6-2-2023

Next-Generation Sequencing Data-Based Association Testing of a Group of Genetic Markers for Complex Responses Using a Generalized Linear Model Framework

Zheng Xu

Song Yan

Cong Wu

Qing Duan

Sixia Chen

See next page for additional authors

Authors

Zheng Xu, Song Yan, Cong Wu, Qing Duan, Sixia Chen, and Yun Li

mathematics

MDPI

Article

Next-Generation Sequencing Data-Based Association Testing of a Group of Genetic Markers for Complex Responses Using a Generalized Linear Model Framework

Zheng Xu [1,*], Song Yan [2,3,4,†], Cong Wu [5], Qing Duan [2,3,6], Sixia Chen [7] and Yun Li [2,3,4,*]

1 Department of Mathematics and Statistics, Wright State University, Dayton, OH 45324, USA
2 Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
3 Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
4 Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
5 Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68508, USA
6 Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA
7 Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA
* Correspondence: zheng.xu@wright.edu (Z.X.); yunli@med.unc.edu (Y.L.); Tel.: +1-937-775-2103 (Z.X.); +1-919-843-2832 (Y.L.)
† Deceased author.

Abstract: To study the relationship between genetic variants and phenotypes, association testing is adopted; however, most association studies are conducted by genotype-based testing. Testing methods based on next-generation sequencing (NGS) data without genotype calling demonstrate an advantage over testing methods based on genotypes in the scenarios when genotype estimation is not accurate. Our objective was to develop NGS data-based methods for association studies to fill the gap in the literature. Single-variant testing methods based on NGS data have been proposed, including our previously proposed single-variant NGS data-based testing method, i.e., UNC combo method. The NGS data-based group testing method has been proposed by us using a linear model framework which can handle continuous responses. In this paper, we extend our linear model-based framework to a generalized linear model-based framework so that the methods can handle other types of responses especially binary responses which is a common problem in association studies. To evaluate the performance of various estimators and compare them we performed simulation studies. We found that all methods have Type I errors controlled, and our NGS data-based methods have better performance than genotype-based methods for other types of responses, including binary responses (logistics regression) and count responses (Poisson regression), especially when sequencing depth is low. We have extended our previous linear model (LM) framework to a generalized linear model (GLM) framework and derived NGS data-based methods for a group of genetic variables. Compared with our previously proposed LM-based methods, the new GLM-based methods can handle more complex responses (for example, binary responses and count responses) in addition to continuous responses. Our methods have filled the literature gap and shown advantage over their corresponding genotype-based methods in the literature.

Keywords: next-generation sequencing; association testing; generalized linear model; joint significance test; variable collapse test; genotype calling; score test; group testing; rare variant

MSC: 62P10; 62J12; 92B15

Citation: Xu, Z.; Yan, S.; Wu, C.; Duan, Q.; Chen, S.; Li, Y. Next-Generation Sequencing Data-Based Association Testing of a Group of Genetic Markers for Complex Responses Using a Generalized Linear Model Framework. *Mathematics* **2023**, *11*, 2560. https://doi.org/10.3390/math11112560

Academic Editor: Sorana D. Bolboacă

Received: 14 April 2023
Revised: 29 May 2023
Accepted: 30 May 2023
Published: 2 June 2023

1. Introduction

Next-generation sequencing (NGS) is a massively parallel sequencing technology used to determine the order of nucleotides in entire genomes or targeted regions of de-

oxyribonucleic acid (DNA) or ribonucleic acid (RNA) which offers ultra-high throughput, scalability, and speed [1]. With fast development in NGS technology, it becomes more and more widely-used in association studies of genetic variables. Compared with traditional sequencing technologies, such as Sanger technology [2], NGS technologies can achieve higher sequencing throughput and lower costs. Enormous amounts of NGS data are collected from NGS platforms (for example, MiniSeq, MiSeq, NextSeq 1000 and NovaSeq 6000 from Illumina) by researchers and these researchers own the datasets and conduct genetic studies based on these NGS datasets [3].

Next-generation sequencing (NGS) data in the format of raw sequencing reads are collected in NGS platforms [4–6]. These platforms typically do not provide genotype data. Researchers have proposed multi-step bio-informatics data-processing pipelines to obtain genotypes based on NGS data. A typical pipeline usually includes quality control (QC), alignment of sequences, variant calling, and genotype calling (GC) [5,7,8]. Genotype calling is the process of determining the genotype for each individual and is typically only performed for positions in which a SNP or a 'variant' has already been called [9]. After obtaining estimated genotypes, researchers conduct regression analysis to study the relationship between phenotype and genotype, as well as other variables (environmental variables, clinical variables, etc.) [9–11].

Regression methods have been used in statistics and bio-statistics including kernel regression [12], spline smoother [13], Alternative Conditional Expectations (ACE) [14], and Additivity Variance Stabilization (AVAS) [15]. Suppose $X \in \mathcal{R}^m$ and $Y \in \mathcal{R}$, and there is a relationship $m(.)$ between X and Y via the model $Y = m(X) + e$, where the error term e has zero mean conditional on X, i.e., $E(e|X) = 0$. Kernel regression can estimate the function $m(.)$ non-parametrically using kernel functions [12]. The spline smoother estimates the relationship between X and Y, i.e., $m(.)$, using splines. That is, the spline smoother assumes that $m(.)$ can be approximated by splines, and then it finds the best spline m as the (spline estimator) of $m(.)$ [13]. Alternating conditional expectations (ACE) is a non-parametric method to find the optimal non-linear transformations of the response variable Y and its predictor variables X's to minimize the fraction of variance in transformed Y not explained by transformed X's assuming an additive model. In mathematics, let $X_1, X_2, \ldots, X_p, Y$ be random variables. Suppose $\theta(Y)$, $\phi_1(X_1)$,$\phi_2(X_2), \ldots, \phi_p(X_p)$ are zero-mean functions. The fraction of variance in transformed Y not explained by transformed X based on additive models are $E[\theta(Y) - \sum_{i=1}^{p} \phi_i(X_i)]^2 / E[\theta^2(Y)]$. ACE is a non-parametric method to find optimal transformations $(\theta, \phi_1, \ldots, \phi_p)$ to minimize this fraction [14]. The Additivity Variance Stabilization (AVAS) method is an improved method over ACE which also aimed to find optimal transformation to maximize the fraction in transformed Y explained by transformed Xs, and AVAS has better performance than ACE when correlation between transformed Xs and transformed Y is small [15]. An application example of the regression algorithm is arterial volume-weighted arterial spin tagging (AVAST), which is a variant of a pseudo-continuous arterial spin labeling acquisition (PCASL) technique to measure the arterial cerebral blood volume (aCBV) and provides useful information about neuronal activation based on functional magnetic resonance imaging (fMRI) brain data [16]. Regression methods based on a linear model (LM) and a generalized linear model (GLM) are widely used for association studies. Responses are phenotypes. Genotypes and other variables, including environmental variables and behavior variables, are explanatory variables/predictors. Depending on different types of responses, different regression models can be adopted. For example, bio-statisticians typically conduct logistics regression and linear regression, respectively, for binary responses and continuous responses. If the response is a count/integer type, a Poisson regression can be adopted. To handle complex responses (various types), the framework of a generalized linear model (GLM) can be adopted, which is better than the linear model (LM) framework because LM can only handle continuous responses as we proposed before [17]. This motivates us to extend our previous linear model framework [17] to a general linear model framework in this article.

Testing methods are different for common variants and rare variants. Common variants refer to genetic variables with minor allele frequency (MAF) greater than a threshold value c, $0.01 \leq c \leq 0.05$ [18,19]. Researchers typically set $c = 0.05$. Rare variants refer

to those with MAF less than c. For common genetic variants, single-variant testing is conducted in association studies. The genetic effect can be specified by an additive model, dominant model, or recessive model [9,20]. Genome-wide association study (GWAS) means to repeat the single-variant testing for all markers genomewide [9,20]. For common variants, Chi-square test or F tests are adopted to test for the joint significance of a group of variables. Markers within a gene can be treated as a group and tested simultaneously. Genome-wide gene-based group testing means repeat the single-gene joint-significance testing one-by-one for all genes in the whole genome.

Because rare variants have a small MAF, say $MAF < 0.05$, they have low variations in genotypes so that the power of testing for a variant may not be enough [18,21]. Rare-variant testing is often a group test instead of single variable test. There are various testing methods for rare variants in the literature. Among these tests, two big categories are often adopted. Category 1 refers to variable collapsing (VC) testing methods. These methods first calculate one variable based on multiple genetic variants, i.e., collapse/merge multiple variants into one, and then conduct association studies using the calculated variable [22,23]. Burden test is a representative method. It conducts association studies between the phenotype and the total number of rare alleles in a group of markers [23,24]. Category 2 contains different versions of Sequence Kernel Association testing (SKAT) methods, including SKAT, MK-SKAT, SKAT-O, and BESKAT [18,22,25,26]. The SKAT method is a representative method in Category 2 and it adopts a linear-model framework (linear regression) to deal with continuous phenotypes, and a logistics-model framework (logistics regression) to deal with binary phenotypes [26]. Continuous type and binary type are two mostly encountered types of phenotypes in association studies, thus this article focuses on these two types; both categories are genotype-based tests. Genotypes are estimated and association studies are performed between phenotypes and estimated genotypes.

There are estimation errors in genotype calling. Genotype accuracy are influenced by a range of factors, including sequencing errors, alignment accuracy, and sequencing depth. When sequencing depth is low, genotype calling can be very imprecise, which can influence the performance of association methods based on estimated genotypes [27,28]. To improve the performance of testing, NGS data-based methods with no genotype calling are recommended. These methods model NGS data directly without genotype calling and have shown better performance [29,30].

Researchers have proposed a range of NGS data-based single-variant association testing methods without genotype calling [29–31], including our previously proposed UNC combo method [30]. When sequencing depth is low and genotype calling is not accurate, these NGS data-based single-variant methods can achieve better performance under the scenario of low sequencing depth, heterogeneous sequencing depths, and imprecise genotype calls [29–31]; however, there are no NGS data-based group testing methods in the literature except our previously proposed linear model (LM)-based group testing methods [17]. Being linear model-based, our previously proposed method can only handle continuous phenotypes. However, in the fields of bio-statistics and bio-informatics, especially association studies, other types of phenotypes, especially binary phenotypes (such as disease status and case/control association studies), are widely encountered. It is greatly desired and necessary to extend our methods to enable handling of other types of phenotypes, especially binary phenotypes. Thus, we extend our method from a linear model (LM)-based framework to a generalized linear model (GLM)-based NGS data-based group testing method so that our proposed methods can handle complex responses, including continuous responses (linear regression), binary responses (logistics regression), and count/integer responses (Poisson regression). The proposed NGS-based group testing methods are expected to have an advantage over genotype-based methods, especially when sequencing is low and genotype calling/estimation is not accurate.

Corresponding to genotype-based group testing methods for common variants (joint significance test) and rare variants (variable collapse test) [22,26,32], we fill the literature gap by proposing their corresponding NGS data-based methods without genotype calling.

That is, for a group of common variants, the joint significance test (JS) based on NGS data is proposed. For a group of rare variants, the variable collapse test (VC) based on NGS are proposed. Compared with our previous work [17], the major contribution of this work is that it can handle a range of types of phenotypes, including continuous, binary, and count/integer phenotypes based on the GLM framework, whereas our previous work [17] can only handle continuous phenotypes based on the LM framework.

In this study, we proposed novel NGS data-based testing methods for association studies based on a generalized linear model (GLM). The proposed methods fill the literature gap as the first NGS data-based group testing methods without genotype calling based on a GLM. We previously used the LM framework to develop association testing methods for a group of genetic variables based on NGS data [17]; however, our previous methods can only handle continuous responses. In this paper, we aimed to extend our linear model-based framework to a generalized linear model (GLM)-based framework so that our methods can handle other types of responses, especially binary responses, which is commonly-faced in association studies. The objectives of this study were to (1) develop our proposed novel NGS data-based testing methods for association studies based on a theoretical framework of generalized linear models (GLM) and (2) show our proposed methods can achieve better testing performance than their corresponding genotype-based methods.

2. Methodology

Denote the sample size of the study as N. For individual i ($1 \leq i \leq N$), the data are (y_i, g_i, x_i). The term y_i, g_i, and x_i, respectively, represent the phenotype, genotypes, and additional covariates. Examples of additional covariates are environmental variables, gender, and age. The genotype can only have values of 0, 1, or 2 for bi-allelic markers.

Suppose our group testing includes d_g genetic variants. We use a row vector to represent the genotypes for individual i, i.e., $g_i = (g_{i1}, g_{i2}, \ldots, g_{i_{d_g}})$, where the genotype at variant j for individual i is g_{ij}. We use the row vector $x_i = (1, x_{i1}, x_{i2}, \ldots, x_{id_x})$ represent the intercept and d_x additional covariates, where the value of additional variable j for individual i is denoted as x_{ij}.

Denote genotype matrix with size $N \times d_g$ to be $g = (g_1, g_2, \ldots, g_N)$. Denote response vector with length n to be $y = (y_1, y_2, \ldots, y_N)$. Denote the matrix with size $N \times (d_x + 1)$ for additional covariates to be $x = (x_1, x_2, \ldots, x_N)$.

2.1. Model Complex Phenotypes Using a GLM Framework

Both the linear model for a quantitative phenotype and the logistic regression model for a binary phenotype conform the framework of the general linear model [33,34]. This motivates us to model complex phenotypes by a generalized linear model (GLM) framework. The GLM model-based derivation is a direct extension of our previous linear model (LM)-based framework for testing of a group of variants [17], which is extended from work on NGS-based single-variant testing [29].

We model the complex phenotype by a GLM [33,34]. For the i-th individual, the probability of observing phenotype y_i is modelled to be

$$p(y_i|x_i, g_i) = p_{\alpha,\beta,\phi}(y_i|x_i, g_i) = \exp\left(\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\right), \quad (1)$$

where the row vector $\alpha \in \mathcal{R}^{d_x+1}$ and $\beta \in \mathcal{R}^{d_g}$. The linear predictor is $\eta_i = \eta_{\alpha,\beta}(x_i, g_i) = \alpha x_i^T + \beta g_i^T$.

Depending on the type of the phenotype under consideration, different specification of the functions $a()$, $b()$, and $c()$, including

- The continuous phenotype, corresponding to a linear regression;
- The binary phenotype, corresponding to a logistics regression;
- The count (integer) phenotype, corresponding to a Poisson regression.

To be more specific, we show how the modelling of the three types of phenotype belongs to our generalised linear model (GLM) framework. First, consider a continuous phenotype as specified in our previous linear model (LM) framework [17], i.e., $y_i \sim N(\eta_i, \sigma^2)$, where $\eta_i = \alpha x_i^T + \beta g_i^T$. Our previously proposed LM framework is a special case of our GLM framework because

$$\begin{aligned} f(y_i|x_i, g_i) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(y_i-\eta_i)^2}{2\sigma^2}) = \exp(\frac{y_i\eta_i - \eta_i^2/2}{\sigma^2} - \frac{y_i^2}{2} - \frac{\ln(2\pi\sigma^2)}{2}) \\ &= \exp(\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)). \end{aligned}$$

Thus, $a(\phi) = \phi$, $\phi = \sigma^2$, $b(\eta_i) = \eta_i^2/2$ and $c(y_i, \phi) = -y_i^2/(2\phi) - \ln(2\pi\phi)/2$ for a continuous phenotype.

Next, consider a binary phenotype modelled using a logistics regression model, i.e., $\pi_i = P(y_i = 1|x_i, g_i)$, and $\ln(\pi_i/(1-\pi_i)) = \eta_i = \alpha x_i^T + \beta g_i^T$. This logistics model is a special case of our GLM framework because

$$\begin{aligned} f(y_i|x_i, g_i) &= \pi_i^{y_i}(1-\pi_i)^{1-y_i} = (\frac{\pi_i}{1-\pi_i})^{y_i}(1-\pi_i) = \exp(\eta_i y_i + \ln(1-\pi_i)) \\ &= \exp(y_i\eta_i - \ln(1+e^{\eta_i})) = \exp(\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)). \end{aligned}$$

Thus, $a(\phi) = 1$, $b(\eta_i) = \ln(1+e^{\eta_i})$ and $c(y_i, \phi) = 0$ for a binary response.

Thirdly, consider a count (integer) phenotype modelled using a Poisson regression model, i.e., $Y_i \sim \text{Poisson}(\lambda_i = e^{\eta_i})$, where $\eta_i = \alpha x_i^T + \beta g_i^T$. This Poisson regression is also a special case of our GLM framework because

$$\begin{aligned} f(y_i|x_i, g_i) &= \frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i} = \frac{e^{\eta_i y_i}}{y_i!}e^{-\exp(\eta_i)} \\ &= \exp(y_i\eta_i - e^{\eta_i} - \ln(y_i!)) = \exp(\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)). \end{aligned}$$

Thus, $a(\phi) = 1$, $b(\eta_i) = e^{\eta_i}$ and $c(y_i, \phi) = -\ln(y_i!)$ for a count (integer) response.

Our proposed GLM framework is a general framework which can handle different types of responses. The probability of observing phenotype y_i is influenced by predictors (x_i and g_i) and parameters (α, β, ϕ). When $a(\phi)$ and $c(y_i, \phi)$ are constant functions with respect to ϕ, we can drop ϕ out and denote the functions as a and $c(y_i)$. For example, in logistics regression, $a(\phi) = 1$, $c(y_i, \phi) = 0$, and in Poisson regression, $a(\phi) = 1$, $c(y_i, \phi) = -\ln(y_i!)$. The parameters are (α, β) not involving ϕ. Then the probability of observing y_i is influenced by (x_i and g_i) and the parameters (α, β). In the following, we will discuss two situations, (1) the situation when parameters are (α, β, ϕ); and (2) the situation when ϕ is dropped out, i.e., parameters are (α, β).

2.2. Uncertain Genotypes

Because sequencing data and phenotype are conditional independent given true genotypes, we model their joint distribution to be

$$p_\theta(y_i, D_i|x_i) = \sum_{g\in\mathcal{G}} f_\theta(y_i|x_i, g)h(g, D_i), \tag{2}$$

where $\theta = (\alpha, \beta, \phi)$ or $\theta = (\alpha, \beta)$ depending on whether ϕ can be dropped out, i.e., whether $a(\phi)$ and $c(y_i, \phi)$ are constant functions with respect to ϕ. For individual i, D_i denotes sequencing reads, y_i denotes the phenotype, and x_i denotes additional covariates. We denote the genotype state space as $\mathcal{G}$, which contains all possible genotype values for g. The term $\sum_{g\in\mathcal{G}}$ in Equation (2) refers to sum over all possible values of g. Because each of

the d_g genetic variants can only has values of 0, 1, and 2, $\mathcal{G} = \{0,1,2\}^{d_g}$. The term $h(g, D_i)$ is short for the probability of NGS data and genotype, i.e., $h(g, D_i) = p(D_i|g)p(g|\hat{f})$. Here, the estimated allele frequency $\hat{f}$ is modelled in Skotte et al. [29]. The log-likelihood function thus can be written as

$$l_{y,D}(\theta) = \sum_{i=1}^{N} log\{p_\theta(y_i, D_i|x_i)\} = \sum_{i=1}^{N} log\{\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i, g)h(g, D_i)\}, \tag{3}$$

where $\theta = (\alpha, \beta, \phi)$ or (α, β) depending on whether ϕ is dropped out.

We are interested in testing for genetic effects based on NGS data. The null hypothesis is $H_0 : \beta = 0$. Under H_0, we note that the density f does not depends on genotypes so that in the likelihood function under H_0, we can pull it out of the summation.

Then, under $H_0 : \beta = 0$, we simplify the log-likelihood function as follows,

$$\begin{aligned} l_{y,D}(\alpha, 0, \phi) &= \sum_{i=1}^{N}(\ln(\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i, g)h(g, D_i))) \\ &= \sum_{i=1}^{N}\{\frac{y_i(\alpha x_i^T) - b(\alpha x_i^T)}{a(\phi)} + c(y_i, \phi)\} + \text{constant in terms of parameters,} \end{aligned}$$

when ϕ is not dropped out and parameters are (α, β, ϕ). When ϕ is dropped out and the parameters are (α, β), the formula is

$$\begin{aligned} l_{y,D}(\alpha, 0) &= \sum_{i=1}^{N}(\ln(\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i, g)h(g, D_i))) \\ &= \sum_{i=1}^{N}\{\frac{y_i(\alpha x_i^T) - b(\alpha x_i^T)}{a} + c(y_i)\} + \text{constant in terms of parameters.} \end{aligned}$$

Detailed derivations are in Supplementary Material S1 in Supplementary Information File.

Because constrained MLE under $H_0 : \beta = 0$ can be obtained using the regression of phenotype on additional covariate x (no genotype are used), this motivates us to use score test to develop the methods. Score test only need to know constrained MLE. Note that under $H_0 : \beta = 0$, the linear predictor $\eta_i = \alpha x_i^T + \beta g^T = \alpha x_i^T$ is influenced only by x_i, not by g.

2.3. Joint Significance Test for a Group of Common Genetic Variants

2.3.1. The Situation When Parameters Are (α, β, ϕ)

We adopt the score test [35,36]. When parameters are (α, β, ϕ), the score is

$$s_{y,D}(\alpha, \beta, \phi) = \begin{bmatrix} \partial l_{y,D}(\alpha, \beta, \phi)/\partial\alpha^T \\ \partial l_{y,D}(\alpha, \beta, \phi)/\partial\beta^T \\ \partial l_{y,D}(\alpha, \beta, \phi)/\partial\phi \end{bmatrix}, \tag{4}$$

where the length of row vector α is $d_x + 1$, the length of row vector β is d_g, and ϕ is a scalar. The analytical formula of $s_{y,D}(\alpha, \beta, \phi)$ is specified in Appendix A with detailed derivations provided in Supplementary Material S2 in Supplementary Information File.

The observed information matrix is

$$o_{y,D}(\alpha, \beta, \phi) = -\begin{bmatrix} \partial^2 l_{y,D}(\alpha, \beta, \phi)/\partial\alpha^T\partial\alpha & \partial^2 l_{y,D}(\alpha, \beta, \phi)/\partial\alpha^T\partial\beta & \partial^2 l_{y,D}(\alpha, \beta, \phi)/\partial\alpha^T\partial\phi \\ \partial^2 l_{y,D}(\alpha, \beta, \phi)/\partial\beta^T\partial\alpha & \partial^2 l_{y,D}(\alpha, \beta, \phi)/\partial\beta^T\partial\beta & \partial^2 l_{y,D}(\alpha, \beta, \phi)/\partial\beta^T\partial\phi \\ \partial^2 l_{y,D}(\alpha, \beta, \phi)/\partial\phi\partial\alpha & \partial^2 l_{y,D}(\alpha, \beta, \phi)/\partial\phi\partial\beta & \partial^2 l_{y,D}(\alpha, \beta, \phi)/\partial\phi^2 \end{bmatrix}.$$

The analytical formula for $o_{y,D}(\alpha, \beta, \phi)$ is specified in Appendix B with detailed derivations provided in Supplementary Material S3 in Supplementary Information File.

Denote the constrained MLE of the parameters (α, β, ϕ) under $H_0 : \beta = 0$ as $\tilde{\theta} = (\tilde{\alpha}, 0, \tilde{\phi})$. The test statistic is

$$R(y, D) = [s_{y,D}(\tilde{\alpha}, 0, \tilde{\phi})]^T [o_{y,D}(\tilde{\alpha}, 0, \tilde{\phi})]^{-1} [s_{y,D}(\tilde{\alpha}, 0, \tilde{\phi})]. \quad (5)$$

To calculate this test statistic, we need to evaluate both $s(y, D)(\alpha, \beta, \phi)$ (score function) and $o(y, D)(\alpha, \beta, \phi)$ (information matrix) at the constrained MLE $\tilde{\theta} = (\tilde{\alpha}, 0, \tilde{\phi})$. The evaluation of $s(y, D)(\alpha, \beta, \phi)$ at the constrained MLE $\tilde{\theta} = (\tilde{\alpha}, 0, \tilde{\phi})$, i.e., $s_{y,D}(\tilde{\alpha}, 0, \tilde{\phi})$ is specified in Appendix C with detailed derivations provided in Supplementary Material S4 in Supplementary Information File. The evaluation of $o_{y,D}(\alpha, \beta, \phi)$ at constrained MLE $\tilde{\theta}$, i.e., $o_{y,D}(\tilde{\alpha}, 0, \tilde{\phi})$ is specified in Appendix D with detailed derivations provided in Supplementary Material S5 in Supplementary Information File.

Under H_0, $R(y, D) \sim \chi^2_{d_g}$. We conducted score test based on testing statistic $R(y, D)$ and p-value of the test can be calculated.

2.3.2. The Situation When Parameters Are (α, β)

When ϕ is dropped out and parameters are (α, β), the score function is

$$s_{y,D}(\alpha, \beta) = \begin{bmatrix} \partial l_{y,D}(\alpha, \beta)/\partial \alpha^T \\ \partial l_{y,D}(\alpha, \beta)/\partial \beta^T \end{bmatrix}, \quad (6)$$

where the length of row vector α is $d_x + 1$, the length of row vector β is d_g. The analytical formula of $s_{y,D}(\alpha, \beta)$ is specified in Appendix E with detailed derivations provided in Supplementary Material S6 in Supplementary Information File.

The observed information matrix is

$$o_{y,D}(\alpha, \beta) = -\begin{bmatrix} \partial^2 l_{y,D}(\alpha, \beta)/\partial \alpha^T \partial \alpha & \partial^2 l_{y,D}(\alpha, \beta)/\partial \alpha^T \partial \beta \\ \partial^2 l_{y,D}(\alpha, \beta)/\partial \beta^T \partial \alpha & \partial^2 l_{y,D}(\alpha, \beta)/\partial \beta^T \partial \beta \end{bmatrix}.$$

The analytical formula of $o_{y,D}(\alpha, \beta)$ is specified in Appendix F with detailed derivations provided in Supplementary Material S7 in Supplementary Information File.

Denote the constrained MLE of the parameters (α, β) under $H_0 : \beta = 0$ as $\tilde{\theta} = (\tilde{\alpha}, 0)$. The test statistic is

$$R(y, D) = [s_{y,D}(\tilde{\alpha}, 0)]^T [o_{y,D}(\tilde{\alpha}, 0)]^{-1} [s_{y,D}(\tilde{\alpha}, 0)]. \quad (7)$$

To calculate this test statistic, we need to evaluate both $s_{y,D}(\alpha, \beta)$ and $o(y, D)(\alpha, \beta)$ at the constrained MLE $\tilde{\theta} = (\tilde{\alpha}, 0)$. The evaluation of $s_{y,D}(\alpha, \beta)$ at the constrained MLE $\tilde{\theta} = (\tilde{\alpha}, 0)$, i.e., $s_{y,D}(\tilde{\alpha}, 0)$ is specified in Appendix G with detailed derivations provided in Supplementary Material S8 in Supplementary Information File. The evaluation of $o_{y,D}(\alpha, \beta)$ at the constrained MLE $\tilde{\theta} = (\tilde{\alpha}, 0)$, i.e., $o_{y,D}(\tilde{\alpha}, 0)$ is specified in Appendix H with detailed derivations provided in Supplementary Material S9 in Supplementary Information File.

Under H_0, the test statistic $R(y, D) \sim \chi^2_{d_g}$. We conduct score test based on $R(y, D)$ and calculate p-values.

2.4. Variable Collapse Test for a Group of Rare Variants

2.4.1. The Situation When Parameters Are (α, β, ϕ)

For rare variants, variable collapse (VC) methods collapse multiple genetic variables into one variable and use it in testing [23,37,38]. Rare genetic variants can be collapsed in different ways, depending on which method is used. Weighted burden test based on genotypes aggregate/collapse p rare variants by a weighted sum with the weight w_j, i.e., $AG_i = \sum_{j=1}^{p} w_j g_{ij}$, where g_{ij} refers to the genotype for the j-th rare variants for individual i. Rare alleles are coded as 0 and wild/reference alleles are coded as 1. The burden test adopts equal weight, i.e., $w_1 = w_2 = \cdots = w_p = 1$. In burden test, $AG_i = \sum_{j=1}^{p} g_{ij}$ so that association study is performed between the total sum of rare alleles for a group of genetic

variants and the phenotype. This means the influence of genotype g_{ij} on the phenotype is through AG_i.

We model the phenotype using a generalized linear model [33,34]. For individual i, the same generalized linear model is used except the change in linear predictor η_i. The probability of observing phenotype y_i is modelled to be

$$p(y_i|x_i, g_i) = p_{\alpha,\beta_0,\phi}(y_i|x_i, g_i) = \exp(\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)), \tag{8}$$

where the row vector $\alpha \in \mathcal{R}^{d_x+1}$ and $\beta_0 \in \mathcal{R}$ is a scalar. The linear predictor in GLM model is

$$\eta_i = \eta_{\alpha,\beta_0}(x_i, g_i) = \alpha x_i^T + \beta_0 AG_i = \alpha x_i^T + \beta_0 \sum_{j=1}^{d_g} w_j g_{ij}. \tag{9}$$

Depending on the type of the responses, we adopt different functions for $a(.)$, $b(.)$, and $c(.,.)$. Note that $AG_i = \sum_{j=1}^{d_g} w_j g_{ij}$ aggregates d_g rare genetic variables into one aggregate variable. The probability of observing y_i is influenced by both (x_i and AG_i) and parameters (α, β_0, ϕ).

Our model for common variants uses the same GLM framework except the change in linear predictor η_i, i.e.,

$$\eta_i = \alpha x_i^T + \beta g_i^T, \tag{10}$$

where $\beta \in \mathcal{R}^{d_g}$. We apply *the chain rule in calculus* to derive the formulae for the rare-variant model (Equation (9)) based on the formulae in the common-variant model, i.e., Equation (10). The connection between the two models is that the effects of rare variants as modeled by β_0 satisfy the condition that $\beta = \beta_0 W$, where $\beta \in \mathcal{R}^{d_g}$ and $\beta_0 \in \mathcal{R}$. For burden test which uses equal weight, i.e., $w_1 = w_2 = \cdots = w_{d_g} = 1$, so that $W = [1, 1, \ldots, 1]$ is a unit row vector and $AG_i = \sum_{j=1}^{d_g} g_{ij}$ [23,37]. Then, $\eta_i = \alpha x_i^T + \beta_0 W g_i^T = \alpha x_i^T + \beta_0 \sum_{j=1}^{d_g} w_j g_{ij}$. Unequal weights can also be adopted, such as $w_j = \beta_0 f_{Beta}(MAF_j, 1, 25)$, where f_{Beta} is the Beta density function. The term MAF_j is MAF for the j-th rare variant [25,26,30].

The same assumption on weights are used in our proposed NGS data-based variable collapse (VC) method. We adopted the assumption of weighted burden test in our test based on NGS data. This assumption has been widely used in VC test based on genotypes in the literature. [23,37]. The formula is $\beta = \beta_0 W$, where the weight $W = (w_1, w_2, \ldots, w_{d_g})$ is a row vector and β_0 is a scalar. For identification purpose, the constraint $\sum_{j=1}^{d_g} w_j = d_g$ is adopted.

In our joint significance (JS) method for a group of common variants, η_i is modelled as $\eta_i = \eta_{\alpha,\beta}(x_i, g_i) = \alpha x_i^T + \beta g_i^T$ and (α, β, ϕ) are parameters and the length of row vector β is d_g. In comparison, in our variable collapse (VC) method for a group of rare variants, the linear predictor η_i is modelled as $\eta_i = \eta_{\alpha,\beta}(x_i, g_i) = \alpha x_i^T + \beta_0 W g_i^T$ and (α, β_0, ϕ) are parameters and β_0 is a scalar. First, under $H_0 : \beta = 0$ or $H_0 : \beta_0 = 0$, the same constrained MLE for α and ϕ is obtained, no matter which log-likelihood function ($l_{y,D}(\alpha, \beta, \phi)$ or $l_{y,D}(\alpha, \beta_0, \phi)$) is used. Thus, the same notation $\tilde{\theta} = (\tilde{\alpha}, 0, \tilde{\phi})$ is used to represent both the constrained MLE in $l_{y,D}(\alpha, \beta, \phi)$ (note the term 0 refers to a row vector containing d_g elements with all elements equal to 0) and the constrained MLE in $l_{y,D}(\alpha, \beta_0, \phi)$ (note that the term 0 is a scalar with the value of 0).

The evaluation of the score function at the constrained MLE for the rare-variant model is obtained as follows using the chain rule.

$$\begin{aligned}
\frac{\partial l_{y,D}(\alpha, \beta_0, \phi)}{\partial \beta_0}|_{\tilde{\theta}} &= W\frac{\partial l_{y,D}(\alpha, \beta, \phi)}{\partial \beta}|_{\tilde{\theta}} = W\{\frac{1}{a(\tilde{\phi})}\sum_{i=1}^{N}[y_i - b'(\tilde{\alpha}x_i^T)]E(g^T|D_i)\} \\
\frac{\partial l_{y,D}(\alpha, \beta_0, \phi)}{\partial \alpha}|_{\tilde{\theta}} &= 0; \ \frac{\partial l_{y,D}(\alpha, \beta_0, \phi)}{\partial \phi}|_{\tilde{\theta}} = 0
\end{aligned}$$

where $E(g^T|D_i) = \{\sum_{g\in\mathcal{G}} h(g,D_i)\}^{-1}\sum_{g\in\mathcal{G}} g^T h(g,D_i) = (\sum_{g\in\mathcal{G}} g^T h(g,D_i))/(\sum_{g\in\mathcal{G}} h(g,D_i))$ is the posterior expectation of the genotype of individual i given sequencing data D_i. Evaluation of the last two functions at constrained MLE is 0 because constrained MLE is obtained by constrained optimization of the likelihood function so that the first-order condition is satisfied, i.e., evaluation of the first derivatives are equal to 0.

Working similarly, for rare-variant models, we obtain the observed information matrix, and evaluate it at the constrained MLE. The formulae are

$$\begin{aligned}
\frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\alpha^T\partial\alpha}|_{\tilde{\theta}} &= \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\alpha^T\partial\alpha}|_{\tilde{\theta}} = -\frac{1}{a(\tilde{\phi})}\sum_{i=1}^{N} b''(\tilde{\alpha}x_i^T)x_i^T x_i \\
\frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\beta_0^2}|_{\tilde{\theta}} &= W\{\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\beta^T\partial\beta}|_{\tilde{\theta}}\}W^T \\
&= W\{\sum_{i=1}^{N}[\frac{(y_i - b'(\tilde{\alpha}x_i^T))^2}{[a(\tilde{\phi})]^2}(E(g^T g|D_i) - E(g^T|D_i)E(g|D_i)) - \frac{b''(\tilde{\alpha}x_i^T)}{a(\tilde{\phi})}E(g^T g|D_i)]\}W^T \\
\frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\phi^2}|_{\tilde{\theta}} &= \sum_{i=1}^{N}[(y_i\tilde{\alpha}x_i^T - b(\tilde{\alpha}x_i^T))(\frac{2[a'(\tilde{\phi})]^2}{[a(\tilde{\phi})]^3} - \frac{a''(\tilde{\phi})}{[a(\tilde{\phi})]^2}) + \frac{\partial^2 c(y_i,\phi)}{\partial\phi^2}|_{\tilde{\theta}}] \\
\frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\alpha^T\partial\beta_0}|_{\tilde{\theta}} &= \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\alpha^T\partial\beta}|_{\tilde{\theta}}W^T = -\frac{1}{a(\tilde{\phi})}\sum_{i=1}^{N} b''(\tilde{\alpha}x_i^T)x_i^T E(g|D_i)W^T \\
\frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\alpha^T\partial\phi}|_{\tilde{\theta}} &= \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\alpha^T\partial\phi}|_{\tilde{\theta}} = 0 \\
\frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\beta_0\partial\phi}|_{\tilde{\theta}} &= W\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\beta^T\partial\phi}|_{\tilde{\theta}} = -W\frac{a'(\tilde{\phi})}{[a(\tilde{\phi})]^2}\sum_{i=1}^{N}(y_i - b'(\tilde{\alpha}x_i^T))E(g^T|D_i) \\
\frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\beta_0\partial\alpha}|_{\tilde{\theta}} &= (\frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\alpha^T\partial\beta_0}|_{\tilde{\theta}})^T;\ \frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\phi\partial\alpha}|_{\tilde{\theta}} = (\frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\alpha^T\partial\phi}|_{\tilde{\theta}})^T \\
\frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\phi\partial\beta_0}|_{\tilde{\theta}} &= \frac{\partial^2 l_{y,D}(\alpha,\beta_0,\phi)}{\partial\beta_0\partial\phi}|_{\tilde{\theta}}
\end{aligned}$$

The test statistic is

$$R(y,D) = [s_{y,D}(\tilde{\alpha},0,\tilde{\phi})]^T o_{y,D}(\tilde{\alpha},0,\tilde{\phi})[s_{y,D}(\tilde{\alpha},0,\tilde{\phi})].$$

Under $H_0 : \beta_0 = 0$, $R(y,D)$ is approximately χ_1^2. Based on the test statistic, we conduct score test and calculate p-value.

2.4.2. The Situation When Parameters Are (α,β)

Consider the situation when ϕ is dropped out and parameters are parameters are (α,β). All setups are the same except for the following changes:

1. The parameters used in rare-variant testing are (α,β_0), where $\beta_0 \in \mathcal{R}$ and the parameters in common-variant testing are (α,β), where $\beta \in \mathcal{R}^{d_g}$;
2. The likelihood functions for rare-variant testing and common-variant testing are, respectively, denoted as $l_{y,D}(\alpha,\beta_0)$ and $l_{y,D}(\alpha,\beta)$;
3. The same notation $\tilde{\theta} = (\tilde{\alpha},0)$ is used to represent the constrained MLE in $l_{y,D}(\alpha,\beta_0)$ (note the term 0 is a scalar of 0) and the constrained MLE in $l_{y,D}(\alpha,\beta)$ (note the term 0 is a zero row vector of length d_g).

Evaluation of the score function at the constrained MLE is as follows.

$$\begin{aligned}\frac{\partial l_{y,D}(\alpha,\beta_0)}{\partial \beta_0}|_{\tilde{\theta}} &= W\frac{\partial l_{y,D}(\alpha,\beta)}{\partial \beta_0}|_{\tilde{\theta}} = W\{a^{-1}\sum_{i=1}^{N}(y_i-\tilde{\alpha}x_i^T)E(g^T|D_i)\}\\ \frac{\partial l_{y,D}(\alpha,\beta_0)}{\partial \alpha}|_{\tilde{\theta}} &= 0\end{aligned}$$

Note that, on the above, the first formula is derived by the chain rule. The second formula is 0 because constrained MLE maximizes the log-likelihood under $H_0 : \beta_0 = 0$, so that the first order condition in constrained optimization is satisfied, i.e., evaluation of the first derivatives at constrained MLE is 0.

Working similarly, we derive the observed information matrix, and evaluate it at the constrained MLE. The derivations are as follows.

$$\begin{aligned}\frac{\partial^2 l_{y,D}(\alpha,\beta_0)}{\partial \alpha^T \partial \alpha}|_{\tilde{\theta}} &= \frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial \alpha^T \partial \alpha}|_{\tilde{\theta}} = -\frac{1}{a}\sum_{i=1}^{N} b''(\tilde{\alpha}x_i^T)x_i^T x_i\\ \frac{\partial^2 l_{y,D}(\alpha,\beta_0)}{\partial \beta_0^2}|_{\tilde{\theta}} &= W\{\frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial \beta^T \partial \beta}|_{\tilde{\theta}}\}W^T\\ &= W\{\sum_{i=1}^{N}[\frac{(y_i-b'(\tilde{\alpha}x_i^T))^2}{a^2}(E(g^T g|D_i)-E(g^T|D_i)E(g|D_i))-\frac{b''(\tilde{\alpha}x_i^T)}{a}E(g^T g|D_i)]\}W^T\\ \frac{\partial^2 l_{y,D}(\alpha,\beta_0)}{\partial \alpha^T \partial \beta_0}|_{\tilde{\theta}} &= \frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial \alpha^T \partial \beta}|_{\tilde{\theta}}W^T = -\frac{1}{a}\{\sum_{i=1}^{N} b''(\tilde{\alpha}x_i^T)x_i^T E(g|D_i)\}W^T\\ \frac{\partial^2 l_{y,D}(\alpha,\beta_0)}{\partial \beta_0 \partial \alpha}|_{\tilde{\theta}} &= (\frac{\partial^2 l_{y,D}(\alpha,\beta_0)}{\partial \alpha^T \partial \beta_0}|_{\tilde{\theta}})^T;\end{aligned}$$

The test statistic is

$$R(y,D) = [s_{y,D}(\tilde{\alpha},0)]^T o_{y,D}(\tilde{\alpha},0)[s_{y,D}(\tilde{\alpha},0)].$$

Under $H_0 : \beta_0 = 0$, $R(y,D)$ is approximately χ_1^2. Based on the test statistic, we conduct a score test and calculate the p-value.

2.5. Software to Implement the Methods

We implement our proposed NGS data-based methods using R software (version 4.2.0). We have uploaded the R script files to implement our methods into the Github folder, which is publicly available via the link: https://github.com/zhengxu0459/NGS.Data.Based.Group.Testing.Based.On.GLM (accessed on 17 May 2023). The Github folder contains six script files to implement our NGS data-based (1) joint significance test and (2) variable collapse test for (i) continuous phenotype, (ii) binary phenotype, and (iii) count phenotype.

2.6. Specification of Simulation Studies

To evaluate the performance of our proposed methods (NGS data-based JS test for common variants and VC test for rare variants) versus literature methods (corresponding methods based on genotypes), we conduct simulation studies. Various setting simulation have been designed.

For common genetic variables and binary response, we evaluated the performance of our JS test based on NGS data versus literature methods (Chi-square test) based on genotype. For rare genetic variables and binary response, we evaluated the performance of our VC test based on NGS data versus literature methods (burden test and SKAT test based on genotypes).

We also conducted simulations for count/integer response. Although continuous response and binary response are two of the most commonly encountered types of phenotype

in association studies, other types of responses are also in association studies, though not as common as continuous type and binary type. We show that the use of a generalized linear model-framework allows us to handle other types of responses in addition to continuous phenotype. Because genotype-based SKAT testing method is only available for continuous phenotype and binary phenotype [25], we did not compare our methods with genotype-based SKAT method for count/integer response in rare-variant testing. For count/integer phenotype and rare genetic variables, we evaluated the performance our NGS data-based VC test versus genotype-based burden test literature. For count/integer phenotype and common genetic variables, we evaluated the performance of our NGS data-based JS test versus the genotype-based literature method (Chi-square test).

The software COSI was used to simulate 100 kb genomic regions based on a coalescent model. We adopt the best-fit model in COSI so that regions can be generate to mimic local recombination rate, LD patterns and European population history of Europeans [39]. We simulate chromosomes in the simulated regions. We used the software ShotGun [40] to generate sequencing data (per base pair error rate = 0.5%). ShotGun is publicly available at the webpage https://yunliweb.its.unc.edu/shotgun.html, (accessed on 17 May 2023). We specify various average sequencing depths, such as $d = 1X, 2X, 4X, 10X$. We classify genetic variables as common or rare depending on whether $MAF \geq 0.05$. Two additional covariates are simulated: $X_1 \sim N(0,1)$ (continuous) and $X_2 \sim$ Bernoulli(0.5) (binary).

To simulate the binary phenotype, we use the logistics model,

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \sum_{j=1}^{d_g} \beta_{gj} G_j + \epsilon, \tag{11}$$

where there are d_g is genetic variables, $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 1$, and $\epsilon \sim N(0,1)$.

To simulate the count/integer phenotype, we use the Poisson model

$$Y \sim \text{Poisson}(\lambda = e^{\eta}), \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \sum_{j=1}^{d_g} \beta_{gj} G_j, \tag{12}$$

where there are d_g genetic variables, $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 1$, and $\epsilon \sim N(0,1)$.

In simulations of both types, we set genetic effects, i.e., values of β_{gj}, differently in different scenarios, which allows us to evaluate Type I errors and perform power analysis (Type II errors).

Under $H_0 : \beta_{g1} = \beta_{g1} = \cdots = \beta_{gd_g} = 0$, we generate 9000 replicates for the evaluation of Type I errors. Type I errors are calculated for all combinations of 3 samples sizes ($n = 300, 500, 1000$) and 4 sequencing depths ($d = 1X, 2X, 4X, 10X$). Results of Type I errors are reported for (1) the JS test of a group of common genetic variables with the binary phenotype and the count/integer phenotype, and (2) the VC test of a group of rare genetic variables with the binary phenotype and the count/integer phenotype.

Then, we conduct simulation studies under the alternative hypothesis H_1, i.e., there are non-zero effects in the d_g genetic variables. In our simulation, we randomly choose multiple genetic variables as causal markers. For common variants, we randomly choose 2 to 5 causal genetic markers. For rare variants, we randomly choose 2 to 10 causal genetic markers. Then we simulate phenotypes based on these causal genetic markers and additional covariates (X_1, X_2). The total genetic effect is between 0 and 1 (Scale Parameter = 0.2 multiplied by the magnitude range of 0 to 5) with individual genetic effect specified to be the total effect divided by the number of causal variables.

Our simulations have used various (1) sequencing depths, (2) sample sizes, (3) number of causal variables, and (4) genetic effects. Based on simulated data, we evaluate the performance of different testing methods for common variants and rare variants, and compare our NGS data-based methods with the corresponding genotype-based methods in literature.

Our NGS data-based joint significance (JS) test use true allele frequencies (AFs) and two methods to estimate allele frequencies as in Skotte et al. [29]. NGS data-based JS test 1, 2, and 3, respectively, refer to JS test based on NGS data using (1) true AFs, (2) estimated AFs using the two-step genotype-based method (first estimate genotypes, and then calculate AFs using the estimated genotypes), and (3) one-step MLE estimator of AFs using the likelihood function of NGS data inSkotte et al. [29]. Both AF estimation methods have been proposed in Skotte et al. [29]. In general, we expect our testing method based on NGS data to have the best performance when true allele frequencies are used, i.e., our JS test 1 based on NGS data; however, JS Test 1 is not feasible because we do not know true AFs in practice; therefore, we need to use estimated AFs. The use of estimated AFs instead of true AFs is expected to make testing performance a little worse but we expect it will still be better than the corresponding genotype-based methods in the literature. Because one-step MLE of AFs is expected to be more accurate than the two-step genotype-based AF estimator, which has been reported in our previous work based on the same simulated sequencing data [17]. Thus, we expect our JS Test 3 based on NGS data to be better than our JS Test 2 based on NGS data.

Similarly, Variable Collapse (VC) Test 1, Test 2, and Test 3 based on NGS data refers to VC tests based on NGS data using (1) true AFs, (2) two-step genotype-based estimated AFs, and (3) one-step MLE of AFs. VC Test 1 is an infeasible estimator because, in practice, we do not know true AFs. We expect that VC Test 3 based on NGS data will be better than VC Test 2 based on NGS data.

2.7. Plan of a Real NGS Data Study

We describe our plan to systematically evaluate our methods in real NGS data. To be more specific, in our real NGS data example, we use the recent expansion of the 1000 Genomes Project (1kGP), which includes 602 trios, as described in Byrska-Bishop et al. [41]. We selected the 1000 Genomes Project (1kGP) for analysis because 1kGP is the largest fully open resource of whole-genome sequencing (WGS) data consented for public distribution without access or use restrictions. Recent expansion of 1kGP has contained 3202 samples, including 602 complete trios, deep sequenced to a depth of 30X, which is a suitable NGS dataset for analysis. The dataset is publicly available at https://www.internationalgenome.org/data-portal/data-collection/30x-grch38 (accessed on 5 May 2023).

To evaluate the performance of our methods under different scenarios of sequencing depths ($d = 1X, 2X, 4X, 10X, 30X$), down-sampling has been conducted to generate sequencing data with depth $d = 1X, 2X, 4X$, and $10X$ using the bioinformatics software samtools [42], accessible at http://www.htslib.org/ (accessed on 5 May 2023), which can work on NGS data in the bam file format and randomly down-sample NGS reads. For example, if we want to generate NGS data with the depth 10X based on 30X data, we set the down-sample ratio to be 1 out of 3, i.e., the ratio of 1:3. To generate NGS data with sequencing depths $d = 1X, 2X, 4X$, and $10X$, the down-sample ratios are, respectively, 1:30, 1:15, 2:15, and 1:3. Because our methods are for unrelated individuals only, we randomly select at most 1 individual for each family to form a dataset with unrelated individuals. The 1000 genome project has provided accurately-estimated genotypes based on deep sequenced NGS data at $d = 30X$ and we use these accurately estimated as true genotypes. The estimated genotypes were obtained by genotype calling on the down-sampled NGS data, i.e., NGS data at depth $d = 1X, 2X, 4X$, and $10X$. Because 1 kGP does not provide phenotype data, we simulate phenotype based on a generalized linear model. Therefore, this real data example makes use of real NGS data and genotype data rather than simulated phenotype data.

In our ongoing project, we will evaluate the performance of our methods and compare with traditional methods in the literature.

3. Results

Results of simulations are summarized as (1) Type I errors in Tables 1–4, and (2) power analysis in Figures 1–4. We evaluate the performance of different testing methods. We compare our testing methods based on NGS data with the corresponding genotype-based methods in the literature.

3.1. Results of Type I Errors

Type I errors in different scenarios (sample size $n = 300, 500, 1000$; depth $d = 1, 2, 4, 10$) are reported. Depending on whether genetic variables are common or rare, different NGS data-based methods (JS test or VC test) are used.

For association between continuous phenotype and common genetic variables, Type I errors of our NGS data-based joint significance (JS) tests using true AFs and two ways of estimating AFs are reported. In Table 1, Type I error for different testing methods for association between binary phenotype and a group of common genetic variables are calculated. Genotype-based Chi-square test conduct a Chi-square test for JS of a group of variables in the logistics regression of phenotype on estimated genotypes and other predictors. Genotype-based methods refers to methods which first estimate genotypes and then conduct association study based on estimated genotypes and phenotype. Our methods based on NGS data directly model the probability of observing phenotype and sequencing data, without the step of genotype estimation. We repeat our simulation study for count/integer phenotype and common genetic variants, and report Type I errors in in Table 2. According to Tables 1 and 2, for both binary response and integer/count response, in most scenarios, all methods control Type I errors as expected.

For association between binary phenotype and rare genetic variables, Type I errors of our NGS data-based variable collapse (VC) tests using true AFs, and two ways of estimating AFs (two-step genotype-based AF estimation and one-step NGS data-based AF estimation). NGS data-based VC Test 1, 2, 3 refer to our testing methods based on NGS data using true AFs and two allele frequency estimators. In Table 3, Type I errors for a group of *rare genetic variables* with a binary phenotype are reported. We evaluate the Type I errors of our NGS data-based VC Test 1, 2, 3, and two genotype-based rare-variant methods (burden test and SKAT test). We repeat our simulation study for count/integer phenotype and rare genetic variants and report simulation results of Type I errors in Table 4. Because genotype-based SKAT testing is only available for continuous response and binary response, we compare our methods with genotype-based burden tests. According to Tables 3 and 4, for both binary response and integer/count response, in most scenarios, all methods control Type I errors as expected.

Table 1. Type I errors of testing methods for a group of common genetic variants and binary phenotype. Genotype-based χ^2 test refers to Genotype-based Chi-square test. The term "NGS JS Test 1, 2, 3" refer to NGS data-based joint significant test with use of (1) true AF, (2) two-step genotype-based estimated AF, and (3) MLE of AFs based on NGS data.

Sample Size	Depth	Genotype-Based χ^2 Test	NGS JS Test 1	NGS JS Test 2	NGS JS Test 3
300	1	0.050	0.050	0.050	0.049
500	1	0.047	0.050	0.053	0.050
1000	1	0.047	0.046	0.048	0.046
300	2	0.050	0.050	0.053	0.049
500	2	0.051	0.051	0.052	0.050
1000	2	0.050	0.050	0.049	0.049
300	4	0.050	0.048	0.049	0.049

Table 1. *Cont.*

Sample Size	Depth	Genotype-Based χ^2 Test	NGS JS Test 1	NGS JS Test 2	NGS JS Test 3
500	4	0.054	0.050	0.050	0.050
1000	4	0.044	0.048	0.048	0.048
300	10	0.054	0.046	0.047	0.047
500	10	0.052	0.047	0.047	0.047
1000	10	0.047	0.051	0.052	0.052

Table 2. Type I errors of testing methods for a group of common genetic variants and count/integer phenotype. Genotype-based χ^2 test refers to Genotype-based Chi-square test. The term "NGS JS Test 1, 2, 3" refer to NGS data-based joint significant test with use of (1) true AFs, (2) estimated AFs based on genotype-based method, and (3) MLE of AFs based on NGS data.

Sample Size	Depth	Genotype-Based χ^2 Test	NGS JS Test 1	NGS JS Test 2	NGS JS Test 3
300	1	0.047	0.052	0.052	0.052
500	1	0.047	0.051	0.052	0.051
1000	1	0.054	0.051	0.051	0.052
300	2	0.053	0.052	0.050	0.052
500	2	0.053	0.050	0.050	0.050
1000	2	0.052	0.045	0.043	0.044
300	3	0.046	0.052	0.052	0.052
500	3	0.052	0.049	0.047	0.048
1000	3	0.050	0.051	0.050	0.051
300	4	0.049	0.049	0.050	0.049
500	4	0.052	0.048	0.047	0.047
1000	4	0.050	0.046	0.046	0.046

Table 3. Type I errors of testing methods for a group of rare genetic variants and binary phenotype. Burden and SKAT refer to genotype-based burden test and SKAT test. The term "NGS VC Test 1, 2, 3" refer to NGS data-based variable collapse test with use of (1) true AFs, (2) estimated AF based on genotype-based method, and (3) MLE of AFs based on NGS data.

Sample Size	Depth	Burden	SKAT	NGS VC Test 1	NGS VC Test 2	NGS VC Test 3
300	1	0.050	0.052	0.045	0.046	0.046
500	1	0.052	0.048	0.048	0.050	0.049
1000	1	0.049	0.053	0.049	0.045	0.049
300	2	0.046	0.051	0.042	0.043	0.043
500	2	0.050	0.050	0.053	0.054	0.053
1000	2	0.054	0.047	0.054	0.053	0.053
300	3	0.048	0.052	0.051	0.051	0.052
500	3	0.043	0.048	0.050	0.049	0.050
1000	3	0.050	0.049	0.052	0.052	0.052

Table 3. *Cont.*

Sample Size	Depth	Burden	SKAT	NGS VC Test 1	NGS VC Test 2	NGS VC Test 3
300	4	0.048	0.052	0.054	0.054	0.054
500	4	0.047	0.054	0.051	0.050	0.051
1000	4	0.051	0.050	0.052	0.052	0.053

Table 4. Type I errors of testing methods for a group of rare genetic variants and count/integer phenotype. Burden refers to genotype-based burden test. The term "NGS VC Test 1, 2, 3" refer to NGS data-based variable collapse test with use of (1) true AFs, (2) estimated AFs based on genotypes, and (3) MLE of AF based on NGS data.

Sample Size	Depth	Burden	NGS VC Test 1	NGS VC Test 2	NGS VC Test 3
300	1	0.049	0.051	0.053	0.050
500	1	0.051	0.051	0.051	0.049
1000	1	0.045	0.050	0.052	0.050
300	2	0.052	0.050	0.050	0.050
500	2	0.054	0.051	0.052	0.051
1000	2	0.052	0.048	0.051	0.048
300	3	0.048	0.047	0.050	0.048
500	3	0.045	0.045	0.047	0.046
1000	3	0.050	0.044	0.045	0.044
300	4	0.047	0.046	0.046	0.046
500	4	0.049	0.047	0.046	0.047
1000	4	0.050	0.047	0.048	0.047

3.2. Results of Power Analyses

Performance of different methods are evaluated in terms of statistical power. Statistical power is the probability of rejecting the null hypothesis under alternative hypothesis.

In Figure 1, we show power of different tests for a binary phenotype and common genetic variables. From top to bottom, the four rows have sequencing depth of 1, 2, 4, and 10. From left to right, the three columns have sample sizes of 300, 500, and 1000. Powers of different tests (genotype-based Chi-square test (red), NGS data-based joint significance test 1 (black), test 2 (purple), and test 3 (green)) are represented as curves. We found that when sequencing depth is low (depth = 1X, 2X), our proposed methods based on NGS data performed better than the genotype-based test in the literature. When sequencing depth increases, the advantage of NGS data-based methods over methods based on genotypes decreases. When sequencing depth is 10X, methods based on NGS data and methods based on genotypes have similar performance. When sample size increases, the power of all tests increases. Comparing the three NGS data-based JS tests, we found JS Test 1 (using true AF) and 3 (using MLE of AF based on NGS data) have similar performance, whereas JS Test 2 (estimate AF using genotype-based method) has slightly worse performance compared with Test 1 and 3; however, NGS data-based JS Test 2 still has better performance than the corresponding genotype-based test in the literature. When sequencing becomes deep, the three NGS data-based tests show similar performance in terms of statistical power.

We repeat power analysis of common genetic variants for count/integer phenotype. Similar findings are obtained for the count/integer phenotype. We found that the power of all tests increases as the sample size increases from 300 to 1000. When sequencing depth is low (depth = 1X, 2X), our NGS data-based methods demonstrate advantages over genotype-based methods in the literature; however, when sequencing depth increases, the magnitude of the advantage decreases. When sequencing depth is at 10X, all methods show similar performances. JS Test 1 (using true AF) and 3 (using MLE of AF based on NGS data [29]) have similar performance, whereas JS Test 2 (estimate AF using genotype-based method) show slightly worse performance compared with Test 1 and 3. We report our results for the count/binary phenotype in Figure 2.

In Figure 3, power of different tests for binary phenotype and a group of *rare genetic variables* are reported. The four rows from up to down are for sequencing depth 1X, 2X, 4X, and 10X. The three columns from left to right are for sample size $n = 300, 500$, and 1000. Powers of different tests are represented as curves and coloured differently. The tests include genotype-based burden test (red), SKAT test (blue), NGS data-based VC Test 1 (black), 2 (purple), and 3 (green). Our proposed methods based on NGS data are found to have better performance than genotype-based methods in the literature when sequencing depth is low (1X and 2X). When sequencing become more deep, the advantages of NGS data-based methods over genotype-based methods decreases. When sequencing depth is 10X, NGS data-based methods and genotype-based methods have similar performance. Comparing the three methods based on NGS data, we find VC Test 1 (using true AF) and 3 (using MLE of AF based on NGS data) have similar performance, whereas Test 2 (estimate AF based on genotypes) is slightly worse compared with Test 1 and 3. When sequencing becomes deep, the difference between the performance of the three VC tests decreases. Comparing two genotype-based methods (Burden and SKAT), we find that the burden test has better performance than SKAT methods because our scenarios assume that all genetic effects in one direction, i.e., positive, which is not the assumption for SKAT, which allows effects in two directions, i.e., positive effects and negative effects for the group of markers.

We repeat our power analysis of rare genetic variants for count/integer phenotype. In Figure 4, the powers of different tests for count/integer phenotype and a group of *rare genetic variables* are shown. Because SKAT test is not available for count/integer phenotype, we compare our NGS-based methods with genotype-based burden test. Similar findings were obtained for count/integer phenotype as we find in binary phenotype.

For binary phenotype and rare variants, we consider the scenario of genetic effects in both directions (positive and negative), which is suitable for the SKAT test. We found that under this scenario, all burden tests (genotype-based and NGS data-based) fail as we expected. This is because burden tests assume genetic effects only in one direction and the burden-test assumption is not satisfied in this scenario. The failure of the burden test under the scenarios of genetic effects in both directions was recognized by researchers.

We leave the work of developing NGS data-based group testing methods for rare variants allowing genetic effects in two directions as our future study since it is beyond the scope of this article. Developing the NGS data-based SKAT test can address this issue and will be studied in the future. This article focused on developing NGS data-based test corresponding to genotype-based burden test and joint significant test in the literature.

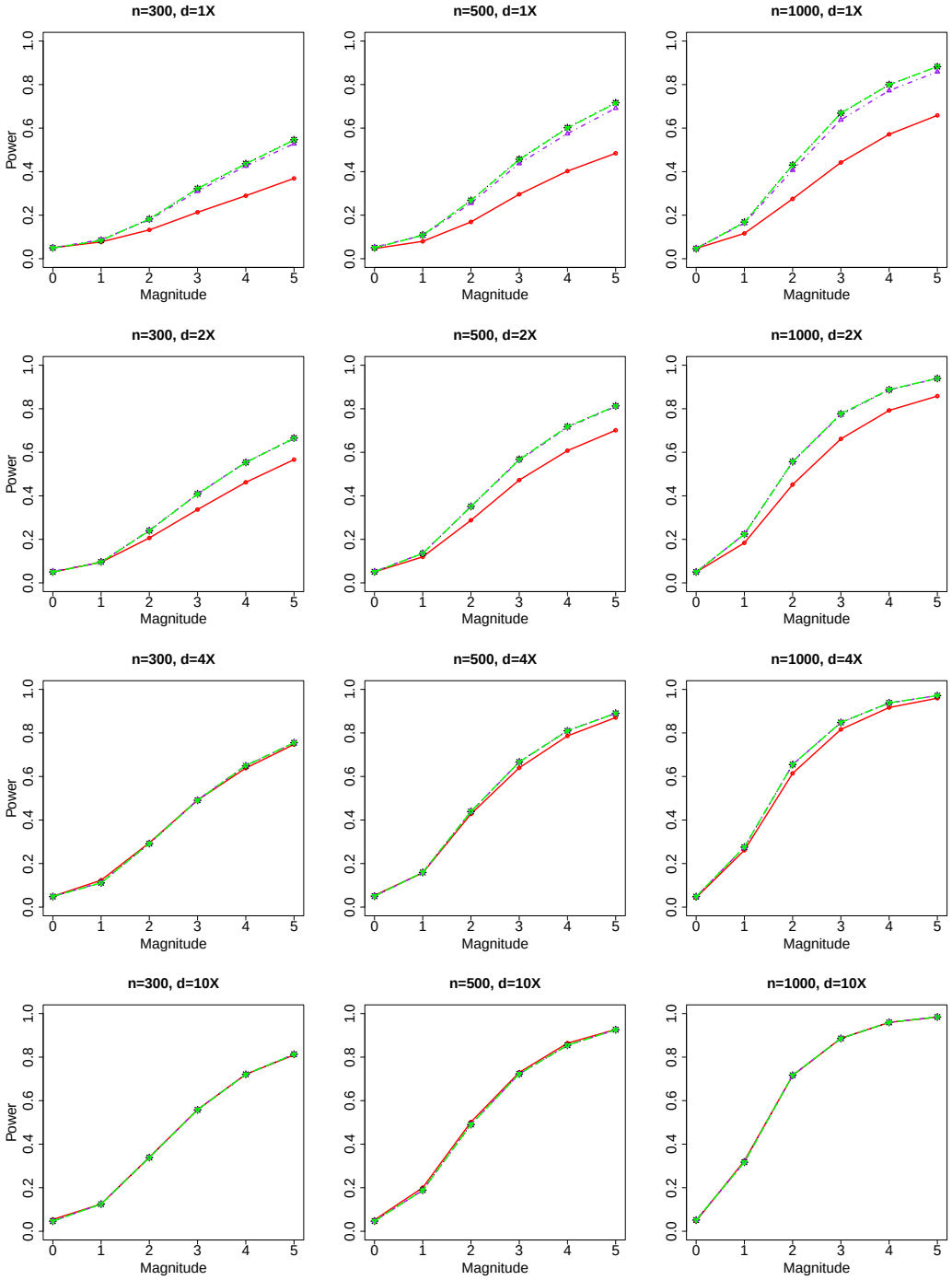

Figure 1. Power curves of tests for a group of common genetic variants and binary phenotype. From up to down, the four rows are for sequencing depth 1X, 2X, 4X, and 10X. From left to right, the three columns are for sample size n = 300, 500, and 1000. The panels show powers of genotype-based Chi-square test (red solid line), NGS data-based joint significance test 1 (black dotted line), test 2 (purple dotted-dashed line), and test 3 (green long-dashed line).

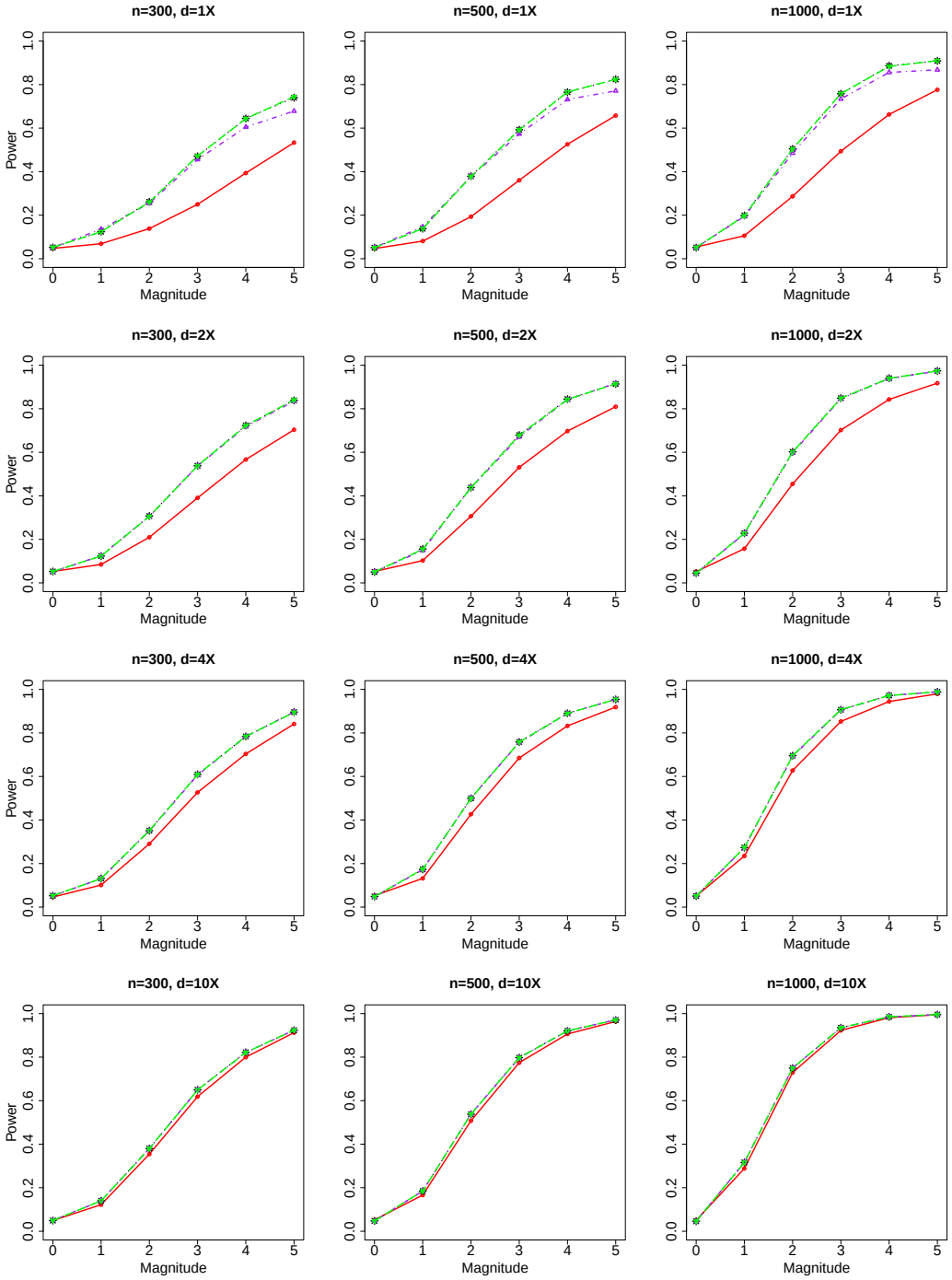

Figure 2. Power curves of tests for a group of common genetic variants and count/integer phenotype. From up to down, the four rows are for sequencing depth 1X, 2X, 4X, and 10X. From left to right, the three columns are for sample size n = 300, 500, and 1000. The panels show powers of genotype-based Chi-square test (red solid line), NGS data-based joint significance test 1 (black dotted line), test 2 (purple dotted-dashed line), and test 3 (green long-dashed line).

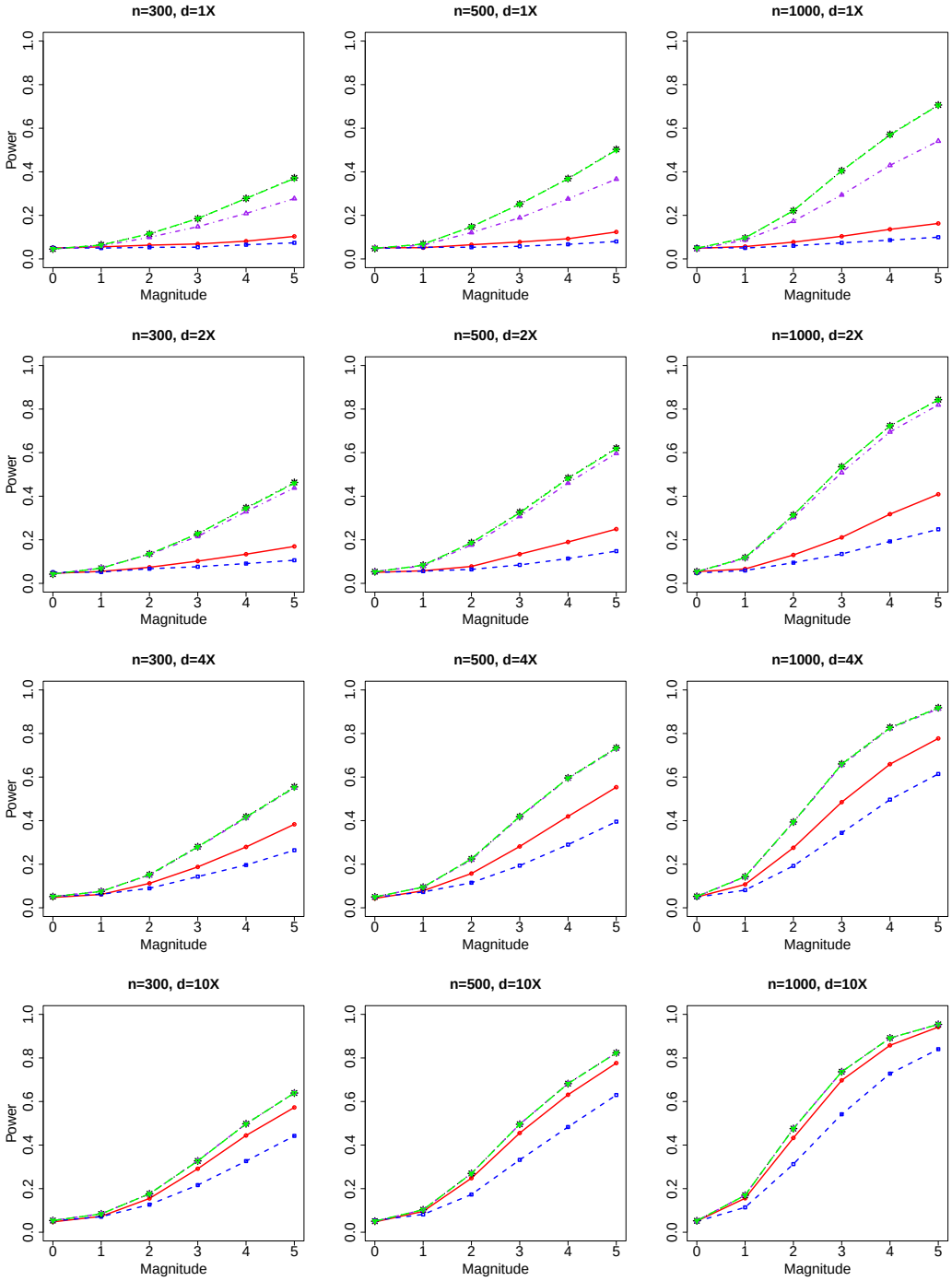

Figure 3. Power curves of tests for a group of rare genetic variants and binary phenotype. From up to down, The four rows are for sequencing depth 1X, 2X, 4X, and 10X. From left to right, the three columns are for sample size n = 300, 500, and 1000. The panels show power of genotype-based burden test (red solid line), genotype-based SKAT test (blue dashed line), NGS data-based variable collapse test 1 (black dotted line), 2 (purple dotted-dashed line), and 3 (green long-dashed line).

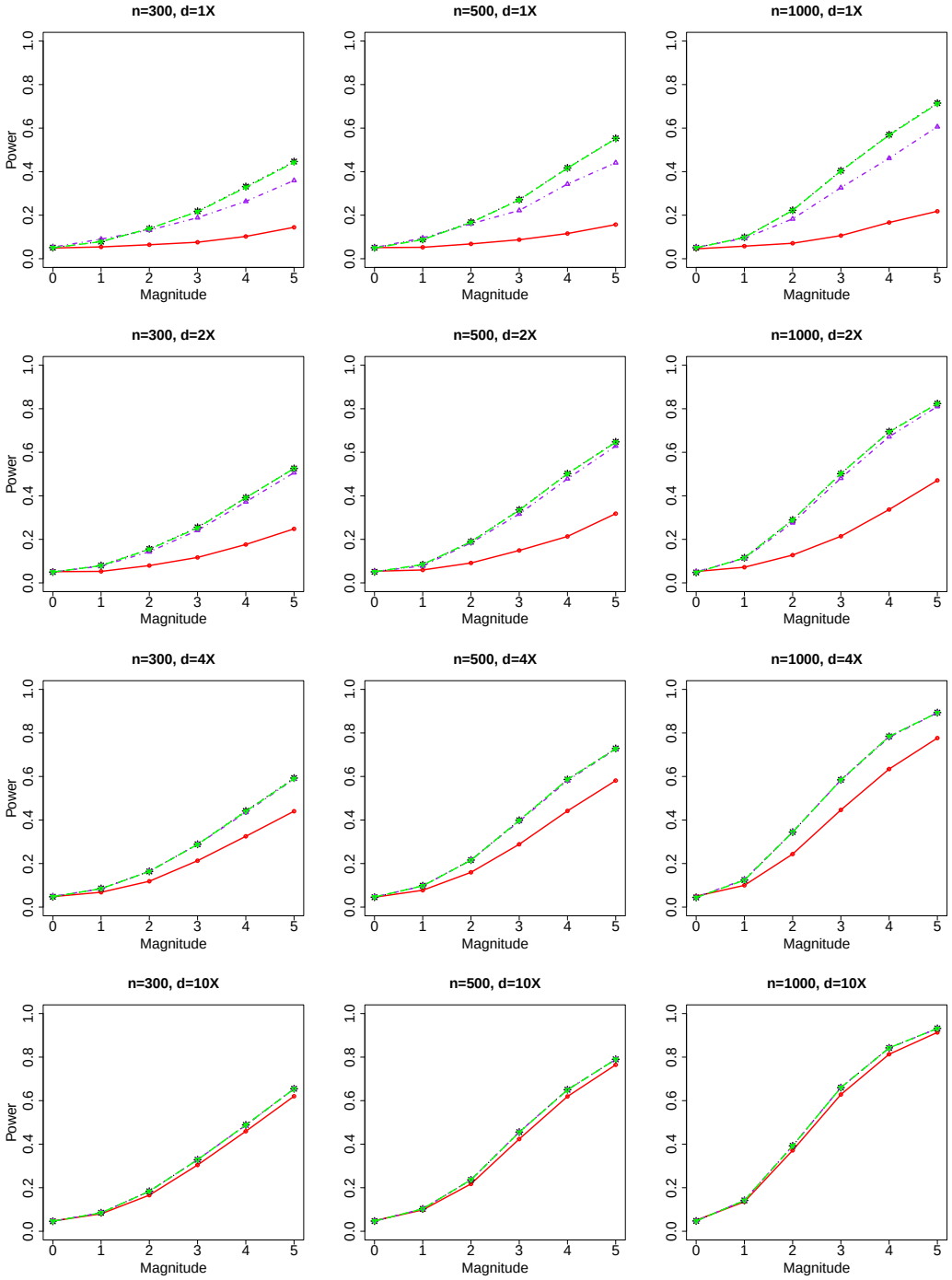

Figure 4. Power curves of tests for a group of rare genetic variants and count/integer phenotype. From up to down, the four rows are for sequencing depth 1X, 2X, 4X, and 10X. From left to right, the three columns are for sample size n = 300, 500, and 1000. The panels show power of genotype-based burden test (red solid line), NGS data-based variable collapse test 1 (black dotted line), test 2 (purple dotted-dashed line), and test 3 (green long-dashed line).

4. Discussion

The major finding of our article is that the proposed methods show advantage over their corresponding methods based on genotype in the literature both in testing for a group of *common* markers and testing for a group of *rare* markers. The main objective of our study was to apply the GLM framework to derive innovative group testing methods based on

NGS data. We fulfil the demand from researchers in bio-statistics, bio-informatics, and biology for developing group testing methods based on NGS data.

Our methods adopt the GLM framework to handle a range of types of phenotypes, including binary responses and continuous responses, which are two mostly encountered types in association studies. Our method extends our previous linear model framework [17] with the analytical capacity for more complex phenotypes in addition to continuous phenotypes.

In the Results section, we show the findings of comparing our methods with their corresponding methods in the literature for binary phenotype and count phenotype. For continuous phenotype, the GLM-based model will reduce to our previously published LM-based model [17]. Our findings for group testing of *common variants* from binary phenotype and count/integer phenotype are similar to our findings from continuous phenotype for group testing of *common variants* published previously [17]. Similarly, our findings for *rare variants* testing from binary phenotype and count/integer phenotype are similar to our findings from *continuous* phenotypes for *rare variant* testing published previously [17].

Our proposed model deals with association studies with unrelated/independent individuals. Association studies can be conducted based on related individuals, such as the situation where multiple individuals from the same family are involved in the study. In that case, a generalized linear mixed (GLMM) model instead of a generalized linear (GLM) model is used. Future studies can be on developing testing methods based on NGS data for related individuals. In an ongoing project, we are working to extend our GLM framework to GLMM framework so that it can handle related individuals.

Our proposed methods are based on a score test. We adopt the score test due to its advantage of fast computation and easy derivation of calculation formulae because it only calculates constrained MLE, which is the optimizer maximizing likelihood function under the null hypothesis. The likelihood ratio test can have improved performance compared with the score test [43,44]. Future studies can be on developing likelihood ratio-based association tests using NGS data which may have improved performance.

Sequencing depth can impact the comparison of performance between methods based on NGS data and methods based on genotypes. When sequencing depth is big (d = 4X and 10X) and genotype estimation is accurate, methods based on NGS data and methods based on genotypes show similar performance. When sequencing depth is small (depth = 1X and 2X) and genotype estimation is not precise, methods based on NGS data can have better performance than genotype-based methods, which are based on *estimated* genotypes [17,29,30]. In practice, given a limited financial budget, low sequencing to include more individuals is preferred to deep sequencing with few individuals sequenced [29,30]. Our proposed methods mainly show an advantage when sequencing is low (depth = 1X and 2X).

Our proposed methods are mainly developed based on a theoretical statistical framework of general linear models. We make use of statistical inference methodology to derive the score tests based on the likelihood function of next-generation sequencing (NGS) data and phenotypes with latent/un-observable genotypes. Then we conducted extensive simulation studies to evaluate the performance of our methods and compare our methods with the traditional methods in the theory. In an ongoing project, we are working on systematically evaluating and comparing our methods based on real NGS data. We separate the development of our methods into two stages. In Stage 1, we conduct theoretical development of our methods and simulation studies. In Stage 2, we further evaluate our methods in real NGS data.

We note that simulations performed in our hypothetical scenarios could be biased so that simulation studies could be biased and real data results can provide stronger verification. In our ongoing project, we aimed to systematically evaluate our method in real NGS data.

We note that verification of our methods can be at four levels. The four levels of verification are:

(1) Statistically theoretical verification to derive our methods theoretically based on GLM framework to show that our methods are statistically theoretically founded;
(2) Simulation studies to evaluate method performance and show that our methods can achieve better performance compared with traditional methods in the literature;
(3) Multiple real NGS data examples to evaluate and compare different methods;
(4) Biology lab verification and biology literature verification to show our methods indeed find some biologically meaningful genes related to the phenotype.

When possible, we recommend the use of higher levels of verification. For example, simulation studies under a specific scenario can be biased so that real data studies can provide stronger verification. We also recommend the use of as many verification levels as possible to provide verification in different perspectives. The current manuscript provides statistical theoretical derivations and simulation studies. The next steps for verification should be Level 3 (real data verification) and 4 (biology lab verification and biology literature verification). In our ongoing project, we are working on verification from real NGS data. In the future, biology lab verification and biology literature verification can be conducted by collaborating with biology experts.

Future directions of our study include (1) extending our GLM-based methods to generalized linear mixed model (GLMM)-based methods so that they can handle related individuals in association studies; (2) systematically evaluating the performance of our methods and comparing our methods with literature methods in real NGS data; (3) evaluating our methods based on the other three genetic effect models, recessive model, dominant model, and heterogeneous effect models [45]; and (4) biology lab verification and biology literature verification by collaborating with biology experts.

To describe the effect of genotype (coded as 0, 1, 2) at a single marker on phenotype, four models are widely used. The effect on linear predictor η in the GLM framework is specified differently in the four models [45], (1) additive model ($effect = \beta_0 + \beta_a g$); (2) recessive model ($effect = \beta_0 + \beta_r I(g = 2)$); (3) dominant model ($effect = \beta_0 + \beta_d I(g \geq 1)$); and (4) heterogeneous effect model ($effect = \beta_0 + \beta_1 I(g = 1) + \beta_2 I(g = 2)$). The indicator function $I(condition)$ is equal to 1 when the condition is satisfied, and is equal to 0 otherwise. The probability of observing the response in the GLM framework is

$$p(y_i|x_i, g_i) = p_{\alpha,\beta,\phi}(y_i|x_i, g_i) = \exp\left(\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\right), \tag{13}$$

and the linear predictor η_i is modelled differently for the recessive model, dominant model and heterogeneous effect model with d_g genetic markers. We model linear predictor η_i in GLM, respectively, as follows,

$$\begin{aligned}
\eta_i &= \alpha x_i^T + \beta_a \{I(G_i = 2)\}^T, \\
\eta_i &= \alpha x_i^T + \beta_d \{I(G_i \geq 1)\}^T, \\
\eta_i &= \alpha x_i^T + \beta_1 \{I(G_i = 1)\}^T + \beta_2 I(G_i = 2)^T,
\end{aligned}$$

where the four row vectors (β_d, β_a, β_1, β_2) have length of d_g. Although our methods are proposed based on the additive model described in Equation (10), the methods can be adapted for the other three genetic-effect models (recessive model; dominant model; heterogeneous effect model) by changing η_i and its first and second derivatives with respect to model parameters.

5. Conclusions

We extend our previously proposed NGS data-based testing methods (joint significance test for a group of common variants and variable collapse test for a group of rare variants) from a linear model (LM) framework to a generalized linear model (GLM) framework so that it can handle a range of types of responses (binary phenotypes; count/integer phenotypes) in addition to continuous phenotypes. Our proposed methods fill the lit-

erature gap. In addition, based on our results from simulation studies reported in the Results section, we found that our methods can achieve better performance than their corresponding genotype-based methods in the literature. Future studies will be conducted to evaluate our methods based on real NGS data.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math11112560/s1, Supplementary Information File.

Author Contributions: Conceptualization, Z.X., S.Y., C.W., Q.D., S.C. and Y.L.; Data curation, C.W.; Formal analysis, Z.X., S.Y., C.W. and S.C.; Funding acquisition, S.C. and Y.L.; Investigation, Z.X. and C.W.; Methodology, Z.X., S.Y., C.W., Q.D., S.C. and Y.L.; Project administration, Z.X. and Y.L.; Resources, Z.X. and Y.L.; Software, Z.X., S.Y., C.W., Q.D., S.C. and Y.L.; Supervision, Z.X. and Y.L.; Validation, Z.X., C.W., Q.D. and Y.L.; Visualization, Z.X., C.W. and Q.D.; Writing—original draft, Z.X. and C.W.; Writing—review and editing, Z.X. and C.W. All authors have read and agreed to the published version of the manuscript.

Funding: Sixia Chen was supported by the Oklahoma Shared Clinical and Translational Resources (U54GM104938) with an Institutional Development Award (IDeA) from NIGMS.

Data Availability Statement: All data used in the study are publicly available.

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

Abbreviations

The following abbreviations are used in this manuscript:

AF	Allele frequency
DNA	Deoxyribonucleic acid
GC	Genotype calling
GLF	Genotype likelihood function
GLM	Generalized linear model
GLMM	Generalized linear mixed model
JS	Joint significance
LM	Linear model
LRT	Likelihood ratio test
MAD	Mean absolute deviation
MSE	Mean squared error
NGS	Next-generation sequencing
RNA	Ribonucleic acid
SKAT	Sequence kernel association test
VC	Variable collapse

Appendix A. Analytical Formula of $s_{y,D}(\alpha, \beta, \phi)$

$$s_{y,D}(\alpha,\beta,\phi) = \sum_{i=1}^{N}[\{p_\theta(y_i, D_i|x_i)\}^{-1} \sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g) \begin{bmatrix} \frac{y_i - b'(\eta_i)}{a(\phi)} x_i^T \\ \frac{y_i - b'(\eta_i)}{a(\phi)} g^T \\ -\frac{y_i\eta_i - b(\eta_i)}{[a(\phi)]^2} a'(\phi) + \frac{\partial c(y_i,\phi)}{\partial \phi} \end{bmatrix} h(g, D_i)],$$

where $\eta_i = \eta(x_i, g) = \alpha x_i^T + \beta g^T$. Detailed derivations have been provided in Supplementary Material S2 in Supplementary Information File.

Appendix B. Analytical Formula of $o_{y;D}(\alpha;\beta;\phi)$

$$o_{y,D}(\alpha,\beta,\phi) = -\begin{bmatrix} \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\alpha^T\partial\alpha} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\alpha^T\partial\beta} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\alpha^T\partial\phi} \\ \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\beta^T\partial\alpha} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\beta^T\partial\beta} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\beta^T\partial\phi} \\ \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\phi\partial\alpha} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\phi\partial\beta} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\phi^2} \end{bmatrix}.$$

$$\begin{aligned}
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\alpha^T\partial\alpha} &= \sum_{i=1}^{N}(-\{p_\theta(y_i,D_i|x_i)\}^{-2})[\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a(\phi)}x_i^T h(g,D_i)][\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a(\phi)}x_i h(g,D_i)] \\
&\quad +\{p_\theta(y_i,D_i|x_i)\}^{-1}\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)[\frac{(y_i-b'(\eta_i))^2}{[a(\phi)]^2}-\frac{b''(\eta_i)}{a(\phi)}]x_i^T x_i h(g,D_i); \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\alpha^T\partial\beta} &= \sum_{i=1}^{N}(-\{p_\theta(y_i,D_i|x_i)\}^{-2})[\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a(\phi)}x_i^T h(g,D_i)][\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a(\phi)}gh(g,D_i)] \\
&\quad +\{p_\theta(y_i,D_i|x_i)\}^{-1}\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)[\frac{(y_i-b'(\eta_i))^2}{[a(\phi)]^2}-\frac{b''(\eta_i)}{a(\phi)}]x_i^T gh(g,D_i); \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\alpha^T\partial\phi} &= \sum_{i=1}^{N}(-\{p_\theta(y_i,D_i|x_i)\}^{-2})[\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a(\phi)}x_i^T h(g,D_i)][\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)(-\frac{y_i\eta_i-b(\eta_i)}{[a(\phi)]^2}a'(\phi)+\frac{\partial c(y_i,\phi)}{\partial\phi})h(g,D_i)] \\
&\quad +\{p_\theta(y_i,D_i|x_i)\}^{-1}\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)[(-\frac{y_i\eta_i-b(\eta_i)}{[a(\phi)]^2}a'(\phi)+\frac{\partial c(y_i,\phi)}{\partial\phi})(\frac{y_i-b'(\eta_i)}{a(\phi)})-\frac{y_i-b'(\eta_i)}{[a(\phi)]^2}a'(\phi)]x_i^T h(g,D_i); \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\beta^T\partial\beta} &= \sum_{i=1}^{N}(-\{p_\theta(y_i,D_i|x_i)\}^{-2})[\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a(\phi)}g^T h(g,D_i)][\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a(\phi)}gh(g,D_i)] \\
&\quad +\{p_\theta(y_i,D_i|x_i)\}^{-1}\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)[\frac{(y_i-b'(\eta_i))^2}{[a(\phi)]^2}-\frac{b''(\eta_i)}{a(\phi)}]g^T gh(g,D_i); \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\beta^T\partial\phi} &= \sum_{i=1}^{N}(-\{p_\theta(y_i,D_i|x_i)\}^{-2})[\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a(\phi)}g^T h(g,D_i)]\{\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)[-\frac{y_i\eta_i-b(\eta_i)}{[a(\phi)]^2}a'(\phi)+\frac{\partial c(y_i,\phi)}{\partial\phi}]h(g,D_i)\} \\
&\quad +\{p_\theta(y_i,D_i|x_i)\}^{-1}\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)((\frac{y_i-b'(\eta_i)}{a(\phi)}g^T)(-\frac{y_i\eta_i-b(\eta_i)}{[a(\phi)]^2}a'(\phi)+\frac{\partial c(y_i,\phi)}{\partial\phi})-(\frac{y_i-b'(\eta_i)}{[a(\phi)]^2}a'(\phi)g^T))h(g,D_i); \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\phi^2} &= \sum_{i=1}^{N}[-\{p_\theta(y_i,D_i|x_i)\}^{-2}(\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)[-\frac{y_i\eta_i-b(\eta_i)}{[a(\phi)]^2}a'(\phi)+\frac{\partial c(y_i,\phi)}{\partial\phi}]h(g,D_i))^2 \\
&\quad +\{p_\theta(y_i,D_i|x_i)\}^{-1}\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\{[-\frac{y_i\eta_i-b(\eta_i)}{[a(\phi)]^2}a'(\phi)+\frac{\partial c(y_i,\phi)}{\partial\phi}]^2+(y_i\eta_i-b(\eta_i))(\frac{2[a'(\phi)]^2}{[a(\phi)]^3}-\frac{a''(\phi)}{[a(\phi)]^2})+\frac{\partial^2 c(y_i,\phi)}{\partial\phi^2}\}h(g,D_i)]; \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\beta^T\partial\alpha} &= (\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\alpha^T\partial\beta})^T; \quad \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\phi\partial\alpha}=(\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\alpha^T\partial\phi})^T; \quad \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\phi\partial\beta}=(\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial\beta^T\partial\phi})^T.
\end{aligned}$$

Detailed derivations have been provided in Supplementary Material S3 in Supplementary Information File.

Appendix C. Evaluation of $s_{y,D}(\alpha,\beta,\phi)$ at Constrained MLE $\tilde{\theta}$, i.e., $s_{y,D}(\tilde{\alpha},0,\tilde{\phi})$

$$s_{y,D}(\tilde{\alpha},0,\tilde{\phi}) = \begin{bmatrix} 0 \\ \sum_{i=1}^{N}\frac{y_i-b'(\tilde{\alpha}x_i^T)}{a(\tilde{\phi})}E(g^T|D_i) \\ 0 \end{bmatrix},$$

where $E(g^T|D_i) = \{\sum_{g\in\mathcal{G}} h(g,D_i)\}^{-1}\sum_{g\in\mathcal{G}} g^T h(g,D_i) = (\sum_{g\in\mathcal{G}} g^T h(g,D_i))/(\sum_{g\in\mathcal{G}} h(g,D_i))$ is the posterior expectation of the genotype given sequencing data. Detailed derivations have been provided in Supplementary Material S4 in Supplementary Information File.

Appendix D. Evaluation of $o_{y,D}(\alpha, \beta, \phi)$ at Constrained MLE $\tilde{\theta}$, i.e., $o_{y,D}(\tilde{\alpha}, 0, \tilde{\phi})$

$$o_{y,D}(\tilde{\alpha}, 0, \tilde{\phi}) = -\begin{bmatrix} \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \alpha^T \partial \alpha}|_{\tilde{\theta}} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \alpha^T \partial \beta}|_{\tilde{\theta}} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \alpha^T \partial \phi}|_{\tilde{\theta}} \\ \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \beta^T \partial \alpha}|_{\tilde{\theta}} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \beta^T \partial \beta}|_{\tilde{\theta}} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \beta^T \partial \phi}|_{\tilde{\theta}} \\ \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \phi \partial \alpha}|_{\tilde{\theta}} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \phi \partial \beta}|_{\tilde{\theta}} & \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \phi^2}|_{\tilde{\theta}} \end{bmatrix}.$$

$$\begin{aligned}
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \alpha^T \partial \alpha}|_{\tilde{\theta}} &= -\frac{1}{a(\tilde{\phi})}\sum_{i=1}^{N} b''(\tilde{\alpha} x_i^T) x_i^T x_i \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \alpha^T \partial \beta}|_{\tilde{\theta}} &= -\frac{1}{a(\tilde{\phi})}\sum_{i=1}^{N} b''(\tilde{\alpha} x_i^T) x_i^T E(g|D_i) \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \alpha^T \partial \phi}|_{\tilde{\theta}} &= 0 \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \beta^T \partial \beta}|_{\tilde{\theta}} &= \sum_{i=1}^{N}[\frac{(y_i - b'(\tilde{\alpha} x_i^T))^2}{[a(\tilde{\phi})]^2}(E(g^T g|D_i) - E(g^T|D_i)E(g|D_i)) - \frac{b''(\tilde{\alpha} x_i^T)}{a(\tilde{\phi})}E(g^T g|D_i)] \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \beta^T \partial \phi}|_{\tilde{\theta}} &= -\frac{a'(\tilde{\phi})}{[a(\tilde{\phi})]^2}\sum_{i=1}^{N}(y_i - b'(\tilde{\alpha} x_i^T))E(g^T|D_i) \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \phi^2}|_{\tilde{\theta}} &= \sum_{i=1}^{N}[(y_i \tilde{\alpha} x_i^T - b(\tilde{\alpha} x_i^T))(\frac{2[a'(\tilde{\phi})]^2}{[a(\tilde{\phi})]^3} - \frac{a''(\tilde{\phi})}{[a(\tilde{\phi})]^2}) + \frac{\partial^2 c(y_i,\phi)}{\partial \phi^2}|_{\tilde{\theta}}] \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \beta^T \partial \alpha}|_{\tilde{\theta}} &= (\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \alpha^T \partial \beta}|_{\tilde{\theta}})^T; \ \ \frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \phi \partial \alpha}|_{\tilde{\theta}} = (\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \alpha^T \partial \phi}|_{\tilde{\theta}})^T; \\
\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \phi \partial \beta}|_{\tilde{\theta}} &= (\frac{\partial^2 l_{y,D}(\alpha,\beta,\phi)}{\partial \beta^T \partial \phi}|_{\tilde{\theta}})^T
\end{aligned}$$

Detailed derivations have been provided in Supplementary Material S5 in Supplementary Information File.

Appendix E. Analytical Formula of $s_{y,D}(\alpha, \beta)$

$$s_{y,D}(\alpha,\beta) = \sum_{i=1}^{N}[\{p_\theta(y_i, D_i|x_i)\}^{-1}\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\begin{bmatrix} \frac{y_i - b'(\eta_i)}{a} x_i^T \\ \frac{y_i - b'(\eta_i)}{a} g^T \end{bmatrix} h(g, D_i)],$$

where $\eta_i = \eta(x_i, g) = \alpha x_i^T + \beta g^T$. Detailed derivations have been provided in Supplementary Material S6 in Supplementary Information File.

Appendix F. Analytical Formula of $o_{y,D}(\alpha, \beta)$

$$o_{y,D}(\alpha,\beta) = -\begin{bmatrix} \frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial \alpha^T \partial \alpha} & \frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial \alpha^T \partial \beta} \\ \frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial \beta^T \partial \alpha} & \frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial \beta^T \partial \beta} \end{bmatrix}.$$

$$\begin{aligned}
\frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\alpha^T\partial\alpha} &= \sum_{i=1}^{N}(-\{p_\theta(y_i,D_i|x_i)\}^{-2})[\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a}x_i^T h(g,D_i)][\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a}x_i h(g,D_i)] \\
&\quad +\{p_\theta(y_i,D_i|x_i)\}^{-1}\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)[\frac{(y_i-b'(\eta_i))^2}{a^2}-\frac{b''(\eta_i)}{a}]x_i^T x_i h(g,D_i); \\
\frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\alpha^T\partial\beta} &= \sum_{i=1}^{N}(-\{p_\theta(y_i,D_i|x_i)\}^{-2})[\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a}x_i^T h(g,D_i)][\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a}g h(g,D_i)] \\
&\quad +\{p_\theta(y_i,D_i|x_i)\}^{-1}\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)[\frac{(y_i-b'(\eta_i))^2}{a^2}-\frac{b''(\eta_i)}{a}]x_i^T g h(g,D_i); \\
\frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\beta^T\partial\beta} &= \sum_{i=1}^{N}(-\{p_\theta(y_i,D_i|x_i)\}^{-2})[\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a}g^T h(g,D_i)][\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)\frac{y_i-b'(\eta_i)}{a}g h(g,D_i)] \\
&\quad +\{p_\theta(y_i,D_i|x_i)\}^{-1}\sum_{g\in\mathcal{G}} f_\theta(y_i|x_i,g)[\frac{(y_i-b'(\eta_i))^2}{a^2}-\frac{b''(\eta_i)}{a}]g^T g h(g,D_i); \\
\frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\beta^T\partial\alpha} &= (\frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\alpha^T\partial\beta})^T.
\end{aligned}$$

Detailed derivations are in Supplementary Material S7 in Supplementary Information File.

Appendix G. Evaluation of $s_{y,D}(\alpha,\beta)$ at Constrained MLE $\tilde{\theta}$, i.e., $s_{y,D}(\tilde{\alpha},0)$

$$s_{y,D}(\tilde{\alpha},0) = \begin{bmatrix} 0 \\ \sum_{i=1}^{N}\frac{y_i-b'(\tilde{\alpha}x_i^T)}{a}E(g^T|D_i)\end{bmatrix},$$

where $E(g^T|D_i) = \{\sum_{g\in\mathcal{G}} h(g,D_i)\}^{-1}\sum_{g\in\mathcal{G}} g^T h(g,D_i) = (\sum_{g\in\mathcal{G}} g^T h(g,D_i))/(\sum_{g\in\mathcal{G}} h(g,D_i))$ is the posterior expectation of the genotype given sequencing data. Detailed derivations are in Supplementary Material S8 in Supplementary Information File.

Appendix H. Evaluation of $o_{y,D}(\alpha,\beta)$ at Constrained MLE $\tilde{\theta}$, i.e., $o_{y,D}(\tilde{\alpha},0)$

$$o_{y,D}(\tilde{\alpha},0) = -\begin{bmatrix} \frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\alpha^T\partial\alpha}|_{\tilde{\theta}} & \frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\alpha^T\partial\beta}|_{\tilde{\theta}} \\ \frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\beta^T\partial\alpha}|_{\tilde{\theta}} & \frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\beta^T\partial\beta}|_{\tilde{\theta}}\end{bmatrix}.$$

$$\begin{aligned}
\frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\alpha^T\partial\alpha}|_{\tilde{\theta}} &= -\frac{1}{a}\sum_{i=1}^{N} b''(\tilde{\alpha}x_i^T)x_i^T x_i;\ \ \frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\alpha^T\partial\beta}|_{\tilde{\theta}} = -\frac{1}{a}\sum_{i=1}^{N} b''(\tilde{\alpha}x_i^T)x_i^T E(g|D_i) \\
\frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\beta^T\partial\beta}|_{\tilde{\theta}} &= \sum_{i=1}^{N}[\frac{(y_i-b'(\tilde{\alpha}x_i^T))^2}{a^2}(E(g^T g|D_i)-E(g^T|D_i)E(g|D_i))-\frac{b''(\tilde{\alpha}x_i^T)}{a}E(g^T g|D_i)] \\
\frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\beta^T\partial\alpha}|_{\tilde{\theta}} &= (\frac{\partial^2 l_{y,D}(\alpha,\beta)}{\partial\alpha^T\partial\beta}|_{\tilde{\theta}})^T.
\end{aligned}$$

Detailed derivation are in Supplementary Material S9 in Supplementary Information File.

References

1. Illumina_Inc. Next Generation Sequencing (NGS). Available online: https://www.illumina.com/science/technology/next-generation-sequencing.html (accessed on 5 October 2022).
2. Men, A.E.; Wilson, P.; Siemering, K.; Forrest, S. Sanger DNA Sequencing. In *Next Generation Genome Sequencing*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2008; Chapter 1, pp. 1–11.
3. Illumina_Inc. Sequencing Platforms. Available online: https://www.illumina.com/systems/sequencing-platforms.html (accessed on 5 October 2022).
4. Mardis, E.R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.* **2013**, *6*, 287–303. [CrossRef] [PubMed]
5. Shendure, J.; Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **2008**, *26*, 1135–1145. [CrossRef] [PubMed]
6. Liu, L.; Li, Y.; Li, S.; Hu, N.; He, Y.; Pong, R.; Lin, D.; Lu, L.; Law, M. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012**, *2012*, 251364. [CrossRef] [PubMed]
7. Long, K.; Cai, L.; He, L. DNA sequencing data analysis. In *Computational Systems Biology*; Springer: Cham, Swizterland, 2018; pp. 1–13.
8. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **2013**, *43*, 11.10.1–11.10.33. [CrossRef] [PubMed]
9. Moore, J.H.; Asselbergs, F.W.; Williams, S.M. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* **2010**, *26*, 445–455. [CrossRef]
10. Lewis, C.M.; Knight, J. Introduction to genetic association studies. *Cold Spring Harb. Protoc.* **2012**, *2012*, pdb-top068163. [CrossRef]
11. Balding, D.J. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **2006**, *7*, 781–791. [CrossRef]
12. Hardle, W.; Werwatz, A.; Muller, M.; Sperlich, S. *Nonparametric and Semiparametric Models*; Springer Series in Statistics; Springer: Cham, Switzerland, 2004.
13. Hastie, T.; Tibshirani, R.; Friedman, J. *Elements of Statistical Learning*; Springer Series in Statistics; Springer: Cham, Switzerland, 2009.
14. Breiman, L.; Friedman, J. Estimating Optimal Transformations for Multiple Regression and Correlation. *J. Am. Stat. Assoc.* **1985**, *80*, 580–598. [CrossRef]
15. Tibshirani, R. Estimating Transformations for Regression via Additivity and Variance Stabilization. *J. Am. Stat. Assoc.* **1986**, *83*, 394–405. [CrossRef]
16. Shah, Y.S.; Hernandez-Garcia, L.; Jahanian, H.; Peltier, S.J. Support vector machine classification of arterial volume-weighted arterial spin tagging images. *Brain Behav.* **2016**, *83*, e00549. [CrossRef]
17. Xu, Z. Association Testing of a Group of Genetic Markers Based on Next-Generation Sequencing Data and Continuous Response Using a Linear Model Framework. *Mathematics* **2023**, *11*, 1285. [CrossRef]
18. Lee, S.; Abecasis, G.R.; Boehnke, M.; Lin, X. Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Genet.* **2014**, *95*, 5–23. [CrossRef] [PubMed]
19. Via, M.; Gignoux, C.; Burchard, E.G. The 1000 Genomes Project: New opportunities for research and social challenges. *Genome Med.* **2010**, *2*, 1–3. [CrossRef] [PubMed]
20. Luo, L.; Boerwinkle, E.; Xiong, M. Association studies for next-generation sequencing. *Genome Res.* **2011**, *21*, 1099–1108. [CrossRef]
21. Auer, P.L.; Lettre, G. Rare variant association studies: Considerations, challenges and opportunities. *Genome Med.* **2015**, *7*, 1–11. [CrossRef]
22. Lin, W.Y. Beyond rare-variant association testing: Pinpointing rare causal variants in case-control sequencing study. *Sci. Rep.* **2016**, *6*, 1–13. [CrossRef]
23. Zhao, J.; Akinsanmi, I.; Arafat, D.; Cradick, T.; Lee, C.M.; Banskota, S.; Marigorta, U.M.; Bao, G.; Gibson, G. A burden of rare variants associated with extremes of gene expression in human peripheral blood. *Am. J. Hum. Genet.* **2016**, *98*, 299–309. [CrossRef]
24. Liu, D.J.; Leal, S.M. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.* **2010**, *6*, e1001156. [CrossRef]
25. 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]
26. Lee, S.; Wu, M.C.; Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **2012**, *13*, 762–775. [CrossRef]
27. Plagnol, V.; Cooper, J.D.; Todd, J.A.; Clayton, D.G. A method to address differential bias in genotyping in large-scale association studies. *PLoS Genet.* **2007**, *3*, e74. [CrossRef] [PubMed]
28. Sham, P.C.; Purcell, S.M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **2014**, *15*, 335–346. [CrossRef]
29. Skotte, L.; Korneliussen, T.S.; Albrechtsen, A. Association testing for next-generation sequencing data using score statistics. *Genet. Epidemiol.* **2012**, *36*, 430–437. [CrossRef]
30. Yan, S.; Yuan, S.; Xu, Z.; Zhang, B.; Zhang, B.; Kang, G.; Byrnes, A.; Li, Y. Likelihood-based complex trait association testing for arbitrary depth sequencing data. *Bioinformatics* **2015**, *31*, 2955–2962. [CrossRef] [PubMed]
31. Korneliussen, T.S.; Albrechtsen, A.; Nielsen, R. ANGSD: Analysis of next generation sequencing data. *BMC Bioinform.* **2014**, *15*, 1–13. [CrossRef] [PubMed]

32. Belonogova, N.M.; Svishcheva, G.R.; Axenovich, T.I. FREGAT: An R package for region-based association analysis. *Bioinformatics* **2016**, *32*, 2392–2393. [CrossRef]
33. McCullagh, P.; Nelder, J. *Generalized Linear Models*, 2nd ed.; Monographs on Statistics and Applied Probability; Chapman and Hall: Boca Raton, FL, USA; CRC Press: Boca Raton, FL, USA, 1989.
34. Dobson, A.; Barnett, A. *Introduction to Generalized Linear Models*, 4th ed.; Chapman and Hall: Boca Raton, FL, USA; CRC Press: Boca Raton, FL, USA, 2018. [CrossRef]
35. Cox, D.R. *Principles of Statistical Inference*; Cambridge University Press: Cambridge, UK, 2006.
36. Young, G.A.; Smith, R.L. *Essentials of Statistical Inference*; Cambridge University Press: Cambridge, UK, 2005; Volume 16.
37. Sul, J.H.; Han, B.; He, D.; Eskin, E. An optimal weighted aggregated association test for identification of rare variants involved in common diseases. *Genetics* **2011**, *188*, 181–188. [CrossRef]
38. Ionita-Laza, I.; Buxbaum, J.D.; Laird, N.M.; Lange, C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.* **2011**, *7*, e1001289. [CrossRef]
39. Schaffner, S.F.; Foo, C.; Gabriel, S.; Reich, D.; Daly, M.J.; Altshuler, D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **2005**, *15*, 1576–1583. [CrossRef]
40. Kang, J.; Huang, K.C.; Xu, Z.; Wang, Y.; Abecasis, G.R.; Li, Y. AbCD: Arbitrary coverage design for sequencing-based genetic studies. *Bioinformatics* **2013**, *29*, 799–801. [CrossRef]
41. Byrska-Bishop, M.; Evani, U.; Zhao, X.; Basile, A.; Abel, H.; Regier, A.; Corvelo, A.; Clarke, W.; Musunuri, R.; Nagulapalli, K.; et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **2022**, *185*, 3426–3440.e19. [CrossRef] [PubMed]
42. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve years of SAMtools and BCFtools. *GigaScience* **2021**, *10*, giab008. [CrossRef] [PubMed]
43. Shao, J. *Mathematical Statistics*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2003.
44. Agresti, A. *Categorical Data Analysis*; Wiley Series in Probability and Statistics; Wiley and Sons: Hoboken, NJ, USA, 2013.
45. Liu, H.M.; Zheng, J.P.; Yang, D.; Liu, Z.F.; Li, Z.; Hu, Z.Z.; Li, Z.N. Recessive/dominant model: Alternative choice in case-control-based genome-wide association studies. *PLoS ONE* **2021**, *16*, e0254947. [CrossRef] [PubMed]