Computational Methods for Compositional Epistasis Detection

by

Lu Cheng

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2022

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Ka Chun Wong
Professor, Dept. of Computer Science,
City University of Hong Kong

Supervisor(s): Mu Zhu
Professor, Dept. of Statistics & Actuarial Science,
University of Waterloo

Internal Member: Kun Liang
Associate Professor, Dept. of Statistics & Actuarial Science,
University of Waterloo

Internal Member: Yeying Zhu
Associate Professor, Dept. of Statistics & Actuarial Science,
University of Waterloo

Internal-External Member: Paul M. Craig
Associate Professor, Dept. of Biology,
University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In genetics, the term "epistasis" refers to the phenomenon that the effect of one gene or single-nucleotide polymorphism (SNP) is dependent on the presence of others. Various possibilities of epistasis exist, and the understanding of them is limited. In recent years, failure of replication for single-locus effects in genome-wide association studies (GWAS) motivates the exploration of epistasis for human complex disease.

This thesis is thus dedicated to the study of computational approaches for two-way compositional epistasis (SNP-SNP interaction) detection. Epistasis of this sort is best described by disease models, which can be simply understood as disease probability patterns associated with the genotype combinations of SNP-pairs. Because the epistasis detection problem requires determination of proper disease models to capture the compositional epistasis effect, it is more complicated than a typical variable selection task.

Three projects are pursued in this thesis. The first two target epistasis that is characterized by a set of "two-locus, two-allele, two-phenotype and complete-penetrance" (TTTC) disease model, and the third one extends to more general epistasis.

There are theoretically $2^9 = 512$ TTTC disease models. For a given SNP-pair, the first step of the problem is to find a proper TTTC model to capture its epistasis effect. It is found that existing methods that use data to determine best-fitting disease models prior to screening may be too greedy. Motivated by this, the first project proposes a less greedy strategy by limiting the search of disease models to a set of prototypes. The prototypes are determined a priori. Specifically, a distance metric is defined and used to cluster all disease models, and then a "representative" from each cluster is selected to form the prototypes. Compared to the existing approaches, the proposed method provides a more satisfying balance between precision and recall in epistasis detection.

If one uses data to determine a best-fitting disease model for a pair of SNPs, the nominal statistical evidence of association between the SNP-pair and the disease outcome is inflated. Therefore, the second project aims to directly correct inflation of this type. To make it feasible for genome-wide studies, a first-order correction method is proposed that can be applied in practice with no additional computational cost. Simulation studies are performed on two popular existing methods, which show that the correction is quite effective in improving an overall epistasis detection.

The TTTC disease models can be viewed as coding two risk levels, i.e., high and low risk. Compared to them, some other disease models code multiple risk levels, which capture more general epistasis patterns. Two methods are proposed in the third project, which are centered on epistasis detection using multi-level risk disease models. One method is inspired by the fused lasso under a regression-based framework, and adopts the post-model selection test to account for inflation incurred during disease model searching. The other one makes sequential split of the genotype combinations of a SNP-pair and uses a stopping criterion to determine the final disease model; after that, it also applies a first-order correction to the testing statistic to effectively account for inflation. It is shown that the two methods with totally different starting framework are equivalent in terms of the disease model searching process. Subsequent simulation studies show that use of multi-level disease models achieves better detection efficiency in terms of a balance between precision and recall than the two-level ones.

In summary, it is a rather complicated task to uncover the underlying mechanism of locus interaction effects, and endeavours are only beginning to be made. The epistasis detection methods in this thesis are practically useful at genome-wide level, which complements the single SNP screening in genome-wide association studies. What's more, the method of first-order correction for inflation is simple and effective, which is practically valuable for the epistasis detection methods involving inflated testing statistics.

Acknowledgements

The PhD study has been the longest journey I have ever taken. There have been many moments when I suffered from anxiety, not knowing if I could ever finish it. I therefore owe great gratitude to all the people who have provided me with help and encouragements along the way.

My greatest gratitude is owed to my supervisor Professor Mu Zhu who has supported me enormously. I had the greatest joy when discussing the research with him, which helped me fullfill my primary and fundamental motivation for the PhD study. Professor Zhu has been extremely understanding, patient and supportive to me in every aspect of my PhD study journey, without which I would not have been able to finish the study.

I would also like to thank the professors in the statistics department who have helped in various aspects of my PhD study, including but not limited to Professor Kun Liang, Yeying Zhu, Ali Ghodsi, Changbao Wu, Shoja'eddin Chenouri, and Paul Marriott.

I want to thank Mary Lou Dufton for her great help in many of the administrative procedures I went through and also her general support for me to carry on.

My sincere thanks also go to many of the good friends I made in University of Waterloo. Their company, encouragement and spiritual support are priceless. Specially, I want to express my great gratitude to my best friends Min Chen and Kexin Ji who have supported me unconditionally like family, and also Linbo Zhang who has kindly helped proofread the thesis and provided inspirational writing discussions.

Dedication

This is dedicated to my parents Xianmin Cheng and Qiong He.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Single-nucleotide polymorphisms (SNPs) (Box 1.1) are the most common variations in the human genome. They underlie individual differences in their susceptibilities to complex disease, and studying them is of fundamental importance in modern medical and health research.

For many years, scientists have tried to identify single-nucleotide polymorphisms (SNPs) associated with various diseases. Among all efforts, genome-wide association studies (GWAs) have been most popular since their inception at around 2005. In a typical GWA study, tagging SNPs are genotyped across the whole human genome and analyzed by comparing the relative frequencies of an allele or genotype between the cases and controls for each SNP [1].

Although GWAS for various complex diseases have been carried out extensively and the findings have been abundant, it is becoming apparent that single genetic variations can explain only very little heritability. This has come to be known as the so-called "missing heritability problem" [2–4]. Various conjectures for the missing heritability exist, such as rare variant effects and interactions between SNPs. In particular, many scientists believe that perhaps SNP-SNP interactions are more prevalent than one had previously thought [5].

Motivated by this, this thesis dedicates effort to the detection of SNP-SNP interactions. The work mainly consists of three parts: the first one is a proposal on using representative genetic disease models for the so-called “compositional epistasis ” detection that is of practical value; the second one is a simple and easy-to-use first-order p-value correction method that can effectively improve the epistasis detection efficiency of popular existing methods such as MDR [6] and Wan et al.’s method [7]; the third one consists of proposed algorithms for SNP-SNP interaction detection that places no restriction on the interaction patterns. The three parts are presented in chapters 2 - 4, respectively.

Before presenting the work, some background information is given in this chapter. The following sections start by introducing the definition, prevalence and types of epistasis, followed by a group of genetic disease models (i.e., “two-locus, two-allele, two-phenotype and complete penetrance” disease models) that can be used to represent the so-called “compositional epistasis ”. After that, some existing methods that achieve the same or similar epistasis detection are reviewed. Lastly, the ideas of the thesis pursuit are persented based on the understanding of the problems and drawbacks of existing methods.

1.1 Epistasis

In genetics, the term “epistasis” refers to the phenomenon that the effect of one gene (or SNP) is dependent on the presence of others. Epistasis was recognized long before the search for disease-associated loci. It goes back about 100 years when Mendel’s laws were being rediscovered, and interactions between genes were observed [8]. Recently, some systematic searches for genetic interactions affecting fitness and quantitative traits in model organisms such as yeast have been done, which confirm that epistasis is extremely prevalent [9–12].

Moore et al. (2003) [13] formulates epistasis as a ubiquitous component of the genetic architecture of human complex disease based on four aspects of reasoning. First, numerous observations of deviations from Mendelian ratios are due to interactions between genes, as have been evidenced since nearly 100 years ago; second, ubiquitous biomolecular interactions occurr in gene regulation, biochemical and metabolic systems, which indicate that interactions between genetic variants are likely to exist; third, the missing heritability problem could

Box 1.1 Glossary

- Genetic Association Studies: methods used to identify correlations between genetic polymorphisms and phenotypes such as diseases.
- Single-nucleotide Polymorphism (SNP): a genetic variant that consists of a single DNA base-pair change.
- Minor Allele Frequency (MAF): the frequency of the allele of a SNP that appears less often in the population.
- Penetrance: the probability of getting the disease for carrying the disease-associated genotype.
  - Complete Penetrance: carrying the disease-associated genotype leads to a 100% chance of developing the condition.
  - Incomplete Penetrance: carrying the disease-associated genotype leads to a <100% chance of developing the condition.
- Disease Model: a model used to map genotypes to the phenotype, could be represented by the penetrance values of each genotype under consideration.
  - Additive: each risk allele increases the risk of disease by one fold.
  - Recessive: two copies of the risk allele required to increase the risk.
  - Dominant: only one copy of the risk allele is required to increase the risk; the additional allele does not increase the risk further.
- Hardy-Weinberg Equilibrium (HWE): allele and genotype frequencies remain constant from generation to generation in a large, randomly mating and homogenous population. E.g., if the allele frequency of a SNP is α and the SNP satisfies HWE, then its genotype frequencies are expected to be $\alpha^2, 2\alpha(1-\alpha)$ and $(1-\alpha)^2$.
- Linkage Disequilibrium (LD): two (usually nearby) SNPs on the same chromosome being inherited together more often than expected by chance. In other words, two SNPs are not independent of each if in LD.
- Quantitative Trait Locus (QTL): genetic loci that contribute to variability in complex quantitative traits, as identified by statistical analysis.

be explained by epistasis effects of some sort; lastly, emerging methods have been yielding findings of gene-gene interaction effects. These reasons imply that epistasis is widespread and rather complex.

Over the years, the term “epistasis” has accumulated a broad range of meanings. In biochemical genetics, the term “functional epistasis” is sometimes used to refer to molecular interactions among proteins (and/or other genetic elements). In population and/or quantitative genetics, the terms “statistical epistasis” and “compositional epistasis” are more often used. Use of “statistical epistasis” is due to Fisher [14] and usually taken to mean deviation from additive genetic effects. “Compositional epistasis” emphasizes the notion of having a masking effect, as such, some researchers [15–17] believe it to be closer to the original meaning of the word “epistasis” when Bateson [8] first coined it in 1909. As Phillips [16] wrote, “compositional epistasis measures the effects of allele substitution against a particular fixed genetic background, while statistical epistasis measures the average effect of allele substitution against the population average genetic background.”

In existing literature, epistasis detection methods mainly aimed at “statistical epistasis”. As indicated by Phillips’ comment above, statistical epistasis is not invariant to study populations due to possible changes in the genetic background. Therefore, researchers have come to realize that its ability to uncover the biological mechanism is probably limited [18, 19]. A very limited amount of research exists that target epistasis more relevant to its biological meaning [7, 20–24]. Nevertheless, compositional epistasis, i.e., epistasis measured in Bateson’s sense, captures the aspects of epistasis relevant to selection dynamics and adaptation [25].

This thesis thus pursues compositonal epistasis detection. Disease models are introduced to conveniently characterize the compositional epistasis.

1.2 Disease Models

A genetic disease model (also called genetic model or disease model, see Box 1.1) is usually used to model an inheritance pattern, which describes how a disease is transmitted among generations. For instance, genetic dominant (only one copy of the allele inherited from either parent would cause the offspring to manifest the phenotype) and recessive models (two copies of the same allele are needed for the offspring to manifest the phenotype) are two common inheritance patterns for simple Mendelian disorders.

In this thesis, disease models concerning two-SNPs are of interest. Two sorts of disease models are introduced. The first one consists of a special group that is simple in form and practically usedful, and the second one is in the most general form and meant for flexible epistasis detection.

1.2.1 TTTC Models

To characterize different compositional epistatic effects, various researchers [17, 21, 26–29] have studied this problem and focused on a set of "two-locus, two-allele, two-phenotype, and complete-penetrance" (TTTC) disease models [30].

Table 1.1 shows a few examples. Often, these disease models can be interpreted as one SNP having a certain masking effect on the other. For instance, the recessive-recessive disease model (see model (c) in Table 1.1) can be viewed as the major allele "A" from one SNP having a masking effect on the causal genotype "bb" from the other SNP, or as the major allele "B" having a similar masking effect on the causal genotype "aa".

In practice, researchers have studied the joint two-locus effect utilizing particular epistatic models such as jointly recessive-recessive, recessive-dominant, dominant-dominant and modifying-effects models [6, 26, 28, 31, 32]. For instance, model (d) in Table 1.1 is known as the "XOR" model and shown to predict human handedness.

Clearly, the TTTC disease models can describe only two-way interactions between two SNPs, and the notion of epistasis itself certainly does not preclude higher-order interactions formed by more than two SNPs. Nonetheless, screening for higher-order interactions is still largely impractical at the genome-wide level. For example, even with 100,000 SNPs, there would be $\binom{100{,}000}{2} \approx 5.0 \times 10^9$ or about 5 *billion* SNP-pairs to screen already if limiting to 2-way

Table 1.1. Examples of TTTC disease models. A “1” means the corresponding genotype combination, e.g., “aabb” in (c), would cause the disease, whereas a “0” means it would not.

(a)				(b)			
	AA	Aa	aa		AA	Aa	aa
BB	**1**	**1**	0	BB	0	**1**	0
Bb	**1**	**1**	0	Bb	**1**	0	**1**
bb	0	0	0	bb	0	**1**	0
(c)				(d)			
	AA	Aa	aa		AA	Aa	aa
BB	0	0	0	BB	0	0	**1**
Bb	0	0	0	Bb	0	0	**1**
bb	0	0	**1**	bb	**1**	**1**	0

interactions only, and $\binom{100{,}000}{3} \approx 1.7 \times 10^{14}$ SNP-triplets to screen if 3-way interactions were to be considered. Therefore, in this thesis, a “narrow” point of view is taken by restricting the consideration to only two-way interactions.

What’s more, TTTC models (TTT models to be exact, or disease models with two different risk-levels) are practically useful, especially when the minor allele frequency (MAF) is low. A TTTC model has two degrees of freedom, which corresponds to two penetrance levels (Box 1.1) that are denoted by “1” and “0” respectively as shown in Table 1.1; whereas a “full model” will have 9 degrees of freedom, one for each of the nine genotype combinations. When the MAF is low, there can be insufficient data for some of the rare genotype combinations, making it hard to obtain reliable parameter estimates. In the extreme case, there may be no data in the sample for a particular genotype combination. Under such circumstances, it is beneficial to reduce the number of parameters or the degree of freedom. Using a TTTC model, one only has to estimate two parameters. By limiting the degree of freedom in this way, the power of the statistical test is expected to be improved.

Thus, when the word “epistasis” is used in this thesis (Chapter 2 & 3 especially, which is dedidated to use of TTTC disease models for epistasis detection), it is largely referring to these TTTC disease models only. Even so, there are still 2^9 possible TTTC disease models in theory [30] for each pair of SNPs, and it is generally not possible to screen them all. But a bad choice of the disease model can be detrimental in that a pair of SNPs may appear highly associated with an outcome under one disease model and not associated under another. For example, studies on single-locus effects have generally confirmed that the power (of detecting an existing effect) is largest when the correct genetic model such as recessive, dominant and additive is specified [33–35], so it is reasonable to expect that the same conclusion will hold for detecting epistatic effects between two SNPs.

Note that, just like some existing methods such as MDR and RS (see Section 1.3), the TTTC models are only used to capture the type of SNP-SNP interaction being considered (sometimes used to approximate a more complicated, non-TTTC model), but it does not mean the proposed method (again, method presented in Chapter 2 especially) in this thesis is only intended for TTTC models. In this regard, use of (or reliance on) these TTTC models is identical to MDR and RS.

1.2.2 General Disease Models

Epistasis effect does not necessarily have to be in the form as represented by a "TTTC" model. Therefore, general disease models are introduced with the aim of more general and flexible epistasis detection.

It is convenient to represent a general disease model by penetrance values. Penetrance is an important notion in genetic studies and refers to the probability of getting the disease for having a particular genotype (Box 1.1). Take dominant or recessive inheritance in simple Mendelian disorders as an example, the penetrance is 1 for the causal genotype(s) and 0 otherwise. A penetrance value of 1 is called complete or full penetrance, whereas reduced or incomplete penetrance refers to the situation in which possessing the disease-associated genotype(s) does not lead to a 100% chance of getting the disease. Reduced penetrance is more common in human complex genetic disorders.

Research on the missing heritability problem shows that small to medium single SNP effects are very probable. Therefore one plausible scenario would be that two SNPs contribute to the disease jointly through both marginal and interaction effects, where the marginal effects are only small to medium in size.

An extreme case is a type of two-locus disease models that code only pure interaction effects. In other words, SNPs that contribute to the disease under such an interaction model have no marginal effects. Previously, researchers have utilized techniques such as genetic programming to discover examples of such type of disease models [36, 37].

Examples of pure interaction disease models are shown in Table 1.2. Take Model (a) as an example. Assume it codes the true underlying interaction mechanism, then for the first SNP, its marginal penetrances are calculated as:

$AA : 0.558 \times 0.8^2 + 0.632 \times 2 \times 0.2 \times 0.8 + 0.546 \times 0.2^2 = 0.58$

$Aa : 0.616 \times 0.8^2 + 0.499 \times 2 \times 0.2 \times 0.8 + 0.674 \times 0.2^2 = 0.58$

$aa : 0.674 \times 0.8^2 + 0.418 \times 2 \times 0.2 \times 0.8 + 0.395 \times 0.2^2 = 0.58$

This shows that different genotypes of the first SNP have the same probabilities of getting the disease, i.e., the marginal effect of the first SNP is zero. A similar argument could be made for the second SNP.

Table 1.2. Examples of disease models without main effects. The values are the penetrances for the genotype combinations, i.e., probabilities of getting the disease for having that particular genotype combination.

(a) h^2=0.01 MAF=0.2				(b) h^2=0.01 MAF=0.4			
	AA	Aa	aa		AA	Aa	aa
BB	0.558	0.616	0.674	BB	0.095	0.122	0.127
Bb	0.632	0.499	0.418	Bb	0.097	0.129	0.010
bb	0.546	0.674	0.395	bb	0.201	0.044	0.122
(c) h^2=0.1 MAF=0.2				(d)h^2=0.1 MAF=0.4			
	AA	Aa	aa		AA	Aa	aa
BB	0.332	0.562	0.573	BB	0.539	0.120	0.258
Bb	0.583	0.112	0.147	Bb	0.165	0.378	0.325
bb	0.399	0.496	0.033	bb	0.123	0.426	0.276

1.3 Existing Methods

Among methods available for choosing a TTTC disease model for each pair of SNPs prior to screening, two popular ones are: the multi-factor dimensionality reduction (MDR) method by

Ritchie *et al.* [6], and the method by Wan *et al.* [7], which is simply referred as the “ratio split” (RS) method throughout. Both of these methods rely on the case-control ratios of different genotype combinations (i.e., AABB, AABb, and so on) in order to decide on a particular disease model to use for a given pair of SNPs.

Specifically, for a given SNP-pair (i, j), the MDR method determines a disease model (DM) $M_{i,j}$ by thresholding the case-control ratios; typically, genotype combinations with ratios ≥ 1 (on a balanced case-control sample) are regarded as high risk. The RS method, on the other hand, first sorts the case-to-control ratios in descending order and evaluates 8 different disease models by sequentially considering the top x genotype combinations as high risk, for $x = 1, 2, ..., 8$. Then, it chooses the one that best predicts the outcome (e.g., disease).

After the DM $M_{i,j}$ is determined, it is refitted by forming a 2×2 cross table

	Risky	Non-Risky	
Case	n_{11}	n_{10}	$n_{1\cdot}$
Control	n_{01}	n_{00}	$n_{0\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 0}$	n

according to how it separates the 9 genotype combinations into two groups: high and low risk.

For $k, \ell = \{0, 1\}$, let

$$\widehat{n}_{k\ell} = (n)\left(\frac{n_{k\cdot}}{n}\right)\left(\frac{n_{\cdot\ell}}{n}\right)$$

denote the expected count of cell-(k, ℓ) under the null hypothesis that the SNP-pair is independent of the outcome. The commonly used chi-squared statistic for testing the null hypothesis is computed as

$$\widehat{\chi}^2_{i,j} = \sum_{k=0}^{1} \sum_{\ell=0}^{1} \frac{(n_{k\ell} - \widehat{n}_{k\ell})^2}{\widehat{n}_{k\ell}}; \tag{1.1}$$

and the SNP-pair is then ranked (against other SNP-pairs) by the nominal p-value,

$$p_{\text{orig}}(i,j) = \Pr\bigl(\chi^2_{(1)} > \widehat{\chi}^2_{i,j}\bigr). \tag{1.2}$$

More details on the two methods are reviewed in Sections 1.3.1 and 1.3.2 below.

In addition to TTTC disease models, there are methods that use more general ones to model two-way epistasis effects. Among these, some assume 3 different levels of risk (i.e., 3 different penetrance values for the disease model), while others use flexible levels. Use of more general disease models may improve the SNP-pair detection in cases where the actual underlying epistasis involve more than two different risk levels. These are reviewed in Section 1.3.3.

1.3.1 MDR

Multifactor-Dimensionality Reduction (MDR) [6, 38] has been popular in detecting interaction effects since its first appearance in 2001, where it was used to detect high-order interaction effects for Sporadic Breast Cancer. There it successfully identified four-locus interactions that were marginally insignificant. For relatively small samples, MDR enjoys advantages such as being nonparametric and model-free. It could be viewed as a constructive induction learning method, because it converts high-dimensional interactions into a single-dimension factor with only two levels.

Application of MDR has detected interactions in a wide range of complex human disease, examples include breast cancer [6], Alzheimer disease [39], hypertension [38], autism [40, 41], asthma [42, 43], type 2 diabetes [44], bladder cancer [45, 46], prostate cancer [47], schizophrenia [48]. Due to its usefulness for binary traits, extensions exist for data samples of imbalanced designs [49] and for quantitative phenotypes and incorporation of environmental factors [50]. Additionally, in [51], the authors discuss the performance of MDR as a filter method for large-scale epistasis detection, where only the first step of the MDR but no cross-validation is applied.

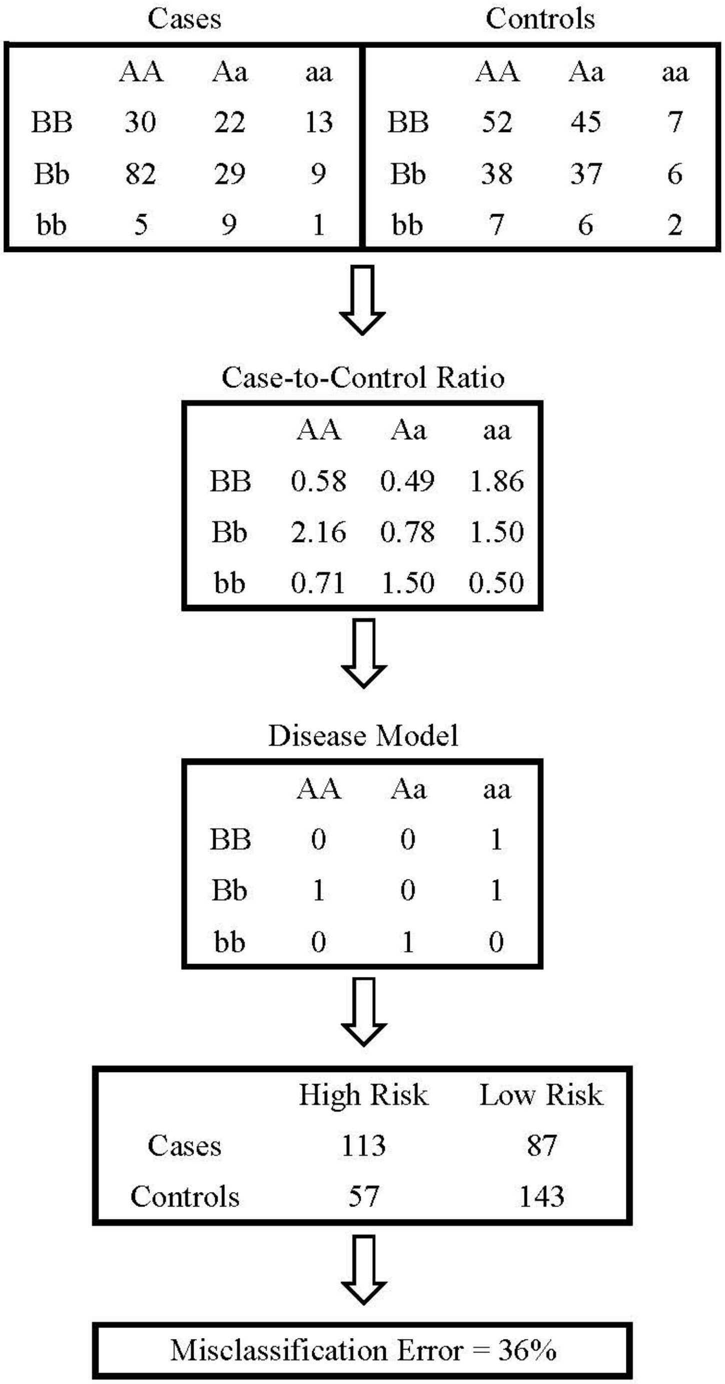
The idea of MDR is relatively simple and can be viewed as having two major parts, i.e., new factor construction and testing, and cross-validation. Assume there are p SNPs in total. For the first part, MDR in principal explores each k-order ($k \leq p$) interactions up to a possible maximum. The following steps depict the finding of a best-fitting k-order interaction.

Step 1 Form all possible $k-$order combinations of SNPs, the total number of which is $\binom{p}{k}$. For each one of them, there are 3^k different genotype combinations. The effect is assessed by the following procedure.

- Find out the number of people in cases and controls for each genotype combination formed by the k SNPs;
- Calculate the cases to controls ratios for each genotype combination;
- Assign the genotypes with ratios exceeding a threshold (e.g., for a balanced case-control sample, the threshold is usually 1 to be 'High Risk'; and the ones with ratios not exceeding the threshold as 'Low Risk').
- Use the assigned status as a classifier for the outcome status. Calculate the misclassification error.

Step 2 Select the combination of SNPs that has the smallest misclassification error as the final model.

An example of the procedure for $k = 2$ is illustrated in figure 1.1.



Cases	AA	Aa	aa	Controls	AA	Aa	aa
BB	30	22	13	BB	52	45	7
Bb	82	29	9	Bb	38	37	6
bb	5	9	1	bb	7	6	2

Case-to-Control Ratio

	AA	Aa	aa
BB	0.58	0.49	1.86
Bb	2.16	0.78	1.50
bb	0.71	1.50	0.50

Disease Model

	AA	Aa	aa
BB	0	0	1
Bb	1	0	1
bb	0	1	0

	High Risk	Low Risk
Cases	113	87
Controls	57	143

Misclassification Error = 36%

The original version of MDR employs a 10-fold cross-validation procedure, i.e., the procedure shown above is carried out 10 times and prediction errors are calculated on the corresponding testing data sets for model selection. To reduce the possibility of poor prediction error estimates due to chance divisions of the data set, the 10-fold cross-validation is repeated ten times, and the prediction errors are averaged. The best model of k-order interaction is selected based on the averaged prediction errors. Lastly, for all selected best k-order ($k = 1, 2, ...$) interactions, the final model is decided based on their prediction errors and selection consistencies, i.e., the number of times the same $k-$order combination is selected in the 10-fold cross-validation. After the final disease model is determined, statistical inference on the significance of a SNP-pair is conducted by refitting the disease model and calculating a p-value from permutation test.

Despite its various advantages, researchers have found that MDR tends to miss the correct interactions and select the wrong ones when the cases to controls ratios are close to that in the whole data, e.g., 1 for a balanced case-control sample, and also for situations in which very few observations exist for some combinations of genotypes. Ritchie et al. (2003) provided detailed studies on the power of MDR in the presence of missing data, genotyping error, phenocopy and genetic heterogeneity [52]. It was found out that MDR was not so affected by the first two factors, was slightly affected by 50% phenocopy and is greatly affected by 50% genetic heterogeneity.

Another significant drawback of MDR is the computational burden. As it explores interactions from order two up to a specified maximum or even the highest order, the procedure is exhaustive in nature. The original and widely used versions of MDR rely heavily on cross-validation procedures (i.e., 10-fold cross-validation) in model building, which adds up to the computation greatly. In addition, permutation for hypothesis testing carried out on the final selected models can also be computationally intensive. For this reason, alternative cross-validation and testing procedures exist, for example, a 5-fold cross-validation procedure was shown to be comparable in power to the original 10-fold one [53], and an extreme value distribution was shown to provide better computational efficacy than the permutation procedure for hypothesis testing [54].

1.3.2 Ratio Split Method

In Wan et al. (2013) [7], the authors have successfully implemented a computationally efficient procedure that applies to the GWAS scale for exact two-way compositional epistasis detection. They have followed Li and Reich (2001)'s 50 unique disease models [30] to guide the search.

Two-stage testing is carried out: the first stage screens out candidate SNP-pairs by testing a limited number of disease models; the second stage tests out the complete compositional epistasis models for selected SNP-pairs from the first stage.

The first stage screening is done by the following steps:

1. Get the frequency distributions of the nine genotypes formed by the pair in the cases and controls sample.

2. Calculate the frequency ratios of cases to controls.

3. Arrange the ratios in an ascending order. Split the ratios into two parts following the order. Generally, this yields eight different splits.

4. For each split, define a new covariate in the way that the genotypes with higher values of ratios are assigned 'High Risk', and those with lower ratios are assigned 'Low Risk'.

5. Test the effect of the new covariate by $\chi^2_{(1)}$ (χ^2-statistic with one degree of freedom). Out of the 8 splits, the one with the highest significance is selected and the corresponding epistasis model is used to capture the effect of the SNP-pair.

6. Candidate SNP-pairs are selected based on a user-specified significance level.

The authors borrow theories from classification trees to show that one of the splits from above minimizes the classification error for a two-class problem (Theorem 1.3.1). Theoretically there are $2^9 = 512$ two-way compositional epistasis models, the above procedure tests only 8, which is a huge reduction.

Theorem 1.3.1. *Suppose there is a categorical variable X taking categorical values from $1, 2, ..., M$ in two classes, class $Y = 0$ and class $Y = 1$. The categories are arranged in the ascending order of $P(Y = 1|X = i)$. Then one of $M - 1$ splits, $L = 1, ..., m$ and $R = m + 1, ..., M$ where $1 \leq m < M$, minimize the misclassification rate.*

An example illustrating the procedure is given below:

	$AABb$	$aaBB$	$aaBb$	$Aabb$	$AaBb$	$AAbb$	$AABB$	$aabb$	$AaBB$
Cases	82	13	9	9	29	5	30	1	22
Controls	38	7	6	6	37	7	52	2	45
Ratios	2.16	1.86	1.50	1.50	0.78	0.71	0.58	0.50	0.49

$\Downarrow$

Split 1	1	0	0	0	0	0	0	0	0
Split 2	1	1	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
Split 8	1	1	1	1	1	1	1	1	0

$\Downarrow$

		1	0	χ^2 Statistic
Split 1	Cases	82	118	
	Controls	38	162	22.012
	...	...	...	...
Split 4	Cases	113	87	
	Controls	57	143	30.946
	...	...	...	...

After the first stage screening, the authors propose a second stage testing of all non-redundant two-way compositional epistasis models (Li and Reich (2001)'s 50 disease models) on the selected SNP-pairs. The disease models with test statistics passing a given significance threshold are considered as the possible interaction patterns.

1.3.3 Multi-level Epistasis Models

Theoretically, two-locus epistasis can be represented by a general disease model with arbitrary penetrances levels. Among existing two-locus epistasis detection methods, different ones assume different levels of risk on the genotype combinations. Regression-based methods essentially assume 9 levels when the variables are treated as categorical, with both main and interaction effects included. MDR, RS and the proposed PTY method in this thesis assume two risk-levels.

A variation of MDR called the model-based MDR [55] assumes three risk-levels: "high", "low" and "no evidence", which was proved to help improve the power of the original MDR considerably. The risk groups are determined by both case-to-control ratios and the χ^2-test of independence. For each genotype of a SNP-pair, the test is performed by comparing its relative frequency of cases and controls to those of the rest genotypes as a whole. Genotypes with the case-to-control ratios larger than 1 and test p-values smaller than a pre-specified threshold are assigned to the high-risk category; genotypes with ratios lower than 1 and p-values smaller than the threshold are assigned to the low-risk category; and the rest ones are assigned to "no evidence". In this way, only those exhibiting significant evidence of high or low risk are identified to be so, whereas the ones with insignificant association or insufficient data samples are treated as "no evidence". As a result, the SNP-pair detection may be improved for better capturing or approximating the epistasis effects.

The "EDCF" (Epistasis Detector based on the Clustering of relatively Frequent items) method [56] also adopts three levels of risk. Unlike model-based MDR that uses the χ^2-test, "EDCF" conducts statistical tests by use of the Binomial distribution to differentiate genotype combinations that more excessively appear in the cases or controls. In more detail, assume

"n_c" and "n_u" denotes the number of cases and controls for a given genotype, "N_C" and "N_U" the number of total cases and controls in the whole sample, and "$p = \frac{N_C}{N_C+N_U}$" the sample prevalence. The null hypothesis is: the penetrance of each genotype is equal to the global sample mean (e.g., 1/2 for a balanced case-to-control sample). Under the null hypothesis, the number of cases should follow a Binomial distribution, i.e., $n_c \sim B(n_t = n_c + n_u, p_a = \frac{N_C}{N_C+N_U})$, therefore, a critical value corresponding to a given significance level α is $\{T_a : Pr(X > T_a|n_t, p_a)\}$. Based on this, the genotype is classified as a relatively frequent item in cases if $n_a > T_a$. In a similar way, a genotype can also be assessed for whether it is a frequent item in controls. If a genotype is not classified as either, it is assigned to the third group of "not enough evidence".

Pan et al. [57] proposed the "DCHE" (Dynamic Clustering for High-order Genome-wide Epistatic Interactions Detecting) method that models epistasis effects by disease models with three to six different levels of risk. The method determines a disease model for a SNP-pair by use of dynamic clustering in the following steps. First, a 2×9 cross-table is created for the distribution of cases and controls across the nine genotype combinations. Second, Pearson's $\chi^2_{(1)}$-test is carried out for all pairs of the genotype combinations based on the correponding 2×2 sub-tables; the pair showing the largest p-value is chosen, and if the p-value is larger than a pre-specified threshold, the two genotype combinations are merged and treated as a new group. The second step is repeated until no merges could be made or that there are only three genotype combination groups left. Lastly, all disease model patterns that have appeared in the merging process and have between three to six levels of risks are refitted against the outcome; p-values by Pearson's χ^2-test are calculated and the disease model with the most significant p-value is selected to model the epistasis effect.

1.4 Discussions and Research Proposals

For either MDR or RS, the DM $M_{i,j}$ that drives the 2×2 table as mentioned at the beginning of Section 1.3 is not just any DM, but one deemed to be "best-fitting" for the underlying

SNP-pair (i, j). Though the concept of "best-fitting" differs for MDR and RS, a post-hoc test of independence based on such a pre-selected $M_{i,j}$ will necessarily bias the test result toward being significant.

In fact, both the MDR and RS methods are essentially greedy and use the disease outcome data twice: first, to determine the disease model for each pair of SNPs; then, to determine whether each pair of SNPs is associated with the outcome. As such, they can be overly adaptive to data, and have a tendency to produce many false positives. The cost of using the data twice is especially pronounced if the sample size is relatively small (which is almost always the case for genome-wide association studies), and/or if the data quality is not so good.

This kind of concern has been reported in the literature [51]. Though extra out-of-sample validation can help mitigate such problems, in the context of genome-wide studies it is computationally prohibitive. For this reason, this thesis does not pursue such kind of validation. In designing the methods, the focus is on performing large-scale epistasis screening similar to what is done for single SNPs in GWAS. Therefore, when using MDR or R-S as comparison methods, only their core ideas in disease model determination and epistasis testing are applied.

1.4.1 Proposal for Prototype Disease Model

To address the problem of tendencies to produce false positives by MDR and RS, the first proposed method is to use a set of pre-selected disease models for epistasis testing. By doing so, it is expected that the problem of being overly adaptive to data is avoided.

The main idea for disease models selection is acquired from observing that some of the TTTC disease models are more similar than others. In Table 1.1, for example, arguably model (a) and (c) are quite different from each other, whereas model (b) and (d) are somewhat similar to each other.

More specifics on how to measure the similarity between two disease models will be explained later (Section 2.4.1). For now, it is observed that all possible disease models can be grouped into a few clusters, and a representative prototype can be selected from each cluster for screening purposes. The set of prototype models can be viewed to place a constraint on the search space, in the sense that only disease models in the prototype set are now “permitted”. This allows the proposed method to be less data-adaptive, while still ensuring that important parts of the search space are not missed out because a prototype from each cluster is included. In what follows, the acronym “PTY” (for “prototype”) is used to refer to the proposed method, especially in tables and figures. All details of PTY, including the similarity measure and selection of disease models, simulation and real data application are given in Chapter 2.

It is worth mentioning that a cluster analysis of all disease models is beneficial in its own right. For example, it may allow people to better understand and characterize different epistatic effects (more on this in Section 2.6), for which there have been a few previous endeavours [27, 29, 30, 58].

1.4.2 Proposal for P-value Adjustment

In the MDR and RS methods, because the testing statistic in Eq. 1.1 are inflated, the p-value calculated through Eq. 1.2 tends to produce false positive discoveries under the usual choice of significance level. With this understanding, the second proposed method aims to derive a more accurate estimation on the distribution of the statistic (i.e., Eq. 1.1) for statistical testing.

Solutions could be to derive the exact distribution for the testing statistic, or to simulate the null distribution and compute an empirical p-value for each SNP-pair. While the former can be mathematically challenging and the distributions differ for different disease model determination procedures, the latter requires too large a number of simulations to be carried out at a genome-wide level. Therefore, a shortcut is proposed and the idea is to use a first-order correction that can be applied in practice with essentially no additional computational cost.

Specifically, the statistic is assumed to still follow χ^2-distribution, but with an elevated degree of freedom. Under this assumption, an empirical approach can be taken to estimate the effective degree of freedom (EDF). In practice, the EDF is found to be different for different SNP-pairs and sample size. Therefore, the aim is to express the EDF as a function of the MAFs of the SNPs and sample size. In this way, the computational cost associated with these "extra" steps to calculate a degree-of-freedom correction is almost negligible.

Despite the assumption that the test statistic still has a chi-squared null distribution, various simulation results have indicated that the proposed first-order correction to the p-value is quite effective at improving the performance of popular methods such as the MDR and the RS for screening pairwise SNP-SNP interactions. The details of this method and results are presented in Chapter 3.

1.4.3 Proposal for Multiple-Level Disease Model

Because the true epistasis may contain multiple risk levels, and that a more accurate specification of the disease model is expected to improve epistasis detection, a natural extension on use of TTTC models is to test the epistasis effect according to its actual number of risk levels. Existing methods reviewed in Section 1.3.3 only provide partial solutions, e.g., the disease models are fixed or still restricted in some way, which are not completely flexible. Motivated by this, this thesis proposes two methods that can determine a disease model of flexible risk levels.

The first idea to achieve this is inspired by fused lasso [59], which can give identical coefficient estimates for neighbour variables. When the problem is formulated in a regression model with indicator variables for the genotypes, a set of cofficient estimates (with some being identical) can be translated to a disease model of multiple levels. To assess the effect of a SNP-pair, the determined disease model is refit and Pearson's χ^2-statistic is calculated. Due to disease model selection in the process, this statistic is inflated if evaluated by the "nominal" degree of freedom of a χ^2-distribution. Therefore, the recent post-model selection test is adopted for the evaluation, which is able to account for the selection.

The second method is inspired by the RS method, which is to continue the splitting till a desirable disease model level is obtained. For a SNP-pair with 9 genotype combinations, theoretically there are $2^9 - 2 = 510$ and $3^9 - 3 \times (2^9) + 3 \times 1^9 = 18150$ two- and three-level disease models. Foreseeably, there are many more of the four and five-level diease models. Due to the large number of possibilities, it is impossible to perform an exhaustive search in the whole model space for the best fitting one. Therefore a greedy algorithm is adopted, which performs the split in a sequential way, i.e., each additional split is built on the existing ones. Alternatively, instead of splitting, a merging procedure can be carried out. Starting with 9 genotypes, the two most similar ones are found by comparing their distribution differences between the cases and controls. Then they are merged to be a new group and the merging procedure continues in this fasion till a desirable disease model level is obtained.

The split and merge procedures require two types of criteria and an evaluation method: 1) criteria to determine the best split or merge place during the sequential search process; 2) stopping criteria to determine the levels of the disease model; 3) and an evaluation method to assess the effects of SNP-pairs. For the evaluation method, the determined disease model is refitted and Pearson's χ^2-statistic is adopted. Due to the searching of disease model in the process, the testing statistics suffer from inflation. Therefore, similar p-value adjustments to that introduced in Section 1.4.2 are carried out.

The details with simulation results are presented in Chapter 4.

Chapter 2

Prototype Disease Model

This chapter presents compositional epistasis detection using TTTC-type models. Specifically, a method called "prototype disease models" (PTY) is proposed. Before the method is presented, an empirical protocal used through out this chapter for SNP-pair screening is given. The method presentation then starts with an introduction on some motivation examples for the PTY method, followed by proposal of a similarity metric used to cluster disease models. After that, the clustering algorithm is introduced with PTY selection. Lastly, simulation studies and real data application are carried out. The work in this chapter has been written as a journal paper and published in Plos One [60].

2.1 Marginal versus sequential screening

Throughout the chapter, the following empirical protocol is used repeatedly to compare different methods. Specifically, either marginal or sequential screening is applied for SNP-pair detection. While the former simply ranks the SNP-pairs, the latter provides a procedure to model aggregated epistasis effects. In the paragraphs below, the TTTC models are used as examples to illustrate the two procedures, which also directly apply to general epistasis models.

For any given pair of SNPs, e.g., (i, j), each method has its own way of determining a "best-fitting" disease model—call it $M_{i,j}$. The ways by MDR and RS are given in 1.3, and the way by PTY is selecting the best-fitting one out of the candidate "prototype" disease models (more on this in 2.6.3). A *nominal* measure of association between the (i, j)-pair and the outcome is then computed as the $\chi^2_{(1)}$-statistic for testing whether the risky/non-risky assignment by $M_{i,j}$ is statistically independent of the outcome (i.e., diseased or non-diseased), which is simply denoted as $\widehat{\chi}^2_{i,j}$. (Explanation why the adjective "nominal" is used to describe these association measures is given later in Section 2.2.3 in more detail.) The pair (i, j) can then be ranked according to $\widehat{\chi}^2_{i,j}$ or considered having been "selected" or "detected" by the method if $\widehat{\chi}^2_{i,j}$ exceed a certain significance threshold (Two threholds are adopted in this chapter, which are expained in Section 2.7.3.). This is referred as the "marginal screening procedure". (One can also use only part of the data to determine $M_{i,j}$, and compute an *out-of-sample* measure of association by testing $M_{i,j}$ against the outcome on the remaining data. For example, MDR is usually applied in this manner when the number of candidate SNPs being studied is relatively small. To reduce variation caused by chance division of the data, however, such a process often needs to be repeated a few times and the resulting measures averaged, thus making it computationally prohibitive for genome-wide screening [51, 61].)

Alternatively, the effects of multiple SNP-pairs can also be combined sequentially. For example, after having selected the top pair—call it (i_1, j_1), each remaining pair (i, j) can be re-assessed by testing whether the *combined* high/low risk assignment by

$$M_{i_1,j_1} \text{ or } M_{i,j} \tag{2.1}$$

is independent of the outcome. $\widehat{\chi}^2_{i,j|\mathcal{H}}$ is used to denote the corresponding test statistic, where $\mathcal{H}$ means the entire *history* of pairs already selected so far(After the top pair has been selected, $\mathcal{H} = \{M_{i_1,j_1}\}$; after two pairs have been selected, $\mathcal{H} = \{M_{i_1,j_1}, M_{i_2,j_2}\}$; and so on.). The pair to be selected next is

$$\underset{M_{i,j} \notin \mathcal{H}}{\arg\max} \quad \widehat{\chi}^2_{i,j|\mathcal{H}}, \tag{2.2}$$

rather than

$$\underset{M_{i,j} \notin \mathcal{H}}{\arg\max} \quad \widehat{\chi}^2_{i,j}. \tag{2.3}$$

This is referred to as the "sequential screening procedure".

2.2 Motivating Examples

An effort needs to be made in determining proper disease models to test for genotype to phenotype mapping. This task could be viewed as a selection of disease models, a prerequisite for selection of SNPs in epistasis detection. Moore et al. (2006) [62] propose a flexible computational framework for detecting and characterizing epistasis, in which a key step is to construct new attributes that capture interaction information.

Both the MDR and RS methods use case-to-control ratios to estimate a proper TTTC-type disease model for the current SNPs under investigation. As discussed in the introduction chapter, these methods may suffer from inflation due to the use of data twice. In this section, some motivating examples are provided which demonstrate weaknesses of existing methods. It needs to be emphasized that these are merely some *examples* of scenarios in which PTY can be seen to have certain advantages over MDR and RS. They are by no means the only or even necessarily the main scenarios. The reason why they are being presented, rather than others, is that they are still relatively easy to describe with a reasonable amount of clarity, whether algebraically (Section 2.2.1), verbally (Section 2.2.2), or both (Section 2.2.3).

2.2.1 A pathological scenario

A pathological scenario is considered here for an easy illustration purpose. Suppose that two pairs of SNPs (e.g., {A/a, B/b}, {C/c, D/d}) are independent (Table 2.1). For $i = 1, 2, ..., 9$, let w_i be the relative frequency of the i-th genotype combination in the first pair, and likewise v_j for the second pair. For simplicity, suppose each genotype combination is either risky ($\in R$) or non-risky ($\in N$). For $k, \ell \in \{0, 1\}$, let $p_{k\ell}$ (Box 1.1) be the penetrance level for individuals having risky combinations from both pairs ($k = \ell = 1$), the first pair only ($k = 1, \ell = 0$), the second pair only ($k = 0, \ell = 1$), or neither ($k = \ell = 0$). Then it can be shown that if

$$p_{11} \sum_{j \in R} v_j + p_{10} \sum_{j \in N} v_j = p_{01} \sum_{j \in R} v_j + p_{00} \sum_{j \in N} v_j \tag{2.4}$$

holds, then the case-control ratios of the first pair are all the same for all its genotype combinations $i = 1, 2, ..., 9$, regardless of whether $i \in R$ or $i \in N$. It is thus a pathological case, in which it would be impossible to rely on the case-control ratios to determine the disease model.

Proof. When the SNPs are not in linkage disequilibrium (Box 1.1, or simply understood as being independent), the population is distributed as:

Genotype Combination Pair 1/Pair 2	Diseased	Non-diseased
Risky/Risky	$p_{11} \sum_{i \in R} w_i \sum_{j \in R} v_j$	$(1 - p_{11}) \sum_{i \in R} w_i \sum_{j \in R} v_j$
Risky/Non-risky	$p_{10} \sum_{i \in R} w_i \sum_{j \in N} v_j$	$(1 - p_{10}) \sum_{i \in R} w_i \sum_{j \in N} v_j$
Non-risky/Risky	$p_{01} \sum_{i \in N} w_i \sum_{j \in R} v_j$	$(1 - p_{01}) \sum_{i \in N} w_i \sum_{j \in R} v_j$
Non-risky/Non-risky	$p_{00} \sum_{i \in N} w_i \sum_{j \in N} v_j$	$(1 - p_{00}) \sum_{i \in N} w_i \sum_{j \in N} v_j$

□

Consider the i-th genotype combination in the first pair, {A/a, B/b}. From the table above, it can seen that, if it is risky ($i \in R$), then its (marginal) case-control ratio is

$$r_R = \frac{p_{11} w_i \sum_{j \in R} v_j + p_{10} w_i \sum_{j \in N} v_j}{(1 - p_{11}) w_i \sum_{j \in R} v_j + (1 - p_{10}) w_i \sum_{j \in N} v_j} = \frac{p_{11} \sum_{j \in R} v_j + p_{10} \sum_{j \in N} v_j}{1 - (p_{11} \sum_{j \in R} v_j + p_{10} \sum_{j \in N} v_j)};$$

whereas if it is non-risky ($i \in N$), the ratio is

$$r_N = \frac{p_{01} w_i \sum_{j \in R} v_j + p_{00} w_i \sum_{j \in N} v_j}{(1 - p_{01}) w_i \sum_{j \in R} v_j + (1 - p_{00}) w_i \sum_{j \in N} v_j} = \frac{p_{01} \sum_{j \in R} v_j + p_{00} \sum_{j \in N} v_j}{1 - (p_{01} \sum_{j \in R} v_j + p_{00} \sum_{j \in N} v_j)}.$$

It is easy to see that if Eq. (2.4) holds, then $r_R = r_N$; that is, the case-control ratio will be the same for the i-th genotype combination in the first pair, regardless of whether $i \in R$ or $i \in N$.

Since both MDR and RS rely on the case-control ratios to determine disease models, their powers (of detecting the relevant pair) can be expected to be greatly affected if Eq. (2.4) holds, even if only approximately.

Remark

For case-control data, both r_R and r_N would be inflated by a factor of $[1 - \mathbb{P}(D)]/\mathbb{P}(D)$, where $\mathbb{P}(D)$ is the disease prevalence. However, this would not affect the previous conclusion.

Simulation Examples

To offer a more concrete illustration, two examples (see Table 2.2) are simulated. In the first one, the true disease models are the same for the two relevant SNP-pairs; in the second, they are different. The penetrance parameters p_{10}, p_{01} and p_{00} are predetermined, and a few different values are explored for the last penetrance parameter, p_{11}, around the value implied by Eq. (2.4). Note that only when p_{11} is equal to the value implied by Eq. (2.4), there would be no signal left; otherwise there is still some weak signal left for the pair to be detectable by using the case-to-control ratios. The simulation was repeated 100 times, with a total of 100 SNPs and a sample size of $n = 800$.

Table 2.1. Analytic framework for Section 2.2.1. Two SNP-pairs (where each w_i, v_j denotes the relative frequency of the respective genotypes) and four penetrance levels ($p_{k\ell}$, $k, \ell \in \{0, 1\}$). Certain relationships among the four penetrance parameters, e.g., Eq. (2.4), can make it impossible to determine an appropriate disease model for the underlying pair based on the case-to-control ratios.

Pair 1				Pair 2				Penetrance		
	BB	Bb	bb		DD	Dd	dd	$\text{Pair}_1 \backslash \text{Pair}_2$	R	N
AA	w_1	w_2	w_3	CC	v_1	v_2	v_3	R	p_{11}	p_{10}
Aa	w_4	w_5	w_6	Cc	v_4	v_5	v_6	N	p_{01}	p_{00}
aa	w_7	w_8	w_9	cc	v_7	v_8	v_9			

Simulation Results

The results are assessed by looking at the number of times each pair is successfully detected by each method. Note the sequential screening procedure is adopted to consider the joint effect of two pairs at the same time, and a relevant pair is considered to have been successfully detected if it is among the top two pairs selected by the method.

The results are shown in Fig. 2.1. Note all three methods have detected the second pair ({C/c, D/d}) perfectly (i.e., 100 times out of 100 replications). This is probably because the effect of the second pair is stronger than that of the first ({A/a, B/b}), e.g., $p_{01} > p_{10}$. For the first pair, the proposed method, PTY, generally has a better detection rate on it than MDR and RS. In particular, as the parameter p_{11} drops (from a value quite different from that implied by Eq. (2.4) to the exact value implied by Eq. (2.4)), both MDR and RS start to deteriorate in their ability to detect the first pair, whereas PTY remains largely unaffected.

Discussion

To better understand Eq. (2.4), notice that it can be rearranged slightly as

$$p_{11} = p_{01} - \frac{\sum_{j \in R} v_j}{\sum_{j \in N} v_j} (p_{10} - p_{00}). \tag{2.5}$$

Table 2.2. Simulated examples for Section 2.2.1. Disease models for the two pairs of SNPs that contribute to the simulated outcome. The penetrance parameters, $(p_{00}, p_{01}, p_{10}, p_{11})$, are chosen so that the case-control ratio is the same for all genotype combinations $i = 1, 2, ..., 9$ in the first pair, {A/a, B/b}.

Example 1: Two SNP-pairs, identical disease models (MAF=0.3).								
	BB	Bb	bb			DD	Dd	dd
AA	0	0	1		CC	0	0	1
Aa	0	1	0		Cc	0	1	0
aa	1	0	0		cc	1	0	0
$(p_{10} = 0.1,\ p_{01} = 0.28,\ p_{00} = 0.01 \overset{\text{Eq. (2.4)}}{\Rightarrow} p_{11} = 0.03.)$								
Example 2: Two SNP-pairs, different disease models (MAF=0.2).								
	BB	Bb	bb			DD	Dd	dd
AA	0	0	0		CC	0	1	1
Aa	0	1	1		Cc	1	0	0
aa	0	1	1		cc	1	0	0
$(p_{10} = 0.09,\ p_{01} = 0.12,\ p_{00} = 0.001 \overset{\text{Eq. (2.4)}}{\Rightarrow} p_{11} = 0.016.)$								

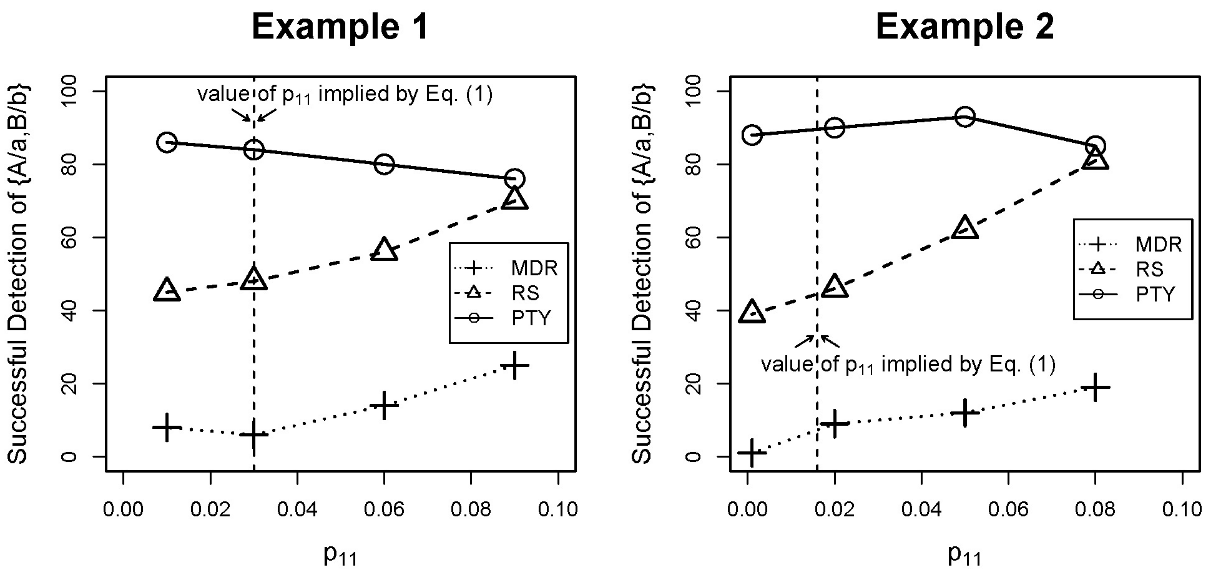


Fig 2.1. Simulated examples for Section 2.2.1. Number of times the first pair, {A/a, B/b}, is successfully detected (out of 100 repetitions) as the parameter p_{11} varied.

Notice it can be typically expected that $p_{10} > p_{00}$, i.e., having a risky genotype in the first pair increases the probability of disease. Hence, Eq. (2.5) implies that $p_{11} < p_{01}$, or that having risky genotypes from both pairs will actually lead to a reduced probability of disease than having risky genotype(s) only from the second pair. This is analogous to the logical operator, “exclusive or” (XOR) (Refer to disease model (d) in Table 1.1 for an example of an “XOR”-type disease model formed by two SNPs. There the genotype “aa” and “bb” are risky for the first and second pair marginally, but “aabb” is non-risky.).

While one can certainly argue that this may be a totally hypothetical scenario that is not likely to occur in real world, it is nonetheless a theoretical possibility against which the proposed method, PTY, is robust.

Remark

Of course, the aforementioned XOR-type relationship means the two pairs, {A/a, B/b} and {C/c, D/d}, are interacting with each other, so there is actually a four-way interaction across the four SNPs involved. Such a high-order interaction still could be detectable by methods such as the MDR or the RS if four-way disease models were considered and screened; but, as stated earlier (Section 1.2.1), this study takes a “narrow” point of view by restricting the consideration to only two-way interactions. Indeed, there is nothing “pathological” about having a high-order interaction; it is only “pathological” when one is restricted to consider only two-way interactions.

2.2.2 Detection of spurious effects

It is also observed that being overly adaptive to data can cause a method to be more easily tricked into detecting spurious epistatic effects, e.g., by SNPs with large individual effects. To demonstrate this, 100 SNPs are simulated on a case-control sample of size $n = 200$. Two pairs of SNPs, i.e., {A/a, B/b} and {C/c, D/d}, contribute to the simulated outcome independently, each according to an additive disease model (see Table 2.3). The SNPs A/a and C/c are simulated to have higher minor allele frequencies (MAFs) than B/c and D/d so that they have larger marginal individual effects than the other two according to the underlying additive disease model.

The simulation is repeated 100 times. The pairs most frequently ranked by each method to be among the top two using the sequential screening procedure (Table 2.4) are counted. Results show that both MDR and RS are more likely to select a spurious pair, {A/a, C/c}, due to the large marginal effects of both of these SNPs. They are much less effective than the proposed method, PTY, in identifying the truly relevant pairs.

Table 2.3. Simulated examples for Section 2.2.2. Disease models for the two pairs of SNPs that contribute to the simulated outcome. Numeric values (e.g., 0.1, 0.2) are penetrance parameters for the corresponding genotype combinations. (MAF=0.5 and 0.3, respectively, for the two SNPs in each pair.)

	BB	Bb	bb		DD	Dd	dd
AA	0	0	0.1	CC	0	0	0.1
Aa	0	0	0.1	Cc	0	0	0.1
aa	0.1	0.1	0.2	cc	0.1	0.1	0.2

Table 2.4. Simulated examples for Section 2.2.2. The number of times different pairs of SNPs were among the top two pairs detected, out of 100 replications. The truly relevant pairs are emboldened.

MDR		RS		PTY	
{A/a, C/c}	75	{A/a, C/c}	74	**{A/a, B/b}**	43
{B/b, D/d}	32	{B/b, D/d}	51	**{C/c, D/d}**	42
{C/c, D/d}	13	{A/a, D/d}	12	{A/a, C/c}	32
{A/a, D/d}	13	{B/b, C/c}	9	{A/a, D/d}	23
{A/a, B/b}	10	**{C/c, D/d}**	8	{B/b, C/c}	9
{B/b, C/c}	10	**{A/a, B/b}**	7	{B/b, D/d}	6
Other Pairs	≤ 9	Other Pairs	≤ 6	Other Pairs	≤ 5

2.2.3 Exaggeration of effects and false positives

In previous sections, it is already stated that both MDR and RS tend to produce many false positives. To demonstrate this point more concretely, another experiment is conducted. 100 SNPs are simulated on a case-control sample of size $n = 400$, except that, this time, *none* of the SNPs is related to the simulated outcome. Then, all three methods, i.e., MDR, RS, and PTY, are followed to assess the resulting $\binom{100}{2} = 4,950$ pairs of SNPs. The distributional properties of the resulting association measures (see Section 2.1) produced by each method for all pairs, $\{\widehat{\chi}^2_{i,j} : 1 \leq i, j \leq 100\}$, are examined.

Fig. 2.2 shows various Q-Q plots of these association measures, produced by different methods under different MAF settings, against the theoretical quantiles of the $\chi^2_{(1)}$-distribution. It can be seen that all methods produced inflated association measures, which would lead to false discoveries if they are assessed against the $\chi^2_{(1)}$-distribution. This is not surprising though. After all, $M_{i,j}$ was not just any disease model but the one deemed "best-fitting" for the underlying pair (i, j). Though the meaning of "best-fitting" differed for the three methods, a post-hoc test of independence based on $M_{i,j}$ was clearly biased toward being significant. This is why the adjective "nominal" is used earlier in Section 2.1 to describe these association measures.

However, the main point here is that PTY suffers the least from this tendency to produce false positives. As the MAF increases, the tendency to produce false positives also becomes more pronounced for both MDR and RS, but not for PTY. To further illustrate this point, the aforementioned "null simulation" is repeated 400 times. For each repetition, the mean value of the (nominal) association measure is computed,

$$\frac{1}{4950} \sum_{i,j} \widehat{\chi}^2_{i,j}, \tag{2.6}$$

across all 4,950 SNP-pairs. The average of these mean values and its standard error over the 400 repetitions are shown in Table 2.5 for each method under different MAF settings. Clearly, this value is more inflated for MDR and especially for RS than it is for PTY.

Table 2.5. Results from simulation study (Section 2.2.3). Average values of the nominal association measures $\{\widehat{\chi}^2_{i,j} : 1 \le i, j \le 100\}$ across all 4,950 SNP-pairs, together with their standard errors, over 400 repetitions.

MAF	MDR	RS	PTY
0.05	1.837 (0.229)	2.115 (0.233)	1.792 (0.214)
0.10	2.457 (0.239)	2.975 (0.230)	2.202 (0.200)
0.40	4.748 (0.289)	5.101 (0.303)	3.033 (0.221)

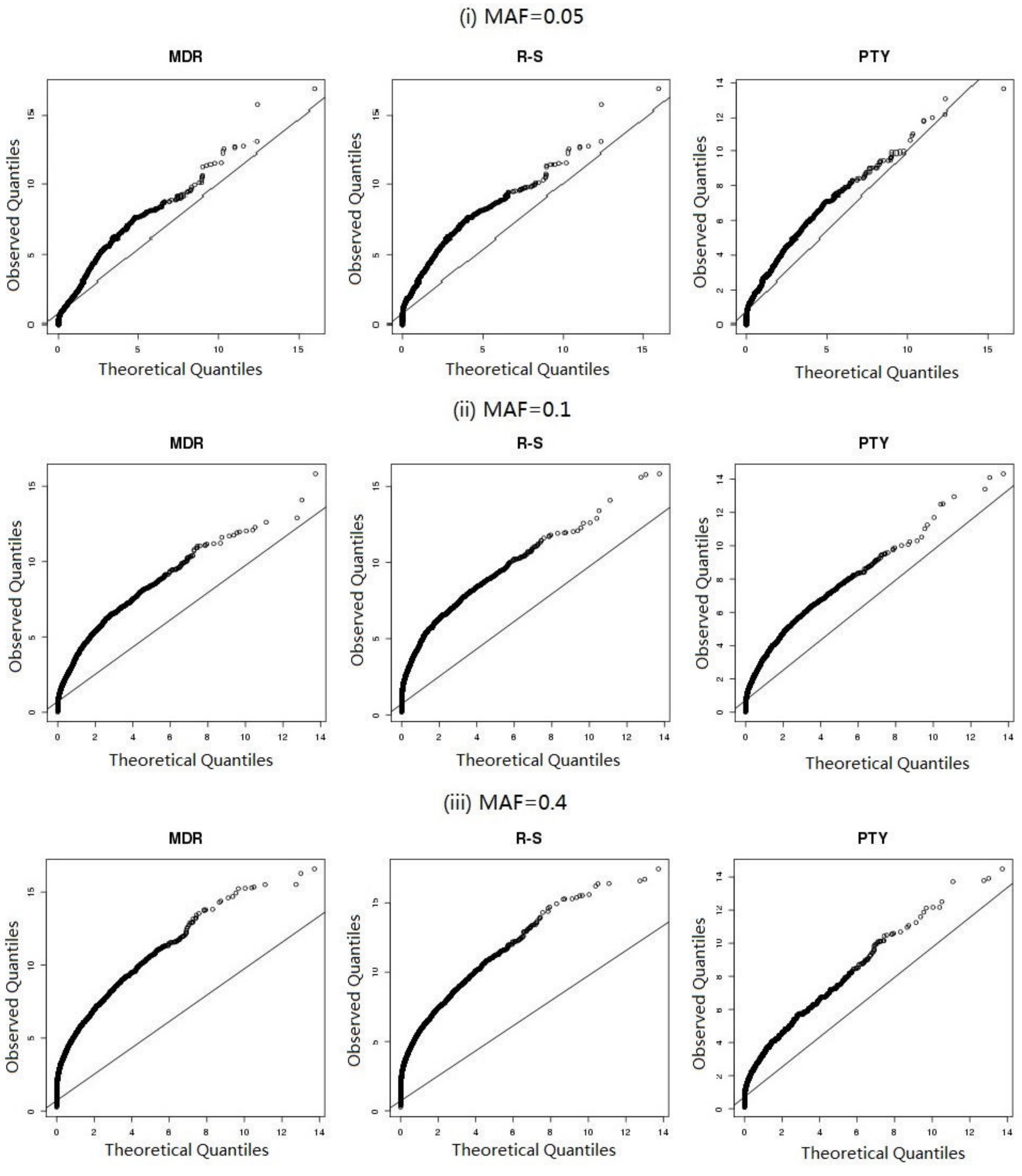


Fig 2.2. Results from simulation study (Section 2.2.3). Q-Q plots of nominal association measures $\{\widehat{\chi}^2_{i,j} : 1 \leq i, j \leq 100\}$ against their theoretical quantiles.

2.3 Overview

Examples in the previous section show that there are situations in which disease models based on the case-to-control ratios lead to reduced power in epistasis detection. Cases include a pathological example in which the case-to-control ratios become non-informative, and an example demonstrating the inclination to false positive selections when large marginal effects of single SNPs are present.

To overcome the problems, an approach without using the case-to-control ratios is proposed, and the idea is to use disease model clustering to select a few prototypes as representatives. Clustering the disease models is natural for characterization when they are similar to each other. Besides, after clustering, those closest to the centers are natural prototype examples that could be used for epistasis testing. Approaching it this way, the disease models to be used are not affected by the disease outcome data, so the epistasis detection is less likely to suffer from a power loss.

Intuitively, if two different disease models have many genotypes with the same or close penetrance values, then they could indeed be viewed as being "similar". Nonetheless, note that taking the genotype frequencies into account is necessary considering that the eventual goal is to do SNP-pair selection using proxy disease models. To understand this, consider the models in Table 2.6. Suppose that $DM0$ is the true disease model, and $DM1$ and $DM2$ are two candidate proxy models. If using $DM1$ as the proxy model, then individuals with the high-risk genotype of AaBb would be predicted as being low-risk; whereas if using $DM2$ as the proxy model, the ones with aabb would be predicted as low-risk. Given a MAF of 0.1, there is expected to be more individuals with AaBb. In other words, using $DM1$ as the proxy model would lead to more incorrect predictions than that of $DM2$. From this sense, it is more reasonable not only to consider the disease model penetrance values for all genotypes, but also the corresponding genotype frequencies.

Based on the above consideration, "similarity" can be defined in such a way that two disease models are regarded as "similar" if they lead to similar results when used to predict the disease outcomes. That is, the numbers of "identical" predictions two disease models

Table 2.6. Disease model examples that are candidate proxy ones for the true model. Assume the SNP pair under study has MAF of 0.1, then the genotype frequency for aabb is lower than that of AaBb, i.e., there would be more individuals who are correctly predicted as having a high-risk genotype by DM2 than that by DM1.)

DM0				DM1				DM2			
	BB	Bb	bb		DD	Dd	dd		FF	Ff	ff
AA	0	0	0	CC	0	0	0	EE	0	0	0
Aa	0	1	0	Cc	0	0	0	Ee	0	1	0
aa	0	0	1	cc	0	0	1	ee	0	0	0

would make across the study subjects are counted to assess their “similarity”. In particular, given two disease models, it is assumed that one of them is the true model, and the other one is used to make the prediction. Then the predictions of disease status for all individuals under the proxy model are obtained. The concordance between the proxy prediction and the true disease status, e.g., the ϕ coefficient [63], is a natural measure of “similarity” between two disease models.

In real-world situations, case-control samples are commonly used. Therefore, it is of interest to derive the measure of disease model distance based on the specific sample to improve the accuracy of using proxy disease models. For example, a disease model that predicts the disease outcome similarly to another one for the population data may not be so for a case-control sample. For instance, for a rare disease, the disease to non-disease ratio is quite small in the population data, whereas it is 1 under a balanced case-control design. To help increase the final epistasis detection, it is desirable to have the proxy model predicting as many individuals the same as possible for the sample instead of for the population.

In the following sections, the approach of prototype disease models are described in more detail. First, a metric is derived to measure the similarity (or equivalently, difference) between two disease models. Then, all disease models are clustered into a few groups and a prototype model is selected from each group. Finally, each pair of SNPs are screened against the set of prototype models. The set of prototype models is decided *a priori*, without considering the disease status of individuals in the data set. This is what makes the proposed approach less greedy, and less data-adaptive, than existing methods such as MDR and RS.

2.4 Similarity measure

2.4.1 Introduction

In the introduction chapter, some intuition has been given that some disease models appear to be more similar than others (E.g., in Table 1.1, model (a) and (c) appear to be quite different from each other, whereas model (b) and (d) are somewhat similar to each other). Such intuition can be formalized in many different ways; for instance, some researchers have used a geometric approach to categorize them [27]. For this study, a more pragmatic approach is taken, as discussed in the previous section.

Suppose the similarity of two disease models denoted by M and M' needs to be measured. It is assessed according to how much they agree in terms of their assignment of individuals into high- and low-risk groups.

Assume there is a group of $n_{\cdot\cdot}$ individuals. According to their genotypes for the SNP-pair under study, M and M' would classify them either as high or low risk.

For $k, \ell = \{0, 1\}$, let $n_{k\ell}$ denote the number of individuals classified to be high-risk by both models ($k = \ell = 1$), by M only ($k = 0, \ell = 1$), by M' only ($k = 1, \ell = 0$), or by neither model ($k = \ell = 0$) (Table 2.7).

Then the Φ-coefficient [63], defined as

$$\Phi = \frac{(n_{11})(n_{00}) - (n_{10})(n_{01})}{\sqrt{(n_{1\cdot})(n_{0\cdot})(n_{\cdot 1})(n_{\cdot 0})}}, \tag{2.7}$$

is a natural quantity to measure the concordance between M and M'. The Φ-coefficient is equal to the Pearson's Correlation Coefficient estimated for two binary variables. As can be seen from the form, high (low) values of n_{11} and n_{00}, and low (high) values of n_{10} and n_{01} lead to high values of Φ, which means that the two models classify many (few) individuals to be in the same high- or low-risk group.

Table 2.7. Assignment of individuals into high- and low-risk groups by two disease models, M and M'.

$M'\backslash M$	High Risk	Low Risk	Total
High Risk	n_{11}	n_{10}	$n_{1\cdot}$
Low Risk	n_{01}	n_{00}	$n_{0\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	$n_{\cdot\cdot}$

2.4.2 Φ-coefficient

For $i = 1, 2, ..., 9$, let G_i denote a genotype combination formed by a pair of SNPs; and let $\mathbb{P}(D|G_i)$ denote the penetrance (or probability of trait/disease) of the particular combination G_i. Suppose that M is the true disease model with penetrance levels denoted below.

$$\mathbb{P}(D|G_i) = \begin{cases} P_1, & M(G_i) = 1, \\ P_0, & M(G_i) = 0; \end{cases} \tag{2.8}$$

whereas M' is a different disease model used to approximate the true model M.

Then, it could be shown that the Φ-coefficient between M and M' can be expressed as

$$\Phi(M', M) = \frac{(W_{11})(W_{00}) - (W_{10})(W_{01})}{\sqrt{\left(\frac{U}{V}W_{11} + W_{01}\right)\left(W_{10} + \frac{V}{U}W_{00}\right)(W_{1\cdot})(W_{0\cdot})}}, \tag{2.9}$$

where

$$\begin{aligned} W_{k\ell} &= \sum_{\substack{M(G_i)=k \\ M'(G_i)=\ell}} \mathbb{P}(G_i) \quad \text{for} \quad k, \ell \in \{0, 1\}; && (2.10) \\ U &= rP_1[1 - \mathbb{P}(D)] + (1 - P_1)\mathbb{P}(D); && (2.11) \\ V &= rP_0[1 - \mathbb{P}(D)] + (1 - P_0)\mathbb{P}(D); && (2.12) \end{aligned}$$

r is the case-control ratio of the sample, and $\mathbb{P}(D)$ is the prevalence of the trait/disease, i.e., $\mathbb{P}(D) = P_1(W_{11} + W_{10}) + P_0(W_{01} + W_{00})$.

Eq. 2.9 shows that

$$d(M', M) = 1 - \Phi(M', M) \tag{2.13}$$

can be used as a distance metric for two disease models.

Proof. The derivation of Eq. (2.9) are divided into a few steps as shown below.

Step 1

For a pair of SNPs assumed to be truly associated with the disease, suppose that M is the true, while M' is another (e.g., prototype), disease model.

Under the true model M, a randomly selected individual can be stratified into 4 different groups with probabilities given by the 2×2 table below.

	$M(G_i) = 1$	$M(G_i) = 0$	All
D	$P_1 W_{1\cdot}$	$P_0 W_{0\cdot}$	$\mathbb{P}(D)$
D^c	$(1 - P_1) W_{1\cdot}$	$(1 - P_0) W_{0\cdot}$	$1 - \mathbb{P}(D)$
All	$W_{1\cdot}$	$W_{0\cdot}$	1

Notice that $W_{1\cdot} = W_{11} + W_{10}$, $W_{0\cdot} = W_{01} + W_{00}$, and

$$\begin{aligned}
\mathbb{P}(D) &= \sum_i \mathbb{P}(D|G_i)\mathbb{P}(G_i) \\
&= \sum_{M(G_i)=1} P_1 \times \mathbb{P}(G_i) + \sum_{M(G_i)=0} P_0 \times \mathbb{P}(G_i) \\
&= P_1 W_{1\cdot} + P_0 W_{0\cdot}.
\end{aligned}$$

However, under a different disease model, M', individuals would be stratified differently—in particular, according to the following table.

	$M'(G_i) = 1$	$M'(G_i) = 0$	All
D	$P_1W_{11} + P_0W_{01}$	$P_1W_{10} + P_0W_{00}$	$\mathbb{P}(D)$
D^c	$(1 - P_1)W_{11} + (1 - P_0)W_{01}$	$(1 - P_1)W_{10} + (1 - P_0)W_{00}$	$1 - \mathbb{P}(D)$
All	$W_{\cdot 1}$	$W_{\cdot 0}$	1

This is because

$$\begin{aligned}
& \mathbb{P}(D \text{ and } M' = 1) \\
= \; & \mathbb{P}(D|M' = 1) \times \mathbb{P}(M' = 1) \\
= \; & \left[\underbrace{\mathbb{P}(D|M' = 1, M = 1)}_{P_1} \mathbb{P}(M = 1) + \underbrace{\mathbb{P}(D|M' = 1, M = 0)}_{P_0} \mathbb{P}(M = 0) \right] \times \mathbb{P}(M' = 1) \\
= \; & P_1 \times \mathbb{P}(M = 1 \text{ and } M' = 1) + P_0 \times \mathbb{P}(M = 0 \text{ and } M' = 1) \\
= \; & P_1W_{11} + P_0W_{01},
\end{aligned}$$

where, for notational convenience, the disease models have been simply written as $M' = 1$ or $M = 1$ rather than $M'(G_i) = 1$ or $M(G_i) = 1$. The other three cells in the table above can be derived in a similar fashion.

Step 2

For a case-control study, the row margins are fixed. Thus for every r case-units and 1 control-unit, respectively, the probabilities in each corresponding row of the previous tables are rescaled, as shown below.

	$M(G_i) = 1$	$M(G_i) = 0$	All
Cases	$\dfrac{r(P_1W_{11} + P_1W_{10})}{\mathbb{P}(D)}$	$\dfrac{r(P_0W_{01} + P_0W_{00})}{\mathbb{P}(D)}$	r
Controls	$\dfrac{(1 - P_1)W_{11} + (1 - P_1)W_{10}}{1 - \mathbb{P}(D)}$	$\dfrac{(1 - P_0)W_{01} + (1 - P_0)W_{00}}{1 - \mathbb{P}(D)}$	1

	$M'(G_i) = 1$	$M'(G_i) = 0$	All
Cases	$\frac{r(P_1W_{11} + P_0W_{01})}{\mathbb{P}(D)}$	$\frac{r(P_1W_{10} + P_0W_{00})}{\mathbb{P}(D)}$	r
Controls	$\frac{(1-P_1)W_{11} + (1-P_0)W_{01}}{1-\mathbb{P}(D)}$	$\frac{(1-P_1)W_{10} + (1-P_0)W_{00}}{1-\mathbb{P}(D)}$	1

Step 3

The tables in Step 2 can now be rearranged according to how the two disease models, M and M', have stratified the $r+1$ case-control units, by summing over W_{11}, W_{10}, W_{01}, and W_{00}, respectively. This gives

	$M(G_i) = 1$	$M(G_i) = 0$
$M'(G_i) = 1$	$\frac{rP_1W_{11}}{\mathbb{P}(D)} + \frac{(1-P_1)W_{11}}{1-\mathbb{P}(D)}$	$\frac{rP_0W_{01}}{\mathbb{P}(D)} + \frac{(1-P_0)W_{01}}{1-\mathbb{P}(D)}$
$M'(G_i) = 0$	$\frac{rP_1W_{10}}{\mathbb{P}(D)} + \frac{(1-P_1)W_{10}}{1-\mathbb{P}(D)}$	$\frac{rP_0W_{00}}{\mathbb{P}(D)} + \frac{(1-P_0)W_{00}}{1-\mathbb{P}(D)}$

Note that each cell contains two terms corresponding to the cases and controls respectively, e.g., there are expected to be $\frac{rP_1W_{11}}{\mathbb{P}(D)}$ units of cases and $\frac{(1-P_1)W_{11}}{1-\mathbb{P}(D)}$ units of controls for which both M and M' would predict as high risk.

The above table can be algebraically simplified to

	$M(G_i) = 1$	$M(G_i) = 0$
$M'(G_i) = 1$	$\frac{UW_{11}}{\mathbb{P}(D)[1-\mathbb{P}(D)]}$	$\frac{VW_{01}}{\mathbb{P}(D)[1-\mathbb{P}(D)]}$
$M'(G_i) = 0$	$\frac{UW_{10}}{\mathbb{P}(D)[1-\mathbb{P}(D)]}$	$\frac{VW_{00}}{\mathbb{P}(D)[1-\mathbb{P}(D)]}$

where

$$U = rP_1 + \mathbb{P}(D) - (r+1)P_1\mathbb{P}(D) = rP_1[1-\mathbb{P}(D)] + (1-P_1)\mathbb{P}(D),$$

and

$$V = rP_0 + \mathbb{P}(D) - (r+1)P_0\mathbb{P}(D) = rP_0[1-\mathbb{P}(D)] + (1-P_0)\mathbb{P}(D).$$

Calculating the Φ-coefficient of the final 2×2 contingency table above gives Equation 2.9, i.e.,

$$\begin{aligned}
\Phi = & \frac{\frac{UW_{11}}{p(D)(1-p(D))}\frac{VW_{00}}{p(D)(1-p(D))} - \frac{UW_{10}}{p(D)(1-p(D))}\frac{VW_{01}}{p(D)(1-p(D))}}{\sqrt{\frac{U(W_{11}+W_{10})}{p(D)(1-p(D))}\frac{V(W_{00}+W_{01})}{p(D)(1-p(D))}\frac{UW_{11}+VW_{01}}{p(D)(1-p(D))}\frac{UW_{10}+VW_{00}}{p(D)(1-p(D))}}} \\
= & \frac{UVW_{11}W_{00} - UVW_{10}W_{01}}{\sqrt{UVW_{1\cdot}W_{0\cdot}(UW_{11}+VW_{01})(UW_{10}+VW_{00})}} \\
= & \frac{W_{11}W_{00} - W_{10}W_{01}}{\sqrt{(\frac{U}{V}W_{11}+W_{01})(W_{10}+\frac{V}{U}W_{11})W_{1\cdot}W_{0\cdot}}}
\end{aligned}$$

□

2.5 Parameters

Some details are given below about how values of various parameters can be obtained in order to compute the expression on the right-hand side of Eq. (2.9).

$W_{k\ell}$**:** Assuming Hardy-Weinberg equilibrium (Box 1.1), the MAFs of the two SNPs can be estimated from the control sample as commonly done in GWAS, and used to determine $\mathbb{P}(G_i)$ for each genotype combination G_i and hence $W_{k\ell}$ as well for $k, \ell \in \{0, 1\}$.

r**:** For any given data set, the case-control ratio r is known, e.g., $r = 1$ for a balanced case-control data set.

$\mathbb{P}(D)$**:** The prevalence, $\mathbb{P}(D)$, of a particular trait/disease can often be obtained from external sources, e.g., published studies and/or expert opinions. (More on this below in Sections 2.6 and 2.7.5.)

P_1, P_0**:** To determine the value of these parameters, a convenient assumption is made that the underlying pair of SNPs is the actual pair associated with the outcome. Then, the prevalence is simply

$$\mathbb{P}(D) = \sum_{i=1}^{9} \mathbb{P}(D|G_i)\mathbb{P}(G_i) \tag{2.14}$$

and the heritability (the amount of genetic contribution to overall phenotype variation [64]) is given by

$$h^2 = \frac{1}{[\mathbb{P}(D)][1 - \mathbb{P}(D)]} \sum_{i=1}^{9} [\mathbb{P}(D|G_i) - \mathbb{P}(D)]^2 \mathbb{P}(G_i). \tag{2.15}$$

Since it has been assumed in Eq. (2.8) that M only has two unique penetrance levels, i.e., each $\mathbb{P}(D|G_i)$ is either P_1 and P_0; they can be uniquely determined from the two equations, (2.14) and (2.15), provided that information about the heritability parameter, h^2, is available . This can often be obtained from external sources as well, the same as the prevalence parameter (More on this below in Sections 2.6 and 2.7.5.). The functional forms are given in Eqs. (2.16).

$$\begin{cases} P_0 = p(D) - h\sqrt{p(D)(1-p(D))}\sqrt{\dfrac{\sum\limits_{M(G_i)=1} \mathbb{P}(G_i)}{\sum\limits_{M(G_i)=0} \mathbb{P}(G_i)}} \\ P_1 = p(D) + h\sqrt{p(D)(1-p(D))}\sqrt{\dfrac{\sum\limits_{M(G_i)=0} \mathbb{P}(G_i)}{\sum\limits_{M(G_i)=1} \mathbb{P}(G_i)}} \end{cases} \tag{2.16}$$

2.6 Clustering

2.6.1 Preparation

There are altogether $2^9 - 2 = 510$ non-trivial TTTC disease models (the trivial ones are those such that $M(G_i) = 1$ or $M(G_i) = 0$ for all G_i). For clustering purposes, there is no need to consider disease models that are symmetric with respect to (i) the exchange of locus, i.e., swapping the two SNPs, or (ii) the exchange of disease status, i.e., flipping the binary values of each $M(G_i)$ from a zero to a one, and vice versa.

There are $2^3 \times 2^3 = 64$ disease models that are invariant under the exchange of locus, an example of which is given in (a) of Table 2.8. This can be seen by considering that there are 2^3 possible combinations of high- and low-risk genotypes along the diagonal cells (e.g., genotypes of AABB, AaBb, aabb), and 2^3 possible ones that are symmetric about the diagonal; therefore, there are $2^3 \times 2^3 = 64$ ones in total. For the rest $512 - 64 = 448$ models, half of them are the same with respect to the exchange of locus, examples of which are models (b) and (c) in Table 2.8. This means that there are $448/2 = 224$ disease models that do not need to be considered for clustering purposes. Excluding these ones, $510 - 224 = 286$ models remain. Among them, half are symmetric under the exchange of disease status. After excluding those, there are 143 disease models to be considered for clustering. The set of models that remain, which are denoted as $\mathcal{M}$, is listed in Appendix A.

Table 2.8. Examples of TTTC disease models that are symmetric about the diagonal and invariant under the exchange of locus.

(a)				(b)				(c)			
	AA	Aa	aa		AA	Aa	aa		AA	Aa	aa
BB	0	0	0	BB	0	1	0	BB	0	0	1
Bb	0	0	1	Bb	0	0	1	Bb	1	0	0
bb	0	1	0	bb	1	0	0	bb	0	1	0

It is worth emphasizing that the reduction of disease models to the set $\mathcal{M}$, due to the symmetry considerations mentioned above, is *only* applicable to the clustering and prototype selection stage. When screening each candidate SNP-pair, prototype disease models that are asymmetric with respect to the exchange of locus, such as M_1^* in Fig. 2.4, are always tested both for {A/a,B/b} and for {B/b,A/a}.

Prototype disease models that are asymmetric with respect to the affection status may or may not need to be tested in both ways, depending on the screening procedure. For instance, for marginal screening (Section 2.1), it is not necessary to test a SNP-pair twice by the prototype disease model itself and its symmetric counterparty because the effect size is the same under the two models. However, for the sequential screening procedure, it is necessary because the direction of effect needs to be considered for each SNP-pair. When testing the SNP-pairs, the reduction has been accounted for to ensure that all models are properly represented in that stage.

2.6.2 Algorithm

In principle, any distance-based clustering algorithm could be used. For this project, the "global K-means" algorithm [65] is adopted. It is an incremental deterministic algorithm that employs k-means as a local search procedure, and is proposed to deal with the initialization problem with the original k-means algorithm. It is directly defined on the distances, so it could be conveniently applied for use here.

Notice that the disease model distances are asymmetric, i.e., in the derivation of the distances, one has been assumed to be the true model and the other one to be the substitute model. Therefore, whenever a distance needs to be assessed, the directional distance of a disease model to the potential prototype one is calculated. In this way, the prototype disease models being selected would be the ones to which all the other models within the same clusters are closest.

The global k-means algorithm starts with one cluster, finds the optimal cluster center, and then runs the k-means algorithm to find the second optimal cluster center while keeping the first one fixed. The detailed steps of the algorithm are given in Table 2.9.

2.6.3 Results

Multidimensional scaling (MDS) is used to transform the distances back to data points with coordinates for an illustration of the disease model clusters. Given pairwise distances, the MDS technique could produce variables whose pairwise distances are as close as the original distance matrix.

While the directional distance metric is used for prototype identification (see Table 2.9 below), a symmetrized distance metric, $d_s(M_i, M_j) \equiv [d(M_i, M_j) + d(M_j, M_i)]/2$, is used

Table 2.9. The global K-means algorithm for identifying prototype disease models.

1. Let $\mathcal{M}$ be the set of all disease models and $\mathcal{M}^*$, the prototype set (initially empty).
2. Evaluate each $M_i \in \mathcal{M}\backslash\mathcal{M}^*$ as a potential new prototype, as follows:

   a. For each $M_k \in \mathcal{M}\backslash\{\mathcal{M}^* \cup M_i\}$, calculate the distances $d(M_i, M_k)$, and $d(M_j^*, M_k)$ for all $M_j^* \in \mathcal{M}^*$ if $\mathcal{M}^*$ is not empty.

   b. Assign M_k either to an existing cluster—e.g., C_j^*, with center M_j^*— or to a potentially new cluster—say C_i, with center M_i—depending on which of $d(M_i, M_k)$ and $d(M_j^*, M_k)$ is the shortest.

   c. After all $M_k \in \mathcal{M}\backslash\{\mathcal{M}^* \cup M_i\}$ are assigned, calculate the total within-cluster distances,

   $$D(M_i) \equiv \sum_{M_k \in C_i} d(M_i, M_k) + \sum_{M_j^* \in \mathcal{M}^*} \sum_{M_k \in C_j^*} d(M_j^*, M_k),$$

   as a result of using M_i as an additional cluster center.

3. Identify a new prototype model as the one that minimizes the total within-cluster distances, i.e.,

   $$M^* \quad = \quad \arg\min_{M_i \in \mathcal{M}\backslash\mathcal{M}^*} \quad D(M_i),$$

   and insert it into the set $\mathcal{M}^* \leftarrow \mathcal{M}^* \cup M^*$.

4. Repeat steps 2-3 until a certain number of prototypes are identified.

for performing MDS so that the resulting 2-dimensional coordinate-map (Fig. 2.3) is more meaningful.

Fig. 2.3 shows the 2-dimensional coordinates of all models $\in \mathcal{M}$ as estimated by the MDS from their pairwise distances, assuming that the MAFs of both SNPs are equal to 0.1, 0.2, 0.3, and 0.4, respectively, while fixing the prevalence and heritability parameters at $\mathbb{P}(D) = h^2 = 0.02$. It is clear from Fig. 2.3 that these disease models form several clusters.

Note that the multivariate technique MDS merely finds the best 2D map that preserves the pairwise distances as much as possible, but the resulting 2D map still does not fully capture the pairwise distances. In fact, for the presented ones, they usually account for only about $50\% - 70\%$ of the total variance in the pairwise distances. Therefore, models that are actually closer to each other may appear farther apart in this 2D map, and vice versa, i.e., there is definitely some loss of information here in the figure.

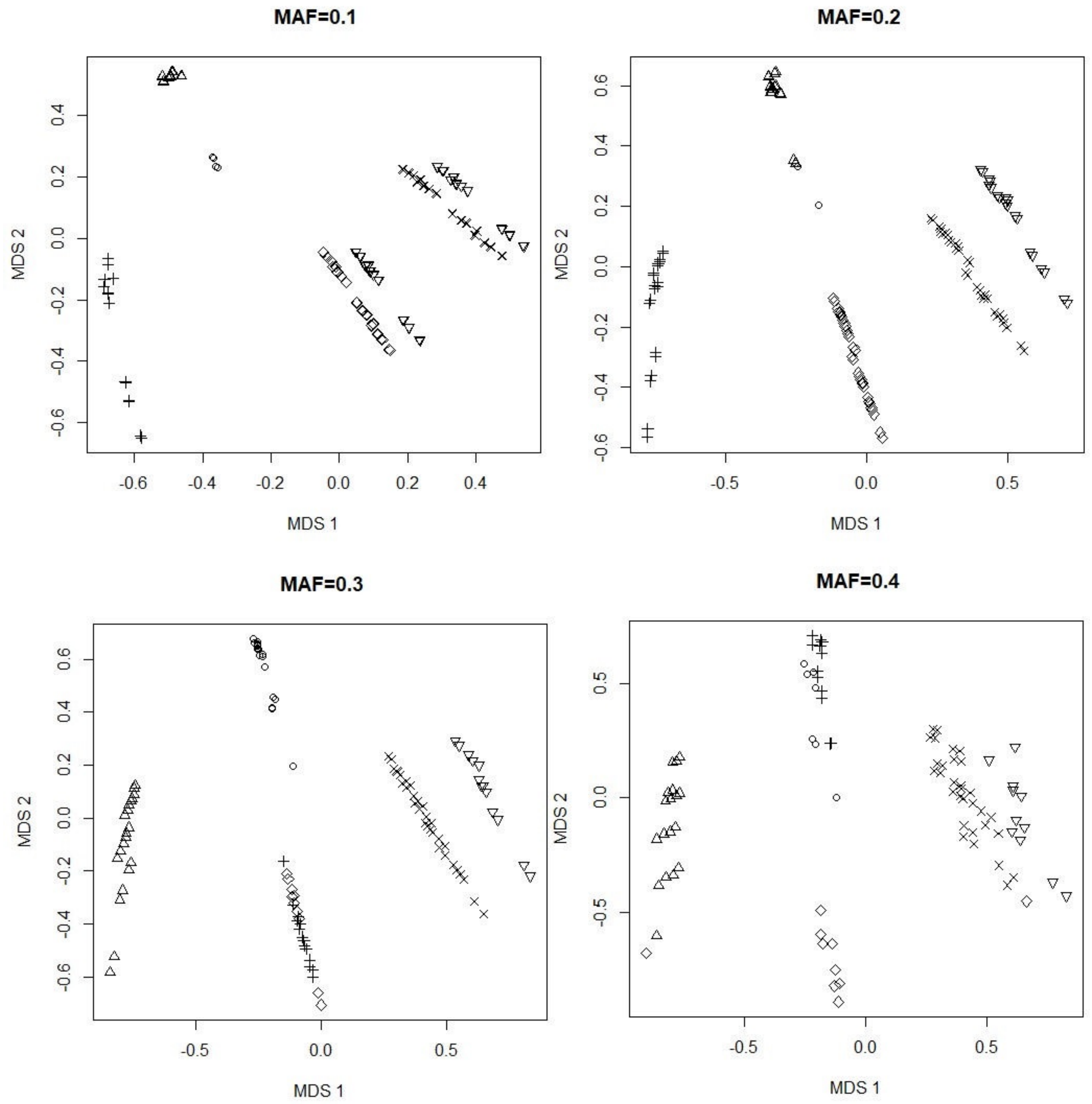
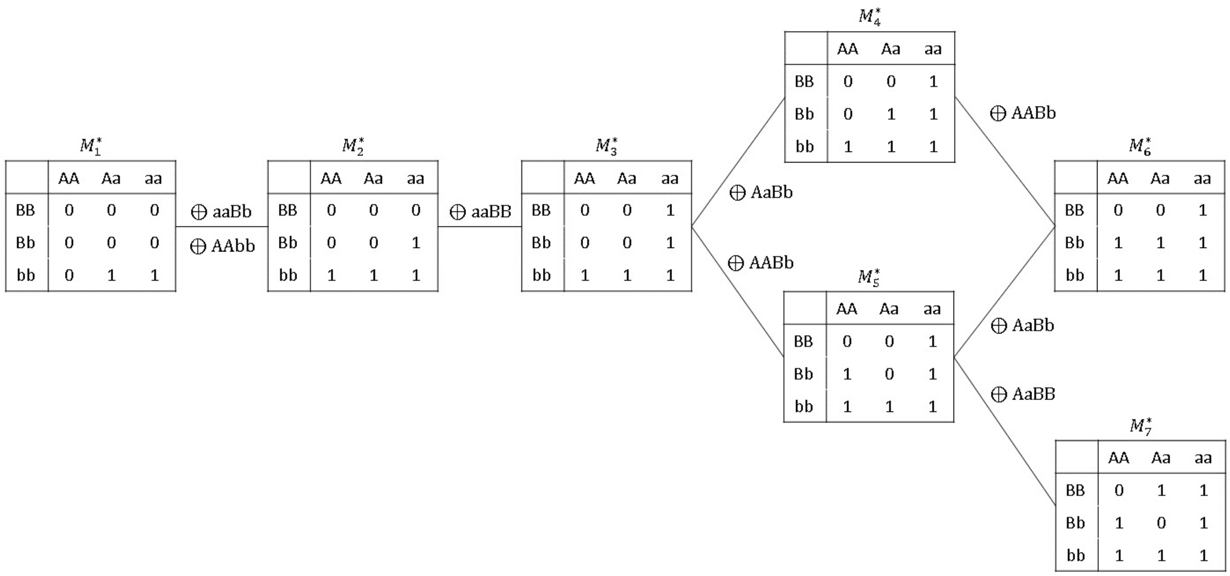


Fig 2.3. A two-dimensional map of disease models in $\mathcal{M}$. The coordinates are estimated by applying the multi-dimensional scaling (MDS) technique to the symmetrized pairwise distances, $d_s(M_i, M_j) \equiv [d(M_i, M_j) + d(M_j, M_i)]/2$, for all $i \neq j$. Models clustered into the same group are depicted by the same symbol (e.g., '+', '∘', '×'). These two-dimensional coordinates explain about 50-70% of the variation in $d_s(\cdot, \cdot)$, so there is some loss of information—in particular, some disease models may be closer to (or farther apart from) each other than how they appear in this map.

From Eq. (2.9) it can be expected that the distance metric will be affected by the MAFs of the underlying SNPs, but Fig. 2.3 shows that the resulting clusters do not change significantly. Therefore, it is not necessary to repeat the prototype selection step for every individual SNP-pair. Instead, the MAF-scale are simply discretized into 6 bins, {0.05, 0.1, 0.2, 0.3, 0.4, 0.45}, and 36 different *sets* of prototypes are created for all 6 × 6 pairwise combinations. For example, when screening a SNP-pair (i, j) with $(\mathrm{MAF}_i, \mathrm{MAF}_j) = (0.068, 0.182)$, the set of prototypes for $(\mathrm{MAF}_i, \mathrm{MAF}_j) = (0.05, 0.2)$ could be used, and so on.

Based on Fig. 2.3, 7 prototypes are selected for each MAF combination. As an illustration, the prototypes for SNP-pairs with $(\mathrm{MAF}_i, \mathrm{MAF}_j) = (0.2, 0.2)$ are displayed in Fig. 2.4 with manual annotations to reveal their relationships with one another. This figure may be interpreted in the follwoing way. For a pair of SNPs that have MAFs around 0.2, these are the primary epistatic effects to consider, and their structural relationships; any other will likely be very similar to one of these—in terms of how they would classify individuals into high- versus low-risk groups. This is also a unique piece of insight from the overall methodology that is not otherwise available from MDR or RS.

Similar plots produced with different values of $\mathbb{P}(D)$ and h^2 have also been examined. While these parameters also affected the distance metric, they did not produce any substantial changes to the clustering. Intuitively, this is because there has to be a fairly drastic warping of the relative distances between objects in order to alter their grouping. This point will be discussed more later in Section 2.7.5. Hence, for this project $\mathbb{P}(D) = h^2 = 0.02$ are simply used.



M_1^*

	AA	Aa	aa
BB	0	0	0
Bb	0	0	0
bb	0	1	1

M_2^*

	AA	Aa	aa
BB	0	0	0
Bb	0	0	1
bb	1	1	1

M_3^*

	AA	Aa	aa
BB	0	0	1
Bb	0	0	1
bb	1	1	1

M_4^*

	AA	Aa	aa
BB	0	0	1
Bb	0	1	1
bb	1	1	1

M_5^*

	AA	Aa	aa
BB	0	0	1
Bb	1	0	1
bb	1	1	1

M_6^*

	AA	Aa	aa
BB	0	0	1
Bb	1	1	1
bb	1	1	1

M_7^*

	AA	Aa	aa
BB	0	1	1
Bb	1	0	1
bb	1	1	1

Fig 2.4. The set of prototype disease models selected by the global K-means algorithm ($K = 7$) for SNP-pairs (i, j) with $(\mathbf{MAF}_i, \mathbf{MAF}_j) \approx (0.2, 0.2)$. The structural relationships between the seven prototypes are manually annotated; the clustering algorithm itself is not capable of making this type of discoveries.

2.7 Simulation

To motivate the proposed method, a few simulated examples in Section 2.2 have already been presented, where the concentration is on evidence that the proposed approach appears to overcome various weaknesses of existing approaches. In this section, the proposed approach is assessed more generally with a number of simulated examples that are commonly examined in the literature.

Note the simulation set-ups, including performance measures and disease model examples, apply commonly throughout this thesis unless otherwise stated.

2.7.1 Set-up and Performance Measure

In each simulation, 100 SNPs are generated, but only the first two determined the simulated outcome according to a particular disease model (more details below in Section 2.7.2).

For each pair of SNPs, the PTY determines a disease model for it by testing out its corresponding "prototype" disease model set and selects the best-fitting one. More specifically, the MAFs of the two SNPs are estimated based on the control sample, and used to map to the MAF bins described in the previous section to obtain the pair's prototypes. After the best-fitting disease model is determined, it is refitted to the SNP-pair by calculating the Pearson's χ^2-statistic and a p-value against the $\chi^2_{(1)}$-distribution for further evaluation (Note that the steps after determining the best-fitting disease model for a SNP-pair is the same for PTY as that for MDR and RS.).

To evaluate the performance of a method, a metric known as the F-measure is used, which is defined as

$$\text{F-measure} \equiv \frac{2 \times (\text{precision}) \times (\text{recall})}{(\text{precision}) + (\text{recall})}, \tag{2.17}$$

where

$$\text{precision} = \begin{cases} \dfrac{1}{\#(\text{pairs detected})}, & \text{if the true pair was detected,} \\ 0, & \text{otherwise;} \end{cases} \tag{2.18}$$

and

$$\text{recall} = \begin{cases} 1, & \text{if the true pair was detected,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.19}$$

Each simulation is repeated for 400 times, and the average F-measure and its standard error are recorded. To avoid excessive computation, the marginal screening procedure for all methods are used; see Section 2.1.

Note that the size of the simulation data set (100 SNPs) is quite different from that in the real data application (around 400,000 SNPs) in terms of the number of SNPs to screen. This is mainly due to computational consideration. It needs to be emphasized here that, while most GWA studies have focused on screening individual SNPs, this study is focusing on pairwise screening rather than individual screening, so 400,000 SNPs would mean "400,000-choose-2 =79,999,800,000 pairs in the real data application. For simulation studies, each simulation is repeated 400 times, so the computational cost is 400 times the "usual amount". That's why a relatively small settings have been used.

Remark

The following are some thoughts that helped the decision making on the choice of the F-measure for evalution.

The F-measure is a widely used criterion in the field of information retrieval, and it is a single numeric metric that balances the trade-off between true positives and false positives. It is adopted over other metrics such as the “balanced accuracy” because the underlying problem is more of an “information retrieval” problem than a “classification” one. To see this, notice there are far more true negatives than true positives, and that detecting the positives (the relevant SNP-pair) is a much more important objective than correctly calling out the negatives.

Take the problem of conducting a Google search as an example. For each query, the majority of the returned results are irrelevant. From a user’s perspective, how many truly irrelevant web pages have been correctly left out of the search results are not important, that is, the user does not care about the true negatives. Therefore, the most important measure should concentrate on the set of detected web pages retrieved by the search engine (true positive). Moreover, because the set of truly irrelevant pages is usually quite large, the true negative rate will also be difficult to distinguish meaningfully for most search engines, e.g., any “reasonable” search engine will have a true negative rate of >99%. Therefore, measures like the “balanced accuracy” actually place an undue amount of emphasis on this rather inconsequential side of the performance. This is also why the information retrieval community tends to largely favour metrics such as the F-measure to those more commonly used by the classification community such as “balanced accuracy”.

The situation of detecting relevant SNP-pairs is very much like performing a Google search in that (i) most pairs are not signals; (ii) it is not very important to get the true negatives right; (iii) it is much more important to care about how many of the detected pairs are true positives or false positives.

2.7.2 Similation Examples

For the prototype disease model approach, the primary focus is to evaluate the ability of different methods to detect different epistatic effects as represented by different disease models.

First, six disease models with main effects (Table 2.10) are included. They were among the most commonly used examples in various studies [24, 66–70]. Here in Table 2.10, these models are parameterized in terms of odds, $\mathbb{P}(D|G_i)/[1-\mathbb{P}(D|G_i)]$, rather than penetrance, $\mathbb{P}(D|G_i)$. The parameters α and θ were determined by simultaneously solving Eq. (2.14) and Eq. (2.15), given the prevalence $\mathbb{P}(D)$ and heritability h^2 of the disease, as well as the MAF of each SNP. The prevalence value is simply fixed as $\mathbb{P}(D) = 0.02$, the heritability values are as shown in Table 2.10. Each of these simulations with MAF=0.1 and 0.4 are repeated for all SNPs. Assuming Hardy-Weinberg equilibrium, the MAF determined $\mathbb{P}(G_i)$ for each genotype combination G_i, leaving α and θ to be the only unknowns in Eq. (2.14) and Eq. (2.15) so that they could be uniquely determined.

Table 2.10. Simulated examples for Section 2.7. Disease models with main effects. The parameters α and θ are uniquely determined given prevalence $\mathbb{P}(D)$, heritability h^2, and MAF. Prevalence is fixed, i.e., $\mathbb{P}(D) = 0.02$, and each simulation is repeated with MAF=0.1 and 0.4 for all SNPs.

(a) Threhold (T) $h^2 = 0.02$				(b) Dominant-Dominant (DD) $h^2 = 0.02$			
	BB	Bb	bb		BB	Bb	bb
AA	α	α	α	AA	α	α	α
Aa	α	α	$\alpha(1+\theta)$	Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$	aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
(c) Modifying Effect (MOD) $h^2 = 0.02$				(d) Exclusive Or (XOR) $h^2 = 0.02$			
	BB	Bb	bb		BB	Bb	bb
AA	α	α	α	AA	α	α	$\alpha(1+\theta)$
Aa	α	α	$\alpha(1+\theta)$	Aa	α	α	$\alpha(1+\theta)$
aa	$\alpha(1+\theta)$	$\alpha(1+\theta)$	$\alpha(1+\theta)$	aa	$\alpha(1+\theta)$	$\alpha(1+\theta)$	α
(e) Multiplicative (ME) $h^2 = 0.015$				(f) Threshold Multiplicative (MET) $h^2 = 0.015$			
	BB	Bb	bb		BB	Bb	bb
AA	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$	AA	α	α	α
Aa	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^3$	Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
aa	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^3$	$\alpha(1+\theta)^4$	aa	α	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^4$

Next, four disease models without main effects (Table 2.11) are included, taken from an earlier study conducted by Ritchie *et al.* [52], in which these disease models were created to have purely epistatic effects in the sense that no marginal effect existed for either SNP involved.

The disease models, T, DD, MOD and XOR, all have two penetrance levels (Table 2.10), and so do the prototype disease models (see Fig. 2.4). However, the simulations were carefully designed to ensure that, while some of these models (e.g., XOR) were relatively close to a prototype, others (e.g., DD) were relatively far from all prototypes, as measured by the metric Φ.

The disease models, ME, MET, and DMN 1-4 all have more than two penetrance levels (Tables 2.10 and 2.11). They have been chosen so that a wider variety of epistatic effects could be studied.

Table 2.11. Simulated examples for Section 2.7. Disease models without main effects, taken from [52], where they were specifically constructed in such a way that there is no individual association between either SNP and the disease.

(a) DMN 1 MAF=0.25, $h^2 = 0.016$				(b) DMN 2 MAF=0.25, $h^2 = 0.04$			
	BB	Bb	bb		BB	Bb	bb
AA	0.08	0.07	0.05	AA	0	0.1	0.09
Aa	0.1	0	0.1	Aa	0.04	0.01	0.08
aa	0.03	0.1	0.04	aa	0.07	0.09	0.03
(c) DMN 3 MAF=0.1, $h^2 = 0.002$				(d) DMN 4 MAF=0.1, $h^2 = 0.015$			
	BB	Bb	bb		BB	Bb	bb
AA	0.07	0.05	0.02	AA	0.09	0.001	0.02
Aa	0.05	0.09	0.01	Aa	0.08	0.07	0.005
aa	0.02	0.01	0.03	aa	0.003	0.007	0.02

Remark

When selecting disease model examples for the simulation study, both two- and multiple-level ones are included to ensure that a wide variety of epistasis mechanisms have been covered. Note that all of these methods in this chapter, i.e., MDR, RS and PTY, merely rely on the TTTC-type models to capture/describe different epistatic effects. However, it does not mean that they can only be applied to analyze data generated by this type. In other words, the usefulness of them are not limited, which is important considering that in real world, any epistasis mechanism could be possible. In the original MDR paper [6], it was shown that approximating more complicated disease models by using the TTTC-type models could actually improve detection because of reduced parameterization.

Moreover, also note that the 0/1 penetrance levels among the 9 cells are merely an indication of the type of SNP-SNP interaction at work, i.e., relatively high and low risk of disease. It does not mean that all of those individuals having the "1" in the genotype will get the disease, or those having the "0" genotype will not get the disease. Therefore, in almost every simulated example throughout, the penetrance level is never complete-penetrance.

2.7.3 Significance Thresholds

To assess the methods, two different thresholds are applied. The nominal association measures produced by different methods for each pair of SNPs (see Section 2.1) are thresholded by their corresponding (nominal) p-values,

$$\widehat{p}_{i,j} \equiv \mathbb{P}(\chi^2_{(1)} > \widehat{\chi}^2_{i,j}), \tag{2.20}$$

and a pair was considered "detected" if $\widehat{p}_{i,j} < \alpha$, where α was a significance threshold. For convenience, simple Bonferroni corrections are applied to determine the threshold α. As there are a total of $\binom{100}{2} = 4,950$ pairs of SNPs, it is natural to consider a threshold of α^{easy} in Eq. (2.21) first.

$$\alpha^{\text{easy}} = 0.05 \div 4,950 \approx 10^{-5}. \tag{2.21}$$

To account for the fact that these nominal association measures are inflated (see Section 2.2.3), a more stringent threshold is also considered to be applied. Notice that for a pair of SNPs, each method has tested a different number of correlated disease models to determine the final one, therefore, there is no straightforward way to pick one that applies fairly for all methods. As a result, a convenient choice of α^{hard} defined in Eq. (2.22) is simply chosen based on the fact that RS would always consider 8 different disease models.

$$\alpha^{\text{hard}} = 0.05 \div 4,950 \div 8 \approx 1.26 \times 10^{-6}, \tag{2.22}$$

Correcting significance thresholds for simultaneous tests of correlated hypotheses is an intricate inferential problem for which there is no good solution yet. It is not clear whether α^{hard} is really the "correct" threshold for RS but, as a rough guideline, one may think that this choice would favour RS slightly. The empirical results below do support this interpretation to some extent.

When assessing the p-value adjustment method in Chapter 3 and results from adjusted p-values in Chapter 4, though, the more "stringent" criteria is not applied because the purpose there is to find a more "appropriate" evaluation criteria to rank SNP-pairs, and so only α^{easy} is used.

2.7.4 Results

The results for all representative models are given in Table 2.12. A relatively large sample size of $n = 600$ is used when the MAF is relatively low (e.g., 0.1), and a relatively small sample size of $n = 300$ is used when the MAF is high (e.g., 0.25, 0.4). This is because, when the MAF is relatively high (low), the underlying signals become stronger (weaker) and easier (harder) to detect. Hence, all the methods would perform quite well (badly), which makes it difficult to differentiate the performance of different methods. For the simulated cases with 100 SNPs, it is found that all methods essentially become indistinguishable when the sample size reached as low as $n = 100$ or as high as $n = 1000$.

As explained previously, the threshold, α^{easy}, only includes a simple Bonferroni correction and does not account for the fact that all of the methods, i.e., MDR, RS or PTY, have usually tested a few disease models already before testing the significance of the SNP-pair against the outcome. Strictly speaking, therefore, the Bonferroni correction alone is not enough, and often leads to inflated false positive rates. Among the three methods, PTY is the least prone to false positives, which explains why its performance is the best under α^{easy}. Generally speaking, the results confirm some practical value to consider a less greedy and less data-adaptive procedure such as the proposed PTY for epistasis detection.

Table 2.12. Results from simulation study (Section 2.7). Average F-measures (and their standard errors) over 400 replications. A star (*) in front of the number indicates the best performer for that simulation.

			$\alpha^{\text{easy}} = 1.00 \times 10^{-5}$						$\alpha^{\text{hard}} = 1.26 \times 10^{-6}$					
n	MAF	Model	MDR		RS		PTY		MDR		RS		PTY	
600	0.1	T	0.012	(0.005)	0.080	(0.013)	*0.083	(0.015)	0.000	(0.000)	*0.046	(0.012)	0.030	(0.010)
		MOD	0.062	(0.009)	0.067	(0.007)	*0.075	(0.010)	0.063	(0.011)	*0.109	(0.013)	0.090	(0.013)
		DD	0.183	(0.016)	0.172	(0.015)	*0.278	(0.019)	0.341	(0.023)	0.372	(0.021)	*0.449	(0.024)
		XOR	0.199	(0.014)	0.198	(0.013)	*0.328	(0.020)	0.283	(0.022)	0.449	(0.022)	*0.531	(0.024)
		ME	0.011	(0.000)	0.011	(0.000)	*0.012	(0.000)	0.013	(0.000)	0.013	(0.001)	*0.015	(0.001)
		MET	0.234	(0.019)	0.245	(0.016)	*0.294	(0.021)	0.335	(0.025)	*0.414	(0.023)	0.350	(0.024)
300	0.4	T	0.167	(0.011)	0.152	(0.010)	*0.223	(0.014)	*0.357	(0.017)	0.346	(0.017)	0.317	(0.018)
		MOD	0.076	(0.007)	0.064	(0.006)	*0.110	(0.010)	0.163	(0.013)	0.154	(0.013)	*0.202	(0.015)
		DD	0.015	(0.001)	0.014	(0.001)	*0.135	(0.009)	0.035	(0.005)	0.032	(0.005)	*0.276	(0.014)
		XOR	0.195	(0.012)	0.164	(0.011)	*0.306	(0.016)	0.441	(0.018)	0.409	(0.018)	*0.561	(0.019)
		ME	0.022	(0.002)	0.020	(0.002)	*0.033	(0.002)	0.043	(0.004)	0.039	(0.003)	*0.086	(0.006)
		MET	0.073	(0.006)	0.074	(0.007)	*0.076	(0.008)	0.132	(0.011)	*0.141	(0.011)	0.103	(0.010)
600	0.25	DMN 1	0.729	(0.015)	0.686	(0.016)	*0.935	(0.009)	0.951	(0.008)	0.939	(0.009)	*0.992	(0.003)
	0.25	DMN 2	0.743	(0.015)	0.705	(0.016)	*0.938	(0.010)	0.959	(0.007)	0.944	(0.008)	*0.972	(0.009)
	0.1	DMN 3	0.700	(0.018)	0.675	(0.018)	*0.832	(0.019)	0.822	(0.021)	*0.852	(0.019)	0.722	(0.026)
	0.1	DMN 4	0.752	(0.015)	0.720	(0.015)	*0.897	(0.014)	0.912	(0.013)	*0.921	(0.012)	0.831	(0.021)

2.7.5 Discussion

Disease Model

Throughout the simulation study, the performance of each screening method has been assessed by its ability to detect the underlying SNP-pair, but not by whether the true disease model is correctly identified. The detection of the relevant SNP-pair is undoubtedly the more fundamental task. Once the relevant SNP-pairs are identified, further studies can be conducted to determine the actual underlying mechanism. Such an approach is certainly not unusual in the context of genome-wide association (GWA) studies. For most GWA studies in the literature, single SNPs are often tested and reported using disease models—e.g., additive, dominant, and so on—that are not necessarily the correct ones. Ascertaining the true disease model is almost never the goal of the initial GWA study; detecting the affected SNPs is.

In fact, this is also why the proposed PTY method works, because one need not always use exactly the true disease model in order to detect a pair of affected SNPs. While using a "very wrong" disease model can negatively affect the chances of detecting an affected SNP-pair, one has a good chance of making the detection as long as the disease models used for screening is "close enough" to the true one. Due to the way the prototype models are selected—i.e., as representative models from each cluster, there is a very good chance that at least one of the models is "close enough" to the true one.

High-order Epistasis

Earlier in the introduction section, it has been stated that screening for higher-order interactions at a genome-wide level is still largely impractical at the present time. Nonetheless, when it becomes possible for doing so, the idea of using prototype disease models will be even more attractive. This is because when higher-order interactions are considered, the number of disease models will grow in combinatorial scale, which leads to an increased tendency for greedy approaches to produce false positives.

Computational Time

In terms of computational time difference for the MDR, RS and the PTY methods, they are comparably on the same scale for the application. As explained in the introduction section, in order for the comparisons to be fair, all three methods (MDR, RS, and PTY) primarily differed in how $M_{i,j}$ is determined; afterward, the same chi-squared test (with the same multiple-testing corrected threshold) was used to determine whether each given SNP-pair (i,j) is significant or not. Therefore, the computational time of the three methods only differ in the step which determines what disease model $M_{i,j}$ to use for (i,j). As such, MDR takes the least amount of time; RS always tests 8 (data-driven) disease models, so its run time is about 8 times the run time of MDR; whereas the run time of PTY is comparable to that of RS, as it requires tests anywhere between 7-14 (prototype) disease models, depending on the specific pair.

Parameters for Similarity Metric

The value of the similarity/distance metric depends on MAF, prevalence, and heritability. So in principle, this should be done "for each analysis", but this is not what is done in practice in this thesis.

The main message of Figure 2.3 is to illustrate that, while the numeric value of the similarity/distance metric itself does change with MAF, the resulting clustering and hence prototype selection do not change significantly. In the paragraphs below, the discussion is devoted to explain that, while the numeric value of the similarity/distance metric can change with prevalence and heritability, the relative distances between disease models are not drastically warped, and the resulting prototype selection is relatively robust to these changes (refer to Figure 2.5).

For each phenotype, the prevalence and heritability parameters are presumably fixed. They can be obtained/estimated from other sources. If not, a vague estimate/guess can be quite safely used, as these parameters do not affect prototype selection too much.

As for the MAF, because the clustering results and hence prototype selection do not change significantly (again, Figure 2.3), it is also not necessary to repeat prototype selection for every SNP-pair. As explained previously, the MAF values are discretized into 6 bins. For each of the $6 \times 6 = 36$ combination bins, the cluster analysis is performed once prior to screening to determine a set of prototypes. At the time of screening, each SNP-pair is first determined (a trivial task) to fall into one of these 36 bins, and the set of prototype models associated with that bin are then used to screen it.

Prototype disease models can be selected in many different ways, although using different sets of prototypes is not expected to make a substantial difference. The specific proposal outlined in this chapter is based on using a particular metric, $\Phi(M', M)$, to quantify the similarity of disease models. This following explores more about the intuitive appeal of this metric, as promised earlier in Section 2.4.1.

Let $r_0 = \mathbb{P}(D)/[1 - \mathbb{P}(D)]$ denote the population-wide case-control ratio. Then, the ratio U/V appearing in the denominator of Eq. (2.9) is simply

$$\frac{U}{V} = \frac{(P_1)r + (1 - P_1)r_0}{(P_0)r + (1 - P_0)r_0} = \frac{r_0 + (r - r_0)P_1}{r_0 + (r - r_0)P_0}. \tag{2.23}$$

This makes it clear that, if $r = r_0$, then $U/V = 1$. In this case, it is easy to see that the denominator of the Φ-coefficient can be interpreted as $\sqrt{\mathbb{V}\text{ar}(M')\mathbb{V}\text{ar}(M)}$. This is because M can be viewed as a Bernoulli random variable mapping various genotype combinations to either 0 or 1, with $\mathbb{P}(M = 1) = W_{1\cdot}$ and $\mathbb{P}(M = 0) = W_{0\cdot}$, so $\mathbb{V}\text{ar}(M) = W_{1\cdot}W_{0\cdot}$. Likewise,

$$\begin{aligned}\mathbb{V}\text{ar}(M') &= W_{\cdot 1}W_{\cdot 0} \\ &= (W_{11} + W_{01})(W_{10} + W_{00}) \\ &= \underbrace{W_{11}W_{10}}_{M=1} + \underbrace{W_{01}W_{10}}_{M \neq M'} + \underbrace{W_{11}W_{00}}_{M=M'} + \underbrace{W_{01}W_{00}}_{M=0}.\end{aligned} \tag{2.24}$$

$\mathbb{V}\text{ar}(M')$ can be decomposed into four terms, as shown above in Eq. (2.24), where each successive term can be seen to measure the variability in M' when $M = 1$, when M and M'

completely disagree, when they completely agree, and when $M = 0$, respectively.

However, for a case-control sample, it is often the case that $r \gg r_0$, in which case Eq. (2.23) implies that $U/V \approx P_1/P_0 > 1$. It can be seen that, in this case, Eq. (2.9) implicitly shows to calculate $\mathbb{V}\mathrm{ar}(M')$, the variance of the potential prototype model M' used to approximate/represent M, differently:

$$\mathbb{V}\mathrm{ar}(M') = \frac{U}{V} W_{11} W_{10} + W_{01} W_{10} + W_{11} W_{00} + \frac{V}{U} W_{01} W_{00}. \tag{2.25}$$

In particular, among genotypes considered to be risky by M (the set for which $M = 1$), the variability in M' should be up-weighted, which reduces their similarity; whereas, among those considered to be non-risky by M (the set for which $M = 0$), the variability in M' should be down-weighted, which increases their similarity. In other words, when considering M' as a potential prototype for representing M, the metric $\Phi(M', M)$ "thinks" it is more important for M' to agree with M on their assignments of the risky genotypes than for them to agree on the non-risky ones. This is intuitively appealing; a concrete numeric example is given in Appendix B.3.

The approximation that $U/V \approx P_1/P_0$ also allows one to see how the parameters, $\mathbb{P}(D)$ and h^2, affect the metric Φ. The solution to Eqs. (2.14)-(2.15) is:

$$P_1 = \mathbb{P}(D) + \sqrt{\frac{W_{0\cdot}}{W_{1\cdot}} \mathbb{P}(D)[1 - \mathbb{P}(D)]h^2}, \quad P_0 = \mathbb{P}(D) - \sqrt{\frac{W_{1\cdot}}{W_{0\cdot}} \mathbb{P}(D)[1 - \mathbb{P}(D)]h^2}. \tag{2.26}$$

Fig. 2.5 contains various views of the odds, P_1/P_0, as a function of the ratio $W_{1\cdot}/W_{0\cdot}$, prevalence $\mathbb{P}(D)$, and heritability h^2. For any given disease model M with a specific ratio $W_{1\cdot}/W_{0\cdot}$, the odds P_1/P_0 is certainly affected by the choices of $\mathbb{P}(D)$ and h^2; but these parameters also affect the odds of other disease models with different $W_{1\cdot}/W_{0\cdot}$-ratios in a similar manner. For example, for fixed h^2, a large (and potentially wrong) choice of $\mathbb{P}(D)$ lowers the odds; whereas for fixed $\mathbb{P}(D)$, a large (and potentially wrong) choice of h^2 elevates it for all disease models. As a result, even though the distances do change between different disease models and their candidate prototypes, the relative distances are not drastically warped. That's why it is observed that the resulting prototypes are fairly robust to different choices of $\mathbb{P}(D)$ and h^2.

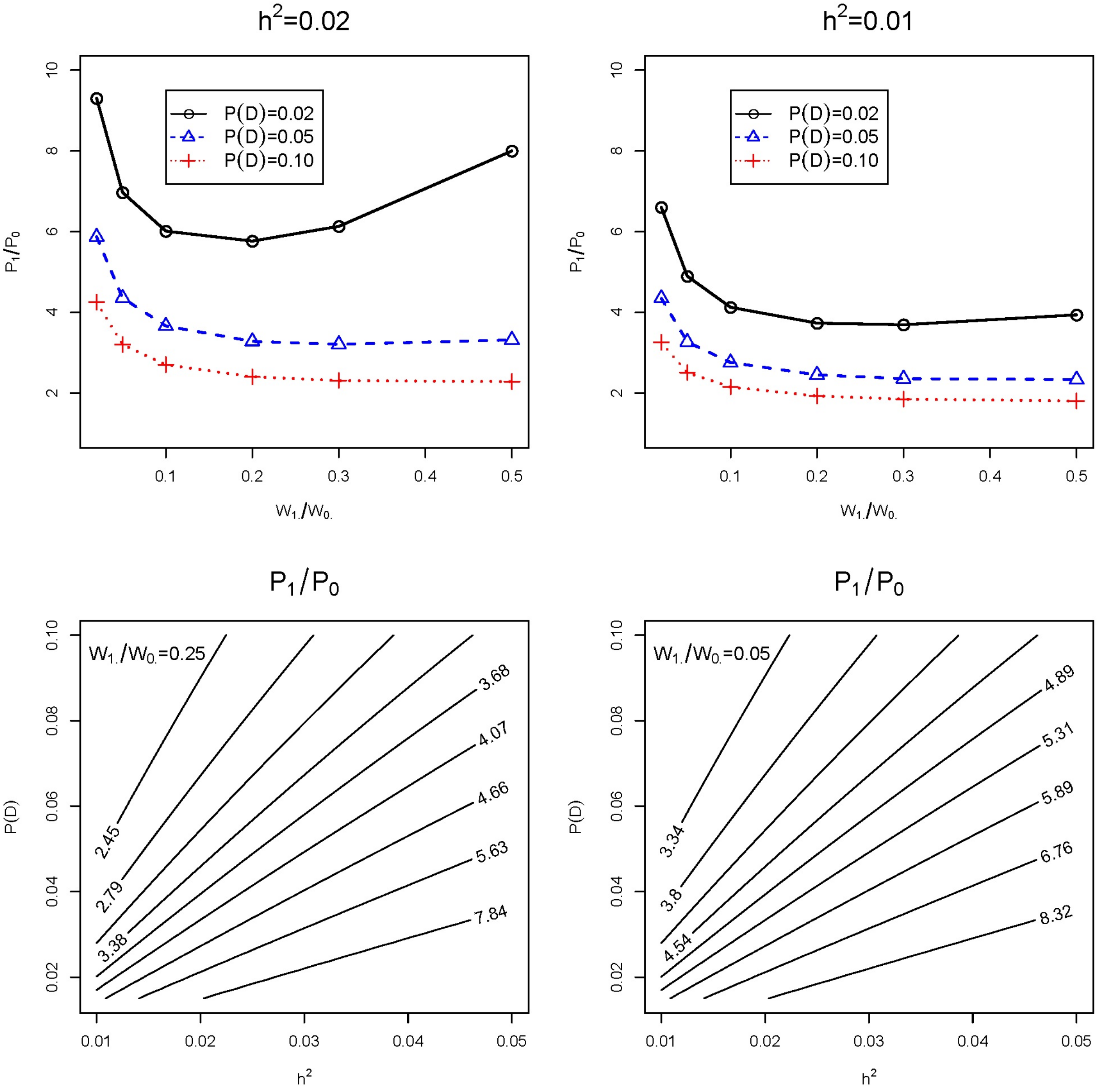


Fig 2.5. Different views of the odds, P_1/P_0, as a function of the ratio $W_{1\cdot}/W_{0\cdot}$, prevalence $\mathbb{P}(D)$, and heritability h^2, where P_1, P_0 are solutions to Eqs. (2.14)-(2.15). While the parameters $\mathbb{P}(D)$ and h^2 do affect the odds P_1/P_0 and hence the metric $\Phi(M', M)$, their impact is similar at different values of $W_{1\cdot}/W_{0\cdot}$ and hence similar for different M.

2.8 Real Data Application

This section reports analysis of the phase I bipolar disorder data from the Wellcome Trust Case Control Consortium (WTCCC) [71]. For the prototype disease model method, because the proposed method aims at screening SNP-pairs for different epistatic effects (rather than individual SNPs for main effects), the complementary value that the proposed method offers is the focus— particularly its ability to find relevant SNPs that other methods may still miss.

The WTCCC project involves genotyping of 500K SNPs on humans of British ancestry. Bipolar disorder is one of seven diseases being studied by the WTCCC, and the shared control samples consist of $1,500$ individuals from the 1958 British Birth Cohort and another $1,500$ individuals selected from blood donors recruited as part of their project.

Identical-twin studies have shown that bipolar disorder has a strong genetic component [72]. Current findings from genome-wise association studies (GWAS) demonstrate that bipolar disorder shares many genetic overlaps with schizophrenia and other major depressive disorders. It is also characteristic of being polygenic, i.e., many variants that coalesce into functional pathways contribute to the disorder with small effects. The current understanding of its neurobiology is that changes in inflammatory cytokines, corticosteroids, neurotrophins, mitochondrial energy generation, oxidative stress, and neurogenesis are all involved in a comprehensive way to explain its various clinical features [73].

2.8.1 Pre-processing

The same data quality control procedures are applied as described in [71]—excluding SNPs with $> 5\%$ missing observations ($> 1\%$ for SNPs with MAF < 0.05), Hardy-Weinberg exact p-value $< 5.7 \times 10^{-7}$, p-value $< 5.7 \times 10^{-7}$ for either a one- or two degree-of-freedom test of association between the two control groups, and genome-wide heterozygosity $< 23\%$ or $> 30\%$, as well as samples with $> 3\%$ missing across all SNPs. In addition, the following SNPs are also filtered out: MAF $< 1\%$, or p-value $< 10^{-7}$ in a univariate test of association, or p-value $< 10^{-5}$ from the test of Hardy-Weinberg equilibrium. The remaining data contained $1,868$

cases (individuals with bipolar disorder), $2,938$ controls, and $405,524$ SNPs. Eliminating "easily detectable" SNPs with "obvious" main effects is not uncommon for studies that focus on the detection of SNP-SNP interactions—for example, the paper by Wan *et al.* [7] that proposed the RS method also did this.

2.8.2 Mapping SNPs to genes

The marginal screening procedure (see Section 2.1) is used to screen and rank all pairs of SNPs. Here, the 100 unique SNPs appearing in the top 85 pairs (nominal p-value $< 10^{-11}$) is focused on. The "Ensembl gene annotation system" [74] as well as SNPnexus [75] are adopted to map these SNPs to the genes in which they most likely reside. Altogether, 75 genes are identified in this manner.

Fig. 2.6 shows the number of SNPs appearing in the top 85 pairs identified by PTY, MDR and RS, respectively. While 15 SNPs were identified by all three methods, 42 were identified by the proposed method alone and they were mapped to 18 unique genes. Five of them—specifically, UNC13A [76], RGS6 [77], DPP10 [78], FGF14 [79] and TLE4 [79]—had been associated with bipolar disorder or related suicide attempts. Moreover, the SNP mapped to FGF14 had a p-value of 0.03 on a univariate test of association, indicating that it would have had no chance of being detected in a genome-wide screening of individual SNPs. Here, it has been detected as a result of pairwise screening that focused on epistatic effects.

Fig. 2.7 shows the largest interaction network based on the 85 genes that are identified. As mentioned above, each of these "detected" SNPs is mapped to a gene in which it resides (or to a nearby gene) using the "Ensemble gene annotation system" and "SNPnexus", and is represented as a node. If a SNP does not map to a recognized gene, it is retained as a node.

In presenting the network, oval is used to refer to the node as a gene and rectangle as the SNP itself (The size of the node is irrelevant, i.e., it is determined by the amount of text inside (e.g., length of the name), rather than anything scientific.). A link is placed between two nodes if the SNPs underlying the nodes are from the same pair detected. So, for example,

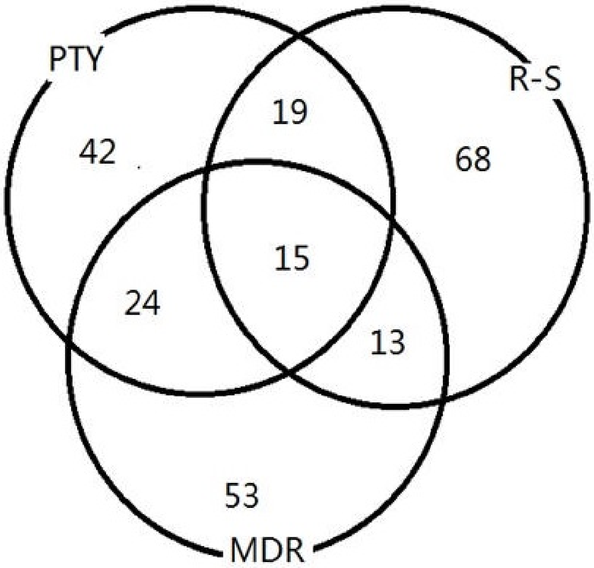


Fig 2.6. Analysis of bipolar disorder data. Venn diagram of unique SNPs appearing in the top 85 pairs detected by PTY, RS, and MDR, respectively. SNPs detected multiple times (e.g., occurring in multiple pairs) were counted only once.

the link between AQP1 and FAH means a pair of SNPs (one of which is mapped to AQP1 and another to FAH) is among the top 85 pairs detected.

As is often the case, these networks contain many disjoint components. The one presented here is the biggest component with the most number of genes/SNPs connected. Therefore, the hub gene, AQP1, appears to have the most connections with other genes/SNPs. AQP1 encodes a small integral membrane protein that functions as a water channel protein and is potentially involved in a human neurological disorder called "central pontine myelinolysis" [80]. The specific SNP that is mapped to this gene (rs4299909) has a p-value of 0.0002 based on a univariate test of association; hence, it would have had no chance of being detected by marginal screening of individual SNPs, either. Here again, it has been detected as a result of pairwise screening that focused on epistatic effects. For this reason, the underlying SNP mapped to AQP1 (rs4299909) is also presented/displayed inside the oval to emphasize this point.

Among other genes in this network, ST6GALNAC5 is known to catalyze the transfer of sialic acid to cell surface proteins, and sialic acid has been suggested as an essential nutrient for brain development and cognition [81]. RGS6 regulates G protein signalling and may modulate neuronal activities; in previous studies, SNPs in this gene have been reported to be associated with schizophrenia [82]. MAN2A1 encodes a glycosyl hydrolase (a common enzyme) and catalyzes the final hydrolytic step in the N-glycan maturation pathway; many SNPs in this gene have been reported to be associated with various phenotypes and diseases, including Alzheimer's disease [83, 84]. TLE4 inhibits the transcriptional activation mediated by PAX5, CTNNB1, and TCF family members in Wnt signalling, which has been suggested to be potentially involved in the pathophysiology of bipolar disorder [85]. FAH encodes the last enzyme in the tyrosine metabolism pathway; the amino acid, tyrosine, is a precursor to neurotransmitters and increases plasma neurotransmitter levels—particularly dopamine and norepinephrine, both important neurotransmitters in the brain [86]. FUT8 encodes an enzyme belonging to the family of fucosyltransferases; a variant in this gene has been reported to influence glutamate concentrations in brains of patients with multiple sclerosis [87]—glutamate is a neurotransmitter accounting in total for well over 90% of the synaptic connections in the human brain.

Out of the 75 genes being identified, the following have also been reported by various independent studies to be associated with bipolar disorder, or suicides related to bipolar disorder: ANK3 [88], CNTNAP2 [89], PTPRN2 [90], DSCAM [76], PSD3 [76], RAPGEF4 [91], CPN1 [92], EPHB2 [79], CAP2 [79], NAV2 [79], and ABCB1 [79].

2.8.3 Gene set enrichment analysis

To further validate the findings, gene set enrichment analysis (GSEA) [93] is also performed on the aforementioned set of 75 genes. GSEA identifies classes of genes (e.g., those involved in specific pathways) that are over-represented in a given gene set (e.g., the ones that haven been discovered by PTY) and may have an association with disease phenotypes by comparing the candidate set against background databases. Gene Ontology [94] is one such database, which annotates and classifies genes in terms of their associated biological processes, cellular components and molecular functions. Other popular databases include KEGG [95] and

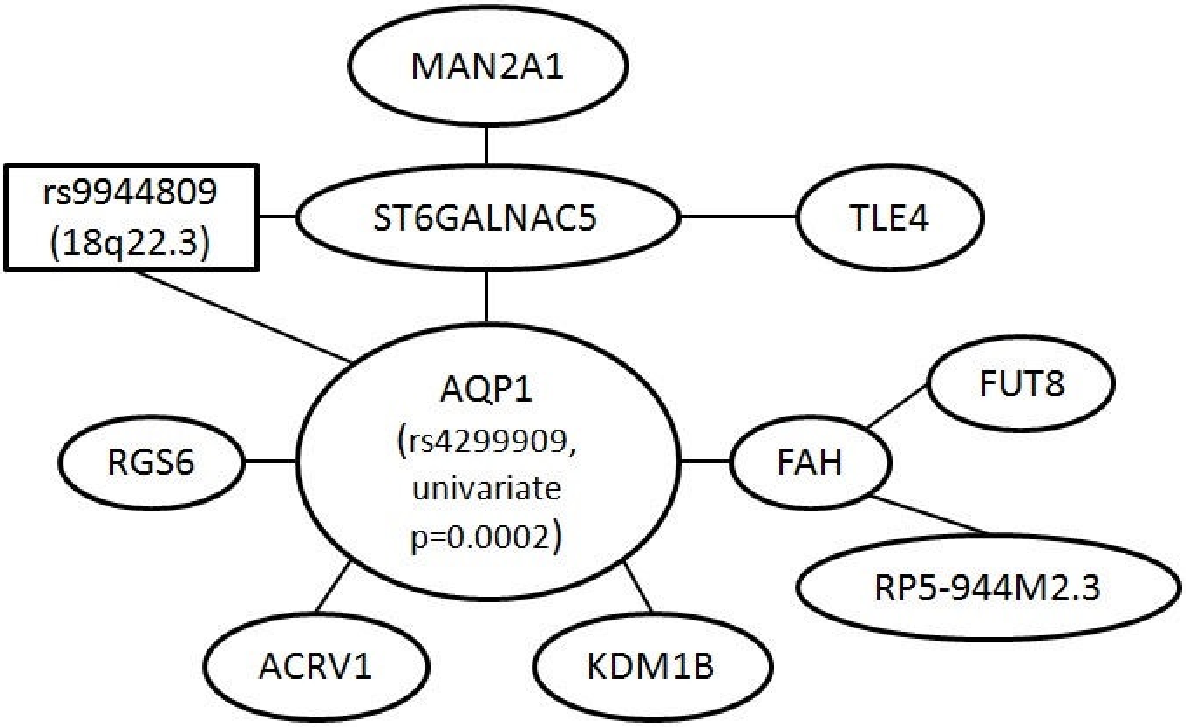


Fig 2.7. Analysis of bipolar disorder data. Largest interaction network formed by genes mapped from SNPs appearing in the top 85 pairs. Each node is either a gene (oval), or a SNP (rectangle) itself if it cannot be mapped to any gene. The size of the node is irrelevant—it is determined by the amount of text inside rather than anything scientific. A link between two nodes means the SNPs underlying the nodes are from the same pair detected, so, for example, a link between AQP1 and FAH means that a pair of SNPs—one of which was mapped to AQP1 and another, to FAH—was among the top 85 pairs detected. The resulting network contains many disjoint components. The one presented here is the biggest component.

Pathway Commons [96]. A tool called WebGestalt [97] is used to compare a candidate gene set to various background databases and determine whether certain gene groups (e.g., those occurring in known pathways) appear statistically more or less often than expected.

Table 2.13 lists the statistically enriched pathways from KEGG (multiple-testing adjusted p-value ≤ 0.05). Many of them have been associated with bipolar disorder or related diseases. For instance, the neurotransmitter dopamine, which is believed to have connections to bipolar disorder, is part of the tyrosine metabolism pathway (line 3). The N-Glycan biosynthesis pathway (line 4) has been reported to be significantly enriched by a study of bipolar disorder

Table 2.13. Analysis of bipolar disorder data. GSEA results from KEGG. O = number of genes in the discovered set; C = total number of genes in the given pathway.

				p-value	
Line	Name	O	C	Nominal	Adjusted
1	metabolic pathways	13	1130	$\ll 0.01$	$\ll 0.01$
2	thyroid cancer	2	29	< 0.01	0.01
3	tyrosine metabolism	2	41	< 0.01	0.01
4	N-glycan biosynthesis	2	49	< 0.01	0.01
5	arginine and proline metabolism	2	54	< 0.01	0.01
6	melanoma	2	71	0.01	0.02
7	ErbB signalling pathway	2	87	0.01	0.02
8	hepatitis C	2	134	0.02	0.03
9	lysosome	2	121	0.02	0.03
10	axon guidance	2	129	0.02	0.03
11	pathways in cancer	3	326	0.02	0.03
12	cell adhesion molecules	2	133	0.02	0.03
13	endocytosis	2	201	0.05	0.05
14	regulation of actin cytoskeleton	2	213	0.05	0.05

in Canadian and UK populations [98]. Both arginine and proline (line 5) have been related to schizophrenia [99]. The ErbB signalling pathway (line 7) regulates a diverse range of physiological responses, such as cell proliferation, migration, differentiation, apoptosis and motility; and insufficient ErbB signalling has been associated with the development of neurodegenerative diseases in humans [100]. The regulations of the lysosome pathway (line 9) and of the actin cytoskeleton pathway (line 14) were found in a transcriptome sequencing and GWA study to be statistically enriched in genes associated with schizophrenia [101].

For comparison, the corresponding results for MDR and RS are provided in Appendix B.1, and enriched pathways from Gene Ontology and Pathway Commons (for PTY identified genes only) are provided in Appendix B.2.

2.8.4 Discussion

Results Validation

Validation of screening results is always tricky for real data. For genetic findings, an ultimate validation can only be done in a real lab, although replication through an independent study using a different sample often provides a certain level of validation as well.

Since it is challenging to perform either type of replication or cross-validation, this study follows various other researchers and makes a best effort to produce some "validating evidence" to corroborate the findings by (i) mapping the findings (raw SNPs) to genes, (ii) conducting a literature search on these genes to see if any of them are known to have any association with bipolar disorder, and (iii) conducting an enrichment analysis of these genes to see if any of the enriched pathways are known to have any such association as well.

These types of analysis are inherent with some limitations. There are many more genetic databases and tools, such as the GTEx tool(s) and/or chromatin data that may be potentially utilized to verify the results better. However, this study does not pursue this direction because the tentative efforts into it imply that it is challenging for statisticians who are not knowledgeable enough about the mapping from SNPs to genes and additional analysis after that. As the computational screening results, including those from any GWA studies, are merely approaches to obtain a suggested list for further investigation, they are perhaps best replicated by independent studies before being taken seriously enough to warrant any further comprehensive functional analysis.

Multiple Testing

In terms of the p-value threshold, the Bonferroni type of correction is applied in the simulation and real data analysis. For a large quantity of variable selection, it might be interesting to consider FDR control. Compared to FDR control, Bonferroni correction is usually considered to be more stringent. However, note that in the simulation, the nominal Bonferroni correction corresponding to α^{easy} is actually too liberal rather than too stringent.

The main reason an even more stringent correction corresponding to α^{hard} is adopted is that multiple disease models are tested for each SNP-pair, though it is not clear what the "right" threshold should be and the α^{hard} is also not necessarily the right one, refer to Section 2.7.3. Therefore, the FDR is not considered because it is less stringent than α^{easy}, whereas the "right" correction should be more stringent (though what it should be theoretically is not exactly clear).

The real data example (bipolar data) is mostly an illustrative case study. There, a few top-ranked discoveries are merely investigated to seek some validation, e.g., by conducting GSEA analysis and looking for existing evidence in the literature that supports the enriched pathways. To limit the scope of such validation exercise (mostly, a literature search), only the top 85 pairs and the unique genes therein are looked at. This cutoff is somewhat arbitrary, and chosen out of convenience rather than anything else. If FDR is adopted, it will significantly enlarge the scope of this validation exercise. That is, a longer list to validate against the literature in this manner, which is not the actual goal for this project.

Chapter 3

First-order Adjustment for Inflated P-values

This chapter presents a method that corrects for the inflation in the nominal $\chi^2_{(1)}$-statistic introduced in Section 1.3. A brief introduction about the inflation is first given, with commonly available solutions. Then the general idea of the proposed correction method is introduced, followed by more details about the solution steps. Lastly, simulation study is performed to evaluate the effectiveness of the proposed method. This piece of work has been written as a book chapter and published by Springer [102].

3.1 Overview

3.1.1 Motivation

In Section 1.3 it has been introduced that the MDR and RS methods use the nominal $\chi^2_{(1)}$-statistic to calculate p-values for a SNP-pair. The MDR and the RS differ in how they declare certain genotype combinations to be risky, and hence in their selection of the disease model (DM). Both methods make their decisions by evaluating the case-control ratios of different genotype combinations. On a balanced case-control sample, the MDR method selects a DM by declaring genotype combinations with case-control ratios exceeding unity (i.e., > 1) to be risky. The RS method first sorts the case-control ratios in descending order, and evaluates 8 different DMs by sequentially considering the top x genotype combinations as being risky, for $x = 1, 2, ..., 8$; then, it chooses the one that best predicts the outcome (i.e., the phenotype).

The p-values are inflated because the statistics are derived from selected disease models based on observed data, as have been evidenced in Section 2.2.3. This section provides some convenient solutions to adjust the p-value calculation to mitigate the inflation for the MDR and RS methods.

3.1.2 Available Solution

Clearly, using the $\chi^2_{(1)}$-distribution is only a valid approximation without the pre-selection step, and a direct way to circumvent the inflation problem is to take the pre-selection of $M_{i,j}$ into account when computing the p-value in Eq. (1.2).

Analytically, it is not easy to derive the exact null distribution of the test statistic for different screening methods. Some theoretical results are available, but they are not easy to apply in practice. For example, Boulesteix's work [103] on "maximally selected chi-square statistics" can be applied to analytically characterize the null distribution of the test statistic taking the pre-screening step into account, where the pre-screening bears much similarity to the RS method. However, the calculation is combinatorial in nature, which makes it not computationally feasible on a genome-wide scale.

Another solution would be to simulate the null distribution and compute an empirical p-value. For example, for a given SNP-pair (i, j), replicates of SNP-pairs with the same minor allele frequencies can be simulated, together with arbitrary case-control labels by independent coin flips. Assume the simulated pairs are indexed by $s = 1, 2, ..., S$. Calculate the test statistic for the SNP-pair (i, j) and those for the simulated null pairs using Eq. (1.1) in Section 1.3, denote as $\widehat{\chi}^2_{i,j}$ and $\widehat{\chi}^2_s(s = 1, 2, ..., S)$ respectively. Then the empirical p-value can be calculated as shown in Eq. (3.1).

$$p_{sim}(i,j) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}\big(\widehat{\chi}_s^2 > \widehat{\chi}_{i,j}^2\big). \tag{3.1}$$

Note that the empirical p-value calculation above requirs a very large number of simulations (S) , especially when the value of $\widehat{\chi}_{i,j}^2$ is large. This is because tail probabilities such as p-values are generally hard to estimate by simulation. The chance of observing a rare event can be quite low over a set of simulations, or it may not be observed at all even over a very large number of simulations. For this reason, it is not feasible to run the simulation for each pair like this in a genome-wide scale. Therefore, the empirical p-value approach to correct for inflation is not preferable for real world application.

Other empirical approaches are also available, for example, permutation test and cross-validation. For each pair (i,j), the case-control labels can be repeatedly permuted. Then for each permutation $s = 1, 2, ..., S$, the resulting test statistic $\widehat{\chi}_s^2$ can be computed. After that, an empirical p-value using Eq. (3.1) can also be calculated. Alternatively, for each pair (i,j), one can only use part of the data to select $M_{i,j}$, and the remaining data to conduct the independence test in Eqs. (1.1)–(1.2). In this way, the nominal p-value in Eq. (1.2) would no longer be biased. However, this process has to be repeated multiple times and the resulting p-values have to be averaged to reduce variation caused by the random division of the data. Because of which, the computational requirement is also too high to be applied in large scale.

In summary, other than some differences in the details, these alternative approaches are all quite similar in spirit; they solve the inflation problem by expensive computation, which makes them impractical for genome-wide studies [51, 61].

3.2 Proposed Method

3.2.1 Method Description

As explained in Section 3.1.2, empirical p-values are challenging to compute for relatively large values of $\widehat{\chi}^2_{i,j}$. That's why parametric models are commonly used for these purposes, and why copula models are so popular in risk management [104]. Given a rare event $R > t_{big}$, where R is a random quantity and t_{big} is a large quantity, by modeling the distribution of R to be $f(r;\theta)$, the tail probability $\mathbb{P}(R > t_{\text{big}})$ can then be estimated by $\int_{t_{\text{big}}}^{\infty} f(r,\widehat{\theta})dr$, where $\widehat{\theta}$ is an estimate of the parameter θ from the simulations $R_1, R_2, ..., R_S$.

The proposed method to correct for inflation amounts to taking such a parametric approach, where f is taken to be a chi-squared distribution, and θ to be its degree of freedom. In more detail, the null distribution of the test statistic in Eq. 1.1 is still assumed to be chi-squared, except that it has an inflated degree of freedom. Usually, the estimation of θ is much easier than $\mathbb{P}(R > t_{\text{big}})$, but it is worth emphasizing that such a parametric approach depends on whether the model $f(r;\theta)$ has been "correctly" specified. The proposal in this section essentially argues that a choice of f is correct to the first order. In the simulation section (Section 3.3), the results are presented which appear to suppport this argument.

Since the chi-squared distribution has its degree of freedom equal to its mean, the effective degree of freedom (EDF) can be computed using the same type of simulation as described in the preceding section (Section 3.1.2), i.e., by

$$\text{EDF} = \frac{1}{S}\sum_{s=1}^{S}\widehat{\chi}^2_s. \tag{3.2}$$

Then an adjusted p-value is calculated as

$$p_{adj}(i,j) = \mathbb{P}\big(\chi^2_{(\text{EDF})} > \widehat{\chi}^2_{i,j}\big). \tag{3.3}$$

Although the form for the null distribution is not known exactly (Section 3.1.2), it is known for sure that the correct one is no longer $\chi^2_{(1)}$ (Section 2.2.3). The proposed adjustment ensures that the null distribution $\chi^2_{(\mathrm{EDF})}$ will at least have the correct first moment by estimating the EDF through simulation, even though it is still not exactly the right null distribution.

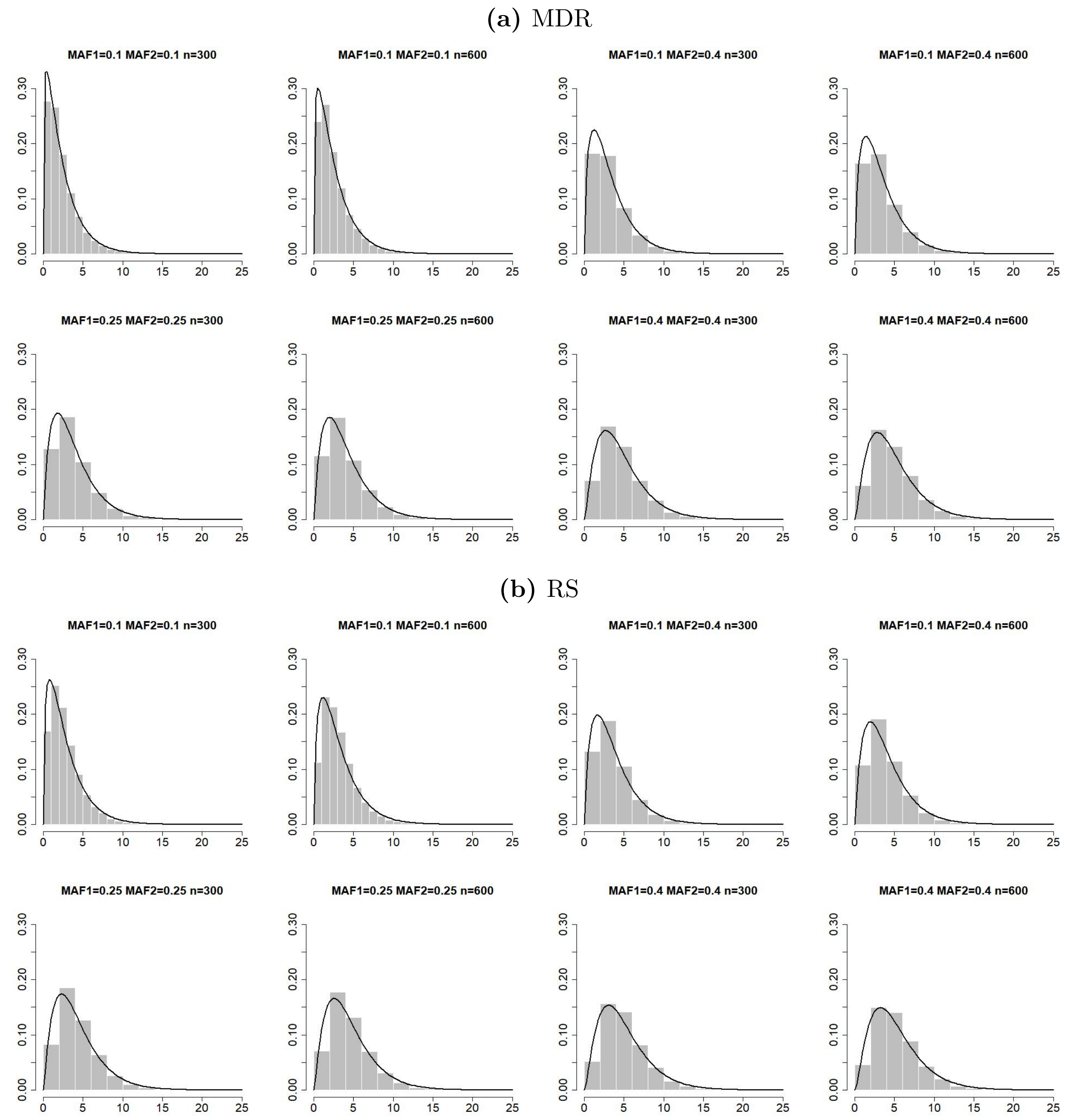


Fig 3.1. Screening simulated null pairs using the MDR or RS method. Histograms of $\{\widehat{\chi}^2_s : s = 1, 2, ..., S\}$ versus the $\chi^2_{(\text{EDF})}$ density functions, where EDF is computed by Eq. (3.2), for some specific combinations of $(\text{MAF}_1, \text{MAF}_2, n)$. While the $\chi^2_{(\text{EDF})}$ density functions are not perfect fits of the underlying histograms, they are reasonable approximations as first-order corrections.

It is in this sense that the proposed method can be regarded as a correction, or adjustment, of the *first order*. Such a first-order corrections can actually make a substantial difference in practice. For example, with an EDF = 3.10, a nominal test statistic of 21.32 would result in an adjusted p-value of $\mathbb{P}(\chi^2_{(3.10)} > 21.32) = 1.018 \times 10^{-4}$. If the first-order correction is not applied and the $\chi^2_{(1)}$-distribution is used, the same test statistic would give a p-value of $\mathbb{P}(\chi^2_{(1)} > 21.32) = 3.887 \times 10^{-6}$. The statistical evidence is inflated by two orders of magnitude from 10^{-4} to 10^{-6}.

Though the method is quite simple, empirically the approximation provided by the first-order correction is quite adequate (see Fig. 3.1). Notice that the main advantage of using Eq. (3.3) over Eq. (3.1), is that there is no need to run a separate simulation for each SNP-pair. More details on this are given in the following sections.

3.2.2 Response Surface Model

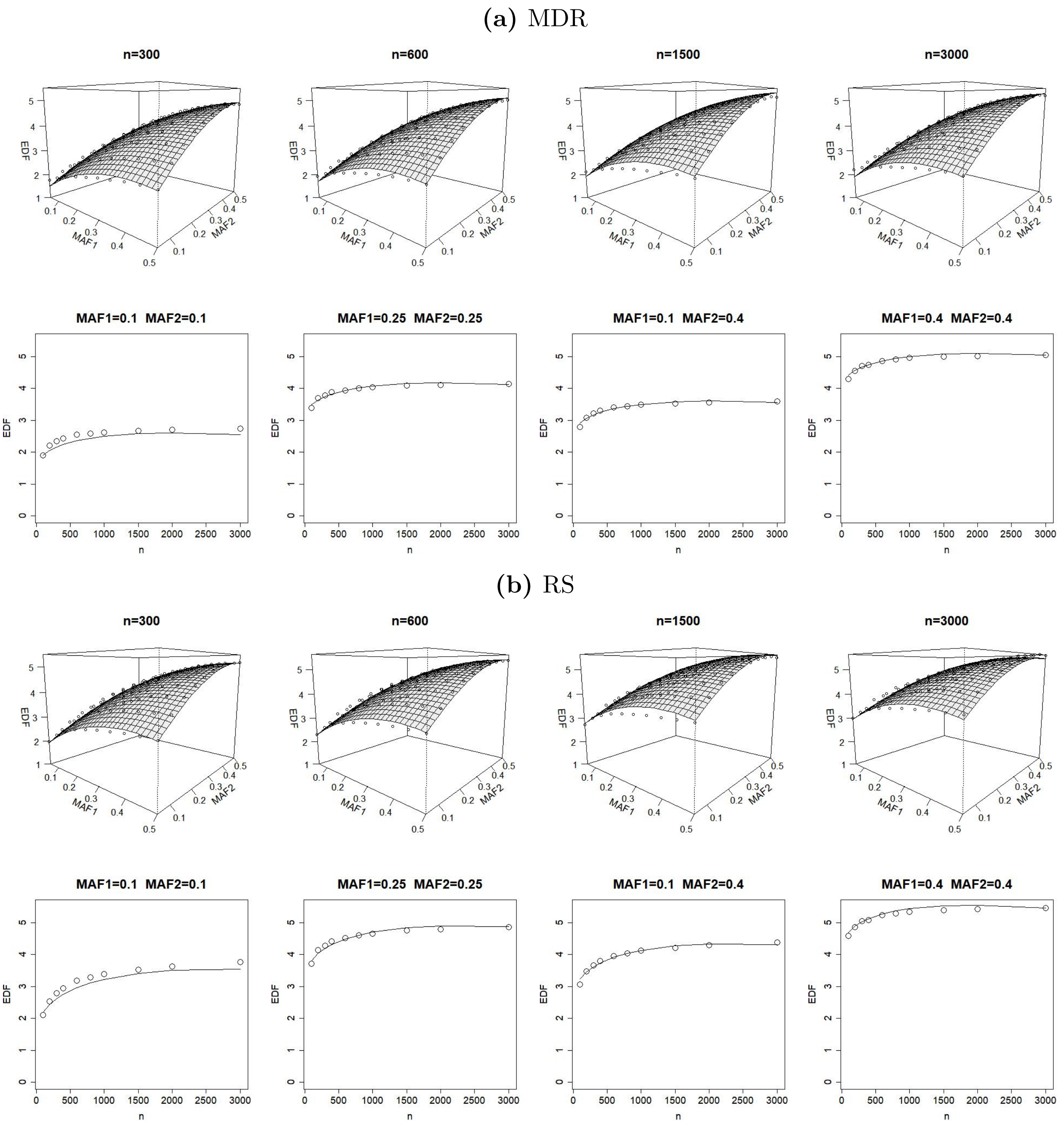
In practice, it is noticed that the null distribution for the nominal $\chi^2_{(1)}$-statistic varies for different SNPs. For this reason, this study aim to derive the EDF as a function of the MAFs of the SNPs. In addition, the sample size for the null data set would also affect the distribution of the null statistic, therefore it is also incorperated into the derivation.

In more detail, it has been empirically observed that the aforementioned EDF is a fairly smooth function of three underlying parameters: MAF_1, MAF_2—the minor allele frequencies of the two respective SNPs; and n—the sample size of the study. Hence, it suffices to compute the EDF for only a few combinations of $(\text{MAF}_1, \text{MAF}_2, n)$, and interpolate everywhere else using a response surface model (RSM) [105] (see Fig. 3.2). This is what makes it a very practical method for large-scale studies.

Specifically, through practice it is found that use of a quadratic RSM is sufficient. In addition, the RSM should also be constrained to be symmetric with respect to the two MAF arguments: MAF_1 and MAF_2. Hence, the final RSM used for interpolation is chosen to be of the following form:

$$\begin{aligned}\mathrm{EDF} \approx \beta_0 + &\underbrace{\beta_1(\mathrm{MAF}_1) + \beta_1(\mathrm{MAF}_2) + \beta_2(\sqrt{n}/100)}_{\text{main effects}} + \\ &\underbrace{\beta_3(\mathrm{MAF}_1)^2 + \beta_3(\mathrm{MAF}_2)^2 + \beta_4(\sqrt{n}/100)^2}_{\text{quadratic terms}} + \\ &\underbrace{\beta_5(\mathrm{MAF}_1)(\mathrm{MAF}_2) + \beta_6(\mathrm{MAF}_1)(\sqrt{n}/100) + \beta_6(\mathrm{MAF}_2)(\sqrt{n}/100)}_{\text{interactions}}, \end{aligned} \tag{3.4}$$

where a commonly-used square-root transform has also been applied to the sample size n.



(a) MDR

(b) RS

Fig 3.2. The estimated response surface model (3.4) for the MDR as shown in Eq. (3.5) and RS screening method as shown in Eq. (3.6). Top: EDF versus $(\mathrm{MAF}_1, \mathrm{MAF}_2)$ for $n = 300, 600, 1500$ and 3000. Bottom: EDF versus n for $(\mathrm{MAF}_1, \mathrm{MAF}_2) = (0.1, 0.1), (0.25, 0.25), (0.1, 0.4)$ and $(0.4, 0.4)$.

3.2.3 Parameter Estimation

Linear regression models are fitted on the simulated null data sets to obtain estimates of the parameters in Eq. 3.4. For the null data, the following settings are applied:

- MAF_1 and $\text{MAF}_2 \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$.
- Sample Size $n \in \{100, 200, 300, 400, 600, 800, 1000, 1500, 2000, 3000\}$. Balanced case-control data sets are simulated, e.g., for the setting of sample size 100, 50 cases and 50 controls are generated.
- Replicates of SNP-pairs: 100,000.

For each combination of MAF_1, MAF_2 and sample size, the MDR and RS procedures are run, and the averages of the testing statistics $\widehat{\chi}^2_{i,j}$ across all replicates of SNP-pairs are calculated and used as the response inputs in the linear regression model, whereas the corresponding MAF_1, MAF_2 and sample size are used as the explanatory variables. As a result, the linear regression models consist of $10 \times 10 \times 10 = 1000$ data points. Results for the regression models are summarized in Table 3.1. RSM model fitting curves and surfaces for some selected sample settings are presented in Figure 3.2. The figures show that the EDF generally increases with the increase of MAF or sample size.

3.2.4 Adjusted P-value Calculation

With the estimated parameters, estimates of the EDFs can be obtained for the MDR and RS methods by plugging in the estimates to Eq. 3.4. For each SNP-pair (i, j) with minor allele frequencies $(\text{MAF}_1, \text{MAF}_2)$, its adjusted p-value, $\widehat{p}_{\text{adj}}(i, j)$ can be computed as follows.

1. Use either the MDR method or the RS method to identify a DM, $M_{i,j}$.
2. Form a 2×2 cross table, as described in Section 1.3, and compute the usual chi-squared statistic, $\widehat{\chi}^2_{i,j}$, from Eq. (1.1).

Table 3.1. Parameter estimates for the MDR and R-S methods under the null distribution. The model equation is given in Eq. (3.4).

Parameter	Estimate	Standard Deviation	P-value	R^2	Adjusted R^2
		MDR			
β_0	0.078	0.039	0.044	0.99	0.99
β_1	7.678	0.105	<2e-16		
β_2	5.099	0.169	<2e-16		
β_3	-9.870	0.154	<2e-16		
β_4	-5.563	0.230	<2e-16		
β_5	5.609	0.194	<2e-16		
β_6	-0.125	0.143	0.383		
		RS			
β_0	-0.047	0.041	0.258	0.99	0.99
β_1	8.439	0.112	<2e-16		
β_2	8.385	0.180	<2e-16		
β_3	-10.403	0.164	<2e-16		
β_4	-7.564	0.244	<2e-16		
β_5	4.499	0.207	<2e-16		
β_6	-2.208	0.152	<2e-16		

3. If the MDR method has been used in Step 1 to identify $M_{i,j}$, compute the effective degree of freedom by (*see* Eq. 3.4 and Table 3.1)

$$\begin{aligned}\widehat{\mathrm{EDF}} = (0.078) + (7.678)(\mathrm{MAF}_1) + (7.678)(\mathrm{MAF}_2) + (5.099)(\sqrt{n}/100) \\ - (9.870)(\mathrm{MAF}_1)^2 - (9.870)(\mathrm{MAF}_2)^2 - (5.563)(\sqrt{n}/100)^2 \\ + (5.609)(\mathrm{MAF}_1)(\mathrm{MAF}_2) \\ - (0.125)(\mathrm{MAF}_1)(\sqrt{n}/100) - (0.125)(\mathrm{MAF}_2)(\sqrt{n}/100). \quad (3.5)\end{aligned}$$

4. If the RS method has been used in Step 1 to identify $M_{i,j}$, compute the effective degree of freedom by (see Eq. 3.4 and Table 3.1)

$$
\begin{aligned}
\widehat{\mathrm{EDF}} = (-0.047) + (8.439)(\mathrm{MAF}_1) + (8.439)(\mathrm{MAF}_2) + (8.385)(\sqrt{n}/100) \\
- (10.403)(\mathrm{MAF}_1)^2 - (10.403)(\mathrm{MAF}_2)^2 - (7.564)(\sqrt{n}/100)^2 \\
+ (4.499)(\mathrm{MAF}_1)(\mathrm{MAF}_2) \\
- (2.208)(\mathrm{MAF}_1)(\sqrt{n}/100) - (2.208)(\mathrm{MAF}_2)(\sqrt{n}/100). \quad (3.6)
\end{aligned}
$$

5. Compute the adjusted p-value as

$$
\hat{p_{adj}}(i,j) = \mathbb{P}\big(\chi^2_{\widehat{\mathrm{EDF}}} > \widehat{\chi}^2_{i,j}\big), \quad (3.7)
$$

and use it—as opposed to the nominal p-value, $p_{orig}(i,j)$, given by Eq. 1.2—to rank the SNP-pair (against other SNP-pairs).

3.3 Simulation Study

The proposed method is evaluated using the same set of simulations and measures as detailed in Section 2.7. The result of the first-order correction is shown in Fig. 3.3. It is clear from the figures that the p-value adjustments are able to signifcantly improve the SNP-pair detection.

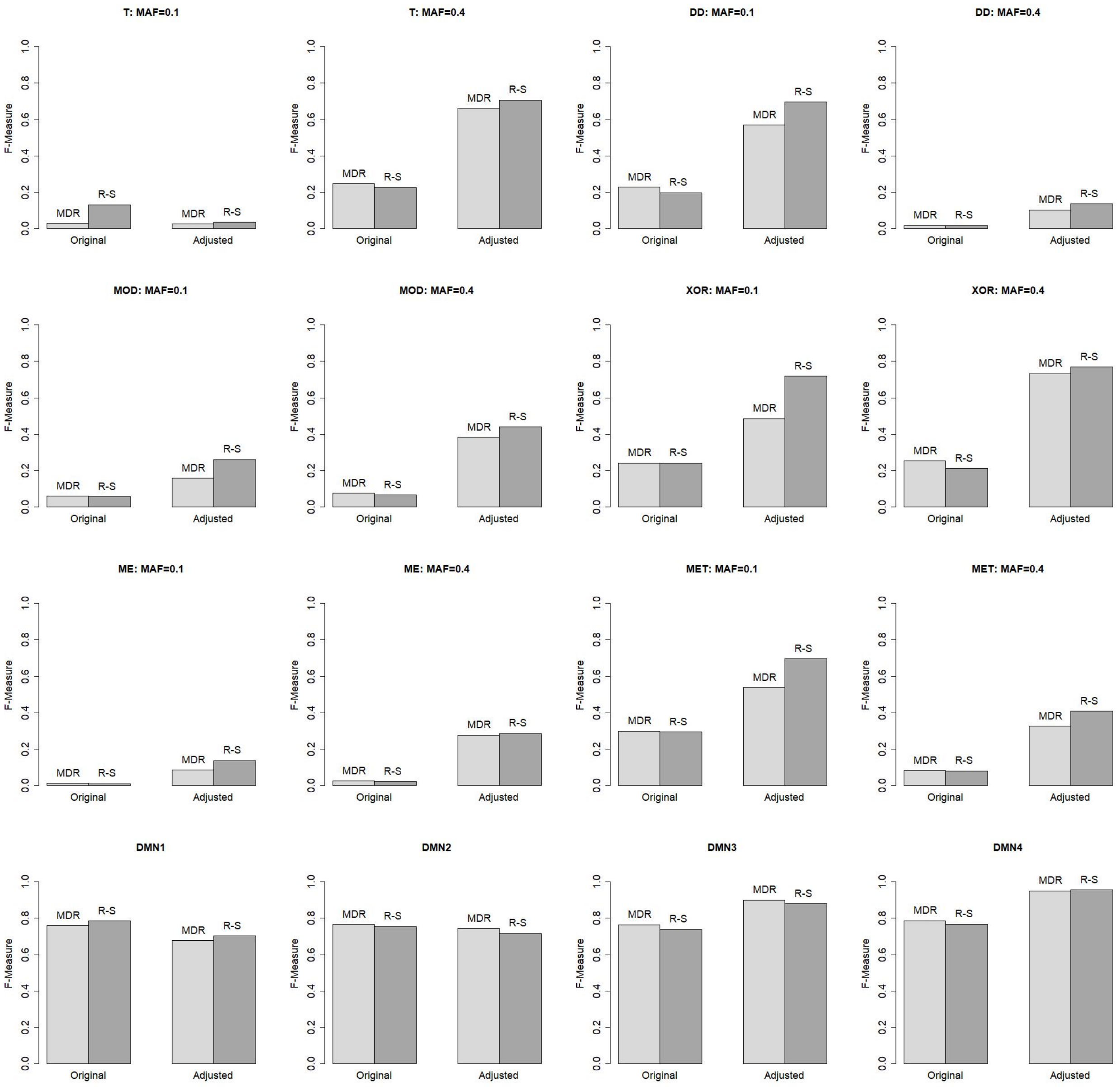


Fig 3.3. Simulation results in terms of F-measure with and without p-value adjustment for the MDR and RS methods. Each simulation has been repeated for 400 times and the average performance is being reported. Using the adjusted p-value given by Eq. (3.7) — as opposed to the nominal p-value given by Eq. (1.2) — generally improves the detection performance of popular screening methods such as the MDR and the RS, sometimes substantially.

Table 3.2 shows the detailed precision, recall and averaged F-measure for each model. It can be seen that in all cases, the adjustment of the p-values improves the detection precision as expected. Although there is a reduction of recall rate accordingly, the overall detection effect as indicated by the F-measure improves in the majority of cases.

Further observation shows that the improvement on the DMN models is not as obvious as the other ones using the first-order correction. Some explanations for this phenomenon are provided below.

For the two-level disease models (DM), because MDR and RS are designed to capture this kind of disease models, there is a tendency for the methods to capture false positive SNP-pairs, which contain one of the true SNP. When there is no DF correction, the chi-square test statistics are inflated. Therefore it is easy for MDR and RS to claim many significant SNP-pairs, which are false positives. This leads to a high recall rate and low precision. When the DF correction is applied, the recall rate decreases and the precision increases, which together lead to increased F-measure as a result. The increase comes from two ways: a) more increase in precision than the decrease of recall; b) a more balanced precision and recall rate to give rise to the F-measure due to its mathematical property.

For the DMN data sets, the true disease models are not two-level anymore. The MDR and RS methods have reduced power to claim significant SNP-pairs because the disease models are now only modelled approximately. In other words, there are fewer false positive pairs (especially the ones that contain one true and one false SNP) being detected compared to two-level DM models. This results in a relatively lower recall rate and a higher precision rate compared to the case of two-level DM, and a higher F-measure. When the DF correction is applied, the recall rate decreases and precision increases, which leads to increased F-measure. However, the increase is not as large as that for two-level models due to two possible reasons: first, the decrease of recall is too large; second, the balance of precision and recall do not change much before and after DF correction, i.e., there is no increase of F-measure due to the mathematical property of it.

Table 3.2. Results from simulation study for the P-value adjustment (Section 2.7). Precision, recall and average F-measures over 400 replications.

			Original			Adjusted		
Method	MAF	Model	Precision	Recall	F-Measure	Precision	Recall	F-Measure
MDR	0.1	T	0.025	0.030	0.027	0.313	0.013	0.024
		MOD	0.033	0.273	0.059	0.152	0.165	0.158
		DD	0.130	0.900	0.228	0.447	0.783	0.569
		XOR	0.155	0.548	0.242	0.634	0.393	0.485
		ME	0.006	1.000	0.011	0.015	0.998	0.029
		MET	0.189	0.718	0.299	0.557	0.520	0.538
MDR	0.4	T	0.141	0.975	0.246	0.527	0.885	0.661
		MOD	0.039	0.993	0.075	0.239	0.953	0.382
		DD	0.009	1.000	0.017	0.054	1.000	0.102
		XOR	0.144	0.998	0.252	0.590	0.963	0.732
		ME	0.012	0.990	0.024	0.072	0.965	0.135
		MET	0.043	0.850	0.081	0.214	0.673	0.325
MDR	0.25	DMN1	0.682	0.858	0.760	0.976	0.675	0.798
	0.25	DMN2	0.677	0.880	0.765	0.976	0.705	0.819
	0.1	DMN3	0.649	0.930	0.764	0.984	0.803	0.884
	0.1	DMN4	0.657	0.980	0.787	0.980	0.908	0.942
RS	0.1	T	0.106	0.170	0.130	0.467	0.018	0.034
		MOD	0.030	0.723	0.057	0.201	0.370	0.261
		DD	0.109	0.983	0.197	0.556	0.930	0.696
		XOR	0.139	0.913	0.241	0.752	0.688	0.719
		ME	0.006	1.000	0.011	0.015	1.000	0.030
		MET	0.175	0.915	0.294	0.681	0.713	0.696
RS	0.4	T	0.128	0.978	0.226	0.606	0.848	0.706
		MOD	0.034	0.993	0.066	0.286	0.943	0.439
		DD	0.008	1.000	0.016	0.072	1.000	0.135
		XOR	0.119	0.998	0.212	0.648	0.945	0.769
		ME	0.011	0.998	0.021	0.087	0.948	0.159
		MET	0.042	0.898	0.080	0.293	0.668	0.407
RS	0.25	DMN1	0.670	0.948	0.785	0.996	0.698	0.821
	0.25	DMN2	0.650	0.898	0.754	0.983	0.663	0.791
	0.1	DMN3	0.607	0.943	0.739	0.992	0.778	0.872
	0.1	DMN4	0.625	0.993	0.767	0.989	0.913	0.949

Chapter 4

General Epistasis Detection

4.1 Introduction

4.1.1 Overview

This chapter aims to develop methods for detecting more general epistasis, which is represented by disease models of multiple penetrance levels. Theoretically, a two-locus epistasis pattern may consist of up to nine different penetrances. As mentioned in Section 1.2.1, using the TTTC-type disease models to capture epistasis of this sort may increase the detection power, especially when the sample size is inadequate. When data observations permit, it is desirable to estimate the epistasis model accurately based on the research results on single-locus effects [33–35]. For instance, if the true model has three different risk levels of getting the disease, then a proper disease model of three penetrance levels is expected to lead to greater power than other ones because it can better capture the true effects.

Similar to the use of TTTC-type disease model for epistasis detection, the general problem in this chapter also requires two steps, i.e., determination of a disease model for a given SNP-pair and test of association between the SNP-pair and disease outcome. The difference is that the latter fits a flexible multi-level risk disease model to the SNP-pairs under study. Roughly speaking, the determination of the disease model is usually less challenging than the association test for the general problem. For instance, the problem can be formulated in a regression model setting where each genotype combination of a SNP-pair is coded as an indicator variable, i.e.,

$$\log\frac{p(y=1)}{p(y=0)} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_9 X_9, \tag{4.1}$$

where

$$X_i = \begin{cases} 1, & \text{if the genotype is } G_i, \\ 0, & \text{otherwise.} \end{cases}$$

Under such a framework, a set of coefficient estimates satisfying certain constraints would correspond to a particular disease model. For example, constraints defined by $\beta_1 = \beta_2 = \beta_3$ and $\beta_4 = \cdots = \beta_9$ indicates that the genotypes of G_1, G_2, G_3 have the same risk probabilities of disease, and so as genotypes of $G_4, \cdots, G_9$. In other words, the coefficient constraints define a two-level risk disease model. Based on this, it is easy to obtain a disease model by setting up some proper constraints and finding the solution to the regression problem defined above. In contrast, the test of association is likely to be more complicated because the determined disease model usually results from model selections; thereby the distribution of the testing statistic would depend on the model selection process and may not be straightforward to be determined.

The first proposed method is inspired by fused lasso [59] and post-model selection test. The fused lasso does neighbour coefficient fusion to achieve the exact same coefficient estimates for input variables under a regression model setting. Such a property makes it possible to obtain coefficient constraints described above, thereby determine a disease model of multiple risk levels. The fused lasso can be achieved by applying ordinary lasso on a set of transformed variables that are the differences of the original explanatory variables. Based on this, forward selection rather than the ordinary lasso is applied on the transformed variables in practice, which can achieve the same goal of multi-level disease model determination. In addition, the recent post-model selection test (PMST) is adopted to determine the number of coefficients to fuse, i.e., for two coefficients being fused, a p-value from PMST can be calculated and used to decide whether the difference between them is statistically significant to reject the fusion. Once a disease model is determined, it is refitted and the Pearson's χ^2-test can be performed to assess the association between the SNP-pair and the disease outcome. Details on this method are presented in Section 4.2.

The second method is derived by extending the RS method (Section 1.3.2), i.e., instead of doing one best split as does RS, it continues to split the genotype combinations till a desirable multi-level disease model is obtained. The method is named as "Sequential Split Procedure (SSP)", where the word of "sequential" is used for the reason that every subsequent split is built upon the previous splits. In this way, a maximum of eight splits are needed, which greatly reduces the computational complexity given the large pool of possible disease models to choose from. Similar to the RS procedure, after that, the risk model is refitted, and Pearson's χ^2-statistics are calculated for SNP-pair ranking. Details on this method are presented in Section 4.3.

The first and second methods are found to be equivalent under mild conditions, details of which are presented in Section 4.4.1. Because the two methods differ in the starting framework and determine disease models in different ways, it is of interest to compare them and understand various factors that contribute to the overall detection.

The third method, "Sequential Merge Procedure (SMP)", is also applied for multi-level risk disease model determination. It bears a similar idea as the second approach, and the only difference is that instead of a sequential split, it does a sequential merge of the genotypes for the purpose. As the whole process is quite similar to the split one, the details of it are presented in Appendix C.2.

Notice that at each step of the SSP (or SMP) method, the rule to determine how to split (or merge) is based on the best fitting P-values. As a result, the tesing statistics are bound to produce inflated p-values if they are evaluated against their nominal degree of freedom. For this reason, similar P-value adjustments as done for the RS method are applied to obtain more proper p-values for the assessment. Details on the adjustment are presented in Section 4.5.

4.1.2 General Simulation Setting

Throughout this chapter, the proposed methods are evaluated using the same simulation examples as introduced in Section 2.7. They cover a wide variety of disease models that are well suited for the investigation of multi-level epistasis detection methods. In more detail, they include four disease models of two-level risk (T, MOD, DD and XOR), one of five-level risk (ME), one of four-level risk (MET) and four of seven to nine different levels of risk (DMN1-DMN4). Ideally, the general epistasis detection methods are expected to produce comparable results as MDR and RS on the two-level risk disease models (T, DD, MOD and XOR), and better results on the multiple-level risk disease models (ME, MET, DMN1-DMN4).

A SNP-pair is claimed to be significantly associated with the outcome if the evaluation p-value is smaller than the given significance threshold. In this chapter, $\alpha^{\text{easy}} = 10^{-5}$ as given in Eq. 2.21 is adopted as the significance threshold. Compared to α^{hard}, α^{easy} takes into account the number of SNP-pairs being evaluated, but not the number of different disease models being tested for each pair. Because the proposed methods in this chapter has included p-value adjustment that corrects for inflation induced by testing of multiple disease models, there is no need to further control type-I error through the significance threshold. Same as before, the evaluation is through F-measure defined in Section 2.7, which assesses both detection power and precision.

4.2 Forward Selection on Transformed Difference Variables with Post-model Selection Test

This section first introduces the idea of fused lasso and its transformation to ordinary lasso. Then a brief review on the post-model selection test (PMST) is presented. After that, the steps to determine a multi-level disease model by forward selection and the post-model selection test are given. Based on the major idea of this method, i.e., "Forward Selection of

Difference variables with Post-Model Selection Test", the acronym "FSD-PST" is used to refer to the method.

4.2.1 Fused Lasso and Transformation to Ordinary Lasso

The fused lasso was first proposed by Tibshirani et al. [59] in the form of a penalized regression

$$\arg_{min}|y - X\beta|^2 + \lambda \sum_{i=1}^{p-1} |\beta_i - \beta_{i+1}| \tag{4.2}$$

where the explanatory variables $X_i, i = 1, 2, ..., p$ are expected to be ordered. The constraint term $\sum_{i=1}^{p-1} |\beta_i - \beta_{i+1}|$ forces the regression model to achieve equal coefficient estimates among neighbour variables, which makes it possible to determine a general disease model with flexible risk levels. For instance, assume that X_i's are the indicator variables representing the ith genotype, which have been ordered in some way. Also assume the model does not include an intercept term. Then a set of coefficient estimates $(\beta_1, \beta_2, \cdots, \beta_9) = (0.1, 0.1, 0.2, 0.2, 0.2, 0.2, 0.3, 0.3, 0.3)$ would indicate that the variables are divided into three groups, and the ones within the same group have the same penetrances. In other words, the coefficient estimates would translate to a disease model of three different risk levels.

The fused lasso problem above can be turned into ordinary lasso [106] by a simple variable transformation. Let $X = (X_1, X_2, ..., X_9)^T$ be the design matrix, where each X_i is the indicator variable for the ith "ordered" genotype formed by a SNP-pair. The transformation is conducted in the following steps:

1. Let

$$D = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}_{8\times 9}$$

denote the fused lasso penalty design matrix on β, i.e., $\sum_{i=1}^{p-1}|\beta_i - \beta_{i+1}| = |D\beta|$.

2. Append D by a vector of $(1,1,...,1)^T$, and denote the new matrix as $\tilde{D}$,

$$\tilde{D} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -1 \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}_{9\times 9}$$

3. The inverse of $\tilde{D}$ is

$$\tilde{D}^{-1} = \frac{1}{9}\begin{pmatrix} -8 & -7 & -6 & \cdots & -1 & 1 \\ 1 & -7 & -6 & \cdots & -1 & 1 \\ 1 & 2 & -6 & \cdots & -1 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 2 & 3 & \cdots & -1 & 1 \\ 1 & 2 & 3 & \cdots & 8 & 1 \end{pmatrix}_{9\times 9}$$

4. Let $\theta = (\theta_1, \theta_2, \cdots, \theta_9)^T = \tilde{D}\beta$, and then $\beta = (\beta_1, \beta_2, \cdots, \beta_9)^T = \tilde{D}^{-1}\theta$.

   The problem in Eq. 4.2 becomes

$$\arg\,\min|y - X\tilde{D}^{-1}\theta|^2 + \lambda\sum_{i=1}^{8}|\theta_i| \tag{4.3}$$

   Note θ_9 does not appear in the penalty term, which indicates the last variable appearing in the transformed design matrix $X\tilde{D}^{-1}$ is not penalized.

5. Write $X\tilde{D}^{-1} = (X_{new1} \quad X_{new2})$ and $\theta = (\theta_{new1}^T, \theta_9)^T$, then

$$X\tilde{D}^{-1}\theta = (X_{new1} \quad X_{new2})(\theta_{new1}, \theta_9)^T = X_{new1}\theta_{new1} + X_{new2}\theta_9$$

and the problem becomes

$$\arg\min|(y - X_{new1}\theta_{new1}) - X_{new2}\theta_9)|^2 + \lambda|\theta_{new1}|_1 \tag{4.4}$$

Fix the value of θ_{new1}^T. Because θ_9 does not appear in the penalty term, the minimization of the above equation with respect to θ_9 is equivalent to minimization of just

$$\arg\min|(y - X_{new1}\theta_{new1}) - X_{new2}\theta_9)|^2$$

From this new minimization problem, it is observable that the solution for θ_9 is

$$\hat{\theta_9} = (X_{new2}^T X_{new2})^{-1} X_{new2}^T (y - X_{new1}\hat{\theta_{new1}})$$

6. Plug back in the solution of θ_9 and denote $P = X_{new2}(X_{new2}^T X_{new2})^{-1} X_{new2}^T$, then the problem in Eq. 4.2 is reformulated as

$$\arg\min|(I - P)y - (I - P)X_{new1}\theta_{new1}|^2 + \lambda|\theta_{new1}|_1 \tag{4.5}$$

   Treat $(I - P)y$ and $(I - P)X_{new1}$ as the new outcome and explanatory variables, the fused lasso is transformed into ordinary lasso.

Based on the transformation, it is easy to see that testing $\beta_i - \beta_{i+1} = 0$ is equivalent to testing $\theta_i = 0, i = 1, 2, ..., 8$. For the new ordinary lasso model, the solution path for θ_{new1}'s is piece-wise linear and easy to be obtained using the "lars" algorithm. Once the solutions are calculated, a crucial next step is to decide on a proper value of λ which determines the final model.

While cross-validation is the most popular way to solve the problem, it is computationally intensive and does not necessarily provide the best solution. The advent of PMST methods [107] provides a new and promising direction. As the name indicates, a PMST is carried out after a model selection procedure, which calculates corrected p-values that have accounted for the selection events for statistical decision making. This implies that it can be applied to nodes on the solution path to calculate corrected p-values to determine λ.

Alternatively, forward selection (FS) other than the ordinary lasso can be applied on the transformed variables of $(I-P)y$ and $(I-P)X_{new1}$. In fact, the procedures described above indicate that any variable selection procedure on the transformed variables with its corresponding PMST could give a multi-level risk disease model. In practice, the forward selection with its tailored PMST is chosen for the task. The choice is due to its simplicity and usefulness, as well as a direct connection of it to the "Sequential Split" procedure to be introduced in Section 4.3.

Notice that the last column of $\tilde{D}^{-1}$ is a constant of 1/9, so $X_{new2} = \frac{1}{9}\sum_{i=1}^{9} X_i = 1/9$, which is also a constant. Therefore, the problem shown in Eq. 4.4 can be viewed as lasso regression on X_{new1} with an intercept term always included in the model. Henceforth, the FS procedure is performed on y and X_{new1} with an intercept in the model.

Notice the transformed variables X_{new1} consist of different contrasts, or in a simpler term, differences of the original variables, so the method is named **F**orward **S**election on **D**ifference variables (FSD).

4.2.2 Review on Exact Post-model Selection Test

The exact PMST by Tibshirani et al. [107] is adopted, which is specially designed for model selection procedures such as forward selection and lasso. It tests the partial regression coefficient of the variable to enter, at each step of the model selection procedure, in a projected linear model on the already selected variables. The test accounts for the adaptive nature of the model selection procedures by conditioning on the selection events such that the information already used during the selection will not be used again in the hypothesis testing. In Tibshirani et al. [107], the authors show that p-values from this type of tests are exact in finite samples under the linear regression model setting. In the paragraphs below, a brief review on the major idea of the PMST by Tibshirani et al. [107] is summarized, which mainly consists of two steps:

1. Perform the model selection and obtain the selection conditions at each selection step.
2. Perform the conditional tests and calculate corrected p-values.

The main idea of the PMST is described based on a typical FS procedure and the following notations:

- $y : (y_1, y_2, \cdots, y_n)^T \sim N(0, \Sigma)$;
- $\Gamma_{m,n}$: $(\Gamma_1^T, \Gamma_2^T, \cdots, \Gamma_m^T)^T$, a real-valued matrix of m rows and n columns;
- $u_{n,1}$: $(u_1, u_2, \cdots, u_n)$, a real-valued vector of length n;
- $v_{n,1}$: a real-valued vector of length n;

The parameter of interest is the coefficient of the variable to enter at each step and is described as $v^T y$. The key idea behind the conditional post-selection hypothesis tests is that the selection events for the forward procedure can be characterized as a set of polyhedral constraints on y, i.e., $\Gamma y \geq u$.

The FS procedure selects variables by repeatedly adding one predictor that most improves the model fitting to the current active set untill all predictors are in the model or the residual error becomes zero. After each addition, the coefficients are recomputed by least-square regression on the active predictors. The parameter of interest is the least-square coefficient estimate for the variable that has entered the model most recently.

Selection Condition

Suppose there are p variables in total. Without loss of generality, assume that y and X_i's have been centered, and that the variables enter the model by the following order: $X_1, X_2, \cdots, X_p$.

This indicates that for the first step, the residual sum of squares for fitting X_1 is smaller than that of the rest variables, i.e.,

$$\|\frac{I - X_1X_1^T}{\|X_1\|_2^2}y\|_2^2 \leq \|\frac{I - X_iX_i^T}{\|X_i\|_2^2}y\|_2^2, \text{ for } i \neq 1.$$

The inequality can be reduced to

$$s_1\frac{X_1^Ty}{\|X_1\|_2} \geq s_i\frac{X_i^Ty}{\|X_i\|_2}, \text{ for } i \neq 1 \text{ and } s_i = \text{Sign}(X_i^Ty),$$

or $\Gamma_1^Ty \geq u_1$, where $\Gamma_1^T = \frac{s_1X_1^T}{\|X_1\|_2} - \frac{s_iX_i^T}{\|X_i\|_2}$ and $u_1 = 0$. Based on the simplified form, the condition can be viewed as measuring the magnitude of the linear predictors X_i^Ty among the vectors y that would result in the FS procedure selecting variable X_1.

For a general step k, assume the variables $A_{k-1} = \{X_1, X_2, \cdots, X_{k-1}\}$ are in the model already and X_k is to be selected next. This means that X_k reduces the current model residual sum of squares the most among $\{X_k, X_{k+1}, \cdots, X_p\}$. Let r denote the model residual after regressing y onto the variable set $X_{A_{k-1}}$, and $\tilde{X}_i$ the residual after regressing X_i onto $X_{A_{k-1}}$, $i \in \{k, k+1, \cdots, p\}$. Again, assume r and $\tilde{X}_i$ have been centered. Then the condition for the kth step selection is

$$\|\frac{I - \tilde{X}_k\tilde{X}_k^T}{\|\tilde{X}_k\|_2^2}r\|_2^2 \leq \|\frac{I - \tilde{X}_i\tilde{X}_i^T}{\|\tilde{X}_i\|_2^2}r\|_2^2, \text{ for } i \in \{k+1, \cdots, p\},$$

which reduces to

$$s_k\frac{X_k^TP_{A_{k-1}}y}{\|P_{A_{k-1}}X_k\|_2} \geq s_i\frac{X_i^TP_{A_{k-1}}y}{\|P_{A_{k-1}}X_i\|_2}, \text{ for } i \in \{k+1, \cdots, p\},$$

where $P_{A_{k-1}} = I - X_{A_{k-1}}(X_{A_{k-1}}^TX_{A_{k-1}})^{-1}X_{A_{k-1}}^T$. Clearly, the condition can also be written in the form of $\Gamma_k^Ty \geq u_k$ for some Γ_k^T and u_k.

Parameter of Interest

At the first step, the parameter of interest to test is $(X_1^T X_1)^{-1} X_1^T y$. In a general step k, it is $e_k^T (X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T y$, where e_k is the kth standard basis vector. It can be seen that the parameter of interest can be represented by $v^T y$ for some vector v of length n.

With these notations, the PMST p-value to calculate is defined as

$$\mathbb{P}_{H_0}(v^T y | \Gamma y \geq u) \tag{4.6}$$

Conditional Test

The following lemmas provide the theoretical basis for testing the statistic of $v^T y | \Gamma y \geq u$, which essentially states that the conditions in the model selection procedures constrain the outcome variable y to be in a polyhedral set.

Lemma 4.2.1 (**Polyhedral selection as truncation**). *For any Σ, v such that $v^T \Sigma v \neq 0$, then*

$$\Gamma y \geq u \Longleftrightarrow \mathcal{V}^{lo}(y) \leq v^T y \leq \mathcal{V}^{up}(y), \mathcal{V}^0(y) \leq 0,$$

where

$$\begin{aligned}
\rho &= \frac{\Gamma \Sigma v}{v^T \Sigma v} \\
\mathcal{V}^{lo}(y) &= \max_{j:\rho_j > 0} \frac{u_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j} \\
\mathcal{V}^{up}(y) &= \min_{j:\rho_j < 0} \frac{u_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j} \\
\mathcal{V}^0(y) &= \max_{j:\rho_j = 0} u_j - (\Gamma y)_j
\end{aligned}$$

Based on this Lemma, the following is true:

$$v^T y|\Gamma y \geq u \Longleftrightarrow v^T y|\mathcal{V}^{lo}(y) \leq v^T y \leq \mathcal{V}^{up}(y), \mathcal{V}^0(y) \leq 0,$$

In other words, when conditional on a selection process $\Gamma y \geq u$, the linear function $v^T y$ would follow a truncated Gaussian distribution. The following lemma provides the theoretical basis for calculating the p-value in Eq. 4.6.

Lemma 4.2.2 (Pivotal Statistic after Polyhedral Selection). *Let $\Phi(x)$ denote the c.d.f. of the standard normal distribution, and $F^{[a,b]}_{\mu,\sigma^2}$ the c.d.f. of a $N(\mu, \sigma^2)$ random variable truncated in $[a, b]$, i.e.,*

$$F^{[a,b]}_{\mu,\sigma^2}(x) = \frac{\Phi(\frac{x-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}.$$

Then for $v^T \Sigma v \neq 0$ and $y \sim N(\theta, \Sigma)$, the statistic $F^{[\mathcal{V}^{lo},\mathcal{V}^{up}]}_{\mu,\sigma^2}(v^T y)$ is a pivotal quantity conditional on $\Gamma y \geq u$:

$$\mathbb{P}(F^{[\mathcal{V}^{lo},\mathcal{V}^{up}]}_{v^T\theta, v^T\Sigma v}(v^T y) \leq \alpha | \Gamma y \geq u) = \alpha,$$

where $\mathcal{V}^{lo}, \mathcal{V}^{up}$ are as given above.

Remark

The reviewed PMST method above is for linear regression models. In Taylor and Tibshirani (2018) [108], the PMST is proposed for l_1 penalized likelihood models, which is applicable to logistic regression. The idea is to use the iteratively reweighted least-squares (IRLS) algorithm to fit the logistic regression model, then the parameter estimates in the logistic model could be expressed as weighted least square estimates (as opposed to least-square estimates in linear regression) that would also take the form of $v^T y$ for some vector v. Additionally, the authors show that the selection conditions could also be expressed as y falling into a polyhedral set, and the PMST p-values for logistic regression can be calculated in a similar fashion to that for linear regression.

However, in practice, linear regression models are chosen for this study, despite the fact that the outcome variable is binary and logistic regression appears to be a more popular choice. The reason is that the calculation of PMST p-values following the IRLS algorithm is too computationally intensive to be applied at a large scale. Additionally, the weighted LS estimates could be quite unstable for SNPs that contain sparse observations in some of their genotypes, which also causes computational challenge in calculating the PMST p-values. Further, the estimated parameters of interest and the corresponding PMST p-values suffer from inaccuracy as a result of the IRLS algorithm. Though the linear regression modelling framework only provides approximate results, it is much easier to be applied, especially at a larger scale.

4.2.3 Application to Disease Model Determination and SNP-pair Ranking

Disease Model Determination

The following notations are borrowed from Section 4.2.1 to demonstrate the "FSD-PST" method in this section, i.e.,

- $X_i, i = 1, 2, \cdots, 9$: indicator variable for the ith ordered genotype by the case-to-control ratios
- $X = (X_1, X_2, \cdots, X_9)$: design matrix for genotypes of the SNP-pair
- $\tilde{D}$: 9×9 transformation matrix

Denote $\tilde{X} = X\tilde{D}^{-1} = (\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_9)$. As mentioned earlier, a linear regression framework is adopted for practical consideration. Based on the transformed modelling form in Eq. 4.4, the transformed variables $\{\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_8\}$ are entered into a simple linear regression model by forward stepwise selection, where an intercept is always included.

Specifically, for the first step, the following regression models are fitted:

$$y = \gamma_{0i}^1 + \gamma_{1i}^1 \tilde{X}_i, i = 1, 2, \cdots, 8.$$

Among the eight candidate variables, the one that leads to the smallest model residual sum of squares is selected. Without loss of generality, assume the selected variable is $\tilde{X}_1$. Then the corrected p-value ($\tilde{p_1}$ as shown below) for the first selected variable is calculated and retained for determining the final disease model.

$$\tilde{p_1} = \mathbb{P}_{H_0:\gamma_{11}^1=0}(\hat{\gamma_{11}^1} | RSS(y, \tilde{X_1}) \leq RSS(y, \tilde{X_i}, i = 2, \cdots, 8)) \tag{4.7}$$

For a general step k, without loss of generality, assume that $\{\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_{k-1}\}$ are already in the model, then the following candidate models are fit:

$$y = \gamma_{0i}^k + \gamma_{1i}^k \tilde{X}_1 + \gamma_{2i}^k \tilde{X}_2 + \cdots + \gamma_{k-1,i}^k \tilde{X}_{k-1} + \gamma_{k,i}^k \tilde{X}_i, i = k, k+1, \cdots, 8.$$

Assume X_k yields the least RSS and is chosen at step k, then the corrected p-value as shown below is calculated and adopted for disease model determination.

$$\tilde{p_k} = \mathbb{P}_{H_0:\gamma_{k,k}^k=0}(\hat{\gamma_{k,k}^k} | RSS(y, (\tilde{X}_{A_{k-1}}, \tilde{X}_k)) \leq RSS(y, (\tilde{X}_{A_{k-1}}, \tilde{X}_i), i = k+1, \cdots, 8)) \tag{4.8}$$

The process is continued till every variable is selected into the model. Denote $\tilde{p} = (\tilde{p_1}, \tilde{p_2}, \cdots, \tilde{p_8})$ as the sequence of PMST p-values from the forward selection procedure. To determine the final disease model for a given SNP-pair, the following criteria are used.

1. Minimum p-value, i.e.,

$$K_{PMSTP} = \{K : \text{ for which } \tilde{p}_K = \text{Min}_{i=1,2,\cdots,8}\ \tilde{p}_i\}, \tag{4.9}$$

2. False discovery rate control (FDR) as proposed by G'Sell [109]. This method can be viewed as an extension of the BH procedure [110] used for FDR control. For the BH procedure, the rejection set can be arbitrary, whereas for the sequence of PMST

p-values, the rejection must always be the first K hypotheses for some K because the PMST p-values are ordered by the way the variables enter the model.

Given a list of p-values $p_i, i = 1, 2, \cdots, m$ and a significance level α, the BH procedure is carried out by first sorting them $(p_i, i = 1, 2, \cdots, m)$ and then finding

$$K_\alpha = max\{k : p_{\{k\}} \leq \frac{\alpha k}{m}\}$$

By rejecting the hypotheses corresponding to the K_α smallest p-values, the BH procedure controls FDR at level α.

The idea of FDR control in G'Sell [109] is to transform the PMST p-values into statistics $q_1 < q_2 < \cdots < q_8$ such that q_i's would behave like a sorted list of p-values that can be used in the BH procedure.

Under the null hypotheses that $\tilde{p}_k \sim U[0, 1]$ i.i.d, let

$$T_i = -log(1 - \tilde{p}_k),$$

then T_i's are distributed as independent exponential random variables. Let

$$R_i = \sum_{j=1}^{i} \frac{T_j}{m - j + 1}.$$

The Renyi representation theorem [111] shows that R_i's have the same distribution as a sorted list of independent standard exponential random variables.

Let

$$q_i = 1 - exp(-R_i).$$

q_i's are distributed like the uniform order statistics. Hence, calculate

$$K_\alpha^q = max\{k : q_{\{k\}} \leq \frac{\alpha k}{m}\}$$

and reject the K_α^q smallest p-values, the procedure can control FDR at α.

In G'Sell [109], the author further applied the following approximations

$$\frac{T_i}{m-j+1} \approx \frac{T_i}{m} \text{ for } j \ll m$$

$$\frac{1-e^{-x}}{x} \to 1 \text{ for } x \to 0, \text{ so } q_i \approx R_i \approx \sum_{j=1}^{i} \frac{T_j}{m} \text{ when } m \to \infty$$

and proposed to conduct FDR control by calculating

$$K_\alpha = max\{k : \frac{1}{k}\sum_{i=1}^{k} T_i \leq \alpha\}$$

With these establishments, FDR control on the PMST p-values at a significance level of α is performed by calculating

$$K_{FDR} = max\{k \in \{1, 2, \cdots, 8\} : -\frac{1}{k}\sum_{i=1}^{k} log(1-\tilde{p_i}) \leq \alpha\}, \tag{4.10}$$

and rejecting all hypotheses corresponding to the PMST p-values before and at K, i.e., $\tilde{p_1}, \tilde{p_2}, \cdots, \tilde{p_K}$.

In other words, the final disease model is selected to have $\hat{K}+1$ levels under this procedure, which is guaranteed to have an FDR controlled at α.

Note that in the above derivation steps, some assumptions have been applied. The first one is that the sequence of p-values are independent, which the PMST p-values do not satisfy. The second one is that the total number of null hypotheses is large relative to the number of non-null ones. This may not apply so well for the SNP-pair genotypes either, because the total hypotheses are only eight (corresponding to differences of nine genotypes for a SNP-pair), so the ratio of null to non-null hypotheses would not exceed seven when any non-null hyothesis exists.

3. Changes in AIC. This criterion is taken from the PMST paper [107] and recommended as a possible strategy for model selection on the PMST p-values. It determines the final disease model in the following steps:

   (a) Refit the selected disease model at each step, and calculate AICs from the corresponding regression models.

   (b) Select the step at which there have been two consecutive rises in the AIC criterion, assume it is step K_0.

   (c) Conduct statistical inference on the selected variables by applying a Bonferroni correction to the p-values of $\tilde{p_1}, \tilde{p_2}, \cdots, \tilde{p}_{K_0}$. The final disease model is selected by the following criteria:

   $$K_{AIC} = max\{k \in \{1, 2, \cdots, K_0\} : \tilde{p_k} \leq \frac{\alpha}{K_0}\}, \tag{4.11}$$

   The authors [107] mention that such a stopping rule defines a polyhedral constraint on the outcome variable that has been included in calculating the PMST p-values. Therefore, the usual statistical inference could be conducted on the PMST p-values when the model is determined by such an adaptively selected model at step K_{AIC}.

4. Changes in BIC. This criterion is all the same as the aforementioned AIC-style criteria except that the AIC is replaced by the BIC criterion.

Once the step K is selected by the above criteria, the final model is:

$$y = \gamma_{0i}^{K} + \gamma_{11}^{K}\tilde{X_{i1}} + \gamma_{12}^{K}\tilde{X_2} + \cdots + \gamma_{1,K}^{K}\tilde{X_K},$$

which means that the transformed variables of $\tilde{X_1}, \tilde{X_2}, \cdots, \tilde{X_K}$ are selected, i.e., splits between the original variables of (X_1, X_2), (X_2, X_3), $\cdots$, (X_K, X_{K+1}) are made sequentially.

In other words, the final disease model is split as

$$G_1 \mid G_2 \mid \quad \cdots \quad \mid G_K \mid G_{K+1} \quad G_{K+2} \quad \cdots \quad G_8 \quad G_9$$

i.e., each $G_i, i \leq K$ is its own group, and $G_i, K+1 < i < 9$ are in one group.

Remark

1. Note that the presented example above is based on the assumption that the forward selection procedure selects the transformed variables in the order of $\tilde{X_1}, \tilde{X_2}, \cdots, \tilde{X_K}$ for convenient illustration. More generally, assume the transformed variables of $\tilde{X_{i_1}}, \tilde{X_{i_2}}, \cdots, \tilde{X_{i_K}}$ are selected in the final model (not necessarily in the order of $\tilde{X_{i_1}}, \tilde{X_{i_2}}, \cdots, \tilde{X_{i_K}}$), where $1 \leq i_1 < i_2 < \cdots < i_K$, then the final disease model is split as

$$G_1 \quad \cdots \quad G_{i_1} \mid G_{i_1+1} \cdots \quad G_{i_2} \mid \quad \cdots \quad \mid G_{i_{K-1}+1} \quad \cdots \quad G_{i_K} \mid G_{i_K+1} \quad \cdots \quad G_9$$

2. Although the p-values from PMSTs (e.g., Eqs. 4.7 and 4.8) have accounted for the model selection process, there is still some inflation about them. This comes from the fact that the genotypes are always ordered by the case-to-control ratios at the beginning, which has relied on the outcome data information.

Association Assessment

Once the final disease model of K levels is determined, it is refit to the SNP-pair by counting the number of cases and controls in each risk group as defined by the disease model. A cross-table of $K \times 2$ is created, and Pearson's χ^2-test with $K-1$ d.f. is performed. The resulting p-values are used directly to assess the SNP-pairs for their association with the outcome.

4.3 Sequential Split Procedure

Inspired by the RS method [7] (section 1.3.2), it is straightforward to continue the splitting of the genotypes for a SNP-pair to generate disease models of multiple risk levels. Because the split procedure is sequential in the sense that each subsequent split is conditional on the previous splits, it is referred to as “sequential split procedure”, or “SSP” for short.

The organization in this section is as follows. First, an overview of the overall splitting procedure is given. Second, details on the statistics used for making the splitting decision and the criteria used for determining the final disease model are introduced. Third, a summary view of the splitting and merging procedures (details on the merging procedures are given in Appendix C.2) are presented with some discussions.

4.3.1 Procedure Overview

The idea of the SSP is to make successive splitting to the genotypes of a given SNP-pair till a proper number of groups are formed. For the first split, there are $2^9 - 2 = 510$ ways to group the genotypes, which reduces to 8 if minimization of misclassification error is desired based on the theory mentioned in Wan et al. (2013)'s method [7]. The reduction is achieved by sorting the 9 genotypes by their case-to-control ratios and splitting the ordered genotypes only. The theory states (Theorem 1.3.1) that the split that minimizes the classification error is guaranteed to be among splits of the ordered genotypes. Inspired by this, it is straightforward to continue splitting the ordered genotypes after the first split. The method is presented in more detail below.

Denote $X = (X_1, X_2, \cdots, X_9)$ as the indicator variables of the genotypes $G_i (i = 1, 2, \cdots, 9)$ formed by two SNPs. Similar to RS, G_i's are first sorted according to their case-to-control ratios. Without loss of generality, assume that the G_i's are already ordered. The sequential split procedure consists of eight sequential steps of nested splits that start from all G_i's being in one group to each one being its own group (refer to 4.12 below for an example).

Example 4.3.1. *An Example of the Proposed General Sequential Splitting Procedure*

$$\begin{array}{ll}
\textit{Step 1:} & \overbrace{X_1 \quad X_2 \quad X_3 \quad X_4}^{g_1} \mid \overbrace{X_5 \quad X_6 \quad X_7 \quad X_8 \quad X_9}^{g_2} \\
\textit{Step 2:} & \overbrace{X_1 \quad X_2 \quad X_3 \quad X_4}^{g_1} \mid \overbrace{X_5 \quad X_6}^{g_2} \mid \overbrace{X_7 \quad X_8 \quad X_9}^{g_3} \\
\textit{Step 3:} & \overbrace{X_1}^{g_1} \mid \overbrace{\quad X_2 \quad X_3 \quad X_4}^{g_3} \mid \overbrace{X_5 \quad X_6}^{g_3} \mid \overbrace{X_7 \quad X_8 \quad X_9}^{g_4} \\
\vdots & \\
\textit{Step 8:} & \overbrace{X_1}^{g_1} \mid \overbrace{X_2}^{g_2} \mid \overbrace{X_3}^{g_3} \mid \overbrace{X_4}^{g_4} \mid \overbrace{X_5}^{g_5} \mid \overbrace{X_6}^{g_6} \mid \overbrace{X_7}^{g_7} \mid \overbrace{X_8}^{g_8} \mid \overbrace{X_9}^{g_9}
\end{array}$$

In general, the sequential split is done in the following steps:

Step 1 Test each one of the eight possible splits (as illustrated below) using a proper statistic (details on the choice of the statistic are presented in Section 4.3.2). Select the one with the best fit.

$$\begin{array}{ll}
\text{Candidate Split 1:} & X_1 \mid X_2 \quad X_3 \quad X_4 \quad X_5 \quad X_6 \quad X_7 \quad X_8 \quad X_9 \\
\text{Candidate Split 2:} & X_1 \quad X_2 \mid X_3 \quad X_4 \quad X_5 \quad X_6 \quad X_7 \quad X_8 \quad X_9 \\
\vdots & \\
\text{Candidate Split 8:} & X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad X_6 \quad X_7 \quad X_8 \mid X_9
\end{array}$$

Step 2 Fix the first split place and explore the rest seven places for a possible second split (see illustration below). Perform a proper test with the first split accounted for (details are presented in Section 4.3.2) and select the best fit.

$$\text{Candidate Split 1:}\quad X_1 \mid X_2 \quad X_3 \quad X_4 \quad \overbrace{\quad\mid\quad}^{\text{1st Split}} \quad X_5 \quad X_6 \quad X_7 \quad X_8 \quad X_9$$

$$\text{Candidate Split 2:}\quad X_1 \quad X_2 \mid X_3 \quad X_4 \quad \overbrace{\quad\mid\quad}^{\text{1st Split}} \quad X_5 \quad X_6 \quad X_7 \quad X_8 \quad X_9$$

$$\vdots$$

$$\text{Candidate Split 7:}\quad X_1 \quad X_2 \quad X_3 \quad X_4 \quad \overbrace{\quad\mid\quad}^{\text{1st Split}} \quad X_5 \quad X_6 \quad X_7 \quad X_8 \mid X_9$$

Step 3 Continue similar procedure as step 2 (while fixing the existing splits) till every G_i is its own group.

Step 4 Refit the disease model selected at each step to quantify the association of the SNP-pair and the outcome. A final disease model is determined by using a "proper" stopping criterion, as shown in Section 4.3.2.

Notice there are two selections in the process: selection of a place to split at each step, and selection of a final disease model from all steps. For the first type of selection, the standardized coefficient difference from fitting logistic regression models are used as the criteria. For the second type of selection, four commonly used model selection criteria are adopted, which are minimum of 1)"nominal" p-value from Pearson's χ^2-test 2) AIC 3) BIC 4) p-value from Likelihood Ratio Test (LRT). Details of these are given in Section 4.3.2.

4.3.2 Testing Statistics and Stopping Criteria

The choice for the place to split at each step is based on logistic regression. When a split is made, the expectation is that the disease probabilities for the two newly split subgroups are significantly different. Based on this idea, the statistic can be derived to test the coefficient difference of the two subgroups of variables.

Step One Splitting

For step 1, suppose the split between X_i and X_{i+1} is being assessed, then the model being fit is

$$\log\frac{p(y=1)}{p(y=0)} = \beta_1\mathcal{X}_{g_1} + \beta_2\mathcal{X}_{g_2}, i = 1, 2, \cdots, 8,$$

where the design matrix is

$$(\mathcal{X}_{g_1} \quad \mathcal{X}_{g_2}) = (\textstyle\sum_{l=1}^{i} X_l \quad \sum_{l=i+1}^{9} X_l) = X \left(\begin{array}{cc} \overbrace{}^{i\,\text{cols}} & \overbrace{}^{9-i\,\text{cols}} \\ 1\cdots 1 & 0\cdots 0 \\ 0\cdots 0 & 1\cdots 1 \end{array} \right)^T$$

Remark. *Because all the X_i's are dummy variables, combining the genotype combinations is equivalent to summing the indicator variables up.*

Based on this logistic regression model, the hypothesis $H_0 : \beta_1 = \beta_2$ can be carried out by

$$Z_i = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\widehat{\text{std}\ (\hat{\beta}_1 - \hat{\beta}_2)}} \sim N(0, 1).$$

A selection of split i_1 over all $i, i = 1, 2, \cdots, 9$ means the following is true

$$|Z^1_{i_1}| > |Z^1_i|, i \neq i_1. \tag{4.12}$$

Testing Statistics for Splitting at Step Two

Assume the first split is at place i_1 (between the i_1th and (i_1+1)th variable). For step 2, the following models are explored,

$$\log\frac{p(y=1)}{p(y=0)} \quad = \quad \beta_1\mathcal{X}_{g_1^{i_1}} + \beta_2\mathcal{X}_{g_2^i} + \beta_3\mathcal{X}_{g_3^i}, \text{for all } i \neq i_1 \tag{4.13}$$

where the superscripts i_1 and i have been used to indicate the spliting place and

$$\begin{pmatrix} \mathcal{X}_{g_2^i} & \mathcal{X}_{g_3^i} \end{pmatrix} = \begin{pmatrix} \sum_{l=i_1+1}^{i} X_l & \sum_{l=i+1}^{9} X_l \end{pmatrix} = X \begin{pmatrix} \overbrace{0\cdots0}^{i_1 \,\text{cols}} & \overbrace{1\cdots1}^{i-i_1 \,\text{cols}} & \overbrace{0\cdots0}^{9-i \,\text{cols}} \\ 0\cdots0 & 0\cdots0 & 1\cdots1 \end{pmatrix}^T \text{for}\, i > i_1$$

$$\begin{pmatrix} \mathcal{X}_{g_2^i} & \mathcal{X}_{g_3^i} \end{pmatrix} = \begin{pmatrix} \sum_{l=1}^{i} X_l & \sum_{l=i+1}^{i_1} X_l \end{pmatrix} = X \begin{pmatrix} \overbrace{1\cdots1}^{i \,\text{cols}} & \overbrace{0\cdots0}^{i_1-i \,\text{cols}} & \overbrace{0\cdots0}^{9-i_1 \,\text{cols}} \\ 0\cdots0 & 1\cdots1 & 0\cdots0 \end{pmatrix}^T \text{for}\, i < i_1$$

Similarly, hypothesis tests of $H_0 : \beta_{g_2^i} = \beta_{g_3^i}$ are carried out based on the asymptotic distribution of the statistic

$$Z_i = \frac{\hat{\beta}_2 - \hat{\beta}_3}{\text{std}\ \widehat{(\hat{\beta}_2 - \hat{\beta}_3)}} \sim N(0, 1).$$

Select i_2 for which

$$|Z_{i_2}| > |Z_i|, i \neq i_1, i_2. \tag{4.14}$$

Testing Statistics for a General Step

Step 3 and beyond could be done similarly, i.e., regression models with dummy variables representing two newly split groups are fit:

$$\log\frac{p(y=1)}{p(y=0)} \quad = \quad \beta_1\mathcal{X}_{g_1^{i_1}} + \cdots + \beta_k\mathcal{X}_{g_k^{i_k}} + \beta_{k+1}\mathcal{X}_{g_{k+1}^i} + \beta_{k+2}\mathcal{X}_{g_{k+2}^i}, \tag{4.15}$$

for all $i \neq i_1, i_2, \cdots, i_k$.

The coefficients are tested for statistically significant differences:

$$Z_i = \frac{\hat{\beta}_{k+1} - \hat{\beta}_{k+2}}{\text{std } \widehat{(\hat{\beta}_{k+1} - \hat{\beta}_{k+2})}} \sim N(0, 1).$$

Select i_{k+1} for which

$$|Z_{i_{k+1}}| > |Z_i|, i \neq i_1, i_2, \cdots, i_k. \tag{4.16}$$

Remark. *The indicator variables in Eq. 4.15 are mutually independent, and perfectly correlated, i.e.,* $\mathcal{X}_{g_1^{i_1}} + \cdots + \mathcal{X}_{g_k^{i_k}} + \mathcal{X}_{g^i_{k+1}} + \mathcal{X}_{g^i_{k+2}} = 1$*, so the model can be reduced to*

$$log\frac{p(y=1)}{p(y=0)} = \beta_{k+1}\mathcal{X}_{g^i_{k+1}} + \beta_{k+2}\mathcal{X}_{g^i_{k+2}}, for\ all\ i \neq i_1, i_2, \cdots, i_k \tag{4.17}$$

Disease Model Level Selection

In each step of the split procedure, there is a corresponding disease model being selected. To determine the final one, the disease model of each level is refitted. Table 4.1 and Eq. 4.18 present the refitted example for a disease model of level $j+1$. Four types of criteria are calculated to make the decision.

Table 4.1. Distribution of individuals for the disease model after the jth step of the SSP method.

	g_1	g_2	$\cdots$	g_{j+1}
Cases	n_{11}	n_{12}	$\cdots$	$n_{1,j+1}$
Controls	n_{21}	n_{22}	$\cdots$	$n_{2,j+1}$

$$\log\frac{p(y=1)}{p(y=0)} = \beta_1 X_{g_1} + \beta_2 X_{g_2} + \cdots + \beta_j X_{g_{j+1}} \tag{4.18}$$

1. Minimum "nominal" p-value: nominal p-values are calculated using the Pearson's χ^2-test, i.e.,

$$p_j = \Pr\big(\chi^2_{(j)} > \widehat{\chi}^2_j\big), j = 1, \cdots, 8,$$

where $\widehat{\chi}^2_j$ is the Pearson's χ^2-statistic after the jth step that is calculated based on Table 4.1. The final disease model is the one with the smallest nominal p-value, i.e.,

$$\text{Selected Level} = K,\ \text{for which}\ p_K = \min_{j=1,2,\cdots,8} p_j$$

2. Minimum AIC: logistic regression model is fitted for each step of the split (see example in Eq. 4.18 for the jth step) and the AIC is calculated. The one with the smallest AIC is selected.

$$AIC_j = -2log(Likelihood) + 2(j+1)$$

$$\text{Selected Level} = K,\ \text{for which}\ AIC_K = \min_{j=1,2,\cdots,8} AIC_j$$

3. Minimum BIC: similar to that of AIC except the criteria is changed to BIC.

$$BIC_j = -2log(Likelihood) + 2log(N)(j+1),$$

where N is the sample size.

$$\text{Selected Level} = K,\ \text{for which}\ BIC_K = \min_{j=1,2,\cdots,8} BIC_j$$

4. Minimum p-value from Likelihood Ratio Test (LRT): the two models before and after a split are nested, so a Likelihood Ratio Test can be carried out to assess the necessity of the split.

$$\begin{aligned}
D_1 &= \text{Deviance at Step 1} \sim \chi^2_{(1)} \\
D_j &= \text{Deviance at Step } (j-1) - \text{Deviance at Step } j \sim \chi^2_{(1)}, j \geq 2 \\
p^D_j &= Pr(\chi^2_{(1)} > D_j), j = 1, 2, \cdots, 8 \\
\text{Selected Level} = K,&\ \text{ for which } p^D_K = \min_{j=1,2,\cdots,8} p^D_j
\end{aligned}$$

Association Assessment

When the final disease model is determined for a SNP-pair (i, j) (assume it has $K + 1$ risk levels), it is refit and Pearson's χ^2-statistic $\hat{\chi}^2_{i,j}$ is calculated. After that, a nominal p-value $\Pr(\chi^2_{(K)} > \hat{\chi}^2_{i,j})$ and p-value by adjusted degree of freedom $\Pr(\chi^2_{(EDF)} > \hat{\chi}^2_{i,j})$ are computed and used to assess the significance of the association between a SNP-pair and the outcome.

Notice the nominal p-values are inflated due to the best-fitting selections in the procedure, therefore in Section 4.5, adjusted p-values similar to those applied for the MDR and RS methods are calculated for the SSP method.

4.3.3 Method Summary

In the application of the SSP or SMP, a heuristic approach is chosen that involves an ordering of the genotypes at its first step and two types of "model selection" thereafter, i.e., the best fitting disease model at each step and the best disease model out of all steps. The complete process is illustrated in Figure 4.1.

Ordering the case-to-control ratios in step one (Figure 4.1) and selecting the best disease models at step two uses the outcome data, so they both lead to inflation in the nominal p-value in step five. For step three, it is not straightforward to know whether any of the criteria leads to inflation. In Section 4.5, a method is proposed to correct for the inflation regardless of where the inflation is from.

Compare to the SSP, FSD-PST (Section 4.2) mainly differs from it in step three (a comparison of the two methods is given in Section 4.4.1). For FSD-PST, its PMST p-value criterion used in step three has partially accounted for the inflation introduced at step two, therefore is expected to lead to better disease model level determination than the "nominal" p-value criteria used by the SSP. Nonetheless, because it does not account for the inflation from step one, the disease model determined by it is still not expected to be optimal. Section 4.5.4 demonstrates that without further adjustment in step five, the PMST p-value generally leads to better performance than the "nominal" p-value that also has no adjustment, but worse performance than the one with an adjustment.

It is important to use a proper stopping criterion in step three to achieve good disease model level estimation. However, there is probably no best or easy solution for it. Therefore, four commonly used criteria are applied to cover a different range of possibilities.

Note that determining a proper disease model among all possible ones could be a much more complicated problem. The essentially greedy sequential selection strategy used by SSP has largely simplified the problem. Without such a design, the search space could be much larger. For example, the total number of all possible two-level disease models is $2^9 - 2 = 510$, and that of three-level ones is $3^9 - 3 \times (2^9) + 3 \times 1^9 = 18150$. Consequently, by largely reducing the searching space, the proposed procedure is not guaranteed to find the correct disease model.

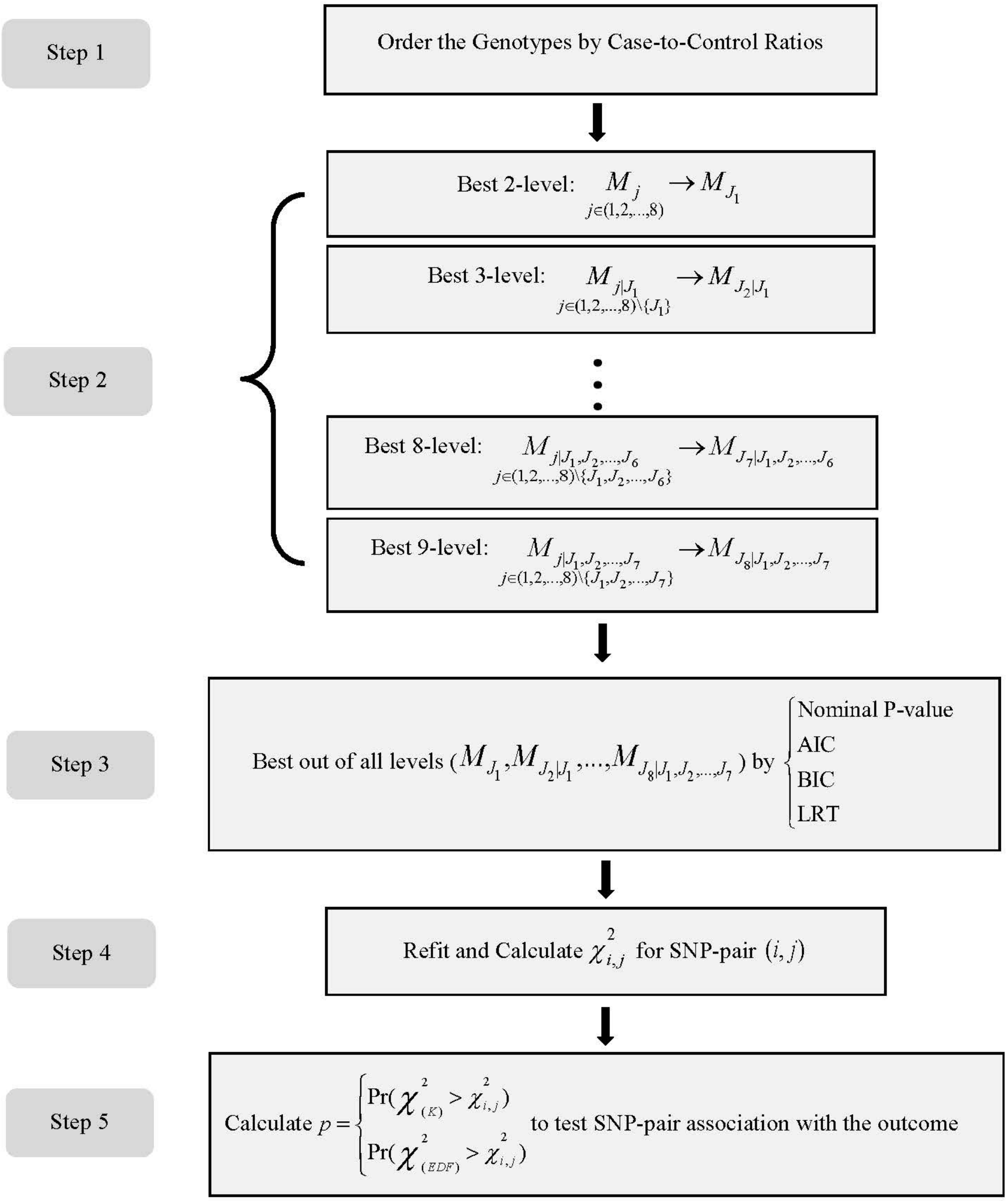

Step 1
Order the Genotypes by Case-to-Control Ratios
Step 2
Best 2-level: $M_{j} \rightarrow M_{J_1}$, $j \in (1,2,\ldots,8)$
Best 3-level: $M_{j|J_1} \rightarrow M_{J_2|J_1}$, $j \in (1,2,\ldots,8) \setminus \{J_1\}$
Best 8-level: $M_{j|J_1,J_2,\ldots,J_6} \rightarrow M_{J_7|J_1,J_2,\ldots,J_6}$, $j \in (1,2,\ldots,8) \setminus \{J_1,J_2,\ldots,J_6\}$
Best 9-level: $M_{j|J_1,J_2,\ldots,J_7} \rightarrow M_{J_8|J_1,J_2,\ldots,J_7}$, $j \in (1,2,\ldots,8) \setminus \{J_1,J_2,\ldots,J_7\}$
Step 3
Best out of all levels $(M_{J_1}, M_{J_2|J_1}, \ldots, M_{J_8|J_1,J_2,\ldots,J_7})$ by
Nominal P-value
AIC
BIC
LRT
Step 4
Refit and Calculate $\chi^2_{i,j}$ for SNP-pair (i,j)
Step 5
Calculate $p = \begin{cases} \Pr(\chi^2_{(K)} > \chi^2_{i,j}) \\ \Pr(\chi^2_{(EDF)} > \chi^2_{i,j}) \end{cases}$ to test SNP-pair association with the outcome

Fig 4.1. Summary steps for the SSP/SMP method. In the chart the disease models are represented by M, with subscripts denoting the index of places for the split or merge to differentiate different models. Subscripts with | in them denote both the current and already selected indexes, e.g., $M_{J_3|J_1,J_2}$ means that the current split or merge is at J_3 and it is built upon two previous splits or merge at J_1 and J_2.

4.4 Theoretical Comparison and Simulation Study

4.4.1 Equivalence of FSD-PST and SSP

In Section 4.2, a method that applies forward selection on $(I-P)y$ and $(I-P)X_{new1}$ (refer to Section 4.2.1 for the definition of P and X_{new1}) is proposed to determine a disease model. In Section 4.3 a sequential ratio split procedure is presented for the same purpose. Although the two methods are motivated from different ideas, i.e., the former from fused lasso under a regression model framework and the latter from the RS method, it is found that they are actually equivalent under minor conditions. Such a finding is helpful to understand the disease model searching methods better, especially for the different steps that contribute to the final disease model selection and SNP-pair ranking. In terms of equivalence, the following result is obtained.

Theorem 4.4.1. *The procedure of forward selection on the transformed difference variables defined in Section 4.2 is equivalent to the sequential ratio split procedure introduced in Section 4.3, if treating the outcome variable y as continuous. The two procedures group genotypes in exactly the same way and order, and the corrected p-values from the post-model selection test as reviewed in Section 4.2.2 would also be the same for them.*

Proof. The proof is done in a few steps. The gist of the proof is given below while details are given in Appendix C.1.

1. Applying forward selection on the transformed variables of $(I-P)y$ and $(I-P)X_{new1}$ is equivalent applying it on y and X_{new1} with an intercept included at each step.

2. The coefficient difference of two newly split groups in SSP is equivalent to the coefficient of the transformed variable representing the difference of split variables in FSD-PST. In more detail, for the first step, $\beta_{g^1_{1_j}} - \beta_{g^1_{2_j}} = \gamma_1$ in

$$\log\frac{p(y=1)}{p(y=0)} = \gamma_0 + \gamma_1\mathbb{X}_1, \tag{4.19}$$

where

$$\mathbb{X}_1 = \frac{1}{2}\sum_{l=1}^{j} X_l - \frac{1}{2}\sum_{l=j+1}^{9} X_l = X(\frac{1}{2}\cdots\frac{1}{2} \quad -\frac{1}{2}\cdots-\frac{1}{2})^T$$

and generally for the kth step, $\beta_{g^k_{1_j}} - \beta_{g^k_{2_j}} = \gamma_k$ in

$$\log\frac{p(y=1)}{p(y=0)} = \gamma_0 + \sum_{i=1}^{k} \gamma_i \mathbb{X}_i,$$

where

$$\mathbb{X}_i = \frac{1}{2}\sum_{l=j_i-1}^{j_i} X_l - \frac{1}{2}\sum_{l=j_i+1}^{9} X_l = X(0\cdots0\,\frac{1}{2}\cdots\frac{1}{2} - \frac{1}{2}\cdots-\frac{1}{2}\,0\cdots0)^T$$

for some $1 \leq j_1 < j_2 < \cdots < j_k \leq 9$. This result holds for both linear and logistic regression.

3. Use $U_j^1, j = 1, 2, \cdots, 8$ to denote the variables of X_{new1}, which are the first eight variables in $X\tilde{D}^{-1}$, and $V_j^1, j = 1, 2, \cdots, 8$ the difference of variables representing the potential split groups at the first step, i.e.,

$$U_j^1 = X\begin{bmatrix} -\frac{9-j}{9} & \\ \vdots & \\ -\frac{9-j}{9} & \\ \frac{j}{9} & \\ \vdots & \vdots \\ \frac{j}{9} & \end{bmatrix}\begin{matrix}\left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} j \text{ rows} \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} 9-j \text{ rows}\end{matrix} \qquad V_j^1 = X\begin{bmatrix} -\frac{1}{2} & \\ \vdots & \\ -\frac{1}{2} & \\ \frac{1}{2} & \\ \vdots & \vdots \\ \frac{1}{2} & \end{bmatrix}\begin{matrix}\left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} j \text{ rows} \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} 9-j \text{ rows}\end{matrix}$$

then the quantities assessed in the forward selection of these two groups of variables at step one can be shown to be equal, i.e.,

$$\frac{(U_j^1 - \bar{U_j^1})^T y}{\| (U_j^1 - \bar{U_j^1})^T \|_2} = \frac{(V_j^1 - \bar{V_j^1})^T y}{\| (V_j^1 - \bar{V_j^1})^T \|_2},$$

which indicate the first step of forward selection satisfies what is stated in the theorem.

4. At the kth step, the relevant quantities being assessed can also be shown to be equal, i.e.,

$$\frac{(U_j^1 - \bar{U}_j^1)^T P_A^{\perp} y}{\| (U_j^1 - \bar{U}_j^1)^T P_A^{\perp} \|_2} = \frac{(V_j^k - \bar{V}_j^k)^T P_A^{\perp} y}{\| (V_j^k - \bar{V}_j^k)^T P_A^{\perp} \|_2},$$

where

$$P_A^{\perp} = I - X_A (X_A^T X_A)^{-1} X_A,$$

and A indexes of already selected variables, V_j^k takes on the following form

$$X(0 \cdots 0 \; -\frac{1}{2} \; \cdots -\frac{1}{2} \; \frac{1}{2} \cdots \frac{1}{2} \; 0 \cdots 0)^T.$$

In fact, $(U_j^1 - \bar{U}_j^1)^T P_A^{\perp} = (V_j^k - \bar{V}_j^k)^T P_A^{\perp}$, which means both the parameter estimates and selection conditions are equivalent, so the corrected p-values from PMST would also be the same for SSP and FSD-PST.

□

4.4.2 Simulation Study

Simulation study results for the FSD-PST method are presented in this section, together with the unadjusted results (as oppose to adjusted results to be presented in Section 4.5) for the SSP method to serve as baseline comparison.

Selected Disease Model Level (DML)

The disease model levels are calculated for the true SNP-pairs that are determined by each of the stopping criteria. Table 4.2 provides the DMLs determined by the SSP and FSD-PST methods, which has also included the true DMLs as a reference; figure 4.2 presents them in bar charts for a clearer view; figure 4.3 presents a selection of them in one line chart for a comparison view. Comparing DMLs and model performance measures between different criteria of the same method could provide a direct perception of how the DMLs affect the SNP-pair detection performance. Additionally, a comparison of DMLs between the SSP and FSD-PST methods helps lend some insight into the usefulness of the post-model selection test for the problem.

Table 4.2. Actual v.s. determined final disease model levels for the true SNP-pairs that are detected by the FSD-PST and SSP methods.

DM	MAF	TRUE	SSP				FSD-PST			
			Nomi P	AIC	BIC	LRT	PMST P	AIC	BIC	FDR
T	0.1	2	3.018	3.673	2.800	2.611	2.250	2.125	2.070	2.185
T	0.4	2	2.62	3.125	2.171	2	2.307	2.687	2.344	2.702
MOD	0.1	2	2.899	3.258	2.484	2.206	2.130	2.153	2.027	2.208
MOD	0.4	2	2.613	3.211	2.199	2	2.223	2.585	2.111	2.650
DD	0.1	2	2.359	2.782	2.086	2	2.370	2.648	2.056	2.930
DD	0.4	2	2.350	3.143	2.155	2	2.153	2.295	2.055	2.444
XOR	0.1	2	2.605	2.997	2.293	2.028	2.264	2.224	2.093	2.262
XOR	0.4	2	2.571	3.153	2.175	2	2.312	2.712	2.239	2.623
ME	0.1	5	2.855	3.803	3.043	2	2.395	2.898	2.803	3.508
ME	0.4	5	3.375	4.523	3.213	2	2.643	3.726	3.443	3.972
MET	0.1	4	2.402	3.402	2.093	2	2.376	2.366	2.068	2.556
MET	0.4	4	3.292	4.01	2.905	2.006	2.673	3.738	3.422	3.739
DMN1	0.25	7	3.146	5.223	2.903	2.052	2.706	3.557	3.440	4.286
DMN2	0.25	8	2.922	3.344	2.583	2	2.425	2.797	2.477	2.831
DMN3	0.1	6	2.997	3.218	2.868	2	2.441	3.087	3.005	3.709
DMN4	0.1	8	3.065	3.423	3.005	2.020	2.495	3.128	3.093	3.768

Figure 4.2 shows that the DMLs selected by different stopping criteria of the SSP method are different, where the difference is observed to be quite consistent across all disease model examples. The AIC tends to select disease models with the highest levels, followed by the nominal p-value and BIC, and the LRT selects the lowest of all. In general, AIC and nominal p-value tend to select disease models of more than two levels, even if the true levels are only two; whereas the LRT tends to select two levels in the majority of cases, even if the true levels are multiple. Nonetheless, a trend for all criteria except LRT is observable that the determined DMLs for disease models with two levels are generally lower than those with

multiple levels. This indicates that the general epistasis detection methods are somewhat effective at determining a more accurate disease model than the TTTC-type methods.

DMLs selected by different criteria of the FSD-PST method are also different, but the differences are not as consistent as those for the SSP method. The FDR criterion selects the highest DMLs in most cases, followed by AIC (note the AIC and BIC criteria applied for the FSD-PST method are different from those for the SSP method, see Section 4.2.3 for details). Same as for SSP, all four criteria for FSD-PST tend to select higher DMLs for disease models with multiple levels than for those with just two levels, which proves their potential usefulness in the general epistasis detection. In particular, such differentiation is most noticeable for the FDR criteria (refer to Figure 4.3 for a clearer view of this point), which tends to suggest that it's relatively the best one among all. Additionally, the FSD-PST method appears to give better DMLs than SSP in approximating the true DMLs. In Figure 4.3, DMLs for two representative stopping criteria are selected and presented for both SSP and FSD-PST. It shows that FSD-PST tends to produce around the same levels of DMLs as SSP for two-level DMs (T, MOD, DD, XOR), and slightly higher level DMLs for the multi-level ones (MET, ME, DMN1-DMN4). In other words, the determined DMLs by FSD-PST give better differentiation of two- and multi-level DMs than SSP.

The PMST p-value for FSD-PST yields DMLs in between that of the minimum (LRT) and maximum (AIC) of the SSP method. In particular, the DMLs by the PMST p-values fluctuate around that of the BIC criteria for the SSP method. Generally, BIC is known to have the desirable property of being a consistent estimator that makes it select the true model asymptotically, but its consistency is built on the condition that the true model is contained among the candidate ones. In the case here, because the SSP is essentially greedy in nature, i.e., the candidate models are only a subset of the full space of models, there is no guarantee that the BIC would select the true model. When the PMST p-value selects close DMLs as BIC, an implication is that it may contain some desirable property worthy of further exploration. This property also applies to the other three criteria for the FSD-PST method, which are built upon the PMST p-values and have accounted for disease model selections in the process.

For all the general epistasis detection methods introduced in this chapter, the determined DMLs by them are still less than perfect in approximating the true ones. It is desirable to improve it more, which is possible by improving both the searching methods and the choices of stopping criteria. However, this is not pursued further in the current work. For the choice of searching methods, all the FSD-PST/SSP/SMP methods are greedy due to the consideration of a large-scale application. Foreseeably, any further improvement that helps for the accuracy (grouping of genotype combinations of a SNP-pair) would inevitably incur a higher computational burden. For the stopping criteria, the most popular ones from traditional model selection procedures (minimum p-value, AIC, BIC, LRT) have already been applied, and a recent one from PMST is also explored. For a complicated data set, the way to find the most appropriate model selection criteria calls for extensive trials of experiments, and it could also take some luck to find one, if any. Therefore the problem is deferred to future work and model performance improvement through other aspects such as a d.f. adjustment on the testing statistics is focused on in this work.

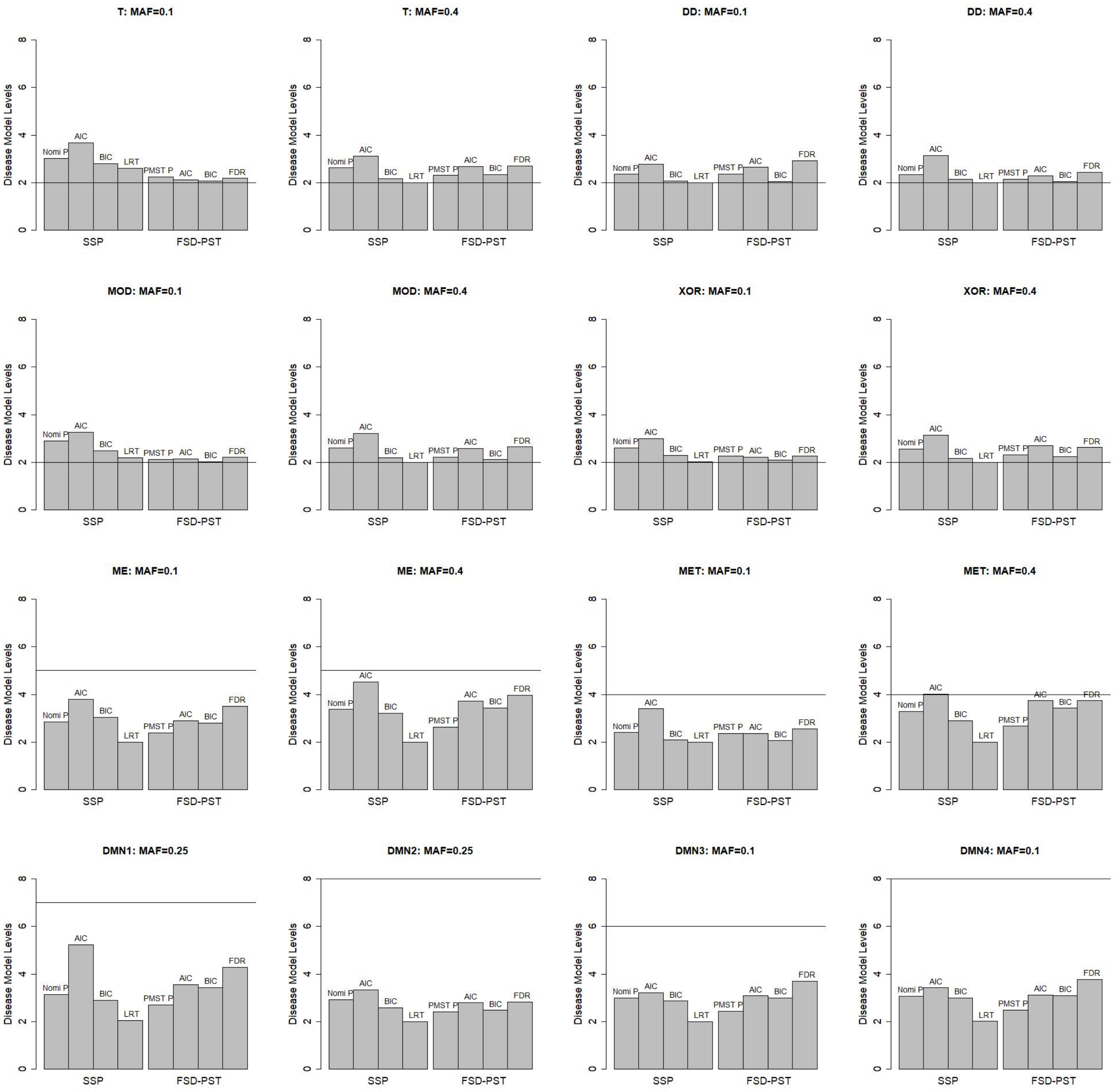


Fig 4.2. Simulation results on the estimated disease model levels by different stopping criteria of the SSP and FSD-PST methods. The values are calculated based on the selected true SNP-pairs. The horizontal line represents the true disease model level. Each simulation has been repeated for 400 times and the averages are reported.

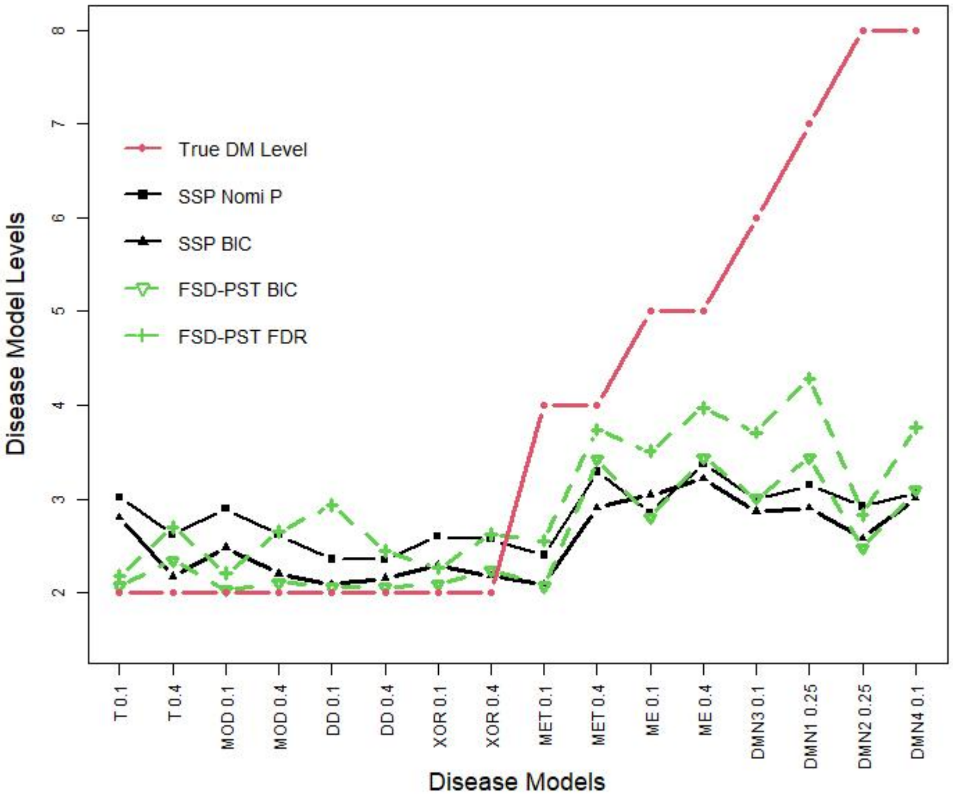


Fig 4.3. Determined disease model levels by SSP and FSD-PST for selected stopping criteria. Two different stopping criteria for each method are selected for a clear view. Different simulation disease models are presented along the X-axis, which are ordered by the true DMLs from low to high. It shows that the FSD-PST method tends to select the same level DMLs for two-level DMLs as SSP, and higher ones for multi-level DMs.

Model Performance

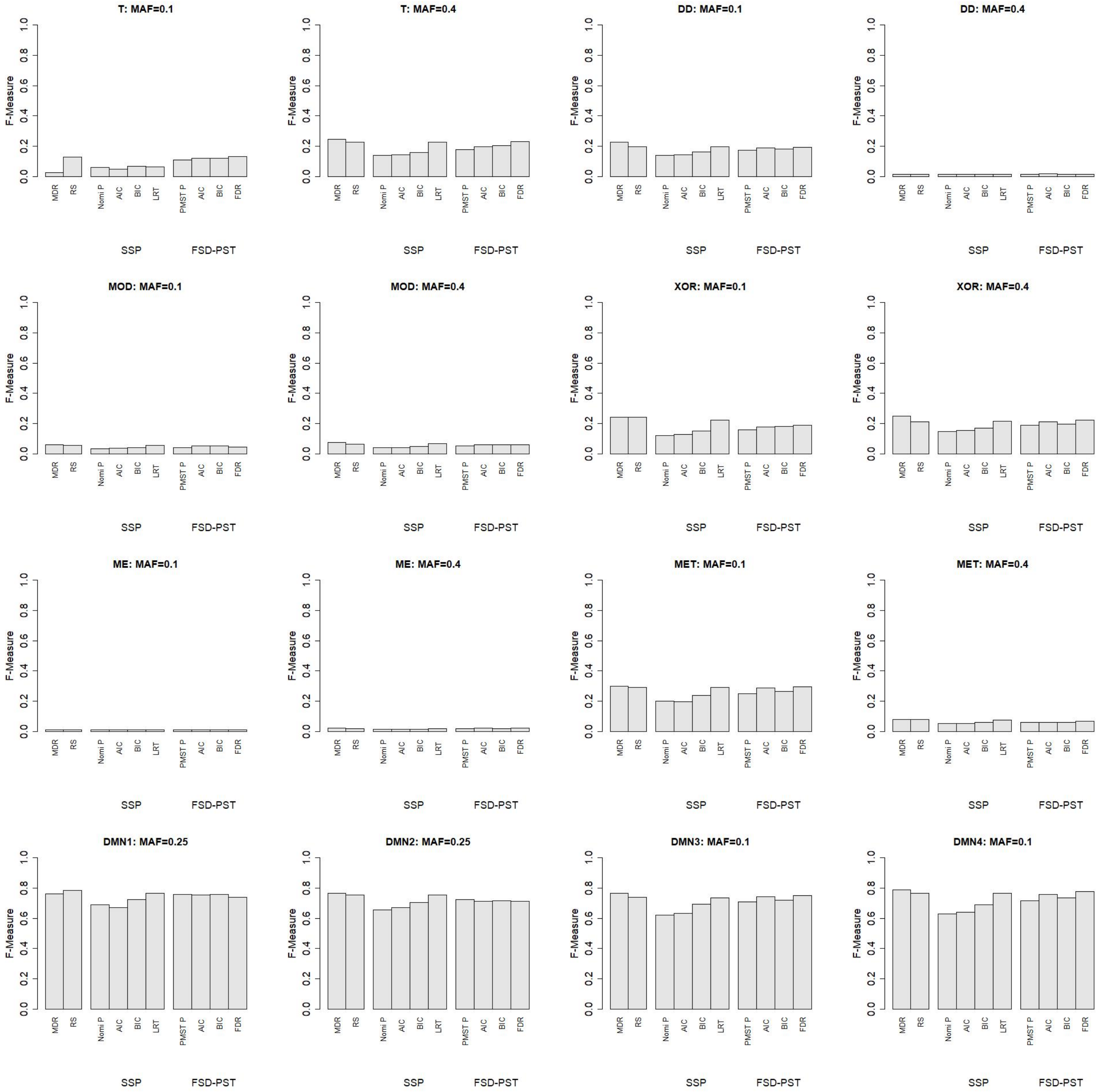


Fig 4.4. Simulation results on F-measure for the FSD-PST and SSP methods (based on unadjusted P-values). The results for MDR and RS are also based on unadjusted p-values. Each simulation has been repeated 400 times and the average performance is reported.

The F-measures are given in Figure 4.4. The results suggest that a good stopping criterion tends to improve SNP-pair detection. Among the explored stopping criteria, some appear to perform somewhat better than others. For FSD-PST, the best performing criteria is FDR, whereas for SSP the best one is LRT. Overall, the general epistasis detection methods of FSD-PST and SSP have comparable performance to MDR and RS under their best performing stopping criteria, but slightly worse performance under the other criteria.

The PMST is useful in improving the DMLs determination and the SNP-pair detection, but only to a certain degree. Notice the PMST p-value criterion for FSD-PST leads to a slightly better F-measures than the nominal p-value criteria for SSP, which indicates that stopping criteria with inflation accounted for is useful. However, as will be shown in Section 4.5.4, when compared to the EDF adjusted F-measure of SSP, the FSD-PST has significantly worse performance. A possible reason is that the criteria of PMST p-value only accounts for inflation incurred in the searching process of disease model, but not that from other causes. Additionally, the FSD-PST method and its stopping criteria contain some imperfect implementation that may be further improved. For instance, the PMST p-value is based on forward selection under a linear regression model framework, which may be refined using logistic regression that is better suited for a binary outcome. Additionally, the adopted FDR procedure is best suited for a sequence of indepedent p-values, which may be refined to better suit the PMST p-values that are dependent. In summary, the PMST is a promising direction to pursue for general epistasis detection, but there is a limit due to practical application constraints.

A trend of consistency between the F-measures and the determined DMLs is observable. For example, the FSD-PST method with its stopping criteria of BIC and FDR appears to give better DMLs than SSP with the nominal p-value and BIC criteria, where “better” is in the sense that the determined DMLs are closer to the true ones, i.e., there is a better differentiation of the two-level and multi-level disease models (see Figure 4.3). As a result, the corresponding F-measures mostly follow the same trend, i.e., the ones by FSD-PST are better than those by SSP, even though slightly. Notice the trend is true for all but the LRT criteria of the SSP method, which is discussed more in the paragraph below.

A comparison of F-measures between the SSP/FSD-PST and MDR/RS methods indicates that the use of simple disease models can still be quite competitive. The LRT criteria, for example, is the only one among all criteria for SSP that performs comparably to the MDR/RS method. It is noticable that LRT leads to simple disease models with DMLs of almost exactly two for all cases, which is the cloest to MDR and RS. The advantage of a simple two-level disease model can be perhaps understood as being a more strict rule that is able to filter out more false positives than the multi-level ones. This is seen from the comparison of recall and precision rates by SSP and MDR/RS. In the majority of simulation examples, the recall rates are close to 100% by all methods, partly because the signal is all inflated. Therefore the F-measure is mainly dependent on the precision. A simple two-level disease model appears to give higher precision than the multi-level ones. The possible reason is: a criterion that tends to select a multi-level disease model for the true SNP-pair also tends to select multi-level ones for the false pairs, which gives rise to the signal of the false pairs. In comparison, when MDR/RS restricts to use of two-level disease models for both the true and false SNP-pairs, the competitive signal from false pairs is also limited, which then results in an overall better selection of the true pairs.

Although the F-measures from LRT (for SSP) and FDR (for FSD-PST) are comparable, their determined DMLs are quite different, i.e., LRT select DMLs of two across all simulation examples, whereas FDR differentiates the two- and multi-level ones more evidently. The difference implies a possibility for improving the choice of stopping criteria. A comparison of the results from the nominal p-value for SSP and the PMST p-value for FSD-PST lends insight into a possible direction to improve. Because the SSP and FSD-PST methods are equivalent in the disease model searching process (under the condition that the outcome variable is treated as the same type by both methods, e.g., continuous or binary, refer to Section 4.4.1), the difference observed in the results can be attributed to the stopping criteria of p-values. The F-measure from the PMST p-value appears to be slightly better than that from the nominal p-value, which suggests that accounting for inflation in the criterion may be useful. Additionally, among all criteria for FSD-PST, the FDR has slightly better performance than the others. Because the criteria are all built on PMST p-values, this result suggests that the performance may be further improved if more appropriate inference methods are applied on the PMST p-values.

4.5 P-value Adjustment

For all methods introduced in previous sections, the refitted Pearson's χ^2-statistics in the last steps are inflated due to disease model searching in the processes. The PMST p-value (Section 4.2) only accounts for the inflation partially, so its usefulness is limited as evidenced by the results in Section 4.4.2. Inspired by the adjustment method for MDR and RS in Chapter 3, this section explores similar p-value adjustment for the SSP and SMP methods to rank the SNP-pairs more reasonably. It does not target any specific cause of the inflation but addresses it for the whole process, which is simple and efficient.

The idea is to run the general epistasis procedures through SNPs unrelated to the disease outcome to obtain the null distribution of the testing statistics, and calculate a more proper $p-$value than the nominal one based on the empirical distribution to rank the SNP-pairs. In practice, the testing statistics from the null data are observed to still follow the general shape of a χ^2-distribution, so they are assumed to still follow the chi-squared distribution but with an elevated degree of freedom. For a given SNP-pair (i, j), since the chi-squared distribution has its degree of freedom equal to its mean, similar effective degree of freedom (EDF) as given in Eq. 3.2 are obtained to calculate the adjusted p-value in Eq. 3.3.

4.5.1 Null Distribution Simulation

The same null data as used in Section 3.2 is adopted, i.e., a replicate of 100,000 SNP-pairs are simulated for different combinations of the two MAFs and sample size, and the disease response variable is randomly assigned to create a balanced case-control sample.

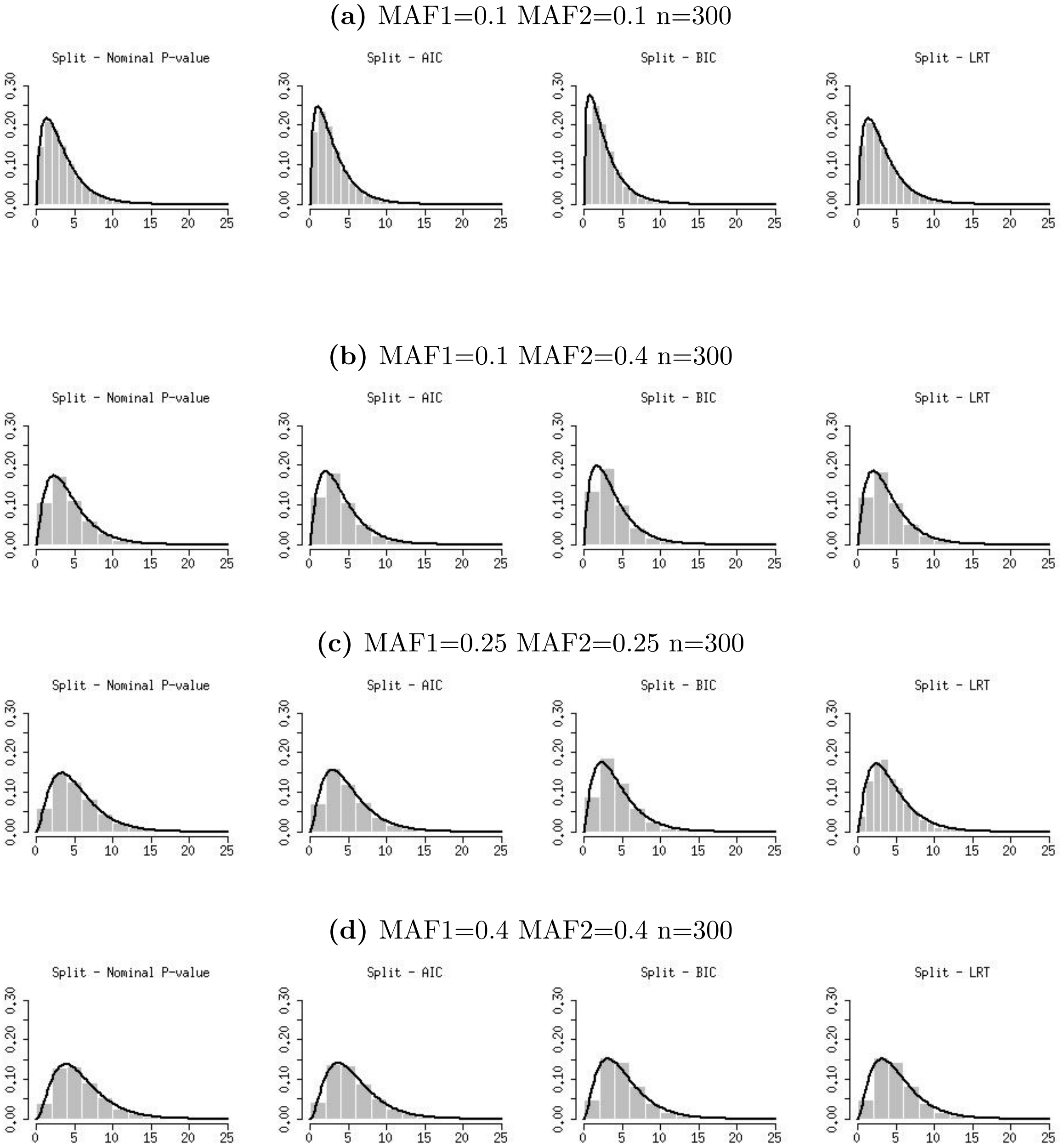
For each combination of MAF_1, MAF_2, sample size n, and a replicate of SNP-pair, the SSP and SMP are run, and the testing statistics (step four of Figure 4.1) are calculated. After that, the averages of χ^2-statistics across all SNP-pair replicates are calculated and retained as the empirical data points for the null distribution estimation. They are put as the response variable in a linear regression model to obtain smooth estimates of the EDF given inputs of MAFs and sample size.

4.5.2 Null Distribution Examples

Figure 4.5 and C.1 provide examples of the distribution for the testing statistics obtained from running the SSP and SMP on the null data. Different combinations of the MAFs and sample size are chosen for a relatively representative illustration. The curves are the density functions of a χ^2-distribution, where the d.f. is estimated to be the average of the empirical χ^2-statistics.

The figures show that the statistics generally follow the shape of a χ^2-distribution. The shapes are not evidently different between the split and merge procedures, but there are slight differences for the statistics obtained from different stopping criteria. In general, the nominal p-value criterion leads to a more dispersed distribution than the other criteria, and the statistics from it appear to align best with a χ^2-distribution. In comparison, the BIC results in the least dispersed distribution, and the statistics from it slightly deviate from the χ^2-distribution occasionally. Overall, it is reasonable to assume that the statistics still follow the χ^2-distribution with a shifted mean.

Figures 4.6 and 4.7 provide comparisons of the original and fitted EDF values against the sample size and MAF. These figures show that the fitted values are quite close to the original points, which is consistent with the model fitting statistics given in Table 4.20 in the next section.



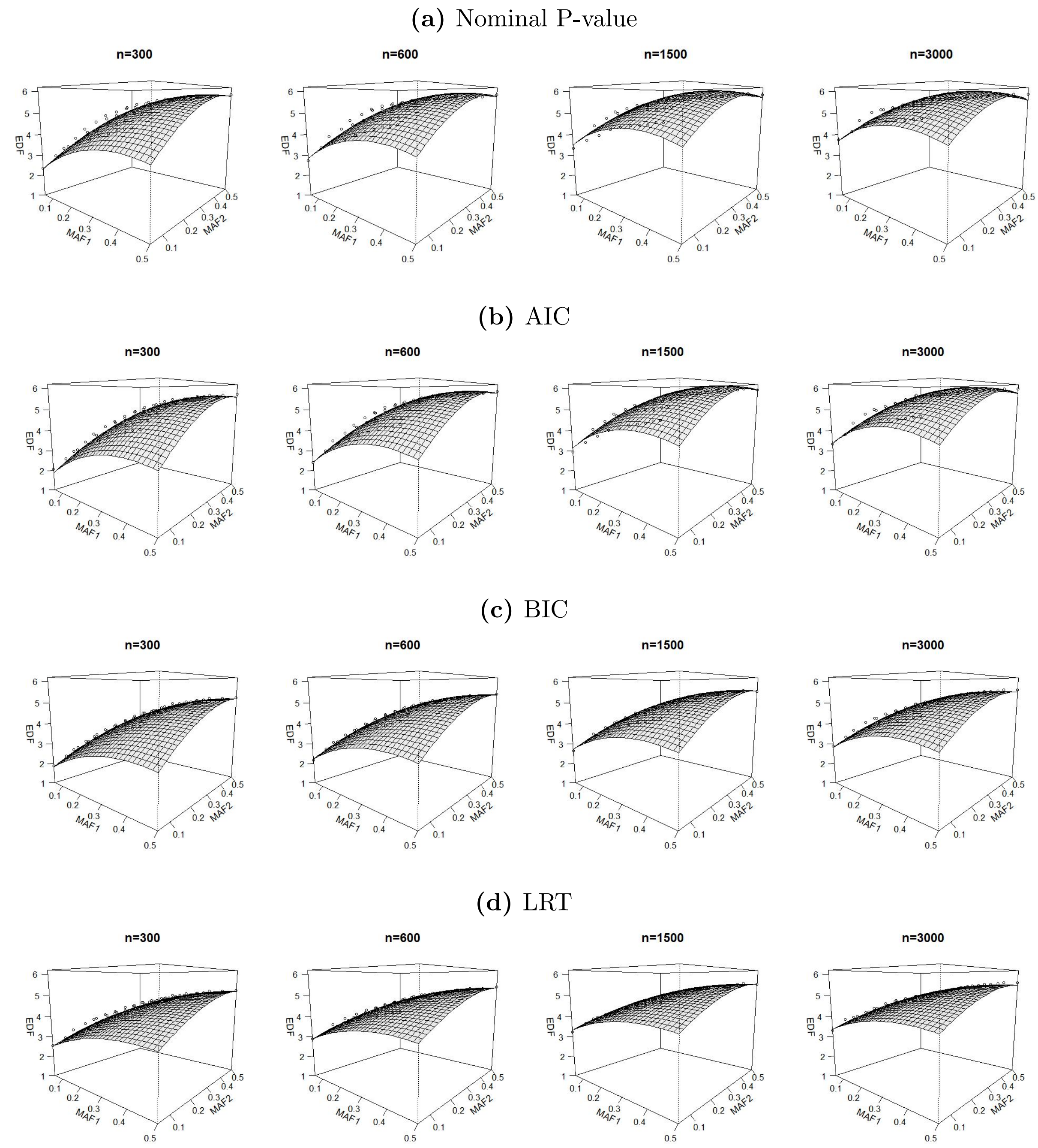
(a) MAF1=0.1 MAF2=0.1 n=300

(b) MAF1=0.1 MAF2=0.4 n=300

(c) MAF1=0.25 MAF2=0.25 n=300

(d) MAF1=0.4 MAF2=0.4 n=300

Fig 4.5. Histograms of $\{\widehat{\chi}_s^2 : s = 1, 2, ..., S\}$ versus the $\chi^2_{(\mathbf{EDF})}$ density functions. , EDF is computed by Eq. 3.2, for some specific combinations of $(\mathrm{MAF}_1, \mathrm{MAF}_2, n)$ and the four stopping criteria. While the $\chi^2_{(\mathrm{EDF})}$ density functions are not perfect fits of the underlying histograms, they are reasonable approximations as first-order corrections.

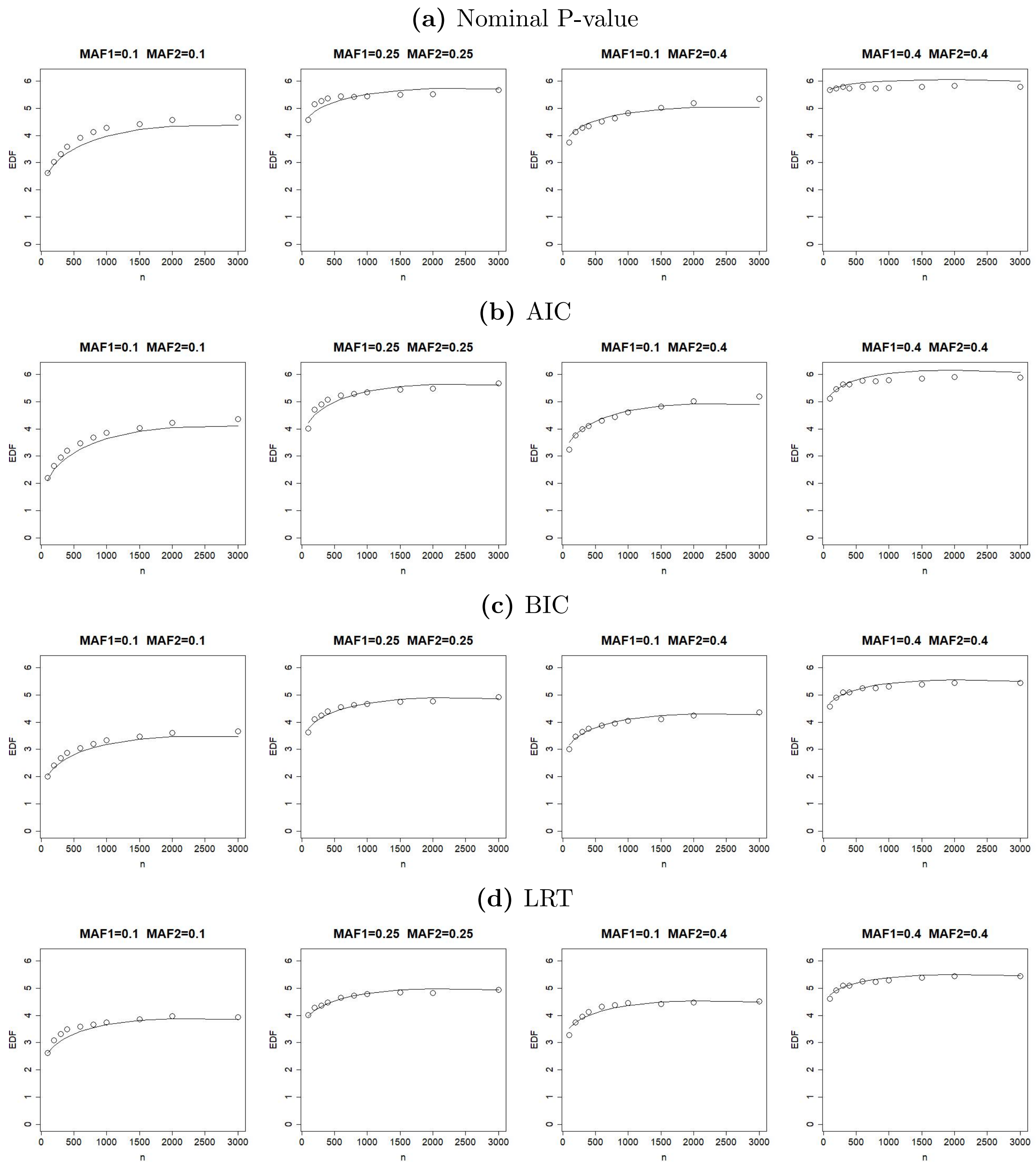

(a) Nominal P-value

(b) AIC

(c) BIC

(d) LRT

Fig 4.6. The estimated response surface model for SSP under different stopping criterion — EDF versus $(\text{MAF}_1, \text{MAF}_2)$ for $n = (300, 600, 1500, 3000)$.



(a) Nominal P-value

(b) AIC

(c) BIC

(d) LRT

Fig 4.7. The estimated response surface model for SSP under different stopping criterion — EDF versus n for $(\text{MAF}_1, \text{MAF}_2) = (0.1, 0.1), (0.25, 0.25), (0.1, 0.4)$ and $(0.4, 0.4)$.

4.5.3 Response Surface Model Results

Linear regression models are fit to obtain smooth estimates of the d.f.. The models take the same form as that for the MDR and RS p-value adjustment, except that they have one additional interaction term between the MAF and sample size to achieve better overall fitting (refer to Eq. 4.20). It is logically reasonable to do so because the multiple steps of disease model selection is a much more complicated process than the MDR and RS methods; therefore it is expected that the inflation is a more complicated function of the MAF and sample size.

The models are fit for the statistics obtained under each of the stopping criteria, and the parameter estimates are presented in Table 4.3. The p-values show that all the terms are significantly associated with the outcomes. The R^2 show that the model could explain a majority of the variance in the outcomes, so the model fitting is acceptable. For different stopping criteria, the model estimates have the same signs but different values. Compared to parameter estimates for the MDR and RS adjustment, the SSP and SMP have different signs and values for β_2 and β_3 , which is likely due to the additional interaction term between the MAF and sample size.

$$\begin{aligned}\text{EDF} \approx \beta_0 + &\underbrace{\beta_1(\text{MAF}_1) + \beta_1(\text{MAF}_2) + \beta_2(\sqrt{n})}_{\text{main effects}} + \\ &\underbrace{\beta_3(\text{MAF}_1)^2 + \beta_3(\text{MAF}_2)^2 + \beta_4(\sqrt{n})^2}_{\text{quadratic terms}} + \\ &\underbrace{\beta_5(\text{MAF}_1)(\text{MAF}_2) + \beta_6(\text{MAF}_1)(\sqrt{n}) + \beta_6(\text{MAF}_2)(\sqrt{n}) + \beta_7(\text{MAF}_1)(n) + \beta_7(\text{MAF}_2)(n)}_{\text{interactions}},\end{aligned} \tag{4.20}$$

Table 4.3. Parameter estimates for the sequential split procedures under the null distribution using the model equation in Eq. 4.20.

Stopping Criteria	Parameter	Estimate	Std	P-value	R^2	Adjusted R^2
Nominal P-value	β_0	-0.556	0.125	1.11×10^{-05}	0.951	0.95
	β_1	12.141	0.277	4.08×10^{-180}		
	β_2	-13.495	0.310	3.41×10^{-179}		
	β_3	13.238	0.836	1.02×10^{-46}		
	β_4	-12.645	1.284	3.72×10^{-21}		
	β_5	3.692	0.391	1.08×10^{-19}		
	β_6	-12.943	1.418	1.36×10^{-18}		
	β_7	11.622	2.178	1.41×10^{-07}		
AIC	β_0	-1.038	0.117	1.29×10^{-17}	0.967	0.966
	β_1	11.786	0.259	4.90×10^{-187}		
	β_2	-13.839	0.29	2.08×10^{-196}		
	β_3	13.119	0.782	3.17×10^{-51}		
	β_4	-12.114	1.201	4.80×10^{-22}		
	β_5	4.033	0.366	1.19×10^{-25}		
	β_6	-7.580	1.326	1.79×10^{-08}		
	β_7	5.112	2.037	1.24×10^{-02}		
BIC	β_0	-0.433	0.070	1.34×10^{-09}	0.985	0.985
	β_1	9.145	0.155	4.91×10^{-238}		
	β_2	-10.759	0.173	1.18×10^{-248}		
	β_3	10.125	0.467	1.80×10^{-75}		
	β_4	-10.027	0.718	4.38×10^{-38}		
	β_5	4.485	0.219	1.19×10^{-69}		
	β_6	-5.000	0.793	5.88×10^{-10}		
	β_7	4.110	1.218	7.92×10^{-04}		
LRT	β_0	0.552	0.078	5.08×10^{-12}	0.971	0.971
	β_1	7.164	0.173	2.16×10^{-170}		
	β_2	-8.053	0.193	1.70×10^{-171}		
	β_3	9.500	0.520	2.14×10^{-58}		
	β_4	-9.809	0.800	1.05×10^{-30}		
	β_5	3.488	0.243	1.03×10^{-39}		
	β_6	-5.096	0.883	1.31×10^{-08}		
	β_7	4.816	1.356	4.18×10^{-04}		

After the parameters are estimated, estimates of the EDFs are obtained by plugging in the estimates to Eq. 4.20, i.e., for each SNP-pair (i, j) with minor allele frequencies $(\mathrm{MAF}_1, \mathrm{MAF}_2)$ and a given sample size, its adjusted p-value is computed by the following steps.

1. Run the SSP (or SMP) and identify a disease model for the SNP-pair using one of the four stopping criteria, $M_{i,j}$.
2. Form the cross table by refitting the identified disease model and compute the usual chi-squared statistic, $\widehat{\chi}^2_{i,j}$.
3. Compute the effective degree of freedom (EDF) by the Eq. 4.20 and the parameters in the corresponding Table 4.3.
4. The adjusted p-value is computed by using the EDF from step three and Eq. 3.3.

4.5.4 Simulation Study

Full simulation study results for the SSP method are presented in this section, together with part of the results from the MDR, RS and FSD-PST methods for comparison.

Two types of adjusted p-values are computed to rank the SNP-pairs for the SSP method, one uses the EDFs from the predictive surface model of Eq. 4.20 in Section 4.5, and the other uses EDFs from an empirical matching. For the latter, the EDF is obtained by matching the MAF_1, MAF_2, and n of the SNP-pair to the values in the simulated null data, where the MAF for each SNP is estimated based on the sample data and rounded to the nearest 5% to match the available points in the simulation setting.

Of special note, p-value by use of EDF for the FSD-PST method is also computed for its minimum p-value criterion, so as to add in more comparisons. The EDF for it is leveraged from the SSP method under the nominal p-value criteria. This is logical to do because the FSD-PST method bears much similarity to the SSP (refer to Section 4.4.1 for details).

The results focusing on the FSD-PST and SSP methods are given in Figures 4.8, 4.9 and 4.10. The results for the SMP method with comparison to the SSP are given in Figures C.2 and C.3 in Appendix C.2.3. Overall, the SSP and SMP methods have quite close results in all cases, both in terms of the d.f. of selected disease models and the F-measure. There is no obvious difference that would make one method more preferable than the other. Figure C.3 also includes a comparison of results between the use of EDFs from the predictive surface model and the empirical matching approach. It shows that there is also no observable difference between them. Based on these observations, the results presented in this section are focused on the FSD-PST and SSP methods with EDFs from the predictive surface model.

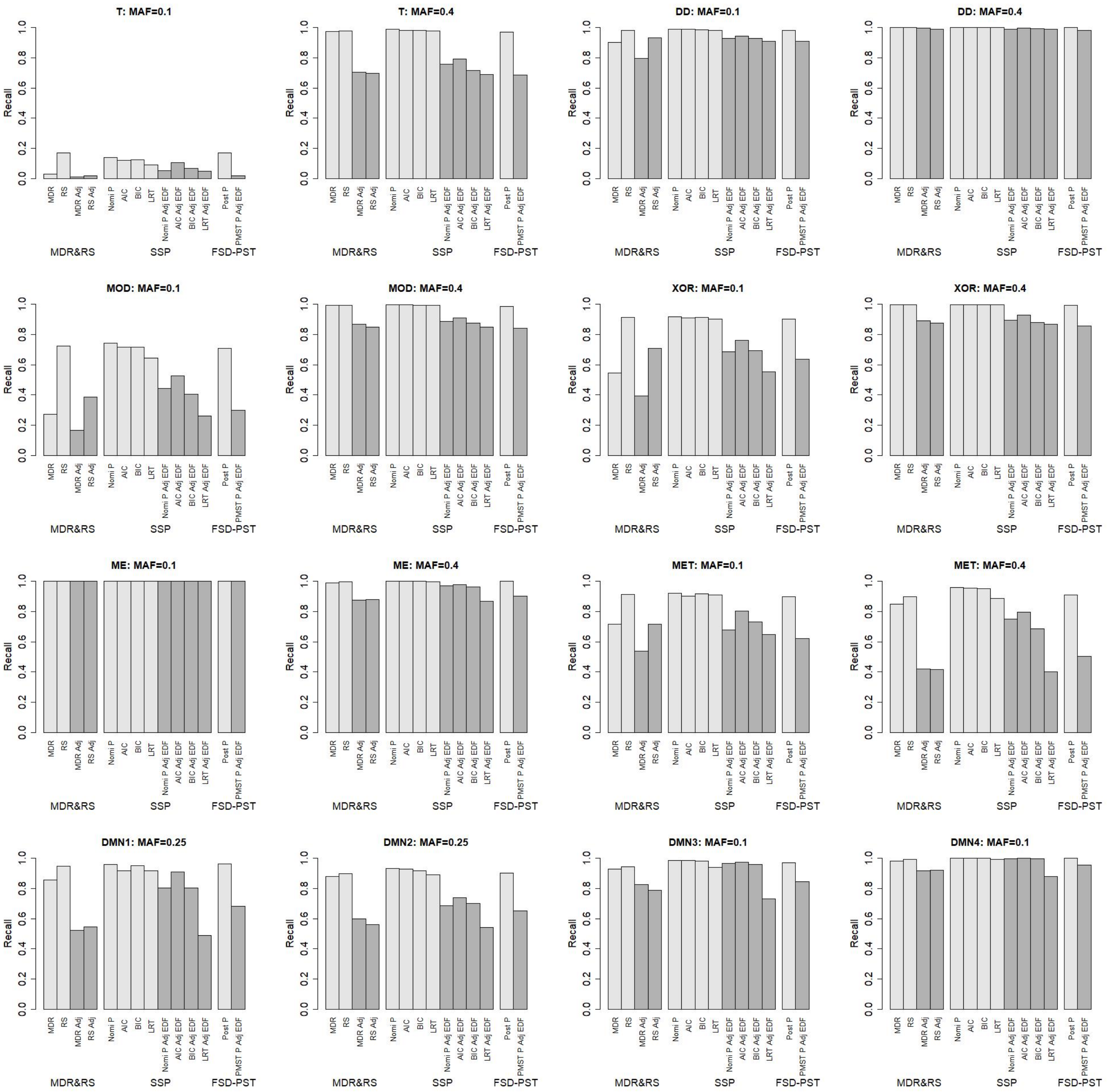


Fig 4.8. Simulation results on recall rates of true SNP-pairs. Each simulation has been repeated 400 times and the average performance is reported.

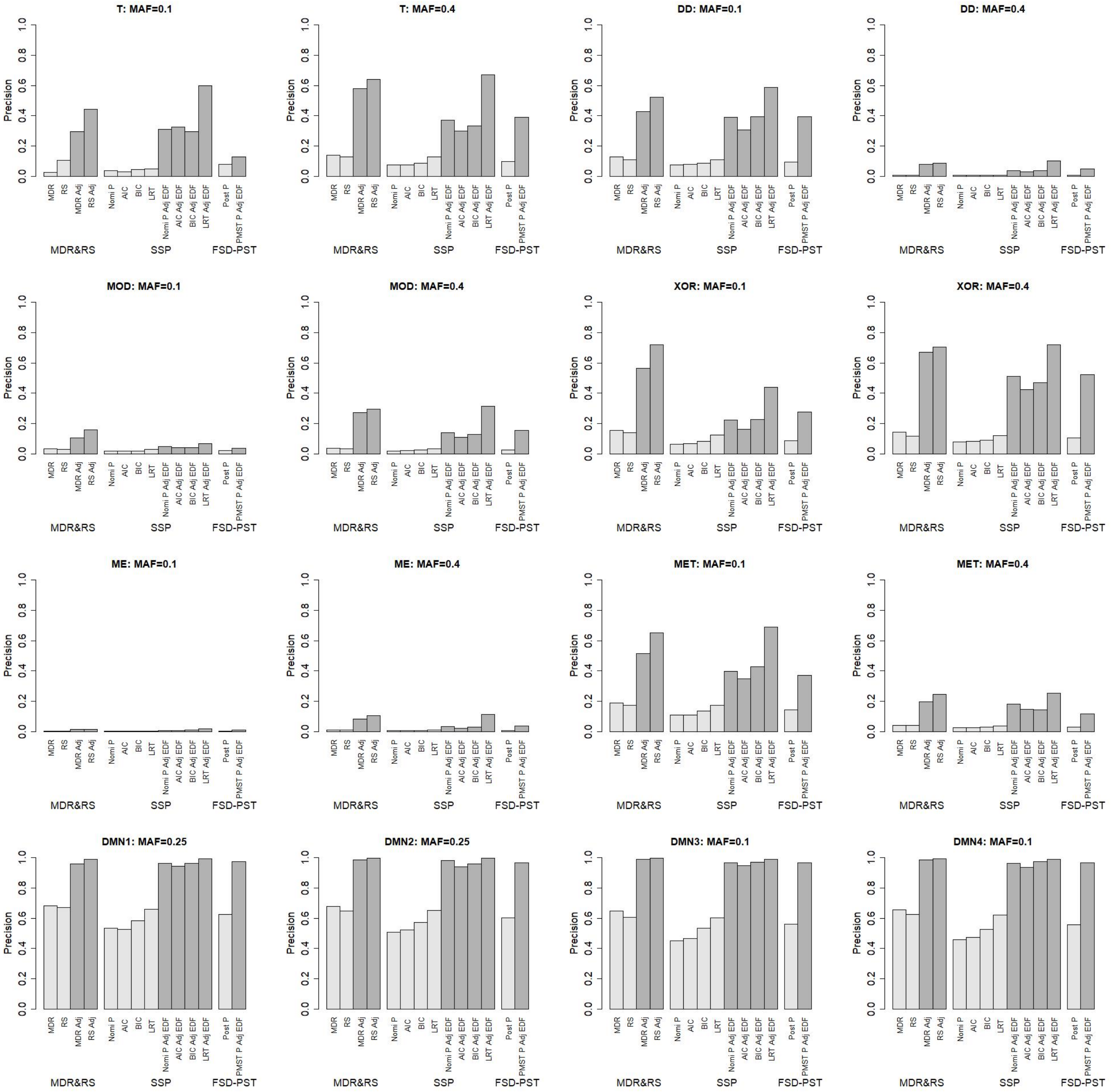


Fig 4.9. Simulation results on Precision of detected SNP-pairs. Each simulation has been repeated 400 times and the average performance is reported.

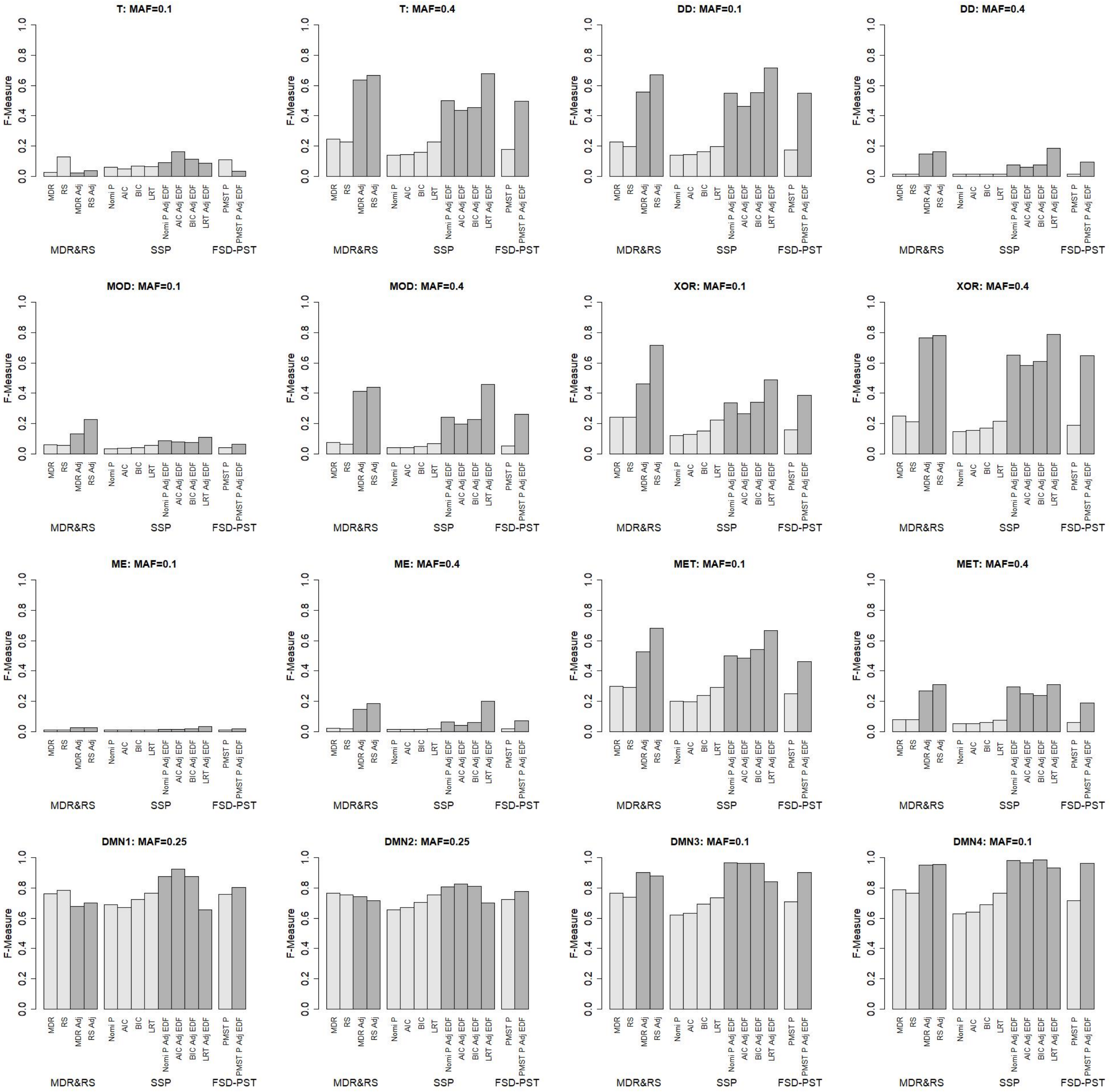


Fig 4.10. Simulation results on F-measure. Each simulation has been repeated 400 times and the average performance is reported.

The model performance in terms of detection power (recall rate), precision and overall efficiency are given in Figures 4.8, 4.9, and 4.10 respectively. The exact values are also given in Table C.2 in Appendix C.3. The results are presented for the SSP, FSD-PST (partially), MDR and RS methods for a comprehensive comparison.

The major findings are the following:

- Similar to MDR and RS, the simple first-order d.f. adjustment for the general epistasis detection methods is able to improve the SNP-pair detection (in terms of F-measure) significantly. This results from a significant increase in the precision after the EDF adjustment. In particular, the magnitude of increase in precision is usually larger than the decrease of recall, which leads to an overall increased F-measure.

- The multi-level disease model approach such as SSP has some advantage over the two-level ones such as MDR and RS. This can be seen from the result that the EDF adjusted F-measures for SSP are in most cases quite close to or even slightly better than that of MDR and RS. In particular, for the two-level DMs of T, MOD, DD, and XOR, as well as the multi-level DMs of ME and MET, the LRT criteria for SSP leads to comparable results as MDR and RS; whereas for the multi-level DMs of DMN 1 - 4, the criteria of nominal p-value, AIC, and BIC for SSP all lead to better F-measures than MDR and RS. In other words, the multi-level disease model approach has the potential to perform better than the two-level approaches when the true DM is multiple level, while still maintaining the chance of not losing too much when the true DM is only two-level.

- A good stopping criterion could have an elevated chance to improve the SNP-pair detection efficiency, but a universally best performing one could be challenging to be found. For instance, the LRT consistently gives the best performance among all criteria of SSP for the two-level DMs (T, MOD, DD, XOR) as well as for the multi-level ones of ME and MET, but its top performance is not maintained for the multi-level DMs of DMN 1 - 4. In contrast, the performance of the other three criteria for SSP are reversed. These results are mostly consistent with the disease model levels being chosen. For T, MOD, DD, and XOR whose true DMLs are two, the LRT criteria give DMLs of almost exactly the true level, so the performance of LRT is optimal. For multi-level DMs of DMN1 - DMN4, the nominal p-value, AIC and BIC criteria for SSP yield disease models of more than two levels, which is linked to better results than MDR and RS. Only the ME and MET are two exceptions from this rule, i.e., although the true DMLs are multiple, the best performance is achieved by the LRT criterion that uses DMLs of quite close to two.

- When an EDF adjustment is applied to the FSD-PST method for its PMST p-value criteria, the F-measure is improved significantly. Note the amount of EDF adjustment is approximate because it is not estimated directly from the FSD-PST method but from the SSP method with the nominal p-value criteria. This might be the reason that the result is slightly worse than that of the nominal p-value for SSP. Foreseeably, if a more proper EDF is applied, the result for FSD-PST may be further improved.

4.5.5 Results Discussion

As expected, the model detection recall rate is quite high before the EDF adjustment due to inflation. When the EDF adjustment is applied, the recall rate is brought down significantly, with varied decrease amounts across different simulation examples. In particular, there is relatively smaller decrease for the SSP/FSD-PST methods than the MDR and RS methods. Among different stopping criteria for the SSP method, the ranking of the recall rates (based on EDF-adjusted p-values) align closely with the estimated DMLs, i.e., the AIC leads to the largest recall rate, followed by the nominal p-value and BIC, and the LRT leads to the smallest recall rate. Because the higher the DMLs, the more complicated the disease model, the alignment between the recall rate and DMLs indicates that use of more complicated disease models lead to higher detection power, no matter if the true disease model is two or multiple levels.

For the majority of simulation examples, the SNP-pair detection precision is quite low ($< 10\%$) before the EDF adjustment. Application of the EDF adjustment leads to a significant increase in the precision, i.e., two to three folds in most cases. Ranking of the precision rates also show a consistent association pattern with the estimated DMLs, i.e, the higher the DMLs, the lower the precision. This indicates that using more complicated models is likely to lead to more noise selections than that of simpler ones.

Looking at the F-measure, no single stopping criterion performs universally better than others. There is a clear trade-off between recall and precision. The competitiveness of using the two-level DMs also illustrates this point, i.e., the advantage of a multi-level DM in its enhanced recall can be easily offset by its disadvantage of decreased precision, regardless of the adopted stopping criterion. Additionally, even though the post-model selection test provides a promising direction to pursue better disease model determination, a universally best performing one is not guaranteed to be found, or that there may be no such a criterion at all. This is because the performance of a criterion may be data dependent. For instance, a consistent estimator such as BIC may be preferable when the sample size is sufficiently large, but the LRT is known to have the highest recall among competitors on finite samples [112].

While the choice of stopping criteria for model performance enhancement is proven to be challenging and inconclusive, the EDF adjustment provides another solution that is practical and efficient. For all methods, the EDF adjustment is shown to improve model performance significantly. For the PMST p-value criteria specifically, it is worth noting that its performance is slightly improved over the nominal p-value of SSP, and EDF adjustment helps improve its performance further. In fact, as can be seen across the majority of the simulation examples, the magnitude of the latter improvement is much larger than the former, which suggests that EDF adjustment is more practically useful than the choice of the stopping rule.

4.6 Summary and Discussion

4.6.1 Methods Summary

Throughout this chapter, three methods have been proposed to do general epistasis detection, i.e., FSD-PST, SSP and SMP, which all use flexible ways to determine multi-level disease models to measure epistasis effects for a SNP-pair. The motivation of this proposal is based on the belief that a correct specification of the disease model is helpful for the true SNP-pair detection.

Similar to MDR and RS, all the methods in this chapter contain two major parts, i.e., determination of disease model and test of SNP-pair association with the outcome. While the focus of the methods is on the first part that is more complicated than MDR and RS, the second part is similar, i.e., SNP-pair association is tested by refitting the determined disease models, and calculating the Pearson's χ^2-statistic.

Due to ordering of the case-to-control ratios at the beginning of these methods, and disease model searching, inflation is introduced and the testing statistics do not follow the χ^2-distribution with the nominal d.f. anymore. Hence, a simple and efficient solution is applied to estimate a more reasonable distribution for the testing statistics.

Inspiration for the FSD-PST method stems from fused lasso. A simple linear transformation of the explanatory and outcome variables can turn the fused lasso into the ordinary lasso, which motivates the use of forward selection on the transformed variables directly to achieve multi-level disease model selection. A nice aspect of the chosen procedure is that it is actually equivalent to the SSP method under minor conditions. Further, the recent post-model selection tests that can account for model selection inflations are explored for their potential usefulness in the general epistasis detection problem.

The SSP method is a direct extension from the RS method, i.e., it makes succesive splits of the genotypes till every genotype is in a separate group. The algorithm is simple and straightforward, which makes it applicable to SNP-pairs at the GWAS level. What makes it more complicated than RS is that it not only needs to determine where to split at each step, but also which the "best" disease model is from all steps.

The disease model searching process for SSP turns out to be exactly the same as that for FSD-PST, if the outcome variable is treated as the same type (e.g., continuous) by both methods. The stopping criteria for FSD-PST are built on PMST p-values with the model selections accounted for, which differ from the stopping criteria for SSP. The difference is helpful to provide some comparison on the choice of criteria for disease model level determination and the subsequent impact on the SNP-pair identification.

4.6.2 Discussion

Based on the formulation for the general epistasis detection problem, the SNP-pair detection efficiency depends on the searching algorithm for the disease model, the selection of stopping criteria for the DMLs determination, and also the EDF adjustment for the testing statistics.

For the searching algorithm, all the FSD-PST, SSP and SMP methods are greedy in nature. This choice is made because both the number of multi-level disease models and SNP-pairs are huge. As such, exhaustive testing is impossible, and any other potentially more powerful algorithm is likely to require significantly larger computational time than the greedy ones. The forward selection algorithm used in FSD-PST is popular for being simple and effective. The idea of the SSP and SMP is also simple and straightforward, and the computational feasibility for them to be applicable at large scale makes them rather appealing. Although derived from different framework, the FSD-PST and SSP methods are shown to be equivalent, which in a way indicates that the choices of applicable greedy algorithms are somewhat limited, and that different greedy algorithms may not produce significantly different results. Overall, the adopted disease model searching algorithms have been chosen carefully out of practical and efficiency considerations, so the space for improvement is expected to be limited.

For the stopping criteria used for SSP and SMP, the commonly used model selection criteria of nominal p-value, AIC, BIC and LRT have been explored. The simulation results show that they behave differently in terms of the selected disease model complexity. The LRT selects the most parsimonious model throughout, and does not show a clear sign of determining a multi-level disease model; whereas the AIC consistantly selects disease models of the highest DMLs. The performance of LRT and AIC are reversed on the disease models with main effects (T, DD, MOD, DD, ME, MET) and disease models without main effects (DMN 1-4). DMN 1-4 are all multi-level risk DMs, so it is easy to understand that AIC performs better on them than the LRT. This is typical result with a tradeoff between recall and precision, which is not expected to be easily improved further.

In terms of disease model complexity, i.e., selected disease model levels, some of the stopping criteria for FSD-PST demonstrate a better sign of differentiation between the two- and multi-level disease models than those for SSP. For example, the BIC and FDR criteria for FSD-PST estimate higher DMLs for the multi-level DMs than the two-level ones, where the differences are noticeably higher than that of the nominal p-value and BIC criteria for SSP (refer to Figure 4.3). This is a desirable property given the study assumption that a more accurate estimate of disease model would result in better SNP-pair detection. Nonetheless, when compared to the (EDF adjusted) F-measures of SSP, the FSD-PST method's performance is significantly worse. Implications from the results are two folds. On the one hand, the better differentiation of two- and multi-level DMs by FSD-PST indicates that the stopping criteria for it are indeed better than those of SSP. Hence a further exploration of the PMST is worth pursuing for better SNP-pair detection. In fact, given the current less-than-perfect implementation, there is known potential to improve the results for FSD-PST further. On the other hand, the gain from better stopping criteria might be quite limited or even not guaranteed.

EDF adjustment for the testing statistics has been shown to be quite effective in increasing the SNP-pair detection precision, which leads to increase in the F-measure. Because the SSP method is approximately the same as FSD-PST, a comparison on the nominal and PMST p-values can directly reveal the effect of accounting for inflation in the stopping criteria. The results show that there is indeed some observable improvement in the use of PMST p-values over the nominal one. However, the improvement is not as phenomenal as the use of EDF adjustment, which suggests that pursuit of the EDF adjustment method is much more rewarding. In practice, in addition to the χ^2-distribution assumption for the testing statistics, the two-parameter *Gamma*-distribution has also been explored. However, use of the more complex distribution did not seem to help significantly.

Looking at the EDF adjusted results, the multi-level disease model methods do not exhibit a definite better performance than the two-level methods of MDR and RS. The results suggest that the use of simple disease models can still be quite appealing. The general epistasis detection methods proposed in this chapter are able to determine flexible disease model levels, and exhibit observably better performance than MDR and RS in some of the multi-level disease model examples (e.g., DMN 1-4), therefore they have complementary values to the two-level disease model methods.

Chapter 5

Discussion and Future Work

5.1 Summary

In this thesis, effort has been dedicated to the detection of SNP-pairs with epistasis effects, which are modeled by two types of approaches, i.e., use of two-level and multi-level disease models. The success of epistasis detection depends on a good estimation of the disease model at its first step, which makes it a different problem from the typical variable selection task.

The prototype disease model approach (PTY) uses two-level DMs to capture epistasis. The idea of it stems naturally from clustering of all the two-locus disease models and selecting representatives to test. It overcomes the limitations observed in the existing approaches of MDR and RS which depend on the outcome data to determine the disease models. Because the DM determination for PTY does not depend on the outcome information, it is found to suffer less from false positive discoveries than MDR and RS. The simulation study shows that it can improve an overall SNP-pair detection measured by the F-measure than MDR and RS. The real-data application also confirms that it has complementary value to the existing MDR and RS methods in finding relevant SNPs through epistasis effects.

Due to the disease model selections in the MDR and RS methods, the association measure of Pearson's χ^2-statistics are observed to be inflated. To address this problem, a first-order p-value correction is proposed that tries to estimate a more appropriate distribution to the statistics. Based on the observation that the testing statistics still generally follow the shape of a χ^2-distribution, an assumption is made that the testing statistics still follow the χ^2-distribution but with a shifted mean. The simulation results show despite the simple correction, the overall SNP-pair detection measured by the F-measure can be significantly improved after the adjustment.

Both the PTY and first-order p-value adjustment methods use DMs of two-level risk, so multi-level DMs are incorporated in the proposed SSP, SMP and FSD-PST methods to capture more general epistasis. The multi-level DM approaches are more complicated than the two-level ones, especially for the step of disease model determination. Because the use of a multi-level DM requires an extra step of the level determination, different commonly used stopping criteria have been explored. The simulation studies show that a good stopping criterion may lead to better SNP-pair detection. The post-model selection test appears to define a better stopping criterion than the other ones, but it only leads to slightly better performance in SNP-pair detection. In comparison, a first-order p-value correction method similar to that applied for the MDR and RS methods has shown to provide significantly better results than the use of a good stopping criterion.

5.2 Future Work

The current effort on the epistasis detection as explored in this thesis has resulted in some promising results, but at the same time also shown the complexity and challenge in the problem. Therefore some future work is needed, examples of which are briefly discussed below.

Firstly, for the PTY method, a disadvantage is that there is bound to be a discrepancy between the prototype disease model used to test a SNP-pair and the true one. Therefore, it may suffer from a power loss for SNP-pair detection. Based on this understanding, it is interesting to reduce the bias of prototype disease models and at the same time maintain as much of its advantage as possible in the future research. To achieve this goal, the research direction could be to combine the idea of PTY and MDR/RS for the determination of a disease model for a SNP-pair, and/or use of "smarter" prototype disease model selections that are more tailored to a specific SNP-pair.

Secondly, from the study on the use of multi-level DMs to detect epistasis, the results support the assumption that a good stopping criterion leads to enhanced SNP-pair detection. However, no universally best performing one has been observed among the criteria used for SSP/SMP. Therefore, further exploration to find a better stopping criterion than the explored ones are needed. Results comparison between the SSP and FSD-PST methods shows that the stopping criteria for FSD-PST may approximate the true DMs better, and also lead to better SNP-pair detection than SSP before EDF adjustment is applied for the later. By conjecture, a stopping criterion that accounts for inflation due to disease model searching might be a good choice. For instance, PMST is a promising direction that falls into this category, and a more proper application of it may lead to better results for the problem. Aside from the stopping criteria, exploring other potentially more powerful or efficient algorithms for the disease model searching provides another research direction. Based on the rich collection of algorithms in the machine learning community, it might be possible to take advantage of them and study their potential to be applied at large scales.

Thirdly, the p-value adjustment method has been shown to be quite effective for the SNP-pair detection, which indicates the importance of proper distribution estimations for the testing statistics. The assumption that the testing statistics still follow the χ^2-distribution in the current approach is somewhat stringent, so the future work can be to relax it to further improve the distribution estimation. Additionally, it would be useful to derive the exact theoretical distribution of the testing statistics. Relevant endeavor on the theoretical derivation already exists for models that are similar to the RS method [103], so it is possible to extend the work to apply to the RS and SSP methods.

It is worth mentioning that all the epistasis detection methods explored in this thesis can be easily extended to the detection of SNP-environment interactions. For instance, the effect of a risk SNP may be different due to different environmental exposures defined by categories. Then such an effect can be identified by following similar steps to that used in this thesis. Pursued this way, an underlying assumption is that the SNP-environment interaction effect is in the form of a specific interaction pattern, which can be represented by some sort of "disease model" similar to the ones described in this thesis. Compared to the typical statistical test of an interaction term, assessing the SNP-environment interaction in this way can improve the risk factor detection and also provide useful insight into the exact effect of the risk factor on the disease outcomes.

Lastly, this study has limited the interaction effects to be between two SNPs, or two-way interactions in statistical terms. In reality, higher-order interactions still possibly exist, so it is imperative to develop methods that can target high-order interactions, specially at the genome-wide level. This is bound to cause much computational challenge, given that three-way interactions already require a number of tests in the magnitude of 10^{14}. It is possible to achieve the task through a multi-step approach, which includes filtering of SNPs at the earlier steps. Then how to design the filtering and testing to yield more efficient identifications remains to be a future research problem. Additionally, due to the large number of required tests, a multiple testing problem exists. As the same SNP is present in many different interaction terms, many of the tests can be highly correlated with each other. Therefore how to achieve the most efficient testing for a large scale of correlated terms in the form of interaction effects is yet to be found.

References

[1] G. M. Clarke, C. A. Anderson, F. H. Pettersson, L. R. Cardon, A. P. Morris, et al. "Basic statistical analysis in genetic case-control studies". In: *Nature protocols* 6.2 (2011), pp. 121–133.

[2] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, et al. "Finding the missing heritability of complex diseases". In: *Nature* 461.7265 (2009), pp. 747–753.

[3] E. E. Eichler, J. Flint, G. Gibson, A. Kong, S. M. Leal, et al. "Missing heritability and strategies for finding the underlying causes of complex disease". In: *Nature Reviews Genetics* 11.6 (2010), pp. 446–450.

[4] S. H. Lee, N. R. Wray, M. E. Goddard, and P. M. Visscher. "Estimating missing heritability for disease from genome-wide association studies". In: *The American Journal of Human Genetics* 88.3 (2011), pp. 294–305.

[5] P. Li, M. Guo, C. Wang, X. Liu, and Q. Zou. "An overview of SNP interactions in genome-wide association studies". In: *Briefings in functional genomics* 14.2 (2015), pp. 143–155.

[6] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, et al. "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer". In: *The American Journal of Human Genetics* 69.1 (2001), pp. 138–147.

[7] X. Wan, C. Yang, Q. Yang, H. Zhao, and W. Yu. "The complete compositional epistasis detection in genome-wide association studies". In: *BMC genetics* 14.1 (2013), p. 7.

[8] W. Bateson, E. Waunders, and R. C. Punnett. “Experimental studies in the physiology of heredity”. In: *Molecular and General Genetics MGG* 2.1 (1909), pp. 17–19.

[9] J. S. Bloom, I. M. Ehrenreich, W. T. Loo, T.-L. V. Lite, and L. Kruglyak. “Finding the sources of missing heritability in a yeast cross”. In: *Nature* 494.7436 (2013), pp. 234–237.

[10] F. W. Albert, S. Treusch, A. H. Shockley, J. S. Bloom, and L. Kruglyak. “Genetics of single-cell protein abundance variation in large yeast populations”. In: *Nature* 506.7489 (2014), pp. 494–497.

[11] M. B. Taylor and I. M. Ehrenreich. “Genetic interactions involving five or more genes contribute to a complex trait in yeast”. In: *PLoS Genet* 10.5 (2014), e1004324.

[12] S. Wilkening, G. Lin, E. S. Fritsch, M. M. Tekkedil, S. Anders, et al. “An evaluation of high-throughput approaches to QTL mapping in Saccharomyces cerevisiae”. In: *Genetics* 196.3 (2014), pp. 853–865.

[13] J. H. Moore. “The ubiquitous nature of epistasis in determining susceptibility to common human diseases”. In: *Human heredity* 56.1-3 (2003), pp. 73–82.

[14] R. A. Fisher. “XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.” In: *Transactions of the royal society of Edinburgh* 52.02 (1919), pp. 399–433.

[15] P. C. Phillips. “The language of gene interaction”. In: *Genetics* 149.3 (1998), pp. 1167–1171.

[16] P. C. Phillips. “Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems”. In: *Nature Reviews Genetics* 9.11 (2008), pp. 855–867.

[17] E. Suzuki and T. J. VanderWeele. “Compositional epistasis: An epidemiologic perspective”. In: *Epistasis*. Springer, 2015, pp. 197–216.

[18] H. J. Cordell. “Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans”. In: *Human molecular genetics* 11.20 (2002), pp. 2463–2468.

[19] X. Wang, R. C. Elston, and X. Zhu. “Statistical interaction in human genetics: how should we model it if we are looking for biological interaction?” In: *Nature Reviews Genetics* 12.1 (2011), pp. 74–74.

[20] J. M. Álvarez-Castro and Ö. Carlborg. “A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis”. In: *Genetics* 176.2 (2007), pp. 1151–1167.

[21] T. J. VanderWeele. “Empirical tests for compositional epistasis”. In: *Nature Reviews Genetics* 11.2 (2010), pp. 166–166.

[22] T. J. VanderWeele. “Epistatic interactions”. In: *Statistical Applications in Genetics and Molecular Biology* 9.1 (2010).

[23] T. J. VanderWeele and N. M. Laird. “Tests for Compositional Epistasis under Single Interaction-Parameter Models”. In: *Annals of human genetics* 75.1 (2011), pp. 146–156.

[24] M. Emily. “IndOR: a new statistical procedure to test for SNP–SNP epistasis in genome-wide association studies”. In: *Statistics in medicine* 31.21 (2012), pp. 2359–2373.

[25] T. F. Hansen. “Measuring gene interactions”. In: *Epistasis*. Springer, 2015, pp. 115–143.

[26] R. J. Neuman, J. P. Rice, and A. Chakravarti. “Two-Locus models of disease”. In: *Genetic epidemiology* 9.5 (1992), pp. 347–365.

[27] I. B. Hallgrímsdóttir and D. S. Yuster. “A complete classification of epistatic two-locus models”. In: *BMC genetics* 9.1 (2008), p. 17.

[28] M. Song and D. L. Nicolae. “Restricted parameter space models for testing gene-gene interaction”. In: *Genetic epidemiology* 33.5 (2009), p. 386.

[29] R. J. Urbanowicz, A. L. Granizo-Mackenzie, J. Kiralis, and J. H. Moore. “A classification and characterization of two-locus, pure, strict, epistatic models for simulation and detection”. In: *BioData mining, Submitted* (2014).

[30] W. Li and J. Reich. “A complete enumeration and classification of two-locus disease models”. In: *Human heredity* 50.6 (2000), pp. 334–349.

[31] D. A. Greenberg. “A simple method for testing two-locus models of inheritance”. In: *American journal of human genetics* 33.4 (1981), p. 519.

[32] L. Goldin and D. Weeks. “Two-locus models of disease: comparison of likelihood and nonparametric linkage methods.” In: *American journal of human genetics* 53.4 (1993), p. 908.

[33] C. M. Lewis. “Genetic association studies: design, analysis and interpretation”. In: *Briefings in bioinformatics* 3.2 (2002), pp. 146–153.

[34] C. Minelli, J. R. Thompson, K. R. Abrams, A. Thakkinstian, and J. Attia. “The choice of a genetic model in the meta-analysis of molecular association studies”. In: *International journal of epidemiology* 34.6 (2005), pp. 1319–1328.

[35] G. Lettre, C. Lange, and J. N. Hirschhorn. “Genetic model testing and statistical power in population-based association studies of quantitative traits”. In: *Genetic epidemiology* 31.4 (2007), pp. 358–362.

[36] J. H. Moore, L. W. Hahn, M. D. Ritchie, T. A. Thornton, and B. C. White. “Application of genetic algorithms to the discovery of complex models for simulation studies in human genetics”. In: *Proceedings of the Genetic and Evolutionary Computation Conference/GECCO. Genetic and Evolutionary Computation Conference.* Vol. 2002. NIH Public Access. 2002, p. 1150.

[37] R. J. Urbanowicz, J. Kiralis, N. A. Sinnott-Armstrong, T. Heberling, J. M. Fisher, et al. “GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures”. In: *BioData mining* 5.1 (2012), pp. 1–14.

[38] J. H. Moore and S. M. Williams. “New strategies for identifying gene-gene interactions in hypertension”. In: *Annals of medicine* 34.2 (2002), pp. 88–95.

[39] E. Martin, M. Ritchie, L Hahn, S Kang, and J. Moore. “A novel method to identify gene–gene effects in nuclear families: the MDR-PDT”. In: *Genetic epidemiology* 30.2 (2006), pp. 111–123.

[40] A. E. Ashley-Koch, H Mei, J Jaworski, D. Ma, M. Ritchie, et al. “An Analysis Paradigm for Investigating Multi-locus Effects in Complex Disease: Examination of Three GABAA Receptor Subunit Genes on 15q11-q13 as Risk Factors for Autistic Disorder.” In: *Annals of Human Genetics* 70.3 (2006), pp. 281–292.

[41] D. Ma, P. Whitehead, M. Menold, E. Martin, A. Ashley-Koch, et al. “Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism”. In: *The American Journal of Human Genetics* 77.3 (2005), pp. 377–388.

[42] I. H. Chan, T. F. Leung, N. L. Tang, C. Y. Li, Y. M. Sung, et al. “Gene-gene interactions for asthma and plasma total IgE concentration in Chinese children”. In: *Journal of allergy and clinical immunology* 117.1 (2006), pp. 127–133.

[43] A. M. Singh, P. E. Moore, J. E. Gern, R. F. Lemanske Jr, and T. V. Hartert. “Bronchiolitis to asthma: a review and call for studies of gene–virus interactions in asthma causation”. In: *American journal of respiratory and critical care medicine* 175.2 (2007), pp. 108–119.

[44] Y. Cho, M. Ritchie, J. Moore, J. Park, K.-U. Lee, et al. “Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus”. In: *Diabetologia* 47.3 (2004), pp. 549–554.

[45] A. S. Andrew, H. H. Nelson, K. T. Kelsey, J. H. Moore, A. C. Meng, et al. “Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility”. In: *Carcinogenesis* 27.5 (2006), pp. 1030–1037.

[46] M. Huang, C. P. Dinney, X. Lin, J. Lin, H. B. Grossman, et al. “High-order interactions among genetic variants in DNA base excision repair pathway genes and smoking in bladder cancer susceptibility”. In: *Cancer Epidemiology Biomarkers & Prevention* 16.1 (2007), pp. 84–91.

[47] J. Xu, J. Lowey, F. Wiklund, J. Sun, F. Lindmark, et al. “The interaction of four genes in the inflammation pathway significantly predicts prostate cancer risk”. In: *Cancer Epidemiology Biomarkers & Prevention* 14.11 (2005), pp. 2563–2568.

[48] S. Qin, X. Zhao, Y. Pan, J. Liu, G. Feng, et al. “An association study of the N-methyl-D-aspartate receptor NR1 subunit gene (GRIN1) and NR2B subunit gene (GRIN2B) in schizophrenia with universal DNA microarray”. In: *European Journal of Human Genetics* 13.7 (2005), pp. 807–814.

[49] D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, et al. “A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction”. In: *Genetic epidemiology* 31.4 (2007), pp. 306–315.

[50] C.-T. Tsai, L.-P. Lai, J.-L. Lin, F.-T. Chiang, J.-J. Hwang, et al. “Renin-angiotensin system gene polymorphisms and atrial fibrillation”. In: *Circulation* 109.13 (2004), pp. 1640–1646.

[51] N. Oki and A. Motsinger-Reif. “Multifactor dimensionality reduction as a filter-based approach for genome wide association studies”. In: *Frontiers in Genetics* 2 (2011), p. 80.

[52] M. D. Ritchie, L. W. Hahn, and J. H. Moore. “Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity”. In: *Genetic epidemiology* 24.2 (2003), pp. 150–157.

[53] A. A. Motsinger and M. D. Ritchie. “The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction”. In: *Genetic epidemiology* 30.6 (2006), pp. 546–555.

[54] K. A. Pattin, B. C. White, N. Barney, J. Gui, H. H. Nelson, et al. “A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction”. In: *Genetic epidemiology* 33.1 (2009), pp. 87–94.

[55] T. Cattaert, M. L. Calle, S. M. Dudek, J. M. Mahachie John, F. Van Lishout, et al. “Model-Based Multifactor Dimensionality Reduction for detecting epistasis in case–control data in the presence of noise”. In: *Annals of human genetics* 75.1 (2011), pp. 78–89.

[56] M. Xie, J. Li, and T. Jiang. “Detecting genome-wide epistases based on the clustering of relatively frequent items”. In: *Bioinformatics* 28.1 (2012), pp. 5–12.

[57] X. Guo, Y. Meng, N. Yu, and Y. Pan. “Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering”. In: *BMC bioinformatics* 15.1 (2014), p. 102.

[58] H. Gao, J. M. Granka, and M. W. Feldman. “On the classification of epistatic interactions”. In: *Genetics* 184.3 (2010), pp. 827–837.

[59] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. “Sparsity and smoothness via the fused lasso”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology* 67.1 (2005), pp. 91–108.

[60] L. Cheng and M. Zhu. “Compositional epistasis detection using a few prototype disease models”. In: *PloS one* 14.3 (2019), e0213236.

[61] H. Mei, D. Ma, A. Ashley-Koch, and E. R. Martin. “Extension of multifactor dimensionality reduction for identifying multilocus effects in the GAW14 simulated data”. In: *BMC genetics*. Vol. 6. 1. Springer. 2005, pp. 1–5.

[62] J. H. Moore, J. C. Gilbert, C.-T. Tsai, F.-T. Chiang, T. Holden, et al. “A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility”. In: *Journal of theoretical biology* 241.2 (2006), pp. 252–261.

[63] G. U. Yule. “On the methods of measuring association between two attributes”. In: *Journal of the Royal Statistical Society* (1912), pp. 579–652.

[64] A. Tenesa and C. S. Haley. “The heritability of human disease: estimation, uses and abuses”. In: *Nature Reviews Genetics* 14.2 (2013), pp. 139–149.

[65] A. Likas, N. Vlassis, and J. J. Verbeek. "The global k-means clustering algorithm". In: *Pattern recognition* 36.2 (2003), pp. 451–461.

[66] C. Dong, X. Chu, Y. Wang, Y. Wang, L. Jin, et al. "Exploration of gene–gene interaction effects using entropy-based methods". In: *European Journal of Human Genetics* 16.2 (2008), pp. 229–235.

[67] Y. Zhang and J. S. Liu. "Bayesian inference of epistatic interactions in case-control studies". In: *Nature genetics* 39.9 (2007), pp. 1167–1173.

[68] X. Wan, C. Yang, Q. Yang, H. Xue, X. Fan, et al. "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies". In: *The American Journal of Human Genetics* 87.3 (2010), pp. 325–340.

[69] M. Y. Park and T. Hastie. "Penalized logistic regression for detecting gene interactions". In: *Biostatistics* 9.1 (2008), pp. 30–50.

[70] J. Marchini, P. Donnelly, and L. R. Cardon. "Genome-wide strategies for detecting multiple loci that influence complex diseases". In: *Nature genetics* 37.4 (2005), pp. 413–417.

[71] W. T. C. C. Consortium et al. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". In: *Nature* 447.7145 (2007), p. 661.

[72] N. Craddock and P. Sklar. "Genetics of bipolar disorder". In: *The Lancet* 381.9878 (2013), pp. 1654–1662.

[73] M. Berk, F Kapczinski, A. C. Andreazza, O. Dean, F. Giorlando, et al. "Pathways underlying neuroprogression in bipolar disorder: focus on inflammation, oxidative stress and neurotrophic factors". In: *Neuroscience & biobehavioral reviews* 35.3 (2011), pp. 804–817.

[74] B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, et al. "The Ensembl gene annotation system". In: *Database* 2016 (2016).

[75] A. Z. Dayem Ullah, N. R. Lemoine, and C. Chelala. "SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update)". In: *Nucleic acids research* 40.W1 (2012), W65–W70.

[76] K.-S. Wang, X.-F. Liu, and N. Aragam. "A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder". In: *Schizophrenia research* 124.1-3 (2010), pp. 192–199.

[77] A Yosifova, T Mushiroda, M Kubo, A Takahashi, Y Kamatani, et al. "Genome-wide association study on bipolar disorder in the Bulgarian population". In: *Genes, Brain and Behavior* 10.7 (2011), pp. 789–797.

[78] R. H. Perlis, J. Huang, S. Purcell, M. Fava, A. J. Rush, et al. "Genome-wide association study of suicide attempts in mood disorder patients". In: *American Journal of Psychiatry* 167.12 (2010), pp. 1499–1507.

[79] V. L. Willour, F. Seifuddin, P. B. Mahon, D. Jancic, M. Pirooznia, et al. "A genome-wide association study of attempted suicide". In: *Molecular psychiatry* 17.4 (2012), pp. 433–444.

[80] B. F. G. Popescu, R. F. Bunyan, Y. Guo, J. E. Parisi, V. A. Lennon, et al. "Evidence of aquaporin involvement in human central pontine myelinolysis". In: *Acta Neuropathologica Communications* 1.1 (2013), pp. 1–12.

[81] B. Wang. "Molecular mechanism underlying sialic acid as an essential nutrient for brain development and cognition". In: *Advances in Nutrition* 3.3 (2012), 465S–472S.

[82] S. Ripke, B. M. Neale, A. Corvin, J. T. Walters, K.-H. Farh, et al. "Biological insights from 108 schizophrenia-associated genetic loci". In: *Nature* 511.7510 (2014), p. 421.

[83] S. Kathiresan, A. K. Manning, S. Demissie, R. B. D'agostino, A. Surti, et al. "A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study". In: *BMC medical genetics* 8.1 (2007), pp. 1–10.

[84] R. Sherva, Y. Tripodis, D. A. Bennett, L. B. Chibnik, P. K. Crane, et al. “Genome-wide association study of the rate of cognitive decline in Alzheimer’s disease”. In: *Alzheimer’s & Dementia* 10.1 (2014), pp. 45–52.

[85] L. W. Hu, E. M. Kawamoto, E. Brietzke, C. Scavone, and B. Lafer. “The role of Wnt signaling and its interaction with diverse mechanisms of cellular apoptosis in the pathophysiology of bipolar disorder”. In: *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 35.1 (2011), pp. 11–17.

[86] D. Rasmussen, B Ishizuka, M. Quigley, and S. Yen. “Effects of tyrosine and tryptophan ingestion on plasma catecholamine and 3, 4-dihydroxyphenylacetic acid concentrations”. In: *The Journal of Clinical Endocrinology & Metabolism* 57.4 (1983), pp. 760–763.

[87] S. E. Baranzini, R. Srinivasan, P. Khankhanian, D. T. Okuda, S. J. Nelson, et al. “Genetic variation influences glutamate concentrations in brains of patients with multiple sclerosis”. In: *Brain* 133.9 (2010), pp. 2603–2611.

[88] M. A. Ferreira, M. C. O’Donovan, Y. A. Meng, I. R. Jones, D. M. Ruderfer, et al. “Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder”. In: *Nature genetics* 40.9 (2008), pp. 1056–1058.

[89] C. O’Dushlaine, E. Kenny, E. Heron, G. Donohoe, M. Gill, et al. “Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility”. In: *Molecular psychiatry* 16.3 (2011), pp. 286–292.

[90] D. Curtis, A. E. Vine, A. McQuillin, N. J. Bass, A. Pereira, et al. “Case-case genome wide association analysis reveals markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes”. In: *Psychiatric genetics* 21.1 (2011), p. 1.

[91] E. N. Smith, C. S. Bloss, J. A. Badner, T. Barrett, P. L. Belmonte, et al. “Genome-wide association study of bipolar disorder in European American and African American individuals”. In: *Molecular psychiatry* 14.8 (2009), pp. 755–763.

[92] P Sklar, J. Smoller, J Fan, M. Ferreira, R. Perlis, et al. “Whole-genome association study of bipolar disorder”. In: *Molecular psychiatry* 13.6 (2008), pp. 558–569.

[93] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, et al. “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles”. In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

[94] G. O. Consortium. “Gene ontology consortium: going forward”. In: *Nucleic Acids Research* 43.D1 (2015), pp. D1049–D1056.

[95] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. “KEGG as a reference resource for gene and protein annotation”. In: *Nucleic Acids Research* 44.D1 (2016), pp. D457–D462.

[96] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, et al. “Pathway Commons, a web resource for biological pathway data”. In: *Nucleic acids research* 39.suppl_1 (2010), pp. D685–D690.

[97] J. Wang, D. Duncan, Z. Shi, and B. Zhang. “WEB-based gene set analysis toolkit (WebGestalt): update 2013”. In: *Nucleic acids research* 41.W1 (2013), W77–W83.

[98] W. Xu, S. Cohen-Woods, Q. Chen, A. Noor, J. Knight, et al. “Genome-wide association study of bipolar disorder in Canadian and UK populations corroborates disease loci including SYNE1 and CSMD1”. In: *BMC medical genetics* 15.1 (2014), pp. 1–13.

[99] Y He, Z. Yu, I Giegling, L Xie, A. Hartmann, et al. “Schizophrenia shows a unique metabolomics signature in plasma”. In: *Translational psychiatry* 2.8 (2012), e149–e149.

[100] K. Roy, J. C. Murtie, B. F. El-Khodor, N. Edgar, S. P. Sardi, et al. “Loss of erbB signaling in oligodendrocytes alters myelin and dopaminergic function, a potential mechanism for neuropsychiatric disorders”. In: *Proceedings of the National Academy of Sciences* 104.19 (2007), pp. 8131–8136.

[101] Z. Zhao, J. Xu, J. Chen, S. Kim, M. Reimers, et al. "Transcriptome sequencing and genome-wide association analyses reveal lysosomal function and actin cytoskeleton remodeling in schizophrenia and bipolar disorder". In: *Molecular psychiatry* 20.5 (2015), pp. 563–572.

[102] L. Cheng and M. Zhu. "First-Order Correction of Statistical Significance for Screening Two-Way Epistatic Interactions". In: *Epistasis: Methods and Protocols.* Ed. by K.-C. Wong. Springer, 2021, pp. 181–190.

[103] A.-L. Boulesteix. "Maximally selected chi-square statistics and binary splits of nominal variables". In: *Biometrical Journal* 48.5 (2006), pp. 838–848.

[104] U. Cherubini, S. Mulinacci, F. Gobbi, and S. Romagnoli. *Dynamic copula methods in finance.* John Wiley & Sons, 2011.

[105] W. A. Jensen. "Response surface methodology: process and product optimization using designed experiments". In: *Journal of Quality Technology* 49.2 (2017), p. 186.

[106] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[107] R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. "Exact post-selection inference for sequential regression procedures". In: *Journal of the American Statistical Association* 111.514 (2016), pp. 600–620.

[108] J. Taylor and R. Tibshirani. "Post-selection inference for-penalized likelihood models". In: *Canadian Journal of Statistics* 46.1 (2018), pp. 41–61.

[109] M. G. G'Sell, S. Wager, A. Chouldechova, and R. Tibshirani. "Sequential selection procedures and false discovery rate control". In: *Journal of the Royal Statistical Society: Series B: Statistical Methodology* (2016), pp. 423–444.

[110] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[111] A. Rényi. "On the theory of order statistics". In: *Acta Mathematica Academiae Scientiarum Hungarica* 4.3-4 (1953), pp. 191–231.

[112] D. L. Solomon. "A note on the non-equivalence of the Neyman-Pearson and generalized likelihood ratio tests for testing a simple null versus a simple alternative hypothesis". In: *The American Statistician* 29.2 (1975), pp. 101–102.

[113] B. Gisabella, V. Y. Bolshakov, and F. M. Benes. "Regulation of synaptic plasticity in a schizophrenia model". In: *Proceedings of the National Academy of Sciences* 102.37 (2005), pp. 13301–13306.

[114] J. Du, J. A. Quiroz, N. A. Gray, S. T. Szabo, C. A. Zarate Jr, et al. "Regulation of cellular plasticity and resilience by mood stabilizers: the role of AMPA receptor trafficking". In: *Dialogues in clinical neuroscience* 6.2 (2004), p. 143.

[115] X. Chen and B. Ganetzky. "A neuropeptide signaling pathway regulates synaptic growth in Drosophila". In: *Journal of Cell Biology* 196.4 (2012), pp. 529–543.

[116] A. J. Forstner, A Hofmann, A Maaser, S Sumer, S Khudayberdiev, et al. "Genome-wide analysis implicates microRNAs and their target genes in the development of bipolar disorder". In: *Translational psychiatry* 5.11 (2015), e678–e678.

[117] E. Brietzke, F. Kapczinski, R. Grassi-Oliveira, I. Grande, E. Vieta, et al. "Insulin dysfunction and allostatic load in bipolar disorder". In: *Expert review of neurotherapeutics* 11.7 (2011), pp. 1017–1028.

[118] D. A. Cousins, K. Butts, and A. H. Young. "The role of dopamine in bipolar disorder". In: *Bipolar disorders* 11.8 (2009), pp. 787–806.

[119] L. Lin, T. G. Lesnick, D. M. Maraganore, and O. Isacson. "Axon guidance and synaptic maintenance: preclinical markers for neurodegenerative disease and therapeutics". In: *Trends in neurosciences* 32.3 (2009), pp. 142–149.

[120] K. S. Cramer and I. J. Miko. "Eph-ephrin signaling in nervous system development". In: *F1000Research* 5 (2016).

[121] Y. Hara, M. Fukaya, K. Hayashi, T. Kawauchi, K. Nakajima, et al. "ADP ribosylation factor 6 regulates neuronal migration in the developing cerebral cortex through FIP3/arfophilin-1-dependent endosomal trafficking of N-cadherin". In: *ENeuro* 3.4 (2016).

[122] A. Gärtner, E. F. Fornasiero, and C. G. Dotti. "Cadherins as regulators of neuronal polarity". In: *Cell adhesion & migration* 9.3 (2015), pp. 175–182.

[123] A. Gärtner, E. F. Fornasiero, S. Munck, K. Vennekens, E. Seuntjens, et al. "N-cadherin specifies first asymmetry in developing neurons". In: *The EMBO journal* 31.8 (2012), pp. 1893–1903.

[124] M. Ide and D. A. Lewis. "Altered cortical CDC42 signaling pathways in schizophrenia: implications for dendritic spine deficits". In: *Biological psychiatry* 68.1 (2010), pp. 25–32.

[125] T. D. Gould and H. K. Manji. "The Wnt signaling pathway in bipolar disorder". In: *The Neuroscientist* 8.5 (2002), pp. 497–511.

[126] A. J. Valvezan and P. S. Klein. "GSK-3 and Wnt signaling in neurogenesis and bipolar disorder". In: *Frontiers in molecular neuroscience* 5 (2012), p. 1.

[127] C. Watkins, A Sawa, and M. Pomper. "Glia and immune cell signaling in bipolar disorder: insights from neuropharmacology and molecular imaging to clinical application". In: *Translational psychiatry* 4.1 (2014), e350–e350.

[128] H. K. Manji, J. A. Quiroz, J. L. Payne, J. Singh, B. P. Lopes, et al. "The underlying neurobiology of bipolar disorder". In: *World Psychiatry* 2.3 (2003), p. 136.

APPENDICES

Appendix A

All 143 Disease Models Used for Clustering

Table A.1. 143 Disease Models

1	0	0	0	13	0	0	0	25	0	0	1	37	0	0	1	49	0	0	0	61	0	0	1
	0	0	0		0	0	0		0	0	1		0	1	0		1	0	0		1	1	0
	0	0	1		1	0	1		1	1	1		1	1	1		0	1	1		0	0	1
2	0	0	0	14	0	0	0	26	0	0	0	38	0	0	1	50	0	0	0	62	0	0	1
	0	0	0		0	0	1		0	1	0		0	1	1		1	0	1		1	1	1
	0	1	0		1	0	0		1	0	0		1	1	0		0	1	0		0	0	0
3	0	0	0	15	0	0	0	27	0	0	0	39	0	0	1	51	0	0	0	63	0	0	1
	0	0	0		0	0	1		0	1	0		0	1	1		1	0	1		1	1	1
	0	1	1		1	0	1		1	0	1		1	1	1		0	1	1		0	0	1
4	0	0	0	16	0	0	1	28	0	0	0	40	0	0	0	52	0	0	1	64	0	0	0
	0	0	1		0	0	0		0	1	1		1	0	0		1	0	0		1	1	0
	0	1	0		1	0	0		1	0	0		0	0	0		0	1	0		0	1	0
5	0	0	0	17	0	0	1	29	0	0	0	41	0	0	0	53	0	0	1	65	0	0	0
	0	0	1		0	0	0		0	1	1		1	0	0		1	0	0		1	1	0

	0	1	1		1	0	1		1	0	1		0	0	1		0	1	1		0	1	1
6	0	0	0	18	0	0	0	30	0	0	1	42	0	0	0	54	0	0	1	66	0	0	0
	0	1	0		0	0	0		0	1	0		1	0	1		1	0	1		1	1	1
	0	0	0		1	1	0		1	0	0		0	0	0		0	1	0		0	1	0
7	0	0	0	19	0	0	0	31	0	0	1	43	0	0	0	55	0	0	1	67	0	0	0
	0	1	0		0	0	0		0	1	0		1	0	1		1	0	1		1	1	1
	0	0	1		1	1	1		1	0	1		0	0	1		0	1	1		0	1	1
8	0	0	0	20	0	0	0	32	0	0	0	44	0	0	1	56	0	0	0	68	0	0	1
	0	1	0		0	0	1		0	1	0		1	0	0		1	1	0		1	1	0
	0	1	0		1	1	0		1	1	0		0	0	0		0	0	0		0	1	0
9	0	0	0	21	0	0	0	33	0	0	0	45	0	0	1	57	0	0	0	69	0	0	1
	0	1	0		0	0	1		0	1	0		1	0	0		1	1	0		1	1	0
	0	1	1		1	1	1		1	1	1		0	0	1		0	0	1		0	1	1
10	0	0	0	22	0	0	1	34	0	0	0	46	0	0	1	58	0	0	0	70	0	0	1
	0	1	1		0	0	0		0	1	1		1	0	1		1	1	1		1	1	1
	0	1	0		1	1	0		1	1	0		0	0	0		0	0	0		0	1	0
11	0	0	0	23	0	0	1	35	0	0	0	47	0	0	1	59	0	0	0	71	0	0	1
	0	1	1		0	0	0		0	1	1		1	0	1		1	1	1		1	1	1
	0	1	1		1	1	1		1	1	1		0	0	1		0	0	1		0	1	1
12	0	0	0	24	0	0	1	36	0	0	1	48	0	0	0	60	0	0	1	72	0	1	0
	0	0	0		0	0	1		0	1	0		1	0	0		1	1	0		1	0	0
	1	0	0		1	1	0		1	1	0		0	1	0		0	0	0		0	0	0
73	0	1	0	85	0	0	0	97	0	0	1	109	0	0	0	121	0	1	1	133	0	1	0
	1	0	0		1	0	0		1	0	0		1	1	0		1	0	0		1	1	1
	0	0	1		1	0	1		1	1	1		1	1	1		1	0	1		1	0	1
74	0	1	0	86	0	0	0	98	0	0	1	110	0	0	0	122	0	1	0	134	0	1	1
	1	0	0		1	0	1		1	0	1		1	1	1		1	0	0		1	1	0
	0	1	0		1	0	0		1	1	0		1	1	0		1	1	0		1	0	0

75	0	1	0	87	0	0	0	99	0	0	1	111	0	0	0	123	0	1	0	135	0	1	1
	1	0	0		1	0	1		1	0	1		1	1	1		1	0	0		1	1	0
	0	1	1		1	0	1		1	1	1		1	1	1		1	1	1		1	0	1
76	0	1	0	88	0	0	1	100	0	0	0	112	0	0	1	124	0	1	0	136	0	1	0
	1	0	1		1	0	0		1	1	0		1	1	0		1	0	1		1	1	0
	0	1	0		1	0	0		1	0	0		1	1	0		1	1	0		1	1	0
77	0	1	0	89	0	0	1	101	0	0	0	113	0	0	1	125	0	1	0	137	0	1	0
	1	0	1		1	0	0		1	1	0		1	1	0		1	0	1		1	1	0
	0	1	1		1	0	1		1	0	1		1	1	1		1	1	1		1	1	1
78	0	1	0	90	0	0	1	102	0	0	0	114	0	0	1	126	0	1	1	138	0	1	0
	1	1	0		1	0	1		1	1	1		1	1	1		1	0	0		1	1	1
	0	0	0		1	0	0		1	0	0		1	1	0		1	1	0		1	1	0
79	0	1	0	91	0	0	1	103	0	0	0	115	0	0	1	127	0	1	1	139	0	1	0
	1	1	0		1	0	1		1	1	1		1	1	1		1	0	0		1	1	1
	0	0	1		1	0	1		1	0	1		1	1	1		1	1	1		1	1	1
80	0	1	0	92	0	0	0	104	0	0	1	116	0	1	0	128	0	1	1	140	0	1	1
	1	1	0		1	0	0		1	1	0		1	0	0		1	0	1		1	1	0
	0	1	0		1	1	0		1	0	0		1	0	0		1	1	0		1	1	0
81	0	1	0	93	0	0	0	105	0	0	1	117	0	1	0	129	0	1	1	141	0	1	1
	1	1	0		1	0	0		1	1	0		1	0	0		1	0	1		1	1	0
	0	1	1		1	1	1		1	0	1		1	0	1		1	1	1		1	1	1
82	0	1	0	94	0	0	0	106	0	0	1	118	0	1	0	130	0	1	0	142	0	1	1
	1	1	1		1	0	1		1	1	1		1	0	1		1	1	0		1	1	1
	0	1	0		1	1	0		1	0	0		1	0	0		1	0	0		1	1	0
83	0	1	0	95	0	0	0	107	0	0	1	119	0	1	0	131	0	1	0	143	0	1	1
	1	1	1		1	0	1		1	1	1		1	0	1		1	1	0		1	1	1
	0	1	1		1	1	1		1	0	1		1	0	1		1	0	1		1	1	1
84	0	0	0	96	0	0	1	108	0	0	0	120	0	1	1	132	0	1	0				

	1 0 0		1 0 0		1 1 0		1 0 0		1 1 1		
	1 0 0		1 1 0		1 1 0		1 0 0		1 0 0		

Appendix B

Additional Material for Chapter 2

B.1 GSEA results from KEGG for genes identified by MDR and RS

For comparison, Table B.1 lists the statistically enriched pathways from KEGG (multiple-testing adjusted p-value < 0.05) for genes identified by MDR and Table B.2, for genes identified by RS. Note that the WebGestalt tool, which we used to perform GSEA (see Section 5.3 of main text), restricts the number of overlapping genes from the candidate and reference lists to be at least 2 — i.e., both O and C must be ≥ 2 in the tables below. This explains why Table B.2 is so short, and corroborates to a certain extent our finding that RS tends to produce the most inflated (nominal) measure of association (see Section 2.3 and Table 6 of main text).

Table B.1. Analysis of bipolar disorder data. GSEA results from KEGG for genes identified by MDR. O = number of genes in the discovered set; C = total number of genes in the given pathway.

				Adjusted
Line	Name	O	C	p-value
1	oocyte meiosis	3	124	0.001
2	adrenergic signaling in cardiomyocytes	3	149	0.002
3	bile secretion	2	71	0.007
4	human T-cell leukemia virus 1 infection	3	255	0.009
5	insulin secretion	2	85	0.010
6	HTLV-I infection	3	258	0.010
7	progesterone-mediated oocyte maturation	2	98	0.013
8	cell cycle	2	124	0.020
9	phospholipase D signaling pathway	2	144	0.027

Table B.2. Analysis of bipolar disorder data. GSEA results from KEGG for genes identified by RS. O = number of genes in the discovered set; C = total number of genes in the given pathway.

				Adjusted
Line	Name	O	C	p-value
1	oocyte meiosis	2	124	0.002

B.2 GSEA results from Gene Ontology and Pathway Commons for genes identified by PTY

Gene Ontology

Table B.3 lists the statistically enriched pathways from Gene Ontology (multiple-testing adjusted p-value < 0.05). Quite a few of them turned out to be related to neurons and/or neuronal activities, which to some extent confirmed the relevance of the gene set we identified. For example, the pathway labelled "regulation of synaptic plasticity" (line 1) is highly relevant to both the pathophysiology and the treatment of bipolar disorder [113]. Animal models have also shown that over-strengthened and/or weakened synapses at different circuits in the brain can disturb brain functions in parallel, causing manic-like or depressive-like behaviors [114]. Similarly, the pathway labelled "neuropeptide signaling" (line 3) is integral to the modulation of membrane excitability, synaptic transmission and synaptic development [115]; while the one labelled “neuron projection” (line 11) has previously been reported to be enriched in a GWAS of bipolar disorder [116].

Table B.3. Analysis of bipolar disorder data. GSEA results from Gene Ontoloty. O = number of genes in the discovered set; C = total number of genes in the given pathway.

				p-value	
Line	Name	O	C	Nominal	Adjusted
1	regulation of synaptic plasticity	4	96	0.0009	0.05
2	generation of neurons	13	1073	0.0009	0.05
3	neuropeptide signalling pathway	4	89	0.0007	0.05
4	optic nerve development	2	8	0.0006	0.05
5	organonitrogen compound biosynthetic process	18	1688	0.0004	0.05
6	oligosaccharide metabolic process	3	37	0.0006	0.05
7	neuron recognition	3	30	0.0003	0.05
8	cell recognition	4	77	0.0004	0.05
9	ion binding	38	5820	0.0004	0.04
10	axon	8	286	$\ll 0.0001$	< 0.01
11	neuron projection	11	651	$\ll 0.0001$	< 0.01

Pathway Commons

Table B.4 lists the statistically enriched pathways from Pathway Commons (multiple-testing adjusted p-values < 0.1). Quite a few of them have been associated with various neural activities, neural disorders, and specifically bipolar disorder itself. For example, insulin (line 1) plays a critical role in the central nervous system, contributing to physiological processes such as neuroplasticity, neuromodulation, and neurotrophism [117]. Animal experiments have also shown that the injection of insulin in the brain increases both the amount and the activity of dopamine transporters, which may play a role in bipolar disorder [118]. The ephrin/eph signaling pathway (lines 5 and 9) affects the structure and connectivity of the dopaminergic pathway [119], and it also coordinates multiple aspects of neural development such as cell migration and axon targeting [120]. The Arf6 downstream pathway (lines 13, 14 and 18) regulates neuronal migration [121]. The N-cadherin (line 51) is important for asymmetrical cellular processes in developing neurons and for regulating neuronal polarity [122, 123]. Altered CDC42 signaling pathways (lines 53 and 54) have been observed in patients schizophrenia [124] and, finally, evidence is also emerging that the Wnt pathway (lines 61, 64 and 68) is important for bipolar disorder [125–127].

In addition, many pathways in this table (e.g., lines 4, 7, 9, 10, 11, 12, and so on) are related to cell signaling. Neurotransmitters are, of course, important signaling molecules; and so are hormones, and many of them (e.g., glucocorticoids, thyroid hormones and gonadal steroids) are known to mediate symptoms observed in mood disorders, e.g., triggering of episodes in the postpartum period [128]. There has also been recent, specific suggestions that some signal transduction pathways may play an integral role in the pathophysiology and treatment of bipolar disorder [128].

Table B.4. Analysis of bipolar disorder data. GSEA results from Pathway Commons. O = number of genes in the discovered set; C = total number of genes in the given pathway.

				p-value	
Line	Name	O	C	Nominal	Adjusted
1	Insulin Pathway	8	1288	0.0015	0.0034
2	Thrombin/protease-activated receptor (PAR) pathway	8	1300	0.0016	0.0034
3	S1P1 pathway	8	1288	0.0015	0.0034
4	IL5-mediated signaling events	8	1292	0.0015	0.0034
5	EphrinB-EPHB pathway	3	60	0.0001	0.0034
6	Signaling events mediated by focal adhesion kinase	8	1288	0.0015	0.0034
7	ErbB receptor signaling network	9	1312	0.0004	0.0034
8	IGF1 pathway	8	1291	0.0015	0.0034
9	Ephrin B reverse signaling	2	30	0.0012	0.0034
10	mTOR signaling pathway	8	1288	0.0015	0.0034
11	LKB1 signaling events	8	1308	0.0016	0.0034
12	PAR1-mediated thrombin signaling events	8	1299	0.0016	0.0034
13	Arf6 downstream pathway	8	1288	0.0015	0.0034
14	Arf6 trafficking events	8	1288	0.0015	0.0034
15	Internalization of ErbB1	8	1288	0.0015	0.0034
16	Signaling events mediated by VEGFR1 and VEGFR2	8	1296	0.0016	0.0034
17	Endothelins	8	1307	0.0016	0.0034

(continued on next page ...)

Table B.4 (... continued from previous page)

				p-value	
Line	Name	O	C	Nominal	Adjusted
18	Arf6 signaling events	8	1288	0.0015	0.0034
19	PDGFR-beta signaling pathway	8	1288	0.0015	0.0034
20	Class I PI3K signaling events mediated by Akt	8	1288	0.0015	0.0034
21	Urokinase-type plasminogen activator (uPA) and uPAR-mediated signaling	8	1288	0.0015	0.0034
22	IL3-mediated signaling events	8	1295	0.0015	0.0034
23	PDGF receptor signaling network	8	1293	0.0015	0.0034
23	Metabolism	7	824	0.0005	0.0034
24	IFN-gamma pathway	8	1296	0.0016	0.0034
25	Alpha9 beta1 integrin signaling events	8	1305	0.0016	0.0034
26	Plasma membrane estrogen receptor signaling	8	1301	0.0016	0.0034
27	Metabolism of amino acids and derivatives	4	188	0.0003	0.0034
28	EGF receptor (ErbB1) signaling pathway	8	1288	0.0015	0.0034
29	ErbB1 downstream signaling	8	1288	0.0015	0.0034
30	Nectin adhesion pathway	8	1295	0.0015	0.0034
31	Amine-derived hormones	2	14	0.0003	0.0034
32	Class I PI3K signaling events	8	1288	0.0015	0.0034
33	Syndecan-1-mediated signaling events	8	1300	0.0016	0.0034
34	Glypican 1 network	8	1299	0.0016	0.0034

(continued on next page ...)

Table B.4 (... continued from previous page)

				p-value	
Line	Name	O	C	Nominal	Adjusted
35	GMCSF-mediated signaling events	8	1292	0.0015	0.0034
36	VEGF and VEGFR signaling network	8	1304	0.0016	0.0034
37	Proteoglycan syndecan-mediated signaling events	9	1345	0.0004	0.0034
38	EGFR-dependent Endothelin signaling events	8	1289	0.0015	0.0034
39	Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met)	8	1293	0.0015	0.0034
40	Sphingosine 1-phosphate (S1P) pathway	8	1311	0.0017	0.0035
41	TRAIL signaling pathway	8	1328	0.0018	0.0036
42	Glypican pathway	8	1338	0.0019	0.0038
43	Beta1 integrin cell surface interactions	8	1351	0.0020	0.0039
44	EPHB forward signaling	2	40	0.0021	0.0040
45	Integrin family cell surface interactions	8	1378	0.0023	0.0043
46	AP-1 transcription factor network	5	623	0.0041	0.0074
47	Integrin-linked kinase signaling	5	656	0.0051	0.0090
48	Signaling by SCF-KIT	2	66	0.0057	0.0099
49	Axon guidance	3	219	0.0062	0.0105
50	Posttranslational regulation of adherens junction stability and dissassembly	3	231	0.0072	0.0120
51	N-cadherin signaling events	3	251	0.0090	0.0144
52	Integration of energy metabolism	2	83	0.0088	0.0144

(continued on next page ...)

Table B.4 (... continued from previous page)

				p-value	
Line	Name	O	C	Nominal	Adjusted
53	CDC42 signaling events	5	757	0.0093	0.0146
54	Regulation of CDC42 activity	5	770	0.0099	0.0153
55	Stabilization and expansion of the E-cadherin adherens junction	3	275	0.0115	0.0171
56	E-cadherin signaling in the nascent adherens junction	3	275	0.0115	0.0171
57	E-cadherin signaling events	3	280	0.0121	0.0177
58	Regulation of nuclear beta catenin signaling and target gene transcription	2	135	0.0221	0.0317
59	Validated transcriptional targets of AP1 family members Fra1 and Fra2	2	136	0.0224	0.0317
60	EGFR1	2	138	0.0230	0.0320
61	Canonical Wnt signaling pathway	2	155	0.0286	0.0392
62	Regulation of p38-alpha and p38-beta	2	164	0.0317	0.0428
63	Developmental Biology	3	433	0.0373	0.0495
64	Noncanonical Wnt signaling pathway	2	182	0.0383	0.0501
65	p53 pathway	2	189	0.0410	0.0520
66	p38 MAPK signaling pathway	2	189	0.0410	0.0520
67	CXCR4-mediated signaling events	2	192	0.0422	0.0527
68	Wnt signaling network	2	200	0.0454	0.0559
69	Glypican 3 network	2	206	0.0478	0.0580
70	Syndecan-4-mediated signaling events	2	209	0.0491	0.0588

(continued on next page ...)

Table B.4 (... continued from previous page)

				p-value	
Line	Name	O	C	Nominal	Adjusted
71	BMP receptor signaling	2	226	0.0564	0.0666
72	Signal Transduction	5	1231	0.0576	0.0671
73	IL1-mediated signaling events	2	234	0.0599	0.0688
74	ATR signaling pathway	2	250	0.0673	0.0763

B.3 A numeric example illustrating the similarity metric $\Phi(M', M)$

Consider a pair of SNPs, each with MAF=0.2. Suppose that the true disease model is M, and that there are two potential prototypes, M' or M'', as shown below. Which candidate is more similar to, and hence a better prototype for, M?

$\mathbb{P}(G_i)$

	AA	Aa	aa
BB	0.4096	0.2048	0.0256
Bb	0.2048	0.1024	0.0128
bb	0.0256	0.0128	0.0016

M

	AA	Aa	aa
BB	0	0	0
Bb	0	1	1
bb	0	1	1

M'

	AA	Aa	aa
BB	0	0	**1**
Bb	0	1	1
bb	**1**	1	**0**

M''

	AA	Aa	aa
BB	0	0	0
Bb	0	1	**0**
bb	0	**0**	**0**

Here, model M' differs from M on the genotypes, aaBB, AAbb and aabb, with

$$W'_{11} = 0.1280, \quad W'_{10} = 0.0016, \quad W'_{01} = 0.0512, \quad W'_{00} = 0.8192;$$

and model M'' differs from M on the genotypes, aaBb, Aabb and aabb, with

$$W''_{11} = 0.1024, \quad W''_{10} = 0.0272, \quad W''_{01} = 0.0000, \quad W''_{00} = 0.8704.$$

The entries for which $M' \neq M$ are emboldened in the table above, and so are those for which $M'' \neq M$.

Overall, M'' has less disagreement with M than does M', as can be seen from the inequality:

$$\sum_{k \neq \ell} W''_{k\ell} = (0.0272 + 0.0000) < (0.0016 + 0.0512) = \sum_{k \neq \ell} W'_{k\ell}.$$

However, while M' has more disagreement with M overall, it actually has more agreement with M than does M'' on the set $\mathcal{G}_1 \equiv \{G_i : M(G_i) = 1\}$—i.e., not only AaBb but also aaBb and Aabb, although it also has additional disagreements with M on the set $\mathcal{G}_0 \equiv \{G_i : M(G_i) = 0\}$—i.e., aaBB, AAbb, whereas M'' agrees with M completely on $\mathcal{G}_0$.

The similarity metric Φ is acutely sensitive to such a difference. On population data ($r = r_0$; see Section 6 of main text), we have

$$\frac{U}{V} = 1 \quad \Rightarrow \quad \Phi(M', M) = 0.273 \quad \text{and} \quad \Phi(M'', M) = 0.294,$$

so M'' would be considered more similar to, and hence a better prototype for, M. However, on case-control data ($r \gg r_0$), the ratio $U/V \approx P_1/P_0 > 1$; see, again, Section 6 of main text. As a concrete example here, suppose $U/V \approx P_1/P_0 = 5$, which is a fairly typical value in practice; then, we get

$$\frac{U}{V} = 5 \quad \Rightarrow \quad \Phi(M', M) = 0.310 \quad \text{and} \quad \Phi(M'', M) = 0.278,$$

so M' would be considered a better prototype for M instead of M''. We think this is a desirable property of our similarity metric Φ, in that it is more important for a prototype candidate M' to agree with M on the set $\mathcal{G}_1$ than on $\mathcal{G}_0$.

Appendix C

Additional Material for Chapter 4

C.1 Proof of the Equivalance between Forward Selection on Transformed Variables and General Ratio Split Procedure

1. *Proof.* The 9th variable in $X\tilde{D}^{-1}$ is $X(1, \cdots, 1)^T = \mathbf{1}_{1\times n}^T$, because as dummy variables, X_i's sum up to one. Therefore it is a constant and in the loss function $|y - X\tilde{D}^{-1}\theta|^2$, it is equivalent to the role of an intercept. □

2. *Proof.* For step 1, $\mathbb{X}_1 = \frac{1}{2}\sum_{l=1}^{j} X_l - \frac{1}{2}(1 - \sum_{l=1}^{j} X_l) = \sum_{l=1}^{j} X_l - \frac{1}{2}$, so the right hand side of Eq. 4.19 becomes

$$\gamma_0 + \gamma_1(\sum_{l=1}^{j} X_l - \frac{1}{2}) = \gamma_0 - \frac{1}{2}\gamma_1 + \gamma_1 \sum_{l=1}^{j} X_l \tag{C.1}$$

   On the other hand, the regression coefficient difference $\beta_{g^1_{1_j}} - \beta_{g^1_{2_j}}$ is from the regression model

$$
\begin{aligned}
\log\frac{p(y=1)}{p(y=0)} &= \beta_{g^1_{1_j}} \sum_{l=1}^{j} X_l + \beta_{g^1_{2_j}} \sum_{l=j+1}^{9} X_l = \beta_{g^1_{1_j}} \sum_{l=1}^{j} X_l + \beta_{g^1_{2_j}} (1 - \sum_{l=1}^{j} X_l) \\
&= (\beta_{g^1_{1_j}} - \beta_{g^1_{2_j}}) \sum_{l=1}^{j} X_l + \beta_{g^1_{2_j}}
\end{aligned}
$$

Compare to Eq. C.1, it indicates $\beta_{g^1_{1_j}} - \beta_{g^1_{2_j}} = \gamma_1$.

For a general step k, notice the $\mathbb{X}_k$ is composed of two groups of dummies that together form one of the split group at step $k-1$. Without loss of generality, assume the split order is $j_1 < j_2 < \cdots < j_{k-1}$, and we are exploring the k split to be at j. Use T_i to denote $\sum_{l=j_{i-1}}^{j_i} X_l, i = 1, 2, \cdots, k-1$, $T_k = \sum_{l=j_{k-1}}^{j} X_l$ and $T_{k+1} = \sum_{l=j+1}^{9} X_l$, then we have

$$
\begin{aligned}
\mathbb{X}_1 &= \frac{1}{2}(\sum_{l=1}^{j_1} X_l - \sum_{l=j_1+1}^{9} X_l) = \frac{1}{2} - (T_2 + \cdots + T_{k+1}) \\
\mathbb{X}_i &= \frac{1}{2}(\sum_{l=j_{i-1}}^{j_i} X_l - \sum_{l=j_i+1}^{9} X_l) = \frac{1}{2}(T_i - T_{i+1} - \cdots - T_{k+1}), 2 \leq i < k \\
\mathbb{X}_k &= \frac{1}{2}(\sum_{l=j_{k-1}}^{j} X_l - \sum_{l=j+1}^{9} X_l) = \frac{1}{2}(T_k - T_{k+1})
\end{aligned}
$$

And so

$$
\begin{aligned}
\sum_{i=1}^{k} \gamma_i \mathbb{X}_i = \frac{1}{2}(\gamma_1 + (\gamma_2 - 2\gamma_1)T_2 + (\gamma_3 - \gamma_2 - 2\gamma_1)T_3 + \cdots + (\gamma_k - \gamma_{k-1} \\
- \cdots - \gamma_2 - 2\gamma_1)T_k - (\gamma_k + \gamma_{k-1} + \cdots + \gamma_2 + 2\gamma_1)T_{k+1}). \qquad \text{(C.2)}
\end{aligned}
$$

On the other hand, $\beta_{g^k_{1_j}} - \beta_{g^k_{2_j}}$ is from the regression model

$$\begin{aligned}
\log\frac{p(y=1)}{p(y=0)} &= \beta_{g^k_{1_{j_1}}} T_1 + \cdots + \beta_{g^k_{1_{j_{k-1}}}} T_{k-1} + \beta_{g^k_{1_j}} T_k + \beta_{g^k_{2_j}} T_{k+1} \\
&= \beta_{g^k_{1_{j_1}}} (1 - T_2 - \cdots - T_{k+1}) + \cdots + \beta_{g^k_{1_{j_{k-1}}}} T_{k-1} + \beta_{g^k_{1_j}} T_k + \beta_{g^k_{2_j}} T_{k+1} \\
&= \beta_{g^k_{1_{j_1}}} + (\beta_{g^k_{1_{j_2}}} - \beta_{g^k_{1_{j_1}}})T_2 + \cdots + (\beta_{g^k_{1_{j_k}}} - \beta_{g^k_{1_{j_1}}})T_k + (\beta_{g^k_{1_{j_{k+1}}}} - \beta_{g^k_{1_{j_1}}})T_{k+1}
\end{aligned}$$

Compare coefficients to the ones in Eq. C.2 we get

$$\begin{cases}
\frac{1}{2}(\gamma_k - \gamma_{k-1} - \cdots - \gamma_2 - 2\gamma_1) = \beta_{g^k_{1_{j_k}}} - \beta_{g^k_{1_{j_1}}} \\
\frac{1}{2}(\gamma_k + \gamma_{k-1} + \cdots + \gamma_2 + 2\gamma_1) = \beta_{g^k_{1_{j_1}}} - \beta_{g^k_{1_{j_{k+1}}}}
\end{cases}$$

Sum the left and right hand side we get $\gamma_k = \beta_{g^k_{1_{j_k}}} - \beta_{g^k_{1_{j_{k+1}}}}$.

□

3. *Proof.* The two variables can be simplified as

$$\begin{aligned}
U_j^1 &= -\frac{9-j}{9}(X_1 + \cdots X_j) + \frac{j}{9}(X_{j+1} + \cdots X_9) := -\frac{9-j}{9}T_1 + \frac{j}{9}(1 - \bar{T}_1) = -T_1 + \frac{j}{9} \\
V_j^1 &= -\frac{1}{2}(X_1 + \cdots X_j) + \frac{1}{2}(X_{j+1} + \cdots X_9) := -\frac{1}{2}T_1 + \frac{1}{2}T_2 = \frac{1}{2} - T_1
\end{aligned}$$

So the center of the two variables are

$$\bar{U}_j^1 = -\bar{T}_1 + \frac{j}{9},$$
$$\bar{V}_j^1 = -\bar{T}_1 + \frac{1}{2}.$$

Taking out the center of them we get

$$
\begin{aligned}
U_j^1 - \bar{U}_j &= -(T_1 - \bar{T}_1) \\
V_j^1 - \bar{V}_j &= -(T_1 - \bar{T}_1)
\end{aligned}
$$

Therefore

$$V_j^1 - \bar{V}_j^1 = U_j^1 - \bar{U}_j^1.$$

□

4. *Proof.* From 3 we know $V_j^1 - \bar{V}_j^1 = U_j^1 - \bar{U}_j^1$, so it suffices to show

$$(V_j^1 - \bar{V}_j^1)^T P_A^{\perp} = (V_j^k - \bar{V}_j^k)^T P_A^{\perp}. \tag{C.3}$$

We have

$$V_j^k = V_j^1 - T_1^k + T_2^k$$

for some T_1^k, T_2^k as summations of the original X_i's (they could be the empty set as well). And so additionally we have

$$\bar{V}_j^k = \bar{V}_j^1 - \bar{T}_1^k + \bar{T}_2^k$$

Plug back to Eq. C.3, we get that it suffices to show

$$(T_2^k - T_1^k - \bar{T}_2^k + \bar{T}_1^k)^T P_A^{\perp} = 0$$

Notice X_A is the design matrix for already selected variables, so T_1^k and T_2^k are summations of dummies variables all included in A. This indicate both T_1^k and T_2^k are linear functions of X_A, so we have

$$(T_2^k - T_1^k - \bar{T_2^k} + \bar{T_1^k})^T = X_A c$$

for some constant vector c.

Therefore

$$(T_2^k - T_1^k - \bar{T_2^k} + \bar{T_1^k})^T X_A = c^T X_A^T X_A$$

And so

$$\begin{aligned}
(T_2^k - T_1^k - \bar{T_2^k} + \bar{T_1^k})^T P_A^{\perp} &= (T_2^k - T_1^k - \bar{T_2^k} + \bar{T_1^k})^T (I - X_A (X_A^T X_A)^{-1} X_A) \\
&= T_2^k - T_1^k - \bar{T_2^k} + \bar{T_1^k} - c^T X_A^T X_A (X_A^T X_A)^{-1} X_A \\
&= c^T X_A^T - c^T X_A^T = 0.
\end{aligned}$$

□

C.2 Sequential Merge Procedure

C.2.1 Procedure Overview

Flexible levels of risk model is achieved by sorting the nine genotype combinations based on the case-to-control ratios, and then sequentially merging the neibours.

Example C.2.1. *General Merging Example*

$$\textit{Step 1:}\quad X_1 \mid \overbrace{X_2 \quad X_3}^{\textit{merge 1}} \mid X_4 \mid X_5 \mid X_6 \mid X_7 \mid X_8 \mid X_9$$

$$\textit{Step 2:}\quad X_1 \mid \overbrace{X_2 \quad X_3 \quad X_4}^{\textit{merge 2}} \mid X_5 \mid X_6 \mid X_7 \mid X_8 \mid X_9$$

$$\textit{Step 3:}\quad X_1 \mid \overbrace{X_2 \quad X_3 \quad X_4}^{\textit{merge 2}} \mid X_5 \mid \quad \overbrace{X_6 \quad X_7}^{\textit{merge 3}} \mid X_8 \mid X_9$$

$$\vdots$$

$$\textit{Step 8:}\quad \overbrace{X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad X_6 \quad X_7 \quad X_8 \quad X_9}^{\textit{merge 8}}$$

In general, the sequential merge is done in the following steps:

Step 1 Test each one of the eight possible splits (see illustration below) and select the one with the best fit.

$$\text{Candidate Merge 1:}\quad \overbrace{X_1 \quad X_2} \mid X_3 \mid X_4 \mid X_5 \mid X_6 \mid X_7 \mid X_8 \mid X_9$$

$$\text{Candidate Merge 2:}\quad X_1 \mid \overbrace{X_2 \quad X_3} \mid X_4 \mid X_5 \mid X_6 \mid X_7 \mid X_8 \mid X_9$$

$$\vdots$$

$$\text{Candidate Merge 8:}\quad X_1 \mid X_2 \mid X_3 \mid X_4 \mid X_5 \mid X_6 \mid X_7 \mid \overbrace{X_8 \quad X_9}$$

Step 2 Fix the first merge place and explore the rest seven places for a possible second merge (see illustration below). Perform a proper test taking into consideration of the first one and select the best.

$$\text{Candidate Merge 1:} \quad \overbrace{X_1 \quad X_2} \mid X_3 \mid \overbrace{X_4 \quad X_5}^{\text{Step 1 Merge}} \mid X_6 \mid X_7 \mid X_8 \mid X_9$$

$$\text{Candidate Merge 2:} \quad X_1 \mid \overbrace{X_2 \quad X_3} \mid \overbrace{X_4 \quad X_5}^{\text{Step 1 Merge}} \mid X_6 \mid X_7 \mid X_8 \mid X_9$$

$$\text{Candidate Merge 3:} \quad X_1 \mid X_2 \mid \overbrace{X_3 \quad \overbrace{X_4 \quad X_5}^{\text{Step 1 Merge}}} \mid X_6 \mid X_7 \mid X_8 \mid X_9$$

$$\vdots$$

$$\text{Candidate Merge 8:} \quad X_1 \mid X_2 \mid X_3 \mid \overbrace{X_4 \quad X_5}^{\text{Step 1 Merge}} \mid X_6 \mid X_7 \mid \overbrace{X_8 \quad X_9}$$

Step 3 Continue the procedure similar to step 2 while fixing the existing merges till all G_i's form one group.

Step 4 When the merging process is finished, refit to quantify the association of the SNP-pair and the outcome by the merge at each step (representing a different level of disease model), and select a final disease model out of all levels by the use of the same stopping criteria as applied for the general splitting procedure, i.e., minimum of

- "nominal" P-value as used in Section 3.2
- AIC
- BIC
- p-value from LRT test

C.2.2 Testing Statistics for a General Merge Step

The testing statitics for derterming the best merge place at each step is chosen to be the standardized coefficient difference of two to-be-merging groups, and the most significant one is selected.

At step k, assume the existing groups are $G_1, G_2, \cdots, G_{10-k}$, then the candidate merges are the following: $(G_1, G_2), (G_2, G_3), \cdots, (G_{9-k}, G_{10-k})$. For the jth candidate merge, the following logistic regression is fit,

$$\log \frac{p(y=1)}{p(y=0)} = \beta_1 \mathcal{X}_{G_1} + \cdots + \beta_j \mathcal{X}_{G_j} + \beta_{j+1} \mathcal{X}_{G_{j+1}} + \cdots + \beta_{9-k} \mathcal{X}_{G_{9-k}}, \tag{C.4}$$

for $j = 1, 2, \cdots, 9-k$, and the testing statistic is

$$Z_j = \frac{\hat{\beta_{j+1}} - \hat{\beta}_j}{\text{std } \widehat{(\hat{\beta_{j+1}} - \hat{\beta}_j})} \sim N(0,1).$$

The merge is chosen to be j_k for which

$$|Z_{j_k}| < |Z_j|, j \in \{1, 2, \cdots, 9-k\} - \{j_k\}.$$

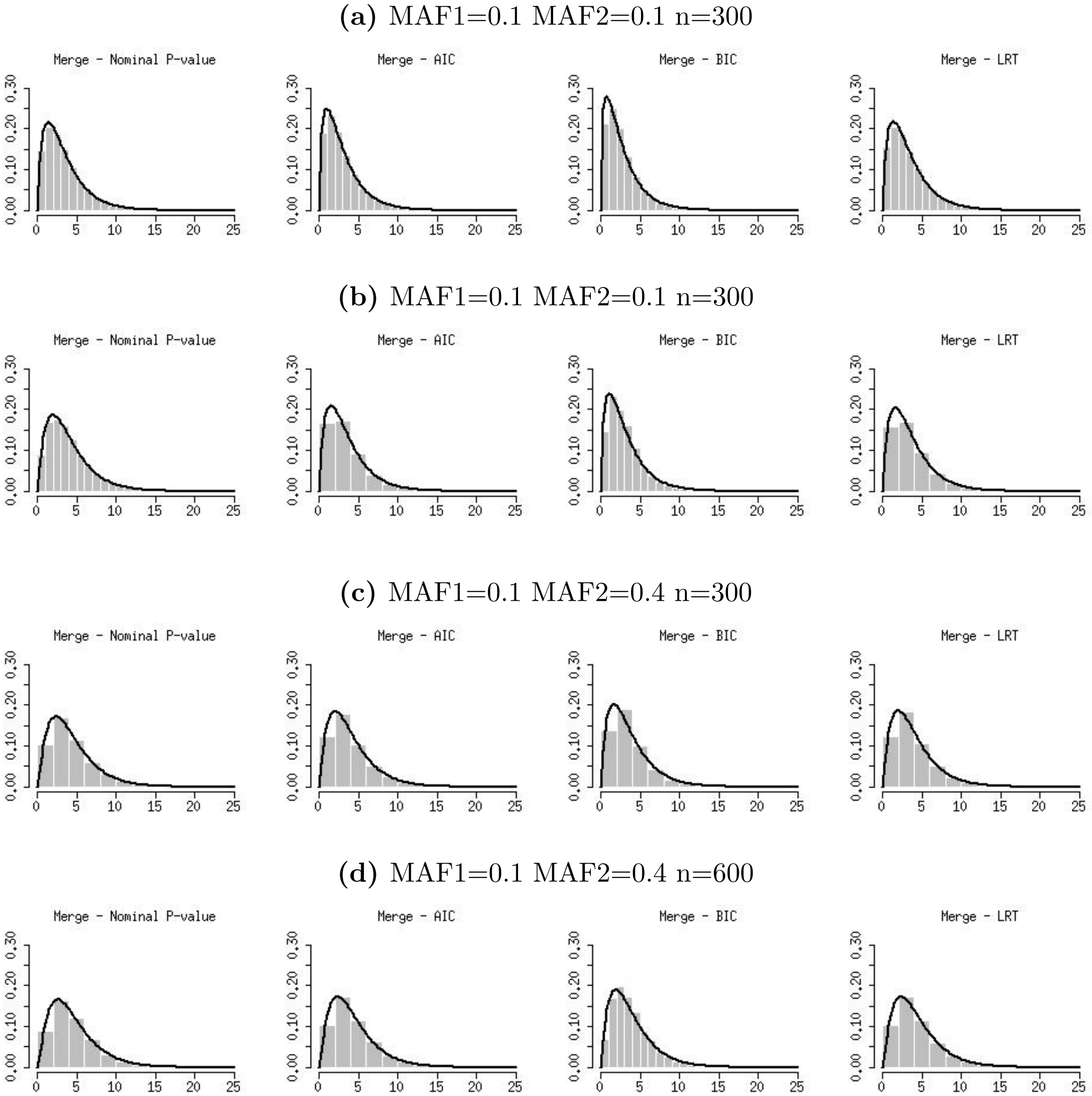
The rest of the general merging steps are the same as the general split procedure, i.e., the final disease model is chosen among models of all levels and refitted to the SNP-pair to obtain a χ^2-statistic, and the statistic is evaluated using adjusted degree of freedom to be used as the SNP-pair ranking measure.

C.2.3 Response Surface Model Results

Figure C.1 provides some histogram examples for the χ^2-statistics generated using the null data, and the response surface model results for SMP are summarized in Table C.1. The statistics are observed to generally follow the shape of χ^2-distribution.

Table C.1. Parameter estimates of the response surface model for the sequential merge procedures under the null distribution. The model equation is given in Eq. 4.20.

Stopping Criteria	Parameter	Estimate	Std	P-value	R^2	Adjusted R^2
Nominal P-value	β_0	-0.647	0.127	5.05E-07	0.951	0.951
	β_1	12.630	0.281	7.62E-185		
	β_2	-13.782	0.314	1.67E-180		
	β_3	13.728	0.847	2.01E-48		
	β_4	-13.250	1.302	2.25E-22		
	β_5	3.929	0.397	2.18E-21		
	β_6	-14.727	1.438	1.25E-22		
	β_7	13.672	2.209	1.19E-09		
AIC	β_0	-1.086	0.117	5.62E-19	0.968	0.967
	β_1	11.872	0.260	4.07E-188		
	β_2	-13.924	0.290	2.64E-197		
	β_3	13.262	0.783	5.44E-52		
	β_4	-12.244	1.203	2.16E-22		
	β_5	4.274	0.366	3.13E-28		
	β_6	-7.797	1.328	7.52E-09		
	β_7	5.301	2.040	9.62E-03		
BIC	β_0	-0.423	0.069	1.89E-09	0.985	0.984
	β_1	9.020	0.153	5.80E-238		
	β_2	-10.578	0.171	6.54E-248		
	β_3	10.103	0.461	1.07E-76		
	β_4	-9.997	0.708	9.74E-39		
	β_5	4.227	0.216	4.90E-65		
	β_6	-5.006	0.782	3.35E-10		
	β_7	4.158	1.202	5.83E-04		
LRT	β_0	0.574	0.078	8.08E-13	0.970	0.969
	β_1	7.038	0.173	5.62E-167		
	β_2	-7.876	0.193	2.81E-167		
	β_3	9.412	0.521	1.91E-57		
	β_4	-9.702	0.801	4.47E-30		
	β_5	3.157	0.244	1.32E-33		
	β_6	-5.015	0.884	2.28E-08		
	β_7	4.759	1.358	4.96E-04		

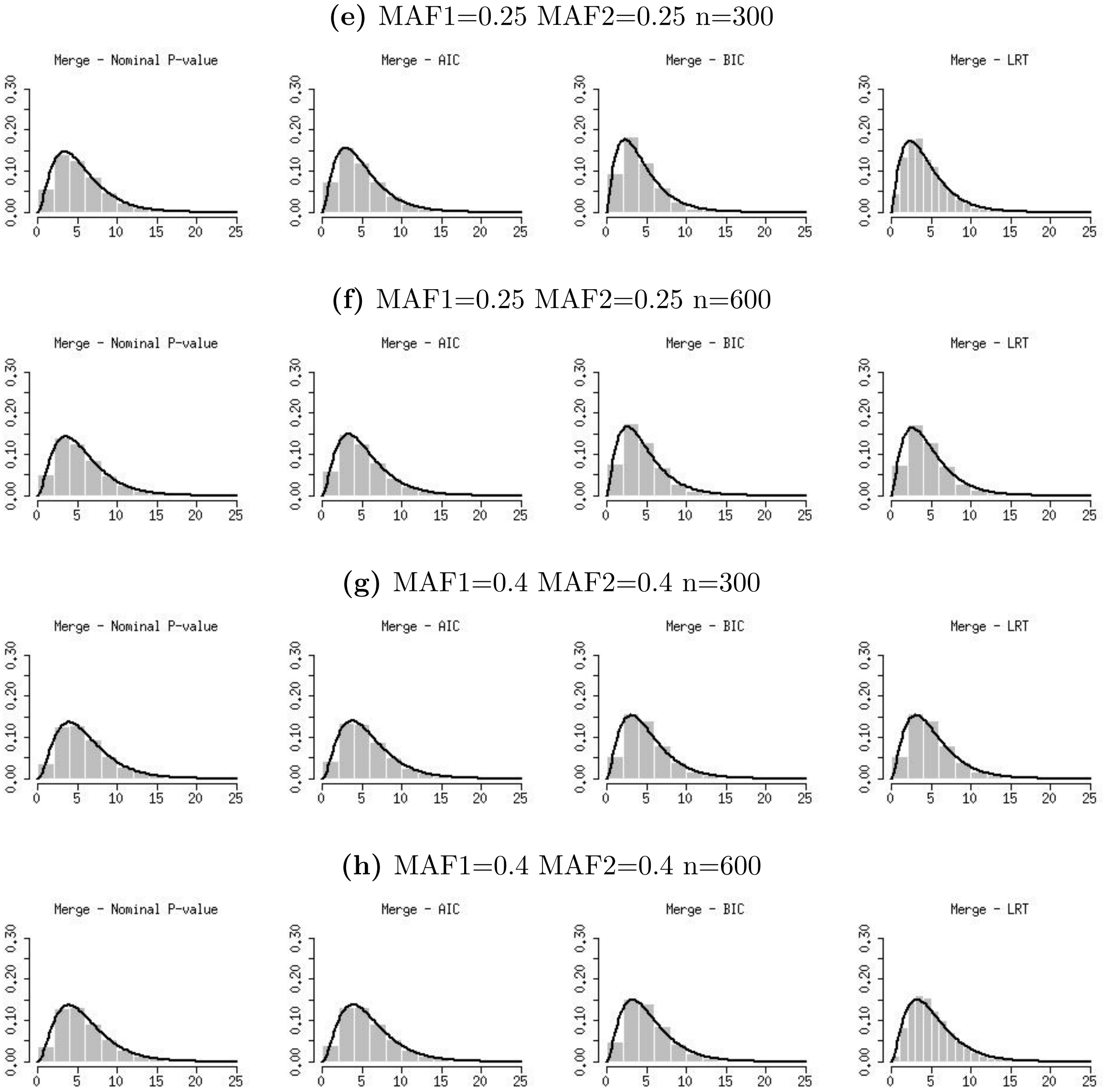


(a) MAF1=0.1 MAF2=0.1 n=300

(b) MAF1=0.1 MAF2=0.1 n=300

(c) MAF1=0.1 MAF2=0.4 n=300

(d) MAF1=0.1 MAF2=0.4 n=600



(e) MAF1=0.25 MAF2=0.25 n=300

(f) MAF1=0.25 MAF2=0.25 n=600

(g) MAF1=0.4 MAF2=0.4 n=300

(h) MAF1=0.4 MAF2=0.4 n=600

Fig C.1. Histograms of $\{\widehat{\chi}_s^2 : s = 1, 2, ..., S\}$ (for the SMP method) versus the $\chi^2_{(\text{EDF})}$ density functions, where EDF is computed by Eq. (3.2), for some specific combinations of $(\text{MAF}_1, \text{MAF}_2, n)$ and four commonly used stopping criterias. While the $\chi^2_{(\text{EDF})}$ density functions are not perfect fits of the underlying histograms, they are reasonable approximations as first-order corrections.

C.3 Simulation Study Results for SSP and SMP

The table C.2 provides the recall, precision, F-measure and determined disease model DF for the SSP and SMP methods under all their respective stopping criteria, with and without EDF adjustment on the SNP-pair association test p-value. Figure C.2 presents the DF values and Figure C.3 presents the F-measures, all in barplots for better view of the results.

Table C.2. Results for the simulation study on the P-value adjustment for the SSP, SMP and FSD-PST methods.

Model	MAF	Method	Stopping Criteria	Precision	Recall	F-measure	DF
T	0.1	Merge	Nomi P	0.045	0.155	0.070	1.968
			AIC	0.042	0.140	0.065	2.375
			BIC	0.046	0.125	0.067	1.720
			LRT	0.065	0.110	0.081	1.568
			Nomi P Adj EDF	0.313	0.053	0.090	3.901
			AIC Adj EDF	0.315	0.103	0.155	3.449
			BIC Adj EDF	0.317	0.063	0.104	3.033
			LRT Adj EDF	0.603	0.045	0.084	3.548
			Nomi P Adj Match	0.405	0.078	0.130	3.638
			AIC Adj Match	0.317	0.115	0.169	3.227
			BIC Adj Match	0.330	0.075	0.122	2.870
			LRT Adj Match	0.597	0.048	0.088	3.379
		Split	Nomi P	0.039	0.140	0.061	2.018
			AIC	0.032	0.123	0.051	2.673
			BIC	0.047	0.125	0.068	1.800
			LRT	0.050	0.090	0.064	1.611
			Nomi P Adj EDF	0.310	0.055	0.093	3.873
			AIC Adj EDF	0.326	0.108	0.162	3.450
			BIC Adj EDF	0.296	0.070	0.113	3.065
			LRT Adj EDF	0.597	0.048	0.088	3.578

			Nomi P Adj Match	0.395	0.075	0.126	3.610
			AIC Adj Match	0.329	0.125	0.181	3.224
			BIC Adj Match	0.298	0.075	0.120	2.877
			LRT Adj Match	0.613	0.048	0.088	3.398
		FST	Post P	0.086	0.175	0.115	1.257
			Post P Adj EDF	0.130	0.020	0.035	3.860
			Post P Adj Match	0.168	0.030	0.051	3.596
T	0.4	Merge	Nomi P	0.072	0.988	0.134	1.623
			AIC	0.074	0.983	0.137	2.046
			BIC	0.084	0.980	0.154	1.209
			LRT	0.130	0.978	0.230	1.000
			Nomi P Adj EDF	0.388	0.753	0.512	5.833
			AIC Adj EDF	0.305	0.793	0.440	5.598
			BIC Adj EDF	0.329	0.750	0.458	4.929
			LRT Adj EDF	0.657	0.695	0.675	4.947
			Nomi P Adj Match	0.389	0.745	0.511	5.910
			AIC Adj Match	0.305	0.798	0.441	5.582
			BIC Adj Match	0.317	0.750	0.446	4.905
			LRT Adj Match	0.653	0.695	0.674	4.916
		Split	Nomi P	0.075	0.988	0.139	1.620
			AIC	0.077	0.983	0.143	2.125
			BIC	0.087	0.980	0.160	1.171
			LRT	0.129	0.978	0.229	1.000
			Nomi P Adj EDF	0.372	0.758	0.499	5.728
			AIC Adj EDF	0.300	0.793	0.435	5.566
			BIC Adj EDF	0.333	0.718	0.455	5.007
			LRT Adj EDF	0.671	0.688	0.679	5.027
			Nomi P Adj Match	0.376	0.755	0.502	5.796

			AIC Adj Match	0.302	0.800	0.438	5.555
			BIC Adj Match	0.322	0.725	0.446	4.978
			LRT Adj Match	0.662	0.693	0.677	4.992
		FST	Post P	0.099	0.970	0.180	1.307
			Post P Adj EDF	0.389	0.685	0.496	5.726
			Post P Adj Match	0.391	0.683	0.497	5.795
MOD	0.1	Merge	Nomi P	0.017	0.753	0.033	1.711
			AIC	0.018	0.738	0.034	1.892
			BIC	0.021	0.723	0.041	1.190
			LRT	0.030	0.708	0.058	1.106
			Nomi P Adj EDF	0.046	0.433	0.083	3.894
			AIC Adj EDF	0.042	0.490	0.077	3.423
			BIC Adj EDF	0.036	0.383	0.066	3.003
			LRT Adj EDF	0.079	0.295	0.125	3.531
			Nomi P Adj Match	0.054	0.453	0.096	3.614
			AIC Adj Match	0.044	0.513	0.080	3.208
			BIC Adj Match	0.043	0.408	0.078	2.849
			LRT Adj Match	0.098	0.325	0.150	3.352
		Split	Nomi P	0.018	0.743	0.035	1.899
			AIC	0.019	0.718	0.037	2.258
			BIC	0.021	0.718	0.041	1.484
			LRT	0.030	0.643	0.057	1.206
			Nomi P Adj EDF	0.048	0.445	0.086	3.866
			AIC Adj EDF	0.043	0.525	0.080	3.440
			BIC Adj EDF	0.041	0.405	0.075	3.023
			LRT Adj EDF	0.069	0.260	0.109	3.554
			Nomi P Adj Match	0.057	0.480	0.102	3.588
			AIC Adj Match	0.044	0.543	0.081	3.221

			BIC Adj Match	0.051	0.440	0.091	2.866
			LRT Adj Match	0.065	0.283	0.106	3.368
		FST	Post P	0.022	0.713	0.043	1.126
			Post P Adj EDF	0.037	0.298	0.065	3.866
			Post P Adj Match	0.044	0.335	0.078	3.578
MOD	0.4	Merge	Nomi P	0.020	0.995	0.040	1.681
			AIC	0.021	0.995	0.041	2.188
			BIC	0.024	0.993	0.048	1.242
			LRT	0.036	0.993	0.070	1.000
			Nomi P Adj EDF	0.143	0.883	0.246	5.839
			AIC Adj EDF	0.113	0.913	0.202	5.606
			BIC Adj EDF	0.122	0.885	0.214	4.939
			LRT Adj EDF	0.311	0.840	0.454	4.953
			Nomi P Adj Match	0.149	0.880	0.255	5.915
			AIC Adj Match	0.109	0.913	0.195	5.587
			BIC Adj Match	0.122	0.888	0.214	4.914
			LRT Adj Match	0.313	0.843	0.457	4.925
		Split	Nomi P	0.021	0.995	0.041	1.613
			AIC	0.022	0.995	0.042	2.211
			BIC	0.025	0.993	0.049	1.199
			LRT	0.034	0.993	0.067	1.000
			Nomi P Adj EDF	0.139	0.885	0.241	5.733
			AIC Adj EDF	0.111	0.910	0.197	5.574
			BIC Adj EDF	0.131	0.875	0.227	5.017
			LRT Adj EDF	0.314	0.848	0.458	5.035
			Nomi P Adj Match	0.138	0.878	0.238	5.800
			AIC Adj Match	0.107	0.910	0.191	5.560
			BIC Adj Match	0.130	0.878	0.226	4.988

			LRT Adj Match	0.318	0.850	0.463	5.001
		FST	Post P	0.027	0.985	0.053	1.223
			Post P Adj EDF	0.155	0.843	0.262	5.731
			Post P Adj Match	0.151	0.835	0.255	5.799
DD	0.1	Merge	Nomi P	0.075	0.988	0.139	1.334
			AIC	0.080	0.988	0.147	1.585
			BIC	0.086	0.985	0.159	1.033
			LRT	0.110	0.983	0.199	1.000
			Nomi P Adj EDF	0.398	0.930	0.557	3.893
			AIC Adj EDF	0.308	0.945	0.465	3.416
			BIC Adj EDF	0.384	0.933	0.544	3.000
			LRT Adj EDF	0.585	0.915	0.714	3.527
			Nomi P Adj Match	0.386	0.935	0.546	3.619
			AIC Adj Match	0.300	0.948	0.456	3.215
			BIC Adj Match	0.382	0.940	0.543	2.863
			LRT Adj Match	0.564	0.920	0.699	3.366
		Split	Nomi P	0.075	0.988	0.139	1.359
			AIC	0.079	0.988	0.146	1.782
			BIC	0.089	0.985	0.163	1.086
			LRT	0.110	0.983	0.197	1.000
			Nomi P Adj EDF	0.390	0.928	0.549	3.863
			AIC Adj EDF	0.306	0.943	0.462	3.430
			BIC Adj EDF	0.395	0.930	0.555	3.020
			LRT Adj EDF	0.589	0.908	0.715	3.545
			Nomi P Adj Match	0.379	0.933	0.539	3.594
			AIC Adj Match	0.300	0.945	0.456	3.225
			BIC Adj Match	0.393	0.940	0.554	2.880
			LRT Adj Match	0.567	0.915	0.700	3.381

		FST	Post P	0.096	0.980	0.175	1.385
			Post P Adj EDF	0.393	0.908	0.549	3.859
			Post P Adj Match	0.390	0.913	0.547	3.591
DD	0.4	Merge	Nomi P	0.007	1.000	0.014	1.950
			AIC	0.007	1.000	0.015	2.105
			BIC	0.008	1.000	0.015	1.175
			LRT	0.008	1.000	0.017	1.000
			Nomi P Adj EDF	0.043	0.990	0.083	5.840
			AIC Adj EDF	0.031	0.995	0.061	5.609
			BIC Adj EDF	0.034	0.990	0.066	4.943
			LRT Adj EDF	0.099	0.988	0.180	4.958
			Nomi P Adj Match	0.040	0.988	0.078	5.916
			AIC Adj Match	0.031	0.995	0.061	5.588
			BIC Adj Match	0.034	0.990	0.066	4.916
			LRT Adj Match	0.097	0.990	0.177	4.928
		Split	Nomi P	0.007	1.000	0.014	1.350
			AIC	0.007	1.000	0.015	2.143
			BIC	0.008	1.000	0.015	1.155
			LRT	0.008	1.000	0.017	1.000
			Nomi P Adj EDF	0.040	0.988	0.077	5.735
			AIC Adj EDF	0.031	0.998	0.060	5.577
			BIC Adj EDF	0.040	0.993	0.076	5.022
			LRT Adj EDF	0.103	0.990	0.187	5.040
			Nomi P Adj Match	0.040	0.988	0.077	5.801
			AIC Adj Match	0.031	0.995	0.060	5.561
			BIC Adj Match	0.038	0.993	0.074	4.991
			LRT Adj Match	0.101	0.990	0.183	5.005
		FST	Post P	0.008	1.000	0.016	1.153

			Post P Adj EDF	0.049	0.980	0.094	5.735
			Post P Adj Match	0.049	0.980	0.094	5.800
XOR	0.1	Merge	Nomi P	0.061	0.923	0.114	1.444
			AIC	0.064	0.915	0.120	1.710
			BIC	0.079	0.913	0.145	1.077
			LRT	0.129	0.910	0.225	1.008
			Nomi P Adj EDF	0.242	0.670	0.355	3.892
			AIC Adj EDF	0.171	0.763	0.280	3.419
			BIC Adj EDF	0.238	0.705	0.356	3.002
			LRT Adj EDF	0.487	0.610	0.542	3.530
			Nomi P Adj Match	0.256	0.710	0.377	3.619
			AIC Adj Match	0.185	0.788	0.300	3.214
			BIC Adj Match	0.248	0.733	0.371	2.859
			LRT Adj Match	0.493	0.643	0.558	3.362
		Split	Nomi P	0.066	0.918	0.123	1.605
			AIC	0.070	0.908	0.129	1.997
			BIC	0.082	0.913	0.150	1.293
			LRT	0.127	0.900	0.223	1.028
			Nomi P Adj EDF	0.224	0.685	0.338	3.863
			AIC Adj EDF	0.162	0.763	0.267	3.433
			BIC Adj EDF	0.226	0.695	0.341	3.022
			LRT Adj EDF	0.439	0.553	0.490	3.549
			Nomi P Adj Match	0.243	0.723	0.363	3.595
			AIC Adj Match	0.182	0.793	0.296	3.226
			BIC Adj Match	0.239	0.728	0.360	2.877
			LRT Adj Match	0.435	0.560	0.490	3.378
		FST	Post P	0.087	0.900	0.158	1.247
			Post P Adj EDF	0.278	0.638	0.388	3.864

			Post P Adj Match	0.290	0.673	0.405	3.596
XOR	0.4	Merge	Nomi P	0.077	0.998	0.144	1.637
			AIC	0.081	0.998	0.149	2.110
			BIC	0.090	0.998	0.164	1.226
			LRT	0.122	0.990	0.217	1.000
			Nomi P Adj EDF	0.529	0.900	0.667	5.840
			AIC Adj EDF	0.425	0.925	0.583	5.607
			BIC Adj EDF	0.462	0.905	0.612	4.940
			LRT Adj EDF	0.717	0.873	0.787	4.955
			Nomi P Adj Match	0.532	0.898	0.668	5.912
			AIC Adj Match	0.423	0.925	0.580	5.584
			BIC Adj Match	0.454	0.910	0.606	4.910
			LRT Adj Match	0.720	0.885	0.794	4.921
		Split	Nomi P	0.079	0.998	0.147	1.571
			AIC	0.084	0.998	0.155	2.153
			BIC	0.093	0.998	0.170	1.175
			LRT	0.121	0.998	0.215	1.000
			Nomi P Adj EDF	0.512	0.895	0.652	5.734
			AIC Adj EDF	0.423	0.930	0.582	5.575
			BIC Adj EDF	0.469	0.880	0.612	5.018
			LRT Adj EDF	0.721	0.868	0.788	5.037
			Nomi P Adj Match	0.515	0.895	0.653	5.797
			AIC Adj Match	0.419	0.930	0.578	5.557
			BIC Adj Match	0.467	0.885	0.612	4.983
			LRT Adj Match	0.718	0.870	0.787	4.997
		FST	Post P	0.105	0.993	0.189	1.312
			Post P Adj EDF	0.522	0.858	0.649	5.733
			Post P Adj Match	0.534	0.858	0.658	5.796

ME	0.1	Merge	Nomi P	0.005	1.000	0.011	3.380
			AIC	0.005	1.000	0.011	2.750
			BIC	0.005	1.000	0.011	1.988
			LRT	0.006	1.000	0.011	1.000
			Nomi P Adj EDF	0.008	1.000	0.016	3.890
			AIC Adj EDF	0.007	1.000	0.015	3.412
			BIC Adj EDF	0.010	1.000	0.019	2.998
			LRT Adj EDF	0.018	1.000	0.036	3.527
			Nomi P Adj Match	0.008	1.000	0.016	3.610
			AIC Adj Match	0.007	1.000	0.014	3.204
			BIC Adj Match	0.010	1.000	0.019	2.855
			LRT Adj Match	0.016	1.000	0.032	3.363
		Split	Nomi P	0.005	1.000	0.011	1.855
			AIC	0.005	1.000	0.011	2.803
			BIC	0.005	1.000	0.011	2.043
			LRT	0.006	1.000	0.011	1.000
			Nomi P Adj EDF	0.008	1.000	0.016	3.860
			AIC Adj EDF	0.007	1.000	0.015	3.427
			BIC Adj EDF	0.010	1.000	0.019	3.020
			LRT Adj EDF	0.019	1.000	0.036	3.545
			Nomi P Adj Match	0.008	1.000	0.016	3.586
			AIC Adj Match	0.007	1.000	0.014	3.215
			BIC Adj Match	0.010	1.000	0.019	2.872
			LRT Adj Match	0.016	1.000	0.032	3.377
		FST	Post P	0.006	1.000	0.011	1.380
			Post P Adj EDF	0.011	1.000	0.021	3.860
			Post P Adj Match	0.011	1.000	0.021	3.586

ME	0.4	Merge	Nomi P	0.007	1.000	0.014	2.343
			AIC	0.007	1.000	0.015	3.318
			BIC	0.008	1.000	0.016	2.088
			LRT	0.011	0.998	0.022	1.000
			Nomi P Adj EDF	0.035	0.970	0.067	5.841
			AIC Adj EDF	0.023	0.975	0.045	5.608
			BIC Adj EDF	0.027	0.968	0.053	4.942
			LRT Adj EDF	0.107	0.858	0.190	4.953
			Nomi P Adj Match	0.035	0.970	0.068	5.916
			AIC Adj Match	0.024	0.978	0.046	5.589
			BIC Adj Match	0.027	0.968	0.052	4.915
			LRT Adj Match	0.109	0.865	0.193	4.922
		Split	Nomi P	0.007	1.000	0.015	2.375
			AIC	0.007	1.000	0.015	3.523
			BIC	0.008	1.000	0.016	2.213
			LRT	0.011	0.998	0.021	1.000
			Nomi P Adj EDF	0.033	0.970	0.064	5.736
			AIC Adj EDF	0.022	0.978	0.043	5.577
			BIC Adj EDF	0.031	0.963	0.060	5.021
			LRT Adj EDF	0.114	0.868	0.201	5.035
			Nomi P Adj Match	0.034	0.970	0.065	5.801
			AIC Adj Match	0.022	0.978	0.043	5.562
			BIC Adj Match	0.030	0.963	0.058	4.990
			LRT Adj Match	0.116	0.878	0.205	4.999
		FST	Post P	0.009	1.000	0.018	1.643
			Post P Adj EDF	0.038	0.903	0.074	5.734
			Post P Adj Match	0.040	0.903	0.076	5.799

MET	0.1	Merge	Nomi P	0.108	0.920	0.194	1.402
			AIC	0.107	0.903	0.192	2.307
			BIC	0.133	0.915	0.232	1.063
			LRT	0.173	0.913	0.290	1.000
			Nomi P Adj EDF	0.406	0.680	0.509	3.891
			AIC Adj EDF	0.346	0.800	0.483	3.417
			BIC Adj EDF	0.418	0.723	0.530	3.000
			LRT Adj EDF	0.680	0.655	0.667	3.528
			Nomi P Adj Match	0.423	0.735	0.537	3.595
			AIC Adj Match	0.345	0.823	0.486	3.194
			BIC Adj Match	0.412	0.733	0.527	2.839
			LRT Adj Match	0.675	0.690	0.683	3.348
		Split	Nomi P	0.112	0.920	0.200	1.402
			AIC	0.112	0.903	0.199	2.402
			BIC	0.138	0.918	0.240	1.093
			LRT	0.174	0.910	0.292	1.000
			Nomi P Adj EDF	0.399	0.678	0.502	3.861
			AIC Adj EDF	0.348	0.803	0.486	3.432
			BIC Adj EDF	0.429	0.730	0.541	3.021
			LRT Adj EDF	0.689	0.648	0.668	3.547
			Nomi P Adj Match	0.417	0.733	0.531	3.569
			AIC Adj Match	0.348	0.823	0.489	3.205
			BIC Adj Match	0.423	0.740	0.538	2.856
			LRT Adj Match	0.673	0.675	0.674	3.362
		FST	Post P	0.144	0.900	0.248	1.256
			Post P Adj EDF	0.371	0.620	0.464	3.854
			Post P Adj Match	0.384	0.670	0.488	3.561

MET	0.4	Merge	Nomi P	0.027	0.960	0.053	2.164
			AIC	0.027	0.958	0.052	2.896
			BIC	0.031	0.953	0.060	1.832
			LRT	0.039	0.883	0.075	1.003
			Nomi P Adj EDF	0.186	0.730	0.297	5.831
			AIC Adj EDF	0.143	0.785	0.242	5.595
			BIC Adj EDF	0.150	0.713	0.248	4.924
			LRT Adj EDF	0.254	0.400	0.311	4.932
			Nomi P Adj Match	0.184	0.728	0.293	5.901
			AIC Adj Match	0.141	0.785	0.240	5.572
			BIC Adj Match	0.148	0.715	0.245	4.895
			LRT Adj Match	0.248	0.403	0.307	4.896
		Split	Nomi P	0.028	0.958	0.054	2.292
			AIC	0.028	0.955	0.054	3.010
			BIC	0.031	0.950	0.060	1.905
			LRT	0.040	0.885	0.076	1.006
			Nomi P Adj EDF	0.184	0.750	0.296	5.726
			AIC Adj EDF	0.149	0.795	0.250	5.564
			BIC Adj EDF	0.145	0.685	0.239	5.001
			LRT Adj EDF	0.254	0.400	0.311	5.011
			Nomi P Adj Match	0.179	0.748	0.289	5.788
			AIC Adj Match	0.145	0.795	0.246	5.547
			BIC Adj Match	0.139	0.685	0.231	4.967
			LRT Adj Match	0.258	0.410	0.316	4.968
		FST	Post P	0.032	0.910	0.063	1.673
			Post P Adj EDF	0.118	0.505	0.191	5.716
			Post P Adj Match	0.115	0.503	0.188	5.780

DMN1	0.1	Merge	Nomi P	0.521	0.963	0.676	2.096
			AIC	0.517	0.928	0.664	4.078
			BIC	0.576	0.948	0.716	1.810
			LRT	0.664	0.925	0.773	1.065
			Nomi P Adj EDF	0.970	0.793	0.872	5.304
			AIC Adj EDF	0.945	0.908	0.926	4.876
			BIC Adj EDF	0.962	0.800	0.874	4.179
			LRT Adj EDF	0.990	0.518	0.680	4.306
			Nomi P Adj Match	0.969	0.805	0.880	5.084
			AIC Adj Match	0.946	0.918	0.931	4.689
			BIC Adj Match	0.964	0.803	0.876	4.089
			LRT Adj Match	0.986	0.528	0.687	4.263
		Split	Nomi P	0.536	0.960	0.688	2.146
			AIC	0.528	0.918	0.671	4.223
			BIC	0.585	0.950	0.724	1.903
			LRT	0.660	0.918	0.767	1.052
			Nomi P Adj EDF	0.964	0.805	0.877	5.226
			AIC Adj EDF	0.943	0.910	0.926	4.873
			BIC Adj EDF	0.964	0.805	0.877	4.232
			LRT Adj EDF	0.992	0.490	0.656	4.352
			Nomi P Adj Match	0.965	0.813	0.882	5.010
			AIC Adj Match	0.946	0.923	0.934	4.686
			BIC Adj Match	0.965	0.815	0.884	4.141
			LRT Adj Match	0.988	0.495	0.659	4.304
		FST	Post P	0.624	0.963	0.757	1.725
			Post P Adj EDF	0.975	0.683	0.803	5.227
			Post P Adj Match	0.974	0.715	0.825	5.004

DMN2	0.1	Merge	Nomi P	0.493	0.935	0.645	1.898
			AIC	0.509	0.933	0.658	2.311
			BIC	0.557	0.923	0.695	1.588
			LRT	0.652	0.895	0.754	1.000
			Nomi P Adj EDF	0.984	0.683	0.806	5.289
			AIC Adj EDF	0.935	0.738	0.824	4.855
			BIC Adj EDF	0.957	0.713	0.817	4.167
			LRT Adj EDF	0.995	0.543	0.702	4.288
			Nomi P Adj Match	0.980	0.690	0.810	5.079
			AIC Adj Match	0.935	0.755	0.835	4.685
			BIC Adj Match	0.954	0.718	0.819	4.090
			LRT Adj Match	0.995	0.543	0.702	4.254
		Split	Nomi P	0.507	0.933	0.657	1.922
			AIC	0.523	0.930	0.670	2.344
			BIC	0.572	0.918	0.704	1.583
			LRT	0.652	0.890	0.753	1.000
			Nomi P Adj EDF	0.980	0.685	0.806	5.210
			AIC Adj EDF	0.938	0.740	0.827	4.852
			BIC Adj EDF	0.960	0.703	0.811	4.214
			LRT Adj EDF	0.995	0.543	0.702	4.334
			Nomi P Adj Match	0.983	0.695	0.814	5.003
			AIC Adj Match	0.931	0.755	0.834	4.682
			BIC Adj Match	0.957	0.708	0.814	4.138
			LRT Adj Match	0.995	0.550	0.709	4.298
		FST	Post P	0.635	0.940	0.758	1.620
			Post P Adj EDF	0.966	0.650	0.777	5.209
			Post P Adj Match	0.961	0.673	0.791	4.993

DMN3	0.25	Merge	Nomi P	0.446	0.985	0.614	1.997
			AIC	0.460	0.985	0.627	2.206
			BIC	0.525	0.983	0.684	1.870
			LRT	0.606	0.940	0.737	1.003
			Nomi P Adj EDF	0.969	0.968	0.968	3.927
			AIC Adj EDF	0.948	0.975	0.961	3.449
			BIC Adj EDF	0.970	0.960	0.965	3.025
			LRT Adj EDF	0.993	0.730	0.841	3.557
			Nomi P Adj Match	0.969	0.968	0.968	3.662
			AIC Adj Match	0.950	0.978	0.963	3.259
			BIC Adj Match	0.969	0.968	0.968	2.898
			LRT Adj Match	0.992	0.740	0.848	3.411
		Split	Nomi P	0.452	0.985	0.620	1.997
			AIC	0.467	0.985	0.634	2.218
			BIC	0.536	0.983	0.693	1.868
			LRT	0.602	0.940	0.734	1.000
			Nomi P Adj EDF	0.967	0.968	0.967	3.897
			AIC Adj EDF	0.947	0.975	0.961	3.463
			BIC Adj EDF	0.970	0.958	0.964	3.047
			LRT Adj EDF	0.990	0.730	0.840	3.575
			Nomi P Adj Match	0.968	0.968	0.968	3.637
			AIC Adj Match	0.946	0.978	0.961	3.270
			BIC Adj Match	0.969	0.965	0.967	2.915
			LRT Adj Match	0.990	0.740	0.847	3.427
		FST	Post P	0.560	0.970	0.710	1.441
			Post P Adj EDF	0.967	0.845	0.902	3.895
			Post P Adj Match	0.966	0.868	0.914	3.638

DMN4	0.25	Merge	Nomi P	0.455	1.000	0.626	2.123
			AIC	0.470	1.000	0.639	2.395
			BIC	0.518	1.000	0.683	1.970
			LRT	0.627	0.993	0.769	1.010
			Nomi P Adj EDF	0.969	0.998	0.983	3.927
			AIC Adj EDF	0.938	1.000	0.968	3.448
			BIC Adj EDF	0.969	0.998	0.983	3.026
			LRT Adj EDF	0.989	0.880	0.931	3.557
			Nomi P Adj Match	0.968	0.998	0.982	3.649
			AIC Adj Match	0.940	1.000	0.969	3.245
			BIC Adj Match	0.969	0.998	0.983	2.888
			LRT Adj Match	0.989	0.883	0.933	3.395
		Split	Nomi P	0.459	1.000	0.629	2.065
			AIC	0.472	1.000	0.641	2.423
			BIC	0.528	1.000	0.691	2.005
			LRT	0.622	0.993	0.765	1.020
			Nomi P Adj EDF	0.963	0.998	0.980	3.897
			AIC Adj EDF	0.936	1.000	0.967	3.462
			BIC Adj EDF	0.973	0.995	0.984	3.048
			LRT Adj EDF	0.989	0.880	0.931	3.573
			Nomi P Adj Match	0.965	0.998	0.981	3.624
			AIC Adj Match	0.939	1.000	0.969	3.256
			BIC Adj Match	0.971	0.998	0.984	2.906
			LRT Adj Match	0.989	0.880	0.931	3.410
		FST	Post P	0.556	1.000	0.714	1.505
			Post P Adj EDF	0.967	0.955	0.961	3.896
			Post P Adj Match	0.966	0.955	0.961	3.628

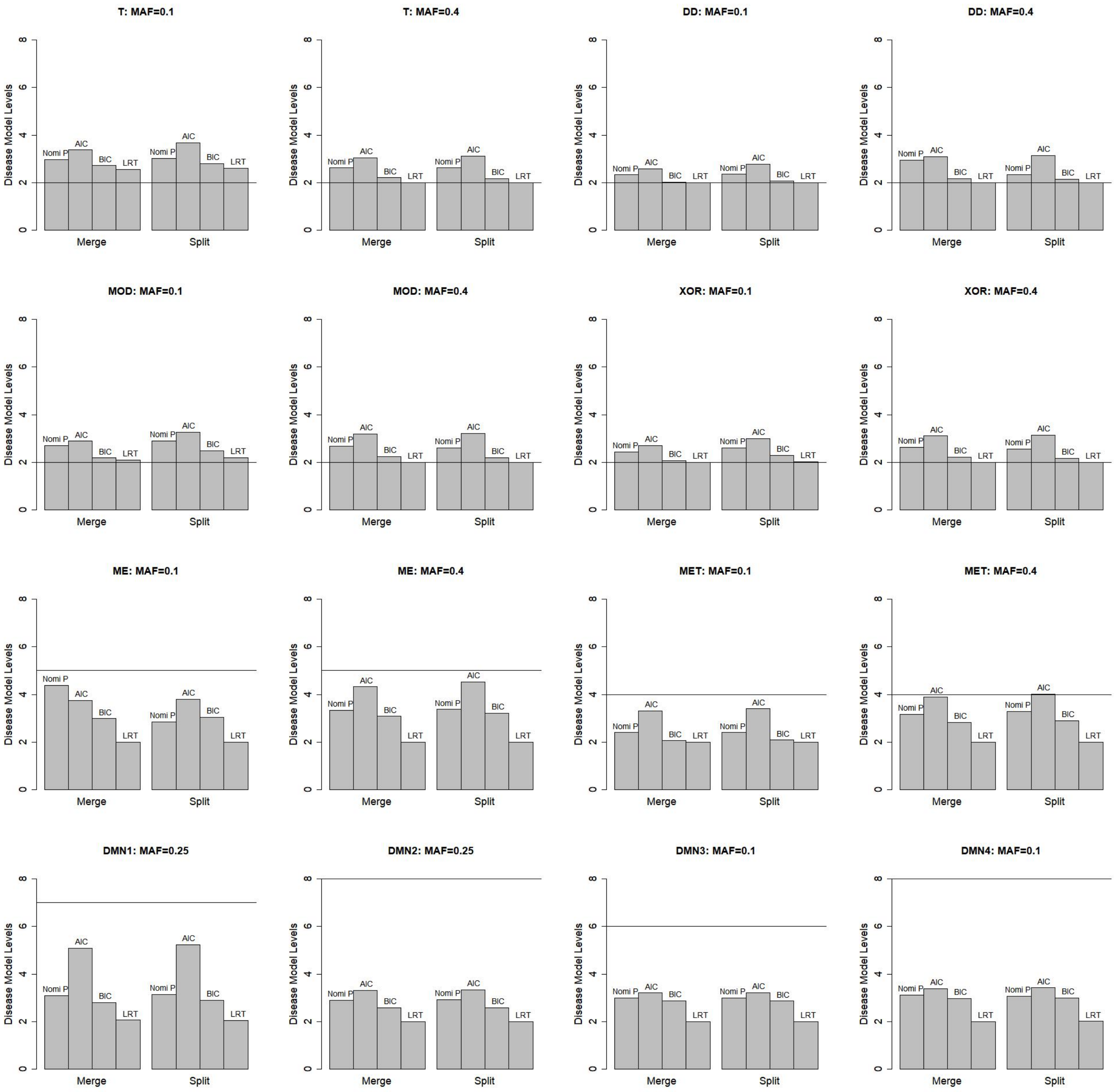


Fig C.2. Simulation results on the estimated disease model levels by different stopping criteria of the SSP and SMP methods that are calculated based on the selected true SNP-pairs. The horizontal line represents the true disease model level. Each simulation has been repeated for 400 times and the averages are reported.

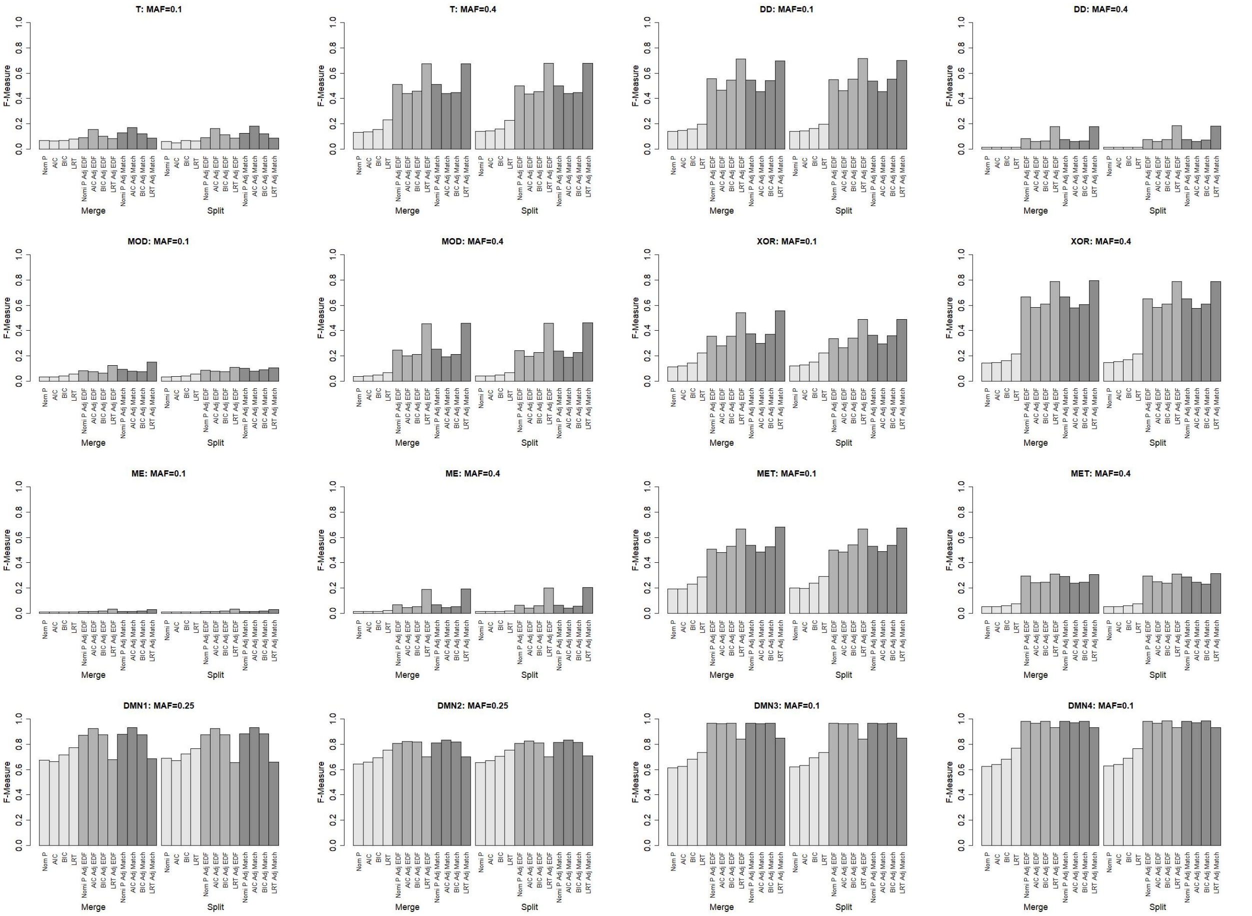


Fig C.3. Simulation results of F-measure for the SSP and SMP methods with and without EDF adjustments. Each simulation has been repeated for 400 times and the average performance is being reported. There are no observable performance difference between SSP and SMP in terms of F-measure. Using the adjusted p-value (3.7), as opposed to the nominal p-value (1.2), generally improves the detection performance, sometimes substantially. The adjustment by matched EDF or EDF from the response surface model does not show any observable difference.