stats

MDPI

Article

A Constrained Generalized Functional Linear Model for Multi-Loci Genetic Mapping

Jiayu Huang, Jie Yang, Zhangrong Gu, Wei Zhu and Song Wu *

Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11790, USA; jiayuhuang1989@gmail.com (J.H.); Jie.Yang@stonybrookmedicine.edu (J.Y.); zhangrong.gu@stonybrook.edu (Z.G.); wei.zhu@stonybrook.edu (W.Z.)
* Correspondence: song.wu@stonybrook.edu

Abstract: In genome-wide association studies (GWAS), efficient incorporation of linkage disequilibria (LD) among densely typed genetic variants into association analysis is a 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 generalized FLM (cGFLM) framework to perform simultaneous association tests on a block of linked SNPs with various trait types, including continuous, binary and zero-inflated count phenotypes. The new cGFLM applies a set of inequality constraints on the FLM to ensure model identifiability under different genetic codings. The method is implemented via B-splines, and an augmented Lagrangian algorithm is employed for parameter estimation. For hypotheses testing, a test statistic that accounts for the model constraints was derived, following a mixture of chi-square distributions. Simulation results show that cGFLM 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 a candidate gene-based COGEND study and a large-scale GWAS data on dental caries risk.

Keywords: GWAS; LD mapping; multi-loci genetic mapping; functional linear model; cGFLM

Citation: Huang, J.; Yang, J.; Gu, Z.; Zhu, W.; Wu, S. A Constrained Generalized Functional Linear Model for Multi-Loci Genetic Mapping. *Stats* **2021**, *4*, 550–577. https://doi.org/10.3390/stats4030033

Academic Editor: Eddy Kwessi

Received: 8 May 2021
Accepted: 22 June 2021
Published: 25 June 2021

1. Introduction

Recent advances in technology have facilitated the development of association studies using a highly dense map of genetic markers, such as single nucleotide polymorphisms (SNPs). Nowadays, up to a million SNPs can be readily genotyped along the human genome for thousands of subjects. Even though analyses based on univariate tests are still popular and useful as pre-screening techniques [1], such methods have a major drawback—the high correlation and composite effect among multiple genetic variants are omitted. To accommodate this challenge, a few more advanced statistical methods that are based on multiple linked SNPs have been developed in recent years [2–5], aiming to improve the power of associations between genetic markers and phenotypic traits of interest.

As SNPs naturally cluster together and form the so-called linkage disequilibrium (LD) blocks, multi-loci association tests that take into account this special genomic structure are usually more powerful than single SNP-based ones [2–5]. Further, due to this locally connected/correlated LD structure, if a casual SNP, even unobserved in some cases, exists in one LD block, its information, i.e., genetic effect, can be partially carried and jointly represented by its neighboring SNPs in the same block, which increases the likelihood of the causal SNPs being detected. Another benefit of testing at the LD block level is on multiple test adjustment in GWAS, including not only significant reduction of the total number of tests, but also improvement on the independence of the tests. Several multi-SNP methods have been proposed, such as principal component analysis [2,6], entropy-based methods [7,8], kernel machine methods (SKAT, SKAT-C and CKAT) [3,5] and two-marker LD mapping [4]. In addition, some variable selection models, such as LASSO [9,10] and

smoothed minimax concave penalized regression (SMCP) [11], were applied to identify a core subset of potential causal variants. Although these methods are useful, most of them overlook marker ordering information in their physical positions.

Functional linear models (FLMs) serve as a good solution to the above-mentioned problems, as it can effectively preserve the intrinsic correlation structure and spatial information of SNPs in an LD [12]. In essence, FLMs take into account the effect that neighboring correlated SNPs in an LD would show similar genetic effects, and treat the regression coefficients of these SNPs with an explicit functional form. The smooth coefficient function can be further expanded in terms of spline bases, which allows for substantial dimension reduction in parameter estimation [12,13]. Both functional principal component analysis (FPCA) and beta-smooth only approaches have been applied to construct the FLMs [12,14–16]. Since the beta-smooth only approach is more straightforward, it was used for the construction of FLM in this study.

One critical issue with FLM is that the fitted coefficient functions are often noisy and hard to interpret. They usually fluctuate dramatically due to several reasons: (1) Strong LD among nearby SNPs causes multicollinearity, which leads to erratic changes in the signs of adjacent functional coefficients; (2) FLMs cannot yield estimates that are exactly zero over regions with no significant association, thus generating unnatural wiggles in the fitted genetic function; (3) population-specific phenomena such as mutation, genetic drift, population structure and variations in allele frequencies result in the LD not decaying 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 detection power due to oversmoothing. In addition, no proper test other than the permutation approach is available for penalty-based methods, which brings heavy computational burden to large-scale studies.

In this paper, we propose a novel constrained generalized functional linear model (cGFLM) for flexible and reasonable multi-loci genetic mapping on a block of correlated genetic variants. The cGFLM retains the merits of FLM such as preserving the spatial and LD information among genetic markers, as well as compressing the high-dimensional problem into aggregate inference about several smoothing components. In addition to these benefits, the cGFLM is more powerful and enables easier interpretation of the functional coefficients. Specifically, we reconstruct FLM by separating the genetic effect into two sign-specific coefficient functions and imposing an equality constraint to encourage overall spatial sparsity. The cGFLM tends to constrain the functional coefficients to be zero in "null regions" where no determinative positive or negative effect is present.

We further extend the cGFLM framework to several types of quantitative traits, such as continuous, binary and count data. We put more focus on count traits as they are common in practice but relatively less mentioned in the literature. Poisson and negative binomial (NB) models provide tractable methods to most count traits, however, in traits with excessive zeros, zero-inflated Poisson (ZIP) or negative binomial (ZINB) models were also developed [17]. Since the ZINB is the most complicated model here, we discuss how to apply the cGFLM to it in more detail. Applications of cGFLM to other models would be similar.

The remainder of the article is organized as follows. In Sections 2 and 3, we introduce a generalized functional linear model (GFLM) framework, discussing the continuous and categorical traits (Section 2) and then zero-inflated negative binomial phenotypic traits (Section 3). In Section 4, we describe the proposed cGFLM and cGFLM-ZINB models, as well as their estimation and testing processes. In Section 5, we present Monte Carlo simulations to validate the proposed test and compare our model with several alternative methods. In Sections 6 and 7, we apply the proposed method to the COGEND (The Collaborative Genetic Study of Nicotine Dependence) and Dental Caries GWAS data. Sections 8 and 9 are the discussion and conclusion, respectively.

2. Generalized Functional Linear Model

Suppose n subjects are sampled, each characterized by p linked SNP markers $j = 1, \ldots, p$, where m_j is the spatial location of SNP marker j with $0 \leq m_1 < \cdots < m_p$. SNP location can be measured by the distance of a SNP from starting position of a block. Denote the SNP genotype at marker position m_j for subject i as $X_i(m_j)$. Suppose the genotypes of a SNP are AA, Aa and aa, which are coded as 0, 1 and 2, according to the number of copies of the allele a. That is,

$$X_i(m_j) = \begin{cases} 0, & \text{if the genotype of SNP } j \text{ at position } m_j \text{ is } AA; \\ 1, & \text{if the genotype of SNP } j \text{ at position } m_j \text{ is } Aa; \\ 2, & \text{if the genotype of SNP } j \text{ at position } m_j \text{ is } aa. \end{cases} \tag{1}$$

Suppose the genetic effect is denoted by $\beta(m)$, a smooth coefficient function over all marker positions m in an LD block. The value of the smooth coefficient function at marker position m_j is then $\beta(m_j)$. Further, suppose a set of other covariates $\boldsymbol{z_i} = (z_{i1}, \ldots, z_{iq})^\top$ are also observed for subject i and let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^\top$ be the $q \times 1$ vector of their corresponding coefficients. The global-featured effect is included as the intercept, denoted as α_0. For the ith subject, let y_i denote its phenotypic response. If y_is are continuous, a functional linear model (FLM) can be formulated as

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

This type of association has been explored in [12,14]. Alternatively, if y_is are categorical, a link function $g(\cdot)$ can be applied to to transform its mean $\mu(y_i) = E(y_i)$ to be $g(\mu(y_i))$. In this case, a generalized functional linear model (GFLM) can be formulated as

$$g(\mu(y_i)) = \alpha_0 + \sum_{u=1}^{q} z_{iu}\alpha_u + \sum_{j=1}^{p} X_i(m_j)\beta(m_j). \tag{3}$$

In the FLM/GFLM above, the genetic effects of a block of SNPs are assumed to be a smooth function. In other words, SNPs located close to each other are expected to show similar effects.

Here we represent $\beta(m)$ with B-splines and express $\beta(m)$ as

$$\beta(m) = \sum_{r=1}^{d} \gamma_r B_r(m) = \boldsymbol{B}(m)\boldsymbol{\gamma}, \tag{4}$$

where $d \geq 1$ is the number of basis functions, $\boldsymbol{\gamma}_{d\times 1} = (\gamma_1, \ldots, \gamma_d)^\top$ is a vector of real-valued coefficients, $(B_1, \ldots, B_d)$ is a set of d basis functions, and $\boldsymbol{B}(m) = (B_1(m), \ldots, B_d(m))$ is a vector for the d basis functions evaluated at position m. Let $\boldsymbol{y}_{n\times 1} = (y_1, \ldots, y_n)^\top, \boldsymbol{Z}_{n\times q} = (\boldsymbol{z_1}, \ldots, \boldsymbol{z_n})^\top$,

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} X_1(m_1) & \cdots & X_1(m_p) \\ \vdots & \vdots & \vdots \\ X_n(m_1) & \cdots & X_n(m_p) \end{pmatrix} \text{and } \boldsymbol{B}_{p\times d} = \begin{pmatrix} B_1(m_1) & \cdots & B_d(m_1) \\ \vdots & \vdots & \vdots \\ B_1(m_p) & \cdots & B_d(m_p) \end{pmatrix}. \tag{5}$$

In matrix form, the GFLM in (3) can be reformulated as

$$\begin{aligned} g(\mu(\boldsymbol{y})) &= \alpha_0 + \boldsymbol{Z\alpha} + \boldsymbol{X\beta} \\ &= \alpha_0 + \boldsymbol{Z\alpha} + \boldsymbol{XB\gamma} \\ &= \begin{pmatrix} \mathbf{1} & \boldsymbol{Z} & \boldsymbol{XB} \end{pmatrix} \begin{pmatrix} \alpha_0 & \boldsymbol{\alpha} & \boldsymbol{\gamma} \end{pmatrix}^\top. \end{aligned} \tag{6}$$

Suppose the responses $\boldsymbol{y}_{n\times 1}$ are i.i.d. samples from a distribution with density $f(y;\boldsymbol{\beta})$. By substituting $\boldsymbol{\beta}$ with $\boldsymbol{B}\gamma$, the number of parameters changes from $(p+q+1)$ to $(d+q+1)$. Then the log-likelihood function can be expressed as

$$l(\alpha_0,\boldsymbol{\alpha},\boldsymbol{\gamma}) = \sum_{i=1}^{n} \log f(y_i;\alpha_0,\boldsymbol{\alpha},\boldsymbol{\gamma}) \tag{7}$$

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

$$(\hat{\alpha}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) = \arg\max_{\alpha_0,\boldsymbol{\alpha},\boldsymbol{\gamma}} l(\alpha_0,\boldsymbol{\alpha},\boldsymbol{\gamma}). \tag{8}$$

Since we are modeling a block of SNPs simultaneously in (3), the hypothesis test of association between genetic variants and phenotypic trait will be made at the block level. That is,

$$\begin{array}{ll} \mathrm{H}_0: & \beta(m_j) = 0, \text{ for all } j = 1,\ldots,p. \\ \mathrm{H}_a: & \text{Not } \mathrm{H}_0. \end{array} \tag{9}$$

Approximately, the hypotheses can also be expressed as:

$$\begin{array}{ll} \mathrm{H}_0: & \gamma_r = 0, \text{ for all } r = 1,\ldots,d. \\ \mathrm{H}_a: & \gamma_r \in \mathbb{R}^d. \end{array} \tag{10}$$

The likelihood ratio test (LRT) can be applied to draw inference about whether a block of SNPs may be associated with the phenotypic trait. Under H_0, the LRT statistic, defined as $-2(l_0 - l_1)$, follows asymptotically a χ^2_d distribution (Chi-square distribution with df = d, the number of basis vectors used in Equation (5)).

3. GFLM for Zero-Inflated Negative Binomial (ZINB) Traits

Most count traits can be modeled with Poisson or negative binomial (NB) distributions. However, when excessive zeroes are present in the data, a zero-inflated Poisson or negative binomial regression framework needs to be employed [17], which utilizes a latent Bernoulli distribution and a Poisson distribution to incorporate explanatory variables. In this section, we describe a GFLM framework for count traits that follow a ZINB distribution.

To illustrate the ZINB model, we use dental caries, more commonly known as tooth decay, as an example. Let Y denote the number of dental caries, and the probability distribution of Y is given as:

$$\begin{array}{c} L \sim \text{Bernoulli}(\pi); \\ Pr(Y = 0|L = 1) = 1; \\ Pr(Y = k|L = 0) \sim NB(\mu,\phi), k = 0,1,\ldots \end{array} \tag{11}$$

Here L is a latent Bernoulli random variable that categorizes caries "risk" into two states: "risk-free" (L = 1) and "risky" (L = 0). π is the probability of $L = 1$, 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,

$$y_i \sim \begin{cases} 0, & \text{with probability } \pi_i; \\ NB(\mu_i,\lambda), & \text{with probability } 1-\pi_i. \end{cases} \tag{12}$$

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

$$y_i = \begin{cases} 0, & \text{with probability } \pi_i + (1-\pi_i)(\frac{\phi}{\phi+\mu_i})^{\phi}; \\ k, & \text{with probability } (1-\pi_i)\frac{\Gamma(\phi+k)}{k!\Gamma(\phi)}(\frac{\phi}{\phi+\mu_i})^{\phi}(\frac{\mu_i}{\phi+\mu_i})^{k}, \text{ k = 1, 2, } \ldots \end{cases} \tag{13}$$

An expectation-maximization (EM) algorithm can be applied to estimate the parameters in the ZINB model. Suppose at step t, the parameter estimates are $(\pi^{(t)}, \mu^{(t)}, \phi^t)$. Then the detailed EM implementation is given as follows.

E-step

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

$$L_i^{(t)} = E_{L|y_i,\pi^{(t)},\mu^{(t)},\phi^{(t)}}(L_i) = \begin{cases} 0, & y_i > 0; \\ \frac{\pi^{(t)}}{\pi^{(t)}+(1-\pi^{(t)})(\frac{\phi^t}{\phi^{(t)}+\mu^{(t)}})^{\phi^{(t)}}}, & y_i = 0. \end{cases} \tag{14}$$

M-step

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

$$l_{Ber}(\pi|\boldsymbol{y}, \boldsymbol{z}^{(t)}) = \sum_{i=1}^{n} L_i^{(t)} \log(\pi) + (1 - L_i^{(t)}) \log(1 - \pi) \tag{15}$$

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

$$l_{NB}(\mu, \phi|\boldsymbol{y}, \boldsymbol{z}^{(t)}) = \sum_{i=1}^{n} (1 - L_i^{(t)}) \log \frac{\Gamma(\phi + y_i)}{y_i! \Gamma(\phi)} (\frac{\phi}{\phi + \mu})^{\phi} (\frac{\mu}{\phi + \mu})^{y_i} \tag{16}$$

The maximization can be performed by the Newton–Raphson algorithm for the two models int the M-step simultaneously.

Using the same notations as in the previous section, the Bernoulli probabilities and the negative binomial means can be transformed using the logit and the log links, respectively, and are modeled as follows:

$$\text{logit}(\pi_i) = \log(\frac{\pi_i}{1 - \pi_i}) = \alpha_0^{Ber} + \sum_{u=1}^{q} z_{iu} \alpha_u^{Ber} + \sum_{j=1}^{p} X_i(m_j) \beta^{Ber}(m_j), \tag{17}$$

and

$$\log(\mu_i) = \alpha_0^{NB} + \sum_{u=1}^{q} z_{iu} \alpha_u^{NB} + \sum_{j=1}^{p} X_i(m_j) \beta^{NB}(m_j). \tag{18}$$

Since usually we do not have strong pre-assumption about genetic components and covariates, the same sets of genetic components and covariates can be used in both Bernoulli and NB models simultaneously. For univariate analyses on single SNPs, it is equivalent to the case that one LD block only contains one SNP, and we can simply set $p = 1$ in the above equations. For multi-loci mapping purposes, we set all SNPs in an LD block as input/explanatory variables.

We substitute the unstructured coefficients to functional coefficients, which are represented by linear combination of B-Splines. That is, Equation (5) is plugged into the coefficients for both the latent Bernoulli model and NB model to form the GFLM-ZINB model. Using the same notation above, the GFLM-ZINB model is reformulated as:

$$\text{logit}(\pi_i) = \begin{pmatrix} 1 & \boldsymbol{Z}_i & \boldsymbol{X_i B^{(Ber)}} \end{pmatrix} \begin{pmatrix} \alpha_0^{Ber} & \boldsymbol{\alpha}^{Ber} & \boldsymbol{\gamma}^{Ber}_{\boldsymbol{d_{Ber}} \times \mathbf{1}} \end{pmatrix}^{\top} = U_i^{Ber} \eta^{Ber}. \tag{19}$$

$$\log(\mu_i) = \begin{pmatrix} 1 & \boldsymbol{Z}_i & \boldsymbol{X_i B^{(NB)}} \end{pmatrix} \begin{pmatrix} \alpha_0^{NB} & \boldsymbol{\alpha}^{NB} & \boldsymbol{\gamma}^{NB}_{\boldsymbol{d_{NB}} \times \mathbf{1}} \end{pmatrix}^{\top} = U_i^{NB} \eta^{NB}.$$

Then we can estimate the parameters using an EM Algorithm again, which is implemented as follows.

1. Let $\hat{L}_i^{(t)} = \frac{\pi_i^{(t)}}{\pi_i^{(t)}+(1-\pi_i^{(t)})(\frac{\phi^t}{\phi^{(t)}+\mu_i^{(t)}})^{\phi^{(t)}}} \mathbb{I}(y_i = 0)$

2. • Perform logistic regression of $\hat{L}_i^{(t)}$ on U^{Ber} to estimate $\eta^{Ber^{(t+1)}}$.
   • Perform weighted negative binomial regression of y on U^{NB} with weights $1-\hat{L}_i^{(t)}$ to obtain estimate $\eta^{NB^{(t+1)}}$ and $\phi^{(t+1)}$.
3. Let $\pi_i^{(t+1)} = \frac{\exp(U_i^{Ber}\eta^{Ber^{(t+1)}})}{1+\exp(U_i^{Ber}\eta^{Ber^{(t+1)}})}$ and $\mu^{(t+1)} = \exp(U_i^{NB}\eta^{NB^{(t+1)}})$, iterate back to step 1.

For hypothesis testing, since we model a block of SNPs simultaneously, the test of association between genetic variants and a 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. The hypothesis test should consider the overall effect of both models. That is, we consider testing at least one coefficients from either model being nonzero. Note that we obtained 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 $\gamma^{Ber}_{d_{Ber}\times 1}$ and $\gamma^{NB}_{d_{NB}\times 1}$ by applying the functional coefficients implemented via the B-spline bases. The hypotheses are formulated as

$$\begin{array}{ll} \mathrm{H}_0: & \gamma_b^{Ber} = 0, \gamma_m^{NB} = 0, \text{ for all } b, n. \\ \mathrm{H}_a: & \text{Not } \mathrm{H}_0. \end{array} \tag{20}$$

The likelihood ratio test (LRT) can 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 = $d_{Ber} + d_{NB}$):

$$\chi^2 = -2(l_0 - l_1) \xrightarrow{d} \chi^2_{d_{Ber}+d_{NB}}. \tag{21}$$

4. Constrained Generalized Functional Linear Model

In practice, we have observed that the fitted coefficient functions in the GFLM can fluctuate dramatically, and the fluctuation makes it difficult to explain the functional patterns and determine the locations of causal SNPs . To address this, here we propose a more reasonable constrained functional linear model (cGFLM). Specifically, we separate the genetic effect into two sign-specific coefficient functions and impose an equality constraint to promote spatial sparsity. The cGFLM is formulated as follows:

$$\begin{aligned} g(\mu(y_i)) &= \alpha_0 + \sum_{u=1}^{q} z_{iu}\alpha_u + \sum_{j=1}^{p} X_i(m_j)\beta^+(m_j) + \sum_{j=1}^{p} X_i(m_j)\beta^-(m_j) \\ &\text{subject to } \beta^+(m_j) \geq 0, \beta^-(m_j) \leq 0, \beta^+(m_j)\cdot\beta^-(m_j) = 0 \text{ for all } j, \end{aligned} \tag{22}$$

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

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 modified coefficient vectors $\gamma^+_{d_1\times 1}, \gamma^-_{d_2\times 1}$, respectively:

$$\beta^+(m) = \boldsymbol{B}^+(m)\gamma^+, \beta^-(m) = \boldsymbol{B}^-(m)\gamma^-. \tag{23}$$

Let $\circ$ denote the Hadamard product of two vectors. In matrix form, the cGFLM is formulated as:

$$\begin{aligned} g(\mu(\boldsymbol{y})) &= \alpha_0 + \boldsymbol{Z\alpha} + \boldsymbol{XB}^+\gamma^+ + \boldsymbol{XB}^-\gamma^- \\ &= \begin{pmatrix} (\mathbf{1} & \boldsymbol{Z}) & \boldsymbol{XB}^+ & \boldsymbol{XB}^- \end{pmatrix} \begin{pmatrix} \gamma_0 & \gamma^+ & \gamma^- \end{pmatrix}^\top = \boldsymbol{U\gamma^*} = \boldsymbol{\eta^*} \\ &\text{subject to } \boldsymbol{B}^+\gamma^+ \geq \mathbf{0}, \boldsymbol{B}^-\gamma^- \leq \mathbf{0}, (\boldsymbol{B}^+\gamma^+) \circ (\boldsymbol{B}^-\gamma^-) = \mathbf{0}. \end{aligned} \tag{24}$$

The log-likelihood function for cGFLM is then

$$l(\gamma_0, \gamma^+, \gamma^-) = \sum_{i=1}^{n} \log f(y_i; \gamma_0, \gamma^+, \gamma^-). \tag{25}$$

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

$$\begin{array}{ll} \text{maximize} & l(\gamma_0, \gamma^+, \gamma^-) \\ \text{subject to} & \boldsymbol{B}^+\gamma^+ \geq \mathbf{0}, \boldsymbol{B}^-\gamma^- \leq \mathbf{0}, (\boldsymbol{B}^+\gamma^+) \circ (\boldsymbol{B}^-\gamma^-) = \mathbf{0}. \end{array} \tag{26}$$

An augmented Lagrangian algorithm (ALA) can be applied to this constrained maximization [18,19]. Let $\gamma^* = (\gamma_0, \gamma^+, \gamma^-)$. Denote the p equality constraints defined by $(\boldsymbol{B}^+\gamma^+) \circ (\boldsymbol{B}^-\gamma^-) = \mathbf{0}$ as $h(\gamma^*) = 0$, and the $2p$ inequality constraints defined by $\boldsymbol{B}^+\gamma^+ \geq \mathbf{0}, \boldsymbol{B}^-\gamma^- \leq \mathbf{0}$ as $g(\gamma^*) \leq 0$. For notation purposes, let $I = p, J = 2p$, then the corresponding augmented Lagrangian for (26) to be minimized is

$$L_\rho(\gamma^*, \lambda_1, \lambda_2) = -l(\gamma^*) + \frac{\rho}{2}\{\sum_{i=1}^{I}[h_i(\gamma^*) + \frac{\lambda_{1i}}{\rho}]^2 + \sum_{j=1}^{J}[\max(0, g_j(\gamma^*) + \frac{\lambda_{2j}}{\rho})]^2\}, \tag{27}$$

where $\boldsymbol{\lambda}_1 \in \mathbb{R}^I, \boldsymbol{\lambda}_2 \in \mathbb{R}^J_+$ and $\rho > 0$. Let $\lambda_{1\min} < \lambda_{1\max}, \lambda_{2\max} > 0, \xi > 1, 0 < \tau < 1, \epsilon_c$ be a small constant value (e.g., 10^{-4}), and $\{\epsilon_t\}$ be a sequence of nonnegative numbers such that $\lim_{t\to\infty} \epsilon_t = 0$. A sketch of the augmented Lagrangian algorithm is given below and more implementation details can be found in [20]:

Step 0. Let $\lambda_{1i}^{(1)} \in [\lambda_{1\min}, \lambda_{1\max}], i = 1, \ldots, I, \lambda_{2j}^{(1)} \in [0, \lambda_{2\max}], j = 1, \ldots, J$ and $\rho_1 > 0$. Let $\gamma^{*(0)}$ in the parameter space Ω be an arbitrary initial point. Set $t \to 1$.

Step 1. Find the approximate minimizer $\gamma^{*(t)}$ of $L_{\rho_t}(\gamma^*, \lambda_1^{(t)}, \lambda_2^{(t)})$ subject to $\gamma^* \in \Omega$, satisfying

$$\|P_\Omega(\gamma^{*(t)} - \bigtriangledown L_{\rho_t}(\gamma^{*(t)}, \lambda_1^{(t)}, \lambda_2^{(t)})) - \gamma^{*(t)}\|_\infty \leq \epsilon_t,$$

where P_Ω is the Euclidean projection onto Ω [21] and $\bigtriangledown L_{\rho_t}(\gamma^{*(t)}, \lambda_1^{(t)}, \lambda_2^{(t)})) = \frac{\partial L_{\rho_t}(\gamma^*, \lambda_1^{(t)}, \lambda_2^{(t)})}{\partial \gamma^*}$ evaluated at $\gamma^* = \gamma^{*(t)}$.

Step 2. Define $V_j^{(t)} = \max\{g_j(\gamma^{*(t)}), -\frac{\lambda_{2j}^{(t)}}{\rho_t}\}, j = 1, \ldots, J$.

$$\rho_{t+1} = \begin{cases} \rho_t, \text{ If t = 1 or } \max\{\|h(\gamma^{*(t)})\|_\infty, \|V^{(t)}\|_\infty\} \leq \tau \max\{\|h(\gamma^{*(t-1)})\|_\infty, \|V^{(t-1)}\|_\infty\} \\ \xi\rho_t, \text{ otherwise} \end{cases}$$

Step 3. Update λ_1s and λ_2s:

$$\lambda_{1i}^{(t+1)} = \min\{\max\{\lambda_{1\min}, \lambda_{1i}^{(t)} + \rho_t h_i(\gamma^{*(t+1)})\}, \lambda_{1\max}\} \text{ for } i = 1, \ldots, I,$$

$$\lambda_{2j}^{(t+1)} = \min\{\max\{0, \lambda_{2j}^{(t+1)} + \rho_t g_j(\gamma^{*(t+1)})\}, \lambda_{2\max}\} \text{ for } j = 1, \ldots, J.$$

Step 4. if $\|\gamma^{*(t)} - \gamma^{*(t-1)}\|_\infty \geq \epsilon_c$, set $t + 1 \to t$ and go to Step 1; else stop. ■

Similar to what has been performed for (10), a likelihood ratio test can be performed to investigate the overall genetic effects represented by a block of SNPs in contiguous genomic regions. However, the null and alternative hypotheses in (10) are updated with regard to the new parameter space and imposed constraints, as follows:

$$\begin{array}{ll} \text{H}_0: & \gamma^+_{d_1\times 1} = \mathbf{0} \text{ and } \gamma^-_{d_2\times 1} = \mathbf{0}. \\ \text{H}_a: & \boldsymbol{B}^+\gamma^+ \geq \mathbf{0}, \boldsymbol{B}^-\gamma^- \leq \mathbf{0}, (\boldsymbol{B}^+\gamma^+) \circ (\boldsymbol{B}^-\gamma^-) = \mathbf{0} \end{array} \tag{28}$$

Since we constrained that the Hadamard product of sign-specific coefficient functions is $\mathbf{0}$, at least one basis coefficient in the positive or 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 [22,23], it has been shown that in nonlinear optimization, the alternative parameter space $\mathbf{\Omega}$ 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 $\mathbf{\Omega}$ taken metrics according to the Hessian matrix $I(\gamma^*)$, and it may follow an asymptotic χ^2_k distribution under H_0 $(k = 0, \ldots, K-1)$. Therefore, 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

$$\bar{\chi}^2 = -2(l_0 - l_1) \xrightarrow{d} \sum_{j=0}^{K} w_j \chi^2_j. \tag{29}$$

For any $c \in \mathbb{R}$, the p-value of the $\bar{\chi}^2$ test statistic is defined as

$$Pr(\bar{\chi}^2 \geq c^2) = \sum_{j=0}^{K} w_j P(\chi^2_j \geq c^2). \tag{30}$$

The mixing probabilities can be calculated using Monte Carlo techniques. The algorithm is given as follows: (1) Take 1000 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 1000 draws in which it has exactly k non-zero coefficients projected on the alternative parameter space. The Monte Carlo technique is easy to implement and able to circumvent complicated numerical integrations.

The LRT can be adapted to the constrained functional coefficients (cGFLM-ZINB) model as follows. Since the parameter estimation is conducted with two independent sets of constraints, the hypothesis test consists of two parts as well, one for the latent Bernoulli distribution and one for the NB distribution.

$$\begin{array}{ll} H_0: & \gamma^{Ber+}_{d_{Ber1}\times 1} = \mathbf{0} \text{ and } \gamma^{Ber-}_{d_{Ber2}\times 1} = \mathbf{0}, \gamma^{NB+}_{d_{NB1}\times 1} = \mathbf{0} \text{ and } \gamma^{NB-}_{d_{Ber2}\times 1} = \mathbf{0}. \\ H_a: & \boldsymbol{B}^{Ber+}\gamma^{Ber+} \geq \mathbf{0}, \boldsymbol{B}^{Ber-}\gamma^{Ber-} \leq \mathbf{0}, (\boldsymbol{B}^{Ber+}\gamma^{Ber+}) \circ (\boldsymbol{B}^{Ber-}\gamma^{Ber-}) = \mathbf{0} \\ & \boldsymbol{B}^{NB+}\gamma^{NB+} \geq \mathbf{0}, \boldsymbol{B}^{NB-}\gamma^{NB-} \leq \mathbf{0}, (\boldsymbol{B}^{NB+}\gamma^{NB+}) \circ (\boldsymbol{B}^{NB-}\gamma^{NB-}) = \mathbf{0} \end{array} \tag{31}$$

Assuming that we used the same number of spline bases for both positive and negative coefficient functions, we have d_{Ber} and d_{NB} degrees of freedom for the latent Bernoulli and the negative binomial models. Since a mixture of chi-square distributions (chi-bar test) is needed for each model, the overall LRT statistic for the above test then follows a mixture of mixture of chi-square distributions. Letting the mixing probabilities be w^{Ber}_j and w^{NB}_k for Bernoulli and NB models and $\sum_{j=0}^{d_{Ber}} w^{Ber}_j = 1$ and $\sum_{k=0}^{d_{NB}} w_k = 1$, we have

$$\bar{\chi}^2_{ZINB} = -2(l_0 - l_1) \xrightarrow{d} \sum_{j=0}^{d_{Ber}} w^{Ber}_j \chi^2_j + \sum_{k=0}^{d_{NB}} w^{NB}_k \chi^2_k. \tag{32}$$

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

$$Pr(\bar{\chi}^2_{ZINB} \geq c^2) = \sum_{j,k} w^{Ber}_j w^{NB}_k P(\chi^2_{j+k} \geq c^2). \tag{33}$$

5. Simulation Studies

To study the statistical properties of the proposed cGFLM, we carried out simulations under different sampling schemes. Genotypic data were simulated under two settings: (1) random LD block with varying structures [9,11], and (2) LD block of gene CHRNA7 [24] borrowed from an existing data (COGEND, the Collaborative Genetic Study of Nicotine Dependence), which represents a real-data genomic structure. For the phenotypes, we considered traits following either a binomial or ZINB distribution. Sample sizes were set to range from 500 to 2000.

To investigate whether cGFLM can correctly control type I error, β_{causal} was set to be zero under the null hypothesis. For empirical power evaluation, we examined two different scenarios. First, we assumed only one causal SNP was located in the LD block, with varying β_{causal}s. Then we considered another interesting setting when two causal loci with reversed sign effects are located in the same LD block, which mimics the scenario when both deleterious and protective SNPs exist in a genomic region. The two causal loci chosen were weakly correlated ($r^2 < 0.01$). The corresponding regression coefficients were set to be $(\beta_{causal1}, \beta_{causal2})$ where $\beta_{causal1} = -\beta_{causal2}$. In this case, we plan to examine if the proposed test can deal with sign-heterogeneous genetic effects. Causal SNPs were removed before analyses to mimic the real data setting that causal SNPs may not be genotyped.

In terms of functional parameters, the order of B-spline basis was set to 4 (degree = 3) to construct cubic curves with desired smoothing properties. Knots were placed evenly in the position domain. In general, the number of spline bases would be determined according to the number of SNPs (p) in an LD block. Data-adaptive choices for the number or the placement of knots can be made via cross-validation, but for simplicity we will 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 robust.

Because the simulation results using the random LD blocks and the LD block structure from CHRNA7 gene are very similar, here we only show results with the random LD blocks, whereas the results with the CHRNA7 gene are located in Appendix A (Tables A1 and A2, Figures A1–A6). For a random LD block, all SNP genotypes within it were generated following the strategy introduced in [9,11]. Briefly, genotypes of p SNPs 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 the ith subject. Let $z_{0.25}$ be the third quartile of standard normal distribution, we have:

$$X_{ij} = \begin{cases} 0, & \text{if } \zeta_{ij} < -z_{0.25}. \\ 1, & \text{if } -z_{0.25} \le \zeta_{ij} < z_{0.25}. \\ 2, & \text{if } \zeta_{ij} \ge z_{0.25}. \end{cases} \tag{34}$$

The covariance matrix was 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 correlated 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 correlations among the 90% "non-tag" SNPs are determined by their physical locations ($Corr(X_{j1}, X_{j2}) = 0.7^{|j_1 - j_2|}$). In this case, we would not violate the assumptions that SNPs are physically adjacent and linked. Further, the LD block structures vary among different randomly generated arrays.

5.1. Simulation Using Binary Traits

The following binary outcomes were simulated based on causal genotypes under the logit model:

$$\text{logit}(\mu(y_i)) = \log\frac{Pr(y_i = 1)}{1 - Pr(y_i = 1)} = \alpha_0 + \boldsymbol{X_i}^\top \boldsymbol{\beta_{causal}} \tag{35}$$

We considered simulation scenarios where $\alpha_0 = 0.2$. 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). We ran 1000 replicates for each scenario for power evaluation, in which a p-value smaller than 0.05 would be declared for significance. 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 the functional linear model (GFLM). p-values for smAT were adjusted by Bonferroni correction, and p-values for SKAT-C and GFLM were calculated by combined sum test and χ^2 test.

We can see from Table 1 that cGFLM can effectively maintain the type I error. Evaluation of empirical power was based on settings when regression coefficient $\beta_{causal} = 0.1, 0.2, 0.3, 0.4$ or 0.5. 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 Figure 1 that the cGFLM showed better power than all other methods. In the second scenario when two causal loci had reverse-sign effects, we included more SNPs in a block so that it would be possible to locate two weakly correlated markers within the region. We used $p = 25$ SNPs in a group and $d_1 = d_2 = 4$ as the number of spline bases. From Figure 2, we can see that cGFLM consistently demonstrates better power than other methods.

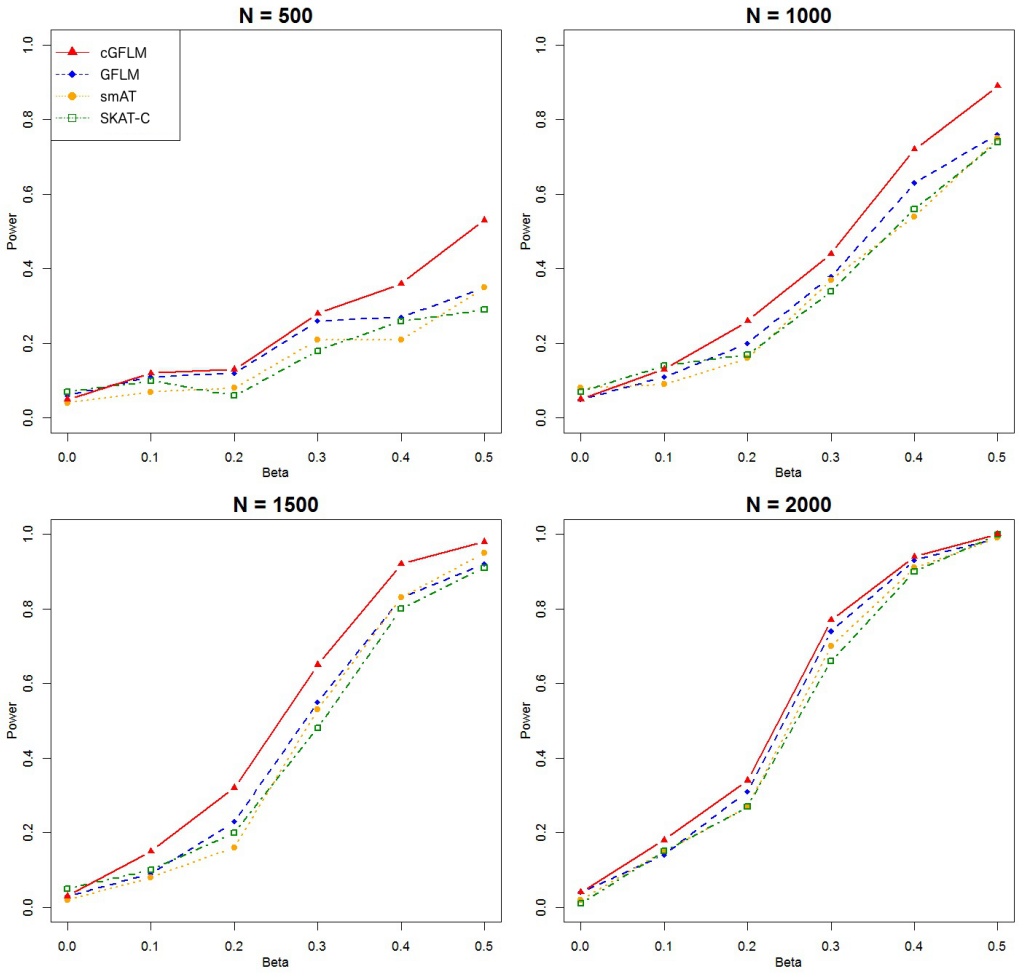

Figure 1. Power simulation for binary outcomes based on random LD blocks, single causal locus.

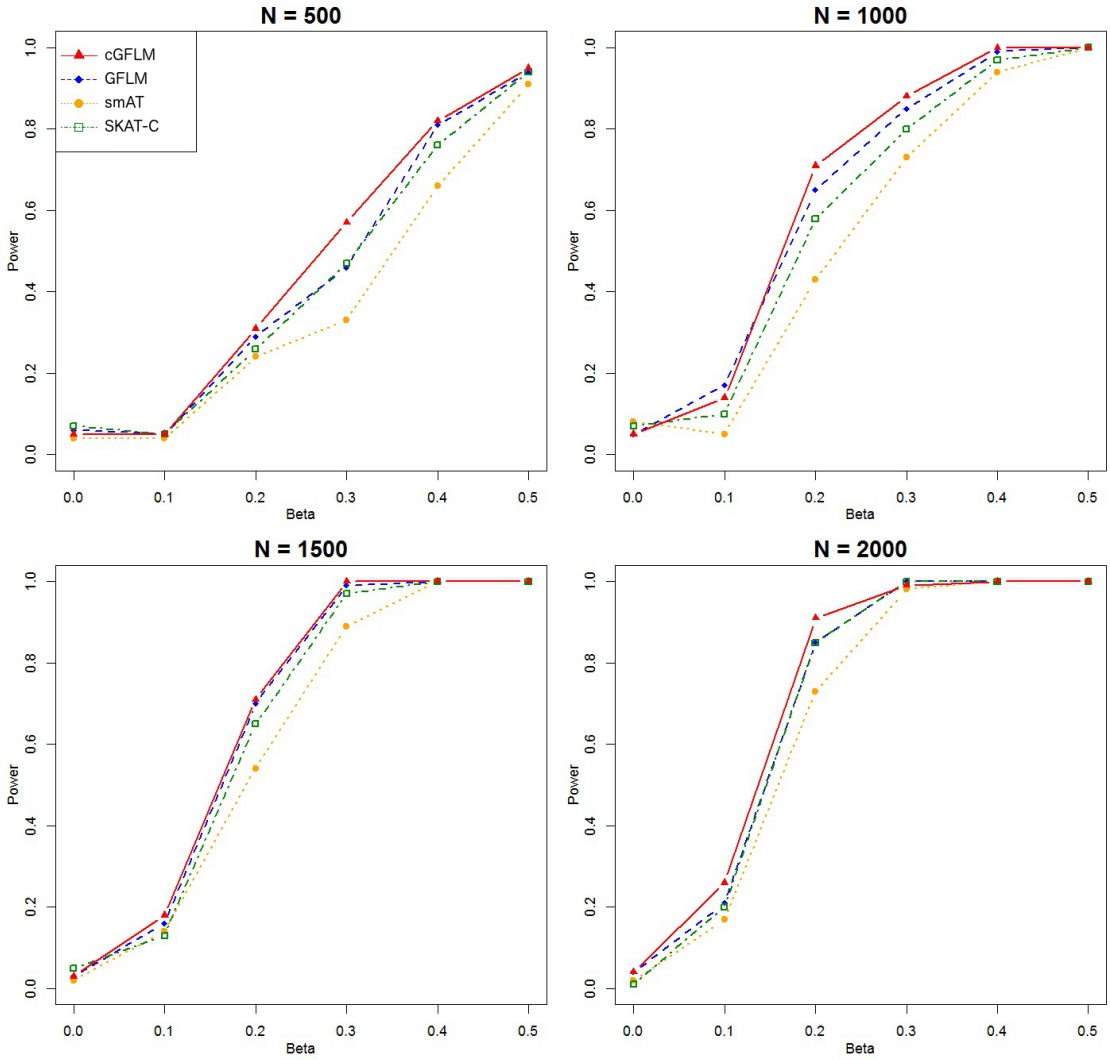

Figure 2. Power simulation for binary outcomes based on random LD blocks, two reverse-sign causal loci.

Table 1. Type I error simulation using cGFLM for binary outcomes based on randomly generated 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

The association patterns of a sample simulation are presented in Figure 3. 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). Compared to smAT and cGFLM, the cGFLM fitted coefficient function is more reasonable and correctly identify the causal loci.

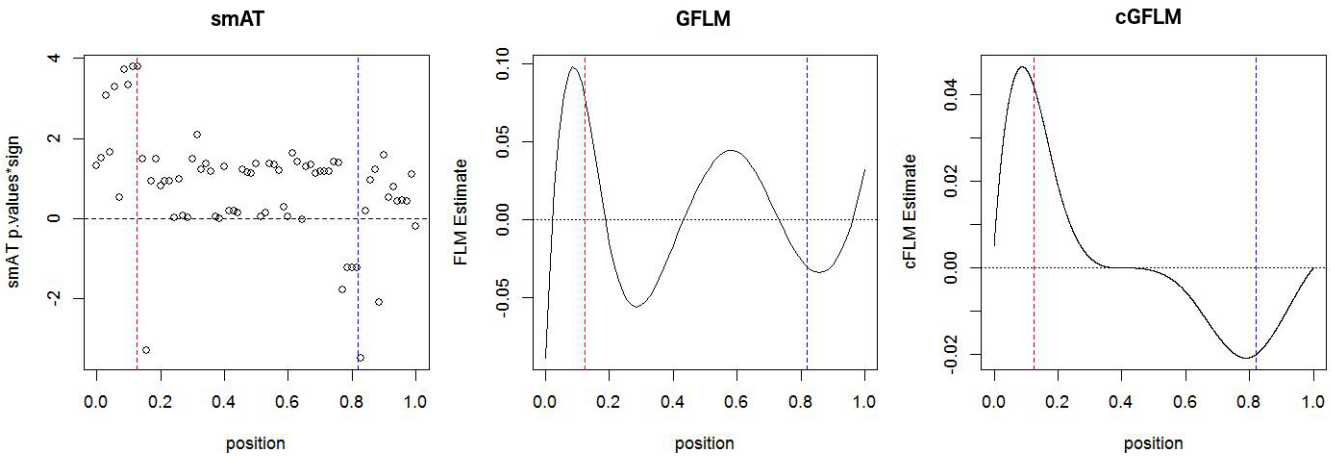

Figure 3. Illustration of fitted genetic mapping patterns using different models. A modified Manhattan plot of the $-\log10$ (p-values) by the sign of the fitted coefficients is used for smAT. For all other methods, the coefficient estimates are plotted. The causal loci is highlighted with dashed lines (left in red for positive effect, right in blue for negative effect).

5.2. Simulation Using ZINB Traits

In this case, we simulate phenotypic traits conditional on causal genotypes under the following ZINB model:

$$\begin{aligned} \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} &= \alpha_0^{Ber} + \boldsymbol{X_i}^\top \boldsymbol{\beta_{causal}^{Ber}} \\ \log(\mu_i) &= \alpha_0^{NB} + \boldsymbol{X_i}^\top \boldsymbol{\beta_{causal}^{NB}} \end{aligned} \tag{36}$$

The outcomes are generated with the latent mixture process as discussed in Section 3. The intercepts, α_0^{Ber} and α^{NB}, were set to be 0.0 in all subsequent simulations. β_{causal} was set to zero under the null hypothesis in the evaluation of type I error. Since the computational burden for ZINB models is much higher than that in the binary outcome model, we reduced the replicated times to 1000 times for each scenario. 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 used similar settings as those in the simulation for binary traits. However, we examined two general scenarios where the genetic effect is in either the latent Bernoulli or the negative binomial models. Then for each general scenario, we first set only one causal SNP in the LD block, affecting either β_{causal}^{Ber} or β_{causal}^{NB}. Then we considered 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 $(\beta_{causal1}^{Ber}, \beta_{causal2}^{Ber})$, or $(\beta_{causal1}^{NB}, \beta_{causal2}^{NB})$ where $\beta_{causal1} = -\beta_{causal2}$. A total of 100 replicates were run for each scenario. We compared the empirical power of our proposed model cGFLM-ZINB with three existing methods: functional linear model with negative binomial traits (GFLM-NB), single marker association test with ZINB traits(smAT-ZINB) and functional linear model with ZINB traits (GFLM-ZINB). p-values for smAT-ZINB were adjusted by Bonferroni correction, and p-values for GFLM-NB and GFLM-ZINB were calculated by the likelihood ratio tests.

Table 2 demonstrated that the proposed cGFLM maintains the type I errors very well. Evaluation of empirical power is based on settings with regression coefficients ranging from 0.1 to 0.5 and sample sizes from 500 to 2000. When the genetic effect was set to be 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 GFLM-NB model compared with other models in Figures 4 and 5. This fortifies our assumption that using 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, GFLM-ZINB and cGFLM-ZINB

(see Figures 6 and 7). The figures demonstrate that ZINB models are more advantageous than the NB regression model when excessive zeros exist. More importantly, the cGFLM-ZINB model consistently shows the best performance among the tested models, in scenarios when one causal locus and two causal loci were set in the LD block. Collectively, these simulations demonstrate the advantages and robustness of our proposed cGFLM model.

Table 2. Type I error simulation using cGFLM-ZINB for ZINB outcomes based on randomly generated 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

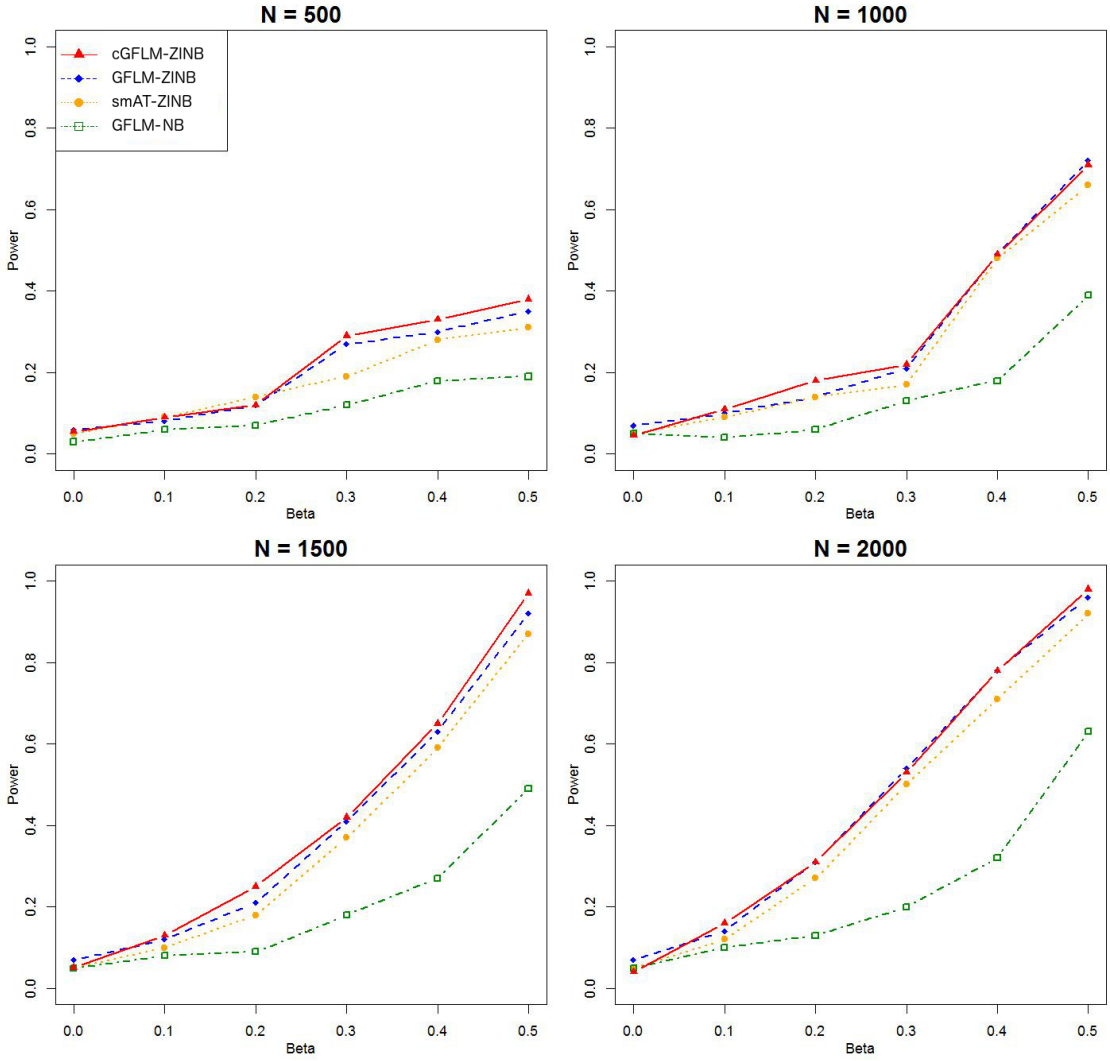

Figure 4. Power simulation for ZINB outcomes based on random LD blocks, single causal locus, effect in latent Bernoulli distribution.

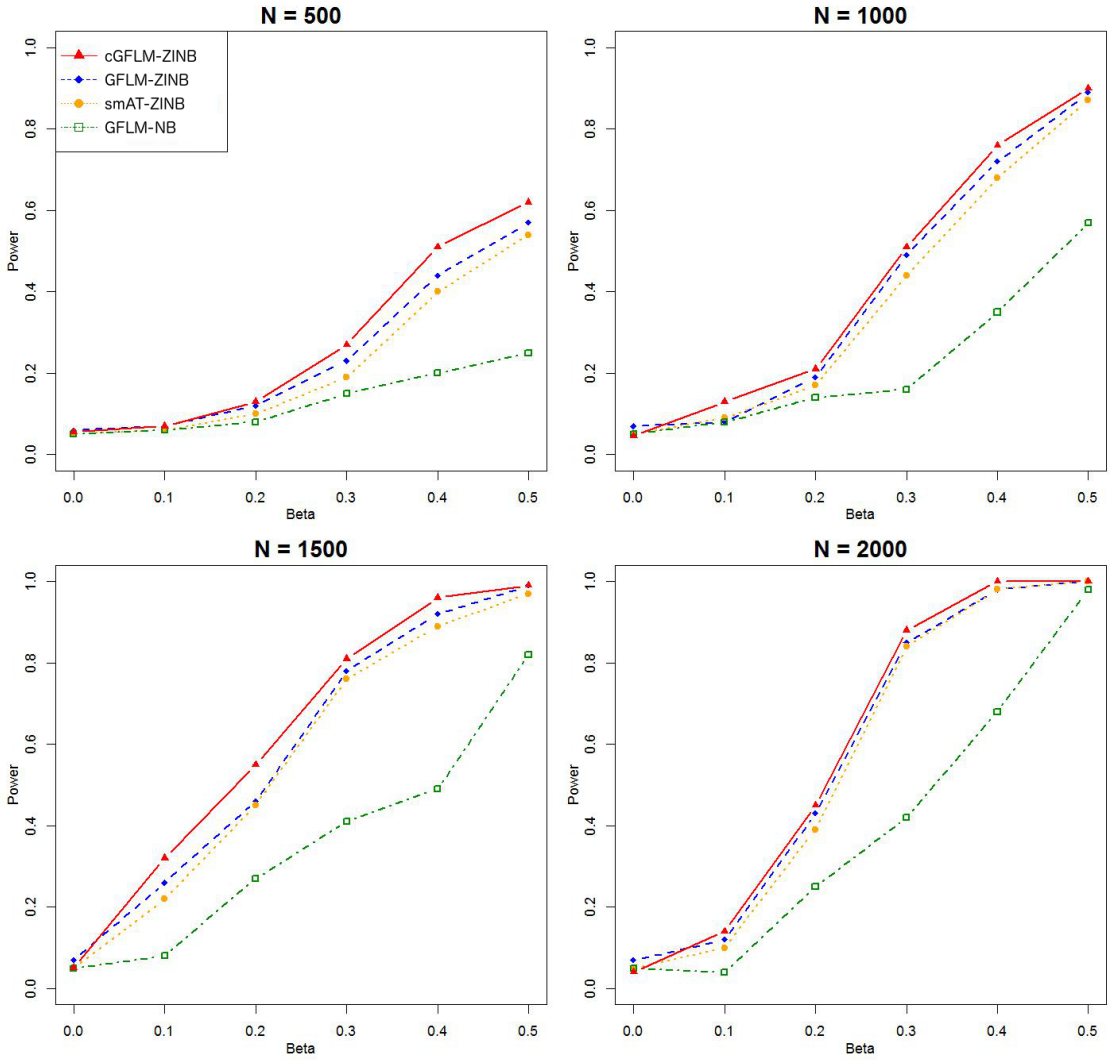

Figure 5. Power simulation for ZINB outcomes based on random LD blocks, two causal loci, effect in latent Bernoulli distribution.

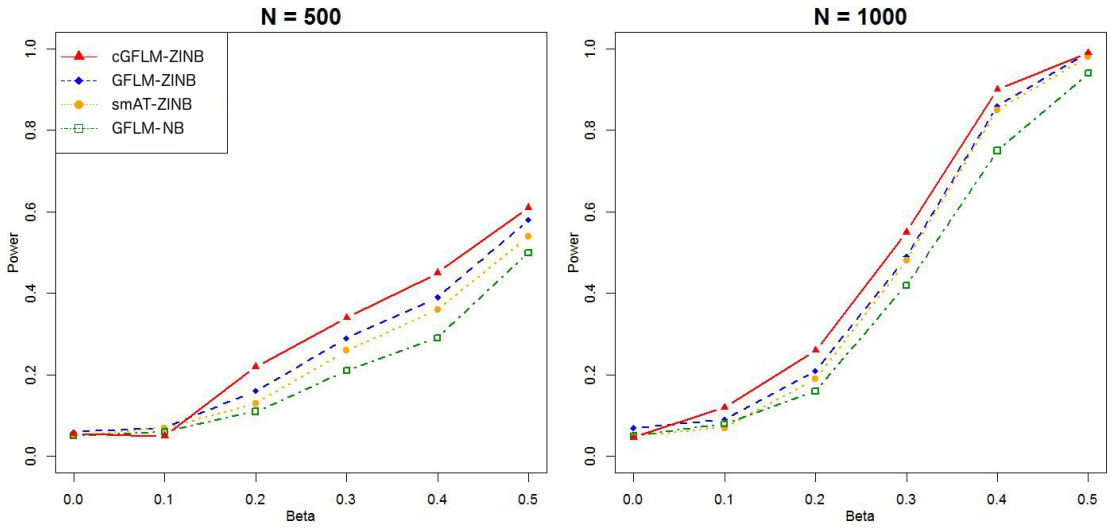

Figure 6. *Cont.*

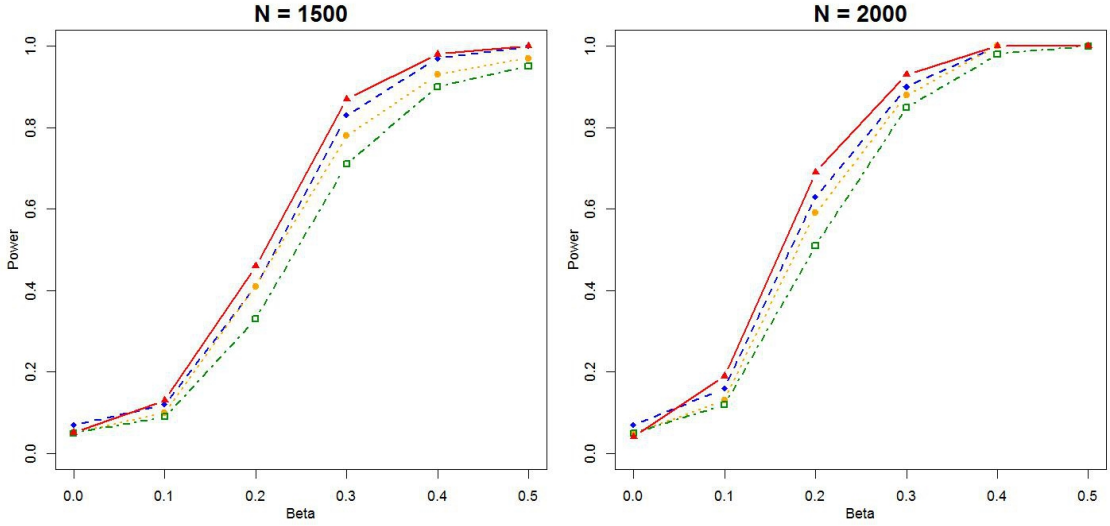

Figure 6. Power simulation for ZINB outcomes based on random LD blocks, single causal locus, effect in NB distribution.

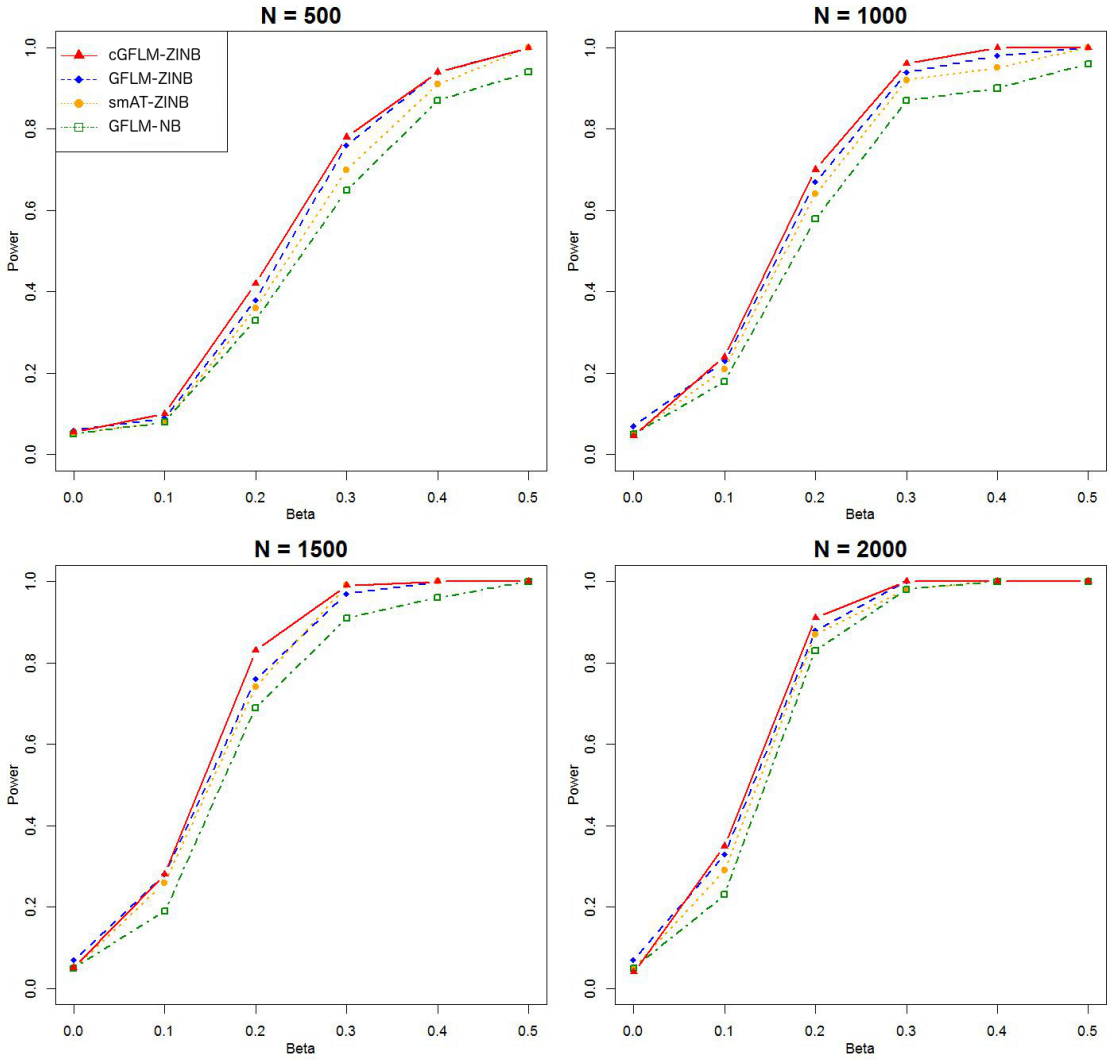

Figure 7. Power simulation for ZINB outcomes based on random LD blocks, two causal loci, effect in NB distribution.

We then applied the cGFLM model to two independent studies to assess its practical usage. The first dataset is a candidate-gene-based SNP study and the other is a GWAS for dental caries.

6. Application 1: COGEND Study

According to the World Health Statistics Report [25], 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 multifactorial, and evidence from various epidemiology studies suggest that genetic factors have a substantial impact on smoking behaviors. Identification of these genetic factors and the development of targeted 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 this study on CHRN candidate genes, a total of 216 SNPs were genotyped for 2022 individuals (1114 cases with nicotine dependence and 908 controls). In the phenotypic data, all cases and controls were current or former smokers who reported smoking more than 100 cigarettes lifetime. 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 FTND $= 0$ or 1 were classified as control. The original SNP set 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, four small blocks with fewer than seven SNPs were excluded for analyses. We applied our proposed method cGFLM, along with smAT, SKAT-C and GFLM to analyze the final dataset, which consists of 191 SNPs in eight LD blocks. Age, gender and race were included as covariates. A Bonferroni significance threshold of $0.05/8 = 6.2 \times 10^{-3}$ is used for cGFLM, GFLM and SKAT-C, and a threshold $0.05/191 = 2.6 \times 10^{-4}$ is used for smAT.

Table 3 summarizes the results. All four methods yielded small p-values (cGFLM: 5.25×10^{-6}; GFLM: 8.24×10^{-6}; SKAT-C: 1.7×10^{-3}; smAT: 1.72×10^{-4}) for the CHRNA5 cluster ("IREB2 + LOC123688 + PSMA4 + CHRNA5 + CHRNA3 + CHRNB4" gene cluster) on chromosome 15. However, only cGFLM and SKAT-C showed significance for the "CHRNB3 + CHRNA6" gene cluster (cGFLM: 8.67×10^{-4}; SKAT-C: 1.3×10^{-3}). For these two blocks, p-values calculated by cGFLM 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 [26–29]. Other LD blocks (candidate gene clusters) are not significantly associated with the phenotypic trait in this cohort.

Table 3. Association tests for COGEND study based on LD blocks using single-marker association tests with Bonferroni correction (smAT), SKAT-C, GFLM and cGFLM.

p-Value LD Block (Genes)	CHR	Length (kb)	# of SNPs	p-Value			
				cGFLM	smAT	SKAT-C	GFLM
CHRND	2	43	10	0.0990	0.4818	0.1920	0.1321
CHRNA9	4	20	11	0.4308	0.7580	0.3941	0.3915
CHRNA2	8	31	11	0.7078	0.7596	0.5698	0.8036
CHRNB3 + CHRNA6	8	101	25	8.67×10^{-4}	0.0064	0.0013	0.0033
CHRNA7	15	126	39	0.2759	0.3382	0.6078	0.5587
CHRNA5 cluster	15	212	72	5.25×10^{-6}	1.72×10^{-4}	0.0017	8.24×10^{-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

7. Application 2: Whole Genome Association Study for Dental Caries

More than 40% children and adolescents, and 90% adults in the US are being affected by dental caries, or more commonly known as tooth decay. Multiple factors are considered to contribute to the risk of having dental caries, such as some environmental factors and social behaviors [30–32]. 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 developing caries [33]. 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) [4,34]. A total of 4020 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 interest were both count traits with excess zeroes (Figure 8), the proposed methods, zero-inflated negative binomial model (smAT-ZINB for single-marker tests) and its application with functional coefficient (GFLM-ZINB and cGFLM-ZINB), were applied to the data set. The final analytic sample consists of 1480 individuals with complete permanent teeth phenotypic data. Age, gender and total number of teeth/surfaces were included as covariates in the analyses.

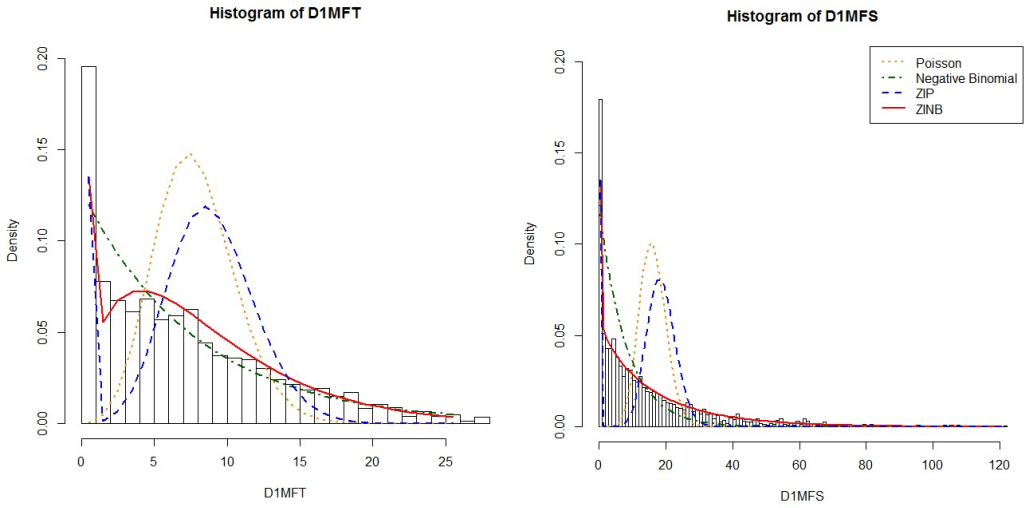

Figure 8. Histograms and fitted densities for traits D1MFT and D1MFS in dental caries study.

Tables 4 and 5 summarize the significant findings. The Manhattan plots for GWAS scans using the ZINB model are presented in Figures 9 and 10. For genome-wide univariate screening, a significance threshold of 1×10^{-7} was used. For the LD block-based analysis, a genome-wide significance threshold of 2.5×10^{-6} was used. 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 the LD block based association tests, the cGFLM-ZINB model identified two significant genetic regions: PKDCC in chromosome 2 (both traits), and the intergenic region between DCN and BTG1 in chromosome 12 (D1MFS). It is worth mentioning these two genetic regions cannot be identified by other competing methods, suggesting that the cGFLM model may provide potentially new insights into understanding the risk of dental caries. More interestingly, the gene PKDCC was found to be associated with craniofacial morphogenesis in previous dental studies [35], further supporting and validating the biological significance of our findings.

Table 4. Significant findings for dental caries GWAS scanning using single-marker association tests based on the ZINB model.

Trait: D1MFT				
SNP ID	**CHR**	**Gene**	**MAF**	***p*-Value**
rs7990965	13	-	0.033	1.02×10^{-12}
rs1058595	10	PHYH	0.063	2.52×10^{-9}
rs12344120	9	-	0.029	4.79×10^{-8}
rs4694666	4	MTHFD2L	0.135	5.48×10^{-8}
rs17078140	3	LIMD1	0.029	5.77×10^{-8}
rs9893536	17	USP32	0.091	8.80×10^{-8}
rs7334525	13	RFC3	0.093	9.52×10^{-8}
Trait: D1MFS				
SNP ID	**CHR**	**Gene**	**MAF**	***p*-Value**
rs7990965	13	-	0.033	2.40×10^{-10}
rs1058595	10	PHYH	0.063	3.29×10^{-9}

Table 5. Significant findings for dental caries association tests based on LD blocks (gene clusters) using the ZINB model (smAT-ZINB with Bonferroni correction, GFLM-ZINB, cGFLM-ZINB).

		Trait: D1MFT				
LD Block	**CHR**	**Length**	**# of**	***p*-Value**		
(Genes)		**(kb)**	**SNPs**	**cGFLM**	**GFLM**	**smAT**
PKDCC	2	70	18	8.41×10^{-7}	8.06×10^{-5}	4.17×10^{-5}
		Trait: D1MFS				
LD Block	**CHR**	**Length**	**# of**	***p*-Value**		
(Genes)		**(kb)**	**SNPs**	**cGFLM**	**GFLM**	**smAT**
Intergenic between DCN and BTG1	12	70	21	6.07×10^{-7}	5.92×10^{-6}	4.41×10^{-3}
PKDCC	2	70	18	1.38×10^{-6}	1.59×10^{-4}	7.73×10^{-6}

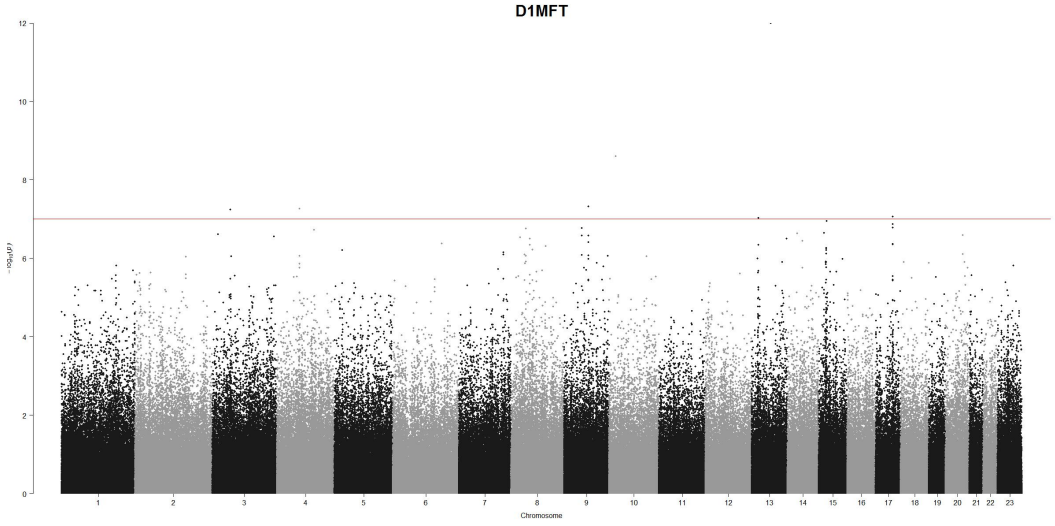

Figure 9. Genome-wide scanning for trait D1MFT using single-marker association tests based on the ZINB model.

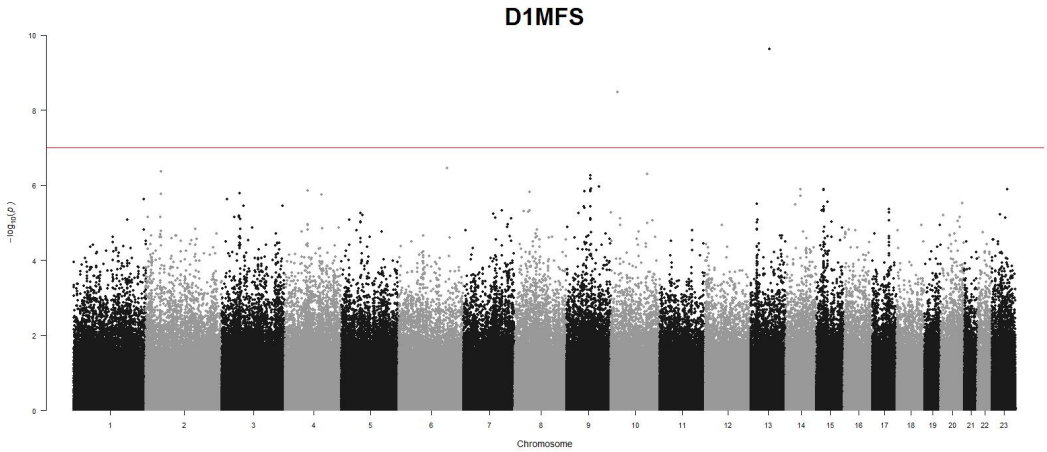

Figure 10. Genome-wide scanning for trait D1MFS using single-marker association tests based on ZINB model.

8. Discussion

Joint analyses of multiple contiguous SNPs, which can form an LD block, are expected to provide better inference about unknown causal variants, since these SNPs all carry partial information about the causal variants and collectively they should be more powerful. In principle, LD between genetic loci should decline with their intervening physical distance, given that more cross-over events occur within longer ranges. Therefore, how to effectively incorporate and take full advantage of the distance/alignment order information of these SNPs is an important yet challenging problem. While FLMs seem to provide a good solution, its estimated coefficient function is usually noisy and hard to interpret, which consequently leads to potential power loss. Improvements to existing methods are then of great interest in order to better detect significant genetic variants.

In this article, we proposed a novel cGFLM for flexible and more reasonable multi-loci mapping strategy. Our model is built upon the general FLM framework by imposing constraints to specify sign-specific effects. Our limited simulations suggest that these constraints encourage spatial sparsity in the estimated coefficient function, which needs to be further validated in other experimental settings. Due to these constraints, the likelihood ratio test statistic does not follow a fixed chi-square distribution, but a null weighted mixture of chi-square distributions. The simulation results show that compared to three competing methods, our proposed cGFLM generally demonstrated better power when effect size is moderate and large, and comparable performance when effect size is small, while at the same time maintaining correct type I errors.

Applications of the cGFLM to two real datasets demonstrate the applicability of our method. Particulary, its application to the GWAS of dental caries risk identifies new genetic regions that are have not been discovered before, validating that our method can potentially provide new insights into large-scale genomic studies. However, it is also important to note that some significant SNPs identified by other methods were missed by cGFLM as well. This is likely because in cGFLM, identification of significant SNPs is based on the collective evidence of a group of linked SNP, and in the case of causal SNPs being weakly linked to its neighboring SNPs, our methods may miss them. Additionally, cGFLM inherently would not work well for variants not in LD with neighboring SNPs, such as singletons and doubletons. Therefore, these observations suggest that our method should be considered as an additional approach in our toolbox for more discoveries, while not replacing existing ones.

When the cGFLM method is applied to large-scale genome-wide scanning of LD blocks, one concern is how to group SNPs into LD blocks. Several software, such as PLINK [36] and LDExplorer [37], have embedded functions to define LD blocks for genomic data, and we can utilize them to help partition the genome. Even with these tools, LD blocks can

vary with different LD thresholds. That is, a weaker LD threshold can lead to larger LD blocks and vice versa. Another concern is about LD blocks with few SNPs, since they are not suitable for functional analysis. They can be either combined into adjacent larger blocks with weak LD, or can be just analyzed with those single-marker methods. Prospectively, in order to discover the core subset of causal genes, further group selection among multiple candidate blocks is of great interest, and regularization methods such as group LASSO and group SCAD can be included as extensions to our current framework. Additionally, in this study cGFLM only considers LD-based blocks for association analyses; however, it would be also of great interest to see if cGFLM can be applied to gene-level analyses as genes are functional units of the genome. Finally, if subpopulations exist in the genotypic sample, stratification can be addressed in our model by including principal components of population variation as additional covariates [38].

9. Conclusions

Our proposed GFLM, cGFLM and their related implementations with the ZINB model (GFLM-ZINB, cGFLM-ZINB) are novel and complex methods. They are specifically designed for multi-loci mapping in naturally formed LD blocks. The models can simultaneously incorporate multiple linked SNPs, including their physical alignments and block-wide linkage structure, and make inference at the LD block level. Simulation studies show that cGFLM and cGFLM-ZINB have desirable performances in terms of both simpler coefficients functions and empirical power gain. The practical usage of cGFLM is demonstrated with a candidate-gene study of nicotine dependence and a GWAS on dental caries. Considering its flexibility and comprehensiveness, cGFLM would be a very attractive method for future gene-based association studies.

Author Contributions: Conceptualization, J.H., J.Y., W.Z. and S.W.; methodology, J.H., Z.G. and S.W.; software, J.H.; Data Analyses and interpretation, J.H. and S.W.; writing—original draft preparation, J.H.; writing—review and editing, J.H., J.Y., W.Z. and S.W.; Critical revision, J.H., J.Y., Z.G., W.Z. and S.W.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable as this was secondary use of publicly available data.

Informed Consent Statement: Not applicable as this was secondary use of publicly available data.

Data Availability Statement: The application datasets are publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Simulation Results Using the LD Block Information in CHRNA7 Gene

Table A1. Type I error simulation using cGFLM for binary outcomes based on the 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

Table A2. Type I error simulation using cGFLM-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

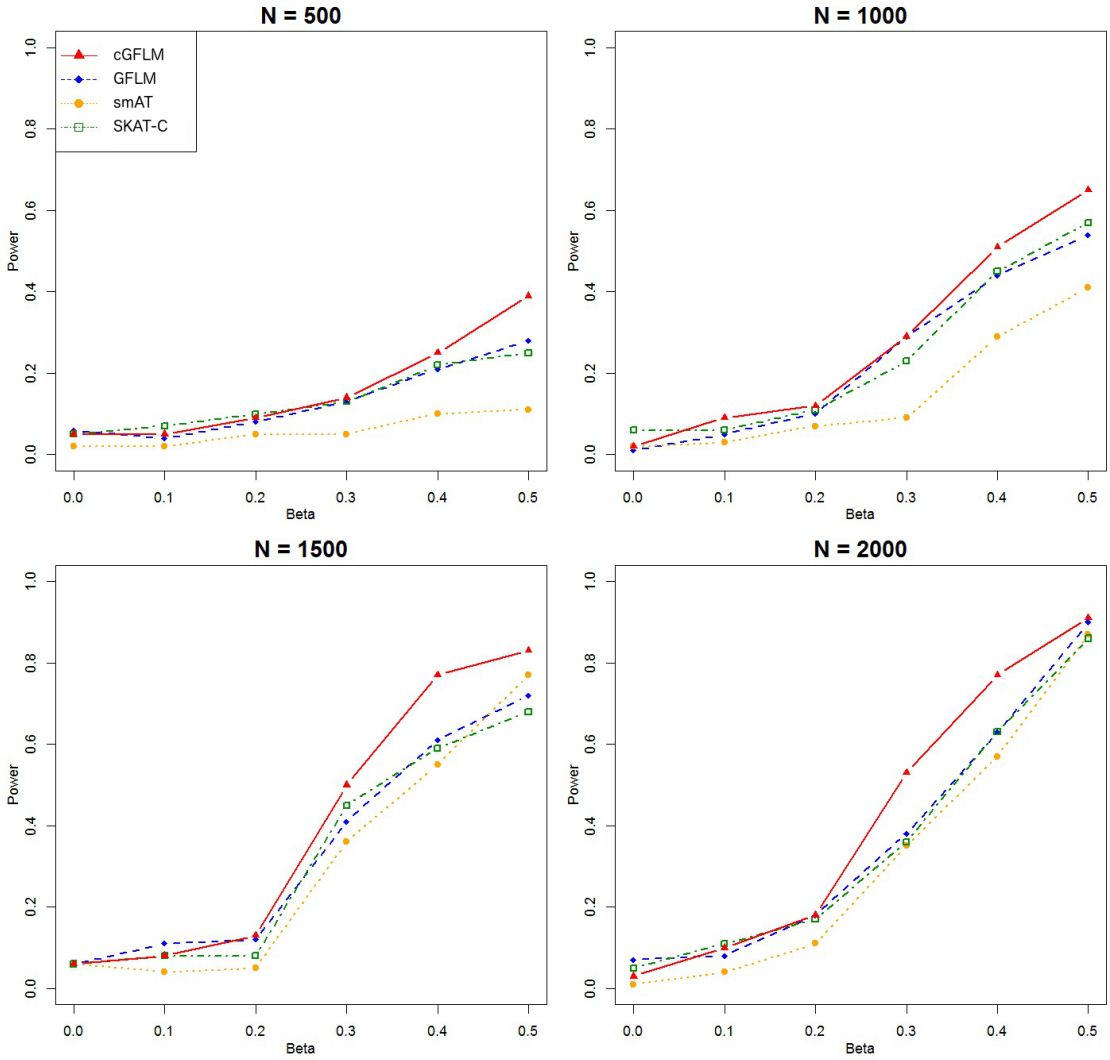

Figure A1. Power simulation for binary outcomes based on the CHRNA7 gene, single causal locus.

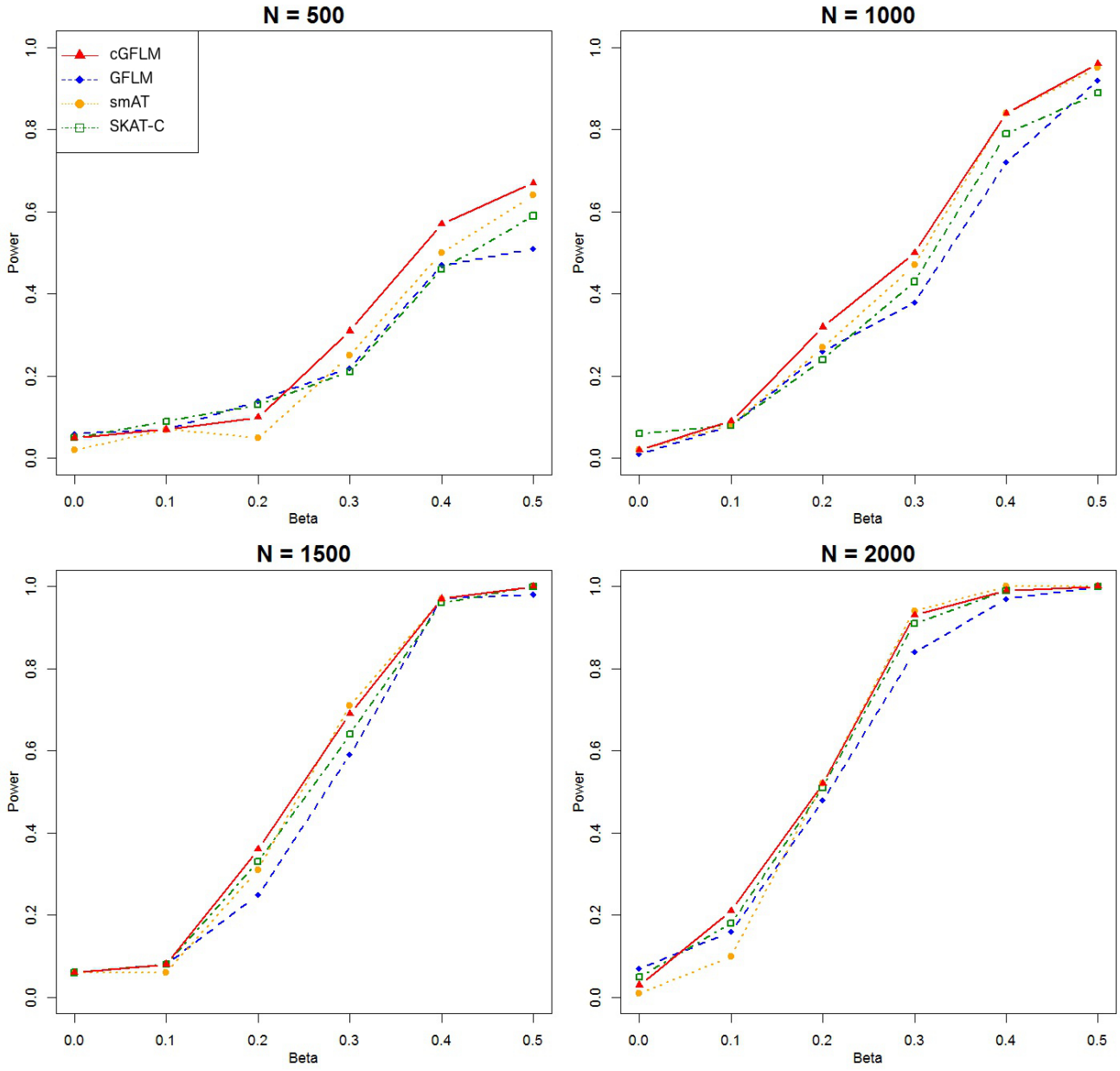

Figure A2. Power simulation for binary outcomes based on the CHRNA7 gene, two reverse-sign causal loci.

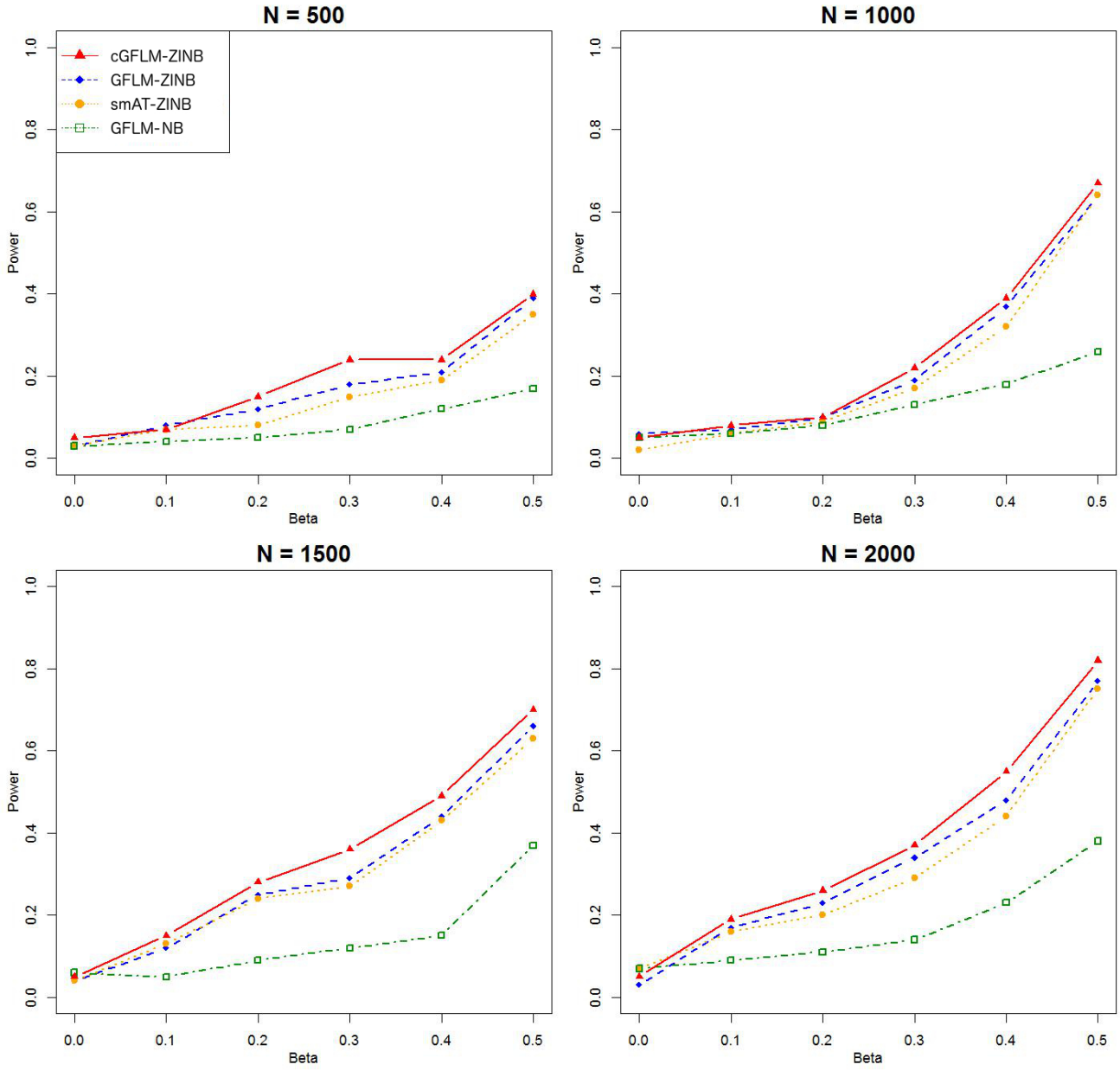

Figure A3. Power simulation for ZINB outcomes based on the CHRNA7 gene, single causal locus, effect in latent Bernoulli distribution.

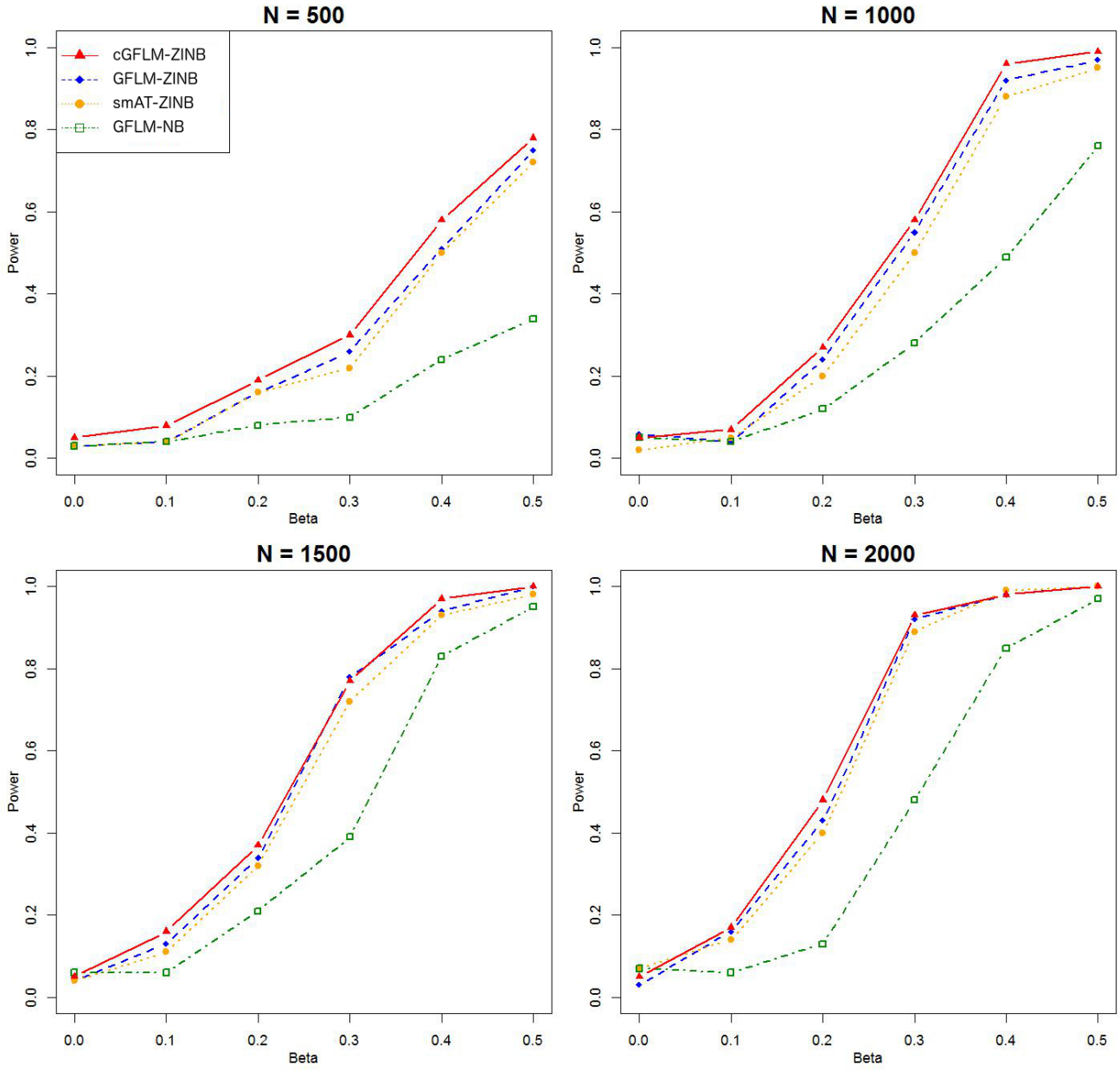

Figure A4. Power simulation for ZINB outcomes based on the CHRNA7 gene, two causal loci, effect in latent Bernoulli distribution.

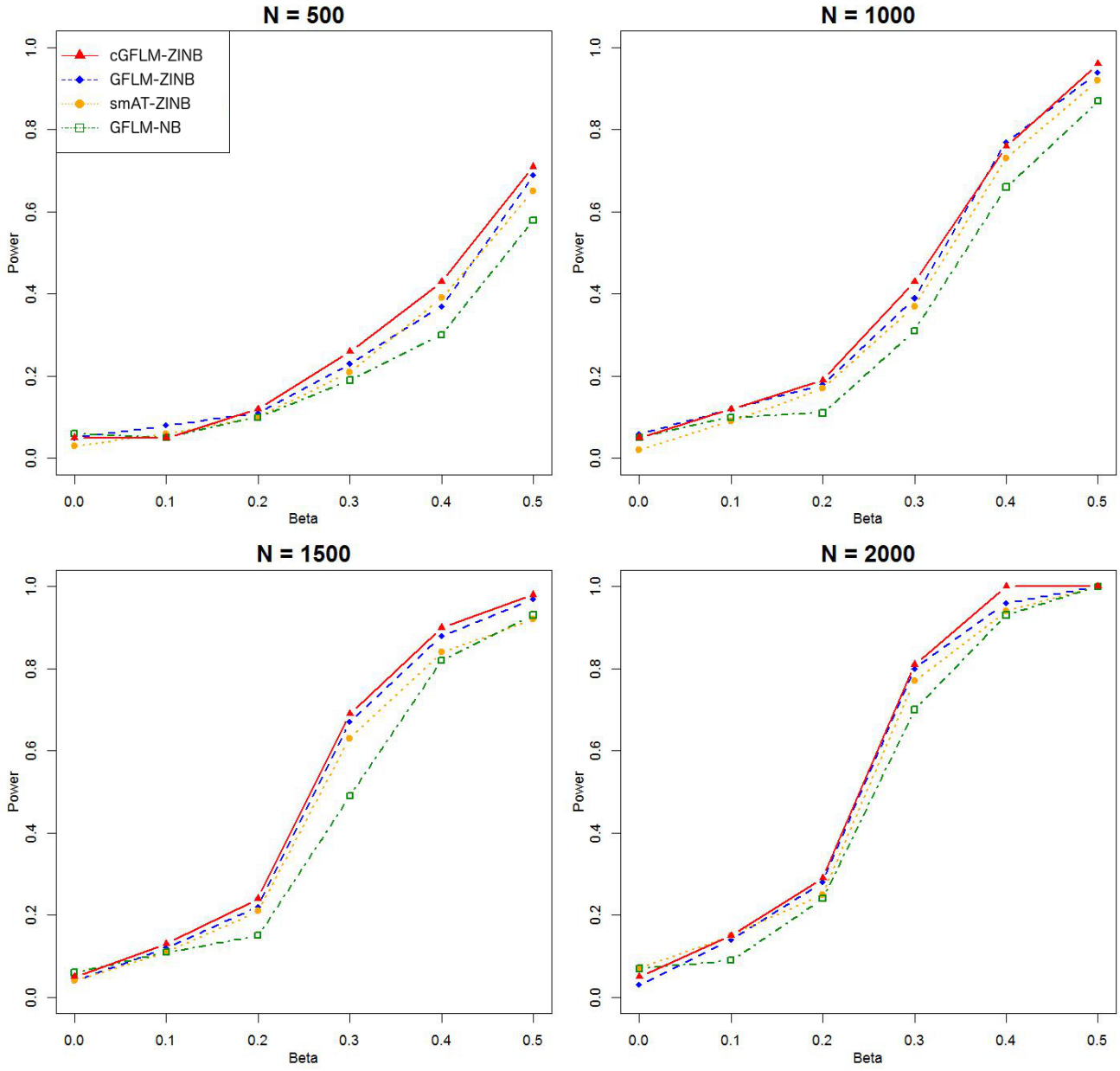

Figure A5. Power simulation for ZINB outcomes based on the CHRNA7 gene, single causal locus, effect in NB distribution.

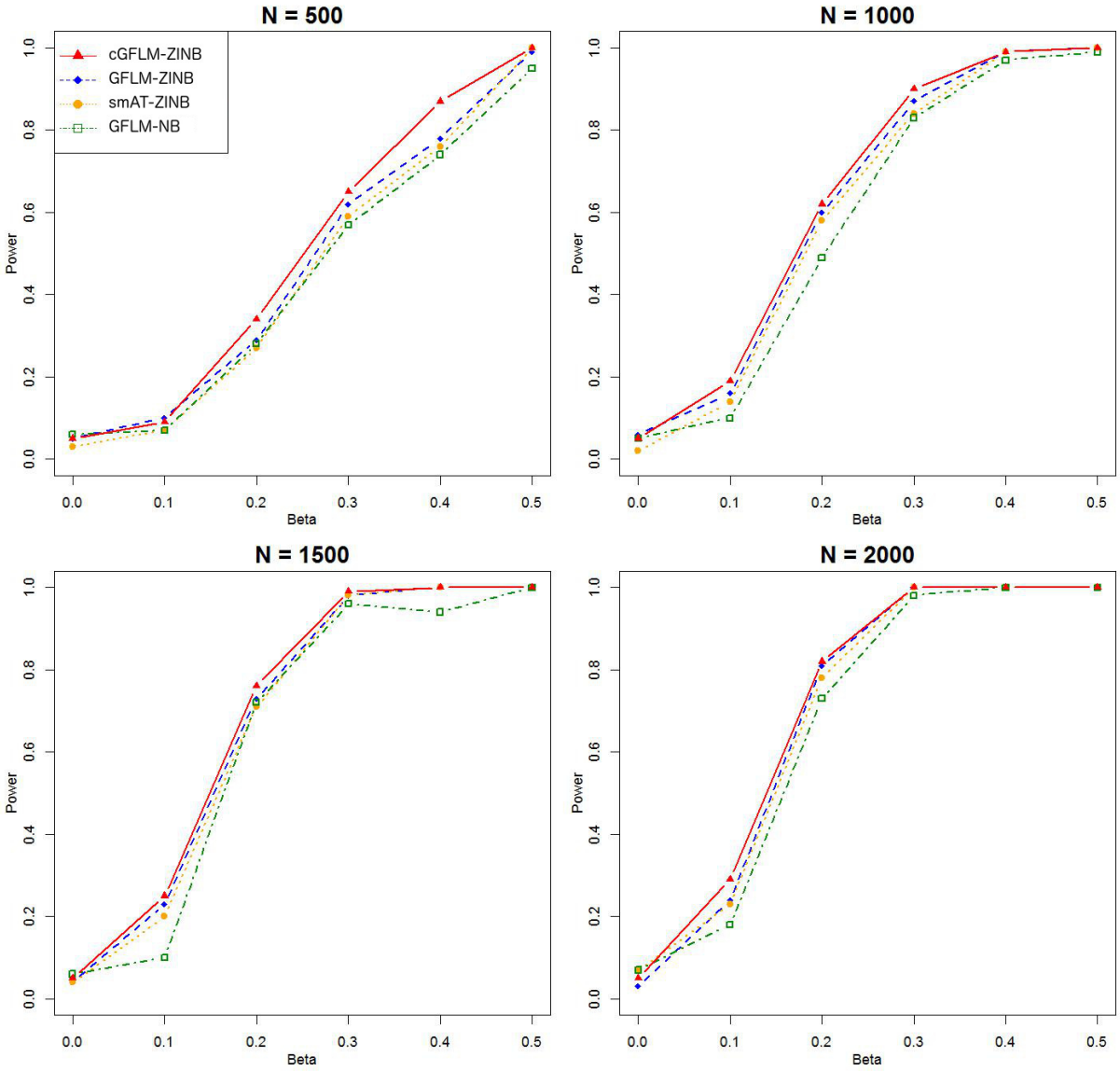

Figure A6. Power simulation for ZINB outcomes based on the CHRNA7 gene, two causal loci, effect in NB distribution.

References

1. Cantor, M.R.; Lange, K.; Sinsheimer, S.J. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *Am. J. Hum. Genet.* **2010**, *86*, 6–22. [CrossRef] [PubMed]
2. Wang, K.; Abbott, D. A principal components regression approach to multilocus genetic association studies. *Genet. Epidemiol.* **2008**, *32*, 108–118. [CrossRef] [PubMed]
3. Ionita-Laza, I.; Lee, S.; Makarov, V.; Buxbaum, J.D.; Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* **2013**, *92*, 841–853. [CrossRef] [PubMed]
4. Yang, J.; Zhu, W.; Chen, J.; Zhang, Q.; Wu, S. Genome-wide Two-marker linkage disequilibrium mapping of quantitative trait loci. *BMC Genet.* **2014**, *15*, 20. [CrossRef] [PubMed]
5. Zhang, H.; Zhao, N.; Mehrotra, D.V.; Shen, J. Composite Kernel Association Test (CKAT) for SNP-set joint assessment of genotype and genotype-by-treatment interaction in Pharmacogenetics studies. *Bioinformatics* **2020**, *36*, 3162–3168. [CrossRef]
6. Tian, Y.; Ma, L.; Cai, X.; Zhu, J. Statistical Method Based on Bayes-Type Empirical Score Test for Assessing Genetic Association with Multilocus Genotype Data. *Int. J. Genom.* **2020**, *2020*, 4708152. [CrossRef]
7. Cui, Y.; Kang, G.; Sun, K.; Qian, M.; Romero, R.; Fu, W. Gene-centric genomewide association study via entropy. *Genetics* **2008**, *179*, 637–650. [CrossRef]

8. Malten, J.; König, I.R. Modified entropy-based procedure detects gene-gene-interactions in unconventional genetic models. *BMC Med. Genom.* **2020**, *13*, 65. [CrossRef]
9. Wu, T.T.; Chen, Y.F.; Hastie, T.; Sobel, E.; Lange, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **2009**, *25*, 714–721. [CrossRef]
10. Yang, S.; Wen, J.; Eckert, S.T.; Wang, Y.; Liu, D.J.; Wu, R.; Li, R.; Zhan, X. Prioritizing genetic variants in GWAS with lasso using permutation-assisted tuning. *Bioinformatics* **2020**, *36*, 3811–3817. [CrossRef]
11. Liu, J.; Wang, K.; Ma, S.; Huang, J. Regularized regression method for genome-wide association studies. *BMC Proc.* **2011**, *5*, S67. [CrossRef]
12. Fan, R.; Wang, Y.; Mills, J.L.; Wilson, A.F.; Bailey-Wilson, J.E.; Xiong, M. Functional linear models for association analysis of quantitative traits. *Genet. Epidemiol.* **2013**, *37*, 726–742. [CrossRef] [PubMed]
13. Cardot, H.; Ferraty, F.; Sarda, P. Spline estimators for the functional linear model. *Stat. Sin.* **2003**, *13*, 571–592.
14. Luo, L.; Zhu, Y.; Xiong, M. Quantitative trait locus analysis for next-generation sequencing with the functional linear models. *J. Med. Genet.* **2012**, *49*, 513–524. [CrossRef] [PubMed]
15. Huang, T.; Saporta, G.; Wang, H.; Wang, S. Spatial functional linear model and its estimation method. *arXiv* **2018**, arXiv:1811.00314.
16. Belonogova, N.M.; Svishcheva, G.R.; Wilson, J.F.; Campbell, H.; Axenovich, T.I. Weighted functional linear regression models for gene-based association analysis. *PLoS ONE* **2018**, *13*, e0190486. [CrossRef]
17. Ridout, M.; Hinde, J.; DeméAtrio, C.G. A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives. *Biometrics* **2001**, *57*, 219–223. [CrossRef]
18. Birgin, E.G.; Martínez, J.M. Improving ultimate convergence of an augmented Lagrangian method. *Optim. Methods Softw.* **2008**, *23*, 177–195. [CrossRef]
19. Birgin, E.; Martínez, J. Complexity and performance of an Augmented Lagrangian algorithm. *Optim. Methods Softw.* **2020**, *35*, 885–920. [CrossRef]
20. Andreani, R.; Birgin, E.G.; Martinez, J.M.; Schuverdt, M.L. Augmented Lagrangian methods under the Constant Positive Linear Dependence constraint qualification. *Math. Programm.* **2008**, *111*, 5–32. [CrossRef]
21. Liu, J.; Ye, J. Efficient Euclidean Projections in Linear Time. In Proceedings of the 26th International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 657–664.
22. Shapiro, A. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* **1985**, *72*, 133–144. [CrossRef]
23. Liu, X. Likelihood ratio test for and against nonlinear inequality constraints. *Metrika* **2007**, *65*, 93–108. [CrossRef]
24. Cameli, C.; Bacchelli, E.; De Paola, M.; Giucastro, G.; Cifiello, S.; Collo, G.; Cainazzo, M.M.; Pini, L.A.; Maestrini, E.; Zoli, M. Genetic variation in CHRNA7 and CHRFAM7A is associated with nicotine dependence and response to varenicline treatment. *Eur. J. Hum. Genet.* **2018**, *26*, 1824–1831. [CrossRef] [PubMed]
25. World Health Organization. *World Health Statistics 2013*; World Health Organization: Geneva, Switzerland, 2013; Volume 1.
26. Saccone, N.L.; Wang, J.C.; Breslau, N.; Johnson, E.O.; Hatsukami, D.; Saccone, S.F.; Grucza, R.A.; Sun, L.; Duan, W.; Budde, J.; et al. The CHRNA5-CHRNA3-CHRNB4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans. *Cancer Res.* **2009**, *69*, 6848–6856. [CrossRef]
27. Culverhouse, R.C.; Johnson, E.O.; Breslau, N.; Hatsukami, D.K.; Sadler, B.; Brooks, A.I.; Hesselbrock, V.M.; Schuckit, M.A.; Tischfield, J.A.; Goate, A.M.; et al. Multiple distinct CHRNB3–CHRNA6 variants are genetic risk factors for nicotine dependence in African Americans and European Americans. *Addiction* **2014**, *109*, 814–822. [CrossRef] [PubMed]
28. Wen, L.; Han, H.; Liu, Q.; Su, K.; Yang, Z.; Cui, W.; Yuan, W.; Ma, Y.; Fan, R.; Chen, J.; et al. Significant association of the CHRNB3-CHRNA6 gene cluster with nicotine dependence in the Chinese Han population. *Sci. Rep.* **2017**, *7*, 9745. [CrossRef]
29. Liu, Q.; Han, H.; Wang, M.; Yao, Y.; Wen, L.; Jiang, K.; Ma, Y.; Fan, R.; Chen, J.; Su, K.; et al. Association and cis-mQTL analysis of variants in CHRNA3-A5, CHRNA7, CHRNB2, and CHRNB4 in relation to nicotine dependence in a Chinese Han population. *Transl. Psychiatry* **2018**, *8*, 1–10. [CrossRef]
30. Ditmyer, M.M.; Dounis, G.; Howard, K.M.; Mobley, C.; Cappelli, D. Validation of a multifactorial risk factor model used for predicting future caries risk with Nevada adolescents. *BMC Oral Health* **2011**, *11*, 18. [CrossRef]
31. Silva, M.J.; Kilpatrick, N.M.; Craig, J.M.; Manton, D.J.; Leong, P.; Burgner, D.P.; Scurrah, K.J. Genetic and early-life environmental influences on dental caries risk: A twin study. *Pediatrics* **2019**, *143*, e20183499. [CrossRef] [PubMed]
32. Lendrawati, L.; Pintauli, S.; Rahardjo, A.; Bachtiar, A.; Maharani, D.A. Risk factors of dental caries: Consumption of sugary snacks among indonesian adolescents. *Pesqui. Bras. Odontopediatria Clin. Integr.* **2019**, *19*. [CrossRef]
33. Bretz, W.A.; Corby, P.M.; Melo, M.R.; Coelho, M.Q.; Costa, S.M.; Robinson, M.; Schork, N.J.; Drewnowski, A.; Hart, T.C. Heritability estimates for dental caries and sucrose sweetness preference. *Arch. Oral Biol.* **2006**, *51*, 1156–1160. [CrossRef]
34. Wang, Q.; Jia, P.; Cuenco, K.T.; Zeng, Z.; Feingold, E. Association signals unveiled by a comprehensive gene set enrichment analysis of dental caries genome-wide association studies. *PLoS ONE* **2013**, *8*, e72653. [CrossRef] [PubMed]
35. Melvin, V.S.; Feng, W.; Hernandez-Lagunas, L.; Artinger, K.B.; Williams, T. A morpholino-based screen to identify novel genes involved in craniofacial morphogenesis. *Dev. Dyn.* **2013**, *242*, 817–831. [CrossRef] [PubMed]

36. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; De Bakker, P.I.; Daly, M.J.; et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef] [PubMed]
37. Taliun, D.; Gamper, J.; Pattaro, C. LDExplorer. 2013. Available online: http://www.eurac.edu/en/research/health/biomed/services/Pages/LDExplorer.aspx (accessed on 20 November 2015).
38. Price, A.L.; Patterson, N.J.; Plenge, R.M.; Weinblatt, M.E.; Shadick, N.A.; Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **2006**, *38*, 904–909. [CrossRef]