Received: 6 October 2020 | Revised: 21 October 2021 | Accepted: 25 October 2021

DOI: 10.1002/sim.9257

RESEARCH ARTICLE

Statistics in Medicine WILEY

A conditional autoregressive model for genetic association analysis accounting for genetic heterogeneity

Xiaoxi Shen[1,4] | Yalu Wen[2] | Yuehua Cui[3] | Qing Lu[4]

[1]Department of Mathematics, Texas State University, San Marcos, Texas

[2]Department of Statistics, University of Auckland, Auckland, New Zealand

[3]Department of Statistics and Probability, Michigan State University, East Lansing, Michigan

[4]Department of Biostatistics, University of Florida, Gainesville, Florida

Correspondence
Qing Lu, Department of Biostatistics, University of Florida, 2004 Mowry Road, 5th Floor CTRB Gainesville, FL 32611-7450, USA.
Email: lucienq@ufl.edu

Funding information
National Institute on Drug Abuse, Grant/Award Number: R01DA043501; U.S. National Library of Medicine, Grant/Award Number: R01LM012848

Abstract

Converging evidence from genetic studies and population genetics theory suggest that complex diseases are characterized by remarkable genetic heterogeneity, and individual rare mutations with different effects could collectively play an important role in human diseases. Many existing statistical models for association analysis assume homogeneous effects of genetic variants across all individuals, and could be subject to power loss in the presence of genetic heterogeneity. To consider possible heterogeneous genetic effects among individuals, we propose a conditional autoregressive model. In the proposed method, the genetic effect is considered as a random effect and a score test is developed to test the variance component of genetic random effect. Through simulations, we compare the type I error and power performance of the proposed method with those of the generalized genetic random field and the sequence kernel association test methods under different disease scenarios. We find that our method outperforms the other two methods when (i) the rare variants have the major contribution to the disease, or (ii) the genetic effects vary in different individuals or subgroups of individuals. Finally, we illustrate the new method by applying it to the whole genome sequencing data from the Alzheimer's Disease Neuroimaging Initiative.

KEYWORDS

genetic heterogeneity, score test

1 | INTRODUCTION

Substantial evidence from a wide range of diseases (eg, breast cancer and hearing loss) indicates that complex diseases are characterized by remarkable genetic heterogeneity.[1] Evolutionary studies also suggest that individually rare mutations generated from each generation create vast genetic heterogeneity in human diseases and could collectively play a substantial role in causing diseases. The recently developed whole-genome sequencing technology generates a deep catalog of genetic variants, especially those rare variants, and allows researchers to comprehensively investigate their role in human diseases. Although new technology holds promise for uncovering novel disease-associated variants, the massive amount of sequencing data and low frequency of rare variants bring tremendous analytical challenges to sequencing data analysis. Further challenge comes from sequencing variants, especially those rare variants, which could be highly heterogeneous: (1) the same gene may harbor many (hundreds of even thousands) different rare mutations; and (2) the same variant may have heterogeneous effects in different individuals or subgroups of individuals.[1] These degrees of genetic heterogeneity often have been neglected in existing statistical frameworks, adding another layer of difficulty to the discovery process.

Many new statistical methods have been proposed to deal with the joint association analysis of single nucleotide variants (SNVs), including rare variants. The burden test,[2-4] as a pioneer in testing the genetic association on sequencing data, collapses all the genetic information through a weighted sum. The burden test performs well if the effects of SNVs are in the same direction and same magnitude. Nevertheless, it is subject to power loss if the assumption fails. Similarity-based methods have been proposed to address this issue. One of the most popular methods is the sequence kernel association test (SKAT),[5] which is a semi-parametric method. SKAT is closely related to the sum of square score (SSU) test,[6] and can detect both uni-directional and bi-directional genetic effects. More recently, a genetic random field model (GenRF)[7] was proposed in analyzing sequencing data for continuous phenotypes, and a generalized genetic random field model (GGRF)[8] was proposed to generalize the random field model for other types of phenotypes. Most of these methods were based on the idea that individuals with similar genotypes tend to have similar phenotypes. Compared with other similarity-based methods (eg, SKAT), GenRF and GGRF have nice asymptotic properties, and can be applied to small-scale sequencing studies without small-sample adjustment.

Most of the existing methods assume the disease under investigation as one unified phenotype with homogeneous genetic causes. When genetic heterogeneity is present, the existing methods will likely yield attenuated estimates for genetic variants with heterogeneous effects, leading to low testing power. To consider the genetic heterogeneity in sequencing studies, we propose a conditional autoregressive (CAR) model. Different from the previous GGRF model, which applies the conditional autoregressive model on the phenotypes directly, we use a linear mixed model with the genetic effect being considered as a random effect to account for heterogeneous genetic effects. By using a score test for variance components, it has advantage of computational efficiency since we only need to obtain estimators under the null hypothesis. On the other hand, it shares a nice asymptotic feature with GenRF and GGRF, which makes it appealing for small sample size studies. Simulation studies also showed that our proposed method can have high power when rare variants play an important role or when variants have different genetic effects among different individuals or subgroups of individuals. Therefore, CAR provides a powerful alternative approach to search for disease-associated variants, especially those rare or having heterogeneous effects.

The remaining article is arranged as follow: In Section 2, we propose a conditional autoregressive model for genetic association analysis of sequencing data and a score test for statistical inference. In Section 3, we conduct simulation studies to compare the performance of our method with two existing methods (ie, SKAT and GGRF) under different scenarios. Finally, in Section 4, we apply our model to the whole-genome sequencing data from the Alzheimer's Disease Neuroimaging Initiative.

2 | METHODS

2.1 | Motivation

Linear mixed models have been commonly used to assess the association of a set of SNVs with a continuous phenotype. It models the effect of each SNV as a random effect and assumes the genetic effects of all SNVs in the set (eg, a gene) follow an arbitrary distribution. By testing the variance component of the random effect, it evaluates the joint effects of SNVs in the set on the phenotype of interest. One of the most popularly used linear mixed model for sequencing studies is the sequence kernel association test (SKAT).[5] It is a semi-parametric method that uses a kernel function to deal with high-dimensional genetic data and uses a score-based variance component test to assess the association. While SKAT has many advantages, such as being robust for the direction and magnitude of genetic effects, it does not consider the heterogeneous effects of genetic variants among individuals or subgroups of individuals (eg, gender and race groups). If the disease of interest undergoes heterogeneous genetic etiological processes (ie, genetic causes differ among individuals), the traditional linear mixed model (eg, SKAT), which typically assumes the genetic effects are similar across all the samples, can suffer from power loss. To consider the genetic heterogeneity in association analysis, we propose a conditional autoregressive model. It considers the genetic effect of an individual as a random effect and therefore accounts for the genetic heterogeneity among individuals.

All SKAT,[5] genome-wide complex trait analysis (GCTA)[9] and CAR, are based on the following linear mixed model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{a} + \epsilon,$$

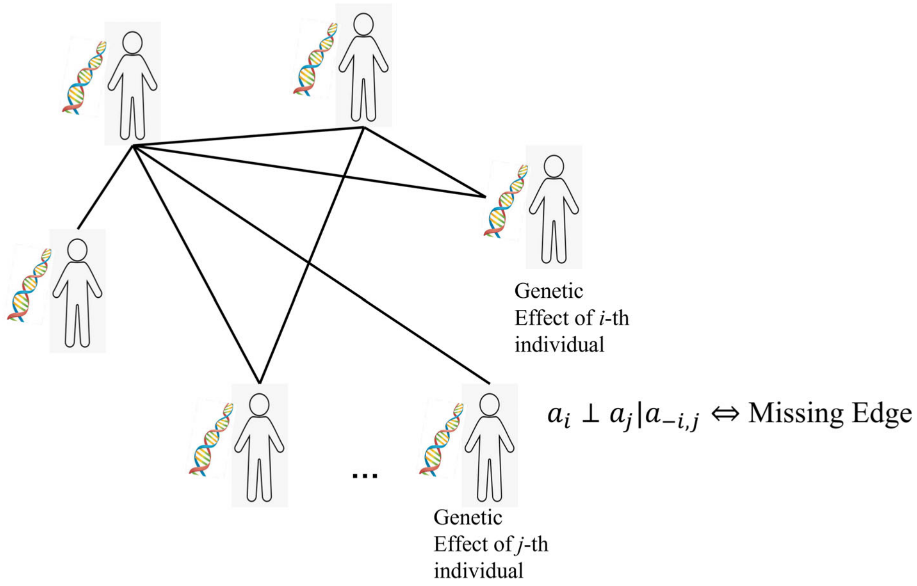

FIGURE 1 An illustration of a Gaussian graphical model. For the CAR model, the graph represents a network of the genetic effects among individuals. In the graph, each node stands for the genetic effect of an individual. The existence of an edge between two individuals indicates dependence of genetic effects between these two individuals. This is characterized by the precision matrix in the joint normal distribution of these genetic effects [Colour figure can be viewed at wileyonlinelibrary.com]

where $\boldsymbol{a}$ is the total genetic random effects of the individuals; $\boldsymbol{X}$ is the design matrix of covariates such as age, gender, and ϵ is the random error. It is natural to assume the total genetic effect $\boldsymbol{a}$ follows a multivariate normal distribution with $\boldsymbol{a} \sim \mathcal{N}_n(\mathbf{0}, \sigma_a^2\boldsymbol{\Sigma})$. A natural question is how to choose the covariance matrix $\boldsymbol{\Sigma}$. In SKAT, $\boldsymbol{\Sigma} = \boldsymbol{GWG}^T$, where $\boldsymbol{G}$ is a matrix of all SNVs and $\boldsymbol{W} = \text{diag}\{w_1, \dots, w_p\}$ is a diagonal matrix containing the weights of the p genetic variants. In GCTA,[9] the (i,j)th element in $\boldsymbol{\Sigma}$ is defined to be

$$\boldsymbol{\Sigma}_{ij} = \frac{1}{p}\sum_{k=1}^{p}\frac{(g_{ik} - 2p_k)(g_{jk} - 2p_k)}{2p_k(1 - p_k)},$$

where p_k is the frequency of the reference allele for the kth SNV. Both methods use a direct way to define the marginal covariance of the genetic effects between two subjects. For CAR model, we consider an indirect way to model the covariance of the genetic effects. Specifically, if we use a graph to represent the connections of the genetic effects among the individuals as shown in Figure 1, where a node in the graph represents a person's genetic effect and an edge represents a link between the genetic effects of two subjects. Due to the Gaussian assumption on the random effect $\boldsymbol{a}$, an edge between nodes i and j is missing if and only if $a_i \perp a_j | \boldsymbol{a}_{-i,j}$, or equivalently $p(a_i|\boldsymbol{a}_{-i}) = p(a_i|\boldsymbol{a}_{-i,j})$, which means that the effects between the ith and the jth individual are independent when the effects of all other individuals are known. So it is worth investigating the conditional distribution of $a_i|\boldsymbol{a}_{-i}$.

A reasonable conditional model can be assumed as follow

$$a_i | a_j, j \neq i \sim \mathcal{N}\left(\sum_{j \neq i} b_{ij} a_j, \tau_i^2\right),$$

By Brook's lemma[10] (details can be found in Appendix A), we can obtain

$$\pi(\boldsymbol{a}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{a}^T\boldsymbol{\Delta}^{-1}(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{a}\right\},$$

where $\boldsymbol{B} = [b_{ij}]$ is a matrix with $b_{ii} = 0$ and $\boldsymbol{\Delta} = \text{diag}\{\tau_1^2, \dots, \tau_n^2\}$. This shows that $\boldsymbol{a} \sim \mathcal{N}_n(\mathbf{0}, (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Delta})$. The first thing we need to ensure is that $\boldsymbol{\Delta}^{-1}(\boldsymbol{I} - \boldsymbol{B})$ is symmetric. A simple sufficient condition for $\boldsymbol{\Delta}^{-1}(\boldsymbol{I} - \boldsymbol{B})$ being symmetric is that $\boldsymbol{\Delta}^{-1}\boldsymbol{B}$ is symmetric, that is,

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2} \text{ for all } i,j.$$

Given a similarity matrix $\boldsymbol{S}$, this can be accomplished by setting $b_{ij} = s_{ij}/\sum_{j\neq i} s_{ij}$ and $\tau_i^2 = \sigma_a^2/\sum_{j\neq i} s_{ij}$ and the CAR model becomes

$$a_i|a_j, j \neq i \sim \mathcal{N}\left(\frac{1}{\sum_{j\neq i} s_{ij}}\sum_{j\neq i} s_{ij}a_j, \frac{\sigma_a^2}{\sum_{j\neq i} s_{ij}}\right). \tag{1}$$

In such case, the joint distribution of $\boldsymbol{a}$ is

$$\pi(\boldsymbol{a}) \propto \exp\left\{-\frac{1}{2\sigma_a^2}\boldsymbol{a}^T(\boldsymbol{D}-\boldsymbol{S})\boldsymbol{a}\right\},$$

where $\boldsymbol{D} = \text{diag}\left\{\sum_{j\neq 1} s_{1j}, \dots, \sum_{j\neq n} s_{nj}\right\}$. However, one issue is that $\boldsymbol{D}-\boldsymbol{S}$ is singular since $(\boldsymbol{D}-\boldsymbol{S})\mathbf{1}_n = \mathbf{0}$. Such impropriety can be remedied by redefining the precision matrix of $\boldsymbol{a}$ as $\boldsymbol{D}-\gamma\boldsymbol{S}$, where γ is chosen to make $\boldsymbol{D}-\gamma\boldsymbol{S}$ nonsingular. In such case, (1) becomes

$$a_i|a_j, j \neq i \sim \mathcal{N}\left(\frac{\gamma}{\sum_{j\neq i} s_{ij}}\sum_{j\neq i} s_{ij}a_j, \frac{\sigma_a^2}{\sum_{j\neq i} s_{ij}}\right), \tag{2}$$

which will be used in the model defined in Section 2.2. In Appendix A, we also showed that $\boldsymbol{D}-\gamma\boldsymbol{S}$ is indeed invertible when $|\gamma| < 1$.

The main difference between the CAR model and the other methods is that the precision matrix of the random effect $\boldsymbol{a}$, which is the inverse of the covariance matrix, is specified first. It is well-known that the (i,j)th element in the precision matrix is in fact the partial covariance between a_i and a_j given all the others. The generalized genetic random fields (GGRF) model[8] has a similar structure. A GGRF model is based on the idea that similar genotypes lead to similar phenotypes, while a CAR model can be interpreted as that similar genotypes leads to similar genetic effects and the variations among phenotypes is related to the variations among genetic effects of individuals.

In Section 2.2, a CAR model will be introduced to account for the genetic heterogeneity. In Section 2.3, a linear score test for testing the variance component will be derived.

2.2 | Model setup

Suppose that there are n subjects. Let y_i be a quantitative trait for the ith subject. To relate the phenotype with genetic variants, we consider the following linear mixed model, also known as the conditional autoregressive (CAR) model:

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + a_i + \epsilon_i, \quad i = 1, \dots, n \tag{3}$$

$$a_i|a_j, j \neq i \sim \mathcal{N}\left(\frac{\gamma}{\sum_{j\neq i} s_{ij}}\sum_{j\neq i} s_{ij}a_j, \frac{\sigma_a^2}{\sum_{j\neq i} s_{ij}}\right) \tag{4}$$

$$\epsilon_1, \dots \epsilon_n \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2), \tag{5}$$

where $\boldsymbol{x}_i = [x_{i1}, \dots, x_{ip}]^T$ is a vector of covariates of the ith subject; $\boldsymbol{\beta}$ is the fixed effect of the covariates; a_i is the genetic random effect, and ϵ_i is the random error of the ith subject; γ measures the overall genetic correlation among all subjects. A larger value of γ implies strong overall genetic correlation among all samples; s_{ij} is the genetic similarity between the ith and the jth subjects, which is measured on a set of SNVs; σ_a^2 measures the variation of the genetic effects.

As we have seen in Section 2.1, the joint distribution of $\boldsymbol{a} = [a_1, \dots, a_n]^T$ can be written as

$$\boldsymbol{a} \sim \mathcal{N}_n(\mathbf{0}, \sigma_a^2(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}),$$

where $\mathcal{N}_n$ denotes an n-dimensional multivariate normal distribution; $\boldsymbol{S}$ is the genetic similarity matrix, and $\boldsymbol{D}$ is a diagonal matrix with diagonal elements being the row sums of $\boldsymbol{S}$.

$$\boldsymbol{S} = \begin{bmatrix} 0 & s_{12} & s_{13} & \cdots & s_{1n} \\ s_{21} & 0 & s_{23} & \cdots & s_{2n} \\ s_{31} & s_{32} & 0 & \cdots & s_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & s_{n3} & \cdots & 0 \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} \sum_{j\neq 1} s_{1j} & & & \\ & \sum_{j\neq 2} s_{2j} & & \\ & & \ddots & \\ & & & \sum_{j\neq n} s_{nj} \end{bmatrix}. \tag{6}$$

Therefore, model (3) can be written into the following matrix form:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{a} + \boldsymbol{\epsilon} \tag{7}$$

$$\boldsymbol{a} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_a^2(\boldsymbol{D} - \gamma \boldsymbol{S})^{-1}) \tag{8}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n\right), \tag{9}$$

where $\boldsymbol{y} = [y_1, \dots, y_n]^T$ is the phenotype and $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]^T$ is the covariate matrix. Since all the genetic information are contained in the total genetic random effects $\boldsymbol{a}$, in order to test the association of a SNV set with the response, it is sufficient to test $H_0 : \sigma_a^2 = 0$.

Let $\boldsymbol{G}_i = [g_{i1}, \dots, g_{iK}]^T$ be the genotypes of K SNVs in a genetic region (eg, a gene) for the ith subject, where $g_{ij}, j = 1, \dots, K$ is coded as additive, that is, $\{0, 1, 2\}$. Given the set of SNVs for subjects i and j, a commonly used kernel function measuring genetic similarity is the weighted identity-by-state (IBS) function,[5] which is defined by

$$s(\boldsymbol{G}_i, \boldsymbol{G}_j) = \sum_{k=1}^{K} \omega_k \{2 - |g_{ik} - g_{jk}|\},$$

where ω_k is the prespecified weight for the kth variant and is usually a function of minor allele frequency (MAF). In our model, we consider the scaled version of the weighted IBS similarity defined by

$$s_{ij} = \sum_{j=1}^{K} \frac{\omega_k \{2 - |g_{ik} - g_{jk}|\}}{2\sum_{k=1}^{K} \omega_k}.$$

For the weights ω_k, we consider four types of weights: unweighted (UW) weights (ie, $\omega_k = 1$), Beta distribution type of weights (BETA), weighted sum statistics type of weights (WSS) and logarithm of MAFs as weights (LOG). The details of these four weights are given in Appendix C. Among these four weights, UW emphasizes the effects of common variants, while WSS focusing on the effects of rare variants. The effects of the other two weights lie in between UW and WSS. LOG gives slightly more weights on common variants than BETA.

To evaluate whether there is a genetic association between a SNV set and the trait of interest, we test $H_0 : \sigma_a^2 = 0$. By introducing the ratio between two variance components $\lambda = \sigma_a^2/\sigma^2$, it is equivalent to test $H_0 : \lambda = 0$ vs $H_1 : \lambda > 0$.

2.3 | Score test for variance component

In order to perform the hypothesis testing, we also need to consider the nuisance parameter γ. In the following discussion, we consider two cases. In the first case, we treat γ as a fixed constant and use the linear score test.[11] In the second case, γ is treated as an unknown nuisance parameter. A maximum statistic[12] is proposed and the corresponding P-value is obtained via a simulation-based method.

2.3.1 | γ as a fixed constant

When γ is fixed as a constant (eg, overall mean of Pearson correlations between SNVs), a linear score test procedure[11] can be used to test $H_0 : \lambda = 0$ vs $H_1 : \lambda > 0$. Based on model (7), we have the marginal distribution of $\boldsymbol{y}$:

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \tilde{\boldsymbol{V}}),$$

where $\tilde{\boldsymbol{V}} = \sigma_a^2(\boldsymbol{D} - \gamma \boldsymbol{S})^{-1} + \sigma^2 \boldsymbol{I}_n = \sigma^2 \left[\boldsymbol{I}_n + \lambda(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\right] := \sigma^2 \boldsymbol{V}(\lambda)$. Let $\boldsymbol{\theta} = [\lambda, \sigma^2, \boldsymbol{\beta}^T]^T$ be the vector of unknown parameters, where σ^2 and $\boldsymbol{\beta}$ are nuisance parameters. The nuisance parameter $\boldsymbol{\beta}$ does not need to be estimated when we use the restricted log-likelihood method. We can find a $q \times n$ matrix $\boldsymbol{K}$ such that $\boldsymbol{KX} = 0$ and $\boldsymbol{KK}^T = \boldsymbol{I}_q$, where $q = n - \text{rank}(\boldsymbol{X})$. Since $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{V}(\lambda))$, we obtain

$$\boldsymbol{y}^* := \boldsymbol{Ky} \sim \mathcal{N}_q(\boldsymbol{0}, \sigma^2\boldsymbol{KV}(\lambda)\boldsymbol{K}^T), \tag{10}$$

where $\boldsymbol{y}^*$ is known as the error contrast since $\mathbb{E}[\boldsymbol{y}^*] = \boldsymbol{0}$, which does not depend on $\boldsymbol{X}$. The restricted log-likelihood function, which provides an unbiased estimator of σ^2, can be formed as

$$\ell(\lambda, \sigma^2|\boldsymbol{y}^*) \propto -\frac{q}{2}\log\sigma^2 - \frac{1}{2}\log|\boldsymbol{KV}(\lambda)\boldsymbol{K}^T| - \frac{1}{2\sigma^2}\boldsymbol{y}^{*^T}(\boldsymbol{KV}(\lambda)\boldsymbol{K}^T)^{-1}\boldsymbol{y}^*. \tag{11}$$

Moreover, by considering the profiled version of the restricted log-likelihood function in (11), we can get the profiled REML of σ^2 for a given λ:

$$\tilde{\sigma}^2(\lambda) = \arg\max_{\sigma^2} \ell(\sigma^2|\boldsymbol{y}^*, \lambda) = \frac{1}{q}\boldsymbol{y}^{*^T}(\boldsymbol{KV}(\lambda)\boldsymbol{K}^T)^{-1}\boldsymbol{y}^*.$$

Back substituting $\tilde{\sigma}^2(\lambda)$ into the restricted log-likelihood function (11), we get the profiled restricted log-likelihood function as follow:

$$\ell_p(\lambda|\boldsymbol{y}^*) \propto -\frac{q}{2}\log[\boldsymbol{y}^{*^T}(\boldsymbol{KV}(\lambda)\boldsymbol{K}^T)^{-1}\boldsymbol{y}^*] - \frac{1}{2}\log|\boldsymbol{KV}(\lambda)\boldsymbol{K}^T|. \tag{12}$$

The profiled restricted log-likelihood function is score and information unbiased and hence can be used for inference.[13] The score statistic $S(\lambda)$ can be calculated as follow:

$$S(\lambda) = \frac{\partial \ell_p}{\partial\lambda} \tag{13}$$

$$= \frac{q}{2}\frac{\boldsymbol{y}^{*^T}(\boldsymbol{KV}(\lambda)\boldsymbol{K}^T)^{-1}\boldsymbol{K}\frac{\partial\boldsymbol{V}(\lambda)}{\partial\lambda}\boldsymbol{K}^T(\boldsymbol{KV}(\lambda)\boldsymbol{K}^T)^{-1}\boldsymbol{y}^*}{\boldsymbol{y}^{*^T}(\boldsymbol{KV}(\lambda)\boldsymbol{K}^T)^{-1}\boldsymbol{y}^*} - \frac{1}{2}\text{tr}\left[(\boldsymbol{KV}(\lambda)\boldsymbol{K}^T)^{-1}\boldsymbol{K}\frac{\partial\boldsymbol{V}(\lambda)}{\partial\lambda}\boldsymbol{K}^T\right] \tag{14}$$

$$= \frac{q}{2}\frac{\boldsymbol{y}^{*^T}(\boldsymbol{KV}(\lambda)\boldsymbol{K}^T)^{-1}\boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T(\boldsymbol{KV}(\lambda)\boldsymbol{K}^T)^{-1}\boldsymbol{y}^*}{\boldsymbol{y}^{*^T}(\boldsymbol{KV}(\lambda)\boldsymbol{K}^T)^{-1}\boldsymbol{y}^*} - \frac{1}{2}\text{tr}\left[(\boldsymbol{KV}(\lambda)\boldsymbol{K}^T)^{-1}\boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\right]. \tag{15}$$

Under the null hypothesis $H_0 : \lambda = 0$, $\boldsymbol{V}(\lambda) = \boldsymbol{V}(0) = \boldsymbol{I}_n$ so that the score statistic under H_0 can be expressed as:

$$S(0) = \frac{q}{2}\frac{\boldsymbol{y}^{*^T}\boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\boldsymbol{y}^*}{\boldsymbol{y}^{*^T}\boldsymbol{y}^*} - \frac{1}{2}\text{tr}\left[\boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\right].$$

To test $H_0 : \lambda = 0$ vs $H_1 : \lambda > 0$, it is equivalent to test $H_0 : S(0) = 0$ vs $H_1 : S(0) \neq 0$. The exact way to calculate the P-value of the test is given by

$$\mathbb{P}(S(0) > s) = \mathbb{P}\left(\sum_{j=1}^{q}\lambda_j Z_j^2 > 0\right), \tag{16}$$

where s is the observed value of $S(0)$, $\lambda_1, \dots, \lambda_q$ are the eigenvalues of the following matrix

$$\boldsymbol{B} = \boldsymbol{K}(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T - \left[\frac{2s}{q} + \frac{1}{q}\operatorname{tr}\left(\boldsymbol{K}(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\right)\right]\boldsymbol{I}_q.$$

$Z_1^2, \dots, Z_q^2$ are independent chi-square-distributed random variables and the P-value can be calculated by using the Davis method.[14] The detailed derivation of (16) can be found in the appendix.

Remark 1. $S(\lambda)$, in fact, does not depend on the choice of $\boldsymbol{K}$. In fact, it can be shown that[15]

$$\boldsymbol{K}^T\left(\boldsymbol{K}\boldsymbol{V}(\lambda)\boldsymbol{K}^T\right)^{-1}\boldsymbol{K} = \boldsymbol{V}(\lambda)^{-1} - \boldsymbol{V}(\lambda)^{-1}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{V}(\lambda)^{-1}\boldsymbol{X}\right)^{-}\boldsymbol{X}^T\boldsymbol{V}(\lambda)^{-1} =: \boldsymbol{P}(\lambda).$$

Then, we can write $S(\lambda)$ as

$$S(\lambda) = \frac{q}{2}\frac{\boldsymbol{y}^T\boldsymbol{P}(\lambda)(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{P}(\lambda)\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{P}(\lambda)\boldsymbol{y}} - \frac{1}{2}\operatorname{tr}\left[(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{P}(\lambda)\right].$$

Under $H_0 : \lambda = 0, \boldsymbol{P}(\lambda) = \boldsymbol{P}(0) = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-}\boldsymbol{X}^T$ and then

$$S(0) = \frac{q}{2}\frac{\boldsymbol{y}^T\boldsymbol{P}(0)(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{P}(0)\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{P}(0)\boldsymbol{y}} - \frac{1}{2}\operatorname{tr}\left[(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{P}(0)\right],$$

and $\boldsymbol{B}$ can be replaced by

$$\tilde{\boldsymbol{B}} = \boldsymbol{P}(0)(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{P}(0) - \left(\frac{2s}{q} + \frac{1}{q}\operatorname{tr}\left[(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{P}(0)\right]\right)\boldsymbol{P}(0).$$

2.3.2 | γ as an unknown nuisance parameter

In practice, γ is usually unknown so that it is necessary to consider how to test $H_0 : \lambda = 0$ vs $H_1 : \lambda > 0$ under such case. One natural idea is to estimate γ and plug in the estimator (eg, MLE). However, since γ is embedded in the variance-covariance matrix of $\boldsymbol{y}$, it is unlikely that the maximum likelihood estimator of γ has an analytic form. Moreover, γ only appears in the alternative model, therefore it is not even possible to evaluate the MLE of γ under H_0. Instead, we construct a maximum score statistic for the association test.[12] More specifically, following the notations we used in Section 2.3.1, we have

$$S_\gamma(0) = \frac{q}{2}\frac{\boldsymbol{y}^{*^T}\boldsymbol{K}(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\boldsymbol{y}^*}{\boldsymbol{y}^{*^T}\boldsymbol{y}^*} - \frac{1}{2}\operatorname{tr}\left[\boldsymbol{K}(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\right],$$

where $\boldsymbol{y}^* \sim \mathcal{N}_q(\boldsymbol{0}, \sigma^2\boldsymbol{K}\boldsymbol{V}(\lambda)\boldsymbol{K}^T)$ with $\boldsymbol{V}(\lambda) = \boldsymbol{I}_n + \lambda(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}$. Now, note that under the null hypothesis $H_0 : \lambda = 0$, we have $\boldsymbol{y}^* \sim \mathcal{N}_q(\boldsymbol{0}, \sigma^2\boldsymbol{I}_q)$ so that $\frac{1}{\sigma}\boldsymbol{y}^* \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{I}_q)$. We form the maximum score test statistic as follows:

$$T = \sup_{\gamma\in\Gamma} S_\gamma(0) \tag{17}$$

$$= \sup_{\gamma\in\Gamma}\frac{q\boldsymbol{y}^{*^T}\boldsymbol{K}(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\boldsymbol{y}^* - \boldsymbol{y}^{*^T}\operatorname{tr}\left[\boldsymbol{K}(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\right]\boldsymbol{I}_q\boldsymbol{y}^*}{2\boldsymbol{y}^{*^T}\boldsymbol{y}^*} \tag{18}$$

$$= \frac{1}{2\left(\frac{1}{\sigma}\boldsymbol{y}^*\right)^T\left(\frac{1}{\sigma}\boldsymbol{y}^*\right)}\sup_{\gamma\in\Gamma}\left(\frac{1}{\sigma}\boldsymbol{y}^*\right)^T\boldsymbol{\Sigma}(\gamma)\left(\frac{1}{\sigma}\boldsymbol{y}^*\right) \tag{19}$$

$$:= (2\|\boldsymbol{Z}\|_2^2)^{-1}\sup_{\gamma\in\Gamma}\boldsymbol{Z}^T\boldsymbol{\Sigma}(\gamma)\boldsymbol{Z}, \tag{20}$$

where $\boldsymbol{Z} = \frac{1}{\sigma}\boldsymbol{y}^* \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{I}_q)$, $\boldsymbol{\Sigma}(\gamma) = q\boldsymbol{K}(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T - \mathrm{tr}\left[\boldsymbol{K}(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\right]\boldsymbol{I}_q$. $\gamma \in \left(1/\mu_{(1)}, 1/\mu_{(n)}\right)$, where $\mu_{(1)} < \mu_{(2)} < \cdots < \mu_{(n)}$ are the ordered eigenvalues of $\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}$ ensuring that the matrix $\boldsymbol{D} - \gamma\boldsymbol{S}$ is nonsingular.[16] Since the parameter γ can be interpreted as a measurement of correlation coefficient, we set $\Gamma = \left(1/\mu_{(1)}, 1/\mu_{(n)}\right) \cap (-1, 1)$. To calculate the P-value, let t be the observed value of T, then

$$\mathbb{P}(T > t) = \mathbb{P}\left((2\|\boldsymbol{Z}\|_2^2)^{-1}\sup_{\gamma\in\Gamma}\boldsymbol{Z}^T\boldsymbol{\Sigma}(\gamma)\boldsymbol{Z} > t\right) \tag{21}$$

$$= \mathbb{P}\left(\sup_{\gamma\in\Gamma}\boldsymbol{Z}^T\left(\boldsymbol{\Sigma}(\gamma) - 2t\boldsymbol{I}_q\right)\boldsymbol{Z} > 0\right) \tag{22}$$

$$= \mathbb{P}\left(\sup_{\gamma\in\Gamma}\sum_{i=1}^{q}\lambda_j(\gamma)Z_j^2 > 0\right), \tag{23}$$

where $\lambda_1(\gamma), \dots, \lambda_q(\gamma)$ are the eigenvalues of

$$\frac{1}{q}(\boldsymbol{\Sigma}(\gamma) - 2t\boldsymbol{I}_q) = \boldsymbol{K}(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T - \left(\frac{2t}{q} + \frac{1}{q}\mathrm{tr}\left[\boldsymbol{K}(\boldsymbol{D} - \gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\right]\right)\boldsymbol{I}_q$$

and $Z_1^2, \dots, Z_q^2$ are independent chi-square-distributed random variables with degrees of freedom 1. We use the following procedures to approximate this probability:

1. Partition the index set Γ as $\gamma_1 < \gamma_2 < \cdots < \gamma_M$. Γ is an open interval, denoted by (L, U). When we implement this step, we choose a small $\epsilon > 0$ and let $\gamma_1 = L + \epsilon$ and $\gamma_M = U - \epsilon$;
2. Generate an $N \times q$ matrix $\boldsymbol{\Upsilon}$ with each element being a χ_1^2 random variable;
3. For each γ_i, $i = 1, \dots, M$,
   Calculate $\lambda_1(\gamma_i), \dots, \lambda_q(\gamma_i)$, which are the eigenvalues of $\boldsymbol{\Sigma}(\gamma_i) - 2t\boldsymbol{I}_q$;
   Calculate $\sum_{j=1}^{q}\lambda_j(\gamma_i)\boldsymbol{\Upsilon}_{kj}$, $k = 1, \dots, N$, where $\boldsymbol{\Upsilon}_{kj}$ is the (k,j)th element of $\boldsymbol{\Upsilon}$;
4. Approximate the P-value by

$$\mathbb{P}\left(\sup_{\gamma\in\Gamma}\sum_{j=1}^{q}\lambda_j(\gamma)Z_j^2 > 0\right) \approx \frac{\#\left\{k : \max_{\gamma_i, i=1,\dots,M}\sum_{j=1}^{q}\lambda_j(\gamma_i)\boldsymbol{\Upsilon}_{kj} > 0\right\}}{N}.$$

3 | SIMULATION RESULTS

Simulation studies were conducted to compare the CAR model with the original GGRF model and the commonly used SKAT method. For the CAR model, we evaluated both scenarios where γ is fixed (denoted by CAR.FIX) and γ is treated as an unknown nuisance parameter (denoted by CAR.SUP). When γ is chosen to be a fixed constant, we set γ to be the overall mean of Pearson correlations between SNVs. In this case, the nonsingularity of $\boldsymbol{D} - \gamma\boldsymbol{S}$ is guaranteed by Proposition 2 in the Appendix. In the simulation, we compared the empirical type I error rates and empirical power of the three methods under various disease models with heterogeneous genetic effects. In addition, we compared the empirical power of the three methods under different percentage of causal SNVs and under the situation when the weights were misspecified. In order to mimic the real structure of sequencing data, the genetic data used for simulation were simulated based on the real sequencing data of Chromosome 17: 7344328-8344327 from the 1000 Genomes Project.[17] The minor allele frequencies (MAF) of SNVs in this region range from 0.046% to 49.954% with a distribution highly skewed to the right. Appendix C summarizes the distribution of the MAF with MAF<0.05. For each setting, we simulated 1000 Monte Carlo replicates to calculate the empirical type I error rates and the empirical power of the three methods. In each replicate, we randomly selected a 30 Kb segment from the region and used all the genetic variants in that segment for the association analysis.

We first examined the type I error performances of the three methods. In the simulation, we considered two significance levels, $\alpha = 0.05$ and $\alpha = 0.01$, under the following null model,

$$y_i = \varepsilon_i \sim \mathcal{N}(0, 1). \tag{24}$$

TABLE 1 Empirical type I error rates under different weights at level $\alpha = 0.05$ and $\alpha = 0.01$ based on 1000 replicates

	$\alpha = 0.05$				$\alpha = 0.01$			
Methods	**UW**	**BETA**	**WSS**	**LOG**	**UW**	**BETA**	**WSS**	**LOG**
GGRF	0.053	0.062	0.055	0.058	0.015	0.010	0.010	0.010
SKAT	0.053	0.069	0.052	0.044	0.008	0.012	0.004	0.012
CAR.FIX	0.050	0.070	0.044	0.053	0.010	0.009	0.012	0.011
CAR.SUP	0.044	0.067	0.044	0.045	0.009	0.007	0.012	0.009

Note: Each cell in the table contains the empirical type I error rate. GGRF, SKAT, CAR.FIX, and CAR.SUP are the generalized genetic random field model,[8] the sequence kernel association test,[5] the conditional autoregressive model with fixed nuisance parameter γ, the conditional autoregressive model with maximum score test statistic, respectively.

TABLE 2 Computation time of one iteration under the null model (24) and BETA weights for GGRF, SKAT, CAR.FIX, and CAR.SUP

	GGRF	**SKAT**	**CAR.FIX**	**CAR.SUP**
Computation time (s)	0.76	0.14	0.91	38.97

Table 1 summarizes the empirical type I error rates of the three methods under four different weights (ie, UW, BETA, WSS, and LOG). As we observe from the results, the empirical type I error rates of GGRF, CAR.FIX, and CAR.SUP are well controlled at the level of 0.05 or 0.01, while the type I error rate for SKAT is conservative under WSS (for $\alpha = 0.05$) and LOG (for $\alpha = 0.01$). Because the score test of CAR and the test procedure of GGRF are exact test procedures without any asymptotic approximations, the type I error rates of the two methods are well controlled.

The computation time of one iteration under the null model (24) and BETA weights for GGRF, SKAT, CAR.FIX, and CAR.SUP is summarized in Table 2. As we can see, the computation time for GGRF, SKAT, and CAR.FIX are comparable. CAR.SUP required more time for computation due to the grid search. We have already seen from Table 1 that CAR.FIX and CAR.SUP have comparative performance in terms of empirical type I error. It will be seen below that they also have similar performance in power. Therefore, we recommend using CAR.FIX for large data sets.

3.1 | Simulation I: Heterogeneous genetic effects among individuals or subgroups

We first considered a complex disease scenario in which each individual has a different genetic effect, which we refer as the heterogeneous genetic effect. The following model was used to simulate the phenotype:

$$y_i = \sum_{k=1}^{K} w_k^* g_{i,k} Z_{i,k} + \varepsilon_i, \quad 1 \leq i \leq n, \tag{25}$$

where K is the number of genetic variants in a 30 Kb segment; $g_{i,k}$ is the genotype of the kth SNV for individual i, coded as additive (ie, $g_{i,k} = 0$ for genotype AA, $g_{i,k} = 1$ for genotype Aa and $g_{i,k} = 2$ for genotype aa). We set the percentage of causal SNVs as 50%. $w_k^* = 0$ if the kth genetic variant is not a causal variant and $w_k^* = \omega_k$ if the kth genetic variant is a causal variant, where ω_k is the weight defined in section 2. $Z_{i,1}, \dots, Z_{i,K}$, $1 \leq i \leq n$ are genetic effects for the ith individual, which follow a normal distribution with mean 0 and standard deviation (SD) σ_Z. $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. random variables distributed as $\mathcal{N}(0, 1)$. Note that in (25), the coefficients of the genetic variants are unique for each individual, and therefore different individuals could have different genetic effects.

When we conducted the simulations, the underlying weights were applied to all of the methods (GGRF, SKAT, CAR.FIX, and CAR.SUP). For GGRF and SKAT method, the kernels were also specified as the weighted IBS kernel, which is the same as the one used in CAR model.

Figure 2 summarizes the empirical power of three methods under different degrees of genetic heterogeneity. As we observe from the simulation results, the CAR model outperforms the other two methods with the increasing level of genetic heterogeneity (ie, increasing σ_Z). For instance, when common variants (ie, UW and LOG) have more contributions to the phenotype, our model has substantial power increase with the increased level of genetic heterogeneity (ie, σ_Z), while

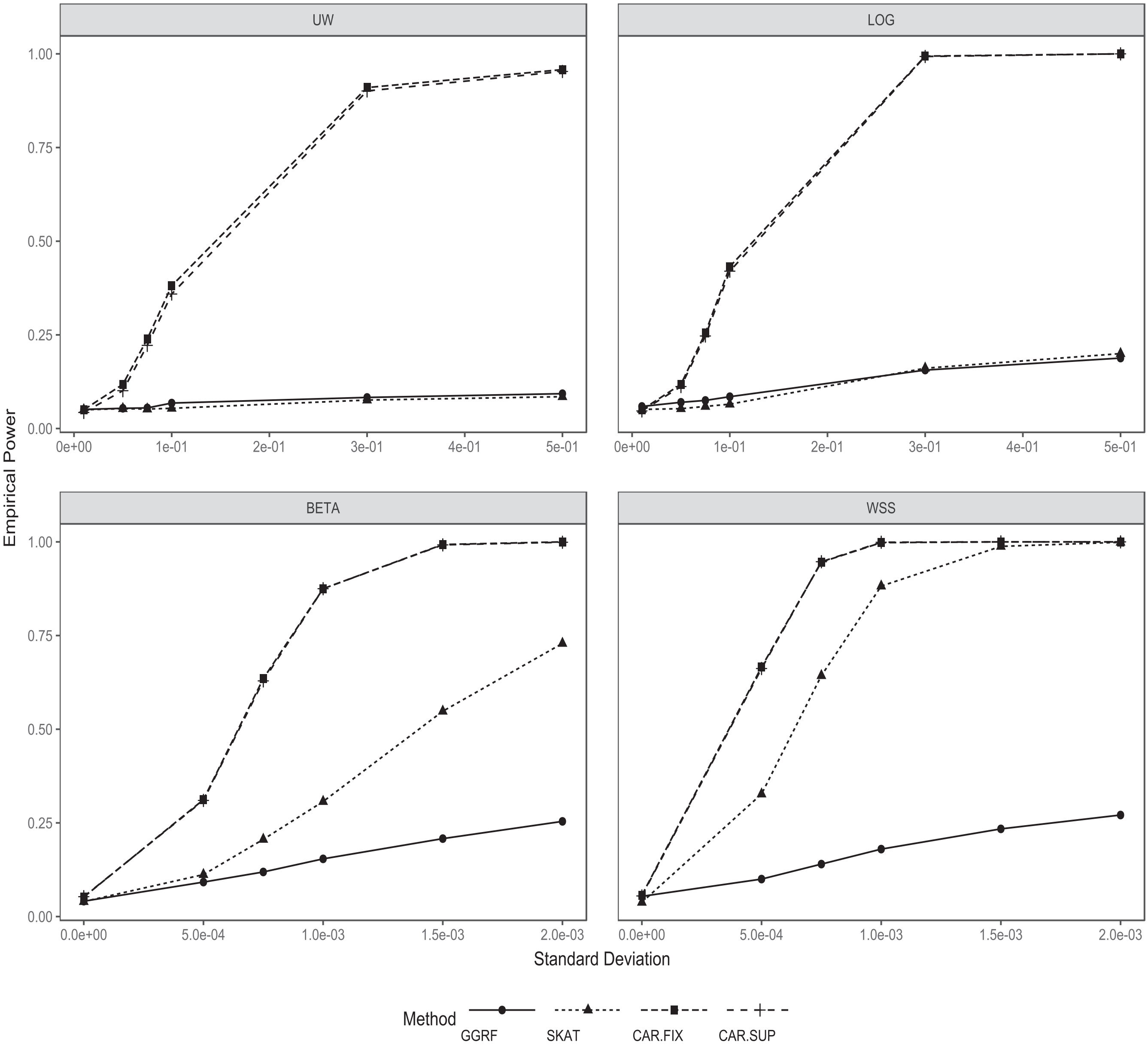

FIGURE 2 Empirical power comparisons of CAR, SKAT, and GGRF by varying the levels of genetic heterogeneity among individuals under four different weights. The x-axis is the effect size σ_Z, used in the simulation. The effect sizes are chosen as 0.01, 0.05, 0.075, 0.1, 0.3, 0.5 for the UW and LOG weights, and the effect sizes are set as 0.0005, 0.00075, 0.001, 0.0015, 0.002 for the BETA and WSS weights

the power of SKAT and GGRF remain low even with the increased genetic effect. When rare variants play an important role (ie, BETA and WSS), GGRF and SKAT have some power increase with the increased genetic effect, but still have lower power than CAR. As also shown in Figure 2, CAR.FIX and CAR.SUP have very similar performance. Therefore, for simplicity, the further analyses were performed based on CAR.FIX, in which we fix the nuisance parameter γ as the overall mean of the correlation matrix of SNVs.

Next, we considered a scenario where genetic heterogeneity exist among subgroups (eg, ethnic groups). Under such case, individuals within a group had homogeneous genetic effects, but the genetic effects among groups were different. In this simulation, we partitioned the total number of individuals of 500 into 8 groups, with the number of individuals in each groups being 200, 100, 50, 25, 40, 35, 45, and 5, respectively. Within each group, we used model (D.1) in Appendix D to simulate the phenotypes, so that the effect sizes were different for different groups. Specifically, we set the effect sizes σ_Z as 0.0001, 0.0005, 0.002, 0.001, 0.003, 0.0001, 0.0005, and 0.002 for the eight groups for the BETA and WSS weights, and 0.01, 0.05, 0.2, 0.1, 0.3, 0.01, 0.05, and 0.2 for the UW and LOG weights.

TABLE 3 Empirical power comparisons of GGRF,[8] SKAT,[5] and CAR based on 1000 Monte Carlo replicates

Methods	UW	BETA	WSS	LOG
GGRF	0.298 (1.505E-02)	0.180 (1.243E-02)	0.189 (1.226E-02)	0.148 (1.140E-02)
SKAT	0.328 (1.497E-02)	0.454 (1.622E-02)	0.805 (1.265E-02)	0.210 (1.287E-02)
CAR	0.378 (1.548E-02)	0.817 (1.248E-02)	0.952 (0.670E-02)	0.445 (1.611E-02)

Note: In the simulation, we simulated eight different subgroups with $\sigma_Z = 0.0001, 0.0005, 0.002, 0.001, 0.003, 0.0001, 0.0005, 0.002$ for BETA and WSS and $\sigma_Z = 0.01, 0.05, 0.2, 0.1, 0.3, 0.01, 0.05, 0.2$ for UW and LOG. The number in the parenthesis is the SD.

TABLE 4 Empirical power comparisons of GGRF,[8] SKAT,[5] and CAR based on 1000 Monte Carlo replicates

Methods	UW	BETA	WSS	LOG
GGRF	0.201 (1.263E-02)	0.149 (1.093E-02)	0.154 (1.161E-02)	0.115 (1.024E-02)
SKAT	0.217 (1.288E-02)	0.374 (1.534E-02)	0.618 (1.545E-02)	0.149 (1.096E-02)
CAR	0.382 (1.576E-02)	0.672 (1.472E-02)	0.769 (1.342E-02)	0.428 (1.544E-02)

Note: In the simulation, we simulated 8 different subgroups with $\sigma_Z = 0.0001, 0.0001, 0.0002, 0.005, 0.0003, 0.0005, 0.0001, 0.005$ for BETA and WSS and $\sigma_Z = 0.01, 0.01, 0.02, 0.5, 0.03, 0.05, 0.01, 0.5$ for UW and LOG. The number in the parenthesis is the SD.

Table 3 summarizes the results of the simulation. Consistent with previous findings, CAR has higher power than the other two methods for all four scenarios. SKAT has good power under the WSS weight, but has reduced power under the other weights. While all three methods have low power when common variants play an important role (ie, UW and LOG), CAR still performs better than SKAT and GGRF.

Since rare mutation occurs in a small number of individuals and could have a larger effect than common variants, we also modified the effect sizes of the eight groups so that groups with a small number of individuals tended to have higher effect sizes. Specifically, we set σ_Z as 0.0001, 0.0001, 0.0002, 0.005, 0.0003, 0.0005, 0.0001, and 0.005 for the eight groups for the BETA and WSS weights, and 0.01, 0.01, 0.02, 0.5, 0.03, 0.05, 0.01, and 0.5 for the UW and LOG weights. Under such setting, the groups with the sizes of 25 and 5 had the highest effect sizes.

From Table 4, we find that all three methods have some power losses compared to the previous cases, but the conclusion is similar to the previous cases. CAR still outperforms the other two under all four scenarios. SKAT performs well when WSS is used, while GGRF has a good performance under the UW weight.

3.2 | Simulation II: Various causal SNV rates

Since the percentage of causal SNVs in a genetic region (eg, a gene) also have impact on the power of the association test, in simulation II, we examined the performances of GGRF, SKAT, and CAR by varying the percentage of causal SNVs. Similar as simulation I, the phenotypes were simulated by using the simulation model (25). In this simulation, we varied the percentages of causal SNVs and investigated the effect of 50%, 40%, 30%, 20%, 10%, 5%, and 1% causal SNV rates on the power performance of the three methods.

Figures 3 and 4 illustrate the empirical power of GGRF, SKAT, and CAR under different causal SNV rates and four weights. Figure 3 summarizes the results under the WSS and BETA weights, that is, the weights focusing more on rare variants. Overall, the CAR model outperforms the other two methods. As the causal SNV rates increase, the empirical power of CAR increases significantly. The similar trend can also be found for SKAT, especially when the WSS weight are used. For the BETA weight, SKAT also attains high empirical power but not as high as that of CAR. GGRF has lower empirical power as compared to CAR and SKAT and its empirical power increases slowly with the increase of causal SNV rates. We also find that under the WSS weight, both SKAT and CAR attain decent performance when the genetic causes are heterogeneous (ie, increasing σ_Z) and the causal SNV rates are moderately high.

Figure 4 summarizes the results under the LOG and UW weights, that is, the weights focusing more on the effects of common variants. Same as the case of rare variants, the CAR model performs the best for all the scenarios. Even the empirical power of CAR under the LOG and UW weights is not high as those under the BETA and WSS weights, it can still reach high power when the causal SNV rate is high. The CAR model could also gain more substantial power than the

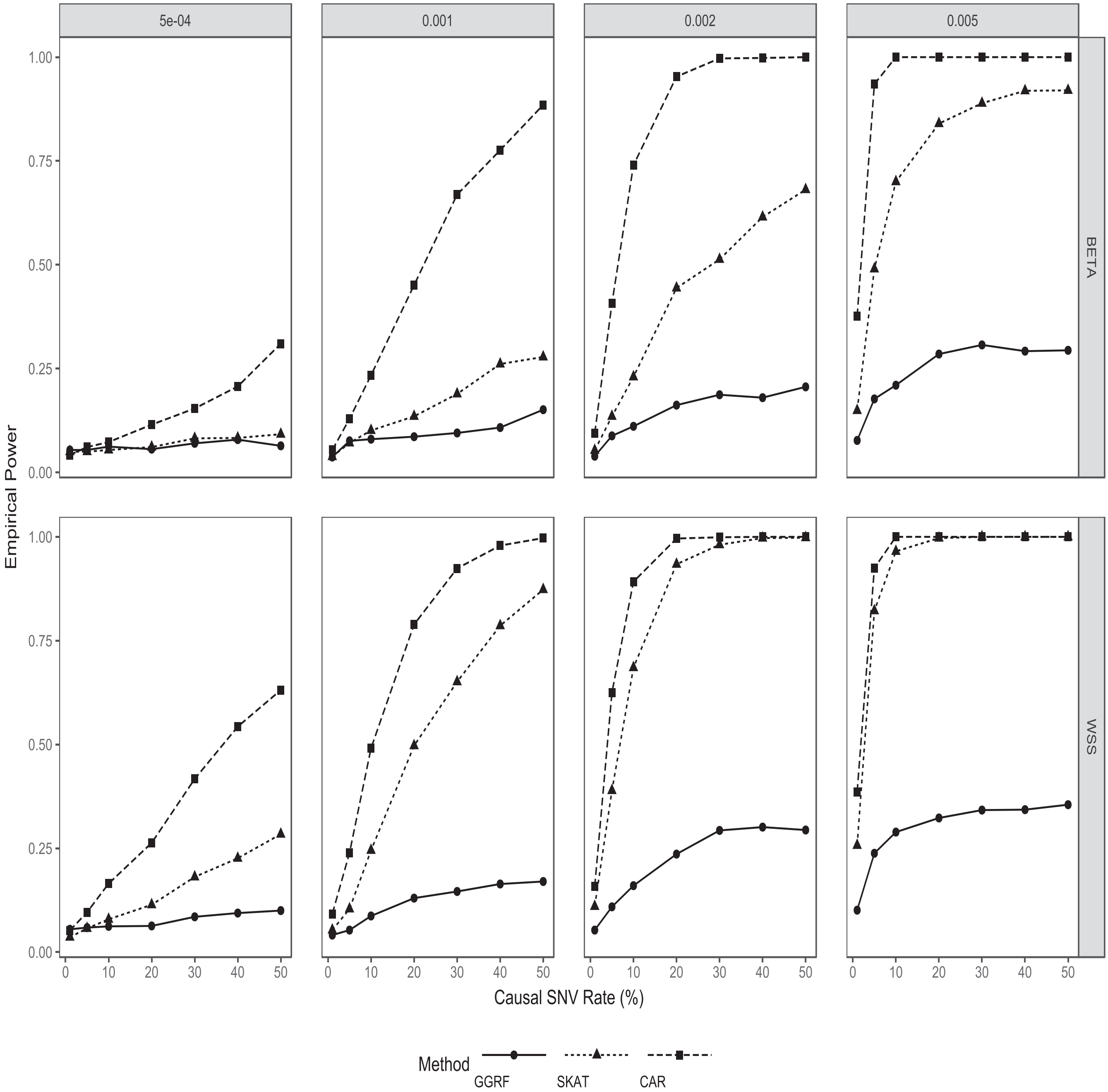

FIGURE 3 Empirical power comparisons of CAR, SKAT, and GGRF with different causal SNV rates under the BETA weight and the WSS weight. The SD σ_Z used in the simulation is gradually increased. $\sigma_Z = 0.0005, 0.001, 0.002, 0.005$ in each column from left to right

other two methods with the increased levels of genetic heterogeneity (ie, increased σ_Z). Neither SKAT nor GGRF performs as well as they do under the WSS or BETA weights, even with high causal SNV rates and high genetic effects. As we can see from Figure 4, if the common variants play an important role in disease phenotypes and have heterogeneous effects, both SKAT and GGRF suffer from extreme power loss, while the CAR model could still obtain moderate or high power.

3.3 | Simulation III: Misspecification of weights

In practice, we do not know whether common variants or rare variants play a more important role in the disease process. Since the performance of GGRF, SKAT, and CAR depend on the prespecified weights, it is important to investigate whether

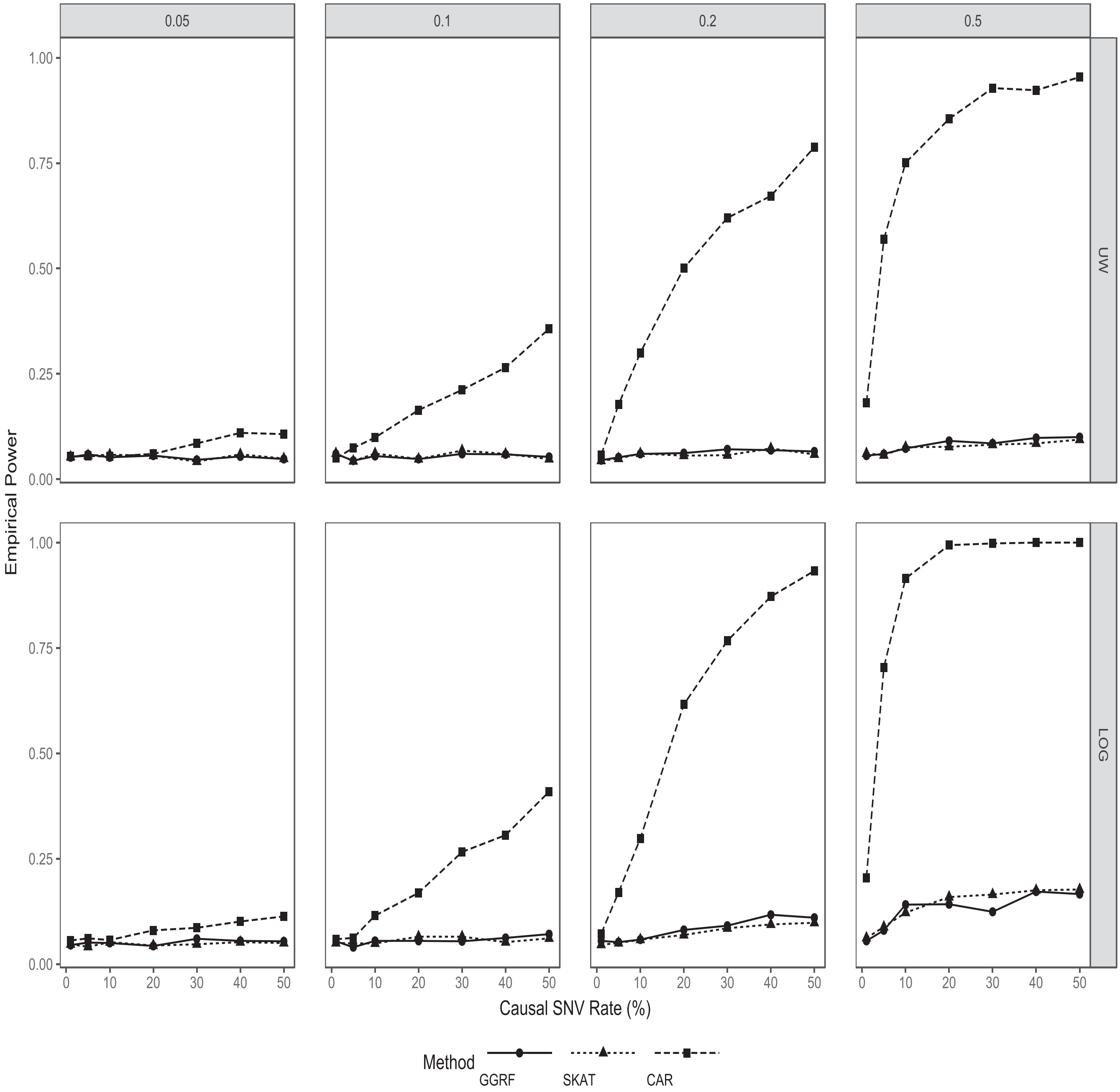

FIGURE 4 Empirical power comparisons of CAR, SKAT, and GGRF with different causal SNV rates under the UW weight and the LOG weight. The SD σ_Z used in the simulation is gradually increased. $\sigma_Z = 0.05, 0.1, 0.2, 0.5$ in each column from left to right

these models could still obtain reasonable power when the weights are misspecified. We used similar simulation settings as before except that we randomly selected 50 SNVs as the causal variants in this particular simulation.

We first focused on rare variants, and simulated phenotypes using either the WSS weight or the BETA weight. When we applied GGRF, SKAT, and CAR to the simulation data, all the weights (UW, BETA, WSS, and LOG) were used so that we were able to evaluate their performances under both misspecified and correctly specified weights.

Figure 5 summarizes the results when the underlying weights in the simulation are WSS and BETA. From Figure 5, since WSS put extremely high weights to the rare variants, all the three methods attain high empirical power when the weights are specified as WSS or BETA. As expected, all the methods have the highest empirical power when the weight function is correctly specified (ie, WSS). We also find that neither SKAT nor GGRF performs well when the weight is misspecified as UW or LOG. On the other hand, as we can see from the figures, as long as there is genetic heterogeneity, CAR has a good power performance even though the weight is misspecified.

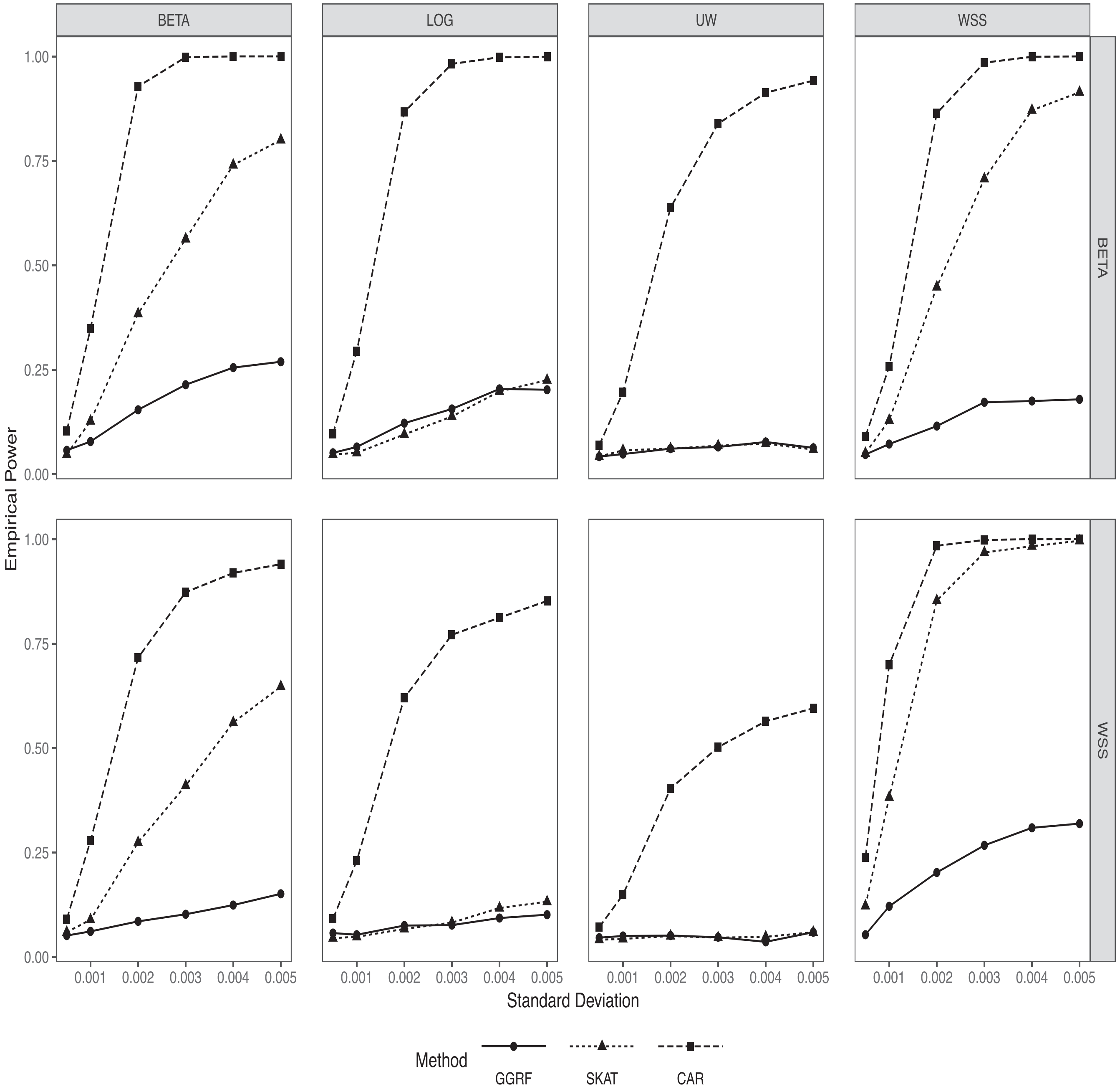

FIGURE 5 Empirical power comparisons of CAR, SKAT, and GGRF with misspecified weights when the true weight is BETA and WSS. In each column from left to right, we used BETA, LOG, UW, and WSS weight, respectively

Same conclusions hold for BETA. Both GGRF and CAR will obtain highest power when the BETA weight is applied, while SKAT reaches highest power when WSS is used. CAR outperforms SKAT and GGRF in all cases. With the increased genetic heterogeneity, CAR retains high power even with misspecified weights. To conclude, in the case when the disease is mainly caused by rare variants, both SKAT and CAR can have high power, but for SKAT, we need to use the right weights (BETA or WSS).

Next, we focused on common variants and simulated phenotypes using UW and LOG. Figure 6 summarizes the results based on LOG and UW. In the first row, the true weight used in the simulation is LOG. Since LOG also puts some weight on rare variants, we find that SKAT has a good performance as compared with its performance under the UW weight. However, SKAT attains highest power by using the WSS weight. This can be explained by the fact that WSS can help to catch the heterogeneous genetic effect. Overall, GGRF has lower power than CAR and SKAT.

The same conclusion holds for the UW weight. As expected, CAR attains the highest power when the true weight (ie, LOG) is specified. When LOG is the underlying weight, we can see from Figure 6 that there is no significant difference of the four weights used in the CAR model. As expected, CAR attains high power when the specified weight functions focus

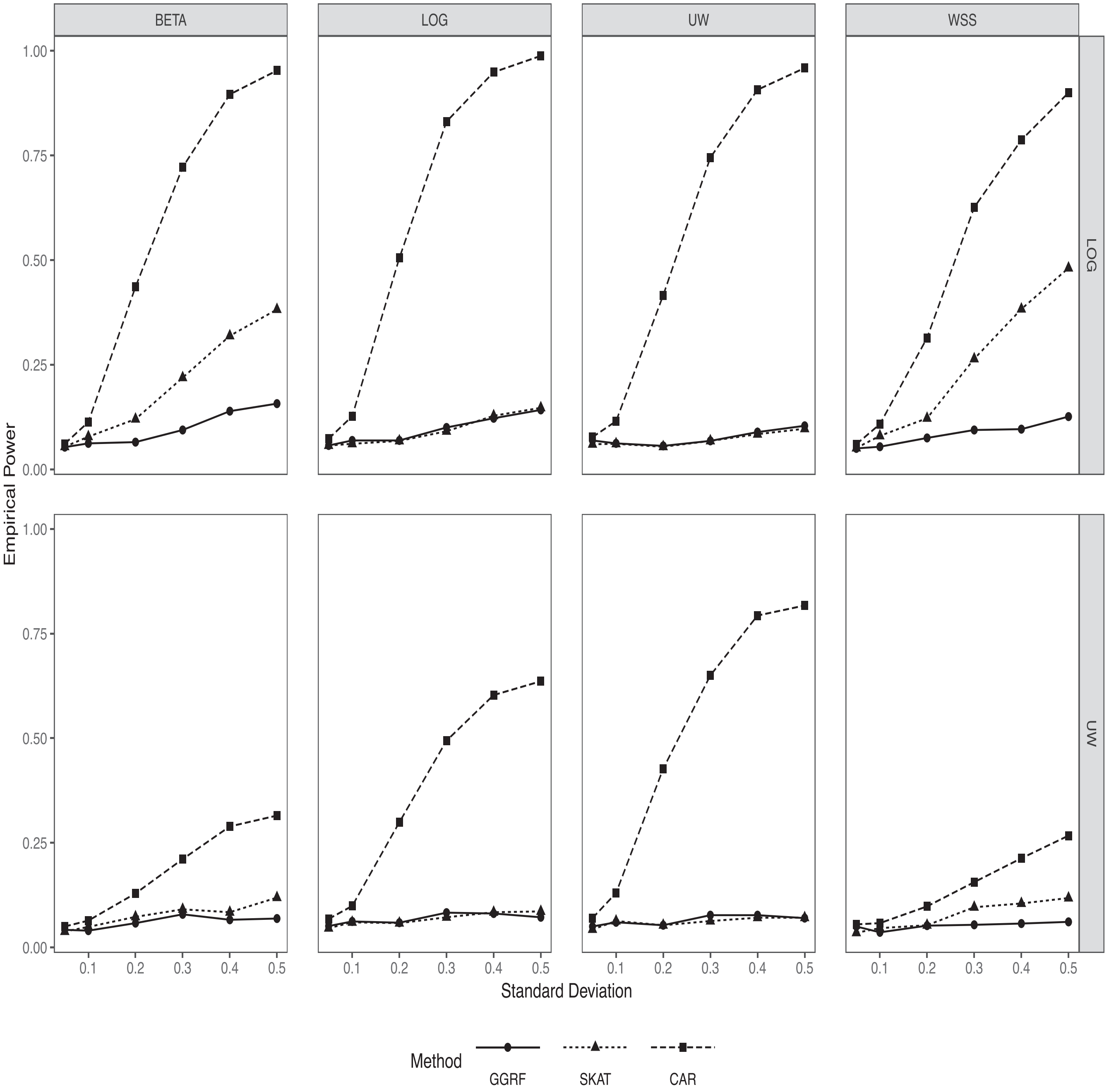

FIGURE 6 Empirical power comparisons of CAR, SKAT, and GGRF with misspecified weights when the true weight is LOG and UW. In each column from left to right, we used BETA, LOG, UW, and WSS weight, respectively

on common variants (ie, UW and LOG). With the increased heterogeneity, CAR can obtain high power with the LOG or UW weights. However, it could suffer from power loss when the weight is misspecified (ie, WSS and BETA). Overall, CAR outperforms SKAT and GGRF, even when the weight is misspecified.

4 | REAL DATA APPLICATIONS

We applied our method to the whole genome sequencing data from Alzheimer's Disease Neuroimaging Initiative (ADNI) and performed a genome-wide gene-based association analysis. A total of 808 samples at the screening and baseline of the ADNI1 and ADNI2 studies have the whole genome sequencing data, from which we extracted 21 069 genes based on the GRCh 37 assembly. ADNI also provides pre-calculated volumes of cortical regions. We chose four of them as the phenotypes of interests, which are hippocampus, entorhinal, whole brain, and ventricles. The motivation of choosing these four volumes as the phenotypes is from previous biological findings. More specifically, the hippocampus, a brain area playing an important role in learning and memory, is especially vulnerable to damage at early stages of Alzheimer's disease (AD).[18] The volume of hippocampus changes over time and could have a large impact on AD.[19] The entorhinal cortex is also crucial in declarative memories. In mild AD patients, the loss in the entorhinal volume is evident. The entorhinal cortex is also highly correlated with the severity of the disease.[20] Similarly, the whole brain volume decreases significantly in patients with AD.[21] The brain ventricles also play an important role in AD and it is well known that ventricular volume is significantly higher in AD patients.[22]

Figure 7 plots the histograms of the four phenotypes used in the analysis. As we can see from the histograms, hippocampus, entorhinal, and whole brain are nearly normally distributed. For ventricles, we first applied a normal quantile transformation to the phenotype and then applied our method.

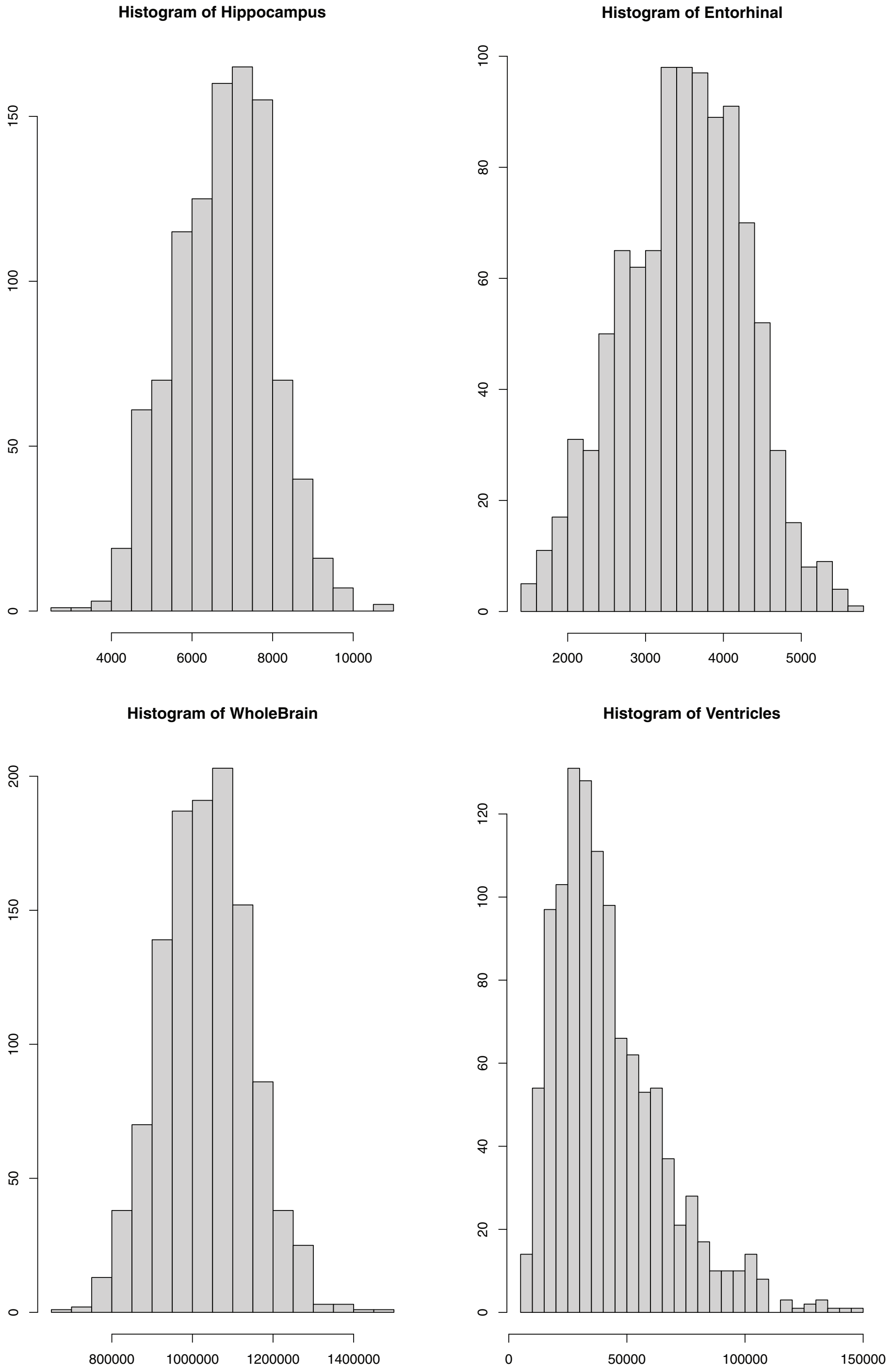

FIGURE 7 Histograms of the phenotypes used for real data analyses

TABLE 5 Top 10 hippocampus-associated genes and their corresponding *P*-values from different methods

CHR	Gene name	CAR	SKAT	GGRF
7	*STK17A*	2.09E-05	6.34E-01	3.79E-02
3	*PEX5L*	2.95E-05	3.02E-02	7.52E-01
11	*LOC105369443*	3.79E-05	9.78E-02	1.06E-01
5	*KIF3A*	1.10E-02	1.65E-04	9.42E-05
5	*IL4*	5.83E-01	1.07E-04	1.22E-04
19	*SYT5*	9.36E-02	2.38E-04	6.25E-04
5	*CANX*	4.31E-01	1.10E-03	2.68E-04
12	*FBRSL1,MIR6763*	2.01E-02	2.44E-04	3.05E-04
20	*LOC100506384*	3.46E-04	8.47E-02	1.89E-01
10	*FRMPD2B*	3.49E-04	3.11E-02	9.61E-03

We restricted our analyses to individuals with Caucasian ancestry due to the small sample size of non-Caucasian samples and the issue of population stratification. In the analysis, age, gender, years of education, marriage status, and *APoEε4* were used as covariates. After removing individuals with missing phenotypes or covariates, a total of 588 subjects remained for the analysis. In this analysis, we also compared our method with SKAT and GGRF. The BETA weight was chosen for all the three methods, and the linear kernel was used in SKAT.

Table 5 summarizes the top 10 genes (based on the *P*-values) associated with hippocampus found by the three methods (CAR, GGRF, and SKAT). We rank the *P*-values for all three methods and summarize the 10 genes with the smallest *P*-values along with their corresponding *P*-values for different methods. We found that the results of GGRF and SKAT are similar and are different from that of CAR. This could be due to the fact that both methods are similar in terms of assuming genetic homogeneity, while CAR is developed based on a different model, which accounts for genetic heterogeneity. In addition to hippocampus, Appendix E also provides results of the top 10 genes related to entorhinal, ventricle and whole brain. Given the limited sample size of the study, none of the association reached statistical significance after multiple-testing adjustment.

5 | DISCUSSION

We have proposed a conditional autoregressive model for genetic association analysis of sequencing data. Our simulations show that the CAR model can obtain high power under the scenario (i) when rare variants are related to the phenotypes and (ii) when genetic variants have different genetic effects among individuals or subgroups of individuals. Moreover, we derive the exact form of the test statistic, which makes the method computationally efficient for large-scale sequencing data analysis. Unlike SKAT, which uses the asymptotic distribution for its test statistic, the exact distribution of the CAR test statistic under the null hypothesis is not conservative. Therefore, no additional small sample size adjustment is needed for the CAR model.

In our simulations, γ is chosen as the average of the genetic correlation coefficients among all individuals. Alternatively, we could also use the average of pairwise linkage disequilibrium (LD) correlation coefficient. Simulation results find that there are no substantial differences between two values (results not shown). However, calculating LD correlation coefficient is more time consuming than estimating the traditional correlation coefficient. It is also interesting to consider the model when $\gamma = 1$, that is, assuming the genetic random effects are highly correlated. When $\gamma = 1$, the genetic random effect is then modeled by an intrinsic random field.[23] In this case, the joint distribution of the genetic random effects is not a proper distribution so that traditional frequentist approach may not work well. One potential solution is to use Bayesian method.[23]

The work introduced in this article focuses on the continuous phenotype with normal distribution. Nevertheless, the model can be extended to the case when the phenotypes follow distributions in the exponential family by using the generalized linear mixed model (GLMM). One potential challenge of extending the CAR model to GLMM is that the test statistic and the likelihood function may not have closed form, which requires an alternative approach (eg, Monte Carlo method) to numerically estimate both of them. This is a future work worth further study.

6 | SOFTWARE

The R code for the proposed CAR model, together with a sample input data set, are available on the website https://msu.edu/~qlu/Software.html.

ACKNOWLEDGEMENTS

The project is supported by the National Institute on Drug Abuse (Award No. K01DA033346), the Faculty Research Development Funds from the University of Auckland, and the National Natural Science Foundation of China. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

AUTHOR CONTRIBUTIONS

Xiaoxi Shen: Conceptualization, Formal Analysis, Writing-Original Draft. **Yalu Wen**: Validation, Writing-Review & Editing. **Yuehua Cui**: Supervision, Writing-Review & Editing. **Qing Lu**: Supervision, Funding Acquisition, Writing-Review & Editing.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Yalu Wen https://orcid.org/0000-0002-0071-5917
Yuehua Cui https://orcid.org/0000-0001-8099-1753
Qing Lu https://orcid.org/0000-0002-7943-966X

REFERENCES

1. McClellan J, King MC. Genetic heterogeneity in human disease. *Cell*. 2010;141(2):210-217.
2. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008;83(3):311-321.
3. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009;5(2):e1000384.
4. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res/Fund Mol Mech Mutagen*. 2007;615(1):28-56.
5. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93.
6. Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. *Genetics*. 2014;197(4):1081-1095.
7. He Z, Zhang M, Zhan X, Lu Q. Modeling and testing for joint association using a genetic random field model. *Biometrics*. 2014;70(3):471-479.
8. Li M, He Z, Zhang M, et al. A generalized genetic random field method for the genetic association analysis of sequencing data. *Genet Epidemiol*. 2014;38(3):242-253.
9. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88(1):76-82.
10. Brook D. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*. 1964;51(3/4):481-483.
11. Qu L, Guennel T, Marshall SL. Linear score tests for variance components in linear mixed models and applications to genetic association studies. *Biometrics*. 2013;69(4):883-892. doi:10.1111/biom.12095
12. Davies RB. Hypothesis testing when a nuisance parameter is present only under the alternatives. *Biometrika*. 1987;74:33-43.
13. Stern SE, Welsh AH. Likelihood inference for small variance components. *Can J Stat*. 2000;28(3):517-532.
14. Davies RB. Algorithm AS 155: The distribution of a linear combination of χ 2 random variables. *J Royal Stat Soc Ser C (Appl Stat)*. 1980;29(3):323-333.
15. McCulloch CE, Searle SR, Neuhaus JM. *Generalized, Linear, and Mixed Models*. 2nd ed. Hoboken, NJ: Wiley; 2008.

16. Banerjee S, Carlin BP, Gelfand AE. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: CRC Press; 2014.
17. Consortium GP. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-1073.
18. Mu Y, Gage FH. Adult hippocampal neurogenesis and its role in Alzheimer's disease. *Mol Neurodegener*. 2011;6(1):1.
19. Schuff N, Woerner N, Boreta L, et al. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain*. 2009;132(4):1067-1077.
20. Juottonen K, Laakso M, Insausti R, et al. Volumes of the entorhinal and perirhinal cortices in Alzheimer's disease. *Neurobiol Aging*. 1998;19(1):15-22.
21. Thambisetty M, Simmons A, Hye A, et al. Plasma biomarkers of brain atrophy in Alzheimer's disease. *PLoS One*. 2011;6(12):e28527.
22. Ferrarini L, Palm WM, Olofsen H, van Buchem MA, Reiber JH, Admiraal-Behloul F. Shape differences of the brain ventricles in Alzheimer's disease. *NeuroImage*. 2006;32(3):1060-1069.
23. Rue H, Held L. *Gaussian Markov Random Fields: Theory and Applications*. Hoboken, NJ: CRC Press; 2005.
24. Horn RA, Johnson CR. *Matrix Analysis*. Cambridge, UK: Cambridge University Press; 1990.

How to cite this article: Shen X, Wen Y, Cui Y, Lu Q. A conditional autoregressive model for genetic association analysis accounting for genetic heterogeneity. *Statistics in Medicine*. 2021;1-26. doi: 10.1002/sim.9257

APPENDIX A. SOME TECHNICAL DETAILS

In this section, we are going to provide some technical details used in the main text. The first one is to use Brook's lemma to show that we can obtain the joint density from conditional density. To begin with, we quote the Brook's lemma.

Lemma 1 (10). *Let $\pi(\boldsymbol{x})$ be the density for $\boldsymbol{x} \in \mathbb{R}^n$ and define $\Omega = \{\boldsymbol{x} \in \mathbb{R}^n : \pi(\boldsymbol{x}) > 0\}$. Let $\boldsymbol{x}, \boldsymbol{x}' \in \Omega$, then*

$$\frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x}')} = \prod_{i=1}^{n} \frac{\pi(x_i|x_1, \dots, x_{i-1}, x'_{i+1}, \dots, x'_n)}{\pi(x'_i|x_1, \dots, x_{i-1}, x'_{i+1}, \dots, x'_n)} \tag{A1}$$

$$= \prod_{i=1}^{n} \frac{\pi(x_i|x'_1, \dots, x'_{i-1}, x_{i+1}, \dots, x_n)}{\pi(x'_i|x'_1, \dots, x'_{i-1}, x_{i+1}, \dots, x_n)} \tag{A2}$$

Next, we establish how to transfer a conditional distribution to a joint distribution.

Proposition 1. *Let $\boldsymbol{a} \in \mathbb{R}^n$ be a random vector with*

$$a_i|a_j, j \neq i \sim \mathcal{N}\left(\sum_{j\neq i} b_{ij}a_j, \tau_i^2\right),$$

then the joint density function of $\boldsymbol{a}$ is given by

$$\pi(\boldsymbol{a}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{a}^T\boldsymbol{\Delta}^{-1}(\boldsymbol{I}-\boldsymbol{B})\boldsymbol{a}\right\},$$

where $\boldsymbol{B} = [b_{ij}]$ is an $n \times n$ matrix with $b_{ii} = 0$ and $\boldsymbol{\Delta} = \mathrm{diag}\{\tau_1^2, \dots, \tau_n^2\}$. This shows that $\boldsymbol{a} \sim \mathcal{N}_n(\boldsymbol{0}, (\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Delta})$.

Proof. Fix $\boldsymbol{a}' = \boldsymbol{0}$, then based on the first equation in the Brook's lemma, we have

$$\frac{\pi(\boldsymbol{a})}{\pi(\boldsymbol{0})} = \prod_{i=1}^{n} \frac{\pi(a_i|a_1, \dots, a_{i-1}, 0, 0)}{\pi(0|a_1, \dots, a_{i-1}, 0, \dots, 0)} \tag{A3}$$

$$= \prod_{i=1}^{n} \frac{(2\pi\tau_i^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\tau_i^2}\left(a_i - \sum_{j=1}^{i-1} b_{ij}a_j\right)^2\right\}}{(2\pi\tau_i^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\tau_i^2}\left(-\sum_{j=1}^{i-1} b_{ij}a_j\right)^2\right\}} \tag{A4}$$

$$= \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\tau_i^2}\left(a_i^2 - 2\sum_{j=1}^{i-1} b_{ij}a_ia_j\right)\right\} \tag{A5}$$

$$= \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{a_i^2}{\tau_i^2} + \sum_{i=1}^{n}\frac{1}{\tau_i^2}\sum_{j=1}^{i-1} b_{ij}a_ia_j\right\}. \tag{A6}$$

Similarly, from the second equation in the Brook's lemma, we have

$$\frac{\pi(\boldsymbol{a})}{\pi(\boldsymbol{0})} = \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{a_i^2}{\tau_i^2} + \sum_{i=1}^{n}\frac{1}{\tau_i^2}\sum_{j=i+1}^{n} b_{ij}a_ia_j\right\}. \tag{A7}$$

Based on (A6) and (A7), we know that $\sum_{i=1}^{n}\frac{1}{\tau_i^2}\sum_{j=1}^{i-1} b_{ij}a_ia_j = \sum_{i=1}^{n}\frac{1}{\tau_i^2}\sum_{j=i+1}^{n} b_{ij}a_ia_j = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\tau_i^2}\sum_{j\neq i} b_{ij}a_ia_j$. Hence the density of $\boldsymbol{a}$ can then be expressed as

$$\pi(\boldsymbol{a}) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{a_i^2}{\tau_i^2} + \frac{1}{2}\sum_{i=1}^{n}\frac{1}{\tau_i^2}\sum_{j\neq i} b_{ij}a_ia_j\right\} \tag{A8}$$

$$= \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{a_i^2}{\tau_i^2} - \sum_{j\neq i}\frac{b_{ij}}{\tau_i^2}a_ia_j\right]\right\} \tag{A9}$$

$$= \exp\left\{-\frac{1}{2}\left[\boldsymbol{a}^T\boldsymbol{\Delta}^{-1}\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{\Delta}^{-1}\boldsymbol{B}\boldsymbol{a}\right]\right\} \tag{A10}$$

$$= \exp\left\{-\frac{1}{2}\boldsymbol{a}^T\boldsymbol{\Delta}^{-1}(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{a}\right\} \tag{A11}$$

where

$$\boldsymbol{B} = \begin{bmatrix} 0 & b_{12} & b_{13} & \cdots & b_{1n} \\ b_{21} & 0 & b_{23} & \cdots & b_{2n} \\ b_{31} & b_{32} & 0 & \cdots & b_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & b_{n3} & \cdots & 0 \end{bmatrix}, \quad \boldsymbol{\Delta} = \begin{bmatrix} \tau_1^2 & & & \\ & \tau_2^2 & & \\ & & \ddots & \\ & & & \tau_n^2 \end{bmatrix}.$$

This shows that $\boldsymbol{a} \sim \mathcal{N}_n(\boldsymbol{0}, (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Delta})$ and the proof is finished. ■

The next result we are going to show is that the precision matrix $\boldsymbol{D} - \gamma\boldsymbol{S}$ in the CAR model is indeed invertible when $|\gamma| < 1$. The proof of this result is based on some facts in matrix analysis. We first provide the definition of a diagonally dominant matrix.

Definition 1 (Diagonally dominant matrix). An $n \times n$ real matrix $\boldsymbol{J}$ is diagonally dominant if

$$\Delta_i(\boldsymbol{J}) := |\boldsymbol{J}_{ii}| - \sum_{j\neq i}|\boldsymbol{J}_{ij}| \geq 0, \quad \text{for } i = 1, \ldots, n. \tag{A12}$$

If the inequality in (A12) is a strict inequality, then $\boldsymbol{J}$ is called a strictly diagonally dominant matrix.

Corollary 1 (24). *Let $\boldsymbol{A}$ is an $n \times n$ matrix. If $\boldsymbol{A}$ is strictly diagonally dominant, then $\boldsymbol{A}$ is nonsingular.*

Based on Corollary 1, it is easy to obtain the desired property.

Proposition 2. *Let $\boldsymbol{D} - \gamma\boldsymbol{S}$ be as defined in the main text. Then it is nonsingular if $|\gamma| < 1$.*

Proof. Based on Corollary 1, it suffices to check that $\boldsymbol{D} - \gamma\boldsymbol{S}$ is strictly diagonally dominant if $|\gamma| < 1$. Note that

$$J := D - \gamma S = \begin{bmatrix} \sum_{j\neq 1} s_{1j} & -\gamma s_{12} & \cdots & -\gamma s_{1n} \\ -\gamma s_{21} & \sum_{j\neq 2} s_{2j} & \cdots & -\gamma s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\gamma s_{n1} & -\gamma s_{n2} & \cdots & \sum_{j\neq n} s_{nj} \end{bmatrix},$$

then since $s_{ij} \geq 0$ for i, j by the definition of similarity, we have

$$\begin{aligned} \Delta_i(\boldsymbol{J}) &= |\boldsymbol{J}_{ii}| - \sum_{j\neq i} |\boldsymbol{J}_{ij}| & \text{(A13)} \\ &= \left|\sum_{j\neq i} s_{ij}\right| - \sum_{j\neq i} |-\gamma s_{ij}| & \text{(A14)} \\ &= \sum_{j\neq i} s_{ij} - |\gamma| \sum_{j\neq i} s_{ij} & \text{(A15)} \\ &= (1 - |\gamma|) \sum_{j\neq i} s_{ij}. & \text{(A16)} \end{aligned}$$

Hence

$$\Delta_i(\boldsymbol{J}) > 0 \Leftrightarrow |\gamma| < 1,$$

which finishes the proof. ■

In the last part, we provide a detailed derivation of (16). It follows from the definition of $S(0)$ in the main text that

$$\mathbb{P}(S(0) > s) = \mathbb{P}\left(\frac{q}{2}\frac{\boldsymbol{y}^{*^T}\boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\boldsymbol{y}^*}{\boldsymbol{y}^{*^T}\boldsymbol{y}^*} - \frac{1}{2}\text{tr}\left[\boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\right] > s\right) \quad \text{(A17)}$$

$$= \mathbb{P}\left(\frac{\boldsymbol{y}^{*^T}\boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\boldsymbol{y}^*}{\boldsymbol{y}^{*^T}\boldsymbol{y}^*} > \frac{2s}{q} + \frac{1}{q}\text{tr}\left[\boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\right]\right) \quad \text{(A18)}$$

$$= \mathbb{P}\left(\boldsymbol{y}^{*^T}\boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\boldsymbol{y}^* > \boldsymbol{y}^{*^T}\left(\frac{2s}{q} + \frac{1}{q}\text{tr}\left[\boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\right]\right)\boldsymbol{I}_q\boldsymbol{y}^* > 0\right) \quad \text{(A19)}$$

$$= \mathbb{P}\left(\boldsymbol{y}^{*^T}\boldsymbol{B}\boldsymbol{y}^* > 0\right) \quad \text{(A20)}$$

$$= \mathbb{P}\left(\left(\frac{\boldsymbol{y}^*}{\sigma}\right)^T\boldsymbol{B}\left(\frac{\boldsymbol{y}^*}{\sigma}\right) > 0\right) \quad \text{(A21)}$$

$$= \mathbb{P}\left(\sum_{j=1}^{q}\lambda_j Z_j^2 > 0\right), \quad \text{(A22)}$$

where $\boldsymbol{B} = \boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T - \left(\frac{2s}{q} + \frac{1}{q}\text{tr}\left[\boldsymbol{K}(\boldsymbol{D}-\gamma\boldsymbol{S})^{-1}\boldsymbol{K}^T\right]\right)\boldsymbol{I}_q$ and the last equality follows since under H_0, $\boldsymbol{y}^* \sim \mathcal{N}_q(\boldsymbol{0}, \sigma^2\boldsymbol{I}_q)$. $\lambda_1, \ldots, \lambda_q$ are eigenvalues of $\boldsymbol{B}$.

APPENDIX B. DEFINITIONS OF WEIGHTS USED IN THE ARTICLE

We considered four different weight functions based on the minor allele frequencies (MAF).[8] Among these four weight functions, unweighted (UW) assigns same weights to both common and rare variants; weighted sum statistics type of weight (WSS), on the other hand, put almost all the weights on rare variants and nearly no weights on the common variants. The Beta distribution type of weights (BETA) and the logarithm of MAFs (LOG) lie between UW and WSS with LOG putting more weights on common variants than BETA.

1. Unweighted (UW)

$$\omega_k = 1, \quad 1 \leq k \leq p$$

2. Beta distribution type of weights (BETA)

$$\omega_k = \text{dbeta}(\text{MAF}_k, 1, 25)^2, \quad 1 \leq k \leq p,$$

   that is, the weight is the square of the probability density of the Beta distribution with parameters 1 and 25.
3. Weighted sum statistics type of weights (WSS)

$$\omega_k = \frac{1}{\text{MAF}_k(1 - \text{MAF}_k)}, \quad 1 \leq k \leq p$$

4. Logarithm of MAFs as weights (LOG)

$$\omega_k = -\log_{10}(\text{MAF}_k), \quad 1 \leq k \leq p$$

APPENDIX C. DISTRIBUTION OF MAF OF SEQUENCING VARIANTS ON CHROMOSOME 17 FROM THE 1000 GENOME PROJECT

In the simulation studies, to mimic the real structure of sequencing data, the simulated data was based on the sequencing data from Chromosome 17: 7344328-8344327 from the 1000 Genome project.[17] Figure C1 summarizes the distribution of the MAF with MAF< 0.05.

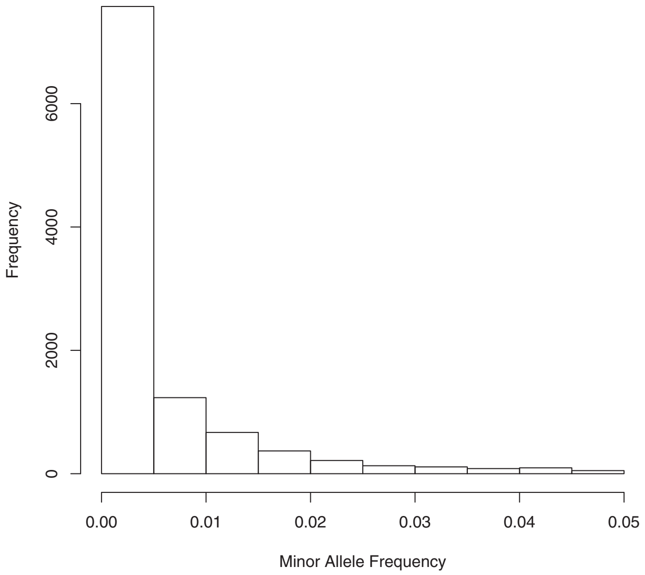

FIGURE C1 The Distribution of minor allele frequency of in sequencing variants on chromosome 17 from the 1000 Genome project

APPENDIX D. EMPIRICAL POWER COMPARISON BETWEEN GGRF, SKAT, AND CAR UNDER THE SCENARIO WHEN ALL INDIVIDUALS HAVE SAME GENETIC EFFECT

In the simulation studies, we also consider the case when all the individuals have the same genetic effect. That is we simulate phenotypes using the following model

$$y_i = \sum_{k=1}^{p} w_k^* g_{i,k} Z_k + \varepsilon_i, \quad 1 \leq i \leq n. \tag{D1}$$

The only difference in model (D1) from model (3.6) in the main text is that in model (D1), Z_k does not depend on the individual i and we also assume that $Z_1, \dots, Z_n \sim \mathcal{N}(0, \sigma_Z^2)$ so that all the individual have the same genetic effect. We regard such situation as homogeneous genetic effects.

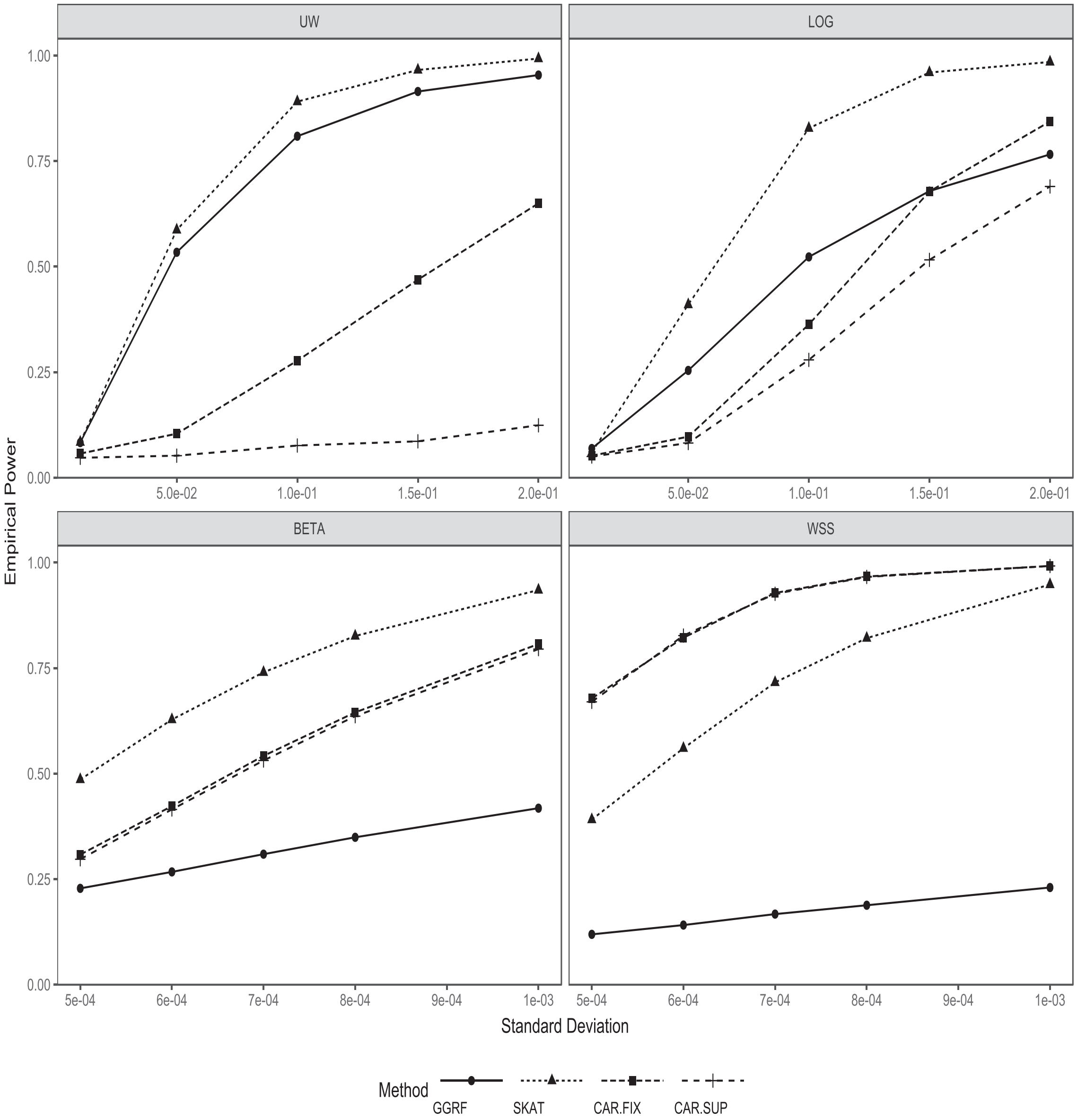

FIGURE D1 Comparison of empirical power under different weights. The x-axis are the effect size σ_Z used in the simulation. For weights UW and LOG, the effect sizes are chosen as 0.01, 0.05, 0.1, 0.15, 0.2 and for weights BETA and WSS, the effect sizes are chosen as 0.0005, 0.0006, 0.0007, 0.0008, 0.001

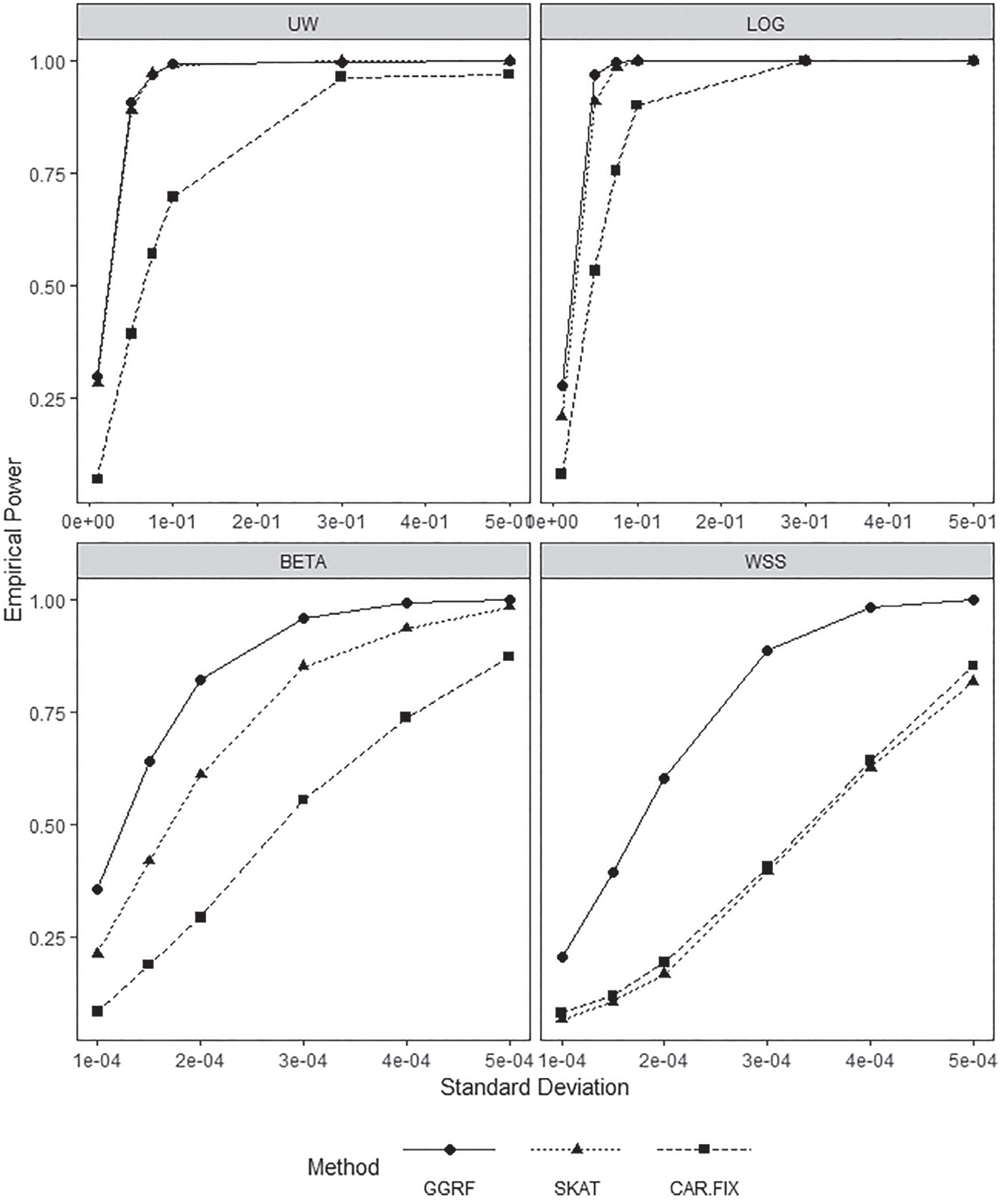

FIGURE D2 Comparison of empirical power under different weights. The *x*-axis are the effect size used in the simulation. For weights UW and LOG, the effect sizes are chosen as 0.01, 0.05, 0.075, 0.1, 0.3, 0.5 and for weights BETA and WSS, the effect sizes are chosen as 0.0001, 0.00015, 0.0002, 0.0003, 0.0004, 0.0005

Figure D1 summarizes the results of simulation under model (D1). From left to right, the weights are used as UW, BETA, WSS, and LOG, respectively. According to the results, we can see that CAR.FIX and CAR.SUP still have similar performances when LOG, BETA, or WSS weights are used, but CAR.SUP has low power increase when UW is used. CAR performs the best when WSS is used and it also has moderate power when BETA is used. However, as we put more weights on the common variants, CAR suffers from power loss as we can see from the first two figures. One thing we can notice that when the effect size σ_Z is large, it still has moderate or high power. SKAT, in this case, has the consistent satisfactory power under all the four scenarios. For GGRF, it has high power when common variants are put more weights, but with more weights putting on rare variants, its power decreases significantly.

We also conducted simulations under the scenario when the genetic effects are one direction. Specifically, we still used the simulation model (D1) to generate the data except that Z_k is fixed at certain values. For UW and LOG, Z_k is chosen among the set $\{0.01, 0.05, 0.075, 0.1, 0.3, 0.5\}$, while for BETA and WSS, Z_k is chosen among the set $\{0.0001, 0.00015, 0.0002, 0.0003, 0.0004, 0.0005\}$ Figure D2.

APPENDIX E. TOP 10 SIGNIFICANT GENES FOUND BY CAR, GGRF, AND SKAT FOR ENTORHINAL, VENTRICLE, AND WHOLE BRAIN

Tables E1–E3 summarize the top 10 significant gene selected by the three methods associated with entorhinal, ventricle, and whole brain respectively.

TABLE E1 Top 10 entorhinal-associated genes and their corresponding *P*-values from different methods

CHR	Gene name	CAR	SKAT	GGRF
14	*GSTZ1*	2.48E-01	2.58E-04	2.10E-05
2	*HPCAL1*	6.62E-03	4.51E-05	3.50E-05
12	*HOXC8*	3.84E-02	1.07E-04	9.95E-01
10	*KCNIP2*	1.24E-04	9.80E-02	2.41E-01
3	*CBLB*	2.30E-01	1.64E-04	1.39E-03
18	*RBFADN*	2.11E-04	2.72E-01	6.89E-01
1	*LOC100130331*	9.66E-02	2.33E-04	4.43E-04
15	*IDH2*	6.73E-01	3.16E-04	2.10E-05
15	*ADAL*	3.29E-04	2.20E-01	1.19E-01
6	*RPL7L1*	8.24E-01	6.05E-04	4.26E-04

TABLE E2 Top 10 ventrical-associated genes and their corresponding *P*-values from different methods

CHR	Gene name	CAR	SKAT	GGRF
5	*LINC01033*	6.63E-01	2.35E-04	9.96E-05
4	*JCHAIN*	7.05E-03	1.65E-04	1.22E-04
9	*CTSL*	1.16E-01	5.67E-04	1.35E-04
12	*ATP23*	2.94E-04	5.53E-01	7.80E-01
1	*LOC100129534,MORN1*	3.11E-04	6.98E-01	6.47E-01
4	*SEL1L3*	4.24E-04	6.68E-02	2.90E-02
3	*PTPRG-AS1*	5.27E-01	4.43e-04	4.94E-04
20	*NRSN2*	5.07E-04	3.09E-02	3.62E-01
12	*MGST1*	5.67E-04	2.19E-01	8.11E-01
13	*RPL21,RPL21P28,SNORA27,SNORD102*	4.17E-04	5.43E-04	5.67E-04

TABLE E3 Top 10 whole brain-associated genes and their corresponding *P*-values from different methods

CHR	Gene name	CAR	SKAT	GGRF
17	*KRTAP16-1*	7.44E-02	3.89E-03	3.12E-05
7	*GALNT11*	8.63E-03	5.96E-05	3.20E-05
9	*NR5A1*	5.25E-05	2.11E-01	3.31E-01
7	*TRIM24*	1.14E-03	6.29E-05	6.49E-05
11	*CD81*	4.46E-02	9.84E-05	8.03E-05
2	*PPP1R7*	1.10E-02	4.81E-04	1.18E-04
21	*TRPM2-AS*	2.51E-02	1.43E-04	3.76E-04
11	*TSSC4*	5.98E-02	1.56E-04	1.19E-04
4	*AMTN*	5.67E-04	1.27E-03	1.69E-04
5	*CANX*	1.99E-01	5.27E-04	2.63E-04