STATISTICAL METHODS FOR GENETICS AND GENOMICS STUDIES

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

MEIJUAN LI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DECEMBER, 2008

Acknowledgement

I would like to express my gradate to all those who gave me the possibility to complete this thesis. My thesis could not have been accomplished without them.

For the first and foremost, I would like thanks my advisor Dr. Cavan Reilly for his guidance, insightful feedback, and brilliant editing, and financial support, which has been a tremendous help for me over the years. The many hours of discussions we had in which he showed his enthusiasm, his patience, and positive attitude towards scientific research. It has been a pleasure to work under his supervision.

I would like to thank the other members of my PH.D. defense committee, Dr. Wei Pan, Dr. Xiaotong Shen, and Dr. Saonli Basu. They have been endlessly merciful in theirs valuable suggestions and reading through my thesis.

I would also like to thank Dr. Tim Hanson, who has always offered me help whenever it is needed through these years. His help, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis.

Special thanks are also extended to the Division of Biostatistics, University of Minnesota for the financial support and the Graduate school, University of Minnesota for the Doctoral Dissertation Fellowship, which allowed me to devote full-time effort to the research and writing of the dissertation in the last year of my PH.D study.

I would also like to extend my thanks to the help and friendship from all other faculty members, staff, and fellow students in the Biostatistics Division. The past

several years studying in University of Minnesota have been the one of most enjoyable experience in my life.

My heartfelt thanks to my parents for theirs love and support. My deepest gratitude I must reserve for my wonderful husband, Huimin Sun, whose love, support, patience, understanding I am very thankful for. Finally, my deepest gratitude I also must reserve for my dearest son, Yujie Sun, whose love, support, patience, and understanding I am deeply thankful for.

Abstract

Genomics study: the data quality from microarray analysis is highly dependent on RNA quality. Because of the lability of RNA, steps involved in tissue sampling, RNA purification, and RNA storage are known to potentially lead to the degradation of RNAs, therefore, assessment of RNA quality is essential. Existing methods for estimating the quality of RNA on microarray either suffer from subjectivity or are inefficient in performance. To overcome these drawbacks, in this dissertation, a linear regression method for assessing RNA quality for a hybridized Genechip is proposed. In particular, our approach used the probe intensities that the Affymetrix software associates with each microarray. The effectiveness and improvements of the proposed method over the existing methods are illustrated by the application of the method to the previously published 19 human Affymetrix microarray data sets for which external verification of RNA quality is available.

Genetics study: although population-based association mapping may be subject to the bias caused by population stratification, alternative methods that are robust to population stratification such as family-based linkage analysis have lower mapping resolution. In this dissertation, we propose association tests for fully observed quantitative traits as well censored data in structured populations with complex genetic relatedness among the sampled individuals. Our methods correct for continuous population stratification by first deriving population structure variables and

kinship matrices through random genetic marker data and then modeling the relationship between trait values, genotypic scores at a candidate marker, and genetic background variables through a semiparametric model, where the error distribution for fully observed data or the baseline survival function for censored data is modeled as a mixture of Polya trees centered around a family of parametric distributions. We also propose multivariate Bayesian statistical models with a Gaussian conditional autoregressive (CAR) framework for multi-trait association mapping in structured populations, where the effects attributable to kinship matrix is modeled via CAR and the population structure variables are included as covariates to adjust populations stratification. We compared our model to the existing structured association tests in terms of model fit, false positive rate, power, precision, and accuracy using real data sets as well as simulated data sets.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Massive quantities of genomic and genetic data and highthroughput technologies are now enabling studies on a vastly larger scale than ever before. Examples include simultaneously monitoring and comparing the activity of tens of thousands of genes and genotyping of hundreds of thousands of polymorphisms in a sample of tissue by a DNA microarray. Advanced computational tools and statistical methods are needed to capture, represent, store, integrate, and analyze the data with the ultimate goal of understanding and modeling living systems. Statistical challenges will continue to rise in an unprecedented way with the explosive torrent of data from large-scale studies at the molecular, cellular, and whole-organism levels.

Gene mapping has become one of the most active research areas in genetics. The goal is to identify genetic variants which underlie the trait of interest. Most current gene mapping in human genetics detects mutations associated with rare genetic disorders that follow Mendelian inheritance patterns. These include myotonic and Duchenne muscular dystrophies, cystic fibrosis, neurofibromatosis type 1, sickle cell anemia, and Huntington's disease. Recently, tests have been developed to detect genetic variants for more complex traits such as heart disease, diabetes, and cancer.

Genomics study

The most common application of microarrays is gene expression analysis. Affymetrix short oligonucleotide microarrays have become the most widespread platform. The

hybridization between fluorescently labeled mRNA in a biological sample and complementary DNA probes, which are affixed to the array, allows rapid quantification of the entire transcriptome. This high-throughput approach has opened up entirely new avenues of research for biologists. Rather than experimentally confirming the hypothesized role of a certain candidate gene in a certain cellular process, they can now use genome-wide comparisons to screen for all genes which might be involved in that process. One of the first examples of such an exploratory approach was the expression profiling study of mitotic yeast cells which determined a set of a few hundred genes involved in the cell cycle by Cho et al. (1998). The study triggered many articles re-analyzing the data or replicating the experiment. Microarrays have also become a central tool in cancer research initiated by the discovery and re-definition of tumor subtypes based on molecular signatures i.e. the expression profile.

However, complex techniques such as microarrays leave many opportunities for errors. After the initial euphoria, the research community soon became aware that findings based solely on microarray measurements were not always as reproducible as they would have expected and that studies with inconclusive results were quite common. With this high-throughput measurement technology becoming established in many branches of life sciences research, scientists in both academic and corporate environments raised their expectations concerning the validity of the measurements. Data quality issues are now frequently addressed. Assessing the quality of microarray

data has emerged as a new research topic for statisticians. For Affymetrix arrays, the commercial software GCOS (2004) includes a quality report with a dozen scores for each microarray. However, the quality picture delivered by the GCOS quality report is incomplete or not sensitive enough, and that it is rarely helpful in assigning causes to poor quality. Some authors have addressed at specific issues. An algorithm for probeset quality assessment has been suggested by Bolstad (2003). Finkelstein (2005) evaluated the Affymetrix quality reports of over 5,000 chips collected by St Jude Children Research Hospital over a period of three years, and linked some quality trends to experimental conditions. Hu et al. (2005) extended traditional effect size models to combine data from different microarray experiments, incorporating a quality measure for each gene in each study. Gautier et al. (2004) investigated the effect of updating the mapping of probes to genes on the estimated expression values. Archer et al. (2006) proposed a mixed-effect model to estimate the 3'/5' ratio and obtain its confidence interval for an individual Affymetrix GeneChip using the interior pixel-level intensity. However, the literature on quality assessment for short oligonucleotide arrays is still sparse, though the importance of the topic has been stressed. We here will propose an approach that is intended to address this issue.

Genetics study

One approach to gene mapping is called linkage analysis, which uses families with a known pedigree structure. Individuals are genotyped at random markers spread

across the genome. If a disease gene, for example, is close to one of the markers then, within the pedigree, the inheritance pattern at the marker will mimic the inheritance pattern of the disease itself. Linkage analysis has been highly successful at identifying genes for simple genetic diseases: i.e., those in which a single major gene is responsible for the disease in a given pedigree, and environmental factors are not very important. The central problem with linkage analysis for fine mapping is the limited number of meioses that have occurred and the cost of collecting family members to allow for a sufficient number of meioses. An alternative approach to gene mapping is "association mapping" or linkage disequilibrium (LD) mapping, taking advantage of events that created association in the relatively distant past. Assuming many generations, and therefore meioses, have elapsed since these events, recombination will have removed association between a quantitative trait loci, and any marker not tightly linked to it. Association mapping thus allows for much finer mapping than linkage analysis (Hästbacka et al., 1992). It has also been argued that in conjunction with new technology for rapid genotyping, this method will ultimately be more powerful than linkage analysis for isolating genes of small effect, which include most of the genes for common diseases. An association between a neutral marker allele and the phenotype occurs when marker alleles are in LD with alleles at a trait locus. Two alleles at distinct loci are in positive LD if they occur together more often than predicted on the basis of their individual frequencies. A

variety of mechanisms generate linkage disequilibrium, and several of these can operate simultaneously. (1) Populations expanding from a small number of founders. The haplotypes present in the founders will be more frequent than expected under equilibrium. Three special cases are noteworthy. First, genetic drift affects LD by this mechanism in that a population experiencing drift derives from fewer individuals than its present size. Second, by considering an individual with a new mutation as a founder, we see that its descendants will predominantly receive the mutation and loci linked to it in the same phase. Linked marker alleles will therefore be in LD with the mutation. (2) Structured populations when allelic frequencies differ at two loci across subpopulations, irrespective of the linkage status of the loci. Admixed populations, formed by the union of previously separate populations into a single panmictic one, can be considered a case of a structured population where substructuring has recently ceased. (3) Negative LD will occur between loci affecting a character in populations under stabilizing or directional selection as a result of the Bulmer effect. (4) Positive LD will occur between loci affecting a character under disruptive selection. (5) When loci interact epistatically, haplotypes carrying the allelic combination favored by selection will also be at higher-than-expected frequencies.

Population-based association mapping has become a powerful, general tool for identifying loci associated with the inheritance of complex traits, however, this approach may be subject to bias caused by population stratification. Several statistical

methods were recently proposed to utilize genomic markers to control for population stratification that may be present in a sample of unrelated individuals, both in the analysis of qualitative traits (Devlin and Roeder, 1999; Bacanu et al., 2000; Devlin et al., 2001; Pritchard et al., 2000a; Pritchard et al., 2000b; Reich and Goldstein, 2001; Satten et al., 2001; Zhang et al., 2002) and in the analysis of quantitative traits (Zhang and Zhao, 2001; Bacanu et al., 2002). These approaches are promising because they may have higher resolution mapping than family-based association designs, and they may be robust against potential population stratification. However, the current available methods for association mapping in structured populations are for the univariate case only. We here propose a multivariate association mapping method via Bayesian conditional autoregressive modeling (CAR). Secondly, the association between a candidate gene and a quantitative trait for censored or fully observed data is often evaluated via a parametric model where inferred population structure variables are included as covariates, although this may be inappropriate for the analysis of many real-life data sets, where increased skewness, kurtosis, and multimodality may hold. Suitable transformation of the response to achieve comfort with the assumptions under a standard parametric family for studies with large numbers of candidate loci may be difficult because the appropriate transformation may vary from one marker to another, and interpretation problems may arise. Enrichment of standard families usually fails to capture all the desirable features. Hence,

association tests that avoid parametric specification and are robust against potential population stratification are desirable; our goals in this dissertation are also to describe statistical methods that is intended to have such desired properties. We propose semiparametric statistical models for the association test between candidate genes and fully observed data as well as for censored data. The proposed methods correct for continuous population stratification using population structure variables and kinship matrix. Both population structure variables and kinship matrix were inferred from random genetic marker data of the sampled individuals. The relationship between trait values, genotypic scores at a candidate gene, and genetic background variables is modeled using a semiparametric approach, where the error distribution for fully observed data or the baseline function for the censored data is modeled as a mixture of Polya trees centered around a family parametric distributions and thus may be viewed as a generalization of standard models in which important, data-driven features, such as skewness and multimodality, are allowed. To illustrate its utility and interpretation, we applied the proposed models to the real data sets as well as the simulated data sets. We compared our proposed models to the existing association tests in terms of model fit, power, and type I error rate, precision, and accuracy.

Chapter 2

Assessing the Quality of Hybridized RNA in Affymetrix GeneChips Using Linear Regression

2.1 Abstract

The quality of data from microarray analysis is highly dependent on the quality of RNA. Because of the lability of RNA, steps involved in tissue sampling, RNA purification, and RNA storage are known to potentially lead to the degradation of RNAs, therefore, assessment of RNA quality and integrity is essential. Existing methods for estimating the quality of RNA hybridized to a GeneChip either suffer from subjectivity or are inefficient in performance. To overcome these drawbacks, in this article, a linear regression method for assessing RNA quality for a hybridized Genechip is proposed. In particular, our approach used the probe intensities from the `.cel` files that the Affymetrix software associates with each microarray. The effectiveness and the improvements of the proposed method over the existing methods are illustrated by the application of the method to the previously published 19 human Affymetrix microarray data sets for which external verification of RNA quality is available.

2.2 Introduction

Hybridization-based DNA microarray technologies have evolved rapidly to become a key high-throughput technology for the simultaneous measurement of the relative expression levels of thousands of individual genes (DeRisi et al., 1996). Over the past several years, microarray technology has been used to explore transcriptional profiles and to obtain molecular expression signatures of the state of activity of diseased cells and patient samples (Xiang et al., 2003). In the field of cancer, microarray analyses have provided information on pathology, progression or resistance to treatment (Pusztai et al., 2003; Chang et al., 2003). However, the reliability of microarray technology to detect transcriptional differences representative of original samples is affected by several factors such as array platform, RNA extraction methods, probe labeling, hybridization conditions, and image analysis (Schuchhardt et al., 2000). In particular, the quality of data from microarray analysis is highly dependent on the quality of the RNA extracted from the tissues which is in turn dependent on the quality of the tissue samples. Because of the lability of RNA, various steps involved in tissue sampling, RNA purification, and RNA storage are known to potentially lead to degradation of the mRNAs by cleavage of RNases. The standard Affymetrix protocol uses as starting material $5-40$ μg of total RNA, which is first reverse transcribed using a T7-Oligo (dT) promoter primer into cDNA molecules from which biotinylated cRNA is synthesized for hybridization onto GeneChip probe arrays. The T7-Oligo

(dT) promoter primer contains an oligo (dT) 24 sequence at its 3' end for specific binding to the poly(A) tail of mRNA and the core sequence of the T7 RNA polymerase promoter at its 5' end. Consequently, the cDNA yield from the sequences near the 5' end of partially degraded mRNA is significantly less than that from the sequences near the poly(A) tail (Swift et al., 2000; Turchin et al., 2006). Hence, tolerance of this platform to degraded RNA samples may lie in the nature of the Affymetrix GeneChip design, which is 3'-biased. To overcome this obstacle, oligonucleotide probes are usually designed to be within the most 3' 600 bp of a transcript. Additional probe sets in the 5' region of the transcript have also been selected for certain housekeeping genes, including GAPDH , ISGF, and β-actin. Signal intensity ratio of the 3' probe set over the 5' probe set is often referred to as the 3'/5' ratio. This ratio gives an indication of the integrity of the starting RNA, efficiency of first strand cDNA synthesis, and/or in vitro transcription of cRNA (Affymetrix, 2001). It has been suggested in the literature that the 3'/5' ratio of a given microarray should be examined for assessing the quality of the mRNA hybridized to the array (Croner et al., 2004; Copois et al., 2007).

A method, referred to as RNA digestion plots, for assessing mRNA integrity is available in the Bioconductor *affy* package in the R programming environment (Gautier et al., 2004). RNA degradation plots which consider all probes across an array show expression as a function of the 5'-3' position of probes. The plot shows

the average intensity of each probe plotted against its interrogation position and the slope of its trend indicates potential RNA degradation of cRNA hybridized to the array. However, the RNA digestion plots can be ineffective for assessing RNA integrity as all mRNA samples could exhibit the same linear trend regardless of the extent of RNA degradation presents (Archer et al., 2006). Therefore, an objective RNA quality assessment method based on a statistical model is desirable. The GeneChip Operating Software (GCOS) estimates the 3'/5' ratio using the perfect match (PM) and mismatch (MM) probe set expression measure (Affymetrix, 2003; Hubbell et al., 2002). Affymetrix suggests that ratios greater than 3 may indicate degraded RNA or insufficient transcription (Affymetrix, 2003). Several methods for summarizing probe-level expression data into probe set summaries have been reported in the literature (Irizarry et al., 2003a,b; Lemon et al., 2002; Li and Wong, 2001). The 3'/5' ratios calculated after summarizing probe-level data using the MAS 5.0 (Affymetrix), robust multiarray average (RMA) (Irizarry et al., 2003a), GC-RMA (Wu et al., 2004), and the PM-only MBEI algorithms (Li and Wong, 2003) may yield different conclusion regarding sample quality depending upon whether the GAPDH or β-actin 3'/5'ratios are compared to the recommended threshold of 3 (Archer et al. 2006). These irresolvable discrepancies between the 3'/5' ratios for different transcripts within the same GeneChip present a difficulty in drawing a final conclusion regarding sample quality. Furthermore, all transcripts but a few are represented by

only one probe set on a GeneChip. Due to the lack of replicates of 3' and 5' probe sets on a GeneChip, it is impossible to provide an estimate of the variance for the 3'/5' ratio when probe set expression summaries are used in the estimation of the 3'/5' ratio. Further, using an arbitrarily selected 3'/5' ratio threshold for determining RNA quality is analogous to the use of fold-change thresholds for identifying differentially expressed genes, and this has been shown to be inferior to statistical methods which take variability into account (Dudoit et al., 2002; Zhang et al., 2002). Establishing a viability threshold for determining RNA quality is therefore problematic, and this becomes especially challenging when different probe set expression summary methods are available for use. To overcome the drawbacks in the above methods, Archer et al. (2006) proposed a mixed-effect model to estimate the 3'/5' ratio and obtain its confidence interval for an individual GeneChip using the interior pixel-level intensity. The model had complex a multi-level nested covariance structure including the random effects for probe set, probe, and pixel. There are several drawbacks to this approach that limit its use in practice. (1) The pixel-level intensity used in this approach is not a routine microarray data type associated with each microarray and requires special data processing to obtain. (2) The effect of the interrogation position of probes was modeled as 40 separate cluster effects, thereby ignoring the linear dependence of probe expression level on the 5'-3' position of probes.

To avoid such drawbacks, we here propose a linear regression approach to model

mRNA degradation. We perform analysis on probe-level mRNA expression data from the `.cel` files arising from a typical microarray experiment. For each control gene, we also model the dependence of the log_2 transformed probe expression level on probe interrogation position as a linear function of log_2 transformed probe position. Our approach to assessing the quality of hybridized RNA in Affymetrix GeneChip experiments differs from previous investigations in a fundamental way.

2.3 Data sources

Here we use the same data sets as those used in Archer's mixed model paper (Archer et al., 2006). For the proposed method of assessing RNA degradation, the 3' and 5' end probe sets for three human housekeeping genes were retained for analysis (Table 2.1) for an individual GeneChip i.e. there are a total of six probe sets which are from three housekeeping genes and located at the two ends of each. For each of six probe sets, there are 20 different probes whose proximities to the 5' end (or 3' end) of the gene are indicated by their unique interrogation position (the position of the 13th ("middle") nucleotide of the probe sequence as it aligns on the consensus/exemplar sequence). We used the probe intensities from the `.cel` files that the Affymetrix software associates with each microarray. No processing of the data was performed on the values obtained from these files (other than taking the logarithm). We conducted the analysis using just the PM probes. Probe sequence information

(i.e. probe sequence composition and location in the transcript) was obtained from the Affymetrix website (`www.affymetrix.com`).

2.3.1 Ovarian tumor samples

Previously published data (Dumur et al., 2004) including five pairs of HG-U133A GeneChips were hybridized as follows. First, total RNA was isolated from multiple 10 μm frozen sections from five snap-frozen ovarian tumor samples using TRIZOL reagent. For each tumor sample, similar-sized aliquots of total RNA were processed with and without a subsequent cleanup process using RNeasy reagents. The RNeasy cleanup should lead to good-quality RNA (GeneChips Ovarian1 G, Ovarian2 G, Ovarian3 G, Ovarian4 G, and Ovarian5 G), whereas lack of the cleanup step should lead to poor-quality RNA (GeneChips Ovarian1 I, Ovarian2 I, Ovarian3 I, Ovarian4 I, and Ovarian5 I). The samples were obtained and processed according to a Virginia Commonwealth University's Institutional Review Board approved protocol. Details confirming the RNA quality, such as absorbency ratios, 28S/18S ratios, and length of cDNA and cRNA synthesis products, are reported elsewhere (Dumur et al., 2004).

2.3.2 Renal cell carcinoma samples

Previously published data consisting of nine Affymetrix GeneChips (4 HG-U133A; 5 HG-Focus) were collected to assess the impact of RNA degradation on microarray gene expression in renal cell carcinoma samples (Schoor et al., 2003) at the University

Table 2.1: The interrogation positions and the locations of the probe set for three housekeeping genes. Length of the transcripts in base pair (bp) is also provided

Gene	Transcript length	Probe set ID	Probe set location	Interrogation positions (max , min)
β-actin	1793bp	HSAC07/X00351_5_at	5'	(62, 572)
		HSAC07/X00351_3_at	3'	(1213, 1726)
ISGF	4157bp	HUMISGF3A/M97935_5_at	5'	(253,784)
		HUMISGF3A/M97935_3_at	3'	(2292, 3689)
GAPDH	1283bp	HUMGAPDH/M33197_5_at	5'	(99, 375)
		HUMGAPDH/M33197_3_at	3'	(920, 1244)

of Tübingen. The goal of this study was to determine the effects of a two-round IVT protocol on 20 ng of partially degraded RNA on gene expression values in comparison to expression values obtained when high-quality RNA is hybridized. The RNA extraction and chemical degradation procedures are described elsewhere (Schoor et al., 2003). The nine samples from this study have been labeled as follows: (1) the first letter indicates cell type, with N indicating normal cells and T indicating tumor cells; (2) the second letter reflects the level of degradation, ranging from A indicating freshly isolated RNA (i.e. no degradation) to D indicating the highest amount of degradation; (3) the third letter reflects the GeneChip used, with U indicating HG-U133A and F indicating HG-Focus. The extent of degradation was confirmed for all samples by electropherogram (Schoor et al., 2003).

2.4 Statistical model

The method described applies to a single array. In this work, we consider gene, probe set location i.e. 3' or 5' end of a transcript, and probe interrogation position as the three covariates that influence probe intensity. In addition, to further account for the impact of the sequence on the measured intensity, the $\log_2$ of the percent GC content ($\log_2$(GC%)) of a probe was included as a covariate, which has been shown to be a more effective method for adjusting for the probe-affinity effect than that proposed by Wu et al., (2004) (Archer et al., 2006). Instead of $\log_2$(GC%), we also considered $\log_2$ percent C content ($\log_2$(C%)) of a probe as a covariate for adjusting for the probe-affinity effect. The three genes are assumed independent. Let y_{ijk} represent the $\log_2$ signal intensity for the i^{th} housekeeping gene (i=1, 2, 3), the j^{th} end of the transcript with $j = 1$ and $j = 2$ indicating the 3' end and the 5' end, respectively, and the k^{th} probe (k=1, 2, ..., 20). Let $x_{1ijk} = 1$ if $j = 1$, and 0 otherwise. Let x_{2ijk} represent the $\log_2$(probe interrogation position), x_{3ijk} represent $\log_2$(GC%) (or $\log_2$(C%)) for the k^{th} probe at the j^{th} end of gene i, respectively, and let ϵ_{ijk} represent the *i.i.d.* random error for the k^{th} probe at the j^{th} end of gene i with $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$. Here we propose the following statistical model to estimate the 3'/5' ratio for an individual GeneChip,

$$y_{ijk} = \mu + \theta_i + \beta_1 x_{1ijk} + \beta_{2i} x_{2ijk} + \beta_3 x_{3ijk} + \epsilon_{ijk} \tag{2.1}$$

Note here, β_1 represents the fixed effect associated with 3' end of the transcript, θ_i represents the fixed effect associated with i^{th} gene with $\sum_{i=1}^{3} \theta_i = 0$, β_{2i} is the linear regression coefficient associated with the $\log_2$(probe interrogation position) for gene i, and β_3 is the linear regression coefficient associated with the $\log_2$(GC%) (or $\log_2$(C%)). The least square method will be used to estimate the model parameters and their associated variance. Since a $\log_2$ transformation of the signal intensities was applied to meet model assumptions (Gaussian), the parameter 2^{β_1}, 3'/5' ratio and it's 95% confidence interval were reported using the original scale by retransforming the estimated value of the parameter to the original scale via antilogs. This transformation was used to compare our results to Archer's, but in practice on can just examine the 95% confidence interval for β_1 to assess RNA quality.

2.5 Results

All analysis was carried out in R and SAS. Table 2.2 and Table 2.3 report the point estimates and the 95% confidence intervals of the 3'/5' ratios for the 10 ovarian GenChips and the 9 renal cell carcinoma GeneChips, respectively. For the ovarian samples, the 95% confidence intervals for all 5 GeneChips hybridized with RNA processed without the RNeasy cleanup step did not include one, suggesting poor RNA quality, as appropriate. For the ovarian samples for which the RNeasy cleanup procedure was used, the 95% confidence intervals appropriately included one for all 5

samples, suggesting adequate sample quality, as expected. For the normal and tumor renal cell samples processed when no degradation was present, the 95% confidence intervals included one for all 5 samples, as appropriate (Table 2.3). For the degraded samples, the 95% confidence intervals exclude one for all but NDU when $\log_2$GC% was used for adjusting probe affinity, suggesting poor RNA quality, again as expected (Table 2.3). Also note that the point estimates of the 3'/5' ratios for all 9 poor quality GeneChips are much higher than the point estimates of their corresponding GeneChips with good quality (Table 2.3 and Table 2.3), as expected. Notice that the 95% confidence intervals for the 3'/5' ratios from the current model are remarkably smaller than those from Archer's mixed model for all 19 GeneChips but one (Ovarian inhibited 2), thereby suggesting that the current model generally provides more precise estimates for the 3'/5' ratios than Archer's mixed model (Figure 2.1).

Furthermore, the point estimates of the 3'/5' ratios from the current model are distinctly different depending on whether degradation is known to be truly present or absent in the sample. For example, the 3'/5' ratios are smaller than 2 for all 5 Ovarian samples with the RNeasy cleanup step and all 5 renal cell samples with good mRNA quality, while the 3'/5' ratios are greater than 2 for all 5 Ovarian samples without the RNeasy cleanup step and for all 4 degraded renal cell samples. However in Archer's model, no such clear-cut separation was observed among the 3'/5' ratios with respect to sample quality. For example, Archer's results showed that the 3'/5'

Table 2.2: The point estimate and the 95% confidence intervals of 3'/5' ratios for the 10 ovarian GeneChips from the linear regression model

GeneChip	With RNeasy cleanup (good)			No RNeasy cleanup (inhibited)		
	2.5%	50%	97.5%	2.5%	50%	97.5%
(A) $\log_2$ percent GC content						
Ovarian 1	0.255	0.593	1.389	8.026	16.056	32.188
Ovarian 2	0.584	1.284	2.825	10.700	21.640	43.596
Ovarian 3	0.560	1.185	2.500	4.924	9.753	19.389
Ovarian 4	0.561	1.191	2.535	7.617	14.850	29.010
Ovarian 5	0.626	1.295	2.681	6.037	12.508	25.950
(B) $\log_2$ percent C content						
Ovarian 1	0.326	0.716	1.578	9.072	17.845	35.120
Ovarian 2	0.737	1.536	3.195	12.139	24.017	47.591
Ovarian 3	0.717	1.420	2.819	5.577	10.859	21.105
Ovarian 4	0.668	1.370	2.805	8.350	16.161	31.310
Ovarian 5	0.765	1.510	2.979	6.460	13.411	29.901

Table 2.3: The point estimate and the 95% confidence intervals of 3'/5' ratios for the 9 renal cell carcinoma GeneChips from linear regression model

GeneChip	Good			Degraded		
	2.5%	50%	97.5%	2.5%	50%	97.5%
(A) $\log_2$ percent GC content						
Normal U133A	0.430	1.075	2.667	0.915	2.184	5.226
Normal Focus	0.735	1.762	4.224	1.098	2.506	5.752
Tumor U133A	0.571	1.300	2.938	1.016	2.174	4.652
Tumor Focus 1	0.695	1.522	3.330	2.316	5.168	11.528
Tumor Focus 2	0.848	1.957	4.517			
(B) $\log_2$ percent C content						
Normal U133A	0.544	1.287	3.045	1.146	2.611	5.952
Normal Focus	0.827	1.978	4.729	1.164	2.708	6.310
Tumor U133A	0.637	1.443	3.266	1.161	2.436	5.101
Tumor Focus 1	0.673	1.563	3.635	2.116	5.127	12.427
Tumor Focus 2	0.850	2.051	4.951			

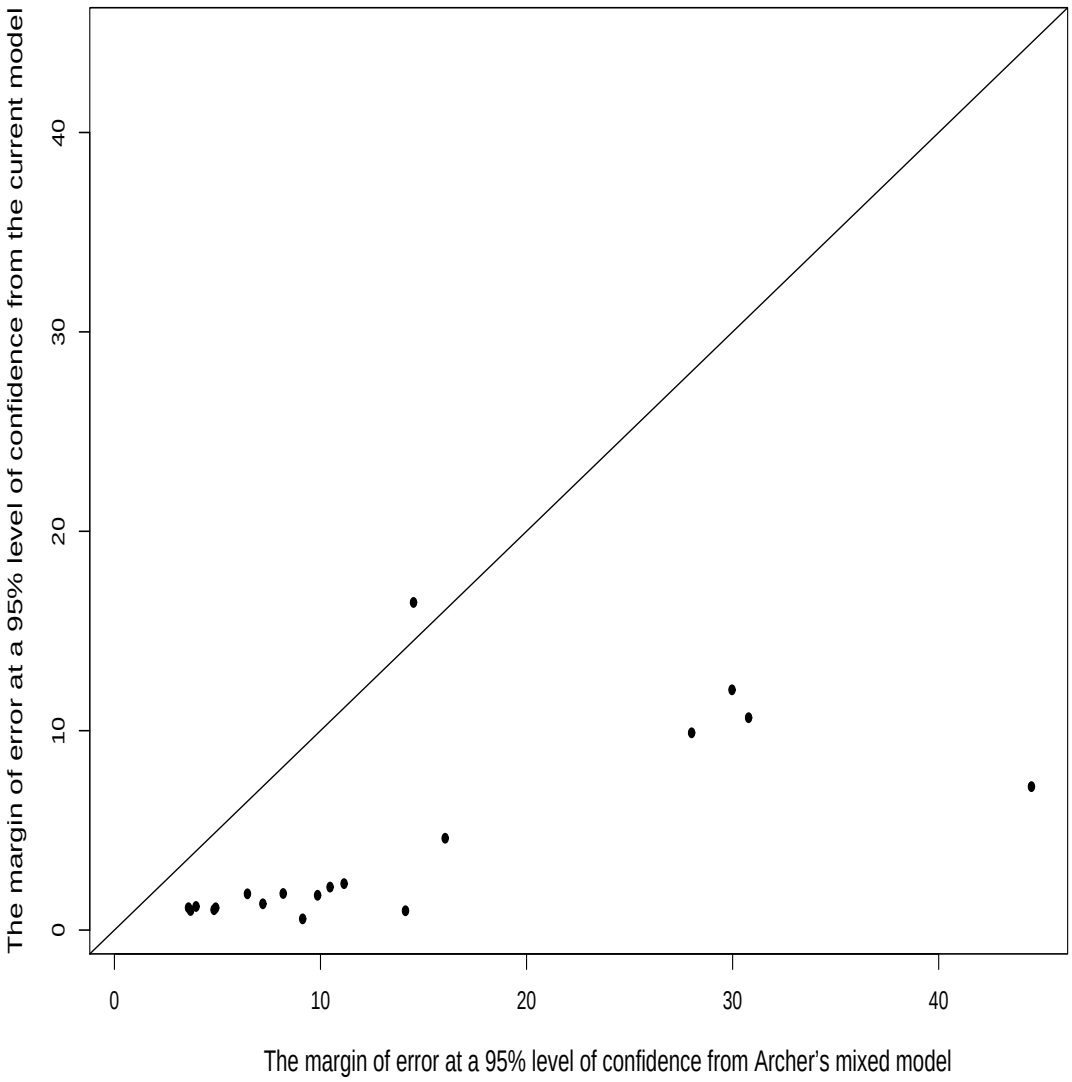

Figure 2.1: The scatter plot of the margin of error at a 95% level of confidence for the current model versus the margin of error at a 95% level of confidence from Archer's model for 19 GeneChips. The diagonal line in the plot demonstrates that the margins of error from the current model are smaller than those from the Archer's model for all 19 GeneChips except for one GeneChip (Ovarian2 I).

ratio for a decayed biological sample with poor mRNA quality, GeneChip NBF, was 5.41 and it was 4.51 for the same biological sample with good quality RNA GeneChip, NAF. This observation leads to the conclusion that the current model leads to more accurate estimates than Archer's mixed model.

2.5.1 Model diagnostics

To determine if measurements from the same probe sets are correlated and how they are correlated, we considered 6 different types of covariance structure: (1) independent, (2) compound symmetry, (3) first order autoregressive, (4) Gaussian spatial i.e. the correlation between two observations at a distance of r probe position apart on the $\log_2$ scale is $\exp(-(r/\varphi)^2)$, where φ is a parameter that is estimated via REML. (5) spherical spatial i.e. the correlation between two observations a distance r apart is $1 - 1.5(r/\varphi) + 0.5(r/\varphi)^3$, and (6) exponential spatial i.e. the correlation between two observations a distance r apart is $\exp(-r/\varphi)$. We used the BIC values to select the best covariance structure. Among 19 GeneChips, 17 out of 19 had the independence model give the best BIC value.

2.5.2 Effect of RNA quality on expression summaries

It is known that the degraded mRNAs hybridized to a GeneChip will yield lower expression levels than high quality mRNAs, however, it has not been shown in the literature to what extent the bias and variability are related to mRNA quality and

their impact on downstream microarray data analysis. To illustrate such impact, here we investigate the effect of mRNA quality related bias on the detection of differential expression and investigate if commonly used normalization procedures will correct this. We used the probe set summaries without normalization, which were generated using the RMA method implemented in the Bioconductor *affy* package in the R programming environment (Gautier et al., 2004). We conducted paired t-tests on those unnormalized probe set summaries i.e. gene expression data for 5 pairs of Ovarian tumor samples (Dumur et al., 2004) for each of the 22,283 probe sets on HG133UA GeneChip. For 15,502 out of 22,283 probe sets, the gene expression level was significantly higher in the Ovarian samples processed using the RNeasy cleanup step (good) than in the ovarian samples processed without the RNeasy cleanup step (inhibited) at a significance level of 0.05, while only 5 out of 22,283 probe sets had significantly lower expression levels in good Ovarian samples than in inhibited ovarian samples. This indicated that false discovery rates (FDR) may exceed 50% even in a sample of size 10. To determine if the mRNA quality related bias will be corrected by commonly used normalization procedures, we used the same five pairs of Ovarian tumor samples, but we normalized the gene expression data by (1) the RMA method implemented Bioconductor *affy* package in the R programming environment (Gautier et al., 2004) and (2) Centering the expression data by subtracting the overall mean expression of an array. For RMA-normalized expression data, we found that

for 1,576 probe sets, the expression level in good quality Ovarian samples was significantly higher than in inhibited ovarian samples at a significance level of 0.05. In addition, 3,152 probe sets had significantly lower expression levels in good Ovarian samples than in inhibited ovarian samples. Almost identical results were observed in the centered-normalized expression data. This observation suggests that the normalization procedures significantly reduced the negative bias due to poor mRNA quality, however, it did not remove the bias completely. In fact, the normalization procedures induced additional positive bias which was not present in the unnormalized data. With a p-value cutoff of 0.05, FDR remained as high as over 20% for the normalized data. Furthermore, at least half of the 4,728 (3152+ 1576) probe sets had expression levels greater than the overall mean expression level of an array for all 10 GeneChips. Hence, the large FDR in the normalized data will not be removed by restricting the data set to those probe sets with higher expression levels.

To investigate the variability of expression levels among samples with the respect to mRNA quality, we also did an F-test using the same five pairs of Ovarian tumor GeneChips (Dumur et al., 2004) for each of 22,283 probe sets on the GeneChips. In the unnormalized data, no probe set had significantly larger variance in the inhibited ovarian samples than in the good quality Ovarian samples, while 831 probe sets had significantly larger variance in the good quality Ovarian samples than in the inhibited ovarian samples at a significance level of 0.05. Similar results were obtained using

the normalized data sets. This indicated that there is no evidence that poor quality mRNAs lead to larger variability of gene expression levels among samples than good quality mRNAs. So inverse weighting by the variance of the probe using weighted least square will not allow one to easily analyze sets of arrays that differ in terms of RNA quality.

2.6 Discussion

The integrity of mRNA transcripts is a principle component of the quality of transcriptional profiling data. Given the sensitivity of RNA to degradation, assessment of RNA quality and integrity is an indispensable prerequisite for all downstream applications in gene expression analysis. The ratio of the 3' to 5' message abundance provides a measure of the quality of the RNA hybridized to a given array. Due to their inability to provide uncertainty measures in the absence of replicate probe sets of control genes on Affymetrix GeneChips, probe set expression summaries used to assess sample quality by comparing the 3'/5' ratios to a predetermined threshold often lead to different conclusions regarding sample quality when several control genes are examined. The RNA digestion plots can be subjective and inconclusive in their interpretation of RNA quality. The mixed model proposed by Archer overcame those drawbacks and provided a framework in which the 3'/5' ratio can be estimated and a confidence interval and hypothesis test can be conducted. However, the pixel-level

intensity used in this approach is not a routine microarray data type and requires special data processing to obtain.

In this article, a linear regression model for assessing RNA quality for a hybridized array, which overcomes the drawbacks in the existing methods is proposed. The proposed model is less complex and easier to implement than the existing methods. In particular, we used the probe intensities from the `.cel` files that the Affymetrix software associates with each microarray. The effectiveness of the proposed method over the existing methods is illustrated by the application of the method to 19 human Affymetrix microarray data sets for which external verification of RNA quality is available. It has been shown in the literature that an unexpected aspect of the asymmetry of G versus C affinities exists, which goes against the zeroth order energetic consideration that G-C and C-G bonds would contribute equally to the binding (Naef et al., 2003). Our results also indicated that percent C content alone was as equally effective as percent GC content for adjusting the probe affinity, while percent G content alone was not as sufficient, thereby indicating that G and C did not contribute equally in probe affinity.

The degraded RNAs hybridized to a GeneChip yield lower expression levels than high quality RNAs in general, however, the variability of gene expression levels among samples does not appear to be linked to RNA quality. The RNA quality related bias can't be completely removed by the commonly used normalization procedures,

thereby having a detrimental impact on down-stream microarray analysis. Assessment of array quality is an essential step in the analysis of data from microarray experiments. Once detected, less reliable arrays are typically excluded from further analysis to avoid misleading results. Here, we present a better statistical tool than the existing methods for RNA quality assessment.

Chapter 3

A Semiparametric Test to Detect Associations between Quantitative Traits and Candidate Genes in Structured Populations

3.1 Abstract

Although population-based association mapping may be subject to the bias caused by population stratification, alternative methods that are robust to population stratification such as family-based linkage analysis have lower mapping resolution. Recently, various statistical methods robust to population stratification were proposed for association studies, using unrelated individuals to identify associations between candidate genes and traits of interest. The association between a candidate gene and a quantitative trait is often evaluated via a regression model with inferred population structure variables as covariates, where the residual distribution is customarily assumed to be from a symmetric and unimodal parametric family, such as a Gaussian, although this may be inappropriate for the analysis of many real-life data sets. In this paper, we proposed a new structured association test. Our method corrects for continuous population stratification by first deriving population structure and kinship matrices through a set of random genetic markers and then modeling the relationship between trait values, genotypic scores at a candidate marker, and genetic background variables through a semiparametric model, where the error distribution is modeled as a mixture of Polya trees centered around a normal family of distributions. We compared our model to the existing structured association tests in terms of model fit, type I error rate, power, precision, and accuracy by application to a real data set as well as simulated data sets.

3.2 Introduction

Population-based association mapping has become a powerful, general tool for identifying loci associated with the inheritance of complex traits. Because the allelic association due to linkage disequilibrium usually operates over shorter genetic distances, association mapping permits higher resolution mapping than does linkage analysis (Hästbacka et al., 1992). However, population-based association mapping may lead to both false positives and failure to detect true associations due to population stratification (Lander et al., 1994). Several statistical methods were recently proposed to utilize genomic markers to correct for potential population stratification in the analysis of qualitative traits (Devlin and Roeder, 1999; Devlin et al., 2001; Price et al., 2006; Pritchard et al., 2000) and in the analysis of quantitative traits (Bacanu et al., 2002; Hoggart et al., 2003; Redden et al., 2006; Zhang et al., 2003). Among them, genomic control (GC) (Devlin et al., 1999; Reich et al., 2001), structured association (SA) (Pritchard et al., 2000), and principal component analysis (PCA) (Price et al., 2006) are the three prevailing methods for dealing with stratification.

The GC method corrects for stratification by adjusting association statistics at each marker by a uniform overall inflation factor (Devlin et al., 1999; Reich and Goldstein, 2001). However, if the actual distribution has thicker tails than predicted, this could lead to a high type I error rate (Pritchard and Donnelly, 2001). The SA method

uses a Bayesian clustering method to assign individuals membership of subpopulations using random genetic marker data. The method is computationally expensive. Assignments of individuals to subpopulations are highly sensitive to the number of subpopulations, which is not well defined. In contrast, the PCA approach is computationally efficient. These methods have been proven useful in a variety of contexts, however, they have limitations. For many real life data having population structure along with diverse levels of familial relatedness within subpopulations, i.e. continuous population stratification, GC, SA, and PCA approaches may lead to either inadequate control for false positives or a loss in power owing to genetic correlation within subpopulations (Yu et al., 2006; Zhao et al., 2007). Yu et al. (2006) recently introduced a unified mixed-model approach in which the fixed population structure effect and random subject specific effects attributable to kinship K with the covariance matrix being $2K\sigma_G^2$, where σ_G^2 is the additive genetic variance, are included to adjust for continuous population stratification.

The association between a candidate gene and a quantitative trait is often evaluated via a linear regression model where inferred population structure variables are included as covariates. In such cases, the residual distribution is customarily assumed to be from a symmetric and unimodal parametric family, such as a Gaussian (Aranzana et al., 2005; Kang et al., 2008; Lettre et al., 2007; Tommasini et al., 2007; Yu et al., 2006; Zhao et al., 2007), although this may be inappropriate

for the analysis of many real-life data sets, where increased skewness, kurtosis, and multimodality may hold. For example, many plant traits are highly skewed due to selection (Kelker et al., 1986). Moreover, for a quantitative trait determined by the sequence variants of a few genes of large effects and normally distributed random error(s), the associated residuals from a single marker and the trait association test (the most common practice in association mapping) are indeed distributed as a mixture of normal distributions with non-identical means, and hence are not normally distributed. Furthermore, exploratory analysis such as residual plots for assessing the error distribution is often impractical in association mapping, in particular for genome-wide scans, because there are far too many markers (one cannot examine 100,000 or more plots). Suitable transformation of the response to achieve comfort with the assumptions under a standard parametric family for studies with large numbers of candidate loci may be difficult because the appropriate transformation may vary from one marker to another, and interpretation problems may arise. Enrichment of standard families usually fails to capture all the desirable features. Hence, association tests that avoid parametric specification for the residual distribution and are robust against potential population stratification are desirable; our goal in this article is to describe a statistical method that is intended to have such desired properties. First, we proposed a new computationally efficient method to infer population structure. DNA-based genetic dissimilarity has been used to measure the degree

of genetic difference between individuals in association studies (Wessel and Schork, 2006; Jakobsson et al., 2008). In our approach, the genetic distance matrix is estimated from a set of random markers, multidimensional scaling (MDS) is performed on this genetic distance matrix. MDS and PCA are both used for dimensionality reduction. When compared to PCA, MDS gives more readily interpretable solutions of lower dimension and does not depend on the assumption of a linear relationship between variables (Lacher, 1987). In our method, the estimated population structure is given by the MDS representations of the distance matrix. Like Yu's mixed model, in addition to population structure, our model also includes kinship K to adjust for continuous population stratification, and the two approaches in our model for uncovering population stratification are complementary; however, unlike Yu's mixed model, the two approaches in our model are orthogonal, so the effects of kinship on a phenotypic variation are estimated after taking the effects of population structure into account. Next, the relationship between trait values, genotypic scores at a candidate gene, and genetic background variables is modeled using a semiparametric regression approach, where the error distribution is modeled as a mixture of Polya trees centered around a family of normal distributions and thus may be viewed as a generalization of standard models in which important, data-driven features, such as skewness and multimodality, are allowed. The mixture of Polya trees prior provides

an intermediate choice between a strictly parametric model and a completely arbitrary model (Ferguson, 1974; Hanson, 2006; Lavine, 1992). To illustrate its utility and interpretation, we applied the proposed model to a previously published data set of association mapping in 95 *Arabidopsis thaliana* lines (Zhao et al., 2007) where the residuals were assumed to be normally distributed. We compared our proposed model to the SA method as well as the PCA method in terms of model fit, power, and type I error rate. We also conducted two different simulation studies to demonstrate the power, precision, and accuracy of the proposed method.

3.3 Methods

Let $y' = (y_1, \ldots, y_n)$ be the n response vector. Let $x_i' = (x_{i1}, \ldots, x_{ip})$ be a p-component column vector of the design matrix for the i^{th} subject, where $x_{i1} = 1$ is an intercept term, x_{i2} is the genotypic score (haplotype, genotype or gene expression level) for individual i at a candidate gene for a given trait, and $x_{i3}, \ldots, x_{ip}$ are the values of the genetic background variables for individual i for $i = 1, \ldots, n$. For a candidate gene with L categorical values, there are $(L-1)$ orthogonal dummy variables corresponding to the candidate gene, and in such cases, $x_{i2}' = (x_{i21}, \ldots, x_{i2(L-1)})$; $\beta_2' = (\beta_{21}, \ldots, \beta_{2(L-1)})$. Let $\beta' = (\beta_1, \beta_2', \ldots, \beta_p)$ be the regression coefficients corresponding to p predictors (including the intercept). A kinship coefficient is often defined as the probability of identity by descent of the markers compared (Ritland,

1996), but estimators based on genetic markers actually estimate a "relative kinship", that can be defined as ratios of differences of probabilities of identity by state (Hardy et al., 2002). Note that with this definition, negative relative kinship coefficients naturally occur between some individuals. This is interpreted as meaning subjects that these are less related than random individuals and the negative kinship coefficients are set to 0 (Hardy et al., 2002; Ritland, 1996; Yu et al., 2006; Zhao et al., 2007). Like Yu et al., we also assume K is symmetric and positive definite and let K_s denote the square root of $2K$.

3.3.1 Methods for inferring population structure

We first consider the SA method for inferring population structure. Let Q be the population structure for n subjects estimated using the software STRUCTURE (Pritchard et al., 2000). Each individual's genetic background is represented by a vector $Q_i' = (Q_{i1}, \ldots, Q_{iJ_a})$, where Q_{ij} is the portion of the individual's genome which originated in subpopulation j for $j = 1, \ldots, J_a$ (thus $\sum_{j=1}^{J_a} Q_{ij} = 1$ for $i = 1, \ldots, n$). For the SA method, $x_{ij} = Q_{i(j-2)}$ for $j = 3, \ldots, (J_a + 1)$.

Second, we consider the PCA method for population structure inference. Let $C = (C_1, \ldots, C_j, \ldots, C_{J_b})$ be the J_b largest principal components of a set of genetic marker data for n individuals, where $C_j' = (C_{1j}, \ldots, C_{nj})$. For this model, $x_{ij} = C_{i(j-2)}$ for $j = 3, \ldots, (J_b + 2)$. We used the Bayesian information criteria (BIC) to choose the number (J_b) of principal components by fitting linear regression models

with l components for $l = 1, \ldots, \leq n$ (the sample size) as the l continuous explanatory variables and a trait of interest as the dependent variable, and J_b is the value of l corresponding to the smallest BIC.

Next, we propose a new method for inferring population structure using data at a set of genetic markers. Suppose that there are N SNP markers which have been genotyped on subjects i and j (for $i = 1, \ldots, n$ and $j = 1, \ldots, n$), let N_{11} be the total number of alleles shared between individuals i and j, then the modified Rogers' genetic distance (GD_{MR}) (Rogers, 1972) between individuals i and j is calculated as:

$$d_{(i,j)} = [(2N - N_{11})/2N]^{0.5}.$$

Note here N may vary among different pairs of subjects if the data is incomplete. MDS (Wickelmaier, 2003) was then performed on this n by n genetic distance matrix D with its $(i,j)^{th}$ element being $d_{(i,j)}$ described above. Let $B = (B_1, \ldots, B_j, \ldots, B_{J_c})$ with $B'_j = (B_{1j}, \ldots, B_{nj})$ be the J_c dimensional MDS representations of genetic distance matrix D. Note here, J_c is determined using BIC by fitting linear regression models with $(B_1, \ldots, B_l)$ for $l = 1, \ldots, \leq r$ (the rank of D) as the l continuous explanatory variables and a trait of interest as the dependent variable, and J_c is set to the value of l corresponding to the smallest BIC. We denoted this method as MDS, and here $x_{ij} = B_{i(j-2)}$ for $j = 3, \ldots, (J_c + 2)$.

3.3.2 Parametric association tests

We first consider Yu's parametric mixed model for association tests in structured population, which may be formulated as a regression model:

$$y_i = \beta_1 + x'_{i2}\beta_2 + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \eta_i + \epsilon_i \tag{3.1}$$

for $i = 1, \ldots, n$, where ϵ_i are assumed to be $i.i.d$ normal random variables with mean 0 and variance σ^2 and independent of the value of $x_{i2}, \ldots, x_{ip}$, and $\eta_i = \sum_{j=1}^{n} K_s(i,j)\gamma_j$. Throughout this paper, we assume $\gamma' = (\gamma_1, \ldots, \gamma_n)$, β, σ are independent in their prior distributions with $\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$. For priors, we assume $\gamma_i \sim \mathrm{N}(0, \tau^{-1})$ for $i = 1, \ldots, n$ and $\tau \sim$ Gamma (a, b) with $a = b = 0.01$, i.e. $\pi(\tau) = \frac{b^a}{\Gamma(b)}\tau^{a-1}\mathrm{e}^{-b\tau}$. This is equivalent to setting $\eta' = (\eta_1, \ldots, \eta_n)$ and $\eta \sim \mathrm{N}(0, 2K\tau^{-1})$. The joint posterior is given by:

$$\begin{aligned}
\pi(\beta, \sigma^2, \gamma, \tau | y) &\propto \prod_i p(y_i|\beta, \sigma^2, \gamma) \frac{1}{\sigma^2} \prod_i p(\gamma_i|\tau)\tau^{a-1}\mathrm{e}^{-b\tau} \\
&\propto \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}\Big(y_i - \beta_1 - x'_{i2}\beta_2 - \sum_{j=3}^{p} \beta_j x_{ij} - \eta_i\Big)^2 \right\} \frac{1}{\sigma^2} \\
&\times \exp^{-\frac{\tau}{2}\sum_{i=1}^{n}\gamma_i^2} \tau^{n/2+a-1}\mathrm{e}^{-b\tau}.
\end{aligned}$$

3.3.3 Semiparametric association tests

Now, we consider semiparametric structured association tests which are different fundamentally from the above parametric models in two aspects. First, the error distribution is modeled as a mixture of Polya trees (MPT) constrained to have median 0 and centered around a normal family of distributions. Second, we restrict kinship effects to the space orthogonal to population structure, thus the effects of kinship only explains the variations in response that are not captured by the population structure. Let P be the population structure inferred by the SA, PCA, or MDS methods. Let $P^c = \mathrm{I} - P(P'P)^{-1}P'$ (I here is an identity matrix), denote $P_K^c = P^c K_s$. The semiparametric association test may be formulated as a regression model:

$$y_i = \beta_1 + x'_{i2}\beta_2 + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \eta_i + \epsilon_i \tag{3.2}$$

for $i = 1, \ldots, n$, where ϵ_i follows a MPT and $\eta_i = \sum_{j=1}^n P_K^c(i,j)\gamma_j$. Specifically, let $G_{\sigma^2} = N(0, \sigma^2)$, and $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$. Consider

$$\epsilon_1, \ldots, \epsilon_n \underset{\sim}{iid} G, \quad G|\sigma^2 \sim PT_M(c, \rho, G_{\sigma^2}),$$

$$\sigma^2 \sim \pi(\sigma^2), \text{ i.e. } G \sim \int PT_M(c, \rho, G_{\sigma^2}) P(d\sigma^2).$$

We briefly describe the MPT prior but leave technical details to Hanson (2006) and Lavine (1992). A Polya tree prior is constructed from a set of partitions $\Pi_M^\theta = \{B_\varepsilon^\theta : \varepsilon \in \bigcup_{l=1}^M \{0,1\}^l\}$ and a family $\mathcal{A}$ of positive real numbers. Here, the partition points are quantiles of the centering family: if j is the base-10 representation of the binary number $\varepsilon = \varepsilon_1, \ldots, \varepsilon_k$ at level k, then $B^\theta_{\varepsilon_1,\ldots,\varepsilon_k}$ is defined to be the interval $(G_\theta^{-1}(j/2^k), G_\theta^{-1}((j+1)/2^k)]$, except the rightmost set is $B^\theta_{11,\ldots,1} = G_\theta^{-1}((2^k-1)/2^k), \infty)$. For example, with $k = 3$, and $\varepsilon = 000$, then $j = 0$ and $B^\theta_{000} = (0, G_\theta^{-1}(1/8)]$, and with $\varepsilon = 010$, then $j = 2$ and $B^\theta_{010} = (G_\theta^{-1}(2/8), G_\theta^{-1}(3/8)]$, etc. Note then that at each level k for $k = 1, \ldots, (M-1)$, the class $\{B_\varepsilon^\theta : \varepsilon \in \{0,1\}^k\}$ forms a partition of the positive reals and furthermore $B^\theta_{\varepsilon_1,\ldots,\varepsilon_k} = B^\theta_{\varepsilon_1,\ldots,\varepsilon_k 0} \cup B^\theta_{\varepsilon_1,\ldots,\varepsilon_k 1}$ for any binary $\varepsilon_1, \ldots, \varepsilon_k$. We take the family $\mathcal{A} = \{\alpha_\varepsilon : \varepsilon \in \bigcup_{j=1}^M \{0,1\}^j\}$ to be defined by $\alpha_{\varepsilon_1,\ldots,\varepsilon_k} = c\rho(k)$ for some $c > 0$ (Hanson 2006). The parameter c is the amount of weight attached to the underlying normal distribution. As c tends to zero, the posterior $G|data$ is almost entirely data driven. As c tends to infinity, we obtain a fully parametric analysis. Here $\rho(.)$ is any increasing positive function of k, and $\rho(k) = k^2$ is used in the vast majority of applications. Given $\prod_M^\theta$ and $\mathcal{A}$, the Polya tree prior is defined up to level M by the random vectors $\mathbf{Z}_M = \{(Z_{\varepsilon_0}, Z_{\varepsilon_1}) : \varepsilon \in \bigcup_{j=0}^{M-1} \{0,1\}^j\}$ through the product of

conditional probabilities

$$G(B^{\theta}_{\varepsilon_1,\ldots,\varepsilon_k}|\mathbf{Z}_M,\theta) = \prod_{j=1}^{k} Z_{\varepsilon_1,\ldots,\epsilon_j}.$$

For $k = 1,\ldots,M$, the vectors $(Z_{\varepsilon_0}, Z_{\varepsilon_1})$ are independent Dirichlet:

$$(Z_{\varepsilon_0}, Z_{\varepsilon_1}) \sim \text{Dirichlet}(\alpha_{\varepsilon_0}, \alpha_{\varepsilon_1}),\ \varepsilon \in \bigcup_{j=0}^{M-1} \{0,1\}^j.$$

Define the vector of probabilities

$$\pi(\mathbf{Z}_M) = (p_1, p_2, \ldots, p_{2^M})'$$

through

$$p_{(k+1)} = G(B^{\theta}_{\varepsilon_1,\ldots,\varepsilon_M}|\mathbf{Z}_M,\theta) = \prod_{j=1}^{M} Z_{\varepsilon_1,\ldots,\varepsilon_j},$$

where $\varepsilon_1,\ldots,\varepsilon_M$ is the base-10 binary representation of the integer k, and here $k = 0,\ldots,(2^M-1)$.

Let $\mu_i = \beta_1 + x'_{i2}\beta_2 + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \eta_i$, and N_i denote the integer part of $2^M\Phi(\frac{y_i-\mu_i}{\sigma}) + 1$ (here Φ is the cdf of N(0,1)). After simplification, the pdf of y_i for $i = 1,\ldots,n$ is given by:

$$f(y_i|\mathbf{Z}_M,\sigma^2,\beta,\gamma) = 2^M p_{N_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{-\frac{1}{2\sigma^2}(y_i-\mu_i)^2\Big\},$$

the likelihood involving n independent subjects is then given by:

$$\mathcal{L}(y;\ \mathbf{Z}_M, \sigma^2, \beta, \gamma) = \prod_{i=1}^{n} 2^M p_{N_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\Big\} \tag{3.3}$$

The joint posterior is then given by:

$$\begin{aligned}
\pi(\mathbf{Z}_M, \sigma^2, \beta, \gamma, \tau | y) &\propto \mathcal{L}(y;\ \mathbf{Z}_M, \sigma^2, \beta, \gamma)\pi(\beta, \sigma^2, \gamma, \tau, \mathbf{Z}_M) \\
&= \prod_{i=1}^{n} 2^M p_{N_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\Big\} \\
&\times \frac{1}{\sigma^2} \exp^{-\frac{\tau}{2}\sum_{i=1}^{n}\gamma_i^2} \tau^{n/2+a-1} \mathrm{e}^{-b\tau} \\
&\times \prod_{j=1}^{M} \prod_{\varepsilon \in \bigcup_{k=0}^{(j-1)} \{0,1\}^k} \frac{\Gamma(2cj^2)}{\Gamma(cj^2)\Gamma(cj^2)} (Z_{\varepsilon_0})^{cj^2} (Z_{\varepsilon_1})^{cj^2}.
\end{aligned}$$

We tried various values for c and M for the MPT prior and we also put a Gamma prior on c. The model appears to be robust to the values of c and M, hence our discussion is based on the case with $c = 1$, $M = 4$, and $\rho(k) = ck^2$, i.e. $\alpha_{\varepsilon_1,\ldots,\varepsilon_k} = c\rho(k) = k^2$. All algorithms use a Metropolis-Hastings (M-H) step for updating the components $(Z_{\varepsilon_0}, Z_{\varepsilon_1})$ one at a time by sampling candidates $(Z^*_{\varepsilon_0}, Z^*_{\varepsilon_1})$ from a Dirichlet $(mZ_{\varepsilon_0}, mZ_{\varepsilon_1})$ distribution, where $m = 20$. Let $\mathcal{L}^* = \mathcal{L}(y;\ \mathbf{Z}^*_M, \sigma^2, \beta, \gamma)$ and

$\mathcal{L} = \mathcal{L}(y;\ \mathbf{Z}_M, \sigma^2, \beta, \gamma)$, this candidate is accepted as $(Z^*_{\varepsilon_0}, Z^*_{\varepsilon_1})$ with probability

$$\rho = \min\left\{1, \frac{\Gamma(mZ_{\varepsilon_0})\Gamma(mZ_{\varepsilon_1})(Z_{\varepsilon_0})^{mZ^*_{\varepsilon_0}-cj^2}(Z_{\varepsilon_1})^{mZ^*_{\varepsilon_1}-cj^2}\mathcal{L}^*}{\Gamma(mZ^*_{\varepsilon_0})\Gamma(mZ^*_{\varepsilon_1})(Z^*_{\varepsilon_0})^{mZ_{\varepsilon_0}-cj^2}(Z^*_{\varepsilon_1})^{mZ_{\varepsilon_1}-cj^2}\mathcal{L}}\right\}.$$

For the M-H step for updating the parameter vector (β, σ^2), we used a multivariate normal random-walk proposal constrained to $(\sigma^2)^* > 0$. The proposal covariance matrix is a scaled version of the estimated covariance matrix of the maximum likelihood estimate from fitting the ordinary linear regression:

$$y_i = \beta_1 + x'_{i2}\beta_2 + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \epsilon_i.$$

The variance scaling factor was set to be a value such that a Markov chain accepted $30\% \sim 40\%$ of the proposed moves. The candidate $(\beta^*, (\sigma^2)^*)$ is accepted with probability

$$\rho = \min\left(1, \frac{\mathcal{L}(y;\ \mathbf{Z}_M, (\sigma^2)^*, \beta^*, \gamma)\sigma^2}{\mathcal{L}(y;\ \mathbf{Z}_M, \sigma^2, \beta, \gamma)(\sigma^2)^*}\right).$$

For the M-H step for updating parameter vector γ, sample $\gamma_i^* \sim \mathrm{N}(\gamma_i, s\tau^{-1})$ for $i = 1, \ldots, n$ (here s is a scaling factor so that a Markov chain accepted $30\% \sim 40\%$ of the proposed moves), accept γ^* with probability

$$\rho = \min\left(1, \frac{\mathcal{L}(y;\ \mathbf{Z}_M, \sigma^2, \beta, \gamma^*)\exp^{-\frac{\tau}{2}\sum_{i=1}^n (\gamma_i^2)^*}}{\mathcal{L}(y;\ \mathbf{Z}_M, \sigma^2, \beta, \gamma)\exp^{-\frac{\tau}{2}\sum_{i=1}^n \gamma_i^2}}\right).$$

We used the full conditional to update of τ by sampling $\tau^* \sim \text{Gamma}(a + 0.5n, b + 0.5\sum_{i=1}^{n} \gamma_i^2)$. For each of the 3 population structure inferring methods, we consider both parametric tests denoted as SA, PCA, and MDS methods, and semiparametric tests denoted as SPSA, SPPCA, and SPMDS methods. We fit both parametric and semiparametric tests using a Bayesian approach. For each of the 6 tests, two Markov chains were run in parallel to monitor the convergence of the Metropolis algorithm with 60,000 iterations per chain for burn-in and 5,000 Monte Carlo samples from 50,000 iterations per chain for inference. C++ software was used for the analysis.

3.3.4 Model Comparison and Diagnostics

Bayes factors (Kass and Raftery, 1995; Han and Carlin, 2001) are difficult to obtain in practice. Instead, for each model we compute the pseudo marginal likelihood (LPML) (Geisser and Eddy, 1979). The LPML is a cross-validated "leave-one-out" measure of a model's ability to predict the data. It is valid for small and large samples and does not suffer from a heuristic justification based on large sample normality. Let $p_1(\cdot)$ and $p_2(\cdot)$ denote probability densities corresponding to models 1 and 2, respectively. The conditional predictive ordinate (CPO) for observation j under model i is given by $\text{CPO}_{ij} = p_i(D_j|D_{-j})$, where $D_{-j} = \{(x_{l1}, \ldots, x_{lp}, y_l)\}_{l \neq j}$. The ratio $\text{CPO}_{1j}/\text{CPO}_{2j}$ measures how well model 1 supports the observation D_j relative to model 2, based on the remaining data D_{-j}. The product of the CPO ratios gives an overall aggregate summary of how well supported the data are by

model 1 relative to model 2 and is called the pseudo Bayes factor:

$$\mathrm{B}_{12} = \prod_{j=1}^{n} \frac{\mathrm{CPO}_{1j}}{\mathrm{CPO}_{2j}}.$$

The log of the product of the n CPO statistics under a given model is the LPML statistic for that model, $\mathrm{LPML}_i = \log \prod_{j=1}^{n} \mathrm{CPO}_{ij}$, and therefore

$$\mathrm{B}_{12} = \exp(\mathrm{LPML}_1 - \mathrm{LPML}_2).$$

3.4 Application to the *Arabidopsis thaliana* data set

The 95 *Arabidopsis thaliana* lines used in this study have been described previously (Zhao et al., 2007), where the errors were assumed to normally distributed. For this specific data set, subjects are the 95 Arabidopsis lines. The trait we are interested in for association mapping is the plant flowering time measured in days from germination to the first opening of flowers with 5 week vernalization and under short-day conditions (8 hour(h) light/16 h dark) at the University of Southern California (SDV). Using the software STRUCTURE (Pritchard et al., 2000), Nordborg et al. (2005) estimated Q with $J_a = 8$ for the 95 Arabidopsis lines. For inference for B for the MDS method and C for the PCA method, we used 5,000 high quality SNP markers, and both J_b and J_c were determined to be 8 as well. The residuals for

SNPs and SDV association tests by the 3 parametric models demonstrated deviations from normality. As an example, random SNP S106.215 and SDV association tests by the parametric models have excessive skewness (1.27 for the PCA model, 1.27 for the MDS model, 1.95 for the SA model) as well as excessive kurtosis (3.85 for the PCA model, 3.86 for the MDS model, 5.12 for the SA model). Therefore, the semiparametric MPT model which relaxes the normality assumption on the error distribution may be more appropriate than the 3 parametric models for this data set.

To compare the performance of the 6 models, we tested the association between SDV and FRIGIDA (*FRI*), a known central regulator of flowering time for unvernalized Arabidopsis plants. The effects of FRI is eliminated by vernalization (Lee and Amasino, 1995), a procedure that accelerates flowering by a long period of cold temperatures, generally below 10°C but above 0°C. There are three genotypes at FRI for the 95 lines, where 1 represents the wild type lines, 2 represents the mutated lines with the deletion of the *Ler* allele, and 3 represents the mutated lines with the deletion of the *Col* allele (Nordborg et al., 2005). In addition, we did the association tests between SDV and random SNP markers, S1.140 and S1.415, respectively.

The LPML statistics from the FRI and SDV association test for the 6 models are summarized in table 3.1. The LPML values vary a great deal among the 6 models. For each of the three population inference methods, the semiparametric test has

Table 3.1: Comparisons of the SA, SPSA, PCA, SPPCA, MDS, and SPMDS models in LPML and 95% credible sets (95% CI) genetic effects of the *Ler* and *Col* deletion alleles of FRI from the association test between FRI and SDV

Model	FRI			S1.140	S1.415
	95%CI$_{\hat{\delta_1}}$	95%CI$_{\hat{\delta_2}}$	LPML	LPML	LPML
SA	(1.7, 51.8)	(-3.6, 34.6)	-408.4	-408.3	-400.2
SPSA	(-15.3, 22.8)	(-4.5, 16.4)	-367.1	-367.2	-365.6
PCA	(-19.1, 24.7)	(-18.1, 17.5)	-401.9	-401.2	-394.4
SPPCA	(-7.9, 22.5)	(-9.2, 13.0)	-370.3	-369.4	-367.9
MDS	(-19.7, 24.6)	(-18.8, 17.9)	-402.1	-401.1	-392.9
SPMDS	(-11.5, 14.7)	(-10.6, 10.6)	-369.5	-368.7	-367.5

LPML stands for log pseudo marginal Likelihood. S1.415 and S1.140 are SNPs; FRI is a known major flowering gene for non-vernalized *Arabidopsis* plants. $\delta_1 =$ the mean difference of flowering times between the wild type group and the *Col* deletion group, and $\delta_2 =$ the mean difference of flowering times between the wild group and the *Ler* deletion group.

larger LPML than the corresponding parametric test. For example, the LPML for proposed SPMDS model is -369.5 compared to -402.1 for the MDS model. The pseudo Bayes factor for comparing the SPMDS model to the MDS model is approximately $\exp((-369.5) - (-402.1)) > 1000$. The SPSA model versus the parametric SA or SPPCA model versus the PCA model also yields a pseudo Bayes factor > 1000. This observation suggests that the proposed semiparametric MPT models which relaxe the normality assumption on the error distribution are better fit than the parametric models for the current data set. The same conclusion is held from the SNPs and SDV association tests (table 3.1).

Also listed in table 3.1 are the effect (δ_1) of the *Col* deletion haplotype and the

effect (δ_2) of the *Ler* deletion haplotype from the FRI and SDV association test by the 6 models. Here δ_1 and δ_2 are estimated by the mean difference in flowering time between the wild type group and the *Col* deletion group, the wild type group and the *Ler* deletion group, respectively. Noted here, SDV is a type of flowering time for vernalized plants, hence SDV and FRI, a gene for non-vernalized plants, are expected to be unrelated. The 95% credible sets for δ_1 covered 0 for all models except the SA model. The 95% credible sets for δ_2 covered 0 for all 6 models, as expected. Figure 3.1 shows the residual predictive density estimates from the SNP S1.415 and SDV association test by the 6 models. The three semiparametric models provide a better predictive density of the residuals than the three corresponding parametric models.

To compare the false positive rates among the 6 models, we randomly selected 500 random SNP haplotype makers with two haplotypes, 0 and 1, for each SNP. These SNPs are distributed over the whole *Arabidopsis thaliana* genome and have minor allele frequencies greater than 5%. We applied the 6 models to each of the 500 random SNPs for testing the association with SDV. We estimated the false positive rate which is defined as $500p$ with p being the proportion of 500 random SNPs whose $(1-\alpha)$ credible sets of the SNP regression coefficients excluded 0 with $\alpha = 0.05$ and 0.01, respectively. As shown in table 3.2, the type I error rate at $\alpha = 0.05$ is the largest for the SA model (10.5%), followed by the PCA model (8.0%) and the

MDS model (7.8%), and the SPSA model has the smallest type I error rate (5.2%), followed by SPMDS model of 5.7%, and SPPCA model of 5.8%. Compared to the parametric models, the semiparametic models show an average of a 36% reduction in the type I error rate. At $\alpha = 0.01$, the SPMDS and SPPCA have a type I error rate of 1.2%, SPSA model has a type I error rate of 1.5%, again close to the expected value of 1%. In contrast, the 3 parametric models which assume normally distributed error have a type I error rate substantially higher than the expected value of 1%.

3.5 Simulation study I

Our goal was to simulate a data set with the same covariate structure as the data from the previous section given a true complex genetic correlation structure among individuals. Our simulations are similar to those carried out by Yu et al. and Zhao et al. in their association studies. Underlying the phenotype simulation, we simulated 500 Arabidopsis flowering genetic loci as follows. For each of the randomly selected 500 SNPs in section 2.5, a fixed additive genetic effect ranging from $\kappa = 0.1s$ to $1.05s$ was added, here $s{=}\sqrt{\hat{\sigma}^2}$, where $\hat{\sigma}^2$ is the usual unbiased estimate of the variance, assuming lines are independent. If we let p be the sample minor allele frequency of a SNP, and $n = 95$ be the sample size, then the percentage (ζ) of the total phenotypic variation explained by this fixed genetic effect can be approximated by $\zeta = p(1-p)\kappa^2/(p(1-p)\kappa^2 + 1 - 1/n)$ (Long et al., 1999; Yu et al., 2006). To

each of the 500 random SNPs, we have 6 simulated data sets; each with the added genetic effect expressed as one of the values $1\%, 2.5\%, 5\%, 7.5\%, 10\%$, and 15%. We denoted those simulated flowering time genetic loci as quantitative trait nucleotides (QTN). For each of 3,000 simulated data sets i.e. 6 genetic effects by 500 SNPs, we applied the 6 models with the same Bayesian implementation as in the analysis from section 3.2.1-3.2.3. For each model and each genetic effect, we calculated the power as $500p$ with p being the proportion of the 500 QTNs whose $(1-\alpha)$ credible set for regression coefficient excluded 0 with $\alpha = 0.05$ and $\alpha = 0.01$, respectively. As shown in table 3.2, while the power steadily increased as the simulated genetic effect increased from 1% to 15% for all models, the three semiparametric models showed consistently higher power than their corresponding parametric models for all the cases considered. At $\alpha = 0.05$, for example, compared to the SA model, the SPSA model showed a 0.1% to 87.5% increase in power for $\zeta = 1\%$ to 15%, while compared to the MDS, the SPMDS showed a 3.5% to 85.4% increase in power. We also estimated the standard deviation (sd) of the regression coefficients corresponding to QTNs. The SA model has the largest mean sd (10.11) averaged across the 500 QTNs and 6 gene effects, followed by the PCA of 9.62, MDS model of 9.57, the three semiparmetric models have smaller sd than any of the parametric models (table 3.2).

3.6 Simulation study II

Our second simulation studies are similar to those carried out by Zhang et al. (2003) in their association mapping studies. One limitation of the simulations based on coalescent models is that these models may not represent population evolutionary histories accurately (Zhang et al. 2003). Therefore, our simulations are based on empirical population genetic data. We assume that sampled individuals were genotyped at a series of unlinked SNP loci with two alleles A and a. There are three genotypes at a SNP marker: aa (or -1), Aa (or 0), and AA (or 1). We also assume that the population under study is a continuous admixture of two ancestral human populations including Biaka and Danes. The allele frequencies of 500 unlinked SNPs with minor allele frequency greater than 5% across the two ancestral populations were extracted from a population genetics database ALFRED (Rajeevan et al., 2003) and we repeated these 500 SNPs 2 times as if we had 1000 SNPs. First, we generated 1000 independent null SNPs with $n = 200$ individuals, those 1000 null SNPs will be used to estimate population structure. Let p_{i1}, p_{i2} represent the probabilities that allele A of the i^{th} individual originated from Biaka, and Danes, respectively, and $p_{i1} + p_{i2}$=1 for $i = 1, \ldots, n$. We simulate 180 out of 200 individuals as follows. We assume $p_{il} \sim$ Uniform$(0, 1)$ independently for each individual i and each SNP l. Therefore, the allele frequency of allele A at marker l for individual i can be written

as

$$p_{il} = q_{l1}p_{i1} + q_{l2}(1 - p_{i1}), \tag{3.4}$$

where q_{l1}, q_{l2} are the allele frequencies of allele A at marker l in Biaka and Danes, respectively. Individual i was assigned genotype -1, 0 or 1 at marker l with probabilities $(1-p_{il})^2$, $2p_{il}(1-p_{il})$, p_{il}^2, respectively for $i = 1, \ldots, 180$ and $l = 1, \ldots, 1000$. For the remaining 20 individuals, we assume they are from two genetically related groups with genetic correlation between individuals being $\rho = 0.8$ for the first group and 0.6 for the second group, and genetically independent between the two groups. For each of the two groups, we assume $p \sim \text{Uniform}(0, 1)$, then $p_i = p + \epsilon_i$ for $i = 1, \ldots, 10$, where ϵ_i is a small quantity generated from $\text{N}(0, 0.001)$ constrained $p_i > 0$, so that the individuals within a group have similar ancestry background. Genotypes -1, 0 or 1 at marker l for 10 individuals are generated according to a multivariate binomial distribution with marginal probabilities being $(1 - p_{il})^2$, $2p_{il}(1 - p_{il})$, p_{il}^2 (estimated by equation 4) and the correlation between any pair of individuals being ρ (Leisch et al. 1998). Next, we simulated 500 additive genetic loci using the simulated null SNP genotype data. We randomly selected 500 out of 1000 SNPs, let x_{il} be the genotype score of individual i at marker l, let $\mu_{01} = 10$, and $\mu_{i0} = \mu_{01}p_{i1}$, the quantitative trait values are generated according to the following model:

$$y_{il} = \mu_{i0} + \mu_1 x_{il} + \epsilon_{il} \tag{3.5}$$

for $i = 1, \ldots, n$ and $l = 1, \ldots, 500$, where ϵ_{il} is a random variable distributed as a skew-normal with mean 0.073 and variance 0.873 i.e. $f(\epsilon_{il}) = \frac{2}{\nu}\phi\Big(\frac{\epsilon_{il}-\theta}{\nu}\Big)\Phi\Big(\lambda\frac{\epsilon_{il}-\theta}{\nu}\Big)$, where $\theta = -1.1$, $\nu = 1.5$, $\lambda = 5$, $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal pdf and cdf, respectively. Note that the error distribution is a unimodal skewed distribution, its median is equal to 0, up to two decimal points. We set $\mu_1 = 0$ and 0.4 for comparing type I error rates and powers, respectively.

We applied the 6 models to each of the 500 data sets at $\mu_1 = 0$ and 500 data sets at $\mu_1 = 0.4$ with the same Bayesian implementation as in the analysis from section 2.1-2.3. As shown in table 3.3, the false positive rates are similar between the three parametric models and the three semiparametric models, which are not significantly different from the expected value of 0.05 at $\alpha = 0.05$ ($95\%\text{CI}_{p=0.05}$ is (0.0309, 0.0691)) or 0.01 at $\alpha = 0.01$ ($95\%\text{CI}_{p=0.01}$ is (0.0013, 0.0187)). The powers for the semiparimetric models are higher than those for the parametric models. On the average, the standard deviation for the candidate gene regression coefficients are smaller for the semiparametric models than for the parametric model. Furthermore, the semiparametric models provide a smaller bias for the candidate gene regression coefficients than the parametric models on the average.

Table 3.2: Comparisons of SA, SPSA, PCA, SPPCA, MDS, and SPMDS models in power to detect the association between simulated QTN and SDV and the mean standard deviation of the regression coefficients for QTNs ($\hat{\beta}_2$s) from simulation study I

	the value of ζ						
Model	0%	1%	2.5%	5%	7.5	10%	15%
Power at 95% credible set							
SA	10.5	23.0	43.6	64.8	77.0	84.4	96.5
SPSA	5.2	44.6	78.8	93.2	96.0	96.4	96.6
PCA	8.0	22.2	42.8	64.4	80.6	88.7	95.5
SPPCA	5.8	37.2	73.4	88.8	94.0	96.8	98.8
MDS	7.8	22.8	42.8	63.8	81.0	88.8	95.6
SPMDS	5.7	42.2	73.0	90.6	94.8	96.8	98.2
Power at 99% credible set							
SA	4.0	11.0	21.8	43.8	60.6	73.0	87.4
SPSA	1.5	26.4	60.8	86.6	93.8	95.2	94.2
PCA	3.2	10.4	20.8	45.4	62.0	76.6	92.2
SPPCA	1.2	20.0	54.2	80.6	90.6	94.2	97.6
MDS	3.2	9.8	21.4	46.2	61.4	77.0	91.8
SPMDS	1.2	23.2	55.8	83.4	91.0	94.2	97.4
Average standard deviation of $\hat{\beta}_2$s							
SA	10.11	10.12	10.11	10.11	10.11	10.12	10.11
SPSA	6.73	6.63	6.54	6.73	6.73	6.85	6.86
PCA	9.60	9.64	9.63	9.62	9.61	9.61	9.62
SPPCA	7.14	7.04	7.11	7.11	7.03	7.08	7.09
MDS	9.58	9.56	9.57	9.57	9.55	9.57	9.59
SPMDS	7.09	6.89	6.96	6.95	6.97	6.89	7.01

Note for $\zeta = 0$, the numbers given in the table are type I error rates. For each of the 6 models, 500 simulated QTNs were used in estimating powers, type I error rate, and the mean sd of $\hat{\beta}_2$ for each of 6 simulated genetic effects.

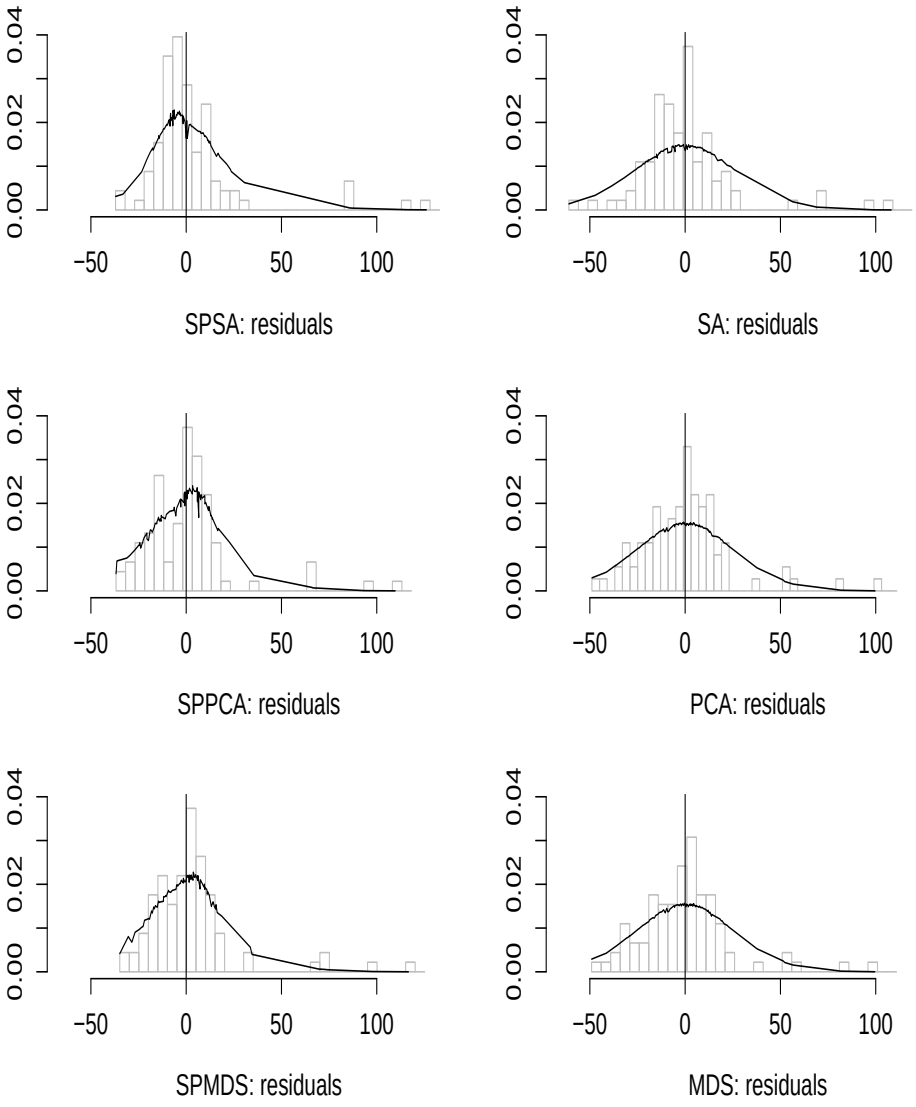

Figure 3.1: The histograms and predictive densities of the residuals from the SNP S1.415 and SDV association test by the SA, SPSA, PCA, SPPCA, MDS, and SPMDS models. The Y-axis is the residual predictive density.

Table 3.3: Comparisons of SA, SPSA, PCA, SPPCA, MDS, and SPMDS models in type I error rate, power, average of biases of candidate gene regression coefficient $\hat{\beta}_2$, and the mean standard deviation of $\hat{\beta}_2$s from simulation study II

Model	Type I error rate		Power		$SD_{\hat{\beta}_2}$	$Bias_{\hat{\beta}_2}$
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$		
SA	5.5	1.2	92.4	80.8	0.113	0.0947
SPSA	5.6	1.3	98.2	94.8	0.097	0.0752
PCA	5.3	1.3	92.6	81.8	0.112	0.0945
SPPCA	5.4	1.5	98.3	96.0	0.098	0.0771
MDS	5.6	1.4	93.0	80.6	0.112	0.0950
SPMDS	5.8	1.5	98.9	97.6	0.096	0.0745

For each of the 6 models, 500 simulated QTNs were used in estimating powers and type I error rate, while 1000 simulated QTNs were used in estimating the mean sd and bias of β_2.

3.7 Conclusion

The current available quantitative trait association mapping methods often assume that the errors are normally distributed. The normality assumption of the underlying error distributions greatly simplifies the form of the likelihood. However, it may be unrealistic. The normality assumption, if violated, may lead to false association detection and a reduction in power (Morton, 1984). The most commonly adopted method to achieve normality involves log-transforming the data. Although this method may give reasonable empirical results, it should be avoided if a more suitable theoretical model can be found (Azzalini and Capitanio, 1999). Furthermore, we applied the MDS model to the log-transformed data set, and the type I error rate from the 500 random SNPs and SDV association tests was much higher (28%

or higher) than that for the SA, MDS, PCA, SPSA, SPPCA SPMDS models using the untransformed data as we presented in this paper. Moreover, the semiparimetric model using a mixture of Polya trees to model the residual error is more flexible and it can be applied to a more broad range of data sets than the log-transformation method or other types of data transformation methods.

In this article, we have developed a semiparametric test to detect associations between candidate markers and quantitative traits using population-based data. Our model is advantageous in terms of the following aspects (1) the error distribution is flexibly modeled as a mixture of Polya trees centered around a family of normal distributions, hence important, data-driven features, such as skewness and multimodality, are allowed (2) both population structure and kinship are included in the model to adjust for continuous population stratification. In comparison to Yu's mixed model, the two aspects in our model for uncovering population stratification are orthogonal, so the effect of kinship on the phenotypic variation is estimated after taking the effects of population structure into account. Moreover, orthogonalizing the population structure and kinship reduces multicollinearity problems. Multicollinearity may lead to a large standard error for the regression coefficients, hence it can reduce the power to detect a true association between a candidate gene and a trait of interest. Compared to the parametric SA, PCA, and MDS methods, the semiparametric methods demonstrate several advantages. An application to the real data set of 95 *Arabidopsis*

lines for the flowering time demonstrated that the semiparametric models fit better the data than the 3 parametric models by comparing pseudo Bayes factor values. Moreover, the semiparametric models achieve a lower rate of false positive associations than the the three parametric models; the semiparametric models also have a better precision than the parametric models. Our simulation results show that the semiparametric models demonstrate higher power and better accuracy compared to the 3 parametric models. Therefore, for the current data set, the proposed semiparametric models are preferred, offering sufficient improvement in goodness of fit, higher power, and better precision in detecting gene effects to offset the increased computational complexity.

Our approach uses both inferred population structure and kinship to adjust for continuous population structure. However, for samples under Hardy-Weinberg equilibrium (HWE) within each subpopulation, the population structure alone is sufficient for adjusting for population structure. If HWE does not hold within each subpopulation; there will be multilevel-genetic correlation among the sampled individuals. In this case, the use of population structure alone will have less power, i.e. both K and population structure will be needed. The limitation of our algorithm is the computational expense incurred due to the use of MCMC in the inference for association stage. For example, for a sample of 100 individuals, our algorithm takes several minutes on a 2.13Ghz Pentium 4 with 1.99GB of RAM PC for one candidate

gene compared to several seconds for the other parametric methods (since MCMC is unnecessary for those methods). Our algorithm is more suited for candidate gene mapping, however it is still possible for genome-wide scans where millions of genetic markers are tested for association with a phenotype if parallel computing is used. Finally, incorporating the uncertainty in inferring population structure in estimating association at a candidate locus might be theoretically superior (Zhang et al., 2006), however, the approach can be computationally enormous. It is almost an impossible task for association tests involving hundreds of candidate genes if a thousand random genetic markers are included in the model for inferring population structure.

Chapter 4

Association Tests for a Censored Quantitative Trait and Candidate Genes in Structured Populations with Multilevel Genetic Relatedness

4.1 Abstract

Several statistical methods for detecting associations between quantitative traits and candidate genes in structured populations have been developed for fully observed phenotypes. However, many experiments are concerned with failure-time phenotypes, which are usually subject to censoring. In this paper, we propose statistical methods for detecting associations between a censored quantitative trait and candidate genes in structured populations with complex multiple levels of genetic relatednees among sampled individuals. The proposed methods correct for continuous population stratification using population structure variables as covariates and the frailty terms attributable to kinship. The relationship between the time-at-onset data and genotypic scores at a candidate marker is modeled via a parametric Weibull frailty accelerated failure time (AFT) model as well as a semiparametric frailty AFT model, where the baseline survival function is flexibly modeled as a mixture of Polya trees centered around a family of Weibull distributions. For both parmetric and semiparametric models, the frailties are modeled via an intrinsic Gaussian conditional autoregressive prior distribution with the kinship matrix being the adjacency matrix connecting subjects. Simulation studies and applications to the *Arabidopsis thaliana* line flowering time data sets demonstrate the advantage of the new proposals over existing approaches.

4.2 Introduction

Time-to-event data as quantitative traits (e.g. age-at-onset of cancer) have been used to identify various disease genes (Boyartchuk et al. 2001; Carter et al. 1992; Miki et al. 1994). For example, in a large linkage study with a total of 3,796 individuals from 263 prostate cancer families, three quantitative trait loci (QTLs) were found to contribute to the variation in the age-at-onset of hereditary prostate cancer (Conlon et al., 2003). Linkage analysis and association mapping are the two most popular approaches to map complex trait (such as age-at-onset) loci. Statistical methods have been developed for mapping time-to-event loci using linkage analysis (Carter et al., 1992; Claus et al., 1990; Li and Zhong 2002, Pankratz et al., 2005; Symons et al., 2002) as well as family-based association mapping. For example, Li and Fan (2000) proposed a linkage disequilibrium-based Cox (1972) model for nuclear family data and used a robust Wald test for the association test between a marker and a disease with variable age of onset. Mokliatchouk et al. (2001) and Shih and Whittemore (2002) developed likelihood-based score statistics to test for the association between a disease and a genetic marker. The score statistic can be written as a weighted sum over family members of their observed minus expected genotypes. Age of onset data can be used in the weight, which is the difference between the observed and expected value, $\delta_i - \Lambda_{i0}(t_i)$ for individual i, where $\Lambda_{i0}(\cdot)$ is the cumulative hazard function, which is assumed to be known from external data sources. Both methods

of Li and Fan (2000) and Shih and Whittemore (2002) assume that the genetic effects on the risk of onset are proportional in the framework of the Cox regression model. Jiang et al. (2006) developed a family-based association test for time-to-onset data by assuming time-dependent differences between the hazard functions among different genotype groups and use of the weighted log-rank approach of Flemming and Harrington (1981). Although population-based association mapping is emerging as a powerful, general tool for identifying loci associated with the inheritance of complex traits, little is available for the analysis of time-to-event outcomes in population-based association studies.

Because the allelic association due to linkage disequilibrium usually operates over shorter genetic distances, association mapping permits higher resolution mapping than does linkage analysis (Hästbacka et al., 1992). However, population-based association mapping may be subject to false positives caused by population structure i.e. a population is subdivided into some subpopulations with different gene frequencies. "Spurious associations" or false positives are associations between phenotypes and markers that are not linked to any causative loci (Lander et al., 1994). Genomic control (GC) (Devlin and Roeder, 1999), Structured association (SA) (Pritchard et al., 2000), and Principal component analysis (PCA) (Price et al., 2006) are the three prevailing statistical methods which use genetic marker data to adjust for population structure. These methods have been proven useful in a variety of contexts, however,

they have limitations. For many real life data sets having population structure along with diverse levels of familial relatedness within subpopulations, i.e. continuous population stratification, the GC, SA, and PCA approaches may lead to either inadequate control for false positives or a loss in power owing to genetic correlation within subpopulations (Yu et al., 2006; Zhao et al., 2007). Yu et al. (2006) recently introduced a unified mixed model approach to association mapping in structured populations for normally distributed and fully observed data. In his approach, the random effects attributable to relative kinship in addition to the fixed effects of population structure variables are included in the model to adjust for continuous population stratification. The covariance matrix of the random effects is $2K\sigma_G^2$, where σ_G^2 is the additive genetic variance, and K is the relative kinship matrix, assumed to be symmetric and positive definite. A kinship coefficient between individuals i and j is often defined as the probability of identity by descent of the genetic markers compared (Ritland, 1996), but estimators based on genetic markers actually estimate a "relative kinship", that can be defined as the ratio of differences of probabilities of identity by state (Hardy et al., 2002) i.e. $K_{(i,j)} = (Q_{ij} - Q_m)/(1 - Q_m)$, where Q_{ij} is the probability of identity in state for genetic markers from individuals i and j, and Q_m is the average probability of identity by state for genetic markers coming from random individuals from the sample. Note that with this definition, negative

relative kinship coefficients naturally occur between some individuals. This is interpreted as meaning that these subjects are less related than randomly selected individuals and the negative kinship coefficients are set to 0 (Hardy et al., 2002; Ritland, 1996; Yu et al., 2006). Li et al. (2008) recently developed a semiparametric approach to association mapping in structured populations for fully observed data. In their approach, the error distribution is flexibly modeled as a mixture of Polya trees (MPT) centered around a family of normal distributions, hence important, data-driven features, such as skewness and multimodality, are allowed. Both the effects attributable to population structure variables X and the effects due to kinship are included in the model, where the kinship effects are restricted to the space orthogonal to the column space of X. Specifically, the kinship effect for subject i is modeled as $\eta_i = \sum_{j=1}^{n} X_K^c(i,j)\gamma_j$ for $i = 1, \ldots, n$ assuming $\gamma \sim \mathrm{N}(0, \mathrm{I}_n\tau^{-1})$ in their prior distributions for $\gamma = (\gamma_1, \ldots, \gamma_n)'$, where K_s is the symmetric square root of K assuming K is positive definite, $X_K^c = (\mathrm{I} - X(X'X)^{-1}X')K_s$ (I here is an identity matrix).

In this article, we have considered the important and challenging problem of detecting the existence of gene(s) for time-to-event traits in population based association mapping studies where the sampled individuals appear to have complex population stratification i.e. diverse levels of familial relatedness within subpopulations. This problem has not yet been addressed in the literature. The accelerated failure

time (AFT) model is widely seen as an alternative to the popular Cox proportional hazard (PH) model when the assumption of proportional hazards is questionable. Even when both models are plausible, the AFT model still has some advantages over the PH model. For example, the covariate effects on the failure time are modeled directly rather than indirectly, as in the PH model (Wei, 1992). Hence, we first propose a parametric Weibull frailty AFT model to model the relationship between the time-at-onset data and genotypic scores at a candidate marker. Since it is difficult to assess the distributional assumptions with censored data, it is preferred to leave the distribution of survival times unspecified or, alternatively, to specify it in a flexible way. Therefore, we also propose to model the association between a candidate gene and time-to-event data via a Bayesian semiparametric frailty AFT model, where the baseline function is flexibly modeled as a MPT centered around a family of Weibull distributions (Ferguson, 1974; Hanson, 2006) and thus may be viewed as a generalization of standard models in which important data-driven features are allowed. A mixture is attractive in that it smoothes over partitioning effects associated with a simple Polya tree. The mixture of Polya trees prior provides an intermediate choice between a strictly parametric model and a completely arbitrary model (Ferguson, 1974; Hanson, 2006; Lavine, 1992). Like Yu et al. and Li et al. approaches to association mapping in structured populations for fully observed data, the proposed approaches here for censored data adjust for continuous population

stratification by including population structure variables as covariates, derived from the individual's genotypes at a series of independent markers. In addition, the proposed methods include the frailty terms, the effects attributable to K, to adjust for continuous population stratification. One limitation of the approaches by Yu et al. and Li et al. is that K has to be positive definite, however K is not guaranteed to be positive definite for real life data (Kang et al., 2008). Our approaches, hence, relax the strict positive definite assumption of K. Specifically, we will model effects attributable to kinship via a Gaussian conditional autoregressive (CAR) model (Besag, 1974). An autoregressive model for a stochastic process x_t in a plane assumes $p(x_t|x_1, \ldots, x_n) = p(x_t|x_{s_1}, \ldots, x_{s_m})$ for some set of neighbors $(s_1, \ldots, s_m)$ that depend on t. In this paper, the autoregressive process is defined through the conditional distribution of subject i's kinship effect γ_i given the rest $\gamma_{j\neq i}$, which depends only on its neighbors. Two subjects are defined to be neighbors if their pairwise relative kinship coefficient $K_{(i,j)}$ is greater than 0, with larger values of the kinship coefficient associated with j's closer to i than those farther away from i. In the proposed approaches, the kinship effects are introduced through a novel CAR prior distributional specification with $2K$ being the adjacency matrix (or connectivity structure) connecting subjects and $(B - W)\tau$ as the precision matrix in the CAR prior, where $W_{(i,j)} = 2K_{(i,j)}$ for $i \neq j$ and 0 otherwise, and B is a diagonal matrix with $B_{(i,i)} = \sum_{j=1}^{n} W_{(i,j)}$. To illustrate the utility and interpretation of our

models, we applied the proposed models to previously published flowering time data sets of association mapping in *Arabidopsis thaliana* lines (Zhao et al., 2007). We compared our proposed models in type I error rate to the parametric Weibull AFT model without considering the effects attributable to kinship, a common practice in population-based association mapping studies. We also conducted simulation studies to demonstrate the type I error rate, power, precision, and accuracy of the proposed methods.

4.3 Methods

For each subject $i = 1, \ldots, n$, let T_i denote the failure time and C_i denote the censoring time. The observed survival data are $t_i = \min(T_i, C_i)$ and $\delta_i = \mathrm{I}(T_i \leq C_i)$, where $\mathrm{I}(\cdot)$ is the indicator function. Let $x_i = (x_{i1}, \ldots, x_{ip})'$ denote a p-dimensional vector of covariates associated with subject i, where x_{i1} is the genotypic score (haplotype, genotype or gene expression level) for individual i at a candidate gene for a given trait, and $x_{i2}, \ldots, x_{ip}$ are the values of the population structure variables for individual i. Note that the model can be easily extended to include other covariates if necessary. Let $\beta = (\beta_1', \beta_2, \ldots, \beta_p)'$ be the regression coefficients corresponding to the p predictors. For a candidate gene with L categorical values, there are $(L-1)$ orthogonal dummy variables corresponding to the candidate gene, and in such cases, $x_{i1} = (x_{i11}, \ldots, x_{i1(L-1)})'$; $\beta_1 = (\beta_{11}, \ldots, \beta_{1(L-1)})'$. We used the PCA method (Price

et al., 2006) to estimate the population structure variables $x_2, \ldots, x_p$ using data at a set of random markers. To decide p, we will fit a parametric Weibull AFT model via a frequentist approach with the first d $(1 \leq d < n)$ largest principal components of marker data as the covariates and the time of onset data as the survival time. The population structure variables will be given by the first $(p-1)$ largest principal components whose regression coefficient has a $p-$value $\leq$ the significance level of α. Let $K_i = (K_{(i,1)}, \ldots, K_{(i,n)})'$ ($K_{(i,j)}$ being the kinship coefficient between individuals i and j), let $\mathcal{D}_i = (x_i, t_i, \delta_i, K_i)$ denote data for the i^{th} subject, and $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^n$. Let $T_i \sim S_{(x_i,K)}(\cdot)$. Let the frailty term, γ_i, be the effect attributable to kinship for subject i for $i = 1, \ldots, n$. Given a baseline survival function S_0, a vector of regression coefficients β, and frailties $\gamma = (\gamma_i, \ldots, \gamma_n)'$, the AFT model defines the survival function $S_{(x,K)}(t)$ for an individual with a p-dimensional covariate vector x and kinship matrix K through the relation:

$$S_{(x_i,K,\gamma)}(t) = S_0(\exp\{-(x_i'\beta + \gamma_i)\}t).$$

In the next three sections, we will introduce parametric and semiparmetric AFT models, particularly, we will consider different priors for vector γ. We fit all three models using a Bayesian approach.

4.3.1 Parametric Weibull AFT model - model 1

We first consider a parametric AFT model for association tests in structured populations with censored data. We include only population structure variables as covariates to account for population stratification without kinship effects in association tests, as is the common practice in population based association studies i.e. $\gamma_i = 0$ for all i. This model may be formulated as

$$S(t_i|x_i) = S_0(\exp\{-(x'_{i1}\beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})\}t_i) \tag{4.1}$$

for $i = 1, \ldots, n$, where $S_0(t) = \mathrm{e}^{-(t/\lambda)^\alpha}$ is the baseline survival function i.e. $S_0(\cdot) = \text{Weibull}(\alpha, \lambda)$. We denote this model as model 1.

4.3.2 Proposed parametric Weibull frailty AFT model - model 2

Second, we consider a parametric Weibull frailty AFT model for association tests in structured populations with censored data i.e. we include population structure variables as covariates, but now use $2K$ as the neighboring structure in the intrinsic CAR distribution to account for multiple levels of genetic relatedness in the sample as discussed in the introduction. This model may be formulated as

$$S(t_i|x_i, K, \gamma) = S_0(\exp\{-(x'_{i1}\beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \gamma_i)\}t_i) \tag{4.2}$$

for $i = 1, \ldots, n$, again $S_0(\cdot) = \text{Weibull}(\alpha, \lambda)$. Here we consider the $\gamma|\tau \sim \text{CAR}(\tau)$ prior. Specifically, let $K_{(i,j)}$ be the kinship coefficient between individuals i and j as defined the introduction, $K_{s(i)} = \sum_{j \neq i} 2K_{(i,j)}$ for $i = 1, \ldots, n$, $\bar{\gamma}_i = \sum_{j \neq i} 2K_{(i,j)}\gamma_j / K_{s(i)}$, $\tau_i = \tau K_{s(i)}$, the full conditional for $\gamma_i|\gamma_{j \neq i}$ is

$$\pi(\gamma_i|\gamma_j, j \neq i) = \frac{\sqrt{\tau_i}}{\sqrt{2\pi}} \exp\left\{-\frac{\tau_i}{2}(\gamma_i - \bar{\gamma}_i)^2\right\}.$$

Note that $\pi(\gamma|\tau)$ is improper, therefore the constraint $\sum_{i=1}^{n} \gamma_i = 0$ will be imposed to ensure the propriety of the full conditionals, $\pi(\gamma_i|\gamma_j, j \neq i)$. Such priors were introduced and have been used quite extensively in the spatial statistics literature (Besag et al., 1995). We denote this model as model 2.

4.3.3 Proposed semiparametric Weibull frailty AFT model - model 3

Now we consider a semiparametric AFT model, we denote this model as model 3, and the associated survivor function is

$$S(t_i|x_i, K, \gamma) = S_0(\exp\{-(x_{i1}'\beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \gamma_i)\}t_i) \tag{4.3}$$

for $i = 1, \ldots, n$, where again $S_0(\cdot)$ is the baseline function. First, we again include population structure variables as covariates, and use $2K$ as the neighboring structure

in CAR distribution to account for multiple levels of genetic relatedness in the sample as described in model 2, but now we consider a MPT prior on $S_0(\cdot)$ i.e.

$$S_0|\theta \sim \mathrm{PT}_M(c, \rho, G_\theta),\ \theta \sim \pi(\theta) \tag{4.4}$$

We briefly describe the prior but leave details to Hanson and Johnson (2002), Lavine (1992), and Walker and Mallick (1999). Let M (a fixed positive integer) denote the number of MPT tree levels, typically $M \leq \log_2(n)$, where n is the sample size. Let G_θ denote a family of Weibull cumulative distribution functions $G_\theta(t) = 1 - \mathrm{e}^{-(t/\alpha)^\lambda}$ for $t > 0$. Here $G_\theta(\cdot)$ is the centering distribution of the Polya tree prior. A Polya tree prior is constructed from a set of partitions $\Pi_M^\theta = \{B_\epsilon^\theta : \epsilon \in \bigcup_{j=1}^M \{0,1\}^j\}$ and a family $\mathcal{A}$ of positive real numbers. Here, the partition points are quantiles of the centering family: if j is the base-10 representation of the binary number $\epsilon = \epsilon_1, \ldots, \epsilon_k$ at level k, then $B^\theta_{\epsilon_1,\ldots,\epsilon_k}$ is defined to be the interval $(G_\theta^{-1}(j/2^k), G_\theta^{-1}((j+1)/2^k)]$, except the rightmost set is $B^\theta_{11,\ldots,1} = G_\theta^{-1}((2^k-1)/2^k), \infty)$. For example, with $k = 3$, and $\epsilon = 000$, then $j = 0$ and $B^\theta_{000} = (0, G_\theta^{-1}(1/8)]$, and with $\epsilon = 010$, then $j = 2$ and $B^\theta_{010} = (G_\theta^{-1}(2/8), G_\theta^{-1}(3/8)]$, etc. Note then that at each level k, the class $\{B_\epsilon^\theta : \epsilon \in \{0,1\}^k\}$ forms a partition of the positive reals and furthermore $B^\theta_{\epsilon_1,\ldots,\epsilon_k} = B^\theta_{\epsilon_1,\ldots,\epsilon_k 0} \cup B^\theta_{\epsilon_1,\ldots,\epsilon_k 1}$ for any binary $\epsilon_1, \ldots, \epsilon_k$. We take the family $\mathcal{A} = \{\alpha_\epsilon : \epsilon \in \bigcup_{j=1}^M \{0,1\}^j\}$ to be defined by $\alpha_{\epsilon_1,\ldots,\epsilon_k} = c\rho(k)$ for some $c > 0$ (Hanson 2006). The parameter c is the amount of weight attached to $G_\theta(\cdot)$. As c tends to

zero, the posterior $S_0|\mathcal{D}$ is almost entirely data driven. As c tends to infinity, we obtain a fully parametric analysis. Given $\prod_M^\theta$ and $\mathcal{A}$, the Polya tree prior is defined up to level M by the random vectors $\mathcal{Y}_M = \{(\mathcal{Y}_{\epsilon_0}, \mathcal{Y}_{\epsilon_1}) : \epsilon \in \bigcup_{j=0}^{M-1}\{0,1\}^j\}$ through the product of conditional probabilities

$$S_0(B^\theta_{\epsilon_1,\ldots,\epsilon_k}|\mathcal{Y}_M, \theta) = \prod_{j=1}^{k} \mathcal{Y}_{\epsilon_1,\ldots,\epsilon_j},$$

for $k = 1, \ldots, M$, where we define $S_0(A)$ to be the baseline measure of any set A. At the coarsest level, $(\mathcal{Y}_0, \mathcal{Y}_1)$ is set to (0.5, 0.5) for identifiability. The remaining vectors $(\mathcal{Y}_{\epsilon_0}, \mathcal{Y}_{\epsilon_1})$ are independent Dirichlet:

$$(\mathcal{Y}_{\epsilon_0}, \mathcal{Y}_{\epsilon_1}) \sim \text{Dirichlet}(\alpha_{\epsilon_0}, \alpha_{\epsilon_1}),\ \epsilon \in \bigcup_{j=1}^{M-1}\{0,1\}^j.$$

Define the vector of probabilities $\mathbf{p}_\mathcal{Y} = (p_\mathcal{Y}(1), p_\mathcal{Y}(2), \ldots, p_\mathcal{Y}(2^M))'$ through

$$p_\mathcal{Y}(j+1) = S_0(B^\theta_{\epsilon_1,\ldots,\epsilon_M}|\mathcal{Y}_M, \theta) = \prod_{i=1}^{M} \mathcal{Y}_{\epsilon_1,\ldots,\epsilon_i},$$

where $\epsilon_1, \ldots, \epsilon_M$ is the base-10 binary representation of the integer j, $j = 0, \ldots, (2^M - 1)$. If we denote $N_\theta(t)$ as the integer part of $2^M G_\theta(t) + 1$, after simplification, the

baseline survival function is

$$S_0(t|\mathcal{Y}_M,\theta) = p_{\mathcal{Y}}(N_\theta(t))[N_\theta(t) - 2^M G_\theta(t)] + \sum_{j=N_\theta(t)}^{2^M} p_{\mathcal{Y}(j)}.$$

The density associated with $S_0(t|\mathcal{Y}_M,\theta)$ is given by

$$f_0(t|\mathcal{Y}_M,\theta) = \sum_{j=1}^{N_\theta(t)} p_{\mathcal{Y}}(j) g_\theta(t) \mathrm{I}_{B_\theta}\{\epsilon_M(j-1)\}(t) = 2^M p_{\mathcal{Y}}(N_\theta(t)) g_\theta(t), \tag{4.5}$$

where $g_\theta(\cdot)$ is the density corresponding to $G_\theta(.)$ and $\epsilon_M(i)$ is the binary representation $\epsilon_1,\ldots,\epsilon_M$ of the integer i. Given S_0 (through $\mathcal{Y}_M$, α, and λ), β, and γ, let $\mu_i = x'_{i1}\beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \gamma_i$, the survival function for subject i under model 3 is

$$S_{(x_i,K)}(t_i|\mathcal{Y}_M,\alpha,\lambda,\beta,\gamma) = S_0(\exp(-\mu_i)t_i|\mathcal{Y}_M,\alpha,\lambda)$$

and the density is

$$f_{(x_i,K)}(t_i|\mathcal{Y}_M,\alpha,\lambda,\beta,\gamma) = f_0(\exp(-\mu_i)t_i|\mathcal{Y}_M,\alpha,\lambda)\exp(-\mu_i).$$

The likelihood is then given by:

$$\mathcal{L}((t,\delta);\mathcal{Y}_M,\beta,\gamma,\alpha,\lambda) = \prod_{i=1}^{n} f_{(x_i,K)}(t_i|\mathcal{Y}_M,\alpha,\lambda,\beta,\gamma)^{\delta_i} S_{(x_i,K)}(t_i|\mathcal{Y}_M,\alpha,\lambda,\beta,\gamma)^{1-\delta_i}.$$

4.3.4 Model implementation

We now describe how we obtain samples from the posterior for model 3. We set $M = 4$, and $\rho(k) = ck^2$, i.e. $\alpha_{\varepsilon_1,\ldots,\varepsilon_k} = c\rho(k) = ck^2$ for the MPT prior. In contrast to Li et al. approach where c is fixed, here we place a Gamma(a_c, b_c) prior on c i.e. $\pi(c) = \frac{b_c^{a_c}}{\Gamma(a_c)} c^{a_c-1}\mathrm{e}^{-b_c c}$. We tried various values of a_c and b_c. For example, we set (a_c, b_c)=(1, 2), (a_c, b_c)=(5, 2), and (a_c, b_c)=(10, 2). The models appear to be robust to these values. Hence our discussion is focus on the case when (a_c, b_c)=(5, 2), a prior specification that allows for substantial variability in S_0 relative to the centering Weibull distribution. Throughout this paper, we assume $\mathcal{Y}_M|c,\ c,\ \beta,\ \alpha,\ \lambda,\ \gamma|\tau,\ \tau$ are independent in prior distributions with $\pi(\alpha, \lambda, \beta) \propto 1$ and $\tau \sim$ Gamma (a, b) with $a = b = 0.1$, i.e. $\pi(\tau) = \frac{b^a}{\Gamma(b)}\tau^{a-1}\mathrm{e}^{-b\tau}$. The joint posterior given data $\mathcal{D}$, and prior $\pi(\mathcal{Y}_M, \alpha, \lambda, \beta, \gamma, \tau)$, is thus proportional to $\mathcal{L}((t, \delta); \mathcal{Y}_M, \alpha, \lambda, \beta, \gamma)\pi(\mathcal{Y}_M|c)\pi(c)\pi(\gamma|\tau)\pi(\tau)$. All algorithms use a Metropolis-Hastings (M-H) step for updating the components $(\mathcal{Y}_{\epsilon_0}, \mathcal{Y}_{\epsilon_1})$ one at a time by sampling candidates $(\mathcal{Y}^*_{\epsilon_0}, \mathcal{Y}^*_{\epsilon_1})$ from a Dirichlet $(m\mathcal{Y}_{\epsilon_0}, m\mathcal{Y}_{\epsilon_1})$ distribution, where we set $m = 30$. We let $\mathcal{L}^* = \mathcal{L}((t, \delta);\ \mathcal{Y}^*_M, \alpha, \lambda, \beta, \gamma)$ and $\mathcal{L} = \mathcal{L}((t, \delta);\ \mathcal{Y}_M, \alpha, \lambda, \beta, \gamma)$, the candidate $(\mathcal{Y}^*_{\epsilon_0}, \mathcal{Y}^*_{\epsilon_1})$ is accepted with probability

$$\rho = \min\left\{1, \frac{\Gamma(m\mathcal{Y}_{\epsilon_0})\Gamma(m\mathcal{Y}_{\epsilon_1})(\mathcal{Y}_{\epsilon_0})^{m\mathcal{Y}^*_{\epsilon_0}-ck^2}(\mathcal{Y}_{\epsilon_1})^{m\mathcal{Y}^*_{\epsilon_1}-ck^2}\mathcal{L}^*}{\Gamma(m\mathcal{Y}^*_{\epsilon_0})\Gamma(m\mathcal{Y}^*_{\epsilon_1})(\mathcal{Y}^*_{\epsilon_0})^{m\mathcal{Y}_{\epsilon_0}-ck^2}(\mathcal{Y}^*_{\epsilon_1})^{m\mathcal{Y}_{\epsilon_1}-ck^2}\mathcal{L}}\right\}.$$

For the M-H step to update c, we sample $c^* \sim N(c, \sigma^2)$ constrained to $c^* > 0$, where σ^2 is chosen such that the acceptance rate of c^* is $30\% \sim 40\%$. The full conditional density of c only depends on $\mathcal{Y}$, that is

$$p(c|\mathcal{Y}, \alpha, \lambda, \beta, \gamma, \tau, \mathcal{D}) \propto \prod_{k=2}^{M} \prod_{\epsilon \in \bigcup_{j=0}^{(k-1)} \{0,1\}^j} \text{Dirichlet}((\mathcal{Y}_{\epsilon_0} \mathcal{Y}_{\epsilon_1})|(ck^2, ck^2))\pi(c).$$

Hence, accept c^* with probability

$$\rho = \min\left\{1, \prod_{k=2}^{M} \prod_{\epsilon \in \bigcup_{j=0}^{(k-1)} \{0,1\}^j} \frac{(\mathcal{Y}_{\epsilon_0})^{c^*k^2} (\mathcal{Y}_{\epsilon_1})^{c^*k^2} \Gamma(2c^*k^2)(\Gamma(ck^2))^2}{(\mathcal{Y}_{\epsilon_0})^{ck^2} (\mathcal{Y}_{\epsilon_1})^{ck^2} \Gamma(2ck^2)(\Gamma(c^*k^2))^2} (c^*/c)^{a_c - 1} \mathrm{e}^{-b_c(c^* - c)}\right\}.$$

For the M-H step for updating the parameter vector $(\alpha, \lambda, \beta,)$, we used a multivariate normal random-walk proposal constrained to $\alpha^* > 0$ and $\lambda^* > 0$. The proposal covariance matrix is a scaled version of the estimated covariance matrix the parametric non-frailty model 1. The variance scaling factor was set to be a value such that a Markov chain accepted $30\% \sim 40\%$ of the proposed moves. The candidate $(\alpha^*, \lambda^*, \beta^*)$ is accepted with probability

$$\rho = \min\left(1, \frac{\mathcal{L}((t, \delta);\ \mathcal{Y}_M, \alpha^*, \lambda^*, \beta^*, \gamma)}{\mathcal{L}((t, \delta);\ \mathcal{Y}_M, \alpha, \lambda, \beta, \gamma)}\right).$$

We use a M-H step for updating each γ_i individually. We sample $\gamma_i^\star \sim \mathrm{N}(\bar{\gamma}_i, \tau_i^{-1})$ for $i = 1, \ldots, n$, where $K_{s(i)} = \sum_{j \neq i} 2K_{(i,j)}$, $\bar{\gamma}_i = \sum_{j \neq i} 2K_{(i,j)}\gamma_j / K_{s(i)}$, and $\tau_i = \tau K_{s(i)}$.

We accept γ_i^* with probability

$$\rho = \min\left(1, \frac{f_{(x_i,K)}(t_i|\mathcal{Y}_M, \alpha, \lambda, \beta, \gamma_i^*)^{\delta_i} S_{(x_i,K)}(t_i|\mathcal{Y}_M, \alpha, \lambda, \beta, \gamma_i^*)^{1-\delta_i}}{f_{(x_i,K)}(t_i|\mathcal{Y}_M, \alpha, \lambda, \beta, \gamma_i)^{\delta_i} S_{(x_i,K)}(t_i|\mathcal{Y}_M, \alpha, \lambda, \beta, \gamma_i)^{1-\delta_i}}\right).$$

We used the full conditional to update of τ by sampling $\tau^* \sim \text{Gamma}(c,\ d)$, where $c = 0.1+0.5(n-1)$, and $d = 0.1+0.5\sum_{i\neq j} 2K_{(i,j)}(\gamma_i-\gamma_j)^2 = 0.1+0.5\sum_{i=1}^{n} K_{s(i)}\gamma_i(\gamma_i-\bar{\gamma}_i)$. For each of the 3 models, two Markov chains were run in parallel to monitor the convergence of the Metropolis algorithm with 100,000 iterations per chain for burn-in and 5,000 Monte Carlo samples from 200,000 iterations per chain for inference. C++ software was used for the analysis.

4.4 Application to the *Arabidopsis thaliana* data set

The 95 *Arabidopsis thaliana* lines used in this study have been described previously (Zhao et al., 2007), where the data was assumed normally distributed. For this specific data set, subjects are the Arabidopsis lines. The traits we are interested in for association mapping are the plant flowering time measured in days from germination to the first opening of flowers at the University of Southern California without vernalization under (1) long-day conditions (16 hour (h) light/8 h dark)(LD) and (2) short days (8h light/16 h dark) (SD). For LD, 14 lines out of 95 did not flower before the study was ended i.e. they were censored, while for SD, 23 lines out of 92 were censored. We used 5,000 high quality SNP halophyte markers for inferring

K and population structure variables via the PCA approach. The population structure variables are given by the first eight largest principal components of the SNP data. The data was originally analyzed by Zhao et al., 2007 using Yu's mixed model ignoring the fact that some lines were censored.

To compare model performance, we tested the association between LD and FRIGIDA (*FRI*), a known central regulator of flowering time for unvernalized Arabidopsis plants. There are 3 haplotypes at FRI for the sampled lines, where 1 represents the wild type lines, 2 represents the mutated lines with the deletion of the *Ler* allele, and 3 represents the mutated lines with the deletion of the *Col* allele (Zhao et al., 2007). In addition, we did the association tests between LD and a randomly selected SNP marker, S1.107, for model comparison.

Listed in table 4.1 are the effect (δ_1) of the *Col* deletion haplotype and the effect (δ_2) of the *Ler* deletion haplotype from the FRI and LD association test by the 3 models. Here δ_1 and δ_2 are estimated by the mean difference in flowering time between the wild type group and the *Col* deletion group, the wild type group and the *Ler* deletion group, respectively. The 95% credible sets for both δ_1 and δ_2 excluded 0 for all 3 models, as expected. The integrated hazard plot for Cox-Snell residuals is often used to assess model fit. Here Cox-Snell residuals are estimated as $-\ln(\hat{S_i}(t_i))$. Figure 4.1 is the integrated hazard plots for Cox-Snell residuals from SNP $S1.107$ and LD association test for each of the 3 models. The plots of the integrated hazards for

the Cox-Snell residuals indicate lack of fit for model 1, the parametric model without the frailties attributable to kinship. Thus model 1 is not appropriate for the current data set. Model 2, the parametric model with the frailties, shows a substantial improvement in model fit over model 1, while model 3, the semiparametric model with the frailties, provides the best model fit for the current data set among the 3 models.

To compare the false positive rates among the 3 models, we randomly selected 500 SNP haplotype makers with two haplotypes, 0 and 1, for each SNP. These SNPs are distributed over the whole *Arabidopsis thaliana* genome and have their minor allele frequencies greater than 5%. We applied the 3 models to each of these 500 SNPs for testing the association with LD. In addition, we applied both approaches by Yu et al. and Li et al. to the 500 SNP data assuming the flowering time, LD, was fully observed. We estimated the false positive rate which is defined as $500q$ with q being the proportion of these 500 SNPs whose 95% credible sets of the SNP regression coefficients excluded 0. The type I error rate is the largest for model 1 (16.8%), which is significantly larger than the expected value of 5%. Model 2 has the type I error rate of 6.2%, which is close to the expected value of 5%. Model 3 has the smallest type I error rate (3.6%), again which is close to the expected value of 5%. These observations indicate that the two proposed models with the frailty terms attributable to kinship included substantially reduce the type I error rate and

improve the model fit over the model without the kinship effects for the current data set. The type I error rates are 12.6% and 7.8% from Yu et al. and Li et al. approaches which ignore the censoring status, respectively, which are substantially larger than the type I error rates from the proposed models 2 and 3, which considered the censoring status. Moreover, the approaches by Yu et al. and Li et al. which ignored the censoring status showed a lack of model fit (not shown), thus they are not appropriate for the current data set.

We also did association test for each of randomly selected 200 SNPs with SD. The type I error rates are 15.0%, 8.5%, 5.5% for model 1, 2, and 3, respectively. The two models with frailties again showed a substantially lower false positive rate than the model without frailties for this data set. In particular, the false positive rate is close to the expected value of 5% for the semiparametric model 3, while the false positive rates are significantly larger than the expected value for the two parametric models.

4.5 Simulation study I

Our goal for this simulation was to simulate a data set with the same covariate structure as the data from the previous section given a true complex genetic correlation structure among individuals. Our simulations are similar to those carried out by Yu et al. and Zhao et al. in their association studies. We simulated m Arabidopsis

Table 4.1: Comparisons of the 3 models in type I error rate, 95% credible sets (95%CI) of $\hat{\delta_1}$ and $\hat{\delta_2}$ for FRI, and $\hat{\beta}_1$s for S1.107 from the association tests with LD

	FRI		S1.107	Type I	
Model	$\hat{\delta_1}$ 95%CI	$\hat{\delta_2}$ 95%CI	$\hat{\beta}_1$ 95%CI	error rate	Power
Model 1	(0.210, 0.826)	(0.541, 1.232)	(-0.254, 0.268)	16.8	92.0
Model 2	(0.162, 0.801)	(0.383, 1.198)	(-0.324, 0.153)	6.2	96.8
Model 3	(0.147, 0.811)	(0.328, 1.219)	(-0.346, 0.176)	3.6	95.2

Here δ_1 and δ_2 are the mean difference of flowering times between the wild type group and the *Col* deletion group, between the wild group and the *Ler* deletion group, respectively. For each of the 3 models, 500 haplotype SNPs and 500 simulated QTNs were used in estimating type I error rate and power, respectively.

flowering time genetic loci with $m = 500$ and 200 for LD and SD, respectively. For each of the randomly selected m SNPs in section 3, a fixed additive genetic effect $\zeta = s\sqrt{\kappa/((1-\kappa)p(1-p))}$ was added to the observed flowering time (Long et al., 1999; Yu et al., 2006), here $\kappa = 0.0045$ and 0.002 for LD and SD, respectively, s=$\sqrt{\hat{\sigma}^2}$ ($\hat{\sigma}^2$ is the usual unbiased estimate of the variance, assuming lines are independent), and p is the sample minor allele frequency of a SNP. Specifically, the new responsible variable value y_i for subject i is $\log(y_i) = \log(t_i) + \zeta x_i$ for $i = 1, \ldots, n$, here t_i and x_i are the observed flowering time and the SNP haplotypic score for individual i, respectively. For each individual, we used the same censoring status as the real data. We denoted those simulated flowering time genetic loci as quantitative trait nucleotides (QTNs). For each of m simulated data sets, we applied the 3 models with the same Bayesian implementation as in the analysis from section 3.

For each model, we calculated the power as mq with q being the proportion of the m QTNs whose 95% credible set for QTN regression coefficient excluded 0. As shown in table 1, the power for detecting LD QTNs for model 1 is 92.0% compared to 96.8% and 95.2% for model 2 and 3, respectively. Hence the two frailty models 2 and 3 have higher power in detecting gene and trait association than model 1 without frailties. The power for detecting SD QTNs for model 1 is 94.0%, while the powers are 97.0% and 97.5% for model 2 and 3, respectively. Again, two frailty models 2 and 3 show higher power in detecting gene and trait association than model 1 without frailties. The semiparametric frailty model 3 and parametric frailty model 2 showed similar power in detecting QTNs.

4.6 Simulation study II

The first simulation study maintains the true complex genetic correlation structure among subjects. However, the limitations are that the bias and mean square error cannot be compared among the models as the true value of the gene effect of a given SNP is unknown. Our second simulation is based on empirical population genetic data and is similar to those carried out by Zhang et al. (2003) and and Li et al. (2008) in their association mapping studies. We assume that sampled individuals were genotyped at a series of unlinked SNP loci with two alleles A and a. There are 3 genotypes at a SNP marker: aa (or -1), Aa (or 0), and AA (or 1). We also

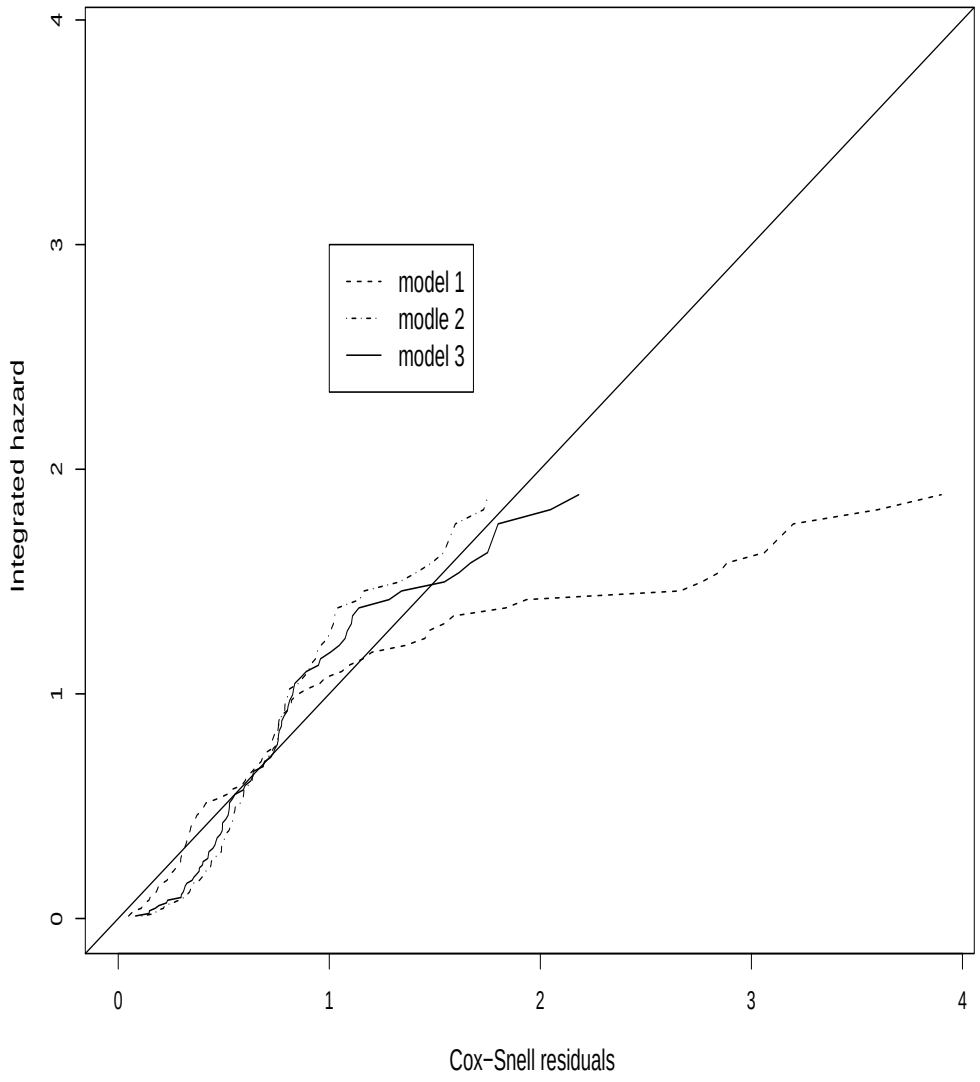

Figure 4.1: Integrated hazard plots for Cox-Snell residuals for the 3 models from SNP $S1.107$ and LD association test.

assume that the population under study is a continuous admixture of two ancestral human populations including Biaka and Danes, and we further assume that some of subjects are genetically related. Specifically, we assume the kinship matrix, K is an n by n matrix, we let

$$K = \begin{pmatrix} A_{110} & 0 & 0 & 0 & 0 \\ 0 & B_{10} & 0 & 0 & 0 \\ 0 & 0 & C_{10} & 0 & 0 \\ 0 & 0 & 0 & D_{10} & 0 \\ 0 & 0 & 0 & 0 & E_{10} \end{pmatrix},$$

where $K_{(i,i)} = 0.5$ for $i = 1, \ldots, n$, A is a square matrix with dimension of 110 with $A_{(i,j)} = 0.0$ for $i = 1, \ldots, 110$, $j = 1, \ldots, 110$, and $i \neq j$, B, C, D, E are square matrices with dimension of 10 and $B_{(i,j)} = 0.35$, $C_{(i,j)} = 0.3$, $D_{(i,j)} = 0.25$, $E_{(i,j)}$=0.20, for $i = 1, \ldots, 10$, $j = 1, \ldots, 10$, and $i \neq j$. Note the addition genetic correlation between subjects is $\rho = 2K$. We simulated individual's SNP genotypes based on both the individual's ancestry background and K. The allele frequencies of 500 unlinked SNPs with their minor allele frequencies greater than 5% across the two ancestral populations were extracted from the population genetics database ALFRED (Rajeevan et al., 2003) and we repeated these 500 SNPs 2 times as if we had 1000 SNPs. First, we generated 1000 independent null SNPs with $n = 150$ individuals. These 1000 null SNPs will be used to estimate population structure variables and K. Let p_{i1} and p_{i2} represent the probabilities that allele A of the

i^{th} individual originated from Biaka, and Danes, respectively, and $p_{i1} + p_{i2}$=1 for $i = 1, \ldots, n$. We simulate 110 out of 150 individuals as follows. We assume $p_{i1} \sim$ Uniform$(0, 1)$ independently for each individual i. Therefore, the allele frequency of allele A at marker l for individual i can be written as

$$p_{il} = q_{l1}p_{i1} + q_{l2}(1 - p_{i1}), \tag{4.6}$$

where q_{l1}, q_{l2} are the population allele frequencies of allele A at marker l in Biaka and Danes, respectively. Individual i was assigned genotype -1, 0 or 1 at marker l with probabilities $(1 - p_{il})^2$, $2p_{il}(1 - p_{il})$, p_{il}^2, respectively for $i = 1, \ldots, 110$ and $l = 1, \ldots, 1000$. We denote those 110 genetically independent individuals as group 1. For the remaining 40 individuals, we assume they are from four different genetically related groups with genetic correlation between individuals within a group being $\rho = 0.7, 0.6, 0.5, 0.4$ for groups 2, 3, 4, 5, respectively and 0 across the groups, i.e. we assume genetic independence between all groups. We assume $p \sim$ Uniform$(0, 1)$ independently for each of the four groups, then $p_{i1} = p + \epsilon_i$ for $i = 1, \ldots, 10$, where ϵ_i is a small quantity generated from N$(0, 0.001)$ constrained $p_{i1} > 0$, so that the individuals within a group have similar ancestry background. Genotypes -1, 0 or 1 at marker l for 10 individuals within each of the four groups are generated according to a multivariate binomial distribution with marginal probabilities being $(1 - p_{il})^2$, $2p_{il}(1 - p_{il})$, p_{il}^2 (where p_{il} is estimated by equation 4.6) and the correlation between

any pair of individuals being ρ (Leisch et al., 1998) as specified above. Next, we simulated 500 additive genetic loci (can easily extended to other genetic models) using the simulated null SNP genotype data. We randomly selected 500 out of the simulated 1000 null SNPs, let x_{il} be the genotype score of individual i at marker l, let $\mu_0 = 2.0$, failure time T is simulated from a Weibull distribution with $T_{il} \sim$ Weibull(α, λ) i.e. $f(T_{il}) = \frac{\lambda}{\alpha}(\frac{T_{il}}{\lambda})^{\alpha-1} \exp^{-(T_{il}/\lambda)^{\alpha}}$, where $\alpha = 2$, $\lambda = 6\exp(\mu_{0i} + \mu_1 x_{il})$ for $i = 1, \ldots, 150$, $l = 1, \ldots, 500$, where $\mu_{0i} = \mu_0 p_{i1} + \gamma_i$, and γ_i was generated according to $\gamma' = (\gamma_1, \ldots, \gamma_n) \sim$ CAR (τ) distribution with $(B - W)\tau$ being the precision matrix; $W_{(i,j)} = 2K_{(i,j)}$ for $i \neq j$ and 0 otherwise; B being a diagonal with $B_{(i,i)} = \sum_{j=1}^{n} W_{(i,j)}$. Thus, our simulation induces both terms, μ_{0i} and x_{il}, as a function of the ancestry background p_{i1} and K, genetic kinship among the individuals. We set $\mu_1 = 0$ and 0.38 for comparing the type I error rate and power, respectively. Uncensored data ($t_{il} = T_{il}$) were considered for 120 individuals. The survival time for randomly selected 30 censored individuals are $t_{il} = \min(T_{il}, 3.759)$ at $\mu_1 = 0$ and $t_{il} = \min(T_{il}, 2.441)$ at $\mu_1 = 0.38$, respectively. The censoring time $C = 3.759$ under null and $C = 2.441$ under alternative are chosen such that $T_{il} > C$ for all 30 censored individuals at $l = 1, \ldots, 500$.

First, we checked whether the estimated kinship coefficients accurately reflect the genetic correlations among individuals. As shown in table 2, the estimated relative

kinship coefficients well reflect the true kinship coefficients. The Pearson correlation coefficients are 0.995, 0.990, 0.989 between the first, the second, and the third largest principal components of the true kinship matrix K and the estimated kinship matrix $\hat{K}$, respectively. This observation again suggests that the estimated kinship coefficients reflect well the true kinship coefficients among the individuals. Secondly, we checked whether the inferred population structure variables $x_2, \ldots, x_p$ accurately reflects K. If $x_2, \ldots, x_p$ fully captures K, then in such case we do not need to include frailties attributable to kinship in our models and population structure variables alone will be enough for adjusting for population stratification. We found that correlation between x_2 i.e. the largest principal component of SNP data, and the largest principal component of K is $r = 0.84$ only, which is much smaller than 0.995, the correlation coefficient between the largest principal components of K and $\hat{K}$. The ordinary linear regression with the largest principal component of K as the response variable and x_2 as the covariate yields a R-square of 0.701, and the R-square increases from 0.701 to 0.952 as the number of the the principal components of SNPs as covariates in the model increases from 1 to 5. However, the R-square only increases from 0.952 to 0.955 as the number of the principal components increases from 5 to 30. The above observations indicate that for data with diverse levels of familial relatedness within subpopulations, the inferred population structure variables alone will

not fully reflect genetic correlations between individuals i.e. we need both population structure variables and the kinship matrix to adequately adjust for continuous populations stratification. We further checked whether inferred population structure variables accurately reflect the membership in population 1 versus population 2 for 150 individuals, and we found that the correlation between x_2 and the true ancestry coefficients, $p_1 = (p_{11}, \ldots, p_{1n})'$, is $r = 0.82$. This is expected, because SNPs genotypes are a function of both ancestry background, p_1, and kinship, K. The ordinary linear regression with the true ancestry coefficients, p_1, as the response variable and the first five largest principal components of SNP marker data as covariates yields a R-square of 0.975 i.e. we need at least five largest principal components of SNPs data to fully reflect the individuals's ancestry background even though the sample is from two ancestry populations. This shows continuous population stratification of the sampled individuals. Finally, we examined whether the inferred population structure variables accurately reflects ancestry for a sample having subpopulations but no diverse levels of genetic relatedness within subpopulations, since this is a common assumption in population based association mapping studies. To do so, we estimated the principal components of 1000 SNPs at 110 individuals of group 1 only, who are genetically independent. The correlation between the first principal component of the SNP data and their true ancestry coefficients, p_1, is $r = 0.992$, which is similar to what has been found from Price's simulation study. This indicates that

Table 4.2: Comparisons of the estimated kinship coefficients $2\hat{K}$ and the true kinship coefficients $2K$ in 5 groups of simulation study II

	group 1	group 2	group 3	group 4	group 5
range of $2\hat{K}$	(0.69, 0.79)	(0.58, 0.69)	(0.45, 0.52)	(0.31, 0.43)	(0, 0.198)
mean of $2\hat{K}$	0.7062	0.6095	0.4845	0.3721	0.008
$2K$	0.7000	0.6000	0.5000	0.4000	0.000

for data with population structure but no diverse levels of familial relatedness within subpopulations, population structure variables alone is adequate for correcting for population stratification. Note also that the first principal component is sufficient for the simulated data set of 110 genetically independent individuals.

When we used the method discussed in the method section, the population structure variables were given by the first five principal components of the SNP data for this simulation data. We applied the 3 models to each of the 500 data sets at $\mu_1 = 0$ and 500 data sets at $\mu_1 = 0.38$ with the same Bayesian implementation as in the analysis from the method section. As shown in table 3, the false positive rates are similar between the parametric model 2 and semiparametric model 3 with frailties, which are not significantly different from the expected value of 5% at $\alpha = 0.05$, 95%$\text{CI}_{p=5\%}$ is (3.09%, 6.91%) or 1% at $\alpha = 0.01$ (95%$\text{CI}_{p=1\%}$ is (0.13%, 1.87%)). Note that the confidence interval for a given α from m independent tests are $\alpha \pm 1.96\sqrt{\alpha(1-\alpha)/m}$. The false positive rate is larger for the parametric model 1 without frailties compared to models 2 and 3 with the frailties attributable to kinship. At $\alpha = 0.05$, the type I

Table 4.3: Comparisons of the 3 model 3 in type I error rate, power, biases, and mean square error (MSE) of candidate gene regression coefficient $\hat{\beta}_1$ from simulation study

	Type I error rate		Power					
	α		α		$\mu_1 = 0$		$\mu_1 = 0.38$	
Model	0.05	0.01	0.05	0.01	Bias	MSE	Bias	MSE
Model 1	7.8	2.6	98.2	94.0	-0.0027	0.0093	0.0121	0.0075
Model 2	5.6	1.2	98.0	93.2	-0.0031	0.0091	0.0136	0.0073
Model 3	3.6	0.8	96.6	92.4	-0.0032	0.0092	0.0132	0.0074

For each of the 3 models, 500 simulated QTNs were used in estimating the power, type I error rate, MSE, and bias of the regression coefficients corresponding to the candidate gene.

error rate for model 1 is 7.8% which is significantly higher than the expected value of 5%, while at $\alpha = 0.01$, the type I error rate for model 1 is 2.6% which is also significant larger than the expected value of 1%. The false positive rates are 5.6% and 3.6% at $\alpha = 0.05$ for models 2 and 3, respectively, while the false positive rates are 1.2% and 0.8% at $\alpha = 0.01$ for model 2 and 3, respectively. The powers of the 3 models are similar. Averaged across 500 simulated additive genetic markers, the mean square error for the candidate gene regression coefficients is the smallest for the parametric frailty model 2 and the largest for model 1 without the frailties. The three models also provide a similar bias for the candidate gene regression coefficients.

4.7 Conclusion and discussion

In this paper, we have developed a parametric frailty survival model and a highly flexible semiparametric frailty survival model in structured populations for time-to-event data that incorporate frailties owing to genetic kinship among individuals. Compared to the existing method, the parametric model without frailties, both proposed models demonstrated a substantial lower type I error rate for the real data sets as well as the simulated data. The advantage of the approaches proposed here is that both the population structure variables $X = (x_2, \ldots, x_p)'$ and the kinship matrix, K, are included in the model to adjust for continuous population stratification. It is expected that these two aspects used for uncovering population stratification may overlap. Hence it is theoretically superior to orthogonalize X and K in order to reduce multicollinearity between the two components uncovering the population stratification, in particular when the correlation between the two aspects is high. However, this will cost 10 times more in computational time if we restrict the frailty terms attributable to K in the space orthogonal to the column space of X. In our real simulation data as well as real data, we found that the Pearson correlation coefficient between the first principal component of K and the first component of X is $0.6 \sim 0.75$, which is moderate. For the two reasons stated above, our approach does not orthogonalize the two sets of effects. Secondly, we modeled the effects attributable

to kinship via a Gaussian conditional autoregressive (CAR) model, therefore our approaches relax the strict positive definite assumption required in both Yu et al. and Li et al. approaches. Consequently, the significant advantage of our models over the approaches by Yu et al. and Li et al. is that our approach can be applied to a broader range of data sets than their approaches. We have also applied both Li et al. and Yu et al. approaches to the real data, the sample trace plots indicated that most of the model parameters did not converge (Figure 4.2), and centering covariates values did not help the convergence. In contrast, our proposed models 2 and 3 as well as model 1 (does not show in Figure) showed an excellent convergence properties for all model parameters. One explanation for the convergence difference in model parameters between Li et al. and Yu et al. approaches and our proposals is that CAR prior in our proposed models can borrow strength from it's neighbors and smoothes over frailties among neighboring subjects, hence it reduces the impact of potential identifiability or weak identifiability problems for frailties. Thirdly, we provided both parametric frailty and semiparametric frailty models, the baseline function in our semiparametric model is flexibly modeled as a mixture of Polya trees centered around a family of Weibull distributions. Although common parametric assumptions are typically made for computational convenience, the MPT model provides a powerful method to relax these strict parametric assumptions in survival and reliability modeling. To

study the robustness of our MPT prior specification to the choice of baseline centering survival function, we replaced our Weibull distribution with the log-logistic function in our real data analysis. The results show that the model appears to be robust with the respect to the baseline function specification. Hanson and Johnson (2002) also found the choice of underlying family to make little difference in density estimation with a MPT prior. The number of tree levels was capped at $M = 4$ in our model, achieving good MCMC mixing and allowing the candidate gene mapping in a reasonable amount of computer time for both real data set as well as the simulated data. Adding levels to the tree allows the MPT to accommodate greater detail, but can also slow MCMC mixing (Hanson, 2006) and greatly increases computational burden. In our algorithm, we updated the components $(\mathcal{Y}_{\epsilon_0}, \mathcal{Y}_{\epsilon_1})$ one at a time by sampling candidates $(\mathcal{Y}^*_{\epsilon_0}, \mathcal{Y}^*_{\epsilon_1})$ from a Dirichlet $(m\mathcal{Y}_{\epsilon_0}, m\mathcal{Y}_{\epsilon_1})$ distribution. Typically $m = 20 \sim 30$ and the larger m usually achieves a better MCMC mixing, we found that $m = 30$ achieves an excellent MCMC mixing. Fourthly, the limitation of the semiparametric model is the computational expense incurred due to the use of MCMC in the inference for association stage. Our algorithm is more suited for candidate gene mapping, however it is still possible for genome-wide scans if parallel computing is used. We found ample reason to prefer the MPT semiparametric model for our data to offset the increased computational complexity. Compared to the two

parametric models, the semiparametric method demonstrates an advantage. An application to the real data set of *Arabidopsis* lines for the flowering time demonstrated that the semiparametric model had a lower type I error rate and showed a better fit for the data than the 2 parametric models. Moreover, our simulation results show that the semiparametric model demonstrated a similar power, false type I error, and accuracy as the parametric model with the frailties, even though the data were simulated from parametric Weibull distributions. Finally, using a simulation study, we here addressed for the first time in the literature whether both the population structure variables and kinship matrix are needed to adjust for populations stratification when sampled individuals appear to have complex population stratification i.e. diverse levels of familial relatedness within subpopulations.

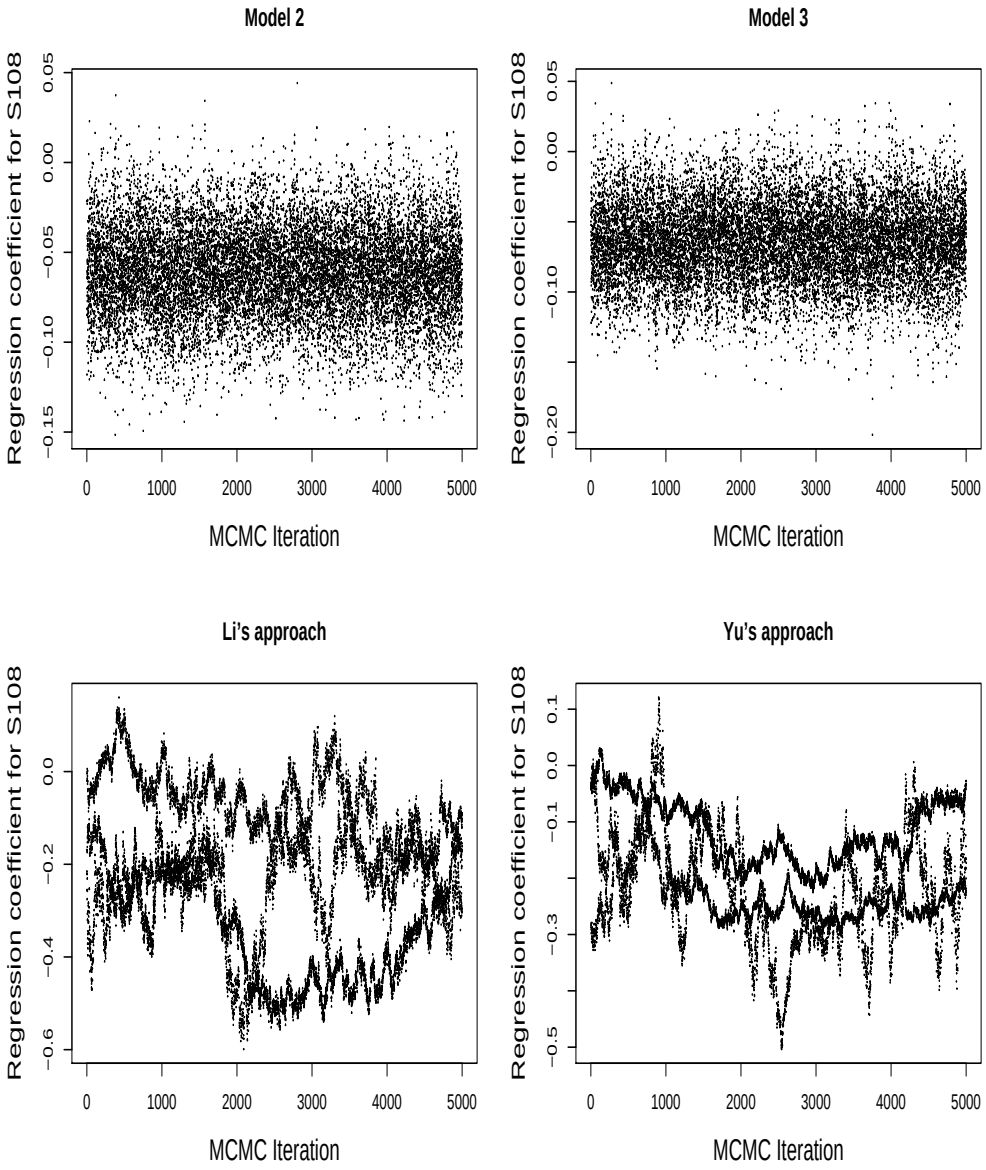

Figure 4.2: Trace plots for the regression coefficient corresponding to SNP, S107 for model 2, model 3, Li et al. model, and Yu et al. model from 3 MCMC chains. For each of 3 chains, there are 5,000 Monte Carlo samples from 200,000 iterations after 100,000 iterations burn-in.

Chapter 5

Multivariate Statistical Models for Association Mapping in Structured Populations

5.1 Abstract

Population-based Linkage-Disequilibrium (LD) mapping permits finer-scale mapping than linkage analysis. However, a population-based association mapping is subject to false positives that arise from population structure and kinship between the samples. While there is interest in simultaneously testing the association between a candidate gene and multiple phenotypes of interest, the current available association mapping methods are limited to univariate traits only. We here present a new method for population-based multi-trait association mapping via Bayesian conditional autoregressive modeling (CAR). The method we developed accounts for population structure and complex relationships between the samples. The method provides a more powerful alternative for association mapping of correlated traits than the currently available methods, which is illustrated via the application of our method to the previously published association mapping for 4 types of *Arabidopsis thaliana* flowering data and simulated data as well.

5.2 Introduction

With the recent improvements in genotyping and DNA sequencing technology that now allow simultaneous genotyping and sequencing of hundreds of thousands of polymorphisms, association mapping (also known as linkage disequilibrium (LD) mapping) is emerging as a powerful, general tool for identifying loci involved in the inheritance of complex traits. Because the allelic association due to linkage disequilibrium usually operates over shorter genetic distances, association mapping permits higher resolution mapping than does linkage analysis (Buckler et al., 2002; Flint-Garcia et al., 2003; Hästbacka et al.,1992). However, a potentially serious obstacle to population based association mapping is that the presence of population structure can result in false positives or spurious associations, that is, associations between phenotypes and markers that are not linked to any causative loci (Lander et al., 1994; Pritchard et al., 2000a; Pritchard et al., 2000b). It is likely to be worse for association mapping when familial relatedness within a subpopulation exists due to recent coancestry (Marchini et al., 2004; Weiss et al., 2000; Yu et al., 2006; Zondervan et al., 2004). Statistical methods have been developed to control for the false positives caused by this population structure. Genomic Control (GC) (Devlin et al., 1999; Reich and Goldstein, 2001) is based on the assumption that the population structure will induce over-dispersion that will affect the test statistics for association. The amount

of over-dispersion can be estimated from the empirical values for test statistics, after which the empirical test statistics can be adjusted on the basis of the estimated over-dispersion. One limitation of this approach is that it assumes a constant effect of stratification or admixture over all loci and thus does not correct appropriately for markers located in regions of adaptive selection (i.e. loci where selection acted or is acting differently on different populations) (Pritchard and Donnelly, 2001). Structured association (SA) proposed by Pritchard et al. (2000a, 2000b) is a popular approach for population structure inference, which has been used both on humans (Kim, et al., 2005; Li, et al., 2006; Rosenberg, et al., 2002; Serre et al., 2008) and other species (Aranzana et al., 2005; Breseghello and Sorrells, 2006; Kuroda et al., 2006; Manel et al, 2004; Pearse et al., 2006; Rosenberg, et al., 2001; Thornsberry et al., 2001; Tommasini1 et al., 2007). This approach uses a model-based clustering method to infer number of subpopulations (K) and population structure matrix (Q) using data at random genetic markers under the assumption of Hardy-Weinberg equilibrium (HWE) and linkage equilibrium within subpopulations. The log posterior probability given the data is computed for each K, $K = 1, 2, \ldots, C$ (C is some finite positive integer determined by users), and the algorithm selects the K which corresponds to the largest log posterior probability conditioned on data. The approach is computationally intensive and it can be impractical on large data sets

(Zhao et al., 2007). Price et al. (2006) recently developed EIGENSTRAT which corrects for population stratification by using principle component analysis (PCA) on a set of genetic markers. The PCA based approach proposed by Price et al. performs similar analyses as the SA method proposed by Pritchard et al. Since the PCA based method does not incorporate any defined genetic model, with only a few hundreds of markers, the PCA based method is more variable than the SA method by Pritchard et al., hence it is not as efficient as the SA method. On the other hand, the PCA based method is computationally more efficient for genome-wide or other large data sets (Serre et al., 2008).

Although these methods have proven useful in a variety of contexts, they have limitations. For example, for samples having population structure along with diverse levels of familial relatedness within subpopulations, GC, SA or PCA approaches which account only for population structure, may lead to either inadequate control for false positives or a loss in power owing to kinship within subpopulations (Malosetti et al., 2007; Yu et al., 2006; Zhao et al., 2007). Yu et al. (2006) recently introduced a unified frequentist mixed-model approach for association mapping that accounts for multiple levels of relatedness in the sample. In this method, random markers are used to estimate both Q and kinship matrix (K), which are then fit into a mixed-model framework to test for marker-trait association on a trait-by-trait basis with the effects of a candidate gene and Q being fixed and subject specific effects

attributable to kniship being random, and the covariance matrix of subject specific random effects is $2K\sigma_G^2$ (here σ_G^2 is a scalar). In application to the real data set of Arabidopsis flowering time association mapping, Zhao et al. (2007) concluded that Yu's mixed model outperformed either the SA or PCA method. Analogous to Yu et al., Malosetti et al. (2007) also presented a mixed model framework that offered two ways of representing genotypic relations: by structuring the variance-covariance matrix of the genotypic effects using pedigree or marker information and by introducing a grouping factor to represent structured association. However, while the covariance matrix of random effects is required to be positive-definite, there is no guarantee that K is positive definite for any given data set. For example, the estimated kinship coefficients could be identical or perfectly linearly correlated for two subjects, thus both K and the covariance matrix, $2K\sigma_G^2$ are singular, hence they are not positive-definite. In these cases, the methods proposed by Yu et al. or Malosetti et al. will fail. One significant advantage of our models over the mixed model approach proposed by Yu et al. and Malosetti et al. is that K in our proposed models is not required to be positive definite, consequently, our approach can be applied to a more broad range of data sets than their approaches.

Gaussian conditional autoregressive (CAR) model (Besag, 1974) has found wide application for modeling spatial correlated random effects in Bayesian linear mixed model. An autoregressive model for a stochastic process x_t in a plane assumes

$p(x_t|x_1,\ldots,x_n) = p(x_t|x_{s_1},\ldots,x_{x_m})$ for some set of neighbors $(x_{s_1},\ldots,x_{x_m})$. In our paper, we referred the autoregressive process as the conditional distribution of subject i's specific effect θ_i given the rest $\theta_{j\neq i}$ depends only on its neighbors, and two subjects are defined to be neighbors if their pairwise kinship coefficient is positive, with larger values of the kinship coefficient be associated with j's closer to i than those farther away from i. We use multivariate Bayesian statistical models for association mapping with the traditional CAR framework. In our approach, the correlations between subject specific random effects are introduced through a CAR prior distributional specification with $2K$ being the adjacency matrix (or connectivity structure) connecting subjects and $(D-W)[\sigma_g^2]^{-1}$ as the precision matrix in the CAR prior, where $W_{(i,j)} = 2K_{(i,j)}$ for $i \neq j$ and 0 otherwise, and D is a diagonal matrix with $D_{(i,i)} = \sum_{j=1}^{n} W_{(i,j)}$. While in the mixed model by Yu et al., the precision matrix due to kinship is marginally modeled as $[2K\sigma_g^2]^{-1}$.

In this paper, we present a unified model framework that enables thorough investigation into the associations between multiple traits and candidate genes, accounting for familial correlation and residual variation that likely arise from unmeasured genetic confounders. We propose two different CAR models which we compare to the popular single trait SA model. In the first model, we assume that subject specific effect is not trait-specific i.e. the same CAR model parameters are used for all the traits of interest. The second model is a multivariate CAR following the suggestion

by Mardia (1988) with genetic correlation attributable to the kinship among subjects to model the variation of the subject effects (Besag et al.,1991; Mollie, 1996). We apply the proposed models to a previously published data set of single trait association mapping in *Arabidopsis thaliana* lines (or varieties). Four types of flowering times are the four response variables of interest for the association test. We will perform association tests for 2 known flowering genes, FLOWERING LOCUS C (FLC) and *FRIGIDA* (*FRI*) (Aranzana et al., 2005, Lee et al., 1995; Michaels et al., 1999; Shindo et al., 2005; Zhao et al., 2007). Both genes are known to be central regulators of flowering time for unvernalized Arabidopsis plants. It is also known that the effects of FLC and FRI are eliminated by vernalization (Aranzana et al., 2005; Lee et al., 1995; Sheldon et al., 1999; Shindo et al., 2005; Zhao et al., 2007), a procedure that accelerates flowering by a long period of cold temperatures, generally below 10°C but above 0°C. The FLC gene expression level at 4 weeks of growth for plants will be used as genetic score in FLC association test. There are three geneotypes at FRI for the 95 lines, where 1 represents the wild type lines, 2 represents the mutated lines with the deletion of *Ler* allele, and 3 represents the mutated lines with the deletion of the *Col* allele (Nordborg et al., 2005). We will also conduct a simulation study to demonstrate the power of the methods.

5.3 Multivariate statistical models

5.3.1 Model Introduction

Let $y_i = (y_{i1}, \ldots, y_{iq})^T$, $i = 1, \ldots, n$ denote the n response vectors of q different univariate trait values, and $y = (y_{11}, \ldots, y_{1q}, \ldots, y_{n1}, \ldots, y_{nq})^T$. Let $x_{ij1}, \ldots, x_{ij(P-1)}$ be the i^{th} row vector of Q estimated using STRUCTURE (Prithcard et al., 2000a), and x_{ijP} be the genetic variation measure (genotype or gene expression level) for individual i at a candidate gene for a given trait, let β_{jp} be the regression coefficient corresponding to x_{jp} for $j = 1, \ldots, q$ and $p = 1, \ldots, P$. For a candidate gene with L genotypes, there are $(L-1)$ orthogonal dummy variables corresponding to the candidate gene, and $\beta_{jP}^T = (\beta_{jP1}, \ldots, \beta_{jP(L-1)})$, $x_{ijP}\beta_{jP} = \sum_{l=1}^{L-1} x_{ijl}\beta_{jPl}$ for $i = 1, \ldots, n$, $j = 1, \ldots, q$. Without loss of generality, we assume L=2 in our discussion. Let $x_{ij}^T = (x_{ij1}, \ldots, x_{ijP})$ be a P-component column vector of predictors for the i^{th} subject and the j^{th} trait such that for subject i and $j = 1, \ldots, q$. The Q matrix has dimension of n rows and M columns, where M denotes the estimated number of subpopulations among the n subjects. The $(i, m)^{th}$ element of Q is the portion of the i^{th} individual's genome which originated in subpopulation m (thus $\sum_{m=1}^{M} Q_{i,m} = 1$ for $i = 1, \ldots, n$). This indicates that only $(M-1)$ of the M columns of Q are linearly independent. Let b_{ij} be the subject specific random effect for subject i and trait j, $\beta_0^T = (\beta_{10}, \ldots, \beta_{q0})$ be the q intercepts, and $\beta_j = (\beta_{j1}, \ldots, \beta_{jP})^T$ be the regression coefficients for the j^{th} trait corresponding to

P predictors. Let $\mu_{ij} = \beta_{j0} + x_{ij}^T \beta_j = \beta_{j0} + \sum_{k=1}^{(P-1)} \beta_{jk} x_{ijk} + \beta_{jP} x_{ijP}$ and $\theta_{ij} = \mu_{ij} + b_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, q$. Writing $\theta_i = (\theta_{i1}, \ldots, \theta_{iq})^T$, $\mu_i = (\mu_{i1}, \ldots, \mu_{iq})^T$, $b_i = (b_{i1}, \ldots, b_{iq})^T$, $\beta = (\beta_1, \ldots, \beta_q)^T$, let 0^T be a P-component row vector of 0's, and $X_i = \begin{pmatrix} x_{i1}^T & 0^T & \ldots & 0^T \\ & \vdots & \vdots & \\ 0^T & \ldots & 0^T & x_{iq}^T \end{pmatrix}$, then we can write $\theta_i = \mu_i + b_i = X_i \beta + b_i$, $i = 1, \ldots, n$. Note that the X_i are matrices of order $q \times Pq$ with the rank for $(\sum_{i=1}^n (X_i - \bar{X})^T (X_i - \bar{X})) = Pq$, where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and the rank of X_i is P given that P predictors are linearly independent. Conditional on $\theta = (\theta_1, \ldots, \theta_n)^T$, $y_{11}, , \ldots, y_{1q}, \ldots, y_{n1}, \ldots, y_{nq}$ are assumed to be independent with pdfs

$$p(y_{ij} | \theta_{ij}, \sigma_j^2) = \frac{1}{\sqrt{2\pi}\sigma_j} \mathrm{e}^{-\frac{1}{2\sigma_j^2}(y_{ij} - \theta_{ij})^2}$$

for $i = 1, \ldots, n$; $j = 1, \ldots, q$. Throughout this paper, we assume priors for β_0, β, and $\sigma^2 = (\sigma_1^2, \ldots, \sigma_q^2)^T$ to be mutually independent with $\beta_0 \sim$ Uniform (R^P), $\beta \sim$ Uniform (R^{Pq}), and $\sigma_j \sim$ Uniform (a, b) for $j = 1, \ldots, q$. In next three sections, we will propose multivariate statistical models. We will introduce various priors for the subject specific effect vector $b_i = (b_{i1}, \ldots, b_{iq})^T$.

5.3.2 Independent univariate SA model - SA

We first consider the case with the q independent univariate SA models i.e. we include the elements of Q as covariates to account for population stratification without subject specific effect and within response correlations in the association test i.e. $b_{ij} = 0$ for all i, j. In this model, θ_{ij} is modeled as $\theta_{ij} = \mu_{ij} = \beta_{j0} + x_{ij}^T\beta_j = \beta_{j0} + \sum_{k=1}^{(P-1)} \beta_{jk}x_{ijk} + \beta_{jP}x_{ijP}$ for $i = 1, \ldots, n$; $j = 1, \ldots, q$. Let I be the indicator function defined as $I_A(x) = 1$ if $x \in A$, 0 otherwise, the joint posterior with the prior specified in above section is

$$\begin{aligned}\pi(\theta, \sigma^2|y) &\propto \prod_{ij} p(y_{ij}|\theta_{ij}, \sigma_j^2) \prod_j I_{(a,b)}(\sigma_j) \\ &\propto \prod_{i,j} \frac{1}{\sigma_j} \exp\Big\{-\frac{1}{2\sigma_j^2}(y_{ij} - \beta_{j0} - \sum_{k=1}^{(P-1)} \beta_{jk}x_{ijk} - \beta_{jP}x_{ijP})^2\Big\} \prod_j I_{(a,b)}(\sigma_j).\end{aligned}$$

5.3.3 Univariate CAR model - UCAR

In this section and the next, we consider 2 different CAR models where we again include elements of Q as covariates, but now use $2K$ as the neighboring structure in CAR distribution to account for multiple levels of genetic relatedness in the sample. A kinship coefficient is often defined as the probability of identity by descent of the markers compared (Ritland, 1996), but estimators based on genetic markers actually estimate a "relative kinship", that can be defined as ratios of differences

of probabilities of identity by state (Hardy et al., 2002). For instance, the kinship coefficient between individuals i and j is defined as $k_{ij} = (A_{ij} - A_r)/(1 - A_r)$, where A_{ij} is the probability of identity by state for random markers from i and j, and A_r is the average probability of identity by state for markers coming from random individuals from the sample (Hardy et al., 2002). Note that with this definition, negative relative kinship coefficients naturally occur between some individuals. This is interpreted as meaning subjects that these are less related than random individuals (Hardy et al., 2002; Ritland, 1996). We set negative values of kinship coefficient to 0. It is also reasonable to assume that the kinship matrix K should be estimated after taking the effects of Q into account. First note that the kinship is defined as the ratios of differences of probabilities of identity by state. Hence, the kinship in our approach is a relative measure, which is adjusted for population stratification in some sense. Let x_{lcia} be an indicator variable ($x_{lcia} = 1$ if the allele on chromosome c at locus l for individual i is a, 0, otherwise), p_{la} be the frequency of allele a at locus l in the reference sample (i.e. the reference allele frequency), and $\sum_{c_j}$ stand for the sum over the homologous chromosomes of individual i. For codominant markers such as SNPs, an unadjusted pairwise kinship coefficient between subject i and j which is not corrected for population structured is computed as (Hardy et al., 2002; Ritland, 1996):

$$k_{ij} = \sum_{l} \Big(\Big(\sum_{a} \sum_{c_j} \sum_{c_i} (x_{lcia} x_{lcja} / p_{la}) / \sum_{c_j} \sum_{c_i} 1 \Big) - 1 \Big) / \sum_{l} (m_l - 1) \qquad (1)$$

where m_l is the number of different alleles found in the sample at locus l ($m_l = 2$ for a SNP). Let subject i and j have Q_{im} and Q_{jm} of the genome originating from the same subpopulation m for $m = 1, \ldots, M$, and let p_{lam} be the estimated frequency of allele a at locus l in ancestral subpopulation m for $m = 1, \ldots, M$ using the software STRUCTURE. If the subject i has $Q_{im_1} = 1$ and subject j has $Q_{jm_2} = 1$ for $m_1 \neq m_2$ i.e. they come from 2 different subpopulations, then the expected adjusted pairwise kinship between the subjects should be 0; and if subject i and j have Q_{im} and Q_{jm} genome originating from the same subpopulation m, then they share $\min(Q_{im}, Q_{jm})$ genome originating from subpopulation m. The pairwise kinship between subject i and j which is corrected for population structure can be estimated as:

$$k_{ij}^{adj} = \sum_{l} \Big(\Big(\sum_{a} \Big(\sum_{c_j} \sum_{c_i} \sum_{m=1}^{M} \min(Q_{im}, Q_{jm}) x_{lcia} x_{lcja} / 4 p_{lam} \Big) \Big) - 1 \Big) / \sum_{l} (m_l - 1) \qquad (2)$$

We call this the adjusted version of K. The marker allele frequencies in ancestral subpopulations were estimated by the software STRUCTURE under the assumptions of HWE and linkage equilibrium within subpopulations, if the assumptions are violated, the allele frequencies for each marker can be strongly biased. The accuracy of

the estimation is also dependent on the model convergence and identifiability. Therefore, the adjusted version of K is more complicated to estimate than the unadjusted version of K.

We will consider different CAR priors for vector b_i. We first consider the case when $b_i = \phi_i \mathbf{1}_q$ (here $\mathbf{1}_q$ is a q-column vector of 1's) for $i = 1, \ldots, n$. This amounts to the assumption that all the components of the vector b_i in a given subject are equal, i.e. the subject specific effect attributable to kinship is not trait type-specific. For $\phi = (\phi_1, \ldots, \phi_n)^T$, we consider the pairwise difference prior for ϕ_i, also known as the intrinsic autoregressive, with joint pdf

$$p(\phi) \propto (\sigma_\phi^2)^{-n/2} \exp\Big\{-\frac{1}{2\sigma_\phi^2} \sum_{i \neq j} w_{ij}(\phi_i - \phi_j)^2\Big\},$$

where w_{ij} are the known weights with $w_{ij} = w_{ji} = 2k_{ij}$ for $i \neq j$ and 0 otherwise, k_{ij} is the kinship coefficient between i^{th} and j^{th} subjects, and σ_ϕ^2 is the genetic variance attributable to kinship. Note that $p(\phi)$ is improper, therefore the constraint $\sum_{i=1}^n \phi_i = 0$ will be imposed to ensure the propriety of the full conditionals, $p(\phi_i|\phi_j, j \neq i)$ (Banerjee et al., 2004). Such priors were introduced and have been used quite extensively in the spatial statistics literature (Besag et al., 1995). Here, we model the θ_{ij} as $\theta_{ij} = \mu_{ij} + \phi_i = \beta_{j0} + \sum_{k=1}^{(P-1)} \beta_{jk} x_{ijk} + \beta_{jP} x_{ijP} + \phi_i$ ($j = 1, \ldots, q$; $i = 1, \ldots, n$). It is assumed that priors for β_0, β, σ^2, and σ_ϕ^2 are mutually independent with $\sigma_\phi \sim$ Uniform (c, d), other priors are specified as in the section 2.1. The

joint posterior under the given set of priors is

$$\begin{aligned}\pi(\theta,\phi,\sigma^2,\sigma_\phi^2|y) &\propto \prod_{i,j} p(y_{ij}|\theta_{ij},\sigma_j^2)\prod_j I_{(a,b)}(\sigma_j)(\sigma_\phi^2)^{-n/2}\\ &\quad\times\exp\Big\{-\frac{1}{2\sigma_\phi^2}\sum_{1\le i\le l\le n} w_{il}(\phi_i-\phi_l)^2\Big\}I_{(c,d)}(\sigma_\phi)\\ &\propto \prod_{i,j}\frac{1}{\sigma_j}\exp\Big\{-\frac{1}{2\sigma_j^2}(y_{ij}-\beta_{j0}-\sum_{k=1}^{(P-1)}\beta_{jk}x_{ijk}-\beta_{jP}x_{ijP}-\phi_i)^2\Big\}\\ &\quad\times\prod_j I_{(a,b)}(\sigma_j)(\sigma_\phi^2)^{-n/2}\exp\Big\{-\frac{1}{2\sigma_\phi^2}\sum_{1\le i\le l\le n} w_{il}(\phi_i-\phi_l)^2\Big\}I_{(c,d)}(\sigma_\phi).\end{aligned}$$

5.3.4 Multivariate CAR model - MCAR

In this section, we consider a special case of the multivariate CAR (MCAR) model first introduced by Mardia (1988), namely the multivariate intrinsically autoregressive (MIAR) distribution (Besag et al., 1991). Let the matrix W reflect the connectivity structure between subjects with its $(i,j)^{th}$ element being w_{ij} as defined in the previous subsection; let D=Diag$(m_1,\ldots,m_n)$ with $m_i=\sum_{j=1}^n w_{ij}$; and let Λ^{-1} be the additive genetic covariance between the different types of traits given the relative kinship among subjects. In the quantitative genetics, the matrix of additive genetic variances and covariances can be expressed as $2KG$, here K is the kinship matrix among the subjects. In our model, G is modeled as Λ^{-1}. Under this framework, conditional on V, the precision matrix of subject specific effects is given by $V^{-1}=(D-W)\otimes\Lambda$ (here $\otimes$ is the Kronecker product). Let $\phi_i=(\phi_{i1},\ldots,\phi_{iq})$ for

$i = 1, \ldots, n$, and define the vector, $\phi^T = (\phi_1, \ldots, \phi_n)$. As with the UCAR model, due to the singularity of $(D - W)$, we will add q centering constrains $\sum_{i=1}^{n} \phi_{ij} = 0$ for $j = 1, \ldots, q$ to ensure the propriety of the full conditionals $p(\phi_i|\phi_j, j \neq i)$ (Banerjee et al., 2004). Following Mardia, under the zero-centered MCAR sets, the ϕ has prior with pdf $p(\phi) \propto |\Lambda|^{n/2} \exp\left(-\frac{1}{2}\phi^T V^{-1}\phi\right)$. Under this model, the θ_{ij} is modeled as $\theta_{ij} = \mu_{ij} + \phi_{ij} = \beta_{j0} + \sum_{k=1}^{(P-1)} \beta_{jk} x_{ijk} + \beta_{jP} x_{ijP} + \phi_{ij}$ (in the previous section $\phi_{ij} = \phi_i$ for all j) for $j = 1, \ldots, q$; $i = 1, \ldots, n$. We assume a Wishart $(m, \boldsymbol{A})$ prior for Λ. Other prior specifications remain the same as in the previous subsection. It is assumed that β, σ^2, and Λ are mutually independent in their prior distributions. Then the joint posterior is given by

$$
\begin{aligned}
\pi(\theta, \phi, \sigma^2, \Lambda|y) &\propto \prod_{i,j} p(y_{ij}|\theta_{ij}, \sigma_j^2) \prod_j I_{(a,b)}(\sigma_j) \\
&\quad \times |\Lambda|^{n/2} \exp\Big\{-\frac{1}{2}\phi^T V^{-1}\phi\Big\} \times |\Lambda|^{(m-q-1)/2} \exp\Big\{-\frac{1}{2}\mathrm{tr}(\Lambda \boldsymbol{A})\Big\} \\
&\propto \prod_{i,j} \frac{1}{\sigma_j} \exp\Big\{-\frac{1}{2\sigma_j^2}(y_{ij} - \beta_{j0} - \sum_{k=1}^{(P-1)} \beta_{jk} x_{ijk} - \beta_{jP} x_{ijP} - \phi_{ij})^2\Big\} \prod_j \\
&\quad \times I_{(a,b)}(\sigma_j) |\Lambda|^{n/2} \exp\Big\{-\frac{1}{2}\phi^T V^{-1}\phi\Big\} \times |\Lambda|^{(m-q-1)/2} \exp\Big\{-\frac{1}{2}\mathrm{tr}(\Lambda \boldsymbol{A})\Big\}.
\end{aligned}
$$

5.4 Model choice

Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC) as a tool to compare complex models. DIC accounts for the fit and the complexity of models

in which the number of parameters is not obviously specified. The fit and complexity of a hierarchical model are shown by $\bar{D}$, the posterior mean of the deviance and the effective number of parameters, P_D, respectively. The DIC is defined as

$$DIC = \bar{D} + P_D = 2E_{\theta|y}(D) - D(E_{\theta|y}(\theta)).$$

DIC can be obtained at the end of MCMC sampling by monitoring θ and $D(\theta)$ during the simulation, and smaller values of DIC are interpreted as more desirable. We will use DIC to select the best model.

5.5 Application to the *Arabidopsis thaliana* data set

Arabidopsis (Arabidopsis thaliana) lines provide an excellent resource to dissect the molecular basis of phenotypic variation such as flowering time. The *Arabidopsis thaliana* lines used in this study have been described in previously published univariate association mapping papers (Aranzana et al., 2005;, Nordborg et al., 2005; Zhao et al., 2007). For this specific example, subjects are the 95 Arabidopsis lines. Although our methods can handle missing data, the DIC in WinBUGS, the software used for Bayesian inference below, is not be correct in the presence of missing observations, hence, we eliminated 5 of the 95 lines in our analysis. For inferring population structure and kinship among the lines, we used the publicly available sequencing data of Nordborg et al. (2005), which included a total of 876 sequence

alignments, representing 0.48 Mbp of the *Arabidopsis thaliana* genome. These sequence fragment-haplotypes were used as random genetic markers for estimation of population structure i.e. a Q matrix. Using the software STRUCTURE v. 2.0, Nordborg et al. (2005) estimated Q with $M = 8$ for the 95 Arabidopsis lines. The kinship matrix K was estimated using the software package SPAGeDi (Hardy et al., 2002) on the basis of 5000 high quality single nucleotide polymorphisms (SNPs) obtained from the 876 sequence fragments. We conducted an association test for the FLC and the FRI flowering time genes to compare model performance. Four types of flowering time, measured in days from germination until the first opening of flowers at the University of Southern California as described in Zhao et al. (2007) were included as the 4 response variables in this analysis. They are (1) long days (16 hour(h) light/8 h dark) without vernalization (LD); (2) short days (8h light/16 h dark) without vernalization (SD); (3) long days (16 h light/8 h dark) with 5 week vernalization (LDV); (4) short days (8h light/16 h dark) with 5 week vernalization(SDV). Note that for LD and SD, there are a few lines were not flowered at the end of experiments and for those lines the total growing days were set to be the observed flowering time. The sample correlation coefficients, assuming measurements from different lines are independent, ranged from 0.66 to 0.89. We have eight covariates included in all three models: one covariate encodes flowering time and seven covariates represent the population structure for adjusting the genetic background among the lines.

Posterior inference is carried out in a Bayesian framework using Markov chain Monte Carlo (MCMC) in WinBUGS. For specification of the hyperparameters, we use uniform priors with $a = 0$ and $b = 200$ for σ and σ_j for $j = 1, \ldots, q$, and $c = 0$ and $d = 200$ for σ_ϕ. For the Wishart vague prior, we set $m = q = 4$ and $\boldsymbol{A}$ with diagonal elements 1000 and off-diagonal elements 5 specified for the prior of Λ.

We eliminated a burn-in of 30,000 iterations to allow the chains to reach their stationary distributions and then ran an additional 30,000 iterations for each of three independent chains. Initial values for regression coefficients and precision parameters were chosen systematically to represent reasonable values in the parameter space. Initial values for subject specific effects, b, are automatically generated from priors by WinBUGS. The model parameters are updated via the Gibbs sampler. The sample trace plots fail to indicate unacceptable convergence of chains. The density plots and Gelman and Rubin's R statistic suggest that the sample values of the MCMC runs should provide adequate posterior inference (Gelman et al., 1992).

FLC and FRI are known flowering loci. We fit all three models to the data and performed model comparison to select the best one using DIC. For two CAR models, we also compared the DICs and the posterior summary statistics for the regression coefficients between the adjusted K and unadjusted K. The results from the association tests between the 4 flowering times and 2 genes (FLC and FRI) show that difference between the adjusted K and unadjusted K are minor. For

Table 5.1: DIC and effective number of parameters P_D for 3 models for FLC or FRI gene and flowering times multivariate association mapping

Model	Covariate	Subject Specific Effect	$\bar{D}$	P_D	DIC
FLC association test					
SA	Q+FLC	no subject specific effect	3210.53	39.70	3250.23
UCAR	Q+FLC	CAR	3076.49	94.72	3171.21
MCAR	Q+FLC	MCAR	3085.46	96.41	3181.87
FRI association test					
SA	Q+FRI	no subject specific effect	3189.75	43.65	3233.40
UCAR	Q+FRI	CAR	3066.94	100.17	3166.11
MCAR	Q+FRI	MCAR	3072.05	102.44	3174.49

Q are covariates representing population structure among lines. The UCAR model is the model with same CAR model parameters for all 4 traits.

example, for the FLC and the 4 trait association tests, under the CAR model, the DIC was 3171.21 for the unadjusted K and 3170.5 for the adjusted K, while under the MCAR model, the DIC was 3181.87 for the unadjusted K and 3175.3 for the adjusted K. For LD and FLC association test, the 95% credible sets using the unadjusted K versus the adjusted K are (4.29, 18.27) vs (4.31, 18.22) for the CAR model, and (4.64, 18.81) against (4.79, 18.66) for the MCAR model, respectively. So we concluded that both the DIC and posterior summaries of regression coefficients are similar using the two versions of K for the two CAR models. Hence our discussion hereafter uses the unadjusted K, which is much simpler in practice. The DICs of the 3 models from the FLC association tests are presented in table 5.1. The DIC and P_D values differ a great deal among the models. It is noted that the SA model has

a higher DIC than any proposed CAR model. The differences between the UCAR model and the MCAR model are so small that they could easily be due to Monte Carlo error, and therefore we consider these two models to be tied in terms of DIC. Similar results hold for the FRI association tests (table 5.1). Therefore, we select the UCAR model due to its greater simplicity.

We presented the 95% credible sets and the median of the regression coefficients corresponding to the flowering genes, FLC and the genetic effects due to the Col and Ler haplotypes of FRI in table 5.2 for all 3 models. Noted here, among the 4 traits we tested, LD, SD are flowering times for plants without vernalization, while LDV and SDV are flowering times for vernalized plants. After accounting for the population structure and the genetic correlation due to kinship, the 95% credible sets for the regression coefficients for the flowering gene FLC on LD and SD did not cover 0 for all 3 models, as expected, thereby confirming the effect of FLC on LD, SD, all types of flowering time for non-vernalization plants. For LDV and SDV, flowering times for vernalized plants, all 3 models have the 95% credible set for the regression coefficient for the FLC gene covering 0, again as expected.

Also listed in table 5.2 are the effect (δ_1) of the *Ler* deletion haplotype and the effect (δ_2) of the *col* deletion haplotype. The *Ler* deletion haplotype is the most common nonfunctional FRI allele and a major determinant for flowering time in the absence of vernalization (Aranzana et al., 2005; Lee et al., 1995; Sheldon et al., 1999;

Shindo et al., 2005; Zhao et al., 2007). Here δ_1 and δ_2 are estimated as the mean difference in flowering time between the wild type group and the *Ler* deletion group, the wild type group and the *col* deletion group, respectively. As expected, for all 3 models, the wild type group has a significantly larger mean flowering time in the absence of vernalization than both *Ler* deletion group and *Col* deletion group, thereby indicating all 3 models successfully detected the true FRI and traits association. For flowering time in the presence of vernalization with short day light i.e. SDV, all 3 models are correctly detecting no effects for both *Ler* and *Col* haplotypes. However, for the flowering time in the presence of vernalization with long day, LDV, while the 95% credible sets of δ_1 inappropriately excludes 0 under the SA model, the credible sets for δ_1 from the UCAR and MCAR models cover 0 which correctly suggested that there is no evidence of association between flowering time and the *Ler* deletion status. For *Col* haplotype, all 3 models are inappropriately detecting an effect on vernalized LDV. Notice also that the 95% credible sets for the regression coefficient for FLC and the 95% credible set for δ_1 and δ_2 are considerably larger for the SA model than the both proposed CAR models under all cases considered except that the MCAR model has larger 95% credible set for δ_2 at LDV than the SA model. Particularly, the UCAR model consistently has the smallest 95% credible sets for all 4 traits among the 3 models.

Figure 5.1 is a scatter-plot of the raw data for LDV vs the fitted LDV values

(solid points). The vertical bars in the scatter plot in Figure 5.1 also demonstrate that the 95% credible sets of the fitted values are narrower for the UCAR and MCAR models compared to the SA model in general, and a total of 31 raw data points are outside of the 95% credible sets of the fitted values in model 1 compared to only 2 data points for the UCAR model and 1 data point for the MCAR model. These observations again indicate that the UCAR and MCAR models give a better fit than the SA model.

Maps of the raw data of SD and the posterior medians of the fitted SD values for the FRI association mapping are shown in Figure 5.2, where the axes are the two-dimensional multidimensional scaling representation of the inverse of kinship matrix among 90 lines. The map shows a slightly different pattern of smoothing among the 3 models, and the two CAR models showed slightly more smoothing than the SA model.

Sensitivity analyses with respect to several prior specifications has been conducted for the UCAR model and the results appear to be robust. The posterior distribution of the regression coefficients from case deletion data sets revealed that no observations were influential points.

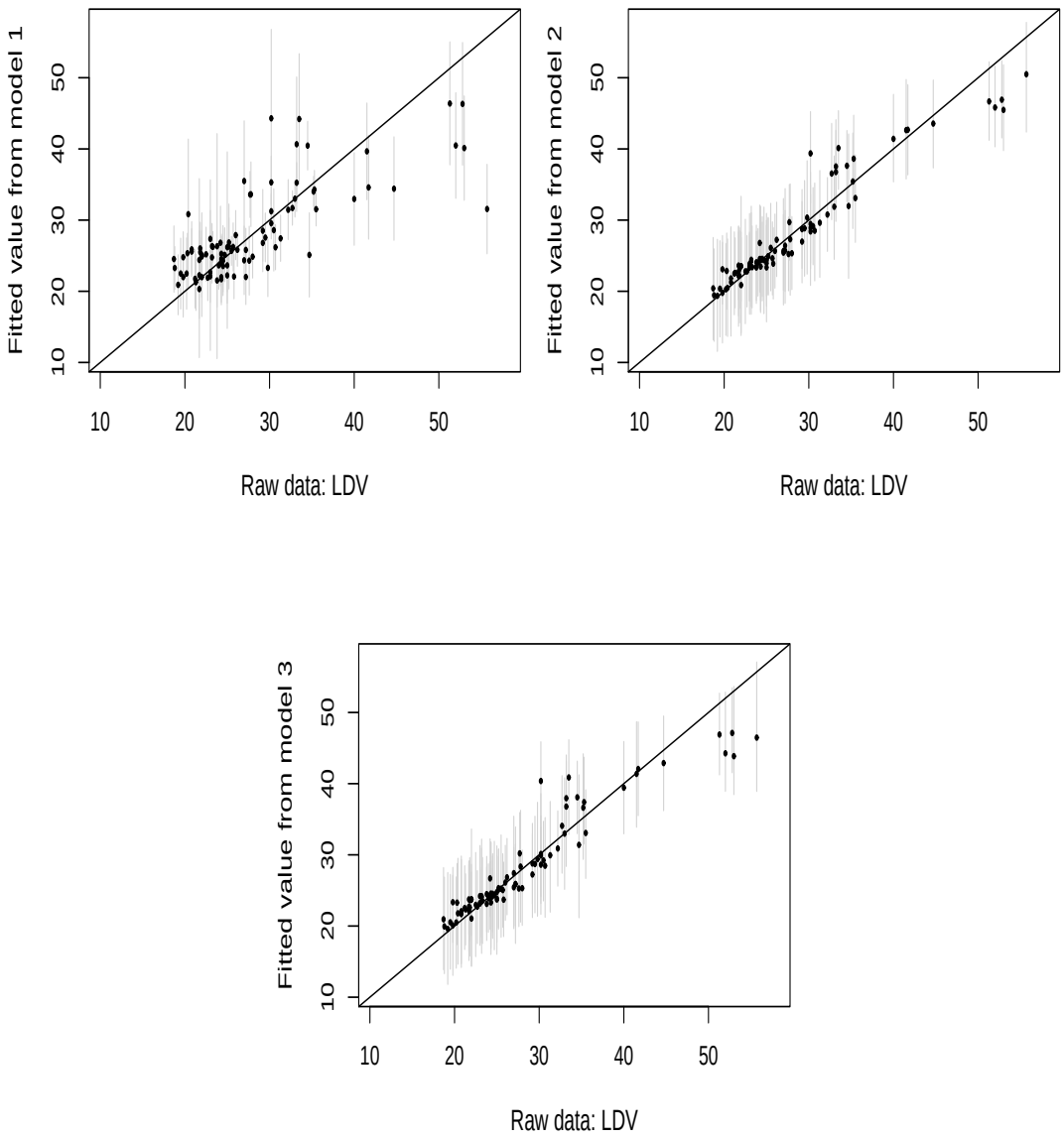

Figure 5.1: Raw data of LDV vs the fitted LDV values (solid point) and the vertical bar in grey are the 95% credible sets of the fitted LDV values from the 3 models.

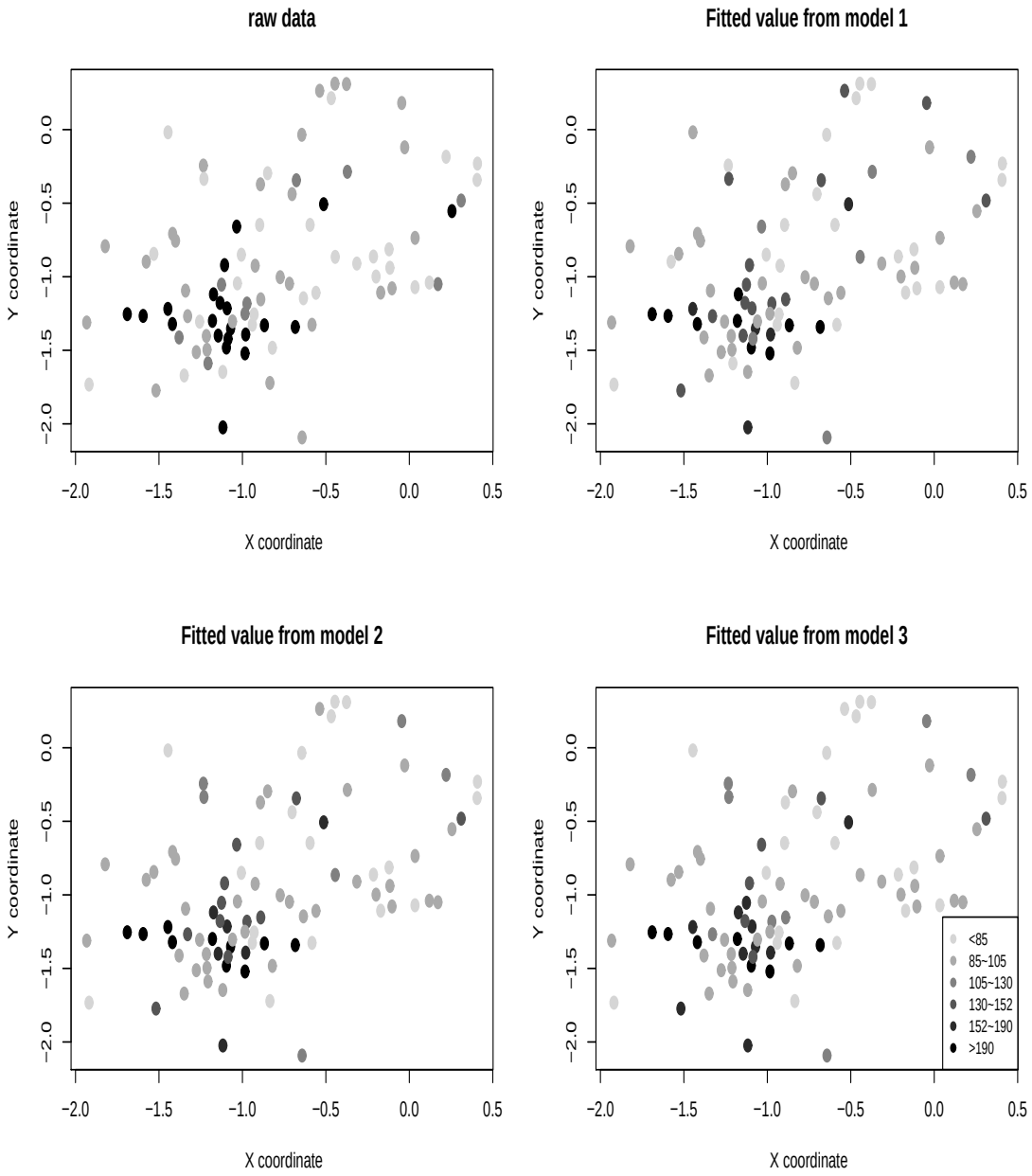

Figure 5.2: Maps of raw data of SD (panel 1) and the fitted SD values from the 3 models (panels 2,3,4)for the FRI association mapping, where axes are the two-dimensional multidimensional scaling representation of the inverse of kinship matrix.

Table 5.2: Posterior summaries of the regression coefficients for FLC and genetic effects *Ler* and *col* for deletion alleles of FRI from multivariate association mapping by the three models

	FLC			**FRI**					
	β_p			δ_1			δ_2		
Traits	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
SA MODEL									
LD	3.9	11.4	19.4	16.6	38.5	60.4	38.1	66.8	95.8
SD	2.7	9.5	16.3	12.0	31.4	50.9	28.2	53.6	79.0
SDV	-6.5	-0.6	5.3	-4.1	13.5	31.2	-0.4	22.9	46.1
LDV	-0.9	0.5	1.8	0.5	4.2	7.4	5.9	10.8	15.7
UCAR MODEL									
LD	4.3	11.3	18.3	15.6	35.8	56.0	36.5	63.3	90.4
SD	3.1	9.1	15.2	10.8	28.9	46.7	26.3	50.0	73.9
SDV	-6.4	-1.0	4.3	-5.9	10.8	27.4	-2.7	19.5	41.7
LDV	-1.0	0.1	1.2	-2.7	1.5	5.6	0.8	7.3	10.5
MCAR MODEL									
LD	4.6	11.8	18.8	13.3	34.4	55.7	33.6	61.7	89.8
SD	4.2	9.9	15.8	7.3	25.5	43.8	21.7	45.8	70.1
SDV	-6.8	-1.2	4.5	-6.1	11.8	29.7	-2.2	21.5	45.4
LDV	-0.9	0.2	1.4	-2.3	1.9	6.0	1.1	7.7	12.7

LD and SD are non-vernalized flowering times, LDV and SDV are vernalized flowering times. Here δ_1= the mean difference of flowering times between the wild type group and the *Ler* deletion group, and δ_2= the mean difference of flowering times between the wild group and the *Col* deletion group.

Table 5.3: The estimated power from the simulated data sets for the SA, UCAR, and MCAR models

Model	$\zeta = 0.25\%$	$\zeta = 0.5\%$	$\zeta = 1\%$	$\zeta = 2\%$	$\zeta = 4\%$	$\zeta = 6\%$
SA	8.2	13.7	22.8	43.7	62.9	81.0
UCAR	9.1	14.6	23.6	50.0	69.2	86.5
MCAR	8.2	15.4	23.6	48.1	67.3	84.6

For each simulated genetic effect, ζ, 91 simulated $QTNs$ were included in the power estimation for each of the 3 models.

5.6 Simulations to investigate relative power

Our goal in simulating data was to simulate a data set with the same covariate structure as the data from the previous section given a true complex genetic correlation structure among subjects. Underlying the phenotype simulation, we simulated 91 Arabidopsis flowering genetic loci as follow. We randomly select 91 SNPs that cover the whole *Arabidopsis thaliana* genome from the list of 17,000 SNPs obtained from the 876 sequence alignments (Nordborg et al., 2005). To ensure that more than 18 lines are in each of the 2 allele categories of a given SNP, we selected 91 SNPs that have a minor allele frequency greater than 20%. We also assumed that the selected SNPs are not genetic loci of the 4 types of flowering times, i.e. they are not associated with any of the 4 response variables, hence we selected the 91 SNPs such that their 95% credible sets for the regression coefficients from the 4 models covered 0. For each of the 4 response variables, a fixed additive genetic effect ranging from $\kappa = 0.1s$ to $1.05s$ was added to those selected SNPs, here $s{=}\sqrt{\hat{\sigma}^2}$, where $\hat{\sigma}^2$ is the usual unbiased estimate of the variance, assuming lines are independent. If we let p be the sample allele frequency of a SNP, and $n = 90$ be the sample size, then the percentage (ζ) of the total phenotypic variation explained by this fixed genetic effect can be approximated by $\zeta = p(1-p)\kappa^2/(p(1-p)\kappa^2 + 1 - 1/n)$ (Long et al., 1999; Yu et al., 2006). To each of the 91 SNPs and for each of 4 traits, we have 6 simulated data sets with the added genetic effect expressed as one of the values

$0.25\%, 0.5\%, 1\%, 2\%, 4\%, 6\%$. We denoted those simulated flowering genetic loci as quantitative trait nucleotides (QTN). We applied the 3 models to each of the 6 by 91 data sets with the same Bayesian implementation as in the analysis from section 3. For each of the 3 models and for each of the 6 fixed genetic effects, we calculated m, the number of QTNs whose 95% credible set for the regression coefficient excluded 0 for any of the 4 traits. The power of a given model is then estimated as $m/(4 \times 91)$. As shown in table 5.3, while the power steadily increased as the simulated genetic effect increased from 0.25% to 6% for all models, the UCAR and MCAR models showed consistently higher power than the SA model. Particularly, compared to the SA model, the proposed UCAR model showed a 4% to 14.6% increase in power, while the proposed MCAR model showed a 0% to 13.3% increase in power, thereby indicating the multivariate analysis using the UCAR and MCAR models increased the power to detect a QTN. Note that the 4 types of flowering times are controlled by the same genes except for the genes that are up-regulated or down-regulated by photoperiod and/or the vernalization process. This may partially explain why the UCAR model performs best among all 3 models.

5.7 Summary and conclusion

We presented a multivariate Bayesian statistical model with the traditional CAR framework for multi-trait association mapping in structured populations, where we

include the population structure as covariates and the kinship matrix as the neighboring structure in CAR distribution to systematically account for multiple levels of relatedness in the sample. In particular, we present a flexible unified model framework that enables thorough investigation into the associations between multiple traits and candidate gene(s), accounting for genetic correlation and residual variation that likely arise from unmeasured confounders. We conducted full Bayesian inference for the proposed models using a MCMC method. We also have shown that the 2 proposed CAR models have smaller DIC and higher power than the SA model. Particularly, the univariate CAR model with the same CAR parameters (UCAR) is shown to be a better association mapping method with greater statistical power and more accurate at localizing genes influencing the flowering times than the SA model for the current data set. In this paper, we have provided two priors for Bayesian multivariate models. While the UCAR model, a special case of the MCAR model, had the best performance for this data set, the MCAR model may be more appropriate for other data sets. The likely reason that the UCAR model performed better for this data sets is that the 4 types of flowering times are controlled by the same genes except for those are in the photoperiod pathway or regulated by vernalization. The biggest difference between the UCAR model and the MCAR model is that for the UCAR model, the genetic variance attributable to kinship is assumed to be identical for all the traits under study, while no such assumption is made for the MCAR model.

In our data set, the estimated genetic variances of the 4 types of flowering times from the MCAR model are almost identical, this again suggested that the UCAR model might be more appropriate for this specific data set. If the genetic variances (either by estimation or with prior knowledge) are different for different traits, the MCAR model will be a better model than the UCAR model in correctly estimating the genetic variances and hence the 95% credible sets of regression coefficients. Since the UCAR model is just a special case of the MCAR model, although the UCAR model will be more efficient for data with traits that share the same genetic variance attributable to kinship. One disadvantage for the MCAR model is that the posterior distribution of Λ is sensitive to its prior specification (Wishart distribution).

A second benefit of our approach is that our model is able to systematically account for multiple levels of relatedness among individuals. While it is intuitively obvious that K captures features of the data that could not possibly be captured by Q, it is not clear that Q should be required in addition to the kinship effect. Nonetheless, we found that K was not sufficient, as did Yu et al. in their association mapping study. Essentially, due to the genetic consequence of local adaptation or diversifying selection among the different populations, a few genes with relatively large phenotypic effects can be accounted for by Q in a gross manner, whereas genes with relatively small phenotypic effect can be accounted for by K on a finer scale. Thus, the two approaches for uncovering population structure are complementary.

Although the adjusted K may be a theoretically superior approach than the unadjusted approach, the unadjusted approach is much simpler in practice and the results from the application to the 4 flowering times and 2 genes study show that difference between the adjusted K and the unadjusted K have minor impact on aspects of the model that are of primary concern. Alternatively, you may orthogonalize population structure variables and K in order to reduce a possible overlap between the two components uncovering the population stratification (Li et al. 2008), in particular when the correlation between the population structure variables and K is high. However, this will cost 10 times more in computational time if we restrict the effects attributable to K in the space orthogonal to the column space of the population structure variables. In summary, our approach as well as Yu's mixed model works better if the subjects have multilevel relatedness, which in turn suggests that our approach is better then Yu's mixed model, because K will often not be positive definite in such data sets. For a random sample from a population under HWE, the expected relative kinship as used here will be 0 for any pair of individuals, while the kinship, the probability of identity by state (or descendent) of the markers compared will be positive for any pair of individuals. For samples under HWE within each subpopulations, then Q is sufficient for adjusting for population structure. However, if HWE does not hold within each subpopulation, for example, the K matrix (relative measure as we used in our analysis) has elements greater than 0.2 (for outbreeds such

as human, note $k = 0.25$ for siblings) or greater than 0.4 (for inbreeds such as wheat) between 2 distinct subjects (or lines), then there is an indication of multilevel-genetic correlation among the sample. In this case, Q alone will make the model less precise in estimating the regression coefficients, i.e. both K and Q will be needed. Using some model selection criteria will also help to determine if K is needed addition to Q. For example, for the data set in our paper, there are 4 lines out of the 90 have the same genotypes at 2,000 SNPs to the other 4 lines plus a few lines with kinship coefficients greater than 0.4, so the model with $K+Q$ correcting for population structure fit better than the model with Q only. Our model directly adjusts each test statistic internally by accounting for multiple levels of relatedness. Consequently, our approach simultaneously improves the detection of gene and trait associations and the estimation of the magnitude of association. While it is well appreciated that multivariate analysis can give higher power for linkage detection and can provide more accurate localization of quantitative trait loci than univariate trait analyses (Turner et al., 2004), no attention has been paid to whether similar gains are possible in an association paradigm. Here we provide a multivariate association mapping approach. Our interest was in developing a procedure which increased our ability to detect the same genetic variant which influenced multiple traits, yet suffered little when correlation between response variables was instead due to background genetic variation or other unmeasured confounders.

Although we use STRUCTURE to infer population stratification, other approaches such as the PCA based method proposed by Price et al. can be also used. Although we have focused on quantitative traits, our methods can be applied to qualitative traits (such as binary data), for which we simply change the likelihood for qualitative data and put a vague prior on βs (except the intercepts which need a flat prior) (Mardia, 1988).

Finally, the method benefits from the explicit availability of the full conditionals and introduction of covariates into the model adds little computational burden to the analysis. Our algorithm is more suited for candidate gene mapping, however, and it is not well suited for genome-wide scan where millions of genetic markers are tested for association with a phenotype.

Chapter 6

References

1. AFFYMETRIX (2001) Microarray Suite. Santa Clara, CA: Affymetrix.

2. AFFYMETRIX (2003) Genechip Expression Analysis Technical Manual. `http://www.affymetrix.com`. Santa Clara, CA: Affymetrix.

3. Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., and Horton, M., et al. (2005) Genome-Wide Association Mapping in *Arabidopsis* Identifies Previously Known Flowering Time and Pathogen Resistance Genes, *PLoS Genetics*, **1(5)**: e60.

4. Archer, K. J., Dumur, C. I., Joel, S. E., and Ramakrishnan, V. (2006) Assessing quality of hybridized RNA in Affymetrix GeneChip experiments using mixed-effects models, *Biostatistics*, **7**, 198-212.

5. Azzalini, A. and Capitanio, A. (1999) Statistical applications of the multivariate skew-normal distribution., *Journal of the Royal Statistical Society, Series B*, **61**, 579-602.

6. Bacanu, S., Devlin, B., and Roeder, K. (2002) Association Studies for Quantitative Traits in Structured Populations, *Genetic Epidemiology*, **22**, 78-93.

7. Banerjee, S., Carlin, B., and Gelfand, A. (2004) Hierarchical modeling and analysis for spatial data, Chapman & Hall/CRC, pp: 129 -174.

8. Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion), *J. Roy. Statistical Society B*, **36**, 192-236.

9. Besag, J. and Kooperberg, C. L. (1995) On conditional and intrinsic autoregressions, *Biometrika*, **82**, 733-746.

10. Besag, J., York, J., and Mollie, A. (1991) Bayesian image restoration, with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics*, **43**, 1-59.

11. Boyartchuk, V. L., Broman, K. W., and Mosher, R. E., et al. (2001) Multigenic control of Listeria monocytogenes susceptibility in mice, *Nat. Genet.*, **27**, 259-260.

12. Buckler, E. S. and Thornsberry, J. M. (2002) Plant molecular diversity and applications to genomics, *Curr. Opin. Plant Biol.*, **5**, 107-111.

13. Carter, B. S., Beaty, T. H, Steinberg, G. D., Childs, B., and Walsh, P. C. (1992) Mendelian inheritance of familial prostate cancer, *Proceedings of National Academy of Sciences USA*, **89**, 3367-3371.

14. Chang, J. C., Wooten, E. C., Tsimelzon, A., Hilsenbeck, S. G., Gutierrez, M. C., Elledge, R., Mohsin, S., Osborn, C. K., Channess, G.C., and Aller, D. C. (2003) Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer, *Lancet*, **362**, 362-369.

15. Claus, E. B., Risch, N. J., and Thompson, W. D. (1990) Using age of onset to distinguish between subforms of breast cancer, *Ann. Hum. Genet.*, **54**,

169-177.

16. Conlon, E. M., Goode, E. L., Gibbs, M., Stanford, J. L., Badzioch, M., Janer, M., Kolb, S., Hood, L., Ostrander, E. A., Jarvik, G. P., and Wijsman, E. M. (2003) Oligogenic segregation analysis of hereditary prostate cancer pedigrees: evidence for multiple loci affecting age at onset, *International Journal of Cancer*, **105**, 630-635.

17. Copois, V., Bibeau, F., Bascoul-Mollevi, C., Salvetat, N., Chalbos, P., Bareil, C., Candeil, L., Fraslon, C., Conseiller, E., Granci, V., Maziere, P., Kramar, A., Ychou, M., Pau, B., Martineau, P., Molina, F., and Del Rio, M. (2007) Impact of RNA degradation on gene expression profiles: Assessment of different methods to reliably determine RNA quality, *Journal of Biotechnology*, **20**, 549-559.

18. Croner, R. S., Guenther, K., Foertsch, T., Siebenhaar, R., Brueckl., W. M., Stremmel, C., Hlubek, F., Hohenberger, W., and Reingruber, B. (2004) Tissue preparation for gene expression profiling of colorectal carcinoma: three alternatives to laser microdissection with preamplification1, *Journal of Laboratory and Clinical Medicine*, **143**, 344-351.

19. Derisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996) Use of a cDNA microarray to

analyse gene expression patterns in human cancer, *Nature Genetics*, **14**, 457-460.

20. Devlin, B. and Roeder, K. (1999) Genomic control for association studies, *Biometrics*, **55**, 997-1004.

21. Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, **12**, 111-139.

22. Dumur, C. I., Nasim, S., Best, A. M., Archer, K. J., Ladd, A. C., Mas, V. R., Wilkinson, D. S., Garrett, C. T., and Ferreira-Gonzalez, A. (2004) Evaluation of quality control criteria in microarray gene expression analysis, *Clinical Chemistry*, **50**, 1994-2002.

23. Ferguson, T. S. (1974) Prior distributions on spaces of probability measures, *The Annals of Statistics*, **2**, 615-629.

24. Flemming, T. R. and Harrington, D. P. (1981) A class of hypothesis tests for one and two sample censored survival data, *Communications in Statistics, Theory and Methods*, **10**, 763-794.

25. Flint-Garcia, S. A., Thornsberry, J. M., and Buckler, E. S. (2003) Structure of linkage disequilibrium in plants, *Annu. Rev. Plant Biol.*, **54**, 357-374.

26. Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004) affy-Analysis of Affymetrix GeneChip data at the probe level, *Bioinformatics*, **20**, 307-315.

27. Geisser, S. and Eddy, W. F. (1979) A predictive approach to model selection, *Journal of the American Statistical Association*, **74**, 153-160.

28. Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 473-483.

29. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004) *Bayesian Data Analysis, 2nd edition*, Boca Raton: Chapman & Hall/CRC Press.

30. Han, C. and Carlin, B. P. (2001) Markov chain Monte Carlo methods for computing Bayes factors: A comparative review, *Journal of the American Statistical Association*, **96**, 1122-1132.

31. Hanson, T. (2006) Inference for mixtures of finite Polya tree models, *Journal of the American Statistical Association*, **101**, 1548-1565.

32. Hanson, T. and Johnson, W. O. (2002) Modeling regression error with a mixture of Polya trees, *Journal of the American Statistical Association*, **97**, 1020-1033.

33. Hardy, O. J. and Vekemans, X. (2002) SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels, *Mol. Ecol. Notes*, **2**, 618-620.

34. Hästbacka, J., de la Chapelle, A., Kaitila, I., Sistonen,P., Weaver, A., and Lander, E. (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland, *Nature Genetics*, **2**, 204-211.

35. Hoggart, C. J., et al. (2003) Control of Confounding of Genetic Associations in Stratified Populations, *Am. J. Hum. Genet.*, **72**, 1492-1504.

36. Hubbell, E., Liu, W. M., and Mei, R. (2002) Robust estimators for expression analysis, *Bioinformatics*, **18**, 1585-1592.

37. Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a) Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research*, **31**, e15.

38. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, D. Y., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, **4**, 249-264.

39. Jakobsson, M., et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations, *Nature*, **451**, 998-1003.

40. Jiang, H., Harrington, D., Raby, B. A., Bertram, L., Blacker, D., Weiss, S .T., and Lange, C. (2006) Family-based association test for time-to-onset data with

time-dependent differences between the hazard functions, *Genetic Epidemiology*, **30**, 124-132.

41. Kang, H. K., Zaitlen, N. A., Wade, C. M., Kirby, A., Daly, M. J., and Eskin, E. (2008) Efficient Control of Population Structure in Model Organism Association Mapping, *Genetics*, **178**, 1709-1723.

42. Kass, R. E. and Raftery, A. E. (1995) Bayes factors, *Journal of the American Statistical Association*, **90**, 773-795.

43. Kelker, D. and Kelker, H. (1986) The effect of skewness on selection in a plant breeding program, *Euphytica*, **99**, 33-54.

44. Lacher, D. (1987) Interpretation of laboratory results using multidimensional scaling and principal component analysis, *Annals of Clinical and Laboratory Science*, **17**, 412-417.

45. Lander, E. S. and Schork, N. J. (1994) Genetic dissection of complex traits, *Science*, **99**, 33-54.

46. Lavine, M. (1992) Some aspects of Polya tree distributions for statistical modeling, *Annals of Statistics*, **20**, 1222-1235.

47. Lee, I. and Amasino, R. M. (1995) Effect of Vernalization, Photoperiod, and Light Quality on the Flowering Phenotype of Arabidopsis Plants Containing the FRIGIDA Gene, *Plant Physiol.*, **108**, 157-162.

48. Leisch, F., Weingessel, A., and Hornik, K. (1998) On the generation of correlated artificial binary data. Working Paper Series, SFB "Adaptive Information Systems and Modelling in Economics and Management Science", *Vienna University of Economics.*

49. Lemon, W. J., Palatini, J., Krahe, R., and Wright, F. A. (2002) Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays, *Bioinformatics*, **18**, 1470-1476.

50. Lettre, G., Lange, C., and Hirschhorn, J. (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits, *Genetic Epidemiology*, **31**, 358-362.

51. Li, H. and Fan, J. (2000) A general test of association for complex diseases with variable age of onset, *Genet Epidemiol*, **19 Suppl**, 1: S43-9.

52. Li, M., Reilly, C., and Hanson, T. (2008) A semiparametric test to detect associations between quantitative traits and candidate genes in structured populations, *Bioinformatics*, **24**, 2356-2362.

53. Li, C. and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error applications, *Genome Biology*, **2**, 1-11.

54. Li, C. and Wong, W. H. (2003) DNA-chip analyzer (dChip). In Parmigiani,

G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L. (eds), *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer, pp. 120-141.

55. Li, H. and Zhong, X. (2002) Multivariate survival models induced by genetic frailties, with application to linkage analysis, *Biostatistics*, **3**, 57-75.

56. Long, A. D. and Langley, C. H. (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits, *Genome Research*, **9**, 720-731.

57. Malosetti, M., van der Linden, C. G., Vosman, B., and van Eeuwijk, F. A. (2007) A mixed model approach to association mapping using pedigree information with an illustration to resistance for Phytophthora infestans in potato, *Genetics*, **175**, 879-889.

58. Mardia, K. V. (1988) Multi-dimensional multivariate Gaussian Markov random fields with application to image processing, *J. Multivariate Anal.*, **24**, 265-284.

59. Marchini, J., Cardon, L., Phillips, M., and Donnelly, P. (2004) The effects of human population structure on large genetic association studies, *Nat. Genet.*, **36**, 512-517.

60. Miki, Y., Swensen, J., Shattuck-Eidens, D., and Futreal, P. A. et al. (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1, *Science*, **166**, 66-71.

61. Michaels, S. D. and Amasino, R. M. (1999) FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering, *Plant Cell*, **11**, 949-956.

62. Mokliatchouka, O., Blackerb, D., and Rabinowit, D. (2001) Association Tests for Traits with Variable Age at Onset, *Human Heredity*, **51**, 46-53.

63. Mollie, A. (1996) Bayesian mapping of disease, In Markov Chain Monte Carlo in Practice. W.R. Gilks, S. Richardson and D.J. Spiegelhalter (eds.), New York: Chapman & Hall, pp: 359-379.

64. Morton, N. E. (1984) Trials of Segregation Analysis by Deterministic and Macro Simulation. In: Chakravarti A. , editor., *Human Population Genetics: The Pittsburgh Symposium, New York: Van Nostrand Reinhold*, 83-107.

65. Naef, F. and Magnasco, M. O. (2003) Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays, *Physical Review E*, **20**, id. 063901.

66. Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., and Zheng, H., et al. (2005) The Pattern of Polymorphism in Arabidopsis thaliana., *PLoS Biology*, **3(7)**: e196.

67. Pankratz, V. S., de Andrade, M., and Therneau, T. M. (2005) Random-effects cox proportional hazards model: General variance components methods for

time-to-event data, *Genetic Epidemiology*, **28**, 97-109.

68. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genetics*, **38**, 904-909.

69. Pritchard, J. K. and Donnelly, P. (2001) Case-control studies of association in structured or admixtured populations, *Theoretical Population Biology*, **60**, 227-237.

70. Pritchard, J. K., Stephens, M., and Donnelly, P. (2000a) Inference of population structure using multilocus genotype data, *Genetics*, **155**, 945-959.

71. Prithcard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000b) Association mapping in structured populations, *Am. J. Hum. Genet.*, **67**, 170-181.

72. Pusztai, L., Ayers, M., Stec, J., and Hortobagyi, G. N. (2003) Clinical application of cDNA microarrays in oncology, *Oncologist*, **8**, 252-258.

73. Rajeevan, H. (2003) ALFRED - the ALlele FREquency Database - update, *Nucleic Acids Research*, **31**, 270-271.

74. Redden, D. T., et al. (2006) Regional admixture mapping and structured association testing: conceptual unificcation and an extensible general linear model, *PLoS Genet.*, **2(8)**: e137.

75. Reich, D. E. and Goldstein, D. B. (2001) Detecting association in a case-control study while correcting for population stratification, *Genetic Epidemiology*, **20**, 4-16.

76. Ritland, K. (1996) Estimators for pairwise relatedness and individual inbreeding coefficients., *Genet. Res.*, **67**, 175-185.

77. Rogers, J. S. (1972) Measures of genetic similarity and genetic distance, *Studies in genetics, VII. Univ. Tex. Publ.*, **2713**, 145-153.

78. Schoor, O., Weinschenk, T., Hennenlotter, J., Corvin, S., Stenzel, A., Rammansee, H. G., and Stevanovic, S. (2003) Moderate degradation does not preclude microarray analysis of small amounts of RNA, *BioTechniques*, **35**, 1192-1201.

79. Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. (2000) Normalization strategies for cDNA microarrays, *Nucleic Acids Res.*, **28**, E47.

80. Serre, D., Montpetit, A., Paré, G, Engert, J. C,, Yusuf, S., et al. (2008) Correction of Population Stratification in Large Multi-Ethnic Association Studies, *PLoS ONE*, **2(1)**, e1382.

81. Sheldon, C., Burn, J., Perez, P., Metzger, P., Edwards, J., Peacocka, J., and Dennisa, E. (1999) The FLF MADS box gene: A repressor of flowering in

Arabidopsis regulated by vernalization and methylation, *Plant Cell*, **11**, 445-458.

82. Shih, M. C. and Whittemore, A. S. (2002) Tests for genetic association using family data, *Genetic Epidemiology*, **22**, 128-145.

83. Shindo, C., Aranzana, M. J., Lister, C., Baxter, C., Nicholls, C., Nordborg, M., and Dean, C. (2005) Role of FRIGIDA and FLC in determining variation in flowering time of Arabidopsis thaliana, *Plant Physiol.*, **138**, 1163-1173.

84. Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion), *J. Roy. Statistical Soc.*, **64**, 583-639.

85. Swift, G. H., Peyton, M. J., and Macdonald, R. J. (2000) Assessment of RNA quality by semi-quantitative RT-PCR of multiple regions of a long ubiquitous mRNA, *BioTechniques*, **28**, 524, 526, 528, 530-531.

86. Symons, R. C., Daly, M. J., Fridlyand, J., Speed, T. P., and Cook, W. D. (2002) Multiple genetic loci modify susceptibility to plasmacytoma-related morbidity in E-v-abl transgenic mice, *Proc. Natl. Acad. Sci., USA*, **99**, 11299-11304.

87. Tommasini, L., Schnurbusch1, T., Fossati, D., Mascher, F., and Keller, B. (2007) Association mapping of Stagonospora nodorum blotch resistance in

modern European winter wheat varieties, *TAG Theoretical and Applied Genetics*, **115**, 697-708.

88. Turchin, A., Guo, C. Z., Adler, G. K., Ricchiuti, V., Kohane, I. S., and Williams, G. H. (2006) Effect of Acute Aldosterone Administration on Gene Expression Profile in the Heart, *Endocrinology*, **147**, 3183-3189.

89. Turner, S. T., Kardia, S. L. R., Boerwinkle, E., and de Andrade, M. (2004) Multivariate linkage analysis of blood pressure and body mass index, *Genet Epidemiol.* **27**, 64-73.

90. Walker, S. G. and Mallick, B. K. (1999) Semiparametric accelerated life time model, *Biometrics*, **55**, 477-483.

91. Wei, L. J. (1992) The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis, *Statistics in Medicine*, **11**, 1871-1879.

92. Weiss, K. and Terwilliger, J. (2000) How many diseases does it take to map a gene with SNPs? *Nature Genetics*, **26**, 151-157.

93. Wessel, J. and Schork, N. J. (2006) Generalized Genomic Distance Based Regression Methodology for Multilocus Association Analysis, *The American Journal of Human Genetics*, **79**, 792-807.

94. Wickelmaier, F. (2003) An Introduction to MDS, `http://acoustics.aau.dk/fw` /mds03.pdf.

95. Wu, Z., Irizarry, R. A., Gentleman, R. C., Martinez-Murillo, F., and Spencer, F. (2004) A model based background adjustment for oligonucleotide expression arrays, *Journal of the American Statistical Association*, **99**, 909-917.

96. Yu, J., Pressoir, G., Briggs, W. H., Vroh, B., Yamasaki, M., Doebley, J. McMullen, M., Gaut, B., Nielsen, D., Holland, J., Kresovich, S., and Buckler, E. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, *Nature Genetics*, **38**, 203-208.

97. Xiang, Z., Yang, Y., Ma, X., and Ding, W. (2003) Microarray expression profiling: analysis and applications, *Curr. Opin. Drug Discov. Dev.*, **6**, 384-395.

98. Zhang, L., Mukherjee, B., Ghosh, M., and Wu, R. L. (2006) Bayesian modeling for genetic association in case-control studies: accounting for unknown population substructure, *Statistical Modelling*, **6**, 352-372.

99. Zhang, L., Wang, L., Ravindranathan, A., and Miles, M. F. (2002) A new algorithm for analysis of oligonucleotide arrays: application to expression profiling in mouse brain regions, *Journal of Molecular Biology*, **317**, 225-235.

100. Zhang, S., Zhu, X., and Zhao, H. (2003) On a Semiparametric Test to Detect

Associations Between Quantitative Traits and Candidate Genes Using Unrelated Individuals, *Genet. Epidemiol.*, **24**, 44-56.

101. Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Chunlao, T., Toomajian, C., Zheng, H. G., Dean, C., Marjoram, P., and Nordborg1 M. (2007) An Arabidopsis Example of Association Mapping in Structured Samples, *PLoS Genet.*, **3(1)**: e4.

102. Zondervan, K. and Cardon, L. (2004) The complex interplay among factors that influence allelic association, *Nature Rev. Genet.*, **5**, 89-100.