OPEN ACCESS Freely available online

PLoS GENETICS

Hierarchical Generalized Linear Models for Multiple Groups of Rare and Common Variants: Jointly Estimating Group and Individual-Variant Effects

Nengjun Yi*, Nianjun Liu, Degui Zhi, Jun Li

Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, Birmingham, Alabama, United States of America

Abstract

Complex diseases and traits are likely influenced by many common and rare genetic variants and environmental factors. Detecting disease susceptibility variants is a challenging task, especially when their frequencies are low and/or their effects are small or moderate. We propose here a comprehensive hierarchical generalized linear model framework for simultaneously analyzing multiple groups of rare and common variants and relevant covariates. The proposed hierarchical generalized linear models introduce a group effect and a genetic score (i.e., a linear combination of main-effect predictors for genetic variants) for each group of variants, and jointly they estimate the group effects and the weights of the genetic scores. This framework includes various previous methods as special cases, and it can effectively deal with both risk and protective variants in a group and can simultaneously estimate the cumulative contribution of multiple variants and their relative importance. Our computational strategy is based on extending the standard procedure for fitting generalized linear models in the statistical software R to the proposed hierarchical models, leading to the development of stable and flexible tools. The methods are illustrated with sequence data in gene *ANGPTL4* from the Dallas Heart Study. The performance of the proposed procedures is further assessed via simulation studies. The methods are implemented in a freely available R package BhGLM (http://www.ssg.uab.edu/bhglm/).

Citation: Yi N, Liu N, Zhi D, Li J (2011) Hierarchical Generalized Linear Models for Multiple Groups of Rare and Common Variants: Jointly Estimating Group and Individual-Variant Effects. PLoS Genet 7(12): e1002382. doi:10.1371/journal.pgen.1002382

Editor: Nicholas J. Schork, University of California San Diego and The Scripps Research Institute, United States of America

Received June 23, 2011; **Accepted** September 29, 2011; **Published** December 1, 2011

Funding: This work was supported in part by NIH research grants: 5R01GM069430-07, R01GM081488, and R00 RR024163. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nyi@ms.soph.uab.edu

Introduction

Many common human diseases and complex traits are highly heritable and are believed to be influenced by multiple genetic and environmental factors. Genome-wide association studies (GWAS) represent a powerful way for discovering disease-associated factors and investigating the genetic architecture of complex diseases [1]. In the past few years, these studies have identified hundreds of common variants (i.e., genetic variants with minor allele frequency (MAF) $>\sim$5%) associated with complex diseases [2]. However, the estimated effect sizes for the identified variants are small (most odds ratios are below 1.5) and explain only a small proportion of the heritability of complex diseases [2,3], motivating research interest in finding 'missing' genetic factors that contribute to the remaining heritability [4,5]. Many explanations for the missing heritability have been suggested [4,5]; one is that many common variants with much smaller effects are yet to be detected, and another is the possible contribution of rare variants (MAF $<$0.5% or 1%) that are poorly captured by previous GWA genotyping arrays. Empirical studies and population genetics theory support the potentially important role of both rare variants and common variants of very small effects [6–10]. Several current studies have implicated association of rare variants with complex diseases and traits [11–18].

Next-generation sequencing technologies have provided unparalleled tools to sequence a large number of individuals in candidate genes, exomes, or even the entire genome, allowing for comprehensive studies of both common and rare variants. In addition to the common problems of handling large numbers of variants, however, detecting disease-associated rare variants and common variants of small effects poses unique statistical challenges [19,20]. As such variants individually contain little variation, statistical methods that detect association between a single variant and disease phenotype provide low power with realistic sample sizes. Therefore, it is necessary to develop sophisticated methods that can effectively combine information across variants and assess the collective effect of multiple variants [4].

Several approaches along this line have been proposed [19,20]. The basic procedure of these methods is to construct a linear combination of multiple variants with fixed weights to summarize the information across the variants and then estimate its association with the phenotype [21–25]. Different weights yield different summaries of the variants and implicate different assumptions about the relative importance of individual variants [24,26]. Further, they implicitly assume that all variants affect phenotype in the same direction. However, there are many examples in which numerous rare variants detected in a gene or region may have disparate or even opposite effects on phenotype [4,11]. Thus, these methods can be suboptimal if the data do not follow the underlying assumptions. Recently, several methods have been proposed to deal with variants with opposite effects [26–32], and to summarize the information across variants using non-linear functions [33,34].

Author Summary

Complex diseases and traits are likely influenced by many common and rare genetic variants and environmental factors. Next-generation sequencing technologies have provided unparalleled tools to sequence a large number of individuals, allowing for comprehensive studies of both common and rare variants. However, detecting disease-associated rare variants and common variants of small effects poses unique statistical challenges. We propose here a comprehensive hierarchical generalized linear model framework for simultaneously analyzing multiple groups of rare and common variants and relevant covariates. The proposed hierarchical generalized linear models introduce a group effect and a genetic score for each group of variants, and jointly they estimate the group effects and the weights of the genetic scores. This framework includes various previous methods as special cases, and it can effectively deal with both risk and protective variants in a group and can simultaneously estimate the cumulative contribution of multiple variants and their relative importance. The methods are illustrated with sequence data in gene *ANGPTL4* from the Dallas Heart Study and are further assessed via simulation studies. The methods have been implemented in a freely available R package BhGLM (http://www.ssg.uab.edu/bhglm/).

All the existing methods have been developed to assess only one group of variants at a time. Since common diseases are likely caused by a complex interplay among many genes and environmental factors, however, it is more appropriate to simultaneously model multiple groups of variants and covariates [19]. The joint analyses would improve the power of detecting causal effects and hence lead to increased understanding about the genetic architecture of diseases. Such methods are also advantageous for studies involving only one candidate gene, because numerous variants detected within a gene can be divided into multiple groups based on their allelic frequencies (common or rare) and functional annotations of the genomic regions they reside in (for example, non-synonymous or synonymous). It has been found in GWAS that the vast majority (80%) of associated variants fall outside coding regions, emphasizing the importance of including both coding and non-coding regions in the search for disease-associated variants [2].

We propose here a comprehensive hierarchical generalized linear model (GLM) framework for simultaneously analyzing multiple groups of rare and common variants and relevant covariates. The proposed hierarchical GLMs introduce a group effect and a genetic score (i.e., a linear combination of main-effect predictors for genetic variants) for each group of variants, and jointly estimate the group effects and the weights of the genetic scores. This framework includes various previous methods as special cases, and can effectively deal with both risk and protective variants in a group and can simultaneously estimate the cumulative contribution of multiple variants and their relative importance. The methods are illustrated with sequence data in gene *ANGPTL4* from the Dallas Heart Study, and are further assessed via simulation studies. Finally, we conclude this article by highlighting some areas of future research.

Methods

Hierarchical GLMs for Multiple Groups of Rare and Common Variants

Suppose that a population-based association study consists of n unrelated individuals, phenotyped for a continuous or discrete disease trait and genotyped for a number of rare and/or common genetic variants in one or multiple candidate genes or genomic regions. The observed values of the response variable are denoted by $y=(y_1, \cdots, y_n)$. We assume that the genetic variants can be divided into K groups, G_k, $k=1, \cdots, K$, and the k-th group G_k contains J_k variants, where $K\geq 1$ and $J_k>1$. The groups can be constructed based on candidate genes in which the variants are located and the types of the variants (e.g., common variants, rare non-synonymous or synonymous coding variants). We assume that some non-genetic variables (e.g., gender indicator, age, etc.) are also measured for each individual and will be included as covariates in the model to control for possible confounding effects.

We extend the hierarchical generalized linear model (GLM) of Yi and Zhi [26] to simultaneously fit covariates and multiple groups of rare and common variants. A generalized linear model consists of three components: the linear predictor η, the link function h, and the data distribution p [35,36]. The linear predictor of individual i is expressed as the multiplicative form:

$$\eta_i = \sum_{j=0}^{J_0} x_{ij}\beta_j + \sum_{k=1}^{K} g_k \left(\sum_{j\in G_k}^{J_k} \alpha_j z_{ij} \right) \quad (1)$$

where β_0 is the intercept, x_{ij} and β_j represent covariate j and its coefficient, respectively, z_{ij} is the main-effect predictor for individual i at genetic variant j in group G_k, equaling to the number of minor alleles for an additive coding (for a rare variant, the additive coding is approximately equivalent to a dominant coding because the frequency of the minor allele is very low), the common coefficient g_k represents the *group effect* for J_k variants in the k-th group, and the individual coefficients α_j can be interpreted as the *weights* or *relative effects* of individual variants.

The common coefficient g_k represents the association between the phenotype and the linear combination $\sum_{j\in G_k}^{J_k} \alpha_j z_{ij}$ of J_k individual main-effect predictors for variants in group G_k. The linear combination provides a way to combine the genetic variation across the J_k individual variants, referred to as *genetic score*. Therefore, the common coefficient g_k represents the cumulative importance of the J_k individual variants in the k-th group, hence referred to as the *group effect*, and the weights α_j, $j\in G_k$, give the relative importance of the individual variants in group G_k.

The mean of the response variable is related to the linear predictor via a link function h:

$$E(y_i|\eta_i) = h^{-1}(\eta_i) \quad (2)$$

The data distribution (likelihood) is expressed as

$$p(y|\eta,\phi) = \prod_{i=1}^{n} p(y_i|\eta_i,\phi) \quad (3)$$

where ϕ is a dispersion (or variance) parameter, and the distribution $p(y_i|\eta_i,\phi)$ can take various forms, including Normal, Gamma, Binomial, and Poisson distributions.

Our main goal is to estimate the group effects g_k and to test the hypotheses $g_k=0$, $k=1, \cdots, K$. We treat the weights α_j's as unknown parameters and estimate them along with the group effects and other parameters from the data. But we cannot simply use classical framework (equivalent to setting uniform distributions on the α_j's from a Bayesian perspective), since this would result in a nonidentifiable model [37,38]. An approach to overcoming the problem is to use an informative prior for α_j. We use the following

hierarchical prior distribution:

$$\begin{aligned} \alpha_j|\tau^2_{\alpha_j} &\sim N\left(\mu_j,\tau^2_{\alpha_j}\right) \\ \tau^2_{\alpha_j}\left|s^2_{\alpha_j}\right. &\sim \text{Inv}-\chi^2\left(1,s^2_{\alpha_j}\right) \\ s^2_{\alpha_j}\left|b_{k[j]}\right. &\sim \text{Gamma}\left(0.5,b_{k[j]}\right) \\ p(\log b_k) &\propto 1 \end{aligned} \tag{4}$$

where the prior means μ_j are prefixed and will be discussed in detail later, and the subscript $k[j]$ indexes the group k that variant j belongs to.

The above hierarchical prior assumes that α_j follows a scale mixture of normals with unknown variable-specific variance $\tau^2_{\alpha_j}$. The prior distribution for τ_{α_j} is a hierarchical formulation of the half-Cauchy distribution, which has desirable properties, such as an infinite spike at the prior mean and very heavy tails, and also facilitates efficient computation [39,40]. An attractive feature of our hierarchical prior is that it is free of user-chosen tuning parameters and introduces group-specific parameters b_k and variable-specific parameters τ_{α_j} and s_{α_j}. The group-specific parameters provide a way to pool the information among variables within a group and also to induce different shrinkage for different groups, while the variable-specific parameters allow different shrinkage for different variables. Yi and Zhi [26] set the scale parameters s_{α_j} to a known value for all the weight parameters and recommended $s_{\alpha_j}=0.5$ as default. However, it would be more reasonable to estimate the scale parameters from the data.

If the number of groups is not large, the group effects g_k usually can be estimated classically. However, low allelic frequencies can yield very small variances for the predictors of g_k, i.e., $\sum_{j\in G_k}^{J_k}\alpha_j z_{ij}$, and as a result the classical procedure can result in numerically instable estimates for the group effects g_k. To overcome this problem, we can place a weakly informative prior on g_k that constrains g_k to a reasonable range [41]. We use the following hierarchical prior distribution:

$$\begin{aligned} g_k|\tau^2_{g_k} &\sim N\left(0,\tau^2_{g_k}\right) \\ \tau^2_{g_k}\left|s^2_{g_k}\right. &\sim \text{Inv}-\chi^2\left(1,s^2_{g_k}\right) \\ s^2_{g_k} &\sim \text{Gamma}(0.5,0.5) \end{aligned} \tag{5}$$

This hierarchical prior distribution includes group-specific parameters $s^2_{g_k}$, which can induce different shrinkage for different group effects g_k. The group-specific parameters $s^2_{g_k}$ are assumed to follow a weakly informative prior Gamma(0.5, 0.5). This weakly informative prior does not strongly shrink g_k towards zero, but can constrain g_k to lie in a reasonable range [41].

For the covariate effects β_j, we also use the above weakly informative prior (5), i.e., $\beta_j \sim N(0,\tau^2_{\beta_j}), \tau^2_{\beta_j} \sim \text{Inv-}\chi^2(1,s^2_{\beta_j}), s^2_{\beta_j} \sim \text{Gamma}(0.5, 0.5)$. For the intercept β_0 and the dispersion parameter ϕ, we can use any reasonable non-informative prior distributions; for example, $p(\beta_0)=N(0,\tau^2_0)$ with τ^2_0 set to a large value, and $p(\log\phi)\propto 1$.

Model Interpretation

Our hierarchical GLMs include multiplicative parameters, a common coefficient g_k for a group of variants and a weight parameter α_j for each variant. As explained earlier, the common coefficient g_k represents the overall association of the J_k individual variants in group k with the disease. In our hierarchical model, the multiplicative term $g_k\sum_{j\in G_k}^{J_k}\alpha_j z_{ij}$ can be expressed as $\sum_{j\in G_k}^{J_k}(g_k\alpha_j)z_{ij}$, and thus the predictor z_{ij} ultimately gets the coefficient $g_k\alpha_j$, which represents the main effect of that variant. The coefficient $g_k\alpha_j$ is affected by the prior mean of α_j. Therefore, we define the adjusted main effects as $g_k(\alpha_j-\mu_j)$, which represent the effects of individual variants.

For the multiplicative model to be useful, we need informative prior distributions on the multiplicative parameters that allow us to distinguish between the group effects and the individual weights. The prior means μ_j and the variances $\tau^2_{\alpha_j}$ in the normal prior distributions of the weights α_j (i.e., $\alpha_j \sim N(\mu_j, \tau^2_{\alpha_j})$) are the key components to interpret our hierarchical model. The variances $\tau^2_{\alpha_j}$ directly control the amount of shrinkage for α_j. If $\tau^2_{\alpha_j}=0$, the coefficient α_j equals the prior mean μ_j. If $\tau^2_{\alpha_j}=\infty$, $g_k\alpha_j$ is actually estimated using least squares and the parameters g_k and α_j cannot be distinguished. If $\tau^2_{\alpha_j}$ is finite, the coefficient α_j is shrunk towards but not identical to the prior mean μ_j. Therefore, the prior distributions bridge the gap between the two extremes of simply using the fixed weighted sum $\sum_{j\in G_k}^{J_k}\mu_j z_{ij}$ of the J_k variants as a predictor ($\tau^2_{\alpha_j}=0$), and including them as J_k independent predictors ($\tau^2_{\alpha_j}=\infty$) [37,38]. This interpretation can be more explicitly understood by the identity

$$g_k\sum_{j\in G_k}^{J_k}\alpha_j z_{ij} = g_k\left(\sum_{j\in G_k}^{J_k}\mu_j z_{ij}+\sum_{j\in G_k}^{J_k}\alpha^*_j z_{ij}\right) \tag{6}$$

where $\alpha^*_j=\alpha_j-\mu_j \sim N(0,\tau^2_{\alpha_j})$. The second term in the right side is controlled by the variances $\tau^2_{\alpha_j}$, and represents the deviation from the fixed weighted sum $\sum_{j\in G_k}^{J_k}\mu_j z_{ij}$. Most of existing methods for analyzing rare variants proceed to construct a linear combination (genetic score) of rare variants with fixed weights [21–25], and thus can be viewed as special cases of our model.

The prior means μ_j represent the prior relative importance of the individual variants and can be specified in various ways. The weights proposed by previous methods [21–25] can be used as the prior means μ_j in our hierarchical model. The simplest way is to set all $\mu_j=1$, resulting in the simple sum $\sum_{j\in G_k}^{J_k} z_{ij}$, and incorporating no prior information about the relative importance of rare variants into the model. But our method can estimate the weights from data and produce different weights to different variants based on their contributions to the phenotype. Therefore, our model uses a previous score (i.e., $\sum_{j\in G_k}^{J_k}\mu_j z_{ij}$) as the baseline, and improves the fit by accounting for the variation among individual variants.

An alternative choice of the prior means is to set all $\mu_j=0$. With this choice, the weights are shrunk towards zero, and variants with small effects can be essentially removed from the model. This seems to be reasonable, especially for the situations with non-functional variants. However, we don't recommend this approach for rare variants for several reasons. First, most of rare and common variants have small effects, but they can be cumulatively important. In order to detect the cumulative effect, therefore, it would be better to include all the small effects in the model.

Second, the estimated group effect can be less interpretable and accurate, if only one or a few variants are included in the model. Third, our hierarchical model can estimate the weights of individual variants from the data, and thus can deal with non-functional variants and disparate effects.

Model Fitting and Inference

Our Bayesian hierarchical GLMs can be fitted using Markov chain Monte Carlo (MCMC) algorithms that fully explore the joint posterior distribution by alternately sampling each parameter from its conditional posterior distribution [36]. However, it is desirable to have a faster computation that provides a point estimate (i.e., the posterior mode) of the coefficients and their standard errors (and thus the p-values). Such an approximate calculation has been routinely applied in statistical practice [41]. We develop our mode-finding algorithm by modifying the standard iterative weighted least squares (IWLS) for fitting classical generalized linear models [42,43].

Our algorithm updates the coefficients α_j and g_k using an iterative procedure. Conditional on the current estimates $\hat{\boldsymbol{g}}_k$, we update α_j by running the generalized linear model with the proposed prior distributions for α_j and other corresponding parameters:

$$\eta_i = \sum_{j=0}^{J_0} x_{ij}\beta_j + \sum_{k=1}^{K}\sum_{j\in G_k}^{J_k} \hat{z}_{ij}\alpha_j \tag{7}$$

where $\hat{z}_{ij} = z_{ij}\hat{\boldsymbol{g}}_k$, and then conditional on the current estimates $\hat{\boldsymbol{\alpha}}_j$, we update g_k by running the generalized linear model with the proposed prior distributions for g_k and other corresponding parameters:

$$\eta_i = \sum_{j=0}^{J_0} x_{ij}\beta_j + \sum_{k=1}^{K} \hat{\boldsymbol{T}}_{ik} g_k \tag{8}$$

where $\hat{\boldsymbol{T}}_{ik} = \sum_{j\in G_k}^{J_k} \hat{\alpha}_j z_{ij}$. We fit these two hierarchical generalized linear models by incorporating a flexible expectation-maximization (EM) algorithm into the iteratively weighted least squares (IWLS) for fitting classical generalized linear models. We describe our EM-IWLS algorithm in detail in Text S1.

We initialize our iterative algorithm by setting the parameters $(\beta,\alpha,g,\phi,\tau^2,s^2,b)$ with some plausible values. For example, we can start with $\beta_j = 0$, $\alpha_j = \mu_j$, $g_k = 1$, $\phi = 1$, $b_k = 0.5$, $s_{\alpha_j} = \tau_{\alpha_j} = 0.5$, and $s_{g_k} = \tau_{g_k} = s_{\beta_j} = \tau_{\beta_j} = 1$. We then update the parameters by iteratively running the hierarchical generalized linear models (7) and (8) until convergence. Instead of doing a nested converged EM-IWLS for each of the two models, we can run one step of the EM-IWLS at each iteration, thus taking less computing time to ultimately achieve convergence by not wasting time running many steps of the EM-IWLS within each iteration. To assess convergence, we use the standard criterion for analysis of classical generalized linear models (as implemented in the R function glm), i.e., $\left|d^{(t)} - d^{(t-1)}\right| / \left(0.1 + \left|d^{(t)}\right|\right) < \varepsilon$, where $d^{(t)}$ is the estimate of deviance (i.e., $-2\log p(y|\eta,\phi)$) at the t^{th} iteration, and ε is a small value (say 10^{-5}).

At convergence of the algorithm, we summarize the inferences using the latest estimates of the coefficients $(\hat{\beta},\hat{\alpha},\hat{g})$ and their standard errors. Based on these outputs, we can calculate approximate p-values as in the classical framework for testing whether a coefficient is significantly different from zero, for example, the hypothesis $g_k = 0$. The adjusted main effects $g_k(\alpha_j - \mu_j)$ are then estimated as $\hat{g}_k(\hat{\alpha}_j - \mu_j)$, and the approximate standard error for $\hat{g}_k(\hat{\alpha}_j - \mu_j)$ can be obtained by using the delta technique:

$$\mathrm{Var}(\hat{g}_k(\hat{\alpha}_j - \mu_j)) = \hat{g}_k^2 \mathrm{Var}(\hat{\alpha}_j) + (\hat{\alpha}_j - \mu_j)^2 \mathrm{Var}(\hat{g}_k) \tag{9}$$

Therefore, we can calculate the approximate p-value for testing the hypothesis $g_k(\alpha_j - \mu_j) = 0$.

Implementation

Our model fitting strategy is based on extending the well-developed IWLS algorithm for fitting classical GLMs to our Bayesian hierarchical GLMs. The IWLS algorithm is executed in the glm function in R (http://www.r-project.org/). We have implemented the EM-IWLS algorithm by inserting the E-step for updating the missing values (i.e., the variances τ_j^2 and the hyperparameters s_j^2 and $b_{k[j]}$) and the steps for calculating the augmented data and the dispersion parameter into the IWLS procedure (see Text S1). We have created a new R function bglm by modifying the glm function to implement our EM-IWLS algorithm that estimates posterior modes and standard deviations for hierarchical GLMs with the prior distributions proposed here (see Text S1) and some other hierarchical priors [44,45]. We have also developed an R function bglm.ex that implements the iterative algorithm described above for fitting our hierarchical multiplicative GLMs. Although described in the context of genetic variants in this paper, the functions bglm and bglm.ex can be used as general tools for routine data analysis using hierarchical GLMs. We have incorporated the functions bglm and bglm.ex into the freely available R package BhGLM (http://www.ssg.uab.edu/bhglm/) that is an extensible and interactive environment for genetic association analysis of common and rare variants and gene-gene and gene-environment interactions.

Alternative Approaches

Our hierarchical multiplicative GLMs include various models as special cases. Although less comprehensive, these reduced models can be useful in some situations, and thus can be used as alternative approaches to analysis of multiple groups of rare and common variants. We here consider two types of reduced models. The first ignores the group effects and directly models the main effects of individual variants. Thus, the linear predictor (1) is reduced to

$$\eta_i = \sum_{j=0}^{J_0} x_{ij}\beta_j + \sum_{k=1}^{K}\sum_{j\in G_k}^{J_k} z_{ij}\alpha_j, \tag{10}$$

and the mean and the distribution of the response variable take the same form of the expressions (2) and (3). In this model, the coefficient α_j represents the main effect of genetic variant j, and follows the hierarchical prior distribution (4) with the prior mean $\mu_j = 0$. This approach can only detect individual variants with strong effects, and is less powerful in situations where the effects of all individual variants are small but they are cumulatively significant.

The second alternative approach is to preset the weights of individual variants using the previous methods [21–25]. Thus, the linear predictor (1) becomes

$$\eta_i = \sum_{j=0}^{J_0} x_{ij}\beta_j + \sum_{k=1}^{K} T_{ik} g_k \tag{11}$$

where $T_{ik}=\sum_{j\in G_k}^{J_k} z_{ij}\mu_j$ with fixed weights μ_j. This model is equivalent to setting the priors as $\alpha_j|\tau^2_{\alpha_j}\sim N(\mu_j,\tau^2_{\alpha_j})$ and $\tau^2_{\alpha_j}\sim$Inv-$\chi^2(+\infty,0)$ (i.e., $\alpha_j\sim N(\mu_j,0)$) and thus is a special case of our hierarchical model. The performance of this method heavily depends on the quality of the fixed weights.

Results

Application: Population-Based Resequencing of *ANGPTL4* and Triglycerides

Description of dataset. Romeo et al. [13] was the first application of resequencing to a large population to examine the role of the adipokine gene *ANGPTL4* in lipid metabolism. The study included the 3,551 participants of the Dallas Heart Study (DHS) from whom fasting venous blood samples were obtained. The DHS is a population-based random sample of Dallas County residents, consisting of 601 Hispanic (H), 1,830 African American (AA), 1,045 European American (EA) and 75 other ethnicities. The 75 participants from other ethnicities will be excluded from our analysis. The phenotype analyzed in our study is the log-transformed plasma levels of triglyceride. The top panel of Figure 1 shows the histogram of this continuous phenotype and the 25th and 75th percentiles. Following the analysis of Romeo et al. [13], we also considered a binary trait, coding individuals in the bottom and top quartiles of the distribution as 0 and 1, respectively, and excluding other individuals from the analysis. Hereafter, we refer these two phenotypes as the *continuous* and *binary* traits. Our analyses included race (a three-level factor), age, sex, and BMI as covariates in the model. We excluded individuals with any missing values of the covariates from the analysis, resulting in 3008 and 1499 individuals in the analyses of the continuous and binary traits, respectively.

Romeo et al. [13] sequenced the seven exons and the intron-exon boundaries of the gene *ANGPTL4*, and identified a total of 93 sequence variations. After removing variants that were not segregating in the sample, the numbers of variants reduced to 82 and 63 for the analyses of the continuous and binary traits, respectively. Most of these variants were rare: only 12 and 13 variants had a minor allele frequency above 1%, and 33 and 26 variants were found only in one object in the two analyses, respectively (see Figure 1).

The methods. We divided the variants into four groups: common non-synonymous, common synonymous, rare non-synonymous, and rare synonymous. We used a minor allele frequency of 1% as the threshold to distinguish between common

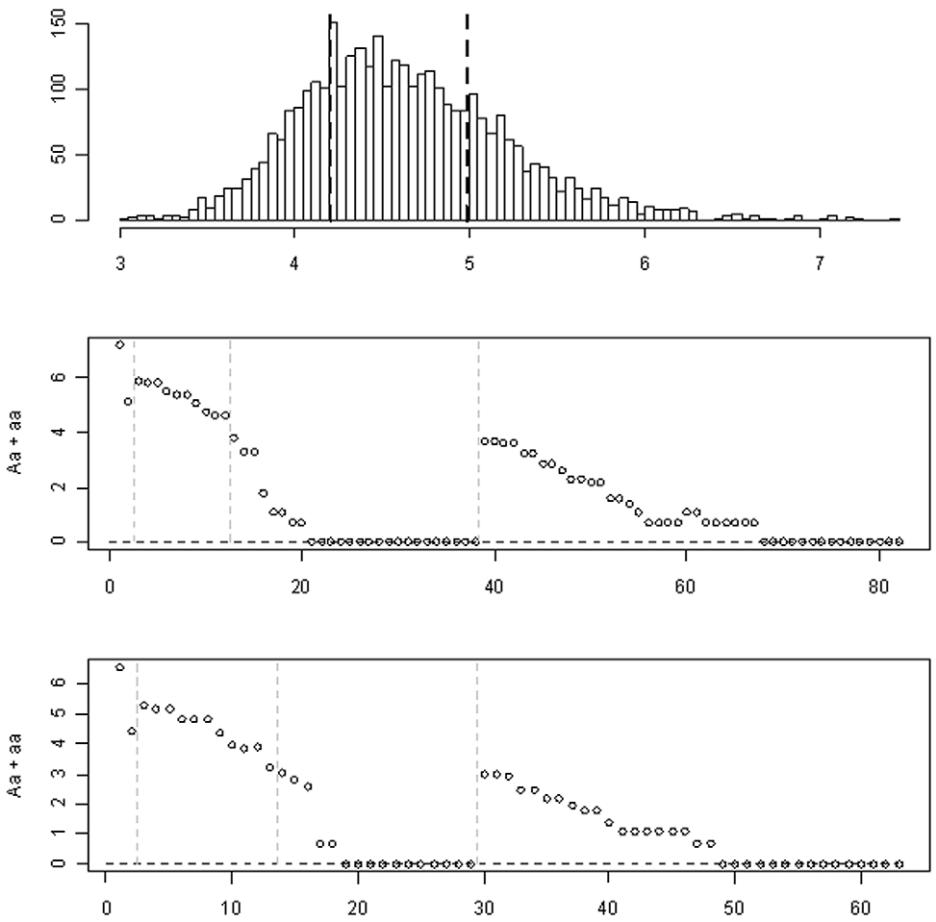

Figure 1. The Dallas Heart Study data set. The top panel: the histogram of the log-transformed plasma levels of triglyceride and the 25th and 75th percentiles (the black dotted lines). The middle panel: the logarithm of the observed count of heterozygotes (Aa) and rare homozygotes (aa) for each variant in the continuous trait analysis. The bottom panel: the logarithm of the observed count of Aa and aa for each variant in the binary trait analysis. The gray dotted lines show the four groups: common non-synonymous, common synonymous, rare non-synonymous, and rare synonymous.
doi:10.1371/journal.pgen.1002382.g001

and rare variants [21]. Figure 1 displays the four groups of variants and the logarithm of the observed count of heterozygotes (Aa) and rare homozygotes (aa) for each variant. The four groups consisted of 2 (2), 10 (11), 26 (16) and 44 (34) variants for the analyses of the continuous (binary) traits, respectively. Since there are only two common non-synonymous variants (i.e., 8155_T266M and 8191_R278Q), we did not estimate their group effect and instead treated them as two covariates in the models. We coded the main-effect predictor of each variant using the additive genetic model, i.e., the number of minor alleles in the observed genotype. The genotypes of the variants contained ~3%–16% missing values. For the missing genotypes, we filled in the variables using the expectation of the observed values in that marker. This simple, but reasonable, imputation method is computationally much more efficient than comprehensive methods using MCMC algorithms or multiple imputations and has been widely used in genetic association studies. The previous studies and the analyses in this work show that this imputation method yields a reasonable result [43].

We first analyzed the data using the hierarchical multiplicative GLMs (Equations 1–3) with the proposed hierarchical prior distributions (Equations 4 and 5). For comparisons, we then used three alternative methods: 1) Setting all the scale parameters s_{α_j} in the hierarchical prior (4) to a known value (e.g., 0.5). This is an extension of Yi and Zhi [26] to multiple groups of variants; 2) Setting the weights of individual variants to fixed values μ_j (see Equation 11). This simply extends the previous Simple-Sum [22,27] and Weighted-Sum methods [24]; 3) Ignoring the group effects and directly estimating the main effects of all individual variants (see Equation 10).

All the analyses simultaneously fitted all the non-genetic variables (i.e., race, age, sex, and BMI), the two common non-synonymous variants (i.e., 8155_T266M and 8191_R278Q) and the three groups of variants. We used a normal regression and a logistic regression for the continuous and the binary traits, respectively. We set the prior means μ_j in two ways; the first is to set 1 for all the variants, and the second is to set 1 for the synonymous variants and the predicted functional scores for the rare non-synonymous variants. The functional scores were calculated using the software PolyPhen [24,46]. The iterative EM-IWLS algorithm started from the plausible initial values described earlier and took 12 (16) iterations to reach convergence for the analysis of the continuous (binary) trait (~0.1 minutes on a P4 desktop computer).

Results of data analyses. Figure 2 shows the results from the analyses of the proposed hierarchical GLMs with prior means μ_j set to 1 for all the grouped variants. All the non-genetic covariates and the common non-synonymous variant 8191_R278Q

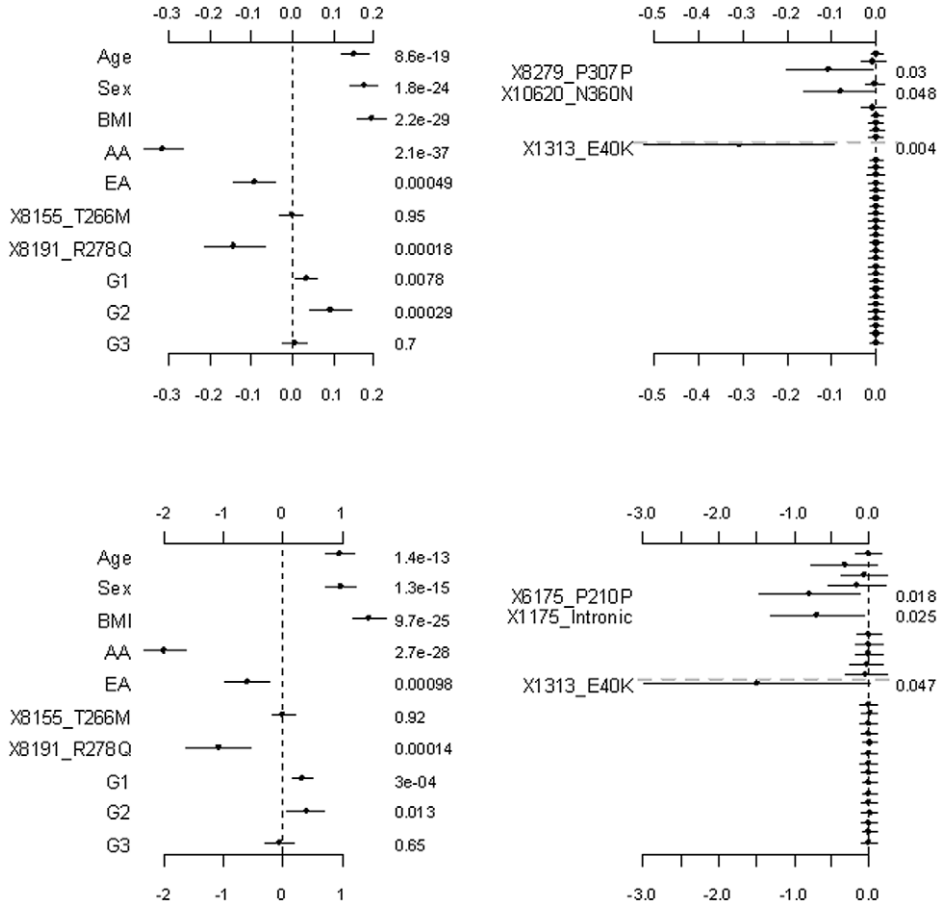

Figure 2. Analyses of the proposed hierarchical GLMs with prior means $\mu_j = 1$ for all variants. The top and bottom panels are for the continuous and binary traits, respectively. The left panel is for the covariates, the two common non-synonymous variants and the three group effects (G1: common synonymous; G2: rare non-synonymous; G3: rare synonymous). The right panel is for the adjusted main effects (the gray dotted line shows the two groups G1 and G2). The points, short lines and numbers at the right side represent estimates of effects, ± 2 standard errors, and p-values, respectively. Only adjusted main effects with p-value <0.1 are labeled.
doi:10.1371/journal.pgen.1002382.g002

were found to significantly affect both the continuous and the binary traits. These effects were also significant under the other models considered (see Figure 3, Figure 4 Figure 5, Figure 6). Although not mentioned in Romeo et al. [13], our analyses detected the minor allele of the variant 8191_R278Q to significantly decrease triglyceride levels, consistent with the finding of King et al. [25]. In addition to the significant non-genetic and genetic covariates, we identified two significant group effects, i.e., those of the common synonymous and the rare non-synonymous variants. These two group effects remained significant even corrected for multiple testing using the method of Benjamini and Hochberg [47]. Our results were fairly consistent with the original analyses; Romeo et al. [13] observed that the number of individuals with nonsynonymous variants in the bottom quartile was significantly greater than the number in the highest quartile, but the number of synonymous variants in the upper and lower tails of the distribution was identical. Since we divided synonymous variants into two groups, our analyses produced additional findings; the group effect of the common synonymous variants was significant, but that of the rare synonymous variants was insignificant.

The group effects in our model should be interpreted with caution; a positive group effect does not necessarily mean that the variants increase the phenotype, because for some variants the weights can be estimated to be negative (for example, the rare variant 1313_E40K in our analyses). The right panel of Figure 2 displays the adjusted main effects of the common synonymous and the rare non-synonymous variants, thus showing which variants are more important. Our analyses identified the rare variant 1313_E40K as the most important. The negative adjusted main effect indicated that the minor allele of 1313_E40K decreases triglyceride levels. Romeo et al. [13] and King et al. [25] also found that the variant 1313_E40K significantly decreased triglyceride levels. Therefore, the proposed method can simultaneously identify significant group effects and individual variants.

Figure 3 shows the results for the proposed hierarchical GLMs with prior means μ_j set to the functional probabilities for the rare non-synonymous variants, which were estimated to range from 0.16 to 1. These models produced qualitatively identical results as the above analyses, but slightly lower p-values for the significant effects. Price et al. [24] showed that incorporating computational predictions of functional importance can increase power for pooled association tests for rare variants.

Figure 4, Figure 5, and Figure 6 display the results from the three alternative approaches. The models setting all the scale parameters s_{α_j} in the hierarchical prior (Equation 4) to 0.5 (as suggested by Yi and Zhi [26]) produced results similar to the previous analyses (Figure 4). However, this alternative method

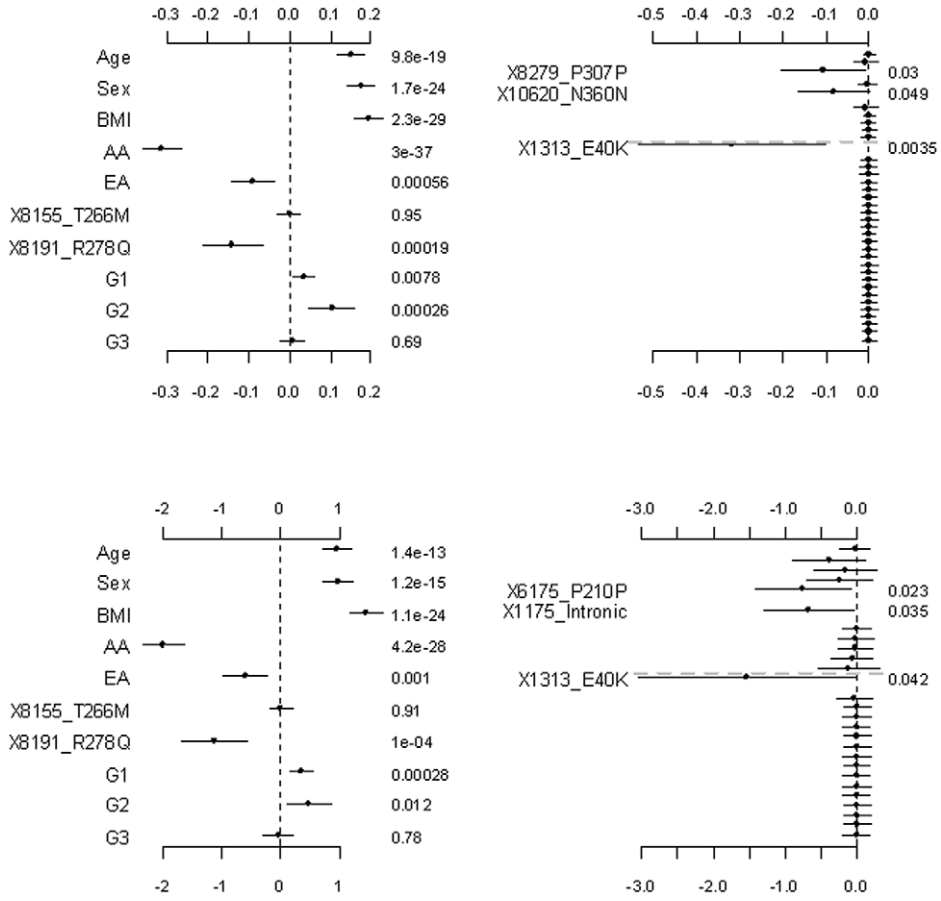

Figure 3. Analyses of the proposed hierarchical GLMs with prior means μ_j being the functional probabilities for the rare non-synonymous variants. The top and bottom panels are for the continuous and binary traits, respectively. The left panel is for the covariates, the two common non-synonymous variants and the three group effects (G1: common synonymous; G2: rare non-synonymous; G3: rare synonymous). The right panel is for the adjusted main effects (the gray dotted line shows the two groups G1 and G2). The points, short lines and numbers at the right side represent estimates of effects, ± 2 standard errors, and p-values, respectively. Only adjusted main effects with p-value <0.1 are labeled.
doi:10.1371/journal.pgen.1002382.g003

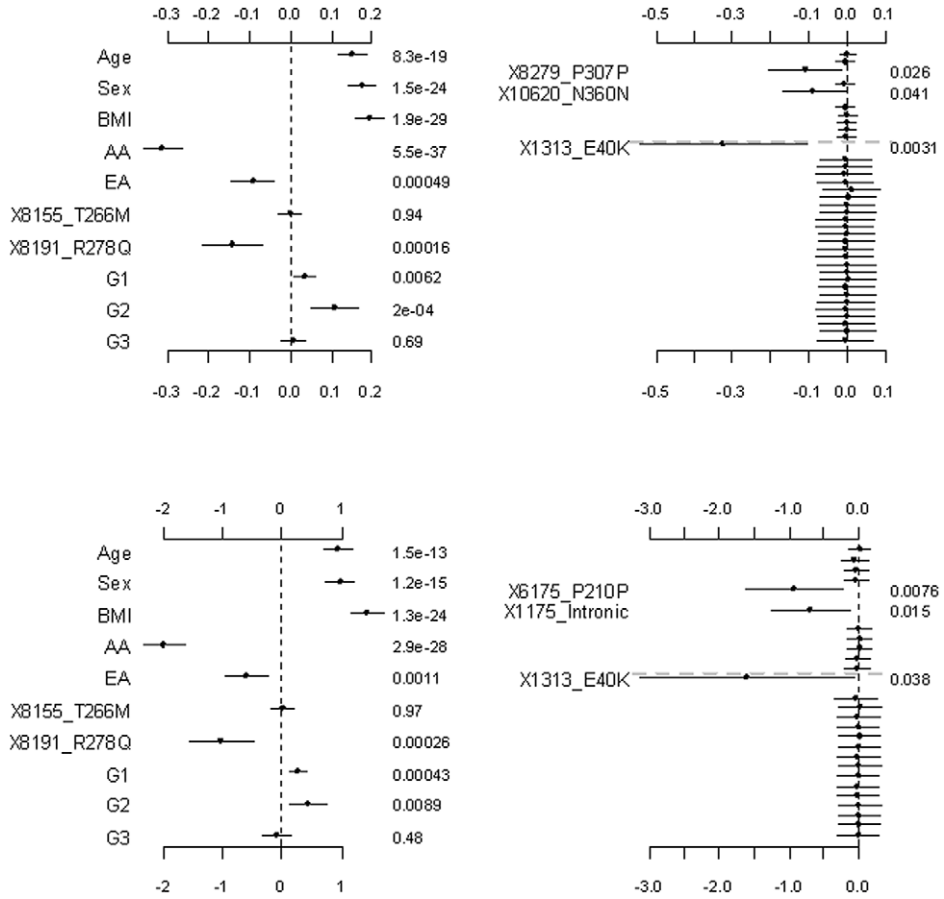

Figure 4. Analyses of the hierarchical GLMs with fixed scale $s_{\alpha_j} = 0.5$. The top and bottom panels are for the continuous and binary traits, respectively. The left panel is for the covariates, the two common non-synonymous variants and the three group effects (G1: common synonymous; G2: rare non-synonymous; G3: rare synonymous). The right panel is for the adjusted main effects (the gray dotted line shows the two groups G1 and G2). The points, short lines and numbers at the right side represent estimates of effects, ± 2 standard errors, and *p*-values, respectively. Only adjusted main effects with *p*-value <0.1 are labeled.
doi:10.1371/journal.pgen.1002382.g004

inflated the standard deviations for the estimates of the weights. The second alternative method preset the weights to 1 for all the variants or the estimated functional probabilities for the rare non-synonymous variants. But, these simpler models did not detect any significant group effects (Figure 5). Our third alternative approach simultaneously fitted the main effects of all the covariates and common and rare variants. As expected, this hierarchical model was able to detect only large effects. The variants 8191_R278Q and 1313_E40K were found to have strong effects in our previous analyses and thus were also detected in this alternative analysis.

Simulation Studies

Simulation design. We used simulations to validate the proposed models and algorithm and to study the properties of the method. Although most published simulation studies of rare variants generated genotypes assuming a population genetics model for the propagation of rare variants, the best way will be to take real sequence data obtained from many individuals and simulate phenotypes based on variants in those sequences, making assumptions only about genetic effects of variants [19]. Thus, we performed simulation studies by taking advantage of the real genotypes of common and rare variants and also the covariates in the above large real dataset.

We evaluated some factors that may affect the performance of the methods:

a) *Sample size:* We considered two sample sizes, including all observations in the real data (n = 3008) or individuals in the bottom and top quartiles of the real continuous phenotype (n = 1499), respectively.

b) *Number of groups and number of variants in each group:* We first considered the three groups of variants (i.e., common synonymous, rare non-synonymous, and rare synonymous) as in our real analyses. We then considered the second scenario with six groups by randomly partitioning each group into two with equal number of variants.

c) *Genetic effect sizes and directions of variants:* For each group of variants, we first assumed the total heritability (h) explained by the variants and the proportion of negative additive effects (*p.neg*) for the variants. We then randomly sampled an additive effect β_j for each variant from the region $[0, \beta_h]$ and changed the sign of β_j with the probability *p.neg*. The upper bound β_h was calculated using the method of Yi and Zhi [26], which controlled the total heritability of each group of variants approximately equal to the assumed value h. We considered several combinations of different h and *p.neg* (see Tables 1 and 2). The assumed total heritabilities were

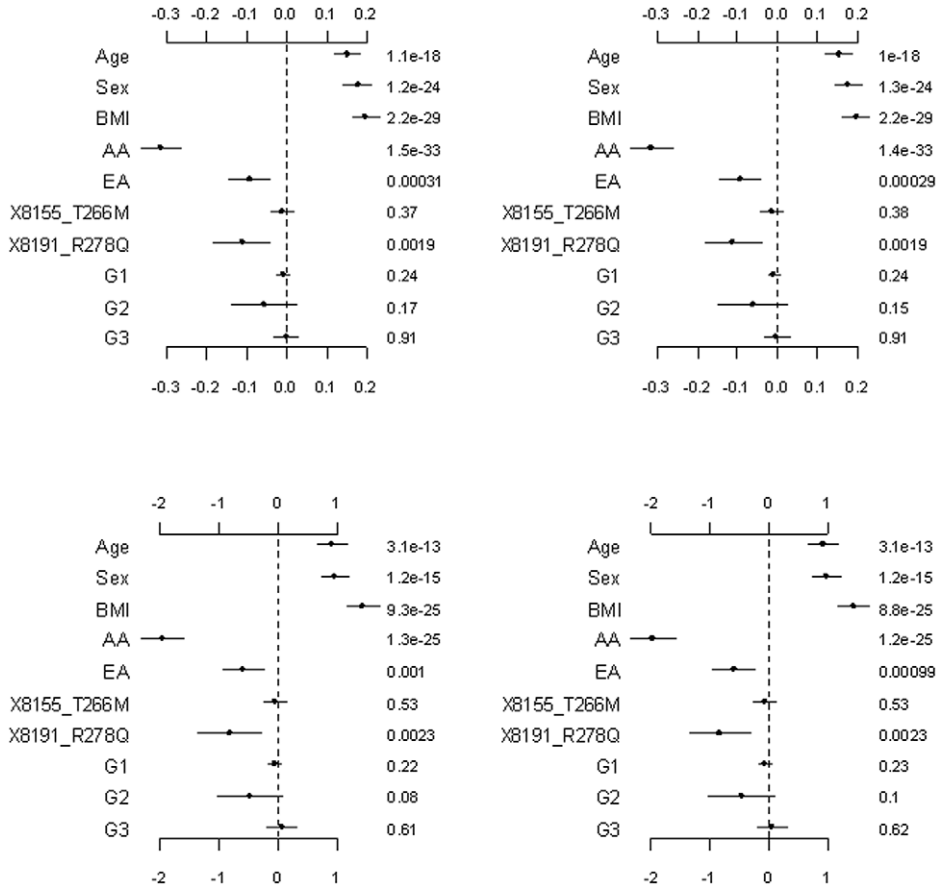

Figure 5. Setting the weights of individual variants to fixed values μ_j. The top and bottom panels are for the continuous and binary traits, and the left and right panels are for Simple-Sum and Weighted-Sum methods, respectively. The points, short lines and numbers at the right side represent estimates of effects, ± 2 standard errors, and *p*-values, respectively.
doi:10.1371/journal.pgen.1002382.g005

approximately equal to the estimated values of significant groups in our real data analyses.

We simulated a continuous and a binary phenotype. As in our real data analyses, we simultaneously fitted all the non-genetic variables (i.e., race, age, sex, and BMI), the two common non-synonymous variants (i.e., 8155_T266M and 8191_R278Q) and the grouped variants. We assumed the non-genetic coefficients and the additive effects of 8155_T266M and 8191_R278Q to be their estimated values in the continuous trait analysis (see the top panel of Figure 2). Given the assumed and the simulated coefficients, we first generated a normal continuous trait by setting the residual standard deviation to the estimated value in the continuous trait analysis ($\approx$0.2), and then set half of individuals with the 50% largest continuous phenotype as 'affected' ($y_i = 1$) and the other individuals as 'unaffected' ($y_i = 0$) to create a binary trait [26].

For each situation, 1000 replicated datasets were simulated. We calculated the frequencies of each effect estimated as significant at the threshold levels of $\alpha = 0.05$, 0.01, and 0.001 over 1000 replicates. These frequencies corresponded to the empirical power for the simulated non-zero effects and the type I error rate for other coefficients, respectively. We compared the proposed method with the three alternative approaches described in the above real data analyses. For the proposed method and the alternative methods with fixed scale parameters or fixed weights, we can calculate powers or type I error rates for all the covariates and the group effects. Since the third alternative approach cannot estimate the group effects, we simply used the minimal p-value to calculate powers or type I error rates for each group of variants. For each situation, the iterative EM-IWLS algorithm started from the plausible initial values described earlier and ran until convergence.

Results of simulations. Figure 7 and Figure 8 display the results at the threshold level of 0.01 from simulations with three groups and sample sizes of 3008 and 1449, respectively. Figure 9 shows the results from simulations with six groups and sample size of 3008. In all the simulations, the non-genetic covariates (race, age, sex, and BMI), which were highly significant in the real data analyses, were detected with high power by all the methods (not shown in the figures). All the methods also had high power to detect the significant common variant 8191_R278Q, and low type I error for the insignificant common variant 8155_T266M, showing that the genetic effects of common variants can be effectively estimated in large-scale studies.

In all the simulation scenarios, the proposed method and the extension of Yi and Zhi [26] were consistently more powerful to detect the simulated group(s) of variants than the other methods. As expected, the power drastically increases with larger sample size and for continuous phenotype. These relationships hold rather generally for the methods that we examined. In the simulations with three groups, the group of common variants was detected with slightly higher power than the group of rare variants (see

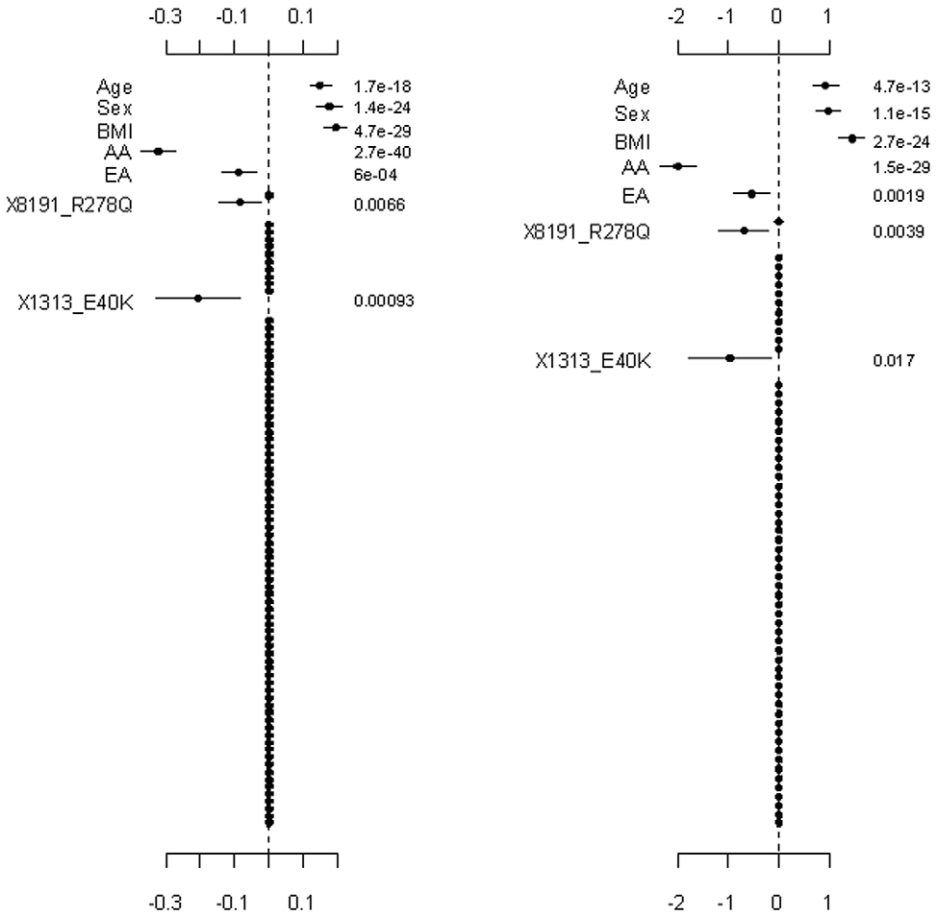

Figure 6. Ignoring the group effects and directly estimating the main effects of individual variants. The left and right panels are for the continuous and binary traits, respectively. The points, short lines and numbers at the right side represent estimates of effects, ± 2 standard errors, and *p*-values, respectively. Only main effects with *p*-value <0.1 are labeled.
doi:10.1371/journal.pgen.1002382.g006

Table 1. Simulations with three groups.

Group		1	2	3
Number of variants		10 (11)	26 (16)	44 (34)
Scenario	a	$h=0\%$	$h=0.5\%$ $p.neg=0$	$h=0\%$
	b	$h=0\%$	$h=0.7\%$ $p.neg=0.4$	$h=0\%$
	c	$h=0.5\%$ $p.neg=0$	$h=0.5\%$ $p.neg=0$	$h=0\%$
	d	$h=0.7\%$ $p.neg=0.4$	$h=0.7\%$ $p.neg=0.4$	$h=0\%$
	e	$h=0\%$	$h=0.7\%$ $p.neg=0$	$h=0.7\%$ $p.neg=0$
	f	$h=0\%$	$h=0.7\%$ $p.neg=0.4$	$h=0.7\%$ $p.neg=0.4$

The number of variants in each group for simulations with n = 3,008 (1,449), the assumed total heritability (*h*) explained by each group of variants, and the proportion of negative additive effects (*p.neg*) for groups with non-zero *h*.
doi:10.1371/journal.pgen.1002382.t001

Figure 7c and 7d, Figure 8c and 8d). However, when the number of common variants was too small, their group effect was detected with lower power than the group of rare variants (see Figure 9a). Our simulations showed that the proposed method and the extension of Yi and Zhi [26] also well controlled the type I error for groups with zero effects. This results from the fact that our

Table 2. Simulations with six groups.

Group		1	2	3	4	5	6
Number of variants		5 (5)	5 (6)	13 (8)	13 (8)	22 (17)	22 (17)
Scenario	a	$h=0\%$	$h=0.5\%$ $p.neg=0$	$h=0\%$	$h=0\%$	$h=0.5\%$ $p.neg=0$	$h=0\%$
	b	$h=0\%$	$h=0.7\%$ $p.neg=0.4$	$h=0\%$	$h=0\%$	$h=0.7\%$ $p.neg=0.4$	$h=0\%$
	c	$h=0\%$	$h=0\%$	$h=0.5\%$ $p.neg=0$	$h=0\%$	$h=0.5\%$ $p.neg=0$	$h=0\%$
	d	$h=0\%$	$h=0\%$	$h=0.7\%$ $p.neg=0.4$	$h=0\%$	$h=0.7\%$ $p.neg=0.4$	$h=0\%$

The number of variants in each group for simulations with n = 3,008 (1,449), the assumed total heritability (*h*) explained by each group of variants, and the proportion of negative additive effects (*p.neg*) for groups with non-zero *h*.
doi:10.1371/journal.pgen.1002382.t002

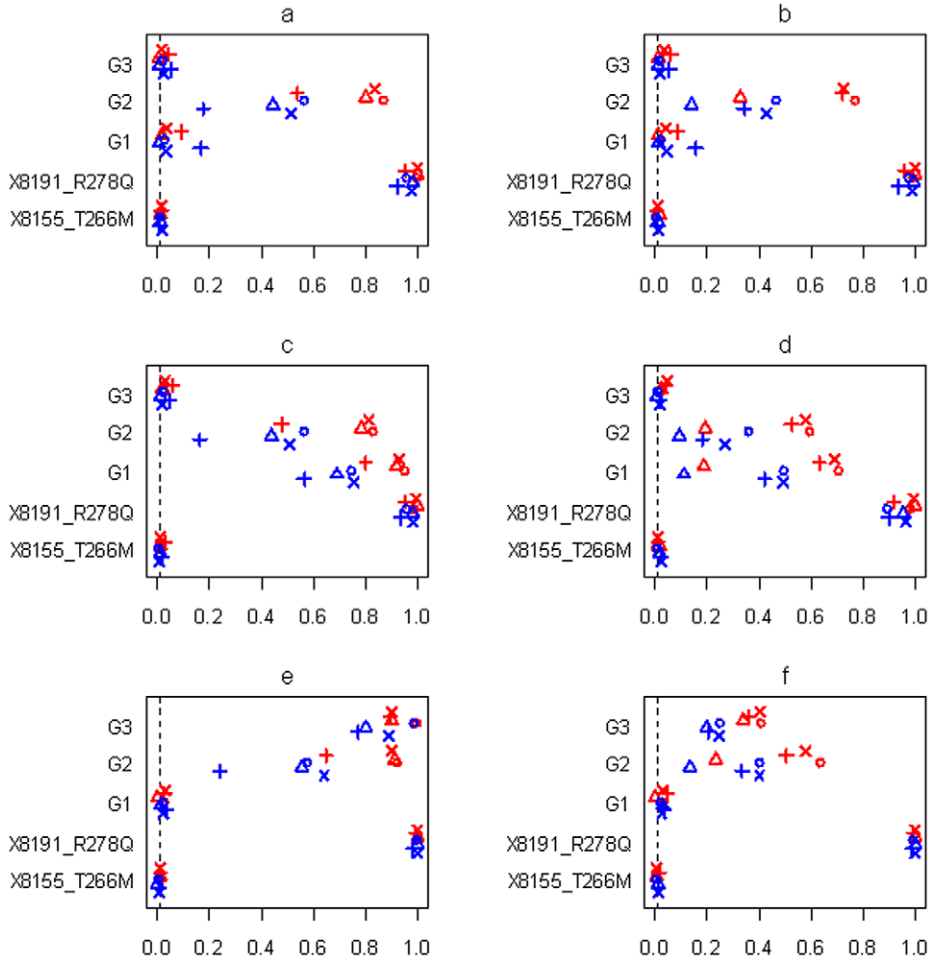

Figure 7. Simulations with sample size n = 3,008 and three groups for the six scenarios (see Table 1). Power or Type I error rate for the proposed method (○), Yi and Zhi (×), Simple-Sum (Δ) and All-Variants (+) under the threshold level of 0.01. X8155_T266M and X8191_R278Q are the two common non-synonymous variants, and G1, G2 and G3 are the three group effects (G1: common synonymous; G2: rare non-synonymous; G3: rare synonymous). Red and blue symbols represent results for continuous and binary traits, respectively. The dashed line is the nominal 0.01 level. doi:10.1371/journal.pgen.1002382.g007

hierarchical priors can shrink the weights of zero-effect variants to the prior means and thus yield the genetic score approximately equal to the simple sum or weighted sum. Our simulations also showed that the method ignoring group effects had the highest type I error rate.

For the groups in which all variants affected the traits in the same direction, the summation of the additive-effect predictors could provide a useful genetic score of these variants, and thus the simple-sum method had reasonable power to detect the group effect (see Figure 7a, 7c, 7e; Figure 8a, 8c, 8e; Figure 9a, 9c). Even in this situation, however, the proposed method and the extension of Yi and Zhi (2011) [26] were still more powerful than the simple-sum method. This may result from the fact that these hierarchical models estimate the weights from data and thus can produce different weights for different variants based on their contributions to the phenotype. The simulations further showed that the proposed method had slightly higher power than Yi and Zhi (2011) [26]. This is likely the results that the proposed method introduces variable-specific shrinkage parameters and thus could estimate the weights of variants more effectively. Finally, we found that the method ignoring group effects had the lowest power. This is expected because with the low total heritability the effects of single variants were very small and could not be detected powerfully.

For the groups in which 60% of variants increase disease risk and others are disease-protective, the summation of the additive-effect predictors provides an inefficient genetic score to summarize the information of the variants, and thus the simple-sum method had low power to detect the association (see Figure 7b, 7d, 7f; Figure 8b, 8d, 8f; and Figure 9b, 9d). These results are expected because using equal weights for disease-causing and disease-protective variants the information across variants can be cancelled out and the true association cannot be detected. However, the proposed method and the extension of Yi and Zhi (2011) [26] still had reasonable power to detect these multiple rare and common variants with opposite effects. These hierarchical models could yield opposite weights for disease-causing and protective variants, and thus avoid cancellation of individual-variant variation. Once again, we found that the proposed method had slightly higher power than Yi and Zhi (2011) [26]. Interestingly, our simulations show that the hierarchical models of all variants were not influenced by opposite effects, because they directly estimate the effects of individual variants; for a higher total heritability of multiple variants, some variants had larger effects and thus could be detectable individually.

We also evaluated power and type I error at several different levels (e.g., $\alpha = 0.05,\ 0.001$). The conclusions described above

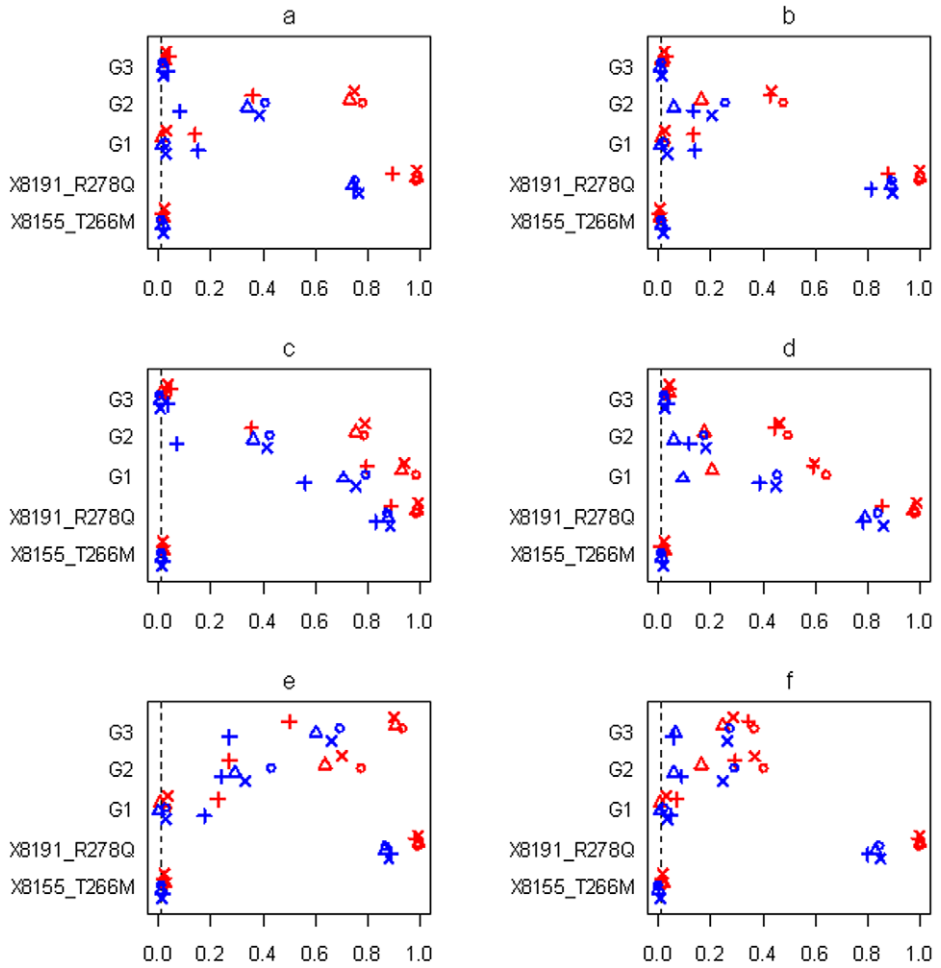

Figure 8. Simulations with sample size n = 1,499 and three groups for the six scenarios (see Table 1). Power or Type I error rate for the proposed method (○), Yi and Zhi (×), Simple-Sum (Δ) and All-Variants (+) under the threshold level of 0.01. X8155_T266M and X8191_R278Q are the two common non-synonymous variants, and G1, G2 and G3 are the three group effects (G1: common synonymous; G2: rare non-synonymous; G3: rare synonymous). Red and blue symbols represent results for continuous and binary traits, respectively. The dashed line is the nominal 0.01 level.
doi:10.1371/journal.pgen.1002382.g008

generally held (data not shown). Yi and Zhi [26] found that the hierarchical GLMs uniformly yielded much lower *p*-values for the simulated group effects than the previous methods. This was also true for the proposed method in our simulation studies. This indicates that our method usually provides stronger evidence of association if the variants really influence the disease.

Discussion

We have proposed here a Bayesian hierarchical generalized linear model framework for simultaneously analyzing multiple groups of rare and common variants and relevant covariates. Since complex diseases and traits are likely influenced by multiple genetic variants and environmental factors, the joint analyses of multiple groups of genetic variants can improve the power of detecting causal effects and lead to increased understanding about the genetic architecture of diseases. The proposed hierarchical generalized linear models introduce a group effect and a genetic score for each group of variants, and jointly estimate the group effects and the weights of the genetic scores. This can produce 'optimal' weights to different variants based on their contributions to the phenotype, yielding an effective summary of the information across variants. The simulation studies show that the proposed method can consistently provide reasonable power even in the presence of both risk and protective variants in a group, and has better power than existing approaches even when all variants act in the same direction. Application of the method to a large published dataset on resequencing of the gene *ANGPTL4* and triglycerides not only confirmed the original findings but also detected new associations.

In addition to the properties described above, our method has several remarkable features. First, the proposed method can simultaneously estimate the group effects of multiple groups of variants and the individual effects of the variants, allowing us to not only identify significant genes (or groups of variants) but also assess the relative importance of single variants. Second, our hierarchical model includes various existing methods for rare variants as special cases. This shows that the proposed method is theoretically more advantageous than the existing methods, and allows us to conveniently analyze data using different ways. Third, any external information about variants, for example, the functional prediction, can be easily incorporated into our hierarchical model by specifying the prior means of the weights for variants. By doing so, our approach has the additional advantage of accounting for uncertainties about the prior assumptions. Fourth, our approach is based on the generalized

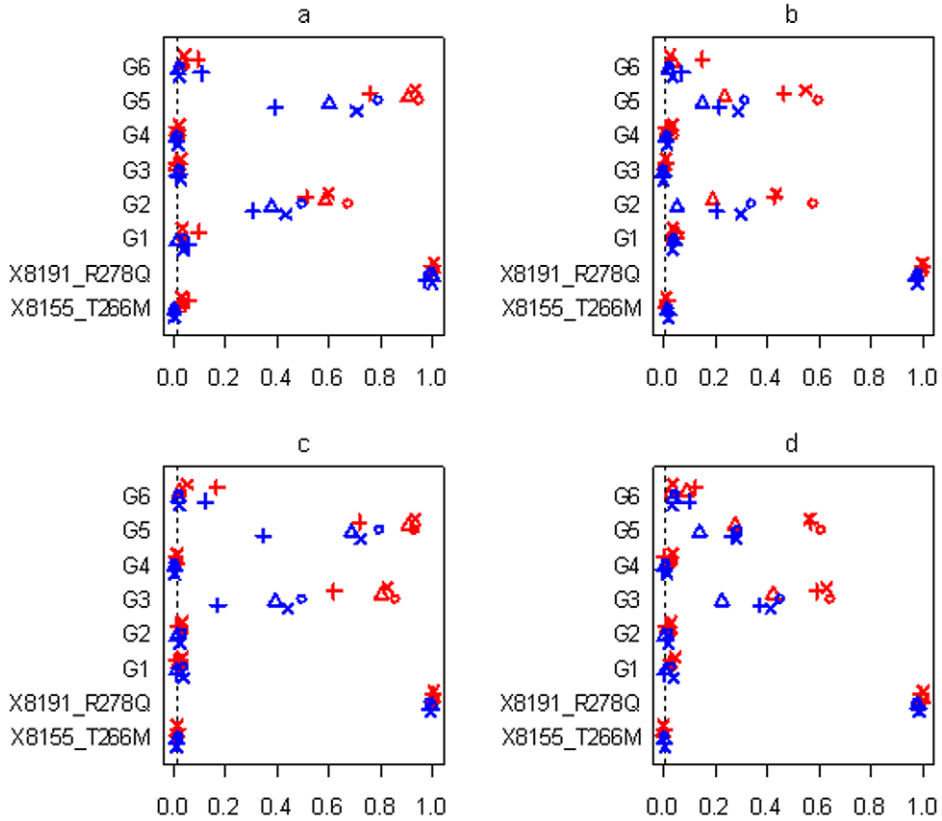

Figure 9. Simulations with sample size n = 3,008 and six groups for the four scenarios (see Table 2). Power or Type I error rate for the proposed method (○), Yi and Zhi (×), Simple-Sum (Δ) and All-Variants (+) under the threshold level of 0.01. X8155_T266M and X8191_R278Q are the two common non-synonymous variants, and G1–G6 are the six group effects. Red and blue symbols represent results for continuous and binary traits, respectively. The dashed line is the nominal 0.01 level.
doi:10.1371/journal.pgen.1002382.g009

linear model framework and thus can deal with various types of continuous and discrete phenotypes and covariates, and can fit any generalized linear models. Finally, the proposed algorithm extends the standard procedure for fitting classical generalized linear models in the general statistical package R to our Bayesian model, leading to the development of stable and flexible software.

Our approach is highly extensible; we have planned several extensions to the proposed method, some of which have been initially implemented in our software BhGLM. The key to our approach is the use of hierarchical prior distributions for the weights and the group effects, so that these multiplicative parameters are identifiable and can be simultaneously estimated from the data. We have proposed to use the hierarchical expression of the half-Cauchy distribution with the innovation of introducing both group- and variable-specific parameters. The half-Cauchy prior is an excellent default choice for many problems [39,40], and has been shown to perform well for our purposes. However, other hierarchical priors or penalized likelihood methods have been developed for high-dimensional data analysis, including lasso [48,49], adaptive Lasso [50], and the elastic net [51]. These methods can be expressed as hierarchical models by assigning certain priors on the variances and other hyperparameters [45,48,52], and can be incorporated into our framework.

Although demonstrated with only several groups of variants, our method can be adapted to deal with large-scale sequencing data involving thousands of exomes or candidate genes. For these high-dimensional settings, we need to modify the prior distributions of the group effects and the computational algorithm. We can place a shrinkage prior on the inverse scale in the gamma prior of $s^2_{g_k}$ and estimate the inverse scale from the data. We can further group the group effects based on pathways that candidate genes belong to, and specify the shrinkage priors by incorporating the second-level hierarchical structure, similar to the hierarchical priors of the weights. We describe our algorithm by simultaneously estimating all weights. This method can be very fast when the number of variables is not very large (say <2000) and has the advantage of accommodating the correlations among all the variables. However, it can be slow or even cannot be implemented when the number of variables is large due to memory storage and convergence problems. We can extend the algorithm to update coefficients group by group; at each of the iteration, the group-at-time algorithm proceeds by cycling through all the groups of parameters and treats the linear predictor of all other groups as an *offset* in the model. This method updates coefficients in a conditional manner, significantly reducing the number of parameters in each M-step of the EM-IWLS algorithm, and thus can deal with large number of variables.

Our third extension could incorporate external gene or pathway level information into the hierarchical model. Candidate genes or pathways studies usually consist of data at different levels, i.e., genetic variants within multiple candidate genes or pathways which may be functionally related [53]. Most of statistical methods for association studies consider only individual-level predictors (i.e., SNPs and covariates) and ignore the hierarchical structure of the data and gene or pathway-level information. Often, rich gene or pathway-level information is available [54], including genomic

annotation or pathway ontologies, functional assays, *in silico* predictions of function or evolutionary conservation [55]. Therefore, there is a growing need to develop sophisticated approaches that model the multilevel variation simultaneously and incorporate gene or pathway-level data into the model [56,57]. Our hierarchical models provide a natural and efficient way to incorporate the external information about candidate genes into the analysis. One way to include the gene-level information in the hierarchical models is to model the prior means of weights and group effects using gene or pathway level predictors [38]. This would allow us to pool the information in the same genes or pathways and thus would provide more effective inference about the genetic effects.

Our fourth extension could incorporate genetic interactions (gene-gene and gene-environment interactions) into the model. Just as interactions must be considered in standard GWA studies [57–59], they are also likely to be important in association studies involving rare variants [19]. In principle, we can extend the proposed model to include additional groups for interactions for each pair of groups of main effects and to define an overall effect and a genetic score for each interaction group. However, it would be necessary to investigate statistical power for detecting interactions for rare variants. Finally, we have planned to extend our method to family-based matched case-control association studies. So far the existing methods for rare variants have focused on population-based studies. However, for rare variants, family-based designs may prove very useful [60]. Not only are they robust against population stratification, but they may also offer increased power due to the increased likelihood of affected relatives to share the same rare disease variants. As the conditional logistic regression commonly used for matched case-control studies can be formulated as a Poisson regression [36], our hierarchical generalized linear models can be applied.

The proposed hierarchical generalized linear models may provide efficient tools for disease risk prediction and personalized medicine. GWA studies have raised expectations for predicting individual susceptibility to common diseases using genetic variants [61,62]. Previous methods using only a limited number of significant variants have typically failed to achieve satisfactory prediction performance [63,64]. Recent studies show that joint analysis of a large number of genetic variants can improve the prediction of complex traits [65–67]. It is understood that a model including as many predictors as possible and fitted appropriately could provide better prediction. Although the previous studies have included many genetic variants in a predictive model, they treat these variants individually and hence could be suboptimal to efficiently use information of genetic variants with small effects and low frequencies. The proposed hierarchical models can better deal with such variants and can integrate external biological knowledge, and therefore may be able to improve the accuracy of prediction.

Supporting Information

Text S1 The EM-IWLS Algorithm for Fitting Hierarchical Generalized Linear Models.
(DOCX)

Acknowledgments

We would like to thank Drs. Jonathan Cohen and Helen Hobbs for access to the Dallas Heart Study dataset. We are grateful for three reviewers for their constructive comments that improve the previous version of the manuscript.

Author Contributions

Conceived and designed the experiments: NY. Performed the experiments: NY. Analyzed the data: NY. Contributed reagents/materials/analysis tools: NY. Wrote the paper: NY NL DZ JL.

References

1. Hardy J, Singleton A (2009) Genomewide association studies and human disease. N Engl J Med 360: 1759–1768.
2. Hindorff L, Sethupathy P, Junkins H, Ramos E, Mehta J, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362–9367.
3. Flint J, Mackay T (2009) Genetic architecture of quantitative traits in mice, flies, and humans. Genome Res 19: 723–733.
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
5. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446–450.
6. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant…or not? Hum Mol Genet 11: 2417–2423.
7. Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40: 695–701.
8. Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19: 212–219.
9. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69: 124–137.
10. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82: 100–112.
11. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, et al. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305: 869–872.
12. Cohen JC, Boerwinkle E, Mosley TH, Jr., Hobbs HH (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med 354: 1264–1272.
13. Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, et al. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 39: 513–516.
14. Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, et al. (2007) Medical sequencing at the extremes of human body mass. Am J Hum Genet 80: 779–791.
15. Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, et al. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40: 592–599.
16. Azzopardi D, Dallosso AR, Eliason K, Hendrickson BC, Jones N, et al. (2008) Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. Cancer Res 68: 358–363.
17. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324: 387–389.
18. Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, et al. (2009) Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J Clin Invest 119: 70–79.
19. Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11: 773–785.
20. Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Genet 44: 293–308.
21. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311–321.
22. Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34: 188–193.
23. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384. doi:10.1371/journal.pgen.1000384.
24. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86: 832–838.
25. King CR, Rathouz PJ, Nicolae DL (2010) An evolutionary framework for association testing in resequencing studies. PLoS Genet 6: e1001202. doi:10.1371/journal.pgen.1001202.
26. Yi N, Zhi D (2011) Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol 35: 57–69.
27. Han F, Pan W (2010) A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. Hum Hered 70: 42–54.

28. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7: e1001322. doi:10.1371/journal.pgen.1001322.
29. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet 7: e1001289. doi:10.1371/journal.pgen.1001289.
30. Pan W, Shen X (2011) Adaptive tests for association analysis of rare variants. Genet Epidemiol.
31. Hoffmann TJ, Marini NJ, Witte JS (2010) Comprehensive approach to analyzing rare genetic variants. PLoS ONE 5: e13584. doi:10.1371/journal.pone.0013584.
32. Li Y, Byrnes AE, Li M (2010) To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. Am J Hum Genet 87: 728–735.
33. Liu DJ, Leal SM (2010) 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 6: e1001156. doi:10.1371/journal.pgen.1001156.
34. Luo L, Boerwinkle E, Xiong M (2011) Association studies for next-generation sequencing. Genome Res.
35. McCullagh P, Nelder JA (1989) Generalized linear models. London: Chapman and Hall.
36. Gelman A, Carlin J, Stern H, Rubin D (2003) Bayesian data analysis. London: Chapman and Hall.
37. Gelman A (2004) Parameterization and Bayesian modeling. Journal of the American Statistical Association 99: 537–545.
38. Gelman A, Hill J (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models. New York: Cambridge University Press.
39. Gelman A (2006) Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1: 515–533.
40. Carvalho C, Polson N, Scott J (2010) The horseshoe estimator for sparse signals. Biometrika 97: 465–480.
41. Gelman A, Jakulin A, Pittau MG, Su YS (2008) A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics 2: 1360–1383.
42. Yi N, Banerjee S (2009) Hierarchical generalized linear models for multiple quantitative trait locus mapping. Genetics 181: 1101–1113.
43. Yi N, Kaklamani VG, Pasche B (2011) Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. Ann Hum Genet 75: 90–104.
44. Armagan A, Dunson D, Lee J (2010) Bayesian generalized double Pareto shrinkage. Biometrika.
45. Kyung M, Gill J, Ghosh M, Casella G (2010) Penalized Regression, Standard Errors, and Bayesian Lassos. Bayesian Analysis 5: 369–412.
46. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249.
47. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57: 289–300.
48. Park T, Casella G (2008) The Bayesian Lasso. Journal of the American Statistical Association 103: 681–686.
49. Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B* 58: 267–288.
50. Zou H (2006) The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association 101: 1418–1429.
51. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67: 301–320.
52. Yi N, Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. Genetics 179: 1045–1055.
53. Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11: 843–854.
54. Rebbeck T, Spitz M, Wu X (2004) Assessing the function of genetic variants in candidate gene association studies. Nat Rev Genet 5: 589–597.
55. Thomas DC, Conti DV, Baurley J, Nijhout F, Reed M, et al. (2009) Use of pathway information in molecular epidemiology. Hum Genomics 4: 21–42.
56. Dunson DB, Herring AH, Engle SM (2008) Bayesian selection and clustering of polymorphisms in functionally related genes. Journal of The American Statistics Association 103: 534–546.
57. Thomas D (2010) Gene-environment-wide association studies: emerging approaches. Nat Rev Genet 11: 259–272.
58. Cordell H (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392–404.
59. Yi N (2010) Statistical analysis of genetic interactions. Genet Res (Camb) 92: 443–459.
60. Zhu X, Feng T, Li Y, Lu Q, Elston RC (2010) Detecting rare variants for complex traits using family and unrelated data. Genet Epidemiol 34: 171–187.
61. Wray N, Goddard M, Visscher P (2008) Prediction of individual genetic risk of complex disease. Curr Opin Genet Dev 18: 257–263.
62. Kraft P, Wacholder S, Cornelis M, Hu F, Hayes R, et al. (2009) Beyond odds ratios–communicating disease risk based on genetic profiles. Nat Rev Genet 10: 264–269.
63. Kraft P, Hunter D (2009) Genetic risk prediction–are we there yet? N Engl J Med 360: 1701–1703.
64. Jakobsdottir J, Gorin M, Conley Y, Ferrell R, Weeks D (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet 5: e1000337. doi:10.1371/journal.pgen.1000337.
65. de los Campos G, Gianola D, Allison DB (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet 11: 880–886.
66. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565–569.
67. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, et al. (2011) Beyond missing heritability: prediction of complex traits. PLoS Genet 7: e1002051. doi:10.1371/journal.pgen.1002051.