBD) as a function of the variance components and disease prevalence (and specifying constraints on the joint IBD probabilities) that can be applied to two linked or unlinked disease loci. Olson (1997; 1999) and Dupuis et al. (1995) also extended the MLS to two loci with a different parameterization. Other joint two-locus linkage methods are presented by Tang and Siegmund (2002), Knapp et al. (1994), and Zinn-Justin and Abel (1998).

Several methods evaluate linkage to multiple loci by using the single-locus evidence for linkage. MacLean et al. (1993) used the single-locus parametric Lod scores from two unlinked regions and examine the correlation between them as a preliminary test for interaction. Cox et al. (1999) proposed a conditional linkage method to detect gene-interaction and heterogeneity by taking into account the

correlation in familial non-parametric linkage (NPL) statistics for two unlinked regions, and reassessing the evidence for linkage at each locus in a pair while accounting for evidence for linkage at the reciprocal locus in stratified samples. Ordered-subset analysis (Hauser et al. 2004) may also be used by stratifying the sample to identify subsets of families linked to two loci (Chang et al. 2006). Liang et al (2001) presented a method in affected sib-pairs similar to that of Cox et al. (1999) to assess evidence for linkage to one region by taking into account evidence for linkage at another region (in the entire family sample), which is also applicable to linked regions (Biernacka et al. 2005). Finally, logistic regression either with familial NPL scores (Langefeld et al. 2001) or allele-sharing probabilities (Holmans 2002) from different regions can be used to detect genetic interactions in linkage analysis.

Tests of association between allelic variants and disease have more power to detect genes of modest effect (Risch and Merikangas 1996). There are a number of extensions of association tests that take into account interactions. In family-based association there are several extensions of the transmission-disequilibrium test to two loci (Morris and Whittaker 1999; Koeleman et al. 2000; Lunetta et al. 2000; Culverhouse et al. 2002; Liu et al. 2002). The marker association sequence χ^2 (MASC) method applied to two loci (Dizier et al. 1994) and conditional logistic regression (Cordell et al. 2004) may also be used in family data. In population samples, regression methods (Millstein et al. 2006), data-mining (Nelson et al. 2001; Ritchie et al. 2001; Tahri-Daizadeh et al. 2003; Culverhouse et al. 2004) and Baysian approaches (Kilpikari and Sillanpaa 2003) are possible.

1.2.3 Power

Incorporating epistasis in gene-mapping models when the true genetic model is a multi-locus model should provide an increase in power to detect the susceptibility

loci, but the amount of increase in power provided by multi-locus linkage analysis remains unclear. Schork et al. (1993), Knapp et al. (1994), and Dizier et al. (1996) show that, under certain genetic models, two-locus linkage analysis has more power to detect the magnitude of genetic effects compared to single-locus analysis, whereas others (Durner et al. 1992; Vieland et al. 1992; Goldin and Weeks 1993; MacLean et al. 1993; Durner et al. 1999) find only a marginal increase in power. These differences may be due in part to different study designs: the number and location of markers relative to disease loci, the linkage methods, and the two-locus and single-locus models compared often vary considerably among studies. However, two-locus analysis can estimate more accurately the genetic locations of the susceptibility loci (Schork et al. 1993; Liang et al. 2001). In general there is at least a modest increase in power to detect the magnitude and location of a genetic effect for two-locus compared to single-locus analyses, but findings will vary across different two-locus models (Todorov et al. 1997; Holmans 2002) and methods (Dupuis et al. 1995; Todorov et al. 1997), and there are limits to the increase in power (Tang and Siegmund 2002).

1.2.4 Search strategies

Detecting and estimating epistatic effects in gene-mapping studies necessitates appropriate corrections for multiple testing (Frankel and Schork 1996). The burden of multiple testing impacts power to detect true interactions, and the trade-off between the two needs to be balanced. Two main strategies to detect epistasis have emerged: 1/ the conditional search, where at least one locus is established and is conditioned upon to improve localization of other susceptibility regions, and 2/ the simultaneous search, which simultaneously scans all possible pairs of interacting loci (Lander and Botstein 1986). Both strategies have been formally examined in linkage (Dupuis et al. 1995; Tang and Siegmund 2002) and association (Marchini et al. 2005) analysis in humans.

Several findings are directly applicable to non-parametric linkage analysis in qualitative traits.

A simultaneous search, or multidimensional scan, involves many statistical tests. This can lead to excessively rigorous thresholds to avoid false-positive results and thereby reduce power to detect a true susceptibility locus. To avoid the reduction in power the search for gene interactions can be reduced to certain portions of the genome of detectable main effects (Holmans 2002) or of biological significance (Colilla et al. 2001; Nath et al. 2001; Xu et al. 2001; Zandi et al. 2001), resulting in a conditional search. Dupuis et al. (1995) examined both search strategies and concluded that a simultaneous search can be useful when the appropriate multilocus genetic model is unknown, but it may give rise to spurious results when at least one region in the gene-pair has a major single-locus effect on the trait (Dupuis et al. 1995). The conditional search is more appropriate when at least one of the loci has a major effect and when heterogeneity, rather than epistasis, is present in the trait (Dupuis et al. 1995; Tang and Siegmund 2002). The conditional search reduces the number of tests performed, but fails to consider interactions among loci with no main effects, and so may miss genetic variants with a real effect on the trait. Detection of genes under purely epistatic genetic models with no main effects would benefit from a simultaneous search since all possible combinations of genes are examined. However, Culverhouse et al. (2002) show that such models result in increased allele-sharing, which will increase the single-locus linkage statistics, though that increase may be marginal and fail to reach the detection threshold in a conditional search.

Multidimensional scans of complex traits in model organisms have identified novel loci that act through epistatic pathways. For example, Sen and Churchill (2001) have investigated strategies for simultaneous scans that demonstrate evidence for

interactions among novel loci that have no significant single-locus effects in experimental (murine) line-crosses. Several other studies have performed two-dimensional scans in animals for loci contributing to growth in the chicken (Carlborg et al. 2003; Carlborg et al. 2004), to diabetes (Kim et al. 2001), circadian rhythms (Shimomura et al. 2001), ethanol consumption (Bachmanov et al. 2002), obesity (Brockmann et al. 2000), and maternal performance (Peripato et al. 2002) in mice, to hypertension (Sugiyama et al. 2001) and running capacity (Ways et al. 2002) in rats, to energy metabolism in *Drosophila* (Montooth et al. 2003), and in plants for loci contributing to grain yield components in rice (Li et al. 1997), to regulation of expression levels in maize (Damerval et al. 1994), and to vernalization responses in oat (Holland et al. 1997). The results identified novel regions that contribute to the traits only through an interaction event, and determined the most likely epistatic or additive model for each pair of contributing loci. The success of studies using model organisms suggests that there is intrinsic merit in considering linkage methods that jointly model multiple susceptibility loci and search for multiple interacting genes simultaneously across the genome (Lander and Botstein 1989; Schork et al. 1993; Dupuis et al. 1995; Carlborg and Haley 2004), and encourages the application of analogous approaches to the numerous existing human linkage datasets.

1.3 Thesis summary

The objective of this study is to examine epistasis in the context of linkage studies of qualitative traits in affected sibling pairs in humans. I extend statistical methods to identify and characterize epistasis in human disease and apply the methods to specific datasets. First, I provide a general overview of most multilocus gene-mapping methods published to date in chapter I. In chapter II, I extend a previously published two-locus linkage method to genome-wide applications in different pairs of

affected relatives for a range of two-locus genetic models (Merloc). Chapter III compares Merloc to previously proposed multilocus linkage methods in the presence of epistasis and heterogeneity. In chapter IV, I apply the approach to detect an interaction between two regions in type II diabetes, for which evidence for epistasis was previously detected using conditional linkage analysis. A search through parametric two-locus models identifies a set of models that are most likely to describe the interaction. In chapters V and VI, I apply Merloc to genome-scan data from affected sibling pairs in two-dimensional (2D) linkage scans of three complex traits. I also perform simulations to establish genome-wide significance thresholds for typical 2D linkage scans in ASPs. Finally, in chapter VII, I present the overall conclusions and future directions arising from this study.

CHAPTER II. Merloc: a two-locus non-parametric linkage method for genome-wide applications

2.1 Introduction

The simultaneous search for pair-wise interactions across the genome requires that we consider all linked and unlinked pairs of loci. The multi-locus test statistics presented by Farrall (1997) and Biernacka et al. (2005) consider linked loci in non-parametric two-locus linkage analysis. The method of Farrall was explored because it was available at the time of the study, it was computationally feasible to extend, and it could provide a likelihood estimate of the fit of underlying genetic models.

Risch (1990a; 1990b; 1990c) introduced a likelihood-ratio test statistic for affected sib-pair (ASP) analysis – the maximum-Lod score (MLS) for single-locus non-parametric linkage analysis. Cordell et al. (1995) extended the single-locus MLS to two unlinked susceptibility loci and Farrall (1997) introduced an extension for two linked loci. In this chapter the method of Cordell et al. (1995) and Farrall (1997) is extended to incorporate affected half-sibling pairs, specified nested two-locus genetic models, and implemented to perform automated genome-wide two-locus scans of complex traits.

2.2 Methods

In this section, I first present the method of Cordell et al. (1995) and Farrall (1997) in section 2.2.1. I present extensions of the method with respect to additional relative pairs in section 2.2.2, and different nested genetic models are explored in section 2.2.3. The final likelihood ratio statistic used in this study is in section 2.2.4

along with the distribution of the test statistic under the null, and software implementation.

2.2.1 Two-locus non-parametric linkage test

Following Risch (1990a), denote the population prevalence of a disease (binary) trait by K and the recurrence risk for a type R relative of a proband by K_R. Let X_1 and X_2 denote two random variables representing the disease status of two relatives of type R. X_i takes value of 1 if the individual is affected and 0 if unaffected. Then K can be expressed as $K = E(X_1)$ and $K_R = E(X_2|X_1 = 1)$. Let λ_R denote the recurrence risk ratio, as an index of familial disease clustering. Then James (1971) defines

$$\lambda_R = K_R / K = 1 + \left(1/K^2\right) Cov(X_1, X_2) \tag{2.1}$$

James (1971) has shown that for a single locus, the covariance of X_1 and X_2, $Cov(X_1,X_2)$, can be derived in terms of additive (V_A) and dominance (V_D) variances attributable to the risk locus. For a single locus for full-sib pairs,

$$Cov(X_1, X_2) = \frac{1}{2} V_A + \frac{1}{4} V_D \tag{2.2}$$

For two-loci, following James (1971), Ewens (1979), and Farrall (1997), the covariance for full-siblings is more complicated because recombination between the two disease risk loci needs to be considered if they are in linkage,

$$\begin{aligned} Cov(X_1, X_2) = {} & \frac{V_{A1} + V_{A2}}{2} + \frac{V_{D1} + V_{D2}}{4} + \frac{1}{8}\left[3 - 2\theta_p - 2\theta_m + 2\theta_p^2 + 2\theta_m^2\right] V_{A1A2} + \\ & + \frac{1}{4}[1 - \theta_p - \theta_m + \theta_p^2 + \theta_m^2] V_{A1D2} + \frac{1}{4}[1 - \theta_p - \theta_m + \theta_p^2 + \theta_m^2] V_{D1A2} + \\ & + \left[\frac{1}{2} - \theta_p + \theta_p^2\right]\left[\frac{1}{2} - \theta_m + \theta_m^2\right] V_{D1D2} \end{aligned} \tag{2.3}$$

where V_{A1} and V_{A2} are the additive variance components (VC) for loci 1 and 2, V_{D1} and V_{D2} are the dominance variances for loci 1 and 2, V_{A1A2} is the additive-by-additive, V_{A1D2} is the additive-by-dominance, V_{D1A2} is the dominance-by-additive, and V_{D1D2} is the dominance-by-dominance epistatic variance; θ_p is the paternal and θ_m is the maternal recombination fraction.

The joint IBD probabilities in the saturated model are a function of the ratio of variance components to trait population prevalence, and the recombination fraction. Cordell et al. (1995) have shown that the IBD sharing probabilities for affected full-sib-pairs under linkage (z_{ij}) can be expressed as

$$z_{ij} = \frac{\alpha_{ij}\lambda_{ij}}{\lambda_S} \tag{2.4}$$

where α_{ij} is the probability of sharing i alleles at locus 1 and j alleles at locus 2 under the hypothesis of no linkage; and λ_{ij} is the risk ratio for sibs sharing i and j alleles at the two loci. To obtain the risk ratio for sib-pairs that share i and j alleles IBD we use the expression for λ_S substituting the covariance for sibs with the covariance for sib-pairs sharing exactly i and j alleles IBD (Table 2.1).

Table 2.1. Covariances for ASP sharing i alleles at locus 1 and j alleles at locus 2.

	IBD at Locus 2, j		
IBD at Locus 1, i	*0*	*1*	*2*
0	0	$\frac{1}{2}V_{A2}$	$V_{A2}+V_{D2}$
1	$\frac{1}{2}V_{A1}$	$\frac{1}{2}V_{A1}+\frac{1}{2}V_{A2}+\frac{1}{4}V_{A1A2}$	$V_{A2}+\frac{1}{2}V_{A1}+V_{D2}+\frac{1}{2}V_{A1A2}+\frac{1}{2}V_{A1D1}$
2	$V_{A1}+V_{D1}$	$V_{A1}+\frac{1}{2}V_{A2}+V_{D1}+$ $+\frac{1}{2}V_{A1A2}+\frac{1}{2}V_{D1A2}$	$V_{A1}+V_{A2}+V_{D1}+V_{D2}+$ $+V_{A1A2}+V_{A1D2}+V_{D1A2}+V_{D1D2}$

If w_{ij} is the estimated probability of observing the marker data for a k^{th} sib-pair, given that the pair shares i alleles IBD at locus 1 and j alleles IBD at locus 2, the likelihood ratio statistic (maximum Lodscore or MLS) for N full sib-pairs is

$$MLS = \log_{10}\left[\frac{\prod_{k=1}^{N}\left(\sum_{i=0}^{2}\sum_{j=0}^{2} z_{ij} w_{ij}\right)}{\prod_{k=1}^{N}\left(\sum_{i=0}^{2}\sum_{j=0}^{2} \alpha_{ij}\right)}\right] \quad (2.5)$$

The two-locus approach presented in this section may be generalized to different types of affected relative pairs (section 2.2.2) and extended to fit different genetic models by imposing restrictions on the variance components (section 2.2.3).

2.2.2 Extension to different types of affected relative pairs

The method presented above is applicable to ASPs for pairs of unlinked and linked loci, as described by Cordell et al. (1995) and Farrall (1997). It can be extended to different types of affected relative pairs by deriving the covariances for two or more loci in different pairs of relatives, as described by Cordell et al. (2000).

For pairs of unlinked loci the extension to affected relatives follows from the general form of the covariance. Briefly, the general expression for the covariance between two relatives (Kempthorne 1957) for multiple unlinked loci is

$$Cov = rV_A + pV_D + r^2V_{AA} + rpV_{AD} + p^2V_{DD} + r^3V_{AAA} + r^2pV_{AAD} + rp^2V_{ADD} + p^3V_{DDD}\cdots \quad (2.6)$$

where r is the coefficient of relationship, which is twice the kinship coefficient (or the coefficient of coancestry), and p is the probability that the relative pair share two alleles IBD. The values of r and p for several types of relative pairs are presented in Table 2.2. Using the general expression for the covariance and the values in Table 2.2 one can derive the two-locus covariances for unlinked loci in different relative pairs.

Table 2.2 Coefficients of relatedness and probabilities of sharing two alleles IBD.

Relative pair	**Coefficient of relatedness (*r*)**	**Probability of IBD 2 (*p*)**
MZ twins	1	1
Parent-offspring	½	0
Full siblings	½	¼
Half-siblings	¼	0
Grandparent-grandchild	¼	0
First-cousins	⅛	0

For linked loci the covariance depends on the recombination fraction and the above approach needs to be adjusted, following James (1971) and Ewens (1979). Ewens (1979) derives the covariance of siblings for linked pairs of loci in terms of the recombination fractions, which can then be used in the two-locus test between linked loci for ASPs (Farrall 1997). This approach is outlined below to derive the covariance for pairs of half-siblings and grandparent-grandchild pairs.

If we consider a pair of half-siblings that are related either through the father (paternal pair) or the mother (maternal pair), the half-sibs may share at most one allele identical-by-descent (IBD), either the paternal allele (denoted as allele *p*) or the maternal allele (allele *m*). Let us define E^i_{pp} as the event that two paternal half sibs have inherited the same allele (allele *p*) at locus *i* from their father. Similarly, for two maternal half-sibs E^i_{mm} denotes the event that the half-sib pair have inherited the same allele (allele *m*) from the mother. Under the assumption of no inbreeding events E^i_{pm} and E^i_{mp} are impossible for any half-sib pair, because alleles *p* and *m* cannot be shared IBD if they are inherited from different parents. Then, for paternal half-sibs E^i_{mm} is also impossible, while for maternal half-sibs the event E^i_{pp} is impossible. Therefore, for two paternal half-sibs

$$P(E^1_{pp}) = P(E^2_{pp}) = \frac{1}{2}$$
$$P(E^1_{pm}) = P(E^2_{pm}) = P(E^1_{mp}) = P(E^2_{mp}) = 0$$
$$P(E^1_{mm}) = P(E^2_{mm}) = 0$$

To derive the half-sibling covariance for two loci, follow the general expression for the covariance obtained by Ewens (1979) using full-siblings,

$$Cov = \frac{1}{2}P\left(E^i_{pp} \cup E^i_{mm}\right)\sum_{i=1}^{2} V_{Ai} + P\left(E^i_{pp} \cap E^i_{mm}\right)\sum_{i=1}^{2} V_{Di} +$$

$$+\frac{1}{4}P\left[\left(E^1_{pp} \cup E^1_{mm}\right)\cap\left(E^2_{pp} \cup E^2_{mm}\right)\right]V_{A1A2} + \frac{1}{2}P\left[\left(E^1_{pp} \cup E^1_{mm}\right)\cap E^2_{pp} \cap E^2_{mm}\right]V_{A1D2} +$$

$$+\frac{1}{2}P\left[\left(E^2_{pp} \cup E^2_{mm}\right)\cap E^1_{pp} \cap E^1_{mm}\right]V_{D1A2} + P\left(E^1_{pp} \cap E^1_{mm} \cap E^2_{pp} \cap E^2_{mm}\right)V_{D1D2} \qquad (2.7)$$

The only term that needs to be computed is

$P\left(E^1_{pp} \cap E^2_{pp}\right)$, which (following expected IBD in ASPs) simplifies to $\frac{1}{2} - \theta_p + \theta_p^2$

where θ_p is the paternal (or male) recombination fraction between the two loci. Therefore, the covariance in paternal half-sibs simplifies to the following:

$$Cov_{HS} = \frac{1}{4}\sum_{i=1}^{2} V_{Ai} + \frac{1}{4}\left(\frac{1}{2} - \theta_p + \theta_p^2\right)V_{A1A2} \qquad (2.8)$$

Similarly, for maternal half-sibs it the covariance is

$$Cov_{HS} = \frac{1}{4}\sum_{i=1}^{2} V_{Ai} + \frac{1}{4}\left(\frac{1}{2} - \theta_m + \theta_m^2\right)V_{A1A2} \qquad (2.9)$$

where θ_m is the maternal (or female) recombination fraction between the two loci.

Likewise, for a grandparent-grandchild pair related via the paternal line, one can set E^i_{pp} as the event that both grandparent and grandchild have inherited the same paternal allele at locus *i* from the son/father in common. Events E^i_{pm}, E^i_{mp}, and E^i_{mm} are then impossible for paternal-grandparent-grandchild pairs. Following the approach above, the only term that needs to be computed is again

$P\left(E^1_{pp} \cap E^2_{pp}\right)$, which simplifies to $\frac{1-\theta_p}{2}$

where θ_p is the paternal (or male) recombination fraction between the two loci.

For paternal-grandparent-grandchild pairs the covariance reduces to

$$Cov_{GG} = \frac{1}{4}\sum_{i=1}^{2} V_{Ai} + \frac{1}{8}(1-\theta_p)V_{A1A2} \tag{2.10}$$

and for maternal-grandparent-grandchild it is

$$Cov_{GG} = \frac{1}{4}\sum_{i=1}^{2} V_{Ai} + \frac{1}{8}(1-\theta_m)V_{A1A2} \tag{2.11}$$

It should be noted that Cordell et al. (2000) have also derived the two-locus (and three-locus) covariances in different pairs of relatives. Their method is similar to the approach described above and is applicable to both linked and unlinked regions. The resulting two-locus covariances in half-sibs (equations 2.8 and 2.9) and grandparent-grandchild pairs (equations 2.10 and 2.11) are the same using either the method described above, or the approach used by Cordell et al. (2000).

To incorporate different pairs of affected relatives in the test statistic, one needs to obtain the allele-sharing probabilities under linkage for these pairs (z_{ij}). Because many of the genome-wide screens in ASPs reveal that some proportion of individuals are half-siblings, the extension to half-siblings was included in the two-locus MLS. To achieve this, first the covariance for half-siblings was substituted into the expression for λ_R to obtain the risk ratio for half-sibs, λ_{HS}. The risk ratios for half-siblings that share exactly i and j alleles IBD at the two hypothetical disease loci (λ_{ij}) are equivalent to those for full sibs who share i =0,1 and j =0,1 alleles (Table 2.1). The IBD sharing probabilities for affected half-sibs (z_{ij}) can be expressed as

$$z_{ij} = \frac{\alpha_{ij}\lambda_{ij}}{\lambda_{HS}} \tag{2.12}$$

To obtain these probabilities for other types of relative pairs, one can follow the procedure for half-sibs and the expressions presented in Table 2.1.

2.2.3 Two-locus genetic models

The most general two-locus model includes 8 variance components parameters that can fit the full range of epistasis models for affected sib-pairs. Specific (nested) genetic models can be fitted to the data by restricting the number of free variance components parameters in the model. For instance, in the case of single locus models, additive and dominance main effects attributed to the locus of interest (e.g. V_{A1} and V_{D1} for locus 1) would be fitted and all other variance components parameters would be fixed at zero.

Two specific two-locus models that model additive and multiplicative penetrance structures (Risch 1990a) have been shown to be nested within the general variance components framework (Cordell et al. 1995; Cordell 2003). The additive model includes locus-specific additive and dominance effects only (i.e. V_{A1}, V_{D1}, V_{A2} and V_{D2}) so that epistasis is ignored in this model; this model can be thought of as a "main-effects-only" model. The multiplicative model is a mathematically simplified model of a fixed degree of epistasis (Hodge 1981; Risch 1990a). If two unlinked loci contribute to the trait multiplicatively, the overall sibling risk ratio, λ_S, is the product of the two risk ratio factors defined in terms of the penetrances for the two contributing loci, where λ_{S1} is the risk ratio for siblings for locus 1 and λ_{S2}, is the risk ratio for locus 2:

$$\lambda_S = \lambda_{S1} \times \lambda_{S2} \tag{2.13}$$

Cordell (2003) has shown how the multiplicative model can be expounded in terms of variance components, so that the four epistatic variance components are expressed as a function of the single locus variance components:

$$
\begin{aligned}
V_{A1A2} &= \left(\frac{1}{K^2}\right)\left(V_{A1}V_{A2}\right) \\
V_{A1D2} &= \left(\frac{1}{K^2}\right)\left(V_{A1}V_{D2}\right) \\
V_{D1A2} &= \left(\frac{1}{K^2}\right)\left(V_{D1}V_{A2}\right) \\
V_{D1D2} &= \left(\frac{1}{K^2}\right)\left(V_{A1}V_{A2}\right)
\end{aligned}
\qquad (2.14)
$$

This formulation suggests a modification that can model a wide range of levels of epistasis, by adding a single parameter (epsilon) as follows:

$$
\begin{aligned}
V_{A1A2} &= \varepsilon\left(\frac{1}{K^2}\right)\left(V_{A1}V_{A2}\right) \\
V_{A1D2} &= \varepsilon\left(\frac{1}{K^2}\right)\left(V_{A1}V_{D2}\right) \\
V_{D1A2} &= \varepsilon\left(\frac{1}{K^2}\right)\left(V_{D1}V_{A2}\right) \\
V_{D1D2} &= \varepsilon\left(\frac{1}{K^2}\right)\left(V_{D1}V_{D2}\right)
\end{aligned}
\qquad (2.15)
$$

Different multilocus models can be incorporated by varying the value of epsilon, e.g. the additive model can be specified by fixing epsilon to 0, while under the multiplicative model epsilon is fixed to 1. Greater degrees of epistasis are modelled with increasing values of epsilon in this epsilon-epistatic genetic model.

To show the impact of different degrees of epistasis on the relative recurrence risk ratios, the recurrence risk ratio in first (λ_1) and second-degree relatives (λ_2) was calculated for a range of two-locus models specified by epsilon (Figure 2.1). To express λ_2 as a function of λ_1 and ε, the underlying genetic model was assumed to be a two-locus model with two genes of equal effects and no dominance (that is, $V_{A1}=V_{A2}\neq 0$, $V_{A1A2}\neq 0$, and $V_{D1}=V_{D2}=V_{A1D2}=V_{D1A2}=V_{D1D2}=0$). To solve for λ_2 in terms of λ_1 and ε, the formulas for the the covariances in full-sibs (equation 2.3) and half-sibs (equations 2.8 and 2.9), and the formula for V_{A1A2} (equation 2.15 above) were

input into the expression for the relative recurrence risk ratio (equation 2.1) with the assistance of Mathematica® to obtain the following relationship between λ_2 (λ_{HS}) and λ_1 (λ_S) with respect to ε,

$$\lambda_2 = \frac{-2 + 3\varepsilon + \varepsilon\lambda_1 + 2\sqrt{1 - \varepsilon + \varepsilon\lambda_1}}{4\varepsilon} \tag{2.16}$$

The resulting relationships fit nearly straight lines and followed the expected trend for the additive and multiplicative models (Risch 1990a). Equation 2.16 was obtained without fixing the variance components or the population prevalence (K) to specific values; that is, the relationship between λ_1 and λ_2 in terms of ε, holds irrespective of the values of V_{A1}, V_{A2}, and K. It appears that for values of epsilon greater than 10^3 the relationship between the relative recurrence risk ratios is very similar to that for $\varepsilon = 10^3$, therefore the range of possible epsilon values was set from 0 to an upper bound of 10^3. The biological interpretation of genetic models for which $\varepsilon > 10^3$, would be that these are models, in which the main effect variance components are very close to 0, given a specific population prevalence for the trait.

The epsilon-epistatic model proposed in equation 2.15 is a simplistic representation of epistasis in the variance components framework. Many genetic models would not be included in the model space shown in Figure 2.1. To cover the space of possible epistatic models in greater depth, one may include different (and more complex) functions of epsilon fitting the different components of the epistatic variance. However, in this framework it is possible to assess how well the epsilon-epistatic model covers the space of genetic models by comparing the fit of the general model to the fit of the epsilon-epistatic model. To relate the general model to the epsilon-epistatic model, the epistatic variances for general model can be thought of as

$$V_{A1A2} = \varepsilon_1\left(\frac{1}{K^2}\right)\left(V_{A1}V_{A2}\right)$$

$$V_{A1D2} = \varepsilon_2\left(\frac{1}{K^2}\right)\left(V_{A1}V_{D2}\right)$$

$$V_{D1A2} = \varepsilon_3\left(\frac{1}{K^2}\right)\left(V_{D1}V_{A2}\right)$$

$$V_{D1D2} = \varepsilon_4\left(\frac{1}{K^2}\right)\left(V_{D1}V_{D2}\right)$$

where there would be three additional degrees of freedom in the general model two-locus test, compared to the epsilon-epistatic model. Chapter III contains a more detailed study of epsilon under models of epistasis and heterogeneity.

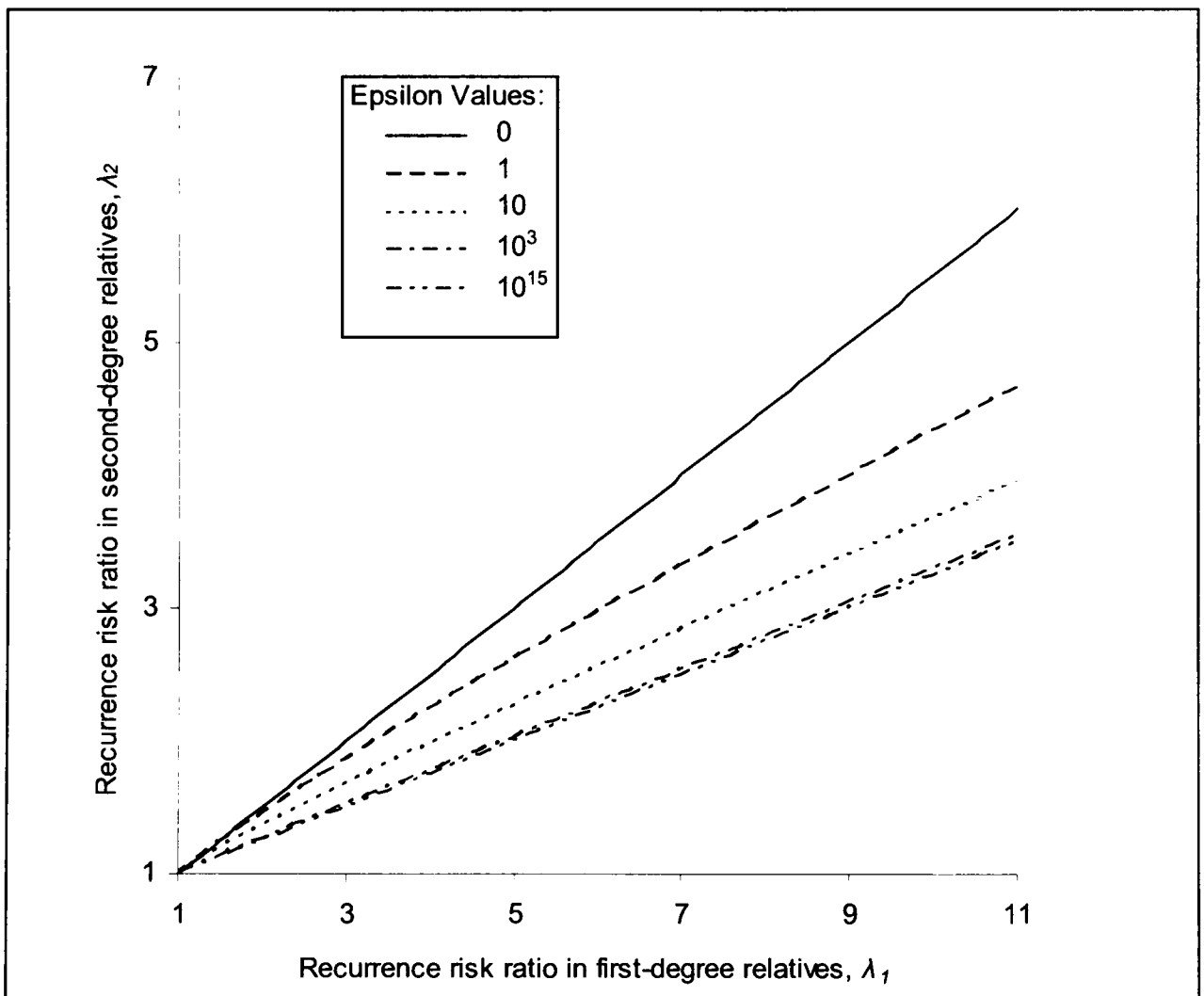


Figure 2.1. The range of plausible epsilon values. The epistatic VC were expressed as a function of the main-effects VC and these were substituted into the covariances for full- and half-sibs to obtain the ratio between λ_{HS} and λ_S in Mathematica. The relationship between λ_2 and λ_1 was obtained by $\lambda_2 = \dfrac{-2+3\varepsilon+\varepsilon\lambda_1+2\sqrt{1-\varepsilon+\varepsilon\lambda_1}}{4\varepsilon}$.

2.2.4 Likelihood ratio statistic: null distribution and software implementation

Test Statistic

The two-locus test statistic presented in this thesis incorporates affected full sibs (ASPs) and half-sibs (HSPs). To calculate the two-locus MLS in affected relative pairs follow the same procedure as for ASPs in equation 2.5. The joint IBD sharing probabilities under linkage (z_{ij}) and under no linkage (α_{ij}) for ASPs and HSPs need to be specified, and w_{ij}, the probability of observing the marker data for relative pair k, given that the pair shares i alleles IBD at locus 1 and j alleles IBD at locus 2, needs to be calculated (see *Software Implemention* below for calculation). The likelihood ratio statistic (MLS) for N affected relative pairs is defined as that in ASPs (equation 2.5),

$$MLS = \log_{10}\left[\frac{\prod_{k=1}^{N}\left(\sum_{i=0}^{2}\sum_{j=0}^{2} z_{ij} w_{ij}\right)}{\prod_{k=1}^{N}\left(\sum_{i=0}^{2}\sum_{j=0}^{2} \alpha_{ij}\right)}\right]$$

Null Distribution of test statistic

Cordell et al. (1995) and Bengtsson (2001) derived a set of constraints for the two-locus allele sharing probabilities in ASPs similar to the triangle constraint of Holmans (1993). It may be possible to calculate the distribution of the two-locus MLS through use of asymptotic theory (Cox and Hinkley 1974), but that would be complicated because of the nonstandard genetic restrictions (Self and Liang 1987). Therefore, significance thresholds were obtained through simulation both in the case of a single pairwise interaction and in the context of a two-dimensional genome scan (results presented in chapter V).

For a single pair-wise interaction, simulations were performed to obtain the distribution of the MLS under the null hypothesis with no genetic effects at either

locus. For pairs of unlinked loci, completely informative markers were simulated in 100 ASPs using 100 000 replicates, by drawing ASPs at random from each possible two-locus IBD sharing configuration (using the G05CCF NAG subroutine for pseudo-random number generation in Fortran, www.nag.co.uk). Single locus MLS and two-locus MLS were obtained under different genetic models. To re-iterate, in the additive (ADD) model the main-effect VC are allowed to vary and the epistatic variance components are set to 0; in the multiplicative (MUL) model the main effect VC are allowed to vary and the epistatic VC are a function of the main-effects VC (equations 2.14); in the epsilon-epistatic (EPS) model the main-effect VC and ε are allowed to vary and the epistatic VC are a function of the main-effect VC and ε (equations 2.15); and in the general (GEN) two-locus model all main effect and epistatic VC are allowed to vary. The resulting thresholds (Table 2.3) for unlinked loci are similar to those obtained by Holmans (1993) for single-locus MLS, and to those obtained by Cordell et al. (1995) for the additive (ADD), multiplicative (MUL), and general (GEN) two-locus models and differences between the models. The thresholds are slightly different to those previously published, reflecting the different amount of information in the samples. The additive and multiplicative models have the same number of free parameters and similar thresholds under the null. The epsilon-epistatic model appears to approximate the general model well because the distribution of the test statistic under the null is similar to that for the general model (Figure 2.2).

Table 2.3. Pointwise significance thresholds for unlinked pairs of loci.

	Test statistic threshold		
Model [a]	**$\alpha = 0.05$**	**$\alpha = 0.01$**	**$\alpha = 0.001$**
GEN	1.306	2.086	3.165
EPS	1.263	2.019	3.096
MUL	1.144	1.847	2.988
ADD	1.148	1.861	2.989
SL	0.723	1.366	2.35
GEN-EPS	0.177	0.424	0.87
GEN-MUL	0.326	0.665	1.232
GEN-ADD	0.344	0.705	1.312
GEN-SL	0.901	1.583	2.584
EPS-MUL	0.189	0.469	0.952
EPS-ADD	0.199	0.536	1.066
EPS-SL	0.863	1.529	2.501

[a] Abbreviations are as follows: GEN - two-locus general, EPS – two-locus epistatic, MUL – two-locus multiplicative, ADD – two-locus additive, and SL – single-locus model.

Simulations were also performed for pairs of linked loci because the distribution of the test statistic under the null depends on θ (Farrall 1997) and the results are presented in chapter V (Figure 5.1A). Finally, the distribution of the test statistic may be examined under different null hypotheses (for example under single-locus models) and further results are presented in chapters III - VI.

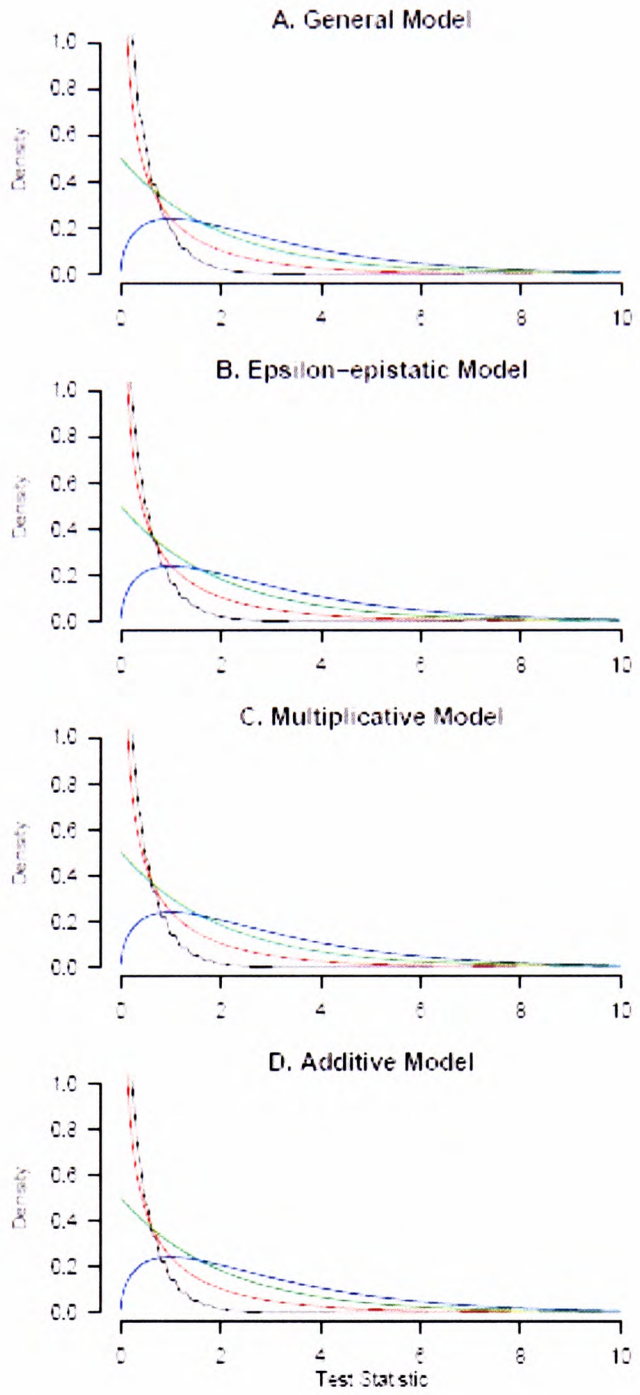


Figure 2.2 Distribution of the two-locus MLS under the null. The fitted density (black line) of the simulated single-point two-locus test statistics obtained under four genetic models - (A) general, (B) epsilon-epistatic, (C) multiplicative, and (D) additive two-locus models. Bandwidth estimates were set to 0.05 in all panels to smooth the density distribution. Superimposed are the density functions from the chi-square distribution with 1 (red line), 2 (green line), and 3 (blue line) degrees of freedom.

Software implementation

The two-locus non-parametric linkage test was implemented in a new computer package, Merloc, which was written in Perl. First, Merloc uses Merlin (Abecasis et al. 2002) as a multipoint likelihood calculating engine to help calculate

the probability (w_{ij}) that a sib-pair shares i and j alleles IBD at the two loci, given the marker data, M, that is $w_{ij} = P(IBD = ij \mid M)$. To calculate w_{ij} in full ASPs, genotypes were set at two indicator markers to reflect the 16 possible parent-specific alternative joint IBD configurations for each sib-pair at a pair of genetic locations, or coordinate (see Figure 2.3A). Each of the two indicator loci was positioned 0.00001 cM away from the actual genetic positions (in the marker map). Merlin (Abecasis et al. 2002) was used to calculate the likelihood for each of the 16 full sib-pair families (for each original ASP in the data) taking into account the real genotype data for markers in the chromosomal locations. The w_{ij} were calculated as the likelihood of the ij^{th} sib-pair dummy family, $L(X_{ij})$, over the sum of the likelihoods for the 16 full-sib families, $\sum_{i,j} L(X_{ij})$, by applying Bayes' theorem following Farrall (1997),

$$w_{ij} = P(IBD = ij \mid M) = \frac{P[(IBD = ij) \cap (M)]}{\sum_{i,j} P[(IBD = ij) \cap (M)]} = \frac{L(X_{ij})}{\sum_{i,j} L(X_{ij})} \qquad (2.17)$$

This approach was used to calculate the w_{ij}'s at each marker coordinate for all affected full sib-pairs. Similarly, for half-sib-pairs 4 half-sib dummy families were constructed to reflect the IBD sharing configurations in each maternal and paternal pair of half-sibs (Figure 2.3B,C), and the w_{ij}'s were obtained using the approach described for full sibs.

The IBD probabilities under no linkage, α_{ij} for pairs of linked loci were calculated in a similar manner using Merlin. In this case, likelihoods were obtained for 16 (for full sibs) or 4 (for half sibs) families, constructed with genotype data at two indicator loci as described in Figure 2.3, but without considering marker genotypes. The two indicator loci were positioned at the required recombination fraction apart.

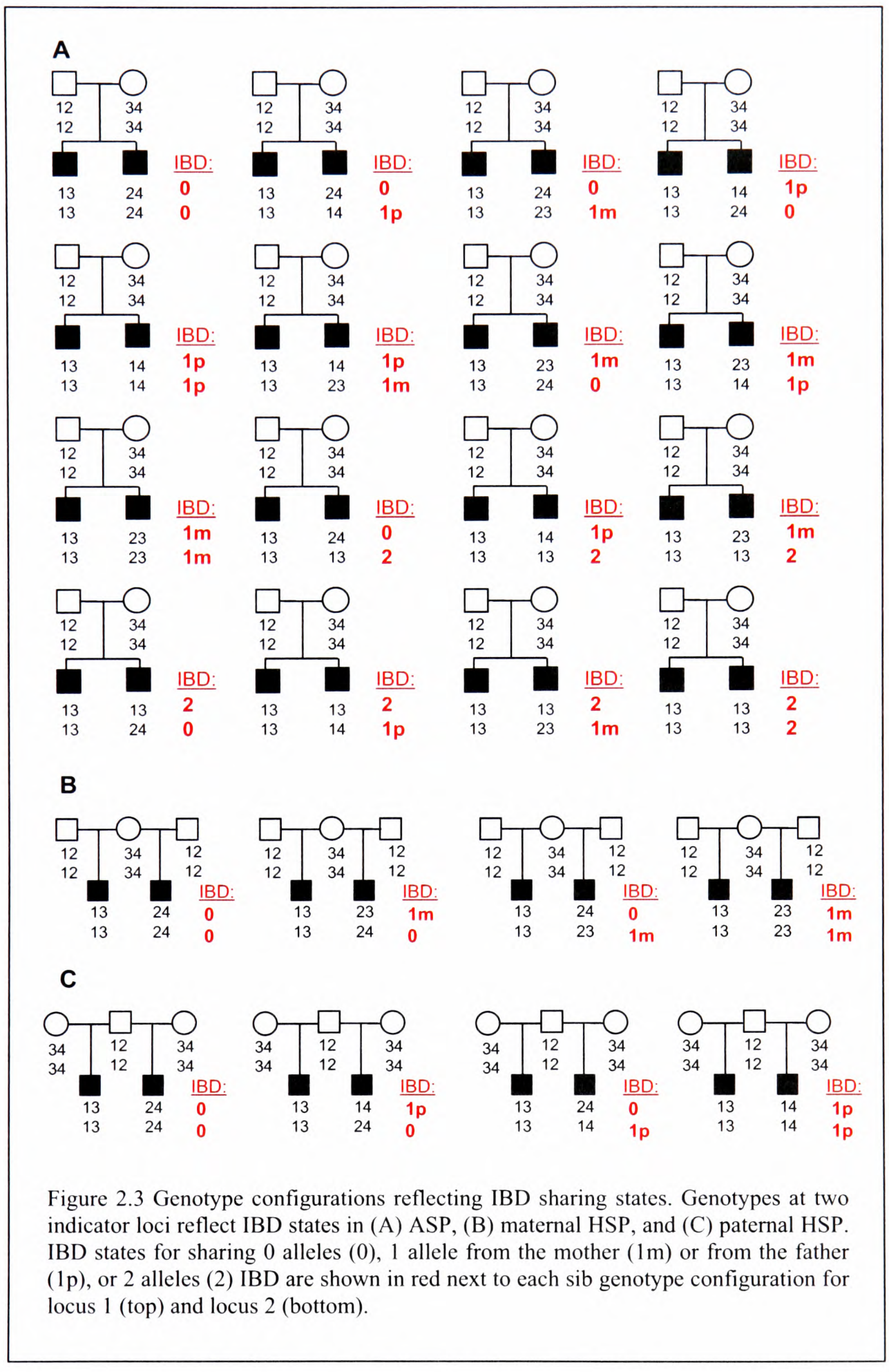


Figure 2.3 Genotype configurations reflecting IBD sharing states. Genotypes at two indicator loci reflect IBD states in (A) ASP, (B) maternal HSP, and (C) paternal HSP. IBD states for sharing 0 alleles (0), 1 allele from the mother (1m) or from the father (1p), or 2 alleles (2) IBD are shown in red next to each sib genotype configuration for locus 1 (top) and locus 2 (bottom).

The obtained probabilities (w_{ij} and α_{ij}) are then internally passed to *twhslog,* a Fortran program that calculates the two-locus MLS. *Twhslog* is based on Twoloc (Farrall 1997) and uses some of the Twoloc subroutines for maximization in the ASPs. The MLS statistic is maximized with respect to the variance components (VC) using numerical methods (NAG maximization libraries). The maximization is performed using NAG Fortran subroutine E04JAF (recently updated to E04JYF), which is a quasi-Newtonian algorithm. The algorithm seeks to minimize a function (the negative of the MLS) subject to constraints on the independent variables (the VC and ε if applicable), using computed function values. The initial starting values (for the VC and ε) were set at 0 and the algorithm moves on the basis of the gradient of change in the MLS to converge to a minimum, using weak and strong convergence criteria. An estimate of whether the final convergence point is a local or a global minimum is obtained.

Convergence was an important technical consideration in this study and to ensure that the maximization procedure worked three main points were examined: 1/ changing the maximization boundaries of the parameters, 2/ re-scaling several of the parameters in the maximization procedure, and 3/ adopting a stepwise manner of estimating initial values for the maximization. The final boundaries of the parameters which gave the most satisfactory results were restrictions on the variance-components to be non-negative and a restriction on epsilon to fall between 0 and 10^3 (from Figure 2.1). In terms of re-scaling the parameters, the final algorithm maximizes the MLS over the logarithms of the VC, and epsilon if required. Finally, the maximization begins by fitting the simple single-locus models to the data and then uses the maximum likelihood estimates of the parameters from these models as initial values for the maximization in the more complex two-locus models. Genetic models are

fitted sequentially in *twhslog* starting with single-locus models at both locations, and then additive, multiplicative, epsilon-epistatic, and general two-locus models.

The approach in Merloc was implemented to automatically scan a two-dimensional (2D) grid of coordinates, for example in a 2D linkage scan of the genome. In a 2D genome-scan initially only the MLS under the general two-locus model is considered. One can then assess 2D coordinates by comparing them to genome-wide significance thresholds (see chapter V), and evaluate the fit of nested two-locus models (additive, multiplicative and epsilon-epistatic) at specific 2D coordinates of interest.

2.3 Application to type 1 diabetes

Two approaches were used to assess the performance of Merloc in the presence of two-locus effects. First, I applied the method to a previously detected interaction between two linked susceptibility loci on chromosome 6 in Type 1 diabetes as 'proof of principle' (section 2.3.1). Second, I used simulations to evaluate Merloc in the presence of two linked and unlinked interacting loci. I first considered two linked susceptibility loci (section 2.3.2) using the data from Type I diabetes. I then performed simulations to assess Merloc when the two disease loci were unlinked to each other under a range of two-locus genetic models and in comparison to other multilocus linkage methods (in chapter III).

2.3.1 Type I diabetes results

Type 1 diabetes (T1D) or insulin dependent diabetes mellitus (IDDM) is a complex genetic trait with an estimated recurrence risk ratio to sibs (λ_S) of 15. To date, at least eighteen contributing loci (IDDM1 – IDDM18) have been mapped to the trait. The major locus appears to be IDDM1, located in the HLA region, with λ_{S1} of

3.42 (Risch et al. 1993). Previous studies have identified IDDM15 (Delepine et al. 1997) on chromosome 6q as a susceptibility gene that is linked to the major locus in diabetes, IDDM1. However, analyzing evidence for the effect of IDDM15 on T1D presents a challenge due to its proximity (close to 40cM) to IDDM1 and the resulting interdependence between the two loci. Delepine et al. (1997) achieved this by developing an IBD method that tests for the presence of a second susceptibility locus linked to an established disease locus. Their IBD approach examined the single-locus marginal IBD sharing and assessed the distortion in IBD sharing at the second susceptibility locus compared to that expected if it did not contribute to the trait.

A genome scan of T1D (Mein et al. 1998) presented no evidence of linkage to IDDM15. Two subsequent studies (Cordell et al. 2000; Biernacka et al. 2005) have since re-analysed these data and have obtained evidence for the IDDM15 as a contributing locus. Here, the data were re-analysed using Merloc in an attempt to obtain the MLS attributed to IDDM15 independent of IDDM1.

Genotype data from 30 microsatellite markers on chromosome 6 were obtained in 356 ASPs (356 families) presented in a genome scan by Mein et al. (1998), from http://www-gene.cimr.cam.ac.uk/todd/public_data/genome_scan_v2.0/. The marker map used in this chapter was obtained from the original publication using sex-averaged cM distances, but analyses were also performed using the Marshfield map (Broman et al. 1999), which gave similar results.

The entire chromosome 6 was analyzed by fixing the first hypothetical disease locus on IDDM1 at 27.3 cM and moving the second locus across chromosome 6 at every marker location and at 3 locations between each marker pair. MLS scores were computed under three genetic models: the general two-locus model (GEN), the single-locus model attributable to IDDM1 (SL1), and the single-locus model attributable to

locus 2 (SL2), which is identical to the single-locus MLS at each cM position along the chromosome.

The results from Merloc are presented in Figure 2.4. The MLS scores obtained under SL2 (linkage due to disease locus 2) are similar to those presented in Mein et al. (1998) as the single locus MLS. The single locus MLS maximizes at 27.3 cM (the previously identified location of IDDM1) to a single-locus MLS of 34.6. The MLS scores obtained under SL1 (linkage due to IDDM1 alone) are at 34.6 across the entire chromosome 6. The MLS scores obtained under the general two-locus model (GEN) peak at 57 cM to a maximum of 36.7. Following Farrall (1997), support for the contribution of IDDM15 in T1D, independent of IDDM1, is calculated as the difference between the MLS under GEN and MLS under SL1, which results in 2.1 at 57 cM. Therefore, the MLS attributed to the locus at location 57 cM (1DDM15) is 2.1. The fit of different two-locus models compared to the general model can be obtained as described in section 2.2.2. The MLS obtained under an additive two-locus model (36.68) best approximated the MLS obtained under the general two-locus case. The maximum-likelihood estimate of epsilon was 0, with a 1-Lod unit support interval of (0-0.9).

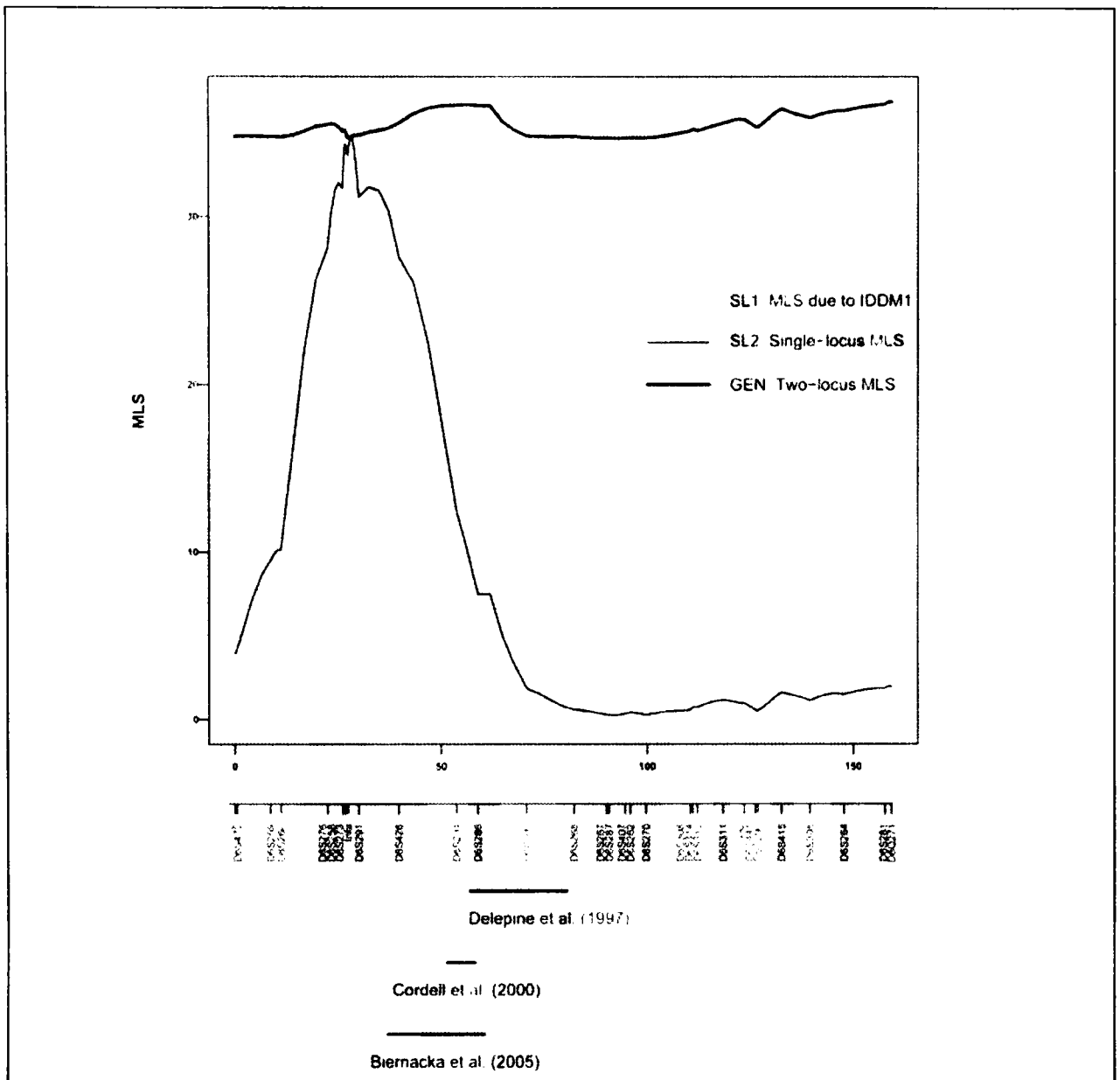


Figure 2.4. T1D chromosome 6 results in the data of Mein et al. (1998). Results from the single-locus and two-locus analyses, the marker map used in the study, and the approximate location of intervals previously denoted as IDDM15 and the corresponding references.

2.3.2 Simulations

Simulations were performed to obtain the expected single locus MLS results under a single-locus and a two-locus model. Under the single locus model, the assumption was that one locus contributed to the trait, at IDDM1, and under the two-locus model two loci, the first at IDDM1 and the second at IDDM15 contributed to the trait. To perform the simulations for the single locus case, the maximum-likelihood estimates of the two-locus allele sharing probabilities were obtained at 57 cM calculated under SL1. For the two-locus case the maximum likelihood estimates of the two-locus allele sharing probabilities at 57 cM were calculated under GEN (Table 2.4). For each model (single-locus and two-locus) 356 ASPs were generated with genotype data at two indicator markers arranged to reflect the IBD configurations (as in Figure 2.3.A). The proportion of ASPs out of the 356 that fell into each joint IBD category were obtained from Table 2.4 for the two genetic models. I then positioned the two indicator markers at locations 27.3 cM and 57 cM on the chromosome 6 marker map and added the first 29 markers along the map. Next, genotype data for the 29 markers was simulated in Simwalk2 (Sobel and Lange 1996), conditional on the observed two-locus IBD probabilities in Table 2.4 (displayed by the indicator marker genotypes). Twenty-nine markers were selected because Simwalk2 could only handle 31 markers at a time. The missing data patterns, allele frequencies, and marker map were identical to those in the data of Mein et al. (1998).

Table 2.4. Maximum-likelihood estimates of IBD probabilities in type 1 diabetes.

		IBD sharing probabilities at locus 2			
		0	**1 (maternal)**	**1 (paternal)**	**2**
IBD estimates at 57 cM for the single-locus model (IDDM1)					
IBD locus 1	0	0.0412	0.0212	0.0212	0.0108
	1 (maternal)	0.0372	0.0725	0.0191	0.0372
	1 (paternal)	0.0372	0.0191	0.0725	0.0372
	2	0.0659	0.1286	0.1286	0.2508
IBD estimates at 57 cM for the two-locus model (IDDM1 and locus 2 at 57					
IBD locus 1	0	0.0207	0.0243	0.0243	0.0195
	1 (maternal)	0.024	0.0735	0.0193	0.0514
	1 (paternal)	0.024	0.0193	0.0735	0.0514
	2	0.0567	0.1243	0.1243	0.2693

Simulations under the two-locus disease model were performed 30 times (Figure 2.5A) and simulations under a single-locus disease model were performed 5 times (Figure 2.5B). The number of simulates was constrained because it would have been computationally unfeasible to perform additional simulations. The reason for simulating genotypes conditional on the observed single-locus and two-locus IBD (rather than perform unconditional simulations) was to simulate the observed genetic effect in the region by using the pattern of IBD distortion, and to keep that genetic effect size constant under a single-gene model or a two-locus model with genetic effects at two linked loci (IDDM1 and IDDM15). In the T1D data, the evidence in support of the presence of IDDM15, can be assessed by comparing the MLS obtained under the two-locus IBD simulations to the MLS obtained in the single-locus IBD simulations for each replicate. The results from Figure 2.4 approximate the two-locus simulations better than the single locus case. The observed single-locus MLS at IDDM15 was 8.75 (Figure 2.4), which is closer to the average single-locus MLS at IDDM15 obtained under the two-locus simulations (9.06), than the MLS under the

single-locus simulations (3.27). There is a smaller peak proximal to IDDM1 in the two-locus simulations, which is similar to the observed single locus MLS.

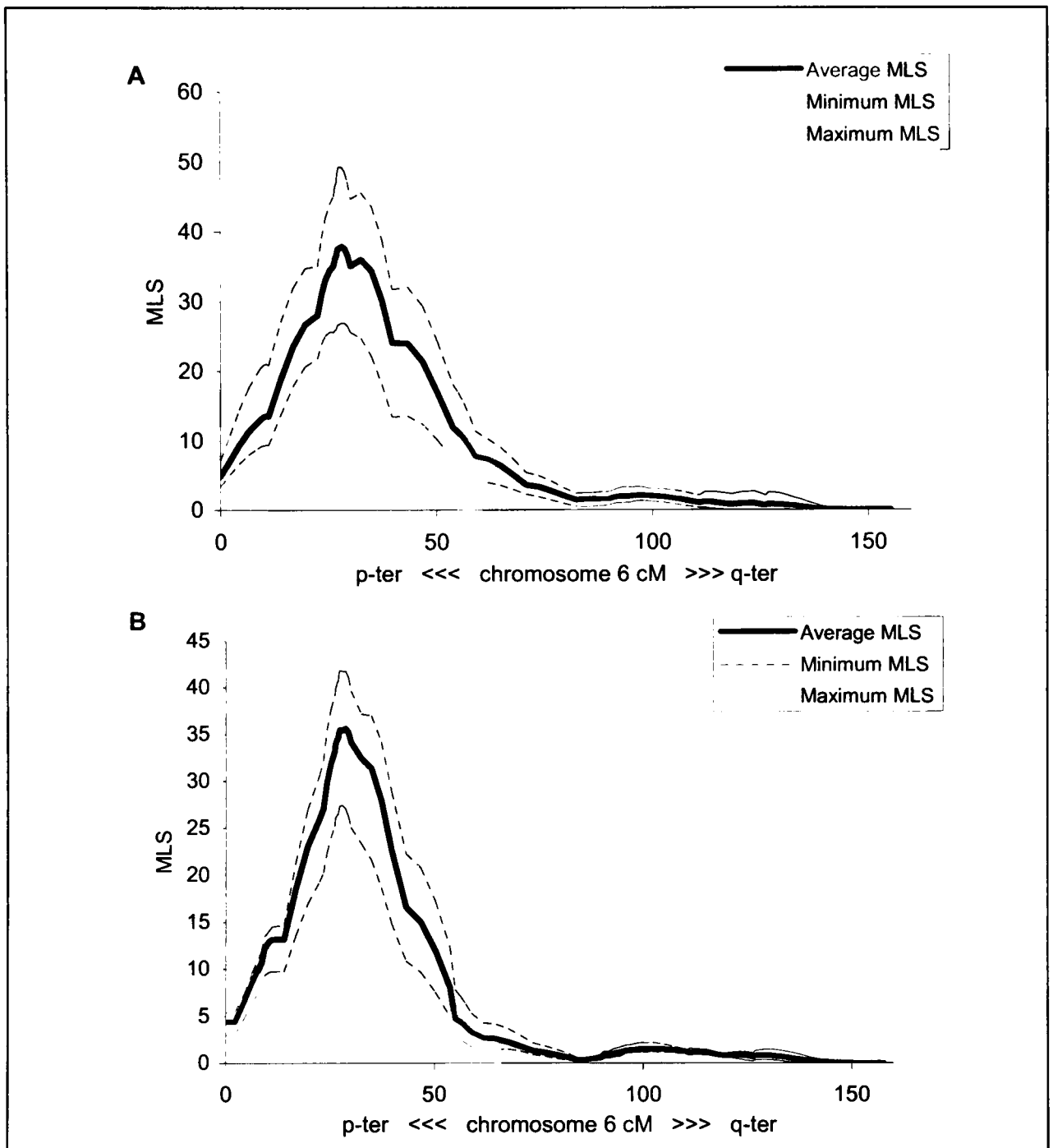


Figure 2.5. T1D chromosome 6 simulation results. Single-locus MLS calculated under (A) the two-locus model (30 replicates) and (B) the single-locus genetic model (5 replicates).

2.4 Discussion

In this chapter, a computational method is described for two-locus non-parametric analysis of genome-scan data in affected relatives under a flexible model of epistasis. The approach detected a previously identified pair of susceptibility loci in T1D, by dissecting the effect of IDDM15 on T1D independent of IDDM1, in the data of Mein et al (1998). The peak MLS attributed to IDDM15 independent of IDDM1 was 2.1 between markers D6S294 and D6S286, which are at the proximal end of the region previously identified as the IDDM15 locus by Delepine et al. (1997) and coincide with the region previously identified by Cordell et al. (2000) and Biernacka et al. (2005) in these data. The results in this study are similar to those obtained by Cordell et al. (2000), but there are slight differences in the peak MLS most likely because the method in this chapter uses all the available marker data along the chromosomes, unlike the approach used by Cordell et al. (2000). In addition, because the two-locus peak involves two linked loci, the results will be sensitive to marker map specification. Recently, Barber et al. (2006) have examined the evidence for IDDM15 in these data assuming sex-specific recombination rates. They obtained similar findings, however, their work highlights the importance of using accurate genetic maps, in particular sex-specific map estimates whenever possible. It is likely that IDDM1 and IDDM15 have additive effects on the trait, because the peak two-locus MLS was obtained under an additive model, which is similar to the finding of Cordell et al. (2000).

The method presented in this chapter is based on the work of Farrall (1997) and Cordell et al. (1995). The ultimate goal is to apply this approach to genome-scan data. There is therefore a need to calculate two-locus MLS statistics for both linked and unlinked susceptibility loci (on- and off-diagonal in a two-dimensional genome

scan grid). Consequently, Merloc, based on the method of Farrall (1997), is aimed primarily at linked loci in the calculation of the allele-sharing probabilities. For consistency, the same approach is used for the calculation of the IBD probabilities for unlinked loci, setting the recombination fraction at one-half. This method is more computationally intensive than simply multiplying the marginal IBD's for unlinked loci, but presents a precise estimate of the real two-locus marker allele-sharing probabilities.

The test statistic used in this study assumes that all pairs of sibs can be treated as independent observations. The families in T1D consist of single affected sib-pairs so there is no bias due to the presence of multiplex sibships. However, one could introduce a weighting criterion to take into account related pairs of affected sibs if the data merited such a modification.

Different nested models can be fit in the two-locus MLS to approximate the general two-locus model. Under the null hypothesis, the distribution of the MLS appears similar across two-locus models (Figure 2.2). In addition, the epsilon-epistatic genetic model appears to approximate the general two-locus model well and includes fewer degrees of freedom. By the principle of parsimony, it is preferable to use the simpler model (the epsilon-epistatic model) because it has fewer parameters, rather than the general model, to model epistasis and test the fit of nested (multiplicative, additive and single-locus) models. In addition, the epsilon-epistatic model allows direct maximum likelihood estimates of the degree of epistasis (ε) in the data.

CHAPTER III. Comparison of multilocus linkage methods

3.1 Introduction

Several linkage methods have been proposed to assess the evidence that multiple regions contribute to a complex trait using linkage analysis. However, there has been little effort to compare the performance of the different approaches in the presence and absence of interactions. In this chapter I compared the performance of Merloc to two other multilocus linkage methods, Genehunter-Plus (Cox et al. 1999), and IBD regression (Holmans 2002) using simulations of single-locus models and two-locus models of epistasis and heterogeneity in samples of affected-sib-pairs.

Genehunter-Plus (GHP) and the IBD regression approach (IBDreg) are both methods of conditional linkage analysis because they evaluate the evidence for linkage at one region while accounting for interactions with a second region. Merloc can perform both joint and conditional two-locus analyses where the allele sharing is estimated jointly at the two loci and is used to assess evidence for linkage to both locations. The three linkage methods were examined under a null genetic model, under 3 single-locus (1L), and 13 two-locus (2L) models. The two-locus models were selected from the literature to represent a wide range of epistasis and heterogeneity (see section 3.2.1 for detail) and represent the space of two-locus models. To compare the performance of the methods I estimated false-positive rates where applicable, power to detect the susceptibility loci and the interaction, and obtained the expected parameter values under different genetic models.

3.2 Methods

Marker genotypes were simulated at markers next to two disease-contributing loci (locus 1, L1, and locus 2, L2) for 100 ASPs (in 100 nuclear families). The two

disease loci were unlinked to each other and either did not contribute to the trait or contributed to the trait under 16 1L and 2L genetic models (see section 3.2.1). One fully-informative marker with four equi-frequent alleles and no missing data was simulated 0.0001 recombination units away from each disease locus (Figure 3.1).

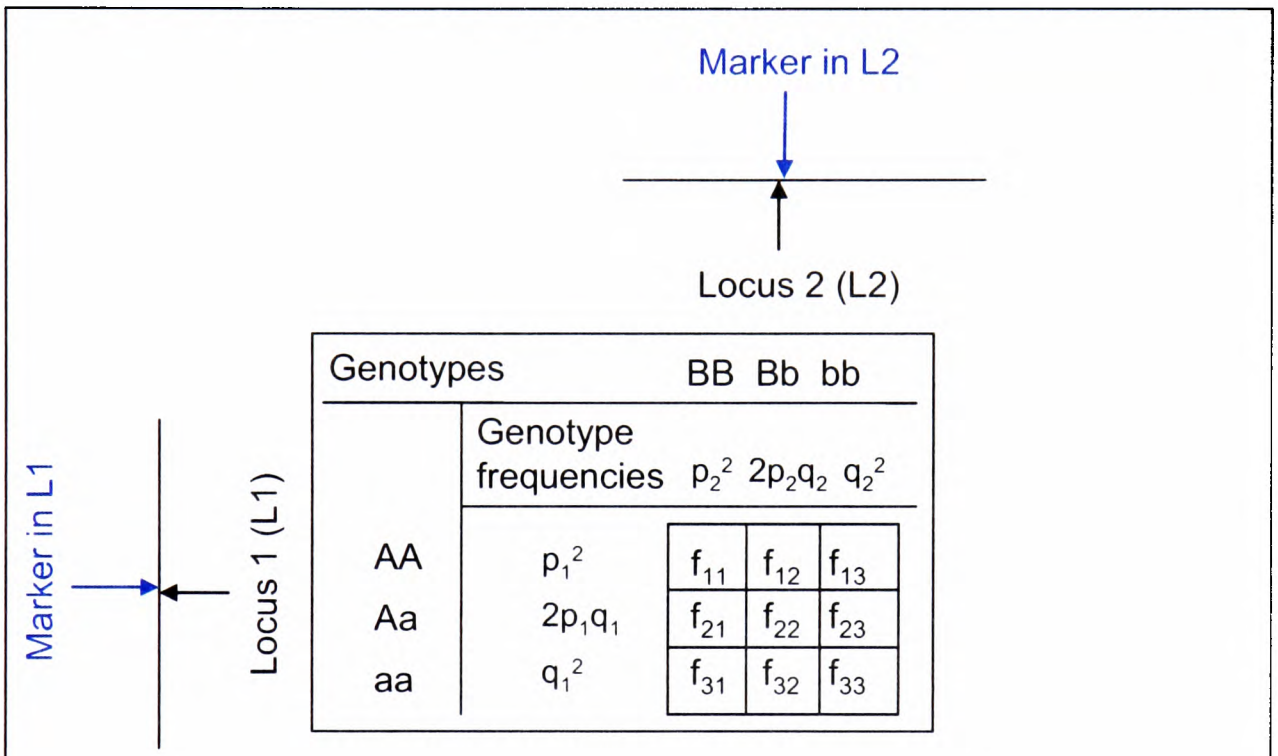


Figure 3.1. Parameters specifying the two-locus parametric models examined. For example, for a single-locus recessive model at the first disease gene $f_{11}=f_{12}=f_{13}=f_{21}=f_{22}=f_{23}=0$ and $f_{31}=f_{32}=f_{33}=f$.

To perform the simulations genotypes were simulated at four markers – two unlinked bi-allelic loci (the disease loci) and two loci with four equi-frequent alleles each (the markers) on top of the disease loci. My Perl script (*get.asp*) initially calls Simulate2 (Terwilliger et al. 1993; Terwilliger and Ott 1994) to simulate four fully informative markers (because I could set the parental genotypes to heterozygotes in Simulate2) and then sifts through the simulated nuclear pedigrees and selects ASP families according to a user-specified genetic model. ASP pedigrees are those in which the genotypes at the two disease loci in the two children (in a pedigree) fall into an affection joint-genotype category according to the underlying genetic model. For example, in a two-locus recessive model, where only the double disease-allele homozygotes are affected, both children in a sib-ship must be double homozygotes (at L1 and L2) to select that sib-ship as an ASP. Using this approach, 1 000 sets of 100 ASPs were selected for each of the single-locus and two-locus genetic models.

3.2.1 Genetic models

Simulations were performed under one null and 16 genetic models specified in terms of the disease-locus genotype penetrances for the trait (Table 3.1). First, I simulated marker genotypes at two unlinked locations under the null hypothesis that neither locus contributes to the trait (m0). For these simulations, I used Simulate2 and selected the first 100 nuclear families as ASP families, and repeated this 1 000 times. Second, I simulated marker genotypes at two unlinked locations under three 1L models: recessive (m1), over-dominant or interference (m2), and dominant (m3) disease mode of inheritance at locus 1. Third, I simulated marker genotypes at two unlinked locations under 13 2L genetic models (m4 – m16). I used *get.asp* to select ASP families in the 1L and 2L simulations.

Table 3.1. Genetic models in the multilocus comparison simulations.

Model	Classification	Penetrance matrix[a]			f^b	Allele frequency[c]	L1 marginal penetrances[d]	L2 marginal penetrances[d]			MLS[e] L1	MLS[e] L2	K	λ_s
m0	Null	-	-	-		-	-	-	-	-			-	-
m1	1L recessive (at L1)	*0*	*0*	*0*	0.36	0.540	0	0.105	0.105	0.105	4.27	0.14	0.105	2.033
		0	*0*	*0*			0							
		f	*f*	*f*			0.360							
m2	1L overdominant (at L1)	*0*	*0*	*0*	0.37	0.830	0	0.104	0.104	0.104	4 68	0.14	0.104	2.022
		f	*f*	*f*			0.370							
		0	*0*	*0*			0							
m3	1L dominant (at L1)	*0*	*0*	*0*	0.32	0.175	0	0.102	0.102	0.102	3.38	0.15	0.102	2.014
		f	*f*	*f*			0.320							
		f	*f*	*f*			0.320							
m4	2L Multiplicative	*0*	*0*	*0*	0.39	0.720	0	0	0	0.202	1.48	1.49	0.105	2.035
		0	*0*	*0*			0							
		0	*0*	*f*			0.202							
m5	2L Multiplicative	*0*	*0*	*0*	0.58	0.695	0	0	0.246	0	3.25	3.30	0.104	2.043
		0	*f*	*0*			0.246							
		0	*0*	*0*			0							
m6	2L Multiplicative	*0*	*0*	*0*	0.39	0.305	0	0	0.202	0.202	1.56	1.57	0.104	2.031
		0	*f*	*f*			0.202							
		0	*f*	*f*			0.202							
m7	2L Multiplicative	*0*	*0*	*0*	0.38	0.575	0	0	0.126	0.126	3.53	0.35	0.103	2.041
		0	*0*	*0*			0							
		0	*f*	*f*			0.311							
m8	2L Heterogeneity	*0*	*0*	*f*	0.35	0.390	0.053	0.053	0.053	0.385	1.51	1.49	0.104	2.033
		0	*0*	*f*			0.053							
		f	*f*	$2f-f^2$			0.385							
m9	2L Heterogeneity	*0*	*f*	*0*	0.31	0.905	0.053	0.053	0.347	0.053	1.22	1.20	0.104	2.024
		f	$2f-f^2$	*f*			0.347							
		0	*f*	*0*			0.053							
m10	2L Additive	*0*	*¼f*	*½f*	0.95	0.110	0.052	0.052	0.290	0.530	0.96	0.99	0.105	2.011
		¼f	*½f*	*¾f*			0.290							
		½f	*¾f*	*f*			0.530							
m11	2L Additive	*0*	*f*	*0*	0.29	0.900	0.052	0.052	0.342	0.052	1.28	1.22	0.104	2.014
		f	*2f*	*f*			0.342							
		0	*f*	*0*			0.052							
m12	2L Epistatic	*0*	*0*	*0*	0.39	0.575	0	0.070	0.070	0.261	2.80	0.58	0.105	2.023
		0	*0*	*f*			0.070							
		f	*f*	*f*			0.390							
m13	2L Epistatic	*0*	*0*	*0*	0.40	0.465	0	0	0.086	0.286	1.37	1.44	0.105	2.030
		0	*0*	*f*			0.086							
		0	*f*	*f*			0.286							
m14	2L Epistatic	*0*	*0*	*f*	0.45	0.365	0.060	0.060	0.060	0.390	1.98	2.07	0.104	2.041
		0	*0*	*f*			0.060							
		f	*f*	*0*			0.390							
m15	2L Epistatic	*0*	*f*	*0*	0.39	0.915	0.060	0.060	0.330	0.060	1.51	1.55	0.103	2.026
		f	*0*	*f*			0.330							
		0	*f*	*0*			0.060							
m16	2L Epistatic	*0*	*0*	*f*	0.72	0.745	0.400	0.400	0.140	0.050	2.247	2.264	0.104	2.019
		0	*½f*	*0*			0.140							
		f	*0*	*0*			0.050							

[a] Two-locus penetrance structure for each genetic model with L1 in rows and L2 in columns. [b] The value of the joint penetrance in the penetrance matrix for each model. [c] Disease allele frequencies at L1 and L2. [d] Estimated marginal penetrances at L1 and L2. [e] Average single-locus MLS.

Two-locus genetic models were classified according to the definitions by Risch (1990a) according to the penetrance relationships described in chapter I (equations 1.6, 1.7, and 1.8), falling into four groups: multiplicative (m4, m5, m6, m7), heterogeneity (m8, m9), additive (m10, m11), and other 'epistatic' (m12 – m16) two-locus models. The 'epistatic' category included five models, that have been

described in the literature as models of epistasis (Neuman and Rice 1992; Frankel and Schork 1996; Todorov et al. 1997; Li and Reich 2000; Holmans 2002; Ritchie et al. 2003; Purcell and Sham 2004).

The genetic models were standardized in terms of genetic effect sizes for comparison purposes. Let there be two biallelic loci contributing to the trait, with three possible genotypes at locus 1 (i=1 for AA, i=2 for Aa, and i=3 for aa) and three genotypes at locus 2 (j=1 for BB, j=2 for Bb, and j=3 for bb). The allele frequencies for the disease alleles are P(a) = q_1 = 1- p_1 at locus 1, and P(b) = q_2 = 1 – p_2 (Figure 3.1). Let the genotype frequency at the first locus be P^1_i, where $P^1_1 = p_1^2$, $P^1_2 = 2\,p_1\,q_1$, $P^1_3 = q_1^2$, and at the second locus be P^2_j, where $P^2_1 = p_2^2$, $P^2_2 = 2\,p_2\,q_2$, $P^2_3 = q_2^2$. Let f_{ij} be the penetrance for the joint genotype ij, with genotype i at locus 1 and j and locus 2. Then the population prevalence (K) can be expressed as

$$K = \sum_{i=1}^{3}\sum_{j=1}^{3} P_i^1 P_j^2 f_{ij} \qquad (3.1)$$

and the recurrence risk K_R of type R relative to the proband as

$$K_R = \frac{1}{K}\left(\sum_{i=1}^{3}\sum_{j=1}^{3} P_i^1 P_j^2 f_{ij} \sum_{r=1}^{3}\sum_{s=1}^{3} \tau_{ir}^1 \tau_{js}^2 f_{rs}\right) \qquad (3.2)$$

where τ^1_{ir} is the probability that a relative will have genotype r given that the proband has genotype i at locus 1. Similarly, τ^2_{js} is the probability of a proband with genotype j and a relative with genotype s at locus 2. To obtain the genotype transition probabilities (τ) for sibs, I followed the approach of Li and Sacks (1954) and obtained the matrix, S, of genotype transition probabilities. S has dimensions $i \times r$ (3×3) and its elements are the genotype transition probabilities τ_{ir} for full-siblings at one locus, where the proband has genotype i and the sibling of the proband has genotype r. From Li and Sacks (1954),

$$S = \frac{1}{4}I + \frac{1}{2}T + \frac{1}{4}O \tag{3.3}$$

where,

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},\quad T = \begin{bmatrix} p & q & 0 \\ \frac{1}{2}p & \frac{1}{2} & \frac{1}{2}q \\ 0 & p & q \end{bmatrix},\quad O = \begin{bmatrix} p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \end{bmatrix}$$

The $i \times r$ matrix (S) of genotype transition probabilities τ_{ir} at a locus (with allele frequencies p and q) for siblings of probands becomes

$$S = \begin{bmatrix} \frac{1}{2}\left(p + \frac{1}{2}p^2 + \frac{1}{2}\right) & \frac{1}{2}q(1+p) & \frac{1}{4}q^2 \\ \frac{1}{4}p(1+p) & \frac{1}{2}(1+pq) & \frac{1}{4}q(1+q) \\ \frac{1}{4}p & \frac{1}{2}p(1+q) & \frac{1}{2}\left(q + \frac{1}{2}q^2 + \frac{1}{2}\right) \end{bmatrix}$$

To standardize the models, I assumed equal disease allele frequencies at the two disease loci, $q_1 = 1 - p_1 = q_2 = 1 - p_2$, population prevalence of the trait (K) of 10%, and overall relative-risk to siblings of affected individuals ($\lambda_s = K_R/K$) set to 2. I then grid-searched for the maximum value of the joint-genotype penetrance f_{ij} that would result in $\lambda_s = 2$, $K = 0.1$, and equal disease allele frequencies, by numerically solving equations (3.1) and (3.2). I allowed for an error of 0.05 units from these values of λ_s and K and incremented over intervals of 0.01 for f_{ij} in the search (Table 3.1). The f_{ij} values used in this chapter (see Table 3.1) are not the unique solution to equations (3.1) and (3.2), but are the maximal values that result in $\lambda_s = 2$, $K = 0.1$.

I also calculated the marginal penetrances at each locus for the two-locus models. The marginal penetrances at locus 1, f_i, can be expressed as

$$f_i = \sum_{j=1}^{3} P_j^2 f_{ij} \tag{3.4}$$

and similarly for locus 2, f_j, where P and f_{ij} are as defined above. The obtained locus-specific population prevalence factors attributed to loci 1 and 2 (K_1 and K_2) estimated from the marginal penetrances and allele frequencies were the same as K as expected.

3.2.2 Multilocus Linkage Methods

I analyzed each replicate of the 100 ASPs with four different analytic methods (Genehunter-two-locus, Merloc, Genehunter-plus, and IBD regression). First, Genehunter-two-locus (GHT) was used to obtain the parametric Lod scores under the true single-locus or two-locus genetic model. I computed the two-locus parametric Lod score by comparing the likelihood that both loci contributed to the trait under the correct genetic model versus the null likelihood that neither gene is linked to the trait. This analysis produced estimates of power under the optimal scenario, which I compared to the power estimates from the other methods.

I used Merloc to obtain the MLS under single-locus and two-locus models using the single-locus (SL), and two-locus additive (ADD), multiplicative (MUL), epsilon-epistatic (EPS), and general (GEN) test-statistics. I examined the distribution of the different MLS statistics (GEN, EPS, MUL, ADD, and SL) for each set of simulations and obtained maximum likelihood estimates of epsilon for each model. I also compared four types of differences in the MLS statistics for each simulate, GEN-ADD, GEN-MUL, GEN-EPS, and GEN-SL. Following the parameterization from chapter II, epistasis may be defined as a deviation from an additive penetrance function: a test for epistasis can therefore be constructed by comparing the fit of the general model with the fit of a restricted additive-effects-only model (GEN-ADD). Alternatively, epistasis can be defined as a departure from a multiplicative penetrance function, and a test can be constructed comparing the general model with the fit of the multiplicative model (GEN-MUL). The degree to which the epsilon-epistatic model approximated the general model in the presence of epistasis or heterogeneity (GEN-EPS) was also estimated. Finally for the purpose of comparison to the next two

multilocus linkage methods presented in this chapter I obtained the GEN-SL1, which represents evidence for linkage at locus 2 in the presence of interaction.

Genehunter-Plus (GHP) was also used to search for epistasis and heterogeneity at each locus. This conditional linkage analysis approach (Cox et al. 1999) models epistasis or heterogeneity as a deviation from a multiplicative penetrance function. A positive correlation between the single-locus familial linkage scores for a pair of loci provides an initial suggestion of epistasis between the loci, while a negative correlation suggests heterogeneity. Spearman correlation coefficients were calculated (in the R statistical environment, http://www.stats.bris.ac.uk/R/) for the familial NPL scores from the markers at L1 and L2 for each replicate, and with the overall Lod scores (over all replicates) per model. The conditional linkage analysis approach explores the familial correlations further by reassessing the evidence for linkage at locus 1 by incorporating the evidence for linkage at locus 2 using family-specific weights. The reciprocal analysis is also performed – assessing the increase in the evidence for linkage at locus 2 while taking into account the linkage at locus 1. Epistatic and heterogeneity weighting schemes were used with discrete weights. Under the epistatic discrete weighting scheme, families with positive NPL scores at a given locus were assigned a weight of 1, and those with zero or negative NPL scores a weight of zero. Under the heterogeneity weighting scheme families with positive NPL scores were assigned a weight of 0, and families with zero or negative NPL score were assigned a weight of 1. Evidence for epistasis or heterogeneity is assessed by comparing the single-locus Lod (Kong and Cox 1997) to a weighted Lod, which accounts for the evidence for linkage at the reciprocal locus by incorporating the family-specific weights. I followed this approach by calculating the difference in Lod scores (Δ Lod = weighted Lod - unweighted Lod). Results are

shown for the Kong and Cox (1997) linear Lod score, using the difference between the weighted and unweighted Lod scores, following weighting under the epistatic scheme at locus 1 (Δ EPI LOD L1) and locus 2 (Δ EPI LOD L2), and following weighting under the heterogeneity scheme at locus 1 (Δ HET LOD L1) and locus 2 (Δ HET LOD L2). The difference in the Lod scores, multiplied by 2log(10) is asymptotically distributed as χ_1^2 under the null hypothesis of no interaction according to Cox et al. (1999).

Finally, I used IBD regression (IBDreg), which performs logistic regression using the estimated IBD probabilities at two loci. This method can be applied to IBD or genotype data at the relevant loci, but for comparison purposes only IBD data was used for a test of linkage. This approach also examines evidence for epistasis or heterogeneity as a departure from a multiplicative genetic model (Holmans 2002). The probability, p, that an ASP shares a given allele IBD at locus 1 is modeled as a logistic regression function of the intercept α (for the IBD distortion from the null at locus 1), which represents linkage at locus 1, and regression coefficient β (for the mean standardized proportion of alleles IBD at locus 2), which represents the interaction between locus 1 and locus 2. Holmans (2002) models probability as

$$p = \frac{e^{\alpha+\beta(x-\bar{x})}}{1+e^{\alpha+\beta(x-\bar{x})}} \tag{3.5}$$

where x is the proportion of shared alleles at locus 1, $\bar{x}$ is the mean of x. A likelihood can be computed as a function of p and may be maximized with respect to both α and β, or just α. Two test statistics, T and S, follow, where T measures linkage at locus 1 allowing for interaction,

$$T = 2\ln\left(\frac{L(\hat{\alpha},\hat{\beta})}{L(\alpha=0,\beta=0)}\right) \tag{3.6}$$

and S, which just measures the effect of the interaction,

$$S = 2\ln\left(\frac{L(\hat{\alpha},\hat{\beta})}{L(\hat{\alpha},\beta=0)}\right) \tag{3.7}$$

Under the null the likelihood ratio test for linkage in the presence of interaction, T, is distributed as χ_1^2 with probability 0.5, and χ_2^2 with probability 0.5, while the likelihood ratio test for the interaction, S, is distributed asymptotically as χ_1^2 (Holmans 2002). I analysed the simulated data for both linkage and interaction, and just interaction at loci 1 and 2, with test statistics T1 and S1 for locus 1 and T2 and S2 for locus 2. The value of the interaction coefficient β was examined for the two-locus models studied and, for significant S, $\beta > 0$ was taken to indicate epistasis, and $\beta < 0$ implied heterogeneity.

3.3 Results

3.3.1 Significance thresholds and Type I error

Single-locus evidence for linkage was initially obtained for each model using the single-locus MLS. The results from the single locus models showed high linkage scores, and the results from the majority of the two-locus models were comparable for two-locus symmetric models (Table 3.1). The Kong and Cox (1997) linear Lod scores were also obtained (results not shown) and were very similar to the single-locus MLS.

The choice of the appropriate null hypothesis is important in establishing significance thresholds for the test statistics used in this chapter and one may apply different null models to establish significance. First, a null genetic model (m0) may applied. Under m0 there are no genetic effects at either of the two loci, and thresholds calculated using this the null hypothesis would be appropriate for detecting joint effects at the two loci. Second, a single-locus genetic model (such as models m1-m3 in this chapter) may be used as the null model. Significance thresholds using this null

hypothesis would be suitable first, for detecting the presence of a secondary gene, given a major linked locus (that is, in conditional analyses), and second, for detecting epistasis. However, the thresholds in this case will depend on the exact single-locus (null) model simulated, and while in this chapter only 3 single-locus models were simulated, in reality there is a wide range of possible single-locus (null) models. Therefore, I did not use models m1-m3 to establish significance thresholds in this chapter, but I have used significance thresholds based on a 'single-locus effect null model' in later chapters when applying the method to genome-scan data. In chapters IV-VI, I used the most likely single-locus model underlying either of the two loci (by sampling from the maximum likelihood estimates of the IBD probabilities under a single-locus model for either locus 1 or locus 2) as the null model to establish the significance of detecting the other locus for conditional analyses (GEN-SL1 and GEN-SL2). Finally, the third set of null hypotheses would include two-locus models simulated under heterogeneity, additive, or multiplicative penetrance structures (as defined in equations 1.6-1.8). Significance thresholds calculated under these models would be useful for detecting the evidence for epistasis, as a deviation from heterogeneity, additivity, or multiplicativity, correspondingly. In this case, there are once again many possible models that fit these penetrance structures and that may be used as null models, but in this chapter I have only simulated two heterogeneity, two addtiive, and four multiplicative models. Therefore, in this chapter I have not used these models as null models to establish significance thresholds, however, I calculated and apply such thresholds to genome-scan data in later chapters. In chapters IV-VI, I use the most likely underlying additive and multiplicative two-locus model (by sampling from the maximum likelihood estimates of the IBD probabilities under an additive and a multiplicative model) as the null model to establish the significance of

epistasis, as deviation from additivity, GEN-ADD, and as a deviation from multiplicativity, GEN-MUL. For the remainder of chapter III all significance thresholds were calculated assuming a null genetic model, m0.

Significance thresholds for the 16 1L and 2L models in GHT were calculated by analyzing the null simulates from m0 under 1L or 2L models (Figure 3.2). Table 3.2 presents the significance thresholds from the remaining three methods, calculated by using simulates from m0. Table 3.2A presents the significance thresholds from Merloc for a range of test statistics that are comparable to the ones obtained in chapter II (Table 2.3). The results are very similar, because both are based on simulations from a pair of markers in 100 fully informative ASPs. The thresholds in chapter II (Table 2.3) are more accurate because they are based on 100,000 replicates rather than 1,000. Tables 3.2B and 3.2C present significance thresholds calculated for GHP and IBDreg, respectively.

Table 3.2 Significance thresholds in Merloc, Genehunter-Plus, and IBD regression.

	Test-statistic thresholds (m0) [a]		
Test Statistic	**α = 0.05**	**α = 0.01**	**α = 0.001**
	(A) Merloc		
GEN	1.40	2.06	2.89
EPS	1.34	2.06	2.89
MUL	1.14	1.92	2.89
ADD	1.17	1.88	2.89
SL	0.70	1.53	2.35
GEN-EPS	0.15	0.38	0.98
GEN-MUL	0.31	0.65	1.02
GEN-ADD	0.34	0.71	1.11
GEN-SL1	0.89	1.43	2.42
EPS-SL1	0.84	1.43	2.42
	(B) Genehunter-plus		
EPI ΔLOD at L1	0.44	0.93	1.86
EPI ΔLOD at L2	0.47	1.14	1.79
HET ΔLOD at L1	0.37	0.81	1.35
HET ΔLOD at L2	0.43	0.89	1.68
	(C) IBD regression		
T at L1 (T1)	6.22	8.38	12.82
T at L2 (T2)	6.14	9.54	14.02
S at L1 (S1)	3.78	6.83	10.40
S at L2 (S2)	3.94	6.57	9.50

[a] The test statistics in (A) are MLS statistics, these in (B) are Kong and Cox (1997) Lod scores, and the test-statistics in (C) are chi-square distributed.

False-positive rates were estimated using simulations and previously published results (Table 3.3). I compared thresholds obtained under the null model (Table 3.2) to results obtained under single-locus models m1, m2, and m3 for test statistics that assess evidence for linkage and interaction to the unlinked locus 2 at $\alpha = 0.05$ (GEN-SL1 and EPS-SL1 in Merloc, Δ HET LOD L2 and Δ EPI LOD L2 in GHP, and T2 and S2 in IBDreg). The false-positive rates in GHP and IBDreg fall at or below 0.05, and in Merloc the statistics assessing evidence for linkage to locus 2 in the presence of interaction (GEN-SL1 and EPS-SL1) have slightly higher Type I error rates of 0.06 to 0.08, but as discussed previously the null model used in this case assumes that there is no genetic effect at either locus. A more suitable null model for assessing the significance of the conditional results (GEN-SL1 and EPS-SL1) would be a single-locus model at locus 2. The appropriate choice of null models may have a greater effect in Merloc, which assesses the IBD at two loci jointly, rather than IBDreg and GHP, which assess only the single-locus IBD and condition upon them. The two-locus test statistics under model m1, m2, and m3 had more than 92% power to detect a genetic interaction, indicating that a significant two-locus MLS should be taken as evidence that at least one susceptibility gene is present.

I also compared the results at $\alpha = 0.05$ to previously published results. I used 2log(10) times the difference in Lod scores (Δ Lod) distributed as χ_1^2 for GHP (Cox et al. 1999), evaluating the S and T results compared to χ_1^2 and χ_2^2 for IBDreg (Holmans 2002), and comparing MLS from Merloc to the results of Cordell et al. (1995, Table 9 for GEN, ADD (HET), and MUL, and the higher of two values in Table 7 for GEN-ADD (GEN-HET) and GEN-MUL). The results from IBDreg and GHP indicate that the published significance thresholds are overly conservative compared to the results from this chapter. The results from Merloc have Type I error

rates just above 5% for the two-locus statistics (GEN, MUL, and ADD). This finding is not surprising because the significance thresholds are slightly lower in Cordell et al. (1995) compared this study (Table 3.2) and in chapter II, which reflects both the different amount of information in the samples simulated and different number of replicates. Surprisingly, the false-positive rates for the difference in MLS statistics are below 0.04 using published thresholds, but around 0.1 using the empirical thresholds from m0, however, a more appropriate null hypothesis for the difference in MLS statistics would be a two-locus additive or multiplicative model, rather than m0.

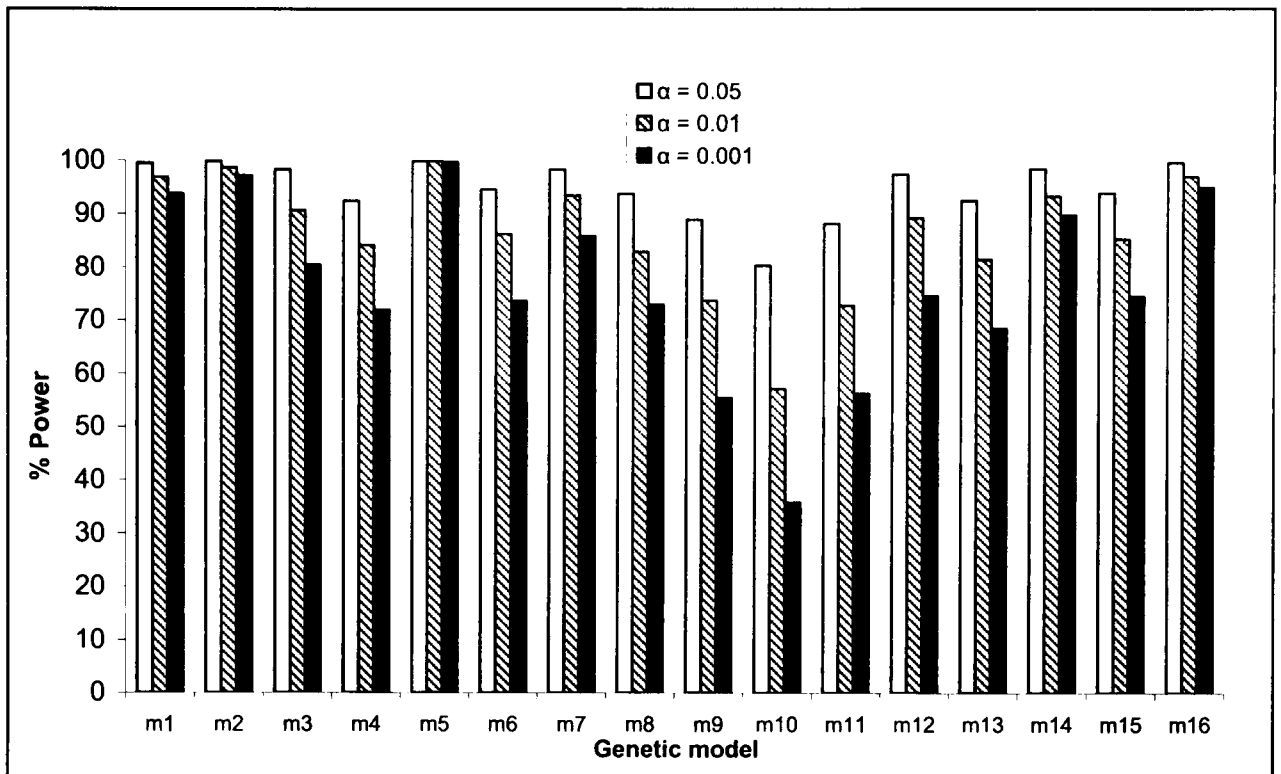
Table 3.3 Type I error and power in Merloc, GHP, and IBDreg for models m0-m3.

	% simulates exceeding m0 α = 0.05 threshold			Published[a]
Test statistic	**m1**	**m2**	**m3**	**m0**
		(A) Merloc		
GEN	97.70	98.30	92.80	7.20
EPS	97.80	98.40	93.00	-
MUL	98.40	98.80	94.10	5.10
ADD	98.30	98.80	93.60	5.30
SL2	4.40	4.30	4.60	4.40
GEN-EPS	10.70	11.90	8.40	-
GEN-MUL	11.10	13.50	7.80	4.00
GEN-ADD	11.60	14.80	9.30	4.10
GEN-SL1	7.50	8.00	6.00	-
EPS-SL1	6.30	6.70	6.00	-
		(B) Genehunter-plus		
Δ EPI LOD at L2	3.40	3.90	3.80	0.90
Δ HET LOD at L2	4.40	5.00	5.90	1.90
		(C) IBD regression		
T at L2 (T2)	3.90	3.50	3.10	4.10
S at L2 (S2)	3.30	4.90	4.50	3.30

[a] Published thresholds were compared to m0 simulates, see text for details.

3.3.2 Power

I evaluated power for each method under the different two-locus models. I first examined power from the parametric two-locus Lod score computed under the correct two-locus model. Figure 3.2 shows the results, with similar power estimates across most genetic model simulations, and slightly lower power observed for the two additive models and the two heterogeneity models.



α = 0.05	-1.1	-1.3	-0.5	0.2	-1.8	0.2	-0.7	0.1	0.3	0.5	0.3	-0.2	0.2	-0.3	0.2	-0.7
α = 0.01	0.5	0.2	0.8	0.9	-0.3	0.9	0.8	0.9	1.1	1.2	1.1	0.9	1.0	0.7	0.9	0.3
α = 0.001	1.0	0.9	1.7	1.5	0.3	1.5	1.5	1.4	1.6	1.8	1.6	1.7	1.5	1.2	1.5	0.9

Figure 3.2 Power estimates from the parametric two-locus Lod score in GHT. Below the power bars are the empirical significance thresholds used for GHT (with recombination fraction of 0.0001 between each marker and disease locus).

Merloc

The results from Merloc are shown in Figure 3.3 for the MLS statistics described in section 3.2.2, and the differences between MLS statistics. The general model, GEN, approximates the parametric two-locus results well (Figure 3.3A). The

models with fewer parameters – the epsilon-epistatic model, EPS, the multiplicative model, MUL, and the additive model, ADD, have very similar power estimates to those of GEN. ADD and MUL have only slightly higher power estimates (by 1-4%) than GEN particularly for the additive and heterogeneity models, and EPS falls between GEN and ADD/MUL.

The power to detect a significant interaction is greatest for models m5 and m16, when one assesses epistasis by using the difference in two-locus MLS statistics GEN-ADD and GEN-MUL (Figure 3.3B). GEN-ADD has on average more power to detect an interaction across all models, but the additive and heterogeneity models (m8-m11) under which ADD approximates GEN and should not detect epistasis. For these cases, the power rates can be interpreted as false-positive rates of detecting epistasis. In that case, GEN-ADD has an average false-positive rate of 18% (13-20%) for the four additive or heterogeneity models examined (m8-m11), and GEN-MUL has a false-positive rate of 32% (22-57%) for the four multiplicative models examined (m4-m7). However, it should pointed out that m0 is not the most appropriate null hypothesis for assessing the effect of epistasis, and in future studies one should use a single-locus model and two-locus additive and multiplicative genetic models as the null models in determining significance thresholds. Therefore, the results presented in this section, which use m0 as the null hypothesis and assess the power (and the false-positive rates) to detect epistasis using the difference in MLS statistics, should be interpreted with caution.

I also compared single-locus evidence for linkage at locus 2 not accounting for interaction (SL2) to evidence for linkage accounting for interaction (GEN-SL1) across the simulated two-locus genetic models (Figure 3.3C). The evidence for linkage at the second locus in a pair, accounting for interaction (GEN-SL1), is highest for

multiplicative model m5, and models m14 and m16. GEN-SL1 performs only slightly better than SL2 across all models.

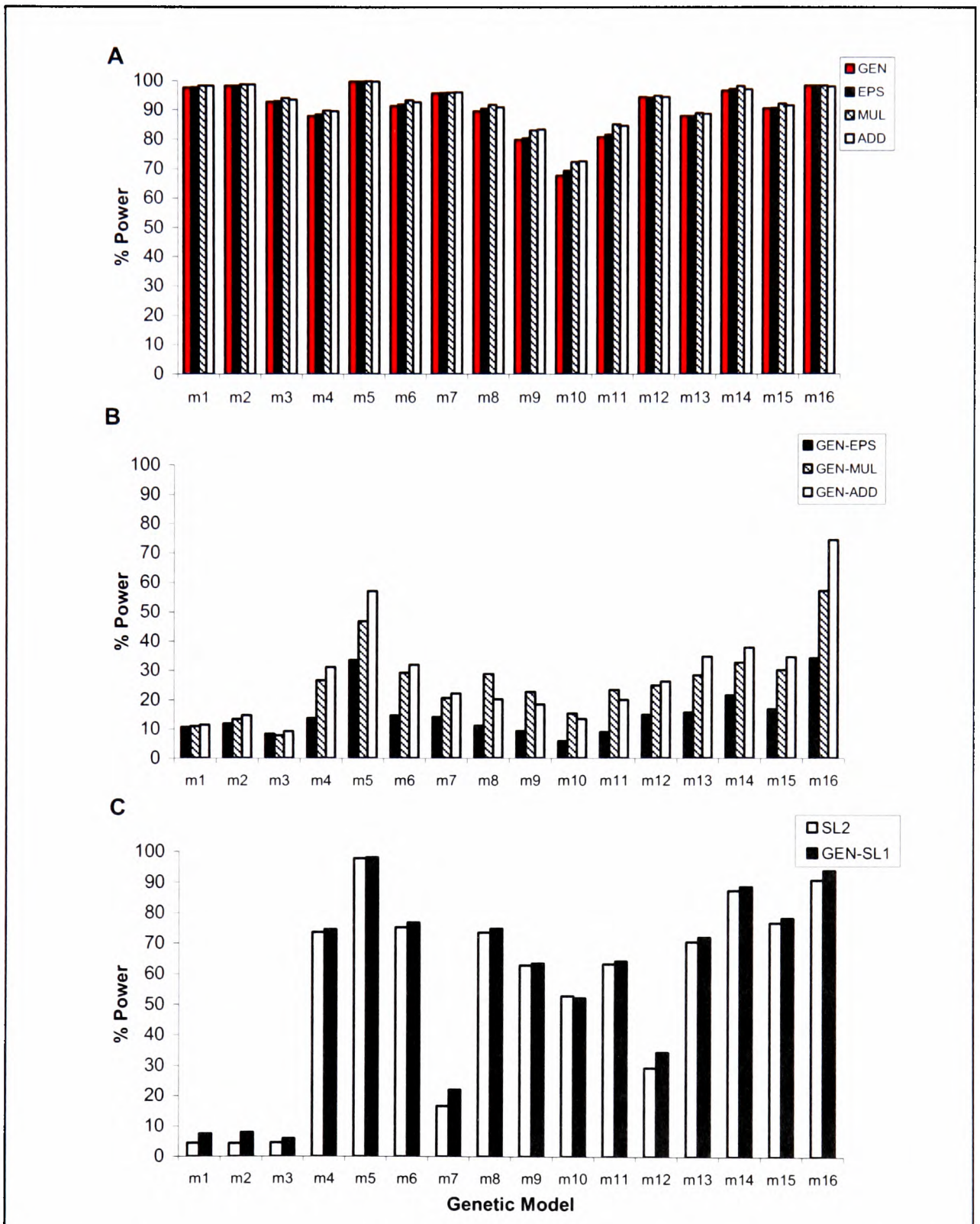


Figure 3.3. Power estimates for MLS statistics in Merloc at $\alpha = 0.05$ compared to m0 as the null model. (A) Power estimates for the general (GEN), epsilon-epistatic (EPS), multiplicative (MUL), and additive (ADD) two-locus MLS statistics. (B) Power estimates for difference in two-locus MLS statistics. (C) Power for linkage at locus 2, not accounting for interaction (single-locus MLS, SL2) and accounting for interactions (GEN-SL1).

Genehunter-Plus

The results for GHP are presented in Figure 3.4. GHP appears to have very low power estimates across all genetic models simulated. Models m4, m5, m6, and m7 represent multiplicative models and fall under the null for detecting interactions in GHP. However the remaining genetic models also have low power, with one exception for model m16, which has power of 15% at $\alpha = 0.05$ (Figure 3.4A).

Cox et al. (1999) suggest first assessing the correlation in the familial NPL scores, and then follow up significant findings with the weighting approach. Therefore, I also examined the GHP results only considering those replicates for which there was a significant positive or negative familial correlation. However, the number of replicates with significant familial correlation was quite low across the 16 models simulated (0.7%-12.5% of the 1000 replicates, Table 3.5). Therefore, I only present findings from models in which at least 5% of replicates had significant familial correlations. Figure 3.4B presents power estimates as the percentage of replicates in which the Δ Lod surpassed the m0 threshold from the total number of replicates with significant familial correlations. It should be noted that while the number of replicates with significant ($P < 0.05$) familial correlation was quite low across the 16 models simulated, the overall number of positive or negative correlations across the models was greater. For example, for additive model m10, 63% of replicates had negative familial correlations and in only 5.1% of replicates the negative familial correlations were significant.

I also examined a different classification of heterogeneity models - models in which the proportion of linked families to one of the two loci was exactly 0.5. I used the single locus models (m1, m2, and m3) and reversed the genetic effect to be attributed to locus 2, instead of locus 1, in the second 50 families in each replicate set.

These new heterogeneity models are denoted m1het, m2het, and m3het. For example, in m1het locus 1 was the disease locus under a recessive gene model in only the first 50 families and locus 2 was the second recessive disease gene, contributing to disease in just the second 50 families. In m2het two overdominant genes contributed to the disease, and in m3het two dominant genes contributed to the disease. I again examined power to detect a significant increase in Δ Lod under the heterogeneity weighting schemes at either locus. Power rates at $\alpha = 0.05$ were 8.7% (for Δ HET LOD L1) and 6.9% (for Δ HET LOD L2) for m1het, 7.7% (Δ HET LOD L1) and 6.3% (Δ HET LOD L2) for m2het, and 9.9% (Δ HET LOD L1) and 8.2% (Δ HET LOD L2) for m3het. The results from heterogeneity models m1het, m2het, and m3het, were very similar to the findings obtained for models m8 and m9.

The power estimates for GHP presented in Figure 3.4 assess epistasis using significance thresholds calculated under m0. As discussed previously a more appropriate null model to obtain significance thresholds and assess power would be a single-locus effect model, or (for GHP in particular) a multiplicative two-locus null model. To examine the appropriate choice of null hypotheses in more detail I obtained power estimates in GHP by using the single-locus models, m1-m3, as the null models in establishing significance thresholds. The resulting power estimates using m1-m3 as the null, were slightly higher than those obtained by using m0. On average, the observed increase in power was only 1.5% over the power obtained using m0 as the null (the greatest increase was observed for m16 from 15% under m0 to 17.7% under m1 as the null). However, overall the results were very similar to those obtained using m0 to calculate power.

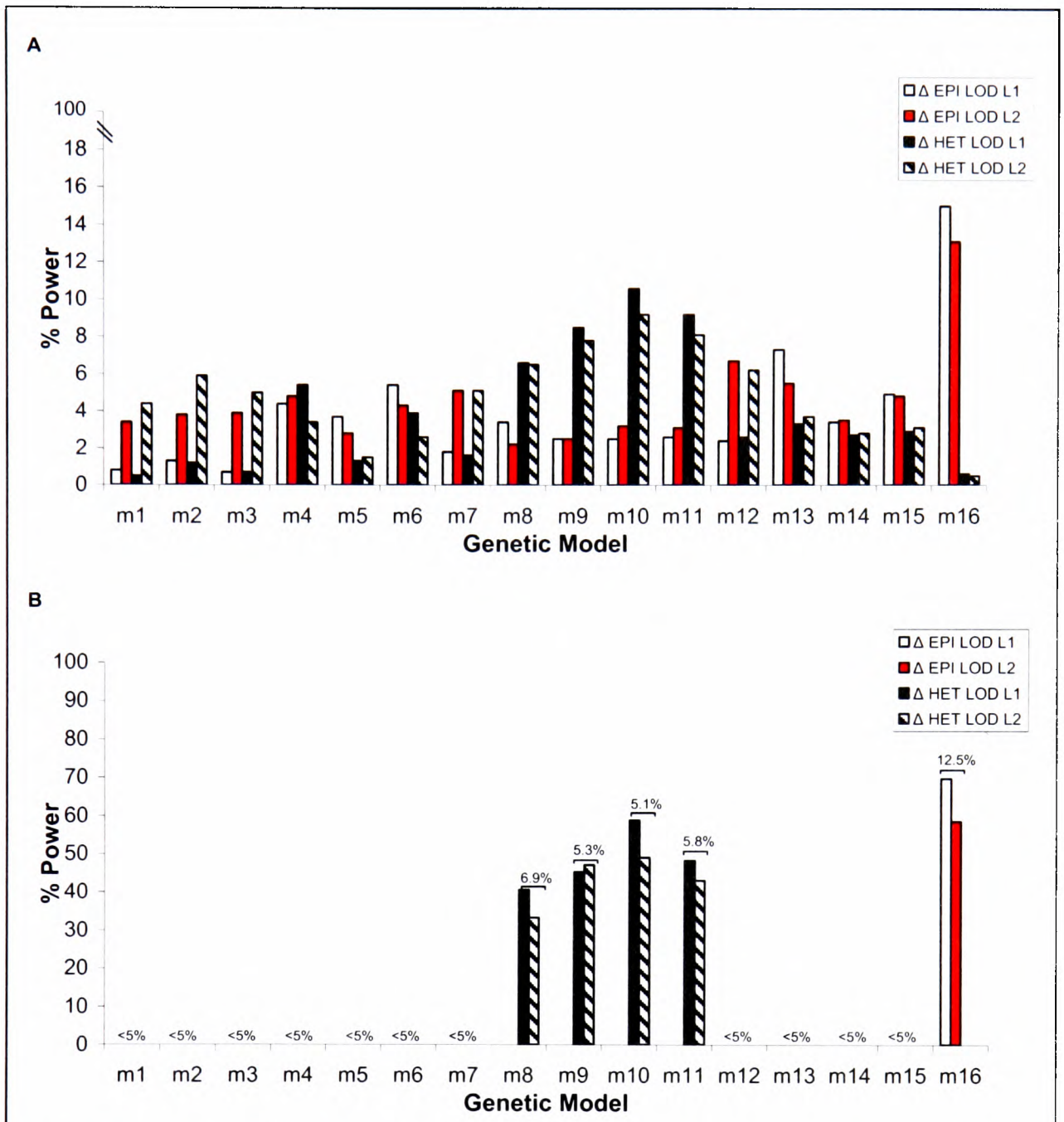


Figure 3.4. Power estimates in GHP at $\alpha = 0.05$. The four test statistics are 1) the difference in the locus2-epistatic-weighted and single-locus Lod score at locus 1 (Δ EPI LOD at L1), 2) the difference in the locus1-epistatic-weighted and single-locus Lod score at locus 2 (Δ EPI LOD at L2), 3) the difference in the locus2-heterogeneity-weighted and single-locus Lod score at locus 1 (Δ HET LOD at L1), and 4) the difference in the locus1-heterogeneity-weighted and single-locus Lod score at locus 2 (Δ HET LOD at L2). (A) Power estimates in all 1000 replicates. (B) Power estimates from GHP for replicates with significant familial correlations. Data only shown for simulated models in which the replicates with significant familial correlations exceeded 5% of the total 1000 replicates (the percentage of such replicates is shown above the bars).

IBD regression

The results from IBDreg are shown in Figure 3.5. Power estimates for the linkage statistic incorporating interaction (T) are higher than estimates for the effect of the interaction alone (S) across all models, as was expected because the significance thresholds for both T and S were calculated under m0. For the 13 two-locus models the power to detect linkage in the presence of interaction is highest for the two asymmetric models, m7 and m12, and then for m5 and m16. The power to detect an interaction event (S) is quite low across all models considered, but using single-locus model or two-locus multiplicative models as the null would be more appropriate in this case in establishing the power to detect an interaction. Models m4, m5, m6, and m7 represents multiplicative models and fall under the null for detecting interactions in GHP. However the remaining genetic models also have low power (for the S statistic), even model m16 for which power is lower than estimates obtained from GHP.

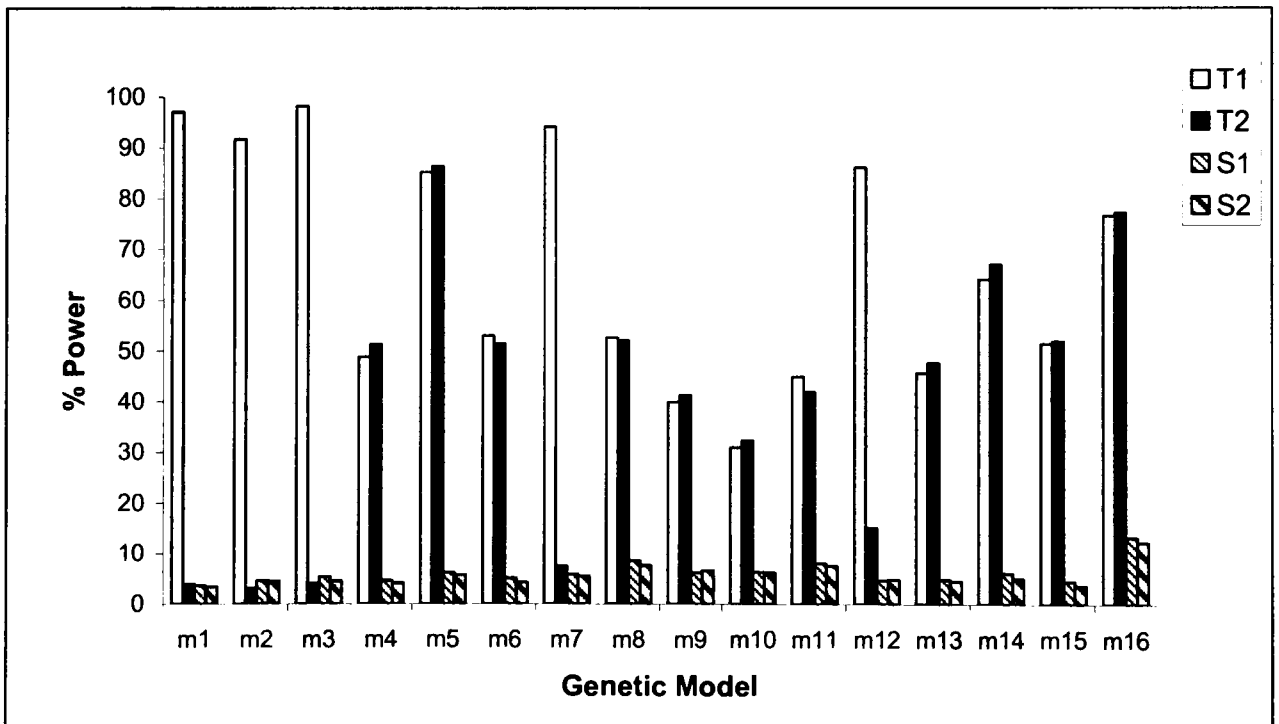


Figure 3.5. Power estimate from IBDreg at $\alpha = 0.05$. The T statistic tests evidence for linkage in the presence of interaction at locus 1 (T1) and locus 2 (T2). The S statistic assesses evidence for interaction only at either locus 1 (S1) or locus 2 (S2).

Comparison across methods

Table 3.4 presents a comparison of the results across the three methods at α = 0.05 in all 1000 replicates per simulated model. Estimates from IBDreg fall between those obtained by Merloc and GHP. For the sake of comparison, GEN-SL1 in Merloc should be used rather than GEN, because both GEN-SL1 and T2 assess evidence for linkage at locus 2 accounting for the interaction with locus 1. The pattern of power to detect linkage in the presence of interaction is similar in IBDreg T2 and Merloc GEN-SL1, but estimates for Merloc are generally higher. GHP does not do as well as the rest of the methods for the two-locus models considered in this study.

Table 3.4. Comparison across Merloc, GHP, and IBDreg.

	Two-locus power estimates at α = 0.05						
	GHP		**IBDreg**		**Merloc**		
Model	**Δ HET LOD L2**	**Δ EPI LOD L2**	**T2**	**S2**	**GEN**	**GEN-ADD**	**GEN-SL1**
Single-locus							
m1	4.4	3.4	3.9	3.3	97.7	11.6	7.5
m2	5.9	3.8	3.1	4.5	98.3	14.8	8
m3	5	3.9	4	4.5	92.8	9.3	6
Multiplicative							
m4	3.4	4.8	51.3	4.1	88	31.2	74.6
m5	1.5	2.8	86.4	5.7	99.8	57.1	98.2
m6	2.6	4.3	51.4	4.3	91.4	32	76.8
m7	5.1	5.1	7.5	5.5	95.8	22.2	22.1
Heterogeneity							
m8	6.5	2.2	52.1	7.6	89.6	20.2	74.8
m9	7.8	2.5	41.2	6.6	79.8	18.4	63.4
Additive							
m10	9.2	3.2	32.3	6.2	67.7	13.5	52.1
m11	8.1	3.1	41.9	7.5	80.8	20	64.2
Other epistatic							
m12	6.2	6.7	15.1	4.9	94.4	26.3	34.3
m13	3.7	5.5	47.6	4.5	88	34.8	72
m14	2.8	3.5	67	5	96.7	37.8	88.5
m15	3.1	4.8	51.9	3.7	90.5	34.6	78.2
m16	0.5	13.1	77.2	12.2	98.3	74.4	93.7

3.3.3 Parameter estimates

I examined the relationship between parameter estimates and genetic model used in the simulations across the three linkage methods (Table 3.5).

Table 3.5 Parameter estimates in Merloc, GHP, and IBDreg.

	Merloc[a]			GHP[b]				IBDreg[c]					
		MLS difference				% ρ* replicates		L1 interacts with L2			L2 interacts with L1		
Model	ε	GEN-ADD	GEN-MUL	ρ	*P*	+ve	-ve	Total β_2	% S1*	β_2*	Total β_1	% S2*	β_1*
Null													
m0	15.453	0.061	0.061	0.034	0.282	2.7	2.5	-0.009	5.0	0.008	-0.009	5.1	-0.027
Single-Locus													
m1	2.108	0.119	0.118	-0.010	0.749	1.8	1.7	0.007	3.6	0.039	0.005	3.3	0.005
m2	2.528	0.141	0.135	0.024	0.441	2.7	1.6	0.026	5.3	0.220	0.028	4.5	0.225
m3	2.215	0.104	0.101	-0.047	0.136	1.9	2.4	-0.004	4.6	-0.091	-0.003	4.5	-0.147
Multiplicative													
m4	0.745	0.312	0.239	0.011	0.719	2.8	1.5	0.016	4.6	0.262	0.016	4.1	0.145
m5	0.655	0.616	0.370	0.007	0.835	2.3	2.3	0.000	6.2	0.137	0.001	5.7	0.131
m6	0.730	0.298	0.251	-0.053	0.091	1.9	2.6	-0.008	5.1	-0.065	-0.006	4.3	-0.129
m7	1.606	0.206	0.187	-0.013	0.680	2.9	2.6	0.004	5.8	0.053	0.007	5.5	0.117
Heterogeneity													
m8	0.110	0.189	0.244	-0.101	0.001	1.1	6.9	-0.182	8.5	-0.694	-0.163	7.6	-0.742
m9	0.163	0.173	0.202	-0.058	0.065	0.9	5.3	-0.152	6.2	-0.712	-0.154	6.6	-0.677
Additive													
m10	0.213	0.134	0.161	-0.112	0.001	1.1	5.1	-0.146	6.3	-0.693	-0.146	6.2	-0.683
m11	0.191	0.178	0.205	-0.067	0.034	0.7	5.8	-0.158	8.0	-0.732	-0.156	7.5	-0.762
Other epistatic													
m12	1.081	0.260	0.227	0.024	0.453	2.2	2.3	0.005	4.7	-0.032	0.002	4.9	0.002
m13	1.475	0.331	0.253	-0.004	0.894	3.4	1.2	0.060	4.9	0.432	0.063	4.5	0.385
m14	0.488	0.358	0.280	-0.058	0.067	1.9	2.2	-0.039	6.1	-0.104	-0.038	5.0	-0.077
m15	1.221	0.332	0.252	-0.012	0.701	2.2	1.9	0.045	4.5	0.160	0.043	3.7	0.114
m16	15.208	0.871	0.489	0.009	0.765	12.5	0.1	0.341	13.2	0.999	0.342	12.2	0.996

* $p < 0.05$

[a] Geometric mean of ε and arithmetic means of the MLS differences

[b] Spearman's ρ and P in 1000 replicates, and the percentage of replicates with significant familial correlations (% ρ* replicates).

[c] Average interaction coefficient in (Total β) 1000 replicates and in (β*) the percent of replicates with $p<0.05$ interactions (% S*).

In Merloc I examined the maximum likelihood estimate of epsilon and the difference in two-locus MLS statitistics (GEN-ADD and GEN-MUL) under each genetic model. The distribution of epsilon per 1000 replicates per model was often bimodal or skewed, therefore, I present the geometric mean of epsilon (Table 3.5). The estimates correlate with the expected value across the two-locus model ($\varepsilon = 0$ for additive and heterogeneity models, $\varepsilon = 1$ for multiplicative, and $\varepsilon > 1$ for epistatic models), however ε is greatly inflated under the null, and under single-locus models. To present the ε estimates in more detail, Figure 3.6 shows a histogram of the 1000 maximum likelihood ε estimates per genetic model. The majority of estimates (49% on average across all models) fell at the upper or lower bounds of the parameter space, that is, at $\varepsilon = 1000$ or $\log_{10}(\varepsilon) = 3$ and at $\varepsilon = 0.001$ or $\log_{10}(\varepsilon) = -3$ (Figure 3.6A). For the two-locus models the number of estimates that fell between $\varepsilon = 90$ and 999 was small (2% on average, but 19% for m0 and 6% for m1-m3) and Figure 3.6B presents the number of ε estimates that maximized between 0 and 90 in more detail.

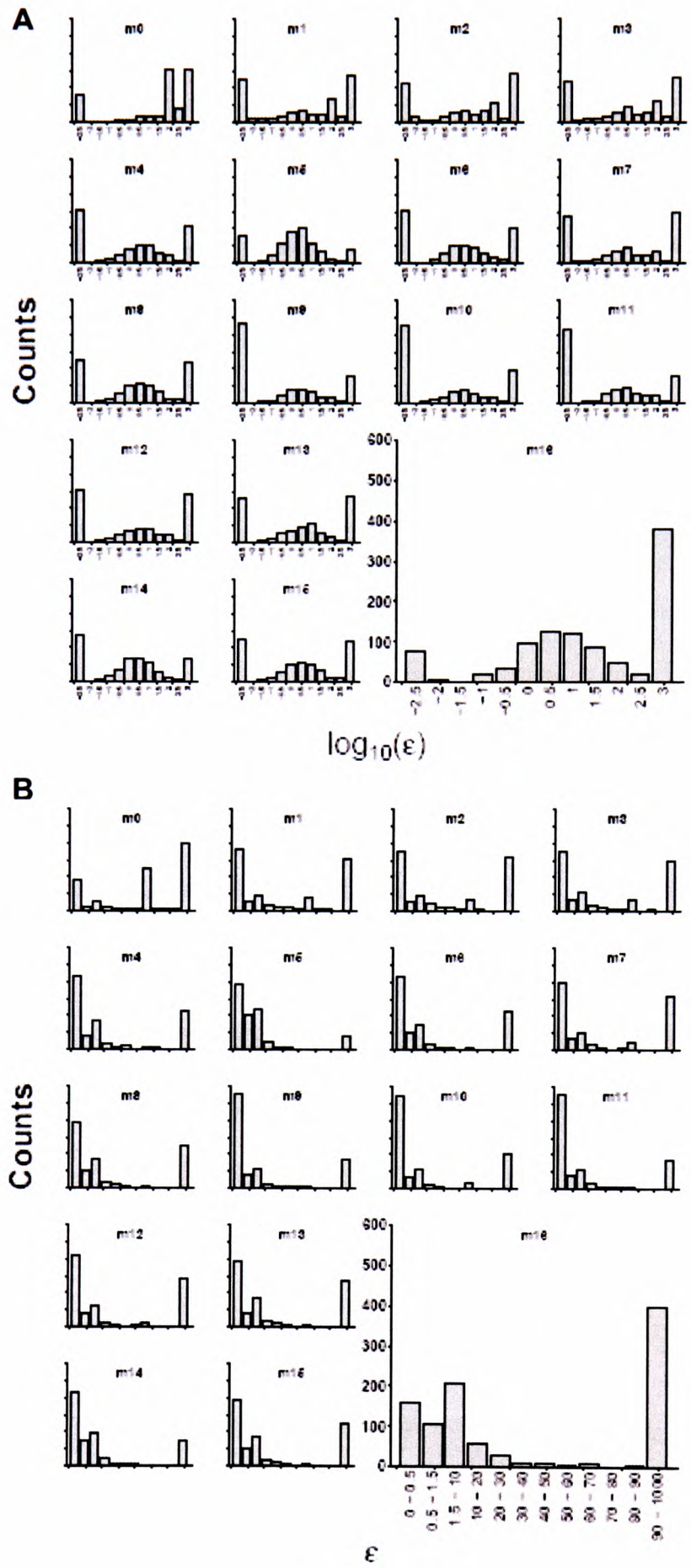


Figure 3.6. Histograms of the estimates of ε. (A) Histogram over the entire parameter range using $\log_{10}(\varepsilon)$ in bins of 0.5 $\log_{10}$ units each (the x-axis shows the upper limit of each bin, the first bin begins at -3 units; the y-axes for all plots are as shown for m16). (B) Histogram of ε with more detail in the 0-90 parameter value range (the x- and y-axes for all plots are as shown for m16).

The difference in two-locus MLS statistics was also obtained for each genetic model simulated (Figure 3.3B). The difference statistics GEN-ADD and GEN-MUL correlate well with two-locus model simulated and have very low values under the null and under single-locus models. For GHP, I chose to present the Spearman's correlations per genetic model simulated. Overall, the number of replicates with significant familial correlations is quite low. However, there are more replicates with significant negative familial correlations for the heterogeneity and additive models (m8-m11), and there are more positive familial correlations for model m16. The same trend is observed for the proportion of overall negative and positive familial correlations (significant or not), and the numbers in that case are much greater. For IBDreg I first examined the overall β for all replicates per model, and there is a noticeable trend. The total β was around zero for the null, for single-models and for multiplicative two-locus models as expected. For the additive and heterogeneity models the total β was around -0.15, for m16 it was 0.34, and for models m12-m15 it was around 0. Next, for replicates which have significant evidence for interaction the β estimates were -0.6 for additive/heterogeneity models and 1 for m16 and these follow Holmans' (2002) observation that $\beta > 0$ can be taken as evidence for epistasis, and $\beta < 0$ as evidence for heterogeneity.

3.4 Discussion

The aim of this chapter was to compare power of different multilocus linkage approaches to detect interaction and localize susceptibility loci in the presence of epistasis and heterogeneity. The results vary according to two-locus model and method applied, but overall the model of strong epistasis (m16) fared well regardless of the mapping approach. The method which stratified the sample appears to have less power than methods which analyzed the unstratified sample across all models.

Finally, a range of parameter estimates was obtained specific to the certain genetic model simulated.

The thirteen two-locus models were examined using the parametric two-locus linkage method to determine power under the optimal case scenario. The conclusion was that most models had very similar high power estimates, with the exception of additive models, which had slightly lower power.

The results from Merloc under the general two-locus test statistic (GEN) supported these findings. The appeal of Merloc is that it provides a joint linkage statistic for a simultaneous test of the involvement of two loci, rather than examining each locus at a time. However, the joint test statistics should be interpreted with caution, because for single locus models the two-locus joint statistics are inflated as shown by the false-positive results (Table 3.4). This suggests that if one searches across all pairs involving at least one locus with strong marginal evidence for linkage, the joint two-locus MLS will be significant even if there is no real second locus present. A similar observation concerning the performance of a different two-locus test in a simultaneous genome-wide scan strategy was made by Dupuis et al. (1995). However, in such cases the GEN-SL1 statistic can be used to confirm the presence of the secondary locus. The Type I error rates for GEN-SL1 under single-locus models are reasonable (6%-8%), but in future should be re-assessed using single-locus effect null models. Therefore this statistic should provide a good indication of whether there are in fact two loci underlying the GEN interaction, or whether the GEN result is a false-positive finding. It is surprising that for the models selected in this study the power to detect linkage to the second locus in the presence of interactions (GEN-SL1) is not much greater than just using the single-locus MLS at locus 2 (SL2, Figure 3.3C). This finding may be explained by the fact that most of the simulated models in

this study have strong marginal evidence for linkage at one or both loci (Table 3.1), or by the fact that we are using a null genetic model as the null. As discussed in chapter I, there have been contradictory findings on the amount of increase in power to localize gene when two-loci are considered. In this study there is evidence for an increase in power for all models considered when interactions are taken into account, but the increase appears to be only marginal.

The difference in two-locus MLS statistics gave expected averages under the two-locus models simulated (Table 3.5). However, the precision of these estimates could perhaps be improved in future studies. Figure 3.3B shows that on average in 18% of cases GEN-ADD surpassed the 0.05 level threshold to identify significant non-additive models when in fact the simulated model was additive. The GEN-MUL performed worse, identifying significant difference from a multiplicative model in 32% of multiplicative model replicates. These results could improve if I had only examined GEN-ADD and GEN-MUL in the proportion of replicates that had significant GEN scores of the total 1000 replicates (similar to the results for GHP in Figure 3.4B). However, in practice for most of the simulated models the majority of replicates had significant GEN scores. These results suggest that one should only use the difference in MLS scores after there is a significant evidence for the involvement of two loci. Alternatively, one may need to re-assess these results using a different null hypothesis specifically modeling additive or multiplicative two-locus effects.

The epsilon-epistatic model (EPS) was introduced in chapter II as an alternative to the general two-locus model. The results from chapter II indicated that the epsilon-epistatic model approximated the general-model well, had fewer free parameters than the general model, and provided a maximum-likelihood estimate of the degree of epistasis. Therefore, the epsilon-epistatic model seemed a suitable, and

perhaps more promising, choice over the general model for an initial two-locus analysis of genome data. However, in this chapter the properties of this model were examined in detail and indicated some advantages, but also some important drawbacks of the model. The main advantage of the EPS model is that appears to fit the general model well under both the null hypothesis, m0, under single-locus models m1-m3, and under the alternative hypothesis of two-locus effects across all two-locus models examined in this chapter (m4-m16). These results from the single-point simulations indicated that EPS is a suitable substitute for GEN. However, because obtaining theoretical thresholds for the test-statistics presented in this thesis is difficult (as discussed in section 2.2.4), in the analysis of genome-scan data significance is declared by using simulation thresholds. Therefore, in the analysis of genome-data the advantage of EPS approximating well the more complex model GEN may be lost. Another important advantage of the EPS model is that it can provide a maximum-likelihood estimate of the strength of epistasis, or the value of ε. However, the results from this chapter indicate that ε is not a very reliable estimator of the degree of epistasis. Epsilon appears to perform moderately well under the two-locus simulates, but has inflated estimates under the null. For the null model, m0, and single-locus models, m1-m3, a sizeable proportion of estimates (30% for m0, 10% for m1-m3) maximized at a specific value of $\varepsilon = 65$. In addition, a high proportion of estimates maximized at the boundary of the parameter space across all models simulated. Both of these maximization points indicated potential convergence problems. As discussed in chapter II, the maximization procedure was examined in detail in this study. Changing the scale to aid the maximization procedure did not improve the results. On the other hand, when we examine just the two-locus models, overall, the estimates of epsilon are generally consistent with the penetrance structure

simulated. In general there were more estimates of ε near 0 for the heterogeneity and additive models, and fewer estimates near 0 for models m5 and m16. Therefore, in data analysis it may of interest to obtain the maximum-likelihood estimate of epsilon, but this finding must be compared to the difference in MLS statistics, and only if both results are consistent can a conclusion about the presence/absence of epistasis be obtained. In summary, while the fit of EPS compared to GEN is very good, the maximum-likelihood estimates of ε may be inconsistent and unreliable. Overall, it is difficult to draw a definite conclusion about whether the EPS model constitutes an significant improvement over the GEN model.

Based on the Merloc results it seems that an appropriate strategy to use Merloc would be to first scan across the regions using the two-locus GEN statistic because it has high power estimates. Once significant pairs of regions are identified using GEN, one should use the difference in MLS, GEN-SL to confirm the presence of both loci in the pair. Next one can obtain the maximum likelihood estimate of epsilon, and GEN-ADD or GEN-MUL, to identify the best-fitting model underlying the interaction. The difference in two-locus MLS should support epsilon in order to draw a conclusion about the most likely underlying two-locus model.

It is not straightforward to compare the results from GHP to the rest of the two-locus methods because GHP is based on analysis in stratified samples. In our simulations of 100 ASPs under two-locus models, the entire sample is linked to both genes, and so for models of epistasis there is nothing to be gained by further stratifying the sample. For models of heterogeneity one would expect an increase in power to detect the loci. Leal and Ott (200) have also looked at power to detect linkage in ASPs stratified according to IBD state, compared to unstratified samples, and have similar findings. The performance of this approach is also likely to be

affected by the sample size, and 100 ASPs may not be a large enough sample to detect significant effects.

The IBD regression results have power estimates that fall between those of Merloc and those of GHP. The two test statistics that measure linkage in the presence of interaction, IBDreg T and Merloc GEN-SL, have very similar patterns across the two-locus trait models. The parameterization of the IBD probabilities is not as complex in IBDreg as it is in Merloc; however, IBDreg has fewer parameters overall and may perform better than Merloc for some subset of models. In addition, this method may be used with genotype data, or with genotype-by-environment interaction data. Another approach which is related to IBDreg, is the method implemented in Lodpal, which models the multilocus IBD probabilities in a conditional logistic likelihood incorporating multiple covariates, genetic or environmental. This method may be considered joint multi-locus analysis, and should be compared to Merloc in future studies.

The two-locus models used in this study were selected to span the range of possible models well and possibly fall at the edges of that parameter space, but it was difficult to assess how well the range of epistatic models was covered. From the analysis results the heterogeneity and additive models and model m16 stand out in the power and parameter estimates across most methods used, and the two asymmetric models, m7 and m12 have higher power estimates in IBDreg. However, for the majority of models (m4-m7, m12-m15) the results are very similar. The implication is that these models may not be very different from each other, even if the penetrance structure appears quite different. Perhaps epistasis can be more appropriately defined using approaches other than the two-locus penetrance structure.

This was further examined to see whether the classification of epistatic models used in this study is the most appropriate approach to categorizing two-locus epistasis. To investigate this, the two-locus genetic model structures were examined from a different perspective. Let us suppose that the two-locus models represent quantitative trait genotype means, rather than discrete trait penetrances (Sham 1998; Culverhouse et al. 2002). One can then obtain the components of the genetic variance attributed to the two loci, and examine the proportion of the genetic variance attributed to the epistatic variance components. I followed the example of Sham (1998) to obtain the variance components for the 13 two-locus models used this chapter (for details and formulation see Appendix A). The results for the variance components are presented in Table 3.6.

Table 3.6. Variance components from 13 two-locus models.

Model	Penetrance matrix			*f*	Allele frequency	Variance Components V_A	V_D	V_{AA}	V_{AD}	V_{DD}	V_I[a]	V_I/V_G
Multiplicative												
m4	*0*	*0*	*0*	0.390	0.720	0.017	0.003	0.007	0.003	0.000	0.009	0.317
	0	*0*	*0*									
	0	*0*	*f*									
m5	*0*	*0*	*0*	0.580	0.695	0.008	0.022	0.001	0.008	0.011	0.020	0.405
	0	*f*	*0*									
	0	*0*	*0*									
m6	*0*	*0*	*0*	0.390	0.305	0.017	0.004	0.006	0.003	0.000	0.009	0.318
	0	*f*	*f*									
	0	*f*	*f*									
m7	*0*	*0*	*0*	0.380	0.575	0.017	0.007	0.002	0.002	0.001	0.005	0.166
	0	*0*	*0*									
	0	*f*	*f*									
Heterogeneity												
m8	*0*	*0*	*f*	0.350	0.390	0.016	0.012	0.000	0.000	0.000	0.000	0.009
	0	*0*	*f*									
	f	*f*	$2f-f^2$									
m9	*0*	*f*	*0*	0.310	0.905	0.019	0.005	0.000	0.000	0.000	0.000	0.008
	f	$2f-f^2$	*f*									
	0	*f*	*0*									
Additive												
m10	*0*	*¼f*	*½f*	0.950	0.110	0.022	0.000	0.000	0.000	0.000	0.000	0.000
	¼f	*½f*	*¾f*									
	½f	*¾f*	*f*									
m11	*0*	*f*	*0*	0.290	0.900	0.019	0.005	0.000	0.000	0.000	0.000	0.000
	f	*2f*	*f*									
	0	*f*	*0*									
Epistatic effects												
m12	*0*	*0*	*0*	0.390	0.575	0.018	0.006	0.000	0.003	0.002	0.006	0.188
	0	*0*	*f*									
	f	*f*	*f*									
m13	*0*	*0*	*0*	0.400	0.465	0.019	0.002	0.003	0.002	0.006	0.010	0.329
	0	*0*	*f*									
	0	*f*	*f*									
m14	*0*	*0*	*f*	0.450	0.365	0.013	0.012	0.003	0.005	0.002	0.011	0.300
	0	*0*	*f*									
	f	*f*	*0*									
m15	*0*	*f*	*0*	0.390	0.915	0.015	0.003	0.007	0.003	0.000	0.010	0.356
	f	*0*	*f*									
	0	*f*	*0*									
m16	*0*	*0*	*f*	0.720	0.745	0.014	0.002	0.005	0.014	0.011	0.030	0.651
	0	*½f*	*0*									
	f	*0*	*0*									

[a] The epistatic variance, V_I, is the sum of the epistatic variance components

The proportion of the genetic variance attributable to epistasis (V_I/V_G) is highest for m16, then m5, then the remaining symmetric multiplicative models (m4, m6) have similar estimates to those of the other epistatic models (m13, 14, m15), after which the asymmetric models (m7 and m12) have similar estimates, and finally that proportion is close to zero for the heterogeneity and additive models. These results

relate to results from the power analyses for the linkage methods, especially for test statistics GEN-ADD and GEN-SL1 in Merloc, and T statistics in IBDreg.

In the simulations of the genetic models I originally started simulating across a range of population prevalences (K = 0.01, 0.05, and 0.1) and λ_s values (λ_s = 1.5, 2, 4, and 10) for the generic models. However, for several of the models the upper bound of the sibling relative risk is restricted (Rybicki and Elston 2000), and therefore I presented results from one case only (λ_s = 2, K = 0.1). It may be of interest to compare results across different genetic effect sizes and prevalence rates. In addition, it may be too simplistic to use just one parameter, λ_s, to describe the genetic effect size (see chapter IV for discussion).

There are several directions in which this study can be extended. First, simulations could be performed under different two-locus epistatic models. One might simulate the penetrance structures used in this study, but with unequal disease allele frequencies. One may also simulate the 'no main effect' epistatic models as discussed by Culverhouse (2002). Second, it is of interest to simulate multiple markers in a disease locus region and examine power to localize the disease gene and estimate the appropriate support intervals for localization of disease genes. Similar to single-locus analyses (Roberts et al. 1999), one could examine the sensitivity of the methods to localize genes within a given interval and investigate the appropriate Lod unit cutoffs to approximate a 95% confidence interval for localization, using the Lod-unit support intervals (further discussion in chapter IV). Third, one could examine how missing data affects power. I have started analyzing these data and the preliminary results indicate that missing parental genotypes greatly affect power estimates in Merloc. Finally, there are other methods that can be applied to our data, but results from which are not presented in this chapter. These methods are: ordered-subset analysis (Hauser

et al. 2004) which is similar to the conditioning approach in GHP, NPL regression (Langefeld et al. 2001) which is similar to the IBD regression in IBDreg, Lodpal (Olson 1999) which is similar to Merloc but with different parameters, and finally GeneFinder (Liang et al. 2001) - a generalized estimating equation approach. I tried to use GeneFinder, but unfortunately could not get the method to converge for the simulations assuming a single marker next to the disease locus. Since the goal of GeneFinder is to identify the most likely position of the disease locus in a marker map, clearly it is suited for simulations with multiple markers per disease gene region.

CHAPTER IV. Interaction between type 2 diabetes susceptibility loci on chromosomes 1q21-q25 and 10q23-q26

4.1 Introduction

Type 2 diabetes (T2D) is a common complex genetic disease of glucose homeostasis, characterized by insulin insensitivity. There is evidence for a substantial genetic component in the trait confirmed by results from genome-wide linkage scans. Two susceptibility loci for T2D have been mapped to chromosomes 1q21-25 (Ehm et al. 2000; Vionnet et al. 2000; Wiltshire et al. 2001) and 10q23-26 (Ghosh et al. 2000; Vionnet et al. 2000; Wiltshire et al. 2001). Both loci contain potential candidate genes and are subjects of extensive fine scale mapping projects.

The evidence for epistasis between these two loci was examined in this chapter in order to characterize interactions between them and to facilitate fine-scale mapping of the underlying genetic variant(s). Two ethnically and demographically matched samples from Britain (Wiltshire et al. 2001) and France (Vionnet et al. 2000) - the only two studies that show evidence for linkage of T2D itself to both loci, were investigated. Two model-free approaches, joint two-locus linkage and conditional linkage analysis, were applied to determine the presence of interactions. The results from the model-free analyses directed a search of parametric two-locus models to define a range of plausible two-locus penetrance models consistent with the data, and provided a reduction in the support intervals for localization of both loci.

4.2 T2D data sets

The subjects for study comprised the 573 full-sib pedigrees analysed in the British Warren 2 type 2 diabetes genome scan (Wiltshire et al. 2001) and the 147

pedigrees previously analysed in the French genome scan for T2D susceptibility (Vionnet et al. 2000). In the present study access was provided to the British sample of 573 families and to the combined British - French sample of 721 families, in which one French family was split in two units because of computational constraints.

In the British study, chromosome 1 had previously been genotyped with 35 microsatellite markers, and the linkage peak on 1q subsequently fine mapped with an additional 17 microsatellites giving a maximum Lod score (Kong and Cox 1997) of 1.98 at D1S2799. Chromosome 10 was analyzed with 22 microsatellites in the primary scan giving a maximum Lod score of 1.99 between D10S1765 and D10S185. A conditional linkage analysis yielded preliminary evidence for epistasis between these two loci (Wiltshire et al. 2001). As part of the present study, the British data were augmented by genotyping a further 6 microsatellites at the 1q21-25 locus, and a further 5 at the 10q23-26 locus, increasing inheritance information from the data.

In the French study, chromosome 1 was genotyped with 31 microsatellite markers (all in common with the British genome scan) together with an additional 16 fine mapping markers around the peak of linkage. These analyses yielded a peak maximum likelihood binomial (MLB) Lod of 2.99 (and a peak maximum Lod score (MLS) of 3.04) at marker APOA2 in a subset of lean patients. Chromosome 10 was genotyped with 23 markers (20 in common with the British study), yielding a peak MLB-Lod of 1.59 (and an MLS=1.24) at D10S1655 in the whole sample of 147 pedigrees.

In the present study, the evidence for epistasis was examined first in the 743 all possible affected sib pairs (meaning all pairs drawn from sib-ships of 2 or more affected individuals) in the British dataset, and subsequently in the combined British-French dataset (containing 1196 all possible affected sib pairs). In both analyses

genetic distances were based on the Marshfield genetic map (Broman et al. 1999), for consistency with the previous fine scale mapping projects for both populations. The order of the markers in the genetic map was consistent with the physical order obtained from the May 2004 release of UCSC (http://genome.ucsc.edu/). In the British sample, genotypes were available for 58 microsatellite markers on chromosome 1 (average marker density of 5 Kosambi cM and peak information content of 0.74), and 27 microsatellites on chromosome 10 (average marker density of 6.2 Kosambi cM and peak information content of 0.68). In the combined British-French sample, genotypes were available for 72 microsatellite markers on chromosome 1 (average marker density of 4 Kosambi cM and peak information content of 0.76), and 30 microsatellites on chromosome 10 (average marker density of 5.86 Kosambi cM and peak information content of 0.7). In the analyses of the combined British - French pedigree dataset, markers that were genotyped in both populations were included twice in the map and given a nominal separation of 0.001cM, with British pedigrees coded missing for the French marker, and vice versa. The rationale was that the two data sets were genotyped on different platforms, by different people, at different times and attempts at reconciling the allele calls between both sets of results were not successful, neither was a systematic comparison of allele calls attempted between the British and French data.

4.3 Analysis

4.3.1 Two-locus analyses

Joint two-locus analysis

Joint two-locus multipoint linkage analysis was performed in the British and British-French combined pedigree datasets using Merloc. The joint two-locus IBD

probabilities were calculated for every pair of markers, the first on chromosome 1 and the second on chromosome 10, to create a two-dimensional grid of linkage coordinates. For the regions on 1q and 10q that spanned the fine-mapped linkage peaks from earlier single-locus multipoint analyses (Wiltshire et al. 2001) - 1q (D1S498 – D1S213) and 10q (D10S537 – D10S217), joint IBD were calculated on top of each pair of markers and at 3 equidistant locations between markers, the first on 1q and the second on 10q.

The evidence for epistasis was examined initially in the British dataset alone in a multipoint analysis with Merloc. The locus on 1q was a broad peak spanning markers D1S196 to D1S218, and achieved a peak MLS statistic of 1.59 (Figure 4.1). It should be noted that Merloc weights all sib-pairs equally in the calculation of the MLS, unlike Allegro (used to generate the single locus linkage evidence in Wiltshire et al. (2001)). The locus on 10q23 achieved a peak single locus MLS statistics of 2.66 at D10S1765. The peak two-locus MLS of 5.49 under the unrestricted general epistasis model was found at D1S196 on 1q23 and 1cM from D10S1765 on 10q23 (Figure 4.1). The peak two-locus MLS obtained under a restricted additive-effects model (with no interaction terms) was 3.84 and under the multiplicative model the two-locus MLS was 4.18 at the same location.

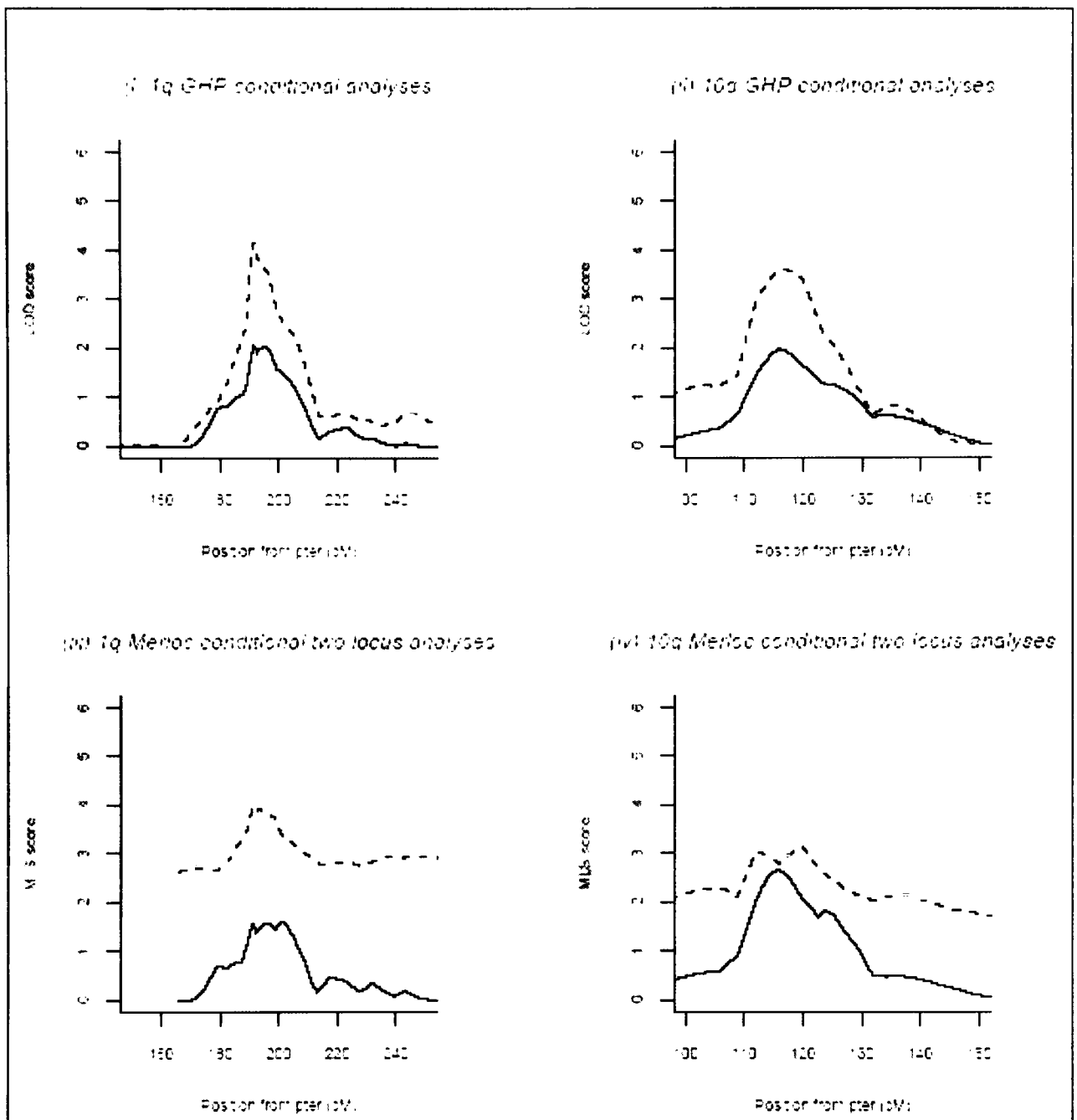


Figure 4.1. Two-locus results in the British data. Conditional linkage analysis in GHP of (A) 1q conditional on evidence for linkage at D10S1765, and (B) 10q conditional on linkage at D1S196. Analyses are shown using proportional weighting (dashed lines) and no weighting (or unconditional analyses, solid lines). Conditional two-locus linkage analyses in Merloc, using GEN-SL, of (C) 1q and (D) 10q under the general two-locus model MLS (dashed lines) at the position of the peak MLS on 10q (1cM from D10S1765) and along 1q (C), and at the position of the peak MLS on 1q (at D1S196) and a range of positions on 10q (D). Single locus MLS analyses of 1q and 10q are shown in solid lines.

To assess the significance of the difference in MLS calculated under different two-locus genetic models (i.e. the evidence for epistasis), simulations were performed under two separate null hypotheses. First, significance was assessed using simulation results from two fully informative, unlinked markers under the null hypothesis of no linkage to the trait and with no epistasis between them, in 100 affected sibling pairs using 100 000 replicates (results presented in chapter II, Table 2.3). Second, simulations were performed using two fully informative unlinked markers linked to the trait under two separate genetic models, for 100 affected sibling pairs. The entire sample of sibs from the British data may be used in these simulations, rather than simulating the null in just 100 affected sibling pairs. Although the number of sibpairs should not greatly affect the results, matching the number of sibpairs is straightforward, not computationally intensive, and should be performed in future analyses. To assess significance of the difference between the fit of the general and additive models 100 000 replicates of fully informative ASPs were generated with two loci acting under an additive model by sampling from the observed joint two-locus IBD distribution calculated under the additive model, at the position of the maximum two-locus Lod score, during the linkage analysis of the actual data (see Table 4.1). Similarly, to simulate loci acting together under the multiplicative model, fully informative ASPs were sampled from the joint two-locus IBD distribution calculated from the observed data under the multiplicative model. This was followed by analysis of 100 000 replicates in Merloc, under the general and multiplicative models.

Table 4.1. Maximum-likelihood estimates of IBD sharing in the British T2D data.

		IBD sharing probabilities at locus 2			
		0	**1 (maternal)**	**1 (paternal)**	**2**
General two-locus model					
IBD at locus 1	0	0.0545	0.0545	0.0545	0.056
	1 (maternal)	0.0561	0.0578	0.0578	0.0729
	1 (paternal)	0.0561	0.0578	0.0578	0.0729
	2	0.0576	0.0611	0.0611	0.1114
Multiplicative two-locus model					
IBD at locus 1	0	0.0501	0.0501	0.0501	0.0687
	1 (maternal)	0.0561	0.0561	0.0561	0.077
	1 (paternal)	0.0561	0.0561	0.0561	0.077
	2	0.0665	0.0665	0.0665	0.0912
Additive two-locus model					
IBD at locus 1	0	0.0506	0.0506	0.0506	0.0706
	1 (maternal)	0.0565	0.0565	0.0565	0.0765
	1 (paternal)	0.0565	0.0565	0.0565	0.0765
	2	0.0664	0.0664	0.0664	0.0864

The observed difference between the general and additive genetic models of 1.65 was significant at $P = 0.00021$ (using significant thresholds from chapter II, Table 2.3) and at $P = 0.00173$ (for simulations under a null hypothesis of two markers linked to the trait and sampled from the observed two-locus IBD distributions), providing evidence for epistasis between these two loci as defined as a departure from additivity. The difference of 1.32 between the general and multiplicative models was significant at $P = 0.00052$ (using significant thresholds from chapter II, Table 2.3) and P = 0.00293 (for simulations under a null hypothesis of two markers linked to the trait and sampled from the observed two-locus IBD distributions) and provides evidence for epistasis as defined as a departure from multiplicativity. The evidence for epistasis was further supported by the maximum-likelihood estimate of epsilon at the two-locus peaks (epsilon maximized at the upper parameter boundary, i.e. epsilon $\geq 10^3$). As pointed out in chapter III, the estimate of epsilon by itself does not provide reliable

evidence for epistasis. However, ε, along with the significant differences in MLS statistics GEN-ADD and GEN-MUL indicates significant evidence for epistasis.

Further support for the findings of epistasis in was obtained in the analyses of the combined British-French dataset (Table 4.2). The two single-locus peak MLS scores were 1.65 and 1.94 at D1S196 and 2cM from D10S1765, respectively. Comparison of the peak two-locus MLS score from the general model (MLS=4.66, at D1S196 and 2cM from D10S1765) with that from the additive model (MLS=3.20 at the same location) and from the multiplicative model (MLS=3.42 at the same location), provide significant evidence ($P<0.01$) for epistasis according to each definition. However, while the evidence for epistasis was sustained in the combined data, it was not enhanced, suggesting that the French data provide weak (if any) evidence for epistasis.

Table 4.2. Merloc results in the T2D data sets on chromosomes 1 and 10.

	Locus 2								
	D10S1686			**D10S1765**			**D10S1753**		
Locus 1	**GEN**	**ADD**	**MUL**	**GEN**	**ADD**	**MUL**	**GEN**	**ADD**	**MUL**
(A) British sample of 743 ASP									
D1S2681	3.89	2.88	3.07	4.41	3.49	3.69	4.14	2.86	3.08
D1S196	5.07	3.29	3.6	5.45	3.89	4.22	5.2	3.28	3.61
D1S2750	4.89	3.16	3.46	5.3	3.76	4.07	5.11	3.14	3.46
(B) British and French combined sample of 1196 ASP [a]									
D1S2681	3.07	2.5	2.6	3.76	2.98	3.11	3.75	2.49	2.65
D1S196	3.98	2.9	3.07	4.59	3.38	3.59	4.59	2.91	3.12
D1S2750	3.83	2.73	2.89	4.45	3.21	3.41	4.5	2.73	2.95

[a] Access to genome data for analyses in this thesis was only provided for the British sample and the combined British - French sample, but not the French sample by itself.

Conditional linkage analysis

The conditional linkage analysis approach (Cox et al. 1999) was applied to these data by Dr Steven Wiltshire from the Wellcome Trust Centre for Human Genetics at the University of Oxford, and the results presented in this section were

obtained by him. The approach was performed as described in chapter III, but using both discrete and proportional epistatic weights using NPL scores calculated in Allegro (Gudbjartsson et al. 2000). The discrete (01) weighting scheme was applied as described in chapter III and in the proportional (PROP) weighting scheme families with positive NPL scores carried their family-specific NPL score as their weight, and those with zero or negative NPL scores were assigned a weight of zero. The significance of a Lod score increase following discrete weighting (Δ EPI LOD) was determined by 10 000 permutations of the family specific discrete weights. Determining significance in the case of proportional weighting is not straightforward (family-specific PROP weights cannot meaningfully be permuted when the pedigrees are not all of the same structure) and so following Cox et al. (1999) *P*-values are not reported in this case.

Initial single-locus multipoint linkage analysis with Allegro on chromosomes 1 and 10 in the British pedigrees alone yielded maximum allele sharing Lod scores (Kong and Cox 1997) of 2.05 at D1S196 on 1q23 and 1.98 at D10S1765 on 10q23 (Figure 4.1). The family-specific NPL scores at these markers were significantly positively correlated (Spearman's ρ=0.136, P=0.0011). Conditional analyses were performed using weights calculated from the family-specific NPL scores at D1S196 and D10S1765. With a 0-1 weighting scheme the evidence for linkage at 1q23 conditional upon that at D10S1765 increased from Lod=2.05 to Lod_{01}=3.46 at D1S196 with P=0.0073 (for the increase). Using a proportional weighting scheme the peak increased to Lod_{PROP}=4.16 at D1S196. The reciprocal analysis (conditioning on the evidence for linkage at D1S196) saw the linkage signal at 10q23 increase from Lod=1.98 to Lod_{01}= 2.58 at D10S1765, with P=0.0401 (for the increase). Following analysis with a proportional weighting scheme, the Lod_{PROP} score increased to 3.59,

approximately 1cM from D10S1765. These statistically significant reciprocal increases in the Lod scores provide evidence for epistasis as defined as a departure from a multiplicative penetrance model according to Cox et al. (1999).

Similar findings were obtained in the combined British-French dataset. Family-specific weights were calculated at the maximally linked markers from the unconditional single locus linkage analyses. During the conditional analyses using these weights, the evidence for linkage at D1S196 on 1q23 increased to Lod_{01}=4.37 (P=0.0031) from an unconditional Lod score of 2.73; the evidence for linkage at D10S176 on 10q23 increased from Lod=1.47, in the unconditional analysis, to Lod_{01}=2.21 (P=0.0435). These confirmed the preliminary evidence of epistasis seen from the significant positive correlations (Spearman's ρ=0.128, P=0.0006) between the family specific NPL scores at the unconditional Lod score maxima at D1S196 and D10S1765.

4.3.2 Two-locus parametric model search

The results from section 4.3.1 indicated that two independent two-locus mapping approaches both provided significant evidence for an interaction between 1q and 10q in T2D in the British data. The findings suggested a model of epistasis in the data that is more complex than the simple formulation described by the multiplicative two-locus model. It is of interest to attempt to identify a set of two-locus parametric models that are consistent with the model-free two-locus results.

To identify a set of 'most-likely' two-locus parametric models a search through the space of plausible genetic models was conducted. Parametric two-locus models were defined as presented in chapter III. If two biallelic disease loci contribute to the trait, locus 1 and locus 2, two-locus parametric models can be defined in terms of the disease allele frequency at the first disease locus - q_1, the disease allele

frequency at the second disease locus - q_2, and the nine two-locus joint-genotype penetrances – f_{11} to f_{33}, as described in chapter III (Figure 3.1). Because the space of all possible two-locus models is too great, restrictions on the parameter values were imposed to reduce the complexity of the search. I only considered models for which the disease allele-frequencies at the two loci were set to the same value ($q_1=q_2$), and the nine penetrances were restricted to take two possible values - zero or a number (*f*) between 0 and 1 (for example for a single-locus recessive model at the first disease gene $f_{11}=f_{12}=f_{13}=f_{21}=f_{22}=f_{23}=0$ and $f_{31}=f_{32}=f_{33}=f$). From the results in section 4.3.1 it seemed highly unlikely that 1q and 10q act under a multiplicative or an additive genetic model. Therefore, the search through penetrance models could further be restricted to models which do not fit single-locus and two-locus additive and multiplicative structures.

There are 512 possible penetrance models with penetrances equal to 0 or *f* (listed by Li and Reich (2000) for $f=1$). If f_{ij} is the joint two-locus penetrance, and f_i and f_j are the corresponding single locus penetrance factors for genotypes *i* and *j* at two disease loci, equations 1.6-1.8 (chapter I) describe the relationships between these factors under three two-locus genetic models, the multiplicative, the additive, and the heterogeneity model (Risch 1990a). To re-iterate,

$$f_{ij} = f_i \times f_j$$

under the multiplicative model, and

$$f_{ij} = f_i + f_j$$

under the additive model, and

$$f_{ij} = f_i + f_j - f_i \times f_j$$

under the heterogeneity model. In the 512 parametric models the penetrances can take only one of two values (0 and *f*), therefore the majority of possible additive models

are not included in the 512 models. Of the 512 models 462 do not fit a single-locus penetrance definition or a two-locus multiplicative definition. Heterogeneity models were not excluded from the search.

From section 4.3.1 at the joint-two-locus analysis peak of 5.5 the maximum likelihood estimate of λ_S was 1.146 under the general two-locus model, assuming a population prevalence, K, of 0.05. I searched through the 462 models to find parameter values (of q_1, q_2, and f_{11} to f_{33}) would result in similar estimates of λ_S and K. Initially, the search through the 462 models was conducted for q_1=q_2=0.001, 0.05, 0.1,..., 0.5, by increments of 0.05 units, and through f = 0.1,..., 1 by increments of 0.1 units. Models were selected for parametric analysis if the population prevalence K was between 0.045 and 0.055, if λ_S was less than 1.2, and if the disease allele frequencies (q_1=q_2) were less than 0.51.

There were 175 models specified in terms of the penetrances and allele-frequencies that fit the population prevalence and sibling recurrence risk ratio specified. All of the selected models had f=0.1, and estimates of q_1=q_2 ranged from 0.25 to 0.5. In the 175 selected models, there were 57 unique penetrance structures. For each of the 175 selected two-locus models the two-locus parametric Lod score was computed in the British data using Genehunter-two-locus (GHT, Strauch et al. 2000). The parametric Lod score in GHT should be a multipoint Lod for a proper comparison to the multipoint two-locus MLS. However, GHT can only perform multipoint analysis on one chromosome and single point analysis at one location on the second chromosome. Therefore, GHT analysis was performed twice for each of the 175 selected parametric models. Initially, multipoint analysis was performed on 1q (in the region D1S498 – D1S213) assuming the second locus was at 18.6 cM on chromosome 10 (corresponding to D10S1765 at the 2D peak in Merloc), and then the

reciprocal multipoint analysis was performed on chromosome 10q (in the region D10S537 – D10S217) assuming that the second locus was at 26.32 cM on chromosome 1 (corresponding to D1S196 at the 2D peak in Merloc). The distribution of parametric Lod scores for the 175 models from both 1q multipoint – 10q singlepoint and 10q multipoint – 1q singlepoint analyses is shown in Figure 4.2. The majority of Lod scores cluster at the upper end of the Lod score values, showing that many parametric models fit the data well.

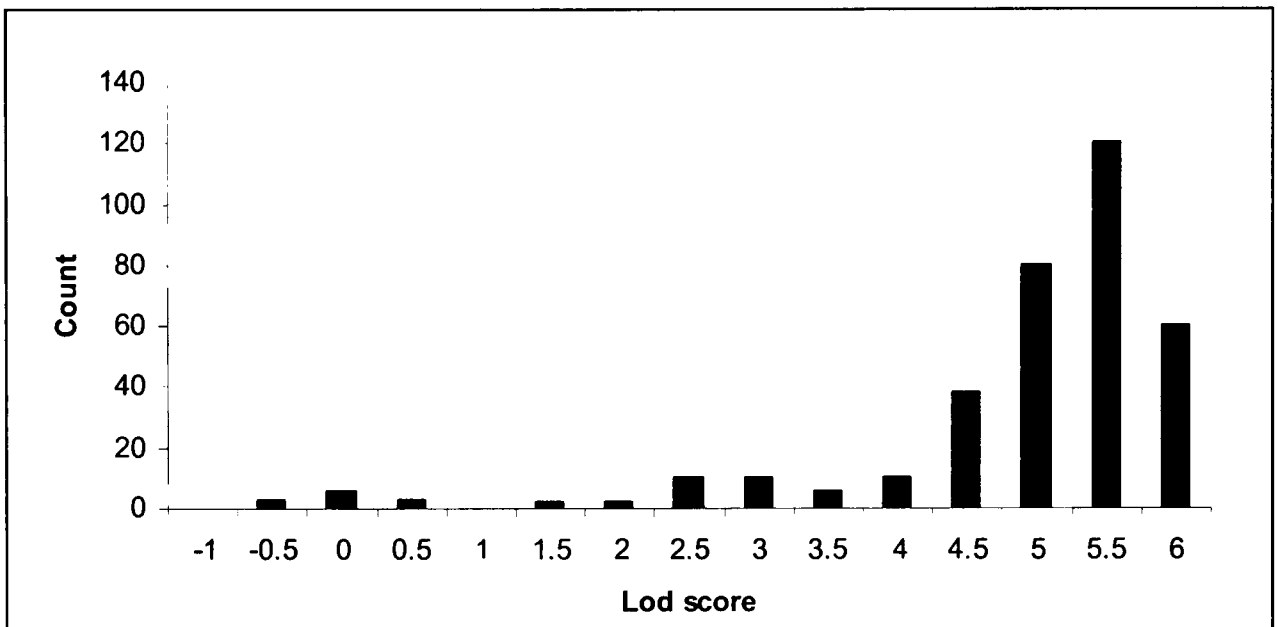


Figure 4.2 Histogram of the two-locus parametric Lod scores. The Lods were obtained for the 175 genetic models selected with $q_1=q_2<0.51$.

The 1-Lod-unit interval was used to select a range of models most compatible with the interaction observed in Merloc. There were 113 parametric models that had a two-locus parametric Lod score in GHT surpassing 4.45 for both multipoint analysis on 1q (and single-point on 10) and multipoint analysis on 10q (and single-point on 1). The 113 parametric models had 40 unique penetrance structures presented in Figure 4.3. There appear to be two distinct groups of models (or penetrance structures): the first comprises structures 132-198 (Figure 4.3) with defining characteristics $f_{11}=f_{22}=0$

and $f_{12}=f_{21}\neq 0$, and the second group of penetrances 231-296 (Figure 4.3) has the opposite characteristics $f_{11}=f_{22}\neq 0$ and $f_{12}=f_{21}=0$.

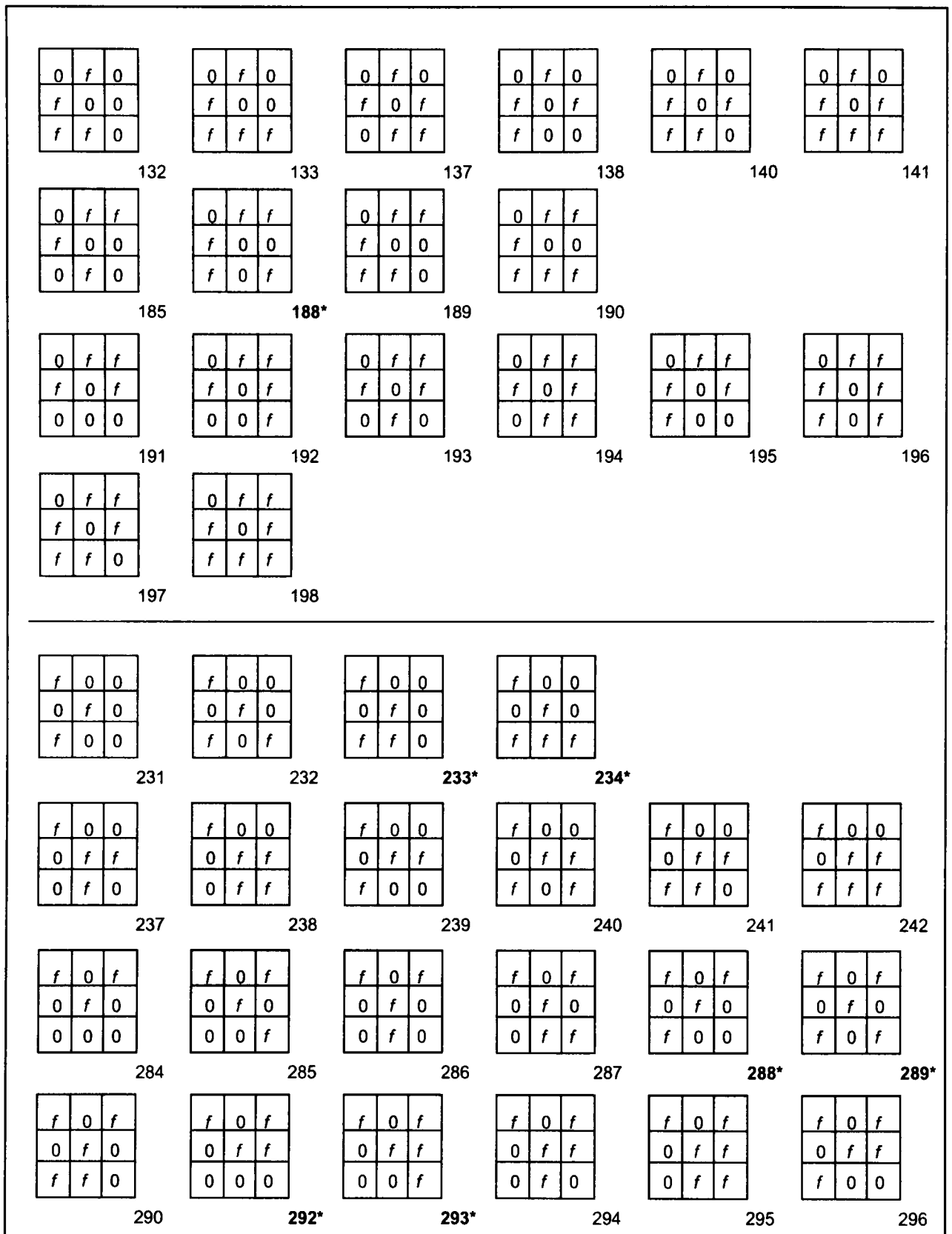


Figure 4.3. Penetrance structures of the 113 parametric models with Lod > 4.45. Seven structures (marked with asterisks in bold typeface) give models which produce parametric Lod > 5.45 with both 1q and 10q multipoint calculations.

The maximum parametric Lod scores obtained in multipoint analysis of either 1q or 10q were 5.97, however, the reciprocal multipoint analysis of 10q or 1q, respectively, did not results in Lod > 5.45. There were 9 parametric models in which the parametric Lod scores for both multipoint analysis on 1q (single-point at 10) and multipoint analysis at 10q (single-point at 1q) surpassed 5.45. These 9 models had 7 unique penetrance structures, 188, 233, 234, 288, 289, 292, and 293 (high-lighted with asterisks in Figure 4.3). The highest parametric Lod scores were obtained for models 288 and 289 with $q_1=q_2=0.25$ and $q_1=q_2=0.3$, resulting in two-locus parametric Lod scores of 5.6.

The aim of this section was to identify a range of 'most-likely' parametric models that were consistent with the results obtained from Merloc and models 188 ($q_1=q_2=0.4$), 233 ($q_1=q_2=0.3$), 234 ($q_1=q_2=0.35$), 288 ($q_1=q_2=0.25$ and $q_1=q_2=0.3$), 289 ($q_1=q_2=0.25$ and $q_1=q_2=0.3$), 292 ($q_1=q_2=0.3$), and 293 ($q_1=q_2=0.35$) appear to give the best fit.

4.3.3 Localization support intervals

In linkage analysis of a given chromosomal region the maximum Lod score can be used as a point estimate of the most likely location of the susceptibility gene. However, sampling error will affect the location estimate relative to the true gene location, and an interval estimator should be used to indicate the presence and magnitude of the sampling error. Interval estimators are generally referred to as confidence intervals in statistics and the probability that a confidence interval will contain the true location is the confidence coefficient (often set at 95%). Support intervals provide an estimate of the statistical confidence region of the linkage statistic (Conneally et al. 1985). Although there can be some imprecision in the localization of genes with multipoint linkage scores (Roberts et al. 1999), such support intervals

nevertheless provide one means of narrowing down the chromosomal region within which to search for the underlying genetic variant(s). In linkage analysis the 1-Lod-unit support interval is often used, corresponding roughly to a 90% or greater confidence region (Dupuis and Siegmund 1999; Ott 1999) depending on the marker density. Other Lod unit cut-offs may be applied, for example a 1.5-Lod-unit intervals corresponds roughly to a 95% confidence region in a dense map of markers (Dupuis and Siegmund 1999), and a 3-Lod-unit cutoff has been suggested (Ott 1999) and is occasionally applied (Craig et al. 1998; Busfield et al. 2002).

One-Lod-unit support intervals were obtained in the British data for both the joint two-locus analysis and the conditional analyses. In the joint two-locus analysis intervals were calculated using the peak two-locus linkage evidence at the 1q and 10q loci, under the general epistasis model. On the chromosome 1, this interval was 16.3cM (falling from 32.9cM during the single locus MLS analysis); on the chromosome 10, the support interval was 12.3cM, falling from 16.0cM during the single locus MLS analysis. The 1-Lod support intervals for the conditional linkage at D1S196 dropped to 8.8cM (with PROP weighting) and 10.0cM (with 0-1 weighting) from the 19.8cM seen during the unconditional analysis of chromosome 1. The 1-Lod support interval for the conditional linkage at D10S1765 dropped from 19.0cM seen during unconditional analysis, to 11.9cM (with PROP weighting) and 16.4cM (with 0-1 weighting).

4.4 Discussion

In this chapter the relationship between T2D susceptibility loci on chromosomes 1q21-25 and 10q23-26 was examined in detail. Both loci were supported by substantial evidence from genome-wide genetic studies. Joint two-locus linkage analysis of British and French pedigrees resulted in significant evidence for

epistasis between these two loci according to two genetic definitions – deviation from additivity and deviation from multiplicativity. The conditional linkage analysis, which provides a test of epistasis in terms of the latter definition only, yielded significant evidence also. The findings increase the probability that these loci represent genuine genetic effects and suggest a model of epistasis that is more complex than the formulation described by the multiplicative two-locus model.

An attempt was made to identify a set of most-likely genetic models consistent with the effects observed in the data. This approach may give an indication of the underlying biology of the interaction. Many assumptions were made to speed up the search through genetic models, some of which may be unrealistic (in particular the lack of phenocopies in the parametric genetic models). In future, one might wish to explore more computationally efficient methods to search a larger space of models, for example using simulated annealing (e.g. Dietter et al. 2005).

It is of interest to note that model 289 or the 'snowflake' model (sometimes also the 'checkerboard' model) was among the models which gave the best fit. The properties of this model have been previously examined for linkage and association test statistics. It is claimed that this model represents an extreme degree of epistasis (Purcell and Sham 2004) and is often selected for study because of its mathematical simplicity (Holmans 2002; Culverhouse et al. 2004). The estimates of variance components for model 289 were calculated following the method used in appendix A and presented in chapter III. If the genotype penetrances in the two-locus models represent quantitative trait genotype means, rather than discrete trait penetrances, one can obtain estimates of the components of the genetic variance attributed to the two loci. The variance components estimates from model 289 were very similar to those obtained at the peak in the British data in the analysis with Merloc under the general

two-locus model (Table 4.3), in contrast to estimates obtained under the multiplicative and additive models at the peak.

Table 4.3. Variance components estimates in the British T2D data.

Variance component	MLE at 1-10 peak in British data[a]			Estimates from model 289
	GEN	MUL	ADD	
$V_{A1} + V_{A2}$	0.054	0.048	0.329	0.047
$V_{D1} + V_{D2}$	0.025	0.766	0.671	0.071
V_{AA}	0.124	0.000	0.000	0.141
$V_{AD} + V_{DA}$	0.417	0.179	0.000	0.424
V_{DD}	0.380	0.007	0.000	0.318

[a] Maximum-likelihood estimates of the variance components obtained at the peak in the British data under 3 two-locus genetic models - the general (GEN), the multiplicative (MUL), and the additive (ADD).

The two-locus parametric model-search approach took a two-stage design where first the effect size was estimated in the data, and then the fit of different models was tested in the same data. The genetic effect size estimate (λ_s) was only used to select a range of models to carry on to the next stage of the test. Ideally, the variance in the maximum-likelihood estimate of λ_s should be allowed for, for example using a bootstrap estimate or the jackknife resampling procedure. This was not pursued in the current study, and instead I chose to include all values less than the estimated λ_s and those that fell 0.05 units above it. However, a resampling estimate of the variance can be obtained and should be taken into account in future analyses.

I think that one potential limitation of the parametric model search was to use λ_s to estimate the size of the genetic effect. The overall λ_s for a trait may not be a reliable estimator of the genetic effect size, because it is a measure of familial aggregation (which includes genetic and environmental effects), it is sensitive to ascertainment bias (Guo 1998), and does not predict the success of genome-wide linkage scans (Altmuller et al. 2001). However, the relative risk attributable to each

locus (or pair of loci as in this study) along with the interaction model among loci should relate to the power to detect susceptibility variants (Risch 1990b). Nevertheless, it seems that the locus-specific λ_s does not relate very well to another measure of genetic effect size, the genotype recurrence risk (Rybicki and Elston 2000). In general, it may be too simplistic to use a single parameter (λ_s) to summarize the overall genetic effect at a locus, because the exact relationship between the magnitude of linkage and λ_s is a function of the underlying genetic model or expected IBD sharing (Cordell 2001). Still, as mentioned above λ_s is only used in this chapter to restrict the number of models tested and therefore including a wide range of λ_s values to select parametric models should not affect the results.

Support intervals for the two-locus regions were obtained using the 1-Lod-unit method and showed a reduction in interval length from the single-locus results. However, it is not clear how to appropriately interpret these results with respect to the statistical confidence of localization of the susceptibility variants. In single-locus analysis for a single marker test with no dominance a 1-Lod unit corresponds to 4.6 chi-square units, and an asymptotic significance level of 0.03 (for a 1 degree of freedom chi-square distribution) giving a 97% confidence interval of localization. For genome-wide single-locus analysis this coverage probability drops to about 90% depending on the density of the markers and strength of the genetic signal (Dupuis and Siegmund 1999). Similarly, in the conditional two-locus analysis (Cox et al. 1999) a series of single-locus tests are performed in stratified data, and so the 1-Lod unit interval should correspond roughly to a 90% confidence region. However, the joint two-locus test is asymptotically distributed as a chi-square with a mixture of degrees of freedom (up to 8), as discussed in chapter II. If we suppose that it is distributed as chi-square with 2 degrees of freedom, then 1-Lod unit or 4.6 chi-square

units results in significance level of 0.1 and a confidence interval of 90% for localization in a single two-locus test. Genome-wide two-locus analysis should reduce this coverage probability similar to the single-locus genome-wide reduction. Hence, a 1-Lod unit interval does not result in equivalent coverage probabilities for the single-locus and joint two-locus analyses, and the probability is likely to be lower in two-locus analysis. Therefore, the reduction in support intervals from the two-locus analyses is not straightforward to interpret. This topic should be addressed in more detail in future studies empirically, or by simulation, or using the theoretical approach of Dupuis and Siegmund (1999) to derive an expression for the coverage probability and the expected length of support intervals in joint two-locus analysis.

The results from this study are pertinent to the ongoing fine-scale mapping of the 1q and 10q loci in several regards. First, parametric joint two-locus linkage analysis was used to identify a set of 'most-likely' two-locus genetic models consistent with the genetic effect sizes seen in our data. These allow more informed, and potentially more powerful, SNP-based association analyses of the 1q and 10q loci. Second, the joint and conditional two-locus analyses yielded narrower 1-Lod support intervals for the linkage evidence on chromosomes 1q and 10q. This provides a potential refinement of the regions for detailed search during LD-mapping, but it is unclear what degree of confidence the 1-Lod-unit intervals correspond to in the joint two-locus case. Finally, the findings suggest a two-locus extension of the single locus approach of assessing the contribution a putatively causal SNP makes the evidence for linkage (Li et al. 2005). It follows that the contributions by candidate SNPs to the linkage evidence at a pair of interacting loci should be assessed jointly, consistent with the epistatic model established during the two locus linkage analysis and confirmed (or refined) during the subsequent association analyses of SNP data.

CHAPTER V. Two-dimensional genome scan of hypertension

5.1 Introduction

Hypertension is the pathological elevation of arterial blood pressure and is a modifiable risk factor for cerebrovascular and coronary heart disease. Genetic factors are implicated, but estimates of the risk ratio to siblings of affected individuals are modest (1.5 to 3.5) and heritability estimates, which vary between populations of differing ancestries and environments, range from 30% to 50% in the UK (Ward 1990). The clinical importance of essential hypertension and the desire to identify susceptibility genes that might provide clues for novel treatments have motivated numerous gene-mapping studies. Genome-wide linkage scans of hypertensions have been performed (Rice et al. 2002; Harrap et al. 2002b; Caulfield et al. 2003; Yang et al. 2003), but show inconsistent results implicating many chromosomal regions (see Appendix B). The presence of epistatic interactions and locus heterogeneity in the underlying genetics of hypertension may explain the lack of replicated linkage to date (Williams et al. 2004b). While several studies have examined interactions between specific genes in hypertension in humans (Staessen et al. 2001; Tsai et al. 2003; Williams et al. 2004b), a genome-wide search for epistasis in the linkage datasets has not been performed.

In this chapter a two-dimensional (2D) linkage scan was performed using Merloc in the BRItish Genetics of HyperTension Study (BRIGHT) dataset, which is one of the largest genetic studies of hypertension with over 2 000 strictly defined ASPs (Caulfield et al. 2003). Genome-wide significance thresholds were established in the context of two-dimensional (2D) genome-scans for typical ASP linkage data. The peaks in the BRIGHT 2D surface pointed to novel loci for hypertension and for

each 2D peak the exact model of epistasis which best fit the pairwise interaction was examined.

5.2 BRIGHT dataset

Ascertainment criteria for the BRIGHT study (http://www.brightstudy.ac.uk/) have been previously detailed (Caulfield et al. 2003). The cohort consisted of 1639 families with at least two extremely hypertensive siblings (each affected individual's blood pressure was greater than the 95th percentile for the general population after adjustment for age and sex), resulting in 2076 affected full-sibling pairs and 66 affected half-sibling pairs, after Relpair (Boehnke and Cox 1997) analyses. Genotypes were obtained for 447 microsatellite markers. The microsatellite markers selected included markers used in the original genome scan (Caulfield et al. 2003) and thirty-four additional markers spaced every 2cM in regions of interest from chromosomes 2, 5, 6 and 9 and every 5cM in regions on chromosomes 8, 11, 13 and 15. Parental genotypes were unavailable for the majority of the families. The average marker spacing was 8 Kosambi cM and the largest inter-marker distance was observed for the most distal marker pair on chromosome 5q (28cM). The Rutgers genetic map (Kong et al. 2004), which integrates physical and genetic mapping data was used in the analysis and the consensus marker order agreed with three releases of UCSC, April 2003, July 2003, and May 2004. The Marshfield (Broman et al. 1999) and deCode (Kong et al. 2002) genetic maps were also used in some of the analyses.

5.3 Determining significance thresholds in a 2D scan

It is necessary to establish appropriate significance thresholds in multidimensional genome scans to determine the overall significance of the results (Frankel and Schork 1996). Genome-wide significance thresholds were obtained via

simulation using 2D genome-wide simulations in typical ASP linkage screens with missing parental data, average marker spacing of 8cM, and partially informative markers.

To obtain the significance thresholds for the entire 2D surface genotypes for 414 markers were simulated using Merlin in 100 ASPs selected from the BRIGHT dataset. The missing data patterns and distribution of sibship sizes for these 100 families (sibling pairs, trios and quartets) were representative of the entire BRIGHT sample. One thousand replicates of the 2D surface were analysed in Merloc under the general two-locus genetic model to obtain the distribution of the 2D surface under the null hypothesis. The general two-locus genetic model, including interaction effects, was compared to a null genetic model, in which neither gene in a pair contributed to the trait. The resulting two-locus maximum Lodscore (MLS) thresholds over the entire 2D surface for the general genetic model were 5.84 and 6.77, for type 1 error rates of 0.05 and 0.01 respectively. The two-locus MLS expected on average once per genome scan, or the threshold for declaring suggestive linkage, was 4.3. These thresholds (4.3, 5.83, and 6.77) were used to assess genome-wide significance over the entire 2D surface, where all 2D coordinates that surpassed suggestive evidence for linkage (two-locus MLS = 4.3) were reported in the 2D results.

The analysis of a 2D genome-wide surface requires consideration of coordinates involving pairs of unlinked loci and coordinates involving pairs of linked loci. In the initial 2D genome-wide results both unlinked and linked pairs of loci are treated similarly, and the same thresholds are used to declare siginificance (5.84 and 6.77, for type 1 error rates of 0.05 and 0.01). However, for linked loci the distribution of the MLS in the absence of linkage is a function of the recombination fraction, θ, that separates the loci. To investigate the relationship between θ and the MLS

thresholds, two completely informative markers located at different distances apart were simulated in 100 ASPs using 1 000 000 replicates (Figure 5.1A). The results indicated that more tightly linked genes have lower nominal significance thresholds. Therefore, two-locus MLS have a different interpretation depending on whether two genes map close together or further apart. One may take this into account and calculate genome-wide significance thresholds particularly for pairs of loci that are are linked. The genome-wide significance thresholds used for linked loci should be comparable to those used for unlinked loci and take into account the recombination fraction separating the pair of loci. There are different approaches to calculating linked genome-wide significance thresholds. I have used the overall 2D thresholds (5.84 and 6.77, in method P_0) and considered three additional methods (P_1, P_2, P_3).

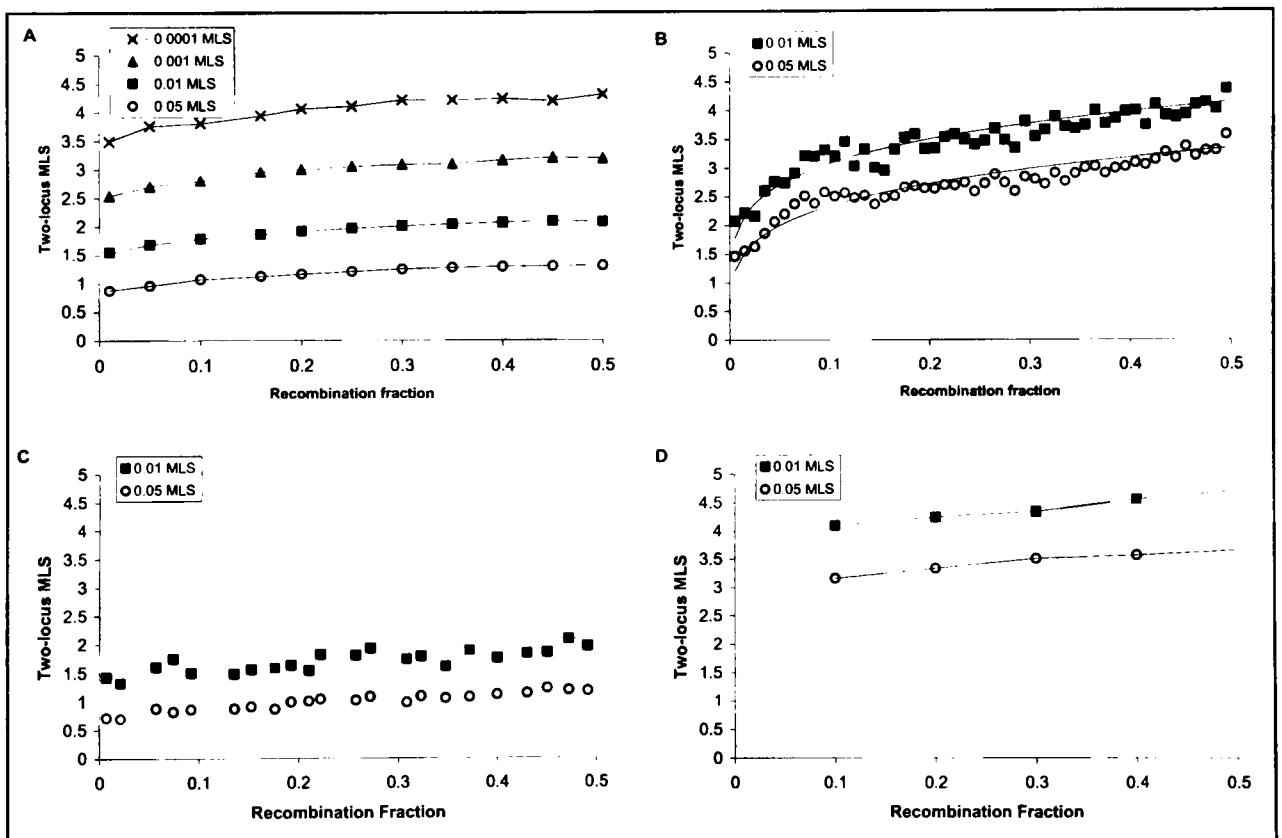


Figure 5.1. 2D significance thresholds for linked pairs of genes. Two-locus MLS thresholds under the general two-locus model as a function of the recombination fraction, θ, in (A) 100 fully informative ASPs with 1,000,000 replicates, (B) in the BRIGHT 2D simulations with 1,000 replicates using all linked coordinates, (C) in the BRIGHT 2D simulations with 1,000 replicates taking one pair of loci from each chromosome to calculate the adjustment Lod, and (D) in the BRIGHT 2D simulations with 1,000 replicates computing the MLS at θ of 0.1, 0.2, 0.3, and 0.4, away from each marker. Least-squares lines were fit for each p-value level using a power function.

In approach P_1 the parts of the surface that comprised unlinked and linked pairs of loci were considered two separate experiments and MLS thresholds were calculated accordingly. For genes that mapped to different chromosomes (representing 95% of the surface) the 2D thresholds for the two-locus MLS were calculated from the 2D genome-wide simulations examining only unlinked coordinates. The thresholds for unlinked pairs of loci were 5.83 and 6.77, for type 1 error rates of 0.05 and 0.01 respectively. For linked genes closely linked loci have lower thresholds (Figure 5.1A) and there are fewer tests in a 2D genome-scan involving two closely linked loci than for coordinates involving loci that map further apart, for example there were only 5 tests for marker pairs at $\theta < 0.01$, compared to over 100 tests at θ between 0.4 and 0.41. Therefore, two-locus MLS thresholds for linked loci can be estimated to take into account θ and the different number of tests performed at each θ value. This was achieved by calculating thresholds by using all the data-points from linked pairs of loci from the 2D genome-wide simulations (Figure 5.1B). This approach is not considered genome-wide in the sense of using the entire 2D grid of coordinates, but only takes into account the 2D simulation results from linked pairs of genes.

However, one may argue that the different number of tests at each θ value should not be relevant when estimating genome-wide significance. To address this question two additional approaches to estimating significance thresholds for linked pairs of loci were also considered (P_2 and P_3). In method P_2, an adjustment in Lod units was made to each of the observed MLS from linked loci in the calculation of the genome-wide *P*-value using MLS thresholds based on the entire 2D surface (5.84 and 6.77), which is predominantly composed of pairs of unlinked loci. The magnitude of adjustment depended on the recombination fraction separating the two linked loci and

was estimated using part of the 2D simulation results from linked pairs of loci taking one coordinate per chromosome (Figure 5.1C). The value of the scaling factor was determined at each value of θ, by calculating Lod (θ=0.5) – Lod (θ<0.5). The value of the scaling factor was determined from the 2D genome-wide simulations and is shown in Figure 5.1C. For example, for two linked loci the MLS of 4 obtained at θ = 0.3 would be adjusted by adding 0.15 units to it (calculated from Figure 5.1C) and the estimated genome-wide significance is calculated as that for an unlinked pairs of loci with MLS of 4.15 in 2D genome simulations, resulting in P_2 = 0.43. However, in this approach it may be too simplistic to adjust the significance by adding a MLS-adjustment to the two-locus MLS for linked regions. The distribution of the MLS under the null may differ greatly for different values of θ and Figure 5.1C only shows the upper 5% and 1% tails in the distribution of the null MLS at each θ, rather than the entire distribution. In addition, the value of the adjustment was established from 22 simulation coordinates, representing one coordinate at each value of θ, and it may be more suitable to use 414 (representing the number of markers) coordinates at each value of θ.

The final method of estimating genome-wide significance thresholds for pairs of linked loci (P_3) suggested in this study was to calculate thresholds using all 414 markers and revisiting the 2D genome-wide simulations. In this method for each simulated marker in the genome the two-locus MLS was calculated for that marker and a location at θ = 0.1, 0.2, 0.3, and 0.4 away (Figure 5.1D). Thus, the resulting relationship will be similar to the results from Figure 5.1B, but with the same number of tests (414) at each θ level. However, only 256 of the 414 markers were located so that all locations θ = 0.1, 0.2, 0.3, and 0.4 away from that marker fell within 20cM of the chromosome, but the resulting thresholds were very similar for either set of 256 or

414 markers. Overall, I believe that method P_0 should be used to declare genome-wide significance for any pair of unlinked or linked regions. For linked regions the additional methods of considering significance are also of interest and in particular P_3.

For each coordinate that surpassed genome-wide suggestive evidence for linkage (two-locus MLS = 4.3 under the general two-locus model) the fit of nested two-locus genetic models was assessed. To examine genetic models in more detail at a particular coordinate 100 000 replicates of 100 completely informative ASPs were simulated under a nested model of interest to evaluate the fit of different nested models compared to the general two-locus model (see chapter IV). To assess significance for the fit of the additive model compared to the general model replicates were generated under an additive model by sampling from the observed two-locus IBD distribution obtained under the additive model (at the peak MLS) during linkage analysis of the BRIGHT data. The same approach was used to assess deviation from multiplicative genetic model.

5.4 BRIGHT Analysis Results

5.4.1 Single-locus results

The results from the one-dimensional genome scan with the Rutgers map are presented in Figure 5.2. The single-locus linkage identified one major peak (MLS = 2.54) on chromosome 5q13.1, and several minor peaks (MLS > 1) on chromosomes 1 (MLS = 1.53), 3 (MLS = 1.27), 6 (MLS = 1.11), 9 (MLS = 1.13), 13 (MLS = 1.56), 15 (MLS = 1.6), and 19 (MLS = 1.51). Overall, no regions surpassed the single-locus genome-wide significance threshold (MLS = 3.14) calculated empirically (Caulfield et al. 2003).

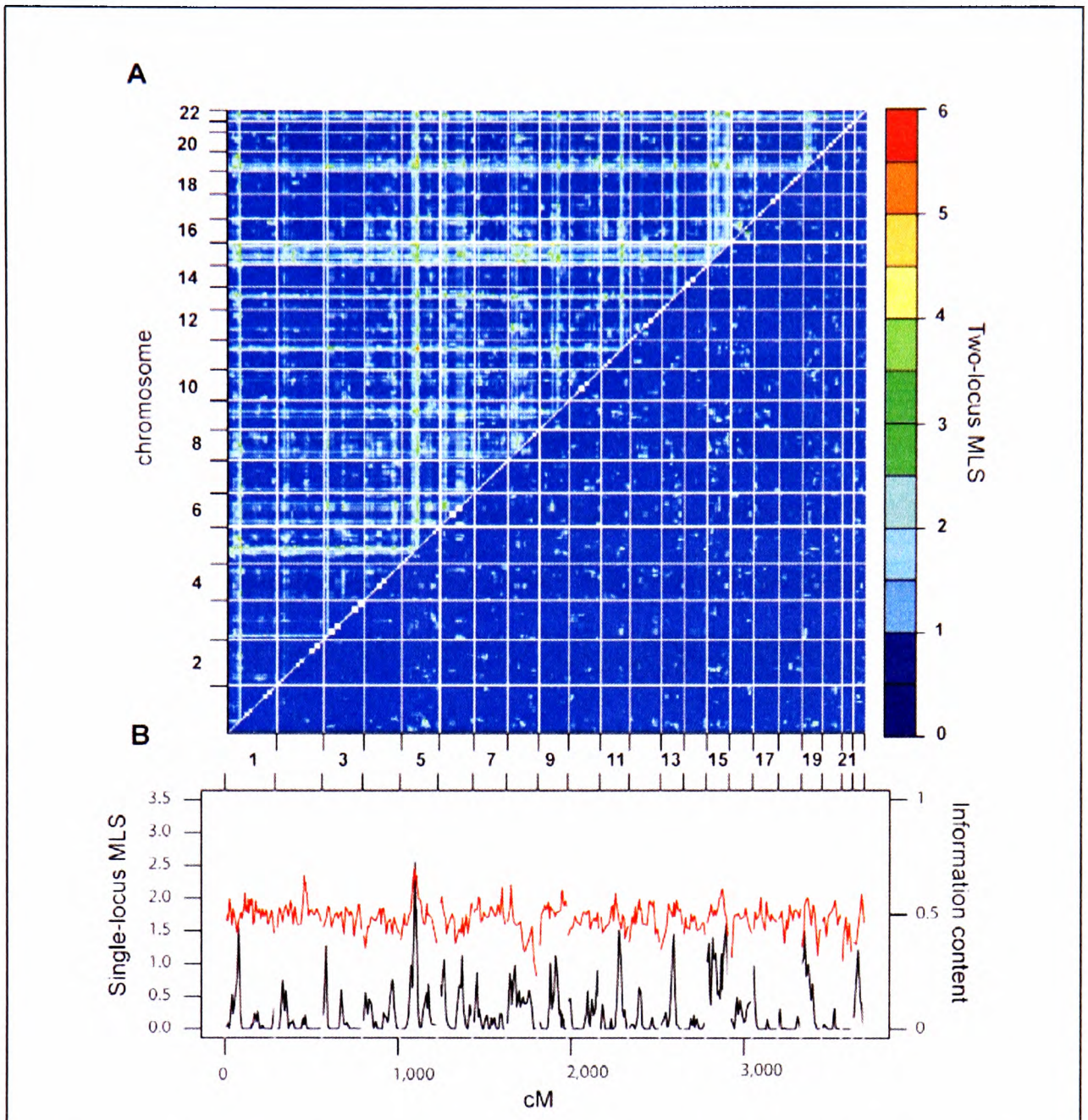


Figure 5.2. Genome scans of essential hypertension. (A) Two-dimensional (2D) genome-wide linkage scan. Two-locus MLS statistics calculated under the general two-locus model are presented above the diagonal. The difference in the MLS for the fit between the general and additive two-locus models is presented below the diagonal. (B) Single-locus genome scan of essential hypertension. The single locus MLS (black line) and the genome-wide information content (red line) for the BRIGHT data.

5.4.2 Two-locus results

A two-dimensional linkage scan was performed by computing the two-locus maximum Lodscore (MLS) at each marker-pair or 2D marker grid coordinate across the genome (Figure 5.2A). Several peaks in the 2D surface identified pairs of loci that

interact and contribute to hypertension susceptibility (Table 5.1). The 2D scan identified regions that had no significant effect on hypertension in the single-locus scan (Figure 5.2B), but that contributed to the phenotype under two-locus models. For each 2D peak, two-locus models were examined in detail, starting with the general model that fit a wide range of epistasis, and restricting the number of free parameters in a stepwise manner to estimate the model that best fit the interaction (Table 5.1). To determine the degree of epistasis in the genetic model the maximum likelihood estimate of ε was used, supported by the difference in MLS computed under different epistatic models. The results from the 2D coordinates that comprised unlinked and linked regions are presented separately along with estimates of genome-wide significance for the peak findings.

For unlinked regions, the highest peak over the 2D surface was obtained for two loci on chromosomes 5 and 11 (Table 5.1). The two loci mapped to chromosomes 5q13 and 11q22, both of which showed suggestive evidence for linkage in the 1D linkage scan (MLS > 1). The two-locus MLS was 5.45 increasing to 5.72 under a fine-scale grid scan at an average spacing of 2 cM (Figure 5.3A), and was marginally genome-wide significant at $P = 0.08$ with a 95% confidence interval of (0.06-0.1). This P-value assesses the significance of detecting joint effects at both loci on chromosomes 5 and 11. There was evidence of epistasis at this coordinate as suggested by significant differences between the MLS under the general (5.45) and additive (3.84) models (difference = 1.61, epistasis $P = 0.0013$) and the general and multiplicative (4.05) two-locus models (difference = 1.4, epistasis $P = 0.0016$), and a two-locus MLS of 5.45 under the epsilon-epistatic model with a maximum-likelihood estimate of epsilon at 100 (1-Lod unit support interval: $4\text{-}10^3$). The second highest unlinked pairwise peak in the surface also involved region 5q13.1 in a pairwise

interaction with a locus on 19q12 (Table 5.1). The Lodscore for this interaction increased from 5.12 under the sparse search to 5.35 under a fine-grid scan (Figure 5.3B), with a genome-wide $P = 0.15$ for detecting the action of the two loci. The best genetic model describing this interaction was a model of strong epistasis (epsilon $\geq 10^3$), supported by a significant difference between the general (5.12) and additive (2.85) models (difference = 2.27, epistasis $P = 0.0001$), and the general and multiplicative models (2.14, epistasis $P = 4\times10^{-5}$). The 2D genome scan also identified a potential interaction between regions 9q22.3 and 15q12 (Figure 5.3C) with a two-locus MLS of 4.8 approximated by a strong epistatic model (Table 5.1), and several other interactions between unlinked regions that reached genome-wide suggestive evidence for linkage (Table 5.2). Of the ten genome-wide suggestive unlinked pairs none fit the additive two-locus model, and only one, 1p33-5q13.1, fit the multiplicative model.

Table 5.1 Most significant peaks from the BRIGHT 2D scan.

			Two-locus MLS				
			GEN[b]	ADD[c]	MUL[d]	ε[e] (1LU SI)	*P* value (95% CI)[f]
Locus 1[a]	**Θ**	**Locus 2[a]**					
Pairs of loci on different chromosomes							
5q13.1		11q22.1					
D5S2019, 2.5	0.5	*D11S898*, 1.5	5.45	4.05	3.84	100 (4-103)	0.08 (0.06-0.10)
5q13.1		19q12					
D5S2019, 2.5	0.5	*D19S414*, 0.5	5.12	2.98	2.85	1000 (21-1000)	0.15 (0.12-0.17)
9q22.3		15q12					
D9S287, 1.1	0.5	1.3	4.8	2.44	2.28	1000 (41-1000)	0.28 (0.26-0.31)
Pairs of linked loci							
16p12.3		16q23.1					
D16S3046, 0.5	0.33	*D16S515*, 0.1	4.37	0.63	0.53	1000 (170-1000)	0.48 (0.45-0.51)

[a] The chromosome location, peak marker under the sparse grid, and single-locus MLS for each region. Two-locus MLS were computed under the [b] general, [c] additive, and [d] multiplicative two-locus models for the sparse grid (non-adjusted MLS shown for syntenic regions). [e] The maximum likelihood estimate of ε and the corresponding 1-Lod unit support interval (1LU SI). [f] The estimated genome-wide *P* value corresponds to the general model MLS under the sparse grid, and is given along with the 95% confidence interval (95% CI) of the estimate. For the linked coordinate the *P* value corresponds to P_0.

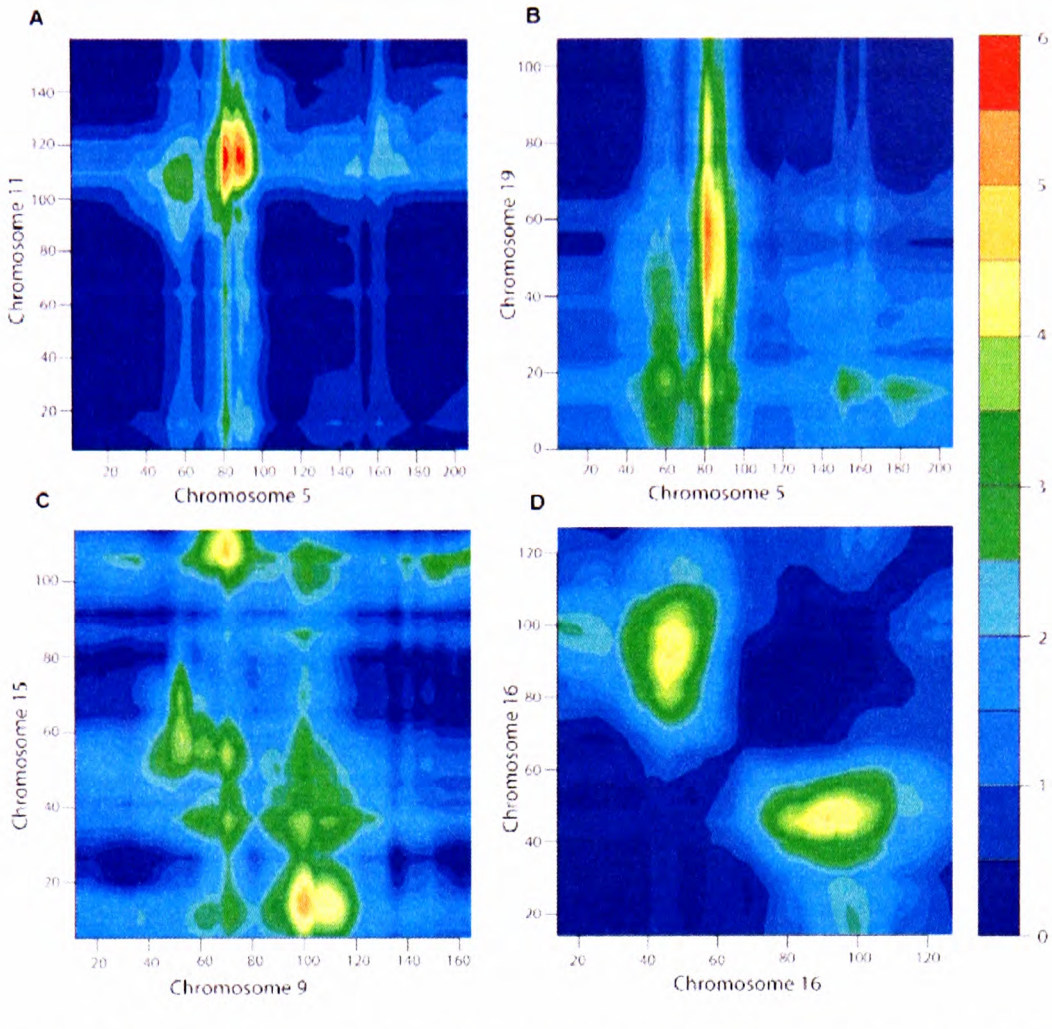


Figure 5.3. Fine-scale examination 2D peaks. (A) Chromosome 5 and chromosome 11, coordinate density is at an average of 1.51cM on chromosome 5 and 2.1cM on chromosome 11. The sparse-grid peak of 5.45 increases to 5.72 under the fine-grid, with a coordinate shift from (chr 5: 80.56, chr 11:109.32) to (chr 5: 90.65, chr 11: 117.86). (B) Chromosome 5 and chromosome 19, coordinate density on chromosome 19 is at an average of 2.06cM. The sparse-grid peak of 5.12 increases to 5.35 under the fine-grid, with a distal coordinate shift on chromosome 19 of 2.34cM. (C) Chromosome 9 and chromosome 15, coordinate density is at an average of 1.54 cM on chromosome 9 and 1.7 cM on chromosome 15. The sparse-grid peak two-locus MLS remains unchanged in value and location. (D) Chromosome 16, coordinate density is at an average of 2.3 cM. The sparse-grid peak of 4.37 increases to 4.50 under the fine-grid, with a proximal coordinate shift of 5.5 cM on 16q23.1.

For pairs of linked regions the most significant result identified two loci on chromosome 16 (16p12.3 and 16q23.1) that had no significant or suggestive effect on hypertension under single-gene models, but contributed to the phenotype under a two-locus model of strong epistasis (Table 5.1). As discussed previously there are different

approaches to interpreting the significance of detecting the joint action of linked pairs of loci. Under the genome-wide 2D significance levels (5.84 and 6.77), the significance was established as $P_0 = 0.477$. Under the other methods, the significance was established as follows. In the first approach, P_1, taking into account the different number of tests performed, the estimated genome-wide *P*-value for this interaction was significant, $P_1 = 0.002$. In the second approach, which used a Lod-scaling factor, the two loci underlying this peak were separated by 0.33 units of recombination, resulting in an adjustment of 0.14 Lod units when evaluating the genome-wide *P* value, and the resulting genome-wide *P*-value was $P_2 = 0.34$. In the final approach, using significance levels directly from Figure 5.1D, the estimated genome-wide P-value was $P_3 < 0.01$, while the interaction was nominally significant at $P < 1\times10^{-4}$. A fine-grid scan of this region at an average 2 cM density for the analysis resulted in a MLS increase from 4.37 to 4.5 (Figure 5.3D). The best-fitting model for this coordinate on chromosome 16 was an extreme epistasis model (epsilon $\geq 10^3$), supported by a highly significant difference between the MLS computed under the general (4.37) and additive (0.53) two-locus models, (difference = 3.84, epistasis $P < 1\times10^{-5}$), and the general and multiplicative (0.63) genetic models (difference = 3.74, epistasis $P < 1\times10^{-5}$). The 2D scan also identified novel regions on chromosomes 5 and 9, which again involved pairs of linked loci (Table 5.2) and were nominally significant. In both cases there was suggestive evidence for linkage to one region in each pair in the 1D scan. On chromosome 9 the single locus peak (9q31.1) interacted with a linked marker with no single-locus evidence for linkage (9p24.2). The estimates of genome-wide significance for the two-locus MLS (4.01) according to the proposed approaches were $P_0 = 0.72$, $P_1 = 0.007$, $P_2 = 0.69$, $P_3 = 0.025$. An epistatic model best fit at this coordinate, with MLS = 4.01 under a model of epistasis (epsilon

≥ 10^3), and significant differences between the two-locus MLS under the general (4.01) and additive (1.12) models (difference = 2.89, epistasis $P<1\times10^{-5}$), and the general and multiplicative (1.17) models (difference = 2.84, epistasis $P<1\times10^{-5}$). On chromosome 5 two regions at 5p14.1 and 5q13.3 interacted under an epistatic model with a general two-locus MLS of 3.75. The estimates of genome-wide significance at this coordinate using the three proposed approaches were $P_0 = 0.86$, $P_1 = 0.008$, $P_2 = 0.78$, $P_3 = 0.025$. An epistatic model approximated this interaction as well, with an epsilon-epistatic MLS of 3.74 and epsilon ≥ 10^3, and significant differences between the general and additive, and the general and multiplicative two-locus MLS.

Table 5.2 Genome-wide suggestive results from the hypertension 2D scan.

		Two-locus MLS						
Locus 1[a]	Locus 2[a]	General[b] (Rutgers)	General[c] (Marshfield)	General[d] (deCode)	Expected 2D peaks[e]	ε	*P* value	1-LU Support Intervals (cM)[f]
Pairs of loci on different chromosomes								
5q13.1	11q22.1							(79.1 – 92.8)
D5S2019, 2.5	*D11S898*, 1.5	5.45	4.81	5.15	0.06	100	0.08	(106.8 – 126.3)
5q13.1	19q12							(78.6 - 90)
D5S2019, 2.5	*D19S414*, 0.5	5.12	4.67	5.13	0.12	1000	0.15	(32.9 - 67.7)
9q22.3	15q12							(93.7 – 116.4)
D9S287, 1.1	*D15S1002*, 1.3	4.8	4.23	5.06	0.37	1000	0.28	(5 – 23.5)
5q13.1	9q22.3							(73.5 – 85.1)
D5S2019, 2.5	*D9S287*, 1.1	4.77	4.32	4.74	0.39	91	0.28	(91.8 – 111.6)
8p12	15q11.2							(50.9 – 74.1)
D8S505, 0.4	*D15S128*, 1.0	4.76	4.55	4.88	0.4	1000	0.28	(0 – 13.6)
1p33	5q13.1							(69.4 – 88.1)
D1S2797, 1.5	*D5S2019*, 2.5	4.6	4.06	4.56	0.55	23	0.36	(76.5 -90.5)
5q13.1	14q32.3							(78.6 - 90.7)
D5S2019, 2.5	*D14S292*, 0.1	4.6	4.61	4.6	0.55	1000	0.36	(104.2 - q-ter)
3p26.1	19p13.3							(4.5 - 30.5)
D3S1304, 1.3	*D19S894*, 1.5	4.33	5.31	2.36	0.93	1000	0.54	(0 – 24.4)
3p12.3	5q13.1							(95.2 - 123.2)
D3S3681, 0.6	*D5S2019*, 2.5	4.31	4.48	4.29	0.98	1000	0.55	(76.5 - 85.1)
1p36.1	19q13.3							(48 - 79.9)
D1S234, 0.4	*D19S420*, 0.4	4.31	4.24	4.23	0.98	1000	0.55	(62.6 - 74.8)
Pairs of linked loci								
16p12.3	16q23.1							(39.7 – 55.2)
D16S3046, 0.5	*D16S515*, 0.1	4.37	6.01	4.26	0.67	1000	0.48	(76.4 - 106)
9p24.2	9q31.1							(p-ter – 16.2)
D9S288, 0.1	*D9S1690*, 1.1	4.01	3.64	4.07	1.56	1000	0.72	(95.6 – 114.1)
5p14.1	5q13.3							(33.4 – 58)
D5S419, 0.2	*D5S424*, 1.8	3.75	3.04	4.1	2.01	1000	0.86	(85.2 – 93.8)

[a] Chromosome location, peak marker under the sparse grid, and single-locus MLS. Two-locus MLS computed under the general two-locus model using the [b] Rutgers, the [c] Marshfield, and the [d] deCode maps. All pairs apart from 1p33 and 5q13.1 show a significant difference ($P < 0.01$) between the fit of nested additive and multiplicative models compared to the general model under the Rutgers map. [e] The expected number of peaks genome-wide with the same or higher Lod as the Rutgers two-locus MLS (or P_0). [f] 1-Lod-unit (1LU) support intervals in Kosambi cM (Rutgers) for locus 1 (top line) and 2 (bottom line) in each pair for the fine-grid MLS.

The 1-Lod-unit (1LU) support intervals were obtained for each 2D interaction peak as a means of narrowing down chromosomal regions within which to search for underlying genetic variants (Table 5.2). For the most significant unlinked and linked pairwise interactions confidence intervals obtained from the 1D scan for loci with single-locus MLS > 1 (5q13.1, 9q31.1, 11q22.1, and 15q12) were compared to support intervals from the 2D scan. On chromosome 5 the support region obtained from the 2D results was 13.7cM (79.1 - 92.8), compared to 23.5cM obtained from the 1D results. The 2D interval on chromosome 9 was 22.7cM (93.7 – 116.4) compared to the 1D support interval of 39cM. On chromosome 11 the 2D support interval was 19.5cM (106.8 – 126.3), falling from 29.3cM under single-locus analysis. Finally, on chromosome 15 the 2D 1LU support interval was 18.5cM (5 – 23.5), which was reduced from 23cM in the 1D results.

5.4.3 Sensitivity to map misspecification

The results of the 2D analysis were examined under three genetic maps, the Rutgers (Kong et al. 2004), the deCode (Kong et al. 2002), and the Marshfield (Broman et al. 1999) maps (Table 5.2). The MLS at the peak coordinates in Rutgers differ across analyses performed assuming the Marshfield and deCode maps. The peak MLS often shifts to a coordinate in the proximity of the original peak when computed under a different genetic map. Of the 2D peaks obtained under the Rutgers map the chromosome 16 peak achieved genome-wide significance under the Marshfield map (using the most stringent approach genome-wide $P_0 = 0.031$). To double-check the peak results under Rutgers, the 2D peaks in Table 5.2 were also analyzed not taking into account multipoint information and similar (but always slightly lower) two-locus Lod scores were obtained using the single-point IBD's for each pair of peak markers.

The variation in two-locus MLS under different genetic maps prompted a re-analysis of the entire surface using the Marshfield map (Figure 5.4A). Overall, most of the peaks remained in the same general location with similar two-locus MLS estimates and small shifts (or none at all) in the specific marker-pairs involved. One exception was region 8p23.1-p22, where the two-locus multipoint MLS under Marshfield for the interaction between D8S277 and D8S552 was 7.8, while the two-locus multipoint MLS under Rutgers for that coordinate was 0.92 (Figure 5.4B). Upon close examination it appears that there is an inversion in the region between the two markers in the deCode genetic map and a literature search revealed that this is a well-characterized polymorphic inversion (Giglio et al. 2001; Sugawara et al. 2003). In addition, a polymorphic duplication, close to the inversion, has also been reported in this region (Barber et al. 1998; Harada et al. 2002). Simwalk2 (Sobel and Lange 1996) and genehunter (Kruglyak et al. 1996) were used to obtain estimates of the observed recombinants for that region in the set (and subsets) of families from BRIGHT. The results supported the recombination rates obtained from the Rutgers map.

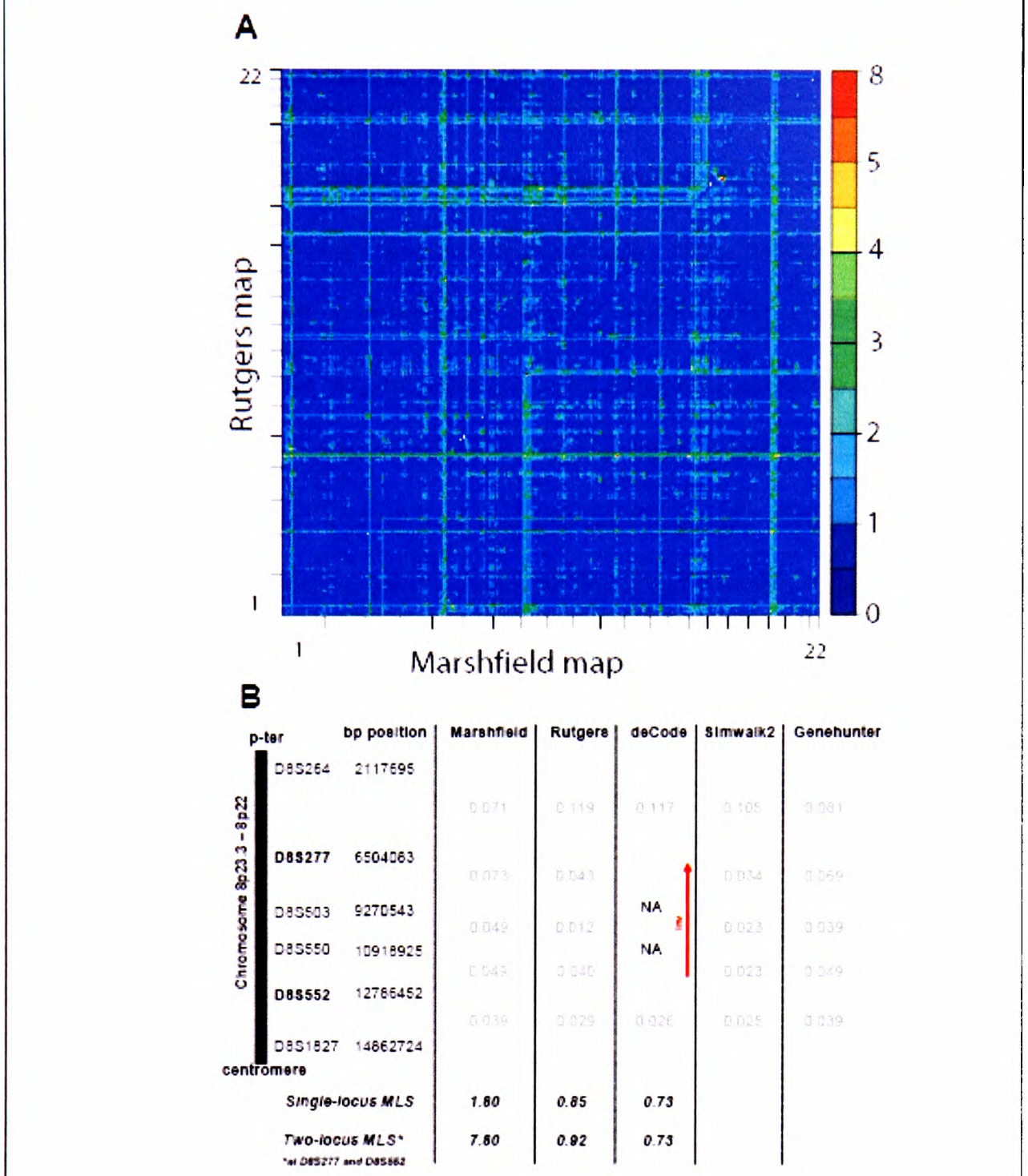


Figure 5.4. BRIGHT genome analysis under the Marshfield genetic map. (A) 2D genome-wide linkage scan. Two-locus MLS are calculated under the general two-locus model. Results computed using the Rutgers map are presented above the diagonal, and results obtained under the Marshfield map are shown below the diagonal. (B) Chromosome 8 results calculated under different genetic maps and observed recombination fractions. Multipoint MLS are presented for results from published genetic maps and deCode distances were interpolated using Marshfield where necessary. Observed recombinants were estimated in BRIGHT in all families with simwalk2 and in the first 1138 families with genehunter (most parental genotypes were unavailable).

5.4.4 Locus-counting results

The method of locus-counting (Wiltshire et al. 2002) is complementary to estimating genome-wide significance levels and may be used to evaluate the joint significance of complex-trait genome scan results. This approach estimates the probability of observing a number of linkage peaks above a pre-defined MLS threshold compared to the number expected under no genetic influence.

First, this method was extended to two loci by counting the number of times that a 2D peak surpassed a given Lod-score threshold per 2D scan by re-examining the 2D simulations, focusing only on coordinates which involved genes located on different chromosomes. Independent linked coordinates were defined as pairs of regions showing evidence for two-locus linkage and separated by at least 40 Kosambi cM at both marker locations. The results indicate that significantly more peaks between markers on different chromosomes contributed to hypertension than expected by chance alone (Figure 5.5). For example, the probability of observing ten peaks above or at a two-locus MLS threshold of 4.31 was only 0.009 under the null hypothesis of no linkage. Based on the findings a two-locus MLS of 4.3 is expected to occur once by chance in a 2D genome-scan, hence 4.3 could be the threshold to designate a two-locus peak as showing 'suggestive' evidence for linkage in a 2D scan with an average spacing of 8cM.

Second, the aim was to determine whether any regions were over-represented amongst the BRIGHT 2D peaks than expected by chance. To achieve this, I counted the number of times that the same region was involved in the top ten peaks per simulated 2D scan for genes on different chromosomes. The simulation results indicated that the same region would be expected to be represented on average 3.01

times in the ten peak coordinates, while chromosome 5q13.1 was represented six times (Table 5.2).

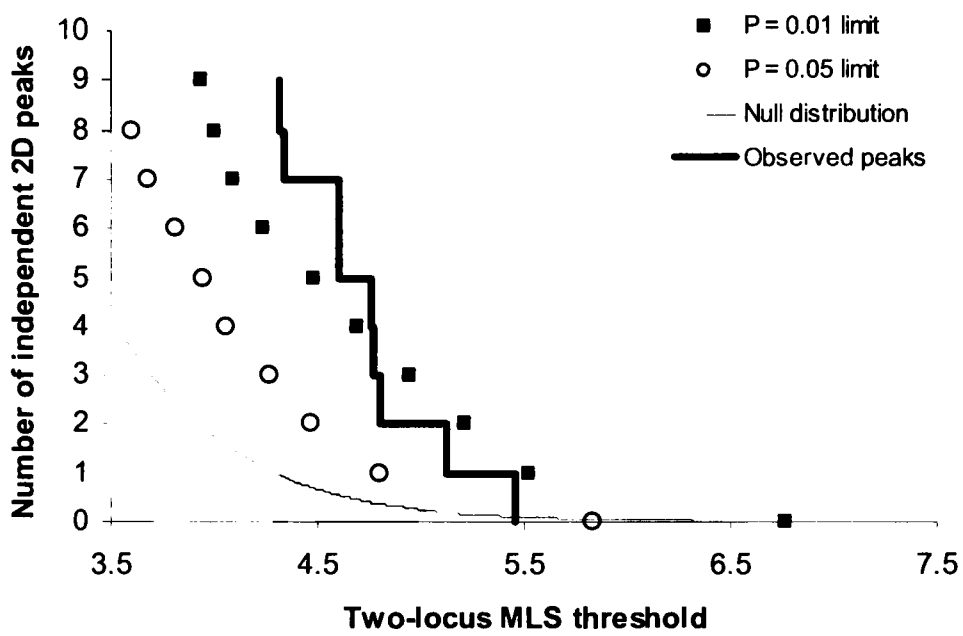


Figure 5.5. Two-locus counting results for unlinked loci. The two lines represent the null distribution of expected number of independent peaks from the 2D simulations and the observed distribution of unlinked 2D peaks in the BRIGHT data, along with significance limits for 0.05 and 0.01 2D genome-wide type 1 error rates.

5.5 Discussion

In this chapter the two-locus model-free approach was applied to the study of essential hypertension, confirming that it is computationally feasible to calculate a 2D linkage grid for typical human genome-scan data. The purpose of performing a 2D linkage scan is twofold: first, to identify novel regions that contribute to the trait via a genetic interaction, and second, to detect interactions between pairs of contributing loci and describe the best genetic model that fits the interaction.

The 2D scan identified regions that had no significant or suggestive (Lander and Kruglyak 1995) effect on hypertension in the single-locus scan, but that contributed to the phenotype under two-locus models. Of particular interest was the interaction of two linked loci on chromosome 16, for which there was no evidence for

linkage in the single-locus analysis. The region on chromosome 16p13.1 was previously implicated in systolic blood pressure in a sample of 274 Australian ASPs (Harrap et al. 2002b). This is consistent with additional findings for linkage of systolic blood pressure to this region (Wong et al. 1999; Yang et al. 2003). A genome-wide scan of both systolic and diastolic blood pressure reports suggestive evidence for linkage to this region in 114 African American families (Rice et al. 2002). In addition, Xu et al. (1999) performed a single-locus linkage scan of systolic blood pressure in 99 low concordant sibling pairs that indicated a linked region on chromosome 16 between 16p13.1 and 16q23.1, maximizing at 16q12.1. The two-locus results are consistent with the hypothesis that these linkage results are due to two separate loci, each mapping on opposite flanks of the single-locus peak, the effects of which superimpose to generate a single-locus peak mid-way between the two loci. Similarly, on chromosome 5 the 1D peak in the data mapped between the two loci identified from the 2D scan. These results again suggest that signals from two separate loci superimpose to generate a single-locus peak between the two contributing loci in the BRIGHT data. There is prior evidence for linkage of systolic blood pressure to a region at 5q14 (Yang et al. 2003) and diastolic blood pressure to 5q15 (Cooper et al. 2002), 15 cM and 23 cM distal to the second locus, respectively. However, previous studies have not implicated either 5p14.1, or either locus on chromosome 9 (9p24.2 and 9q31.1) identified from the 2D scan results.

The highest peak over the entire 2D surface was obtained between chromosomes 5q13 and 11q22. There is evidence for linkage to both regions from the 1D scan and previous studies have shown linkage to regions distal of the locus on chromosome 5 (Cooper et al. 2002; Yang et al. 2003), and have also implicated 11q21 (Rice et al. 2002). The two-locus analysis indicated that these loci interact

epistatically, however, the evidence for epistasis on chromosome 16 is stronger. It is possible that the underlying interaction model involving chromosomes 5 and 11 is a three-locus genetic model, with an interaction between 5p14.1, 5q13.3, and 11q22, although no significant evidence for epistasis was obtained between 5p14.1 and 11q22. Chromosome 5q13.1 was involved in several of the highest 2D peaks in the surface, including the interaction with a locus on 19q12 which is 20cM distal to a previously reported linkage signal for systolic blood pressure (Cooper et al. 2002). The regions involved in the 2D significant and suggestive interactions were examined to define support intervals for localization of contributing loci using the 1-Lod-unit support intervals.

The 1-Lod-unit support intervals from the 2D most significant pairwise interactions allowed us to somewhat narrow down regions containing putative susceptibility loci. However, on average the regions were still quite broad with many potential candidate genes underlying the peaks and so it may not be worthwhile to examine candidate genes under the 2D peaks at this point in the study. Nevertheless, some preliminary ideas may be of interest in order to highlight any obvious candidates. Identifying genes in the same or related pathways, for example using KEGG (Ogata et al. 1999), could be a first step in identifying the genes that interact. The results from the KEGG analysis were obtained from Dr P. Munroe from the BRIGHT consortium at Barts and The London School of Medicine and Dentistry in London. For the interacting loci found on chromosome 16 KEGG analysis revealed genes involved in oxidative metabolism, glycerolphospholipid metabolism and aminoacyl-tRNA biosynthesis. Alternatively, genes could be selected for specific analysis if there were data indicating a role in blood pressure regulation. For some of the epistatic loci detected in the 2D analysis there are interesting potential candidates

which might be further explored, for example sodium-hydrogen exchanger, 11 betahydroxysteroid dehydrogenase, Nedd 4 like E3 ligase, epithelial sodium channel subunits, and a dopamine receptor (results obtained from Dr P. Munroe).

I have also searched the list of known genes in the regions of interest and examined any candidates for potential contribution to hypertension (Appendix C). I selected the list of known genes directly underling the 2D peaks 5q13-11q22 and 16p12-16q22 and attempted to search for interactions involving these genes in one published molecular interaction database, BIND (Bader et al. 2001; Alfarano et al. 2005). Initially, I screened the list of known genes for any obvious candidates for hypertension (not necessarily involved in interactions) and, in addition to the genes identified from the KEGG analyses, I came across one other obvious candidate on 16p12.3 – SAH, SA (acyl-CoA synthetase) hypertension-associated homolog isoform 1, which is involved in metabolism and the homologue of which is associated with hypertension in rats (Iwai et al. 1994). Interaction analyses in BIND (Appendix C, Table C1) revealed a molecular interaction between MMP1, matrix metallopeptidase 1 (interstitial collagenase) on 11q22.3, and PAR1, coagulation factor II (thromin) receptor on 5q13 (Boire et al. 2005), which is of interest to hypertension because protease-activated receptors (PARs) are a class of G protein-coupled receptors that play critical roles in thrombosis, inflammation, and vascular biology. Several interactions were also identified for the two regions on chromosome 16 (Appendix C, Table C1) involving E2F transcription factor 4 and four gene promoters (Cam et al. 2004). However, I believe that these results are still hypothetical and a more comprehensive strategy would be to fine map all the regions first allowing for epistasis, for example in a logistic-regression framework (Cordell et al. 2004), and compare the evidence for interactions across linkage and association studies. It will be

necessary to explore the relationship between statistical and biological epistasis in more detail to be able to interpret the 2D findings in a biological context (Cordell 2002; Elston et al. 2005; Moore and Williams 2005).

Multidimensional grid searches involve multiple testing, making it crucial to control the overall type 1 error (Frankel and Schork 1996). Lander and Botstein (1989) suggested using an n-fold higher threshold for the linkage test statistic before declaring significance in an n-dimensional scan, while Holland (1997) proposed to correct only for the number of tests that involve different linkage groups. Choosing an excessively conservative threshold reduces the power to find significant interactions or any linkage at all. To avoid the reduction in power, the search for interactions could be restricted to a limited number of preselected portions of the genome that have detectable main effects (Lark et al. 1995; Holmans 2002), or are plausible biological candidates (Fijneman et al. 1996). This approach will reduce the number of tests performed, but would fail to detect interactions among loci that have non-significant main effects (Marchini et al. 2005). Although most of the 2D coordinates identified in this study included at least one region which shows suggestive single-locus evidence for linkage (MLS > 1), there is also evidence for interactions among regions (chromosome 16) that did not have significant or suggestive single-locus effects on the trait. This is consistent with results from previous multidimensional scans based on genotype data in model organisms, which detected evidence for epistatic loci with no marginal effects.

The 2D scan strategy requires that we resolve how to consider pairs of genes that map close together on the same chromosome as opposed to genes that localize further apart, or on different chromosomes. Under the null hypothesis of no linkage the distribution of the test-statictic depends on the recombination fraction. Under two-

locus additive and epistatic models the power to detect a second susceptibility gene in the presence of a disease locus increases for greater recombination fractions between the two loci (Farrall 1997), and a bias in estimating the locations of the two genes may be observed if they map close together (Biernacka et al. 2005). In addition, there are fewer tests involving two closely linked loci. It therefore appears that a two-locus MLS has a different interpretation depending on whether two genes map close together or further apart. There are different approaches to establish genome-wide significance criteria for pairs of linked loci, which would be equivalent to unlinked 2D thresholds and would take recombination into account. In this chapter a number of approaches were applied, of which the most stringent I believe is the adjustment in two-locus MLS from pairs of linked loci while establishing the *P*-values based on thresholds calculated from the entire surface (P_2), because it produces results most similar to the 'real' genome-wide thresholds taking into account the entire surface, P_0. I believe that approach P_3 is also of interest because it combines P_0 and P_1 to produce genome-wide thresholds only in the context of linked pair of loci. However, there may be more powerful methods of interpreting the overall significance of the results. This analysis suggests further development, for example by examining the performance of different two-locus mapping methods that account for two linked regions (Delepine et al. 1997; Biswas et al. 2003; Biernacka et al. 2005) and can be applied to this section of the surface.

The two-locus linkage method is also very sensitive to map miss-specification. This result is unsurprising because it has been previously shown that varying the genetic map affects multipoint linkage results in single-locus analysis (Daw et al. 2000). However, the BRIGHT peak coordinates were generally consistent across analyses under different genetic maps. An ideal genetic map would be based on

accurate marker ordering information from genome sequencing studies and recombination fraction estimates for adjacent markers from as many meioses as possible. According to these criteria the Rutgers map is currently the best compromise because it is based on a recent release of the human genome sequence and incorporates information from both the DeCode and Marshfield genetic studies.

The Marshfield genetic map was used to re-examine the entire surface and the deCode genetic map was used to examine part of the surface, rather than then entire genome, because if a marker was present in deCode and was in the correct physical order in 2004, then this marker would most likely also appear in the Rutgers map. Because the Rutgers map uses meioses from CEPH and deCode the genetic distances for that marker are likely to be very similar in the deCode and Rutgers genetic maps. The results from region 8p23.1-p22 in analyses from different genetic maps were quite striking and I believe are due to the presence of a polymorphic inversion in that region. Inversions could lead to lower observed recombination rates due to the lower frequency of crossovers in individuals heterozygous for the inversion. The 8p23 inversion appears to be more frequent in individuals of European ancestry compared to those of African ancestry (Giglio et al. 2001; Sugawara et al. 2003). Previous studies of fine- and broad-scale recombination patterns across the genome have proposed that the differences observed in recombination rates in different ethnic groups in this region are likely to be due to the polymorphic inversion (Jorgenson et al. 2005; Serre et al. 2005). In contrast, the region of interest on chromosome 16 does not seem to be involved in any polymorphic intrachromosomal rearrangements (although there is evidence for past genomic rearrangements involving this region (Loftus et al. 1999)), the two loci are further apart in the genetic maps (compared to the markers on 8p), and recombination estimates from different samples and methods

appear to be more consistent (Jorgenson et al. 2005; Serre et al. 2005). Differences and inconsistencies across published genetic maps have been recorded (Nievergelt et al. 2004) and are likely to be due at least to some extent to polymorphic chromosomal deletions, inversions, and duplications. At present, genetic map studies should take this into account, perhaps using a database of common chromosomal rearrangements and human sequence data.

Complex traits likely involve interactions among more than two loci and so the search of multilocus models involving more than two loci in the data is of great interest. In this respect there are two aspects of statistical inference that were explored in the context of systematic 2D scans. The locus-counting strategy was extended to two loci, and an estimate of how often one expects to see the same region appear in the ten highest unlinked peaks was obtained. It appears that there were significantly more 2D peaks than expected under the null in the 2D surface, and region 5q13.1 was involved in more interactions than expected by chance alone. These approaches could highlight loci that may form part of networks of etiological variants and might identify frameworks of epistatic interactions involved in genetic pathways (Segre et al. 2005).

In conclusion, it is computationally feasible to calculate 2D linkage grids for typical human genome-scan data to potentially enable the detection of loci significantly involved in hypertension that have no apparent effect in single-locus scans. These results therefore provide a compelling rationale for re-examining existing genome-wide linkage datasets using the 2D strategy to provide an even greater insight into the complex interaction of genetic factors involved in common human disease.

CHAPTER VI. Two-dimensional linkage scans of complex traits

6.1 Introduction

The simultaneous search strategy for epistatic interactions appears to be a promising approach for identifying loci that contribute to complex trait. Results from two-dimensional scans in model organisms and the two-dimensional linkage scan of hypertension (HT) from chapter V have shown that 2D scans can identify loci that interact epistatically. In this chapter the 2D linkage scan strategy was further explored by performing 2D linkage scans in two additional complex traits, autism and type 2 diabetes (T2D) and comparing 2D results across three complex traits.

6.2 Autism

Autism is a severe neurodevelopmental disorder, characterized by impairment in social interaction, communication, and repetitive patterns of behaviour. Autism is a narrow diagnosis which represents one extreme of the spectrum of autism spectrum disorders (which include Asperger Syndrome, Rett syndrome, Pervasive developmental disorder, and others). The prevalence estimates are approximately 0.1-0.2% for autism and 0.6% for autism spectrum disorders with a male to female bias in prevalence rates (Chakrabarti and Fombonne 2005). There is strong evidence for a genetic basis in autism (see section 6.4) supported by encouraging convergent linkage results in several chromosome regions from genome-wide scans (see Veenstra-Vanderweele et al. 2004). There is also evidence for a complex genetic involvement in the aetiology of autism, in particular, estimates of relative recurrence risk ratios imply epistasis and statistical modelling suggests the involvement of 3 or 4 loci, although as many as 15 loci may be involved (Risch et al. 1999). With this evidence

in mind, a 2D linkage scan was performed to search for pairwise interactions across the genome in autistic ASPs.

The subjects in the study comprised the 99 pedigrees from the International Molecular Genetic Study of Autism Consortium (IMGSAC) with 87 affected sib-pairs (IMGSAC 1998). The proband in each family was diagnosed with autism and the affected sibling(s) either had autism or a broader phenotype falling under the autism spectrum disorder. After re-examination of the data, genotypes were available for 83 ASPs and most of their parents in 79 pedigrees, consisting of 377 individuals. The sample size available for the autism study was very small and as a result these data have very low power to detect joint action of susceptibility loci. Genotypes were obtained for 370 microsatellite markers spanning the autosomal genome at an average spacing of 8.92 Kosambi cM with the largest intermarker distance of 47 cM on 2p14, under the deCode genetic map. A more detailed examination was performed on chromosome 7 to examine evidence for the presence of two linked susceptibility loci.

6.2.1 2D scan

One-dimensional and two-dimensional linkage scans were performed in the IMGSAC data (Figure 6.1). The 1D genome-wide results identify several peaks of suggestive linkage results (single-locus MLS > 1) on chromosomes 2 (MLS = 2.78), 4 (MLS = 1.63), 7 (MLS = 1.83), 8 (MLS = 1.05), 9 (MLS = 2.30), 10 (MLS = 1.72), 14 (MLS = 2.47), 16 (MLS = 1.71), 17 (MLS = 1.74), 19 (MLS = 1.34), and 22 (MLS = 1.31). Overall, no region reached genome-wide significance criteria as defined by Lander and Kruglyak (1995). The 2D genome-wide results were obtained for the sparse 2D marker-grid, by computing the two-locus MLS under the general model at each pair of markers in the genome. Several peaks in the 2D surface identified pairs of loci that contributed to autism (Table 6.1).

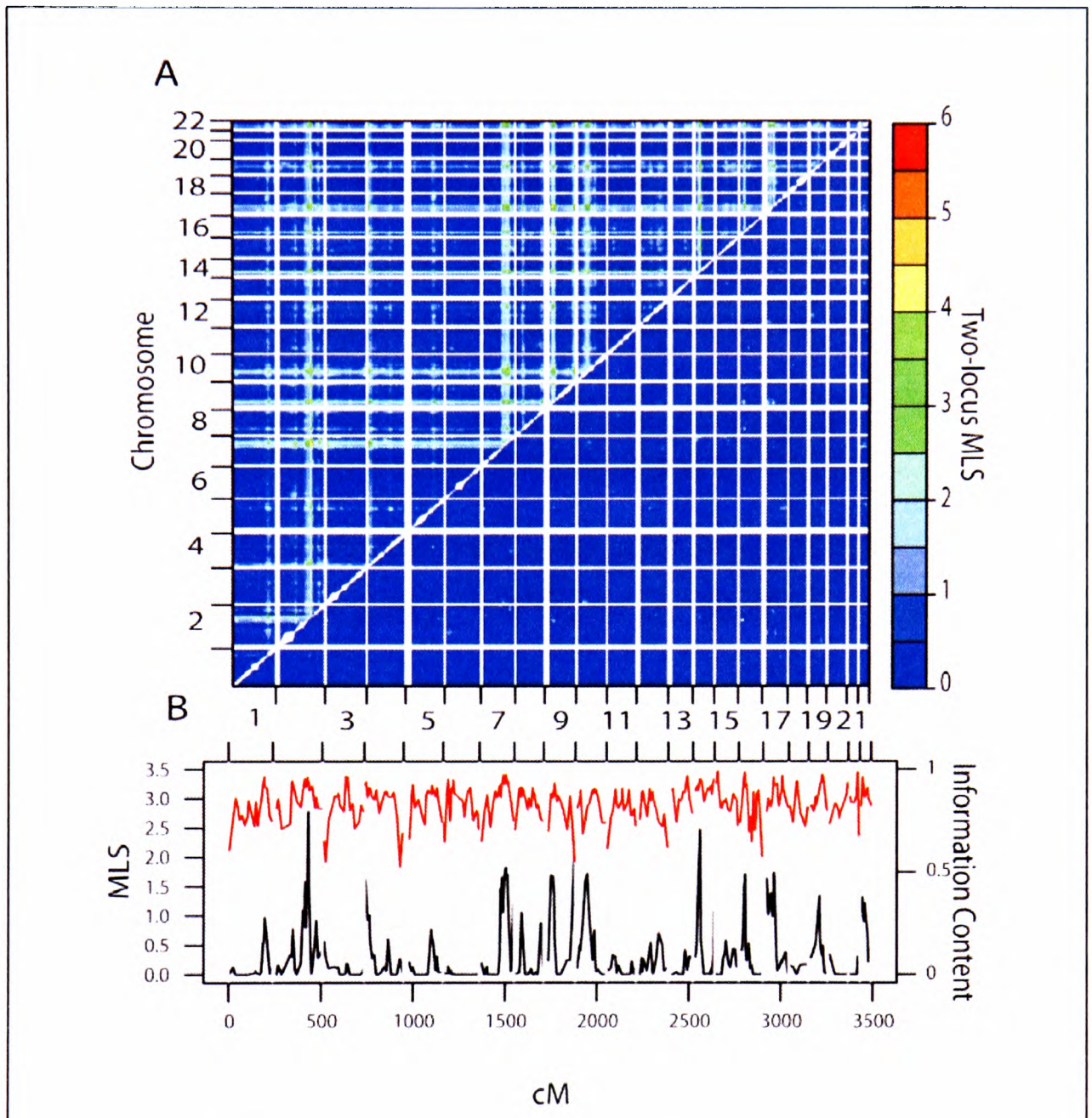


Figure 6.1. Autism genome-wide linkage scans. (A) Two-dimensional (2D) genome-wide linkage scan. Two-locus MLS calculated under the general two-locus model are presented above the diagonal. The difference in the MLS for the fit between the general and additive two-locus models is presented below the diagonal. (B) Single-locus genome scan. The single locus MLS (black line) and the genome-wide information content (red line) for the 83 ASPs in the IMGSAC data.

As discussed in chapter IV it is crucial to determine the appropriate significance thresholds in the context of a 2D linkage scan. The 2D genome-wide simulation results presented in section 5.3 should be helpful in this respect, however,

the simulations were performed on a sample representative of the BRIGHT data in which most parental genotypes were unavailable. In contrast, in the IMGSAC data parental genotypes were obtained in the majority of families. The 2D genome-wide thresholds for the autism 2D scan should be adjusted to reflect the availability of parental data. Ideally, one would perform 2D genome-wide simulations suited to the autism data, similar to those used for the BRIGHT data in section 5.3. However, it was computationally prohibitive to simulate the 1000 2D autism surfaces and instead I used the results from BRIGHT 2D simulations in the autism analysis. To achieve this I inflated the two-locus MLS thresholds obtained from the simulations in section 5.3 by 0.2 Lod units for determining significance in the autism data. The choice of 0.2 Lod units was somewhat arbitrary and was in part based on twice the maximum difference observed in results from single-locus analysis. Wiltshire et al. (2002, Table 3) find that the suggestive and significant single-locus Lod (Kong and Cox 1997) thresholds rise from 1.556 and 2.956 (no parental data) to 1.627 and 2.992 when parental data is available, respectively. Holmans (1993, Tables 2 and 3) reports that the increase in single-locus MLS thresholds across different sizes of tests and numbers of alleles ranges from 0 to 0.071, when parental data becomes available. In two-locus linkage the amount of Lod units inflation in the thresholds probably depends on the position in the test-statistic distribution, i.e. 0.2 may be appropriate for the upper tail of the distribution at $P = 0.05$ and perhaps $P = 0.01$, but for suggestive linkage a smaller inflation should probably be used and 0.2 may be too stringent. The aim here was not to obtain exact significance thresholds, but to choose cut-offs to declare interesting results that are roughly comparable between samples.

Table 6.1. Genome-wide suggestive results from the autism 2D scan.

		Two-locus MLS					
Locus 1[a]	**Locus 2[a]**	**GEN**	**ADD**	**GEN-ADD**	**MUL**	**GEN-MUL**	**ε**
Pairs of loci on different chromosomes							
2q31.3 *D2S2310, 2.8*	14q13.1 *D14S49, 2.5*	5.75	4.59	1.17	5.25	0.5	15.7
2q31.3 *D2S2310, 2.8*	9q34.3 *D9S158, 2.3*	5.4	4.8	0.6	5.08	0.32	2.2
2q31.3 *D2S2310, 2.8*	16p13.1 *D16S3102, 1.7*	5.34	5.31	0.03	4.49	0.85	0
9p22.2 *D9S157, 1.7*	17q11.2 *HTTINT2, 1.7*	5.1	2.62	2.47	3.43	1.67	1000
2q31.3 *D2S2310, 2.8*	7q32.3-33 *D7S640, 1.8*	5.08	5.06	0.02	4.61	0.47	0
9q34.3 *D9S158, 2.3*	16p13.2 *D16S407, 1.5*	5.07	3.07	2	3.83	1.24	1000
2q31.3 *D2S2310, 2.8*	9p22.2 *D9S157, 1.7*	5.06	3.96	1.09	4.47	0.58	354
7q36.3 *D7S550, 1.2*	14q12 *D14S1034, 2.1*	5.05	2.53	2.52	3.25	1.8	1000
2q31.3 *D2S2310, 2.8*	4p16.3 *D4S2936, 1.6*	4.97	4.97	0	4.41	0.57	0
2q31.3 *D2S2310, 2.8*	17q11.2 *HTTPR, 1.7*	4.92	4.41	0.51	4.52	0.4	1.2
9p21.3 *D9S171, 1.7*	14q13.1 *D14S49, 2.5*	4.85	4.31	0.54	4.13	0.71	0.03
9q34.3 *D9S158, 2.3*	14q13.1 *D14S49, 2.5*	4.81	4.6	0.2	4.8	0.04	0.76
2q31.3 *D2S2310, 2.8*	10p11.22 *D10S208, 1.7*	4.77	4.32	0.45	4.5	0.27	2.65
9p21.3 *D9S171, 1.7*	16p13.2 *D16S407, 1.5*	4.75	2.52	2.23	3.2	1.55	1000
9q34.3 *D9S158, 2.3*	17q11.2 *HTTPR, 1.7*	4.7	3.39	1.32	4.03	0.67	1000
9q34.3 *D9S158, 2.3*	22q11.21 *D22S311, 1.3*	4.66	4.66	0	3.61	1.05	0
2q31.3 *D2S2310, 2.8*	22q11.23 *D22S315, 0.9*	4.54	3.78	0.76	3.67	0.86	1000
10p12.1 *D10S197, 1.6*	14q13.1 *D14S49, 2.5*	4.53	4.46	0.07	4.07	0.46	0
Pairs of linked loci							
9p22.2 *D9S157, 1.7*	9q34.3 *D9S158, 2.3*	4.56	3.37	1.2	3.99	0.57	1000

[a] The chromosome location, peak marker under the sparse grid, and single-locus MLS are shown for each region.

At a suggestive linkage threshold of 4.5 and significant linkage threshold of 6 there were 19 pairs of regions showing suggestive evidence for linkage and no region

pairs with genome-wide significant evidence for linkage (Table 6.1). Eighteen of the 19 pairs were unlinked and one interaction on chromosome 9 involved two loci at opposite ends of the chromosome. Of the 18 unlinked peak pairs, six (2q31.3-4p16.3, 2q31.3-7q33, 2q31.3-16p13.1, 9p21.3-14q13.1, 9q34.3-22q11.21, 10p12.1-14q13.1) were approximated by the additive two-locus genetic model (epsilon = 0, GEN-ADD = 0), four (2q31.3-9q34.3, 2q31.3-17q11.2, 9q34.3-14q13.1, 2q31.3-10p11.22) were approximated well either by the multiplicative model (epsilon = 1, GEN-MUL = 0) or a model similar to the multiplicative model (1<epsilon<5, GEN-MUL<GEN-ADD, GEN-MUL<0.5), and nine interacted under a model of strong epistasis (Table 6.1). The 19 pairs consisted of interactions between ten chromosomal regions (regions on the same chromosome were at least 40 Kosambi cM apart). The region with the highest single-locus MLS, 2q31.3, was part of 9 of the 19 interacting pairs. Overall, the same regions participated in most of the pairwise suggestive interactions.

6.2.2 Chromosome 7

There is evidence for the involvement of chromosome 7q in autism supported by results from several independent research groups (IMGSAC 1998; Philippe et al. 1999; Risch et al. 1999). However, the linkage in this region spans a broad interval that may include multiple underlying disease contributing genes. Therefore, chromosome 7q was examined in more detail in the IMGSAC data by collecting additional family data, resulting in a final set of 219 ASPs obtained from 207 nuclear families with 958 individuals. The extra families were genotyped for the original 25 microsatellite markers on chromosome 7 included in the 2D linkage scan. Previous analyses in these data indicated sex-specific and parent-of-origin effects at two closely linked locations under the chromosome 7q peak (Lamb et al. 2005) raising the possibility of the potential involvement of two loci underlying the single-locus peak.

Support for the presence of two autism susceptibility loci on chromosome 7 was assessed using Twoloc (Farrall 1997) and Merloc. Analyses in Twoloc were performed because Twoloc uses Vitesse (O'Connell and Weeks 1995), which can take account of the sex-specific differences in recombination fraction between the two loci, and accuracy in the recombination distances will help evaluate the evidence for linkage more precisely. However, Twoloc was limited to analyses in a maximum of 8 markers. Therefore, Merloc was also applied to this data, but only using sex-averaged recombination rates, because Merlin could not handle sex-specific recombination rates at the time. Analyses in Merloc were performed on top of each of the 25 markers on chromosome 7, and at 3 locations between pairs of markers, so as to form a 2D grid of coordinates

Analyses in Merloc using sex-averaged recombination under the two-locus general model generated a peak MLS of 3.34 at the intercept between marker D7S477, at 111cM, and the interval D7S530-D7S640 at 134cM (Figure 6.2). The MLS generated under a single locus model were 1.94 at D7S477 and 2.51 in the interval D7S530-D7S640 (133cM). Linkage support for the presence of a secondary locus at 111cM, independent of the disease locus at 134cM, can be calculated as the difference, i.e. 0.83, between the two-locus general MLS and single-locus MLS. The data were also analyzed using Twoloc with sex-specific recombination fractions from the deCode genetic map for more accurate resolution. This analysis was restricted to a subset of 8 markers at the peak and resulted in MLS of 3.31 under the two-locus general model, and MLS of 2.13 at 111cM (D7S477) and 2.31 at 134cM (D7S530-D7S640) under the single locus model. The best fit of nested two-locus models was observed for the additive model, with a two-locus additive MLS of 3.31.

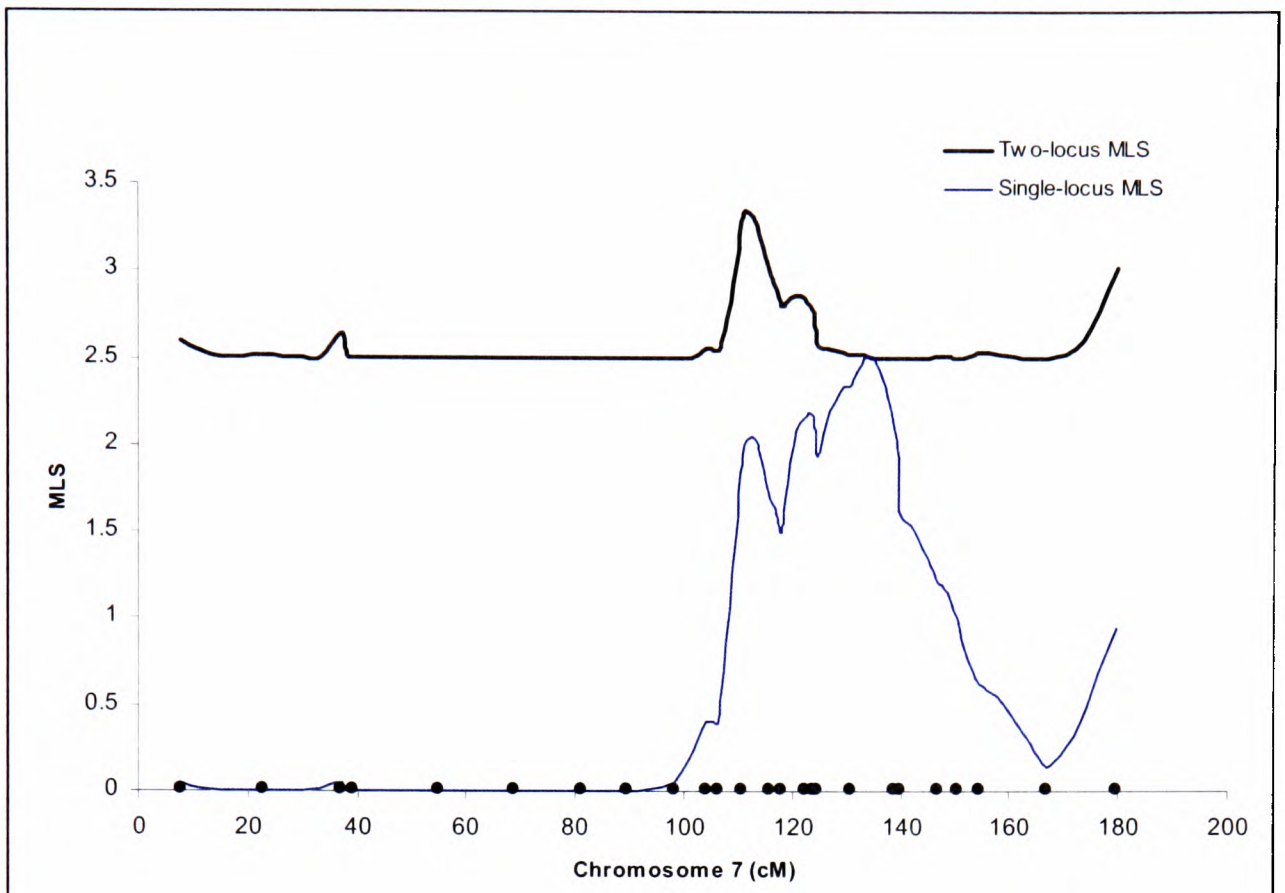


Figure 6.2. Autism chromosome 7 results. Increase in the MLS for a two-locus model, assuming a primary susceptibility locus at 134cM, versus the single-locus results. Results are obtained in 219 ASPs in Merloc assuming deCode sex-averaged recombination rates. The positions of the markers are represented by dots on the x axis.

The single-locus linkage peak on chromosome 7q spanned a broad interval (about 40 cM) and previous analyses have attempted to localize the linkage signal more precisely, by examining the IMGSAC data for sex-specific and parent-of-origin effects (Lamb et al. 2005). Evaluation of parent-of-origin contributions in this region in the entire sample of 219 ASPs lent support to the hypothesis that there may be two discrete loci (separated by $\theta = 0.1$) underlying linkage of autism to chromosome 7, with parent-of-origin specific effects (Lamb et al. 2005). However, the results of two-locus analysis using Merloc suggest that the majority of evidence for linkage comes from locus 2 (D7S530-D7S640), and the additional support for linkage at locus 1 (D7S477), assuming that locus 2 is linked, is only MLS = 0.83. To evaluate the significance of the two-locus MLS in this case (and of GEN-SL1), simulations should

be performed conditional on the observed IBD probabilities under a single-locus model, similar to the approach used in section 2.3.2 in type I diabetes, which is computationally demanding. The small increase in the two-locus findings at 7q, compared to the single-locus results, indicates that the evidence for the involvement of two susceptibility loci under the 7q peak should be interpreted with caution. On the other hand, the two loci are very close together ($\theta = 0.11$ for males, and $\theta = 0.09$ for females) and if there are two genes acting under an additive model at this coordinate the power to detect the loci is likely to be low. The results of Farrall (1997) indicate that the power of detecting the secondary locus in the presence of a major locus (with single locus MLS = 3.3) is just above 50% for an additive model with a test size of 0.001, in a sample of 336 ASPs, and an overall λ_S of 2. If there was a real secondary locus at location 111 cM linked to the primary locus at 134 cM in the autism data, the power to detect it may be less than 50%. Therefore, while there is no strong evidence for the presence of two linked susceptibility loci under the chromosome 7q peak in autism, this hypothesis cannot be excluded.

6.3 Type 2 Diabetes

Type 2 diabetes (T2D) is a common complex disease of glucose homeostasis with a clear genetic predisposition and rapidly increasing prevalence worldwide. There is a wide spectrum of prevalence rates in different populations across ethnic groups. In Caucasians the prevalence of the disease is approximately 2.4% (see Barroso 2005). The evidence for a genetic component in T2D (see section 6.4) is supported by over 30 published genome-wide linkage scans. Evidence for epistasis has previously been reported in T2D in Mexican Americans (Cox et al. 1999) between chromosome 2 and 15 and in Finns (Ghosh et al. 2000) between chromosomes 2 and 20, encouraging a genome-wide search for pairwise interactions in different samples.

The subjects in this study comprised the 573 nuclear pedigrees from the British Warren 2 type 2 diabetes genome scan (Wiltshire et al. 2001), which consisted of 2532 individuals with 756 all possible affected sib-pairs. Parental data were not available in the majority of families. Genotype data were available for 490 microsatellite markers spanning the autosomal genome, which included markers used in the original study (Wiltshire et al. 2001) and additional markers genotyped on 1q, 10q, and on several other regions. The Rutgers genetic map was used in the analysis of the T2D 2D linkage scan. The resulting average marker spacing was approximately 7.22 cM across the genome with the largest inter-marker distance of 20.1 cM on chromosome 18. Chromosomes 1 and 10 were examined in more detail and the results were compared to those obtained with the Marshfield genetic map (chapter IV).

6.3.1 2D scan

One-dimensional and two-dimensional genome linkage scans were performed in the Warren 2 data (Figure 6.3). The 1D genome-wide results identify several peaks of suggestive linkage (single-locus MLS > 1) on chromosomes 1 (MLS = 1.34), 2 (MLS = 1.13), 5 (MLS = 1.55), 6 (MLS = 2.17), 8p (MLS = 1.21), 8q (MLS = 2.08), 9 (MLS = 1.19), 10 (MLS = 2.47), 11 (MLS = 2.10), 12 (MLS = 1.24), and 22 (MLS = 1.43). Overall, no region reached genome-wide significance criteria as defined by Lander and Kruglyak (1995). The 2D genome-wide linkage scan was performed by computing the two-locus MLS under the general model at each pair of markers in the genome. Several peaks in the 2D surface identified pairs of loci that contributed to T2D (Table 6.2).

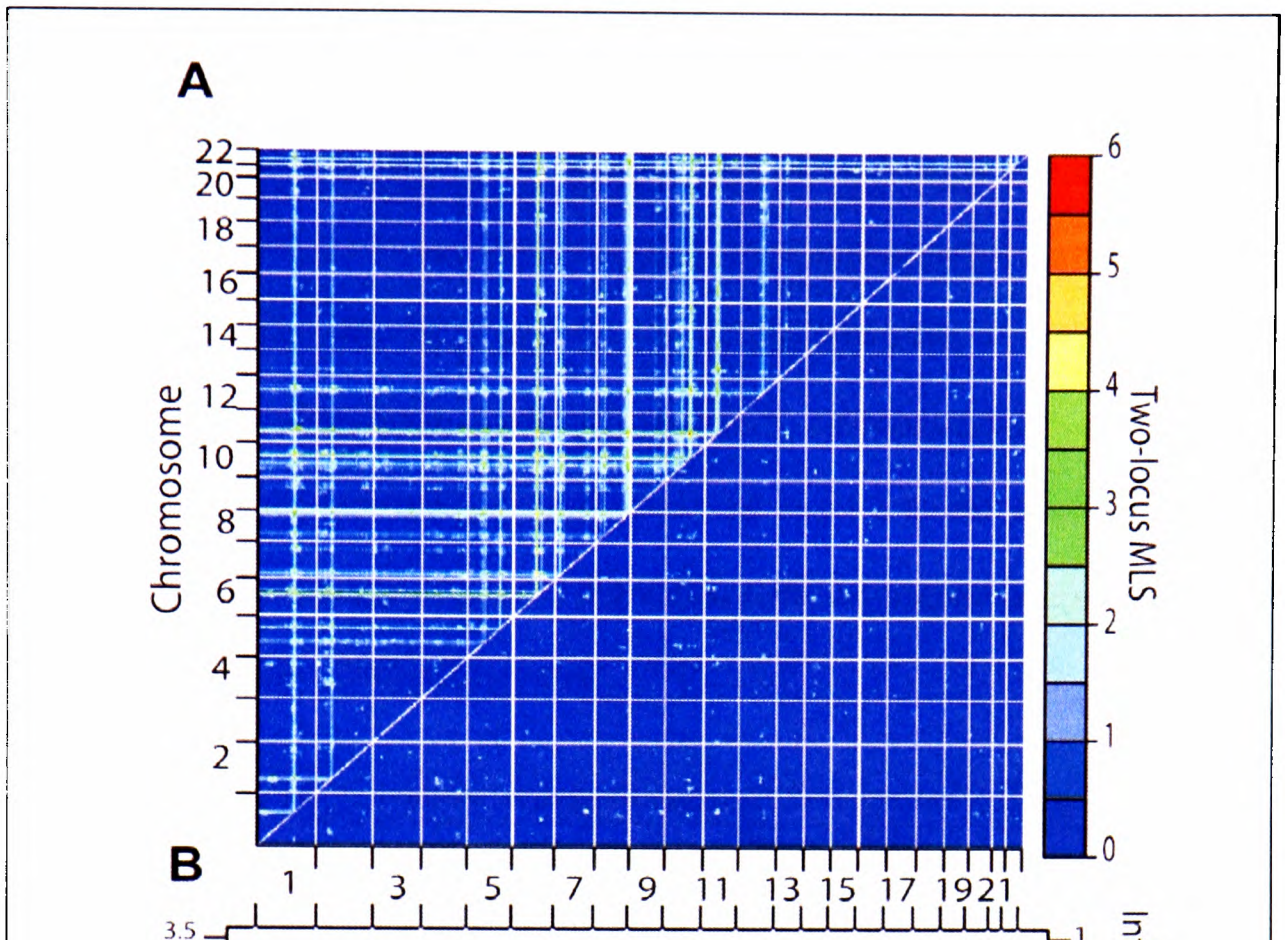


Figure 6.3. T2D genome-wide linkage scans. (A) Two-dimensional (2D) genome-wide linkage scan. Two-locus MLS calculated under the general two-locus model are presented above the diagonal. The difference in the MLS for the fit between the general and additive two-locus models is presented below the diagonal. (B) Single-locus genome scan. The single locus MLS (black line) and the genome-wide information content (red line) for the 756 ASPs in the Warren 2 data.

To determine significance criteria in the T2D 2D scan the MLS thresholds from the BRIGHT simulations (section 5.3) were used, because the missing data patterns were similar between the BRIGHT and Warren 2 data. At a suggestive linkage threshold of 4.3 there were 13 unlinked pairs of regions showing suggestive evidence for linkage and no pairs with genome-wide significant evidence for linkage. No evidence of two-locus linkage was obtained for linked pairs of loci. The thirteen

pairs of interactions consisted of 12 regions. Of the 13 pairs, one (8q24.21-10q23.31) was approximated very well by the multiplicative model (epsilon = 1, GEN-MUL = 0), two (6q16.3-10q23.31 and 6q16.3-11p13) were approximated well by a model where the maximum-likelihood estimates of epsilon suggested a model of epistasis similar to the multiplicative model (1<epsilon<5, GEN-MUL<GEN-ADD, GEN-MUL<0.5), and the remaining 10 pairs interacted under a model of strong epistasis (Table 6.2). Chromosome 10q23.31, the region with the highest single-locus MLS, was part of 9 of the 13 interacting pairs.

Table 6.2. Genome-wide suggestive results from the T2D 2D scan.

		Two-locus MLS						
Locus 1[a]	**Locus 2[a]**	**GEN**	**ADD**	**GEN-ADD**	**MUL**	**GEN-MUL**	**ε**	***P* value[b]**
10q23.31 *D10S1765, 2.5*	11p13 *D11S914, 2.1*	5.19	4.27	0.91	4.55	0.64	24	0.13
5q32 *D5S436, 1.6*	8q24.23 *D8S272, 1.5*	4.94	2.74	2.2	3.01	1.93	1000	0.21
1q24.2 *D1S452, 1.3*	10q23.31 *D10S1765,*	4.92	3.52	1.4	3.79	1.12	411	0.22
6q16.3 *D6S434, 2.2*	10q23.31 *D10S1765,*	4.72	4.54	0.18	4.64	0.08	2.84	0.31
2p21 *D2S391, 0.9*	10q23.31 *D10S1765,*	4.66	3.1	1.56	3.34	1.33	303	0.34
8q24.21 *D8S284, 2.1*	10q23.31 *D10S1765,*	4.55	4.52	0.03	4.55	0	1.06	0.4
5q11.2 *D5S1968, 0.75*	10q23.31 *D10S1765,*	4.5	3	1.5	3.22	1.28	1000	0.43
8p21.3 *D8S258, 1.2*	10q23.31 *D10S1765,*	4.49	3.39	I.1	3.62	0.86	1000	0.43
8q24.21 *D8S284, 2.1*	11p13 *D11S914, 2.1*	4.47	4	0.47	4.18	0.29	9.25	0.44
10q23.31 *D10S1765, 2.5*	12q21.33 *D12S351, 1.2*	4.43	3.49	0.94	3.71	0.72	164	0.47
7p15.3 *D7S493, 0.9*	10q23.31 *D10S1765,*	4.42	3.22	1.2	3.34	1.08	19.6	0.47
6q16.3 *D6S434, 2.2*	7q34 *D7S684, 0.7*	4.37	2.64	1.73	2.86	1.5	1000	0.51
6q16.3 *D6S434, 2.2*	11p13 *D11S914, 2.1*	4.32	4.14	0.18	4.25	0.07	3.84	0.54

[a] The chromosome location, peak marker under the sparse grid, and single-locus MLS are shown for each region. [b] The genome-wide P value corresponds to the two-locus general model MLS and is calculated using thresholds calculated from the BRIGHT simulations.

The results obtained for the interaction between chromosomes 1 and 10 were compared to the results from chapter IV for the British data. The findings in chapter IV were obtained under the Marshfield genetic map, while the results in this chapter are calculated under the Rutgers genetic map. The data for comparison consisted of the same families and genotypes (the Warren 2 data), but the diagnostic criteria for the data in this chapter were updated, providing additional ASPs for analysis. The single-locus peaks were reduced from chapter IV, on chromosome 1 the peak changed from 1.56 at D1S196 (164.3 Mb) under the Marshfield map to 1.32 at D1S452 (167.3 Mb) under the Rutgers map, and the single-locus peak on chromosome 10 changed from 2.66 at D10S1765 under the Marshfield map to 2.47 at the same location under the Rutgers map. The two-locus peak slightly changed in location (from D1S196 and D10S1765 in Marshfield to D1S452 and D10S1765 in Rutgers) and its magnitude reduced (5.45 in Marshfield to 4.92 in Rutgers) under the Rutgers map reflecting the reduction in the single-locus results. These findings once again emphasize the sensitivity of multipoint linkage analysis to map specification, both in the single-locus and two-locus case.

6.4 Comparison of 2D scans across complex traits

It is of interest to compare the results of 2D simultaneous linkage scans across different complex traits. The 2D scan results from autism, hypertension (HT), and type 2 diabetes (T2D) were examined in an attempt to identify traits that have successful outcomes in 2D linkage scans and explore their characteristics, in particular estimates of genetic contribution to the aetiology of each disease. Complex traits with evidence of a strong genetic component and multiple contributing genes would presumably do well in a simultaneous search.

In order to achieve this, estimates of the magnitude of the genetic contribution in each disease were obtained. Of the three traits examined in this chapter and chapter V, autism, a relatively rare and early-onset trait, appears to have the best evidence for a strong genetic component. A decade ago autism was used as an example of a complex trait with very high estimates of a familial component. High heritability estimates of 80-90%, MZ:DZ concordance ratio of 15-25, and sibling recurrence risk ratios of 75-100 were reported (Smalley et al. 1988; Risch et al. 1999). However, more recent studies (see Szatmari et al. 1998; Chakrabarti and Fombonne 2005) report lower estimates, resulting in λ_S of around 15. This is possibly because the prevalence of the trait in the general population is higher than previously thought, which perhaps reflects better diagnostic criteria and social awareness of the trait. Another factor which contributes to these differences is the variation in sample size across studies (see Szatmari et al. 1998). Following the most recent review of family studies (Szatmari et al. 1998) relative risk to first-degree relatives is 2.2% (95% CI = 1.1-3.3%), that to second-degree relatives is 0.18% (95% CI = 0.03-0.33%), and that to third degree relatives is 0.12% (0.01-0.23%). The high MZ:DZ concordance ratio and the rapid fall-off in the relative recurrence rate of autism with genetic relatedness suggests that multiple genetic factors contribute to the trait. Multiple gene models have been fitted to data and a model of 3-4 loci gave a good fit (Pickles et al. 1995), and models with 15 or more genes have also been suggested (Risch et al. 1999).

In contrast, the two other traits examined - essential hypertension and type 2 diabetes are both late onset common complex traits with rising prevalence worldwide, different disease frequencies across populations, and potentially significant involvement non-genetic factors. Hypertension is a complex trait in which genetic and environmental factors are likely to both play an important role. Genetic factors are

implicated, but estimates of the risk ratio to siblings of affected individuals are modest (1.5 to 3.5) and heritability estimates range from 30% to 50% in the UK (Ward 1990) and up to 57% (Levy et al. 2000) elsewhere in Europe. Twin studies report MZ correlation of 0.55-0.81 and DZ correlation of 0.25-0.54 for blood pressure (systolic or diastolic; see Ward 1990) and concordance rates for diagnosis of hypertension of 0.44 for MZ twins and 0.25 for DZ twins (Williams et al. 2004a). In type 2 diabetes there is also evidence for the involvement of both genetic components and environmental factors. Twin studies report MZ concordance rates of 0.2-0.91, and DZ concordance of 0.1-0.43 (see Barroso 2005) and heritability estimates of around 79% (Kaprio et al. 1992) in European and US populations. Sibling recurrence-risk ratio estimates range from 1.2-1.8 in US populations (Weijnen et al. 2002) to 3.5 in European populations (Rich 1990), and risk ratio to second-degree relatives in European populations is reported to be 1.5 (Rich 1990). In both HT and T2D, there is at least some evidence that environmental factors contribute to these traits from adoption studies or by examining the difference in concordance between DZ twins and siblings, while in autism this difference is less pronounces and environmental factors are likely to play less of an important role in comparison.

Table 6.3 Comparison of unlinked pairwise suggestive results across 2D scans.

Trait	ASP	K_p	H^2	λ_R-1 decrease[a]	Number of 2D peaks	Number of regions[b]	Maximum peaks per region (region)
Autism	83	0.002	0.8-0.9	15.13	18	10	9 (2q31.3)
Type 2 diabetes	756	0.05	0.7-0.8	5	13	12	9 (10q23.31)
Hypertension	2076	0.05	0.3-0.6		10	10	6 (5q13.1)

[a] The estimates were obtained across different relatives pairs from Szatmari et al. (1998) for autism (averaged across 3rd to 2nd, and 2nd to 1st degree relatives) and from Rich (1990) for T2D (one estimate from 2nd to 1st degree relatives). [b] This column represents the number of independent regions (separated by at least 38 Kosambi cM) that are observed in the suggestive 2D peaks.

At a first glance it appears that the number of 2D linkage pairs indicating evidence for suggestive linkage is higher in autism compared to hypertension and T2D, perhaps reflecting the evidence for a stronger genetic component in autism (Table 6.3). However, it is difficult to appropriately interpret this finding, because a number of factors complicate the comparison. First, the sample sizes vary greatly between the three traits examined and will impact the results. While under no linkage ASP sample size may not greatly affect the distribution of Lod scores, under linkage sample size should affect power to detect the two genes, similar to findings for single-locus ASP analysis (see Risch and Merikangas 1996; Cordell 2001). For a more suitable comparison one may repeat the analyses by only examining 83 ASPs in each of the HT and T2D scans and dropping the parental genotypes from the autism scan. However, the genetic effect sizes expected in these traits are small and it is possible that all the findings obtained in the 3 2D scans are false positives, because none of the scans showed genome-wide significant evidence for linkage (the only borderline result was the interaction on chromosome 16 in hypertension, depending on how one interprets the findings from linked pairs of regions). Second, I only used approximate significance cut-offs for declaring genome-wide suggestive linkage in a 2D scan when parental genotypes were available, and these thresholds may not be entirely correct. Third, the information content in the autism data was much higher than in the other two data-sets (as a consequence of the availability of parental genotypes), which will also complicate the comparison. Finally, in the T2D data, a higher proportion of siblings were not independent (came from sib-ships of 3 or more affected sibs), compared to HT and autism. Comparison of single-locus multipoint Lod (Kong and Cox 1997) and MLS statistics in the Warren 2 data highlighted several regions (including 11p13 and 12q21), at which the difference in MLS and Lod results is

pronounced. As discussed in chapter II, one could introduce a weighting criterion in the calculation of the two-locus MLS to take the ASP non-independence into account. Single-locus analysis that incorporates weighting has been shown to produce conservative results, and the most appropriate weighting scheme remains to be established (Greenwood and Bull 1999).

The evidence for a stronger genetic component in autism is reflected by the single-locus genome-wide MLS results, where there are higher single-locus MLS scores obtained in autism (peak single-locus MLS of 2.78, 2.47, and 2.3), compared to T2D (peak MLS of 2.47, 2.17, and 2.1), and HT (peak MLS of 2.54, 1.6, and 1.56). The results from the 2D scans indicate that regions with strong single-locus evidence for linkage are often found in the peak suggestive 2D findings. This is not surprising, because such regions may harbour real susceptibility genes, and the two-locus MLS for interactions involving such regions will always be equal to or higher than the single-locus MLS, suggesting that different null hypotheses should be examined. I think that for future analyses of 2D genome-scan linkage data the general two-locus test-statistic should still be used to identify peak coordinates based on whether that statistic surpasses a pre-defined threshold. However, the next stage should be to assess the significance of the increase in the 2D MLS over the single-locus MLS for each region in the peak pairs, i.e. assessing the significance of GEN-SL1 and GEN-SL2 by simulation. This will be especially relevant for interactions involving at least one region with strong main effects.

As discussed previously, the two goals of a 2D scan are to identify susceptibility regions and to narrow down the type of interaction for each pair of contributing loci. Assessing the significance of GEN-SL1 and GEN-SL2 would address the first goal and quantify the evidence for a susceptibility locus in a given

region, in the presence of interactions. Once, this is achieved the exact model of interaction may be examined by testing the fit of different nested two-locus models (as shown in chapters IV and V). In autism, for example, the effect sizes of the underlying regions are higher than in the other two traits, but there appears to be less evidence for strong epistasis across the genome. This is illustrated by the fact that on the general two-locus MLS surface in Figure 6.1A, most of the peaks occur on ridges, i.e. most of the interacting regions have at least some single-locus evidence for linkage. However, it is the additive or multiplicative two-locus models (and models that have low values of epsilon) that seem to fit most of the pairwise interactions ('strong-epistasis' fit only 9 of the 19 pairs) as seen below the diagonal on Figure 6.1A. In contrast, in HT and T2D there were fewer regions with strong single-locus evidence for linkage, but more regions interacted under a strong-epistatic model in the peak 2D results (9 of 10 in HT and 10 of 13 for T2D). The 2D surfaces for these traits show a less discernible pattern, although the majority of peaks still occur on ridges.

An estimate of the contribution of epistasis in a complex trait may be obtained from the rate of decrease of $\lambda_R - 1$ in sets of relatives of decreasing degrees of relatedness and the MZ:DZ twin concordance ratio (Risch 1990a). Most studies modelling relative recurrence risks have assumed a multiplicative model in multiple loci of equal effects (Farrall and Holder 1992; Risch et al. 1999). As illustrated in Figure 2.1 the rate of fall-off in $\lambda_R - 1$ will depend on the intensity of epistasis in the underlying model, suggesting that the number of loci may be overestimated if intense epistasis exists, and unequal gene effects may exist. Relative recurrence risk measures could only be obtained for autism and T2D, and were nt available for HT. The rates of drop-off in $\lambda_R - 1$ were approximately 15.13 for autism, and 5 for T2D (Table 6.3), indicative of a complex multilocus aetiology. The MZ:DZ concordance rates across

the three traits were about 15-25 for autism and around 2 for T2D and HT (estimates varied according to population examined). From the 2D linkage results it appears that while more pairs of interacting regions are involved in autism (at a similar level of significance as for the other traits), most of these regions acted together under an additive or more simplistic genetic model, instead of models of strong epistasis as suggested by the MZ:DZ concordance and $\lambda_R - 1$ drop-off. It is possible that higher-order interactions are involved in autism in scenarios where several regions act additively and are connected to complexes which interact epistatically (Figure 6.4).

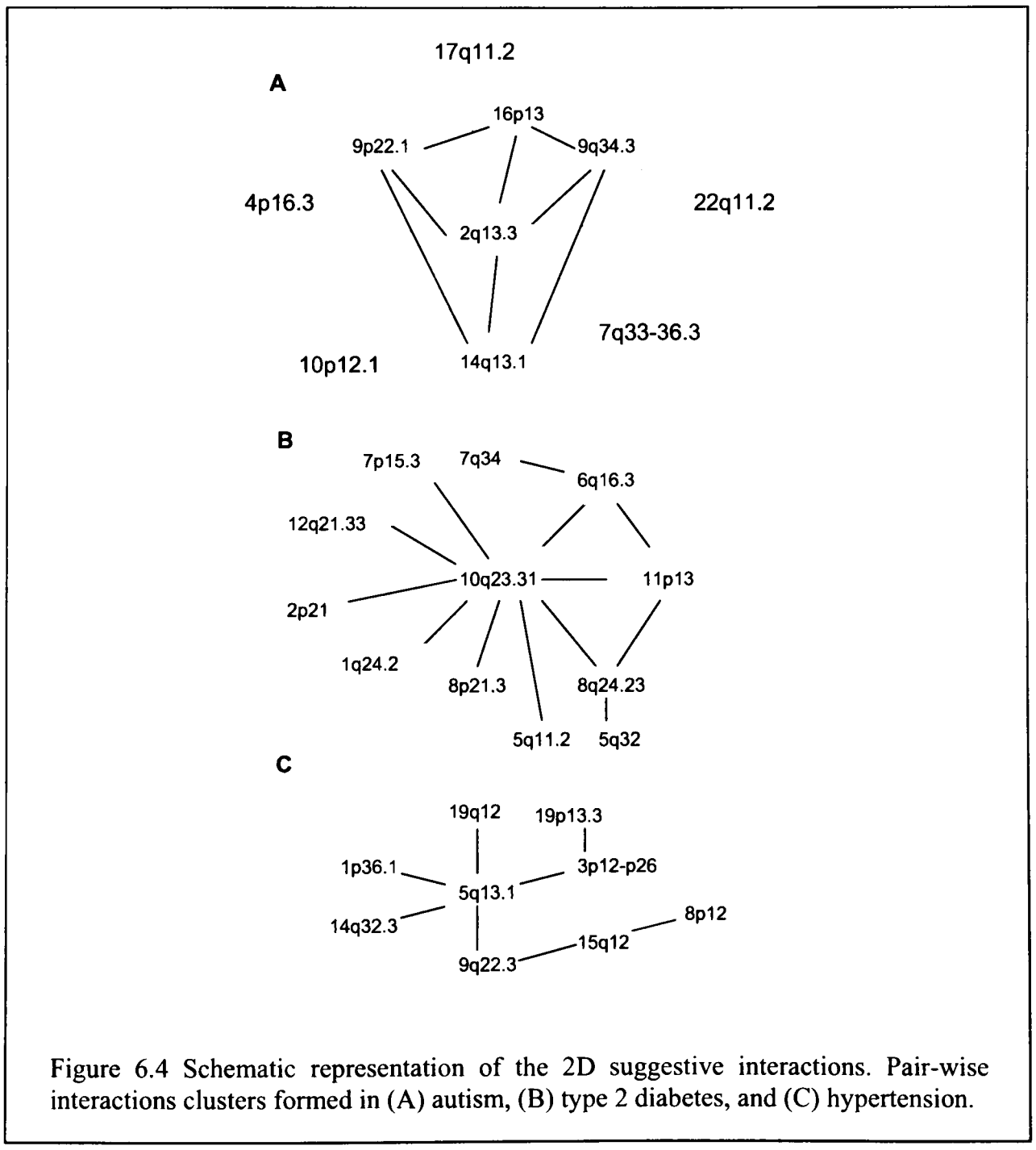


Figure 6.4 Schematic representation of the 2D suggestive interactions. Pair-wise interactions clusters formed in (A) autism, (B) type 2 diabetes, and (C) hypertension.

The possibility of obtaining large two-locus Lod scores in the absence of strong evidence for linkage at the single-locus level should be examined in more detail in multilocus linkage mapping. As discussed in chapter V, others have suggested that this finding is unlikely to occur often and have recommended restricting analyses to regions with at least some evidence of linkage (Holmans 2002). However, these studies used different methods of incorporating interactions, for example, Holmans (2002) tested for linkage at locus 1 allowing for interactions at a second locus using logistic regression, rather than testing both loci jointly. On the other hand, two-dimensional scans in model organisms do find significant evidence for pairwise interactions in the absence of main effects. The majority of these scans use the genotypes or genotype scores, rather than IBD sharing data. It is possible to obtain large epistatic effects in the absence of single-locus effects in association analysis (Culverhouse et al. 2002), but for multilocus linkage methods this may depend on whether one performs conditional (Cox et al. 1999; Holmans 2002) or joint (Cordell et al. 1995, present study) two-locus linkage analysis. It has been shown for linkage analysis that even if the true disease model is one of parwise epistasis with no-main effects, the expected IBD probabilities will be distorted at the contributing loci and there will be at least some linkage signal in the single-locus scans (Culverhouse et al. 2002). Conditional and joint analyses take these into account differently and it would be interesting to examine this difference formally, for example by adding a pairwise no-main effects epistatic model (as defined by Culverhouse et al. 2002) to the two-locus simulation models studied in chapter III.

Another comparison of interest would be to contrast the 2D results from this study to results from applications of conditional linkage analysis methods in search for pair-wise interactions across the genome. Only one study (Chang et al. 2006),

which was recently published, examined complete genome-wide evidence for interactions in prostate cancer using ordered-subset analysis. In the prostate cancer 2D linkage scan the top six genome-wide epistatic pairs consisted of 12 different regions, none of which included the single-locus peak, unlike the results from this thesis. Other groups have also looked at genome-wide approaches to detect interaction, but have not exactly performed a genome-wide search, instead restricting the number of interacting pairs considered either by only considering regions with some single-locus evidence for linkage (NPL > 2; Nath et al. 2001; Zandi et al. 2001), or pairs of regions which show significant correlation in the distribution of familial NPL scores (Colilla et al. 2001). These three studies examined data from two asthma genome-wide scans, and although the trait and samples were similar or identical there was some variability in the results. Due to the study design, the results from Nath et al. (2001) and Zandi et al. (2001) always involved at least one region with single-locus evidence for linkage, but Colilla et al. (2001) found that the strongest familial correlations were obtained for regions which had very weak single-locus effects. Overall, the same regions appeared more than once in the peak pair-wise results from each study, but there seems to be no overlap in pair-wise results between studies.

Tremendous effort over the past decade has generated databases of genotype data for human families affected by complex diseases. These data have been predominantly analysed using single disease gene models, and have found little consistent evidence for the involvement of particular genes. Two-dimensional linkage approaches can add considerable value to such genome-scan data by identifying loci that interact epistatically.

CHAPTER VII. Extensions and Conclusion

The aim of this thesis was to examine epistasis in linkage analysis in human complex traits. This was achieved by reviewing linkage methods that consider pair-wise epistasis in chapter I, extending a two-locus non-parametric approach in chapter II, comparing different multilocus linkage methods in chapter III, and two-locus linkage analysis of genome-scan data in chapters IV, V, and VI. I believe that there are many directions in which this study can be extended and improved. Two immediate extensions follow from this work – extensions to examine epistasis among more than two loci in linkage analysis and extensions to pair-wise association in searching for interacting disease variants. I will discuss these in the next two sections.

7.1 Linkage approaches to examine more than two genes

Complex traits likely involve interactions among more than two loci, so it is of considerable interest to extend the search to multilocus models. However, higher dimension scans become statistically and computationally prohibitive. For certain traits it has been argued that a two-dimensional search can capture a significant amount of the underlying trait complexity (Sen and Churchill 2001). On the other hand, when higher-order interactions exist, it is best to use the correct underlying multiple gene model (Marchini et al. 2005). Multilocus linkage methods in humans have considered 3 loci (Cordell et al. 2000) and may be extended to higher order interactions. One possibility of searching for higher order interactions without performing simultaneous linkage scans is to model the expected contribution of each gene and combine the estimated effects and single-locus IBD probabilities in a multilocus model.

Initially, an estimate of the number of contributing loci and the underlying distribution of their gene effect sizes need to be specified. Results from QTL mapping experiments across organisms are consistent with an exponential distribution of gene effect sizes. It appears that for most traits relatively few genes of moderate to major effects account for most of the genetic variance and a large number of loci of minor effects explain the remaining genetic variance, supporting oligogenic or polygenic models of disease. Studies by Hayes and Goddard (2001), Zeng (1992), Xu (2003), and others have examined the distribution of gene effects empirically and by simulations and have established that the distribution of QTL effects appears to fit an L-shaped gamma distribution, with parameters specific to the trait and organism considered.

To obtain an estimate of the number of contributing loci several approaches have been proposed. In qualitative traits, Farrall and Holder (1992) introduce a maximum-likelihood method in the framework of Risch (1990a) to determine the optimal number of genes underlying a trait. A similar method is that of Risch et al. (1999) applied to autism. However, there are a number of parameter restrictions associated with this analysis, demonstrated by Schliekelman and Slatkin (2002) who expand the work of Craddock et al. (1995), and propose a different likelihood estimation of the number of underlying loci. It should also be noted that these approaches assume that all contributing loci act under a multiplicative model and have equal gene effect sizes, while the underlying distribution of gene effects is probably not uniform and intense epistasis is likely to exist. If there is quantitative information on the trait one can also use extensions of the Castle-Wright estimator (reviewed in Lynch and Walsh 1998) by Zeng (1992), also applicable to outbred populations (Lande 1981), or the estimator proposed by Otto and Jones (2000), for both of which

the expected distribution of gene effects must be specified, but again these estimators assume no epistasis among the underlying genes. Finally, for particular complex traits, such as hypertension for example, the number of additive contributing genes has been estimated with respect to inbreeding using a blood pressure dataset (Rudan et al. 2003), assuming that hypertension is mediated by recessive or partially recessive QTL alleles, which would be influenced by the increased homozygosity found in inbred populations. The model of distribution of locus effects suggests that 8-16 QTL of larger effect together account for a maximum of 25% of the variance, while the remaining 75% of the variation is mediated by at least 400 QTL of very small effects.

The next stage would be to combine the individual single locus effects into a multilocus model. Assuming a multiplicative genetic model the multilocus MLS for unlinked loci is the sum of the individual single-locus MLS. However, computing the full multi-locus likelihood under an additive or general model would be more complicated. One could assume a multiplicative multi-locus model with 15 genes contributing to hypertension. The gene effect sizes would be drawn at random from the exponential distribution with specified parameters, and the number of genes and allele frequencies would have to satisfy the parameter constraints illustrated by Schliekelman and Slatkin (2002). Another constraint would be imposed with respect to the overall λ_S for the trait. If all the genes acted under a multiplicative model, the individual locus λ_{Si} multiplied together give the overall λ_S for the trait. Under the additive model the individual locus λ_{Si} could sum with a weighting factor to give the overall λ_S (or instead of λ_S using the deviation of sharing 0 alleles IBD from the null (δ_{S0}): then the sum of the individual δ_{S0i} could give the overall δ_{S0}). Under the general model the individual locus λ_{Si} might vary between 1 and the overall λ_S. One may want to evaluate either the fit of different draws from the exponential distribution to

evaluate the number and effect sizes of genes underlying the trait (see Risch et al. 1999; Pritchard 2001), or evaluate the fit of the multiplicative and additive models compared to the general model in the multi-locus case. However, the calculation of the overall likelihood of the multi-locus model would assume a multiplicative model, in the first instance at least, because it is not straightforward to combine the single locus effects into a multi-locus MLS under the additive or general models.

Another possibility of searching for higher-order interactions would be to take a stepwise-conditioning search. For example, Storey et al. (2005) explore 2D linkage strategies in yeast, and propose a new 'sequential search' method of simultaneously mapping multiple loci, which appears to be more powerful than the exhaustive 2D linkage scan. Their novel approach is specific to analysis of multiple (quantitative) phenotypes in haploids and measures the probability that a locus is included in the true genetic model given the data, without necessarily specifying the true genetic model. The method is based on model selection algorithms, where the goal is to identify the most likely subset of loci in the best model, and combines them with composite interval mapping methods. These ideas could be extended to diploid organisms in quantitative trait linkage mapping studies in forward stepwise regression analysis.

The concept of a sequential search can also be extended to the method used in this thesis. For a sample of ASPs for a discrete trait, one may begin the search for higher-order interactions by first performing a 2D linkage scan. The pair-wise peaks surpassing a liberal threshold could then be selected and each pair-wise combination of loci underlying a 2D peak may be collapsed into a new single 'super-locus'. The next stage would be to test all pair-wise interactions between the new super-loci and the rest of the genome; then collapse the peak pair-wise findings again, and continue

to perform series of incomplete 2D scans and collapse peak pair-wise findings into new 'super-loci'. This idea is related to that of Zeng (1994) in model organisms where the search for epistasis is performed in a series of 1D genome-wide scans with covariates. The difficult part would be to determine how to collapse the peak two-locus findings into new 'super-loci'. In Merloc, it would be the joint marker IBD probabilities that would have to be collapsed. For example, if IBD_{ij} is the two-locus IBD probability that an ASP share *i* alleles at locus 1 and *j* alleles at locus 2, and IBD_k is the single-locus IBD of sharing k alleles at a 'super-locus', two methods of collapsing two-locus IBD might be considered. In method 1 one would collapse IBD_{00} to IBD_0, IBD_{ij} to IBD_1 (where i=0 and j=1,2 or i=1,2 and j=0, or i=j=1), and IBD_{22} to IBD_2; method 2 would be to collapse IBD_{ij} (where i=0 and j=0,1, or i=0,1 and j=0) to IBD_0, IBD_{ij} to IBD_1 (where i=0 and j=2, or i=2 and j=0, or i=j=1), and IBD_{ij} (where i=2, and j=1,2, or i=1,2 and j=2) to IBD_2. One would have to extensively test the different collapsing methods by simulation, for example by starting with a 3L model in chapter III and testing how the different collapsing methods perform.

Finally, there are multi-locus linkage methods that can be applied to higher order simultaneous scans. The method in this study can be extended to 3 or more loci (see Cordell et al. 2000) and other methods may be extended as well, preferably using joint modeling of multiple loci rather than conditional linkage approaches. For example, the conditional logistic likelihood approach of Olson (1997; 1999) can take into account multiple covariates (which could be genetic factors), as could the logistic regression approach of Holmans (2002); and neural networks have also been used to detect multi-locus involvement in linkage (Lucek et al. 1998). The number of parameters will increase dramatically with each dimension and appropriate significance thresholds, accounting for multiple testing, need to be established at each

level of complexity. It would also be computationally challenging to establish the multi-locus joint IBD probabilities for more than 2-3 loci, unless they are approximated by multiplying the marginal single-locus IBD probabilities (see Cordell et al. 2000). It remains to be established whether simultaneous multi-locus linkage scans for more than 2-3 loci will contribute much to the understanding of disease genetics.

7.2 Association studies

In single locus analysis association tests appear to have more power to detect susceptibility variants of modest genetic effects relative to linkage analysis (Risch and Merikangas 1996), specifically for association using TDT and linkage analysis in ASPs. By extension a similar increase in power should be observed in two-locus analyses using association over linkage. For example, Tang and Siegmund (2002) observe that there are limits to the increase in power from correctly modelling epistasis in non-parametric linkage over single-locus linkage, but argue that this increase in power will be much more substantial in association tests, because association analysis examines the direct correlation between genotype and phenotype. Moreover, Culverhouse et al. (2002) show that association can detect large epistatic effects in the complete absence of main effects. Finally, a study of genome-wide search strategies for multi-locus association tests indicated that methods which incorporate epistasis are more powerful than single-locus approaches, but the relative power of the simultaneous compared to the conditional search strategy depends on the exact underlying genetic model of interaction, disease allele frequencies, and LD between marker and disease locus (Marchini et al. 2005).

Tests of association aimed at detecting pair-wise contributions from variants at two unlinked loci have been developed for both case-control and family-based

samples. The family-based methods consider association of two variants in the presence of linkage, such as the TDT two-locus extensions. Another study related to these methods is the IBD regression of Holmans (2002) at one locus conditional on genotypes at the second locus. Holmans (2002) shows that this test is more powerful than the IBD regression conditional on IBD status at the second locus, which is a linkage only test. It is possible to extend the method considered in this thesis in a similar manner to include tests of association in the presence of two-locus linkage. One simple way of testing this would be to estimate how the evidence for linkage is partitioned according to genotype at candidate SNP marker(s) in the region(s) of linkage (see Horikawa et al. 2000). The two-locus MLS can be calculated in the entire sample and in sample subsets by partitioning the number of ASPs according to their genotype (or joint genotypes if two candidate SNPs at the two putative loci are considered). The observed proportions of the two-locus MLS statistics in the genotype partitions could then be compared to those expected under the null hypothesis.

It would be perhaps more interesting to extend this idea to the work of Sun et al. (2002) and Li et al. (2005), which formally test whether a SNP explains the evidence of linkage in single locus analysis. In the two-locus case one would then test whether the pair-wise combination of SNP genotypes explains (either completely or to some extent) the two-locus linkage signal. One possibility to do this would be to introduce a weighting factor into the two-locus MLS calculation, which weighs the contribution of the ASP to the MLS. The weights may reflect which genotype configuration (for the 3-by-3 two-locus SNP genotypes) the sib-pair falls into, and as such would be similar to the idea of stratifying the linkage signal by Horikawa et al. (2000). Another possibility may be to extend the method of Li et al. (2005) directly to two-loci by modelling the LD between the SNP marker and the unobserved disease

variant. This would be more complicated to extend because one would have to model the LD between the SNP and the unobserved disease variant at each locus, as well as the interaction between the two SNPs.

One could extend this idea further into analysing genome-scan data with current SNP linkage panels that consist of tens of thousands of SNP markers. First, a two-dimensional linkage scan across the genome would be performed using the SNP linkage scans and accounting for the LD between the SNP markers for example by clustering together markers that are in pair-wise LD measured above a certain r^2 threshold. At this stage one could use the extension of Merlin that can model LD (Abecasis and Wigginton 2005), and then calculating the two-locus MLS is straightforward because Merloc uses Merlin to calculate family likelihoods to establish two-locus IBD probabilities. The next stage in the analysis would be for each pair of chromosomes to drop each pair-wise SNP combination from the linkage analysis (where SNP1 is on the first chromosome and SNP2 is on the second chromosome in the pair) and test whether that combination of SNP markers explains all (or some) of the two-locus linkage signal. The latter would be achieved either by introducing the weighting in the two-locus MLS calculation to take into account the SNP joint genotypes, or by extending the work of Sun et al. (2002) and Li et al. (2005) as discussed above.

There are other methods for detecting epistasis in association tests, which have been extended to more than two loci, mostly using data reduction strategies (see Hoh and Ott 2003; Jansen 2003). Some examples include combining multi-locus effects via the sum of the individual single-locus associations (Hoh et al. 2001), other examples jointly model multiple loci (Nelson et al. 2001; Ritchie et al. 2001). I think that these methods should be explored in more detail in future studies, because they

are more likely to have the potential to detect higher-order interactions than multi-locus linkage analysis.

7.3 Conclusion

Epistasis appears to be an important biological feature underlying many complex traits. Linkage analysis incorporating epistatic interactions could contribute significant information gain in terms of the biological processes involved in the trait aetiology. The results from the 2D linkage scans in this thesis were not overwhelmingly significant, but have provided suggestive evidence for the involvement of novel loci in these traits. Multidimensional scans are computationally feasible and may identify regions involved in the disease susceptibility through a complex interplay of genetic factors. These analyses should be performed in the numerous existing linkage datasets that have been collected in a wide array of complex diseases, because they may identify novel loci, identify interaction pathways, and help narrow down regions for a detailed fine-mapping search through linkage-disequilibrium analysis.

Map specification appears to play a very important role in the magnitude of the multipoint single-locus, and in particular two-locus Lod scores. The bias to map misspecification is likely to be inflated in two-locus analysis especially if using multipoint IBDs and for two-locus analysis between linked regions. For the two-locus case it appears that not only the order of markers affects the findings, but also differences in the inter-marker distances (in particular instances when markers appear further apart than they really are). In my opinion this sensitivity deserves more attention in current linkage studies. It seems crucial to compare map order and distances across physical and genetics maps for chromosomal regions awaiting multipoint linkage analysis. Furthermore, this observation has implications for linkage

analysis studies that have been published to date, and suggests that it might be of interest to re-examine genome scans using different genetic maps. There is a need for a comprehensive web resource to combine genetic and physical map information from all studies in humans and possibly model organisms, similar to Cartographer (Metzidis et al. 2003) and the study of Nievergelt et al. (2004). This resource may be further developed by incorporating fine-scale population-based estimates of recombination rates, linkage-disequilibrium data, and cytogenetic findings.

Analytical methods for gene-mapping can further our understanding of individual molecular defects and their interactions with one another and with environmental risk factors. The current thesis describes methods for detecting interactions among genes in complex human traits. The findings may contribute to identification of genetic variants, which are involved in the trait through complex molecular pathways and biological complexes. Future studies will be necessary to confirm the presence of these interactions, and examine higher-order interactions among multiple genetic and environmental factors. Ultimately, it will be necessary to link all these findings and establish the complete network of interacting factors underlying each complex disease.

APPENDIX A. Variance components in 13 two-locus models

To obtain the variance components for the two-locus models in chapter III I followed Sham (1998), pp 200-211. If we assume that the 13 two-locus models are represented in terms of the quantitative trait genotypic means, instead of trait penetrances, we can derive the variance components attributed to the two loci.

Let us have two biallelic genes, gene 1 with alleles A and a and allele frequencies (p_A) and (p_a = 1-p_A), and gene 2 with alleles B and b and allele frequencies p_B and $p_b = 1 - p_B$. Then the population mean can be calculated from the allele frequencies and the genotypic means and is equivalent to the population prevalence K for qualitative traits (see section 3.2.1, equation 3.1), which is 0.1 for the two-locus models used in this study. Following Sham (1998) obtain the population mean corrected genotype means, f_{ij},

$$f_{ij} = f_{ij}^{uncorrected} - K \tag{A.1}$$

where $f_{ij}^{uncorrected}$ is described in Table 3.1. Then the marginal genotypic means for the genotypes AA, Aa, and aa at locus 1, and BB, Bb, bb at locus 2 are:

$$f_{AA..} = p_B^2 f_{AABB} + 2p_B p_b f_{AABb} + p_b^2 f_{AAbb} \tag{A.2}$$

$$f_{Aa..} = p_B^2 f_{AaBB} + 2p_B p_b f_{AaBb} + p_b^2 f_{Aabb} \tag{A.3}$$

$$f_{aa..} = p_B^2 f_{aaBB} + 2p_B p_b f_{aaBb} + p_b^2 f_{aabb} \tag{A.4}$$

$$f_{..BB} = p_A^2 f_{AABB} + 2p_A p_a f_{AaBB} + p_a^2 f_{aaBB} \tag{A.5}$$

$$f_{..Bb} = p_A^2 f_{AABb} + 2p_A p_a f_{AaBb} + p_a^2 f_{aaBb} \tag{A.6}$$

$$f_{..bb} = p_A^2 f_{AAbb} + 2p_A p_a f_{Aabb} + p_a^2 f_{aabb} \tag{A.7}$$

The marginal means of alleles A, a, B, and b can be obtained from the marginal genotypic means, for example for allele A the marginal mean is

$$f_{A...} = p_A f_{AA..} + p_a f_{Aa..} \tag{A.8}$$

In order to obtain the epistatic variance components, we need to calculate marginal means involving alleles at difference loci. For example the mean of marginal genotype A.B. is

$$f_{A.B.} = p_A p_B f_{AABB} + p_a p_B f_{AaBB} + p_A p_b f_{AABb} + p_a p_b f_{AaBb} \tag{A.9}$$

We then define genic effects ($\mu_A, \mu_a, \mu_B, \mu_b$), dominance interactions ($\mu_{AA}, \mu_{Aa}, \mu_{aa}, \mu_{BB}, \mu_{Bb}, \mu_{bb}$), and additive-by-additive ($\mu_{AB}, \mu_{Ab}, \mu_{aB}, \mu_{ab}$), additive-by-dominance ($\mu_{AAB}, \mu_{AAb}, \mu_{AaB}, \mu_{Aab}, \mu_{aaB}, \mu_{aab}, \mu_{ABB}, \mu_{ABb}, \mu_{Abb}, \mu_{aBB}, \mu_{aBb}, \mu_{abb}$), and dominance-by-dominance ($\mu_{AABB}, \mu_{AABb}, \mu_{AAbb}, \mu_{AaBB}, \mu_{AaBb}, \mu_{Aabb}, \mu_{aaBB}, \mu_{aaBb}, \mu_{aabb}$) interactions. For example, for allele A some of these quantities are,

$$\mu_A = f_A \tag{A.10}$$

$$\mu_{AA} = f_{AA..} - 2\mu_A \tag{A.11}$$

$$\mu_{AB} = f_{A.B.} - \mu_A - \mu_B \tag{A.12}$$

$$\mu_{AAB} = f_{AAB.} - \mu_{AA} - 2\mu_{AB} - 2\mu_A - \mu_B \tag{A.13}$$

$$\mu_{AABB} = f_{AABB} - 2\mu_{AAB} - 2\mu_{ABB} - 4\mu_{AB} - \mu_{AA} - \mu_{BB} - 2\mu_A - 2\mu_B \tag{A.14}$$

The variance components are the expectations of the squares of the genic effects (V_A), of the dominance interactions (V_D), of the additive-by-additive interactions (V_{AA}), of the additive-by-dominance interactions (V_{AD}), and of the dominance-by-dominance interactions (V_{DD}), and can then be calculated as follows,

$$V_A = p_A^2 p_B^2 (2\mu_A + 2\mu_B)^2 + 2p_A^2 p_B p_b (2\mu_A + \mu_B + \mu_b)^2 + p_A^2 p_b^2 (2\mu_A + 2\mu_b)^2 +$$
$$+ 2p_A p_a p_B^2 (\mu_A + \mu_a + 2\mu_B)^2 + 2p_A p_a p_B^2 (\mu_A + \mu_a + 2\mu_B)^2 +$$
$$+ 4p_A p_a p_B p_b (\mu_A + \mu_a + \mu_B + \mu_b)^2 +$$
$$+ p_a^2 p_B^2 (2\mu_a + 2\mu_B)^2 + 2p_a^2 p_B p_b (2\mu_a + \mu_B + \mu_b)^2 + p_a^2 p_b^2 (2\mu_a + 2\mu_b)^2 \quad \text{(A.15)}$$

$$V_D = p_A^2 p_B^2 (\mu_{AA} + \mu_{BB})^2 + 2p_A^2 p_B p_b (\mu_{AA} + \mu_{Bb})^2 + p_A^2 p_b^2 (\mu_{AA} + \mu_{bb})^2 +$$
$$+ 2p_A p_a p_B^2 (\mu_{Aa} + \mu_{BB})^2 + 4p_A p_a p_B p_b (\mu_{Aa} + \mu_{Bb})^2 + 2p_A p_a p_B^2 (\mu_{Aa} + \mu_{bb})^2 +$$
$$+ p_a^2 p_B^2 (\mu_{aa} + \mu_{BB})^2 + 2p_a^2 p_B p_b (\mu_{aa} + \mu_{Bb})^2 + p_a^2 p_b^2 (\mu_{aa} + \mu_{bb})^2 \quad \text{(A.16)}$$

$$V_{AA} = p_A^2 p_B^2 (4\mu_{AB})^2 + 2p_A^2 p_B p_b (2\mu_{AB} + 2\mu_{Ab})^2 + p_A^2 p_b^2 (4\mu_{Ab})^2 +$$
$$+ 2p_A p_a p_B^2 (2\mu_{AB} + 2\mu_{aB})^2 + 2p_A p_a p_B^2 (2\mu_{Ab} + 2\mu_{ab})^2 +$$
$$+ 4p_A p_a p_B p_b (\mu_{AB} + \mu_{Ab} + \mu_{aB} + \mu_{ab})^2 +$$
$$+ p_a^2 p_B^2 (4\mu_{aB})^2 + 2p_a^2 p_B p_b (2\mu_{aB} + \mu_{ab})^2 + p_a^2 p_b^2 (4\mu_{ab})^2 \quad \text{(A.17)}$$

$$V_{AD} = p_A^2 p_B^2 (2\mu_{AAB} + 2\mu_{ABB})^2 + 2p_A^2 p_B p_b (2\mu_{ABb} + \mu_{AAB} + \mu_{AAb})^2 + p_A^2 p_b^2 (2\mu_{AAb} + 2\mu_{Abb})^2 +$$
$$+ 2p_A p_a p_B^2 (2\mu_{AaB} + \mu_{ABB} + \mu_{aBB})^2 + 2p_A p_a p_B^2 (2\mu_{Aab} + \mu_{Abb} + \mu_{abb})^2 +$$
$$+ 4p_A p_a p_B p_b (\mu_{AaB} + \mu_{Aab} + \mu_{ABb} + \mu_{aBb})^2 +$$
$$+ p_a^2 p_B^2 (2\mu_{aaB} + 2\mu_{aBB})^2 + 2p_a^2 p_B p_b (\mu_{aaB} + \mu_{aab} + 2\mu_{aBb})^2 + p_a^2 p_b^2 (2\mu_{aab} + 2\mu_{abb})^2 \quad \text{(A.18)}$$

$$V_{DD} = p_A^2 p_B^2 \mu_{AABB}^2 + 2p_A^2 p_B p_b \mu_{AABb}^2 + p_A^2 p_b^2 \mu_{AAbb}^2 +$$
$$+ 2p_A p_a p_B^2 \mu_{AaBB}^2 + 4p_A p_a p_B p_b \mu_{AaBb}^2 + 2p_A p_a p_B^2 \mu_{Aabb}^2 +$$
$$+ p_a^2 p_B^2 \mu_{aaBB}^2 + 2p_a^2 p_B p_b \mu_{aaBb}^2 + p_a^2 p_b^2 \mu_{aabb}^2 \quad \text{(A.19)}$$

APPENDIX B. Genome-scans for hypertension and variation in blood pressure.

Table B.1 Significant and suggestive results from 25 genome scans of hypertension or variation in blood pressure. Significance criteria follow previous definitions (Lander and Kruglyak 1995; Rao and Province 2000).

Locus	cM[a]	Trait[b]	Sample[c]	Origin	Study
1p	65-95	SBP	274 ASP	Australia	(Harrap et al. 2002b)
1p	87-120	SBP	206 nuclear families	Quebec	(Rice et al. 2000)
1q	168-170	DBP	160 extended families	USA	(Thiel et al. 2003)
1q	192	HT	401extended families	USA	(Hunt et al. 2002)
2p	26.5-27	HT	1 large family	Sardinia	(Angius et al. 2002)
2p	57-59	SBP	69 DSP	USA	(Krushkal et al. 1999)
2p	63	HT	599 families	USA	(Rao et al. 2003)
2p	86	SBP	114 extended families	USA	(Rice et al. 2002)
2p	96-115	SBP	206 nuclear families	Quebec	(Rice et al. 2000)
2p	103	DBP	10 extended families	USA	(Atwood et al. 2001)
2p	104	DBP	196 families	Nigeria	(Cooper et al. 2002)
2p	115-118	HT	91 families	Sweden, Finland	(von Wowern et al. 2003)
2q	140-165	HT	1273 ASP	China	(Zhu et al. 2001)
2q	184	HT	47 ASP	Finland	(Perola et al. 2000)
2q	205-224	SBP, DBP	28 large families	Amish	(Hsueh et al. 2000)
2q	250	HT	61 ASP	Australia	(Harrap et al. 2002a)
3p	5	SBP	99 extended families	USA	(Rice et al. 2002)
3p	5.5	SBP	99 LSP	China	(Xu et al. 1999)
3p	16	DBP	196 families	Nigeria	(Cooper et al. 2002)
3q	119	DBP	160 extended families	USA	(Thiel et al. 2003)
3q	165	HT	47 ASP	Finland	(Perola et al. 2000)
3q	201	SBP	99 extended families	USA	(Rice et al. 2002)
4p	13-43	SBP	18 extended families	Holland	(Allayee et al. 2001)
4q	95-132	SBP	274 ASP	Australia	(Harrap et al. 2002b)
5p	14-46	SBP	206 nuclear families	Quebec	(Rice et al. 2000)
5q	98	SBP	330 pedigrees	UK	(Yang et al. 2003)
5q	102	DBP	196 families	Nigeria	(Cooper et al. 2002)
5q	188-192	SBP	69 DSP	USA	(Krushkal et al. 1999)
6p	34-42	ΔSBP, ΔDBP	498 ASP	USA	(Pankow et al. 2000)
6q	80-102	DBP	18 extended families	Holland	(Allayee et al. 2001)
6q	89	SBP	401 extended families	USA	(Hunt et al. 2002)
6q	134-155	SBP	69 DSP	USA	(Krushkal et al. 1999)
7p	58	HT	401 extended families	USA	(Hunt et al. 2002)
7p	81	DBP	196 families	Nigeria	(Cooper et al. 2002)
7q	109	DBP	196 families	Nigeria	(Cooper et al. 2002)
7q	127	HT	401 extended families	USA	(Hunt et al. 2002)
7q	135-150	SBP	206 nuclear families	Quebec	(Rice et al. 2000)
8p	26	SBP	330 pedigrees	UK	(Yang et al. 2003)
8q	86.7	SBP	206 nuclear families	Quebec	(Rice et al. 2000)
8q	94	SBP	330 pedigrees	UK	(Yang et al. 2003)
8q	164	DBP	10 extended families	USA	(Atwood et al. 2001)

Table B.1 Continued

Locus	cM[a]	Trait[b]	Sample[c]	Origin	Study
9q	163	low BP	184 LSP	USA, Hawaii, Taiwan	(Ranade et al. 2003)
10p	30	HT	661 ASP, 580 DSP	USA, Hawaii, Taiwan	(Ranade et al. 2003)
10q	76	DBP	196 families	Nigeria	(Cooper et al. 2002)
11p	35	SBP	1109 DZ twin pairs	UK	(de Lange et al. 2004)
11q	63	SBP	207 DSP	China	(Xu et al. 1999)
11q	85	SBP	99 extended families	USA	(Rice et al. 2002)
11q	126	HT	169 ASP	UK	(Sharma et al. 2000)
12	NA	DBP	160 extended families	USA	(Thiel et al. 2003)
12q	83	HT	401 extended families	USA	(Hunt et al. 2002)
12q	95	DBP	114 extended families	USA	(Rice et al. 2002)
14q	41-45	HT	91 families	Sweden, Finland	(von Wowern et al. 2003)
14q	92	HT	661 ASP	China	(Ranade et al. 2003)
15q	84-101	SBP	69 DSP	USA	(Krushkal et al. 1999)
15q	103	HT	401 extended families	USA	(Hunt et al. 2002)
15q	105	DBP	99 LSP	China	(Xu et al. 1999)
16p	40-62	SBP	274 ASP	Australia	(Harrap et al. 2002b)
16p	44	SBP	330 pedigrees	UK	(Yang et al. 2003)
16q	64	SBP	99 LSP	China	(Xu et al. 1999)
16q	65	DBP	1109 DZ twin pairs	UK	(de Lange et al. 2004)
17p	23.5	SBP	258 ASP	China	(Xu et al. 1999)
17q	60-76	SBP, DBP	332 extended families	USA	(Levy et al. 2000)
17q	70	SBP	1109 DZ twin pairs	UK	(de Lange et al. 2004)
17q	90-100	SBP	332 extended families	USA	(Levy et al. 2000)
18	NA	DBP	160 extended families	USA	(Thiel et al. 2003)
18p	7	DBP	332 extended families	USA	(Levy et al. 2000)
18q	66-89	ΔSBP	408 ASP	USA	(Pankow et al. 2000)
18q	80-94	HT	120 extended families	Iceland	(Kristjansson et al. 2002)
18q	116	SBP	10 extended families	USA	(Atwood et al. 2001)
19p	0-10	SBP	18 extended families	Holland	(Allayee et al. 2001)
19p	03-Jul	SBP	206 nuclear families	Quebec	(Rice et al. 2000)
19p	47	SBP	196 families	Nigeria	(Cooper et al. 2002)
19p	48.5	SBP	114 extended families	USA	(Rice et al. 2002)
19q	78	SBP	196 families	Nigeria	(Cooper et al. 2002)
21q	37	SBP	10 extended families	USA	(Atwood et al. 2001)
22q	32	HT	47 ASP	Finland	(Perola et al. 2000)
Xp	43	HT	47 ASP	Finland	(Perola et al. 2000)
Xp	37-52	SBP	274 ASP	Australia	(Harrap et al. 2002b)
None	NA	HT	989 ASP	USA	(Kardia et al. 2003)

[a] cM position were obtained from genetic maps used in the corresponding study. NA is not available. [b] Abbreviation for for phenotypes are as follows: SBP is systolic blood pressure; DBP is diastolic blood pressure; ΔSBP is change in SBP; ΔDBP is change in DBP; low BP is low blood pressure. [c] Abbreviation in samples are defined as follows: ASP are concordant affected sib-pairs; DSP are discordant sib-pairs; LSP are concordant sib-pairs with low blood pressure.

APPENDIX C. Candidate genes at peak coordinates in the BRIGHT 2D scan

The most significant peaks in the BRIGHT 2D linkage scan were examined to identify potential gene candidates that may be involved in interactions contributing to hypertension. The two peaks studied were 5q13.1 - 11q22 and 16p12.3 – 16q23.1, and for each peak I selected the peak microsatellite marker (at the multipoint fine-grid 2D peak or closest to it) and chose the two flanking markers on either side to delineate the search interval. I then queried the UCSC database build 17 of the human genome from May 2004 to identify a list of the known genes in each of the search intervals. There were altogether 942 known genes identified in these intervals, of which I excluded all records for hypothetical proteins. Occasionally there were also several entries denoting the same protein, and if these are excluded as well, there were altogether 490 entries in the four search intervals. Of the 490 gene records 45 were in D5S647 - D5S1982 on 5q13.1, 137 in D11S4175 - D11S908 on 11q22, 100 in D16S3103-D16S3068 on 16p12.3, and 208 in D16S503-D16S516 on 16q23.1. There were many genes in these intervals that could hypothetically be involved in pathways which contribute to hypertension, however, the most obvious candidate I believe was SAH in D16S3103-D16S3068, which is the SA hypertension-associated homolog isoform 1 that belongs to the ATP-dependent AMP-building enzyme family, and the homologue of which is associated with hypertension in rats (Iwai et al. 1994).

The 490 gene entries were then queried against the BIND database of molecular interactions (www.bind.ca). I queried the gene symbol from UCSC in a text query of all fields in BIND restricting the search to results from the 'Homo Sapiens' taxonomy (in the 'taxname' field), thus excluding all synthetic constructs and

interactions identified in other organisms, and then only examined the interacting partners that had a record in NCBI or UCSC containing a chromosomal location. I included all possible interaction types, which encompassed interactions, complexes, pathways, protein-protein interactions, nucleic acid interactions, genetic interactions, and small molecule interactions. For each gene I obtained a list of molecular interactions and I collected the interacting partner's identifier. The identifier was queried in NCBI (and UCSC if necessary) to obtain a chromosomal location. For molecules that were involved in a large number of interactions (over 40) I automated the search over all the interacting partners' identification numbers, by linking the 'GI' number to the 'geneSymbol' in the 'Known Genes' table from UCSC human genome build17, May 2004. This was done for all genes in each the four search intervals and for each entry I searched for hits that included an interaction in the reciprocal partner's region for the exact cytogenetic band location as the one obtained from two-locus results. For example, a gene in D5S647 - D5S1982 could interact in BIND with a gene on chromosome 11q22 that was not necessarily in D11S4175 - D11S908. The results from this search are presented in Table C.1.

There was one interesting interaction obtained for pair 5q13.1 – 11q22. I started the search by looping over all entries in interval D5S647 - D5S1982 on 5q13.1. I obtained one hit in BIND (BIND id: 214922) involving 5q13.1 and 11q22, an interaction between MMP1, matrix metallopeptidase 1 (interstitial collagenase) on 11q22.3, and PAR1, coagulation factor II (thromin) receptor on 5q13 (Boire et al. 2005). This result is potentially of interest because protease-activated receptors (PARs) are a class of G protein-coupled receptors that play critical roles in thrombosis, inflammation, and vascular biology, and may be interesting candidates

for hypertension. I next searched over all entries in interval D11S4175 - D11S908 on 11q22, and found the same interaction record (BIND id: 214922).

Several interactions were observed on chromosome 16, all involving the E2F transcription factor 4, which is involved in the control of cell cycle and action of tumour suppressor proteins. It should be noted that E2F4 was involved in approximately 1400, interactions, the majority of which were with promoters of genes. For the chromosome 16 search, I first looped over all entries in D16S3103-D16S3068 on 16p12.3. I obtained a result in BIND involving 16p13 and 16q21-q22 (BIND id:194029), an interaction between the promoter of LOC81691 (exonuclease NEF-sp) and E2F transcription factor 4 (Cam et al. 2004). A second hit was the interaction involving 16p13 and 16q21-q22 (BIND id: 194088), again involving E2F4, E2F transcription factor 4, and this time the PLK1 promoter (serine/threonine-protein kinase PLK1 - polo-like kinase 1) on 16p13 (Cam et al. 2004). Next, I looped over all the entries in D16S503-D16S516 on 16q23.1. I obtained a result (BIND id: 194017) for the interaction between the promoter of KIF22 – kinesin family member 22 on 16p11.2 and the E2F transcription factor 4 p107/p130-binding protein (Cam et al. 2004). I also observed the two results (BIND id:194029 and BIND id: 194088) obtained from searching over entries in 16p12.3. The fourth hit for this pair of regions on chromosome 16 involved an interaction (BIND id: 194565) between the promoter of Znf267 – zinc finger protein 267 on 16p11.2 and E2F transcription factor 4 p107/p130-binding protein (Cam et al. 2004).

Table C.1 Summary of molecular interactions in BIND involving genes underlying the 5q13-11q22 and 16p12-16q22 2D linkage peaks.

	Molecule A					Molecule B						
BIND ID	**Location**	**Name**	**Type**	**bp position**	**Description**	**Location**	**Name**	**Type**	**Location**	**Description**	**Interaction evidence**	**Reference**
214922	5q13.3	PAR1	protein	76047547	Coagulation factor II receptor precursor.	11q22.2	MMP-1	protein	102165861	Matrix metalloproteinase 1 preproprotein.	in vitro direct-cleavage assay	Boire et al. (2005)
194029	16p12.3	LOC81691 promoter	DNA	20725334	exonuclease NEF-sp	16q22.1	E2F4	protein	65783569	E2F transcription factor 4	cross linking	Cam et al. (2004)
194088	16p12.1	PLK1 promoter	DNA	23597702	Polo-like kinase 1	16q22.1	E2F4	protein	65783569	E2F transcription factor 4	cross linking	Cam et al. (2004)
194017	16p11.2	KIF22 promoter	DNA	29709559	kinesin family member 22	16q22.1	E2F4	protein	65783569	E2F transcription factor 4	cross linking	Cam et al. (2004)
194565	16p11.2	ZNF267 promoter	DNA	31792661	zinc finger protein 267	16q22.1	E2F4	protein	65783569	E2F transcription factor 4	cross linking	Cam et al. (2004)

Finally, I also searched BIND for interactions linked to hypertension. To achieve this I queried 'hypertension' in a text query of all fields in BIND, again restricting the search to results from 'Homo Sapiens' taxonomy (in 'taxname) and examined all results with records in NCBI containing the chromosomal location. The aim was to identify pairs of genes where both partners fell in the regions identified in Table 5.2. In the BIND results for hypertension, pairwise interactions involving ACE and Calpain-1 were frequent, but no pairs of regions where both partners fell in regions identified as interacting in Table 5.2 were identified. The only potentially relevant result was a gene on 16p13.3 - SOCS-1 (suppressor of cytokine signaling 1), interacting with the insulin receptor on 19p13.3-p13.2 (BIND id:12763 for the interaction).

ASSOCIATED PUBLICATIONS

Chapter II (section 2.2) and Chapter V:

Bell JT, Wallace C, Dobson R, Wiltshire S, Mein C, Pembroke J, Brown M, Clayton D, Samani N, Dominiczak A, Webster J, Mark Lathrop G, Connell J, Munroe P, Caulfield M, Farrall M (2006) Two-dimensional genome scan identifies novel epistatic loci for essential hypertension. Hum Mol Genet 15: 1365-1374.

Chapter IV (excluding section 4.3.2):

Wiltshire S [§], **Bell JT** [§], Groves CJ, Dina C, Hattersley AT, Frayling TM, Walker M, Hitman GA, Vaxillaire M, Farrall M, Froguel P, McCarthy MI Epistasis between type 2 diabetes susceptibility loci on chromosomes 1q21-25 and 10q23-26 in northern Europeans. Ann Hum Genet [in press]

[§] These authors contributed equally to this work.

Chapter VI (section 6.2.2):

Lamb JA, Barnby G, Bonora E, Sykes N, Bacchelli E, Blasi F, Maestrini E, Broxholme J, **Tzenova J**, Weeks D, Bailey AJ, Monaco AP (2005) Analysis of IMGSAC autism susceptibility loci: evidence for sex limited and parent of origin specific effects. J Med Genet 42:132-137.

REFERENCES

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97-101

Abecasis GR, Wigginton JE (2005) Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 77:754-767

Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33:D418-424

Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69:936-950

Araujo H, Bier E (2000) sog and dpp exert opposing maternal functions to modify toll signaling and pattern the dorsoventral axis of the Drosophila embryo. Development 127:3631-3644

Bachmanov AA, Reed DR, Li X, Li S, Beauchamp GK, Tordoff MG (2002) Voluntary ethanol consumption by mice: genome-wide analysis of quantitative trait loci and their interactions in a C57BL/6ByJ x 129P3/J F2 intercross. Genome Res 12:1257-1268

Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW (2001) BIND--The Biomolecular Interaction Network Database. Nucleic Acids Res 29:242-245

Barber JC, Joyce CA, Collinson MN, Nicholson JC, Willatt LR, Dyson HM, Bateman MS, Green AJ, Yates JR, Dennis NR (1998) Duplication of 8p23.1: a cytogenetic anomaly with no established clinical significance. J Med Genet 35:491-496

Barber MJ, Todd JA, Cordell HJ (2006) A multimarker regression-based test of linkage for affected sib-pairs at two linked loci. Genet Epidemiol 30:191-208

Barroso I (2005) Complex disease: pleiotropic gene effects in obesity and type 2 diabetes. Eur J Hum Genet 13:1243-1244

Bateson W (1909) Mendel's principles of heredity. Cambridge University Press, Cambridge

Beaudoin N, Serizet C, Gosti F, Giraudat J (2000) Interactions between abscisic acid and ethylene signaling cascades. Plant Cell 12:1103-1115

Bengtsson O (2001) PhD Thesis. Two-locus affected sib-pair identity by descent probabilities. Goteborg University, Goteborg, Sweden

Biernacka JM, Sun L, Bull SB (2005) Simultaneous localization of two linked disease susceptibility genes. Genet Epidemiol 28:33-47

Biswas S, Papachristou C, Irwin ME, Lin S (2003) Linkage analysis of the simulated data - evaluations and comparisons of methods. BMC Genet 4 Suppl 1:S70

Boehnke M, Cox NJ (1997) Accurate inference of relationships in sib-pair linkage studies. Am J Hum Genet 61:423-429

Boer MP, Ter Braak CJ, Jansen RC (2002) A penalized likelihood method for mapping epistatic quantitative trait Loci with one-dimensional genome searches. Genetics 162:951-960

Boire A, Covic L, Agarwal A, Jacques S, Sherifi S, Kuliopulos A (2005) PAR1 is a matrix metalloprotease-1 receptor that promotes invasion and tumorigenesis of breast cancer cells. Cell 120:303-313

Brockmann GA, Kratzsch J, Haley CS, Renne U, Schwerin M, Karle S (2000) Single QTL effects, epistasis, and pleiotropy account for two-thirds of the phenotypic F(2) variance of growth and obesity in DU6i x DBA/2 mice. Genome Res 10:1941-1957

Brodie BI (2000) Why evolutionary genetics doesn't always add up. In: Wade M, Brodie BI, Wolf J (eds) Epistasis and the evolutionary process. Oxford University Press, Oxford

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1999) Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 63:861-869

Busfield F, Duffy DL, Kesting JB, Walker SM, Lovelock PK, Good D, Tate H, Watego D, Marczak M, Hayman N, Shaw JT (2002) A genomewide search for type 2 diabetes-susceptibility genes in indigenous Australians. Am J Hum Genet 70:349-357

Cam H, Balciunaite E, Blais A, Spektor A, Scarpulla RC, Young R, Kluger Y, Dynlacht BD (2004) A common set of gene regulatory networks links metabolism and growth inhibition. Mol Cell 16:399-411

Carlborg O, Andersson L (2002) Use of randomization testing to detect multiple epistatic QTLs. Genet Res 79:175-184

Carlborg O, Andersson L, Kinghorn B (2000) The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. Genetics 155:2003-2010

Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5:618-625

Carlborg O, Kerje S, Schutz K, Jacobsson L, Jensen P, Andersson L (2003) A global search reveals epistatic interaction between QTL for early growth in the chicken. Genome Res 13:413-421

Caulfield M, Munroe P, Pembroke J, Samani N, Dominiczak A, Brown M, Benjamin N, Webster J, Ratcliffe P, O'Shea S, Papp J, Taylor E, Dobson R, Knight J, Newhouse S, Hooper J, Lee W, Brain N, Clayton D, Lathrop GM, Farrall M, Connell J (2003) Genome-wide mapping of human loci for essential hypertension. Lancet 361:2118-2123

Chakrabarti S, Fombonne E (2005) Pervasive developmental disorders in preschool children: confirmation of high prevalence. Am J Psychiatry 162:1133-1141

Chang BL, Lange EM, Dimitrov L, Valis CJ, Gillanders EM, Lange LA, Wiley KE, Isaacs SD, Wiklund F, Baffoe-Bonnie A, Langefeld CD, Zheng SL, Matikainen MP, Ikonen T, Fredriksson H, Tammela T, Walsh PC, Bailey-Wilson JE, Schleutker J, Gronberg H, Cooney KA, Isaacs WB, Suh E, Trent JM, Xu J (2006) Two-locus genome-wide linkage scan for prostate cancer susceptibility genes with an interaction effect. Hum Genet 118:716-724

Charcosset A, Causse M, Moreau L, Gallais A (1994) Investigation into the effect of genetic background on QTL expression using three connected maize recombinant inbred lines (RIL) populations. In: Ooijen JWv, Jansen J (eds) Biometrics in Plant Breeding: Applications of Molecular Markers. CPRO-DLO, Wageningen, The Netherlands, pp 75-84

Chase K, Adler FR, Lark KG (1997) EPISTAT: a computer program for identifying and testing interactions between pairs of quantitative trait loci. Theor Appl Genet 94:724-730

Clark AG, Feldman MW, Christiansen FB (1981) The estimation of epistasis in components of fitness in experimental populations of Drosophila melanogaster I. A two-stage maximum likelihood model. Heredity 46:321-346

Colilla S, Tsalenko A, Pluznikov A, Cox NJ (2001) Genome-wide approaches for identifying interacting susceptibility regions for asthma. Genet Epidemiol 21 Suppl 1:S266-271

Conneally JH, Edwards JH, Kidd KK, Lalouel JM, Morton NE, Ott J, White R (1985) Report of the committee on methods of linkage analysis and reporting. Cytogenetics and Cell Genetics 40:356-359

Cooper RS, Luke A, Zhu X, Kan D, Adeyemo A, Rotimi C, Bouzekri N, Ward R (2002) Genome scan among Nigerians linking blood pressure to chromosomes 2, 3, and 19. Hypertension 40:629-633

Cordell HJ (2001) Sample size requirements to control for stochastic variation in magnitude and location of allele-sharing linkage statistics in affected sibling pairs. Ann Hum Genet 65:491-502

Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11:2463-2468

Cordell HJ (2003) Affected-sib-pair data can be used to distinguish two-locus heterogeneity from two-locus epistasis. Am J Hum Genet 73:1468-1471; author reply 1471-1463

Cordell HJ, Barratt BJ, Clayton DG (2004) Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. Genet Epidemiol 26:167-185

Cordell HJ, Todd JA, Bennett ST, Kawaguchi Y, Farrall M (1995) Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes. Am J Hum Genet 57:920-934

Cordell HJ, Todd JA, Hill NJ, Lord CJ, Lyons PA, Peterson LB, Wicker LS, Clayton DG (2001) Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. Genetics 158:357-367

Cordell HJ, Wedig GC, Jacobs KB, Elston RC (2000) Multilocus linkage tests based on affected relative pairs. Am J Hum Genet 66:1273-1286

Cox DR, Hinkley DV (1974) Asymptotic Theory. In: Theoretical Statistics. Chapman&Hall, London, pp 279-363

Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A (1999) Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. Nat Genet 21:213-215

Craddock N, Khodel V, Van Eerdewegh P, Reich T (1995) Mathematical limits of multilocus models: the genetic transmission of bipolar disorder. Am J Hum Genet 57:690-702

Craig HD, Gunel M, Cepeda O, Johnson EW, Ptacek L, Steinberg GK, Ogilvy CS, Berg MJ, Crawford SC, Scott RM, Steichen-Gersdorf E, Sabroe R, Kennedy CT, Mettler G, Beis MJ, Fryer A, Awad IA, Lifton RP (1998) Multilocus linkage identifies two new loci for a mendelian form of stroke, cerebral cavernous malformation, at 7p15-13 and 3q25.2-27. Hum Mol Genet 7:1851-1858

Crow JF (1970) Genetic loads and the cost of natural selection. In: Kojima KI (ed) Mathematical topics in population genetics. Springer, Heidelberg

Crow JF, Kimura M (1970) An introduction to population genetic theory. Burgess Publishing Company, Minneapolis, MN

Culverhouse R, Klein T, Shannon W (2004) Detecting epistatic interactions contributing to quantitative traits. Genet Epidemiol 27:141-152

Culverhouse R, Suarez BK, Lin J, Reich T (2002) A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 70:461-471

Damerval C, Maurice A, Josse JM, de Vienne D (1994) Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics 137:289-301

Delepine M, Pociot F, Habita C, Hashimoto L, Froguel P, Rotter J, Cambon-Thomsen A, Deschamps I, Djoulah S, Weissenbach J, Nerup J, Lathrop M, Julier C (1997) Evidence of a non-MHC susceptibility locus in type I diabetes linked to HLA on chromosome 6. Am J Hum Genet 60:174-187

Dietter J, Lenzen K, Sander T, Strauch K (2005) Estimating two-locus disease model parameters in a lod score analysis with Genehunter-Twolocus. Genetic Epidemiology 29:243-243

Dizier MH, Babron MC, Clerget-Darpoux F (1994) Interactive effect of two candidate genes in a disease: extension of the marker-association-segregation chi(2) method. Am J Hum Genet 55:1042-1049

Dizier MH, Babron MC, Clerget-Darpoux F (1996) Conclusion of LOD-score analysis for family data generated under two-locus models. Am J Hum Genet 58:1338-1346

Dizier MH, Clerget-Darpoux F (1986) Two-disease locus model: sib pair method using information on both HLA and Gm. Genet Epidemiol 3:343-356

Dupuis J, Brown PO, Siegmund D (1995) Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. Genetics 140:843-856

Dupuis J, Siegmund D (1999) Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151:373-386

Durner M, Greenberg DA, Hodge SE (1992) Inter- and intrafamilial heterogeneity: effective sampling strategies and comparison of analysis methods. Am J Hum Genet 51:859-870

Durner M, Vieland VJ, Greenberg DA (1999) Further evidence for the increased power of LOD scores compared with nonparametric methods. Am J Hum Genet 64:281-289

Ehm MG, Karnoub MC, Sakul H, Gottschalk K, Holt DC, Weber JL, Vaske D, Briley D, Briley L, Kopf J, McMillen P, Nguyen Q, Reisman M, Lai EH, Joslyn G, Shepherd NS, Bell C, Wagner MJ, Burns DK (2000) Genomewide search for type 2 diabetes susceptibility genes in four American populations. Am J Hum Genet 66:1871-1881

Elston RC, Song D, Iyengar SK (2005) Mathematical assumptions versus biological reality: myths in affected sib pair linkage analysis. Am J Hum Genet 76:152-156

Ewens WJ (1979) Mathematical Population Genetics. Springer-Verlag, New York

Farrall M (1997) Affected sibpair linkage tests for multiple linked susceptibility genes. Genet Epidemiol 14:103-115

Farrall M (2003) Reports of the death of the epistasis model are greatly exaggerated. Am J Hum Genet 73:1467-1468; author reply 1471-1463

Farrall M, Holder S (1992) Familial recurrence-pattern analysis of cleft lip with or without cleft palate. Am J Hum Genet 50:270-277

Fasoulas AC, Allard RW (1962) Nonallelic gene interactions in the inheritance of quantitative characters in bailey. Genetics 47:899-907

Fijneman RJ, de Vries SS, Jansen RC, Demant P (1996) Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility to lung cancer in the mouse. Nat Genet 14:465-467

Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb 52:399-433

Frankel WN, Schork NJ (1996) Who's afraid of epistasis? Nat Genet 14:371-373

Freire-Maia N (1990) Five landmarks in inbreeding studies. Am J Med Genet 35:118-120

Garrod AE (1902) The incidence of alcaptonuria: a study in chemical individuality. Lancet 2:1616-1620

Ghosh S, Watanabe RM, Valle TT, Hauser ER, Magnuson VL, Langefeld CD, Ally DS, et al. (2000) The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. 1. An autosomal genome scan for genes that predispose to type 2 diabetes. Am J Hum Genet 67:1174-1185

Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, Ohashi H, Voullaire L, Larizza D, Giorda R, Weber JL, Ledbetter DH, Zuffardi O (2001) Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. Am J Hum Genet 68:874-883

Goldin LR, Weeks DE (1993) Two-locus models of disease: comparison of likelihood and nonparametric linkage methods. Am J Hum Genet 53:908-915

Greenwood CM, Bull SB (1999) Down-weighting of multiple affected sib pairs leads to biased likelihood-ratio tests, under the assumption of no linkage. Am J Hum Genet 64:1248-1252

Gudbjartsson DF, Jonasson K, Frigge ML, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nat Genet 25:12-13

Guo SW (1998) Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting. Am J Hum Genet 63:252-258

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315-324

Harada N, Takano J, Kondoh T, Ohashi H, Hasegawa T, Sugawara H, Ida T, Yoshiura K, Ohta T, Kishino T, Kajii T, Niikawa N, Matsumoto N (2002) Duplication of 8p23.2: a benign cytogenetic variant? Am J Med Genet 111:285-288

Harrap SB, Wong ZY, Stebbing M, Lamantia A, Bahlo M (2002b) Blood pressure QTLs identified by genome-wide linkage analysis and dependence on associated phenotypes. Physiol Genomics 8:99-105

Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD, Boehnke M (2004) Ordered subset analysis in genetic linkage mapping of complex traits. Genet Epidemiol 27:53-63

Hayes B, Goddard ME (2001) The distribution of the effects of genes affecting quantitative traits in livestock. Genet Sel Evol 33:209-229

Hayman BI, Mather K (1955) The description of genetic interactions oin continuous variation. Biometrics 11:69-82

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 61:748-760

Hodge SE (1981) Some epistatic two-locus models of disease. I. Relative risks and identity-by-descent distributions in affected sib pairs. Am J Hum Genet 33:381-395

Hoh J, Ott J (2003) Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 4:701-709

Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res 11:2115-2119

Holland JB (1998) EPISTACY: a SAS program for detecting two-locus epistatic interactions using genetic marker information. J Hered 89:374-375

Holland JB, Moser HS, O'Donoughue LS, Lee M (1997) QTLs and epistasis associated with vernalization responses in oat. Crop Sci 37:1306-1316

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362-374

Holmans P (2002) Detecting gene-gene interactions using affected sib pair analysis with covariates. Hum Hered 53:92-102

Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, et al. (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat Genet 26:163-175

IMGSAC (1998) A full genome screen for autism with evidence for linkage to a region on chromosome 7q. International Molecular Genetic Study of Autism Consortium. Hum Mol Genet 7:571-578

Iwai N, Ohmichi N, Hanai K, Nakamura Y, Kinoshita M (1994) Human SA gene locus as a candidate locus for essential hypertension. Hypertension 23:375-380

James JW (1971) Frequency in relatives for an all-or-none trait. Ann Hum Genet 35:47-48

Jannink JL, Jansen R (2001) Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics 157:445-454

Jansen RC (1993) Interval mapping of multiple quantitative trait loci. Genetics 135:205-211

Jansen RC (2003) Studying complex biological systems using multifactorial perturbation. Nat Rev Genet 4:145-151

Jansen RC, Stam P (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447-1455

Jorgenson E, Tang H, Gadde M, Province M, Leppert M, Kardia S, Schork N, Cooper R, Rao DC, Boerwinkle E, Risch N (2005) Ethnicity and human genetic linkage maps. Am J Hum Genet 76:276-290

Kao CH, Zeng ZB, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152:1203-1216

Kaprio J, Tuomilehto J, Koskenvuo M, Romanov K, Reunanen A, Eriksson J, Stengard J, Kesaniemi YA (1992) Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. Diabetologia 35:1060-1067

Kempthorne O (1954) The correlation between relatives in a random mating population. Proc R Soc London Ser B 242:103-113

Kempthorne O (1957) An introduction to genetic statistics. John Wiley & Sons, New York

Khazanehdari KA, Borts RH (2000) EXO1 and MSH4 differentially affect crossing-over and segregation. Chromosoma 109:94-102

Kilpikari R, Sillanpaa MJ (2003) Bayesian analysis of multilocus association in quantitative and qualitative traits. Genet Epidemiol 25:122-135

Knapp M, Seuchter SA, Baur MP (1994) Two-locus disease models with two marker loci: the power of affected-sib-pair tests. Am J Hum Genet 55:1030-1041

Koeleman BP, Dudbridge F, Cordell HJ, Todd JA (2000) Adaptation of the extended transmission/disequilibrium test to distinguish disease associations of multiple loci: the Conditional Extended Transmission/Disequilibrium Test. Ann Hum Genet 64:207-213

Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179-1188

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241-247

Kong X, Murphy K, Raj T, He C, White PS, Matise TC (2004) A combined linkage-physical map of the human genome. Am J Hum Genet 75:1143-1148

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

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439-454

Lamb JA, Barnby G, Bonora E, Sykes N, Bacchelli E, Blasi F, Maestrini E, Broxholme J, Tzenova J, Weeks D, Bailey AJ, Monaco AP (2005) Analysis of IMGSAC autism susceptibility loci: evidence for sex limited and parent of origin specific effects. J Med Genet 42:132-137

Lande R (1981) The Minimum Number of Genes Contributing to Quantitative Variation between and within Populations. Genetics 99:541-553

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241-247

Lander ES, Botstein D (1986) Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc Natl Acad Sci U S A 83:7353-7357

Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199

Langefeld CD, Davis CC, Brown WM (2001) Nonparametric linkage regression. I: Combined Caucasian CSGA and German genome scans for asthma. Genet Epidemiol 21 Suppl 1:S136-141

Lark KG, Chase K, Adler F, Mansur LM, Orf JH (1995) Interactions between quantitative trait loci in soybean in which trait variation at one locus is conditional upon a specific allele at another. Proc Natl Acad Sci U S A 92:4656-4660

Lathrop GM, Ott J (1990) Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. Am J Hum Genet Suppl 47:A188

Leips J, Mackay TF (2000) Quantitative trait loci for life span in Drosophila melanogaster: interactions with genetic background and larval density. Genetics 155:1773-1788

Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, Cupples LA, Myers RH (2000) Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study. Hypertension 36:477-483

Li CC, Sacks L (1954) The derivation of joint distribution and correlation between relatives by the use of stoachastic matrices. Biometrics 10:347-360

Li M, Boehnke M, Abecasis GR (2005) Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. Am J Hum Genet 76:934-949

Li W, Reich J (2000) A complete enumeration and classification of two-locus disease models. Hum Hered 50:334-349

Li Z, Pinson SR, Park WD, Paterson AH, Stansel JW (1997) Epistasis for three grain yield components in rice (Oryza sativa L.). Genetics 145:453-465

Liang KY, Chiu YF, Beaty TH, Wjst M (2001) Multipoint analysis using affected sib pairs: incorporating linkage evidence from unlinked regions. Genet Epidemiol 21:105-122

Liu Y, Tritchler D, Bull SB (2002) A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits. Genet Epidemiol 22:26-40

Loftus BJ, Kim UJ, Sneddon VP, Kalush F, Brandon R, Fuhrmann J, Mason T, Crosby ML, Barnstead M, Cronin L, Deslattes Mays A, Cao Y, Xu RX, Kang HL, Mitchell S, Eichler EE, Harris PC, Venter JC, Adams MD (1999) Genome duplications and other features in 12 Mb of DNA sequence from human chromosome 16p and 16q. Genomics 60:295-308

Long AD, Mullaney SL, Mackay TF, Langley CH (1996) Genetic interactions between naturally occurring alleles at quantitative trait loci and mutant alleles at candidate loci affecting bristle number in Drosophila melanogaster. Genetics 144:1497-1510

Lucek P, Hanke J, Reich J, Solla SA, Ott J (1998) Multi-locus nonparametric linkage analysis of complex trait loci with neural networks. Human Heredity 48:275-284

Lunetta KL, Faraone SV, Biederman J, Laird NM (2000) Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. Am J Hum Genet 66:605-614

Luschnig S, Krauss J, Bohmann K, Desjeux I, Nusslein-Volhard C (2000) The Drosophila SHC adaptor protein is required for signaling by a subset of receptor tyrosine kinases. Mol Cell 5:231-241

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Association, Ltd, Sunderland, MA

MacLean CJ, Sham PC, Kendler KS (1993) Joint linkage of multiple loci for a complex disorder. Am J Hum Genet 53:353-366

Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37:413-417

Mather K, Jinks JL (1982) Biometrical Genetics: the study of continuous variation, 3rd edn. Chapman & Hall, London

Mein CA, Esposito L, Dunn MG, Johnson GC, Timms AE, Goy JV, Smith AN, Sebag-Montefiore L, Merriman ME, Wilson AJ, Pritchard LE, Cucca F, Barnett AH, Bain SC, Todd JA (1998) A search for type 1 diabetes susceptibility genes in families from the United Kingdom. Nat Genet 19:297-300

Mendel G (1865) Experiments in plant hybridisation by Gregor Mendel; Mendel's original paper in English translation with commentary and assessment by the late Sir Ronald A. Fisher, together with a reprint of W. Bateson's biographical notice of Mendel. Oliver and Boyd, 1965, Edinburgh

Metzidis A, Sammalisto S, Perola M, Peltonen L, Saharinen J (2003) Cartographer: A tool to generate marker maps based on the physical, genomic location of markers. American Journal of Human Genetics 73:467-467

Millstein J, Conti DV, Gilliland FD, Gauderman WJ (2006) A testing framework for identifying susceptibility genes in the presence of epistasis. Am J Hum Genet 78:15-27

Moore JH (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 56:73-82

Morgan TH (1910) Sex Limited Inheritance in Drosophila. Science 32:120-122

Morris A, Whittaker J (1999) Generalization of the extended transmission disequilibrium test to two unlinked disease loci. Genet Epidemiol 17 Suppl 1:S661-666

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277-318

Nath SK, Chen CH, Schork NJ (2001) Two-trait-locus linkage analyses of asthma susceptibility. Genet Epidemiol 21 Suppl 1:S278-283

Nelson MR, Kardia SL, Ferrell RE, Sing CF (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 11:458-470

Neuman RJ, Rice JP (1992) Two-locus models of disease. Genet Epidemiol 9:347-365

Nievergelt CM, Smith DW, Kohlenberg JB, Schork NJ (2004) Large-scale integration of human genetic and physical maps. Genome Res 14:1199-1205

O'Connell JR, Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat Genet 11:402-408

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27:29-34

Olson JM (1997) Likelihood-based models for genetic linkage analysis using affected sib pairs. Hum Hered 47:110-120

Olson JM (1999) A general conditional-logistic model for affected-relative-pair linkage studies. Am J Hum Genet 65:1760-1769

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

Otto SP, Jones CD (2000) Detecting the undetected: estimating the total number of loci underlying a quantitative trait. Genetics 156:2093-2107

Pedersen JC, Berg K (1989) Interaction between Low-Density Lipoprotein Receptor (LDLR) and Apolipoprotein-E (ApoE) alleles contributes to normal variation in lipid levels. Clin Genet 35:331-337

Penrose LS (1935) The detection of autosomal linkage in data which consists of pairs of brothers and sisters of unspecified parentage. Ann Eugen 6:133-138

Philippe A, Martinez M, Guilloud-Bataille M, Gillberg C, Rastam M, Sponheim E, Coleman M, Zappella M, Aschauer H, Van Maldergem L, Penet C, Feingold J, Brice A, Leboyer M (1999) Genome-wide scan for autism susceptibility genes. Paris Autism Research International Sibpair Study. Hum Mol Genet 8:805-812

Phillips PC (1998) The language of gene interaction. Genetics 149:1167-1171

Pickles A, Bolton P, Macdonald H, Bailey A, Le Couteur A, Sim CH, Rutter M (1995) Latent-class analysis of recurrence risks for complex phenotypes with

selection and measurement error: a twin and family history study of autism. Am J Hum Genet 57:717-726

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124-137

Puniyani A, Feldman MW (2005) A semi-symmetric two-locus model. Theor Popul Biol

Purcell S, Sham PC (2004) Epistasis in quantitative trait locus linkage analysis: interaction or main effect? Behav Genet 34:143-152

Rao DC, Province MA (2000) The future of path analysis, segregation analysis, and combined models for genetic dissection of complex traits. Hum Hered 50:34-42

Rebai A, Blanchard P, Perret D, Vincourt P (1997) Mapping quantitative trait loci controlling silking date in a diallel cross among four lines of maize. Theor Appl Genet 95:451-459

Rice T, Rankinen T, Chagnon YC, Province MA, Perusse L, Leon AS, Skinner JS, Wilmore JH, Bouchard C, Rao DC (2002) Genomewide linkage scan of resting blood pressure: HERITAGE Family Study. Health, Risk Factors, Exercise Training, and Genetics. Hypertension 39:1037-1043

Rich SS (1990) Mapping genes in diabetes. Genetic epidemiological perspective. Diabetes 39:1315-1319

Risch N (1990a) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222-228

Risch N (1990b) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229-241

Risch N (1990c) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242-253

Risch N, Ghosh S, Todd JA (1993) Statistical evaluation of multiple-locus linkage data in experimental species and its relevance to human studies: application to nonobese diabetic (NOD) mouse and human insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 53:702-714

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

Risch N, Spiker D, Lotspeich L, Nouri N, Hinds D, Hallmayer J, Kalaydjieva L, et al. (1999) A genomic screen of autism: evidence for a multilocus etiology. Am J Hum Genet 65:493-507

Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405:847-856

Ritchie MD, Hahn LW, Moore JH (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24:150-157

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138-147

Roberts SB, MacLean CJ, Neale MC, Eaves LJ, Kendler KS (1999) Replication of linkage studies of complex traits: an examination of variation in location estimates. Am J Hum Genet 65:876-884

Rudan I, Rudan D, Campbell H, Carothers A, Wright A, Smolej-Narancic N, Janicijevic B, Jin L, Chakraborty R, Deka R, Rudan P (2003) Inbreeding and risk of late onset complex disease. J Med Genet 40:925-932

Rybicki BA, Elston RC (2000) The relationship between the sibling recurrence-risk ratio and genotype relative risk. Am J Hum Genet 66:593-604

Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) A bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144:805-816

Scanga SE, Ruel L, Binari RC, Snow B, Stambolic V, Bouchard D, Peters M, Calvieri B, Mak TW, Woodgett JR, Manoukian AS (2000) The conserved PI3'K/PTEN/Akt signaling pathway regulates both cell size and survival in Drosophila. Oncogene 19:3971-3977

Schliekelman P, Slatkin M (2002) Multiplex relative risk and estimation of the number of loci underlying an inherited disease. Am J Hum Genet 71:1369-1385

Schork NJ, Boehnke M, Terwilliger JD, Ott J (1993) Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet 53:1127-1136

Segre D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. Nat Genet 37:77-83

Self SG, Liang KY (1987) Asymptotic Properties of Maximum-Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. Journal of the American Statistical Association 82:605-610

Sen S, Churchill GA (2001) A statistical framework for quantitative trait mapping. Genetics 159:371-387

Serre D, Nadon R, Hudson TJ (2005) Large-scale recombination rate patterns are conserved among human populations. Genome Res 15:1547-1552

Sham PC (1998) Statistics in human genetics. Arnold, Hodder Headline Group, London

Shook DR, Johnson TE (1999) Quantitative trait loci affecting survival and fertility-related traits in Caenorhabditis elegans show genotype-environment interactions, pleiotropy and epistasis. Genetics 153:1233-1243

Sillanpaa MJ, Arjas E (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148:1373-1388

Sillanpaa MJ, Arjas E (1999) Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. Genetics 151:1605-1619

Smalley SL, Asarnow RF, Spence MA (1988) Autism and genetics. A decade of research. Arch Gen Psychiatry 45:953-961

Sobel E, Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 58:1323-1337

Staessen JA, Wang JG, Brand E, Barlassina C, Birkenhager WH, Herrmann SM, Fagard R, Tizzoni L, Bianchi G (2001) Effects of three candidate genes on prevalence and incidence of hypertension in a Caucasian population. J Hypertens 19:1349-1358

Stephens DA, Fisch RD (1998) Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics 54:1334-1347

Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genomewide expression in yeast. Plos Biology 3:1380-1390

Strauch K, Fimmers R, Kurz T, Deichmann KA, Wienker TF, Baur MP (2000) Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: application to mite sensitization. Am J Hum Genet 66:1945-1957

Sturtevant AH (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. Journal of Experimental Zoology 14:43-59

Sugawara H, Harada N, Ida T, Ishida T, Ledbetter DH, Yoshiura K, Ohta T, Kishino T, Niikawa N, Matsumoto N (2003) Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23. Genomics 82:238-244

Sun L, Cox NJ, McPeek MS (2002) A statistical method for identification of polymorphisms that explain a linkage result. Am J Hum Genet 70:399-411

Szatmari P, Jones MB, Zwaigenbaum L, MacLean JE (1998) Genetics of autism: overview and new directions. J Autism Dev Disord 28:351-368

Tahri-Daizadeh N, Tregouet DA, Nicaud V, Manuel N, Cambien F, Tiret L (2003) Automated detection of informative combined effects in genetic association studies of complex traits. Genome Res 13:1952-1960

Tang HK, Siegmund D (2002) Mapping multiple genes for quantitative or complex traits. Genet Epidemiol 22:313-327

Templeton AR (2000) Epistasis and complex traits. In: Wade M, Brodie BI, Wolf J (eds) Epistasis and the evolutionary process. Oxford University Press, Oxford, pp 41-57

Terwilliger JD, Ott J (1994) Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore

Terwilliger JD, Speer M, Ott J (1993) Chromosome-based method for rapid computer simulation in human genetic linkage analysis. Genet Epidemiol 10:217-224

Tiwari HK, Elston RC (1997) Deriving components of genetic variance for multilocus models. Genet Epidemiol 14:1131-1136

Tiwari HK, Elston RC (1998) Restrictions on components of variance for epistatic models. Theor Popul Biol 54:161-174

Todorov AA, Borecki IB, Rao DC (1997) Linkage analysis of complex traits using affected sibpairs: effects of single-locus approximations on estimates of the required sample size. Genet Epidemiol 14:389-401

Tsai CT, Fallin D, Chiang FT, Hwang JJ, Lai LP, Hsu KL, Tseng CD, Liau CS, Tseng YZ (2003) Angiotensinogen gene haplotype and hypertension: interaction with ACE gene I allele. Hypertension 41:9-15

van der Veen JH (1959) Tests of non-allelic interaction and linkage for quantitative characters in generations derived from two diploid pure lines. Genetica 30:201-232

Veenstra-Vanderweele J, Christian SL, Cook EH, Jr. (2004) Autism as a paradigmatic complex genetic disorder. Annu Rev Genomics Hum Genet 5:379-405

Vieland VJ, Hodge SE, Greenberg DA (1992) Adequacy of single-locus approximations for linkage analyses of oligogenic traits. Genet Epidemiol 9:45-59

Vieland VJ, Huang J (2003) Two-locus heterogeneity cannot be distinguished from two-locus epistasis on the basis of affected-sib-pair data. Am J Hum Genet 73:223-232

Vionnet N, Hani El H, Dupont S, Gallina S, Francke S, Dotte S, De Matos F, Durand E, Lepretre F, Lecoeur C, Gallina P, Zekiri L, Dina C, Froguel P (2000)

Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. Am J Hum Genet 67:1470-1480

Wang DL, Zhu J, Li ZK, Paterson AH (1999) Mapping QTLs with epistatic effects and QTL × environment interactions by mixed linear model approaches. Theor Appl Genet 99:1255-1264

Ward R (1990) Familial aggregation and genetic epidemiology of blood pressure. In: Laragh JH, Brenner BM (eds) Hypertension: Pathophysiology, Diagnosis and Management. Raven, New York, pp 81-100

Weijnen CF, Rich SS, Meigs JB, Krolewski AS, Warram JH (2002) Risk of diabetes in siblings of index cases with Type 2 diabetes: implications for genetic studies. Diabet Med 19:41-50

Williams FM, Cherkas LF, Spector TD, MacGregor AJ (2004a) A common genetic factor underlies hypertension and other cardiovascular disorders. BMC Cardiovasc Disord 4:20

Williams SM, Ritchie MD, Phillips JA, 3rd, Dawson E, Prince M, Dzhura E, Willis A, Semenya A, Summar M, White BC, Addy JH, Kpodonu J, Wong LJ, Felder RA, Jose PA, Moore JH (2004b) Multilocus analysis of hypertension: a hierarchical approach. Hum Hered 57:28-38

Wiltshire S, Cardon LR, McCarthy MI (2002) Evaluating the results of genomewide linkage scans of complex traits by locus counting. Am J Hum Genet 71:1175-1182

Wiltshire S, Hattersley AT, Hitman GA, Walker M, Levy JC, Sampson M, O'Rahilly S, et al. (2001) A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. Am J Hum Genet 69:553-569

Wong ZY, Stebbing M, Ellis JA, Lamantia A, Harrap SB (1999) Genetic linkage of beta and gamma subunits of epithelial sodium channel to systolic blood pressure. Lancet 353:1222-1225

Wright S (1977) Evolution and the genetics of populations. Vol. 3. Experimental results and evolutionary deductions. University of Chicago Press, Chicago

Xu J, Meyers DA, Ober C, Blumenthal MN, Mellen B, Barnes KC, King RA, Lester LA, Howard TD, Solway J, Langefeld CD, Beaty TH, Rich SS, Bleecker ER, Cox NJ (2001) Genomewide screen and identification of gene-gene interactions for asthma-susceptibility loci in three U.S. populations: collaborative study on the genetics of asthma. Am J Hum Genet 68:1437-1446

Xu S (2003) Estimating polygenic effects using markers of the entire genome. Genetics 163:789-801

Xu X, Rogus JJ, Terwedow HA, Yang J, Wang Z, Chen C, Niu T, Wang B, Xu H, Weiss S, Schork NJ, Fang Z (1999) An extreme-sib-pair genome scan for genes regulating blood pressure. Am J Hum Genet 64:1694-1701

Yang X, Wang K, Huang J, Vieland VJ (2003) Genome-wide linkage analysis of blood pressure under locus heterogeneity. BMC Genet 4 Suppl 1:S78

Yi N, Xu S (2002) Mapping quantitative trait loci with epistatic effects. Genet Res 79:185-198

Zandi PP, Klein AP, Addington AM, Hetmanski JB, Roberts L, Peila R, Shrestha S, Shaw CK, Kiat HC, Langefeld CD, Beaty TH (2001) Multilocus linkage analysis of the German asthma data. Genet Epidemiol 21 Suppl 1:S210-215

Zeng ZB (1992) Correcting the bias of Wright's estimates of the number of genes affecting a quantitative character: a further improved method. Genetics 131:987-1001

Zeng ZB (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc Natl Acad Sci U S A 90:10972-10976

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457-1468

Zeng ZB, Kao CH, Basten CJ (1999) Estimating the genetic architecture of quantitative traits. Genet Res 74:279-289

Zinn-Justin A, Abel L (1998) Two-locus developments of the weighted pairwise correlation method for linkage analysis. Genet Epidemiol 15:491-510