Scalable mixed model methods for set-based association studies on large-scale categorical data analysis and its application to exome-sequencing data in UK Biobank

Authors

Wenjian Bi, Wei Zhou, Peipei Zhang, Yaoyao Sun, Weihua Yue, Seunggeun Lee

Correspondence

wenjianb@pku.edu.cn (W.B.), lee7801@snu.ac.kr (S.L.)

To analyze rare variants, Bi et al. proposed POLMM-GENE, an approach that is scalable for large-scale sequencing datasets. POLMM-GENE fully utilizes the categorical nature of phenotypes, which avoids inflated type I error rates or power loss. It can identify gene-phenotype associations, providing valuable insights into missing trait heritability.

Bi et al., 2023, The American Journal of Human Genetics *110*, 762–773
May 4, 2023

https://doi.org/10.1016/j.ajhg.2023.03.010

CellPress

ARTICLE

Scalable mixed model methods for set-based association studies on large-scale categorical data analysis and its application to exome-sequencing data in UK Biobank

Wenjian Bi,[1,2,3,14,*] Wei Zhou,[4,5,6,14] Peipei Zhang,[7,8] Yaoyao Sun,[9,10] Weihua Yue,[9,10,11,12] and Seunggeun Lee[13,*]

Summary

The ongoing release of large-scale sequencing data in the UK Biobank allows for the identification of associations between rare variants and complex traits. SAIGE-GENE+ is a valid approach to conducting set-based association tests for quantitative and binary traits. However, for ordinal categorical phenotypes, applying SAIGE-GENE+ with treating the trait as quantitative or binarizing the trait can cause inflated type I error rates or power loss. In this study, we propose a scalable and accurate method for rare-variant association tests, POLMM-GENE, in which we used a proportional odds logistic mixed model to characterize ordinal categorical phenotypes while adjusting for sample relatedness. POLMM-GENE fully utilizes the categorical nature of phenotypes and thus can well control type I error rates while remaining powerful. In the analyses of UK Biobank 450k whole-exome-sequencing data for five ordinal categorical traits, POLMM-GENE identified 54 gene-phenotype associations.

Introduction

Whole-exome- and whole-genome-sequencing data in large cohorts and biobanks enable detection of rare and ultra-rare variants, which can help explain missing trait heritability that common variant-based genome-wide association studies (GWASs) cannot account for.[1] For rare variants (minor allele frequency [MAF] $< 1\%$), set-based tests such as the burden test, sequencing kernel association test (SKAT), and SKAT-O are more powerful than single-variant tests. Recently, big biobanks have started to generate large-scale sequencing data. For example, UK Biobank (UKBB) has released whole-exome sequencing data for most of the participants in public. Linked to extensively collected electronic health records and self-reported questionnaires, the resource enables genetic association studies of common and rare variants for thousands of human diseases and traits.[2]

Ordinal categorical data are widely collected in surveys and questionnaires to characterize human behavior, satisfaction, and psychiatric status. One strategy for analyzing an ordinal categorical phenotype is transforming it into a quantitative or binary trait and then using well-developed association test methods, such as SAIGE and SAIGE-GENE+.[2] However, since these strategies either ignore the categorical nature of the phenotype or lose parts of the phenotypic information, they can have either inflated type I error rates or power loss.[3] To address this issue, Bi et al. developed (proportional odds logistic mixed model) POLMM, a single-variant test method based on proportional odds logistic mixed model.[3] Using efficient mixed model approach and saddlepoint approximation (SPA), POLMM can analyze large-scale biobank data while controlling for sample relatedness and unbalanced phenotypic distribution.[3]

As the ongoing release of large-scale whole-exome- and genome-sequencing data continues, scalable set-based approaches are needed for ordinal categorical data analysis to study rare variant associations. In this paper, we extend POLMM to POLMM-GENE for set-based tests. POLMM-GENE consists of the following important features: (1) it utilizes a proportional odds logistic mixed model to accurately model ordinal categorical phenotypes while controlling for sample relatedness; (2) it supports the burden test, SKAT, and SKAT-O and thus is powerful in a wide range of scenarios of different proportions of causal variants and effect directions; (3) it can incorporate multiple maximal MAF cutoffs and multiple annotations to

[1]Department of Medical Genetics, School of Basic Medical Sciences, Peking University, Beijing, China; [2]Center for Medical Genetics, School of Basic Medical Sciences, Peking University, Beijing, China; [3]Department of Biomedical Informatics, School of Basic Medical Sciences, Peking University, Beijing, China; [4]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; [5]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA; [6]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA; [7]Department of Biochemistry and Biophysics, School of Basic Medical Sciences, Peking University Health Science Center, Beijing, China; [8]Key Laboratory for Neuroscience, Ministry of Education/National Health and Family Planning Commission, Peking University, Beijing, China; [9]Peking University Sixth Hospital, Peking University Institute of Mental Health, Beijing, China; [10]NHC Key Laboratory of Mental Health (Peking University), National Clinical Research Center for Mental Disorders (Peking University Sixth Hospital), Beijing, China; [11]Henan Key Lab of Biological Psychiatry, the Second Affiliated Hospital of Xinxiang Medical University, Xinxiang, Henan, China; [12]Chinese Institute for Brain Research, Beijing, China; [13]Graduate School of Data Science, Seoul National University, Seoul, Korea
[14]These authors contributed equally
*Correspondence: wenjianb@pku.edu.cn (W.B.), lee7801@snu.ac.kr (S.L.)
https://doi.org/10.1016/j.ajhg.2023.03.010.

improve power; (4) it uses saddlepoint approximation and ultra-rare variants collapsing to control type I error rates, which is essential, especially if the phenotypic distribution is unbalanced; (5) the core algorithm is written in C++ to increase computational efficiency while strictly controlling for memory usage regardless of the set size, both of which are important for cloud computation, e.g., in the UK Biobank Research Analysis Platform (RAP). With all of these features, POLMM-GENE is a valid approach for testing associations between rare variants and ordinal categorical data in large-scale cohorts and biobanks, such as in UK Biobank.

Material and methods

Proportional odds logistic mixed model and test statistics

We consider a proportional odds logistic mixed model (POLMM) to relate ordinal categorical phenotypes to covariates and genotypes while adjusting for sample relatedness.[3] We let n denote the sample size and J the number of category levels of the phenotype. For the i-th subject, where $i \le n$, we let $y_i = 1, 2, \ldots, J$ denote the ordinal categorical phenotype and let $G_i = (g_{i1}, g_{i2}, \ldots, g_{iM})$ denote the genotype or dosage values of the M genetic variants in a target set (e.g., a gene).

$$\text{Logit}(\nu_{ij}) = \epsilon_j - \eta_i = \epsilon_j - X_i^T\beta - G_i^T\gamma - b_i, 1 \le i \le n, 1 \le j \le J \quad \text{(Equation 1)}$$

where $\nu_{ij} = \Pr(y_i \le j | X_i, G_i, b_i)$ is the cumulative probability of the ordinal phenotype $y_i \le j$ conditional on a p-dimensional vector of covariates X_i and an M-dimensional vector of genotype G_i. We used the cutpoints ϵ: $\epsilon_1 < \ldots < \epsilon_J = \infty$ to categorize the data, and coefficients β and γ are fixed effect sizes of the covariates and genotype, respectively. To adjust for sample relatedness, we incorporate an n-dimensional random effect vector $b = (b_1, \cdots, b_n)^T$ following a multivariate normal distribution $N(0, \tau V)$, where τ is a variance component parameter and V is an $n \times n$ dimensional genetic relationship matrix (GRM). If $J = 2$, the phenotype is binary and the Equation 1 is a logistic mixed model as in SAIGE and GMMAT.[4–6] Although POLMM is based on the proportional odds assumption, previous studies indicate that it is still valid with respect to single-variant tests when the assumption is violated.[3,7] Assessing the model residuals can be useful for evaluating the null model fitting.

POLMM-GENE contains two steps. In step 1, we used the average information restricted maximum likelihood (AI-REML) algorithm[3] to fit the null model with $\gamma = 0$. Both dense GRM and sparse GRM are supported. As demonstrated in previous studies, we recommend using a sparse GRM because of its high computational efficiency and no loss of power in most scenarios.[3] In step 2, we calculate test statistics of the burden test, SKAT, and SKAT-O on the basis of which p values are estimated for each set.

For subject i, we define a $J \times 1$ vector $\tilde{y}_i = (y_{i1}, \cdots, y_{ij})^T$ as an equivalent representation of the ordinal categorical phenotype $y_i = 1, 2, \ldots, J$: if $y_i = j$, then $y_{ij} = 1$ and the other elements in $\tilde{y}_i$ are 0. We let μ_{ij} denote the mean of y_{ij} and $\hat{\mu}_{ij}$ the fitted value μ_{ij} under the null hypothesis. Suppose $\tilde{Z} = (e_1, \cdots, e_1, e_2, \cdots, e_2 \cdots, e_n, \cdots, e_n)^T$, where e_i is an $n \times 1$ vector with unity in the i-th coordinate and 0's elsewhere. We let an $n(J-1) \times n(J-1)$ block diagonal matrix Ψ be the covariance matrix of $\tilde{y}$ as follows.

$$\Psi = \begin{bmatrix} \Psi_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Psi_n \end{bmatrix}, \Psi_i = \begin{bmatrix} \mu_{i1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mu_{i(J-1)} \end{bmatrix} - \mu_i\mu_i^T, \mu_i = \left(\mu_{i1}, \cdots, \mu_{i(J-1)}\right)^T.$$

We define an $n(J-1) \times n(J-1)$ diagonal matrix

$$R = diag\left(R_{11}, R_{12}, \cdots, R_{1(J-1)}, R_{21}, R_{22}, \cdots, R_{2(J-1)}, \cdots, R_{n1}, R_{n2}, \cdots, R_{n(J-1)}\right),$$

where $R_{ij} = 1/\mu_{ij} \cdot \partial\mu_{ij}/\partial\eta_i - 1/\mu_{ij} \cdot \partial\mu_{ij}/\partial\eta_i$. The score statistics for the variant m is

$$S_m = \sum_{i=1}^{n}\sum_{j=1}^{J-1}\left[g_{im}R_{ij} \cdot \left(y_{ij} - \hat{\mu}_{ij}\right)\right]$$

The variance-covariance matrix of $\boldsymbol{S}$ can be obtained by extending a single variance estimation procedure in POLMM.[3] We define $\tilde{V} = \tilde{Z}V\tilde{Z}^T$, $\Sigma = R^{-1}\Psi^{-1}R^{-1} + \tau\tilde{V}$, and $P = \Sigma^{-1} - \Sigma^{-1}\tilde{Z}X(X^T\tilde{Z}^T\Sigma^{-1}\tilde{Z}X)^{-1}X^T\tilde{Z}^T\Sigma^{-1}$. Suppose that the genotype matrix of the M variants in the set is $\boldsymbol{G} = (G_1, G_2, \cdots, G_M)$, then, following the similar derivation as in POLMM, the variance-covariance matrix of the score vector $S = (S, \ldots, S_M)^T$ is $\overline{\boldsymbol{G}}^{\boldsymbol{T}}\tilde{Z}^TP\tilde{Z}\overline{\boldsymbol{G}}$, where

$$\overline{\boldsymbol{G}} = \boldsymbol{G} - X\left(X^T\tilde{Z}^TR\Psi R\tilde{Z}X\right)^{-1}X^T\tilde{Z}^TR\Psi R\tilde{Z}\boldsymbol{G}.$$

The Burden test and SKAT statistics can be written as

$$Q_B = \left(\sum_{j=1}^{m}\omega_jS_j\right)^2, Q_S = \sum_{j=1}^{m}\omega_j^2S_j^2,$$

where ω_j is the weight for each variant. In simulation studies and real data analysis, we used beta(MAF, 1, 25) distribution to upweight rarer variants.[5] The SKAT-O method combined the burden test and SKAT using the following framework:

$$Q_\rho = (1-\rho)Q_B + \rho Q_S,$$

where ρ is a tuning parameter with a range of [0, 1]. Since the optimal ρ is unknown, SKAT-O applies the minimal p values over a grid of ρ as a test statistic.[8]

Robust burden test, robust SKAT, and robust SKAT-O

We propose a robust set-based testing approach, POLMM-GENE, which includes the burden test, SKAT, and SKAT-O. Under the null hypothesis, the vector $S = (S_1, \ldots, S_M)^T$ asymptotically follows a multivariate normal distribution with a variance-covariance matrix $\Psi = \overline{G}^T\tilde{Z}^TP\tilde{Z}\overline{G}$. Given the variance-covariance matrix, the R package SKAT can be used to calculate set-based p values of the burden test, SKAT, and SKAT-O.

If the phenotypic distribution is unbalanced, the score statistics distribution for rare variants is highly skewed, which could cause inflated type I error rates. Zhao et al.[8] and Zhou et al.[5] have demonstrated that using SPA to adjust the variance-covariance matrix can address this issue for binary traits. POLMM-GENE extends this framework to account for phenotype imbalance of ordinal categorical traits.

For each variant m, if the score statistics S_m lies within two standard deviations of the mean, the normal approximation with a variance of Ψ_{mm} generally performs well and no adjustment is needed.[3] If the score statistics S_m is beyond two standard deviations of the mean, we use SPA to calculate single-variant p value $p_{spa,m}$, which we will use to calibrate the variance of S_m. The

detailed derivation of SPA for single-variant analysis can be seen in previous work.[3] Suppose that score statistics S_m follows a normal distribution and the estimated variance of S_m is V_m, then S_m^2/V_m follows the chi-square distribution with one degree of freedom. We adjust the variance so that the p value is the same as $p_{spa,m}$, in which the adjusted variance is

$$V_{spa,m} = \frac{S_m^2}{\chi^2_{quantile}\left(1 - p_{spa,m}\right)},$$

where $\chi^2_{quantile}$ is the quantile function of the chi-square distribution with one degree of freedom. Then, we update the variance-covariance matrix $\Psi_{spa} = D \cdot \Psi \cdot D$, where D is a diagonal matrix whose m-th diagonal element is $\sqrt{V_m/\Psi_{mm}} = \sqrt{\max(V_{spa,m}, \Psi_{mm})/\Psi_{mm}}$. We calculate region-based p value on the basis of the assumption that $S \sim MVN(0, \Psi_{spa})$. On the other side, if we standardize the score statistics as $\tilde{S}_m = S_m/\sqrt{V_m}$, then $S \sim MVN(0, \Psi_{spa})$ is equivalent to the assumption that

$$\tilde{S} = (\tilde{S}_1, \ldots, \tilde{S}_M)^T \sim MVN(0, \tilde{\Psi}_{spa}),$$

where $\tilde{\Psi}_{spa} = \tilde{D} \cdot \Psi \cdot \tilde{D}$ and matrix $\tilde{D}$ is diagonal in which the m-th diagonal element is $1/\sqrt{\Psi_{mm}}$.

As the sample size in sequencing data increases, the number of rare and ultra-rare variants increases substantially. SAIGE-GENE+ collapses ultra-rare variants whose minor allele counts (MACs) are less than a pre-defined cutoff (e.g., $\leq$10) to a single marker.[9] The collapsing process reduces the number of variants in a set to test and thus increases the computational efficiency. Meanwhile, the collapsing approach helps to better control the type I error rate while remaining powerful. POLMM-GENE also adopts this strategy to collapse ultra-rare variants.[9]

A common practice for set-based associations is to test all rare (MAF $\leq$ 1%) protein-altering variants. However, this approach can lose power if association signals are enriched in ultra-rare variants or certain functional annotation classes. Previous studies have demonstrated that incorporating multiple MAF cutoffs and functional annotations in the exome-wide set-based association tests can increase power and thus help identify novel gene-phenotype associations.[9,10] POLMM-GENE supports incorporating multiple MAF cutoffs and functional annotations to calculate p values with a computationally efficient approach. Then, we can use the Cauchy combination or minimum p value approaches to combine these p values.[11,12]

Numeric simulations

We conducted extensive simulation studies to evaluate the performance of POLMM-GENE in terms of type I error rates and powers for ordinal categorical trait analysis. We simulated 100,000 samples with 50,000 unrelated samples and 12,500 families. Each family has two parents and two full siblings.[3] To mimic the allele frequency distribution and linkage disequilibrium (LD) structure in real data, we simulated sequencing data by using the WES data with 100,000 unrelated White British samples in UK Biobank (Field ID: 23155). We used ANNOVAR for gene annotation and defined loss-of-function (LoF) variants as those annotated as frameshift deletion, frameshift insertion, non-frameshift deletion, non-frameshift insertion, splicing, stop gain, and stop loss. We randomly selected 1,000 genes, and the distribution of MAC of all variants in the selected 1,000 genes is shown in Figure S1. For each gene, we used three maximal MAF cutoffs of 1%, 0.1%, and 0.05% and two annotation groups of LoF and LoF+missense to define six sets of rare variants. We used Cauchy combination to calculate a combined p value.[9,11] In simulation studies, ultra-rare variants with MAC $\leq$ 10 were collapsed to a single marker.

Ordinal categorical phenotypes were simulated following Equation 1. We simulated two covariates, one of which follows a standard normal distribution and the other one follows a Bernoulli distribution with a probability of 0.5. The effect sizes β of the two covariates are 0.5 and the variance component parameter $\tau = 1$. To evaluate type I error rates, we set $\gamma = 0$, that is, the effect sizes of all genetic variants are 0. We selected the cutpoint vector ϵ to simulate ordinal phenotypes with three category levels. We considered balanced phenotypic distribution of 1:1:1, unbalanced phenotypic distribution of 10:1:1, and extremely unbalanced phenotypic distribution of 30:1:1. For each sample size distribution, we simulated 4,000 datasets of phenotypes and conducted a total of 4×10^6 gene-level tests (i.e., 4,000 phenotypes × 1,000 genes). To evaluate power, we set effect sizes of the causal variants as $-\log10(\text{MAF}) \times 0.5$ and simulated γ by using nine scenarios including three different settings of causal variants proportions and three different settings of effect size directions. For each scenario, we simulated ten datasets of phenotypes and conducted a total of 10,000 tests (i.e., 10 phenotypes × 1,000 genes). In addition to the proposed POLMM-GENE, we also evaluated SAIGE-GENE+ (RAW), SAIGE-GENE+ (INT), and SAIGE-GENE+ (BINA). SAIGE-GENE+ (RAW) used the raw categorical trait as a quantitative phenotype and SAIGE-GENE+ (INT) additionally performed an inverse normalization prior to analysis. SAIGE-GENE+ (BINA) transformed the categorical trait into a binary phenotype (see legend of Figure 1).

Application to UK Biobank data

We used POLMM-GENE to conduct exome-wide analyses for the following five ordinal categorical phenotypes:

- alcohol intake frequency (Field ID: 1558; 380,046 White British subjects),
- comparable height size at age 10 (Field ID: 1697; 374,448 White British subjects),
- comparable body size at age 10 (Field ID: 1687; 374,133 White British subjects),
- morning/evening person chronotype (Field ID: 1180; 339,879 White British subjects),
- cognitive symptoms severity (Field ID: 120042; 130,449 White British subjects).

The sample size distribution in each category level is shown in Figure S2. We made a sparse GRM for White British subjects by using hard-called genotype data (Field ID: 22418). In step 1, we incorporated gender, birth year, top ten PCs, and batch data (Field ID: 23160) as covariates and then fitted a null mixed model. In step 2, we used population-level exome OQFE variants, 450k release (Field ID: 23149), to conduct gene-level analyses on the UK Biobank RAP.

We collapsed ultra-rare variants whose MAC $\leq$ 10 to a single marker in the analysis of alcohol intake frequency, comparable height at age of 10, and morning/evening person chronotype. For cognitive symptoms severity, because of its smaller sample size, we set the MAC cutoff at 5. We used three maximal MAF cutoffs of 1%, 0.1%, and 0.01% and two annotation groups of LoF and LoF+Missense to group variants in each gene. Gene annotations were conducted with ANNOVAR.

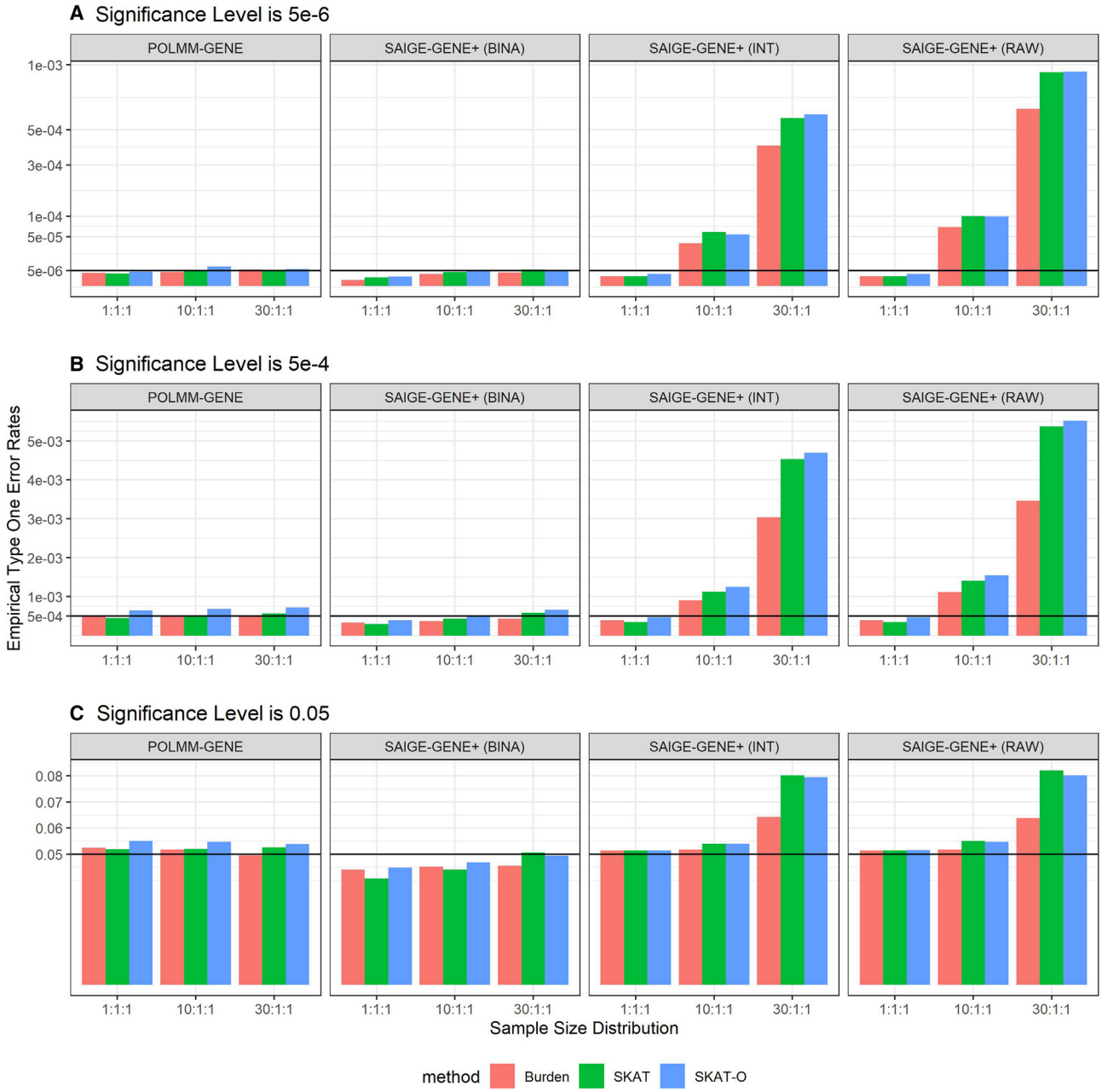

Figure 1. Empirical type I error rates of POLMM-GENE, SAIGE-GENE+ (BINA), SAIGE-GENE+ (RAW), and SAIGE-GENE+ (INT) under sample size distributions of 1:1:1, 10:1:1, and 30:1:1

(A–C) Empirical type I error rates at (A) significance level of 5e−6, (B) significance levels of 5e−4, and (C) significance levels of 0.05. SAIGE-GENE+ (RAW) considered the raw categorical trait as a quantitative phenotype; SAIGE-GENE+ (INT) additionally performed an inverse normalization prior to analysis; SAIGE-GENE+ (BINA) considered the categorical data as a binary phenotype by grouping individuals in the last two categories.

Results

False positive rate and statistical power

Simulation studies showed the empirical type I error rates of POLMM-GENE, SAIGE-GENE+ (BINA), SAIGE-GENE+ (RAW), and SAIGE-GENE+ (INT) at three significance levels of 5e−6, 5e−4, and 0.05 (Figure 1). For POLMM-GENE and SAIGE-GENE+ (BINA) approaches, the burden test, SKAT, and SKAT-O all reasonably controlled type I error rates regardless of the sample size distributions and significance levels. Meanwhile, SAIGE-GENE+ (RAW) and SAIGE-GENE+ (INT) could not control type I error rates when the sample size distribution was unbalanced or extremely unbalanced. For example, if the sample size distributions were 10:1:1 and 30:1:1, the empirical type I error rates of SAIGE-GENE+ (RAW) at the significance level $\alpha = 5\times10^{-6}$ were greater than 7.1e−5 (i.e., $14.2\times\ \alpha$) and 6.4e−4 (i.e., $128\times\ \alpha$), respectively. Even with the inverse normalization transformation (INT), SAIGE-GENE+ (INT) was still not reliable if the sample

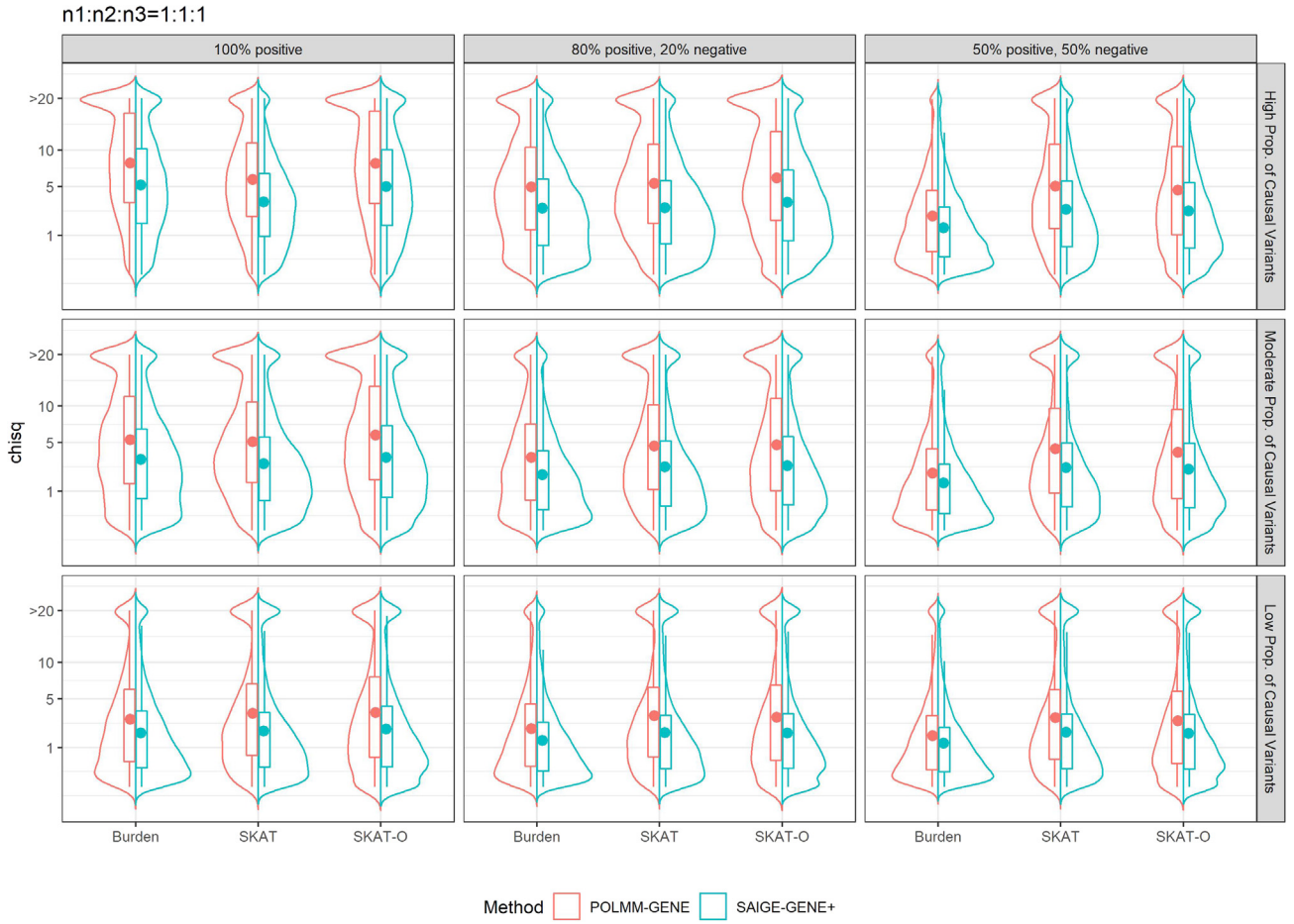

Figure 2. Distribution of chi-square statistics of POLMM-GENE and SAIGE-GENE+ (BINA)
The sample size distribution of the three categorical levels was n1:n2:n3 = 1:1:1. SAIGE-GENE+ (BINA) considered the categorical data as a binary phenotype (n1:n2+n3 = 1:2). A total of nine scenarios include three settings of causal variants proportional and three settings of the effect directions. For high proportion of causal variants, we simulated 80% of LoF and 50% of missense variants as causal variants; for moderate proportion of causal variants, we simulated 50% of LoF and 20% of missense variants as causal variants; for low proportion of causal variants, we simulated 20% of LoF and 10% of missense variants as causal variants.

size distribution was unbalanced or extremely unbalanced, in which the empirical type I error rates were greater than 3.8e−5 (i.e., 7.6 × α) and 4.0e−4 (i.e. 80 × α), respectively.

To evaluate the empirical power of POLMM-GENE and SAIGE-GENE+ approaches, we demonstrate the empirical distribution of the chi-square statistics derived from p values in Figures 2 and S3–S6. The chi-square statistics ≥ 20 (the corresponding p value = 7.7e−6) were collapsed. The p value comparison for POLMM-GENE and SAIGE-GENE+ (BINA) is shown in Figure S7. Compared to SAIGE-GENE+ (BINA), the median chi-square statistics of POLMM-GENE were increased by 1.43–2.13 times, 1.17–1.54 times, and 1.10–1.39 times if the sample size distributions were 1:1:1, 10:1:1, and 30:1:1, respectively. This result suggests that POLMM-GENE is more powerful than SAIGE-GENE+ (BINA). This is expected as SAIGE-GENE+ (BINA) combined two levels of categories, which can cause the loss of phenotypic information. For POLMM-GENE, in line with previously reported for binary and quantitative traits,[13] the burden test was more powerful than SKAT if the proportion of causal variants was high and most of the causal variants were of the same effect direction. Otherwise, SKAT was more powerful than the burden test. As a combination of the burden test and SKAT, SKAT-O always performed the best or close to the best. The simulation results are expected and consistent with previous studies,[14] all of which suggest that SKAT-O is an optimal approach for set-based testing.

Figure S8 demonstrated the empirical power comparison at significance level of 2.5×10^{-6}, which also showed that POLMM-GENE was more powerful than SAIGE-GENE+ (BINA). When the sample size distribution was unbalanced or extremely unbalanced, SAIGE-GENE+ (RAW) and SAIGE-GENE+ (INT) identified more significant findings. This was consistent with the simulation results that these two approaches cannot control type I error rates well in these scenarios.

Application to UK Biobank data

The computation time and cost in UK Biobank RAP to analyze 450k whole-exome-sequencing data are shown in

Table S1. In general, analyzing one chromosome requires less than 8 h, and the total cost for one phenotype is less than 5 Great British pounds. Fifty-four genes were identified for the five phenotypes (Table 1 and Figure 3) at the exome-wide significant threshold p value < 2.5e−6, Most of these genes (45/54) were associated with comparative height size at age 10. The below gives more details of the identified associations, including both previously reported associations, such as an association between gene *PER2* (MIM: 603426)/*PER3* (MIM: 603427) and chronotype, and potentially novel associations, such as an association between gene *GIGYF1* (MIM: 612064) and alcohol intake frequency.

ADH1C (MIM: 103730) and *GIGYF1* (MIM: 612064) were identified to be associated with alcohol intake frequency, in which p values are 1.11e−15 and 1.06e−6, respectively. Gene *ADH1C* encodes class I alcohol dehydrogenase (ADH), gamma subunit. The associations between *ADH1C* and alcohol dependence[15] and alcohol-associated diseases (i.e., alcoholic liver cirrhosis [MIM: 215600],[16] upper aerodigestive tract cancer [MIM: 133239][17]) have been widely reported in previous studies. ADH enzyme is involved in alcohol metabolism and affects the response to alcohol. Genetic variation in *ADH1C* was relevant to the ethanol elimination rate in Whites individuals.[18] *In vitro* kinetic studies suggested that the ADH enzyme encoded by the ADH1C*1 allele metabolizes ethanol to acetaldehyde (AA) 2.5 times faster than that encoded by the ADH1C*2 allele.[19]

GIGYF1 encodes a member of the GYF family of adaptor proteins. It links to activated insulin receptors and insulin-like growth factor-1 (IGF-1) by the growth factor receptor-bound 10 (GRB10) and negatively regulates receptor signaling, metabolic responses, and IGF1-induced mitogenesis.[20,21] In a study conducted on participants in the UK Biobank and an independent validation cohort from the Geisinger Health System, *GIGYF1* predicted LoF variants associated with increased levels of glucose and risk of diabetes (MIM: 125853).[22] A positive association between moderate alcohol consumption and insulin sensitivity has been reported in a previous clinical study,[23] which may explain the association of alcohol consumption with *GIGYF1* genetic variants.

Genes significantly associated with morning/evening person (chronotype) include *PER3* (p value = 3.03e−18), *MTNR1B* (MIM: 600804) (p value = 3.9e−12), and *PER2* (p value = 4.3e−10). *PER2* and *PER3* are core circadian clock genes. A transcription/translation feedback loop (TTFL) forms the core of the molecular circadian clock mechanism. *PER2* and *PER3* are transcriptional repressors that form the negative limb of the feedback loop and interact with a heterodimer to inhibit its activity and thereby negatively regulate their own expression. Genetic variants of *PER2* and *PER3* are involved in many common sleep disorders[24,25] and circadian phenotypes.[26,27] The two genes may show specific diurnal preference, while *PER2* (rs934945) was once reported associated with "morning alertness" and *PER3* (rs2640909) associated with "morningness."[28]

MTNR1B, encoding a high-affinity receptor for melatonin, is an inhibitory G-protein-coupled receptor and may be involved in the neurobiological effects of melatonin. Melatonin is naturally secreted by the pineal gland during the biological night in humans. Its primary physiological function is to convey information on the timing and length of the night to the rest of the body.[29] Lane et al. suggested that rs10830963 variation in *MTNR1B* may influence dim-light melatonin offset through changes in sleep timing or that *MTNR1B* variation may influence sleep timing through changes in the timing of the melatonin profile.[30] It is worth mentioning that the discovery that genetic variation in *MTNR1B* is a risk factor for impaired fasting glucose and diabetes aroused the interest in circadian disruption on glucose metabolism.[31] To deeply explore the physiological function of *MTNR1B* on circadian disruption would contribute to the diabetes management.

MRGPRX1 (MIM: 607227) was identified to be associated with cognitive symptoms severity (p value = 1.89e−7). *MRGPRX1* is a protein-coding gene. Sensory neuron-specific Mas-related G protein-coupled receptors-X1 (MRGPR-X1) are primate-specific proteins with a putative role in nociception and pruritus.[32,33] This receptor is selectively enriched in dorsal root ganglion neurons and is activated by a variety of endogenous peptides. Previous studies found that the naturally occurring mutations of this gene can alter the pharmacology of MRGPR-X1,[34] but limited signaling pathways have been identified related to MRGPR-X1 and cognition performance.

For the phenotype of "comparative body size at age 10," we identified genes of *MC4R* (MIM: 155541) and *CALCR* (MIM: 114131). *MC4R* encodes the melanocortin-4 receptor, a key component of the leptin-melanocortin pathway, which plays a central role in the regulation of energy homeostasis and body weight.[35–37] Genetic variants in *MC4R* were identified among individuals with severe obesity (MIM: 618408) from early childhood.[38,39] The importance of the *MC4R* in maintaining energy homeostasis has made it a compelling target for the potential treatment of obesity diseases. *CALCR* encodes a receptor of calcitonin and amylin, which belongs to a subfamily of seven transmembrane-spanning G protein-coupled receptors.[40] Amylin is a peptide hormone, which has been shown to promote satiety, delayed gastric emptying, and weight control in individuals with type 2 diabetes.[41,42] Furthermore, fine-mapping analysis of 645,626 individuals showed protein-truncating variants in *CALCR* were associated with higher BMI and obesity risk in humans.[43] Ablation of Calcr-expressing neurons in the nucleus tractus solitarius has been recently shown to abrogate the long-term suppression of food intake in mice models.[44]

The phenotype of "comparative height size at age 10" contributes to the most significant gene-level associations. Using the same dataset in UK Biobank (subject selection is slightly different), Backman et al.[2] also conducted genome-wide gene-based association studies on comparative height size and body size at age 10. Backman et al.

Table 1. Genes identified by POLMM-GENE that reached the exome-wide significant threshold with p values < 2.5e−6

	LoF: p value (number of variants)		LoF+Missense: p value (number of variants)		
Genes	**Ultra-rare variants (MAC ≤ 10)**	**Rare variants (MAF < 1%)**	**Ultra-rare variants (MAC ≤ 10)**	**Rare variants (MAF < 1%)**	**Cauchy p value**
Alcohol intake frequency					
ADH1C (MIM: 103730)	4.64e−01 (n = 18, MAC = 40)	4.98e−16 (n = 22)	8.49e−01 (n = 155, MAC = 405)	2.61e−16 (n = 178)	1.11e−15
GIGYF1 (MIM: 612064)	3.02e−07 (n = 98, MAC = 164)	5.74e−06 (n = 113)	2.05e−02 (n = 626, MAC = 1,604)	2.55e−02 (n = 767)	1.06e−06
Morning/evening person (chronotype)					
PER3 (MIM: 603427)	2.26e−02 (n = 92, MAC = 229)	2.45e−05 (n = 105)	4.30e−02 (n = 611, MAC = 1,513)	5.05e−19 (n = 759)	3.03e−18
MTNR1B (MIM: 600804)	1.39e−01 (n = 24, MAC = 52)	1.84e−01 (n = 26)	1.05e−02 (n = 180, MAC = 511)	6.50e−13 (n = 218)	3.90e−12
PER2 (MIM: 603426)	4.10e−10 (n = 80, MAC = 148)	2.15e−10 (n = 85)	1.66e−01 (n = 568, MAC = 1,373)	5.88e−05 (n = 675)	4.30e−10
Cognitive symptoms severity (ultra-rare variants are defined as MAC ≤ 5)					
MRGPRX1 (MIM: 607227)	4.77e−02 (n = 7, MAC = 10)	6.31e−08 (n = 9)	8.43e−01 (n = 86, MAC = 182)	7.15e−01 (n = 122)	1.89e−07
Comparative body size at age 10					
MC4R (MIM: 155541)	3.35e−07 (n = 11, MAC = 27)	1.11e−23 (n = 17)	2.04e−05 (n = 122, MAC = 280)	1.03e−22 (n = 168)	1.98e−33
EPHB2 (MIM: 600997)	2.50e−01 (n = 24, MAC = 55)	3.10e−01 (n = 28)	5.28e−02 (n = 416, MAC = 1,160)	2.58e−07 (n = 504)	1.55e−06
CALCR (MIM: 114131)	3.02e−06 (n = 44, MAC = 138)	1.03e−05 (n = 56)	6.45e−04 (n = 234, MAC = 616)	1.59e−04 (n = 289)	1.59e−06
Comparative height size at age 10					
ZFAT (MIM: 610391)	4.56e−10 (n = 46, MAC = 89)	2.58e−10 (n = 48)	9.83e−10 (n = 505, MAC = 1,339)	7.77e−30 (n = 628)	4.66e−29
ACAN (MIM: 155760)	1.96e−06 (n = 43, MAC = 75)	1.37e−06 (n = 47)	5.23e−03 (n = 934, MAC = 2,519)	1.04e−27 (n = 1,208)	6.26e−27
SCMH1 (MIM: 616396)	9.98e−06 (n = 34, MAC = 69)	1.04e−05 (n = 35)	6.94e−02 (n = 274, MAC = 650)	2.18e−26 (n = 328)	1.31e−25
ADAMTS17 (MIM: 607511)	3.36e−04 (n = 75, MAC = 186)	1.00e−09 (n = 86)	3.48e−06 (n = 649, MAC = 1,798)	1.41e−13 (n = 832)	5.41e−22
NPR3 (MIM: 108962)	6.55e−02 (n = 30, MAC = 71)	7.82e−03 (n = 34)	5.76e−02 (n = 273, MAC = 674)	7.00e−22 (n = 312)	4.20e−21
NPR2 (MIM: 607072)	1.70e−11 (n = 30, MAC = 80)	9.74e−11 (n = 30)	1.43e−13 (n = 358, MAC = 894)	3.34e−18 (n = 437)	8.28e−18
GH1 (MIM: 139250)	2.60e−01 (n = 11, MAC = 13)	2.60e−01 (n = 11)	1.64e−02 (n = 118, MAC = 311)	3.84e−06 (n = 157)	1.06e−17
STC2 (MIM: 603665)	4.23e−01 (n = 5, MAC = 12)	4.23e−01 (n = 5)	3.40e−02 (n = 119, MAC = 357)	9.65e−17 (n = 144)	2.89e−16
FBN2 (MIM: 612570)	5.49e−02 (n = 73, MAC = 119)	8.57e−02 (n = 77)	5.97e−02 (n = 1195, MAC = 2,893)	1.83e−14 (n = 1406)	1.10e−13
ADAMTS10 (MIM: 608990)	5.47e−05 (n = 32, MAC = 48)	1.25e−04 (n = 33)	1.45e−10 (n = 460, MAC = 1,106)	2.17e−04 (n = 565)	4.80e−13
PDE3B (MIM: 602047)	2.83e−03 (n = 63, MAC = 139)	9.02e−12 (n = 78)	4.92e−02 (n = 472, MAC = 1,147)	5.82e−10 (n = 584)	5.33e−11
PIEZO1 (MIM: 611184)	2.03e−06 (n = 229, MAC = 509)	1.52e−04 (n = 266)	9.62e−06 (n = 2047, MAC = 5,529)	9.09e−08 (n = 2,646)	1.64e−10
MC3R (MIM: 155540)	5.68e−02 (n = 12, MAC = 24)	1.62e−02 (n = 14)	3.68e−01 (n = 156, MAC = 441)	2.93e−11 (n = 183)	1.76e−10

(*Continued on next page*)

Table 1. ***Continued***

Genes	LoF: p value (number of variants)		LoF+Missense: p value (number of variants)		Cauchy p value
	Ultra-rare variants (MAC $\leq$ 10)	Rare variants (MAF $<$ 1%)	Ultra-rare variants (MAC $\leq$ 10)	Rare variants (MAF $<$ 1%)	
DDR2 (MIM: 191311)	5.31e−02 (n = 22, MAC = 46)	5.31e−02 (n = 22)	3.74e−04 (n = 284, MAC = 693)	2.44e−10 (n = 339)	3.65e−10
GHSR (MIM: 601898)	6.07e−04 (n = 21, MAC = 56)	7.30e−07 (n = 25)	6.52e−04 (n = 189, MAC = 468)	2.69e−10 (n = 234)	8.08e−10
SMIM29 (MIM: 611419)	6.43e−02 (n = 7, MAC = 15)	5.65e−01 (n = 12)	4.02e−01 (n = 69, MAC = 186)	1.86e−10 (n = 93)	1.12e−09
GRAMD2A (MIM: 620181)	6.75e−01 (n = 24, MAC = 61)	3.14e−06 (n = 28)	1.47e−01 (n = 125, MAC = 326)	2.16e−10 (n = 159)	1.29e−09
HSD11B2 (MIM: 614232)	8.21e−01 (n = 23, MAC = 39)	6.11e−01 (n = 28)	1.47e−02 (n = 189, MAC = 480)	2.19e−10 (n = 233)	1.31e−09
FGFR3 (MIM: 134934)	9.59e−01 (n = 27, MAC = 61)	1.15e−01 (n = 31)	1.08e−03 (n = 421, MAC = 1,135)	2.36e−10 (n = 533)	1.41e−09
PDE11A (MIM: 604961)	1.49e−01 (n = 68, MAC = 182)	4.31e−06 (n = 87)	4.86e−02 (n = 452, MAC = 1,201)	5.65e−10 (n = 562)	3.39e−09
IHH (MIM: 600726)	4.93e−05 (n = 10, MAC = 14)	4.93e−05 (n = 10)	2.68e−05 (n = 186, MAC = 443)	2.53e−09 (n = 225)	3.75e−09
MICA (MIM: 600169)	9.89e−01 (n = 11, MAC = 37)	4.02e−04 (n = 18)	6.96e−01 (n = 144, MAC = 352)	9.78e−10 (n = 193)	5.87e−09
NPAS4 (MIM: 608554)	1.84e−02 (n = 9, MAC = 11)	1.21e−01 (n = 12)	9.99e−01 (n = 297, MAC = 670)	1.62e−09 (n = 354)	9.72e−09
SCUBE3 (MIM: 614708)	3.66e−06 (n = 47, MAC = 82)	3.66e−06 (n = 47)	1.52e−05 (n = 396, MAC = 966)	4.42e−03 (n = 476)	9.79e−09

For the phenotype of "comparative height size at age 10," the significance level is 1e−8. LoF ultra-rare variants (URVs) and LoF+Missense URVs are the marker of the collapsed ultra-rare variants in the annotation of LoF and LoF+Missense, respectively. Genome position is based on the GRCh38 reference. LoF rare variants and LoF+Missense rare variants are SKAT-O test p values with MAF cutoffs 1%. Cauchy p values are calculated to combine six SKAT-O p values with three MAF cutoffs, including 0.01%, 0.1%, and 1%, and two annotation groups, including LoF and LoF+Missense.

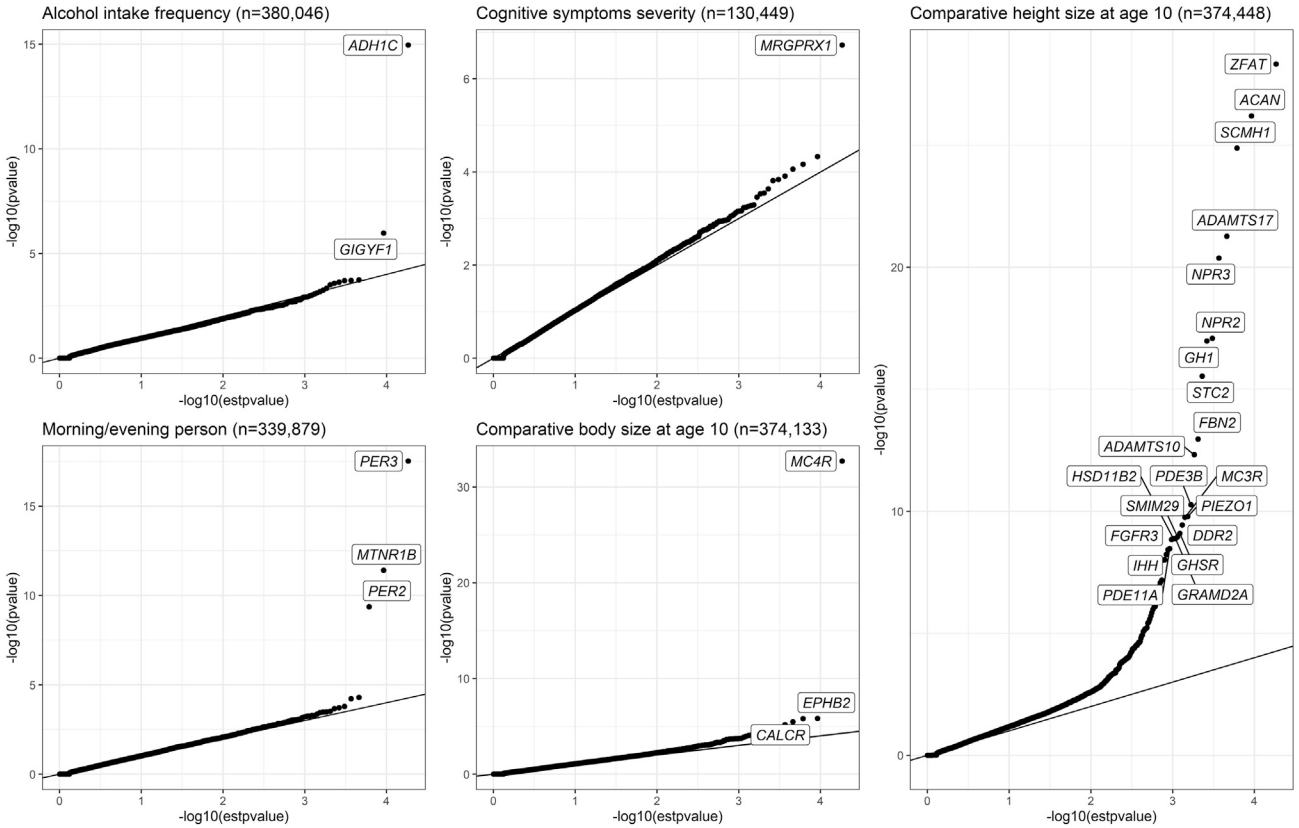

Figure 3. QQ plot of Cauchy combination SKAT-O p values for five ordinal categorical phenotype analyses
For comparative height size at age 10, genes with p values < 5e−9 were labeled.

tested for association between 3,994 traits and individual variants in 18,811 genes, as well as with aggregations of variants in each gene, considering either putative loss-of-function (pLoF) or pLoF and deleterious missense variants jointly, all of which sum up to 2.3 billion association tests. At a significance level of 2.18e−11 (0.05/2.3 billion), *ZFAT* (MIM: 610391), *SCUBE3* (MIM: 614708), *NPR3* (MIM: 108962), *STC2* (MIM: 603665), *NPR2* (MIM: 607072), *PDE3B* (MIM: 602047), *ACAN* (MIM: 155760), *ADAMTS17* (MIM: 607511), and *ADAMTS10* (MIM: 608990) were identified. Backman et al. transformed the original categorical phenotypes into binary traits (taller/average versus shorter and taller versus average/shorter) and then used the burden test for analysis.[2] Via POLMM-GENE, seven of the nine genes (*NPR3*, *STC2*, *ZFAT*, *NPR2*, *ACAN*, *ADAMTS17*, *ADAMTS10*) identified by Backman et al. reached the same significance level. The p values of the other two genes *SCUBE3* and *PDE3B* were 9.79e−9 and 5.33e−11, respectively, which also reached the genome-wide significance level. In addition to the nine genes, we also identified 36 genes at a significance level of 2.5e−6, three of which reached the significance level of 2.18e−11 (*SCMH1* [MIM: 616396], p value = 1.31e−25; *FBN2* [MIM: 612570], p value = 1.1e−13; *GH1* [MIM: 139250], p value = 1.06e−17). The associations between these three genes and height have been previously reported.[45,46] With Cauchy combination, POLMM-GENE gives an optimal unified p value by combining p values from multiple MAF cutoffs and multiple annotations, which reduces the number of tests and thus reduces false negative results caused by the multiple comparison problem.

In addition to the proposed POLMM-GENE, we also evaluated SAIGE-GENE+ (RAW), SAIGE-GENE+ (INT), and SAIGE-GENE+ (BINA). More details and results can be seen in Figures S9–S12. The real data analysis was consistent with the simulation results. SAIGE-GENE+ (BINA) was less powerful than POLMM-GENE as a result of the information lost when transforming a categorical trait to a binary trait (see Figure S12). The p values of SAIGE-GENE+ (RAW) and SAIGE-GENE+ (INT) showed a slight inflation. For example, when analyzing categorical trait of "alcohol intake frequency," the genome control (gc) lambda for these two approaches were 1.157 and 1.154, respectively.

Discussion

In this study, we develop POLMM-GENE, a set-based testing approach to associating an ordinal categorical phenotype to a set of multiple rare and ultra-rare variants. POLMM-GENE is scalable to analyze large-scale biobank data with hundreds of thousands of samples and can adjust for sample relatedness. Simulation studies demonstrated that POLMM-GENE could better control type I error rates while gaining higher

power than set-based methods that treat raw ordinal categorical phenotypes as quantitative or binary traits. We applied POLMM-GENE to analyze ordinal categorical phenotypes by using UK Biobank 450k whole-exome-sequencing data. The real data analyses identified several well-known gene-trait associations, including alcohol consumption and *ADH1C*, chronotype and *PER3*, etc. In addition, we also identified several promising findings, such as the association between cognitive symptoms severity and *MRGPRX1*.

It is expected that more and more biobank-scale whole-exome- and whole-genome-sequencing data will be accessible in the next decade. SKAT-O is an optimal unified approach by incorporating both SKAT and the burden test. Simulation studies demonstrate that SKAT-O is always the best or close to the best in all scenarios, which suggests the superior performance of SKAT-O. As the increase of sample size in sequencing data analysis, the number of rare and ultra-rare variants increases significantly, which could result in inflated type I error rates for SKAT and SKAT-O. POLMM-GENE uses SPA and ultra-rare variant-collapsing methods and thus is robust even if the MAF cutoff is small and phenotypic distribution is highly skewed. In addition, POLMM-GENE uses the Cauchy combination to efficiently combine p values calculated with multiple functional annotations, which reduces the multiple comparison burden while remaining powerful to fully utilize annotation information.

In summary, we have proposed POLMM-GENE, a robust set-based method for ordinal categorical data analysis. The method is scalable for biobank data analysis, can adjust for sample relatedness, and is accurate even if the phenotypic distribution is unbalanced. With all of these features, POLMM-GENE is a valid method developed for conducting robust and powerful set-based tests for the ordinal categorical phenotypes in large-scale biobank data and will contribute to the identification of genetic components of complex traits. POLMM-GENE is implemented in an R package, GRAB, and the UK Biobank analysis results are publicly available (see web resources).

Data and code availability

The POLMM-GENE approach is available in an R package, GRAB, from https://wenjianbi.github.io/grab.github.io/. The UK Biobank data analyses were accessed under the accession number 78795.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.03.010.

Acknowledgments

This research was supported by National Key Research and Development Program of China (2021YFF1201201, W.B.), National Natural Science Foundation of China (62273010, W.B.), National Human Genome Research Institute of the National Institutes of Health (T32HG010464 and K99HG012222-01, W.Z.), and the Brain Pool Plus Program (BP+, Brain Pool+) through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2020H1D3A2A03100666, S.L.). UK Biobank data were accessed under the accession number 78795. This research is supported by high-performance computing platform of Peking University.

Declaration of interests

The authors declare no competing interests.

Received: October 24, 2022
Accepted: March 13, 2023
Published: April 4, 2023

Web resources

ANNOVAR (June 8, 2020), https://annovar.openbioinformatics.org/en/latest/
GRAB (0.0.3.3), https://wenjianbi.github.io/grab.github.io/
SAIGE (1.0.9), https://github.com/saigegit/SAIGE
UK Biobank analysis results, https://doi.org/10.5281/zenodo.7112241
UKBB RAP, https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform

References

1. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. USA *111*, E455–E464.
2. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature *599*, 628–634. https://doi.org/10.1038/s41586-021-04103-z.
3. Bi, W., Zhou, W., Dey, R., Mukherjee, B., Sampson, J.N., and Lee, S. (2021). Efficient mixed model approach for large-scale genome-wide association studies of ordinal categorical phenotypes. Am. J. Hum. Genet. *108*, 825–839. https://doi.org/10.1016/j.ajhg.2021.03.019.
4. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341. https://doi.org/10.1038/s41588-018-0184-y.
5. Zhou, W., Zhao, Z., Nielsen, J.B., Fritsche, L.G., LeFaive, J., Gagliano Taliun, S.A., Bi, W., Gabrielsen, M.E., Daly, M.J., Neale, B.M., et al. (2020). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nat. Genet. *52*, 634–639. https://doi.org/10.1038/s41588-020-0621-6.
6. Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Celedón, J.C., et al. (2016). Control for population structure and relatedness for

binary traits in genetic association studies via logistic mixed models. Am. J. Hum. Genet. *98*, 653–666. https://doi.org/10.1016/j.ajhg.2016.02.012.

7. Holtbrügge, W., and Schumacher, M. (1991). A comparison of regression models for the analysis of ordered categorical data. Appl. Stat. *40*, 249–259.
8. Zhao, Z., Bi, W., Zhou, W., VandeHaar, P., Fritsche, L.G., and Lee, S. (2020). UK biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. Am. J. Hum. Genet. *106*, 3–12. https://doi.org/10.1016/j.ajhg.2019.11.012.
9. Zhou, W., Bi, W., Zhao, Z., Dey, K.K., Jagadeesh, K.A., Karczewski, K.J., Daly, M.J., Neale, B.M., and Lee, S. (2022). SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. Nat. Genet. *54*, 1466–1469. https://doi.org/10.1038/s41588-022-01178-w.
10. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., Aslibekyan, S., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat. Genet. *52*, 969–983. https://doi.org/10.1038/s41588-020-0676-4.
11. Liu, Y., and Xie, J. (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. J. Am. Stat. Assoc. *115*, 393–402. https://doi.org/10.1080/01621459.2018.1554485.
12. Liu, Y., Chen, S., Li, Z., Morrison, A.C., Boerwinkle, E., and Lin, X. (2019). ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. Am. J. Hum. Genet. *104*, 410–421. https://doi.org/10.1016/j.ajhg.2019.01.002.
13. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
14. Lee, S., Wu, M.C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. Biostatistics *13*, 762–775.
15. Sanchez-Roige, S., Palmer, A.A., Fontanillas, P., Elson, S.L., 23andMe Research Team; and the Substance Use Disorder Working Group of the Psychiatric Genomics Consortium, Adams, M.J., Howard, D.M., Edenberg, H.J., Davies, G., et al. (2019). Genome-wide association study meta-analysis of the alcohol use disorders identification test (AUDIT) in two population-based cohorts. Am. J. Psychiatry *176*, 107–118. https://doi.org/10.1176/appi.ajp.2018.18040369.
16. Tolstrup, J.S., Grønbaek, M., Tybjaerg-Hansen, A., and Nordestgaard, B.G. (2009). Alcohol intake, alcohol dehydrogenase genotypes, and liver damage and disease in the Danish general population. Am. J. Gastroenterol. *104*, 2182–2188. https://doi.org/10.1038/ajg.2009.370.
17. Visapää, J.P., Götte, K., Benesova, M., Li, J., Homann, N., Conradt, C., Inoue, H., Tisch, M., Hörrmann, K., Väkeväinen, S., et al. (2004). Increased cancer risk in heavy drinkers with the alcohol dehydrogenase 1C*1 allele, possibly due to salivary acetaldehyde. Gut *53*, 871–876. https://doi.org/10.1136/gut.2003.018994.
18. Martínez, C., Galván, S., Garcia-Martin, E., Ramos, M.I., Gutiérrez-Martín, Y., and Agúndez, J.A.G. (2010). Variability in ethanol biodisposition in whites is modulated by polymorphisms in the ADH1B and ADH1C genes. Hepatology *51*, 491–500. https://doi.org/10.1002/hep.23341.
19. Bosron, W.F., and Li, T.K. (1986). Genetic polymorphism of human liver alcohol and aldehyde dehydrogenases, and their relationship to alcohol metabolism and alcoholism. Hepatology *6*, 502–510. https://doi.org/10.1002/hep.1840060330.
20. Giovannone, B., Lee, E., Laviola, L., Giorgino, F., Cleveland, K.A., and Smith, R.J. (2003). Two novel proteins that are linked to insulin-like growth factor (IGF-I) receptors by the Grb10 adapter and modulate IGF-I signaling. J. Biol. Chem. *278*, 31564–31573. https://doi.org/10.1074/jbc.M211572200.
21. Zhao, Y., Stankovic, S., Koprulu, M., Wheeler, E., Day, F.R., Lango Allen, H., Kerrison, N.D., Pietzner, M., Loh, P.R., Wareham, N.J., et al. (2021). GIGYF1 loss of function is associated with clonal mosaicism and adverse metabolic health. Nat. Commun. *12*, 4178. https://doi.org/10.1038/s41467-021-24504-y.
22. Deaton, A.M., Parker, M.M., Ward, L.D., Flynn-Carroll, A.O., BonDurant, L., Hinkle, G., Akbari, P., Lotta, L.A., et al.; Regeneron Genetics Center; and DiscovEHR Collaboration (2021). Gene-level analysis of rare variants in 379,066 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. Sci. Rep. *11*, 21565. https://doi.org/10.1038/s41598-021-99091-5.
23. Schrieks, I.C., Heil, A.L.J., Hendriks, H.F.J., Mukamal, K.J., and Beulens, J.W.J. (2015). The effect of alcohol consumption on insulin sensitivity and glycemic status: a systematic review and meta-analysis of intervention studies. Diabetes Care *38*, 723–732. https://doi.org/10.2337/dc14-1556.
24. Emeklİ, R., İsmaİloğullari, S., Bayram, A., Akalin, H., Tuncel, G., and Dündar, M. (2020). Comparing expression levels of PERIOD genes PER1, PER2 and PER3 in chronic insomnia patients and medical staff working in the night shift. Sleep Med. *73*, 101–105. https://doi.org/10.1016/j.sleep.2020.04.027.
25. Yang, M.Y., Lin, P.W., Lin, H.C., Lin, P.M., Chen, I.Y., Friedman, M., Hung, C.F., Salapatas, A.M., Lin, M.C., and Lin, S.F. (2019). Alternations of circadian clock genes expression and oscillation in obstructive sleep apnea. J. Clin. Med. *8*, 1634. https://doi.org/10.3390/jcm8101634.
26. Zhang, L., Hirano, A., Hsu, P.K., Jones, C.R., Sakai, N., Okuro, M., McMahon, T., Yamazaki, M., Xu, Y., Saigoh, N., et al. (2016). A PERIOD3 variant causes a circadian phenotype and is associated with a seasonal mood trait. Proc. Natl. Acad. Sci. USA *113*, E1536–E1544. https://doi.org/10.1073/pnas.1600039113.
27. Hida, A., Kitamura, S., Katayose, Y., Kato, M., Ono, H., Kadotani, H., Uchiyama, M., Ebisawa, T., Inoue, Y., Kamei, Y., et al. (2014). Screening of clock gene polymorphisms demonstrates association of a PER3 polymorphism with morningness-eveningness preference and circadian rhythm sleep disorder. Sci. Rep. *4*, 6309. https://doi.org/10.1038/srep06309.
28. Ojeda, D.A., Perea, C.S., Niño, C.L., Gutiérrez, R.M., López-León, S., Arboleda, H., Camargo, A., Adan, A., and Forero, D.A. (2013). A novel association of two non-synonymous polymorphisms in PER2 and PER3 genes with specific diurnal preference subscales. Neurosci. Lett. *553*, 52–56. https://doi.org/10.1016/j.neulet.2013.08.016.
29. Arendt, J. (1994). Melatonin and the Mammalian Pineal Gland (Springer Science & Business Media).
30. Lane, J.M., Chang, A.M., Bjonnes, A.C., Aeschbach, D., Anderson, C., Cade, B.E., Cain, S.W., Czeisler, C.A., Gharib, S.A., Gooley, J.J., et al. (2016). Impact of common diabetes risk variant in MTNR1B on sleep, circadian, and melatonin

physiology. Diabetes *65*, 1741–1751. https://doi.org/10.2337/db15-0999.

31. Mason, I.C., Qian, J., Adler, G.K., and Scheer, F.A.J.L. (2020). Impact of circadian disruption on glucose metabolism: implications for type 2 diabetes. Diabetologia *63*, 462–472. https://doi.org/10.1007/s00125-019-05059-6.
32. Li, Z., Tseng, P.Y., Tiwari, V., Xu, Q., He, S.Q., Wang, Y., Zheng, Q., Han, L., Wu, Z., Blobaum, A.L., et al. (2017). Targeting human Mas-related G protein-coupled receptor X1 to inhibit persistent pain. Proc. Natl. Acad. Sci. USA *114*, E1996–E2005. https://doi.org/10.1073/pnas.1615255114.
33. Liu, Q., Tang, Z., Surdenikova, L., Kim, S., Patel, K.N., Kim, A., Ru, F., Guan, Y., Weng, H.J., Geng, Y., et al. (2009). Sensory neuron-specific GPCR Mrgprs are itch receptors mediating chloroquine-induced pruritus. Cell *139*, 1353–1365. https://doi.org/10.1016/j.cell.2009.11.034.
34. Heller, D., Doyle, J.R., Raman, V.S., Beinborn, M., Kumar, K., and Kopin, A.S. (2016). Novel probes establish Mas-related g protein-coupled receptor X1 variants as receptors with loss or gain of function. J. Pharmacol. Exp. Ther. *356*, 276–283. https://doi.org/10.1124/jpet.115.227058.
35. Krashes, M.J., Lowell, B.B., and Garfield, A.S. (2016). Melanocortin-4 receptor-regulated energy homeostasis. Nat. Neurosci. *19*, 206–219. https://doi.org/10.1038/nn.4202.
36. Siljee, J.E., Wang, Y., Bernard, A.A., Ersoy, B.A., Zhang, S., Marley, A., Von Zastrow, M., Reiter, J.F., and Vaisse, C. (2018). Subcellular localization of MC4R with ADCY3 at neuronal primary cilia underlies a common pathway for genetic predisposition to obesity. Nat. Genet. *50*, 180–185. https://doi.org/10.1038/s41588-017-0020-9.
37. Hinney, A., Volckmar, A.L., and Knoll, N. (2013). Melanocortin-4 receptor in energy homeostasis and obesity pathogenesis. Prog. Mol. Biol. Transl. Sci. *114*, 147–191. https://doi.org/10.1016/b978-0-12-386933-3.00005-4.
38. Vaisse, C., Clement, K., Guy-Grand, B., and Froguel, P. (1998). A frameshift mutation in human MC4R is associated with a dominant form of obesity. Nat. Genet. *20*, 113–114. https://doi.org/10.1038/2407.
39. Lotta, L.A., Mokrosiński, J., Mendes de Oliveira, E., Li, C., Sharp, S.J., Luan, J., Brouwers, B., Ayinampudi, V., Bowker, N., Kerrison, N., et al. (2019). Human gain-of-function MC4R variants show signaling bias and protect against obesity. Cell *177*, 597–607.e9. https://doi.org/10.1016/j.cell.2019.03.044.
40. Moore, E.E., Kuestner, R.E., Stroop, S.D., Grant, F.J., Matthewes, S.L., Brady, C.L., Sexton, P.M., and Findlay, D.M. (1995). Functionally different isoforms of the human calcitonin receptor result from alternative splicing of the gene transcript. Mol. Endocrinol. *9*, 959–968. https://doi.org/10.1210/mend.9.8.7476993.
41. Schmitz, O., Brock, B., and Rungby, J. (2004). Amylin agonists: a novel approach in the treatment of diabetes. Diabetes *53*, S233–S238. https://doi.org/10.2337/diabetes.53.suppl_3.s233.
42. Hollander, P.A., Levy, P., Fineman, M.S., Maggs, D.G., Shen, L.Z., Strobel, S.A., Weyer, C., and Kolterman, O.G. (2003). Pramlintide as an adjunct to insulin therapy improves long-term glycemic and weight control in patients with type 2 diabetes: a 1-year randomized controlled trial. Diabetes Care *26*, 784–790. https://doi.org/10.2337/diacare.26.3.784.
43. Akbari, P., Gilani, A., Sosina, O., Kosmicki, J.A., Khrimian, L., Fang, Y.-Y., Persaud, T., Garcia, V., Sun, D., Li, A., et al. (2021). Sequencing of 640,000 exomes identifies *GPR75* variants associated with protection from obesity. Science *373*, eabf8683. https://doi.org/10.1126/science.abf8683.
44. Cheng, W., Gonzalez, I., Pan, W., Tsang, A.H., Adams, J., Ndoka, E., Gordian, D., Khoury, B., Roelofs, K., Evers, S.S., et al. (2020). Calcitonin receptor neurons in the mouse nucleus tractus solitarius control energy balance via the non-aversive suppression of feeding. Cell Metab. *31*, 301–312.e5. https://doi.org/10.1016/j.cmet.2019.12.012.
45. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.
46. Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., et al. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. Nat. Genet. *53*, 1415–1424. https://doi.org/10.1038/s41588-021-00931-x.

The American Journal of Human Genetics, Volume *110*

Supplemental information

Scalable mixed model methods for set-based association studies on large-scale categorical data analysis and its application to exome-sequencing data in UK Biobank

Wenjian Bi, Wei Zhou, Peipei Zhang, Yaoyao Sun, Weihua Yue, and Seunggeun Lee

Figure S1. Minor allele counts distribution of the randomly selected 1,000 genes.

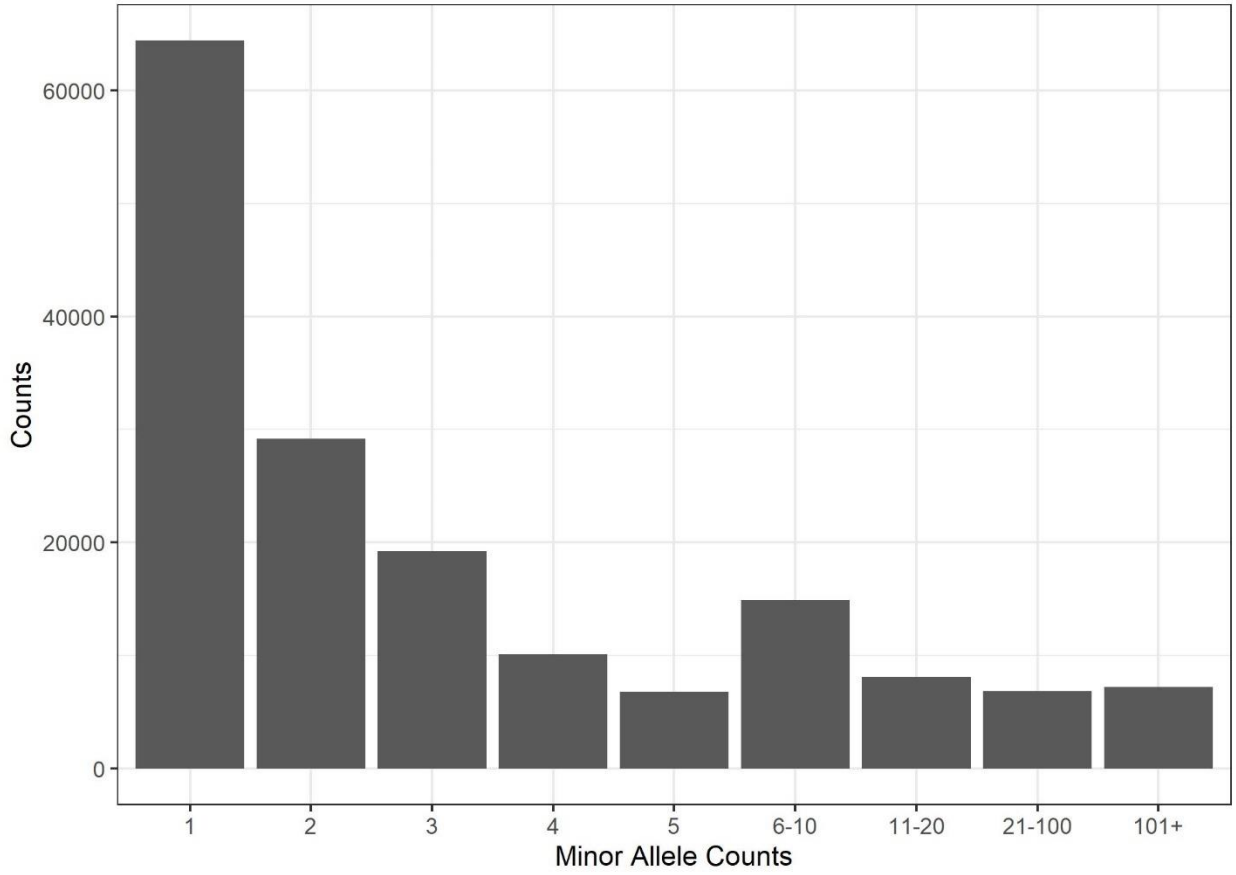

Figure S2. Sample size distribution for the four phenotypes in UK Biobank data analysis

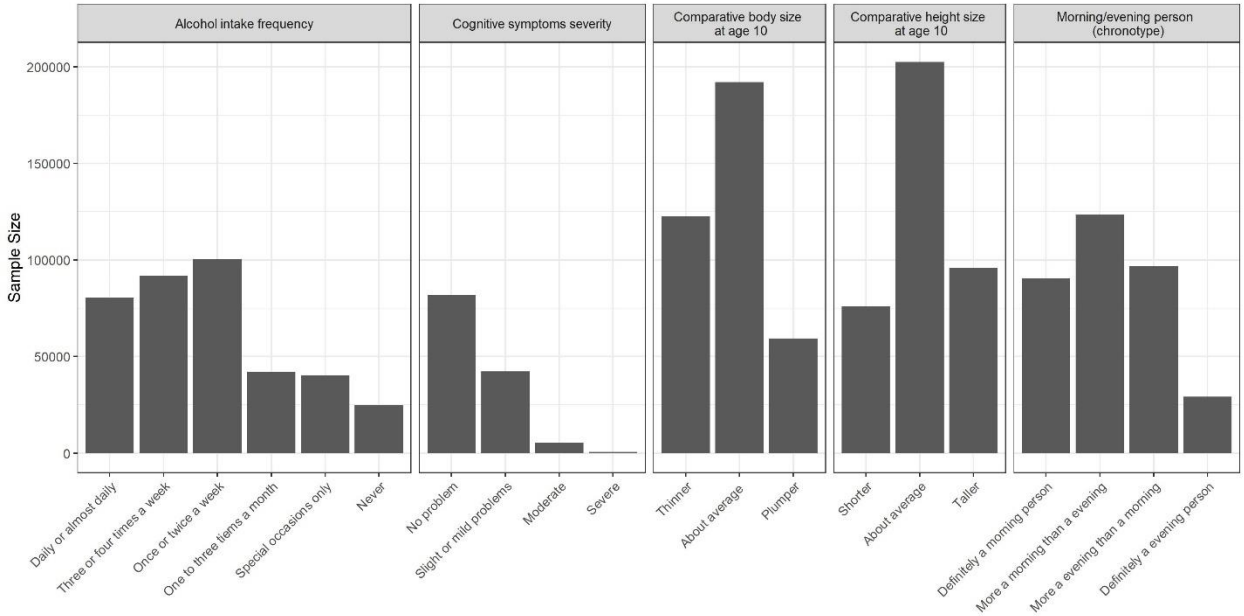

Figure S3. Distribution of chi-square statistics of POLMM-GENE and SAIGE-GENE+ (BINA). The sample size distribution of the three categorical levels is n1:n2:n3=10:1:1. SAIGE-GENE+ considered the categorical data as a binary phenotype (n1:n2+n3=10:2=5:1). A total of 9 scenarios include 3 settings of causal variants proportional and three settings of the effect directions. For high proportion of causal variants, we simulated 80% of LoF and 50% of missense variants as causal variants; for moderate proportion of causal variants, we simulated 50% of LoF and 20% of missense variants as causal variants; for low proportion of causal variants, we simulated 20% of LoF and 10% of missense variants as causal variants.

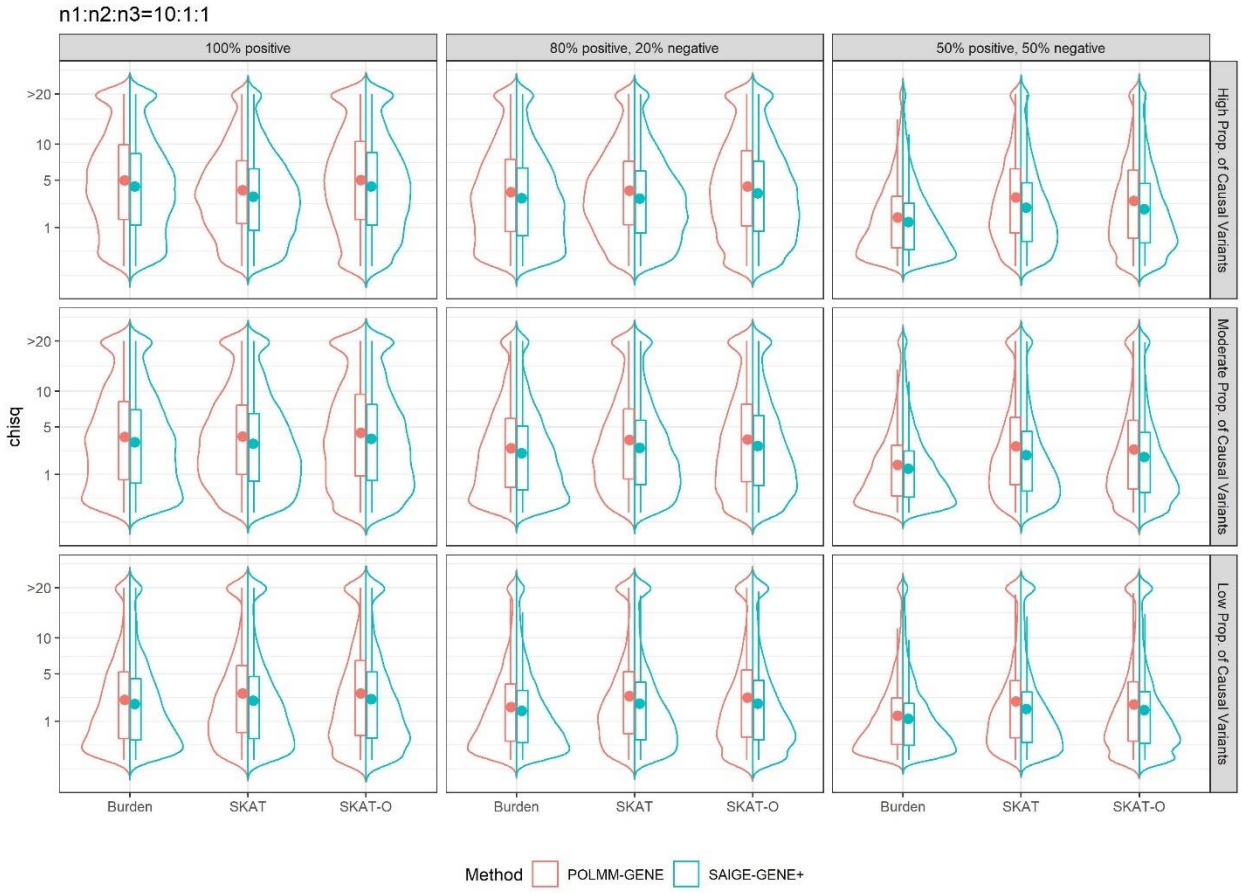

Figure S4. Distribution of chi-square statistics of POLMM-GENE and SAIGE-GENE+ (BINA). The sample size distribution of the three categorical levels is n1:n2:n3=30:1:1. SAIGE-GENE+ considered the categorical data as a binary phenotype (n1:n2+n3=30:2=15:1). A total of 9 scenarios include 3 settings of causal variants proportional and three settings of the effect directions. For high proportion of causal variants, we simulated 80% of LoF and 50% of missense variants as causal variants; for moderate proportion of causal variants, we simulated 50% of LoF and 20% of missense variants as causal variants; for low proportion of causal variants, we simulated 20% of LoF and 10% of missense variants as causal variants.

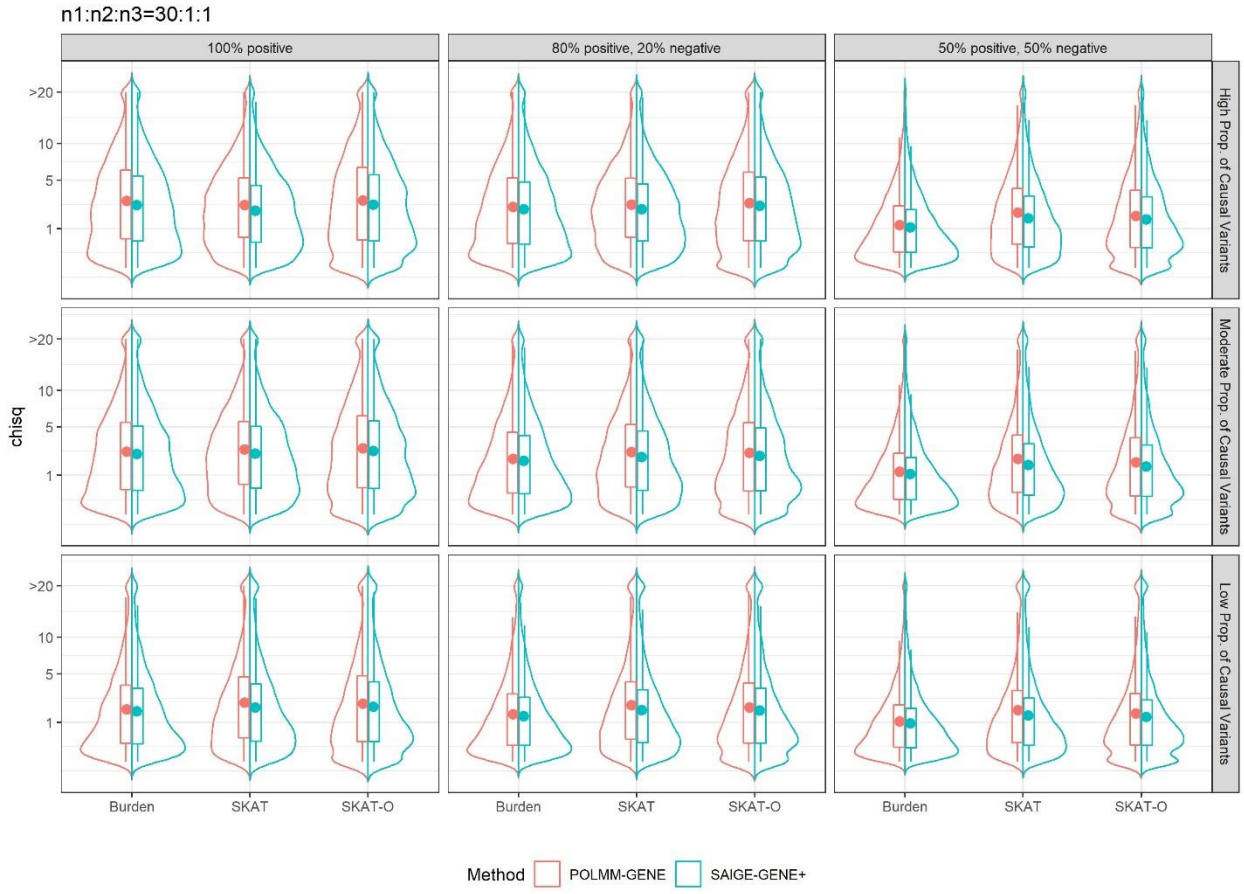

Figure S5. Distribution of chi-square statistics of POLMM-GENE and SAIGE-GENE+ (RAW). SAIGE-GENE+ (RAW) considered the categorical data as a raw quantitative phenotype of 1, 2, and 3. A total of 9 scenarios include 3 settings of causal variants proportional and three settings of the effect directions. For high proportion of causal variants, we simulated 80% of LoF and 50% of missense variants as causal variants; for moderate proportion of causal variants, we simulated 50% of LoF and 20% of missense variants as causal variants; for low proportion of causal variants, we simulated 20% of LoF and 10% of missense variants as causal variants.

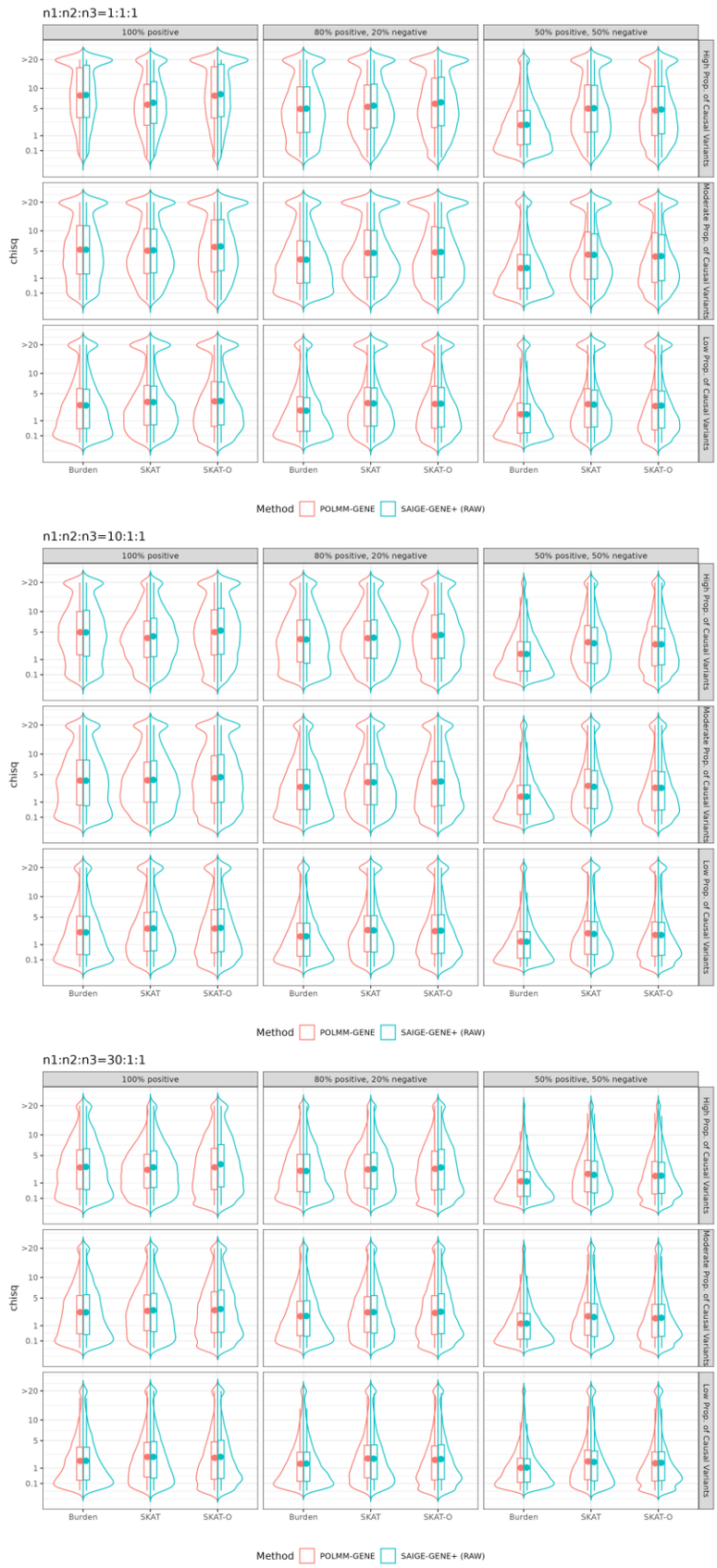

Figure S6. Distribution of chi-square statistics of POLMM-GENE and SAIGE-GENE+ (INT). SAIGE-GENE+ (INT) considered the categorical data as a quantitative phenotype of 1, 2, and 3. Inverse normalization transformation is conducted for phenotype prior to analysis. A total of 9 scenarios include 3 settings of causal variants proportional and three settings of the effect directions. For high proportion of causal variants, we simulated 80% of LoF and 50% of missense variants as causal variants; for moderate proportion of causal variants, we simulated 50% of LoF and 20% of missense variants as causal variants; for low proportion of causal variants, we simulated 20% of LoF and 10% of missense variants as causal variants.

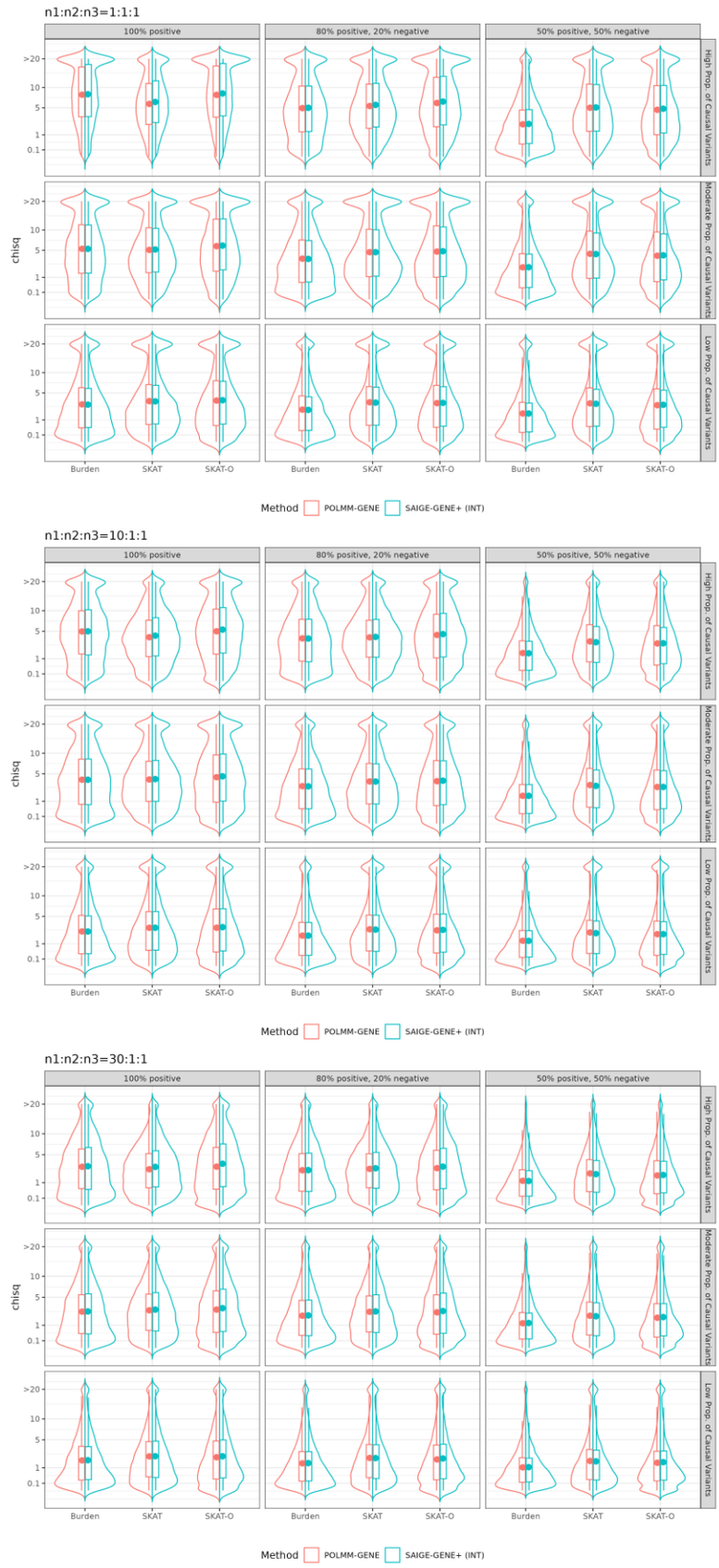

Figure S7. Comparison of p-values using POLMM-GENE and SAIGE-GENE+ (BINA). The sample size distribution of the three categorical levels is n1:n2:n3=1:1:1. SAIGE-GENE+ considered the categorical data as a binary phenotype (n1:n2+n3=1:2). A total of 9 scenarios include 3 settings of causal variants proportional and three settings of the effect directions. For high proportion of causal variants, we simulated 80% of LoF and 50% of missense variants as causal variants; for moderate proportion of causal variants, we simulated 50% of LoF and 20% of missense variants as causal variants; for low proportion of causal variants, we simulated 20% of LoF and 10% of missense variants as causal variants.

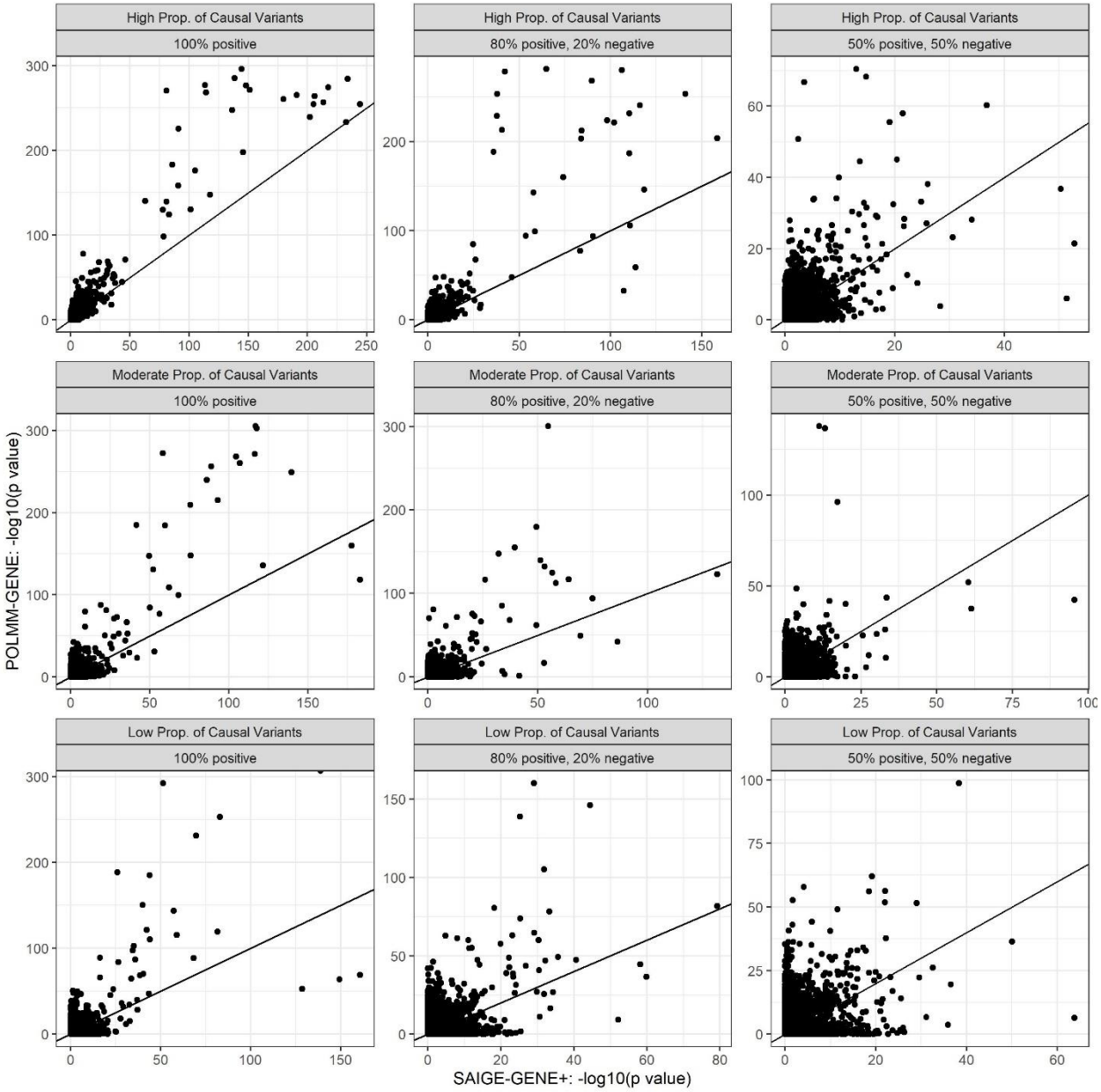

Figure S8. Empirical power of POLMM-GENE, SAIGE-GENE+ (BINA), SAIGE-GENE+ (RAW), and SAIGE-GENE+ (INT) at a significance level of 2.5e-6.

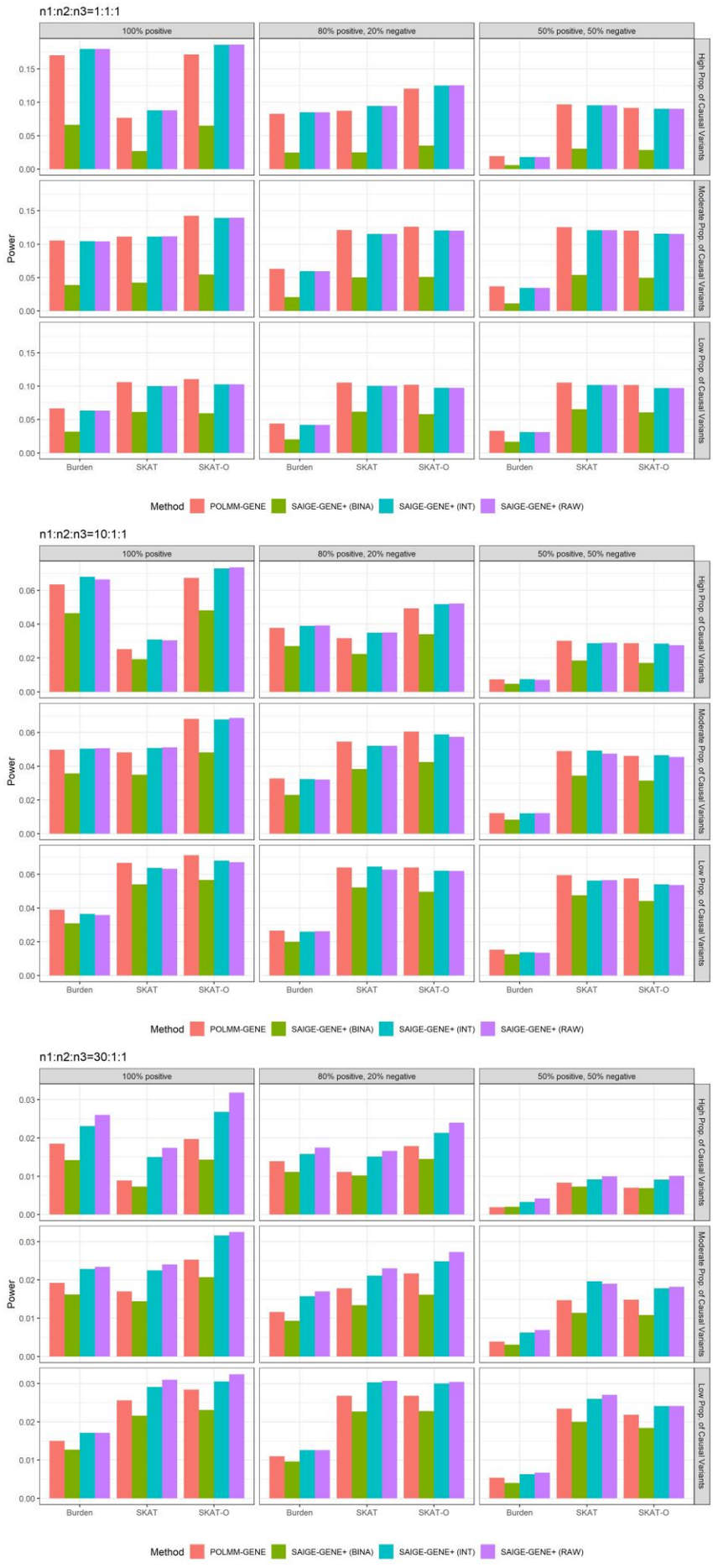

Figure S9. SAIGE-GENE+ (BINA): QQ plot of Cauchy combination SKAT-O p-values for five ordinal categorical phenotype analyses. For comparative height size at age 10, genes with p-values < 5e-9 were labeled. The categorical traits were transformed to a binary trait prior to analysis. For trait of "alcohol intake frequency", categories "daily or almost daily" and "three or four times a week" were grouped, and other categories were grouped. For trait of "cognitive symptoms severity", categories "slight or mild problems", "moderate", and "severe" were grouped. For trait of "comparative body size at age 10", categories "about average" and "plumper" were grouped. For trait of "comparative height size at age 10", categories "shorted" and "about average" were grouped. For trait of "morning/evening person", categories "definitely a morning person" and "more a morning than a evening" were grouped, and the other categories were grouped.

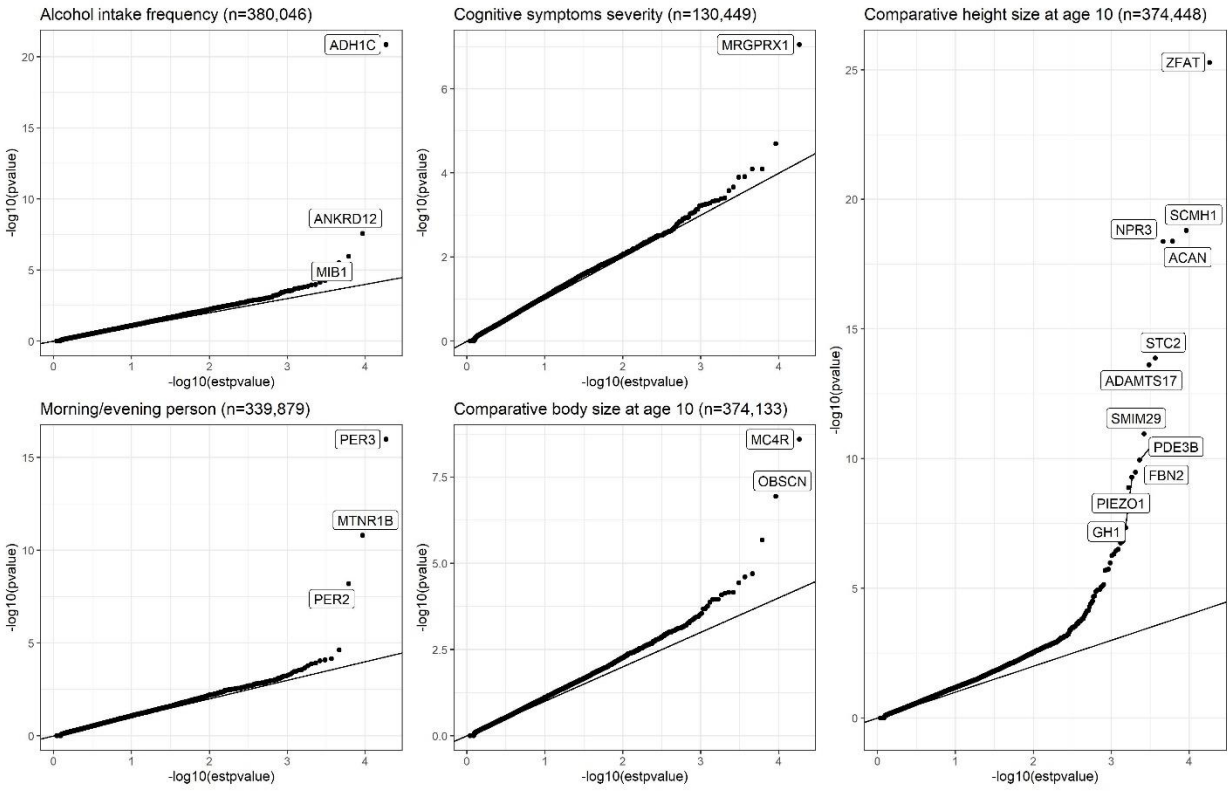

Figure S10. SAIGE-GENE+ (RAW): QQ plot of Cauchy combination SKAT-O p-values for five ordinal categorical phenotype analyses. For comparative height size at age 10, genes with p-values < 5e-9 were labeled. The traits were recorded as $1,2,\ldots,m$ where m is the number of categories.

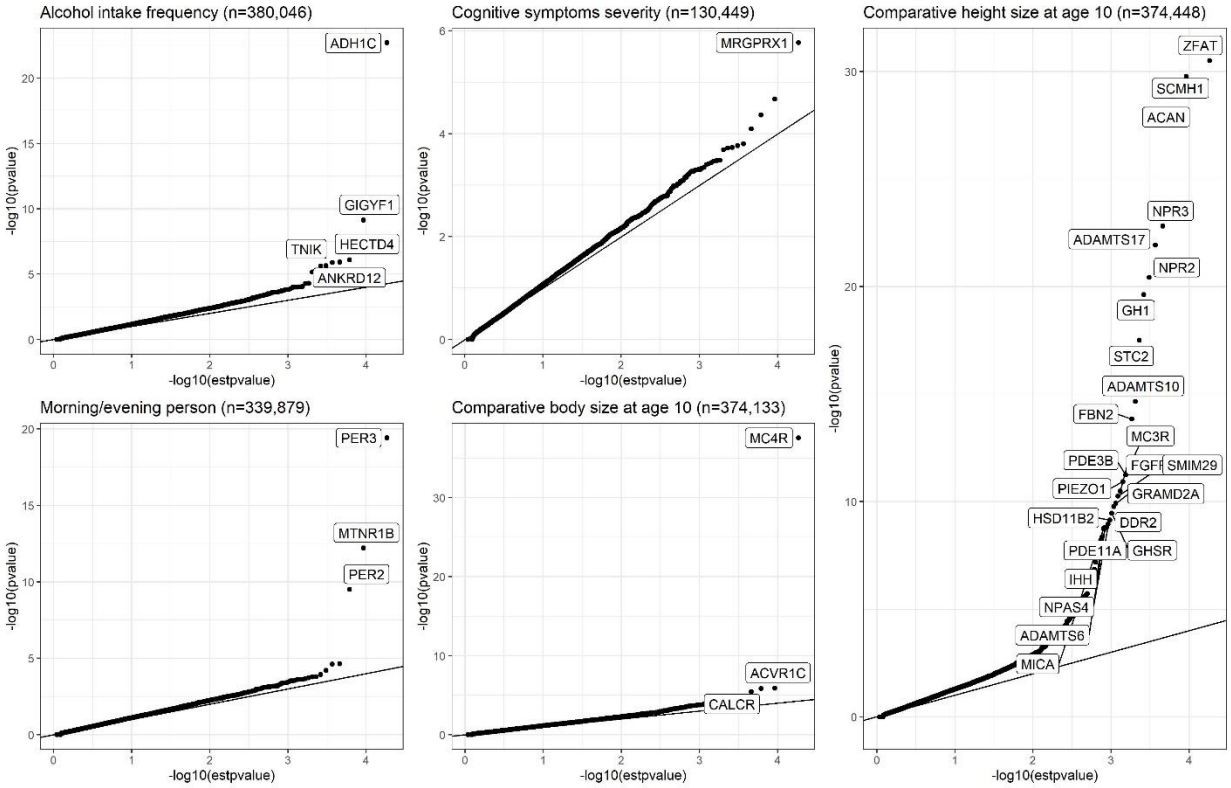

Figure S11. SAIGE-GENE+ (INT): QQ plot of Cauchy combination SKAT-O p-values for five ordinal categorical phenotype analyses. For comparative height size at age 10, genes with p-values < 5e-9 were labeled. The traits were recorded as $1,2,\ldots,m$ where m is the number of categories.

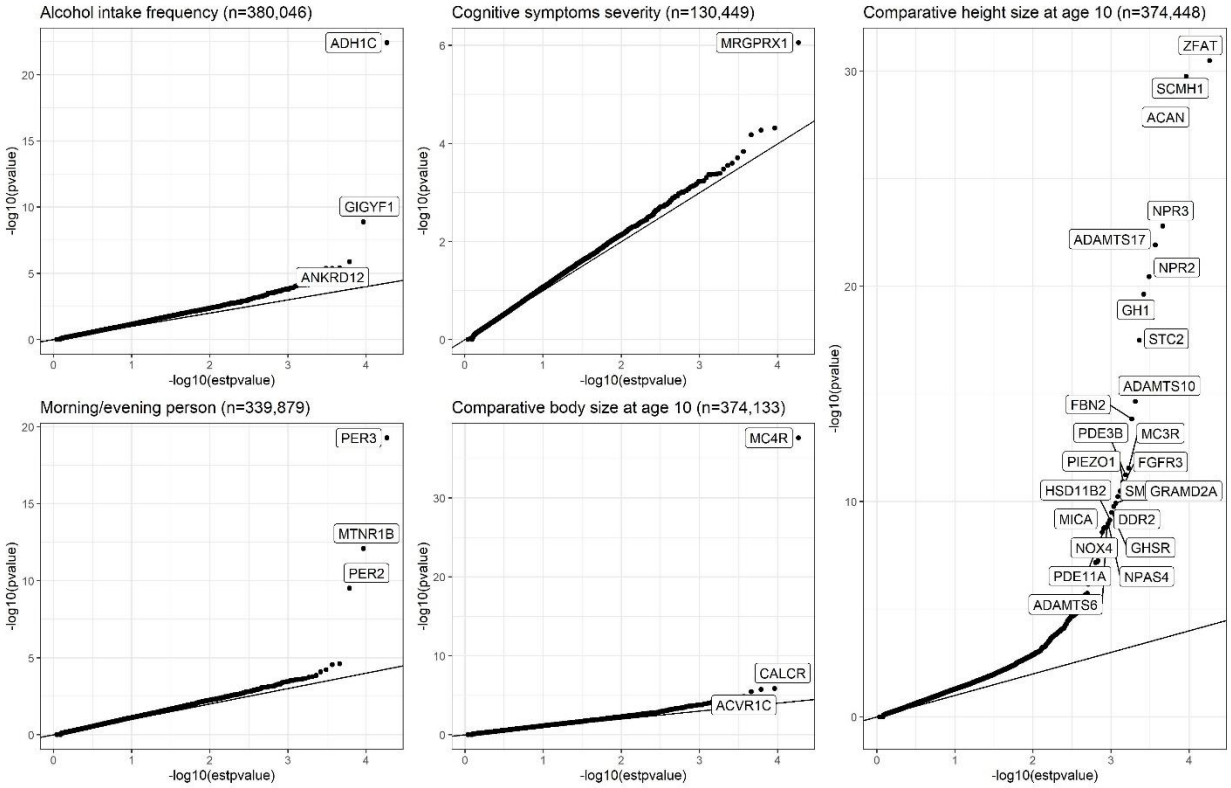

Figure S12. Comparison of POLMM-GENE and SAIGE-GENE+ approaches when analyzing "Comparative height size at age 10"

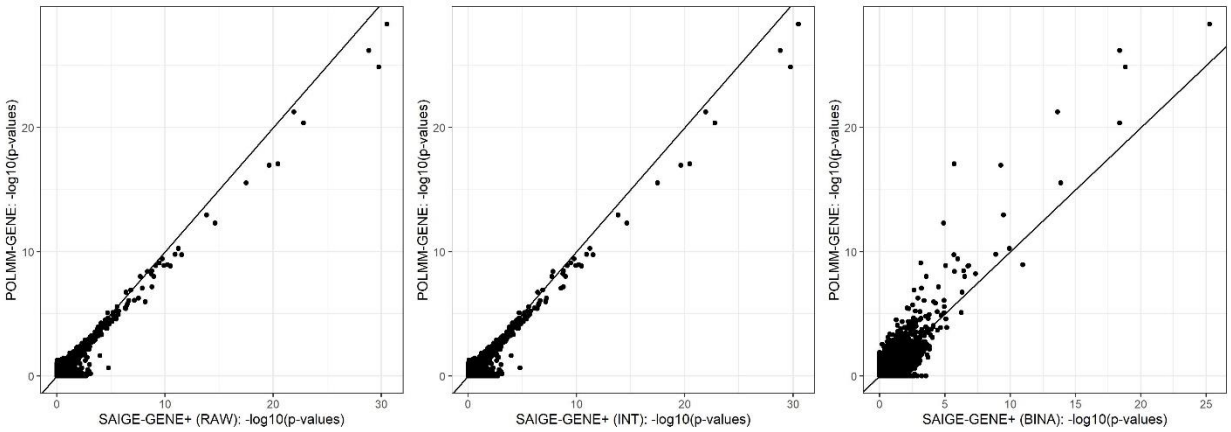

Table S1. Computation time and cost of the 5 ordinal categorical phenotypes analysis in UK Biobank RAP.

		Alcohol intake frequency		Morning/evening person (chronotype)		Comparative height size at age 10		Cognitive symptoms severity		Comparative body size at age 10	
chrom	instance type	Computation time	Cost	Computation time	Cost	Computation time	Cost	Computation time	Cost	Computation time	Cost
chr1	mem1_ssd2_v2_x4	14:05:53	£0.94	7:46:13	£0.52	7:45:53	£0.51	3:37:38	£0.24	7:58:04	£0.53
chr2	mem1_ssd2_v2_x5	9:35:13	£0.64	5:25:19	£0.36	5:26:52	£0.36	2:22:11	£0.16	6:16:15	£0.42
chr3	mem1_ssd2_v2_x4	8:05:40	£0.54	4:18:55	£0.28	4:14:07	£0.28	2:00:11	£0.13	4:29:32	£0.30
chr4	mem2_ssd2_v2_x2	12:24:50	£0.66	3:44:32	£0.12	3:51:08	£0.13	1:46:22	£0.06	3:55:12	£0.13
chr5	mem2_ssd2_v2_x2	7:51:12	£0.26	4:17:44	£0.14	4:08:02	£0.14	2:01:42	£0.07	4:10:49	£0.14
chr6	mem1_ssd2_v2_x4	7:13:12	£0.48	3:52:57	£0.26	3:42:33	£0.24	1:44:56	£0.11	3:44:54	£0.25
chr7	mem1_ssd2_v2_x4	6:44:47	£0.45	3:39:05	£0.24	3:46:44	£0.25	1:36:48	£0.10	3:42:59	£0.24
chr8	mem2_ssd2_v2_x2	6:30:03	£0.22	3:28:35	£0.11	3:27:20	£0.11	1:34:19	£0.05	3:23:22	£0.11
chr9	mem2_ssd2_v2_x2	7:48:58	£0.26	4:08:28	£0.14	4:19:22	£0.14	2:00:00	£0.07	4:19:49	£0.14
chr10	mem2_ssd2_v2_x2	6:48:18	£0.23	3:51:18	£0.13	3:45:04	£0.12	1:51:24	£0.06	3:54:31	£0.13
chr11	mem1_ssd2_v2_x4	9:20:37	£0.62	4:46:54	£0.32	4:54:08	£0.32	2:10:05	£0.14	4:51:50	£0.32
chr12	mem1_ssd2_v2_x4	7:45:54	£0.52	4:00:31	£0.26	3:52:35	£0.26	1:43:22	£0.11	3:47:06	£0.25
chr13	mem2_ssd2_v2_x2	3:27:18	£0.11	1:45:17	£0.06	1:42:03	£0.06	0:51:32	£0.03	1:43:48	£0.06
chr14	mem2_ssd2_v2_x2	6:04:14	£0.20	3:00:39	£0.10	3:02:43	£0.10	1:26:58	£0.05	3:00:47	£0.10
chr15	mem2_ssd2_v2_x2	6:18:19	£0.21	3:28:58	£0.11	3:21:35	£0.11	1:34:35	£0.05	3:14:50	£0.11
chr16	mem1_ssd2_v2_x4	7:14:18	£0.48	3:46:29	£0.25	3:47:20	£0.25	1:42:28	£0.11	3:41:26	£0.24
chr17	mem1_ssd2_v2_x4	8:30:11	£0.56	4:26:48	£0.29	4:24:21	£0.29	2:01:29	£0.13	4:36:51	£0.30
chr18	mem2_ssd2_v2_x2	2:54:58	£0.10	1:30:12	£0.05	1:33:30	£0.05	0:45:34	£0.02	1:39:17	£0.05
chr19	mem1_ssd2_v2_x4	11:04:40	£0.74	5:53:23	£0.39	5:39:49	£0.37	2:30:19	£0.16	6:50:59	£0.45
chr20	mem2_ssd2_v2_x2	4:51:36	£0.16	2:28:39	£0.08	2:26:22	£0.08	1:13:21	£0.04	2:30:49	£0.08
chr21	mem2_ssd2_v2_x2	2:03:10	£0.07	1:08:40	£0.04	1:05:50	£0.04	0:35:45	£0.02	1:08:04	£0.04
chr22	mem2_ssd2_v2_x2	4:30:42	£0.15	2:14:31	£0.07	2:17:47	£0.08	1:07:09	£0.04	2:18:44	£0.08
Total			£8.60		£4.32		£4.29		£1.95		£4.47
The allocated instance: "on-demand" for job of chr4 to analyze "Alcohol intake frequency", "spot" for the other jobs											