The Pennsylvania State University

The Graduate School

College of Medicine Public Health Sciences

STATISITCAL MODELS FOR HIGH DIMENSIONAL SCREENING OF GENETIC AND EPIGENETIC EFFECTS

A Dissertation in

Biostatistics

by

Kirk Gosik

Doctor of Philosophy

February 2017

The dissertation of Kirk Gosik was reviewed and approved* by the following:

Rongling Wu

Distinguished Professor of Public Health Sciences and Statistics

Thesis Advisor, Chair of Committee

Vernon Chinchilli

Distinguished Professor and Chair of Public Health Sciences

Lan Kong

Associate Professor of Public Health Sciences

James Broach

Distinguished Professor and Chair of Biochemistry and Molecular Biology

Abstract

Knowledge about how changes in gene expression are encoded by expression quantitative trait loci (eQTLs) is a key to construct the genotype-phenotype map for complex traits or diseases. Traditional eQTL mapping is to associate one transcript with a single marker at a time, thereby limiting our inference about a complete picture of the genetic architecture of gene expression. Here, I present innovative applications of variable selection approaches to systematically detect main effects and interaction effects among all possible loci on differentiation and function of gene expression and other phenotypes of interest. Forward-selection-based procedures were particularly implemented to tackle complex covariance structures of gene-gene interactions. Simulation studies were performed on each of the models to assess the computational properties of each model. Applications of the models were also performed on real datasets. The first was a reanalysis of a published genetic and genomic dataset collected in a mapping population of Caenorhabditis elegans, gaining new discoveries on the genetic origin of gene expression differentiation, which could not be detected by a traditional one-locus/one-transcript analysis approach. The next dataset was of Mei Tree growth, analyzing the genetic control of the height and diameter during the developmental process. The underlying genotypes and epistasis that impact the process of these developments were considered as candidates for the selection of the procedure.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

There are several techniques used for studying genetics and mapping the results. Some of the more populor techniques include cross-breeding experiments or, in the case of humans, the examination of family histories, known as pedigrees. More recently, CRISPR/Cas9 can be used to mimic mitotic recombination to help map out genes as well. (Sadhu et al. (2016))

Construction of genetic maps are a variety of techniques used to show relative positions between genes or other sequence features of the genome and the phenotype that is controlled by such sequences. Genes are very useful markers but they are by no means ideal. One problem, especially with larger genomes such as those of vertebrates and flowering plants, is that a map based entirely on genes is not very detailed.(Brown (2006)) Genes have long areas of non-coding regions between them and therefore result in large gaps from gene to gene. This is further complicated because not every gene has allelic forms that can be easily or conventiently distinguished. With these

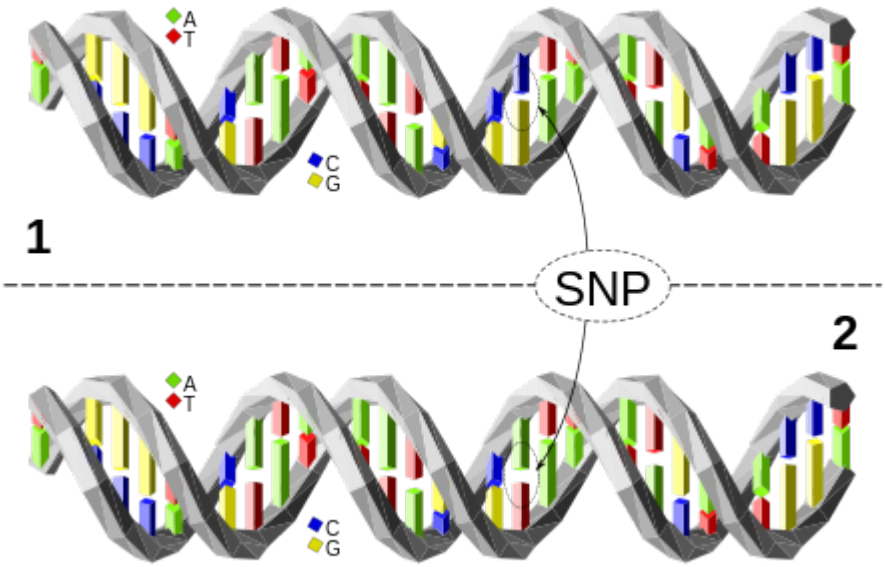

Figure 1.1: SNP Picture

considerations in mind gene maps may not be comprehensive enough and other markers may be needed.

According to brown, mapped features that are not genes are called DNA markers. As with gene markers, a DNA marker must have at least two alleles to be useful. There are three types of DNA sequence feature that satisfy this requirement: restriction fragment length polymorphisms (RFLPs), simple sequence length polymorphisms (SSLPs), and single nucleotide polymorphisms (SNPs). (Brown (2006)) The genetic markers that have been emphaised in this work are single nucleotide polymorphisms. Attempting to be at the highest levels of resolution for identifying quantitative traits, using SNPs are the most specific case. This will give exact location of the nucletid that may be impacting the genetic control over the phenotype.

There are several goals to genetic mapping and association studies that identify certain regions of the genome that contain genes involved in specifying a quantitative trait, refered to as quantitative trait loci (QTLs). One main goal is to estimate the genetic effects of these loci. The relationship

between the genetic effects of QTLs and the phenotypic value of quantitative traits can be described by a linear model (Collard et al. (2005), Xu (2007)). Typically, because of the high throughput nature of the data there are a large number of markers across the whole genome, and most of the markers may have very little or next no effect on the phenotype under study. The models can be very sparse, with most cases, the number of genetic markers or variables is bigger than the sample size, especially when interactions among markers are considered. This makes a model is oversaturated and furher model selection techniques may be required to capture the necessary information. Dong et al. (2015)

1.2 Some Exisiting Methods

Numerous methods exist and are being developed to measure and find quantitative trait loci (QTL) effects. These methods can broadly fall into three main categories. These categories are Least-Square methods, maximum likelihood and Bayesian approaches. (Wu et al. (2007)) Each method has advantages and considerations that you would need to be aware before conducting analyses to find QTL effects from the given markers. A brief discussions on a few of the methods are given to highlight some areas of consideration and how the methods proposed can handle such considerations.

Marker Regression would fall in the category of Least Squares approaches. If looking at one marker analysis general t-test and ANOVA procedures can be used to analyze the relationship. It is not recommended however for use in general practice because you do not know how dense the markers are measured. QTL interval mapping would be preferred in such an analysis because the methods take account for missing genotype data that may not have been measured. When estimating a QTL

position through maximum likelihood methods, like interval mapping, positions of other possible QTLs could affect the detection of the true position. Neighboring QTLs could possibly flatten the likelihood in instances where there are multiple QTLs on the same chromosome. This would make an effect look less significant at a given location than it actually is. Another possibility is that in the search over the interval you may find an area where the likelihood could reach a peak but could be a "ghost" QTL. This is where an effect is observed because a neighboring QTL is skewing the results at the particular position you are looking in and the result is a false discovery of the position. Marker Regression has been shown to improve interval mapping, which is call Composite Interval Mapping. This is where the QTL position found is also combined in a linear regression where the covariates are the other markers in the dataset. By including the markers as covariates the other position in the chromosome are accounted for in the analysis and false discovery is reduced.

The analysis of interval mapping and single marker analyses has shown to be effective but it limits our inference to one marker at a time as a possible loci that controls a trait. Using Marker Regression however you can incorporate multiple markers in a single analysis to test for possible QTL for a given trait. It is cautioned that running such an analysis is only an approximate test because the null hypothesis is there is no difference between the marker levels and therefore a non-mixture distribution but the alternative is a mixture of distributions. The assumptions regression would make of the errors within the marker type to be normally distributed may not be entirely met if the QTL's fall between the marker regions. However Whittaker et al. (1996) have shown that a direct regression of phenotypes on marker types, provides the same information about location of QTL-effects without having to step to all positions on the interval. With this information using the entire marker set in a regression analysis would provide a nice, computationally efficient way to map out the genetic architecture of a trait.

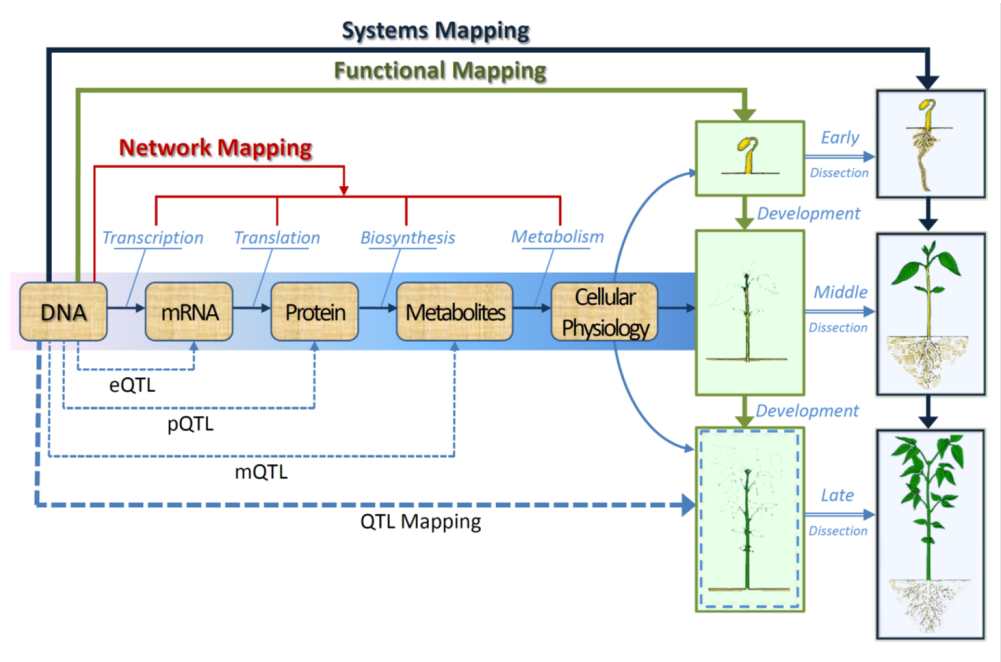

Figure 1.2: Systems Map

1.3 Chapter Overview

The main goal of this paper is to propose an improved selection procedures which use regression techniques to approach high scale variable selection problems such as the ones arising in epistatic analysis

The variable selection procedure for QTLs mapping can be seen as one of deciding which subset of variables have effects on phenotypes, and identifying out all possible effects of those markers.

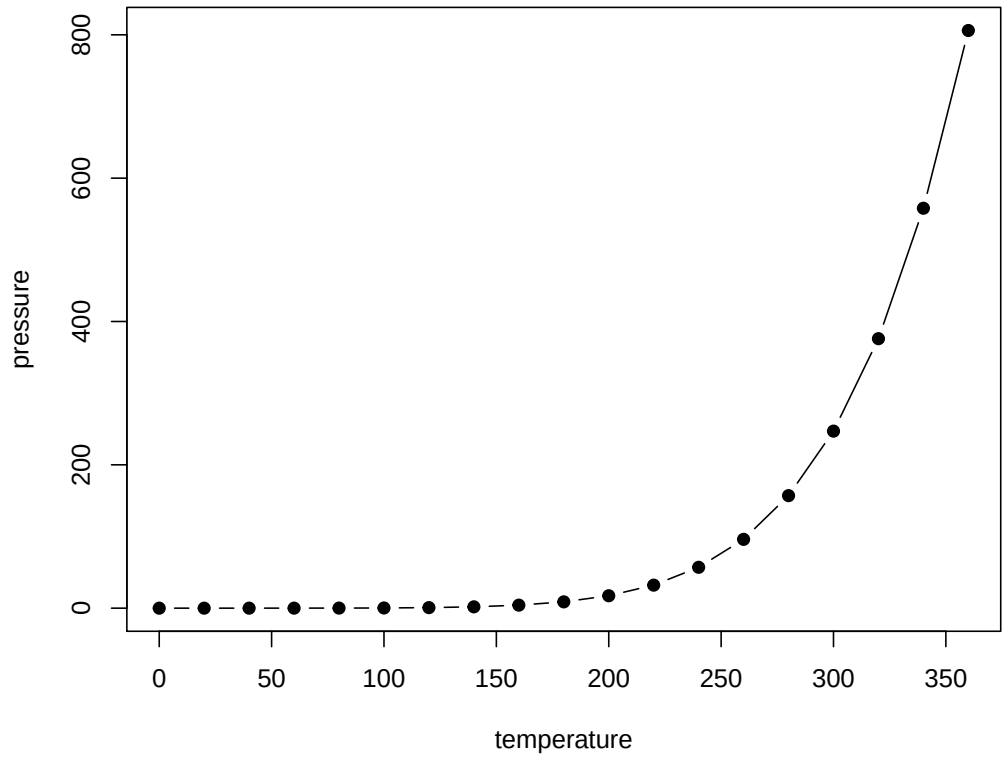

Figure 1.3: Here is a nice figure!

1.3.1 HighDeQTL

1.3.2 Higher Order Epistasis

1.3.3 iForm Funcional Mapping

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 1. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter 2.2.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))

plot(pressure, type = 'b', pch = 19)
```

Table 1.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 1.3. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 1.1.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2016) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

Chapter 2

High Dimensional eQTL

2.1 Motivation

Since activation or inhibition of gene expression causes change in phenotypic formation, the identification of expression quantitative trait loci (eQTLs) that regulate the pattern of gene expression is essential for constructing a precise genotype-phenotype map (Emilsson et al. (2008); Cookson et al. (2009); Nica and Dermitzakis (2013)). With the advent and development of various biotechnologies, it has become possible that genome-scale marker and expression data can be generated, providing an important fuel to systematically study the biological function of any types of cellular components in an organism (Kim et al. (2014); Fairfax et al. (2014); Lee et al. (2014)). Several genome-wide association studies (GWAS) have been initiated to map a complete set of eQTLs for the abundance of genome-wide transcripts whose expression levels are related to biological or clinical traits (Nica and Dermitzakis (2013);Li et al. (2013); Koopmann et al. (2014)). Statistical analysis and modeling are playing an increasing role in mapping and identifying the underlying

eQTLs from massive amounts of observed data (Kendziorski et al. (2006); Chun and Keleş (2009); Sun (2012); Flutre et al. (2013)).

A typical eQTL mapping approach is to associate a gene transcript with a single marker such as single nucleotide polymorphism (SNP). By analzying the significance of all these markers one by one adjusted for multiple testing, one can count significant loci that contribute to variation of expression by the gene. This marginal approach based on a simple regression model has been instrumental for the identifcation of eQTLs in a variety of organisms (Rockman et al. (2010); Kim et al. (2014)). However, there are two major limitations for the results by such a marginal analysis: First, it does not take into account the dependence of different markers, thus a significant association detected by one marker may be due to the other markers that are linked with it. The marginal marker analysis cannot separate the confounding effect of eQTLs due to marker-marker dependence or linkage (Wu et al. (2007)). Second, an eQTL may act through its interaction with other eQTLs and environmental factors. Because of their paramount importance in affecting complex diseases and traits, gene-gene interactions, or epistatic effects, and gene-environment interactions have been studied intensively in modern biological and medical research (Cheverud and Routman (1995); Moore (2003); van Eeuwijk et al. (2010); Mackay (2014))

These two limitations can be overcome by analyzing all markers and their pairwise interactions simultaneously through formulating a high-dimensional regression model. Although it can infer a complete picture of the genetic architecture of gene expression, this endeavor is highly challenged by the curse of dimensionality, i.e., the number of predictors far exceeds the number of observations. The past decade has witnessed the tremendous development of variable selection models for high-dimensional data analysis, such as LASSO (Tibshirani (1996)), SCAD (Fan and Li (2001)), Dantzig selector (Candes and Tao (2007)), elastic net (Zhao and Yu (2006)), minimax concave penalty

(MCP) (Zhang et al. (2010)) among others. Many methods possess favorable theoretical properties such as model selection consistency (Zhao and Yu (2006)) and oracle properties (Fan and Lv 2011). When the number of predictors is much larger than the number of observation, sure screening is a more realistic goal to achieve than oracle properties or selection consistency (Fan and Lv (2008); Wang (2009)). Sure screening assures that all important variables are identified with a probability tending to one, hence achieving effective dimension reduction without information loss and providing a reasonable starting point for low-dimensional methods to be applied.

More recently, Hao and Zhang (Hao and Zhang (2014)) extended variable selection approaches to jointly model main and interaction effects from high-dimensional data. Based on a greedy forward approach, their model can identify all possible interaction effects through two algorithms iFORT and iFORM which have been proved to possess sure screening property in an ultrahigh-dimensional setting. In this article, we implement and reform Hao and Zhang's model to map the genetic architecture of eQTL actions and interactions for gene expression profiles. This model is modified to accommodate to the feature of a genetic mapping or GWAS design in which molecular markers as genetic predictors are discrete although some additional continuous predictors can also be considered. We expand Hao and Zhang's regression model to include discrete components. Also, for an F2 or a natural population with three genotypes at each locus, we need to estimate a total of eight genetic effects for a pair of markers, which are additive and dominant effects at each locus, and additive $\times$ additive, additive $\times$ dominant, dominant $\times$ additive and dominant $\times$ dominant effects between the two loci (Kempthorne (1968)). Thus, if the number of markers is p, a total number of predictors including all main and two-way interaction terms is $2p^2$. For a typical moderate-sized mapping study, in which several thousands of markers are genotyped on a few hundred individuals, consideration of pair-wise genetic interactions will quickly make the dimension of predictors an ultrahigh one.

By modeling all markers jointly at one time under an organizing framework, the modified model can detect all possible significant eQTLs and their epistasis. An eQTL can be either a cis-QTL, coming from the same physical location as the gene expression, or a trans-QTL, coming from other areas of the genome. Our model can more precisely discern these two different types of eQTLs and their interactions than traditional marginal analysis. By reanalyzing a published data collected in a mapping population of C. elegans (Rockman et al. (2010)), the new model has validated previous results by the marginal approach, meanwhile obtained new discoveries on the genetic origin of gene expression differentiation, which could not be detected in a traditional way.

2.2 Methods

2.2.1 Experimental design

Consider an experimental population for genetic studies of complex traits, such as the backcross and F2 initiated from two inbred lines, full-sib family derived from two outcrossing parents, or random samples drawn from a natural population. These types of populations are used specifically for different species. Although they have different levels of complexities for statistical modeling, the genetic dissection of different populations underlies a similar principle. For the purpose of simplicity, we consider a backcross design in which there are only two genotypes at each marker.

Suppose the backcross contains n progeny, each of which is genotyped by p markers, such as single nucleotide polymorphisms (SNPs), distributed over different chromosomes. The number of SNPs p should be large enough to completely cover the entire genome at an adequate depth so that we can possibly capture all possible genetic variants. An increasing body of evidence suggests that

significant SNPs associated with complex traits or diseases are more likely to be eQTLs (Li et al. (2013)). Hence the identification of eQTLs is an important first step toward the genetic dissection of end-point phenotypes. For this reason, we assume that genome-wide gene transcripts have been available for the assumed study population. Assume that all progeny are recorded for the same organ by microarray, leading to expression abundance data of m gene transcripts. We purport to identify all possible genetic variants including main effects and interaction effects of SNPs that contribute to each gene transcript.

2.2.2 Adaptation of iFORM procedure

Hao and Zhang Hao and Zhang (2014) formulated an interaction forward selecting procedure under the marginality principle (iFORM). The marker and gene transcript data of the study population can be denoted as $(X_i, Y_i)(i = 1, , n)$ which are independent and identically distributed copies of (X,Y), where $X = (X_1, .., X_p)^T$ is a p-dimensional predictor vector and Y is the response, expressed by a linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \ + \beta_p X_p + \epsilon(\#eq : lin - mod) \tag{2.1}$$

The 's are the coefficients for the genetic effects of each marker. Like most genome-wide datasets, the number of markers here grossly outnumbers the number of observations, p >> n. Therefore, selection procedures would need to be implemented in order to fit a linear regression model such as (1). We are already at the point of high-dimensional data but if we want to include epistatic effects between different markers as predictors as well it would increase the amount of predictors by (pˆ2+p)/2 . The resulting linear model would grow to be,

$$Y = \beta_0 + \beta_1 X_1 + \ + \beta_p X_p + \gamma_{11} X_1^2 + \gamma_{12} X_1 X_2 + \ + \gamma_{pp} X_p^2 + \epsilon(\#eq : lin - mod2) \tag{2.2}$$

where 's are the coefficients for the epistatic effects for all the quadratic and two-way interactions between the markers. For convenience we will assume that the markers and the transcripts are standardized before running the selection procedure. Therefore, E(X_ij)=0,Var(X_ij)=1,E(Y_i)=0 and Var(Y_i)=1 for i=1,…,n; j=1,…,p. Also, the quadratic and two-way interaction effects will be centered which we will write as Z_i=(…,X_ik X_il-E(X_ik X_il),…)ˆT. By doing so we would eliminate the need for an intercept in regression model (2). This would reduce the model to the form,

$$Y = X^T \beta + Z^T \gamma(\#eq : lin - mod3) \tag{2.3}$$

Some notations that will be used to define the elements of Hao and Zhang (2014) iFORM procedure are as follows.

$P_1 = 1, 2, , p \ P_2 = (k, l) : 1klp.$ which are the index sets for the linear and two-way interactions terms, respectively. The significant main effects for the markers and their interaction effects are $T_1 = j : \beta_j \neq 0, j \in P_1, T_2 = (j, k) : \beta_{jk} \neq 0, (j, k) \in P_2$. For any model M, |M| will be used to denote the number of predictors contained in the model. The true model size would be indicated by $|T_1| = p_0$ and $|T_2| = q_0$ or together would be |T|=d_0=p_0+q_0. For the procedure, three sets will be used throughout. The sets are M for the model set, C for the candidate set of predictors and S for the solution set of predictors currently selected in the model.

There are two principles that are used in the selection procedure when considering interactions as

candidates for selection into the final model. The first is considering the principle of marginality. The principle states that it is inappropriate to model interaction terms when the main effects contributing to the interaction have either not been included in the model or are deleted because their effects become marginal by the inclusion of the interaction effect. The second principle important to the procedure is the heredity principle. The strong case of the principle states that an interaction effect should not be considered unless both the contributing main effects are in the model (Zhao and Yu (2006)). This would translate to $\gamma_{jk} 0 only if \beta_j, \beta_k \neq 0 \forall 1 j, kp$ for model (2). By including both principles during the selection process it allows for dynamically including both main effects and interactions effects. The interaction effects can only be considered between the main effects currently selected into the solution set of the model according the discussed principles. A more formal description of the procedure is given below.

2.2.3 iFORM

Hao and Zhang (2014) formulated an interaction forward selecting procedure under the marginality principle (iFORM). The procedure's initial step starts with the empty set for both the solution set and the model set, $S_0 = \emptyset$ and $M_0 = \emptyset$. The candidate set contains all main effects at the beginning, $C_0 = P_1$, for each of the markers as a possible eQTL. Typical forward selection procedures are carried out to start the selection. Each marker is tested individually using a marker regression. The marker that results in the lowest residual sum of squares is the marker selected from the candidate set into the solution set as an eQTL. This is then iterated again for a selection of another marker into the model set. Once there are at least two main effects selected into the solution set, using the strong heredity principle, the quadratic and two-way interactions are then created and placed into the candidate set as possible eQTLs for selection in the next step. This process

continues selecting main effects or the newly created interaction effects into the solution set. If another main effect is selected into the solution set, then the candidate set grows with the creation of all possible two-way interactions of the main effects that are currently in the solution set. This is continued until a designated stopping value, say d. For the number of predictors placed into the model set from the solution set the Bayesian information Criterion was used, $BIC_2(\hat{M}) = log(\hat{\sigma}\hat{M}) + n^{-1}|\hat{M}| \star (log(n) + 2 \star log(d^{\star}))$, where $\sigma^2\hat{M}$ is the sample variance for the given model, $|\hat{M}|$ is the size of the model or the number of predictors selected into the given model, and n is the sample size. The d^* term is the number of predictors in the full model. This was proposed as BIC_2 by Chen and Chen (2008) which they derived to help control the false discovery rate in high dimensional data situations. They also showed that it was selection consistent if $d^{\star} = O(n^{\xi})$ for some $\xi > 0$. The only difference between the traditional BIC calculation and the BIC_2 is the additional term involving $2log(d*)$. Ignoring the BIC, the most the number of steps in the solution path is of size n. The parameter d controls the overall length of the solution path. In practice, the exact number of predictors to include, say d_0, in the true model is unknown. We want to make d large enough to include d_0 but not so large as to fit the model to the point where it becomes oversaturated. Using the BIC_2 should help avoid such a matter as well. It is reasonable to assume that d_0 is much smaller than n in high dimensional sparse regression problems (Fan and Lv (2008)). Since this is the case, for the purposes of our model, d was set to be no larger than $n/log(n)$. Generally, the BIC_2 should reach minimum, indicating the optimal stopping point, before the designated stopping value, is reached.

2.2.4 Some considerations

There were some considerations and pre-processing steps taken before the iFORM procedure was implemented. The first consideration was to see if there were any exact duplicate markers in the dataset. One drawback that could arise with marker datasets when attempting to run multiple linear regression is the possibility of duplicate markers in the dataset. If two different markers would happen to have exactly the same genotypes for each subject it would show up as an exact linear combination of each other if both markers were to be placed in the linear model. Including redundant markers in a linear model would not add any additional information and therefore should not be included in the candidate set during the selection procedure. This also reduces the dimension slightly when there are duplicate markers in the dataset.

Another consideration made is the type of coding used for the genotypes. At any given eQTL, the jth eQTL, say, there are two possible genotypes: Q_j Q_j and Q_j q_j, making the total number of possible QTL genotypes in the population 2ˆm. The goal of a genetic model is to relate the 2ˆm possible genotypic values to a set of genetic parameters, such that these parameters are interpretable in terms of main and epistatic effects of the m eQTL. A genetic model is to use orthogonal contrast scales because it is consistent in the sense that the effect of a eQTL is consistently defined whether the genetic model includes one, two, three, or more eQTL (Kao and Zeng (2002)). The orthogonal contrasts for the genetic model can be expressed by $x_{ij} = \left[\frac{-1}{2} if homozygote Q_j Q_j, \frac{-1}{2} if heterozygote Q_j q_j\right]$ Typically in an inbred line backcross population a given genotype is coded with a 0 and 1. However there are two draws backs to this coding when considering the selection procedures discussed above. The first issue comes with not including an intercept in model (2). If this is the case each of the predictors would need to be centered making the coding to -1/2 and 1/2 instead of 0 and 1. Besides meeting the assumptions of the model that

the predictors are centered, it is also beneficial for the interaction effects as well. If the coding would remain at 0's and 1's, the interaction coding would also consist of 0's and 1's. This could propose a problem because three out of the four scenarios of epistasis between markers would result in a coding of 0 for the level in the interaction effect. This has the potential to falsely skew the data of no additive effect for interactions terms because of the sparseness of coding. By centering the coding to $(-1/2, 1/2)$, it would result in an interaction effect being coded as $(-1/4, 1/4)$. This coding would happen for different scenarios for each of the levels. The -1/4 could arise when the interaction is made up of a homozygote interacting with a heterozygote genotype. A coding of 1/4 would arise by either a homozygote interacting with another homozygote genotype, or when a heterozygote interacts with another heterozygote genotype.

2.3 Application

2.3.1 Simulation Results

Simulations studies were conducted to test the theoretical properties of the selection procedures and the results **(Tables 1 -3)1.1**. The results were compared to several other commonly used methods for eQTL mapping. In each of the examples the response was generated from model (2) with =1,2,and 3 for the random error with a sample size of $n = 200$. The X_i's were all independently and identically distributed realizations generated from $Binomial(0.5)$ and then orthogonal contrasts were made making each $x_i j(-1/2, 1/2)$. The true $= (3, 0, 0, 3, 0, 3, 3, 0_4 93)$, therefore making $T_1 = 1, 4, 6, 7$ and $p_0 = 4$. The relevant interactions were set to the pairs $T_2 = (1, 6), (1, 7), (4, 7), (4, 7)$ and $q_0 = 4$ all with $_j k = 3$ where $(j, k)T_2$. There were several methods compared during each of the simulations **(Tables 1 -3)1.1**. The methods that were used to model the data were single

marker analysis, forward selection involving only main effects (FS), forward selection involving all main effects and interaction (FS2) and the iFORM procedure. Several outcomes were evaluated to compare across each of the models. The outcomes are separated into three parts. The first part focuses on the selection of main effects, the second part focuses on the selection of interaction effects and the third part is the overall model performance. Simulations of M=100 replicates were run and the outcomes considered include

Convergence Probability (Cov) $_(m = 1)^M I(TT)/M$ Percentage of correct zeros (Cor0) $_(m = 1)^M{}_(j = 1)^p I((_j) = 0, _j = 0)/[M(p - p_0)]$ *Percentage of incorrect zeros (Inc0)* $_(m = 1)^M{}_(j = 1)^p I((_j) = 0, _j 0)/[M(p_0)]$ Exact Selection probability (Exact) $_(m = 1)^M I(T = T)/M$ *The average model size* Mean Square Error (MSE) *Adjusted R-square* Computation Time in seconds

In each instance of the simulation, the iFORM procedure was closest to the simulated data, indicated as Oracle. Single marker analysis was conducted on each of the main effects individually and the significant markers were then designated as eQTLs. When comparing the single marker analysis, we can see it rarely designated the full set of main effects as significant from the simulated data. Also, no consideration for interactions could be assessed in single marker analysis. The iFORM procedure contains the identified main effects over 90% of the time across all simulations. The procedure also includes interaction selection. The interaction screening shares a similar success rate where the interaction effects are correctly selected over 90% of the time as well. Focusing on the computation time, we observed only a few seconds, on average, increase than running single marker analysis. The final models selected by the iFORM procedure had similar adjusted R-square values as the Oracle results, on average. Looking at the exact selection percentage, we can see that the vast majority of the time the correct predictors were selected and indicated as significant each time. To compare the interaction screening effectiveness, forward selection was implemented

on both the main effects and interactions effects. The time it took to create the design matrix in order to implement forward selection was not included in the computation time. As can be seen from the results, using forward selection on the full set of main effects and pair-wise interactions took substantially longer to run on average than any of the other methods, including the iFrom procedure. Another drawback to implementing forward selection on such a large set seemed to come with over fitting the model. The selection included the maximum number of predictors allowed by the designated stopping value and did not use the BIC criteria for final model selection. This resulted in 19 additional predictors selected **(Tables 1 -3)1.1**. This increased the adjusted R-square value of the final model, however this is suspected because of over fitting the data and not to be a true prediction of the response.

2.3.2 Real Data Analysis

Rockman et al. (2010) reported an eQTL mapping study of C. elegans using 208 recombinant inbred advanced intercross lines (RIAIL) from a cross between the laboratory strain, N2, and a wild isolate from Hawaii, CB4856. Abundances of 20,000 gene transcripts were measured by microarray in developmentally synchronized young adult hermaphrodites of these lines, providing a genome-wide coverage of C. elegans from WormBase, a public C. elegans genome database. The microarray data was preprocessed through a normal–exponential convolution background correction and normalized using quantile standardization. Although they are closely related, the two strains used for the cross are considered relatively divergent for C. elegans. The two strains differ roughly at approximately 1 base pair per 900. Their RIAILs were genotyped at 1454 ordered single-nucleotide polymorphism (SNP) markers that cover the whole genome of C. elegans including five autosomes (denoted as I – V) and one sex chromosome (denoted as X).

Rockman et al. (2010) used a classic interval mapping approach to detect 2309 eQTLs by testing and scanning associations of each SNP with each gene transcript over the entire genome. Rockman et al.'s analysis allowed a rectangular map of eQTL positions gene positions to be constructed (Fig. 1), from which one can identify cis-eQTLs on the diagonal and trans-eQTLs off the diagonal. However, because their association analysis was conducted individually for each SNP, the detection of eQTLs was based on the marginal effects of individual eQTLs, which may lead to two issues being unsolved. First, of those eQTLs detected for the same gene transcript, some may include confounded effects by others. Second, the effects of genetic epistasis may take place but were not detected. By analyzing all SNPs simultaneously under a single framework, the high-dimensional model, iFORM, implemented in this study can more precisely characterize the genetic machineries underlying variation in each gene transcript. More specifically, we treat each transcript as a response with all SNP markers and their interactions as predictors by building a big regression model. Significant predictors were then selected based on the iFORM procedure. A final model including both main and interaction effects can be evaluated by calculating adjusted R-square values

Figure 2 illustrates the map of how a particular gene transcript is controlled by its eQTLs through main effects and interaction effects. For clarity of our presentation, we only chose one representative gene transcript from each chromosome. For example, gene transcript A_12_P103290 located at position 2069088 – 2069147 of chromosome I was detected to be controlled by main effects due to X2_13516256 eQTLs on chromosomes II and X4_15632637 eQTLs on chromosome IV and X2_13516256:X4_15632637 interactions between some of these eQTLs on these two chromosomes.

iFORM provides the estimates of each effect (either main effect or interaction effect), standard errors of each estimate and the significance tests of each effect. As an example, Table 4 gives the result of how gene transcript A_12_P103290 can be predicted by its eQTLs and their interactions.

It can be seen that the final predictive model (adjusted R2 = 0.896) contains 14 markers which exert their main effects and/or interaction effects on the transcript. Of the 14 final markers, a half shows significant main effects ($p < 0.05$), with several (i.e., X_14636404, X4_15568674, X4_15632637 and X_14542103) explaining about 5% heritability (defined as a proportion of genetic variance due to a predictor over the total phenotypic variance). Of these final markers, we identified eight significant epistatic interactions. Each epistasis accounts for 4.6 – 5.5% heritability (Table 4).

It is interesting to note that all predictors jointly contribute to 62.6% heritability for transcript A_12_P103290, of which main effects account for 26.7% and epistatic effects account for 35.9%. It is very surprising that epistasis contributes to more than a half of heritability. Of the eight epistatic interactions, only one occurs due to the interaction between two significant eQTLs, X_14542103 and X4_13532205 (Table 4). All the remaining is due to interactions between one significant eQTL and one non-significant marker. Some eQTLs, such as X_14542103 and X_14636404, produce epistasis with a greater frequency than others. Despite their involvement in the final predictive model, some markers were tested to be insignificant in terms of both main and interaction effects, suggesting that they regulate a gene transcript in a subtle but important fashion. In summary, iFORM can not only provide an estimate of the overall heritability of gene transcript A_12_P103290 (i.e., the sum of individual heritabilities explained by each predictor), but also chart a detailed picture of how each genetic variant contributes to transcript variation. In particular, iFORM can characterize epistasis and its role in trait control, thus equipped with a capacity to retrieve so-called missing heritabilities (Manolio et al. (2009)), a significant issue arising from current genome-wide association studies.

Through analyzing associations between all markers and each transcript by iFORM, we can identify the difference of cis- and trans-eQTLs for a particular transcript. For example, of the eQTLs affecting A_12_P103290, we detected that X1_2068168 is a cis-eQTL, whereas all others are

trans-eQTLs (Table 4). We list the number and distribution of these two types of eQTLs and the pattern of how they interact with each other to determine gene transcripts (Table 5). By detecting cis-eQTLs and trans-eQTLs, iFORM detected that genetic interactions take place mostly between trans-eQTLs.

2.4 Discussion

With the recent development of genotyping and sequencing techniques, the collection of genome-wide genetic and genomic data from any tissue of an organism has been made much easier and more efficient. Because of this, genetic studies of complex diseases or traits have developed during the past decade to a point at which we can draw a complete picture of genetic architecture for disease or trait formation and progression by genome-wide association studies (GWAS) (Mackay et al. 2009). Traditional marginal analysis based on simple regression has been instrumental for the detection of important genetic variants or quantitative trait loci in a variety of organisms, but its bottleneck has emerged quickly due to its limitation in precisely and comprehensively charting genetic control landscapes. Many GWAS studies published are bothered by missing heritabilities because of their incapacity to detect genome-wide epistasis and genotype environment interactions (Manolio et al. (2009)).

Epistasis is a phenomenon by which the influence of a gene on the phenotype depends critically upon the context provided by other genes (Cheverud and Routman (1995)). It has been increasingly recognized that epistasis is an important source for trait variation (Moore (2003); Carlborg and Haley (2004); Cordell (2009)), thus inclusion of epistasis would enhance the prediction accuracy of phenotypic performance and shed more light on the global genetic architecture of trait control

(Mackay 2014). However, epistasis is extremely hard to detect as an interaction term, whose inclusion may complicate the inference of the predictive model (Carlborg and Haley (2004); Mackay (2014)). Thanks to recent progresses in high-dimensional data modeling, we have been able to implement several cutting-edge statistical models for systematical detection and characterization of genome-wide epistasis.

Hao and Zhang (2014) proposed a new high-dimensional model, iFORM, that can tackle an issue of interaction selection simultaneously from a large pool of continuous predictors. This model is based on forward-selection-based procedures, characteristic of computational feasibility and efficiency. The authors further proved that the detection of interactions by iFORM is consistent, even if the dimension increases exponentially for a sample size. As one of the first attempts to introduce high-dimensional models into genetic studies, we modified iFORM to accommodate to the discrete nature of molecular markers. Our simulation studies indicate that iFORM can provide reasonably accurate and precise estimates of genetic main effect and interaction effects. Also, it shows greater power to detect significant genes and their interactions which may not be detected by traditional single marker analysis.

We applied iFORM to re-analyze gene expression data in an eQTL mapping study (Rockman et al. (2010)). While our results confirmed those by the traditional approach, the new model provides some new findings including new eQTLs and epistasis, thus allowing a complete set of genetic variants to be characterized. As an important tool to understand the genetic mechanisms underlying both complex traits and diseases, eQTL mapping has been widely used to identify key regulatory pathways toward endophenotype and end-point phenotypes (Schadt et al. (2005); Emilsson et al. (2008); Cookson et al. (2009); Pickrell et al. (2010); Nica and Dermitzakis (2013)). A typical eQTL study may not only include a large number of molecular markers as like in a GWAS, but also record

tens of thousands of gene transcripts throughout the entire genome. Our current version of iFORM can only take into account one gene transcript as a response at a time, thus having a limitation to model the correlation and dependence among different genes. It is our next step to formulate a multivariate multiple regression model by which to test how an individual predictor, main effect or epistatic effect, pleiotropically affects correlated expression profiles of different genes.

Given the complexity of biological phenomena, pair-wise epistasis may be insufficient to explain phenotypic variation. Imielinski and Belta (2008) argued that high-order interactions among more than two genes may provide a key pathway toward complex traits. Three-way interactions have been detected in trait control (McMullen et al. (1998); Stich et al. (2007)). A model for modeling three-way interactions has been developed in a case-control GWAS design (Wang et al. (2010)) and a genetic mapping setting (Pang et al. (2013)). It is crucial to extend iFORM to map main effect, two-way epistasis and three-way epistasis in an eQTL mapping study although no substantial change is needed in the computational algorithm, except for an enlarged test set and extra computing time. Our work is based on a backcross population in which there are only two genotypes at a locus. The backcross population can facilitate our estimation and test of genetic effects owing to a smaller number of parameters at each locus or locus pair, but its utility is very limited in the F2 design of model systems and natural populations of outcrossing species such as humans. A more general model of iFORM should consider three genotypes at each locus, which provides estimates of additive and dominant effects at each locus and four types of epistasis, i.e., additive additive, additive dominant, dominant additive and dominant dominant, between each pair of loci (Kempthorne (1968)). Each of these epistatic types may affect a phenotype through a different pathway.

With continuous falling of sequencing price, we will have desirable opportunities to study the dynamic behavior and pattern of gene expression profiles across time and space scales (Viñuela

et al. (2010); Ackermann et al. (2013)). Many previous studies suggest that gene expression during cell and organ development may follow a particular form, which can be quantified by mathematical equations (Kim et al. (2010)). For example, abundance of gene expression may change periodically in human's brain during circadian clock. Many researchers used Fourier's series approximation to model the periodic changes of gene expression by estimating the period and amplitude of the cycles (Li et al. (2013)). By integrating Fourier series into iFORM, we will be able to map dynamic eQTLs for gene expression and make a quantitative prediction of temporal and spatial patterns of genetic control by eQTLs.

Chapter 3

High-order Epistatic Networks

3.1 Motivation

Start of My Paper

Quantitative traits are very difficult to study because these traits are controlled by many genes that interact in a complicated way (Nelson et al. (2013); Mackay (2014)). Genome-wide mapping and association studies increasingly available due to next-generation high-throughput genotyping techniques have proven to be useful for characterizing gene-gene interactions, coined epistasis, that contribute to phenotypic variation (Cordell (2009); Van Steen (2011); Wei et al. (2014)). Powerful statistical methods have been developed to analyze all possible markers simultaneously, from which to search for a complete set of epistasis for quantitative traits (Li et al. (2014); Gosik et al. (2016)). The joint analysis of all markers is particularly needed to chart an overall picture of genetic interactions, in comparison with computationally less expensive marginal analysis.

Epistasis reported in the current literature is mostly due to interactions between two genes. However, a growing body of evidence shows that genetic interactions involving more than two loci play a pivotal role in regulating the genetic variation of traits (Wang et al. (2010); Dowell et al. (2010); Pang et al. (2013); Taylor and Ehrenreich (2014)). For example, in a mapping population deriving from crossing two chicken lines, three-locus interactions were detected to determine body weight (Pettersson et al. (2011)). A mapping study established by two yeast strains identified genetic interactions involving five or more loci for colony morphology (Taylor and Ehrenreich (2014)). Other studies have demonstrated that high-order epistasis is of critical importance in regulating metabolic networks in yeast (Weinreich et al. (2013)) and Escherichia coli and Saccharomyces cerevisiae (Imielinski and Belta (2008); He et al. (2010b)), whereas lower-order (pairwise) epistasis may be insufficient to explain metabolic variation for these organisms.

The theoretical models of high-order epistasis have well been established by mathematical biologists (Hansen and Wagner (2001); Beerenwinkel et al. (2007)). These models provided a foundation to interpret high-order epistasis from a biological standpoint. A few statistical models have been derived to estimate and test high-order epistasis in case-control designs (Wang et al. (2015)) and population-based mapping settings (Pang et al. (2013)). Wang et al. (2015) developed a Bayesian version of detecting high-order interactions for both continuous and discrete phenotypes. However, these models were based on a marginal analysis, thus less powerful to illustrate a global view of genetic control mechanisms due to high-order epistasis.

In this article, we deploy a variable selection procedure within a genetic mapping or association setting to characterize the genetic architecture of complex traits composed of main effects of individual genes, pairwise epistasis between two genes, and three-way epistasis among three genes. The model was built on greedy interaction screening forward selection developed under the marginality

principle (named iFORM) by Hao and Zhang (2014). These approaches, proved to possess sure screening property for ultrahigh-dimensional modeling, have been implemented to model the genetic architecture of main effects and pairwise epistasis due to eQTLs for gene transcripts (Gosik et al. (2016)). Here, we extend the implementation of iFORM to systematically capture three-way interactions that are expressed among all possible markers studied. To show the statistical power of the extended model, we performed computer simulation studies. The model was further validated through analyzing a real data of genetic mapping for shoot growth in a woody plant, mei (Prunus mume). The model should be used in any other mapping or association studies of quantitative traits.

3.2 Methods

3.2.1 Mapping and association studies

Genetic mapping and association studies are two types of designs used to dissect quantitative traits. The former is based on a controlled cross derived from distinct parents, whereas the latter samples different genotypes from a pool of accessions or a natural population. In both types of design, a set of individuals are sampled to be phenotyped for quantitative traits of interest and genotyped by molecular markers distributed throughout the entire genome. For a particular genetic experiment, the number of markers is much larger than that of samples, thus, it is impossible to estimate the genetic effects of all markers simultaneously using traditional regression models. This issue becomes much intractable when we aim to estimate genetic interactions of different orders. To tackle the issue of the number of predictors >> the number of samples, several variable selection approaches have been implemented in association studies. One approach is forward selection which was shown

to be robust for estimating pairwise interactions of predictors (Hao and Zhang 2014). With sure screening properties and controlling for false positives, this approach, named iFORM, performs very well in capturing important information in explaining the response variable. On top of these nice theoretical properties it is computationally efficient by using ordinary least squares calculations and only requiring a predetermined set up steps. Here, we extended the iForm procedure to include HGI's to capture more relevant information. In the following sections, the notation and model set-up will be introduced. After this theoretical properties will be explored. Finally simulated and real data analysis will be conducted to help confirm the theoretical properties and show the feasibility of using the model for screening across whole genomes to more precisely explain phenotypes of interest.

3.2.2 Epistatic model

Consider a linear model that underlies the true genotype-phenotype relationship. Assume that the phenotype, as the response of the model, is controlled by a set of p SNPs that act singly and/or interact with each other. These main and interaction effects of markers, i.e., the predictors of the model, need to be estimated. Let $Y = (y1, , yn)^T$ denote the phenotypic value of n samples from a mapping or association population. When considering pairwise and three-way interactions, the linear model is expressed as

$$Y = \alpha + X^T\beta + Z^T\gamma + W^T\eta + \epsilon \tag{3.1}$$

where $X = (X_1, , X_p)^T$ is the design matrix that specifies the genetic effects of each $marker = (1, , p), Z = (X_jX_k)^T(1jkp)$ is the design matrix that specifies the epistatic effects between two

markers, expressed in $\gamma, W = (X_j X_k X_l)^T (1jklp)$ is the design matrix that specifies the epistatic effects among three markers, expressed in , and $\epsilon\ N(0, \sigma^2)$ is the residual error normally distributed with mean zero and variance ˆ2. We denote the index sets for the linear, order-2 and order-3 effects in equation (1), respectively, as $P_1 = 1, 2, , p\ P_2 = (j,k) : 1jkp\ P_3 = (j,k,l) : 1jklp$ With the significant main, order-2 interaction and order-3 interaction effect sets being, $T_1 = j :_j 0, jP_1$ $T_2 = (j,k) :_j k0, (j,k)P_2\ T_3 = (j,k,l) :_j kl0, (j,k,l)P_3$

The true size of T_1, will be p_1 and similarly for T_2 and T_3 will have sizes p_2 and p_3 respectively. There will be a total of 3 sets referred to throughout the procedure, the candidate set C, the selection set S and the model set, M. The candidate set is the set of all possible predictors at a given step in the selection process. The selection set contains the predictors that have previously been selected from the candidate set from each iteration of the procedure. Finally, the model set is the final model that is fit from the selection set at the end of the procedure. The BIC is used to determine the optimal cutoff for the final model size.

3.2.3 iForm with high-order epistasis

The iForm procedure is a forward selecting procedure. In traditional forward selection the procedure starts with the empty set and then iterates through the entire set of possible predictors in C and selects the best predictor and includes it in S at the end of each step. The best predictor can be determined in many ways but usually is defined by the predictor that results in the least amount of error. For our purposes we use the residual sum of squares. This continues with selecting the best predictor from C at each step until a designated stopping criterion is met or until some information criterion is met. Common information criteria used for selecting predictors to be in M are AIC, BIC, R^2 and Mallow's C_p statistic.

The iForm procedure for high-order epistatic detection parallels the forward selection procedure, but C will grow dynamically with the creation of order-2 and order-3 interaction effects between main effects that were included from previous iterations of the procedure. There are three steps to the model selection. The first step is to initialize the 3 sets mentioned above. The sets, S and M are set to the empty set while the candidate set, C, is first set to P_1, all the main effects. The next step starts the forward selection procedure selecting predictors from C. The selected predictor will be a main effect at the first step. At subsequent steps, after interaction effects are included, selected predictors could be either be a main effect, order-two or order-three interaction effect. The final step involves repeating the second step until a designated stopping criterion is met. This can be a certain amount of predictors to be considered in the final model, or it can be based off of other factors such as the sample size. The designated stopping criterion will be denoted as d. For our purposes we use d as a function of the sample size, $d = n/log_2 a(n)$. The procedure will run up until d iterations, and the optimal model will then be constructed from the selection set. This is done by an information criterion. Here we used the Bayesian Information Criterion proposed by Chen and Chen (2008) denoted as the BIC_2. This was derived by them to control the false discovery rate in high dimensional model selections.

$$BIC_2(\hat{M}) = log(\hat{_M}^2) + n^{-1}|\hat{M}| * (log(n) + 2 * loga(d^*)) \tag{3.2}$$

Once the selection procedure is done and there are d predictors in the selection set the BIC is used to determine the cutoff value for the optimum number of predictors in the model set. Then linear regression is performed on the model set.

Two guiding principles are used to help dynamically select the main effects and epistasis effects

throughout the procedure. The first is the marginality principle, which states that an effect will not be removed from the model once it has been selected. A previous selected effect may become marginal by the inclusion of subsequent effects. This especially can be the case when an interaction effect is included. One of the parent effects may become less significant or even not significant at all by considering both in the model. The next principle we state as the heredity principle but has also been referred to in other work as the hierarchy principle (Bien et al 2013 and Lim and Hastie 2014).

The heredity (hierarchy) principle help reduce the search space by making the assumption that previously selected main effects would be involved in the interaction effects. By considering this principle it substantially reduces the search space making this feasible for ultra-high dimensional situations. Even larger than ram datasets can be used with efficient memory mapping of the dataset while running the procedure. The weak version of the heredity principle for three-way interactions states that at least one of the main effects needs to be selected into the model to consider an interaction effect that contains that predictor. Considering a moderately high set of predictors say p = 5000, if trying to include all order-2 interactions upfront, will make the candidate set be as high as 12,498,000. This alone could exceed most ram requirements of standard computers. This is before even stepping up to order-3 interactions. The weak heredity principle would decrease the candidate set substantially. Assuming a sample size of $n = 200$, would give a cut off of $n/log_2(n) = 200/log_2(200) = 26$ steps in the procedure. This would give a maximum of approximately 135,000 candidate predictors. This gives a 100 fold decrease in the candidate set. This could substantially make ultra-high dimensional analysis more feasible and also speed it up in the process. This is the weak case. If considering the strong case the decrease in candidate space is even more apparent. Aside from the efficiency by lowering the search space of the candidate set, the heredity principle is usually taken into account by researchers when selecting models involving

the consideration for interaction effects.

3.2.4 Theoretical Properties

The theoretical properties of the iForm procedure with high-order epistasis follow closely with the forward selection procedure. Hao and Zhang (2014) summarize forward selection nicely as follows. At each step, the response is regressed on the most correlated covariate, and the residual is calculated and used as the new response in next step. After the most correlated covariate (say, X_1) is selected, all other covariates are regressed on X_1, and then the covariates are substituted by the corresponding normalized residuals, which are used as the new covariates in next step. By viewing forward selection in this sense the computational complexity of the procedure depends upon the size of the candidate set. The candidate set in the iForm's case does grow dynamically at each step, by at most the number of predictors currently selected in C for each step. If we denote the current size of the candidate set as m then each iteration of the procedure grows with complexity of O(nm), where n is the sample size. Leaving the selection unrestricted we would not be able to fit more than n predictors for a linear model and therefore n would be the most main effects that would be able to be selected. Considering the weakest form of the heredity principle at the current iteration there would be at most $p + (n(n-1)(n-2))/6$ predictors in the candidate set. This would make the total complexity of the selection procedure to be $nO(n(p + n(n-1)(n-2))) = O(n^3p + n^5)$. This makes the total complexity grow linearly as p grows.

The theoretical properties of the iForm procedure show sure screening properties (Fan and Lv (2008)). By this we mean that all the import predictors, whether that is a main effect or epistatic effect will be selected with probability tending to 1. This is important to capture as much of the signal as possible through all the noise that comes with $p >> n$ or ultra-high dimensional situations.

It is also important not to 'over-fit' the model with unnecessary predictors that actually explain more noise in the data that the model is being fitted on than the actual signal you would like to pick up on.

To show the property from above the following conditions would need to be met for regulatory purposes. Hao and Zhang (2014) showed how under these conditions sure screening properties for interaction models like FS2 and iForm are satisfied. This also applies to order-3 interaction models like FS3 and iForm with higher order epistasis, like we do with the high-order epistasis model. The following assumptions need to be met for these conditions. The first is that the $X = (X_1, , X_p)^T$ are jointly and marginally normal with independent normally distributed error. Next we would need the eigenvalues of the covariance matrix to be positive and bounded by two constants $0 < \tau_m in < 1 < \tau_{max} < \infty$, such that $\sqrt{\tau}_m in < \lambda_m in(\Sigma)\lambda_m ax(\Sigma) < \sqrt{\tau}_m ax/4$. Also, the genetic effects, β need a certain level of signal strength. This we would assume to be $|(|\beta|)|C_\beta$ for some positive constant C_β and $\beta_m in \nu\beta\eta^{-\xi_m in}$,with $\beta_m in = mina(\beta)$. Lastly, there needs to be a certain level of sparsity to the number of important effects. Denoting the total number of important effects as d_0, and positive constants ξ, ξ_0 and ν we would need $log(p)\nu n^{\xi}, d_0\nu n^{(}\xi_0)$ and $+ 6_0 + 12_m in < 1/2$. The conditions stated are accepted standards in the literature when studying ultra-high dimensional situations. (Hao and Zhang (2014) , Fan and Lv (2008); Sun et al. (2013)).

3.3 Application

3.3.1 Simulation Studies

To study the numeric properties of the selection procedure, simulation studies were conducted. To capture relevant data structures, there were several different scenarios considered. For each scenario 50 predictors were generated with a sample size of 300 observations. The data was split into training and a testing set to study both the fitted properties of the model as well as the generalizability of the model. There were a variety of metrics obtained to assess the suitability of each model utilized in the simulations. The first metrics that were taken into account were the rates for the true positives, false positives, true negatives and false negatives. Since we have a variety of levels to each of the models each of the rates were evaluated for the different hierarchical levels. Some of the models only have main effects and/or two-way interactions, therefore the rates were only given for the area applicable to model and the rest were reported as NA. The generalizability of the models was also assessed by withholding 100 random observations as a test set. All the data was generated from the same scenario and then 100 of the observations were randomly selected and stored for out of sample measures. The data was generated from the given scenario and randomly split before assessing the models. The exact same training and testing sets were used to fit and assess each of the models in order to make as fair of a comparison as possible. Each scenario was replicated 100 times and measures were averaged over all replicates. The two measures assessed were mean square error and the coefficient of determination. The analogous in-sample measures were also calculated for comparison. The models being compared in the simulation studies are Forward Selection, Forward Selection with all order-2 interactions (FS2), Forward Selection with all order-2 interactions (FS3), iForm strong heredity order-2, iForm weak heredity order-2, iForm

stong heredity order-3, iForm weak heredity order-3, Glinternet (Bien et al. (2013)), and finally hierNet (Lim and Hastie (2015))

Covering a variety of settings the following scenarios were evaluated and compared. The first is where the data were generated from the interactions of the model follow a strong heredity (hierarchy) with $\sigma = 1$. The second, the data is generated to have the interactions in follow a weak heredity (hierarchy) with $\sigma = 1$. The third scenario is anti-heredity (hierarchical) where the interaction effects are only among predictors not present as main effects in the model. Finally the last scenario on generates data that come from pure interactions between predictors with no main effects present in the model used to generate the data.

For the first scenarios where the truth obeys strong heredity where all of the parent main effects need to be selected before interactions are selected. The models that appeared to do the best in this simulation were forward selection on all order-3 interactions included from the beginning (FS3), iForm order-3 weak heredity and iForm order-3 strong heredity (Table 1). The FS3 took over a 40 fold increase in time to run. The other comparison models, glinternet and hierNet seemed to perform well on the training set but not as well on the testing set. This would indicate that some overfitting was occurring with those types of regularization models. The next scenario was when the truth obeys weak heredity. With the underlying model obeying the weak heredity, the iForm order-3 strong heredity version dropped off in performance slightly. However, the FS3 and iForm order-3 remained as top performers (Table 2). The third scenario assessed was from an underlying model with an anti-heredity structure. Both main effects and interaction effects were used in the model to generate the data. However the interactions included in the model were of combinations of main effects in the candidate set, that were not in the model. The iForm seems to drop in performance with this scenario (Table 3). This is to be expected because it is in direct violation of

the underlying assumptions of the model hierarchy. Even with these violations of the heredity it still performed reasonably well. Lastly, making the scenario a little more extreme, the underlying model generating the data was only of interactions. There were no main effects included in the model. The results of this scenario are shown in Table 4. Performance appeared to drop off for all models explored in the simulation.

In the scenarios where the data was assumed to follow some form of a hierarchical structure for the epistasis effects the iForm procedure for higher-order epistasis effects appeared to perform the best. Not only did it result in selecting the correct model, the false positive rate was also among the lowest. The out of sample error was also among the lowest between each of the models compared. With the procedure using OLS calculations, it also performed the fastest out of the models including epistasis effects. All of the combined show the promise of the iForm procedure for GWAS type studies. With the other scenarios, the underlying structure of the data does not follow a typical intuition about the structure of data in biology.

3.3.2 Worked Example

We validated the biological usefulness of the model by analyzing a mapping data for a woody plant, mei (Prunus mume). Originated in China, mei has been cultivated for its ornamental flowers for thousands of years (Sun et al. (2013), Sun et al. (2014a)). Its many desirable properties, such as cold-hardiness, colors and flavors, are appraised as a symbol of persistence and beauty in Chinese culture. Recent sequencing of its genome has made it an ideal model system to study the genetics and evolution of woody plants (Zhang et al. (2013)). To improve the growth rigor and form of mei important to its ornamental value, a cross was made between two distinct cultivars, Fenban (female parent) and Kouzi Yudie (male parent), aimed to select superior genotypes from hybrids.

To the end, an F1 mapping population of 190 hybrids was established and further genotyped for 4,934 SNP markers over eight mei chromosomes.

To test genotypic differences in growth performance, each of these hybrids was grafted on an established root stock using multiple budding scions. Next spring, buds on the scions sprouted into shoots. The lengths and diameters of 10 randomly selected shoots were measured once every two weeks during an entire growth season from March to October. It was found that both shoot length and diameter growth was well fitted to the three-parameter growth equation expressed as

$$g(t) = a/[1 + b * exp(-rt)] \tag{3.3}$$

where $g(t)$ is the amount of shoot growth at time t, a is the asymptotic value of growth when time tends to be infinite, b is a parameter that reflects the amount of growth at time 0, and r is the relative growth rate. These three parameters determine the overall form of growth curve jointly, although they function differently. Thus, by estimating these parameters for individual hybrids using a nonlinear least squares approach, we can draw the growth curve of each hybrid. Differences in growth curve among hybrids may be controlled by specific genes or quantitative trait loci (QTLs). Although tremendous efforts have been made to map growth QTLs and their epistasis (Ma et al. (2002b);Wu and Lin (2006); Li and Sillanpää (2012)), none has characterized the contribution of high-order epistasis although it has been thought to regulate growth processes.

By treating the estimates of growth parameters for individual hybrids as "phenotyopic traits," we used iFORM to map growth QTLs and QTL-QTL interactions. Of 4,934 markers, 00 are the testcross markers at which markers are segregating due to only one heterozygous parent and 00 are the intercross markers whose segregation results from the heterozygosity of both parents. For

a testcross marker, there is only one main genetic effect, whereas an intercross marker contains additive and dominant main effects. Thus, a pair of testcross markers produces only type of epistasis, but a pair of intercross markers forms four types of epistasis, additive × additive, additive × dominant, dominant × additive and dominant × dominant. For two markers with one from the testcross and the other from the intercross, there are two types of epistasis, i.e., additive × additive and additive × dominant (Tong et al. (2011)). The number and type of epistasis can be characterized for any three markers accordingly. Here, the iFORM was implemented in a way that allows both marker markers to be modeled and analyzed simultaneously.

To demonstrate the possible importance of high-order epistasis, we analyze the data by assuming that growth parameters are controlled by low-order epistasis only and by both low- and high-order epistasis, respectively. The weak heredity (hierarchical) was used to screen every SNP and possible interaction of the main effects selected and the rest of the SNPs left in the candidate set. It was not restricted to the strong case where both main effects had to be in the model for the interaction to be considered. For the pairwise epistatic model, this grew the candidate set to almost 20,000 predictors to choose from. It turned out that 5 predictors were chosen, i.e., four main additive effects of markers, AATTC_nn_np_2517, AATTC_nn_np_2815, CATG_nn_np_3479 and CATG_nn_np_1284 and one epistatic effect due to markers AATTC_nn_np_2815 and AATTC_lm_ll_3034, for growth parameter r of shoot length (Table 4). The main effect of marker AATTC_lm_ll_3034 was detected to be insignificant. These main and epistatic effects together explained 32.41% of the total variance of parameter r.

When opening up the iForm procedure to the possibility to creating higher order interactions to be placed into the candidate set, a more complete picture of the phenotypical variation was revealed. The amount of predictors included in the final model grew to 12, with one of them

being three-way interactions among markers AATTC_nn_np_2815, AATTC_lm_ll_3034 and AATTC_nn_np_1615.The adjusted R2 jumped up to over 70% (Table 4). This astonishing jump in predictive power is an exemplar case as to the importance of higher-order interactions in genetic models. Not only did higher-order interactions become one of the most significant predictors in the model selected, it also allowed for other order-two interactions and main effects to be kept in the model that were previously left out. At the next step of every iteration, the new candidate effect was conditioned on everything previously selected. With the conditional effect of the higher-order interaction it enabled for other lost effects to be modeled as well.

The purpose of the mei genetic project is to study the genetic control of shoot growth form. Here, we further analyze how three-way interactions detected by our model affect growth form. Assume that there are three testcross markers, A (with two alleles A, a), B (with two alleles B, b), and C (with two alleles C, c), which interact jointly to affect shoot growth. The three markers form eight genotypes AABBCC, AABBCc, AABbCC, AABbCc, AaBBCC, AaBBCc, AaBbCC and AaBbCc whose genotypic means at time t are partitioned into different components, respectively, expressed as

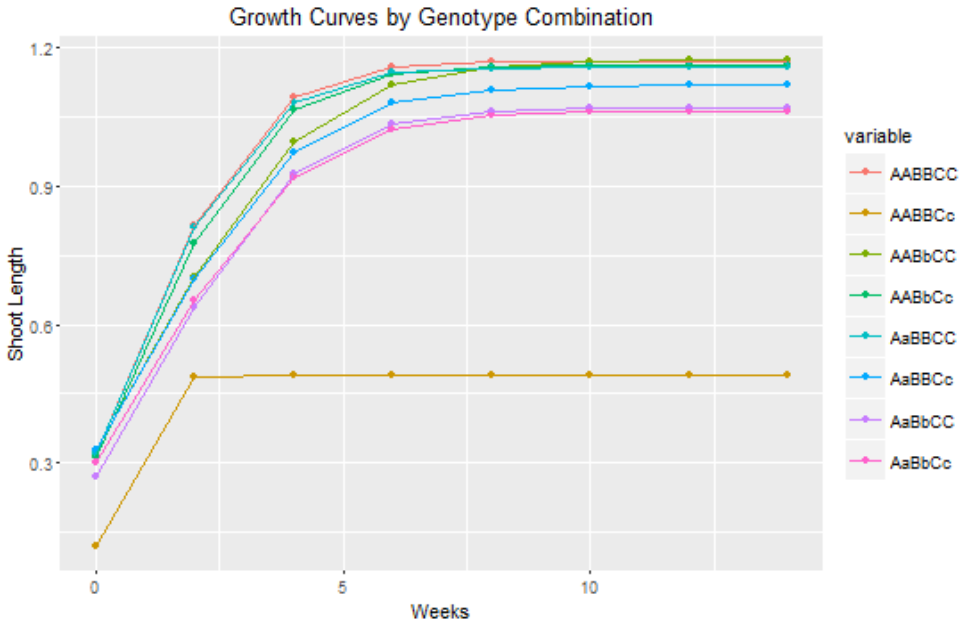

Figure 3.1: Growth Curve Comparison

$$\mu_{111}(t) = \mu(t) + \alpha_1(t) + \alpha_2(t) + \alpha_3(t) + i_{12}(t) + i_{13}(t) + i_{23}(t) + i_{123}(t)$$

$$\mu_{112}(t) = \mu(t) + \alpha_1(t) + \alpha_2(t)\alpha_3(t) + i_{12}(t)i_{13}(t)i_{23}(t)i_{123}(t)$$

$$\mu_{121}(t) = \mu(t) + \alpha_1(t)\alpha_2(t) + \alpha_3(t)i_{12}(t) + i_{13}(t)i_{23}(t)i_{123}(t)$$

$$\mu_{122}(t) = \mu(t) + \alpha_1(t)\alpha_2(t)\alpha_3(t)i_{12}(t)i_{13}(t) + i_{23}(t) + i_{123}(t)$$

$$\mu_{211}(t) = \mu(t)\alpha_1(t) + \alpha_2(t) + \alpha_3(t)i_{12}(t)i_{13}(t) + i_{23}(t)i_{123}(t)$$

$$\mu_{212}(t) = \mu(t)\alpha_1(t) + \alpha_2(t)\alpha_3(t)i_{12}(t) + i_{13}(t)i_{23}(t) + i_{123}(t)$$

$$\mu_{221}(t) = \mu(t)\alpha_1(t)\alpha_2(t) + \alpha_3(t) + i_{12}(t)i_{13}(t)i_{23}(t) + i_{123}(t)$$

$$\mu_{222}(t) = \mu(t)\alpha_1(t)\alpha_2(t)\alpha_3(t) + i_{12}(t) + i_{13}(t) + i_{23}(t)i_{123}(t) \tag{3.4}$$

where $\mu(t)$ is the population mean at time t; $\alpha_1(t)$, $\alpha_2(t)$, and $\alpha_3(t)$ are the genetic effects of markers A, B and C at time t, respectively; $i_{12}(t)$, $i_{13}(t)$, and $i_{23}(t)$ are the pairwise epistatic

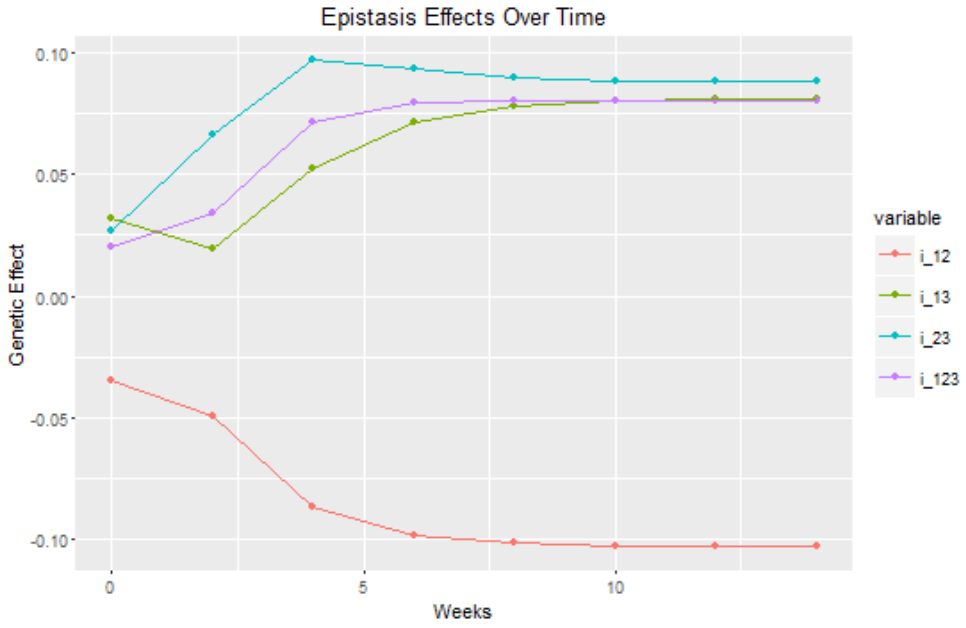

Figure 3.2: Epistasis Comparison

effects between markers A and B, A and C and B and C at time t, respectively; and $i_{123}(t)$ is the three-way epistatic effect among three the markers at time t. From the above equations, we solve the pairwise and three-way epistatic effects as

$$
\begin{aligned}
i_{12}(t) &= [(\mu_{111}(t) + \mu_{112}(t) + \mu_{221}(t) + \mu_{222}(t))(\mu_{121}(t) + \mu_{122}(t) + \mu_{211}(t) + \mu_{212}(t))] \\
i_{13}(t) &= [(\mu_{111}(t) + \mu_{121}(t) + \mu_{212}(t) + \mu_{222}(t))(\mu_{112}(t) + \mu_{122}(t) + \mu_{211}(t) + \mu_{221}(t))] \\
i_{23}(t) &= [(\mu_{111}(t) + \mu_{122}(t) + \mu_{211}(t) + \mu_{222}(t))(\mu_{112}(t) + \mu_{121}(t) + \mu_{212}(t) + \mu_{221}(t))] \\
i_{123}(t) &= [(\mu_{111}(t) + \mu_{122}(t) + \mu_{212}(t) + \mu_{122}(t))(\mu_{112}(t) + \mu_{121}(t) + \mu_{211}(t) + \mu_{222}(t))] \qquad (3.5)
\end{aligned}
$$

Each genotype can draw a growth curve using its growth parameters (a, b, r) estimated from raw data, from which we can chart the curves of pairwise and three-way epistatic effects using equation (4). Three markers AATTC_nn_np_2815, AATTC_lm_ll_3034 and AATTC_nn_np_1615 that

produce a significant three-way interaction for parameter x of shoot length display pronounced differences in growth curve (3.1). The epistasis of low- and high-order performs differently to affect growth form, with three-way interactions playing a more remarkable role than pairwise epistasis (3.2).

3.4 Discussion

Genetic interactions have been thought to contribute to a significant portion of genetic variance for quantitative traits of critical importance to evolutionary biology, agriculture and medicine (Nelson et al. 2013; Mackay 2014). While pairwise interactions have been a major focus of quantitative genetic studies, there has been growing evidence that genetic interactions involving three or more loci play an important role in affecting the phenotypic differentiation of traits (Wang et al. 2010; Dowell et al. 2010; Pettersson et al. 2011; Pang et al. 2013; Weinreich et al. 2013; Taylor and Ehrenreich 2014; Taylor and Ehrenreich 2015). Because of its complexity due to a network of interactions, the detection of high-order epistasis is extremely difficult (Mackay 2014). More importantly, interpretation of high-order epistasis and its contribution to overall genetic architecture can be better made by jointly analyzing all possible low- and high-order interactions among genes. This has added an extra challenge to statistical modeling and detection of this important phenomenon. Thanks to the recent development of statistical models for high-dimensional variable selection, we have reformed a statistical modeling framework for detecting high-order epistasis by focusing on three-way interactions.

Our model extends Hao and Zhang's (2014) forward selection-based algorithm iFORM that has proven to be robust and efficient for computing and detecting two-way interactions between pre-

dictors (including continuous predictors). A favorable property of iFORM is its capacity to detect interactions even if the dimension of predictors is extremely high relative to a sample size used. The fundamental assumption used by iFORM is the heredity principle, i.e., the existence of interactions between a pair of variables that each has at least weak main effects. After extending it to characterize three-way interactions, this assumption can be relaxed for the third variable; i.e., even if there is no detectable main effect for the third marker, then extended iFORM can still detect the three-way interaction. This property may explain the reason why high-order epistatic model outperforms low-order epistatic model, as demonstrated from the detection of significant genetic interactions in a real data of a woody plant, mei (Prunus mume). It was found from a recent study that loci participating in high-order genetic interactions may not individually have measurable effects (Bloom et al. 2013). As a result, our model can be used as a general tool to detect genetic interactions of various orders and, therefore, elucidate the overall picture of genetic architecture by capturing the so-called missing heritability.

The model was investigated by simulation studies whose result help users to determine an optimal design of mapping or association studies in terms of sample size, phenotyping precision and the number of markers. Its application to P. mume genetic mapping leads to the detection of key loci and their interactions expressed at the low- and high-order levels for the growth form of shoots. …. the R packages was created and made available through CRAN (Comprehensive R Archive Network)[https://cran.r-project.org/]. We packed iFORM/eQTL in R with the source code is available at Center for Statistical Genetics (website)[http://statgen.psu.edu/software/]

Chapter 4

iForm Functional Mapping (A computational method)

4.1 Motivation

As we have seen and also has been noted by several researchers while conducting biometric analysis (Jinks and Mather (1982); Hill and Mackay (2004); Wu (1996)) or molecular dissection (Mackay et al. (2009), Park et al. (2010)) is that quantatitve traits are very complex and much is still needed to be learned. The researchers cited note that the traits are most likely polygenic, including gene-gene interactions and other sources of interaction effects. (Cheverud and Routman (1995); Moore (2003); van Eeuwijk et al. (2010); Mackay (2014)) Higher order interactions of complex traits are not well studied because of their difficulty to detect in mapping studies as well. The lack of data shoud not be construed as proof that this order of interaction does not exist. (Taylor and Ehrenreich (2015)). The difficulty in detection leads a way for new computational methods to be

developed and approaches to describe how to distinguish such effects. As noted in chapter 3, new theoretical models of high-order epistasis have well been established by mathematical biologists (Hansen and Wagner (2001); Beerenwinkel et al. (2007)). These models provided a foundation to interpret high-order epistasis from a biological standpoint. A few statistical models have been derived to estimate and test high-order epistasis in case-control designs (Wang et al. (2015)) and population-based mapping settings (Pang et al. (2013))

Growth and developmental traits are mostly better described by a functional process (Hernandez (2015), Muraya et al. (2017)), it is more biologically meaningful to map these traits as growth curves (Sun and Wu (2015)). There have been a few different approaches that have integrated growth equations into genetic mapping via the likelihood function, leading to the birth of a so-called functional mapping models (Ma et al. (2002a); Wu and Lin (2006); Li and Sillanpää (2015); Muraya et al. (2017)). These style of appoaches can allow for the developmental change of genetic control to be characterized across both time and as well as space (He et al. (2010a); Li and Wu (2010)). Treating the phenotype as a complex trait it would be likely it would follow a more functional or dynamic process. This information could be lost or greatly limited by treating the response as a single static predictor. Modeling the longitudinal structures in this fashion, functional mapping has proven to be of great statistical power in gene identification and the utilization of sparse phenotypic data (Hou et al. (2006)). In an attempt to capture all relevant information and be as parsimonuous as possible principles from biophysical and biochemical processes were considered. The logistic growth equations are both biologically relevant (West et al. (2001); Sun et al. (2014b)) and have few parameters that can be mapped to growth QTLs by estimating these parameters for each genotype and interactions between genotypes.

There are many approaches for gene mapping with genome-wide association studies (GWAS) being

one of the most popular one, achieving a considerable success since their first publication in 2005 (\cite{klein2005complement). Analytical approaches are constantly being developed to perform GWAS studies. There are a few areas of challenges in statistical modeling and analysis of genetic data that account for the complexity of phenotypic information. Generally GWAS studies associate genetic markers with static, single valued phenotypes. As we have discussed, most analysis revolve around pointwise estimates and do not always take the entirety of the system during the analysis. Incorporating selecitons are starting to become more common but further work is this area still needs to be explored. Extending the forward selecting procedure previously state in 2 and 3 in order to handle a functional phenotype would be very beneficial with GWAS level studies. A few challenges do arise while considering to conduct a genome-wide association study (GWAS) on interacting traits measured at a sequence of time points. The model needs to be flexible enough to fit different situations, independence of the error structure needs to be maintained or accounted for with the time dependencies and finally computational efficiency needs to be good enough to fit such complex models. All of these issues combined make it a difficult exertion to take on but with computational power increasing, it is becoming more feasible to handle.

In applications like the scenario described where we have a functional value phenotype and a high dimensional predictor space with dynamically considering interaciton effects it may be too restrictive to suppose that the effect of all of the predictors is captured by a simple linear fit. Reframing the regression problem to help code in the longitudinal data into the structure of the data in a biologically meaningful manner and making some sparsity assumptions about the number of significant genetic and epistatic effects that affect the phenotype will help in the developement of such a model to tackle such a task.

4.2 Methods

4.2.1 Regression by linear combination of basis functions

One common approach to regression problems is to frame the model as a linear combination of basis functions. In typical multiple regression the design matrix would be the values of the observed predictors and these would be used to fit the model, usually with a least squares approach. The goal then being to fit the expected value of the phenotype of interest in terms of the values of the predictors. This would result in a linear model of the form,

$$Y = \beta_0 + \beta_1 X_1 + \ + \beta_p X_p + \epsilon(\#eq : lin - mod) \tag{4.1}$$

This model is nice for a single response but can be too restrictive at times. With a functional response over time, having a model with more flexibility could more accurately estimate the phenotype espeically when considering a functional phenotype like a growth model. A fit like the one mentioned would only restrict growth to be a straight line and that may not be applicable in real world applications. By treating the problem as linear combination of basis functions, the general form would look like,

$$f(x) = \sum_{i=0}^{P} \theta_i \phi_i(x) \tag{4.2}$$

where the ϕ are the basis functions of the researchers choosing. Under this format you can choose any function that would fit the need of the given problem and has relevance to the application area.

A common choice is to use polynomial regression, where /*phi* would be the predictors raised to different degrees in order to invoke a non-linear relationship into the model. This works well but it comes with some draw backs. The first being that for each degree consider, it could grow the predictor set even larger. Instead of just one effect for each predictor you could have up to the order of the polynomial effects for each predictor. With the predictor set being at a high dimensional level already, this may not be something feasible to do. The other area of concern is that it would give a way for higher correlation between effects in the model. This would violate the intial assumptions of the model.

Standard polynomial regression is just one case of using basis functions in linear regression. There are many transformations that are able to be performed to invoke nicer properities to the data. The basis functions that are going to be focused on in this work are orthogonal polynomials. This would be a special case of polynomial regression that would alleviate some of the drawbacks mentioned above. Orthogonal polynomials by definition are othogonal to each other and therefore would not have any correlation between predictors when used as basis functions. Also as an advantage, polynomial regression can be used to make similar types of interest as other types of multiple regression analysis. It does this while modeling a non-linear relationship between the phenotype and genetic markers without having to use complex optimization methods. Ordinary leasat squares would still apply in this framework, making it more computational effecient as well. One specific class of othogonal polynomials that will be used are the Legendre polynomials because of the nice properites they posses.

4.2.2 Legendre Polynomials

```
library(ggplot2)

library(data.table)

library(RColorBrewer)



Legendre<-function( t, np.order=1,tmin=NULL, tmax=NULL )

{

  u <- -1

  v <- 1

  if (is.null(tmin)) tmin<-min(t)

  if (is.null(tmax)) tmax<-max(t)

  nt <- length(t)

  ti    <- u + ((v-u)*(t-tmin))/(tmax - tmin)

  np.order.mat <- matrix(rep(0,nt*np.order),nrow=nt)

  if(np.order >=1)

    np.order.mat[,1] <- rep(1,nt)

  if (np.order>=2)

    np.order.mat[,2] <- ti

  if (np.order>=3)

    np.order.mat[,3] <- 0.5*(3*ti*ti-1)

  if (np.order>=4)

    np.order.mat[,4] <- 0.5*(5*ti^3-3*ti)

  if (np.order>=5)
```

```
    np.order.mat[,5] <- 0.125*(35*ti^4-30*ti^2+3)
  if (np.order>=6)
    np.order.mat[,6] <- 0.125*(63*ti^5-70*ti^3+15*ti)
  if (np.order>=7)
    np.order.mat[,7] <- (1/16)*(231*ti^6-315*ti^4+105*ti^2-5)
  if (np.order>=8)
    np.order.mat[,8] <- (1/16)*(429*ti^7-693*ti^5+315*ti^3-35*ti)
  if (np.order>=9)
    np.order.mat[,9] <- (1/128)*(6435*ti^8-12012*ti^6+6930*ti^4-1260*ti^2+35)
  if (np.order>=10)
    np.order.mat[,10] <- (1/128)*(12155*ti^9-25740*ti^7+18018*ti^5-4620*ti^3+315*ti)
  if (np.order>=11)
    np.order.mat[,11] <- (1/256)*(46189*ti^10-109395*ti^8+90090*ti^6-30030*ti^4+3465*ti^2-63)
  return(np.order.mat)
}


L <- data.frame(t = seq(0, 1, by = 0.01), Legendre(seq(0, 1, by = 0.01), 10))
L_melt <- melt(L, id = "t")


ggplot(L_melt, aes(x = t, y = value, group = variable, color = brewer.pal(10, "Set3")[variable]
  geom_line() + theme(legend.position="none")
```

The definition of the Legendre polynomials are the solutions for n = 0, 1, 2, … (with the normalization Pn(1) = 1) form a polynomial sequence of orthogonal polynomials called the Legendre

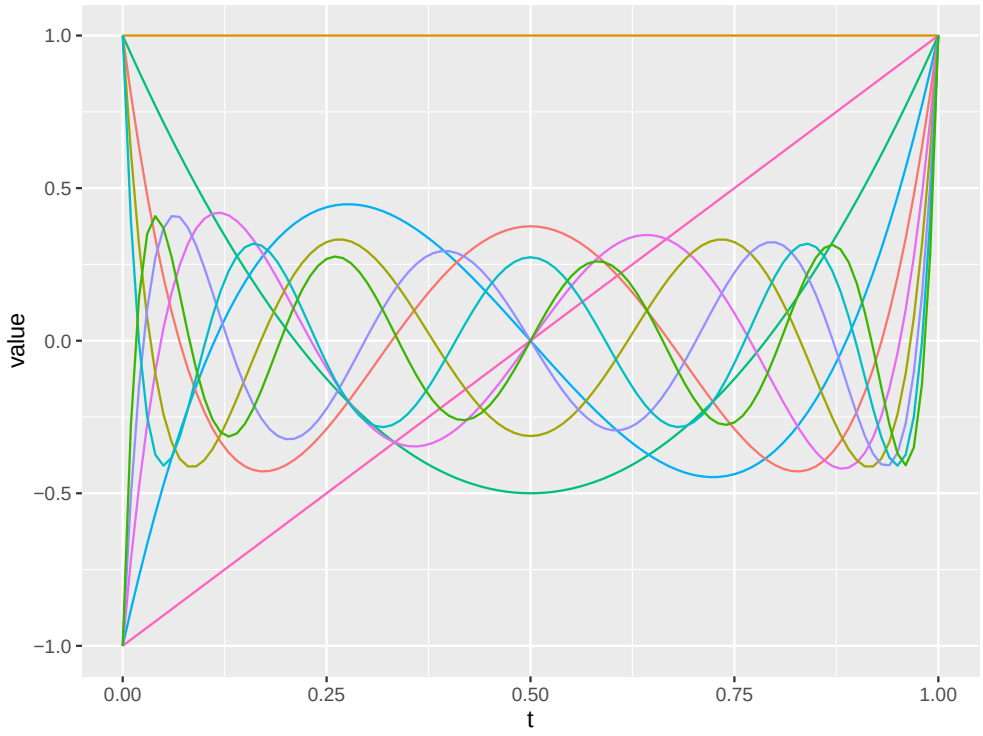

Figure 4.1: First 10 Legendre Polynomials

polynomials. Each Legendre polynomial Pn(x) is an nth-degree polynomial. It may be expressed using Rodrigues' formula:

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}[(x^2-1)^n] \tag{4.3}$$

An important property of the Legendre polynomials is that they are orthogonal with respect to the L2-norm on the interval −1 x 1:

$$\int_1^{-1} P_m(x)P_n(x)dx = \frac{2}{2+1}\delta_{mn} \tag{4.4}$$

δ_{mn} denotes the Kronecker delta equalt to 1 if m = n and 0 otherwise. These polynomials can be

generated by using the following recursively. Each Legendre polynomial would be the next order n in the expression below.

$$\begin{aligned} P_n(x) &= \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}^2(x-1)^{n-k}(x+1)^k \\ &= \sum_{k=0}^{n}\binom{n}{k}\binom{-n-1}{k}\left(\frac{1-x}{2}\right)^k \\ &= 2^{-n}\sum_{k=0}^{n}x^k\binom{n}{k}\binom{\frac{n+k+1}{2}}{k} \end{aligned} \tag{4.5}$$

With nature of the Legendre orthogonal polynomials, it was advantages for both dimension reduction and also handling unevenly spaced, missing or non-uniform time measurements from different subjects in the dataset. By seeing which polynomial curve fits the given phenotype, it removes some of the challenges when fitting the model. Different orders of the polynomial are tried throughout the procedure to allow for flexibility in the fitting the genetic variation from the mean curve for each of the genotypes or epistasis between genotypes considered in the model.

4.2.3 Model

The layout of the underlying model is first fit to an asymptotic growth model described by a logistic curve of the form,

$$\mu(t) = a/(1 + b * exp(-r * t))$$

It is biologically meaningful to implement a growth equation, like a logistic curve, to describe growth

trajectory West et al. (2001). Here the population is described by a mean growth curve by this growth equation where a, b and r are growth parameters each provide a biological interpretation, with a being the asymptotic growth, b being the initial amount of growth and r being the relative growth rate. Timevarying additive and dominant effects of significant SNPs are modeled by the Legendre orthogonal polynomial used in quantitative genetic studies, mentioned above. (Jiang et al. (2015), Olori et al. (1999), Li and Wu (2010)). This representation can be expressed as

$$\alpha_j(t) = (L_0(t), L_1(t), ..., L_s(t)) * (u_{j0}, u_{j1}, ..., u_{js})^T$$

$$\beta_j(t) = (L_0(t), L_1(t), ..., L_{s'}(t)) * (v_{j0}, v_{j1}, ..., v_{js'})^T$$

where $L_0(t), L_1(t), ..., L_s(t)$ and $L_0(t), L_1(t), ..., L_{s'}(t)$ are the LOP of orders s and s', respectively; and $u_{j0}, u_{j1}, ..., u_{js}$ and $v_{j0}, v_{j1}, ..., v_{js'}$ are the vectors of time-invariant additive and dominant effects, respectively. Orders s and s', selected from information criteria, for the purposes of this procedure the Bayesian information criterion (BIC_2), orignally developed by Chen and Chen (2008) was implemented. A nice feature that comes from modeling the fit in this manner is that the dimension of response phenotypic data is reduced through LOP modeling. (Li and Wu (2010), Jiang et al. (2015), Li and Wu (2010), Ahn et al. (2010), Das et al. (2011)). Writing the model out more explicitly we would have,

$$
\begin{aligned}
y(t) = \mu(t) + \sum_{j=1}^{J} \alpha_j(t)\xi_j + \sum_{k=1}^{K} \beta_k(t)\zeta_k & \\
& + \sum_{I_1<I_2=1}^{I} \gamma_I^{aa}(t)\xi_{I_1}\xi_{I_2} \\
& + \sum_{I_1<I_2=1}^{I} \gamma_I^{ad}(t)\xi_{I_1}\zeta_{I_2} \\
& + \sum_{I_1<I_2=1}^{I} \gamma_I^{da}(t)\zeta_{I_1}\xi_{I_2} \\
& + \sum_{I_1<I_2=1}^{I} \gamma_I^{dd}(t)\zeta_{I_1}\zeta_{I_2} \\
& + \epsilon(t)
\end{aligned}
\tag{4.6}
$$

4.2.4 Applying the iForm Procedure

An outline of the selection procedure used following the model described is as follows. At first the mean growth curve is estimated for the presented data following the logist growth curve. This could be adjusted depending on the funcitonal process the researcher is studying. Once the mean curve is fit the selection procedure is intialized in a similar fashion as mentioned in previous chapters @ref{highdeqtl}. Both the solution set and the model set are assigned to the empty set, $S_0 = \emptyset$ and $M_0 = \emptyset$. The candidate set starts off containing all main effects for the additive and dominant effects of each SNP. The selection procedure then begins and each SNP is assessed and the best fitting candidate is then placed in the selection set. While assessing each candidate SNP, an additional search is performed for the best fitting polynomial fit up to a pre-specified order that is determined at the begining of the procedure. The orthogonal polynomials are used to assist in fitting the genetic effects for each marker or epistatic interaction between the markers. This would allow the genetic effect some flexibility over time and give a more representative fit. This could

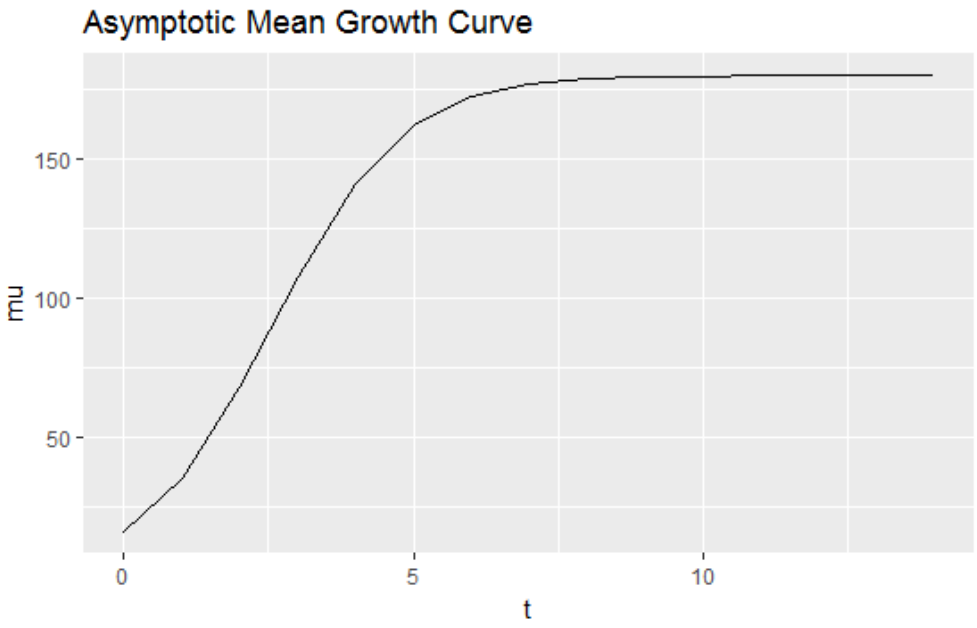

Figure 4.2: Example Growth Curve

also be used with other functional models or other types of non-linear functions that characterizes the biological systems being evaluated. We are treating the polynomials as a basis function for the regression problem and therefore the Residual Sum of Squares is calculated similar to multiple regression but replacing the design matrix with the necessary basis functions. This continues until a designated stopping value is reached. The BIC_2 is then used to find the optimal fit given the selection procedure performed. The following graphics show how the process works at each step

4.3 Application

4.3.1 Simulation Studies

As statistical issues become more complex they are going to be more analytically intractable and computational methods will need to close that gap to show the effectiveness of new models and

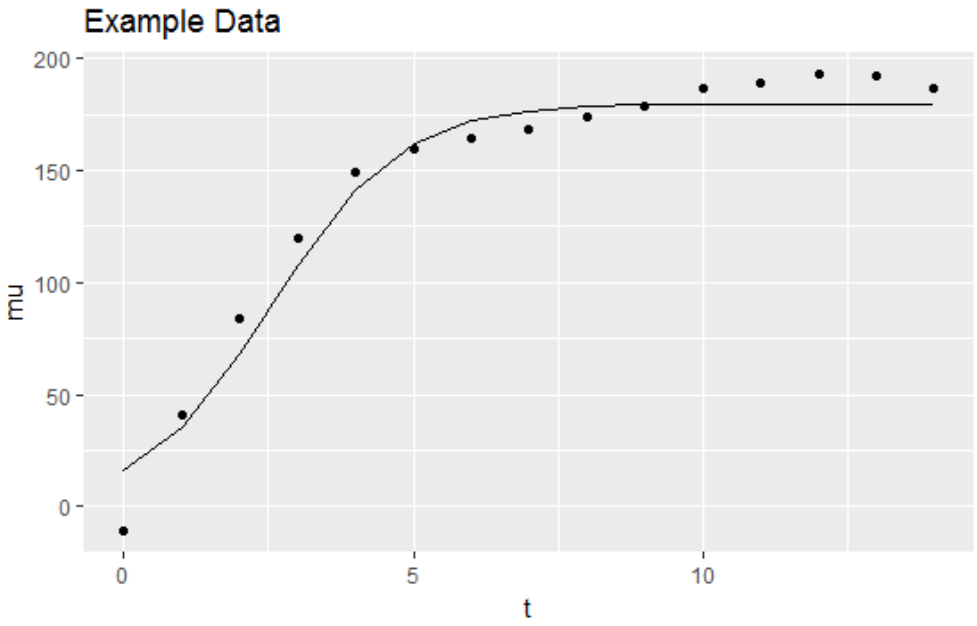

Figure 4.3: Example Data with Growth Cruve

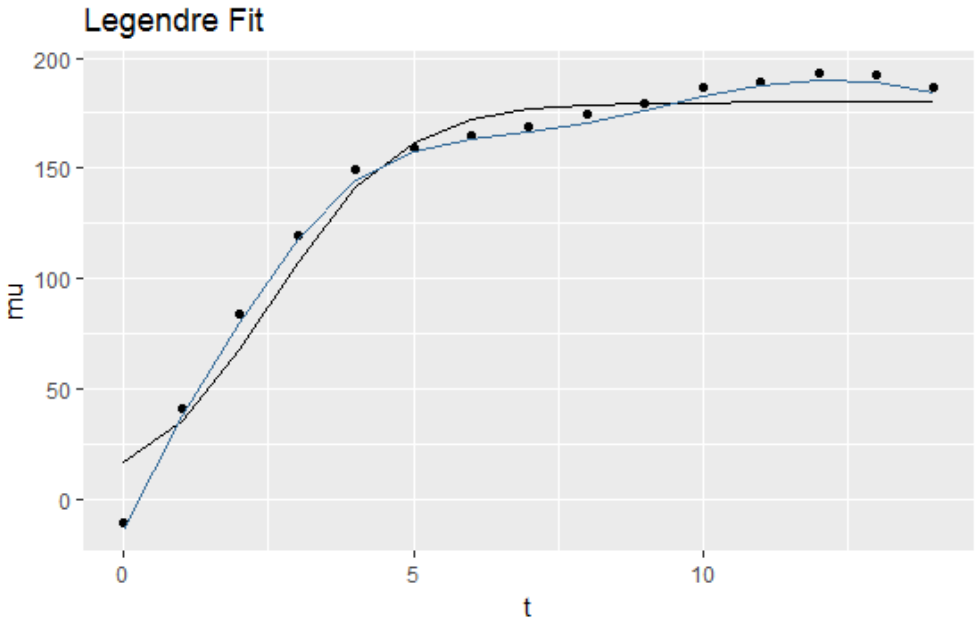

Figure 4.4: Additional Legendre Fit to Data

procedures.Simulation studies were performed to ascertain the validity of the model. Rates at which correct markers/epistasis were selected and overall model true model size was assessed. Data was original generated from a mean curve following the growth equation described above. It was then sampled from a multivariate normal distribution with the mean vector following the generated and with correlated errors over time. Significant effects were also included in the model to simulate different marker levels. These effects could be main effects of SNPs, or epistatic effects of interaction between SNPs. There were a total of 4 main effects and three interaction effects simulated. This simulation was replicated 100 times and then the selection procedure was ran. It performed well with 93% of the replicates selecting all main and interaction effects and the other 7% of the models select a majority of the effects. The model size simulated was of size 7. The average model size of the replicates fit was around 13.5. This would indicate some slight over-fitting but all important effects were included in the model. Some considerations for this need to be looked into further for future work.

4.3.2 Worked Example

As a comparison to what was done before in 3 the same mei tree dataset used in this chapter was studyied again. This serves as a comparison to previous performance and also to see if any new discoveries can be made by incorporating the time component and fitting the growth parameters simulatneously throughout the procedure. The previous model only fit one parameter of the growth equation as a time and assessed the genetic markers that had a significant impact of this parameter. The parameter focused on was the rate parameter,r for the shoot height of the progeny. The initial results running the analysis with a single predictor and using the selection procedure are,

As a comparison here are the results for simulatneaously fitting the growth curve and allowing

for a more flexible genetic effect to be fit to the data. As you can see there are overlapping markers identified in the models. This shows the robustness of the new selection technique to be consistent with previous models. You can also see that the fit of the overall model has also increased. By including all effects at once, you gain more statistical power and it boost the adjusted R square value from 0.71 to above 0.9. This boost in model performance could be partially due to some over-fitting like we observed in the simulation studies and therefore a very strict bonferroni correction was implmented to assess whether individual markers were truly significant. Even with a strict cut-off we stil observed 3 epistatic predictors to be highly significant. This shows the importance of inlcuding such terms while performing such a GWAS. The other area to note is the highly significant intercept term, which in our case is the result of the growth curve fit before implementing the selection procedure. This inidicates also the importance of including biologically relevant information in the model to help better understand the genetic architecture being studied of the phenotype.

4.4 Discussion

The new model proposed has some very nice features and seems to perform well in the given application. Considering the complexity of working high dimensional data coupled with including epsitatic effects as well as a functional component to the phenotype, it performs relatively efficiently. Using OLS type calculations aided in this efficiecy. This was enabled by the framing of the regression problem as a linear combination of basis functions. The agreeable properties of orthogonal polynomials helped ensure model assumptions are being met as closely as possible in order to use the calcuations. Also by including the biologically meaningful logistic growth equations it helps fit a baseline fit to the data and would better allow for individual genetic effects to be found throughout

the selection. These would of course have to be lab verified in order to assess the true biological mechanisms at play for the genetic control of the phenotype.

Some possible areas of concern for fitting the genetic effects to the polynomials could be that the effect does not conform to the polynomial curve under consideration. This can be alleviated by using other types of basis functions such as splines, but this would also cost a huge amount in computational efficiency with such a high dimensional data set. You need to be as efficient as possible when implementing a GWAS style study with the inclusion of possible epistasis. False positives could be of concern as well throughout the selection. The flexibility of the model and the wide range of polynomial fits that are considered could result in artifacts in the data to be picked up on. This is partially alleviated by using improved selection criteria like the BIC_2 and also having stricter than conventional cut-off values for significance testing. However these would still need to be assessed further through computational techniques such as cross-validation and bootstrapping. The selection procedure is effective as a screening tool for exploratory data analysis and hypothesis generation. Lab verification would be another area that would help validate the findings even futher. The comparison to previously run models is promising step in assesing the validity of the model at hand.

Chapter 5

Conclusions

5.1 Summary

In my disseration I have applied, adapted and extended the forward selection procedure under the marginality assumption first proposed by Hao and Zhang (2014) that is able to reduce search space substantially by making some reasonable assumptions about the data. These assumptions can be relaxed a little to broaden the scope of the space. The procedure ahs been applied in both simulated and real datasets in all areas of study. There are many advantages to using such a procedure with some being it is a computationally efficient way for high dimensional data situations that arise from high throughput data, especially when considering epistasis between gene markers or other type of interaction effects in the model. Guided search space based off of commonly used principles in model selection that are relevant in real world situations. It is able to perform a GWAS with included epistatic interactions a fairly quick manner. Able to relax assumptions to widen the search space. This may decrease the speed of the algorithm but makes the search more comprehensive.

Also it is able to screen for complex scenarios that would be hard to find in a lab setting otherwise. Having said that with the flexiblity of the initial model, false posiive rates could be a concern and need to be taken into consideration. Adjustement to the significance level and the model selection criteria have been made to account for this. This would still need to be followed up on independtly by checking with other datasets and/or lab verification.

Functional componenets were also considered to such a selection procedure. This also increases the computational complexity of the model and therefore decreases some efficeiency gains previously seen. Use biologically relevant information to guide the fitting of the model. Initially considers well established growth equations as a baseline for the underlying structure. It then performs a GWAS level analysis with considering epstatic interactions throughout the process.

5.2 Discussion

The selection procedures proposed attempt to address problems that are complex and have many moving parts to the procedure. Using a flexible model is very helpful to fulfill such requirements but this could have it be prone to over fitting at times if not well controlled. Need to heavily consider this as part of interpretating the model. Using stricter selection criteria and corrections to multuple testing for significance is extremely important. The main purpose is to use a screening to guide future research, use for exploratory data analysis and hypothesis genration.

With the complexity and expense that comes with genetic mapping, especially with a functional traits that needs repeatedly measured over time. Correlation structures are inherent in this data that need be addressed and usually done so at a heavy computational cost. Framing the problem as a regression problem and beingn able to use OLS calculations help to reduce the cost. Verification

steps are important steps especially with something as intricate as epistatic effects between gene markers.

Further investigations are needed to confirm or modify our findings by QTL mapping in natural populations.

5.3 Future Steps

More extensive explaination of future aims will be necessary for the final write up.

5.3.1 Aim 1

Incorporating other mean curves for the intercept term could help extend and relate to other areas of biology that follow a functional trait. There are also other types of orthogonal polynomials that could be explored as well. Using others polynomials would allow for other fits to the data that may be more applicable in other scenarios. Also using other basis functions in general could open up opportunties for other areas of application.

5.4 Aim 2

Other interesting areas would be to consider different levels of interactions with other omics data. Gene-gene interactions considered are only a portion of the picture and this type of modeling could also handle more levels of interactions that occur in a biological system. One area that I am particularly interested in applying would be in methylation studies of the gene expression.

This level of interaction is very important and a selection procedure like the one proposed could help screen and generate of possible effects and hypothesis to continue to look into. It would also apply in gene-environment interactions. This does not have to be restricted to just within the biological system under study. Environmental factors are important areas that could vastly impact the development of a phenotype.

5.4.1 Aim 3

Statistical areas that could as be considered to extend the model would be to include multivariate reponses to the system. For example having gene and protein expression being a bivariate reponse and to see how genetic markers and epistasis between the markers would better predict this by taking the correlation between those two response variables into account. Other statistical considerations would be to extend selection criteria to help further reduce the possibility of false positves being selected given a growing dataset like the one that occurs while dynamically including interaction effects throughout the selection. Making the model even more computational efficient would always be benefical as well. The faster a model can accurately run the more likely a researcher will use it. It will also help process all the high throughput data that is being constantly developed.

5.4.2 Closing Remarks

Continuing on, my aim would be to work with datasets of this scale and incorporate the types of statistical methods mentioned and machine learning techniques to aid in analysis. The results could help gain larger insights into the genomic/epigenetic architecture of biological systems. On top of the importance of a functional component to the phenotype, considering other types of multivariate

responses would be interesting to study in context of such a system. Integrating different level of omics data and the challenges that arise with such complicated and large datasets has interested me throughout my PhD work. Translating such a complex system into usable information that can be shared in order to prevent and fight disease would be ideal research for me. This type of research would need both methodological development as well as application of existing statistical and machine/deep learning techniques to handle the magnitude of the problem.

Chapter 6

Appendix

This needs more work. Only includes first part of the appendix pdf

Three-way Interactions

- Computational Complexity and Practical Issues
  - Page 10 in iFor paper
  - FS has a cost of $O(nm)$ for each step
    - iForm two-way has at most $p + \frac{n(n+1)}{2}$ or $m \leq p + n(n+1)$ holds for any step
    - Overall Complexity $nO\left(n\left(p + n(n+1)\right)\right) = O(n^2 p + n^4)$
    - D controls length of procedure, they tried $\frac{n}{4}, \frac{n}{3}, \frac{n}{2}$
- Theoretical Results
- Long-term concern about two-stage models because of theoretical validity
  - Hao and Zhang proved that two-stage model captures main effects in ultra-high dimensional situations
- Screening Consistency
  - Page 14

Naively, we can use any one-stage variable selection tool to fit (1.1) directly (as long as computation is feasible), ignoring the hierarchical structure. Though the model consistency or screening consistency result (Zhao & Yu, 2006; Wang, 2009; Fan & Lv, 2011) could be generalized to the context of interaction selection, the extension of earlier proofs is not straightforward due to heavy tails of interaction effects. Actually, all the existing proof technique would require some regularity conditions on the eigenvalues of $\Sigma_{(2)}$. Next, we establish the screening consistency of FS2 under conditions that are related only to $\Sigma_{(1)}$.

Bibliography

Ackermann, M., Sikora-Wohlfeld, W., and Beyer, A. (2013). Impact of natural genetic variation on gene expression dynamics. *PLoS genetics*, 9(6):e1003514.

Ahn, K., Luo, J., Berg, A., Keefe, D., and Wu, R. (2010). Functional mapping of drug response with pharmacodynamic–pharmacokinetic principles. *Trends in pharmacological sciences*, 31(7):306–311.

Beerenwinkel, N., Pachter, L., Sturmfels, B., Elena, S. F., and Lenski, R. E. (2007). Analysis of epistatic interactions and fitness landscapes using a new geometric approach. *BMC evolutionary biology*, 7(1):60.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111.

Brown, T. A. (2006). *Genomes*. Garland science.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351.

Carlborg, Ö. and Haley, C. S. (2004). Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, pages 759–771.

Cheverud, J. M. and Routman, E. J. (1995). Epistasis and its contribution to genetic variance components. *Genetics*, 139(3):1455–1461.

Chun, H. and Keleş, S. (2009). Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*, 182(1):79–90.

Collard, B., Jahufer, M., Brouwer, J., and Pang, E. (2005). An introduction to markers, quantitative trait loci (qtl) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica*, 142(1-2):169–196.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184–194.

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404.

Das, K., Li, J., Wang, Z., Tong, C., Fu, G., Li, Y., Xu, M., Ahn, K., Mauger, D., Li, R., et al. (2011). A dynamic model for genome-wide association studies. *Human genetics*, 129(6):629–639.

Dong, Z., Wang, J., and Wang, Z. (2015). Accurate estimation of quantitative trait locus effects with epistatic by improved variational linear regression. *arXiv preprint arXiv:1503.05628*.

Dowell, R. D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D. A., Rolfe, P. A., Heisler, L. E., Chin, B., et al. (2010). Genotype to phenotype: a complex problem. *Science*, 328(5977):469–469.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Hel-

gason, A., Walters, G. B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428.

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., et al. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, 343(6175):1246949.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eqtl analysis in multiple tissues. *PLoS Genet*, 9(5):e1003486.

Gosik, K., Kong, L., Chinchilli, V. M., and Wu, R. (2016). iform/eqtl: an ultrahigh-dimensional platform for inferring the global genetic architecture of gene transcripts. *Briefings in bioinformatics*, page bbw014.

Hansen, T. F. and Wagner, G. P. (2001). Epistasis and the mutation load: a measurement-theoretical approach. *Genetics*, 158(1):477–485.

Hao, N. and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301.

He, Q., Berg, A., Li, Y., Vallejos, C. E., and Wu, R. (2010a). Mapping genes for plant structure, development and evolution: functional mapping meets ontology. *Trends in genetics*, 26(1):39–46.

He, X., Qian, W., Wang, Z., Li, Y., and Zhang, J. (2010b). Prevalent positive epistasis in escherichia coli and saccharomyces cerevisiae metabolic networks. *Nature genetics*, 42(3):272–276.

Hernandez, K. M. (2015). Understanding the genetic architecture of complex traits using the function-valued approach. *New Phytologist*, 208(1):1–3.

Hill, W. G. and Mackay, T. F. (2004). Ds falconer and introduction to quantitative genetics. *Genetics*, 167(4):1529–1536.

Hou, W., Garvan, C. W., Littell, R. C., Behnke, M., Eyler, F. D., and Wu, R. (2006). A framework to monitor environment-induced major genes for developmental trajectories: implication for a prenatal cocaine exposure study. *Statistics in medicine*, 25(23):4020–4035.

Imielinski, M. and Belta, C. (2008). Exploiting the pathway structure of metabolism to reveal high-order epistasis. *BMC systems biology*, 2(1):40.

Jiang, L., Liu, J., Zhu, X., Ye, M., Sun, L., Lacaze, X., and Wu, R. (2015). 2higwas: a unifying high-dimensional platform to infer the global genetic architecture of trait development. *Briefings in bioinformatics*, page bbv002.

Jinks, J. L. and Mather, K. (1982). *Biometrical Genetics; The Study of Continous Variation*. Chapman and Hall.

Kao, C.-H. and Zeng, Z.-B. (2002). Modeling epistasis of quantitative trait loci using cockerham's model. *Genetics*, 160(3):1243–1261.

Kempthorne, O. (1968). The correlation between relatives on the supposition of mendelian inheritance. *American journal of human genetics*, 20(4):402.

Kendziorski, C., Chen, M., Yuan, M., Lan, H., and Attie, A. (2006). Statistical methods for expression quantitative trait loci (eqtl) mapping. *Biometrics*, 62(1):19–27.

Kim, B.-R., McMurry, T., Zhao, W., Wu, R., and Berg, A. (2010). Wavelet-based functional

clustering for patterns of high-dimensional dynamic gene expression. *Journal of Computational Biology*, 17(8):1067–1080.

Kim, Y., Xia, K., Tao, R., Giusti-Rodriguez, P., Vladimirov, V., Van Den Oord, E., and Sullivan, P. (2014). A meta-analysis of gene expression quantitative trait loci in brain. *Translational psychiatry*, 4(10):e459.

Koopmann, T. T., Adriaens, M. E., Moerland, P. D., Marsman, R. F., Westerveld, M. L., Lal, S., Zhang, T., Simmons, C. Q., Baczko, I., Dos Remedios, C., et al. (2014). Genome-wide identification of expression quantitative trait loci (eqtls) in human heart. *PLoS One*, 9(5):e97380.

Lee, M. N., Ye, C., Villani, A.-C., Raj, T., Li, W., Eisenhaure, T. M., Imboywa, S. H., Chipendo, P. I., Ran, F. A., Slowikowski, K., et al. (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*, 343(6175):1246980.

Li, J., Zhong, W., Li, R., and Wu, R. (2014). A fast algorithm for detecting gene–gene interactions in genome-wide association studies. *The annals of applied statistics*, 8(4):2292.

Li, L., Kabesch, M., Bouzigon, E., Demenais, F., Farrall, M., Moffatt, M. F., Lin, X., and Liang, L. (2013). Using eqtl weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Frontiers in genetics*, 4:103.

Li, Y. and Wu, R. (2010). Functional mapping of growth and development. *Biological Reviews*, 85(2):207–216.

Li, Z. and Sillanpää, M. J. (2012). Estimation of quantitative trait locus effects with epistasis by variational bayes algorithms. *Genetics*, 190(1):231–249.

Li, Z. and Sillanpää, M. J. (2015). Dynamic quantitative trait locus analysis of plant phenomic data. *Trends in plant science*, 20(12):822–833.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654.

Ma, C.-X., Casella, G., and Wu, R. (2002a). Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics*, 161(4):1751–1762.

Ma, J. F., Shen, R., Zhao, Z., Wissuwa, M., Takeuchi, Y., Ebitani, T., and Yano, M. (2002b). Response of rice to al stress and identification of quantitative trait loci for al tolerance. *Plant and Cell Physiology*, 43(6):652–659.

Mackay, T. F. (2014). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, 15(1):22–33.

Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10(8):565–577.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

McMullen, M., Byrne, P., Snook, M., Wiseman, B., Lee, E., Widstrom, N., and Coe, E. (1998). Quantitative trait loci and metabolic pathways. *Proceedings of the National Academy of Sciences*, 95(5):1996–2000.

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human heredity*, 56(1-3):73–82.

Muraya, M. M., Chu, J., Zhao, Y., Junker, A., Klukas, C., Reif, J. C., and Altmann, T. (2017). Genetic variation of growth dynamics in maize (zea mays l.) revealed through automated non-invasive phenotyping. *The Plant Journal*.

Nelson, R. M., Pettersson, M. E., and Carlborg, Ö. (2013). A century after fisher: time for a new paradigm in quantitative genetics. *Trends in Genetics*, 29(12):669–676.

Nica, A. C. and Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Phil. Trans. R. Soc. B*, 368(1620):20120362.

Olori, V., Hill, W., McGuirk, B., and Brotherstone, S. (1999). Estimating variance components for test day milk records by restricted maximum likelihood with a random regression animal model. *Livestock Production Science*, 61(1):53–63.

Pang, X., Wang, Z., Yap, J. S., Wang, J., Zhu, J., Bo, W., Lv, Y., Xu, F., Zhou, T., Peng, S., et al. (2013). A statistical procedure to map high-order epistasis for complex traits. *Briefings in bioinformatics*, 14(3):302–314.

Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature genetics*, 42(7):570–575.

Pettersson, M., Besnier, F., Siegel, P. B., and Carlborg, Ö. (2011). Replication and explorations of high-order epistasis using a large advanced intercross line pedigree. *PLoS Genet*, 7(7):e1002180.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772.

Rockman, M. V., Skrovanek, S. S., and Kruglyak, L. (2010). Selection at linked sites shapes heritable phenotypic variation in c. elegans. *Science*, 330(6002):372–376.

Sadhu, M. J., Bloom, J. S., Day, L., and Kruglyak, L. (2016). Crispr-directed mitotic recombination enables genetic mapping without crosses. *Science*, 352(6289):1113–1116.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717.

Stich, B., Yu, J., Melchinger, A. E., Piepho, H.-P., Utz, H. F., Maurer, H. P., and Buckler, E. S. (2007). Power to detect higher-order epistatic interactions in a metabolic pathway using a new mapping strategy. *Genetics*, 176(1):563–570.

Sun, L., Wang, Y., Yan, X., Cheng, T., Ma, K., Yang, W., Pan, H., Zheng, C., Zhu, X., Wang, J., et al. (2014a). Genetic control of juvenile growth and botanical architecture in an ornamental woody plant, prunus mume sieb. et zucc. as revealed by a high-density linkage map. *BMC genetics*, 15(1):S1.

Sun, L. and Wu, R. (2015). Mapping complex traits as a dynamic system. *Physics of life reviews*, 13:155–185.

Sun, L., Yang, W., Zhang, Q., Cheng, T., Pan, H., Xu, Z., Zhang, J., and Chen, C. (2013). Genome-wide characterization and linkage mapping of simple sequence repeats in mei (prunus mume sieb. et zucc.). *PloS one*, 8(3):e59562.

Sun, L., Ye, M., Hao, H., Wang, N., Wang, Y., Cheng, T., Zhang, Q., and Wu, R. (2014b). A model framework for identifying genes that guide the evolution of heterochrony. *Molecular biology and evolution*, 31(8):2238–2247.

Sun, W. (2012). A statistical framework for eqtl mapping using rna-seq data. *Biometrics*, 68(1):1–11.

Taylor, M. B. and Ehrenreich, I. M. (2014). Genetic interactions involving five or more genes contribute to a complex trait in yeast. *PLoS Genet*, 10(5):e1004324.

Taylor, M. B. and Ehrenreich, I. M. (2015). Higher-order genetic interactions and their contribution to complex traits. *Trends in genetics*, 31(1):34–40.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tong, C., Wang, Z., Zhang, B., Shi, J., and Wu, R. (2011). 3funmap: full-sib family functional mapping of dynamic traits. *Bioinformatics*, 27(14):2006–2008.

van Eeuwijk, F. A., Bink, M. C., Chenu, K., and Chapman, S. C. (2010). Detection and use of qtl for complex traits in multiple environments. *Current opinion in plant biology*, 13(2):193–205.

Van Steen, K. (2011). Travelling the world of gene–gene interactions. *Briefings in bioinformatics*, page bbr012.

Viñuela, A., Snoek, L. B., Riksen, J. A., and Kammenga, J. E. (2010). Genome-wide gene expression regulation as a function of genotype and age in c. elegans. *Genome research*, 20(7):929–937.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524.

Wang, J., Joshi, T., Valliyodan, B., Shi, H., Liang, Y., Nguyen, H. T., Zhang, J., and Xu, D. (2015). A bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. *BMC genomics*, 16(1):1011.

Wang, Z., Liu, T., Lin, Z., Hegarty, J., Koltun, W. A., and Wu, R. (2010). A general model for multilocus epistatic interactions in case-control studies. *PLoS One*, 5(8):e11384.

Wei, W.-H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733.

Weinreich, D. M., Lan, Y., Wylie, C. S., and Heckendorn, R. B. (2013). Should evolutionary geneticists worry about higher-order epistasis? *Current opinion in genetics & development*, 23(6):700–707.

West, G. B., Brown, J. H., and Enquist, B. J. (2001). A general model for ontogenetic growth. *Nature*, 413(6856):628–631.

Whittaker, J., Thompson, R., and Visscher, P. (1996). On the mapping of qtl by regression of phenotype on marker-type. *Heredity*, 77(1):23–32.

Wu, R. (1996). Detecting epistatic genetic variance with a clonally replicated design: models for lowvs high-order nonallelic interaction. *Theoretical and Applied Genetics*, 93(1-2):102–109.

Wu, R. and Lin, M. (2006). Functional mapping—how to map and study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics*, 7(3):229–237.

Wu, R., Ma, C., and Casella, G. (2007). *Statistical genetics of quantitative traits: linkage, maps and QTL*. Springer Science & Business Media.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.3.8.

Xu, S. (2007). An empirical bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics*, 63(2):513–521.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhang, Y.-Y., Fischer, M., Colot, V., and Bossdorf, O. (2013). Epigenetic variation creates potential for evolution of plant phenotypic plasticity. *New Phytologist*, 197(1):314–322.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

High Definition eQTL Tables and Figures

Figure Legends

Figure 1 The distribution of eQTLs for each transcript abundance phenotype in *C. elegans*, located at the genomic positions of the transcripts. Those eQTLs on the diagonal are cis-eQTLs, whereas those off the diagonal are trans-eQTLs. Adapted from Rockman et al. (2010).

Figure 2 Circos plot illustrating the pattern of how a particular gene transcript is regulated by eQTLs on different chromosomes in *C. elegans*.

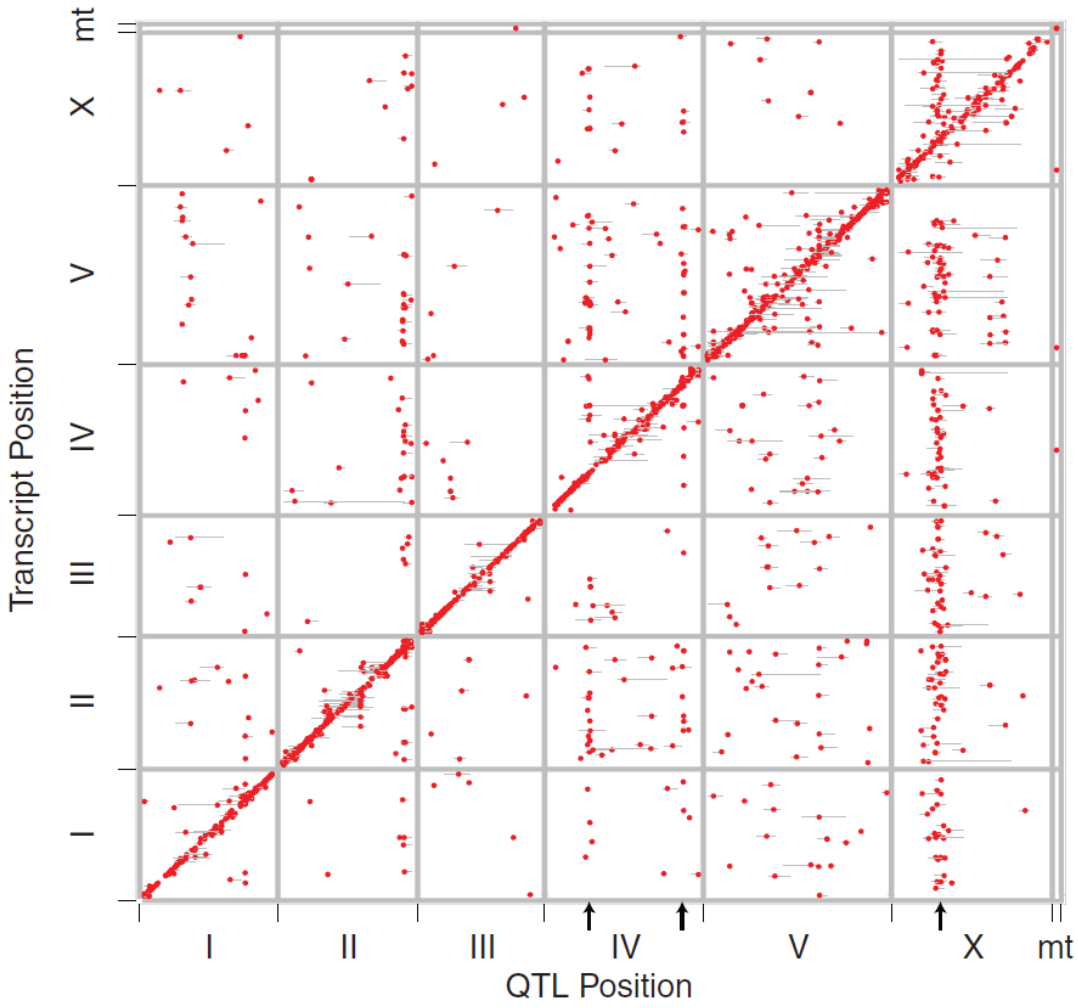
Transcript Position
mt
X
V
IV
III
II
I
I
II
III
IV
V
X
mt
QTL Position

Figure 1

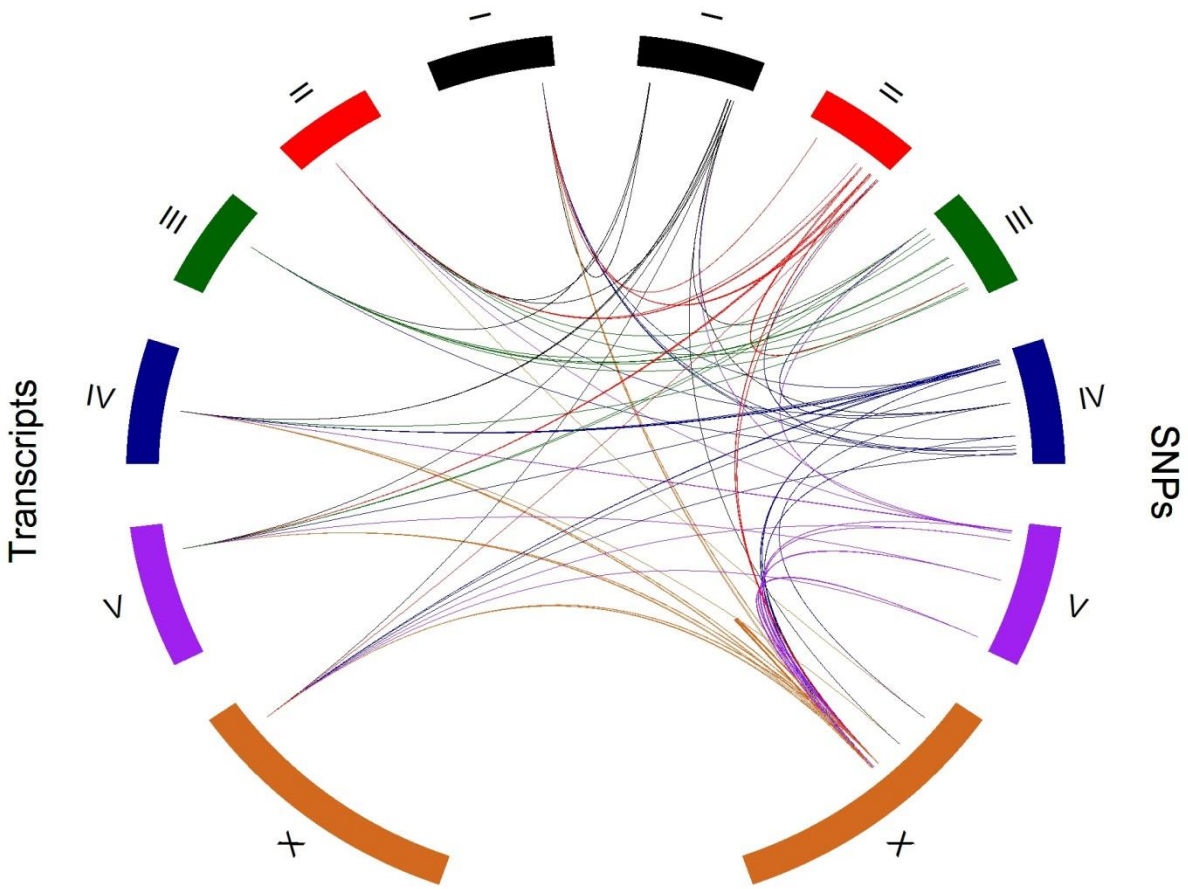

Figure 2

Table 1 Results of simulation 1 with $\sigma = 1$ for the random error with independent predictors.

$\boldsymbol{\sigma = 1}$	**Main Effects**				**Interactions Effects**				**Model**			
Method	**Cov**	**Cor0**	**Inc0**	**Exact**	**Cov**	**Cor0**	**Inc0**	**Exact**	**Size**	**MSE**	$\boldsymbol{Adj - R^2}$	**Time**
Single Marker	0	1	0.25	0	NA	NA	NA	NA	3	23.63	0.216	0.824
FS	0.85	0.953	0.0625	0.85	NA	NA	NA	NA	27	10.23	0.660	3.47
FS2	1	0.996	0	1	0.95	0.981	0	0	27	0.302	0.989	72.31
iFORM	0.9	0.999	0.05	0.9	0.9	1	0	0.9	7.55	2.93	0.894	4.08
Oracle	1	1	0	1	1	1	0	1	8	1.023	0.965	NA

Outcomes include the convergence Probability (Cov) $\sum_{m=1}^{M} I\left(\mathcal{T} \subset \hat{\mathcal{T}}\right)/M$, percentage of correct zeros identified (Cor0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\widehat{\beta_j} = 0, \beta_j = 0)/[M(p - p_0)]$, percentage of incorrect zeros identified (Inc0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\widehat{\beta_j} = 0, \beta_j \neq 0)/[M(p_0)]$, the exact selection probability (Exact) $\sum_{m=1}^{M} I\left(\mathcal{T} = \hat{\mathcal{T}}\right)/M$, average model size, Mean Square Error for the model (MSE), the adjusted R-square of the model, and the computational time in seconds.

Table 2 Results of simulation 2 with $\sigma = 2$ for the random error with independent predictors.

$\sigma = 2$	Main Effects				Interactions Effects				Model			
Method	**Cov**	**Cor0**	**Inc0**	**Exact**	**Cov**	**Cor0**	**Inc0**	**Exact**	**Size**	**MSE**	$\boldsymbol{Adj - R^2}$	**Time**
Single Marker	0.02	0.999	0.529	0.029	NA	NA	NA	NA	1.97	27.02	0.178	0.69
FS	0.8	0.953	0.05	0.8	NA	NA	NA	NA	27	11.49	0.651	3.22
FS2	1	0.996	0	1	0.98	0.98	0	0	27	1.17	0.964	68.2
iFORM	0.97	0.998	0.007	0.97	0.95	1	0	0.93	8.7	4.41	0.865	3.84
Oracle	1	1	0	1	1	1	0	1	8	3.92	0.880	NA

Outcomes include the convergence Probability (Cov) $\sum_{m=1}^{M} I(\mathcal{T} \subset \hat{\mathcal{T}})/M$, percentage of correct zeros identified (Cor0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\widehat{\beta_j} = 0, \beta_j = 0)/[M(p - p_0)]$, percentage of incorrect zeros identified (Inc0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\widehat{\beta_j} = 0, \beta_j \neq 0)/[M(p_0)]$, the exact selection probability (Exact) $\sum_{m=1}^{M} I(\mathcal{T} = \hat{\mathcal{T}})/M$, average model size, Mean Square Error for the model (MSE), the adjusted R-square of the model, and the computational time in seconds.

Table 3 Results of simulation 3 with $\sigma = 3$ for the random error with independent predictors.

$\sigma = 3$	Main Effects				Interactions Effects				Model			
Method	**Cov**	**Cor0**	**Inc0**	**Exact**	**Cov**	**Cor0**	**Inc0**	**Exact**	**Size**	**MSE**	$\boldsymbol{Adj - R^2}$	**Time**
Single Marker	0	0.999	0.612	0	NA	NA	NA	NA	1.65	33.90	0.138	0.69
FS	0.82	0.953	0.043	0.827	NA	NA	NA	NA	27	14.44	0.633	3.22
FS2	1	0.997	0	1	0.98	0.97	0	0	27	2.69	0.931	68.2
iFORM	0.89	0.995	0.060	0.896	0.96	1	0	0.94	7.93	11.10	0.713	3.83
Oracle	1	1	0	1	1	1	0	1	8	8.98	0.771	NA

Outcomes include the convergence Probability (Cov) $\sum_{m=1}^{M} I(\mathcal{T} \subset \hat{\mathcal{T}})/M$, percentage of correct zeros identified (Cor0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\widehat{\beta_j} = 0, \beta_j = 0)/[M(p - p_0)]$, percentage of incorrect zeros identified (Inc0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\widehat{\beta_j} = 0, \beta_j \neq 0)/[M(p_0)]$, the exact selection probability (Exact) $\sum_{m=1}^{M} I(\mathcal{T} = \hat{\mathcal{T}})/M$, average model size, Mean Square Error for the model (MSE), the adjusted R-square of the model, and the computational time in seconds.

Table 4 Estimated main and epistatic effects of eQTLs by iFORM on gene transcript A_12_P103290 on chromosome I , in a comparison with the result by traditional single marker analysis.

eQTL	iFORM				Single Marker Analysis		
	Effect	**SE**	***p*-value**	**Herit-ability(%)**	**Effect**	**SE**	***p*-value**
X1_2068168 (cis-eQTL)	-0.197	0.035	0.000	0.060	-0.210	0.074	0.005
X2_13516256	-0.069	0.039	0.080	0.007	-0.138	0.075	0.068
X2_13813025	-0.023	0.052	0.665	0.001	-0.095	0.074	0.201
X2_13694563	0.074	0.062	0.238	0.008	-0.089	0.075	0.242
X2_2482896	0.064	0.027	0.017	0.006	-0.057	0.076	0.453
X_15500580	0.073	0.064	0.253	0.008	0.010	0.076	0.899
X_14636404	-1.768	0.092	0.000	4.794	0.024	0.076	0.751
X4_16403215	0.028	0.040	0.489	0.001	0.087	0.075	0.252
X4_15568674	-1.972	0.134	0.000	5.964	0.094	0.075	0.213
X4_1873297	0.044	0.026	0.086	0.003	0.095	0.071	0.185
X4_15632637	1.960	0.143	0.000	5.892	0.104	0.075	0.169
X4_13532205	0.064	0.028	0.024	0.006	0.111	0.075	0.143
X_15820520	-0.014	0.055	0.796	0.000	0.146	0.075	0.054
X_14542103	1.786	0.087	0.000	4.892	0.162	0.075	0.031
X2_13516256.X4_15632637	-3.799	0.268	0.000	5.534	NA	NA	NA
X2_13516256.X4_15568674	3.753	0.276	0.000	5.401	NA	NA	NA
X_15820520.X_14636404	-3.771	0.172	0.000	5.453	NA	NA	NA
X_15820520.X_14542103	3.691	0.172	0.000	5.224	NA	NA	NA
X_14636404.X4_1873297	-3.534	0.163	0.000	4.789	NA	NA	NA
X_14636404.X4_13532205	-3.567	0.166	0.000	4.879	NA	NA	NA
X_14542103.X4_1873297	3.629	0.164	0.000	5.050	NA	NA	NA
X_14542103.X4_13532205	3.469	0.167	0.000	4.614	NA	NA	NA

Higher Order Epistasis Networks Tables and Figures

Table 1 Simulation results when the truth obeys strong heredity

Model	T1_tpr	T1_tnr	T1_fpr	T1_fnr	T2_tpr	T2_tnr	T2_fpr	T2_fnr	T3_tpr	T3_tnr	T3_fpr	T3_fnr	Train_MSE	Train_Rsq	Test_MSE	Test_Rsq	Run Time
forward_select	1.000	0.999	0.001	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.330	0.727	3.490	0.711	0.757
iform_order_2_weak	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.128	0.907	1.252	0.895	5.896
iform_order_2_strong	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.102	0.909	1.198	0.900	1.557
forward_select_2	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.086	0.910	1.198	0.900	25.481
forward_select_3	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.992	0.918	1.121	0.906	471.881
iform_order_3_weak	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.020	0.916	1.135	0.905	11.346
iform_order_3_strong	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.968	0.920	1.060	0.911	1.872
glinternet	1.000	0.559	0.441	0.000	1.000	0.982	0.018	0.000	0.000	1.000	0.000	1.000	1.246	0.898	1.446	0.880	208.167
hierNet	1.000	0.697	0.303	0.000	1.000	0.976	0.024	0.000	0.000	1.000	0.000	1.000	0.906	0.925	1.421	0.882	27.521
Oracle	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.953	0.921	1.050	0.912	NA

Table 2 Simulation results when the truth obeys weak heredity

Model	T1_tpr	T1_tnr	T1_fpr	T1_fnr	T2_tpr	T2_tnr	T2_fpr	T2_fnr	T3_tpr	T3_tnr	T3_fpr	T3_fnr	Train_MSE	Train_Rsq	Test_MSE	Test_Rsq	Run Time
forward_select	1.000	0.999	0.001	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.326	0.731	3.480	0.716	4.355
iform_order_2_weak	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.119	0.910	1.200	0.901	8.342
iform_order_2_strong	1.000	0.992	0.008	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.580	0.872	1.707	0.859	2.952
forward_select_2	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.083	0.912	1.167	0.904	38.872
forward_select_3	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.979	0.921	1.089	0.910	569.983
iform_order_3_weak	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.003	0.919	1.079	0.911	13.054
iform_order_3_strong	1.000	0.992	0.008	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.578	0.872	1.705	0.859	2.787
glinternet	1.000	0.469	0.531	0.000	1.000	0.980	0.020	0.000	0.000	1.000	0.000	1.000	0.906	0.927	1.425	0.883	29.975
hierNet	1.000	0.657	0.343	0.000	1.000	0.973	0.027	0.000	0.000	1.000	0.000	1.000	0.856	0.931	1.412	0.884	33.302
Oracle	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.940	0.924	1.034	0.915	NA

Table 3 Simulation results when the truth is anti-heredity

Model	T1_tpr	T1_tnr	T1_fpr	T1_fnr	T2_tpr	T2_tnr	T2_fpr	T2_fnr	T3_tpr	T3_tnr	T3_fpr	T3_fnr	Train_MSE	Train_Rsq	Test_MSE	Test_Rsq	Run Time
forward_select	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.284	0.729	3.510	0.714	1.005
iform_order_2_weak	1.000	0.996	0.004	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.140	0.741	3.435	0.719	7.866
iform_order_2_strong	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.284	0.729	3.510	0.714	2.386
forward_select_2	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.081	0.911	1.171	0.904	29.095
forward_select_3	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.989	0.918	1.095	0.910	548.617
iform_order_3_weak	1.000	0.997	0.003	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.155	0.739	3.448	0.719	13.216
iform_order_3_strong	1.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.284	0.729	3.510	0.714	2.703
glinternet	1.000	0.290	0.710	0.000	1.000	0.971	0.029	0.000	0.000	1.000	0.000	1.000	0.844	0.931	1.578	0.871	26.564
hierNet	1.000	0.142	0.858	0.000	1.000	0.915	0.085	0.000	0.000	1.000	0.000	1.000	0.307	0.975	2.216	0.819	3.417
Oracle	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.952	0.921	1.031	0.915	NA

Table 4 Simulation results when the truth is constructed of pure interactions

Model	T1_tpr	T1_tnr	T1_fpr	T1_fnr	T2_tpr	T2_tnr	T2_fpr	T2_fnr	T3_tpr	T3_tnr	T3_fpr	T3_fnr	Train_MSE	Train_Rsq	Test_MSE	Test_Rsq	Run Time
forward_select	NaN	0.980	0.020	NaN	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.316	0.025	3.445	-0.039	1.177
iform_order_2_weak	NaN	0.972	0.028	NaN	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.007	0.115	3.181	0.040	5.840
iform_order_2_strong	NaN	0.979	0.021	NaN	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.294	0.031	3.429	-0.034	2.081
forward_select_2	NaN	1.000	0.000	NaN	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.117	0.669	1.170	0.644	26.396
forward_select_3	NaN	1.000	0.000	NaN	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.005	0.703	1.081	0.671	530.362
iform_order_3_weak	NaN	0.975	0.025	NaN	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.043	0.106	3.209	0.032	9.461
iform_order_3_strong	NaN	0.979	0.021	NaN	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	3.294	0.031	3.429	-0.034	2.265
glinternet	NaN	0.429	0.571	NaN	1.000	0.983	0.017	0.000	0.000	1.000	0.000	1.000	1.002	0.699	1.445	0.561	145.078
hierNet	NaN	0.147	0.853	NaN	1.000	0.955	0.045	0.000	0.000	1.000	0.000	1.000	0.672	0.802	1.758	0.467	4.491
Oracle	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.968	0.713	1.022	0.689	NA

Table 5 The detection of epistasis for the relative growth rate (*r*) of shoot length in the full-sib family of mei tree by a low-order epistatic model

Coefficient	Estimate	SE	T-value	P-value
(Intercept)	0.18285	0.07613	2.402	0.0174 *
AATTC_nn_np_2517_a	0.40013	0.06509	6.147	5.13e-09 ***
AATTC_nn_np_2815_a	0.15792	0.06837	2.310	0.0221 *
CATG_nn_np_3479_a	0.23433	0.05285	4.434	1.63e-05 ***
CATG_nn_np_1284_a	0.22200	0.05313	4.179	4.61e-05 ***
AATTC_nn_np_2815_a×AATTC_lm_ll_3034_a	0.45783	0.09244	4.953	1.71e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3504 on 176 degrees of freedom
Multiple R-squared: 0.3428, Adjusted R-squared: 0.3241
F-statistic: 18.36 on 5 and 176 DF, p-value: 1.189e-14

Table The detection of epistasis for the relative growth rate (*r*) of shoot length in the full-sib family of mei tree by a high-order epistatic model

Coefficient	Estimate	SE	T-value	P-value
(Intercept)	0.16859	0.05801	2.906	0.00415 **
AATTC_nn_np_2517_a	0.27773	0.04396	6.318	2.27e-09 ***
AATTC_nn_np_2815_a	0.26382	0.05295	4.983	1.54e-06 ***
CATG_nn_np_3479_a	0.20767	0.03467	5.990	1.23e-08 ***
CATG_nn_np_1284_a	0.04522	0.04265	1.060	0.29055
AATTC_nn_np_2815_a×AATTC_lm_ll_3034_a	1.82572	0.17925	10.185	< 2e-16 ***
AATTC_nn_np_2815_a×AATTC_hk_hk_278_a	0.25935	0.03888	6.671	3.48e-10 ***
CATG_lm_ll_3153_a	0.14877	0.03491	4.262	3.36e-05 ***
CATG_nn_np_1284_a×AATTC_nn_np_554_a	0.22994	0.05104	4.505	1.23e-05 ***
AATTC_nn_np_2815_a.AATTC_lm_ll_3034_a×AATTC_nn_np_1615_a	-1.51714	0.19060	-7.960	2.39e-13 ***
AATTC_nn_np_2815_a×AATTC_nn_np_929_a	-0.30805	0.05477	-5.624	7.57e-08 ***
AATTC_hk_hk_479_d	0.16044	0.03443	4.660	6.37e-06 ***
AATTC_nn_np_2517_a×CATG_hk_hk_648_a	0.14537	0.02840	5.118	8.33e-07 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2268 on 169 degrees of freedom
Multiple R-squared: 0.7356, Adjusted R-squared: 0.7168
F-statistic: 39.19 on 12 and 169 DF, p-value: < 2.2e-16

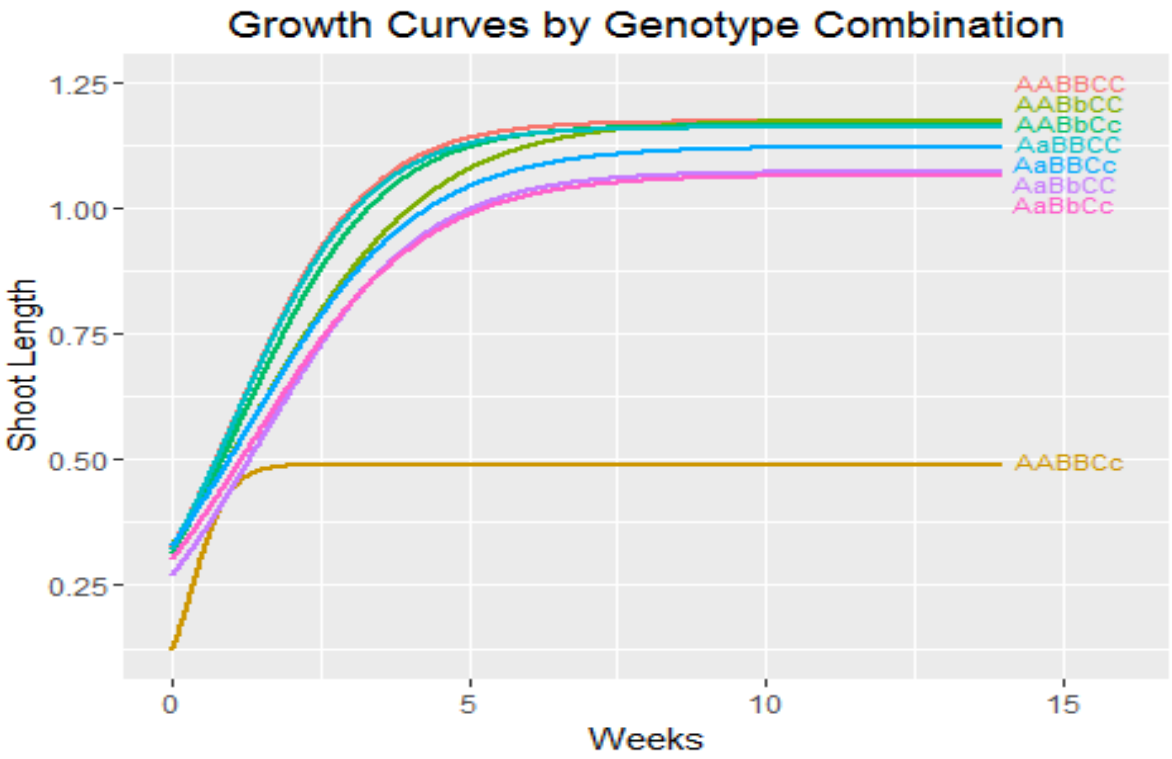

Figure 1 Growth curves of shoot length in mei drawn from estimated growth parameters at three loci of significant high-order epistasis

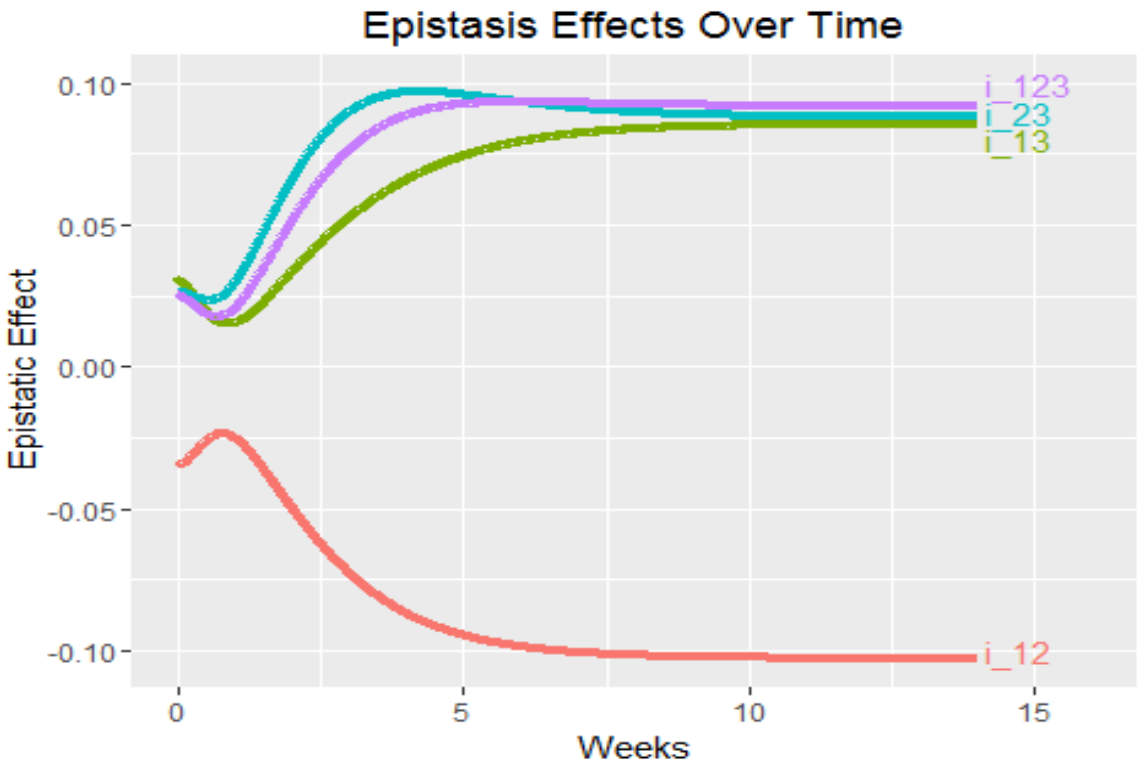

Figure 2 Curves of epistatic effects on shoot length growth in mei at three significant loci

Three-way Interactions

- Computational Complexity and Practical Issues
    - Page 10 in iFor paper
    - FS has a cost of $O(nm)$ for each step
        - iForm two-way has at most $p + \frac{n(n+1)}{2}$ or $m \leq p + n(n+1)$ holds for any step
        - Overall Complexity $nO\left(n\left(p + n(n+1)\right)\right) = O(n^2p + n^4)$
        - D controls length of procedure, they tried $\frac{n}{4}, \frac{n}{3}, \frac{n}{2}$
- Theoretical Results
- Long-term concern about two-stage models because of theoretical validity
    - Hao and Zhang proved that two-stage model captures main effects in ultra-high dimensional situations
- Screening Consistency
    - Page 14

Naively, we can use any one-stage variable selection tool to fit (1.1) directly (as long as computation is feasible), ignoring the hierarchical structure. Though the model consistency or screening consistency result (Zhao & Yu, 2006; Wang, 2009; Fan & Lv, 2011) could be generalized to the context of interaction selection, the extension of earlier proofs is not straightforward due to heavy tails of interaction effects. Actually, all the existing proof technique would require some regularity conditions on the eigenvalues of $\Sigma_{(2)}$. Next, we establish the screening consistency of FS2 under conditions that are related only to $\Sigma_{(1)}$.

- C1
    - $X_{i1}, \dots, X_{ip}$ are jointly and marginally standard normal

In this section we work on the total covariance matrix Σ and show it is determined by the covariance of the matrix $\Sigma^{(1)}$ of main effects under the Gaussian assumption

For X_j the main effects and $Z_{kl} = X_k X_l - E(X_k X_l)$ for $(k,l) \in \mathcal{P}_2$ the interactions and $W_{rst} = X_r X_s X_t - E(X_r X_s X_t)$ for $(r,s,t) \in \mathcal{P}_3$ for order 3 effects

Lemma 1
Under the normality condition (C1), for $\forall\, j,k,l,r,s,t$

1. $cov(X_j, Z_{kl}) = 0$
2. $cov(X_j, W_{rst}) = 0$
3. $cov(Z_{kl}, W_{rst}) = 0$

$$\Sigma = \begin{pmatrix} \Sigma^{(1)} & 0 & 0 \\ 0 & \Sigma^{(2)} & 0 \\ 0 & 0 & \Sigma^{(3)} \end{pmatrix}$$

Proof:

1. $cov(X_j, Z_{kl}) = cov(X_j,\ X_k X_l) = E(X_j X_k X_l) - E(X_j)\, E(X_k X_l) = 0$
2. $cov(X_j, W_{rst}) = cov(X_j,\ X_r X_s X_t) = E(X_j X_r X_s X_t) - E(X_j)\, E(X_r X_s X_t) = 0$
3. $cov(Z_{kl}, W_{rst}) = cov(X_k X_l,\ X_r X_s X_t) = E(X_k X_l X_r X_s X_t) - E(X_k X_l) E(X_r X_s X_t) = 0$

This holds if the joint density of
$X_1, \dots, X_p$ *is symmetric with repsec to the origin point* 0

Lemma 2 (still need to extend to order-three interactions)

Generic Formula:

$$E\left\{\prod_{i=1}^{n} X_i^{a_i}\right\} = \sum_{I \in S_a} d_{a,I} \left(\prod_{i=1}^{n}\prod_{j=1}^{n} \varphi_{ij}^{I_{ij}}\right)\left(\prod_{j=1}^{n} \mu_j^{I_{aj}}\right)$$

$$d_{a,I} = \frac{\prod_{k=1}^{n} a_k!}{2^{M_I}\left(\prod_{i=1}^{n}\prod_{j=1}^{n} I_{ij}!\right)\left(\prod_{j=1}^{n} I_{aj}!\right)}$$

Under the normality condition (C1)

$cov(Z_{ij}, Z_{kl}) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$ Bar & Dittrich, 1971

$$\begin{aligned} E(X_iX_jX_kX_l) &= E(X_iX_j)E(X_kX_l) + E(X_iX_k)E(X_jX_l) + E(X_iX_l)E(X_jX_k) \\ &\quad - 2E(X_i)E(X_j)E(X_k)E(X_l) = 0 + E(X_iX_k)E(X_jX_l) + E(X_iX_l)E(X_jX_k) - 0 \\ &= E(X_iX_k)E(X_jX_l) + E(X_iX_l)E(X_jX_k) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} \end{aligned}$$

$$\begin{aligned} &E(X1X2X3X4X5X6) \\ &= \rho_{12}\rho_{34}\rho_{56} + \rho_{12}\rho_{35}\rho_{46} + \rho_{12}\rho_{36}\rho_{45} + \rho_{13}\rho_{24}\rho_{56} + \rho_{13}\rho_{25}\rho_{46} + \rho_{13}\rho_{26}\rho_{45} \\ &+ \rho_{14}\rho_{23}\rho_{56} + \rho_{14}\rho_{25}\rho_{36} + \rho_{14}\rho_{26}\rho_{35} + \rho_{15}\rho_{23}\rho_{46} + \rho_{15}\rho_{24}\rho_{36} + \rho_{15}\rho_{26}\rho_{34} \\ &+ \rho_{16}\rho_{23}\rho_{45} + \rho_{16}\rho_{24}\rho_{35} + \rho_{16}\rho_{25}\rho_{34} \\ &= 0 + 0 + 0 + 0 + \rho_{13}\rho_{25}\rho_{46} + \rho_{13}\rho_{26}\rho_{45} + 0 + \rho_{14}\rho_{25}\rho_{36} + \rho_{14}\rho_{26}\rho_{35} \\ &+ \rho_{15}\rho_{23}\rho_{46} + \rho_{15}\rho_{24}\rho_{36} + 0 + \rho_{16}\rho_{23}\rho_{45} + \rho_{16}\rho_{24}\rho_{35} + 0 \end{aligned}$$

(7.1)

Let $A = (A_{ij})$ be an N x N matrix. In linear algebra, a K x K submatrix is called a principal submatrix if it is of the form $A_I = \left(A_{l_i,l_j}\right)$ where $\mathcal{I}$ is an index set $\mathcal{I} = \{1 \le l_1 < \cdots < l_K \le N\}$. Here with slight abuse of this conception, we allow arbitrary order of the index set $\mathcal{I}$. For example, let $\mathcal{I} = \{2,1\}$ and $A_{\mathcal{I}} = \begin{pmatrix} A_{22} & A_{21} \\ A_{12} & A_{11} \end{pmatrix}$ is still called a principal submatrix in this paper.

Based on the formula (7.1), we can decompose $\Sigma^{(2)}$ to a sum $\Sigma_1^{(2)} + \Sigma_2^{(2)}$. In fact, we have

Lemma 3

Both $\Sigma_1^{(2)}$ and $\Sigma_2^{(2)}$ $are\ principal\ submatrices\ of$ $\Sigma^{(1)} \otimes \Sigma^{(1)}$

Proof: The Kronecker product (Laub, 2005) $\Sigma^{(1)} \otimes \Sigma^{(1)}$ is a $p^2 x\ p^2$ matrix whose rows and columns are both indexed by the set $\mathcal{P}_1 \times \mathcal{P}_1$. The entry corresponding to the index $(ij, kl) is\ \sigma_{ij}\sigma_{kl}$. By formula (7.1), both $\Sigma_1^{(2)}$ and $\Sigma_2^{(2)}$ $are\ \frac{p(p+1)}{2} \times \frac{p(p+1)}{2}$ $principal\ submatrices\ of$ $\Sigma^{(1)} \otimes \Sigma^{(1)}$

Lemma 4: *Under C1 and C2a we have*

$$2\tau_{min} < \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) < \tau_{max}/2 \qquad (7.2)$$

Proof: By Laub (2005) Theorem 13.12, the eigenvalues of $\Sigma^{(1)} \otimes \Sigma^{(1)}$ are $\lambda_i\lambda_j, 1 \leq i, j \leq p$, if the eigenvalues of $\Sigma^{(1)}$ are $\lambda_1, \ldots, \lambda_p$. Therefore under condition C2a we have

$$\tau_{min} < \lambda_{min}\left(\Sigma^{(1)} \otimes \Sigma^{(1)}\right) \leq \lambda_{max}\left(\Sigma^{(1)} \otimes \Sigma^{(1)}\right) < \tau_{max}/4$$

By Lemma 3, the eigenvalues of $\Sigma_1^{(2)} and\ \Sigma_2^{(2)}$ are also bounded by $\tau_{min}\ and\ \tau_{max}/4$, so

$$2\tau_{min} < \lambda_{min}\left(\Sigma^{(2)}\right) \leq \lambda_{max}\left(\Sigma^{(2)}\right) < \tau_{max}/2$$

It is straight forward to get (7.2)

Appendix B. A Bernstein Inequality and Its Application

Lemma 5

Let $W_1, \dots, W_n$ be independent random variables with mean zero and variances bounded by $\sigma^2 \geq 1$. Assume for some $0 < \alpha < 1$,

$$E\left(|W_i|^{3(1-\alpha)} e^{t|W_i|^{\alpha}}\right) \leq \mathrm{A}, \text{for all } 1 \leq i \leq n, \qquad 0 \leq t \leq T$$

Then for $x > \left(\frac{2\mathrm{A}}{T^2}\right)^{\frac{1}{(1-\alpha)}}$,

$$\boldsymbol{P}\left(\left|\sum_{i=1}^{n} W_i\right| \geq x\right) \leq 2 * \exp\left\{-\frac{x}{2\left(n\sigma^2 + \frac{x^{2-\alpha}}{T}\right)}\right\} + \sum_{i=1}^{n} \boldsymbol{P}(|W_i| \geq x)$$

Proof: Let $W_i^* = W_i \cdot I_{(-\infty,x]}(W_i)$. *Then*

$$\boldsymbol{P}\left(\sum_{i=1}^{n} W_i \geq x\right) \leq \boldsymbol{P}\left(\sum_{i=1}^{n} W_i^* \geq x\right) + \sum_{i=1}^{n} \boldsymbol{P}(W_i \geq x).$$

For $W_i^* \geq 0$, *we have*

$$e^{tW_i^*} \leq 1 + tW_i^* + \frac{t^2}{2} W_i^{*2} + \sum_{k=3}^{\infty} \frac{t^k}{k!} |W_i|^{k\alpha+3(1-\alpha)} x^{(k-3)(1-\alpha)}$$

Note that (7.6) is true also for $W_i^* < 0$ because of the montonicity of function $f(u) = e^u - 1 - u - \frac{u^2}{2}$

It is easy to get $E|W_i|^{k\alpha+3(1-\alpha)} \leq \frac{k!\mathrm{A}}{T^k}$ from (7.3). Moreover, we have $E(W_i^*) \leq 0,\ Var(W_i^*) \leq \sigma^2$ from definition. Taking the expectation of (7.6)

$$\begin{aligned} E\left(e^{tW_i^*}\right) &\leq 1 + \frac{t^2\sigma^2}{2} + \sum_{k=3}^{\infty} \frac{2\mathrm{A}}{T^2 x^{1-\alpha}} * \frac{1}{2} * \left(\frac{x^{1-\alpha}}{T}\right)^{k-2} t^k \\ &\leq 1 + \frac{t^2\sigma^2}{2} + \frac{t^2}{2} \sum_{k=3}^{\infty} \left(\frac{tx^{1-\alpha}}{T}\right)^{k-2} \\ &\leq 1 + \frac{t^2\sigma^2}{2\left(1 - \frac{t^{1-\alpha}}{T}\right)}, \\ &\quad when \left|\frac{tx^{1-\alpha}}{T}\right| < 1 \end{aligned}$$

$$Let\ t = \frac{x}{n\sigma^2 + \frac{x^{2-\alpha}}{T}}.\ By\ the\ Markov\ inequality$$

$$\begin{aligned} P\left(\sum_{i=1}^{n} W_i^* \geq x\right) &\leq e^{-tx} E\left(e^{t\sum_{i=1}^{n} W_i^*}\right) \\ &\leq e^{-tx} \prod_{i=1}^{n} E\left(e^{tW_i^*}\right) \\ &\leq e^{-tx} \left(1 + \frac{t^2\sigma^2}{2\left(1 - \frac{tx^{1-\alpha}}{T}\right)}\right)^n \end{aligned}$$

$$\leq \exp\left\{\frac{x^2}{n\sigma^2 + \frac{x^{2-\alpha}}{T}}\right\}\left(1 + \frac{x^2}{2n\left(n\sigma^2 + \frac{x^{2-\alpha}}{T}\right)}\right)^n$$

$$\leq \exp\left\{\frac{x^2}{2\left(n\sigma^2 + \frac{x^{2-\alpha}}{T}\right)}\right\}$$

Therefore,

$$\boldsymbol{P}\left(\sum_{i=1}^{n} W_i \geq x\right) \leq \boldsymbol{P}\left(\sum_{i=1}^{n} W_i^* \geq x\right) + \sum_{i=1}^{n} \boldsymbol{P}(W_i \geq x).$$

Lemma 6

Under condition (C1) and (C2), $for\ m = o\left(n^{\frac{1}{3}-\frac{1}{3}\xi}\right), \boldsymbol{\mathcal{M}} \subset \mathcal{P}_1,$

$$\boldsymbol{P}\left(\tau_{min} \leq \min_{|\mathcal{M} \leq m|}(\lambda_{min}(\widehat{\Sigma_{\mathcal{M}}})) \leq \max_{|\mathcal{M} \leq m|} \lambda\left(\widehat{\Sigma_{\mathcal{M}}}\right) \leq \tau_{max}\right) \to 1.$$

Furthermore, under condition (C4), (7.8) holds for $m = O\left(n^{2\xi_0 + 4\xi_{min}}\right) = o\left(n^{\frac{1}{3}-\frac{1}{3}\xi}\right)$

Lemma 7

Let $W_1, \ldots, W_n$ be independent random variables with zero mean and such that $E\left(e^{T_0|W_i^{\alpha}|}\right) \leq \mathrm{A}_0$ $for\ constants\ T_0 > 0,\ \mathrm{A}_0 > 0\ and\ 0 < \alpha < 1.$ Then for $\alpha\ sequence\ \alpha_n \to \infty\ with\ \alpha_n = o\left(n^{\frac{\sigma}{2(2-\alpha)}}\right), there\ exists\ constants\ c_1, c_2\ \ such\ that$

$$P\left(|W_1 + \cdots + W_n| \leq \sqrt{n\alpha_n}\right) \leq c_1 \exp(-c_2\alpha_n^2)$$

Proof:

The condition $E\left(e^{T_0|W_i^{\alpha}|}\right) \leq \mathrm{A}_0$ implies $Var(W_i) \leq \sigma^2$,
$E\left(|W_i|^2 e^{T|W_i^{\alpha}|}\right) \leq \mathrm{A}\ and\ E\left(|W_i|^{3(1-\alpha)} e^{T|W_i|^{\alpha}}\right) \leq$
A for some constant $\sigma^2, T, and\ \mathrm{A}.\ By\ Lemma\ 5, we\ have$

$$\boldsymbol{P}\left(\left|\sum_{i=1}^{n} W_i\right| \geq x\right) \leq 2\exp\left\{-\frac{x^2}{2\left(n\sigma^2 + \frac{x^{2-\alpha}}{T}\right)}\right\} + \sum_{i=1}^{n} \boldsymbol{P}(|W_i| \geq x)$$

Let $x = \sqrt{n}\alpha_n,\ Then$

$$\exp\left\{-\frac{x^2}{2\left(n\sigma^2 + \frac{x^{2-\alpha}}{T}\right)}\right\} = \exp\left\{-\frac{n\alpha_n^2}{2\left(n\sigma^2 + \frac{n^{\frac{2-\alpha}{2}}\alpha_n^{2-\alpha}}{T}\right)}\right\} = \exp\left\{-\frac{\alpha_n^2}{2\sigma^2 + o(1)}\right\}$$

On the other hand, by the Markov Inequality

$$\boldsymbol{P}(|W_i| \geq x) = \boldsymbol{P}\left(W_i^2 e^{T|W_i|^\alpha} \leq x^2 e^{Tx^\alpha}\right) \leq \mathrm{A}x^{-2}\exp(-Tx^2) \leq \frac{\mathrm{A}}{n\alpha_n^2}\exp\left(-\frac{T\alpha_n^2}{o(1)}\right)$$

Hence, $\sum_{i=1}^{n} P(|W_i| \geq x) \leq \frac{\mathrm{A}}{n\alpha_n^2}\exp\left(-\frac{T\alpha_n^2}{o(1)}\right)$. And (7.9) is easily obtained.

Remark 1. We are interested in the case that $W_i = X_{ij}X_{ik}X_{il}, where\ X_{ij}, X_{ik}, X_{il}$ are joint normal and marginally standard normal. It is easy to see that W_i satisfies $E\left(e^{\frac{1}{4}|W_i|^{\frac{2}{3}}}\right) \leq \sqrt{2}\ and\ Var(W_i) \leq 30$. Therefore, (7.9) holds for $c_1 = 3, c_2 = \frac{1}{61}\ when\ n$ is sufficiently large.

In order to show Theorem 2, we have to obtain an analogue of Lemma 6 for arbitrary submodel $\mathcal{M}$. We start from a generalization of Lemma A3 in Bickel & Levina (2008)

Lemma 8

Let $W_1, \dots, W_n$ be independent random variables with zero mean and such that $E\left(e^{T_0|W_i^\alpha|}\right) \leq \mathrm{A}_0 \; for \; constants \; T_0 > 0, \; \mathrm{A}_0 > 0 \; and \; 0 < \alpha < 1.$ Then there exists constants $c_3, c_4, for \; 0 < \epsilon \leq 1$

$$P(|W_1 + \cdots + W_n| \geq n\epsilon) \leq c_3 \exp(-c_4 n^\alpha \epsilon^2)$$

Proof:

The condition $E\left(e^{T_0|W_i^\alpha|}\right) \leq \mathrm{A}_0$ implies
$Var(W_i) \leq \sigma^2, E\left(|W_i|^2 e^{T|W_i^\alpha|}\right) \leq \mathrm{A} \; and \; E\left(|W_i|^{3(1-\alpha)} e^{T|W_i^\alpha|}\right) \leq$
$\mathrm{A} \; for \; some \; constants \; \sigma^2, T \; and \; \mathrm{A}. \; When \; \alpha < 1, by \; Lemma \; 5,$

$$\boldsymbol{P}\left(\left|\sum_{i=1}^n W_i\right| \geq x\right) \leq 2 * \exp\left\{-\frac{x}{2\left(n\sigma^2 + \frac{x^{2-\alpha}}{T}\right)}\right\} + \sum_{i=1}^n \boldsymbol{P}(|W_i| \geq x)$$

$Let \; x = n\epsilon. \; Then$

$$\exp\left\{-\frac{x^2}{2\left(n\sigma^2 + \frac{x^{2-\alpha}}{T}\right)}\right\} = \exp\left\{-\frac{n^2\epsilon^2}{2\left(n\sigma^2 + \frac{n^{2-\alpha}\epsilon^{2-\alpha}}{T}\right)}\right\}$$
$$= \exp\left\{-\frac{n^\alpha \epsilon^2}{2n^{\alpha-1}\sigma^2 + \frac{2\epsilon^{2-\alpha}}{T}}\right\}$$
$$\leq \exp\left\{-\frac{n^\alpha \epsilon^2}{o(1) + \frac{2}{T}}\right\}$$

On the other hand, by the Markov inequality

$$\boldsymbol{P}(|W_i| \geq x) = \boldsymbol{P}\left(W_i^2 e^{T|W_i|^\alpha} + x^2 e^{Tx^\alpha}\right) \leq \mathrm{A}x^{-2} \exp\{-Tx^\alpha\} \leq \frac{\mathrm{A}}{n^2\epsilon^2} \exp\{-Tn^\alpha \epsilon^\alpha\}.$$

Hence, $\sum_{i=1}^n \boldsymbol{P}(|W_i| \geq x) \leq \frac{\mathrm{A}}{n\epsilon^2} \exp\left\{-\frac{1}{2}Tn^\alpha\epsilon^\alpha\right\} \exp\left\{-\frac{1}{2}Tn^\alpha\epsilon^\alpha\right\} \leq o(1) \exp\left\{-\frac{1}{2}Tn^\alpha\epsilon^\alpha\right\}.$ And (7.10) is easily obtained.

When $\alpha = 1, E\left(e^{T_0|W_i^\alpha|}\right) \leq \mathrm{A}_0$ implies $\sum_{k=0}^\infty \frac{1}{k!} T_0^k E\left(|W_i|^k\right) \leq \mathrm{A}_0.$ So
$E\left(|W_i|^k\right) \leq \frac{1}{2} k! \left(\frac{1}{T_0}\right)^{k-2} \frac{2\mathrm{A}_0}{T_0^2} \; for \; k \geq 2.$ By Bernstein's Inequality, Lemma 2.2.11 in van der Vaart Wellner (1996), we have

$$\boldsymbol{P}\left(\left|\sum_{i=1}^n W_i\right| \geq n\epsilon\right) \leq 2\exp\left\{-\frac{n^2\epsilon^2}{2\left(\frac{2n\mathrm{A}_0}{T_0^2} + \frac{n\epsilon}{T_0}\right)}\right\} \leq 2\exp\left\{\frac{n\epsilon^2}{\frac{4A_0}{T_0^2} + \frac{2}{T_0}}\right\}$$

Lemma 9

Under condition (C1) and (C2), for $0 < \epsilon < 1, we\ have$

$$P\left(\left|\sum_{i=1}^{n} X_{si}X_{sj} - \sigma_{ij}\right| \geq n\epsilon\right) \leq C_1 \exp(-C_2 n\epsilon^2)$$

$$P\left(\left|\sum_{i=1}^{n} X_{si}X_{sj}X_{sk} - 0\right| \geq n\epsilon\right) \leq C_3 \exp\left(-C_4 n^{\frac{2}{3}}\epsilon^2\right)$$

$$P\left(\left|\sum_{i=1}^{n} X_{si}X_{sj}X_{sk}X_{sl} - \sigma_{ij}\sigma_{kl} - \sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk}\right| \geq n\epsilon\right) \leq C_5 \exp\left(-C_6 n^{\frac{1}{2}}\epsilon^2\right)$$

Where $C_1, \ldots, C_6\ are\ constants.$

Proof:
We show the last inequality here. The first two are similar.

$$Let\ W_s = X_{si}X_{sj}X_{sk}X_{sl} - \sigma_{ij}\sigma_{kl} - \sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk})$$

$$E\left(e^{\frac{1}{4}|W_s|^{\frac{1}{2}}}\right) = E\left(e^{\frac{1}{4}|X_{si}X_{sj}X_{sk}X_{sl} - \sigma_{ij}\sigma_{kl} - \sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk}|^{\frac{1}{2}}}\right)$$

$$(\because (a+b)^{\frac{1}{2}} \leq a^{\frac{1}{2}} + b^{\frac{1}{2}}) \leq E\left(e^{\frac{1}{4}|X_{si}X_{sj}X_{sk}X_{sl} - \sigma_{ij}\sigma_{kl} - \sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk}|^{\frac{1}{2}}}\right)$$

$$(\because |\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} \leq 3) \leq e^{\frac{\sqrt{3}}{4}} E\left(e^{\frac{1}{4}|X_{si}X_{sj}X_{sk}X_{sl}|^{\frac{1}{2}}}\right)$$

$$(\because |abcd \leq \frac{a^2+b^2+c^2+d^2}{4}) \leq e^{\frac{\sqrt{3}}{4}} E\left(e^{\frac{1}{4}\frac{X_{si}^2+X_{sj}^2+X_{sk}^2+X_{sl}^2}{4}}\right)$$

$$(again\ abcd \leq \frac{a^2+b^2+c^2+d^2}{4}) \leq e^{\frac{\sqrt{3}}{4}} E\left(\left[e^{\frac{X_{si}^2}{4}} + e^{\frac{X_{sj}^2}{4}} + e^{\frac{X_{sk}^2}{4}} + e^{\frac{X_{sl}^2}{4}}\right]/4\right)$$

$$= \sqrt{2} e^{\frac{\sqrt{3}}{4}}$$

The inequality follows directly from the last lemma.