mathematics

MDPI

Article

Robust Linear Trend Test for Low-Coverage Next-Generation Sequence Data Controlling for Covariates

Jung Yeon Lee [1], Myeong-Kyu Kim [2] and Wonkuk Kim [3,*]

[1] Department of Psychiatry, New York University School of Medicine, New York, NY 10016, USA; JungYeon.Lee@nyulangone.org
[2] Department of Neurology, Chonnam National University Medical School, Gwangju 61469, Korea; mkkim@jnu.ac.kr
[3] Department of Applied Statistics, Chung-Ang University, Seoul 06974, Korea
* Correspondence: wkim@cau.ac.kr; Tel.: +82-2-820-6688

Received: 30 December 2019; Accepted: 5 February 2020; Published: 8 February 2020

Abstract: Low-coverage next-generation sequencing experiments assisted by statistical methods are popular in a genetic association study. Next-generation sequencing experiments produce genotype data that include allele read counts and read depths. For low sequencing depths, the genotypes tend to be highly uncertain; therefore, the uncertain genotypes are usually removed or imputed before performing a statistical analysis. It may result in the inflated type I error rate and in a loss of statistical power. In this paper, we propose a mixture-based penalized score association test adjusting for non-genetic covariates. The proposed score test statistic is based on a sandwich variance estimator so that it is robust under the model misspecification between the covariates and the latent genotypes. The proposed method takes advantage of not requiring either external imputation or elimination of uncertain genotypes. The results of our simulation study show that the type I error rates are well controlled and the proposed association test have reasonable statistical power. As an illustration, we apply our statistic to pharmacogenomics data for drug responsiveness among 400 epilepsy patients.

Keywords: allele read counts; low-coverage; mixture model; next-generation sequencing; sandwich variance estimator

1. Introduction

Genome-wide association study (GWAS) is a powerful tool for screening a high-dimensional genome data set and selecting candidate genetic variants such as single nucleotide polymorphisms (SNPs) in genetic association studies. Next-generation sequencing (NGS) technology is widely used to produce a large amount of genetic information in a fast way. In the past decade, there have been numerous studies using NGS data such as rare variants association study [1,2], pharmacogenomics [3,4], machine learning and deep learning applications [5,6], and big data analysis [7,8]. Many NGS experiments are based on low-coverage sequencing with a large sized sample since there is a trade-off between sample size and sequencing depth in the NGS experiments [9,10]. For the low-coverage NGS data, a high uncertainty of the inferred genotypes is common; however, it causes biased and unreliable results on genetic association analyses. In genetic research based on NGS data, therefore, it is important to obtain accurate genotypes to perform an association analysis.

A number of researchers have worked on the effects of genotype misclassification in genetic association studies. There are two types of genotype misclassifications: differential and non-differential misclassifications, determined by whether the misclassification mechanism differs in the case and

control groups or not. In summary, non-differential misclassifications result in a loss of statistical power and differential misclassifications distort type I error rates in a genetic case-control association study [11–14].

While there have been many research on improving the accuracy of genotypes such as the joint genotype calling algorithms across all samples were suggested to increase the accuracy of genotype calls [15–17], several researchers have tried to develop new association statistics accounting for the genotype errors. Their approaches are based on the raw measurements rather than inferred genotypes. In statistical genetics literature, Kim et al. [18] extended a chi-squared test of independence and developed a mixture likelihood based association test using the continuous measurements for copy number polymorphisms. Barnes et al. [19] proposed a mixture model linear trend test for the continuous copy number measurements. In NGS experiments, a likelihood ratio test based on allele read counts of pooled samples was proposed to test independence of genetic variants with a binary phenotype [20]. Gordon et al. [21] proposed a likelihood ratio test of the binomial mixture model of allele read counts with known error parameters. Kim et al. [13,22] proposed an extended version of Cochran–Armitage (CA) trend test and a multi-variant linear trend test for next-generation sequences data by using binomial mixture models. For a case-parent trio design, the binomial mixture model was applied to develop extended transmission disequilibrium tests (TDTs) based on read counts and read depths and to provide power analysis and sample size formulas [23]. All these approaches do not require genotype calls that can be highly uncertain when the read depth or coverage is low. However, none of these previous research has addressed how to include covariates in their mixture-based association studies.

When the covariates are independent of the latent genotypes, the extension of the mixture model based association tests is straightforward. However, if the latent genotype variable is associated with other covariates, then a likelihood based approach requires a model specification between the genotype variable and the other covariates as opposed to the previous research [16–23]. To our knowledge, this is the first study that investigates a genetic case-control association test controlling for covariates in low-coverage NGS experiments. Since we do not know the true model, we apply a sandwich variance estimator to develop a robust genetic association test statistic.

2. Materials and Methods

2.1. Mixture Model Accounting for Covariates

Let $\mathbf{w}$ be a covariate vector. Let y be a random variable indicating the case-control status of an individual such that $y = 1$ if a subject is in the case group and $y = 0$, otherwise. Let $\mathbf{z} = (z_0, z_1, z_2)$ denote an unobservable latent genotype vector, where $\sum_{g=0}^{2} z_g = 1$ and $z_g = 1$ if and only if the genotype is equal to g. Let x and v denote the minor allele read count and the read depth, respectively. The probability function is given by

$$\begin{aligned} p(y, x, v, \mathbf{w}) &= \sum_{\mathbf{z}} p(y, x, v, \mathbf{w}, \mathbf{z}) \\ &= \sum_{\mathbf{z}} p(x|v, \mathbf{w}, \mathbf{z}, y) p(y|v, \mathbf{w}, \mathbf{z}) p(\mathbf{z}|\mathbf{w}, v) p(\mathbf{w}, v) \\ &= p(\mathbf{w}, v) \sum_{\mathbf{z}} p(x|v, \mathbf{z}, y) p(y|\mathbf{z}, \mathbf{w}) p(\mathbf{z}|\mathbf{w}). \end{aligned} \tag{1}$$

If the probability function of the read count x does not depend on the phenotype y, that is, $p(x|v, \mathbf{z}, y) = p(x|v, \mathbf{z})$, then it is called a non-differential error model. We apply a binary logit model to the case-control phenotype response variable y that is the same model for Cochran–Armitage trend test when perfect genotypes are available:

$$p(y|\mathbf{z}, \mathbf{w}) = \frac{e^{y(\beta \mathbf{s}^T \mathbf{z} + \beta_w^T \mathbf{w})}}{1 + e^{\beta_0 + \beta \mathbf{s}^T \mathbf{z} + \beta_w^T \mathbf{w}}}. \tag{2}$$

We assume a binomial error model to the allele read counts as in previous research [13,16,20,21,23]:

$$p(x|v,\mathbf{z},y) = \binom{v}{x}\left(\mathbf{u}_\epsilon^T\mathbf{z}\right)^x\left(1-\mathbf{u}_\epsilon^T\mathbf{z}\right)^{v-x}, \tag{3}$$

where $\mathbf{u}_\epsilon = (\epsilon, 1/2, 1-\epsilon)^T$. For a differential error model, we can use $\mathbf{u}_\epsilon = y(\epsilon_1, 1/2, 1-\epsilon_1)^T + (1-y)(\epsilon_0, 1/2, 1-\epsilon_0)^T$. When perfect genotypes are available, we do not need the conditional probability of the genotype $\mathbf{z}$ given covariates $\mathbf{w}$ to perform genetic association tests since the logistic regression model is a conditional model given the genotypes and covariates. In this work, we assume a multinomial logit model for the latent genotype given the covariates as follows:

$$p(\mathbf{z}|\mathbf{w}) = \frac{\sum_{g=0}^2 z_g e^{\gamma_g^T\mathbf{w}}}{\sum_{m=0}^2 e^{\gamma_m^T\mathbf{w}}}, \tag{4}$$

where $\gamma_0 = (0,0,0)^T$ to remove over-parametrization. Other statistical models without the assumptions of a multinomial logit model may also be used for the relationship between covariates and latent genotypes, where we do not know the true model.

The likelihood function L and the log-likelihood function ℓ are written as

$$\begin{aligned} L &= \prod_{k=1}^N\left[\sum_{\mathbf{z}_k} p(y_k|\mathbf{z}_k,\mathbf{w}_k)p(x_k|v_k,\mathbf{z}_k,y_k)p(\mathbf{z}_k|\mathbf{w}_k)p(\mathbf{w}_k,v_k)\right] \\ &= \prod_{k=1}^N\sum_{i=0}^2\left\{\left(\frac{e^{y_k(\beta s_i+\beta_w^T\mathbf{w}_k)}}{1+e^{\beta s_i+\beta_w^T\mathbf{w}_k}}\right)\left(\binom{v_k}{x_k}(u_{\epsilon i})^{x_k}(1-u_{\epsilon i})^{v_k-x_k}\right)\left(\frac{e^{\gamma_i^T\mathbf{w}_k}}{\sum_{m=0}^2 e^{\gamma_m^T\mathbf{w}_k}}\right)p(\mathbf{w}_k,v_k)\right\}, \quad (5) \\ \ell &= \sum_{k=1}^N\log\left[\sum_{i=0}^2\left\{\left(\frac{e^{y_k(\beta s_i+\beta_w^T\mathbf{w}_k)}}{1+e^{\beta s_i+\beta_w^T\mathbf{w}_k}}\right)\left((u_{\epsilon i})^{x_k}(1-u_{\epsilon i})^{v_k-x_k}\right)\left(\frac{e^{\gamma_i^T\mathbf{w}_k}}{\sum_{m=0}^2 e^{\gamma_m^T\mathbf{w}_k}}\right)\right\}\right] \\ &\quad + \sum_{k=1}^N\log\binom{v_k}{x_k}p(\mathbf{w}_k,v_k). \quad (6) \end{aligned}$$

The error parameter ϵ is commonly small and hence the estimate of ϵ is often equal to zero. The zero estimate of the error parameter results in a divergent information matrix. It prevents us from calculating Rao's score test statistic. In order to overcome this issue, we include a beta density penalty term to prevent from zero estimate of the error parameter. The penalized log-likelihood function is given by

$$\ell_p = \ell + C\log\left[\epsilon^{a_\epsilon}(1-\epsilon)^{b_\epsilon}\right]. \tag{7}$$

During this work, we choose $C = 1$ as in [24,25]. The penalized complete-data likelihood function is given by

$$L_C = \prod_{k=1}^N\prod_{i=0}^2\left[\frac{e^{y_k(\beta s_i+\beta_w^T\mathbf{w}_k)}}{1+e^{\beta s_i+\beta_w^T\mathbf{w}_k}}\times\binom{v_k}{x_k}(u_{\epsilon i})^{x_k}(1-u_{\epsilon i})^{v_k-x_k}\epsilon^{\frac{a_\epsilon}{N}}(1-\epsilon)^{\frac{b_\epsilon}{N}}\times\frac{e^{\gamma_i^T\mathbf{w}_k}}{\sum_{m=0}^2 e^{\gamma_m^T\mathbf{w}_k}}\right]^{z_{ik}} \tag{8}$$

The complete data log-likelihood function is written as

$$\begin{aligned} \ell_C &= \sum_{k=1}^N\sum_{i=0}^2 z_{ik}\left[y_k(\beta s_i+\beta_w^T\mathbf{w}_k) - \log\left(1+e^{\beta s_i+\beta_w^T\mathbf{w}_k}\right)\right] \\ &\quad + \sum_{k=1}^N\sum_{i=0}^2 z_{ik}\left[x_k\log(u_{\epsilon i}) + (v_k-x_k)\log(1-u_{\epsilon i})\right] + a_\epsilon\log\epsilon + b_\epsilon\log(1-\epsilon) \quad (9) \\ &\quad + \sum_{k=1}^N\sum_{i=0}^2 z_{ik}\left[\gamma_i^T\mathbf{w}_k - \log\left(\sum_{m=0}^2 e^{\gamma_m^T\mathbf{w}_k}\right)\right]. \end{aligned}$$

2.2. Derivation of EM Algorithm under H_0

We apply the Expectation–Maximization (EM) algorithm [26] to estimate the parameters in our mixture model. Given data and the (r)-th step estimated parameters, the $(r+1)$-th E-step of the EM algorithm is written as

$$\begin{aligned} Q^{(r+1)} &= \sum_{k=1}^{N}\sum_{i=0}^{2}\tau_{ik}^{(r)}\left[y_k(\beta s_i+\beta_w^T\mathbf{w}_k)-\log\left(1+e^{\beta s_i+\beta_w^T\mathbf{w}_k}\right)\right] \\ &+\sum_{k=1}^{N}\sum_{i=0}^{2}\tau_{ik}^{(r)}\left[x_k\log(u_{\epsilon i})+(v_k-x_k)\log(1-u_{\epsilon i})\right]+a_\epsilon\log\epsilon+b_\epsilon\log(1-\epsilon) \\ &+\sum_{k=1}^{N}\sum_{i=0}^{2}\tau_{ik}^{(r)}\left[\gamma_i^T\mathbf{w}_k-\log\left(\sum_{m=0}^{2}e^{\gamma_m^T\mathbf{w}_k}\right)\right], \end{aligned} \tag{10}$$

where

$$\tau_{ik}^{(r)} = \frac{\left(\frac{e^{y_k(\beta^{(r)}s_i+\beta_w^{(r)T}\mathbf{w}_k)}}{1+e^{\beta^{(r)}s_i+\beta_w^{(r)T}\mathbf{w}_k}}\right)\left((u_{ei}^{(r)})^{x_k}(1-u_{ei}^{(r)})^{v_k-x_k}\right)\left(\frac{e^{\gamma_i^{(r)T}\mathbf{w}_k}}{\sum_{m=0}^{2}e^{\gamma_m^{(r)T}\mathbf{w}_k}}\right)}{\sum_{g=0}^{2}\left[\left(\frac{e^{y_k(\beta^{(r)}s_g+\beta_w^{(r)T}\mathbf{w}_k)}}{1+e^{\beta^{(r)}s_g+\beta_w^{(r)T}\mathbf{w}_k}}\right)\left((u_{eg}^{(r)})^{x_k}(1-u_{eg}^{(r)})^{v_k-x_k}\right)\left(\frac{e^{\gamma_g^{(r)T}\mathbf{w}_k}}{\sum_{m=0}^{2}e^{\gamma_m^{(r)T}\mathbf{w}_k}}\right)\right]}. \tag{11}$$

We note that the posterior probability of subject k belonging to genotype class i depends on the all parameters. In M-step, the $(r+1)$-th estimates of the parameters are obtained by maximizing $Q^{(r+1)}$:

$$\frac{\partial Q^{(r+1)}}{\partial\beta} = \sum_{k=1}^{N}\sum_{i=0}^{2}\tau_{ik}^{(r)}s_i\left(y_k-\pi_{ik}\right)=0 \tag{12}$$

$$\frac{\partial Q^{(r+1)}}{\partial\beta_w} = \sum_{k=1}^{N}\sum_{i=0}^{2}\tau_{ik}^{(r)}\mathbf{w}_k\left(y_k-\pi_{ik}\right)=0 \tag{13}$$

$$\frac{\partial Q^{(r+1)}}{\partial\epsilon} = \sum_{k=1}^{N}\left[\tau_{0k}^{(r)}\left(\frac{x_k}{\epsilon}-\frac{v_k-x_k}{1-\epsilon}\right)+\tau_{2k}^{(r)}\left(\frac{v_k-x_k}{\epsilon}-\frac{x_k}{1-\epsilon}\right)\right]+\frac{a_\epsilon}{\epsilon}-\frac{b_\epsilon}{1-\epsilon}=0 \tag{14}$$

$$\frac{\partial Q^{(r+1)}}{\partial\gamma_i} = \sum_{k=1}^{N}\mathbf{w}_k\left(\tau_{ik}^{(r)}-p_{ik}\right)=0, \tag{15}$$

where we use notations $\pi_{ik}=\pi_{ik}(\beta,\beta_w)=\frac{e^{\beta s_i+\beta_w^T\mathbf{w}_k}}{1+e^{\beta s_i+\beta_w^T\mathbf{w}_k}}$ and $p_{ik}=p_{ik}(\gamma_1,\gamma_2)=\frac{e^{\gamma_i^T\mathbf{w}_k}}{\sum_{m=0}^{2}e^{\gamma_m^T\mathbf{w}_k}}$ for simplicity. From Equation (14), we derive a closed form iteration formula to update the allele read error parameter ϵ:

$$\epsilon^{(r+1)} = \frac{\sum_{k=1}^{N}\left[\tau_{0k}^{(r)}x_k+\tau_{2k}^{(r)}(v_k-x_k)\right]+a_\epsilon}{\sum_{k=1}^{N}\left[(\tau_{0k}^{(r)}+\tau_{2k}^{(r)})v_k\right]+a_\epsilon+b_\epsilon}. \tag{16}$$

There is no closed form iteration formulas to update other parameters β,β_w,γ_i. The M-step for β,β_w, and γ can be obtained by the Newton–Raphson method. The Hessian matrix of $Q^{(r+1)}$ is given by

$$\frac{\partial^2 Q^{(r+1)}}{\partial\beta^2} = -\sum_{k=1}^{N}\sum_{i=0}^{2}\tau_{ik}^{(r)}s_i^2\left[\pi_{ik}(1-\pi_{ik})\right] \tag{17}$$

$$\frac{\partial^2 Q^{(r+1)}}{\partial\beta\partial\beta_w} = -\sum_{k=1}^{N}\sum_{i=0}^{2}\tau_{ik}^{(r)}s_i\mathbf{w}_k\left[\pi_{ik}(1-\pi_{ik})\right] \tag{18}$$

$$\frac{\partial^2 Q^{(r+1)}}{\partial \beta_w \partial \beta_w^T} = -\sum_{k=1}^{N}\sum_{i=0}^{2} \tau_{ik}^{(r)} \mathbf{w}_k \mathbf{w}_k^T \left[\pi_{ik}(1-\pi_{ik})\right] \quad (19)$$

$$\frac{\partial^2 Q^{(r+1)}}{\partial \gamma_i \partial \gamma_i^T} = -\sum_{k=1}^{N} \mathbf{w}_k \mathbf{w}_k^T \left[p_{ik}(1-p_{ik})\right] \quad (20)$$

$$\frac{\partial^2 Q^{(r+1)}}{\partial \gamma_i \partial \gamma_j^T} = \sum_{k=1}^{N} \mathbf{w}_k \mathbf{w}_k^T \left[p_{ik} p_{jk}\right] \quad (21)$$

$$\frac{\partial^2 Q^{(r+1)}}{\partial \gamma_i \partial \beta_w^T} = \frac{\partial^2 Q^{(r+1)}}{\partial \gamma_i \partial \beta} = 0 \quad (22)$$

Let $M = \text{diag}\left(\sum_{i=0}^{2} \tau_{ik}\pi_{ik}(1-\pi_{ik})\right)$ be an $N \times N$ diagonal matrix. Let $W = (w_{ik})$ be the $N \times p$ matrix of covariates. Let μ be an $N \times 1$ vector of $\mu_k = \sum_{ik} \tau_{ik}\pi_{ik}$ and Y be an $N \times 1$ vector of y_k. Initially, we set $\beta^{[0]} = \beta^{(r)}$ and update the parameter estimate by

$$\beta^{[t+1]} = \beta^{[t]} + (W^T M W)^{-1} W^T (Y - \mu). \quad (23)$$

Let $D_{11} = \text{diag}\,(p_{1k}(1-p_{1k}))$, $D_{12} = D_{21} = -\text{diag}\,(p_{1k}p_{2k})$, and $D_{22} = \text{diag}\,(p_{2k}(1-p_{2k}))$. Let $\tau_i = (\tau_{ik})$ be the $N \times 1$ vector and $p_i = (p_{ik})$ be the $N \times 1$ vector. Initially, set $\gamma_i^{[0]} = \gamma_i^{(r)}$ and update the parameters γ_i by

$$\begin{pmatrix} \gamma_1^{[t+1]} \\ \gamma_2^{[t+1]} \end{pmatrix} = \begin{pmatrix} \gamma_1^{[t]} \\ \gamma_2^{[t]} \end{pmatrix} + \begin{pmatrix} W^T D_{11} W & W^T D_{12} W \\ W^T D_{21} W & W^T D_{22} W \end{pmatrix}^{-1} \begin{pmatrix} W^T(\tau_1 - p_1) \\ W^T(\tau_2 - p_2) \end{pmatrix}. \quad (24)$$

In order to obtain $\beta_w^{(r+1)}$ and $\gamma_i^{(r+1)}$, we stop the iterations in the M-step for β and γ_i when $||\beta^{[t+1]} - \beta^{[t]}||^2 + ||\gamma_1^{[t+1]} - \gamma_1^{[t]}||^2 + ||\gamma_2^{[t+1]} - \gamma_2^{[t]}||^2 \leq tol^2$ or the number of iterations reaches the prespecified maximum number of iterations. In our work, we set $tol = 10^{-6}$ and fix the maximum iteration as 1000.

2.3. Hypothesis Tests of Genetic Association Controlling for Covariates

To test genetic association between the latent genetic variables and the binary response variable while controlling covariates, we employ Rao's score test. There are several advantages for the use of the score test. Cochran-Armitage trend test with perfect genotypes is a score test, and we extend this test to when the genotypes are highly uncertain. The score test requires less computational cost compared to the likelihood ratio test since it requires the parameter estimates only under the null hypothesis of no association. The score function calculated in previous section is given by

$$S = \sum_{k=1}^{N}\sum_{i=0}^{2} \tau_{ik(0)} s_i \left(y_k - \frac{e^{\beta_{w(0)}^T \mathbf{w}_k}}{1 + e^{\beta_{w(0)}^T \mathbf{w}_k}} \right) \quad (25)$$

where the subscript (0) denotes the estimated parameter under the null hypothesis. Another important issue to be considered when we include the covariates in a low-coverage next-generation sequencing genetic association study is a model misspecification of the latent genotypes on the covariates. To overcome this model misspecification problem, we employ the sandwich variance estimator [27]. In this work, we derive a robust generalized score test using the sandwich variance–covariance estimator. In general, one of the difficulties in applying the sandwich estimator in practice is that it requires analytic derivation for the covariance matrix of the proposed model. For simplicity in our derivation of the sandwich variance estimator, θ denotes the vector of all parameters $\theta = (\beta, \beta_w, \gamma, \epsilon)$,

and $\phi = (\beta_w, \gamma, \epsilon)$ denotes the parameter vector except β, and hence $\theta = (\beta, \phi)$. The sandwich variance estimator for the score function S under H_0 is given by

$$v_s = V_{\beta\beta} - V_{\beta\phi}J_{\phi\phi}^{-1}J_{\phi\beta} - J_{\beta\phi}J_{\phi\phi}^{-1}V_{\phi\beta} + J_{\beta\phi}J_{\phi\phi}^{-1}V_{\phi\phi}J_{\phi\phi}^{-1}J_{\phi\beta}, \tag{26}$$

where $V = E_{f_0}\left[\frac{\partial \ell}{\partial \theta}\frac{\partial \ell}{\partial \theta}^T\right]$ and $J = -E_{f_0}\left[\frac{\partial^2 \ell}{\partial \theta \partial \theta^T}\right]$ under the unknown true distribution f_0. For simplicity, we may use h_{ik} during derivation of the sandwich variance estimator:

$$h_{gk} = \left(\frac{e^{y_k(\beta s_g + \beta_w^T \mathbf{w}_k)}}{1 + e^{\beta s_g + \beta_w^T \mathbf{w}_k}}\right)\left(\left(u_{\epsilon g}\right)^{x_k}\left(1 - u_{\epsilon g}\right)^{v_k - x_k}\right)\left(\frac{e^{\gamma_g^T \mathbf{w}_k}}{\sum_{m=0}^{2} e^{\gamma_m^T \mathbf{w}_k}}\right), \tag{27}$$

so that the likelihood function is written as

$$\ell = \sum_{k=1}^{N} \log\left[\sum_{g=0}^{2} h_{gk}\right] + C. \tag{28}$$

The relationship between J and V can be written as

$$\begin{aligned} J &= \frac{1}{N}\sum_{k=1}^{N}\left[\sum_{g=0}^{2}\tau_{gk}\frac{\partial}{\partial\theta}\log h_{gk}\right]\left[\sum_{g=0}^{2}\tau_{gk}\frac{\partial}{\partial\theta^T}\log h_{gk}\right] \\ &\quad -\frac{1}{N}\sum_{k=1}^{N}\sum_{g=0}^{2}\tau_{gk}\left[\left(\frac{\partial}{\partial\theta}\log h_{gk}\right)\left(\frac{\partial}{\partial\theta^T}\log h_{gk}\right) + \frac{\partial^2}{\partial\theta\partial\theta^T}\log h_{gk}\right] \\ &= V - \frac{1}{N}\sum_{k=1}^{N}\sum_{g=0}^{2}\tau_{gk}\left[\left(\frac{\partial}{\partial\theta}\log h_{gk}\right)\left(\frac{\partial}{\partial\theta^T}\log h_{gk}\right) + \frac{\partial^2}{\partial\theta\partial\theta^T}\log h_{gk}\right] \end{aligned} \tag{29}$$

If there is no model misspecification, we have $J = V$ and the robust score test statistic is reduced to Rao's score test statistic. We denote the difference $R = V - J$ so that $R = \frac{1}{N}\sum_{k=1}^{N}\sum_{g=0}^{2}\tau_{gk}\left[\left(\frac{\partial}{\partial\theta}\log h_{gk}\right)\left(\frac{\partial}{\partial\theta^T}\log h_{gk}\right) + \frac{\partial^2}{\partial\theta\partial\theta^T}\log h_{gk}\right]$. The components of $\frac{\partial}{\partial\theta}\log h_{gk}$ are calculated by

$$\frac{\partial}{\partial\beta}\log h_{gk} = s_g[y_k - \pi_k] \tag{30}$$

$$\frac{\partial}{\partial\beta_w}\log h_{gk} = \mathbf{w}_k[y_k - \pi_k] \tag{31}$$

$$\frac{\partial}{\partial\epsilon}\log h_{gk} = \delta_g(0)\left[\frac{X_k}{\epsilon} - \frac{V_k - X_k}{1-\epsilon}\right] + \delta_g(2)\left[\frac{V_k - X_k}{\epsilon} - \frac{X_k}{1-\epsilon}\right] + \frac{a_\epsilon}{N\epsilon} - \frac{b_\epsilon}{N(1-\epsilon)} \tag{32}$$

$$\frac{\partial}{\partial\gamma_i}\log h_{gk} = \mathbf{w}_k\left[I(g=i) - p_{ik}\right], \tag{33}$$

where $\delta_g(i) = 1$ if $g = i$ and $\delta_g(i) = 0$ if $g \neq i$. It is straightforward to calculate V from the above first derivatives. The second term $\frac{\partial^2}{\partial\theta\partial\theta^T}\log h_{gk}$ of R has components as

$$\frac{\partial^2}{\partial\beta^2}\log h_{gk} = -s_g^2\pi_k(1-\pi_k) \tag{34}$$

$$\frac{\partial^2}{\partial\beta_w\partial\beta_w^T}\log h_{gk} = -\mathbf{w}_k\mathbf{w}_k^T\pi_k(1-\pi_k) \tag{35}$$

$$\frac{\partial^2}{\partial\beta_w\partial\beta}\log h_{gk} = -\mathbf{w}_k s_g\pi_k(1-\pi_k) \tag{36}$$

$$\frac{\partial^2}{\partial\gamma_i\partial\gamma_i^T}\log h_{gk} = -\mathbf{w}_k\mathbf{w}_k^T p_{ik}(1-p_{ik}) \tag{37}$$

$$\frac{\partial^2}{\partial\gamma_i\partial\gamma_{3-i}^T}\log h_{gk} = \mathbf{w}_k\mathbf{w}_k^T p_{ik}p_{3-i,k} \tag{38}$$

$$\frac{\partial^2}{\partial\epsilon^2}\log h_{gk} = -\left(\delta_g(0)\left[\frac{X_k}{\epsilon^2}+\frac{V_k-X_k}{(1-\epsilon)^2}\right]+\delta_g(2)\left[\frac{V_k-X_k}{\epsilon^2}+\frac{X_k}{(1-\epsilon)^2}\right] +\frac{a_\epsilon}{N\epsilon^2}+\frac{b_\epsilon}{N(1-\epsilon)^2}\right), \tag{39}$$

where $i = 1$ or 2. All other second derivatives that are not presented are equal to zero. Using these first and second derivatives of $\log h_{gk}$, we can obtain the components of the difference matrix R as follows:

$$R_{\beta\beta} = \frac{1}{N}\sum_{k=1}^{N}\sum_{g=0}^{2}\tau_{gk}s_g^2\left[(y_k-\pi_k)^2-\pi_k(1-\pi_k)\right] \tag{40}$$

$$R_{\beta_w\beta} = \frac{1}{N}\sum_{k=1}^{N}\sum_{g=0}^{2}\tau_{gk}s_g\mathbf{w}_k\left[(y_k-\pi_k)^2-\pi_k(1-\pi_k)\right] \tag{41}$$

$$R_{\beta_w\beta_w} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_k\mathbf{w}_k^T\left[(y_k-\pi_k)^2-\pi_k(1-\pi_k)\right] \tag{42}$$

$$\begin{aligned}R_{\epsilon\epsilon} = \frac{1}{N}\sum_{k=1}^{N}\Bigg(&\tau_{0k}\left[\frac{X_k+a_\epsilon/N}{\epsilon}-\frac{V_k-X_k+b_\epsilon/N}{1-\epsilon}\right]^2+\tau_{1k}\left[\frac{a_\epsilon}{N\epsilon}-\frac{b_\epsilon}{N(1-\epsilon)}\right]^2\\ &+\tau_{2k}\left[\frac{V_k-X_k+a_\epsilon/N}{\epsilon}-\frac{X_k+b_\epsilon/N}{1-\epsilon}\right]^2-\tau_{0k}\left[\frac{X_k+a_\epsilon/N}{\epsilon^2}+\frac{V_k-X_k+b_\epsilon/N}{(1-\epsilon)^2}\right]\\ &-\tau_{1k}\left[\frac{a_\epsilon}{N\epsilon^2}+\frac{b_\epsilon}{N(1-\epsilon)^2}\right]-\tau_{2k}\left[\frac{V_k-X_k+a_\epsilon/N}{\epsilon^2}+\frac{X_k+b_\epsilon/N}{(1-\epsilon)^2}\right]\Bigg)\end{aligned} \tag{43}$$

$$\begin{aligned}R_{\beta\epsilon} = \frac{1}{N}\sum_{k=1}^{N}[y_k-\pi_k]\Bigg(&\tau_{0k}s_0\left[\frac{X_k+a_\epsilon/N}{\epsilon}-\frac{V_k-X_k+b_\epsilon/N}{1-\epsilon}\right]+\tau_{1k}s_1\left[\frac{a_\epsilon}{N\epsilon}-\frac{b_\epsilon}{N(1-\epsilon)}\right]\\ &+\tau_{2k}s_2\left[\frac{V_k-X_k+a_\epsilon/N}{\epsilon}-\frac{X_k+b_\epsilon/N}{1-\epsilon}\right]\Bigg)\end{aligned} \tag{44}$$

$$\begin{aligned}R_{\beta_w\epsilon} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_k[y_k-\pi_k]\Bigg(&\tau_{0k}\left[\frac{X_k+\epsilon/N}{\epsilon}-\frac{V_k-X_k+b_\epsilon/N}{1-\epsilon}\right]+\tau_{1k}\left[\frac{a_\epsilon}{N\epsilon}-\frac{b_\epsilon}{N(1-\epsilon)}\right]\\ &+\tau_{2k}\left[\frac{V_k-X_k+a_\epsilon/N}{\epsilon}-\frac{X_k+b_\epsilon/N}{1-\epsilon}\right]\Bigg)\end{aligned} \tag{45}$$

$$R_{\gamma_i\gamma_i} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_k\mathbf{w}_k^T\left[(\tau_{ik}-p_{ik})(1-2p_{ik})\right] \tag{46}$$

$$R_{\gamma_1\gamma_2} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_k\mathbf{w}_k^T\left[p_{1k}(p_{2k}-\tau_{2k})+p_{2k}(p_{1k}-\tau_{1k})\right] \tag{47}$$

$$R_{\gamma_i\beta} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_k(y_k-\pi_k)\left[\tau_{ik}s_1-p_{ik}\sum_{g=0}^{2}\tau_{gk}s_g\right] \tag{48}$$

$$R_{\gamma_i\beta_w} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_k\mathbf{w}_k^T(y_k-\pi_k)\left[\tau_{ik}-p_{ik}\right] \tag{49}$$

$$\begin{aligned}R_{\gamma_1\epsilon} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_k\Bigg(&-p_{1k}\tau_{0k}\left[\frac{X_k+a_\epsilon/N}{\epsilon}-\frac{V_k-X_k+b_\epsilon/N}{1-\epsilon}\right]\\ &+(1-p_{1k})\tau_{1k}\left[\frac{a_\epsilon}{N\epsilon}-\frac{b_\epsilon}{N(1-\epsilon)}\right]\\ &-p_{1k}\tau_{2k}\left[\frac{V_k-X_k+a_\epsilon/N}{\epsilon}-\frac{X_k+b_\epsilon/N}{1-\epsilon}\right]\Bigg)\end{aligned} \tag{50}$$

$$R_{\gamma_2\epsilon} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{w}_k\left(-p_{2k}\tau_{0k}\left[\frac{X_k+a_\epsilon/N}{\epsilon}-\frac{V_k-X_k+b_\epsilon/N}{1-\epsilon}\right]-p_{2k}\tau_{1k}\left[\frac{a_\epsilon}{N\epsilon}-\frac{b_\epsilon}{N(1-\epsilon)}\right]\right.$$
$$\left.+(1-p_{2k})\tau_{2k}\left[\frac{V_k-X_k+\epsilon/N}{\epsilon}-\frac{X_k+b_\epsilon/N}{1-\epsilon}\right]\right) \quad (51)$$

Therefore, our proposed robust score test statistic Z_R can be written as

$$Z_R = \frac{S}{\sqrt{Nv_s}}, \quad (52)$$

which asymptotically has a standard normal distribution under H_0.

Another common approach to obtain p-values is to use Monte Carlo permutation method based on the score vector or function. However, the Monte Carlo permutation p-value calculation given a very small Bonferroni's corrected level of significance needs high computational expenses since it requires at least 10^7 or 10^8 permuted resamples. In this work, we employ the asymptotic permutation p-value calculation. The score function is given by

$$S = \sum_{k=1}^{N}\sum_{i=0}^{2}\tau_{ik(0)}s_i\left(y_k-\frac{e^{\beta^T_{w(0)}\mathbf{w}_k}}{1+e^{\beta^T_{w(0)}\mathbf{w}_k}}\right)$$
$$= \sum_{k=1}^{N} r_k e_k \quad (53)$$

where the subscript (0) denotes the estimated parameter under the null hypothesis. We define a score $r_k = \sum_{i=0}^{2}\tau_{ik(0)}s_i$ associated with subject k and the kth residual $e_k = \left(y_k-\frac{e^{\beta^T_{w(0)}\mathbf{w}_k}}{1+e^{\beta^T_{w(0)}\mathbf{w}_k}}\right)$. We can permute the residuals e_k's to calculate the permutation p-value for adjusting covariate effects. The asymptotic permutation test statistic Z_{AP} for a large sample size is given by

$$Z_{AP} = \frac{S-N\cdot\bar{r}\cdot\bar{e}}{\sqrt{\frac{1}{N-1}\left[\sum_{i=1}^{N}e_i^2-N(\bar{e})^2\right]\left[\sum_{i=1}^{N}r_i^2-N(\bar{r})^2\right]}} \quad (54)$$

where $\bar{r}=\frac{1}{N}\sum_{i=1}^{N}r_i$ and $\bar{e}=\frac{1}{N}\sum_{i=1}^{N}e_i$. The simple linear rank test statistic Z_{AP} asymptotically has a standard normal distribution under the null hypothesis [28].

3. Results

3.1. Simulation Study

In this section, we simulate data from the following process:

$$P(Y=1|w)f(w) = \sum_{i=0}^{2}P(Y=1|G=i,w)P(G=i)f(w) \quad (55)$$

For simplicity, we assume genetic relative risk $R_i=\frac{P(Y=1|G=i,w)}{P(Y=1|G=0,w)}$, for $i=1,2$, does not depend on the covariate W. We assume that the genotype frequency $\pi_i=P(G=i)$ satisfies Hardy–Weinberg equilibrium (HWE), so that $P(G=0)=p^2$, $P(G=1)=2pq$, and $P(G=2)=q^2$, where q is the minor allele frequency. Then, the prevalence is given by

$$\phi = \int P(Y=1|w)f(w)dw$$
$$= \int\left[p^2f(w|G=0)+2pqR_1f(w|G=1)+q^2R_2f(w|G=2)\right]P(Y=1|G=0,w)dw \quad (56)$$

We consider two scenarios when generating covariates w: (1) $f(w|G=i)$ is equal to a standard normal $N(0,1)$ for all $i=0,1,2$, called by a single normal, and (2) $f(w|G=i)$ has a normal distribution with mean μ_i and standard deviation $\sigma=1$, we call this a normal mixture. For the single normal model,

$$\phi \quad = \quad \left[p^2+2pqR_1+q^2R_2\right]\int P(Y=1|G=0,w)f(w)dw \tag{57}$$

We finally assume $P(Y=1|G=0,w)=\frac{e^{\alpha+\beta_w w}}{1+e^{\alpha+\beta_w w}}$. During the simulation study, we compute α by numerical integration given prevalence ϕ and other parameters.

3.1.1. Simulation Study for Null Distribution

To evaluate the type I error rate of the proposed test statistic, we perform simulations with 5000 replicates per each parameter setting. We fixed the proportion of cases as 0.5. The parameter settings that we consider are:

(i) Prevalence (ϕ): 0.1, 0.3
(ii) Coverage (v): 4, 30
(iii) Minor allele frequency (q): 0.05, 0.3
(iv) Total sample size (n): 500, 1000, 1500
(v) Covariate (w_1): single normal or normal mixture with mean $\mu=(0,\frac{1}{2},\frac{1}{2})$ given genotype $(0,1,2)$
(vi) Regression coefficient β_w: 0, 1

We consider prevalence $\phi=0.3$ that may be large in a genetic association study. It is chosen to reflect pharmacogenomics data that we use in the real data analysis.

Figure 1 shows boxplots of the null simulations. The permutation method appears to have more variability of the empirical rejection rates over different configurations and to have the smaller empirical rejection rates compared to the proposed robust score test based on the sandwich variance estimator. When the sample size was small as 500 and the coverage was 4×, the permutation-based test had less than 2.5% rejection rate though the desired value is 5%. The smallest empirical rejection rate for the proposed robust test was greater than 3.5%, and it appears the empirical rejection rates become closer to 5% as the sample size increases. If the coverage is 30× or higher, then the estimated posterior probabilities in our approach are close to zero-or-one and most inferred genotypes are quite clear. When the coverage was 30×, our proposed test seems to well control the type I error rates regardless of other parameter settings as expected. Table 1 shows the empirical rejection rates under the null settings by combining our simulation results for the lower level of significance.

Table 1. Empirical rejection rates under null settings for level $1\times10^{-2}, 1\times10^{-3}, 1\times10^{-4}$, and 1×10^{-5}.

Method (cvrg)	$\mathbf{1\times10^{-2}}$	$\mathbf{1\times10^{-3}}$	$\mathbf{1\times10^{-4}}$	$\mathbf{1\times10^{-5}}$
Permutation (4×)	7.13×10^{-3}	6.14×10^{-4}	4.44×10^{-5}	0
Permutation (30×)	7.73×10^{-3}	7.08×10^{-4}	6.11×10^{-5}	5.56×10^{-6}
Sandwich (4×)	8.34×10^{-3}	6.89×10^{-4}	4.44×10^{-5}	5.56×10^{-6}
Sandwich (30×)	1.02×10^{-2}	1.01×10^{-3}	8.75×10^{-5}	8.33×10^{-6}

3.1.2. Simulation Study for Statistical Power

We used the same parameter settings as in the null simulation study. Additionally, we set multiplicative genetic relative risks vector $(1, 1.5, 1.5^2)$ in the alternative parameter configurations. In the alternative simulations, we calculated empirical rejection rates under Bonferroni corrected level of significance, that is, 5×10^{-8}. Figure 2 shows the boxplots of empirical power under various alternative settings. We removed the results when the sample size was 500 or the minor allele frequency was 0.05 since all the rejection rates were small in Figure 2. It appears interesting that the power of the

proposed test when the coverage was 4× and the sample size was 1500 is higher than the power of the test when the coverage was 30× and the sample size was 1000. If the two design costs are similar, then the low-coverage with more samples seems more effective than the high-coverage with less samples.

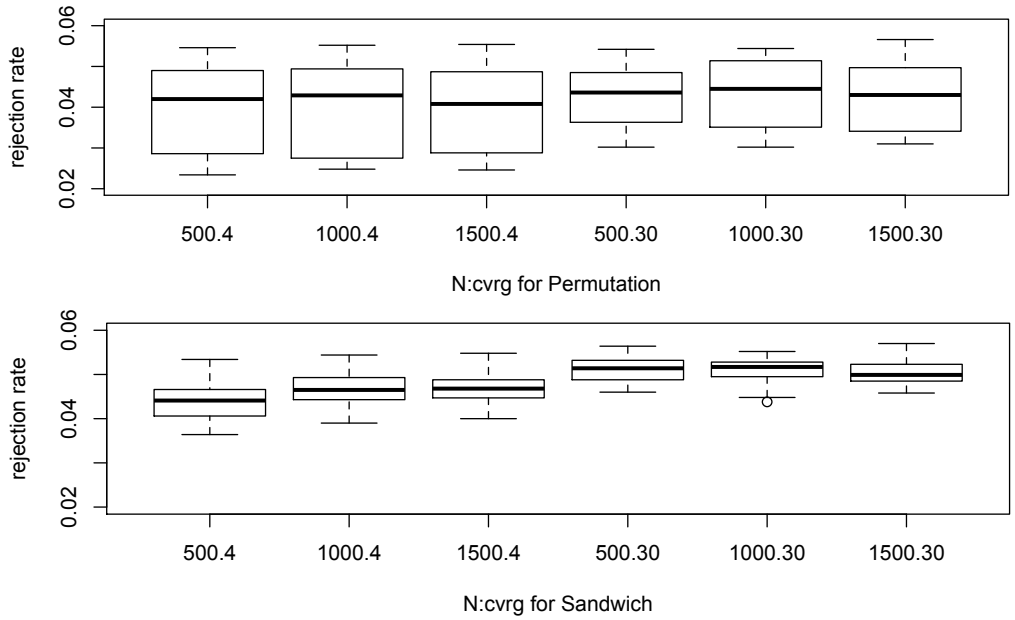

Figure 1. Boxplot of the empirical rejection rates under the null hypothesis.

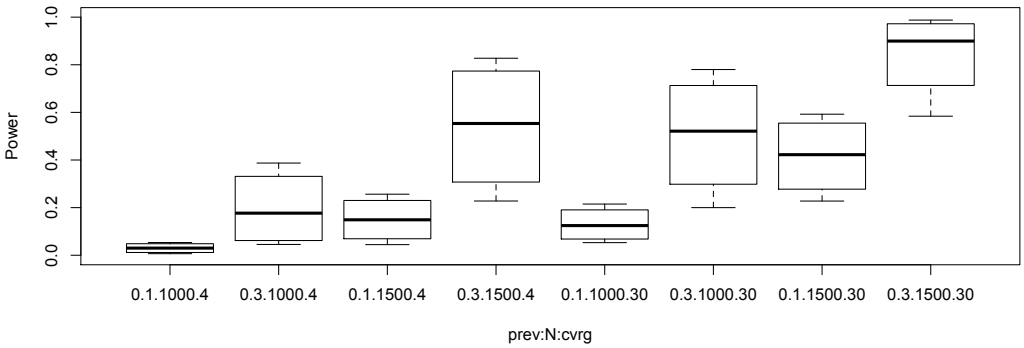

Figure 2. Boxplots of statistical power of the proposed robust test under the alternative settings. The level of significance was set as 5×10^{-8}. The notation 0.1.1000.4 represents prevalence 0.1, total sample size 1000, and coverage 4×.

Table 2 summarizes statistical power of our proposed method and a naive approach. The naive approach uses uncertain genotypes by the maximum posterior probability classification rule [29]. The standard logistic regression was applied to the uncertain genotypes. As expected, the proposed robust method shows higher power than the naive approach when the sequencing coverage is as low as 4×. When the sequencing coverage is high as 30×, two approaches show similar performance in terms of statistical power.

Table 2. Empirical rejection rates under alternative hypothesis. The level of significance was set as 5×10^{-8}.

Coverage	Total Sample Size	Covariate	β_w	Naive	Proposed
4	1000	Normal mixture	0	0.102	0.113
4	1000	Normal mixture	1	0.233	0.261
4	1000	Single normal	0	0.190	0.277
4	1000	Single normal	1	0.269	0.374
4	1500	Normal mixture	0	0.398	0.429
4	1500	Normal mixture	1	0.657	0.701
4	1500	Single normal	0	0.626	0.741
4	1500	Single normal	1	0.736	0.840
30	1000	Normal mixture	0	0.384	0.355
30	1000	Normal mixture	1	0.617	0.603
30	1000	Single normal	0	0.622	0.637
30	1000	Single normal	1	0.734	0.760
30	1500	Normal mixture	0	0.792	0.761
30	1500	Normal mixture	1	0.959	0.954
30	1500	Single normal	0	0.933	0.939
30	1500	Single normal	1	0.978	0.978

3.2. Real Data Analysis

The proposed robust generalized score test was applied to the pharmacogenomics data consisting of 400 epilepsy patients [22]. The data were collected from several epilepsy clinics in Korea and were genotyped for whole-exomes by NGS experiments [30]. All study participants followed the criteria in [31] if the participants had drug-resistant (case group) or drug-responsive (control group) epilepsy. We defined the drug resistance as the occurrence of at least four unprovoked seizures during the past one year at the time of recruitment, with trials of two or more appropriate antiepileptic drugs (AEDs) at maximal tolerated doses. Patients who underwent surgical treatment for drug-resistant epilepsy were classified as having drug-resistant epilepsy, regardless of the surgical outcome. We excluded some patients from the study if they were frequently in poor compliance with AED therapy and had reported seizures with a questionable semiology. In addition, we defined the drug responsiveness as complete freedom from seizures for at least one year up to the date of the last follow-up visit.

We included two non-genetic covariates in our association analysis. The two covariates were age of patient and duration of epileptic seizures. The age variable was definitely independent of genetic information, whereas duration variable may be associated with genetic variables. Due to the relatively small sample size 400, we did not expect to find a significantly associated SNP controlling for the two covariates. Therefore, instead of reporting a genome-wide association study, we illustrated the results of a SNP with low read depths and a SNP with high read depths. For the low read depths example, we selected a SNP from chromosome 1, which is *rs*3811406. The distribution of read depths for the SNP was summarized in Table 3. More than 10% of the sample had five or less read depths and more than 30% of the sample had 10 or less read depths at the SNP. When applying our proposed mixture-based association test, the test statistic value was $z_R = 2.864$ and the p-value was $p = 4.183 \times 10^{-3}$, while the standard logistic regression analysis using pooled genotype calls had $z = 2.601$ and the p-value $p = 9.30 \times 10^{-3}$ that was more than twice the p-value of the proposed robust test.

Table 3. Distribution of read depths at *rs*3811406.

Read Depth v	$v \leq 5$	$5 < v \leq 10$	$10 < v \leq 30$	$v > 30$	Total
Frequency	43	86	95	176	400
Proportion	0.1075	0.215	0.2375	0.44	1

In addition, we applied our proposed test to SNP *rs*4915154 at which all patients had 13× or higher read depths and 85% patients had 25× or higher read depths. For this SNP, the proposed robust test statistic was $z_R = 2.940$ with p-value $= 3.28 \times 10^{-3}$ and the multiple logistic regression with the pooled genotype calls reported $z = 2.963$ with p-value $= 3.05 \times 10^{-3}$. The two results were quite close, as expected, due to high read depths at the SNP.

4. Discussion and Conclusions

In the present study, we developed the mixture-based genetic association tests adjusting the effects of non-genetic covariates in low-coverage NGS data. In order to construct a robust test statistic under model misspecification, we derived the sandwich variance estimator of the mixture model. The proposed test statistic is calculated from allele read counts and read depths instead of inferred genotypes so that we can apply this association test to low-coverage NGS data controlling for non-genetic covariates without external imputation or elimination of uncertain genotypes. Another important issue that we addressed in the present study is that the proposed test takes account of potential dependence between latent genotypes and the non-genetic covariates. Regarding computational cost, our proposed method is efficient because it is a generalized score test that uses the estimates of the parameters only under the null hypothesis of no association. When the sequencing depth is 4×, it takes around 1.2 s for sample size 500, 4 s for sample size 1000, and 9 s for sample size 1500 to simulate a dataset and to calculate both test statistics Z_{AP} and Z_R. When the sequencing depth is 30×, it takes approximately 0.13 s for sample size 500, 0.3 s for sample size 1000, and 0.53 s for sample size 1500. Time for these computations is measured based on a single core work of a 3.5 GHz Intel Xeon processor. As illustrated in the real data analysis section, the read depth is not a fixed constant. Therefore, the computational time for real data is usually less than that for the coverage 4× simulation setting. We used statistical software R, which is known to be slow. It would be computationally beneficial to run our proposed methods in other faster program languages for a high-dimensional genome-wide association study.

We applied the penalized likelihood method to avoid singularity of information matrix when calculating the proposed score test statistic. Therefore, the penalty term is not necessary for a non-zero estimate of the error parameter. During our work, we fixed the degree of penalization $C = 1$, $a_\epsilon = 0.01$, and $b_\epsilon = 0.99$ that implies 1% of allele read error as prior information. This parameter choice does not affect the proposed test statistic much since the likelihood function is merely changed when the sample size is greater than 500. It may be of interest to find optimal values for the parameters of the penalty term.

The simulation study confirms that the type I error rates of the proposed test are well controlled under the various parameter settings. The proposed robust test appears to perform better than the permutation based approach. Simulation results indicate that coverage 4× with sample size 1500 shows higher power as compared to coverage 30× with sample size 1000. Our method can be applied to an NGS experimental design by simulations to select coverage and sample size given a fixed amount of budget.

We presented a real data example in which the proposed test and multiple logistic regression results are similar to one another if the sequencing depth is high, whereas the test results may differ when the sequencing depth is low. This might have been caused because the proposed test is an extension of the multiple logistic regression with the unobserved latent genotype predictor. If the sequencing depth is high enough to call accurate genotypes, then our probability model becomes identical to the probability model of the multiple logistic regression. It would be more beneficial to compare with the previous methods by evaluating our proposed methods using a larger sized public dataset.

In this work, we focused on a single variant association test while controlling covariates. By adopting a multivariate mixture model, the proposed method can be extended to the multi-variant

genetic association test including covariates. We can also extend the present method to differential genotype misclassifications.

Author Contributions: Conceptualization, W.K.; methodology, J.Y.L. and W.K.; software, J.Y.L. and W.K.; formal analysis, J.Y.L. and W.K.; data curation, M.-K.K. and W.K.; writing—original draft preparation, J.Y.L. and W.K.; writing—review and editing, J.Y.L., M.-K.K., and W.K.; project administration, W.K.; funding acquisition, M.-K.K. and W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07050012) and was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (HI15C1559).

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

Abbreviations

The following abbreviations are used in this manuscript:

EM	Expectation–Maximization
GWAS	Genome-wide association study
HWE	Hardy–Weinberg equilibrium
maf	Minor allele frequency
NGS	Next-generation sequence
SNP	Single nucleotide polymorphism
TDT	Transmission disequilibrium test

References

1. Wu, M.C.; 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*, 82–93. [CrossRef]
2. Cirulli, E.T.; White, S.; Read, R.W.; Elhanan, G.; Metcalf, W.J.; Tanudjaja, F.; Fath, D.M.; Sandoval, E.; Isaksson, M.; Schlauch, K.A.; et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat. Commun.* **2020**, *11*, 542. [CrossRef]
3. Lakiotaki, K.; Kanterakis, A.; Kartsaki, E.; Katsila, T.; Patrinos, G.P.; Potamias, G. Exploring public genomics data for population pharmacogenomics. *PLoS ONE* **2017**, *12*, e0182138. [CrossRef] [PubMed]
4. Patrinos, G.P.; Giannopoulou, E.; Katsila, T.; Tsermpini, E.E.; Mitropoulou, C. Integrating next-generation sequencing in the clinical pharmacogenomics workflow. *Front. Pharmacol.* **2019**, *10*, 384.
5. Celesti, F.; Celesti, A.; Wan, J.; Villari, M. Why Deep Learning Is Changing the Way to Approach NGS Data Processing: A Review. *IEEE Rev. Biomed. Eng.* **2018**, *11*, 68–76. [CrossRef]
6. Le, N.Q.K.; Yapp, E.K.Y.; Nagasundaram, N.; Chua, M.C.H.; Yeh, H.Y. Computational identification of vesicular transport proteins from sequences using deep gated recurrent units architecture. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 1245–1254. [CrossRef]
7. Tripathi, R.; Sharma, P.; Chakraborty, P.; Varadwaj, P.K. Next-generation sequencing revolution through big data analytics. *Front. Life Sci.* **2016**, *9*, 119–149. [CrossRef]
8. Cirillo, D.; Valencia, A. Big data analytics for personalized medicine. *Curr. Opin. Biotechnol.* **2019**, *58*, 161–167. [CrossRef]
9. Sims, D.; Sudbery, I.; Ilott, N.E.; Heger, A.; Ponting, C.P. Sequencing depth and coverage: Key considerations in genomic analyses. *Nat. Rev. Genet.* **2014**, *15*, 121–132. [CrossRef]
10. Song, K.; Li, L.; Zhang, G. Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. *Sci. Rep.* **2016**, *6*, 35736. [CrossRef]
11. Gordon, D.; Finch, S.J.; Nothnagel, M.; Ott, J. Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms. *Hum. Hered.* **2002**, *54*, 22–33. [CrossRef]
12. Ahn, K.; Haynes, C.; Kim, W.; Fleur, R.S.; Gordon, D.; Finch, S.J. The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. *Ann. Hum. Genet.* **2007**, *71*, 249–261. [CrossRef]

13. Kim, W.; Londono, D.; Zhou, L.; Xing, J.; Nato, A.Q.; Musolf, A.; Matise, T.C.; Finch, S.J.; Gordon, D. Single-variant and multi-variant trend tests for genetic association with next-generation sequencing that are robust to sequencing error. *Hum. Hered.* **2012**, *74*, 172–183. [CrossRef]
14. Hou, L.; Sun, N.; Mane, S.; Sayward, F.; Rajeevan, N.; Cheung, K.; Cho, K.; Pyarajan, S.; Aslan, M.; Miller, P. Impact of genotyping errors on statistical power of association tests in genomic analyses: A case study. *Genet. Epidemiol.* **2017**, *41*, 152–162. [CrossRef]
15. Consortium, .G.P. An integrated map of genetic variation from 1092 human genomes. *Nature* **2012**, *491*, 56.
16. Le, S.Q.; Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* **2011**, *21*, 952–960. [CrossRef]
17. Li, Y.; Sidore, C.; Kang, H.M.; Boehnke, M.; Abecasis, G.R. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res.* **2011**, *21*, 940–951. [CrossRef]
18. Kim, W.; Gordon, D.; Sebat, J.; Kenny, Q.Y.; Finch, S.J. Computing power and sample size for case-control association studies with copy number polymorphism: Application of mixture-based likelihood ratio test. *PLoS ONE* **2008**, *3*, e3475. [CrossRef]
19. Barnes, C.; Plagnol, V.; Fitzgerald, T.; Redon, R.; Marchini, J.; Clayton, D.; Hurles, M.E. A robust statistical method for case-control association testing with copy number variation. *Nat. Genet.* **2008**, *40*, 1245. [CrossRef]
20. Kim, S.Y.; Li, Y.; Guo, Y.; Li, R.; Holmkvist, J.; Hansen, T.; Pedersen, O.; Wang, J.; Nielsen, R. Design of association studies with pooled or un-pooled next-generation sequencing data. *Genet. Epidemiol.* **2010**, *34*, 479–491. [CrossRef]
21. Gordon, D.; Finch, S.J.; De La Vega, F. A new expectation-maximization statistical test for case-control association studies considering rare variants obtained by high-throughput sequencing. *Hum. Hered.* **2011**, *71*, 113–125. [CrossRef]
22. Kim, W.; Kim, Y.H. Genetic association tests when a nuisance parameter is not identifiable under no association. *Commun. Stat. Appl. Methods* **2017**, *24*, 663–671. [CrossRef]
23. Kim, W. Transmission Disequilibrium Tests Based on Read Counts for Low-Coverage Next,-Generation Sequence Data. *Hum. Hered.* **2015**, *80*, 36–49. [CrossRef]
24. Chen, H.; Chen, J.; Kalbfleisch, J.D. A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Stat. Soc. Ser. B* **2001**, *63*, 19–29. [CrossRef]
25. Zhou, H.; Pan, W. Binomial mixture model-based association tests under genetic heterogeneity. *Ann. Hum. Genet.* **2009**, *73*, 614–630. [CrossRef]
26. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–22.
27. White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **1982**, *50*, 1–25. [CrossRef]
28. Sidak, Z.; Sen, P.K.; Hajek, J. *Theory of Rank Tests*; Academic Press: San Diego, CA, USA, 1999.
29. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*; Wiley: New York, NY, USA, 1962.
30. Kang, K.W.; Kim, W.; Cho, Y.W.; Lee, S.K.; Jung, K.Y.; Shin, W.; Kim, D.W.; Kim, W.J.; Lee, H.W.; Kim, W. Genetic characteristics of non-familial epilepsy. *PeerJ* **2019**, *7*, e8278. [CrossRef]
31. Kim, M.-K.K.; Moore, J.H.; Kim, J.K.; Cho, K.H.; Cho, Y.W.; Kim, Y.S.; Lee, M.C.; Kim, Y.O.; Shin, M.H. Evidence for epistatic interactions in antiepileptic drug resistance. *J. Hum. Genet.* **2011**, *56*, 71–76. [CrossRef]