Stony Brook University

Constrained Functional Linear Model for Multi-loci Genetic Mapping

A Dissertation presented

by

Jiayu Huang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Concentration - Statistics)

Stony Brook University

January 2016

Stony Brook University

The Graduate School

Jiayu Huang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Song Wu - Dissertation Advisor
Assistant Professor, Department of Applied Mathematics and Statistics

Wei Zhu - Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Jie Yang - Committee Member
Assistant Professor, Department of Family, Population and Preventive Medicine

Wadie Bahou - Outside Committee Member
Professor, Department of Medicine

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Constrained Functional Linear Model for Multi-loci Genetic Mapping

by

Jiayu Huang

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Concentration - Statistics)

Stony Brook University

2016

In genome-wide association studies (GWAS), the efficient incorporation of linkage disequilibria (LD) among dense-typed linked genetic variants into analysis to improve the association power is critical yet challenging problem. Functional linear models (FLM), which impose a smoothing structure on the coefficients of correlated covariates, are advantageous in genetic mapping of multiple variants with high LD. Here we propose a novel constrained FLM (cFLM) framework to perform simultaneous association tests on a block of linked SNPs with various traits, including continuous, binary and zero-inflated count phenotypes. The new cFLM applies a set of inequality constraints on the FLM to ensure model identifiability under different

genetic codings. The method is implemented via B-splines, with an augmented Lagrangian algorithm is employed for parameter estimation. For hypotheses testing, a test statistic that accounts for the model constraints has been derived, following a mixture of chi-square distributions. Simulation results show that cFLM is effective in identifying causal loci and gene clusters compared to several competing methods based on single markers and SKAT-C. We applied the proposed method to analyze the COGEND data and a large-scale GWAS data on dental caries risk.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SNP	single-nucleotide polymorphism
LD	linkage disequilibrium
GWAS	genome-wide association studies
LD blocks	linkage disequilibrium blocks
smAT	single marker association tests
FLM	functional linear model
cFLM	constrained functional linear model
NB	negative binomial
ZINB	zero-inflated negative binomial
COGEND	The Collaborative Genetic Study of Nicotine Dependence

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor, Professor Song Wu, for his constant help and support throughout my graduate study. It was Professor Wu's warm guidance and encouragement that led me into the relentless pursuit of academic excellence in the past four years. I would also like to thank my committee chair, Professor Zhu, for kindly providing the data and insights in this research. Additionally, I would like thank Professor Yang for her generous help in the research efforts of my work and dissertation, and Professor Bahou for his important comments to this thesis.

I would like extend my thanks to Dr. Pati for offering the valuable research opportunity and funding support in the Pediatrics in the past two years. I also want to thank my lab mates and friends, especially Jiawen and Xue, for their enduring help and company. I would like to reserve the special word of thanks to Bing, who is always by my side, no matter rain or shine.

Lastly, I want to express my deepest gratitude to my parents, Youji Huang and Zheng Guo. Without their love and support, I would never have been able to come this far.

This dissertation is dedicated to my parents.

Chapter 1 Background

1.1 Genetics Background

DNA is a nucleic acid that contains information on the development and functioning of almost all known living organisms except for certain viruses. The DNA segments carrying such genetic information are called genes. DNA consists of two chains of subunits twisted around one another forming a double strand helix. The subunits of each strand are nucleotides, each of which may be one of the four bases, which are adenine (A), thymine(T), guanine(G) and cytosine(C). Typically, A pairs with T and G pairs with C. Inside the nucleus of a living cell, genes are arranged in a linear order along the chromosomes, consituting the entire genome. Particularly, a human genome consists of 23 pairs of chromosomes, i.e., 46 chromatics.

A **single-nucleotide polymorphism (SNP)** is a DNA sequence variation occurring when a single nucleotide A, T, C or G at a specific position in the genome differs between members of a biological species. Figure 1 showed an example of SNPs. As we can see, in the first SNP, nucleotide C in chromosome a was replaced by nucleotide T in the same place of chromosome c. Positions with such properties are called C/T polymorphism. In this case, we say that this SNP has two alleles. SNPs occur normally throughout the genome, with a frequency of about one in every 300 nucleotides. For human genome in a size of three billion bases, there are roughly 10 million SNPs,

acting as biological markers to help locate true genetic variants that are associated with a phenotype of interest. When SNPs occur within a gene or in a regulatory region, they may play a more direct role by affecting the gene's function. By extracting SNPs from the same chromosome, a combination of bases from nearby regions can be constructed, which is called a haplotype (See Figure 1).

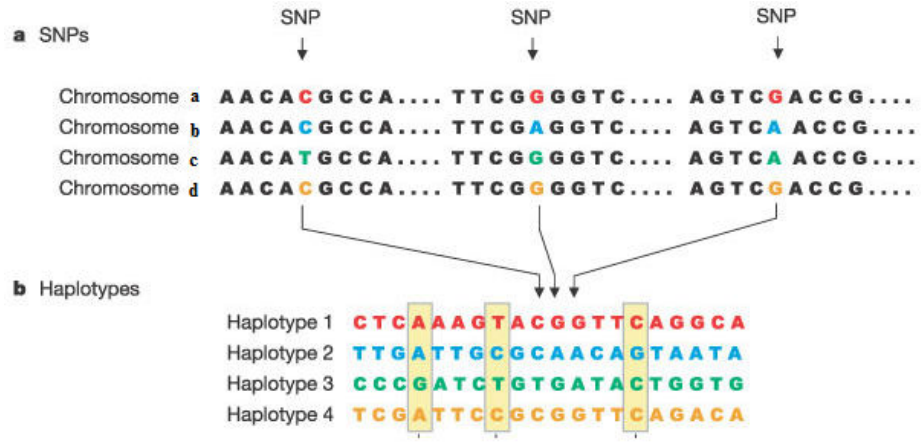

Figure 1: Illustration of SNPs and haplotypes in human genome. Cited from http://hapmap.ncbi.nlm.nih.gov/ (2013).

In general, almost all the common SNPs have only two alleles, so most studies usually focus on the information provided by such biallelic SNPs. For example, assume that one SNP locus has A/C alleles. The genotypes for this locus can be AA, AC and CC. AA and CC are called homozygous, and AC heterozygous. When under random mating, both allele and genotype frequencies of a nature population would remain constant from generation to generation. This phenomenon is called the Hardy-Weinberg Equilibrium

(HWE). In the case of a single locus of two alleles, A and a, with frequencies given by $P(A) = p, P(a) = 1 - p$, respectively, their genotype frequencies at each generation are:

$$P(AA) = p^2, P(Aa) = 2p(1-p), P(aa) = (1-p)^2$$

These theoretical frequencies provide information about a baseline against changes that could be analysed.

Linkage disequilibrium (LD) is another important genetic phenomenon arising from the co-inheritance of alleles at nearby loci on the same chromosome. It suggests the non-random association of alleles at two or more loci. The degree of LD depends on the difference between observed and expected haplotype frequencies. If one locus has alleles A and a with frequencies p_A and p_a, $p_A + p_a = 1$, and a second has alleles B and b with frequencies p_B and p_b, $p_B + p_b = 1$. The haplotypes for two loci can be AB, aB, Ab or ab, and the frequencies of each combination are expressed in terms of haplotype frequencies, i.e. p_{AB} for haplotype AB, p_{aB} for aB, p_{Ab} for Ab, p_{ab} for ab. If the two SNPs are unlinked, the expected haplotype frequencies are the product of the constituent allele frequencies. For example, the AB haplotype should have frequency $p_{AB} = p_A p_B$. The deviation of the observed haplotype frequency from the expected is the coefficient of linkage disequilibrium, denoted by:

$$D = p_{AB} - p_A p_B.$$

Table 1: Relationship between haplotype frequencies, allele frequencies and coefficient of linkage disequilibrium D.

	A	a	Total
B	$p_{AB} = p_A p_B + D$	$p_{aB} = p_a p_B - D$	p_B
b	$p_{Ab} = p_A p_b - D$	$p_{ab} = p_a p_b + D$	p_b
Total	p_A	p_a	1

Here, $D \neq 0$ implies linkage disequilibrium. Table 1 illustrates the relationship between the haplotype frequencies and the allele frequencies with D.

Sometimes D can be difficult to interpret because its range depends on allele frequency and it is not symmetrical about zero. Therefore, it is usually rescaled to be within a range from 0 to 1. One of the standardized measures is

$$D' = \frac{D}{D_{\max}}, \quad \text{where} \quad D_{\max} = \begin{cases} \max(-p_A p_B, -p_a p_b), & D < 0; \\ \min(p_A p_b, p_a p_B), & D > 0. \end{cases}$$

Another standardized measure is given by the correlation coefficient between pairs of loci, denoted as

$$r = \frac{D}{\sqrt{p_A p_B p_a p_b}}.$$

Over a series of generations, in an unstructured population, we can assume that only strong correlations between SNP markers closely linked to each other will remain. LD decays exponentially over time due to recombination. If at some time we observe linkage disequilibrium between two loci,

it may disappear in the future. However, the stronger the link between the two loci, the smaller it will be, the rate of convergence of D to zero.

A large number of studies have shown that smaller distance between SNPs suggests stronger correlation (Kawakami et al., 2014; Talas et al., 2012). That is, **linkage disequilibrium usually decays over distance**. Figure 2 depicts the relationship between LD (r^2) and genetic distance between SNP loci pairs using a simulated data set. We can see that the value of r^2 decreases when the genetic distance increases. Figure 3 shows how the LD between causal SNP and nearby markers are employed in linkage disequilibrium mapping. Markers must be mapped at a density compatible to the distances in which LD extends in the population (Rafalski 2002). The decay of LD with genetic distance over the genome is specified by a smooth curve with peaks at the trait loci. We could see that the steeper the decay is, the higher density of markers is required to locate the correct causal SNP.

1.2 Genome-wide Association Studies

Recent advances in technology have facilitated the development of association studies using a highly dense map of genetic markers. Nowadays, up to a million SNPs can be genotyped along the whole genome for thousands of sampled subjects. To investigate how genetic variants, like SNPs, may contribute to certain phenotypes including diseases, the Genome-Wide Association Studies (GWAS) have been developed and seen huge progress in last few years. The basic idea of GWAS is to screen and test the associations of a large

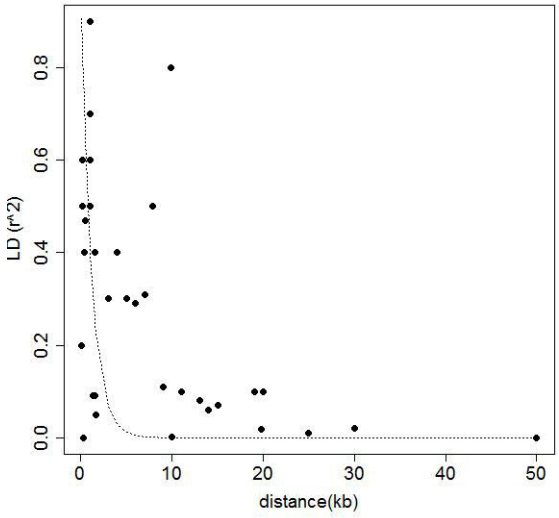

Figure 2: Relationship between linkage disequilibrium (r^2) and genetic distance between SNP loci pairs

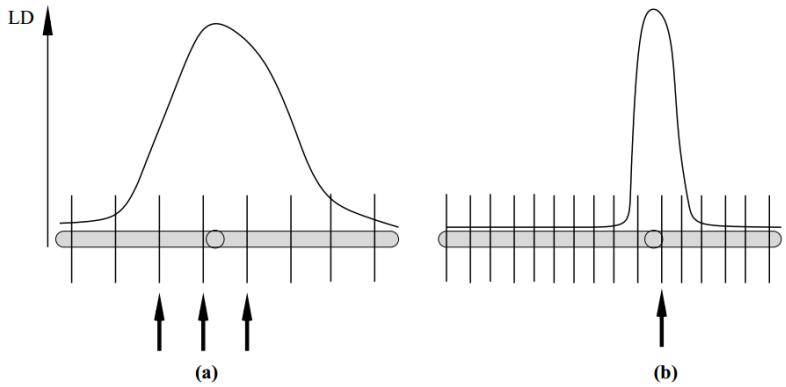

Figure 3: Influence of density of markers on the true causal variant detection, the steeper the decay of LD is, the higher the density of markers is required. The graph is cited from Wu (2008).

number of SNPs at the genome level to a given disease using high-throughput genotyping technologies. Nearly 12 million unique SNPs have been assigned SNP IDs, i.e. the rs number, in the National Center for Biotechnology Informations dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/) with a summary of allele frequencies.

GWAS does require to genotype SNPs with sufficient density to capture a large proportion of the common variants and the causative gene, with regard to the decay of LD. On the other hand, considerable number of unrelated individuals should also be included in order to gain adequate power to detect genetic variants of modest effects.

Typically, GWAS has 4 steps:

1. Selection of large number of individuals with certain phenotypic traits.

2. DNA genotyping and quality control of assays and data, check for genotyping errors.

3. Statistical tests for association between the SNPs and the phenotypic traits, usually based on single markers.

4. Replication of identified associations in an independent population sample and examination of functionality of identified SNPs and their surrounding regions.

Figure 4B displays the LD pattern of the associations in the CHRNA5 gene region identified by a GWAS on nicotine dependence. The LD measure

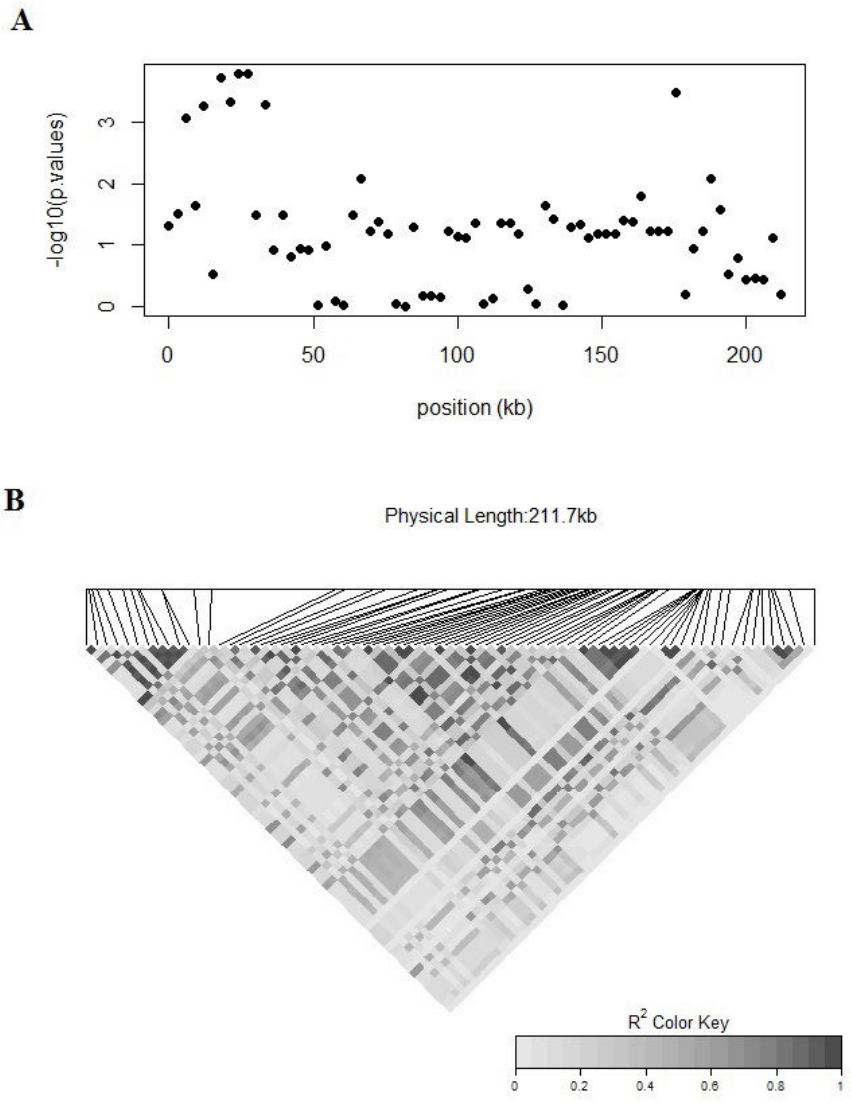

Figure 4: Associations in the CHRNA7 Gene Region Identified by a Genome-wide Association Study of COGEND (The Collaborative Genetic Study of Nicotine Dependence) (2011). A, The $-\log_{10} P$ values for association with smoking behavior in CHRNA5 gene cluster are plotted for each SNP genotyped in the region. B, Pairwise linkage disequilibrium estimates (measured as r^2) between SNPs in CHRNA5 gene cluster are plotted for the region.

was given by the pairwise allelic squared correlations r^2, which are plotted in the form of a heat map. Higher r^2 values are indicated by darker shading. Notice that the whole chromosomal region can be roughly divided into several contiguous DNA segments, or blocks. Since such bunches of SNPs present high LD correlations inside the blocks, we may also call these blocks the **linkage disequilibrium blocks** (abbreviated as **LD blocks**).

Typical GWAS analyses mainly utilize linear regressions with SNP genotypes applied SNP by SNP, i.e. hundreds or thousands of linear regression models are fitted. When SNPs are tested one by one, it is easy to assign a p-value to a SNP by conducting a likelihood ratio test. If we ignore non-genetic predictors such as age, sex and diet, then the only relevant parameters are the intercept μ and the slope β of the SNP. The parameters can be estimated by OLS or MLE. The null hypothesis $\beta = 0$ can be tested by permutation tests, or likelihood ratio test methods. The test statistic is asymptotically distributed as a χ^2-distribution with 1 degree of freedom. Importantly, the p-values must be corrected for multiple testing, either by a Bonferroni correction or some version of a FDR correction. Figure 4A shows the log 10 p-values for association of CHRNA5 gene cluster with phenotypic trait in GWAS. We can see that such scatter plots can identify suspected SNPs over a chromosomal region by analysing the p-values of the coefficients fitted by univariate linear models.

However, traditional GWAS ignores the functional genetic effects from the surrounding markers while assessing one particular SNP. Additionally,

Table 2: A sample GWAS data set

Subject	Marker				Response
	M_1	M_2	$\cdots$	M_p	y
1	2	2	$\cdots$	0	1
2	2	1	$\cdots$	1	0
3	2	0	$\cdots$	2	0
4	1	2	$\cdots$	2	1
5	1	1	$\cdots$	0	0
6	1	0	$\cdots$	1	0
7	0	2	$\cdots$	1	1
8	0	1	$\cdots$	2	0
9	0	0	$\cdots$	0	1

while there are more than 10 million SNPs in the human genome, the densest SNP platform can only genotype about 1 million of them. In some cases, if we are lucky, the causal SNPs may be already included in the genotyped SNPs. Nevertheless, in most situations, we might not be that lucky since a large number of SNPs are still not included by the current biological technology. In this case, by using LD, we can make use of the information provided by the adjacent SNPs that are highly correlated to a causal SNP. In fact, the traditional SNP by SNP single marker association test (smAT) method is inadequate to detect the accurate functionality of causal SNPs and is rather time-consuming. Nowadays, more and more multi-loci genetic mapping methods in GWAS help improve identifying associations with many strongly correlated SNPs in a wide chromosomal region among the SNPs, due in part to LD. This renders it intuitive to discover some suspected gene blocks (or LD blocks) that may be associated with the phenotypic trait.

1.3 Existing Methods in Multi-loci Association Studies

Association tests based on multiple linked SNPs are believed to be more powerful than single SNP-based tests. Since SNPs are locally connected due to linkage disequilibrium, they have natural grouping structure and form so-called linkage disequilibrium (LD) blocks in contiguous genomic regions, as shown in Figure 4B . By incorporating this correlation structure, inferences about the underlying causal variants become possible and flexible. Partial information about unsampled causal SNPs can be jointly represented by neighboring sampled markers, which could possibly increase the likelihood of capturing unobserved genetic effects. Integrating a block of SNPs simultaneously into the model also enables multi-loci genetic mapping which relaxes the isolated assumptions of individual marker tests. Several multi-SNP methods have been proposed, such as entropy-base methods (Cui et al., 2008), kernel machine methods (SKAT, SKAT-C) (Ionita-Laza et al., 2013), and two-marker LD mapping (Yang et al., 2014). In addition, some variable selection models, such as LASSO (Wu et al., 2009) and smoothed minimax concave penalized regression (SMCP) (Liu et al., 2011), were developed for identifying a core subset of potential causal variants. While these methods are useful, most of them overlooked marker ordering information in their physical positions. Meanwhile, even though SKAT and SMCP consider pairwise LD in their models, the rich information on block-wide correlation structure is still not well recognized.

1.3.1 Basic Linear Regression Models

Let y_i be the phenotypic trait for subject i, α_0 be the intercept term, $\boldsymbol{z}_i, \boldsymbol{\alpha}$ the vectors of covariates and their corresponding coefficients, x_{ij} the jth predictor variable, i.e. the genotype coding of the jth SNP for subject i and β_j its corresponding regression coefficient. When the response y is quantitative, under the framework of multiple linear regressions, the relationship with predictor variables $x_1, \ldots, x_p$ and covariates $\boldsymbol{z}$ is model in the way in the way as

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{z}_i + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \tag{1.1}$$

where ϵ_i is the error term for the ith subject. We could interpret the β_js as the genetic effect of the jth marker SNP. The problem was then formulated as a variable selection problem by choosing the significant effects according to their p-values.

On the other hand, the generalized linear models can be used to apply analyses when y_i is not quantitative. Using a simple link function g the model is then usually formulated as

$$g(\mu(y_i)) = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{z}_i + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \tag{1.2}$$

By running the forward, backward or stepwise model selections, we could find a subset of predictor variables linking the corresponding genetic effects

to the phenotypic responses.

Although the basic linear models provide statisticians an interpretable solution in GWAS, some novel developments in biotechnology may make the situations more complex than ever before. Since millions of SNPs can be typed on samples involving thousands of individuals nowadays, the sheer scale of data creates new problems in data analysis. In fact, with hundreds of thousands of predictors, the standard methods of multiple regression break down. These methods involve matrix inversion or the solution of linear equations for a huge number of p predictors. Since these operations scale as p^3, it comes as nosurprise that geneticists have chosen for univariate linear regression based on single marker. Besides, some simulation studies have shown that under certain circumstances, basic linear models can not well control the type I error well and usually have little power.

1.3.2 Penalized Regression Models

Some penalized regression methods have been proposed to solve the issues raised by basic linear models in the recent years. We will review two of them in this section. The LASSO approach performs well in high-dimension cases where p is large, and the smooth minimax concave penalty approach offers a more biologically meaningful view in the GWAS analysis.

LASSO Penalized Regression

Wu et al. (2009) proposed the lasso penalized logistic regression for case-control GWAS studies. The motivation of their work is as follows. In lin-

ear regression, imposition of a lasso penalty renders continuous model selection straightforward. Lasso penalized regression is particularly advantageous when the number of predictors far exceeds the number of observations. The method evaluates the performance of LASSO penalized logistic regression in case-control disease gene mapping with a large number of SNPs predictors. The strength of the lasso penalty can be tuned to select a predetermined number of the most relevant SNPs and other predictors. For a given value of the tuning constant, the penalized likelihood is quickly maximized by cyclic coordinate ascent. Once the most potent marginal predictors are identified, their two-way and higher order interactions can also be examined by lasso penalized logistic regression.

The lasso penalty is an effective device for continuous model selection, especially in problems where the number of predictors p far exceeds the number of observations n. The parameter vector $\theta = (\mu, \beta_1, \ldots, \beta_p)^T$ is usually estimated by maximizing the log-likelihood in ℓ_1-regression one replaces squares by absolute values. Lasso penalized regression is implemented by maximizing the modified objective function

$$L(\theta) = \mathcal{L}(\theta) - \lambda \sum_{j=1}^{p} |\beta_j|. \tag{1.3}$$

Note that the intercept is ignored in the lasso penalty $\lambda \sum_{j=1}^{p} |\beta_j|$. The tuning constant λ controls the strength of the penalty, which shrinks each β_j toward the origin and enforces sparse solutions.

However, the LASSO approach does have some major issues. First, it doesn't have a reliable asymptotic calculation of the overall group-wise p-value. Generally permutation test will be employed to assess significance of association in group/block level. (James et al., 2009) Second, it tends to overselect unimportant variables. Therefore, direct application of the LASSO to GWAS tends to generate findings with high false positive rates. Another limitation of the LASSO is that, if there is a group of variables among which the pairwise correlations are high, then the LASSO tends to select only one variable from the group and does not care which one is selected.

Smooth Minimax Concave Penalized Regression

In the above cases, the loss function of penalized methods did not deal with the linkage disequilibrium information among a group of SNPs. This would undermine the accuracy of results and cause potential loss of power. Liu et al.(2011) adapted the following penalty to include the possible LD information among nearby SNPs.

$$\frac{\lambda_2}{2}\sum_{j=1}^{p-1}\zeta_j(|\beta_j| - |\beta_{j+1}|)^2.$$

where the weight ζ_j is a measure of LD between SNP j and SNP $(j+1)$. This penalty encourages $|\beta_j|$ and $|\beta_{j+1}|$ to be similar to an extent inversely proportional to the LD strength between the two corresponding SNPs. Adjacent SNPs in weak LD are allowed to have larger difference in their $|\beta|$s than if they are in stronger LD. The effect of this penalty is to encourage

smoothness in $|\beta|$s for SNPs in strong LD. By using this penalty, we expect a better delineation of the association pattern in LD blocks that harbor disease variants while reducing randomness in $|\beta|$s in LD blocks that do not.

On the other hand, the minimax concave penalty (MCP) is used for the purpose of SNP selection. Zhang et al. (2010) proposed a flexible criterion that attenuates the effect of shrinkage in LASSO regression that leads to bias. The MCP is denoted by

$$\rho(t; \lambda_1, \gamma) = \lambda_1 \int_0^{|t|} (1 - x/(\gamma\lambda_1))_+ dx,$$

where λ_1 is a penalty parameter and γ is a regularization parameter that controls the concavity of ρ. Assume the unpenalized log-likelihood is defined by $\mathcal{L}(\theta)$, the Smooth Minimax Concave Penalized regression (SMCP) in a working model can be expressed as maximizing the criterion

$$L(\theta) = \mathcal{L}(\theta) - \sum_{j=1}^{p} \rho(|\beta_j|; \lambda_1, \gamma) - \frac{\lambda_2}{2} \sum_{j=1}^{p-1} \zeta_j (|\beta_j| - |\beta_{j+1}|)^2 \qquad (1.4)$$

The authors derived a coordinate descent algorithm for computing the solution to $L(\theta)$. This algorithm optimizes a target function with respect to one parameter at a time, iteratively cycling through all parameters until convergence is reached.

The SMCP penalized regression is suitable for multi-loci mapping in a dense set of SNPs by incorporating the LD information. However, it only

considers the lag-one correlation of neighboring SNPs instead of considering the higher level group-wise linkage disequilibrium in a SNP block. On the other hand, although SMCP is capable of dealing with a large number of SNPs simutaneously, it only yields single SNP-based p-values instead of an overall p-value at the block or gene set level. Since our proposed methods aim to perform probabilistic significance analyses at the block level, it is not reasonable to compare the power of SMCP and our method in this study.

1.3.3 Sequence Kernel Association Test

The sequence kernel association test (SKAT) is another multi-loci genetic mapping method proposed by Wu et al. (2011) and Ionita-Laza et al. (2013). It is a SNP set (e.g., a gene or a LD block) level test for association between a set of genetic variants and binary or quantitative phenotypes. SKAT aggregates individual score test statistics of SNPs in a SNP set and efficiently computes SNP-set level p-values, while adjusting for covariates, such as principal components to account for population stratification. For dichotomous outcomes, based on (1.2), we have the following logistic regression model.

$$g(\mu(y_i)) = \log \frac{p(y_i = 1)}{p(y_i = 0)} = \alpha_0 + \boldsymbol{\alpha}' \boldsymbol{z}_i + \boldsymbol{\beta}' \boldsymbol{X}_i$$

To test the overall effect of a SNP set, the null hypothesis is $\mathrm{H}_0 : \beta_j = 0$ for all $j = 1, \ldots, p$. Nonetheless, the standard likelihood ratio test approach

is usually underpowered to detect causal genetic variant. The SKAT method tries to use a variance component method to enhance the power. Assume that each β_j follows an unknown distribution with mean zero and variance $w_j\tau$, where τ is a variance component and w_j is a pre-specified weight for SNP j. The following variance-component score test statistic in the mixed effect model is used:

$$Q = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})'\boldsymbol{K}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \tag{1.5}$$

where $\boldsymbol{K} = \boldsymbol{XWX'}$, $\hat{\mu}$ is the predicted mean of y under H_0, that is $\hat{\mu} = \text{logit}^{-1}(\hat{\alpha}_0 + \boldsymbol{Z}\hat{\alpha})$ for binary traits. Here $\boldsymbol{W} = \text{diag}(w_1, \ldots, w_p)$ contains the weights of the p genetic variants. In fact, $\boldsymbol{K}$ is an $n \times n$ matrix with the (i, i')-th element equal to $K(X_i, X_{i'}) = \sum_{j=1}^{p} w_j X_{ij} X_{i'j}$. The operator $K(\cdot, \cdot)$ is a kernel function, and $K(X_i, X_{i'})$ measures the genetic similarity between subjects i and i' in the region via the p markers. This form of $K(\cdot, \cdot)$ is called the weighted linear kernel function. Epistatic effects can be modeled by using other kernel functions. The power of the SKAT depends on choices of weights w_j. Decreasing the weight of noncausal variants and increasing the weight of causal variants can produce improved power. Because in practice we do not know which variants are causal, $w_j = Beta(MAF_j; a_1, a_2)$ which is a beta function is usually used. Under H_0, the score statistic Q follows a mixture of χ^2 distributions, i.e. the $\bar{\chi}^2$ distribution where the weights can be calculated by the Davies Algorithm (Ionita-Laza et al., 2013).

The SKAT method is computationally efficient, and the exploration of

local correlation structure does consider the effect of linkage disequilibrium in a SNP set. Incorporating flexible weights in the mixed model also help greatly in boosting power. However, the SKAT method still does not consider the alignment of SNPs and does not use LD information to a higher order other than pairwise correlation. In a multi-loci genetic mapping framework, the SKAT-C method for combined common and rare variants (Ionita-Laza et al., 2013) is applicable for power comparison with our proposed method.

Chapter 2 Constrained Functional Linear Models With Simple Link Functions

2.1 Generalized Functional Linear Model

2.1.1 Motivation and model formulation

As mentioned in the first chapter, regression-based methods for association studies try to extract the causal SNP by studying the relationship between a phenotypic trait and the genotypes of a large number of SNPs. Our goal is to build up a regression model that could explain the phenotypic trait with the SNP marker genotypes gathered from the sample. Association tests based on multiple linked SNPs are believed to be more powerful than single SNP-based tests. By incorporating the LD correlation structure, inferences about the underlying causal variants become more plausible and flexible. Partial information about unsampled causal SNPs can be jointly represented by neighboring sampled markers, which could potentially increase the likelihood of capturing unobserved genetic effects. Integrating a block of SNPs simultaneously into the model also enables multi-loci genetic mapping which relaxes the isolated assumptions of individual marker tests. However, most of the existing multi-marker regression methods overlooked the alignment of marker physical positions. Even though some of these methods (SKAT and SMCP) considered pairwise LD in their models, the rich information about block-wide correlation structure is still under utilized.

In the presence of high degree of LD, we could use a smooth coefficient function over the markers to simulate the SNP effects. Imposing the structure of smooth coefficient function has two prominent advantages. On one hand, the smoothing structure could incorporate the spatial information for SNP markers. The spatial distance here may mean the physical genomic distance, or it could be the biological distance that is related to LD between markers. On the other hand, the strong correlation of SNP genotypes in a LD block can be well addressed. Adjacent SNPs, or SNPs in high LD, are expected to show similar effects under this structure. This result corresponds to the nature of genetic effects.

Functional linear models (FLM) serve as a good solution to the above-mentioned problems. Unlike conventional methods that discard a large amount of information due to lack of model complexity, FLM preserves the intrinsic correlation structure and spatial information in the data. In essence, FLM can model genetic effects from a functional prospective, and explicitly take into account the effect that neighboring correlated SNPs in LD would show similar genetic effects. That is, regression coefficients of the genetic markers can be structured as a smooth continuum over their contiguous positions (Ramsay, 2006). When multiple causal variants exist in one LD block, the peaks/valleys of the fitted coefficient function would indicate the potential loci of causal variants. Additionally, the smooth coefficient function can be expanded in terms of spline bases, which allows for substantial dimension reduction in parameter estimation (Cardot et al., 2003). Both functional

principal component analysis (FPCA) and beta-smooth only approaches can be applied to construct the FLM (Fan et al., 2013; Luo et al., 2012). Since beta-smooth only is more straightforward, it will be used for the construction of FLM hereafter. By treating cumulative genetic effect as a continuous function over marker positions rather than discrete realizations, the FLM utilized block-wide LD information more efficiently and is expected to show higher power in association analysis.

Suppose n subjects are sampled, each characterized by a block of p linked SNP markers. Let m_j be the spatial location of SNP marker j, $j = 1, \ldots, p$, assuming $0 \leq m_1 < \cdots < m_p \leq M$. SNP location can be measured by the distance of a SNP from the starting position of a block. Denote $X_i(m_j)$ as the SNP genotype at marker position m_j for subject i. The genotype of a SNP is coded as 0, 1, or 2 according to the number of copies of a reference allele. The genetic effect is denoted by $\beta(m)$, a smooth coefficient function over all marker positions m. The value of the smooth coefficient function at marker position m_j is denoted by $\beta(m_j)$. Let $\boldsymbol{z_i} = (z_{i1}, \ldots, z_{iq})^\top$ stand for the $q \times 1$ vector of covariates for subject i, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^\top$ be the $q \times 1$ vector of covariate coefficients. The global-featured effect is included as the intercept, denoted by α_0. For the ith subject, let y_i denote the phenotypic response. If the response is quantitative, an i.i.d. normal error term ϵ_i with mean zero and variance σ_ϵ^2 is imposed in the model. The functional linear

model (FLM) can be formulated as

$$y_i = \alpha_0 + \sum_{u=1}^{q} z_{iu}\alpha_u + \sum_{j=1}^{p} X_i(m_j)\beta(m_j) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma_\epsilon^2). \tag{2.1}$$

Functional linear models for association analysis with quantitative traits have been substantially studied in Luo et al. (2012) and Fan et al. (2013). Alternatively, if the phenotypic response has an error distribution other than a normal distribution, we use a link function $g(\cdot)$ to model the mean $\mu(y_i) = E(y_i|\boldsymbol{x_i}, \boldsymbol{z_i})$ of the response. In this case, the generalized functional linear model (GFLM) is formulated as

$$g(\mu(y_i)) = \alpha_0 + \sum_{u=1}^{q} z_{iu}\alpha_u + \sum_{j=1}^{p} X_i(m_j)\beta(m_j). \tag{2.2}$$

2.1.2 B-spline Basis

Many nonparametric smoothing methods can be applied to estimate the coefficient function $\beta(m)$, including the kernel-local smoothing and B-spline smoothing. B-spline is well suited for our model. It constitutes an appealing choice for the basis function in estimating the smooth coefficient function. A B-spline is the maximally differentiable interpolative basis function that has compact support. They can be evaluated in a numerically stable and efficient way by the de Boor algorithm (De Boor, 1978). The application of B-splines has several advantages. For example, they have limited parameter covariance and built-in smoothness and continuity. What's more, they are

embedded in popular statistical softwares such as R and SPlus, which may be very convenient for practical use. It is defined as follows.

Let $\xi_0 = a$ and $\xi_{k+1} = b$. Define new knots $\tau_1, \ldots, \tau_M$ such that

$$\tau_1 \leq \tau_2 \leq \ldots \leq \tau_M \leq \xi_0,$$

$\tau_{j+M} = \xi_j$ for $j = 1, \ldots, k$, and

$$\xi_{k+1} \leq \tau_{k+M+1} \leq \ldots \leq \tau_{k+2M}.$$

The choice of extra knots is arbitrary; usually one takes $\tau_1 = \ldots = \tau_M = \xi_0$ and $\xi_{k+1} = \tau_{k+M+1} = \ldots = \tau_{k+2M}$. We define the basis functions recursively as follows. First we define

$$B_{i,1} = \begin{cases} 1 & \text{if } \tau_i \leq x < \tau_{i+1}, \\ 0 & \text{otherwise.} \end{cases}$$

for $i = 1, \ldots, k + 2M - 1$. Next, for $m \leq M$, we define

$$B_{i,m} = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1} + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}$$

for $i = 1, \ldots, k + 2M - m$. It is understood that if the denominator is 0, then the function is defined to be 0.

The functions $\{B_{i,4}, i = 1, \ldots, k + 4\}$ are a basis for the set of natural

cubic splines. They are called the B-spline basis functions of degree 3. In general, for a B-spline of degree d with $M-1$ interior knots, $(M+d)$ basis functions are needed to span the linear space formed by additional constraints at the endpoints, thus the dimension of the linear space formed by natural cubic B-spline function is $(M+d-2) = (M+1)$. To describe a curve, another important issue for B-spline is the location of knots, since different locations of knots yield different shapes of the spline functions. Usually the knots are placed uniformly over the interval for its simplicity. An alternative way is to place the knots unevenly for its flexibility.

The advantage of the B-spline basis function is that they have compact support which makes it possible to speed up calculations. See James and Hastie (2001) and Wasserman (2006) for details. Figure 5 shows the cubic B-spline basis using nine equally spaced knots on $(0, 1)$.

2.1.3 Estimation

By the virtue that linear combinations of B-splines produce smooth curves, it is convenient to represent $\beta(m)$ in terms of B-spline basis. We can express $\beta(m)$ as

$$\beta(m) = \sum_{r=1}^{d} \gamma_r B_r(m) = \boldsymbol{B}(m)\boldsymbol{\gamma}, \tag{2.3}$$

where $d \geq 1$, $\boldsymbol{\gamma}_{d\times 1} = (\gamma_1, \ldots, \gamma_d)^\top$ are real-valued coefficients, and $\boldsymbol{B}(m) = (B_1(m), \ldots, B_d(m))$ is the matrix of values for B-spline basis functions. Then

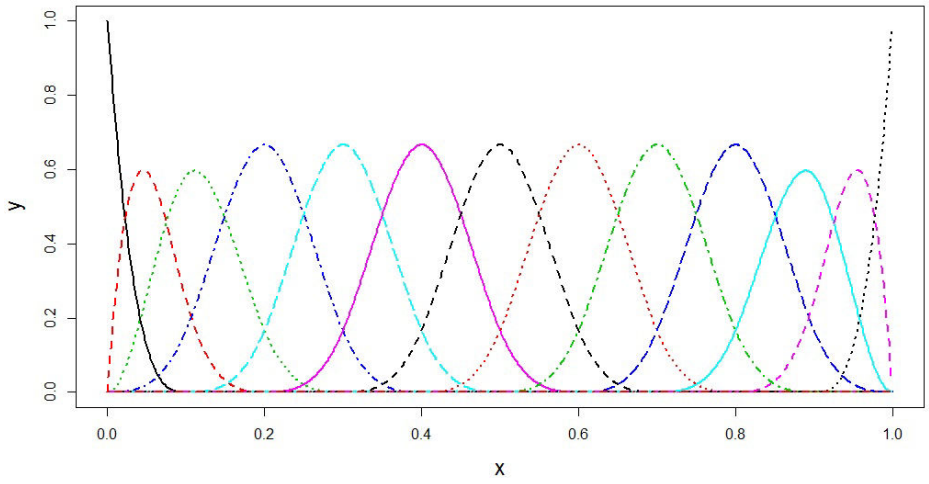

Figure 5: B-spline basis of df=3 with nine equally spaced knots on $(0, 1)$.

the formulation of GFLM using B-spline basis is:

$$\begin{aligned} g(\mu(y_i)) &= \alpha_0 + \sum_{u=1}^{q} z_{iu}\alpha_u + \sum_{j=1}^{p} X_i(m_j)\beta(m_j) \\ &= \alpha_0 + \sum_{u=1}^{q} z_{iu}\alpha_u + \sum_{j=1}^{p} X_i(m_j)(\sum_{r=1}^{d} B_r(m_j)\gamma_r) \\ &= \alpha_0 + \sum_{u=1}^{q} z_{iu}\alpha_u + \sum_{r=1}^{d} \gamma_r(\sum_{j=1}^{p} X_i(m_j)B_r(m_j)) \end{aligned} \tag{2.4}$$

Let $\boldsymbol{y}_{n\times 1} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{Z}_{n\times q} = (\boldsymbol{z_1}, \ldots, \boldsymbol{z_n})^\top$

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} X_1(m_1) & \cdots & X_1(m_p) \\ \vdots & \vdots & \vdots \\ X_n(m_1) & \cdots & X_n(m_p) \end{pmatrix}, \boldsymbol{B}_{p\times d} = \begin{pmatrix} B_1(m_1) & \cdots & B_d(m_1) \\ \vdots & \vdots & \vdots \\ B_1(m_p) & \cdots & B_d(m_p) \end{pmatrix}.$$

Using the matrix notations given, the GFLM can be reformulated as

$$g(\mu(\boldsymbol{y})) = \alpha_0 + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{B}\boldsymbol{\gamma} = \begin{pmatrix} \mathbf{1} & \boldsymbol{Z} & \boldsymbol{X}\boldsymbol{B} \end{pmatrix} \begin{pmatrix} \alpha_0 & \boldsymbol{\alpha} & \boldsymbol{\gamma} \end{pmatrix}^{\top}. \quad (2.5)$$

Suppose the response vector $\boldsymbol{y}_{n\times 1}$ has independent entries from a distribution in exponential family with density $f(y;\boldsymbol{\beta})$. By substituting $\boldsymbol{\beta}$ with $\boldsymbol{B}\gamma$, the dimensionality of parameter estimation changes from $(p+q+1)$ to $(d+q+1)$, where usually $d \ll p$. Suppose the loglikelihood function can be expressed as

$$\mathcal{L}(\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log f(y_i; \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \quad (2.6)$$

The maximum likelihood estimators (MLE) for parameters can be computed by maximizing the loglikelihood function

$$(\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \arg\max_{\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\gamma}} \mathcal{L}(\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\gamma}; \boldsymbol{y}). \quad (2.7)$$

which can be solved by the iteratively reweighted least squares algorithm.

2.1.4 Hypothesis Test

Since we are modeling a block of SNPs simultaneously in GFLM, the hypothesis test of association between genetic variants and phenotypic trait

will be made at the block level.

$$\begin{aligned} \mathrm{H}_0 &: \quad \beta(m_j) = 0, \text{ for all } j = 1, \ldots, p. \\ \mathrm{H}_a &: \quad \text{Not } \mathrm{H}_0. \end{aligned} \tag{2.8}$$

As mentioned above, we have attained dimension reduction by changing the estimator of interest from the p-dimensional $\beta(m_j), j = 1, \ldots, p$ to the d-dimensional $\boldsymbol{\gamma}_{d\times 1}$. The null hypothesis is equivalent to

$$\begin{aligned} \mathrm{H}_0 &: \quad \gamma_r = 0, \text{ for all } r = 1, \ldots, d. \\ \mathrm{H}_a &: \quad \text{Not } \mathrm{H}_0. \end{aligned} \tag{2.9}$$

The likelihood ratio test (LRT) will be applied to draw inference about whether a block of SNPs may be associated with the phenotypic trait. The LRT statistic can be calculated as the deviance between the null model and the saturated model, both adjusted for covariates. Under H_0, the LRT statistic asymptotically follows a χ^2_d distribution (Chi-square distribution with df $= d$).

2.1.5 Smoothing Penalty

Sometimes in functional linear models we hope to control the smoothness of $\beta(m)$ so that the correct association can be addressed. The benefits of using smoothing splines with B-spline basis to fit the coefficient function is discussed by Marx and Eilers (1999). For practical application, estimating method may vary with regard to the type of response or other issues. If the

phenotypic trait is a categorical variable, we should use the penalized likelihood for estimation. We can use the integration of second degree derivative of the fitted function as smooth penalty.

We continue to use the matrix notations given in the previous part. For categorical responses, the generalisation of the FLM is

$$g(\mu(\boldsymbol{y})) = \begin{pmatrix} \mathbf{1} & \boldsymbol{Z} & \boldsymbol{XB} \end{pmatrix} \begin{pmatrix} \alpha_0 & \boldsymbol{\alpha} & \boldsymbol{\gamma} \end{pmatrix}^\top = \boldsymbol{U}\boldsymbol{\gamma}^*.$$

Traditional approach attempts to maximize $\mathcal{L}$ in equation (2.6), while subject to the requirement of smoothness. Let $\boldsymbol{\Omega}$ be the $d \times d$ matrix whose (i,j)th element is $\Omega_{ij} = \int \{|B_i(m)|''^T |B_j(m)|''\} dm$. For notational convenience, we also denote

$$\boldsymbol{\Theta}_{(1+q+d)\times(1+q+d)} = \begin{pmatrix} \mathbf{0}_{(1+q)\times(1+q)} & \mathbf{0}_{(1+q)\times d} \\ \mathbf{0}_{d\times(1+q)} & \boldsymbol{\Omega}_{d\times d} \end{pmatrix}$$

The penalized log-likelihood now maximizes

$$\begin{aligned} \mathcal{L}_{pen} &= \mathcal{L}(\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{y}) - \frac{1}{2}\lambda \boldsymbol{\gamma}^\top \boldsymbol{\Omega} \boldsymbol{\gamma}^\top \\ &= \mathcal{L}(\boldsymbol{\gamma}^*, \boldsymbol{y}) - \frac{1}{2}\lambda \boldsymbol{\gamma}^{*\top} \boldsymbol{\Theta} \boldsymbol{\gamma}^{*\top} \end{aligned} \tag{2.10}$$

The factor $\frac{1}{2}$ is a small trick to get rid of a factor 2 that appears when differentiating the penalty. Maximization of the penalized log-likelihood in (2.10) can be done through scoring algorithm. Upon convergence we obtain

the maximum likelihood (ML) estimates. The preceding variance formulas are only asymptotically correct if λ is chosen a priori. Then the estimators can be computed instantly.

- **Selection of smoothing parameter**

For practical implementation of smoothing splines, one has to select an adequate smoothing parameter λ and the basis functions. It can be seen from equation (2.10) that too large a λ gives an excessive penalty for the roughness of $\beta(m)$, thus resulting in an oversmoothed estimator $\hat{\beta}(m)$. Conversely, too small a λ results in an undersmoothed $\hat{\beta}(m)$.

In practice, the smoothing parameters can be selected by cross-validation method. We use a form of cross-validation in which single subjects are deleted one at a time. Smoothing in the current context could have a number of different objectives, especially when $\beta(m)$ has more than one component. Denote by λ the smoothing parameter of any linear estimator of this section. We consider averaged log-likelihood as the metric to measure our objective. If we perform a 10-fold cross validation, let $y^{(iF)}$ be test observations that are in the i^{th} fold of subjects, $y^{(-iF)}$ be all other leaved out 9 folds of subjects. Let $\hat{\mathcal{L}^{(-iF)}}(\lambda, \boldsymbol{\gamma}^*, \boldsymbol{y}^{(-iF)})$ be the maximized penalized log-likelihood using the leaved out subjects with smoothing parameter λ. The cross-validation average loss metric is defined as

$$CV(\lambda) = \frac{1}{10} \sum_{iF=1}^{10} \hat{\mathcal{L}^{(-iF)}}(\lambda, \boldsymbol{\gamma}^*, \boldsymbol{y}^{(iF)}) \tag{2.11}$$

Then our cross-validation smoothing parameter, λ_{CV} is the maxmizer of $CV(\lambda)$. For smoothing splines estimator, the cross-validation smoothing parameters consist of λ_{CV}. Intuitively, the extra number of smoothing parameters in smoothing splines can be used to allow for possibly different smoothness of the nonparametric components.

In our simulation studies, we skipped the Functional Linear Model with smoothing penalty since previous studies have shown that they have very similar empirical power. On the other hand, we have noticed that even with smoothing penalty, the fitted coefficient function is still hard to interpret due to fluctuating wiggles and incapability of identifying zero-effect regions. In this case, we only focus on the Generalized Functional Linear Model featureing the same estimation problem, similar empirical power but much lighter computational burden.

2.2 Constrained Functional Linear Model

2.2.1 Motivation and Model Formulation

One critical problem with FLM is that the fitted coefficient functions are noisy and hard to interpret. They usually fluctuate dramatically due to several possible reasons: (1) Strong LD among nearby SNPs causes multicollinearity, which leads to erratic changes in the signs of adjacent functional coefficients; (2) The FLM is rarely capable of producing estimates that are exactly zero over regions with no apparent relationship, thus generating un-

natural wiggles in the fitted genetic function; (3) Population-specific phenomena such as mutation, genetic drift, population structure, and variations in allele frequencies result that the LD does not exactly decay with distance. Excessive local fluctuation may be relieved by adding a smoothness penalty in the model or controlling the number of spline bases. However, these methods are still not able to identify the null regions in which the coefficient function should be zero and may suffer from loss of predictive power due to oversmoothing.

As mentioned in previous section, the fitted coefficient function in FLM usually fluctuates dramatically. The fluctuation makes it difficult to explain the functional patterns and distinguish where a causal SNP is located. To ease this situation, an interpretable constrained functional linear model (cFLM) is proposed on the basis of FLM. We separate the genetic effect into two sign-specific coefficient functions and impose an equality constraint to promote spatial sparsity. The cFLM is formulated as follows:

$$g(\mu(y_i)) = \alpha_0 + \sum_{u=1}^{q} z_{iu}\alpha_u + \sum_{j=1}^{p} X_i(m_j)\beta^+(m_j) + \sum_{j=1}^{p} X_i(m_j)\beta^-(m_j)$$
$$\text{subject to } \beta^+(m_j) \geq 0, \beta^-(m_j) \leq 0, \beta^+(m_j) \cdot \beta^-(m_j) = 0 \text{ for all } j \tag{2.12}$$

where $\beta^+(m), \beta^-(m)$ are smooth coefficient functions.

2.2.2 Estimation

We express $\beta^+(m), \beta^-(m)$ in terms of B-spline bases $\boldsymbol{B}^+_{1\times d_1}(m)$, $\boldsymbol{B}^-_{1\times d_2}(m)$ and the new coefficient vectors $\boldsymbol{\gamma}^+_{d_1\times 1}, \boldsymbol{\gamma}^-_{d_2\times 1}$, respectively:

$$\beta^+(m) = \boldsymbol{B}^+(m)\boldsymbol{\gamma}^+, \beta^-(m) = \boldsymbol{B}^-(m)\boldsymbol{\gamma}^-, \tag{2.13}$$

In matrix notation, the constrained Functional Linear Model (cFLM) is formulated as:

$$\begin{aligned} g(\mu(\boldsymbol{y})) &= \alpha_0 + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{B}^+\boldsymbol{\gamma}^+ + \boldsymbol{X}\boldsymbol{B}^-\boldsymbol{\gamma}^- \\ &= \begin{pmatrix} (\mathbf{1} & \boldsymbol{Z}) & \boldsymbol{X}\boldsymbol{B}^+ & \boldsymbol{X}\boldsymbol{B}^- \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma_0} & \boldsymbol{\gamma}^+ & \boldsymbol{\gamma}^- \end{pmatrix}^\top = \boldsymbol{U}\boldsymbol{\gamma}^* = \boldsymbol{\eta}^* \\ &\text{subject to } \boldsymbol{B}^+\boldsymbol{\gamma}^+ \geq \mathbf{0}, \boldsymbol{B}^-\boldsymbol{\gamma}^- \leq \mathbf{0}, (\boldsymbol{B}^+\boldsymbol{\gamma}^+) \circ (\boldsymbol{B}^-\boldsymbol{\gamma}^-) = \mathbf{0} \end{aligned} \tag{2.14}$$

The revised loglikelihood function for cFLM is

$$\mathcal{L}(\boldsymbol{\gamma_0}, \boldsymbol{\gamma}^+, \boldsymbol{\gamma}^-) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\gamma_0}, \boldsymbol{\gamma}^+, \boldsymbol{\gamma}^-) \tag{2.15}$$

In order to obtain the MLEs for parameters, the following nonlinear optimization problem with inequality/equality constraints need to be solved:

$$\begin{aligned} \text{maximize} \quad & \mathcal{L}(\boldsymbol{\gamma_0}, \boldsymbol{\gamma}^+, \boldsymbol{\gamma}^-) \\ \text{subject to} \quad & \boldsymbol{B}^+\boldsymbol{\gamma}^+ \geq \mathbf{0}, \boldsymbol{B}^-\boldsymbol{\gamma}^- \leq \mathbf{0}, (\boldsymbol{B}^+\boldsymbol{\gamma}^+) \circ (\boldsymbol{B}^-\boldsymbol{\gamma}^-) = \mathbf{0} \end{aligned} \tag{2.16}$$

The Augmented Lagrangian Algorithm (ALA) (Birgin and Martínez, 2008) will be applied to this constrained maximization problem. Denote $h(\boldsymbol{\gamma}^*) = 0$ as the I-dimensional equality constraints, $g(\boldsymbol{\gamma}^*) \leq 0$ as the J-dimensional inequality constraints, let the augmented lagrangian be

$$L_\rho(\boldsymbol{\gamma}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) = -\mathcal{L}(\boldsymbol{\gamma}^*) + \frac{\rho}{2}\{\sum_{i=1}^{I}[h_i(\boldsymbol{\gamma}^*) + \frac{\lambda_i}{\rho}]^2 + \sum_{j=1}^{J}[\max(0, g_i(\boldsymbol{\gamma}^*) + \frac{\mu_i}{\rho})]^2\}, \tag{2.17}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^I, \boldsymbol{\mu} \in \mathbb{R}^J_+$ and $\rho > 0$. Let $\lambda_{\min} < \lambda_{\max}, \mu_{\max} > 0, \xi > 1, 0 < \tau < 1$. Let $\{\epsilon_k\}$ be a sequence of nonnegative numbers such that $\lim_{k\to\infty} \epsilon_k = 0$. Let $\lambda_i^1 \in [\lambda_{\min}, \lambda_{\max}], i = 1, \ldots, I, \mu_j^1 \in [0, \mu_{\max}], j = 1, \ldots, J$ and $\rho_1 > 0$. Let $\gamma^{*0} \in \Omega$ be an arbitrary initial point. Initialize $k \to 1$.

Augmented Lagrangian Algorithm

Step 1. Find the approximate minimizer $\gamma^{*(k)}$ of $L_{\rho_k}(\gamma^*, \lambda^{(k)}, \mu^{(k)})$ subject to $\gamma^* \in \Omega$, satisfying

$$\|P_\Omega(\gamma^{*(k)} - \bigtriangledown L_{\rho_k}(\gamma^{*(k)}, \lambda^{(k)}, \mu^{(k)})) - \gamma^{*(k)}\|_\infty \leq \epsilon_k,$$

where P_Ω is the Euclidean projection onto Ω.

Step 2. Define $V_i^{(k)} = \max\{g_i(\gamma^{*(k)}), -\frac{\mu_i^{(k)}}{\rho_k}\}, i = 1, \ldots, p$. If $k = 1$ or

$$\max\{\|h(\gamma^{*(k)})\|_\infty, \|V^k\|_\infty\} \leq \tau \max\{\|h(\gamma^{*(k-1)})\|_\infty, \|V^{(k-1)}\|_\infty\},$$

define $\rho_{k+1} = \rho_k$. Otherwise, define $\rho_{k+1} = \xi\rho_k$

Step 3. Compute $\lambda_i^{(k+1)} \in [\lambda_{\min}, \lambda_{\max}], i = 1, \ldots, I$ and $\mu_j^{(k+1)} \in [0, \mu_{\max}], j =$

$1, \ldots, J$. Set $k+1 \to k$ and go to Step 1. In practice, the first-order safeguarded estimates of Lagrange multipliers will be used:

$$\lambda_i^{(k+1)} = \min\{\max\{\lambda_{\min}, \lambda_i^{(k)} + \rho_k h_i(\gamma^{*(k+1)})\}, \lambda_{\max}\}$$

for $i = 1, \ldots, I$, $\mu_j^{(k+1)} = \min\{\max\{0, \mu_j^{(k+1)} + \rho_k g_i(\gamma^{*(k+1)})\}, \mu_{\max}\}$ for $j = 1, \ldots, J$.

2.2.3 Hypothesis Test

Similar to what has been done for FLM, we perform the likelihood ratio test to investigate the overall genetic effects represented by a block of SNPs in contiguous genomic regions. However, the null and alternative hypotheses in equation (2.8) are upon revision with regard to the new parameter space and imposed constraints.

$$\begin{aligned} &\mathrm{H}_0: \ \boldsymbol{\gamma}^+_{d_1 \times 1} = \mathbf{0} \text{ and } \boldsymbol{\gamma}^-_{d_2 \times 1} = \mathbf{0}. \\ &\mathrm{H}_a: \ \boldsymbol{B}^+\boldsymbol{\gamma}^+ \geq \mathbf{0}, \boldsymbol{B}^-\boldsymbol{\gamma}^- \leq \mathbf{0}, (\boldsymbol{B}^+\boldsymbol{\gamma}^+) \circ (\boldsymbol{B}^-\boldsymbol{\gamma}^-) = \mathbf{0} \end{aligned} \tag{2.18}$$

The LRT statistic is the deviance between the null model G_0 and the saturated model G_1, both adjusted for covariates. The alternative K-dimensional parameter space $\boldsymbol{\Omega}$ to test against is defined by the inequality and equality constraints $\boldsymbol{B}^+\boldsymbol{\gamma}^+ \geq \mathbf{0}, \boldsymbol{B}^-\boldsymbol{\gamma}^- \leq \mathbf{0}, (\boldsymbol{B}^+\boldsymbol{\gamma}^+) \circ (\boldsymbol{B}^-\boldsymbol{\gamma}^-) = \mathbf{0}$. Since we constrained that the Hadamard product of sign-specific coefficient functions is $\mathbf{0}$, for each non-negative estimate of basis coefficient in the positive part,

at least one basis coefficient in the negative part would be constrained to zero. Therefore, the dimension of the alternative parameter space should be $K = \max(d_1, d_2)$. According to Shapiro (1985) and Liu (2007), it has been shown in nonlinear optimization that $\mathbf{\Omega}$ can be approximated at the null estimate by a polyhedral convex cone defined by the gradient vectors of the constraint functions. If the unconstrained true parameter value is an interior point of $\mathbf{\Omega}$, the test statistic has an asymptotic χ^2_K distribution under H_0. Otherwise when the unconstrained parameter estimate does not fall in the admissible parameter space, the test statistic is defined by the projection of the unconstrained estimate on the k-dimensional boundary of $\mathbf{\Omega}$ according to the Hessian Matrix $I(\gamma^*)$. In this case, it has an asymptotic χ^2_k distribution under H_0 ($k = 0, \ldots, K-1$).

Under general cases, the LRT statistic asymptotically follows a mixture of chi-square distributions with mixing probabilities w_j such that $\sum_{j=0}^{K} w_j = 1$, denoted as

$$\bar{\chi}^2 = G_0 - G_1 = -2(l_0 - l_1) \xrightarrow{d} \sum_{j=0}^{K} w_j \chi^2_j. \tag{2.19}$$

The p-value of the $\bar{\chi}^2$ test statistic is defined as

$$P(\bar{\chi}^2 \geq c^2) = \sum_{j=0}^{K} w_j P(\chi^2_j \geq c^2) = \sum_{j=0}^{\max(d_1,d_2)} w_j P(\chi^2_j \geq c^2). \tag{2.20}$$

The mixing probabilities can be calculated using Monte Carlo Tech-

niques (Wolak, 1989). The algorithm is as follows: (1) Take 1,000 draws from a multivariate normal distribution with mean zero and covariance matrix equaling to the Hessian matrix $I(\gamma^*)$; (2) For each draw compute and count the number of sign-agree elements of the vectors that fall in the k-dimensional boundaries ($k = 0, \ldots, K$) of the admissible parameter space. In this case w_j is computed as the proportion of the 1,000 draws that has exactly k non-zero coefficients projected on the alternative parameter space. Monte Carlo technique is easy to implement and able to circumvent complicated numerical integrations. However, the resulting mixing probabilities are not exact.

2.3 Simulation Studies

2.3.1 General Strategies

To study the finite sample performance of the proposed cFLM, we carried out various sets of simulation under different sampling schemes. For the genotypic file, we first simulated random LD blocks with varying structures. Then we borrowed the LD structures from existing datasets to mimic real gene analyses. For the phenotypic file, we simulated binary outcomes conditional on causal genotypes under the logistic model:

$$\text{logit}(\mu(y_i)) = \log\frac{p(y_i = 1)}{1 - p(y_i = 1)} = \alpha_0 + \boldsymbol{X_i}^\top \boldsymbol{\beta_{causal}} \tag{2.21}$$

We considered simulation scenarios where $\alpha_0 = 0.2$ and sample size ranging from 500 to 2,000 to mimic the characteristics of realistic data.

To investigate whether cFLM can control type I error, β_{causal} was set to zero under the null hypothesis. Each scenario was replicated 10,000 times in order to observe the type I error rates under small genome-wide thresholds (nominal $\alpha = 0.05, 0.01, 0.005$ and 0.001).

For empirical power evaluation, we examined two different settings. First we assumed only one causal SNP was located in the LD block, having varying regression coefficient β_{causal}. Then we considered another crucial setting when two causal loci with reversed sign effects were both found in the LD block. This mimics the scenario when both deleterious and protective SNPs were in a genomic region. The two causal loci chosen were weakly correlated ($r^2 < 0.01$). The corresponding regression coefficients were set to $(\beta_{causal1}, \beta_{causal2})$ where $\beta_{causal1} = -\beta_{causal2}$. In this case we can check if the proposed test can deal with sign-heterogeneous genetic effects. Our model assumeed that the LD structure was mainly contributing to the power gain, hence the causal SNPs would be removed before the first set of analyses. For comparison purpose, we also checked scenarios where causal SNPs were kept so that we can see if the model can handle such settings. We ran 1,000 replicates for each scenario. Table 3 is the summary of parameter settings in the simulations. A p-value smaller than 0.05 would be declared significant.

In terms of functional parameters, the order of B-spline basis was set to 4 (degree = 3) to construct cubic curves with desired smoothing proper-

Table 3: Parameter settings for power evaluation with binary outcomes

Scenario	Outcome	Block structure	Causal SNP genotyped	# of Causal SNPs
SC1	Binary	random	No	1
SC2	Binary	random	No	2
SC3	Binary	random	Yes	1
SC4	Binary	random	Yes	2
SC5	Binary	CHRNA7	No	1
SC6	Binary	CHRNA7	No	2
SC7	Binary	CHRNA7	Yes	1
SC8	Binary	CHRNA7	Yes	2

ties. Knots were placed evenly in the position domain. The number of spline bases would be determined by the number of SNPs (p) in a LD block. Data-adaptive choices for the number or the placement of knots can be made via cross-validation, however for simplicity we shall not provide further discussions here. Empirically, we suggest using the maximum of 4 and the integer part of $p/6$ as the number of bases so that it is possible to capture clustering genetic effects in the fitted function. Sensitivity analyses using a broad range of parameters were performed to make sure our results are valid.

We compared the empirical power of our proposed model with three existing methods: single marker association test (smAT), SKAT for the combined effect of rare and common variants (SKAT-C), and functional linear

model (FLM). P-values for smAT were adjusted by min P Bonferroni correction. P-values for SKAT-C and FLM were calculated by combined sum test and F test.

2.3.2 Simulation using random LD blocks

The first set of simulation used randomly generated LD blocks. To make our method comparable to the other methods, SNP arrays were produced following the strategy introduced in Wu et al. (2009). The genotypes of a SNP array were generated based on a random p-dimensional multivariate normal matrix $\zeta_{n\times p}$ with mean $\mathbf{0}$ and covariance $\Sigma_{p\times p}$. Assuming that SNPs have equal allele frequencies, the following rule would be applied to generate the genotype of the jth SNP for ith subject.(Let $z_{0.25}$ be the third quartile of standard normal distribution.)

$$X_{ij} = \begin{cases} 0, & \text{if } \zeta_{ij} < -z_{0.25}. \\ 1, & \text{if } -z_{0.25} \leq \zeta_{ij} < z_{0.25}. \\ 2, & \text{if } \zeta_{ij} \geq z_{0.25}. \end{cases} \tag{2.22}$$

The correlation matrix will be defined as follows. For each block, 10% of the SNPs were selected as "tag SNPs". They were highly correlated with each other ($Corr(X_{j1}, X_{j2}) = 0.8$), moderately correalted with 30% of other SNPs ($Corr(X_{j1}, X_{j2}) = 0.5$), and weakly correlated with the remaining 60% SNPs ($Corr(X_{j1}, X_{j2}) = 0.2$). The 90% "non-tag" SNPs are correlated according

to their physical location ($Corr(X_{j1}, X_{j2}) = 0.7^{|j_1 - j_2|}$). In this case, we would not violate the assumptions that SNPs are physically adjacent and linked. Also, the LD block structures varies among different randomly generated arrays.

For type I error simulation, we used $p = 25$ SNPs in a group and $d_1 = d_2 = 4$ as the number of spline bases. We can see from Table 4 that cFLM maintained the type I error under this random sampling scheme. The quantile-quantile (Q-Q) plots of the observed p-values against expected p-values in log10 scale are presented in Figure 6.

Table 4: Type I error simulation using cFLM for binary outcomes based on random LD blocks

nominal α	N=500	N=1000	N=1500	N=2000
0.05	0.0500	0.0473	0.0466	0.0482
0.01	0.0109	0.0076	0.0099	0.0082
0.005	0.0054	0.0045	0.0046	0.0033
0.001	0.0010	0.0009	0.0011	0.0006

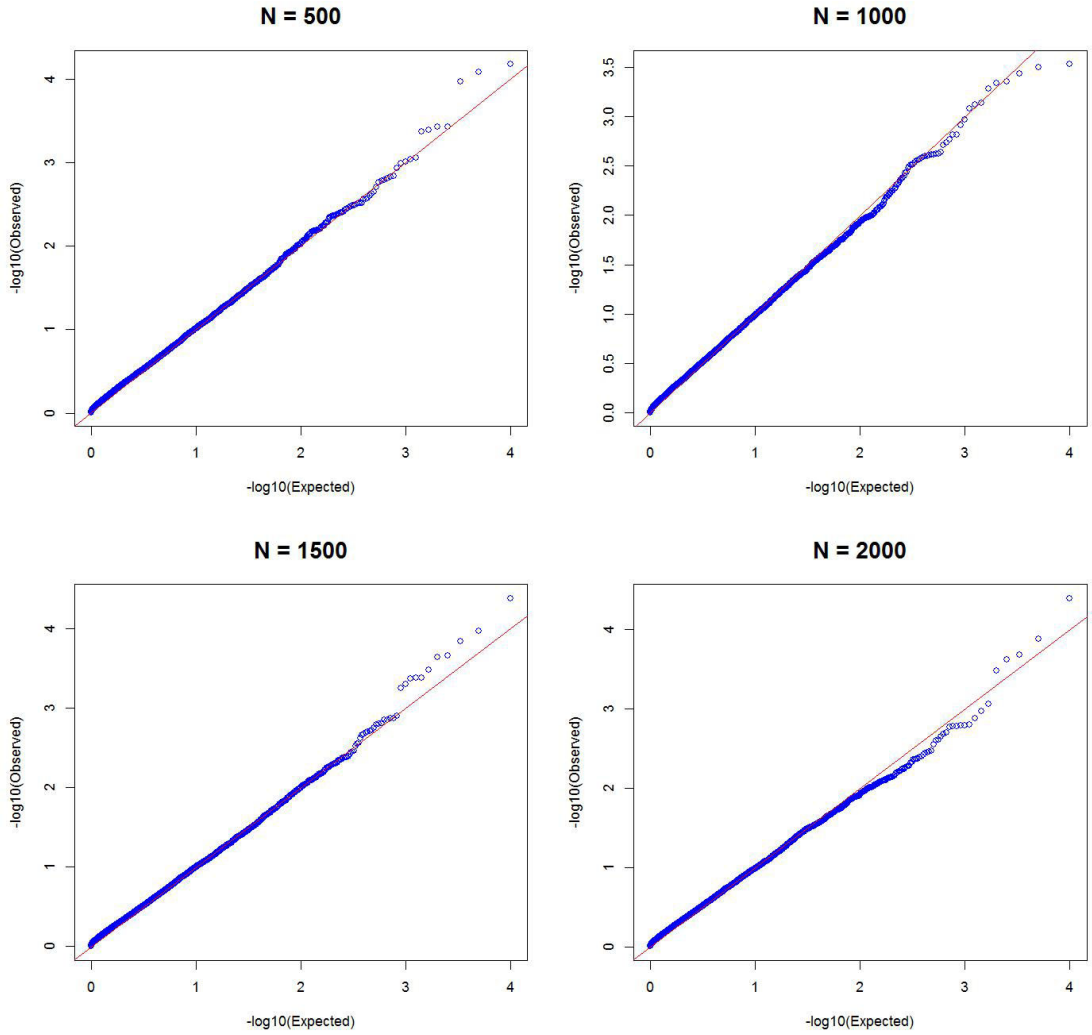

Figure 6: Type I error simulation using cFLM for binary outcomes based on random LD blocks, Q-Q plots of p-values.

Evaluation of empirical power was based on settings when regression coefficient $\beta_{causal} = 0.1, 0.2, 0.3, 0.4, 0.5$, respectively. For power calculation with single causal locus, we used the same settings as that in type I error simulation. When causal SNPs were not genotyped, we can see from SC1 (Figure 7) that the cFLM outperform other methods. In the second scenario when two causal loci had revere-sign effect, we included more SNPs in a block so that it would be possible to locate two weakly-correlated markers within the region. In this case, each block contained $p = 36$ SNPs and $d_1 = d_2 = 6$ was applied as the number of spline bases, which is similar to the setting for later simulation using CHRNA7 gene. From SC2 (Figure 8), we can see that cFLM consistently demonstrates greater power than other methods. When causal SNPs were genotyped in SC3 and SC4, smAT showed better empirical power as expected (Figure 9, Figure 10). However, cFLM had very similar power as smAT in most cases. This set of analyses demonstrates the robustness of the proposed cFLM under circumstances whenever causal loci were genotyped or not.

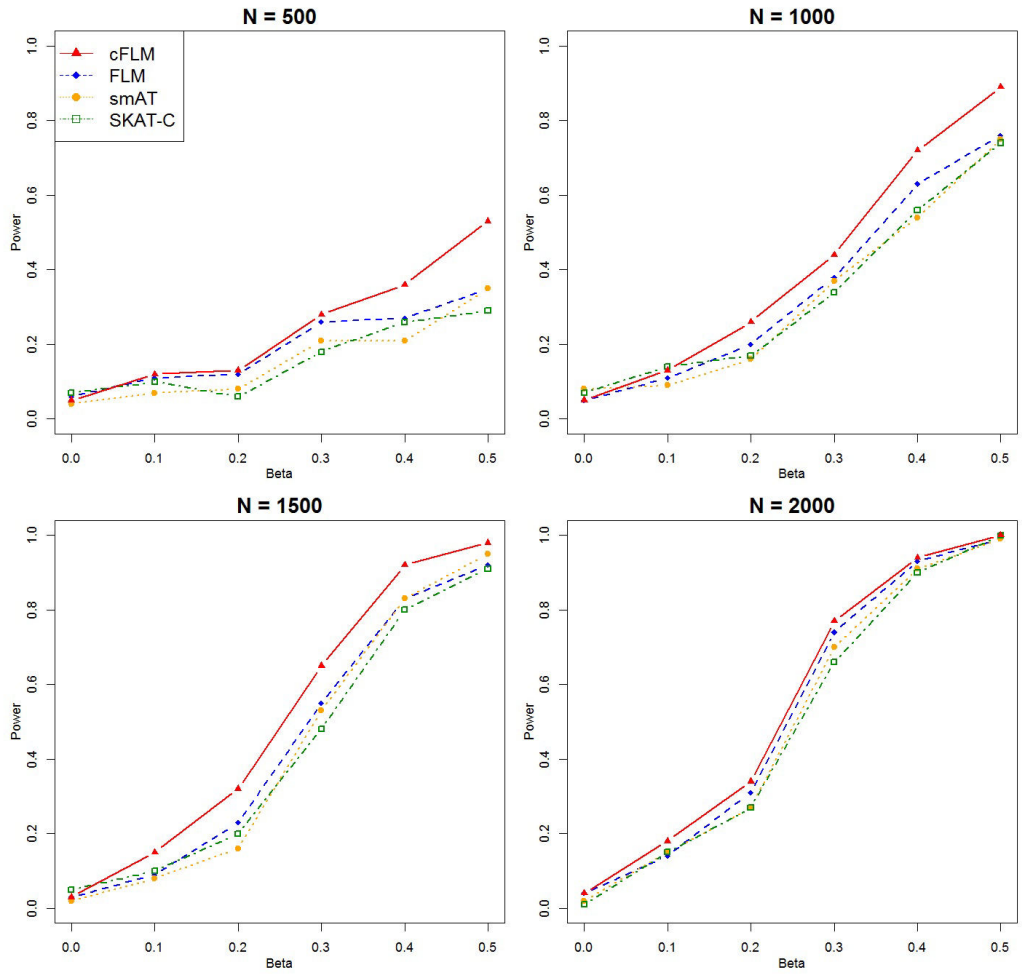

Figure 7: Power simulation for binary outcomes based on random LD blocks: single causal locus, causal SNP not genotyped, SC1.

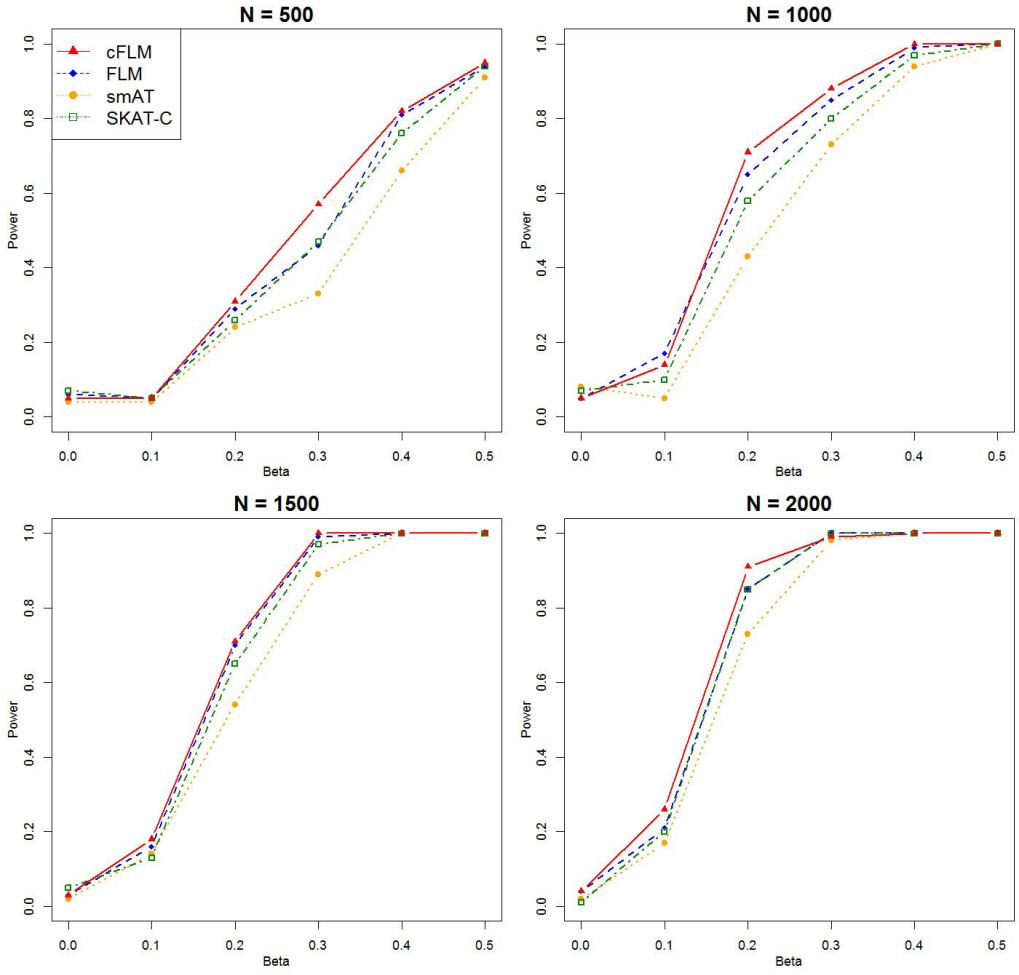

Figure 8: Power simulation for binary outcomes based on random LD blocks: two reverse-sign causal loci, causal SNPs not genotyped, SC2.

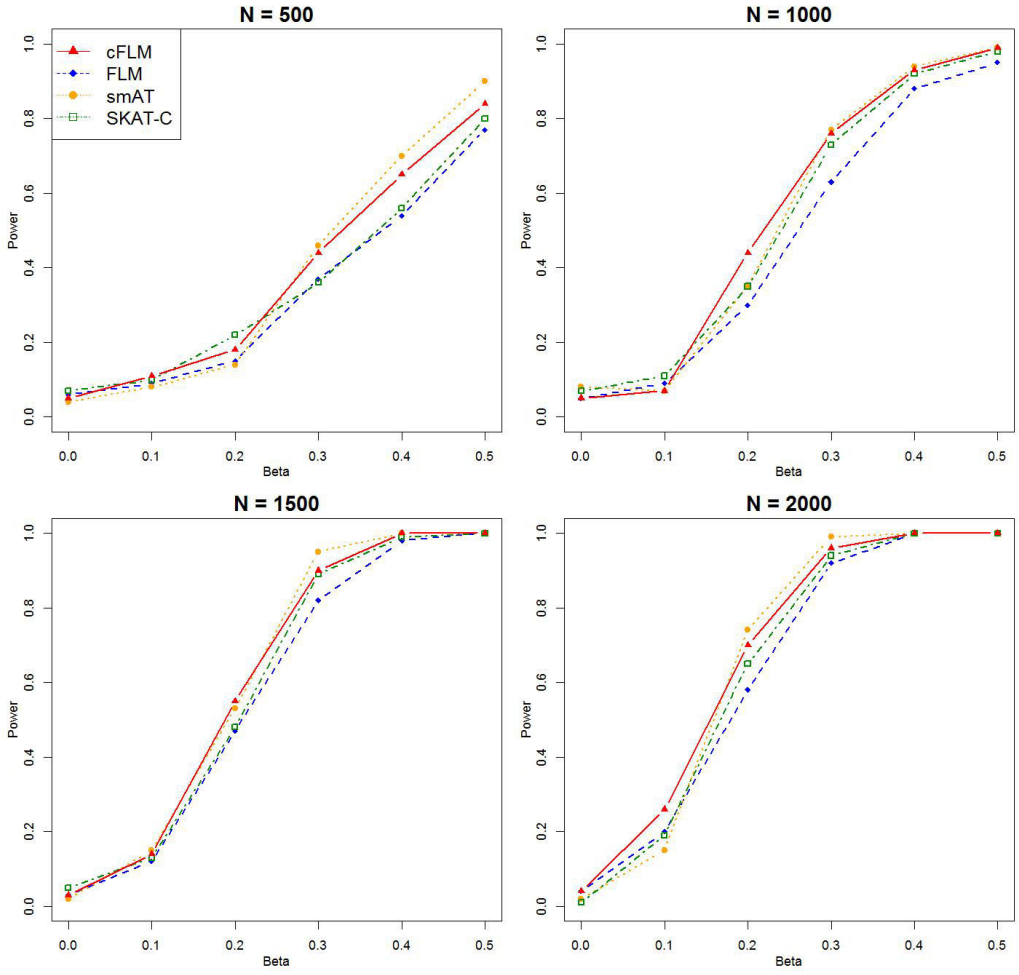

Figure 9: Power simulation for binary outcomes based on random LD blocks: single causal locus, causal SNP genotyped, SC3.

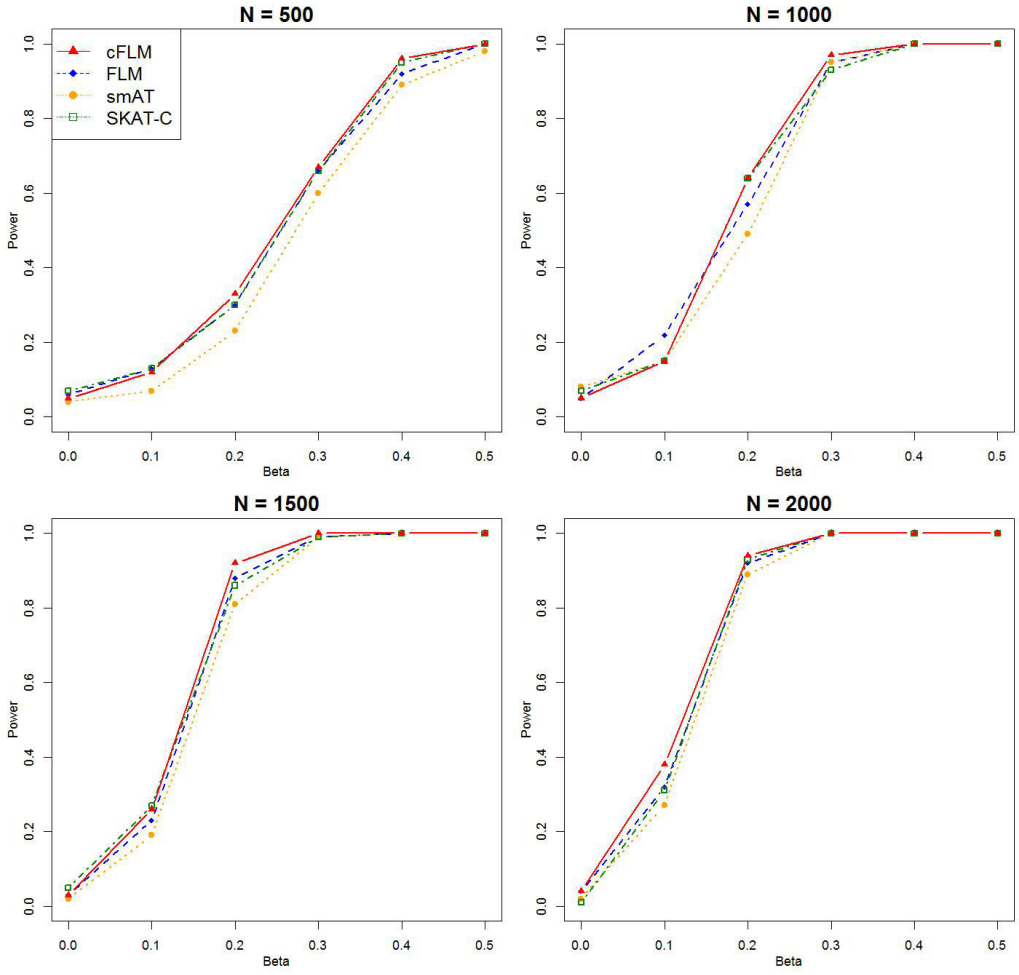

Figure 10: Power simulation for binary outcomes based on random LD blocks: two reverse-sign causal loci, causal SNPs genotyped, SC4.

2.3.3 Simulation using the CHRNA7 gene (15q13.3)

To make the genotypic files more realistic, the second set of simulation used the SNP array of the CHRNA7 gene obtained from COGEND (The Collaborative Genetic Study of Nicotine Dependence) data. A total of 2,022 individuals were included in this study. The CHRNA7 gene is located at chromosome 15, with a length of about 126 kb. A total of 39 SNPs were genotyped. We used $d_1 = d_2 = 6$ as the number of bases.

Results for type I error simulation are shown in Table 5. The cFLM well maintained the nominal type I error with various thresholds (0.05, 0.01, 0.005, 0.001). The Q-Q plots of the observed p-values against expected p-values in log10 scale are available in Figure 11.

Table 5: Type I error simulation using cFLM for binary outcomes based on CHRNA7 gene

nominal α	N=500	N=1000	N=1500	N=2000
0.05	0.0508	0.0512	0.0493	0.0506
0.01	0.0100	0.0099	0.0102	0.0093
0.005	0.0056	0.0053	0.0048	0.0046
0.001	0.0011	0.0009	0.0011	0.0007

We used the same strategy for power calculation as that in the previous simulation with random LD blocks. The power graphs are shown in Figure 13 and 14. For the single causal locus case, the proposed cFLM method has greater power when effect size is relatively large ($\beta_{causal} \geq 0.3$). When the effect size is small ($\beta_{causal} \leq 0.2$), all methods are comparable. For the two

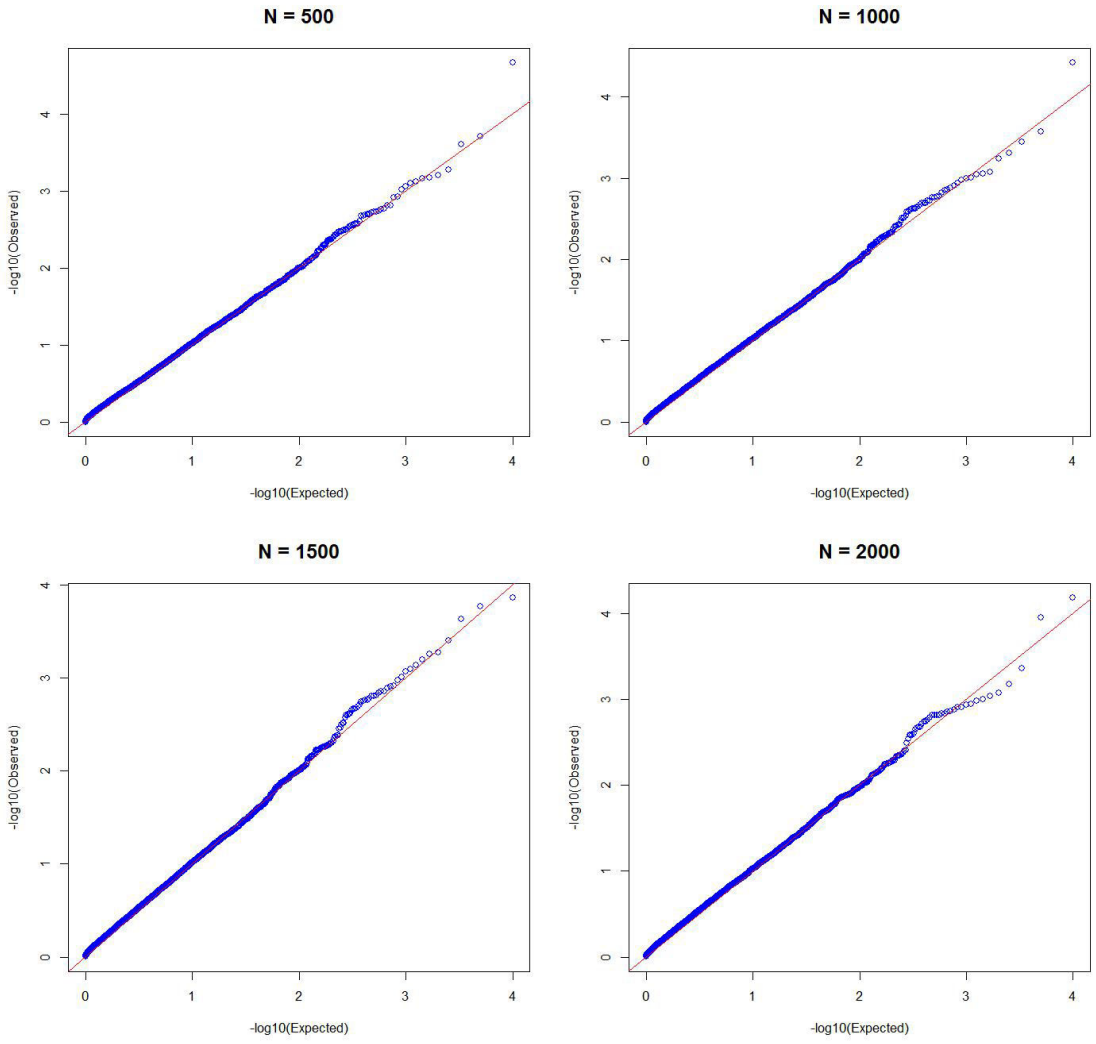

Figure 11: Type I error simulation using cFLM for binary outcomes based on CHRNA7 gene, Q-Q plots of p-values

causal loci case, the cFLM shows advantageous performance when the effect size is large. The smAT is also powerful in this case since the CHRNA7 gene has some very-highly correlated SNPs ($r^2 > 0.9$) that resemble the existence of causal SNPs in the block.

The association patterns of a sample simulation based on the CHRNA gene cluster are presented in Figure 12. For smAT, a modified manhattan plot of the -log10(p-values) by the sign of the fitted coefficients is used. For all other methods, the coefficient estimates are plotted. The causal loci is highlighted with dashed lines (left in red for positive effect, right in blue for negative effect). Compare to smAT and FLM, the cFLM fitted coefficient function is interpretable and able to identify the causal loci.

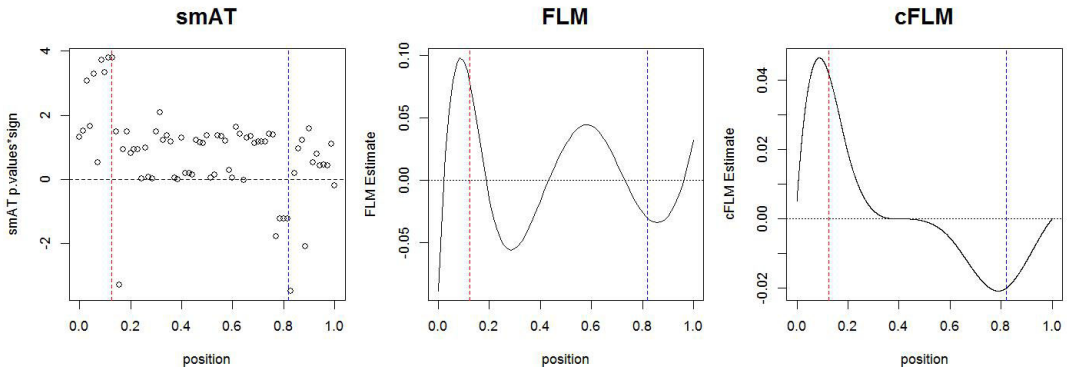

Figure 12: Illustration of fitted genetic mapping patterns using different models.

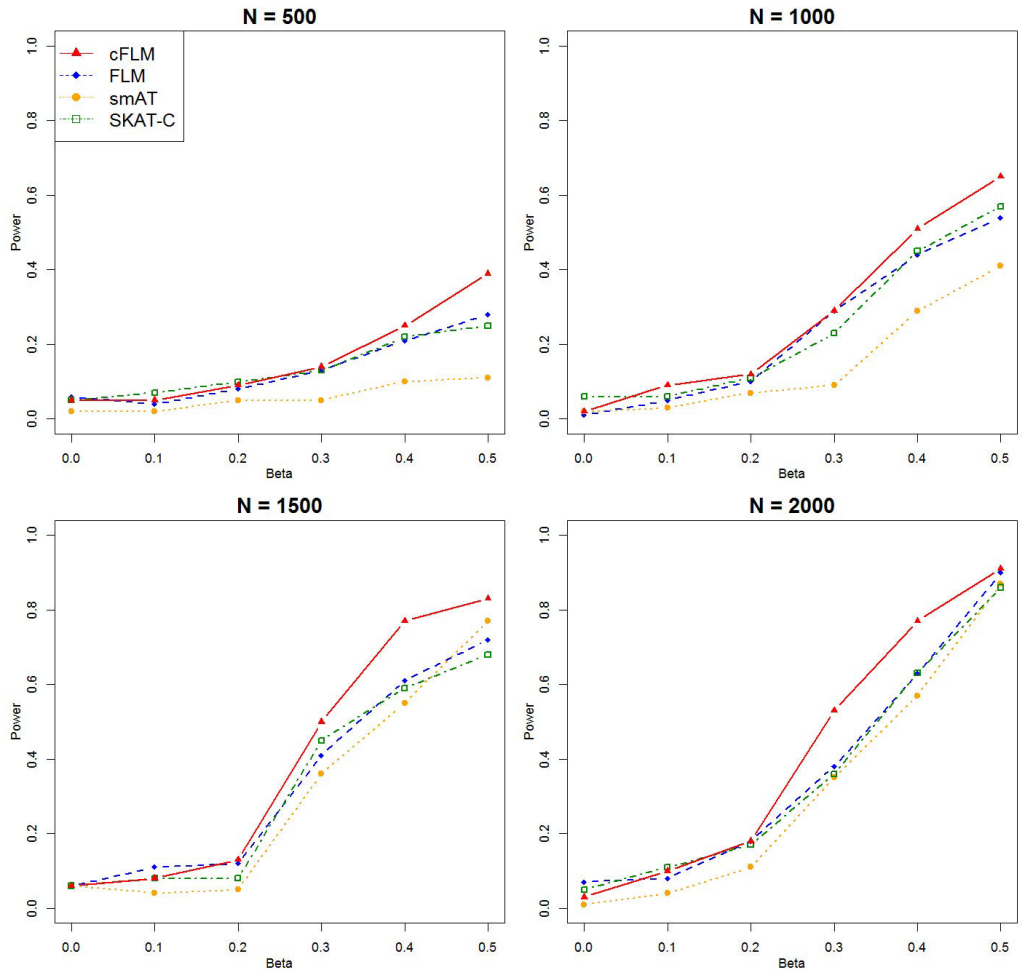

Figure 13: Power simulation for binary outcomes based on the CHRNA7 gene: single causal locus, causal SNP not genotyped, SC5.

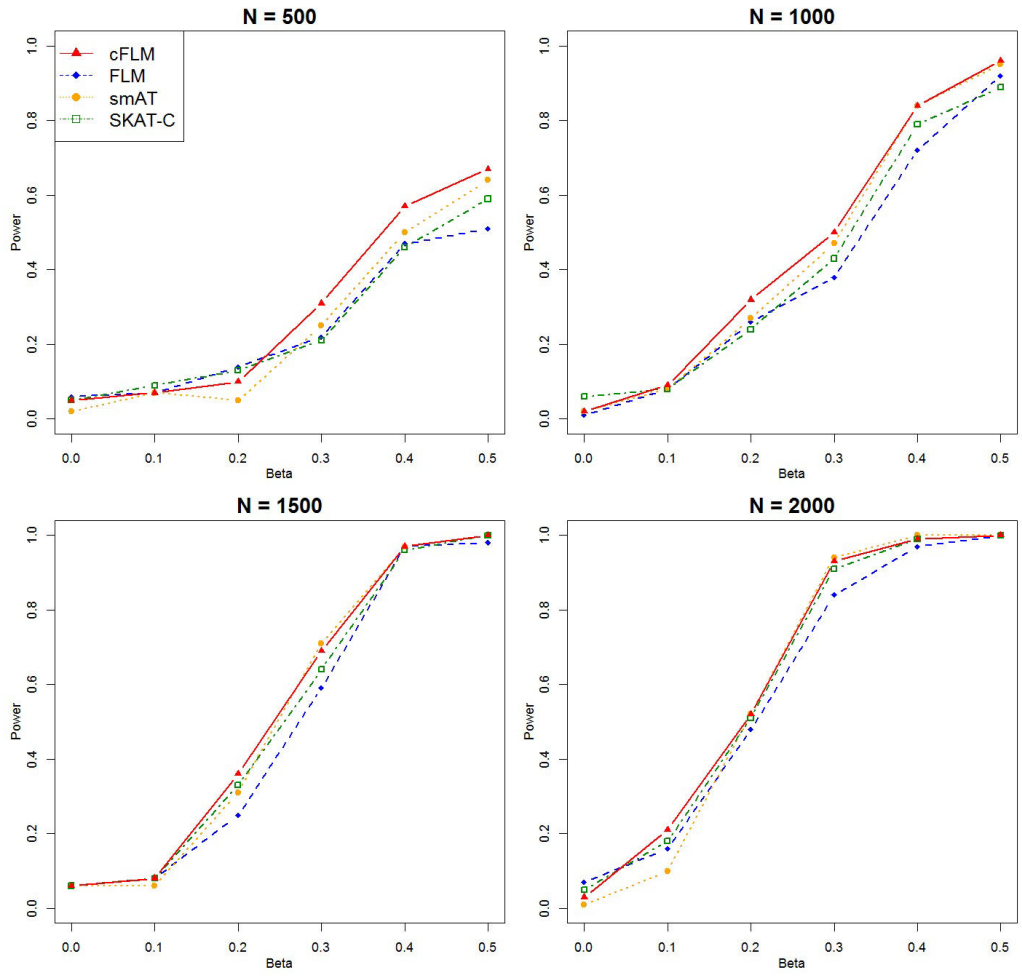

Figure 14: Power simulation for binary outcomes based on the CHRNA7 gene: two reverse-sign causal loci, causal SNPs not genotyped, SC6.

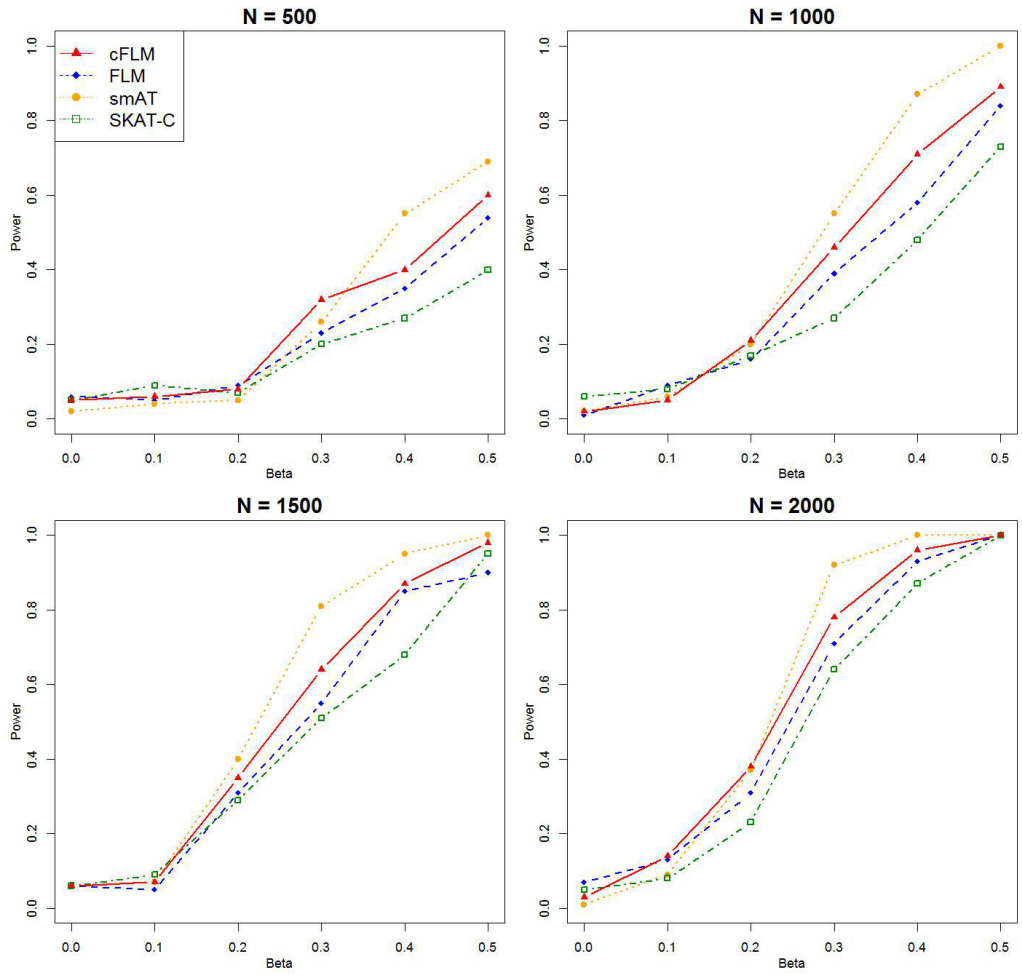

Figure 15: Power simulation for binary outcomes based on the CHRNA7 gene: single causal locus, causal SNP genotyped, SC7.

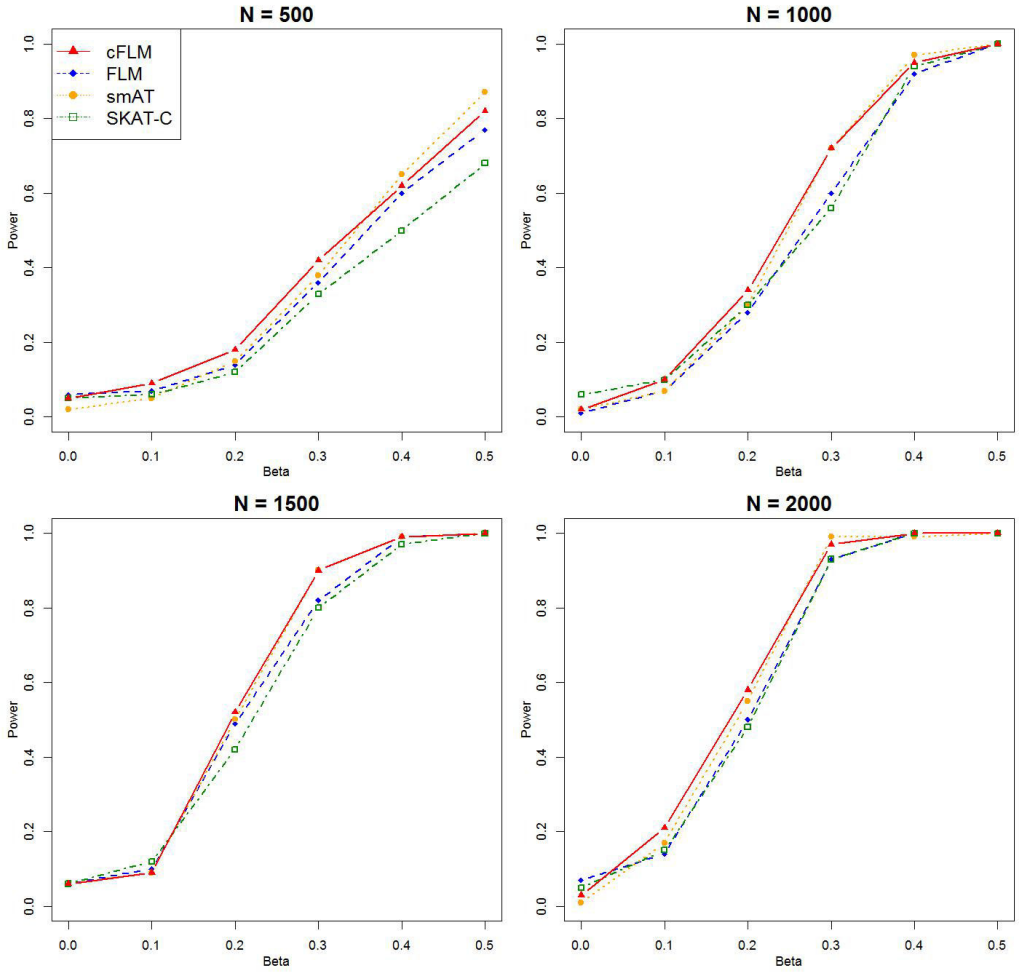

Figure 16: Power simulation for binary outcomes based on the CHRNA7 gene: two reverse-sign causal loci, causal SNPs genotyped, SC8.

2.4 Empirical Studies - COGEND

According to the World Health Statistics Report 2013, cigarette smoking is the single biggest cause of preventable mortality worldwide, causing more than 5 million deaths per year and accounting for one in 10 adult deaths. Nicotine dependence, the primary psychoactive component in tobacco, profoundly impacts people's ability to cease tobacco smoking. The etiology of nicotine dependence is found to be multifactorial, such as environmental factors related to cultural perceptions and peer smoking. However, compelling evidence from twin studies suggests that genetic factors also have a substantial impact on smoking behaviors. It is essential to identify genetic components that are associated with nicotine dependence, so that the development of corresponding treatments could be promoted to further reduce smoking related morbidity and mortality.

The Collaborative Genetic Study of Nicotine Dependence (COGEND) is a nationwide project aiming to detect the genetic mechanisms and environmental features of nicotine dependence. In our part of study on CHRN candidate genes, a total of 216 SNPs were genotyped for 2,022 individuals (1,114 cases and 908 controls). In the phenotypic file, all cases and controls were current or former smokers who reported smoking more than 100 cigarettes lifetime. When recruiting new subjects, rates of current nicotine dependence were defined by the Fagerstrom Test for Nicotine Dependence (FTND). Subjects having FTND ≥ 4 were classified as nicotine dependent

(case). Subjects having lifetime FNTD = 0 or 1 were classified as control. The original genotypic file was divided into 12 LD blocks according to their physical locations and LD structure, all of which consist of one or more contiguous gene regions. Since functional models are not well-suited for LD blocks having small number of SNPs, 4 small blocks with fewer than 7 SNPs were excluded for analyses. We applied our proposed method cFLM, along with smAT, SKAT-C and FLM to analyze the final dataset which consists of 191 SNPs in 8 LD blocks. Age, gender and race were included as covariates.

Table 6 summarizes the results. Assuming there are 20,000 genes per genome, the gene-based Bonferroni threshold for genome-wide significance is $p < 0.05/20000 = 2.5 \times 10^{-6}$. Since LD blocks usually contain one or more genes, such a significance level is suitable for block-based multiple tests adjustment. Among all 8 candidate LD blocks, none of them reached genome-wide Bonferroni significance as the sample size is limited. The CHRNA5 cluster in chromosome 15 demonstrates a rather small p-value (cFLM: $p = 5.25 \times 10^{-6}$) and is suggestively associated with nicotine dependence. The "CHRNB3 + CHRNA6" gene cluster (cFLM: $p = 8.67 \times 10^{-4}$) may also be considered for potential association with the phenotypic trait. For these two blocks, p-values calculated by cFLM are much smaller than those calculated by other methods. It is also worth mentioning that both gene clusters have been shown to be associated with nicotine dependence in previous studies (Saccone et al., 2009; Culverhouse et al., 2014). Other candidate LD blocks are not significantly associated with the phenotypic trait in this cohort.

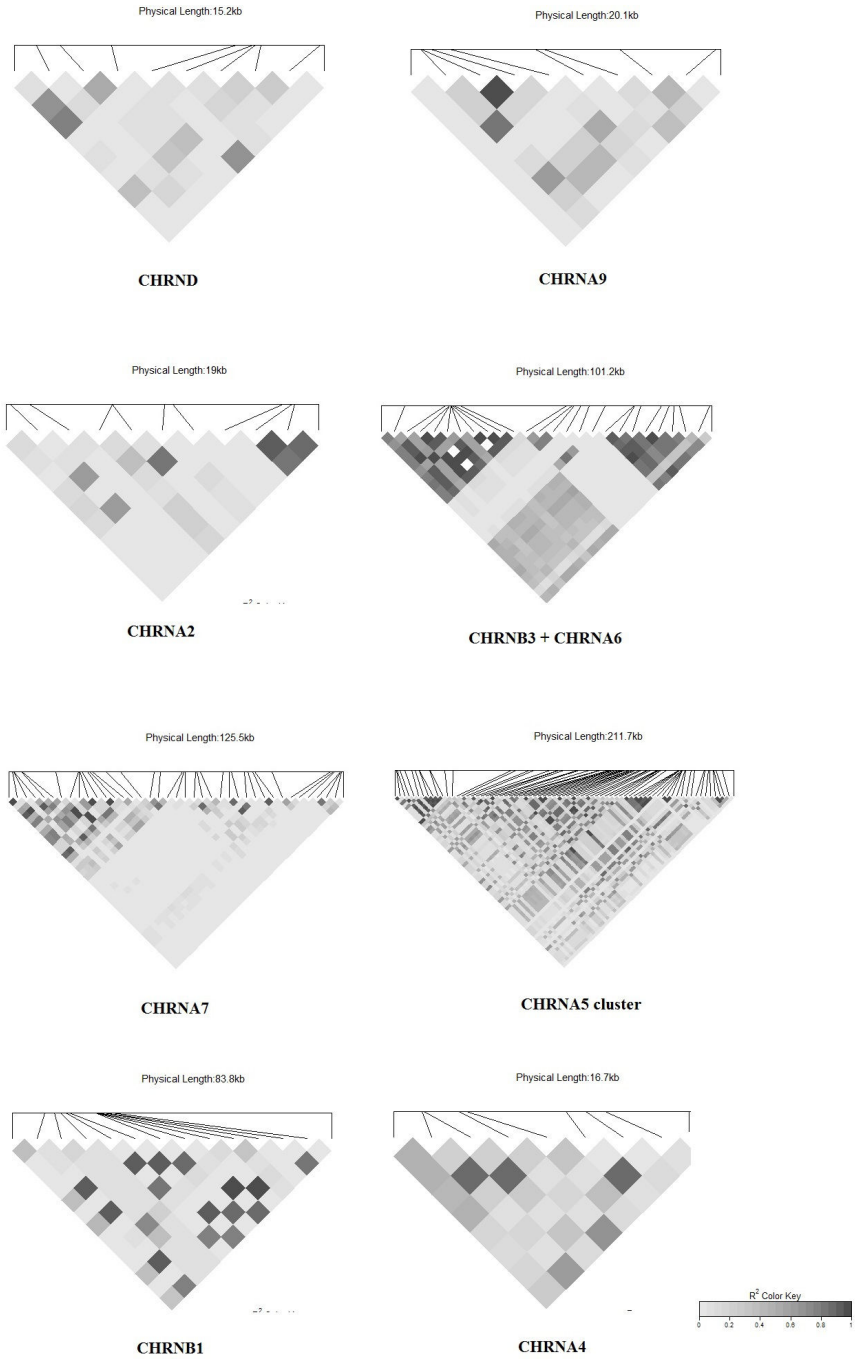

Figure 17: Linkage disequilibrium heat map for candidate LD blocks (gene regions) in COGEND.

Table 6: Association tests for COGEND study based on LD blocks (gene clusters).

LD block (Genes)	CHR	Length (kb)	# of SNPs	p-value cFLM	smAT	SKAT-C	FLM
CHRND	2	15	10	0.0990	0.4818	0.1920	0.1321
CHRNA9	4	20	11	0.4308	0.7580	0.3941	0.3915
CHRNA2	8	19	11	0.7078	0.7596	0.5698	0.8036
CHRNB3 + CHRNA6	8	101	25	8.67e-4	0.0064	0.0013	0.0033
CHRNA7	15	126	39	0.2759	0.3382	0.6078	0.5587
CHRNA5 gene cluster	15	212	72	5.25e-6	1.72e-4	0.0017	8.24e-6
CHRNB1	17	84	14	0.0137	0.3975	0.0274	0.0602
CHRNA4	20	17	9	0.0556	0.1134	0.0255	0.1409

Chapter 3 Constrained Functional Linear Models for Zero-inflated Count Traits

3.1 Zero-inflated Count Responses

3.1.1 Distribution of Count Response

Apart from quantitative and binary phenotypic traits, another type of data often observed in real experiments is count data where the phenotypic trait of interest is measured in counts. For example, sylleptic branches on the main stem in interspecific poplar hybrids (Ma et al., 2008), and the number of cholesterol gallstones formed in mice (Wittenburg et al., 2003) are typical examples of phenotypes measured in counts.

A natural way to analyze regular count data is to fit a Poisson regression distribution to the count response. The probability mass function (PMF) of the Poisson distribution is:

$$f_{\text{Pois}}(k|\lambda) = \frac{\exp(-\lambda)\lambda^k}{k!}. \tag{3.1}$$

The most important feature of the Poisson distribution is that it has equal mean and variance:

$$\text{Mean}_{\text{Pois}}(k) = \text{Var}_{\text{Pois}}(k) = \lambda.$$

The log link is usually used to model the Poisson distribution parameter λ, say,

$$\log(\lambda_i) = x_i\beta.$$

However, since the Poisson distribution has the restriction that variance has to be equal to the mean, the problem of overdispersion will happen when the actual variance is greater than the actual mean in real count responses. In the case of highly right-skewed response, the problem is more prominent. If dispersion occurs, ignoring it will result in biased parameter estimates, which may lead to incorrect conclusions and inferences The most common alternative to Poisson distribution is the negative binomial (NB) Distribution. The negative binomial distribution is a generalized Poisson distribution by introducing a dispersion parameter. From another perspective, it is a continuous mixture of Poisson distributions where the mixing distribution of the Poisson rate follows a gamma process:

$$\begin{aligned}
f_{NB}(k;\phi,\mu) &= \int_0^\infty f_{\text{Pois}(\lambda)}(k)\cdot f_{Gamma(\phi,\frac{\phi}{\mu})}(\lambda)d\lambda \\
&= \int_0^\infty \frac{\lambda^k}{k!}e^{-\lambda}\cdot\lambda^{\phi-1}\frac{e^{-\lambda\phi/\mu}}{(\frac{\mu}{\phi})^\phi\Gamma(\phi)}d\lambda \\
&= \frac{\phi^\phi\mu^{-\phi}}{k!\Gamma(\phi)}\int_0^\infty \lambda^{\phi+k-1}e^{-\lambda/(\frac{\mu}{\phi+\mu})} \\
&= \frac{\phi^\phi\mu^{-\phi}}{k!\Gamma(\phi)}(\frac{\mu}{\phi+\mu})^{\phi+k}\Gamma(\phi+k) \\
&= \frac{\Gamma(\phi+k)}{k!\Gamma(\phi)}(\frac{\mu}{\phi+\mu})^k(\frac{\phi}{\phi+\mu})^\phi,
\end{aligned} \tag{3.2}$$

where $\mu > 0, \phi > 0, k = 0, 1, 2, \ldots$. Note that the mean and variance of the negative binomial distribution is derived as:

$$\text{Mean}_{\text{NB}}(k) = \mu,$$

$$\text{Var}_{\text{NB}}(k) = \mu(\frac{\phi + \mu}{\phi}) = \mu + \frac{\mu^2}{\phi} > \mu. \tag{3.3}$$

Therefore, the variance in the negative binomial distribution is always greater than the mean. In this case the problem of overdispersion can be well addressed. The dispersion parameter ϕ is introduced to measure the dispersion of a count process, which is often treated as a nuisance parameter. The log link is also commonly used to model the mean μ in negative binomial regression, i.e,

$$\log(\mu_i) = x_i\beta. \tag{3.4}$$

The dispersion parameter ϕ is usually treated as a nuisance parameter in estimation.

For both the Poisson and the negative binomial models, we can use a simple link function to model the phenotypic trait. By employing methods discussed in the previous chapter, we can perform functional linear regression analyses in GWAS for count traits, i.e. FLM-Poisson and FLM-NB model can be fitted when the count traits obey such distributions.

3.1.2 Distribution of Zero-inflated Count Response

In several cases, count data often have excessive number of zero outcomes than are expected in Poisson or negative binomial distribution. For example, in healthcare setting a large proportion of zeroes in psychiatric outpatient service utilization is reported within a year. Dental caries, which happen rarely among the population, often distributed differently from a Poisson or NB distribution. Such zero-inflation phenomenon is a special type of overdispersion. If traditional Poisson or Negative Binomial models were fitted, it is probably incapable of detecting the underlying and unobserved zero-inflated effect and leads to loss of power in detecting such genetic causal variants. Therefore, specific zero-inflated models should be developed to handle zero-inflated data.

To better illustrate the mechanism of zero-inflated count process, we take the number of dental caries as an example. Assume that an unobserved factor, denoted by the latent variable L, caused the "risk" of occurrence of caries to move randomly between two states. A subject is in the "risk-free" state ($L = 1$) if he/she inherently does not have the latent risk of having any caries. On the other hand, a subject is in the "risky" state ($L = 0$) if he/she carries the latent factor that caused dental caries inherently. In the language of probability, the latent variable L follows a Bernoulli distribution. Conditionally, dental caries are not likely to occur in the "risk-free" state, thus the conditional probability of having 0 caries in this sate is 1. On the

other hand, in the "risky" state caries occur conditionally according to a count process, which, still brings the possibility of carrying zero defect. Let Y denote the number of dental caries. The probabilistic distribution of Y is:

$$\begin{aligned} L &\sim \text{Bernoulli}(p); \\ P(Y = 0|L = 1) &= 1; \\ P(Y = k|L = 0) &\sim \textbf{Count Process}, k = 0, 1, \ldots \end{aligned} \tag{3.5}$$

A commonly used model is the zero-inflated Poisson (ZIP) regression, which mixes a distribution that degenerates at zero with a Poisson distribution, by allowing the incorporation of explanatory variables in both the point-mass distribution at zero and the Poisson distribution.

As an alternative to the ZIP model, zero-inflated negative binomial (ZINB) model is beneficial from the nature of NB distribution if the count data continue to suggest additional overdispersion. The ZINB model is obtained by mixing a distribution degenerate at zero with a NB distribution. Using the same notations above, assume that in the "risky" state, the outcome follows a negative binomial (NB) distribution. The probabilistic distribution of Y is:

$$\begin{aligned} L &\sim \text{Bernoulli}(p); \\ P(Y = 0|L = 1) &= 1; \\ P(Y = k|L = 0) &\sim NB(\mu, \phi), k = 0, 1, \ldots \end{aligned} \tag{3.6}$$

where p is the probability of $Z = 1$ in the Bernoulli distribution, μ and ϕ are

the mean and dispersion parameter of the NB distribution, respectively. In a ZINB regression model, let y_i denote the observation of number of caries for subject i, $i = 1, \ldots, n$. Then,

$$y_i \sim \begin{cases} 0, & \text{with probability } p_i; \\ NB(\mu_i, \phi), & \text{with probability } 1 - p_i. \end{cases} \tag{3.7}$$

We can see that the zeroes may come from two sources: the conditionally deterministic distribution and the conditional NB distribution. Thus, the occurrence of dental caries y_i is:

$$y_i = \begin{cases} 0, & \text{with probability } p_i + (1 - p_i)(\frac{\phi}{\phi+\mu_i})^{\phi}; \\ k, & \text{with probability } (1 - p_i)\frac{\Gamma(\phi+k)}{k!\Gamma(\phi)}(\frac{\phi}{\phi+\mu_i})^{\phi}(\frac{\mu_i}{\phi+\mu_i})^{k}, \text{ k = 1, 2, } \ldots \end{cases} \tag{3.8}$$

The expectation-maximization (EM) algorithm is a convenient way to estimate the parameters in mixture models like the ZINB model. To quantify the expected conditional log likelihood we have the following calculations:

$$P(L = 1|y, p^{(t)}, \mu^{(t)}, \phi^{(t)}) = \begin{cases} 0, & y > 0; \\ \frac{p^{(t)}}{p^{(t)}+(1-p^{(t)})(\frac{\phi^{(t)}}{\phi^{(t)}+\mu^{(t)}})^{\phi^{(t)}};}, & y = 0. \end{cases}$$

$$P(L = 0|y, p^{(t)}, \mu^{(t)}, \phi^{(t)}) = \begin{cases} 1, & y > 0; \\ \frac{(1-p^{(t)})(\frac{\phi^{(t)}}{\phi^{(t)}+\mu^{(t)}})^{\phi^{(t)}}}{p^{(t)}+(1-p^{(t)})(\frac{\phi^{(t)}}{\phi^{(t)}+\mu^{(t)}})^{\phi^{(t)}};}, & y = 0. \end{cases}$$

The conditional mean of L given the current estimate of parameters $p^{(t)}$

and $\mu^{(t)}$ is:

$$L^{(t)} = E_{L|y,p^{(t)},\mu^{(t)},\phi^{(t)}}(L) = \begin{cases} 0, & y > 0. \\ \frac{p^{(t)}}{p^{(t)}+(1-p^{(t)})(\frac{\phi^t}{\phi^{(t)}+\mu^{(t)}})^{\phi^{(t)}}}; & y = 0; \end{cases} \tag{3.9}$$

which corresponds to the expected conditional log-likelihood that does not need to be calculated. The M-step in the EM Algorithm only requires terms depending on p when we maximize for p, or only terms depending on μ andϕ if we maximize for μ and ϕ, as these two parts are additive as shown below:

$$\begin{aligned} & E_{L|y,p^{(t)},\mu^{(t)},\phi^{(t)}}[\mathcal{L}(p,\mu,\phi|\boldsymbol{y},\boldsymbol{L})] \\ =& \sum_{i=1}^{n}\sum_{l=0}^{1} P(L_i = l|y_i,p^{(t)},\mu^{(t)},\phi^{(t)})\mathcal{L}(p,\mu,\phi|y_i,L_i) \\ =& (\sum_{i=1}^{n} L_i^{(t)}\log(p) + (1-L_i^{(t)})\log(1-p)) \\ & + (\sum_{i=1}^{n}(1-L_i^{(t)})\log\frac{\Gamma(\phi+y_i)}{y_i!\Gamma(\phi)}(\frac{\phi}{\phi+\mu})^{\phi}(\frac{\mu}{\phi+\mu})^{y_i}). \end{aligned} \tag{3.10}$$

EM Algorithm

E-step For subject $i, i = 1,\ldots,n$, estimate L_i by its conditional mean

$$L_i^{(t)} = E_{L|y_i,p^{(t)},\mu^{(t)},\phi^t}(L_i) = \begin{cases} 0, & y_i > 0; \\ \frac{p^{(t)}}{p^{(t)}+(1-p^{(t)})(\frac{\phi^t}{\phi^{(t)}+\mu^{(t)}})^{\phi^{(t)}}}, & y_i = 0. \end{cases} \tag{3.11}$$

M-step

- Find $p^{(t+1)}$ by maximizing

$$\mathcal{L}_{Ber}(p|\boldsymbol{y}, L^{(t)}) = \sum_{i=1}^{n} L_{l,i}^{(t)} \log(p) + (1 - L_{l,i}^{(t)}) \log(1-p). \tag{3.12}$$

- Find $\mu^{(t+1)}, \phi^{(t+1)} i$ by maximizing

$$\mathcal{L}_{NB}(\mu, \phi|\boldsymbol{y}, L^{(t)}) = \sum_{i=1}^{n} (1 - L_{l,i}^{(t)}) \log \frac{\Gamma(\phi + y_i)}{y_i! \Gamma(\phi)} (\frac{\phi}{\phi + \mu})^{\phi} (\frac{\mu}{\phi + \mu})^{y_i}. \tag{3.13}$$

The maximization can be performed by the Newton-Raphson Algorithm for the two distributions simultaneously.

In regression setting, the Bernoulli probabilities and the negative binomial means can be modeled by their canonical links, i.e. the logit and the log functions. Without strong assumption about genetic components and covariates that may impact the response for these two distributions, the same sets of genetic components and covariates are entered into both Bernoulli model and NB model simultaneously. Using the same notations as in the previous section, the link functions are denoted by:

$$\text{logit}(p_i) = \log(\frac{p_i}{1-p_i}) = \alpha_0^{Ber} + \sum_{u=1}^{q} z_{iu} \alpha_u^{Ber} + \sum_{j=1}^{p} X_i(m_j) \beta^{Ber}(m_j), \tag{3.14}$$

and

$$\log(\mu_i) = \alpha_0^{NB} + \sum_{u=1}^{q} z_{iu} \alpha_u^{NB} + \sum_{j=1}^{p} X_i(m_j) \beta^{NB}(m_j).$$

For whole genome screening purpose, we can set $p = 1$ to perform univariate

analyses. For multi-loci mapping purpose, we set the genotypes in a LD block as input explanatory variables.

3.2 Functional Linear Model with Zero-inflated Negative Binomial Responses

3.2.1 Model Formulation

Similar to that in GFLM, we substitute the unstructured coefficients in the ZINB model with the functional coefficients being represented by a linear combination of B-Splines, as follows:

$$\beta^{Ber}(m) = \sum_{r=1}^{d_{Ber}} \gamma_r B_r(m) = \boldsymbol{B}^{Ber}(m)\boldsymbol{\gamma}^{Ber},$$

$$\beta^{NB}(m) = \sum_{r=1}^{d_{NB}} \gamma_r B_r(m) = \boldsymbol{B}^{NB}(m)\boldsymbol{\gamma}^{NB}. \tag{3.15}$$

That is, Equation (3.15) features the functional coefficients for both the Bernoulli and NB models. Using the same notation as in the above, the FLM-ZINB model is reformulated as:

$$\begin{aligned}\text{logit}(\boldsymbol{p}) &= \begin{pmatrix} \mathbf{1} & \boldsymbol{Z} & \boldsymbol{X} \end{pmatrix} \begin{pmatrix} \alpha_0^{Ber} & \boldsymbol{\alpha}^{Ber} & \boldsymbol{B}^{Ber}\boldsymbol{\gamma}^{Ber}_{\boldsymbol{d_{Ber}\times 1}} \end{pmatrix}^\top \\ &= \begin{pmatrix} \mathbf{1} & \boldsymbol{Z} & \boldsymbol{X}\boldsymbol{B}^{\boldsymbol{Ber}} \end{pmatrix} \begin{pmatrix} \alpha_0^{Ber} & \boldsymbol{\alpha}^{Ber} & \boldsymbol{\gamma}^{Ber}_{\boldsymbol{d_{Ber}\times 1}} \end{pmatrix}^\top = \boldsymbol{U}^{Ber}\boldsymbol{\gamma}^{Ber*},\end{aligned}$$

and

$$\log(\boldsymbol{\mu}) = \begin{pmatrix} \mathbf{1} & \boldsymbol{Z} & \boldsymbol{X} \end{pmatrix} \begin{pmatrix} \alpha_0^{NB} & \boldsymbol{\alpha}^{NB} & \boldsymbol{B}^{NB}\boldsymbol{\gamma}^{NB}_{\boldsymbol{d_{NB}\times 1}} \end{pmatrix}^\top$$
$$= \begin{pmatrix} \mathbf{1} & \boldsymbol{Z} & \boldsymbol{X}\boldsymbol{B}^{\boldsymbol{NB}} \end{pmatrix} \begin{pmatrix} \alpha_0^{NB} & \boldsymbol{\alpha}^{NB} & \boldsymbol{\gamma}^{NB}_{\boldsymbol{d_{NB}\times 1}} \end{pmatrix}^\top = \boldsymbol{U}^{NB}\boldsymbol{\gamma}^{NB*}. \tag{3.16}$$

3.2.2 Estimation and EM Algorithm

After substituting $\boldsymbol{p}$ and $\boldsymbol{\mu}$ in Equation (3.10) by Model (3.16), we can perform parameter estimation using the EM Algorithm to maximize the log-likelihood of the FLM-ZINB as mentioned above. The general implementation is as follows.

EM Algorithm Implementation

1. Let $\hat{L_i^{(t)}} = \frac{p^{(t)}}{p_i^{(t)}+(1-p_i^{(t)})(\frac{\phi^t}{\phi^{(t)}+\mu_i^{(t)}})^{\phi^{(t)}}}\mathbb{I}(y=0)$

2. - Perform logistic regression of $\hat{L_i^{(t)}}$ on U^{Ber} to estimate $\gamma^{(t+1)}_{Ber*}$.
   - Perform weighted negative binomial regression of y on U^{NB} with weights $1-\hat{L_i^{(t)}}$ to obtain estimate $\gamma^{(t+1)}_{NB*}$.

3. Let $\boldsymbol{p}^{(t+1)} = \frac{\exp(U^{Ber}\gamma^{(t+1)}_{Ber*})}{1+\exp(U^{Ber}\gamma^{(t+1)}_{Ber*})}$ and $\boldsymbol{\mu}^{(t+1)} = \exp(U^{NB}\gamma^{(t+1)}_{NB*})$, iterate back to step 1.

The Newton-Raphson Algorithm can be employed for maximization of the two distributions simultaneously.

3.2.3 Hypothesis Test

For hypothesis testing, since we model a block of SNPs simultaneously, the test of association between genetic variants and phenotypic trait will be made at the block level. In the meantime, we have two sets of parameters, one for the latent Bernoulli model and one for the negative binomial model, respectively. The hypothesis test should consider the overall effect of both the Bernoulli and the NB models. That is, we consider the likelihood test on testing whether any of the coefficients in the two parameters sets is non-zero. On the other hand, we attained dimension reduction by changing the estimator of interest from the $2p$-dimensional $\beta^{Ber}(m_j)$ and $\beta^{NB}(m_j), j = 1, \ldots, p$ to the $d_{Ber} + d_{NB}$-dimensional $\boldsymbol{\gamma}^{Ber}_{d_{Ber}\times 1}$ and $\boldsymbol{\gamma}^{NB}_{d_{NB}\times 1}$ by applying the functional coefficients implemented via the B-spline bases. The null hypothesis is equivalent to

$$\begin{aligned} &\mathrm{H}_0: \quad \boldsymbol{\gamma}_r^{Ber} = 0, \boldsymbol{\gamma}_r^{NB} = 0, \text{ for all } r. \\ &\mathrm{H}_a: \quad \text{Not } \mathrm{H}_0. \end{aligned} \tag{3.17}$$

The likelihood ratio test (LRT) will be applied to draw inference about whether a block of SNPs may be associated with the phenotypic trait. Under H_0, the LRT statistic asymptotically follows a $\chi^2_{d_{Ber}+d_{NB}}$ distribution (Chi-square distribution with df $= d_{Ber} + d_{NB}$):

$$\chi^2 = G_0 - G_1 = -2(l_0 - l_1) \xrightarrow{d} \chi^2_{d_{Ber}+d_{NB}}. \tag{3.18}$$

3.3 Constrained Functional Linear Model with Zero-inflated Negative Binomial response

3.3.1 Model Formulation

Similar to cFLM, we express $\beta^{Ber+}(m), \beta^{Ber-}(m), \beta^{NB+}(m), \beta^{NB-}(m)$ in terms of B-spline bases $\boldsymbol{B}^{Ber+}_{1\times d_Ber1}(m)$, $\boldsymbol{B}^{Ber-}_{1\times d_Ber2}(m)$, $\boldsymbol{B}^{NB+}_{1\times d_{NB1}}(m)$, $\boldsymbol{B}^{NB-}_{1\times d_{NB2}}(m)$ and the new coefficient vectors $\boldsymbol{\gamma}^{Ber+}_{d_{Ber1}\times 1}, \boldsymbol{\gamma}^{Ber-}_{d_{Ber2}\times 1}, \boldsymbol{\gamma}^{NB+}_{d_{NB1}\times 1}, \boldsymbol{\gamma}^{NB-}_{d_{NB2}\times 1}$, respectively:

$$\begin{aligned}
&\beta^{Ber+}(m) = \boldsymbol{B}^{Ber+}(m)\boldsymbol{\gamma}^{Ber+}, \beta^{Ber-}(m) = \boldsymbol{B}^{Ber-}(m)\boldsymbol{\gamma}^{Ber-} \\
&\beta^{NB+}(m) = \boldsymbol{B}^{NB+}(m)\boldsymbol{\gamma}^{NB+}, \beta^{NB-}(m) = \boldsymbol{B}^{NB-}(m)\boldsymbol{\gamma}^{NB-}
\end{aligned} \tag{3.19}$$

The ZINB model with constrained functional coefficients (cFLM-ZINB) applies the complimentarity constraints on both the latent Bernoulli model and the NB model, respectively. Based on Equation (3.16), the formulation of cFLM is as follows:

$$\begin{aligned}
\text{logit}(\boldsymbol{p}) &= \begin{pmatrix} \mathbf{1} & \boldsymbol{Z} & \boldsymbol{X}\boldsymbol{B}^{Ber+} & \boldsymbol{X}\boldsymbol{B}^{Ber-} \end{pmatrix} \begin{pmatrix} \alpha_0^{Ber} & \boldsymbol{\alpha}^{Ber} & \boldsymbol{\gamma}^{Ber+} & \boldsymbol{\gamma}^{Ber-} \end{pmatrix}^\top \\
&= \begin{pmatrix} (\mathbf{1} & \boldsymbol{Z}) & \boldsymbol{X}\boldsymbol{B}^{Ber+} & \boldsymbol{X}\boldsymbol{B}^{Ber-} \end{pmatrix} \begin{pmatrix} \gamma_0^{Ber} & \boldsymbol{\gamma}^{Ber+} & \boldsymbol{\gamma}^{Ber-} \end{pmatrix}^\top = \boldsymbol{U}^{Ber}\boldsymbol{\gamma}^{Ber*} \\
&\text{subject to } \boldsymbol{B}^{Ber+}\boldsymbol{\gamma}^{Ber+} \geq \mathbf{0}, \boldsymbol{B}^{Ber-}\boldsymbol{\gamma}^{Ber-} \leq \mathbf{0}, \\
&\qquad (\boldsymbol{B}^{Ber+}\boldsymbol{\gamma}^{Ber+}) \circ (\boldsymbol{B}^{Ber-}\boldsymbol{\gamma}^{Ber-}) = \mathbf{0},
\end{aligned}$$

and

$$\begin{aligned}\log(\boldsymbol{\mu}) &= \begin{pmatrix} \mathbf{1} & \boldsymbol{Z} & \boldsymbol{X}\boldsymbol{B}^{NB+} & \boldsymbol{X}\boldsymbol{B}^{NB-} \end{pmatrix} \begin{pmatrix} \alpha_0^{NB} & \boldsymbol{\alpha}^{NB} & \boldsymbol{\gamma}^{NB+} & \boldsymbol{\gamma}^{NB-} \end{pmatrix}^\top \\ &= \begin{pmatrix} (\mathbf{1} & \boldsymbol{Z}) & \boldsymbol{X}\boldsymbol{B}^{NB+} & \boldsymbol{X}\boldsymbol{B}^{NB-} \end{pmatrix} \begin{pmatrix} \gamma_0^{NB} & \boldsymbol{\gamma}^{NB+} & \boldsymbol{\gamma}^{NB-} \end{pmatrix}^\top = \boldsymbol{U}^{NB}\boldsymbol{\gamma}^{NB*} \\ &\text{subject to } \boldsymbol{B}^{NB+}\boldsymbol{\gamma}^{NB+} \geq \mathbf{0}, \boldsymbol{B}^{NB-}\boldsymbol{\gamma}^{NB-} \leq \mathbf{0}, \\ &\qquad (\boldsymbol{B}^{NB+}\boldsymbol{\gamma}^{NB+}) \circ (\boldsymbol{B}^{NB-}\boldsymbol{\gamma}^{NB-}) = \mathbf{0}. \end{aligned} \tag{3.20}$$

The fitted coefficients $(\boldsymbol{\gamma}^{Ber+}, \boldsymbol{\gamma}^{Ber-})$ and $(\boldsymbol{\gamma}^{NB+}, \boldsymbol{\gamma}^{NB-})$ represent genetic effect found associated with the latent Bernoulli model and the negative binomial model, respectively.

3.3.2 Estimation

As to parameter estimation, although the EM algorithm derived in the previous sections are still applicable to be extended to the cFLM-ZINB by adding the complimentarity constraints in the estimation process. However, since the complimentarity constraints are nonlinear constraints and the maximization process is a nonlinear constrained optimization problem, the speed for performing two sets of nonlinear constrained optimization in EM algorithm is not time-efficient. In this case, we found it more sensible to directly maximize the log-likelihood of cFLM-ZINB model because it only requires one set of nonlinear optimization. The log-likelihood function for the cFLM-

ZINB model is

$$
\begin{aligned}
\mathcal{L}_{\text{cFLM-ZINB}} &= \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\gamma}^{Ber*}, \boldsymbol{\gamma}^{NB*}) \\
&= \log \prod_{i=0}^{n} \{(p_i + (1-p_i)(\frac{\phi}{\phi+\mu_i})^{\phi}) \mathbb{I}(y_i = 0) \\
&\quad + ((1-p_i)\frac{\Gamma(\phi+y_i)}{y_i!\Gamma(\phi)}(\frac{\phi}{\phi+\mu_i})^{\phi}(\frac{\mu_i}{\phi+\mu_i})^{y_i}) \mathbb{I}(y_i > 0)\} \\
&= \log \prod_{i:y_i=0} (\frac{\exp(U_i^{Ber}\gamma^{Ber*})}{1+\exp(U_i^{Ber}\gamma^{Ber*})} + \frac{1}{1+\exp(U_i^{Ber}\gamma^{Ber*})}(\frac{\phi}{\phi+\exp(U_i^{NB}\gamma^{NB*})})^{\phi}) \\
&\quad + \log \prod_{i:y_i>0} (\frac{1}{1+\exp(U_i^{Ber}\gamma^{Ber*})} \frac{\Gamma(\phi+y_i)}{y_i!\Gamma(\phi)} \frac{\phi^{\phi}\exp(U_i^{NB}\gamma^{NB*}y_i)}{(\phi+\exp(U_i^{NB}\gamma^{NB*}))^{\phi+y_i}}) \\
\text{subject to } & \boldsymbol{B}^{Ber+}\boldsymbol{\gamma}^{Ber+} \geq \mathbf{0}, \boldsymbol{B}^{Ber-}\boldsymbol{\gamma}^{Ber-} \leq \mathbf{0}, (\boldsymbol{B}^{Ber+}\boldsymbol{\gamma}^{Ber+}) \circ (\boldsymbol{B}^{Ber-}\boldsymbol{\gamma}^{Ber-}) = \mathbf{0} \\
& \boldsymbol{B}^{NB+}\boldsymbol{\gamma}^{NB+} \geq \mathbf{0}, \boldsymbol{B}^{NB-}\boldsymbol{\gamma}^{NB-} \leq \mathbf{0}, (\boldsymbol{B}^{NB+}\boldsymbol{\gamma}^{NB+}) \circ (\boldsymbol{B}^{NB-}\boldsymbol{\gamma}^{NB-}) = \mathbf{0}.
\end{aligned}
\tag{3.21}
$$

In order to obtain the MLEs for parameters, the following nonlinear constrained optimization problem with linear/nonlinear constraints need to be solved:

$$
\begin{aligned}
\text{maximize} \quad & \mathcal{L}_{\text{cFLM-ZINB}} \\
\text{subject to} \quad & \boldsymbol{B}^{Ber+}\boldsymbol{\gamma}^{Ber+} \geq \mathbf{0}, \boldsymbol{B}^{Ber-}\boldsymbol{\gamma}^{Ber-} \leq \mathbf{0}, (\boldsymbol{B}^{Ber+}\boldsymbol{\gamma}^{Ber+}) \circ (\boldsymbol{B}^{Ber-}\boldsymbol{\gamma}^{Ber-}) = \mathbf{0} \\
& \boldsymbol{B}^{NB+}\boldsymbol{\gamma}^{NB+} \geq \mathbf{0}, \boldsymbol{B}^{NB-}\boldsymbol{\gamma}^{NB-} \leq \mathbf{0}, (\boldsymbol{B}^{NB+}\boldsymbol{\gamma}^{NB+}) \circ (\boldsymbol{B}^{NB-}\boldsymbol{\gamma}^{NB-}) = \mathbf{0}.
\end{aligned}
\tag{3.22}
$$

The Augmented Lagrangian Algorithm (ALA) will be applied to this con-

strained maximization problem, as illustrated in the previous chapter.

3.3.3 Hypothesis Test

Similar to what has been done for the FLM, we perform likelihood ratio test to investigate the overall genetic effects represented by a block of SNPs in contiguous genomic regions. However, the null and alternative hypotheses in (3.23) are upon revision with regard to the new parameter space and imposed constraints:

$$\begin{aligned}
\mathrm{H}_0: \quad & \boldsymbol{\gamma}^{Ber+}_{d_{Ber1}\times 1} = \mathbf{0} \text{ and } \boldsymbol{\gamma}^{Ber-}_{d_{Ber2}\times 1} = \mathbf{0}, \boldsymbol{\gamma}^{NB+}_{d_{NB1}\times 1} = \mathbf{0} \text{ and } \boldsymbol{\gamma}^{NB-}_{d_{Ber2}\times 1} = \mathbf{0}. \\
\mathrm{H}_a: \quad & \boldsymbol{B}^{Ber+}\boldsymbol{\gamma}^{Ber+} \geq \mathbf{0}, \boldsymbol{B}^{Ber-}\boldsymbol{\gamma}^{Ber-} \leq \mathbf{0}, (\boldsymbol{B}^{Ber+}\boldsymbol{\gamma}^{Ber+}) \circ (\boldsymbol{B}^{Ber-}\boldsymbol{\gamma}^{Ber-}) = \mathbf{0} \\
& \boldsymbol{B}^{NB+}\boldsymbol{\gamma}^{NB+} \geq \mathbf{0}, \boldsymbol{B}^{NB-}\boldsymbol{\gamma}^{NB-} \leq \mathbf{0}, (\boldsymbol{B}^{NB+}\boldsymbol{\gamma}^{NB+}) \circ (\boldsymbol{B}^{NB-}\boldsymbol{\gamma}^{NB-}) = \mathbf{0}.
\end{aligned} \tag{3.23}$$

In the cFLM-ZINB model, the parameter estimation consists of two independent sets of constraints. Therefore, the hypothesis test is constructed with regard to the latent Bernoulli and NB models, respectively. A mixture of chi-square distribution (chi-bar square test) is performed for each part. The overall LRT statistic therefore follows a mixture of the mixtures of chi-square distribution, which is slightly different from the that proposed in the previous chapter. To simplify our setting, we set the number of B-spline bases in positive and negative genetic effect coefficient function to be the same, i.e. $d_{Ber1} = d_{Ber2} = d_{Ber}$ and $d_{NB1} = d_{NB2} = d_{NB}$.

The likelihood ratio test statistic asymptotically follows a mixture of the mixtures of chi-square distributions with mixing probabilities w_j^{Ber} and w_k^{NB}

such that $\sum_{j=0}^{d_{Ber}} w_j^{Ber} = 1$ and $\sum_{k=0}^{d_{NB}} w_k^{NB} = 1$, denoted as:

$$\bar{\chi}^2_{ZINB} = G_0 - G_1 = -2(l_0 - l_1) \xrightarrow{d} \sum_{j=0}^{d_{Ber}} w_j^{Ber} \chi_j^2 + \sum_{k=0}^{d_{NB}} w_k^{NB} \chi_k^2. \tag{3.24}$$

The p-value of the $\bar{\chi}^2$ test statistic is defined as

$$P(\bar{\chi}^2_{ZINB} \geq c^2) = \sum_{j,k} (w_j^{Ber} w_k^{NB}) P(\chi^2_{j+k} \geq c^2). \tag{3.25}$$

The mixing probabilities can be calculated using Monte Carlo Techniques. The algorithm is similar to that described in the previous chapter.

3.4 Simulation Studies

3.4.1 Parameter Settings

We used similar simulation approaches similar to those introduced in the previous chapter. However, since two sets of parameters representing genetic effects affecting the latent Bernoulli and NB models need to be estimated, respectively, we also compared our proposed models' performances with traditional count data model. Therefore, in addition to smAT-ZINB, FLM-ZINB, cFLM-ZINB, we calculated the power for the FLM-NB model, in order to demonstrate that the zero-inflated model is superior for being able to model count outcomes with excessive zeroes.

For the genotypic file, similarly we simulated both random LD blocks and the CHRNA7 gene structures to mimic the real gene analyses. For the phenotypic file, we simulated ZINB outcomes conditional on causal genotypes based on the following model:

$$\begin{aligned} \text{logit}(p_i) = \log \frac{p_i}{1-p_i} &= \alpha_0^{Ber} + \boldsymbol{X_i}^\top \boldsymbol{\beta_{causal}^{Ber}} \\ \log(\mu_i) &= \alpha_0^{NB} + \boldsymbol{X_i}^\top \boldsymbol{\beta_{causal}^{NB}}. \end{aligned} \tag{3.26}$$

The outcomes were simulated following the latent mixture model as discussed in the previous sections.

Sample size was set to range from 500 to 2,000. β_{causal} was set to zero under the null hypothesis in type I error simulation. Since the computational

burden for ZINB model is much higher than that for binary outcome model, we reduced the replicates to 1,000 times for each scenario. The type I error rates were investigated under small genome-wide thresholds (nominal $\alpha = 0.05, 0.01$ and 0.005).

For assessment of empirical power, we also used similar settings as in the previous chapter. However, we examined two general scenarios where the effects happen in either the latent Bernoulli or the negative binomial models. Then, for each general scenario, we first set that only one causal SNP was located in the LD block, having varying regression coefficient β^{Ber}_{causal} or β^{NB}_{causal}. Then we considered set two causal loci with reversed sign effects in the LD block. The two causal loci chosen were weakly correlated ($r^2 < 0.01$). The corresponding regression coefficients were set to $(\beta^{Ber}_{causal1}, \beta^{Ber}_{causal2})$, or $(\beta^{NB}_{causal1}, \beta^{NB}_{causal2})$ where $\beta_{causal1} = -\beta_{causal2}$.

We ran pivotal analyses to determine the suitable setting of intercept for ZINB outcome simulation. By looking at Table 7, we found that only when $\alpha^{Ber}_0 = 0.0$ and $\alpha^{NB}_0 = 1.0$ the model had correct estimates of regression parameters. This phenomenon is probably triggered by the setting that the latent zero-inflated Bernoulli distribution should not interfere with the NB process when it's already adequately right skewed. Based on the pivotal analyses, the intercepts were always set to $\alpha^{Ber}_0 = 0.0$ and $\alpha^{NB}_0 = 1.0$ in any subsequent simulations.

Table 7: Pivotal analyses of intercept settings for ZINB model simulation

True Value				Estimate (Variance)				Nominal α	
α_0^{Ber}	α_0^{NB}	β_{causal}^{Ber}	β_{causal}^{NB}	$\hat{\alpha}_0^{Ber}$	$\hat{\alpha}_0^{NB}$	$\hat{\beta}_{causal}^{Ber}$	$\hat{\beta}_{causal}^{NB}$	0.05	0.01
0.0	0.0	0.0	0.0	-1.4379 (16.423)	-0.0250 (0.0069)	-0.5676 (32.141)	0.0196 (0.0086)	0.060	0.010
0.0	1.0	0.0	0.0	-0.0152 (0.0274)	0.9949 (0.0076)	-0.0005 (0.0114)	0.0054 (0.0046)	0.045	0.010
1.0	0.0	0.0	0.0	-1.2259 (453.98)	-0.0006 (0.0132)	-0.1268 (120.74)	-0.0356 (0.1316)	0.040	0.005
0.0	0.0	0.2	0.0	-0.3091 (1.9862)	-0.0447 (0.0593)	0.3124 (0.3635)	-0.0096 (0.0071)	-	-
0.0	1.0	0.2	0.0	-0.0003 (0.0239)	1.0025 (0.0067)	0.2021 (0.0099)	0.0022 (0.0047)	-	-
1.0	0.0	0.2	0.0	0.8489 (1.3093)	-0.0265 (0.1276)	0.1610 (0.2628)	0.0093 (0.0192)	-	-
0.0	0.0	0.0	0.2	-2.4829 (41.616)	-0.0580 (0.0785)	-0.4224 (8.3156)	0.1936 (0.0080)	-	-
0.0	1.0	0.0	0.2	-0.0158 (0.0288)	0.9864 (0.0079)	0.0087 (0.0077)	0.1966 (0.0036)	-	-
1.0	0.0	0.0	0.2	0.8619 (1.3146)	-0.0161 (0.1002)	-0.0102 (0.1813)	0.1894 (0.0181)	-	-

Table 8: Parameter settings for power evaluation with ZINB outcomes

Scenario	Outcome	Block structure	Genetic effect process	# of Causal SNPs
SC9	ZINB	random	Bernoulli	1
SC10	ZINB	random	Bernoulli	2
SC11	ZINB	random	NB	1
SC12	ZINB	random	NB	2
SC13	ZINB	CHRNA7	Bernoulli	1
SC14	ZINB	CHRNA7	Bernoulli	2
SC15	ZINB	CHRNA7	NB	1
SC16	ZINB	CHRNA7	NB	2

Since our main purpose was to check if the LD structure modeled by smooth functions would be able to contribute to the gain of empirical power, we only examined scenarios when causal SNPs were removed in order to avoid overburdening our computation resources. We ran 100 replicates for each scenario. A p-value smaller than 0.05 would be used to declare significant. Table 8 summarizes the parameter settings in different simulation scenarios.

In terms of functional parameters, we used the same strategies described in the previous chapter. For convenience, we set the number of spline bases the same in both the latent Bernoulli model and the negative binomial model.

3.4.2 Simulation using Random LD Blocks

The generation of genotypic file used the same parameter settings as in the simulation with binary outcomes, which tried to mimic the decay of LD in a idealized LD block. Both type I error and empirical power were investigated under such settings. Results in table 9 and Figure 18 demonstrate that the proposed cFLM-ZINB model can maintain the type I error under the random sampling scheme for genetic files.

Table 9: Type I error simulation using cFLM for ZINB outcomes based on random LD blocks.

nominal α	N=500	N=1000	N=1500	N=2000
0.05	0.055	0.046	0.051	0.040
0.01	0.009	0.009	0.012	0.007
0.005	0.003	0.007	0.008	0.004

Evaluation of empirical power was based on settings where regression coefficients ranged from 0.1 to 0.5 and sample size ranged from 500 to 2,000. Under the first setting when the genetic effect was set to occur in the latent Bernoulli model, we can observe the apparent failure of using a negative binomial (NB) regression model, by looking at the significantly lower power when using the FLM-NB model in SC9 and SC10 (Figure 19 and 20). When the genetic effect was set to occur in the NB model under SC11 and SC12 (Figure 21 and 22), the FLM-NB model is still underpowered but not as significantly as that in SC9 and SC10. While using ZINB models, smAT-ZINB, FLM-ZINB and cFLM-ZINB demonstrated similar powers, which were all

superior to the FLM-NB regression model. The cFLM-ZINB model generally had the greatest power among these ZINB models, under both scenarios when one causal locus or two causal loci were set in the LD block.

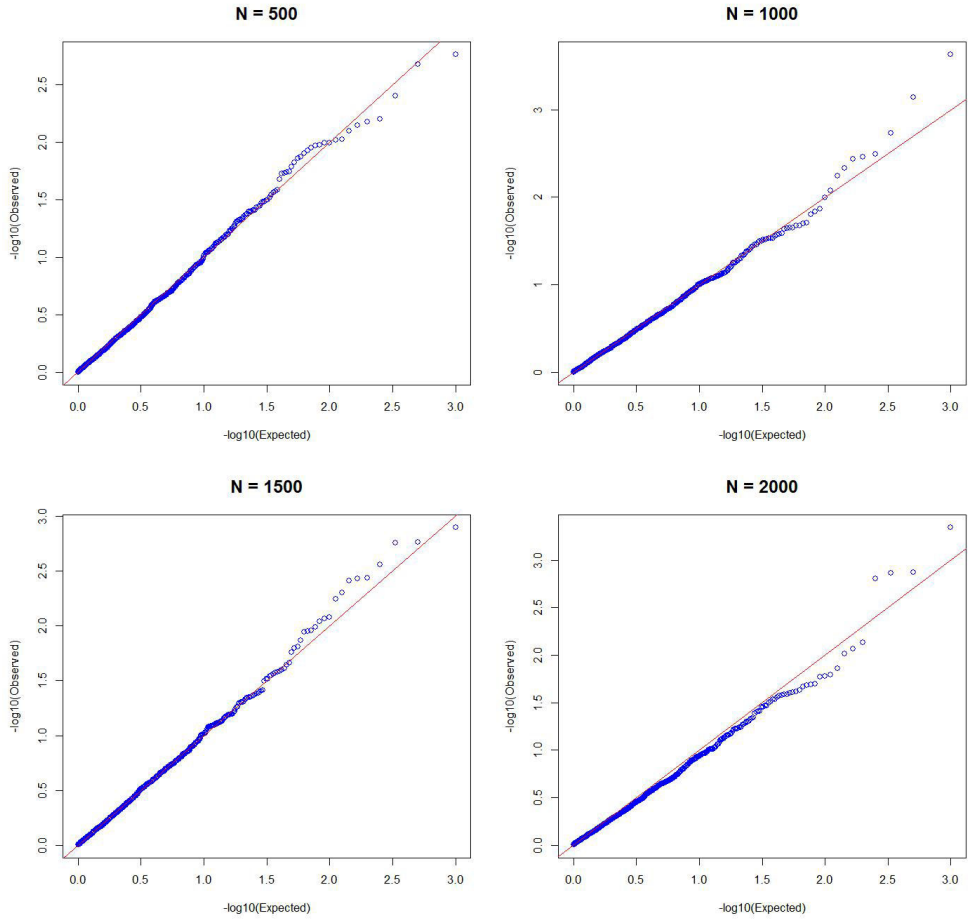

Figure 18: Type I error simulation using cFLM for ZINB outcomes based on random LD blocks, Q-Q plots of p-values.

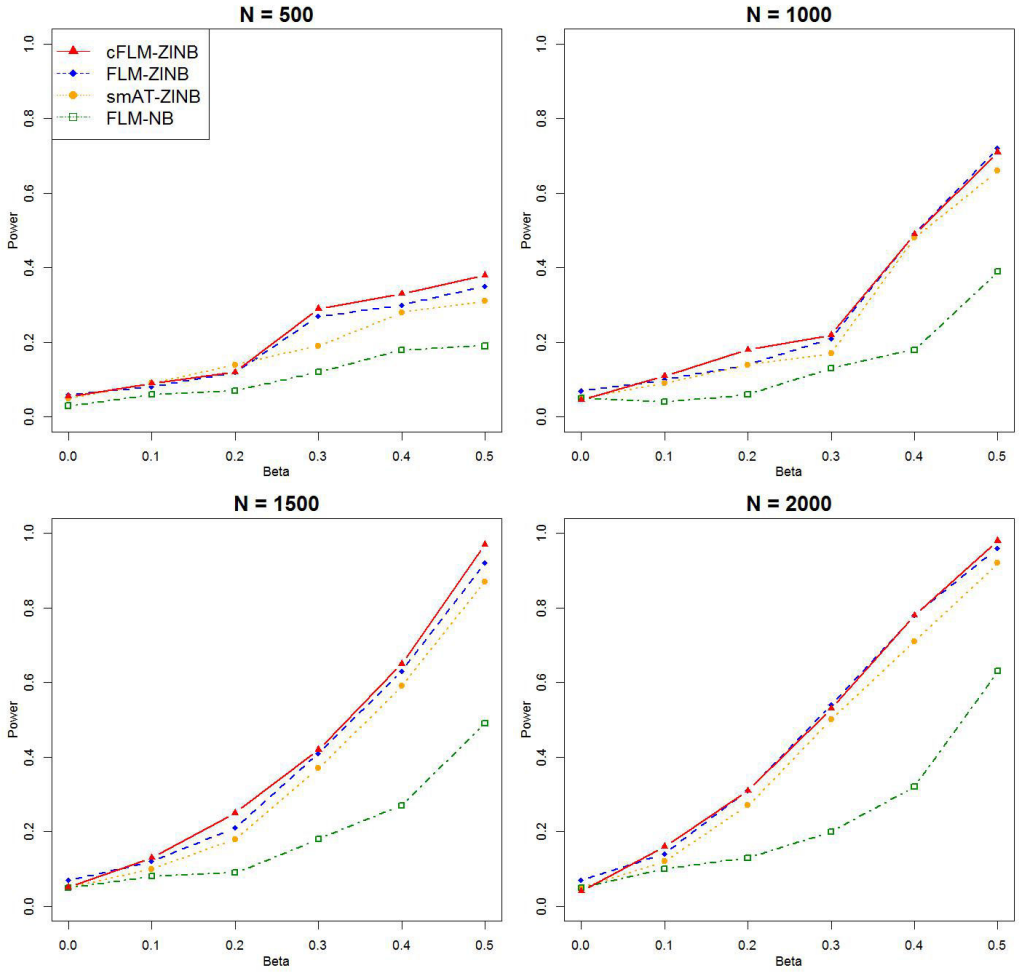

Figure 19: Power simulation for ZINB outcomes based on random LD blocks: single causal locus, effect in latent Bernoulli distribution, SC9.

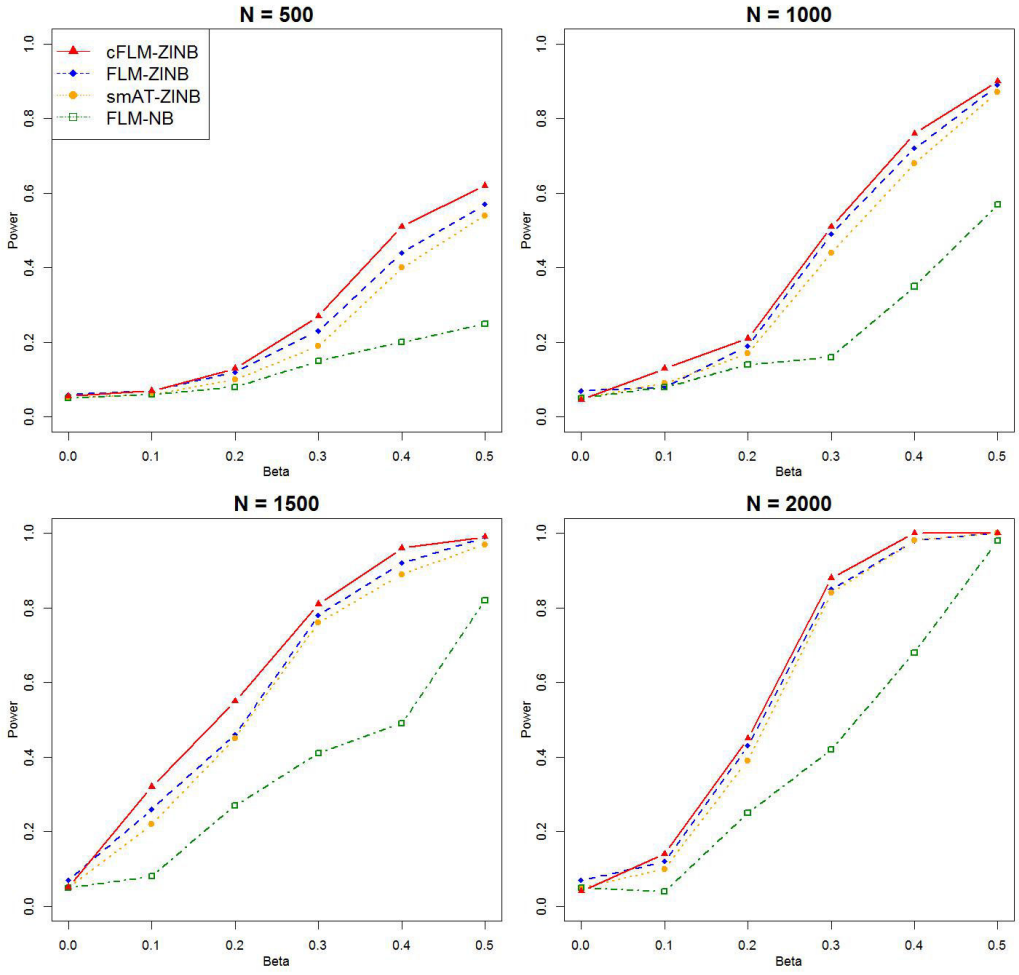

Figure 20: Power simulation for ZINB outcomes based on random LD blocks: two causal loci, effect in latent Bernoulli distribution, SC10.

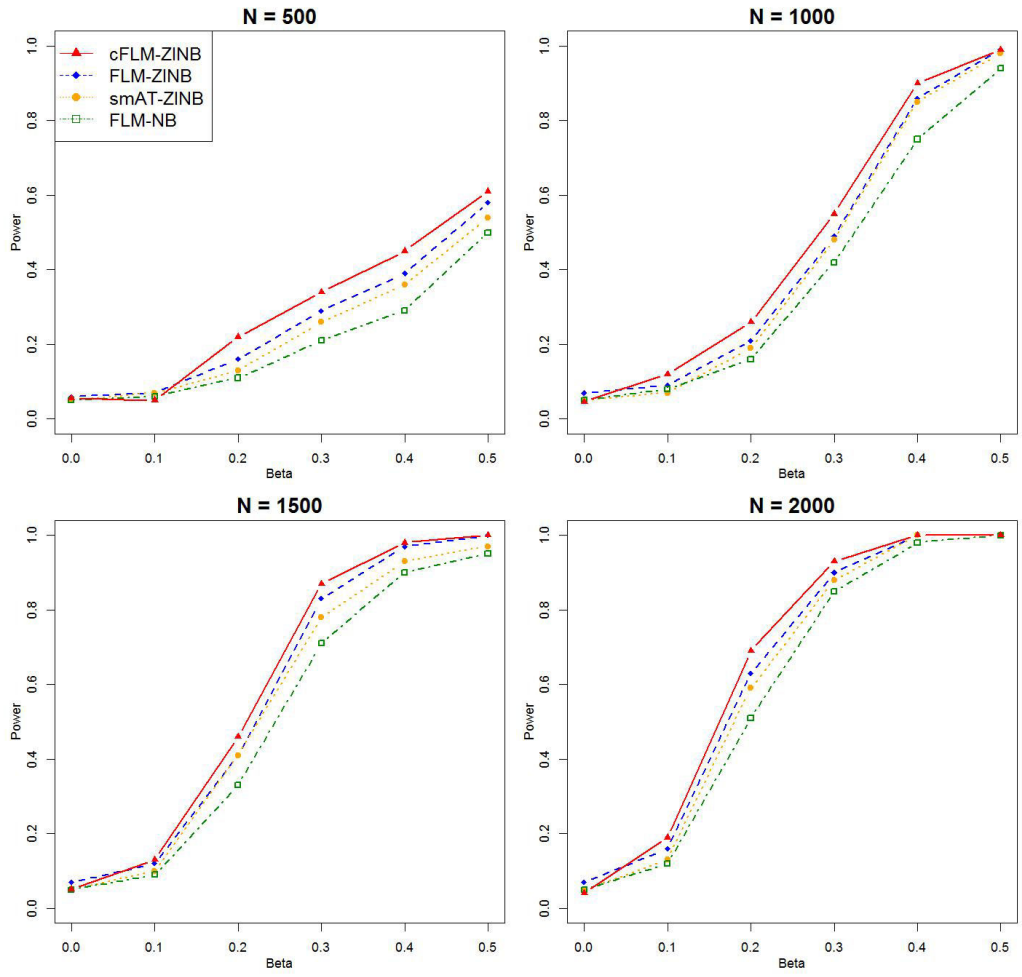

Figure 21: Power simulation for ZINB outcomes based on random LD blocks: single causal locus, effect in NB distribution, SC11.

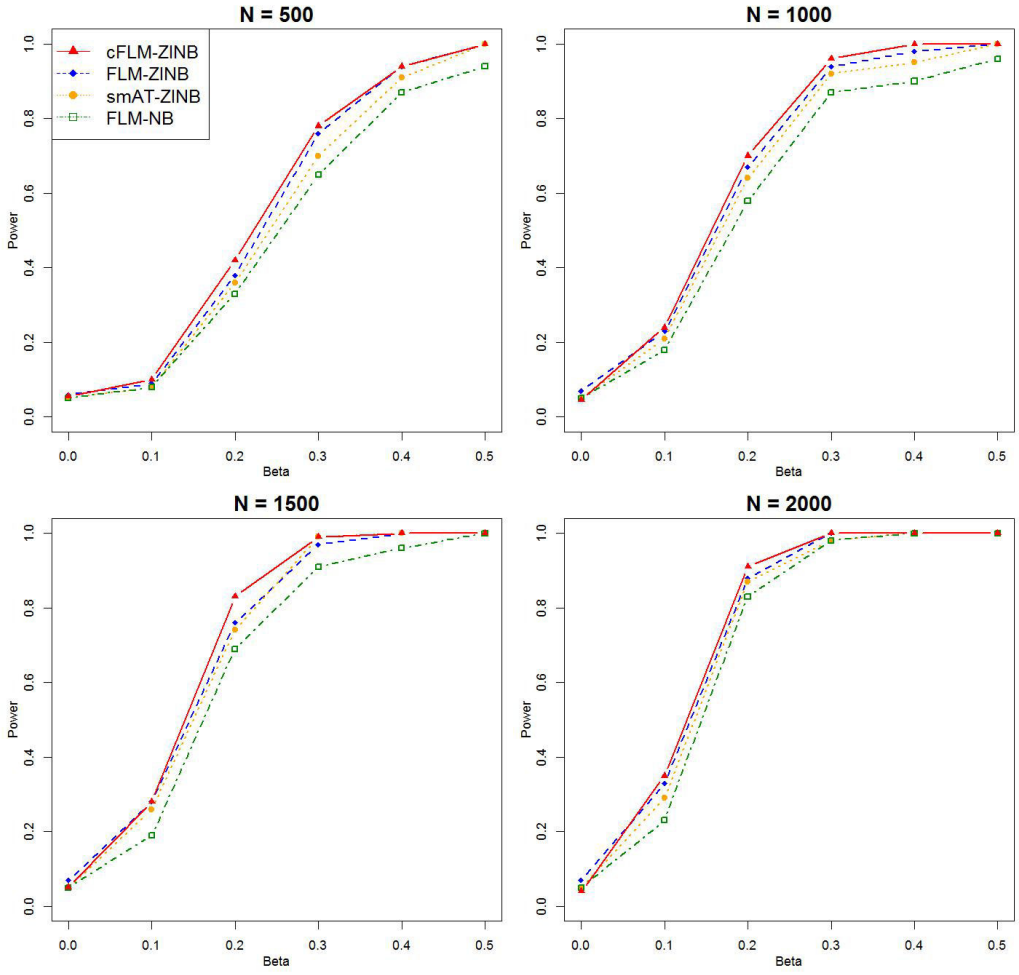

Figure 22: Power simulation for ZINB outcomes based on random LD blocks: two causal loci, effect in NB distribution, SC12.

3.4.3 Simulation using the CHRNA7 gene (15q13.3)

The entry of genotypic file in this set of simulations used information from the CHRNA7 gene. Table 10 and Figure 23 show that the proposed cFLM-ZINB can control type I error for simulations in this setting.

Table 10: Type I error simulation using cFLM-ZINB for ZINB outcomes based on the CHRNA7 gene.

nominal α	N=500	N=1000	N=1500	N=2000
0.05	0.037	0.046	0.045	0.042
0.01	0.011	0.011	0.006	0.008
0.005	0.005	0.007	0.002	0.006

Under the first scenario setting when the genetic effect was set to occur in the latent Bernoulli model, same as simulation using random LD blocks, we can observe significant failure of negative binomial regression models when the FLM-NB model simulations were run in SC13 and SC14 (Figure 24 and 25). This fortifies our assumption that the usnig a simple NB distribution will lead to loss of power when modeling genetic effects affecting excess zero in zero-inflated count process. While using ZINB models, similar performances were observed for smAT-ZINB, FLM-ZINB and cFLM-ZINB in SC13 - SC16 (Figure 24, 25, 26 and 27). The figures demonstrate that the ZINB models are more advantageous than the NB regression model. Generally, the cFLM-ZINB model had the best performance among these ZINB models under scenarios when one causal locus and two causal loci were entered into the LD block.

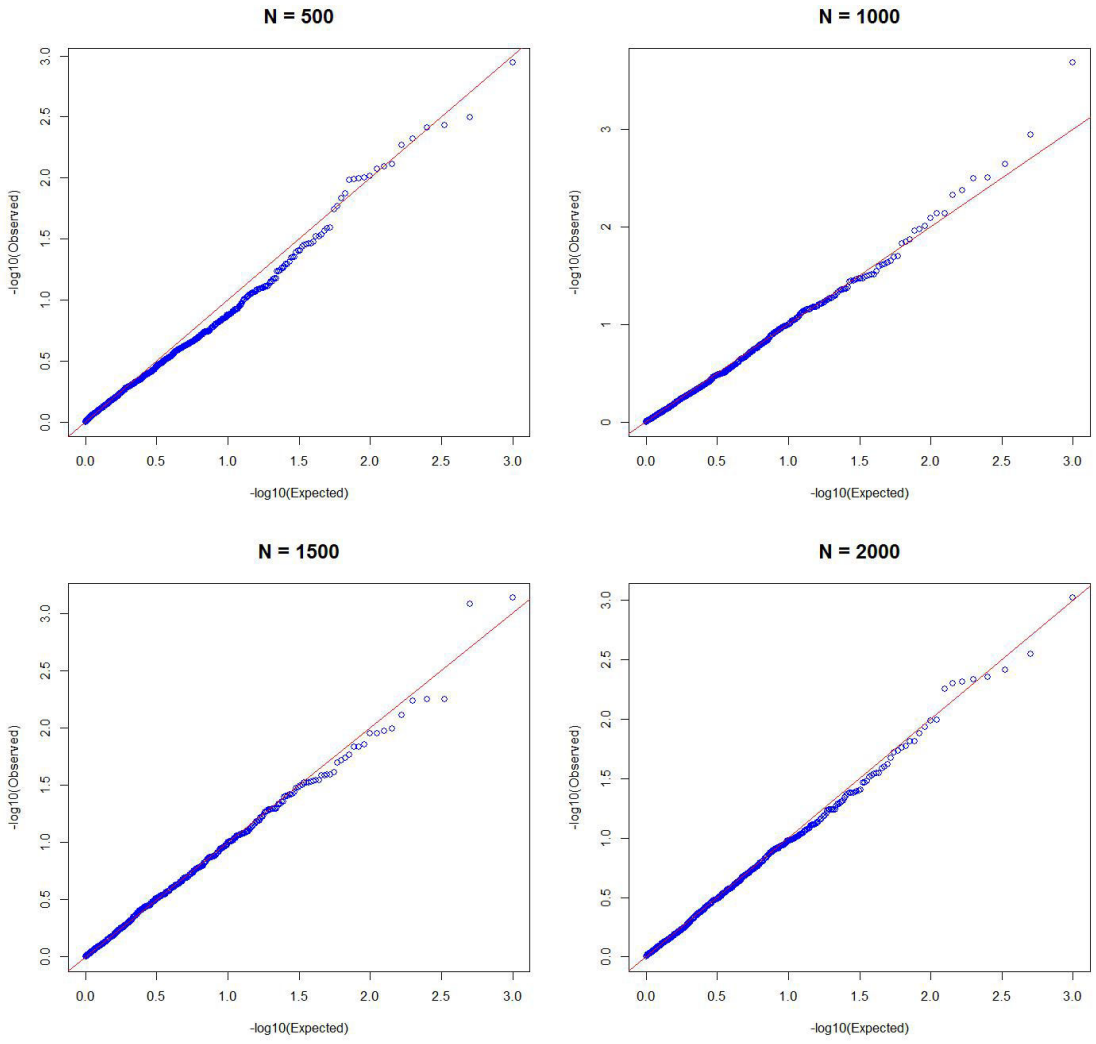

Figure 23: Type I error simulation using cFLM-ZINB for ZINB outcomes based on the CHRNA7, Q-Q plots of p-values.

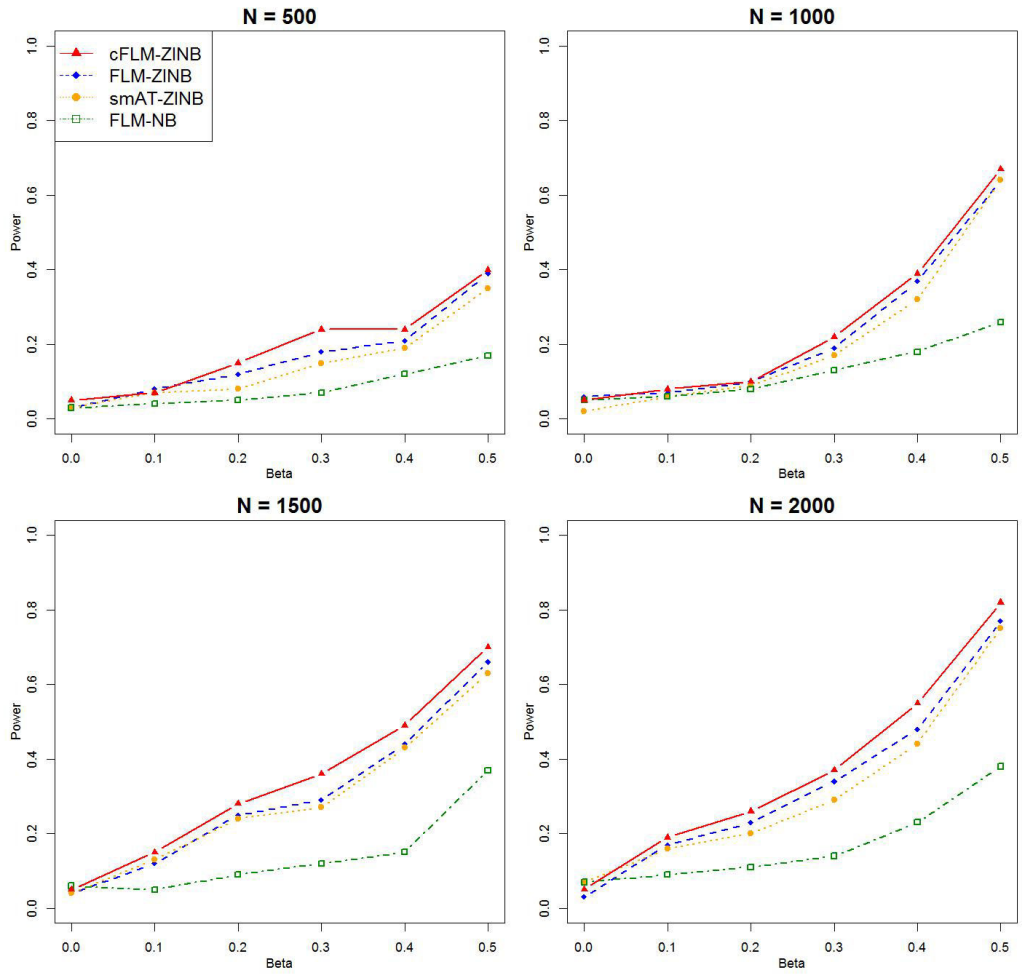

Figure 24: Power simulation for ZINB outcomes based on the CHRNA7 gene: single causal locus, effect in latent Bernoulli distribution, SC13.

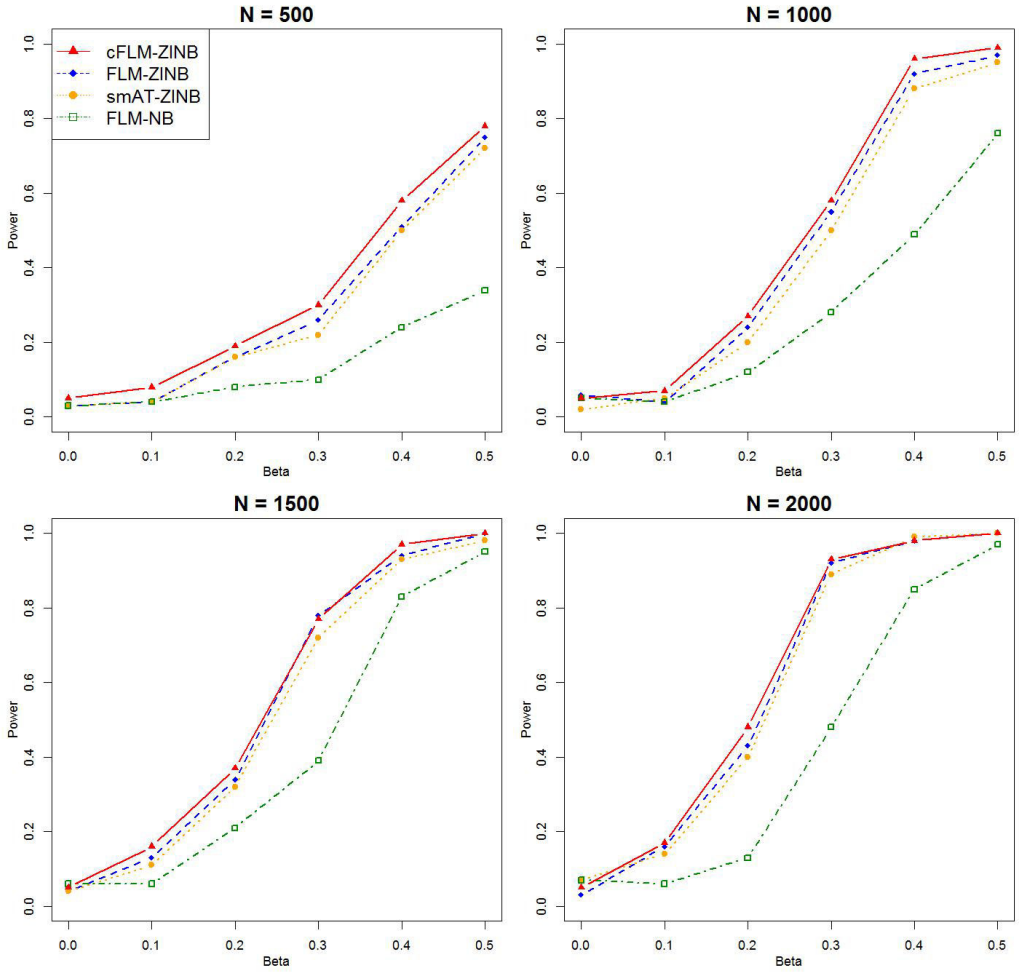

Figure 25: Power simulation for ZINB outcomes based on the CHRNA7 gene: two causal loci, effect in latent Bernoulli distribution, SC14.

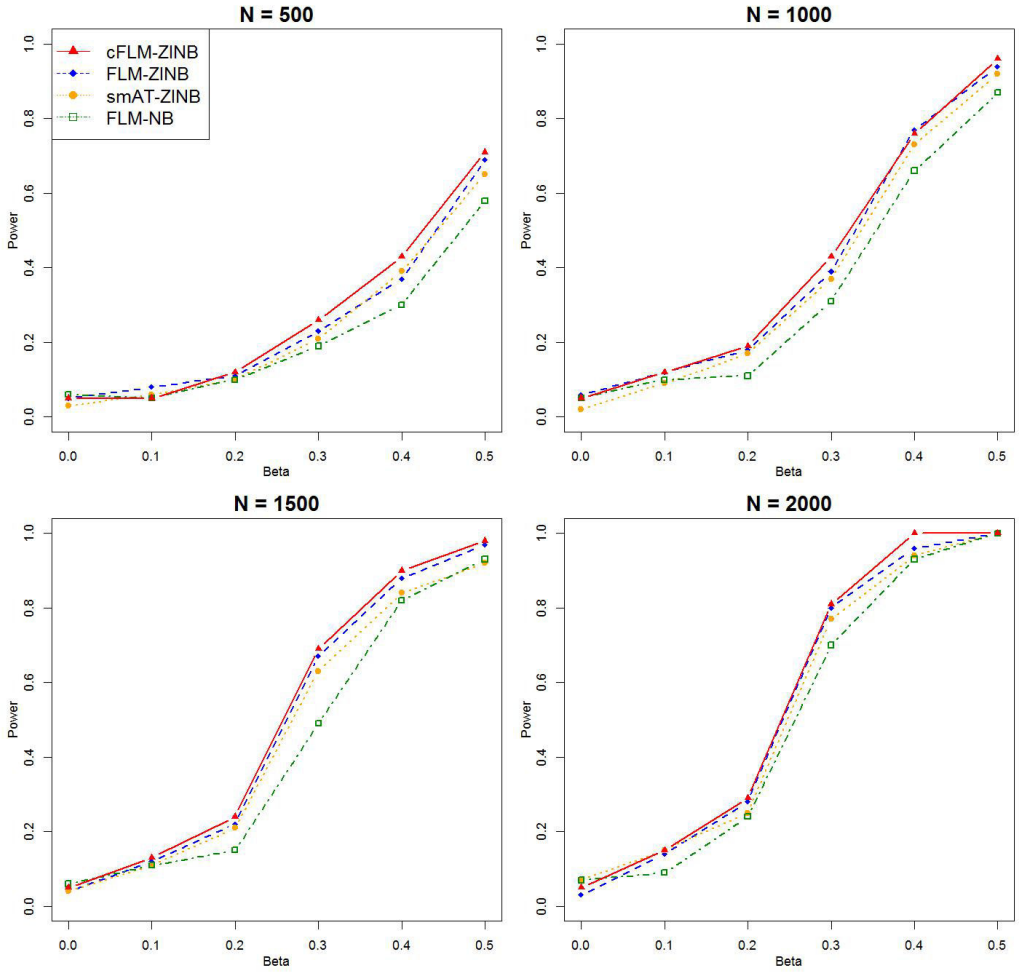

Figure 26: Power simulation for ZINB outcomes based on the CHRNA7 gene: single causal locus, effect in NB distribution, SC15.

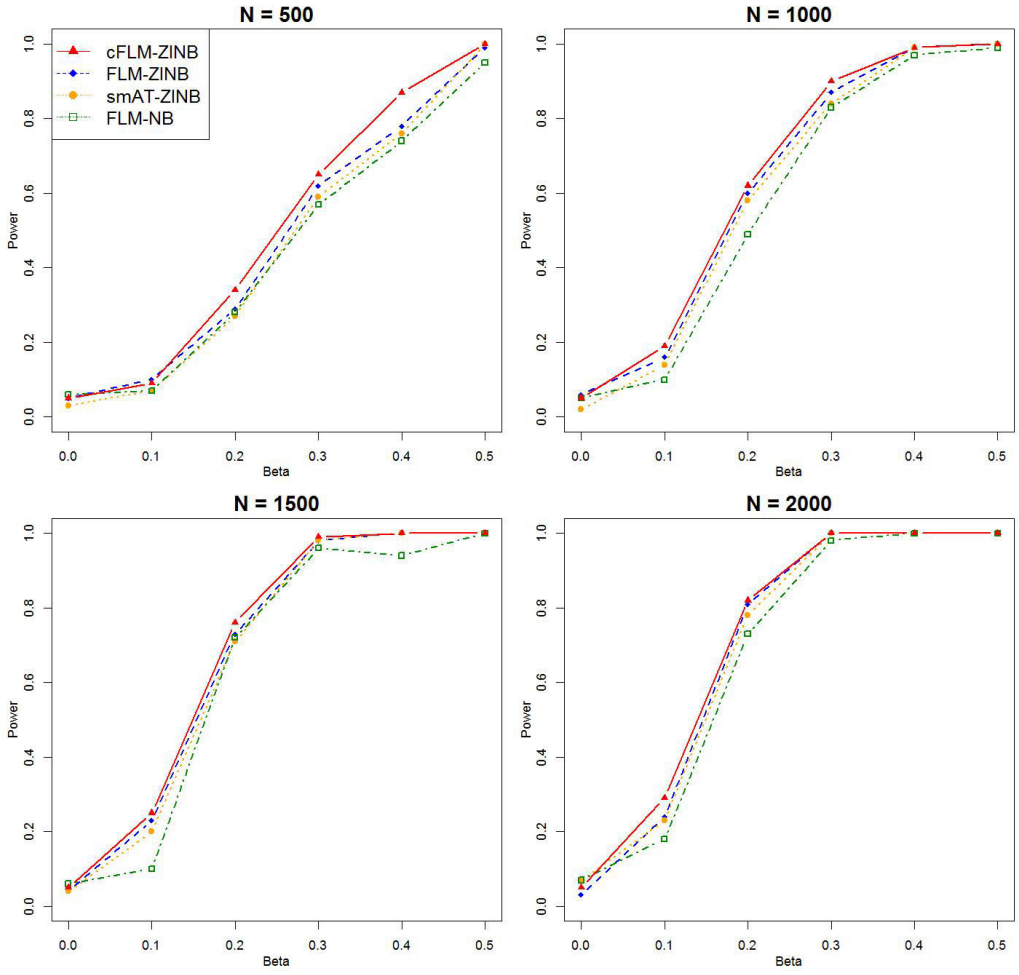

Figure 27: Power simulation for ZINB outcomes based on the CHRNA7 gene: two causal loci, effect in NB distribution, SC16.

3.5 Empirical Studies - Dental Caries

More than 40% children and adolescents, and 90% adults in the US are being affected by dental caries, more commonly known as tooth decay. Even though overall caries prevalence has declined over the last few decades, mean caries rates in children ages 2-11 has increased dramatically over the past few years. Multiple factors are considered to contribute to the risk of having dental caries, such as some environmental factors and social behaviors (Ditmyer et al., 2011). Evidence has shown that some individuals are more susceptible to caries while some others are more resistant, almost irrelevant to the environmental risk factors they are exposed to, suggesting that genetic factors may play crucial roles in the risk of having caries (Bretz et al., 2006). According to several previous studies, the heritability of dental caries were evaluated to be as high as 60%.

To better understand the genetic mechanisms of the risk of dental caries, a GWAS study has been conducted as part of the Gene Environment Association Studies initiative (deposited in dbGaP Study Accession: phs000095.v2.p1) (Yang et al., 2014; Wang et al., 2013). 4,020 individuals were genotyped with a large panel of SNPs (610,000) and examined with multiple outcomes. Our study focused on traits related to caries in permanent teeth. Two indexes, D1MFT and D1MFS which quantifies the total permanent tooth/surface caries with white spots, were included in the analyses. Since the outcomes of interests were both count traits with excess zeroes (Table 11, Figure 28),

the proposed methods, zero-inflated negative binomial model (smAT-ZINB for single-marker tests) and its application with functional coefficient (FLM-ZINB and cFLM-ZINB), were applied to the data set. The final analytic sample consists of 1,480 individuals with complete permanent teeth phenotypic data. Age, gender and total number of teeth/surfaces were included as covariates in the analyses.

Table 12 and Table 13 summarize the significant findings. The Manhattan plots for GWAS scans using ZINB model are presented in Figure 30 and Fig. 31. For genome-wide univariate screening purpose, the threshold of $p < 1 \times 10^{-7}$ was used to declare significance. Same as in COGEND study, the LD block-based Bonferroni threshold for genome-wide significance is $p < 2.5 \times 10^{-6}$. Several SNPs were identified significantly associated with D1MFT and D1MFS in genome-wide scan, of which rs7990965 in chromosome 13 and rs1058595 in chromosome 10 demonstrate consistent significance for both traits. In LD block based association tests, gene PKDCC in chromosome 2 is significantly associated with both traits while the intergenic region between DCN and BTG1 in chromosome 12 is associated with D1MFS. The fitted coefficient functions using cFLM-ZINB model for traits D1MFT and D1MFS based on gene PKDCC are presented in Figure 29. The coefficient patterns are consistent for both traits. It is worth mentioning that gene PKDCC was discovered to be associated with craniofacial morphogenesis in previous dental studies (Melvin et al., 2013).

Table 11: P-values for Kolmogorov-Smirnov (KS) test between distribution of traits and fitted densities of different count models

	Trait:	**D1MFT**	
Poisson	ZIP	Negative Binomial	ZINB
<2.2e-16	<2.2e-16	1.65e-05	0.4914
	Trait:	**D1MFS**	
Poisson	ZIP	Negative Binomial	ZINB
<2.2e-16	<2.2e-16	<2.2e-16	0.9116

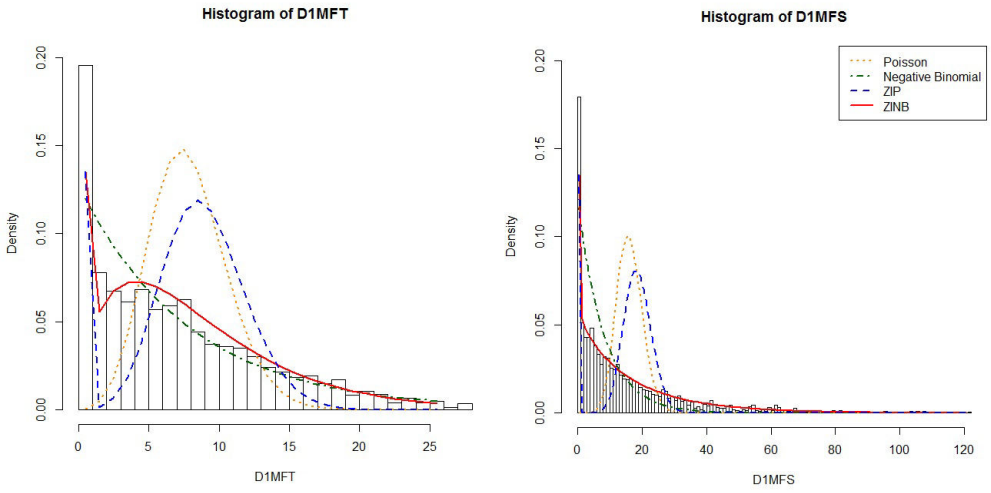

Figure 28: Histograms and fitted densities for traits D1MFT and D1MFS.

Table 12: Significant findings for dental caries GWAS scanning using single-marker association tests based on ZINB model

Trait: D1MFT				
SNP ID	CHR	Gene	MAF	p-value
rs7990965	13	-	0.033	1.02e-12
rs1058595	10	PHYH	0.063	2.52e-09
rs12344120	9	-	0.029	4.79e-08
rs4694666	4	MTHFD2L	0.135	5.48e-08
rs17078140	3	LIMD1	0.029	5.77e-08
rs9893536	17	USP32	0.091	8.80e-08
rs7334525	13	RFC3	0.093	9.52e-08

Trait: D1MFS				
SNP ID	CHR	Gene	MAF	p-value
rs7990965	13	-	0.033	2.40e-10
rs1058595	10	PHYH	0.063	3.29e-09

Table 13: Significant findings for dental caries association tests based on LD blocks (gene clusters) using ZINB model

		Trait:	**D1MFT**			
LD block (Genes)	CHR	Length (kb)	# of SNPs	p-value		
				cFLM	FLM	smAT
PKDCC	2	70	18	8.41e-07	8.06e-05	4.17e-05
		Trait:	**D1MFS**			
LD block (Genes)	CHR	Length (kb)	# of SNPs	p-value		
				cFLM	FLM	smAT
Intergenic between DCN and BTG1	12	70	21	6.07e-07	5.92e-06	4.41e-03
PKDCC	2	70	18	1.38e-06	1.59e-04	7.73e-06

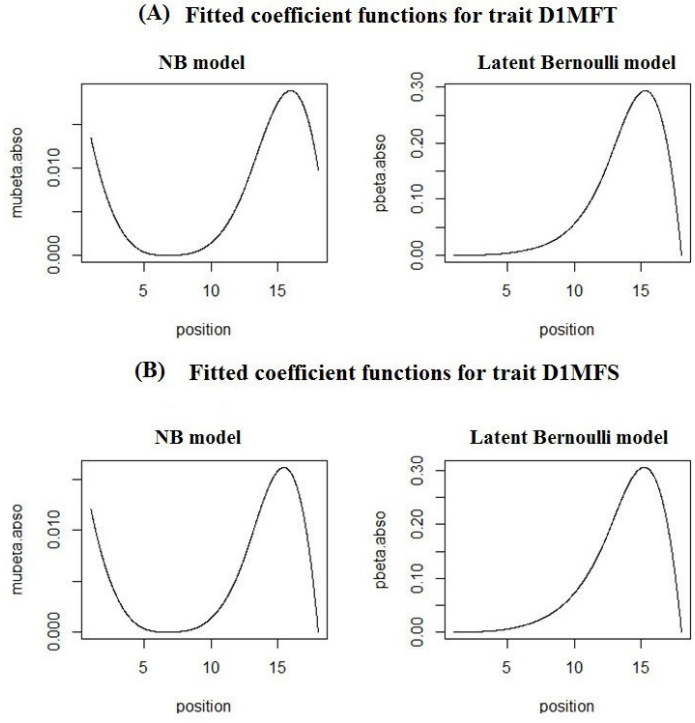

Figure 29: Fitted coefficient functions using cFLM-ZINB model for traits D1MFT and D1MFS based on the PKDCC gene.

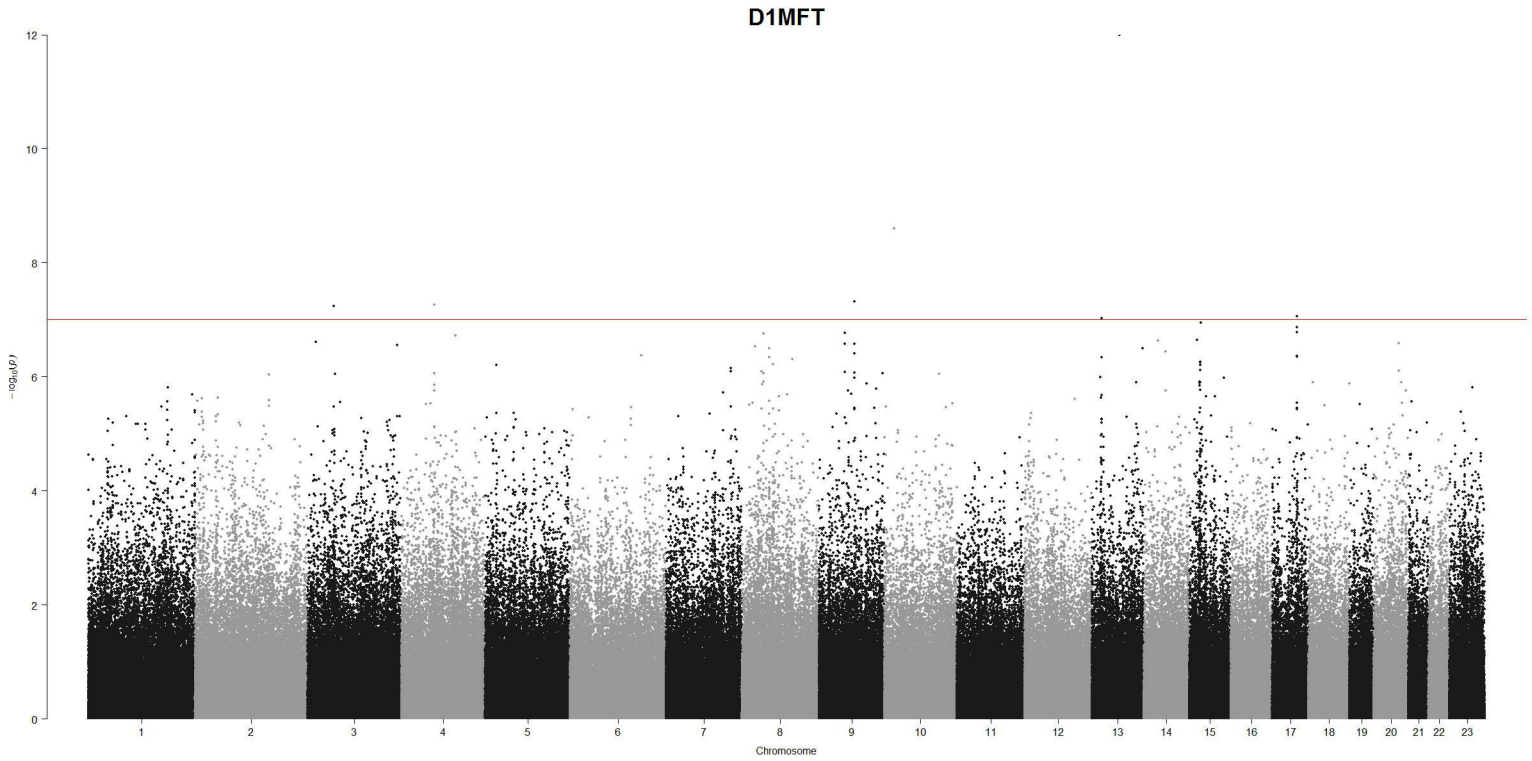

Figure 30: Genome-wide scanning for trait D1MFT using single-marker association tests based on ZINB model.

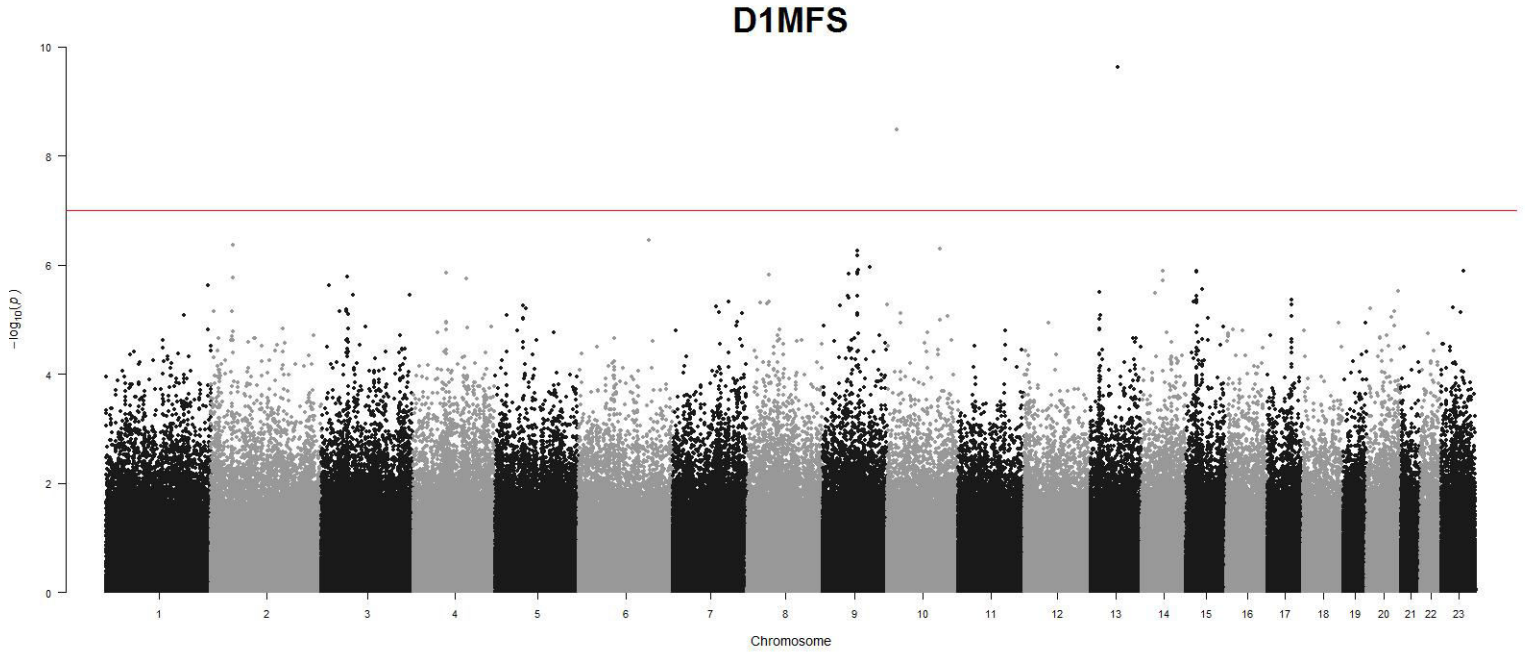

Figure 31: Genome-wide scanning for trait D1MFS using single-marker association tests based on ZINB model.

Chapter 4 Discussion

Association analyses with multiple linked SNPs are expected to be more advantageous than those single-marker tests, because they can account for linkage disequilibrium and composite effects of multiple SNPs. Due to the possible high correlation among contiguous SNP markers, LD blocks consisting of one or more gene regions are formed along the genome. Joint analyses of multiple SNPs on such LD blocks may promote inference about unknown causal variants, since a causal variant is usually in linkage with multiple neighboring SNPs, all of which carrying partial information about the corresponding phenotypic traits. Besides, taking into account the alignment order of SNP markers is crucial because linkage disequilibrium between genetic loci should decline with distance biologically. While the functional linear model is able to incorporate all the aforementioned genetic information into a complex model design, its estimated coefficient function is usually noisy and hard to interpret, reducing lack of causal loci identifiability and power of the test. Improvements to existing methods are of great interest in order to better detect significant genetic variants.

In this dissertation, we proposed a novel constrained functional linear model (cFLM) for flexible and interpretable multi-loci mapping in LD blocks. Our model is built upon the functional linear model (FLM) and is able to accommodate different types of outcomes (normal, binary, ZINB, etc). We reconstruct the FLM by imposing constraints to specify sign-specific effect

and encourage spatial sparsity in the estimated coefficient function. Test of association significance is grounded on the likelihood ratio test statistic following a null weighted mixture of chi-square distribution. Simulation studies were carried out under both random and real gene sampling schemes. We also examined scenarios when (1) only one risk locus was linked to the LD block, and (2) both risk and protective loci were linked to the block. Results show that the proposed tests could well control the type I error. Compared to the competing methods, cFLM generally demonstrated better power when effect size is moderate and large, and comparable performance when effect size is small. Results are similar for all four combinations of simulation settings, suggesting that our method is robust and consistent. We applied the proposed model to a real dataset of nicotine dependence study. The genotypic file includes candidate blocks of CHRN gene clusters where SNPs have high within-block correlation. The observed p-values for two suggestive LD blocks calculated by the cFLM model were much smaller than those computed by other methods. We also applied the proposed ZINB model with (constrained) functional coefficients to a GWAS of dental caries risk. Several SNPs and LD blocks were detected to be associated with the zero-inflated count traits. Since our model will be more powerful when sample size increases, more significant findings can be expected in larger scale studies that will quickly become the new norm.

The proposed methods can be applied to large-scale genome-wide scanning of LD blocks. One concern about the genome-wide application is how

to group SNPs into LD blocks. Luckily several softwares such as PLINK (Purcell et al., 2007) and LDExplorer (2013) have embedded functions to define LD blocks for genomic data. Another concern is about small LD blocks which are not suitable for functional analysis. It may be sensible to combine them with adjacent blocks since nearby blocks are also in linkage. On the other hand, with the advent of next generation sequencing techniques, SNPs can be genotyped much more densely so that this issue will be vastly eased. Prospectively, in order to discover the core subset of causal genes, further group selection among multiple candidate blocks is of great importance. Penalization methods such as group LASSO (Yuan and Lin, 2006) may be directly applicable to our model. Alternatively, machine learning methods such as Neural Network and Random Forest (Botta et al., 2014) are also worth exploring. Finally in practice, if subpopulations exist in the genotypic sample, stratification can be easily adjusted in our model by including principal components of population variation as additional covariates (Price et al., 2006).

Chapter 5 Future Extensions

In this dissertation, the generalized functional linear models (FLM), constrained functional linear models (cFLM) and their extension to model zero-inflated negative binomial (ZINB) outcomes were proposed for multi-loci genetic mapping in genetic regions called linkage disequilibrium blocks (LD blocks) where SNPs are physically close and highly correlated. In order to obtain a more flexible and comprehensive analyses in GWAS, the proposed methods can be extended in the following directions in the future.

(1) Gene-Environment Interaction. In this study, we did not consider any gene-environment interaction effects in the model, while such features are well known to be crucial in association studies. To incorporate the gene-environment interaction effect in the functional linear model framework, we can consider the following model:

$$g(\mu(y_i)) = \alpha_0 + \sum_{u=1}^{q} z_{iu}\alpha_u + \sum_{j=1}^{p} X_i(m_j)\beta(m_j) + \sum_{j,u} z_{iu}X_i(m_j)f(m_j,u), \quad (5.1)$$

where $f(\cdot,\cdot)$ is the two-dimension smooth parameter surface that represents the interaction effect between covariate and genes. This type of varying coefficient functional linear model was primarily discussed in Wu et al. (2010) and it is similar to spatial correlation problems. However, much more details need to be discussed on the parameter estimation, hypothesis tests and additional constraints imposition. Extension to incorporating the gene-environment ef-

fect will greatly enhance the flexibility of our methods.

(2) Genome-wide Block Selection. In order to tect the core subset of causal LD blocks, it is important to develop group selection methods among large number of candidate blocks based on our method. One of the well known methods, the group LASSO (Yuan and Lin, 2006) can be applied to the proposed models. For example, assume K total groups were included in the model and $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp_k})$ is the regression coefficient vector of group k, the sparse group LASSO penalty is formulated as follows.

$$\lambda \sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_2 \tag{5.2}$$

where λ is a tuning parameter. The penalty is a mixture of L_1 and L_2 regularization methods. It encourages sparsity in coefficients among different groups. That is, groups not selected by group LASSO has all SNP level coefficients equal to zero. This extension will be well suited for large-scale GWAS where many candidate gene regions may be found.

(3) Epistasis Study. It is widely acknowledged that genes form a network tend to function simultaneously. Currently, our proposed methods do not consider block-block (gene-gene) interaction effects. It will be interesting to see how the functional modeling framework can incorporate such epistasis effects.

References

The genetic architecture of smoking and smoking cessation. http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000404.v1.p1, 2011. Accessed: 2015-12-21.

International hapmap project. http://hapmap.ncbi.nlm.nih.gov/, 2013. Accessed: 2015-12-21.

E. G. Birgin and J. M. Martínez. Improving ultimate convergence of an augmented lagrangian method. *Optimization Methods and Software*, 23 (2):177–195, 2008.

V. Botta, G. Louppe, P. Geurts, and L. Wehenkel. Exploiting snp correlations within random forest for genome-wide association studies. *PloS one*, 9(4), 2014.

W. A. Bretz, P. M. Corby, M. R. Melo, M. Q. Coelho, S. M. Costa, M. Robinson, N. J. Schork, A. Drewnowski, and T. C. Hart. Heritability estimates for dental caries and sucrose sweetness preference. *Archives of oral biology*, 51(12):1156–1160, 2006.

H. Cardot, F. Ferraty, and P. Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, 13(3):571–592, 2003.

Y. Cui, G. Kang, K. Sun, M. Qian, R. Romero, and W. Fu. Gene-centric genomewide association study via entropy. *Genetics*, 179(1):637–650, 2008.

R. C. Culverhouse, E. O. Johnson, N. Breslau, D. K. Hatsukami, B. Sadler, A. I. Brooks, V. M. Hesselbrock, M. A. Schuckit, J. A. Tischfield, A. M. Goate, et al. Multiple distinct chrnb3–chrna6 variants are genetic risk factors for nicotine dependence in african americans and european americans. *Addiction*, 109(5):814–822, 2014.

C. De Boor. A practical guide to splines. *Mathematics of Computation*, 1978.

M. M. Ditmyer, G. Dounis, K. M. Howard, C. Mobley, and D. Cappelli. Validation of a multifactorial risk factor model used for predicting future caries risk with nevada adolescents. *BMC Oral Health*, 11(1):18, 2011.

R. Fan, Y. Wang, J. L. Mills, A. F. Wilson, J. E. Bailey-Wilson, and M. Xiong. Functional linear models for association analysis of quantitative traits. *Genetic epidemiology*, 37(7):726–742, 2013.

I. Ionita-Laza, S. Lee, V. Makarov, J. D. Buxbaum, and X. Lin. Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*, 92(6):841–853, 2013.

G. M. James and T. J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 533–550, 2001.

G. M. James, J. Wang, and J. Zhu. Functional linear regression that's interpretable. *The Annals of Statistics*, pages 2083–2108, 2009.

T. Kawakami, N. Backström, R. Burri, A. Husby, P. Ólason, A. M. Rice, M. Ålund, A. Qvarnström, and H. Ellegren. Estimation of linkage disequilibrium and interspecific gene flow in ficedula flycatchers by a newly developed 50k single-nucleotide polymorphism array. *Molecular ecology resources*, 14(6):1248–1260, 2014.

J. Liu, K. Wang, S. Ma, and J. Huang. Regularized regression method for genome-wide association studies. In *BMC proceedings*, volume 5, page S67. BioMed Central Ltd, 2011.

X. Liu. Likelihood ratio test for and against nonlinear inequality constraints. *Metrika*, 65(1):93–108, 2007.

L. Luo, Y. Zhu, and M. Xiong. Quantitative trait locus analysis for next-generation sequencing with the functional linear models. *Journal of medical genetics*, 49(8):513–524, 2012.

C.-X. Ma, Q. Yu, A. Berg, D. Drost, E. Novaes, G. Fu, J. S. Yap, A. Tan, M. Kirst, Y. Cui, et al. A statistical model for testing the pleiotropic control of phenotypic plasticity for a count trait. *Genetics*, 179(1):627–636, 2008.

V. S. Melvin, W. Feng, L. Hernandez-Lagunas, K. B. Artinger, and T. Williams. A morpholino-based screen to identify novel genes involved in craniofacial morphogenesis. *Developmental Dynamics*, 242(7):817–831, 2013.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

J. O. Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

N. L. Saccone, J. C. Wang, N. Breslau, E. O. Johnson, D. Hatsukami, S. F. Saccone, R. A. Grucza, L. Sun, W. Duan, J. Budde, et al. The chrna5-chrna3-chrnb4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in african-americans and in european-americans. *Cancer research*, 69(17):6848–6856, 2009.

A. Shapiro. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, 72(1):133–144, 1985.

F. Talas, T. Würschum, J. C. Reif, H. K. Parzies, and T. Miedaner. Association of single nucleotide polymorphic sites in candidate genes with aggressiveness and deoxynivalenol production in fusarium graminearum causing wheat head blight. *BMC genetics*, 13(1):14, 2012.

D. Taliun, J. Gamper, and C. Pattaro. Ldexplorer. http://www.eurac.edu/en/research/health/biomed/services/Pages/LDExplorer.aspx, 2013. Accessed: 2015-11-20.

Q. Wang, P. Jia, K. T. Cuenco, Z. Zeng, and E. Feingold. Association signals unveiled by a comprehensive gene set enrichment analysis of dental caries genome-wide association studies. *PloS one*, 8(8):e72653, 2013.

L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

H. Wittenburg, M. A. Lyons, R. Li, G. A. Churchill, M. C. Carey, and B. Paigen. Fxr and abcg5/abcg8 as determinants of cholesterol gallstone formation from quantitative trait locus mapping in mice. *Gastroenterology*, 125(3):868–881, 2003.

F. A. Wolak. Local and global testing of linear and nonlinear inequality constraints in nonlinear econometric models. *Econometric Theory*, 5(01): 1–35, 1989.

M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

S. Wu. *A robust approach for genetic mapping of complex traits*. PhD thesis, University of Florida, 2008.

T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

Y. Wu, J. Fan, H.-G. Müller, et al. Varying-coefficient functional linear regression. *Bernoulli*, 16(3):730–758, 2010.

J. Yang, W. Zhu, J. Chen, Q. Zhang, and S. Wu. Genome-wide two-marker linkage disequilibrium mapping of quantitative trait loci. *BMC genetics*, 15(1):20, 2014.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.