Received: 14 June 2021 | Revised: 25 November 2021 | Accepted: 14 February 2022

DOI: 10.1002/sam.11578

WILEY

RESEARCH ARTICLE

A tree-based gene–environment interaction analysis with rare features

Mengque Liu[1] | **Qingzhao Zhang**[2] | **Shuangge Ma**[3]

[1]School of Journalism and New Media, Xi'an Jiaotong University, Shanxi Xi'an, China

[2]Department of Statistics and Data Science, School of Economics, Wang Yanan Institute for Studies in Economics, and Fujian Key Lab of Statistics, Xiamen University, Fujian Xiamen, China

[3]Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA

Correspondence
Qingzhao Zhang, School of Economics, Xiamen University, Xiamen, Fujian, China.
Email: qzzhang@xmu.edu.cn
Shuangge Ma, Department of Biostatistics, Yale University, New Haven, CT 06520, USA.
Email: shuangge.ma@yale.edu

Funding information This research was supported by the National Natural Science Foundation of China (11971404), 111 Project (B13028), National Institutes of Health (CA204120), and National Science Foundation (1916251).

Abstract

Gene–environment (G-E) interaction analysis plays a critical role in understanding and modeling complex diseases. Compared to main-effect-only analysis, it is more seriously challenged by higher dimensionality, weaker signals, and the unique "main effects, interactions" variable selection hierarchy. In joint G-E interaction analysis under which a large number of G factors are analyzed in a single model, effort tailored to rare features (e.g., SNPs with low minor allele frequencies) has been limited. Existing investigations on rare features have been mostly focused on marginal analysis, where various data aggregation techniques have been developed, and hypothesis testings have been conducted to identify significant aggregated features. However, such techniques cannot be extended to joint G-E interaction analysis. In this study, building on a very recent tree-based data aggregation technique, which has been developed for main-effect-only analysis, we develop a new G-E interaction analysis approach tailored to rare features. The adopted data aggregation technique allows for more efficient information borrowing from neighboring rare features. Similar to some existing state-of-the-art ones, the proposed approach adopts penalization for variable selection, regularized estimation, and respect of the variable selection hierarchy. Simulation shows that it has more accurate identification of important interactions and main effects than several competing alternatives. In the analysis of NFBC1966 study, the proposed approach leads to findings different from the alternatives and with satisfactory prediction and stability performance.

KEYWORDS

gene–environment interaction analysis, penalized joint regression, rare features, tree-based aggregation

1 | INTRODUCTION

Gene–environment (G-E) interactions have important implications beyond main G and E effects for understanding and modeling complex human diseases. Compared to main-effect-only analysis, G-E interaction analysis is uniquely challenged by higher data dimensionality, weaker signals, and the unique "main effects, interactions" hierarchy (which postulates that a G-E interaction term cannot be identified if the corresponding main G effect is not identified). G-E interaction analysis can be classified as marginal and joint. In marginal analysis one or a small number of G measurements are analyzed at a time, and thus many analyses are needed. In comparison,

© 2022 Wiley Periodicals LLC.

in joint analysis, a large number of G measurements are analyzed in a single model. In the past decade, we have witnessed significant developments in G-E interaction analysis methodology, computation, theory, and application. For reviews and representative studies, we refer to [1–3]. In this article, we conduct joint G-E interaction analysis and note that joint and marginal analyses are two different analysis paradigms, have different implications, and cannot replace each other, although joint analysis may better fit the biology of complex diseases. For recent developments in joint G-E interaction analysis, we refer to [4, 5].

Our literature review suggests that, in most of the existing joint G-E interaction analyses, attention has been on "simple" data, for example, continuously distributed gene expressions [6] and single nucleotide polymorphisms (SNPs) with moderate to high MAFs (minor allele frequencies). Comparatively, attention to rare features, for example, SNPs with low MAFs (often defined as MAF < 5%) and certain methylation data, has been limited. Rare features are not uncommon in practice. In Figure 1 (Data S1), for the NFBC1996 data to be analyzed in Section 4, we show the genotype distributions of the rare features (post screening). Published studies have established that "ordinary" statistical methods lose power with rare features [7, 8], and that as features get increasingly rare, an unreasonably large sample size will be needed to detect their effects. Here, it is noted that such conclusions have been drawn for main-effect-only methods, most of which conduct marginal analysis. However, it is sensible to expect similar conclusions for interaction analysis. Some early studies inappropriately drop rare features from analysis [9]. With the development of personalized medicine, the significance of rare features for complex human diseases has been firmly recognized [10–12]. Its theoretical basis is that features that strongly predispose to diseases are likely to be deleterious and thus kept at low frequencies by purifying selection [13, 14]. Examining rare features can assist identifying subpopulations that may benefit from targeted treatment.

In main-effect-only analysis, it has been recognized that the most effective and possibly the only feasible strategy for identifying rare features is pooling. That is, as opposed to identifying the individual effects of rare features, the combined effects of "related" rare features, for example those in the same genetic region, are identified. Popular data pooling/collapsing strategies include gene-based bins [15, 16], windows of a fixed length [17], windows of a fixed number of variants [18, 19], and others. A common limitation of these approaches is that they do not take into account the directions of features' effects on a response variable. Generically, methods for analyzing rare features can be classified into four main categories: burden tests using linear statistics [20, 21], variance-components-type tests using quadratic statistics [22, 23], hybrid methods combining burden and variance-components-type tests [15, 24], and other dimension-reduction-based approaches. Examples in the last category include [25, 26], which conduct unsupervised clustering to create denser features. Another example is a penalization method called ConvexConcave Rare variant Selection (CCRS) [27]. However, it has been found that, even after applying the aforementioned aggregation methods, a large portion of aggregated rare features may still be too sparse, and they may still have to be discarded. Here, we note that the aforementioned and many other approaches are limited to marginal analysis in the hypothesis testing framework and are not directly applicable to joint analysis. Recognizing limitations of the existing data aggregation techniques, in a recent study, Yan and Bien [28] develop a more effective strategy for aggregating and selecting rare features, which leverage side information (additional prior information) in the form of a tree. A tree-based parameterization strategy is introduced to translate the feature aggregation problem into a sparse modeling one. Statistical and numerical investigations show that this approach can significantly improve over the existing ones. This flexible, data-adaptive, and tree-based aggregation approach is integrated into a log-contrast regression model in Reference [29]. It is noted that this approach has only been applied to main-effect-only analysis.

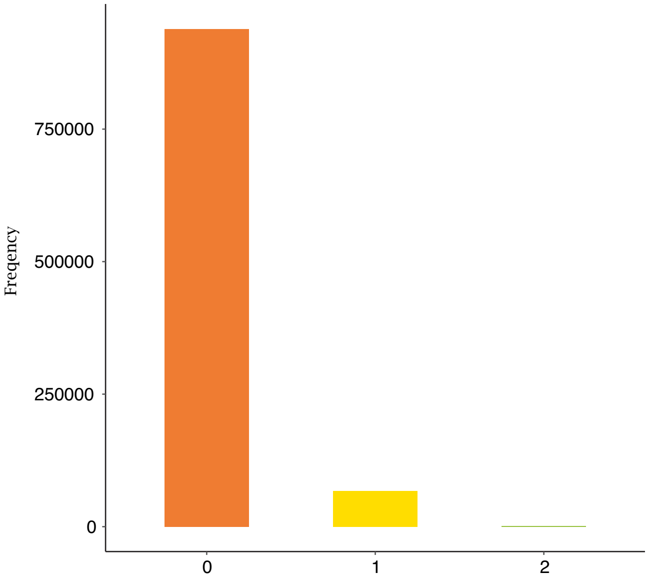

FIGURE 1 A small example of aggregating features within branches

With the high significance of G-E interaction analysis, there has been some effort on detecting interactions between rare features and E variables. For example, Lu and others [30] propose an aggregated statistic, which is derived from the MAF-based logistic principal component analysis (MLPCA). A limitation of this approach is that the adopted unsupervised technique is not ideal to indicate how genetic variants are modified by environment factors to affect disease risk and traits. Zhao and others [31] aggregate genetic and G-E interaction information across markers and construct score tests to identify important G-E interactions. Yang and others [32] develop a family of data-adaptive G-E interaction tests in the framework of adaptive powered score testing. It is noted that these works mostly belong to the marginal analysis paradigm. For joint analysis, Lin and others [33] develop a variance component score test within the induced generalized linear mixed model (GLMM) framework and apply ridge regression to estimate the nuisance main effects. Lim and others [34] adopt a kernel-based method to leverage joint information across rare variants under the GLMM framework. However, in these studies, there has been no attention to the "main effects, interactions" variable selection hierarchy [35, 36].

In this article, we consider joint G-E interaction analysis where a significant number of candidate G features are rare. Although certain individual components of this analysis share some common ground with the existing studies, overall, this study complements and advances published literature in the following aspects. Unlike most of the existing G-E interaction studies, there is special attention to rare features. It differs from most of the existing rare feature studies by conducting joint analysis (which differs significantly from "marginal analysis + hypothesis testing") and by accommodating interactions (and the accompanying unique challenges in particular the "main effects, interactions" variable selection hierarchy). It also advances from many existing pooling studies for rare features by adopting the cutting-edge tree-based aggregation technique [32] and from Reference [32] by conducting joint interaction analysis. In addition, the proposed approach can directly go beyond rare features and be applied to other types of data that also have individual weak effects, and hence data integration is needed.

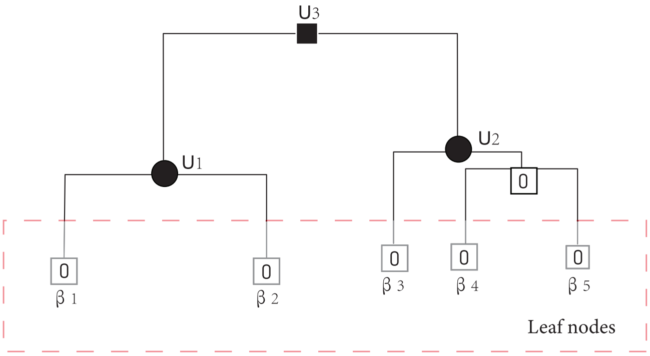

FIGURE 2 Aggregating $\boldsymbol{\beta}$ (left) and $\boldsymbol{\xi}_{\mathbf{k}}(k=1,\ldots,3$; right) in $\mathcal{T}$

2 | METHODS

2.1 | Data and model

Y is denoted as the disease outcome/phenotype. In what follows, we consider a continuously distributed outcome and corresponding linear regression. The proposed approach can be directly applied to other

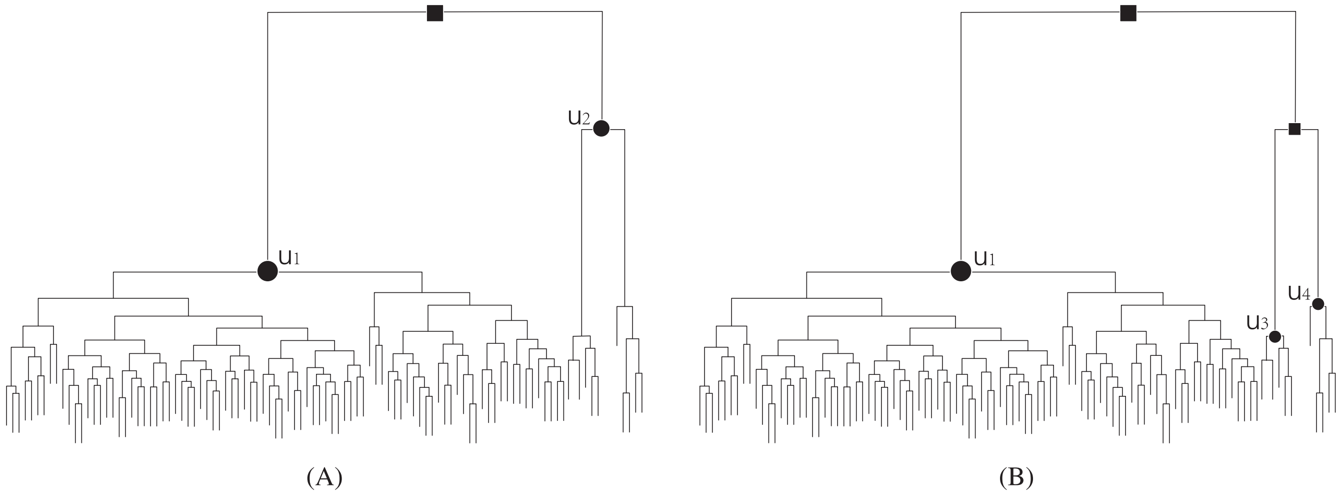

FIGURE 3 The tree structure $\mathcal{T}$ of p leaves with $(p,m,s)=(200{,}20{,}0.4)$. Gray leaves have zero effects, leaves with the other colors have nonzero effects, and leaves with the same color have the same effects

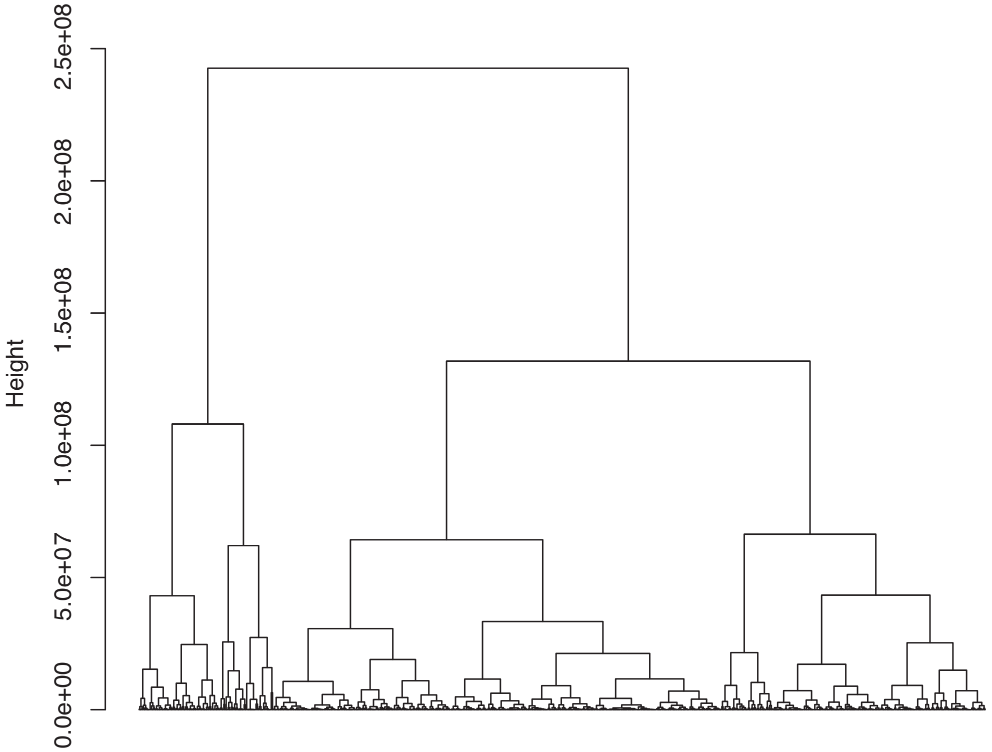

FIGURE 4 NFBC1996 data: distribution of genotypes for the rare single nucleotide polymorphisms (SNPs) after screening

types of outcomes/phenotypes by adopting corresponding regression models and likelihood functions. $Z = (Z_1, \dots, Z_p)'$ is denoted as the p rare features. In our data analysis, we consider SNPs with low MAFs. Further, $X = (X_1, \dots, X_q)'$ is denoted as the q clinical/environmental risk factors. Following strong advocate in the recent literature, we also consider the interactions with demographic and clinical variables. It is also possible to limit interactions to narrowly defined E factors. Consider the joint regression model:

$$Y = \sum_{k=1}^{q} \alpha_k X_k + \sum_{j=1}^{p} \left(\beta_j Z_j + \sum_{k=1}^{q} \eta_{kj} X_k Z_j \right) + \varepsilon, \qquad (1)$$

where α_k's, β_j's, and η_{kj}'s are the regression coefficients for the main E effects, main G effects, and their interactions, respectively. ε is the random error. With proper normalization, the intercept term has been omitted. There are multiple ways of respecting the "main effects, interactions" variable selection hierarchy. Here, we adopt the decomposition strategy [37], where $\eta_{kj} = \beta_j \xi_{kj}$. Then, model (1) can be rewritten as:

$$Y = \sum_{k=1}^{q} \alpha_k X_k + \sum_{j=1}^{p} \left(\beta_j Z_j + \sum_{k=1}^{q} \beta_j \xi_{kj} X_k Z_j \right) + \varepsilon.$$

Denote $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)'$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$, and $\xi_k = (\xi_{k1}, \dots, \xi_{kp})'$. Assume n iid observations $\{(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i), i = 1, \dots, n\}$. Denote $\boldsymbol{y}$ as the n-vector composed of y_i's, $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}^{(k)}$ as the matrices composed of x_i's, z_i's, and $\boldsymbol{w}_i^{(k)} = (x_{ik}z_{i1}, \dots, x_{ik}z_{ip})'$s, respectively. In the matrix form, the least squares objective function is $L(\theta) = \frac{1}{2n} \left\| \boldsymbol{y} - \boldsymbol{X}\alpha - \boldsymbol{Z}\beta - \sum_{k=1}^{q} \boldsymbol{W}^{(k)} (\beta \odot \xi_k) \right\|_2^2$, where $\theta = (\alpha', \beta', \xi_1', \dots, \xi_q')'$, $\| \cdot \|_2$ is the l_2 norm, and $\odot$ is the component-wise product.

We note that the data and model settings have been extensively adopted in the literature, with the difference that Z represents rare features. It is expected that other loss functions, for example, the robust ones, can also be adopted.

2.2 | Estimation

With data aggregation, one of the most critical steps is to determine the regions within which rare features are pooled. Quite a few approaches have been developed for this purpose. Some utilize biological information, for example, functionalities of SNPs. However, this may be not sufficiently effective as the functions of many SNPs, especially those in noncoding regions, are unknown. Another family of approaches utilizes information on features' physical locations, which is usually known. When SNPs are densely measured, those physically close can be in high linkage disequilibrium (LD) and have similar biological functions and/or statistical effects [38]. In our numerical study, for SNP data, we follow [28] and conduct hierarchical clustering analysis of the physical locations of SNPs

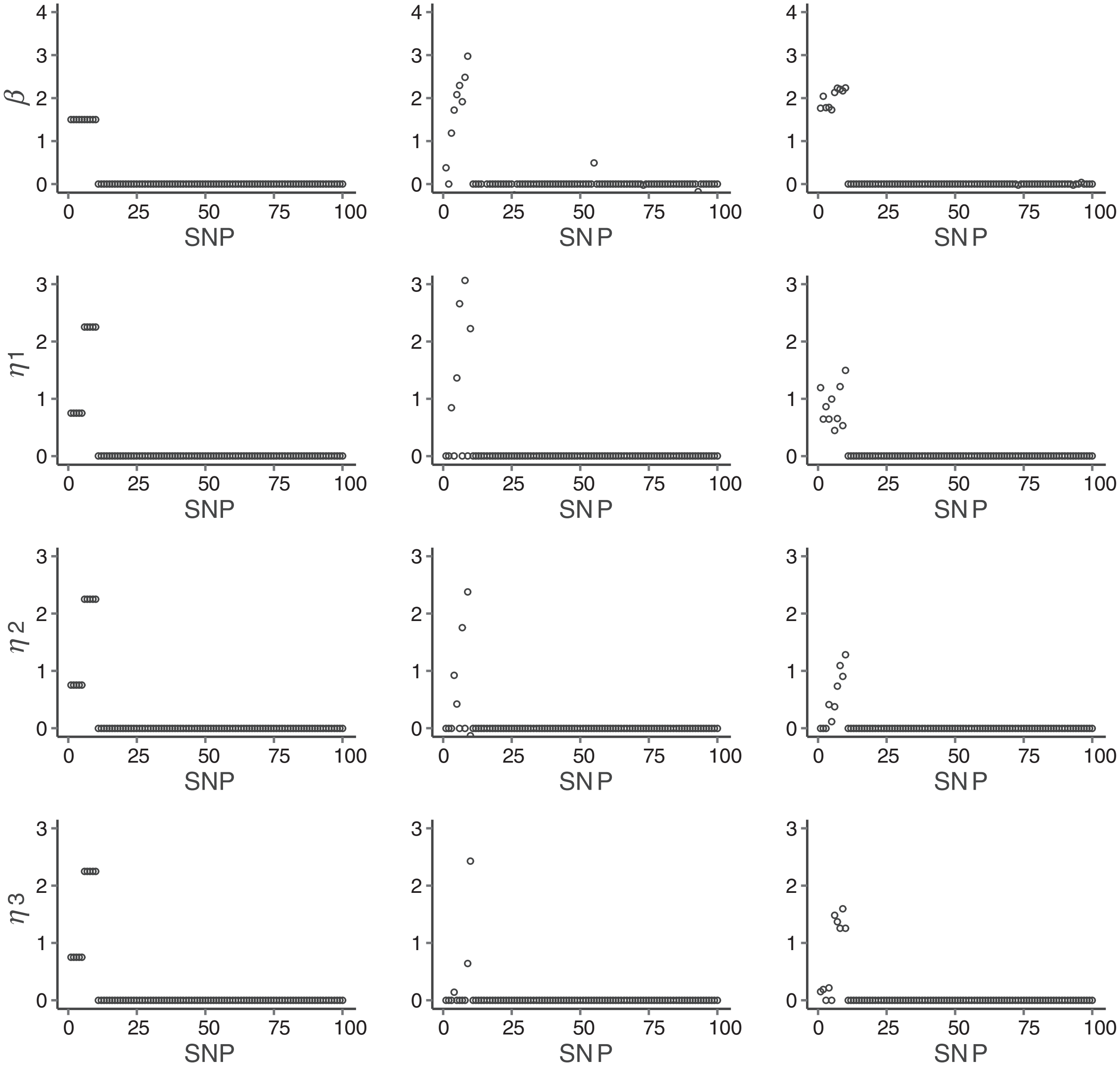

FIGURE 5 Toy example: true (left) and estimated (center: Lasso; right: proposed) main G effects and interactions

to form a tree $\mathcal{T}$, as showcased in Figures 2, 3, and 4. The consideration is that features physically close to each other tend to have related biological functions, which has been established for SNP and some other types of data. We refer to Reference [28] for more discussions on the tree construction. Advancing from Reference [28], we also incorporate interactions and propose densifying β and $\xi_{\mathbf{k}}$ using the same tree structure.

Let u be a node, which is a branching point in a tree. A node is called a leaf node, if it has no additional nodes coming out of it. For example, in Figure 2, those in the red box are leaf nodes. The ancestor(u) and descendant(u) are denoted as the ancestors and descendants of node u in $\mathcal{T}$, respectively. The set of nodes in the path from the root of $\mathcal{T}$ to the jth leaf can be written as ancestor$(j) \cup \{j\}$. Assign a parameter γ_{0u} (γ_{ku}) to each node u in $\mathcal{T}$. Similar to [28], we can conduct a tree-based parameterization to associate β_j and ξ_{kj} with $\mathcal{T}$. Specifically, β_j and $\xi_{kj}(k = 1, \dots, q)$ are decomposed into the sum of all the parameters on the path:

$$\beta_j = \sum_{u \in \text{ancestor}(j) \cup \{j\}} \gamma_{0u}, \quad \xi_{kj} = \sum_{u \in \text{ancestor}(j) \cup \{j\}} \gamma_{ku}. \tag{2}$$

When $\gamma_{0\,\text{descendant}(u)} = 0$ $(\gamma_{k\,\text{descendant}(u)} = 0)$, β_j's (ξ_{kj}'s) associated with the leaves lying beneath node u are equal. For example, with the tree in Figure 2, coefficients of all the nodes beneath nodes u_1 and u_2 are zero. According to (2), β_j's are aggregated into two groups: $\beta_1 = \beta_2 = \gamma_{0u_1} + \gamma_{0u_3}$

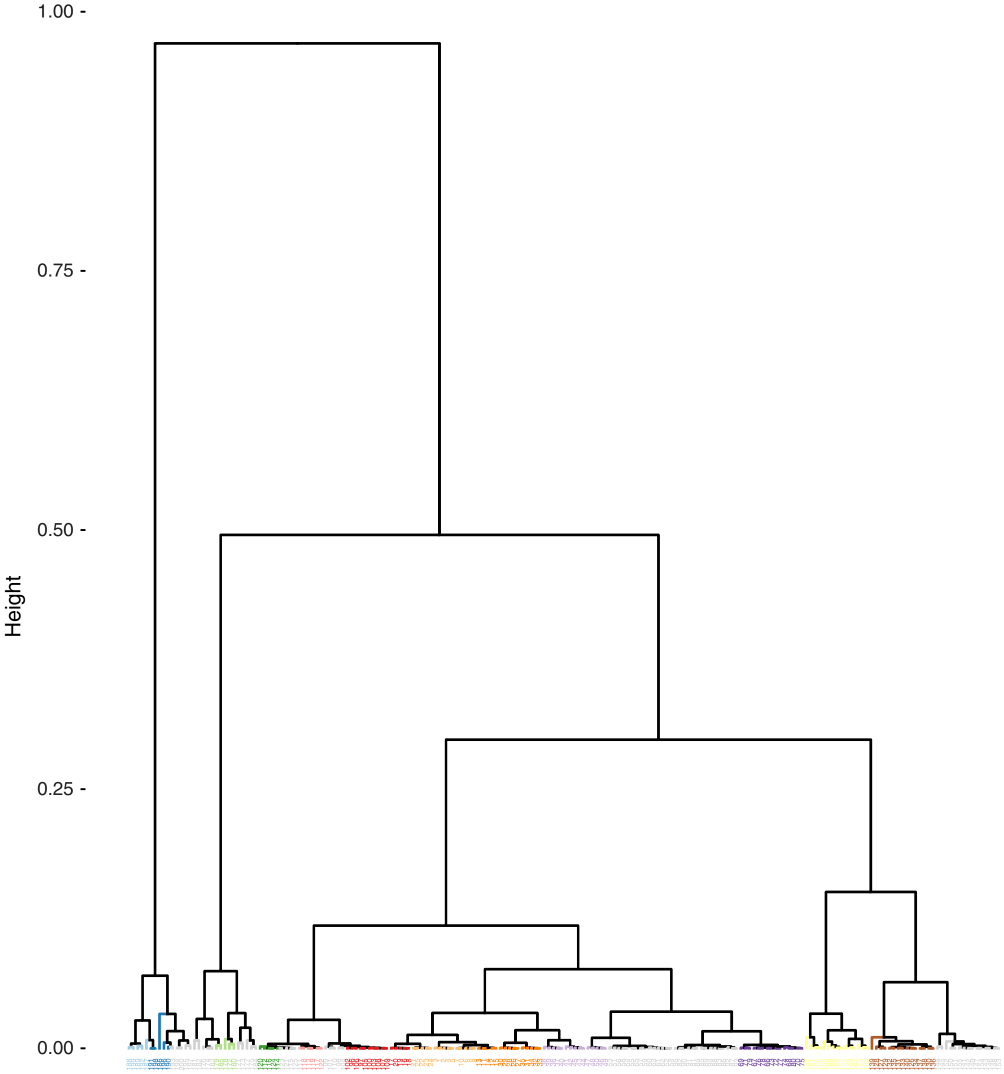

FIGURE 6 Simulation: estimation error with $(n, p, q, s) = (200{,}200{,}5{,}0.4)$

and $\beta_3 = \beta_4 = \beta_5 = \gamma_{0u_2} + \gamma_{0u_3}$. As such, feature aggregation can be achieved by introducing sparsity to $\gamma_0(\gamma_k)$.

For regularized estimation and selection of important interactions and main effects, we propose the penalized objective function:

$$Q_n(\boldsymbol{\theta}, \boldsymbol{\Gamma}) = L(\boldsymbol{\theta}) + a\lambda \sum_{\ell=1}^{|\mathcal{T}|} \left[\omega_{0\ell} \, |\gamma_{0\ell}| + \sum_{k=1}^{q} \omega_{k\ell} \, |\gamma_{k\ell}| \right] + (1-a)\lambda \sum_{j=1}^{p} \left[\widetilde{\omega}_{0j} \, |\beta_j| + \sum_{k=1}^{q} \widetilde{\omega}_{kj} \, |\xi_{kj}| \right],$$

$$s.t. \;\; \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\gamma}_0, \boldsymbol{\xi}_k = \boldsymbol{A}\boldsymbol{\gamma}_k (k = 1, \ldots, q), \qquad (3)$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q)' \in \mathbb{R}^{(q+1)\times|\mathcal{T}|}$, $\lambda \geq 0$ and $a \in [0, 1]$ are tuning parameters, $\omega_{0\ell}, \omega_{k\ell}, \widetilde{\omega}_{0j}, \widetilde{\omega}_{kj}$ are covariate-specific weights (more details below), $|\mathcal{T}|$ denotes the number of nodes in $\mathcal{T}$, $\boldsymbol{A} \in \{0, 1\}^{p\times|\mathcal{T}|}$ is a matrix with elements $\boldsymbol{A}_{jr} := 1_{\{u_r \in \text{ancestor}(j) \cup \{j\}\}} = 1_{\{j \in \text{descendant}(u_r) \cup \{u_r\}\}}$, and $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\gamma}_0$ and $\boldsymbol{\xi}_k = \boldsymbol{A}\boldsymbol{\gamma}_k$ are the compact forms of (2). Similar to other penalized interaction analyses, interactions, and main effects with nonzero coefficients are identified as being important for the response.

Rationale The overall strategy is similar to other penalizations, with the first term quantifies lack-of-fit—it can be revised to accommodate other data types/models. The two penalty terms induce different types of sparsity, which are controlled by λ and balanced by a. The second penalty is relatively "simple" and has been considered in the existing penalized G-E interaction studies. In particular, the Lasso penalty is directly imposed to $\boldsymbol{\beta}$ and $\boldsymbol{\xi_k}$, identifying important main effects and interactions. With the decomposition strategy, the variable selection hierarchy is guaranteed. The weights $\widetilde{\omega}_{0j}, \widetilde{\omega}_{kj}$ lead to weighted (adaptive) penalization. For choosing weights, we refer to Reference [28] and many other publications. The most straightforward choice, which is adopted in our numerical study, is to set the weights equal to 1. The most significant advancement over the existing G-E interaction analysis is

TABLE 1 Simulation Scenario 1

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	0.97(0.07)	0.03(0.01)	1.00(0.00)	0.21(0.13)	4.63(0.54)	0.21(0.08)
L1_dense	0.62(0.08)	0.04(0.01)	0.78(0.07)	0.22(0.04)	6.43(0.77)	3.12(1.08)
L1_ag_h	0.84(0.09)	0.13(0.05)	1.00(0.04)	0.68(0.16)	5.68(1.06)	3.69(0.72)
Lasso	0.74(0.06)	0.02(0.00)	0.95(0.04)	0.19(0.03))	6.43(0.8)	2.64(1.00)
Proposed	0.82(0.13)	0.01(0.01)	1.00(0.00)	0.08(0.07)	4.85(0.58)	0.82(0.47)
$m = 40$						
Oracle Lasso	0.98(0.03)	0.03(0.01)	1.00(0.00)	0.09(0.06)	5.62(0.86)	0.39(0.16)
L1_dense	0.61(0.09)	0.05(0.02)	0.77(0.08)	0.24(0.05)	7.12(0.82)	3.72(1.56)
L1_ag_h	0.75(0.11)	0.15(0.08)	0.99(0.06)	0.84(0.05)	7.03(1.68)	5.66(1.15)
Lasso	0.74(0.07)	0.03(0.01)	0.94(0.06)	0.19(0.03)	7.05(0.76)	2.79(1.21)
Proposed	0.89(0.09)	0.01(0.01)	1.00(0.01)	0.09(0.04)	5.80(0.56)	1.06(0.42)
$m = 60$						
Oracle Lasso	0.94(0.06)	0.02(0.01)	1.00(0.01)	0.04(0.03)	6.42(0.98)	0.59(0.34)
L1_dense	0.61(0.09)	0.04(0.01)	0.75(0.10)	0.22(0.03)	6.74(0.66)	3.47(1.20)
L1_ag_h	0.56(0.18)	0.14(0.12)	1.00(0.00)	0.9(0.08)	8.04(1.82)	6.22(1.09)
Lasso	0.73(0.06)	0.03(0.00)	0.93(0.05)	0.19(0.03)	6.54(0.64)	2.66(1.15)
Proposed	0.88(0.09)	0.02(0.01)	1.00(0.01)	0.08(0.03)	6.48(0.48)	1.13(0.39)
$m = 80$						
Oracle Lasso	0.91(0.05)	0.02(0.01)	1.00(0.02)	0.02(0.02)	6.88(0.74)	0.81(0.44)
L1_dense	0.59(0.08)	0.05(0.01)	0.74(0.09)	0.23(0.04)	7.38(0.70)	4.35(1.29)
L1_ag_h	0.57(0.16)	0.15(0.18)	0.98(0.05)	0.93(0.06)	7.85(2.03)	6.48(1.76)
Lasso	0.71(0.07)	0.03(0.01)	0.93(0.05)	0.19(0.03)	7.24(0.59)	3.53(1.52)
Proposed	0.89(0.07)	0.03(0.01)	1.00(0.02)	0.12(0.04)	7.07(1.84)	1.56(0.62)
$m = 100$						
Oracle Lasso	0.86(0.05)	0.02(0.01)	0.97(0.03)	0.02(0.02)	7.05(0.87)	1.23(0.55)
L1_dense	0.60(0.07)	0.05(0.01)	0.76(0.08)	0.23(0.04)	7.20(0.60)	4.56(1.57)
L1_ag_h	0.20(0.16)	0.15(0.16)	0.95(0.09)	0.85(0.13)	8.62(1.66)	7.71(1.22)
Lasso	0.73(0.07)	0.03(0.01)	0.94(0.04)	0.18(0.03)	7.16(0.64)	3.53(1.43)
Proposed	0.88(0.06)	0.03(0.01)	0.97(0.03)	0.13(0.04)	7.09(0.57)	1.84(0.87)

Note: In each cell, mean (SD) based on 500 replicates.

the first term. Penalty is imposed to $\gamma_{0\ell}$ and $\gamma_{k\ell}$, which, with the constraint defined in (2), induces *fusion* to the coefficients in $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_k$. This fusion is built on the tree structure (as showcased in Figures 2 and 3). In particular, following [28], we leave the root ($\gamma_{k|\mathcal{T}|}$ for $k = 0, \ldots, q$) unpenalized with $\left\{\omega_{k|\mathcal{T}|} = 0\right\}_{\{k=0,1,\ldots,q\}}$. This allows all features to be aggregated into one single group with coefficients fused to a nonzero value. Under $\mathcal{T}$, nearby features, which are expected to have similar effects, are put into the same data aggregating sets. Their effects are fused to be similar, which allows nearby rare features to borrow strength from their neighbors. The aggregated effects can be considerably larger than the individual ones, making them more likely to be identified. It is noted that, with the proposed penalty, $\gamma_{0\,\mathrm{descendant}(u)}$ ($\gamma_{k\,\mathrm{descendant}(u)}$) is encouraged but not forced to be zero. As such, with this fusion/data aggregation technique, features in the same aggregating sets not necessarily have the same coefficients, making this approach more flexible than, for example, those directly adding up rare features.

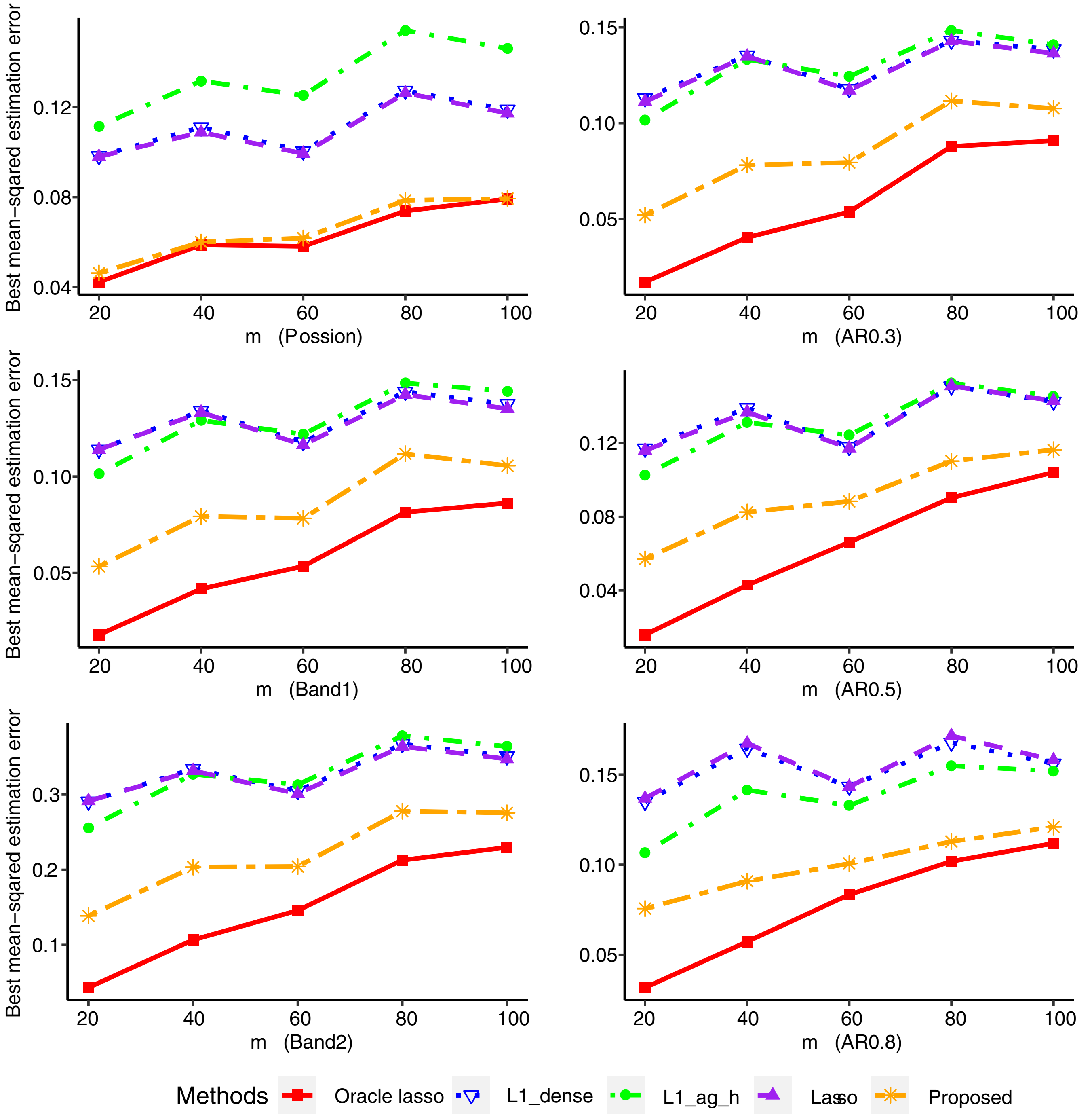

FIGURE 7 Pairwise LD analysis of rare single nucleotide polymorphisms (SNPs) (post screening). Top: LD decay plot; bottom: LD heatmap

A toy example To better appreciate working characteristics of the proposed method, we simulate a small dataset with $n = 100$ and $p = 100$. The tree structures for the main G effects and (components of) G-E interactions are shown in Figure 3. The true aggregating sets are determined based on Figure 3. In particular, the main G effects β_j' are aggregated into two groups, corresponding to nodes u_1 and u_2. All the leaves under u_1 have coefficients zero. β_j' under node u_2 are set to be 1.5. ξ_{kj}'s are aggregated into three groups, corresponding to nodes u_1, u_3, u_4. ξ_{kj}'s under node u_1 are set as 0, and those under nodes u_3 and u_4 are set as 0.75 and 2.25, respectively. Finally, the G-E interactions are calculated as $\eta_{kj} = \beta_j \xi_{kj}$. There are in total 10 main G effects and 30 G-E interactions with nonzero coefficients, and they satisfy the variable selection hierarchy. We graphically show the true regression coefficients in the left column of Figure 5. The SNP measurements are simulated from a Poisson(0.02) distribution and truncated at 2 if

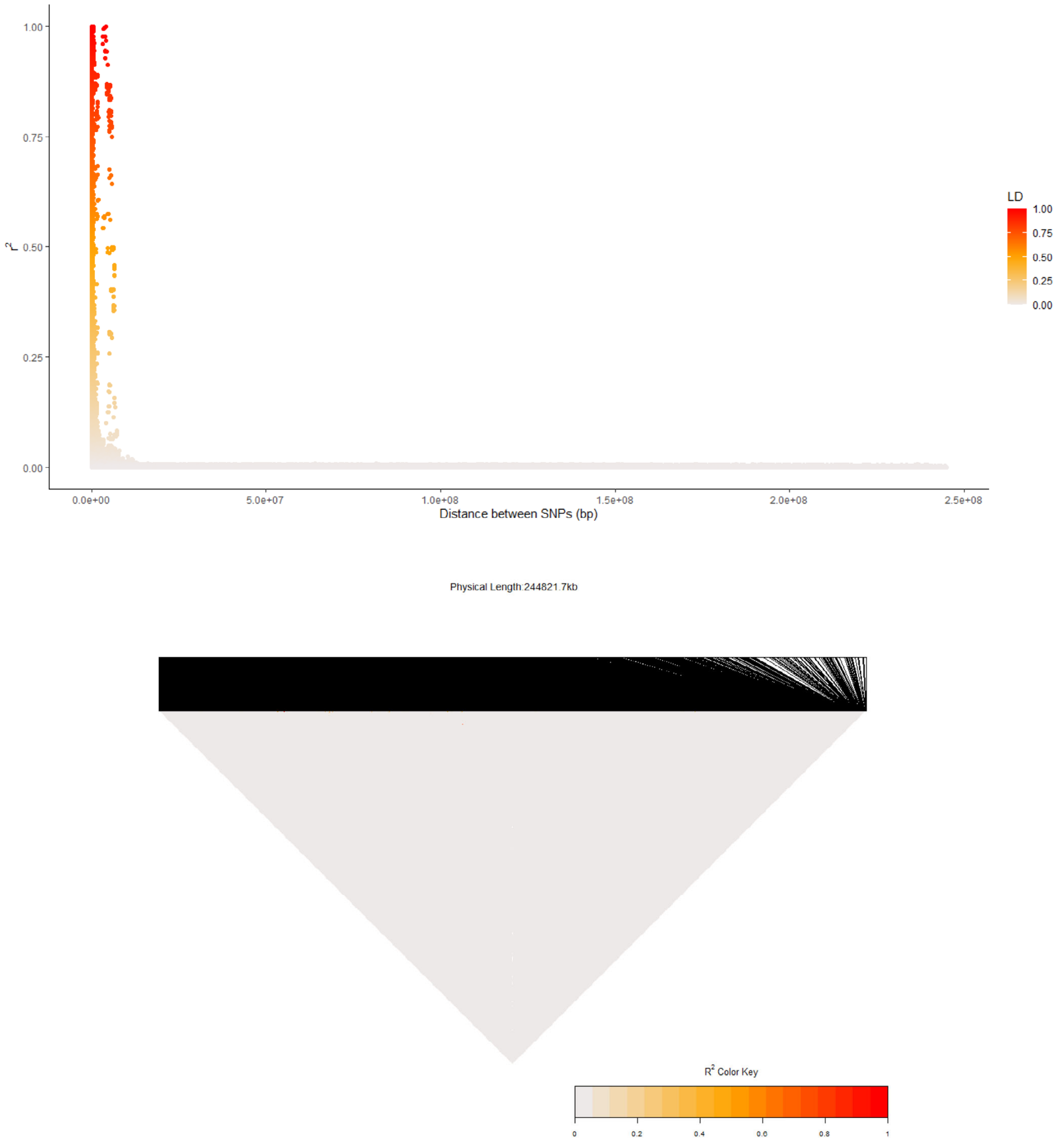

FIGURE 8 Pairwise LD analysis of all single nucleotide polymorphisms (SNPs) (post screening). Top: LD decay plot; bottom: LD heatmap

needed. We then simulate three E variables as having a Bernoulli distribution with probability of success 0.7. The response variable is generated from a linear regression model with a standard normally distributed random error. Beyond the proposed approach, we also consider the Lasso approach as a benchmark, which shares the same penalization framework as the proposed approach but does not conduct data aggregation. The estimation results using the proposed and Lasso approaches are graphically presented in Figure 5. By borrowing strength and effectively aggregating data, the proposed approach is observed to have significantly better identification and estimation

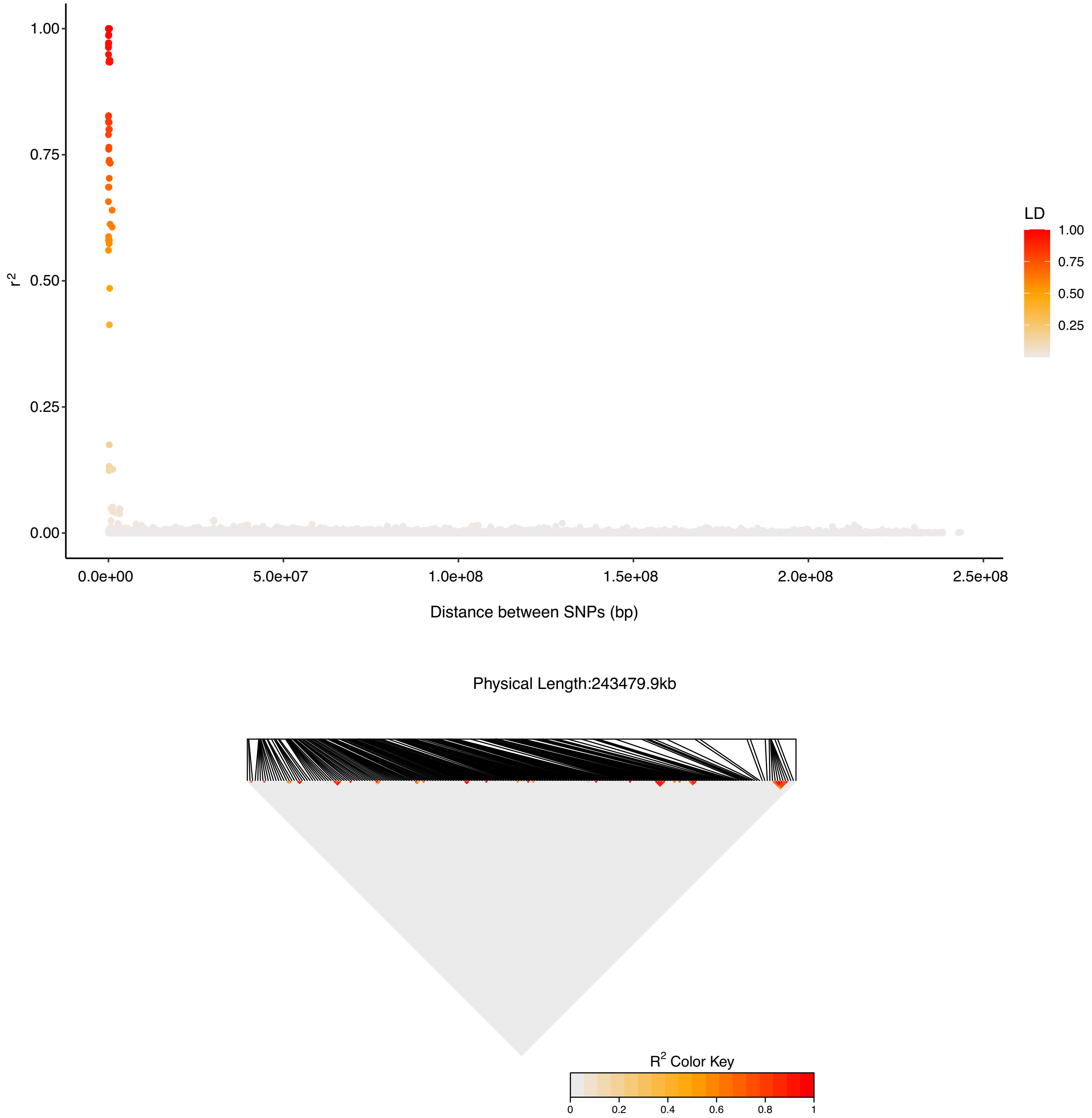

FIGURE 9 NFBC1996 data analysis: tree $\mathcal{T}$ of single nucleotide polymorphisms' (SNPs') physical positions (post screening)

accuracy. More definitive results based on larger-scale simulations are presented below in Section 3.

2.3 | Computation

With fixed tuning parameters, the optimization of (3) can be conducted using an iterative coordinate descent (CD) algorithm, which optimizes the objective function with respect to one of the three (sets of) vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and ξ_k's at a time and iteratively cycles through all of the parameters until convergence. Let $\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}$, and $\xi_k^{(t)}$ denote the estimates of $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and ξ_k at iteration t, respectively. The proposed algorithm proceeds as follows:

Step 1 Initialize $t = 0$, $\boldsymbol{\beta}^{(t)} = \mathbf{0}$, $\xi_k^{(t)} = \mathbf{0}$, and $\boldsymbol{\alpha}^{(t)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$.

Step 2 Update $t = t + 1$. With ξ_k and $\boldsymbol{\alpha}$ fixed at $\xi_k^{(t-1)}$ and $\boldsymbol{\alpha}^{(t-1)}$, optimize (3) with respect to $\boldsymbol{\beta}$. Let $\widetilde{\boldsymbol{y}}^{(t)} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}^{(t-1)}$ and $\widetilde{\boldsymbol{Z}}^{(t)} = \boldsymbol{Z} + \sum_{k=1}^{q} \boldsymbol{W}^{(k)} \odot \left(\mathbf{1}_{n\times 1}\left(\xi_k^{(t-1)}\right)'\right)$ with

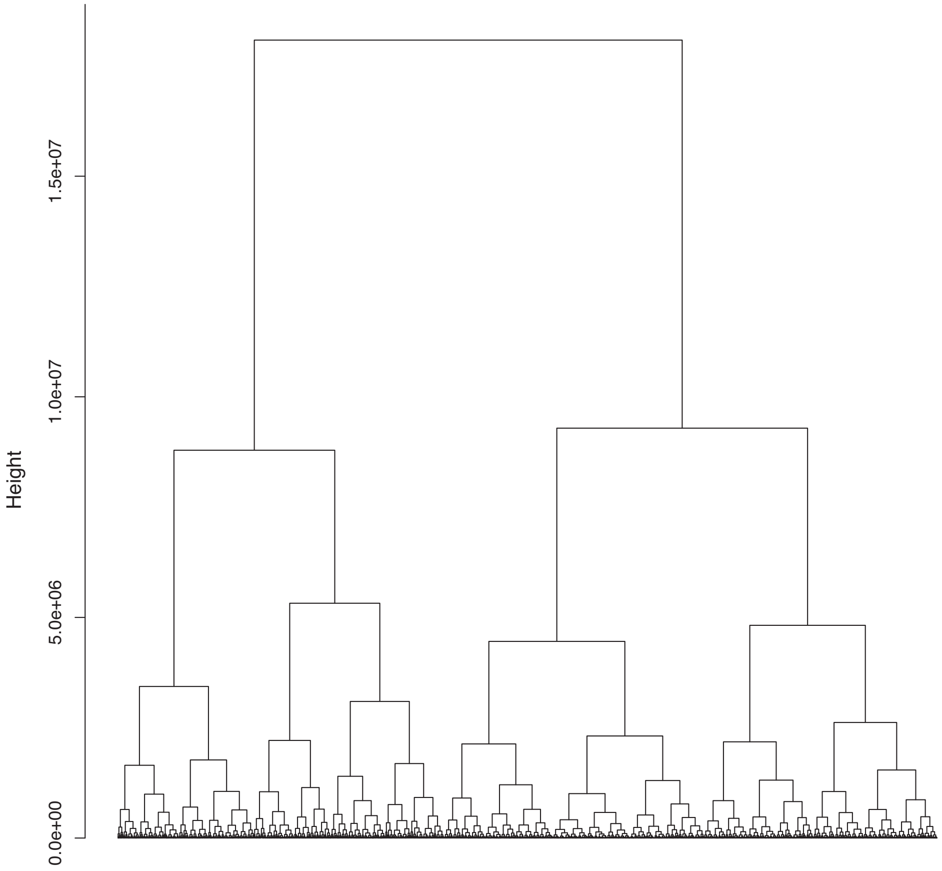

FIGURE 10 NFBC1996 data analysis: tree $\mathcal{T}$ of single nucleotide polymorphisms' (SNPs') physical positions (post screening), under the alternative screening approach

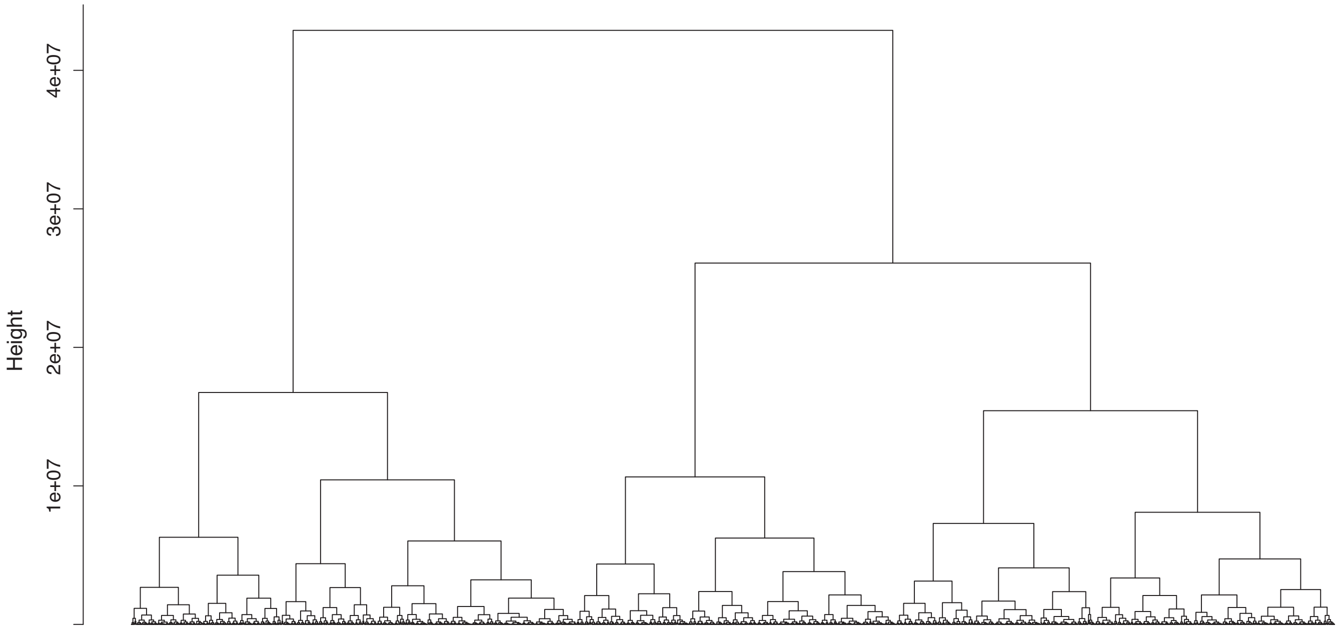

FIGURE 11 NFBC1996 data analysis: tree $\mathcal{T}$ of single nucleotide polymorphisms' (SNPs') physical positions (post screening), with glucose as the response variable

TABLE 2 Analysis of the NFBC1996 data using the proposed approach: identified main effects and interactions

SNP	Main effect	Gender	CRP	Glucose	TC	HDL
rs12208417	0.015	0.045	0.014	—	0.005	0.046
rs6488338	−0.017	−0.037	0.074	—	−0.014	—
rs2453244	−0.015	−0.044	−0.007	0.022	−0.067	—
rs11042023	−0.008	−0.057	—	—	—	0.027
rs2511841	−0.012	−0.081	0.014	0	−0.028	0.076
rs4965685	0.016	0.037	—	0.005	—	0.029
rs489487	0.01	0.05	—	−0.011	—	—
rs7306908	−0.016	−0.069	—	−0.015	—	—
rs10949732	0.017	0.039	−0.095	—	—	0.013
rs4575188	0.011	0.036	0.016	—	0.017	—
rs4720078	−0.016	−0.048	—	−0.012	−0.003	0.013
rs7039156	−0.018	−0.034	—	—	−0.022	−0.011
rs1676996	0.017	0.053	—	0.042	0.026	0.01
rs1386894	0.009	0.057	—	—	−0.014	—
rs1180819	−0.015	−0.051	0.005	—	−0.008	−0.053
rs10512052	0.014	0.042	−0.041	—	0.011	—
rs1237044	−0.016	−0.03	0.007	—	—	−0.109
rs10773484	−0.013	−0.065	—	−0.01	—	—
rs1284412	0.012	0.029	0.055	0.051	—	—
rs2571249	0.016	0.052	0.05	—	−0.027	−0.001
rs4149570	0.009	0.031	—	—	—	−0.005
rs1372555	−0.013	−0.049	−0.011	0.001	—	−0.035
rs3782631	0.013	0.06	0.01	—	—	—
rs2121671	0.015	0.048	−0.004	−0.011	−0.016	—
rs1870591	−0.011	−0.084	—	—	—	0.008
rs10508924	0.015	0.043	0.018	−0.018	—	−0.011
rs2834889	−0.011	−0.025	0.019	—	0.019	−0.003
rs7186722	−0.022	−0.069	—	0.006	0.046	0.094
rs3092379	0.013	0.038	—	—	0.008	0.006
rs2150855	−0.017	−0.051	−0.005	—	0.017	0.046
rs516783	0.013	0.058	—	0.009	—	−0.009
rs7506974	0.012	0.022	—	−0.049	—	0.009
rs3898586	0.014	0.061	0.04	—	0.003	—
rs344386	−0.024	−0.045	−0.02	—	—	0.038

$\mathbf{1}_{n\times1} = (1, \ldots, 1)_{n\times1}$. Then

$$\beta^{(t)} = \arg\min_{\beta} \frac{1}{2n}\left\|\widetilde{\boldsymbol{y}}^{(t)} - \widetilde{\boldsymbol{Z}}^{(t)}\boldsymbol{\beta}\right\|_2^2 + \lambda\left\{a\sum_{\ell=1}^{|\mathcal{T}|} w_{0\ell}\,|\gamma_{0\ell}| + (1-a)\sum_{j=1}^{p}\widetilde{w}_{0j}\,|\beta_j|\right\}$$
$$s.t.\ \ \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\gamma}_0. \tag{4}$$

To simplify notation, we consider the representative setting with $w_{0|\mathcal{T}|} = 0$ and $\left\{w_{0\ell} = 1, \widetilde{w}_{0j} = 1\right\}_{\{l\neq|\mathcal{T}|\,j\in\{1,\ldots,p\}\}}$. Problem (4) can be efficiently solved with the consensus ADMM algorithm [28]. Taking the form of a decomposition-coordination procedure, it combines the benefit of dual decomposition and augmented Lagrangian methods for constrained optimization.

Step 3 With $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ fixed at $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\alpha}^{(t-1)}$, optimize (3) with respect to $\xi = \left(\xi_1, \ldots, \xi_q\right)$. Let $\breve{\boldsymbol{y}}^{(t)} = \boldsymbol{y} -$

TABLE 3 Simulation Scenario 2

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	1.00(0.02)	0.05(0.03)	1.00(0.00)	0.65(0.13)	3.01(1.08)	0.31(0.17)
L1_dense	0.14(0.07)	0.00(0.00)	0.42(0.19)	0.02(0.01)	7.03(1.00)	6.64(3.60)
L1_ag_h	0.89(0.09)	0.08(0.04)	0.98(0.08)	0.61(0.16)	6.18(0.94)	3.84(1.08)
Lasso	0.14(0.07)	0.00(0.00)	0.42(0.19)	0.02(0.01)	6.94(0.99)	6.46(3.36)
Proposed	0.67(0.20)	0.01(0.01)	0.94(0.09)	0.18(0.08)	5.03(1.00)	3.11(2.38)
$m = 40$						
Oracle Lasso	0.95(0.06)	0.05(0.02)	0.99(0.04)	0.57(0.11)	4.58(1.11)	0.96(0.70)
L1_dense	0.12(0.07)	0.00(0.00)	0.36(0.19)	0.02(0.02)	7.47(0.95)	7.39(3.07)
L1_ag_h	0.85(0.13)	0.24(0.08)	1.00(0.00)	0.76(0.15)	7.19(0.94)	5.61(1.93)
Lasso	0.12(0.07)	0.00(0.00)	0.36(0.19)	0.02(0.02)	7.4(0.89)	7.37(2.96)
Proposed	0.56(0.22)	0.01(0.01)	0.90(0.09)	0.18(0.09)	5.99(1.01)	4.41(1.99)
$m = 60$						
Oracle Lasso	0.88(0.08)	0.03(0.02)	0.98(0.04)	0.48(0.11)	5.33(1.38)	1.93(1.18)
L1_dense	0.14(0.07)	0.00(0.00)	0.42(0.19)	0.02(0.02)	7.19(0.77)	6.86(2.41)
L1_ag_h	0.60(0.21)	0.26(0.15)	0.94(0.12)	0.83(0.11)	7.76(1.32)	7.13(4.07)
Lasso	0.14(0.07)	0.00(0.00)	0.42(0.19)	0.02(0.02)	7.22(0.76)	6.75(2.38)
Proposed	0.50(0.19)	0.02(0.01)	0.87(0.11)	0.22(0.08)	6.21(0.68)	4.92(2.09)
$m = 80$						
Oracle Lasso	0.75(0.10)	0.04(0.03)	0.96(0.06)	0.41(0.11)	6.25(1.03)	2.97(1.27)
L1_dense	0.09(0.06)	0.00(0.00)	0.30(0.18)	0.02(0.01)	7.79(0.52)	7.98(2.60)
L1_ag_h	0.42(0.18)	0.30(0.17)	0.86(0.12)	0.82(0.15)	8.39(0.91)	7.98(2.48)
Lasso	0.09(0.06)	0.00(0.00)	0.30(0.18)	0.02(0.01)	7.76(0.50)	7.85(2.53)
Proposed	0.34(0.17)	0.01(0.01)	0.76(0.15)	0.25(0.10)	7.03(0.66)	6.63(2.21)
$m = 100$						
Oracle Lasso	0.69(0.11)	0.04(0.03)	0.93(0.06)	0.40(0.12)	6.31(1.13)	3.06(1.38)
L1_dense	0.11(0.08)	0.00(0.00)	0.33(0.19)	0.02(0.02)	7.61(0.69)	7.86(3.35)
L1_ag_h	0.41(0.13)	0.26(0.13)	0.84(0.13)	0.76(0.18)	8.42(1.04)	8.34(2.78)
Lasso	0.11(0.08)	0.00(0.00)	0.33(0.19)	0.02(0.02)	7.60(0.80)	7.91(3.40)
Proposed	0.38(0.17)	0.02(0.01)	0.78(0.13)	0.25(0.09)	6.73(0.70)	5.75(2.59)

Note: In each cell, mean (SD) based on 500 replicates.

$\boldsymbol{X\alpha}^{(t-1)} - \boldsymbol{Z\beta}^{(t)}$ and $\left(\widetilde{\boldsymbol{W}}^{(k)}\right)^{(t)} = \boldsymbol{W}^{(k)} \odot \left(\mathbf{1}_{n\times 1}\left(\boldsymbol{\beta}^{(t)}\right)'\right)$. Then

$$\boldsymbol{\xi}^{(t)} = \arg\min \frac{1}{2n}\left\|\breve{\boldsymbol{y}}^{(t)} - \sum_{k=1}^{q}\left(\widetilde{\boldsymbol{W}}^{(k)}\right)^{(t)}\xi_k\right\|_2^2$$

$$+ \lambda\left(a\sum_{\ell=1}^{|\mathcal{T}|}\sum_{k=1}^{q} w_{k\ell}\,|\gamma_{k\ell}| + (1-a)\sum_{j=1}^{p}\sum_{k=1}^{q}\widetilde{w}_{kj}\,|\xi_{kj}|\right)$$

$$s.t.\ \xi_k = \boldsymbol{A\gamma}_k,\ k = 1, \ldots, q.$$

The algorithm is similar to that in **Step 2**.

Step 4 Compute $\boldsymbol{\alpha}^{(t)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\left(\boldsymbol{y} - \boldsymbol{Z\beta}^{(t)} - \sum_{k=1}^{q} \mathbf{W}^{(k)}\left(\boldsymbol{\beta}^{(t)} \odot \xi_k^{(t)}\right)\right)$.

Step 5 Repeat **Steps 2–4** until convergence. In our numerical study, convergence is concluded if $\frac{\left|Q_n(\boldsymbol{\theta}^{(t)},\boldsymbol{\Gamma}^{(t)}) - Q_n(\boldsymbol{\theta}^{(t-1)},\boldsymbol{\Gamma}^{(t-1)})\right|}{\left|Q_n(\boldsymbol{\theta}^{(t-1)},\boldsymbol{\Gamma}^{(t-1)})\right|} < 10^{-4}$.

TABLE 4 Simulation Scenario 3

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	1.00(0.01)	0.06(0.05)	1.00(0.00)	0.70(0.14)	3.11(0.98)	0.36(0.24)
L1_dense	0.16(0.07)	0.00(0.00)	0.43(0.17)	0.02(0.01)	7.03(1.02)	7.92(3.93)
L1_ag_h	0.91(0.09)	0.09(0.04)	1.00(0.04)	0.67(0.17)	5.73(1.27)	3.51(1.59)
Lasso	0.16(0.07)	0.00(0.00)	0.43(0.17)	0.02(0.01)	6.95(0.99)	7.67(3.91)
Proposed	0.68(0.21)	0.01(0.01)	0.94(0.09)	0.18(0.07)	5.24(1.09)	3.49(2.02)
$m = 40$						
Oracle Lasso	0.95(0.06)	0.05(0.05)	0.99(0.03)	0.59(0.12)	4.65(1.38)	1.29(0.77)
L1_dense	0.14(0.06)	0.00(0.00)	0.39(0.15)	0.02(0.01)	7.69(0.92)	8.36(3.42)
L1_ag_h	0.79(0.20)	0.23(0.10)	0.99(0.06)	0.81(0.10)	7.03(1.23)	6.18(2.52)
Lasso	0.14(0.06)	0.00(0.00)	0.39(0.15)	0.02(0.01)	7.66(1.15)	8.55(5.08)
Proposed	0.55(0.21)	0.02(0.01)	0.87(0.14)	0.20(0.08)	6.25(0.85)	5.00(1.74)
$m = 60$						
Oracle Lasso	0.84(0.10)	0.03(0.02)	0.96(0.06)	0.50(0.10)	5.18(0.89)	1.89(1.00)
L1_dense	0.14(0.06)	0.00(0.00)	0.38(0.16)	0.02(0.01)	7.14(0.64)	6.73(2.76)
L1_ag_h	0.57(0.17)	0.24(0.13)	0.96(0.11)	0.83(0.14)	7.16(0.85)	6.54(3.20)
Lasso	0.14(0.06)	0.00(0.00)	0.38(0.16)	0.02(0.01)	7.13(0.68)	6.70(2.78)
Proposed	0.48(0.18)	0.02(0.01)	0.85(0.12)	0.23(0.09)	5.98(0.83)	4.82(2.10)
$m = 80$						
Oracle Lasso	0.77(0.09)	0.03(0.02)	0.96(0.05)	0.42(0.10)	6.65(1.16)	3.58(2.22)
L1_dense	0.11(0.06)	0.00(0.00)	0.30(0.16)	0.02(0.01)	7.93(0.68)	9.09(3.69)
L1_ag_h	0.48(0.16)	0.38(0.18)	0.87(0.13)	0.81(0.17)	8.90(1.11)	10.33(5.29)
Lasso	0.11(0.06)	0.00(0.00)	0.30(0.16)	0.02(0.01)	7.99(0.89)	8.87(3.45)
Proposed	0.37(0.16)	0.02(0.01)	0.78(0.12)	0.25(0.09)	6.97(0.71)	6.99(3.32)
$m = 100$						
Oracle Lasso	0.68(0.10)	0.04(0.02)	0.91(0.08)	0.41(0.10)	6.30(0.98)	3.18(1.38)
L1_dense	0.11(0.08)	0.00(0.00)	0.32(0.20)	0.02(0.02)	7.72(0.66)	8.29(3.28)
L1_ag_h	0.44(0.13)	0.30(0.14)	0.89(0.12)	0.81(0.17)	8.35(1.21)	9.28(5.37)
Lasso	0.11(0.08)	0.00(0.00)	0.32(0.20)	0.02(0.02)	7.66(0.63)	8.00(3.06)
Proposed	0.36(0.18)	0.02(0.01)	0.76(0.15)	0.25(0.10)	6.79(0.64)	6.36(2.63)

Note: In each cell, mean (SD) based on 500 replicates.

The proposed objective function is bounded from below. In each iteration step, its value decreases. As such, convergence is guaranteed. It is satisfactorily achieved with a moderate number of iterations in all of our numerical studies. The tuning parameters (λ, a) are chosen using a modified BIC criterion with the degree of freedom defined as the effective number of parameters [39]. With simple updates, the proposed computational algorithm is affordable. For one simulation replicate (details described below), computation can be accomplished within 3 min on a regular desktop. To facilitate numerical analysis within and beyond this study, we have developed R code and made it publicly available at http://github.com/shuanggema/.

3 | SIMULATION

We consider a total of six scenarios to examine the dependence of performance on distributional properties,

TABLE 5 Simulation Scenario 4

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	0.99(0.02)	0.05(0.04)	1.00(0.00)	0.69(0.14)	3.41(1.90)	1.02(1.19)
L1_dense	0.17(0.07)	0.01(0.00)	0.41(0.10)	0.02(0.01)	8.57(2.02)	9.95(4.83)
L1_ag_h	0.91(0.13)	0.10(0.05)	0.98(0.12)	0.65(0.16)	6.99(1.89)	6.14(3.6)
Lasso	0.17(0.07)	0.01(0.00)	0.41(0.10)	0.02(0.01)	8.46(1.86)	9.63(4.38)
Proposed	0.64(0.16)	0.01(0.01)	0.90(0.12)	0.19(0.06)	6.86(2.82)	5.83(4.37)
$m = 40$						
Oracle Lasso	0.90(0.09)	0.06(0.04)	0.99(0.03)	0.65(0.10)	6.12(1.96)	3.57(3.17)
L1_dense	0.16(0.06)	0.01(0.00)	0.38(0.13)	0.02(0.01)	8.92(1.75)	12.11(6.68)
L1_ag_h	0.88(0.12)	0.29(0.08)	0.99(0.06)	0.83(0.11)	7.43(1.90)	11.15(6.73)
Lasso	0.16(0.06)	0.01(0.00)	0.38(0.13)	0.02(0.01)	9.07(1.89)	12.71(7.06)
Proposed	0.55(0.16)	0.02(0.01)	0.89(0.12)	0.22(0.08)	7.04(1.52)	9.00(7.59)
$m = 60$						
Oracle Lasso	0.82(0.09)	0.04(0.03)	0.97(0.06)	0.54(0.09)	6.00(1.46)	4.43(3.64)
L1_dense	0.15(0.06)	0.01(0.00)	0.39(0.10)	0.03(0.01)	7.66(0.80)	9.24(3.39)
L1_ag_h	0.72(0.18)	0.35(0.10)	0.98(0.06)	0.84(0.12)	7.48(1.00)	9.96(4.20)
Lasso	0.15(0.06)	0.01(0.00)	0.39(0.1)	0.03(0.01)	7.86(1.23)	9.59(4.18)
Proposed	0.53(0.12)	0.02(0.01)	0.90(0.09)	0.22(0.09)	6.61(0.92)	6.69(3.21)
$m = 80$						
Oracle Lasso	0.69(0.11)	0.05(0.03)	0.94(0.08)	0.46(0.11)	7.32(1.46)	6.43(2.90)
L1_dense	0.14(0.07)	0.01(0.00)	0.35(0.14)	0.03(0.01)	8.80(1.36)	12.12(5.52)
L1_ag_h	0.60(0.16)	0.50(0.16)	0.93(0.1)	0.90(0.10)	8.78(1.23)	13.18(6.26)
Lasso	0.14(0.07)	0.01(0.00)	0.35(0.14)	0.03(0.01)	8.67(1.11)	11.97(5.70)
Proposed	0.46(0.14)	0.03(0.01)	0.84(0.10)	0.24(0.09)	7.33(0.96)	9.55(4.85)
$m = 100$						
Oracle Lasso	0.63(0.11)	0.04(0.02)	0.89(0.08)	0.43(0.08)	7.50(1.53)	5.79(3.04)
L1_dense	0.13(0.05)	0.01(0.00)	0.36(0.12)	0.03(0.02)	8.68(1.58)	10.78(5.38)
L1_ag_h	0.56(0.14)	0.38(0.15)	0.94(0.10)	0.88(0.10)	9.01(1.72)	14.04(10.54)
Lasso	0.13(0.05)	0.01(0.00)	0.36(0.12)	0.03(0.02)	8.69(1.39)	10.26(3.83)
Proposed	0.41(0.13)	0.03(0.01)	0.80(0.12)	0.26(0.09)	7.54(1.22)	8.53(3.28)

Note: In each cell, mean (SD) based on 500 replicates.

especially correlation. To mimic data analyzed in the next section, we simulate G variables with properties similar to SNPs. Under Scenario 1, the G variables are independently generated from a Poisson(0.02) distribution and truncated at 2 if needed. The five E variables are generated from a Bernoulli distribution with probability of success 0.7. Under Scenarios 2–6, we first generate p continuous variables from multivariate normal distributions, and then dichotomize the continuous variables at the 0.98 and 0.995 percentiles to generate the three-level G measurements. The multivariate normal distributions have marginal means 0 and variances 1. Two correlation structures with different parameters are considered and referred to as Band1, Band2, AR(0.3), AR(0.5), and AR(0.8). Here, Band1 and Band2, the two banded correlation structures, have correlation coefficients of variables j and k as $0.3^{|j-k|}I(j - k| < 2)$ and $0.3^{|j-k|}I(j - k| = 2) + 0.5^{|j-k|}I(j - k| < 2)$, respectively. The three auto-regressive structures correspond to weak, moderate, and strong correlations, respectively. We note that such correlation

TABLE 6 Simulation Scenario 5

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	1.00(0.01)	0.05(0.03)	1.00(0.00)	0.68(0.14)	2.76(0.82)	0.33(0.32)
L1_dense	0.11(0.09)	0.00(0.00)	0.34(0.22)	0.02(0.01)	6.67(0.97)	5.54(2.47)
L1_ag_h	0.91(0.07)	0.07(0.03)	1.00(0.00)	0.59(0.18)	6.28(0.82)	3.73(0.84)
Lasso	0.11(0.09)	0.00(0.00)	0.34(0.22)	0.02(0.01)	6.64(1.03)	5.57(2.45)
Proposed	0.65(0.24)	0.01(0.01)	0.95(0.09)	0.18(0.11)	4.73(0.82)	2.49(1.13)
$m = 40$						
Oracle Lasso	0.93(0.08)	0.05(0.03)	0.98(0.05)	0.55(0.13)	4.26(0.81)	1.07(0.76)
L1_dense	0.10(0.08)	0.00(0.00)	0.29(0.19)	0.02(0.02)	7.3(0.7)	6.89(2.40)
L1_ag_h	0.79(0.13)	0.22(0.07)	0.99(0.06)	0.77(0.12)	6.74(1.00)	4.95(1.61)
Lasso	0.10(0.08)	0.00(0.00)	0.29(0.19)	0.02(0.02)	7.21(0.67)	6.79(2.32)
Proposed	0.54(0.24)	0.02(0.01)	0.88(0.14)	0.25(0.12)	5.68(0.84)	4.04(1.82)
$m = 60$						
Oracle Lasso	0.89(0.08)	0.04(0.02)	0.99(0.04)	0.49(0.13)	5.17(1.26)	1.95(1.39)
L1_dense	0.14(0.07)	0.00(0.00)	0.41(0.17)	0.02(0.01)	7.15(0.73)	7.03(3.15)
L1_ag_h	0.57(0.18)	0.24(0.12)	0.92(0.13)	0.78(0.12)	7.45(1.15)	6.36(2.38)
Lasso	0.14(0.07)	0.00(0.00)	0.41(0.17)	0.02(0.01)	7.11(0.68)	6.84(3.00)
Proposed	0.57(0.16)	0.02(0.01)	0.87(0.09)	0.29(0.11)	5.99(0.64)	4.77(2.14)
$m = 80$						
Oracle Lasso	0.75(0.09)	0.03(0.01)	0.96(0.07)	0.4(0.09)	6.57(1.37)	2.85(1.36)
L1_dense	0.09(0.08)	0.00(0.00)	0.28(0.18)	0.02(0.02)	7.79(0.69)	7.64(2.67)
L1_ag_h	0.42(0.19)	0.29(0.18)	0.9(0.13)	0.85(0.12)	8.62(1.09)	8.43(2.50)
Lasso	0.09(0.08)	0.00(0.00)	0.28(0.18)	0.02(0.02)	7.74(0.70)	7.41(2.61)
Proposed	0.39(0.20)	0.02(0.01)	0.83(0.14)	0.27(0.11)	6.87(0.73)	5.87(2.35)
$m = 100$						
Oracle Lasso	0.69(0.10)	0.04(0.03)	0.92(0.06)	0.41(0.11)	6.20(1.20)	3.20(1.56)
L1_dense	0.10(0.07)	0.00(0.00)	0.30(0.18)	0.02(0.02)	7.56(0.65)	7.48(2.09)
L1_ag_h	0.40(0.15)	0.24(0.13)	0.88(0.12)	0.82(0.13)	8.41(1.25)	8.25(2.52)
Lasso	0.10(0.07)	0.00(0.00)	0.30(0.18)	0.02(0.02)	7.52(0.65)	7.43(2.20)
Proposed	0.38(0.17)	0.02(0.01)	0.81(0.11)	0.28(0.10)	6.84(0.69)	6.20(2.21)

Note: In each cell, mean (SD) based on 500 replicates.

structures have been considered in quite a few 3 studies. For the E variables, we first generate five continuous variables from a multivariate normal distribution with marginal means 0, marginal variances 1, and an AR(0.3) correlation structure. Then, two variables are dichotomized at 0 to create two binary variables, leading to three continuous and two binary E variables. Under all scenarios, the G variables have low (MAF 1%–5%) and very low (MAF < 1%) frequencies. In practical data analysis, more common variants are expected. Here, we focus on rare variants whose effects are more difficult to quantify. The proposed approach is expected to have better performance for variants that are less rare.

The nonzero main effects and interactions are generated as follows. For the SNPs, based on their adjacency (correlation) information, the true tree structure $\mathcal{T}$ of the p leaves is shown in Figure 6. These leaves form m aggregating sets (clusters) with varying sizes, which are indexed by $\boldsymbol{B}^{*}$. This construction is similar to that in [28]. To generate the main G and G-E interaction effects, we first

TABLE 7 Simulation Scenario 6

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	1.00(0.02)	0.05(0.04)	1.00(0.00)	0.67(0.15)	2.96(1.00)	0.36(0.22)
L1_dense	0.16(0.07)	0.01(0.00)	0.48(0.15)	0.02(0.01)	7.00(1.34)	7.57(3.81)
L1_ag_h	0.91(0.09)	0.08(0.04)	1.00(0.00)	0.63(0.18)	5.31(0.77)	3.91(0.91)
Lasso	0.16(0.07)	0.01(0.00)	0.48(0.15)	0.02(0.01)	7.02(1.38)	7.43(3.82)
Proposed	0.65(0.16)	0.01(0.01)	0.95(0.06)	0.13(0.07)	4.95(1.01)	2.95(1.90)
$m = 40$						
Oracle Lasso	0.96(0.05)	0.05(0.03)	0.99(0.04)	0.60(0.13)	4.44(1.22)	1.29(0.82)
L1_dense	0.15(0.07)	0.00(0.00)	0.40(0.16)	0.02(0.01)	7.52(0.68)	8.04(2.91)
L1_ag_h	0.83(0.14)	0.25(0.11)	0.99(0.04)	0.78(0.14)	7.02(1.16)	6.28(2.71)
Lasso	0.15(0.07)	0.00(0.00)	0.40(0.16)	0.02(0.01)	7.66(1.04)	8.15(3.24)
Proposed	0.49(0.14)	0.01(0.01)	0.87(0.11)	0.16(0.08)	6.24(0.90)	4.85(2.04)
$m = 60$						
Oracle Lasso	0.86(0.12)	0.04(0.03)	0.98(0.06)	0.52(0.13)	5.61(1.53)	2.37(1.75)
L1_dense	0.13(0.07)	0.00(0.00)	0.39(0.18)	0.02(0.01)	7.16(0.65)	7.82(3.50)
L1_ag_h	0.60(0.21)	0.27(0.13)	0.95(0.12)	0.84(0.13)	7.52(0.97)	8.22(3.82)
Lasso	0.13(0.07)	0.00(0.00)	0.39(0.18)	0.02(0.01)	7.18(0.67)	7.79(3.45)
Proposed	0.46(0.17)	0.01(0.01)	0.85(0.11)	0.18(0.07)	6.09(0.89)	5.57(3.10)
$m = 80$						
Oracle Lasso	0.77(0.11)	0.04(0.02)	0.96(0.07)	0.42(0.10)	7.01(1.82)	4.39(3.22)
L1_dense	0.11(0.07)	0.01(0.00)	0.37(0.17)	0.02(0.02)	8.08(0.70)	9.37(3.90)
L1_ag_h	0.48(0.15)	0.38(0.17)	0.89(0.14)	0.84(0.17)	8.95(1.29)	9.81(3.44)
Lasso	0.11(0.07)	0.01(0.00)	0.37(0.17)	0.02(0.02)	8.05(0.76)	9.32(3.96)
Proposed	0.40(0.17)	0.02(0.01)	0.80(0.14)	0.19(0.08)	7.08(0.70)	7.12(3.37)
$m = 100$						
Oracle Lasso	0.71(0.10)	0.04(0.02)	0.91(0.06)	0.44(0.12)	6.66(1.29)	3.85(2.50)
L1_dense	0.10(0.07)	0.00(0.00)	0.28(0.19)	0.02(0.02)	7.67(0.57)	7.87(3.17)
L1_ag_h	0.44(0.14)	0.27(0.16)	0.89(0.12)	0.81(0.16)	8.25(1.06)	8.14(2.77)
Lasso	0.10(0.07)	0.00(0.00)	0.28(0.19)	0.02(0.02)	7.58(0.59)	7.73(3.14)
Proposed	0.31(0.17)	0.01(0.01)	0.74(0.15)	0.18(0.09)	6.93(0.69)	6.08(2.59)

Note: In each cell, mean (SD) based on 500 replicates.

generate a matrix $\boldsymbol{A}_{B^*} \in \mathbb{R}^{p\times m}$ with binary components $\boldsymbol{A}_{B^*jl} := 1_{\{j\in l\,cluster\}}$. Then, the coefficient vectors are generated via these aggregating sets as: $\boldsymbol{\beta}^* = \boldsymbol{A}_{\boldsymbol{B}^*}\widetilde{\boldsymbol{\beta}}^*$, $\boldsymbol{\xi}_k^* = \boldsymbol{A}_{\boldsymbol{B}^*}\widetilde{\boldsymbol{\xi}}^*_{(k)}$, where $\widetilde{\boldsymbol{\beta}}^*, \widetilde{\boldsymbol{\xi}}^*_k \in \mathbb{R}^m$ have $m \times s$ elements zeroed out, and the remaining elements are independently drawn from a Uniform(0.8, 1.5) distribution. Here, s controls the true level of sparsity. For the main E effects, their nonzero coefficients α_k^*'s are generated from Uniform (0.8,1.2). The response $\mathbf{y} \in \mathbb{R}^n$ values are simulated from (1) with independent Gaussian errors and variances $\sigma^2 = \sum_{i=1}^n \left(\boldsymbol{x}_{i'}\boldsymbol{\alpha}^* + \boldsymbol{z}_{i'}\boldsymbol{\beta}^* + \sum_{k=1}^q x_{ik}\boldsymbol{z}_{i'}\left(\boldsymbol{\beta}^* \odot \boldsymbol{\xi}_k^*\right)\right)^2/(5n)$. The above data generation satisfies the "main effects, interactions" hierarchical structure and aggregative effects of the nearby G features.

We set $n = 200$, $p = 200$, $q = 5$, $s = 0.4$. It is noted that the combined number of unknown parameters is much larger than the sample size. We consider a sequence of m values up to $p/2$. The proposed approach is applied based

TABLE 8 NFBC1996 data analysis: numbers of main G effects and interactions identified by different approaches (diagonal) and their overlaps (off-diagonal)

		Lasso	L1_dense	L1_a_h	Proposed
Main G effects	Lasso	31	22	20	27
	L1_dense	–	29	20	24
	L1_a_h	–	–	23	21
	Proposed	–	–	–	34
Interactions	Lasso	84	50	34	68
	L1_dense	–	65	33	54
	L1_a_h	–	–	40	36
	Proposed	–	–	–	107

on the tree $\mathcal{T}$. To gauge its performance, we further consider the following alternatives. The first is Oracle Lasso, under which the true aggregation structure $\boldsymbol{XA}_{B^*}$ is known, and Lasso (which is the proposed approach with $a = 0$) is applied for regularized selection and estimation. The second is L1_dense, which applies Lasso after first discarding all features with MAF < 1%. It represents approaches that focus on dense features. The third is L1_ag_h, which applies Lasso to features aggregated in the same clusters after the tree is cut at a certain height. This approach conducts feature aggregation based on $\mathcal{T}$, however, in a relatively "brutal" manner. It represents approaches that first conduct clustering, then group features in an unsupervised way, and finally conduct modeling and estimation based on the postaggregation features. Lastly, we also consider the Lasso approach as for the toy example. For each setting, we simulate 500 replicates.

We evaluate identification performance using the true-positive rate (TPR) and false-positive rate (FPR) for main G effects (M:TPR and M:FPR) and interactions (I:TPR and I:FPR) separately. Here, it is noted that the proposed approach conducts penalized variable selection as opposed to hypothesis testing. As such, it does not have a direct false discovery rate control. Nevertheless, TPR and FPR are still highly informative performance measures. We further evaluate estimation performance using the root sum of squared errors (RSSE) defined as $\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2$. In addition, we calculate the best mean-squared estimation error as a function of m, that is, $\min_\Lambda \left\|\hat{\boldsymbol{\theta}}(\Lambda) - \boldsymbol{\theta}^*\right\|_2^2/(p + q + pq)$, where Λ represents a method's tuning parameter(s). For evaluating prediction performance, for each simulation replicate, we generate an independent testing dataset with size 200 and compute the prediction mean squared error (PMSE).

Summary results are presented in Tables 1 and 3–7 (Data S1). Across all of the simulation scenarios, the proposed approach is observed to perform similarly to the Oracle Lasso and has superior performance compared to the other alternatives. Specifically, it can more accurately identify both the true main effects and interactions while having a small number of false positives. For example, in Table 1, under Scenario 1 and $m = 40$, the proposed approach has (M:TPR, M:FPR, I:TPR, I:FPR) = (0.89, 0.01, 1, 0.09), compared to (0.61, 0.05, 0.77, 0.24) for L1_dense, (0.75, 0.15, 0.99, 0.84) for L1_ag_h, and (0.74, 0.03, 0.94, 0.19) for Lasso. This result and those alike can establish the effectiveness of accommodating rare features (when compared to L1_dense), data aggregation (when compared to Lasso), and more effective data aggregation (when compared to L1_ag_h). We also observe the superiority of the proposed approach in estimation. For example, in Table 6 (Data S1), under Scenario 5 and $m = 60$, the proposed approach has RSSE = 5.99, compared to 7.15(L1_dense), 7.45(L1_ag_h), and 7.11(Lasso). This superiority is further shown in Figure 7 (Data S1). In particular, under Scenario 1 with $\boldsymbol{Z}$ simulated from a Poisson(0.02) distribution, the proposed approach performs nearly as well as the oracle. As m increases, borrowing strength from neighbors decreases, and so estimation performance deteriorates. L1_dense performs very similarly to Lasso. Last but not least, the proposed approach also has satisfactory prediction performance. For example, in Table 1, under Scenario 1, the PMSEs are 3.12 (L1_dense), 3.69 (L1_ag_h), 2.64 (Lasso), and 0.82 (proposed) (Figure 8).

4 | DATA ANALYSIS

To demonstrate the practical applicability of the proposed approach, we analyze the individual-level data from the NFBC (Northern Finland Birth Cohorts) study [40]. This study was conducted to very broadly examine risk factors involved in preterm birth and intrauterine growth retardation, as well as the consequences of these early adverse

TABLE 9 Analysis of the NFBC1996 data using the proposed approach under an alternative marginal screening: identified main effects and interactions

SNP	Main effect	Gender	CRP	Glucose	TC	HDL
rs12222221	−0.253	0.001	−0.001	0.031	−0.002	—
rs4585672	−0.270	−0.001	0.003	0.001	0.006	—
rs6743144	0.230	0.008	0.001	0.028	−0.010	—
rs12548107	0.285	0.009	0.002	−0.021	0.037	—
rs1470829	−0.329	−0.008	—	0.016	0.007	—
rs6127943	0.216	0.007	−0.002	0.002	0.006	—
rs4735825	0.291	0.009	−0.004	−0.007	−0.028	−0.001
rs1882681	−0.462	−0.007	0.001	0.061	−0.030	−0.001
rs4870024	0.300	0.008	—	−0.008	0.039	—
rs937557	−0.236	0.001	—	−0.019	−0.018	—
rs177195	−0.245	0.011	—	−0.008	0.006	—
rs17552964	0.251	0.005	—	—	0.016	0.001
rs3771327	−0.292	0.012	0.001	−0.027	0.024	—
rs2833383	−0.421	0.007	−0.004	−0.005	0.016	−0.001
rs4077636	−0.285	0.002	0.003	0.035	−0.016	—
rs4512398	−0.244	0.003	0.004	−0.004	−0.006	—
rs1025404	−0.288	0.008	−0.005	−0.002	−0.024	—
rs2961725	−0.244	−0.003	0.001	−0.011	−0.011	−0.001
rs1934127	−0.272	0.002	0.002	−0.030	0.028	−0.001
rs1293770	0.223	0.009	—	0.023	0.008	—
rs1407593	0.239	0.010	0.003	−0.007	−0.010	—
rs10906021	0.225	0.004	−0.003	0.013	0.031	0.001
rs12475063	−0.231	0.004	0.004	0.032	0.032	0.001
rs2306970	−0.226	0.008	0.001	−0.015	−0.047	−0.001
rs2868975	0.213	0.008	0.001	0.010	−0.014	—
rs6737978	−0.305	0.005	−0.001	0.017	−0.043	—
rs881204	−0.337	−0.006	−0.001	−0.009	−0.005	—
rs1886434	−0.244	0.009	−0.001	0.014	0.034	0.001
rs7962035	0.282	0.005	0.002	0.054	0.024	0.001
rs11854565	0.354	0.005	0.001	−0.003	0.033	−0.001
rs7209713	−0.307	0.002	−0.001	−0.027	0.010	—
rs2016327	−0.259	0.003	−0.002	−0.003	0.009	0.001
rs4422244	0.225	0.007	−0.002	−0.034	0.002	0.002
rs11812486	−0.300	0.012	0.002	−0.006	0.075	—
rs1345981	−0.293	0.001	0.001	0.023	−0.009	—
rs6122682	−0.233	—	−0.002	−0.003	—	0.001
rs1920083	−0.234	—	0.001	0.008	0.012	—
rs987648	−0.292	—	0.002	−0.016	0.016	−0.001
rs7989689	−0.264	—	0.001	−0.008	0.005	0.001
rs3887251	−0.218	—	—	0.017	−0.009	—
rs1202657	−0.066	—	—	—	—	—

TABLE 10 NFBC1996 data analysis under an alternative marginal screening: numbers of main G effects and interactions identified by different approaches (diagonal) and their overlaps (off-diagonal)

		Lasso	L1_dense	L1_a_h	Proposed
Main G effects	Lasso	47	17	17	23
	L1_dense	–	24	10	10
	L1_a_h	–	–	24	12
	Proposed	–	–	–	41
Interactions	Lasso	102	25	14	28
	L1_dense	–	34	15	16
	L1_a_h	–	–	26	12
	Proposed	–	–	–	163

outcomes on subsequent morbidity. The data collected from Northern Finland forms a unique resource, allowing to study the emergence of diseases, which can be caused by genetic, biological, social, and behavioral risk factors. The NFBC1966 dataset contains 10 traits and 364,590 SNPs for 5402 individuals whose expected year of birth is 1966. In our data analysis, the response variable of interest is BMI (body mass index), which is an important phenotype and critical biomarker for many illness conditions. For the G factors, we consider SNPs. And for the E factors, we consider gender, C-reactive protein (CRP), glucose, total cholesterol (TC), and high-density cholesterol (HDL). We note that these factors are not environmental in the narrow sense. Rather, they are clinical biomarkers. In the recent literature [41], the interactions between clinical/demographic variables and G factors analyzed under the G-E interaction analysis framework and have attracted strong interest. Such analysis can reveal the interplay between clinical/demographic variables and G factors on disease outcomes and other biomarkers.

Data processing is first conducted, following similar procedures as in published studies [42]. In particular, we exclude individuals that have discrepancies between reported sex and sex determined using the X chromosome. Further, individuals with missingness in the response and E variables or with genotype missing call-rates > 5% are excluded. A SNP is removed from analysis if its MAF < 1% or missing call-rate > 1%, or it fails the Hardy–Weinberg equilibrium test. The SNP data quality control is conducted using PLINK [43]. These processing procedures lead to data on 5123 individuals and 319,147 SNPs. In principle, the proposed approach can be directly applied. Considering the limited sample size, we further conduct a prescreening to improve estimation. In particular, we split the data into two parts with sizes 2:3. Marginal regression analysis is conducted in smaller part, under which one SNP is analyzed at a time using regression. The 5000 SNPs with the smallest marginal p-values are selected for downstream analysis. In Figure 9 (for the rare SNPs) and 8 (for all of the SNPs) in the Data S1, we examine the LD structures for the SNPs that have passed screening. It is observed that a relatively small number of rare SNPs have high LD values, which may limit the power of information borrowing (Figures 10 and 11).

The larger part of the data is analyzed using the proposed and alternative approaches. It is recognized that this may lead to a smaller sample size and loss of power, compared to the analysis of the whole data. However, separating the screening and analysis data can lead to more objective analysis and comparison and has been adopted in many published studies. The tree $\mathcal{T}$ is constructed using hierarchical clustering and the physical locations of SNPs and shown in Figure 4 (Data S1). For all approaches, tuning parameters are selected using the modified BIC criterion [39]. The proposed approach identifies 34 main G effects and 107 interactions. The detailed estimation results are provided in Table 2. The summary comparison results are presented in Table 8 (Data S1). It is observed that the proposed approach identifies more effects. This is sensible as, with fusion, it can pull some SNPs that otherwise may not be identified. It is also noted that, with the complexity of BMI, more main G effects and interactions (than identified by the proposed and alternative methods) may be involved. In general, penalization approaches, including the proposed, can only identify the relatively strong effects. Alternative techniques will be needed for the identification of weaker effects (Tables 9 to 14).

The alternative methods miss the rare SNP rs6488338, which has MAF = 0.046 and belongs to gene CD163. Published studies have suggested that gene CD163 is associated with pregravid obesity. The adipose tissue expression of gene CD163 is elevated in obesity and type 2 diabetes, and this gene is a novel immune marker for metabolic inflammation [44]. Many other findings are also biologically meaningful. For example, LINGO2 is a protein coding

TABLE 11 Analysis of the NFBC1996 data with Glucose as the response variable using the proposed approach: identified main effects and interactions

SNP	Main effect	Gender	CRP	TC	HDL	LDL
rs10504197	0.023	0.041	−0.017	—	−0.020	—
rs1551547	0.022	0.061	—	—	−0.011	—
rs7460495	−0.031	−0.033	−0.032	—	—	0.010
rs2290526	−0.030	−0.064	—	—	—	—
rs10108007	0.017	0.026	—	—	−0.005	—
rs11998308	0.023	0.049	—	—	—	—
rs9692725	−0.027	−0.052	—	—	—	0.033
rs10091115	0.032	0.042	0.023	0.031	—	—
rs979843	−0.026	−0.047	—	—	−0.043	−0.020
rs7836768	0.025	0.024	0.050	—	—	—
rs9643401	−0.044	−0.058	−0.017	—	0.009	−0.008
rs1896135	0.019	0.045	—	—	—	—
rs2380540	−0.032	−0.033	0.056	—	—	0.028
rs2380607	0.018	0.020	—	—	0.005	—
rs12678469	0.015	0.018	—	0.003	—	—
rs1383978	−0.021	−0.032	—	0.007	—	—
rs1031177	0.018	0.030	0.005	—	—	−0.007
rs959974	−0.023	−0.033	—	—	—	—
rs12334848	0.028	0.027	−0.016	−0.053	—	−0.036
rs2941456	0.013	0.018	—	—	—	—
rs998731	−0.022	−0.016	—	—	—	—
rs6473219	−0.018	−0.018	—	—	—	—
rs272610	0.014	0.013	—	—	—	—
rs4961056	−0.024	−0.048	—	0.009	0.004	—
rs7818882	0.039	0.017	−0.019	−0.007	0.044	−0.083
rs1507883	−0.024	−0.021	0.024	—	—	—
rs1382101	−0.016	−0.016	—	—	—	—
rs1487796	0.038	0.083	0.076	—	—	—
rs13262606	0.021	0.038	−0.009	—	—	—
rs551496	0.015	0.013	—	—	—	—
rs1374633	−0.031	−0.072	0.028	—	-	0.039
rs2890805	0.031	0.053	0.022	−0.032	0.062	—
rs3104966	−0.016	−0.017	—	—	—	—
rs4263730	−0.024	−0.006	−0.038	—	0.032	—
rs2587000	−0.036	−0.045	0.059	0.037	0.017	—
rs2513399	−0.036	−0.036	−0.118	−0.033	—	—
rs2513402	0.038	—	−0.071	0.036	0.010	0.023
rs998980	0.002	—	—	—	—	—

TABLE 12 NFBC1996 data analysis with Glucose as the response variable: numbers of main G effects and interactions identified by different approaches (diagonal) and their overlaps (off—diagonal)

		Lasso	L1_dense	L1_a_h	Proposed
Main G effects	Lasso	42	12	18	31
	L1_dense	–	31	21	11
	L1_a_h	–	–	33	16
	Proposed	–	–	–	38
Interactions	Lasso	55	11	15	51
	L1_dense	–	31	17	15
	L1_a_h	–	–	30	18
	Proposed	–	–	–	86

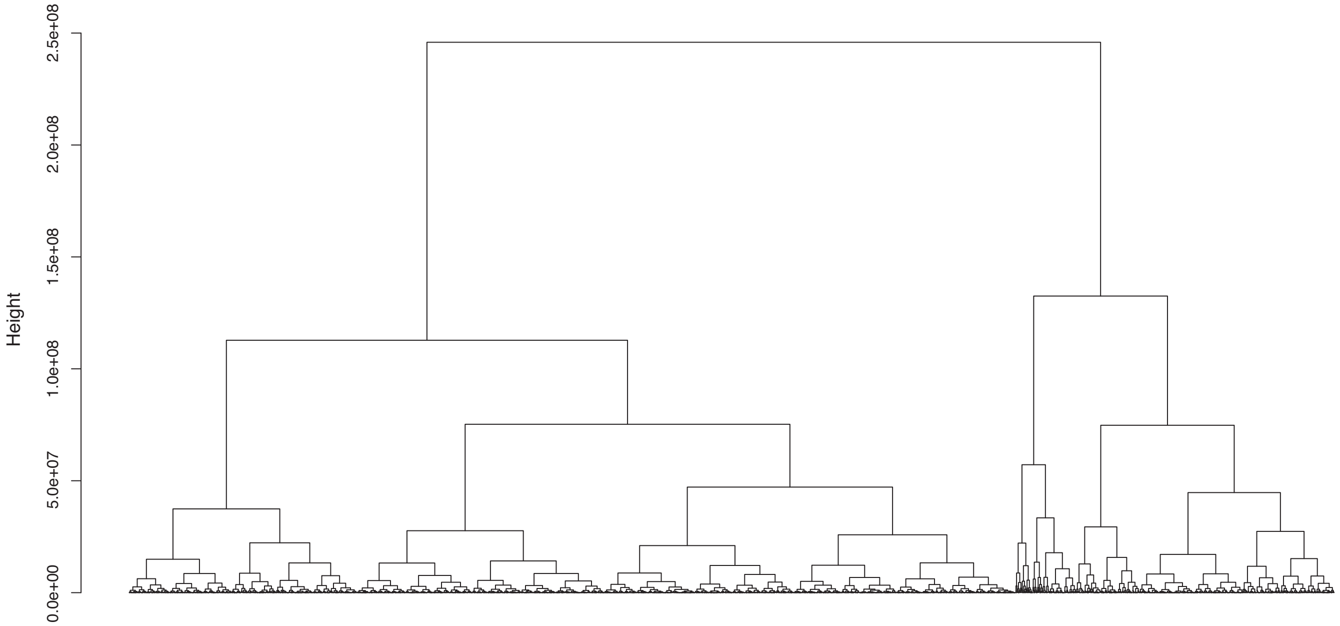

FIGURE 12 NFBC1996 data analysis: tree $\mathcal{T}$ of single nucleotide polymorphisms' (SNPs)' physical positions (post screening), with insulin as the response variable

gene that is an important part of the cellular membrane. It has been linked to obesity in GWAS [45]. Additionally, the variants of LINGO2 have been linked to essential tremor in Parkinson's disease [46]. Thus, its linkage to obesity via interactions with DA signaling seems possible [47]. RBFOX1 is an important RNA-binding protein mediating the incorporation of microexons into many transcripts associated with neurological patterning and tissue development. Its association with obesity has been suggested [48]. Gene *ANO2* has been suggested as playing a role in the pathophysiology of childhood obesity [49]. Studies [50] have revealed that ANO2 is a Ca^{2+}−activated chloride channel in vagal afferents of nodose neurons and a major determinant of CCK-induced satiety, body weight control, and energy expenditure, making it a potential therapeutic target in obesity. TNFRSF1 genotypes have been identified as significantly associated with sTNFR1 plasma levels in obese women [51]. It has been suggested that TNFRSF1A polymorphism can have functional significance in obesity. In addition, genes *TAF4B*, *PCSK5*, *LDLRAD4*, and *TENM4* have also been associated with obesity by GWAS [52].

With real data, it is hard to objectively evaluate the accuracy of identification. To provide further insight and "indirect" support, we apply a resampling-based approach and evaluate prediction performance and stability. Specifically, the dataset is randomly divided into a training and testing set, with sizes 9:1. The model/parameters are estimated using only the training set and then used to

TABLE 13 Analysis of the NFBC1996 data with insulin as the response variable using the proposed approach: identified main effects and interactions

SNP	Main effect	Gender	CRP	TC	HDL	LDL
rs2220326	0.021	0.021	—	—	—	—
rs9598811	−0.023	−0.018	—	——	—	—
rs359379	0.049	0.092	—	−0.020	0.002	−0.055
rs359361	−0.035	−0.032	—	—	—	0.056
rs396985	0.060	0.120	—	—	−0.111	—
rs967958	−0.041	−0.051	0.027	—	—	−0.023
rs1928955	0.020	0.017	0.000	—	−0.013	—
rs1036995	−0.037	0.003	0.072	0.018	−0.026	0.015
rs7324254	−0.019	−0.008	—	—	—	—
rs11148750	−0.023	−0.019	—	—	−0.003	0.003
rs1384607	−0.071	−0.182	−0.009	0.064	0.064	—
rs9541273	−0.029	−0.041	—	—	—	—
rs9571979	0.035	0.084	—	—	—	—
rs7993187	0.055	0.096	—	—	−0.044	—
rs1341525	−0.040	−0.006	—	0.118	—	0.006
rs9572442	−0.025	−0.031	—	—	—	—
rs9542369	0.040	0.087	—	—	—	—
rs1114564	0.040	0.025	—	−0.040	—	—
rs9599903	0.031	0.063	—	—	—	—
rs936457	−0.046	−0.016	—	−0.067	—	−0.065
rs7333339	−0.047	−0.052	0.093	—	—	—
rs287553	−0.028	−0.066	—	—	—	—
rs1324061	0.041	0.045	—	—	−0.034	0.033
rs4597197	0.049	0.065	0.048	—	—	—
rs1505149	0.056	0.121	—	—	0.041	—
rs9574389	−0.021	−0.040	—	—	—	-
rs2274554	0.043	0.054	−0.016	—	−0.018	—
rs1998452	−0.023	−0.014	—	—	—	—
rs7336627	0.035	0.058	—	—	—	——
rs1335852	−0.046	−0.122	—	−0.075	—	—
rs9602002	−0.019	−0.028	—	—	—	—
rs985035	0.030	0.033	—	—	—	—
rs988474	−0.089	−0.085	0.122	—	0.087	−0.117
rs1334166	−0.039	−0.091	—	—	—	—
rs7333936	0.032	0.021	—	0.002	—	0.050
rs184385	0.040	0.031	—	−0.061	−0.090	—

(Continues)

TABLE 13 (Continued)

SNP	Main effect	Gender	CRP	TC	HDL	LDL
rs12855484	−0.052	−0.018	—	−0.041	—	—
rs9516496	0.022	0.040	—	—	—	—
rs7321486	−0.042	−0.023	−0.030	—	0.041	—
rs913427	0.029	0.044	—	—	—	−0.014
rs9556889	0.041	0.032	−0.096	—	—	—
rs679363	−0.039	−0.060	−0.002	0.041	—	—
rs1556799	−0.034	−0.045	—	—	—	—
rs1998550	−0.048	−0.060	−0.003	0.045	−0.025	0.005
rs701556	0.045	0.033	—	—	—	−0.073
rs1571513	−0.063	−0.025	−0.100	−0.083	0.020	—
rs1730649	0.054	0.067	0.031	—	0.112	−0.014
rs2067741	0.042	—	−0.085	0.040	−0.044	0.022
rs937872	0.051	—	0.216	—	0.002	—
rs9317872	−0.055	—	−0.013	—	0.093	0.020
rs9572541	0.052	—	−0.091	0.005	0.045	0.005
rs9300342	−0.020	—	−0.025	—	—	—
rs1998535	0.040	—	—	−0.031	—	−0.046
rs516872	−0.059	—	—	—	0.194	—
rs9572146	−0.011	—	—	—	—	—
rs4405440	−0.003	—	—	—	—	—

TABLE 14 NFBC1996 data analysis with insulin as the response variable: numbers of main G effects and interactions identified by different approaches (diagonal) and their overlaps (off-diagonal)

		Lasso	L1_dense	L1_a_h	Proposed
Main G effects	Lasso	50	18	13	35
	L1_dense	–	43	21	9
	L1_a_h	–	–	52	17
	Proposed	–	–	–	56
Interactions	Lasso	102	24	30	64
	L1_dense	–	90	35	16
	L1_a_h	–	–	99	29
	Proposed	–	–	–	120

make prediction for samples in the testing set, where prediction performance is evaluated using prediction mean squared error (PMSE). This procedure is repeated 1000 times. The training set estimates are also used to evaluate stability. This approach has been extensively adopted in the literature. The squared roots of the average PMSEs are 1.057 (L1_ag_h), 1.058 (L1_dense), 1.052 (Lasso), and 1.046 (Proposed). In the stability evaluation, we compute the OOI (observed occurrence index) for each effect. Briefly, the OOI is the probability of a specific effect being identified across replicates and measures the stability of finding. For the identified main G effects, the mean OOI values are 0.599 (L1_ag_h), 0.636 (L1_dense), 0.604 (Lasso), and 0.638 (Proposed). And for the identified interaction effects, the mean OOI values are 0.544 (L1_ag_h), 0.572 (L1_dense), 0.552

(Lasso), and 0.553 (Proposed). The proposed approach has competitive prediction performance and selection stability (Figure 12).

5 | ADDITIONAL ANALYSIS

To complement the above analysis, additional analysis is conducted and reported in the Data S1. In the first set of analysis, we repeat the above analysis under a different marginal screening approach, with which we select a block of consecutive SNPs. In the second set of analysis, we consider two alternative response variables. The findings have the same patterns as above. The proposed approach is able to make biologically sensible findings with satisfactory prediction and stability performance.

6 | DISCUSSION

In this article, we have developed a new G-E interaction analysis approach, taking advantage of the most recent development in data aggregation. The proposed approach can complement and advance from the existing approaches by effectively accommodating rare features, conducting joint analysis, more effectively aggregating nearby features, and others. It is built on the existing penalized joint G-E interaction analysis and state-of-the-art data integration [28] and has a sensible formulation. Simulation has demonstrated its competitive performance. In the NFBC data analysis, it has generated findings different from the alternatives and with satisfactory prediction and stability performance.

This study can be extended in multiple directions. As briefly described above, it can be (almost) directly applied to other data types/models. A closer examination of the proposed estimation suggests that it may not be specific to SNPs, physical locations (for tree construction), or rare features. When it is expected that certain features may share similar effects, and when a similarity measure can be defined statistically or functionally, the proposed approach may be applicable. In some genetic studies, multiple responses that share related genetic basis are jointly analyzed. In the NFBC1966 study, there are some traits that may share main G effects and interactions. It will be of interest to extend the proposed method to the collective analysis of multiple response variables. It may also be of interest for future research to establish theoretical properties, which may follow from Reference [28] and the existing theoretical studies on penalized G-E interaction analysis. In data analysis, the prediction and stability evaluation can provide some indirect support to the validity of our analysis. It is of interest further examine and validate the findings.

ACKNOWLEDGMENTS

We thank the editor and reviewers for their careful review and insightful comments, which have led to a significant improvement of this article.

DATA AVAILABILITY STATEMENT

Data analyzed in this study is publicly available.

ORCID

Shuangge Ma https://orcid.org/0000-0001-9001-4999

REFERENCES

1. W. T. Boyce, M. B. Sokolowski, and G. E. Robinson, *Genes and environments, development and time*, Proc. Natl. Acad. Sci. 117 (2020), no. 38, 23235–23241.
2. D. Thomas, *Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies*, Annu. Rev. Public Health 31 (2010), 21–36.
3. F. Zhou, J. Ren, X. Lu, S. Ma, and C. Wu, *Gene–environment interaction: Avariable selection perspective*, Methods and Protocols, Epistasis, 2021, 191–223.
4. M. Wu and S. Ma, *Robust genetic interaction analysis*, Brief. Bioinform. 20 (2019), no. 2, 624–637.
5. F. Zhou, J. Ren, G. Li, Y. Jiang, X. Li, W. Wang, and C. Wu, *Penalized variable selection for lipid–environment interactions in a longitudinal lipidomics study*, Genes 10 (2019), no. 12, 1002.
6. C. Wu, Y. Jiang, J. Ren, Y. Cui, and S. Ma, *Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures*, Stat. Med. 37 (2018), no. 3, 437–456.
7. L. Bomba, K. Walter, and N. Soranzo, *The impact of rare and low-frequency genetic variants in common disease*, Genome Biol. 18 (2017), no. 1, 1–17.
8. R. Mukherjee, N. S. Pillai, and X. Lin, *Hypothesis testing for high-dimensional sparse binary regression*, Ann. Stat. 43 (2015), no. 1, 352–381.
9. M. E. Tabangin, J. G. Woo, and L. J. Martin, *The effect of minor allele frequency on the likelihood of obtaining false positives*, BMC Proc. 3 (2009), no. S7, 1–4.
10. Y. Li, N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, T. Jiang, H. Jiang, A. Albrechtsen, G. Andersen, H. Cao, T. Korneliussen, N. Grarup, Y. Guo, I. Hellman, X. Jin, Q. Li, J. Liu, X. Liu, T. Sparsø, M. Tang, H. Wu, R. Wu, C. Yu, H. Zheng, A. Astrup, L. Bolund, J. Holmkvist, T. Jørgensen, K. Kristiansen, O. Schmitz, T. W. Schwartz, X. Zhang, R. Li, H. Yang, J. Wang, T. Hansen, O. Pedersen, R. Nielsen, and J. Wang, *Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants*, Nat. Genet. 42 (2010), no. 11, 969–972.
11. S. Nejentsev, N. Walker, D. Riches, M. Egholm, and J. A. Todd, *Rare variants of ifih1, a gene implicated in antiviral responses, protect against type 1 diabetes*, Science 324 (2009), no. 5925, 387–389.

12. J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, et al., *Evolution and functional impact of rare coding variation from deep sequencing of human exomes*, Science 337 (2012), no. 6090, 64–69.
13. D. B. Goldstein, A. Allen, J. Keebler, E. H. Margulies, S. Petrou, S. Petrovski, and S. Sunyaev, *Sequencing studies in human genetics: Design and interpretation*, Nat. Rev. Genet. 14 (2013), no. 7, 460–470.
14. G. V. Kryukov, L. A. Pennacchio, and S. R. Sunyaev, *Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies*, Am. J. Hum. Genet. 80 (2007), no. 4, 727–739.
15. A. Derkach, J. F. Lawless, D. Merico, A. D. Paterson, and L. Sun, *Evaluation of gene-based association tests for analyzing rare variants using genetic analysis workshop 18 data*, BMC Proc. 8 (2014), no. 1, 1–6.
16. C. Mallaney and Y. J. Sung, *Rare variant analysis of blood pressure phenotypes in the genetic analysis workshop 18 whole genome sequencing data using sequence kernel association test*, BMC Proc. 8 (2014), no. 1, 1–6.
17. J. Xuan, L. Yang, and Z. Wu, *Higher criticism approach to detect rare variants using whole genome sequencing data*, BMC Proceed. BioMed. Central 8 (2014), 1–6.
18. M. Agne, C. H. Huang, I. Hu, H. Wang, T. Zheng, and S. H. Lo, *Considering interactive effects in the identification of influential regions with extremely rare variants via fixed bin approach*, BMC Proc. 8 (2014), no. 1, 1–6.
19. H. C. Yang and H. W. Li, *Analysis of homozygosity disequilibrium using whole-genome sequencing data*, BMC Proceed. BioMed. Central 8 (2014), 1–5.
20. A. P. Morris and E. Zeggini, *An evaluation of statistical approaches to rare variant analysis in genetic association studies*, Genet. Epidemiol. 34 (2010), no. 2, 188–193.
21. Y. J. Sung, J. Basson, and D. C. Rao, *Whole genome sequence analysis of the simulated systolic blood pressure in genetic analysis workshop 18 family data: Long-term average and collapsing methods*, BMC Proceed. BioMed. Central 8 (2014), 1–5.
22. B. M. Neale, M. A. Rivas, B. F. Voight, A. David, D. Bernie, O. M. Marju, K. Sekar, S. M. Purcell, R. Kathryn, and M. J. a. Daly, *Testing for an unusual distribution of rare variants*, PLoS Genet. 7 (2011), no. 3, e1001322.
23. M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, *Rare-variant association testing for sequencing data with the sequence kernel association test*, Am. J. Hum. Genet. 89 (2011), no. 1, 82–93.
24. A. Derkach, J. F. Lawless, and L. Sun, *Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests*, Genet. Epidemiol. 37 (2013), no. 1, 110–121.
25. L. Luo, E. Boerwinkle, and M. Xiong, *Association studies for next-generation sequencing*, Genome Res. 21 (2011), no. 7, 1099–1108.
26. T. Wang and H. Zhao, *A dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms*, Biometrics 73 (2017), no. 3, 792–801.
27. A. Yazdani, A. Yazdani, and E. Boerwinkle, *Rare variants analysis using penalization methods for whole genome sequence data*, BMC Bioinform. 16 (2015), no. 1, 1–8.
28. X. Yan and J. Bien, *Rare feature selection in high dimensions*, J. Am. Stat. Assoc. 534 (2021), 887–900.
29. J. Bien, X. Yan, L. Simpson, and C. L. Müller, *Tree-aggregated predictive modeling of microbiome data*, Sci. Rep. 11 (2021), no. 1, 1–13.
30. M. Lu, H. S. Lee, D. Hadley, J. Z. Huang, and X. Qian, *Logistic principal component analysis for rare variants in gene-environment interaction analysis*, IEEE/ACM Trans. Comput. Biol. Bioinform. 11 (2014), no. 6, 1020–1028.
31. G. Zhao, R. Marceau, D. Zhang, and J.-Y. Tzeng, *Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression*, Genetics 199 (2015), no. 3, 695–710.
32. T. Yang, H. Chen, H. Tang, D. Li, and P. Wei, *A powerful and data-adaptive test for rare-variant–based geneenvironment interaction analysis*, Stat. Med. 38 (2019), no. 7, 1230–1244.
33. X. Lin, S. Lee, D. C. Christiani, and X. Lin, *Test for interactions between a genetic marker set and environment in generalized linear models*, Biostatistics 14 (2013), no. 4, 667–681.
34. E. Lim, H. Chen, J. Dupuis, and C. T. Liu, *A unified method for rare variant analysis of gene-environment interactions*, Stat. Med. 39 (2020), no. 6, 801–813.
35. J. Bien, J. Taylor, and R. Tibshirani, *A lasso for hierarchical interactions*, Ann. Stat. 41 (2013), no. 3, 1111–1141.
36. N. Hao, Y. Feng, and H. H. Zhang, *Model selection for high-dimensional quadratic regression via regularization*, J. Am. Stat. Assoc. 113 (2018), no. 522, 615–625.
37. M. Wu, Q. Zhang, and S. Ma, *Structured gene-environment interaction analysis*, Biometrics 76 (2020), no. 1, 23–35.
38. D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander, *Linkage disequilibrium in the human genome*, Nature 411 (2001), no. 6834, 199–204.
39. W. Pan, and X. Shen, *Penalized model-based clustering with application to variable selection*. J. Mach. Learn. Res. 8 (2007), no. 5, 1145–1164.
40. C. Sabatti, S. K. Service, A. L. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, et al., *Genome-wide association analysis of metabolic traits in a birth cohort from a founder population*, Nat. Genet. 41 (2009), no. 1, 35–46.
41. N. P. Torres-Aguila, C. Carrera, E. Muiño, N. Cullell, J. Cárcel-Márquez, C. Gallego-Fabrega, J. González-Sánchez, A. Bustamante, P. Delgado, L. Ibañez, L. Heitsch, J. Krupinski, J. Montaner, J. Martí-Fàbregas, C. Cruchaga, J. M. Lee, I. Fernandez-Cadenas, and Acute Endophenotypes Group of the International Stroke Genetics Consortium (ISGC), *Clinical variables and genetic risk factors associated with the acute outcome of ischemic stroke: A systematic review*, J. Stroke 21 (2019), no. 3, 276–289.
42. X. Shi, Y. Jiao, Y. Yang, C. Y. Cheng, C. Yang, X. Lin, and J. Liu, *Vimco: Variational inference for multiple correlated outcomes in genome-wide association studies*, Bioinformatics 35 (2019), no. 19, 3693–3700.
43. M. E. Rentería, A. Cortes, and S. E. Medland, "*Using plink for genome-wide association studies (gwas) and data analysis*," *Genome-wide association studies and genomic prediction*, Springer, 2013, pp. 193–213.
44. S. Sindhu, R. Thomas, P. Shihab, E. Al Shawaf, A. Hasan, M. Alghanim, K. Behbehani, and R. Ahmad, *Changes in the*

adipose tissue expression of cd86 costimulatory ligand and cd163 scavenger receptor in obesity and type-2 diabetes: Implication for metabolic disease, J. Glycom. Lipidom. 5 (2015), no. 134, 2153-0637.

45. S. Homma, T. Shimada, T. Hikake, and H. Yaginuma, *Expression pattern of lrr and ig domain-containing protein (lrrig protein) in the early mouse embryo*, Gene Expr. Patterns 9 (2009), no. 1, 1–26.
46. Y. Wu, K. Prakash, T. Rong, H. Li, Q. Xiao, L. Tan, W. Au, J. Ding, S. Chen, and E. Tan, *Lingo2 variants associated with essential tremor and parkinson's disease*, Hum. Genet. 129 (2011), no. 6, 611–615.
47. J. R. Speakman, *Functional analysis of seven genes linked to body mass index and adiposity by genome-wide association studies: A review*, Hum. Hered. 75 (2013), no. 2–4, 57–79.
48. N. Fernàndez-Castillo, G. Gan, M. M. van Donkelaar, M. Vaht, H. Weber, W. Retz, A. Meyer-Lindenberg, B. Franke, J. Harro, A. Reif, et al., *Rbfox1, encoding a splicing regulator, is a candidate gene for aggressive behavior*, Eur. Neuropsychopharmacol. 30 (2020), 44–55.
49. A. G. Comuzzie, S. A. Cole, S. L. Laston, V. S. Voruganti, K. Haack, R. A. Gibbs, and N. F. Butte, *Novel genetic loci identified for the pathophysiology of childhood obesity in the hispanic population*, PLoS One 7 (2012), no. 12, e51954.
50. R. Wang, Y. Lu, M. Z. Cicha, M. V. Singh, C. J. Benson, C. J. Madden, M. W. Chapleau, and F. M. Abboud, *Tmem16b determines cholecystokinin sensitivity of intestinal vagal afferents of nodose neurons*, JCI Insight 4 (2019), no. 5, e122058.
51. A. Mavri, D. Bastelica, M. Poggi, P. Morange, F. Peiretti, M. Verdier, I. Juhan Vague, and M. C. Alessi, *Polymorphism a36g of the tumor necrosis factor receptor 1 gene is associated with pai-1 levels in obese women*, Thromb. Haemost. 97 (2007), no. 01, 62–66.
52. W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant, H. Parkinson, and L. M. Schriml, *Disease ontology 2015 update: An expanded and updated database of human diseases for linking biomedical knowledge through disease data*, Nucleic Acids Res. 43 (2015), no. D1, D1071–D1078.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: M. Liu, Q. Zhang, and S. Ma, *A tree-based gene–environment interaction analysis with rare features*, Stat. Anal. Data Min.: ASA Data Sci. J. (2022), 1–27. https://doi.org/10.1002/sam.11578

in joint analysis, a large number of G measurements are analyzed in a single model. In the past decade, we have witnessed significant developments in G-E interaction analysis methodology, computation, theory, and application. For reviews and representative studies, we refer to [1–3]. In this article, we conduct joint G-E interaction analysis and note that joint and marginal analyses are two different analysis paradigms, have different implications, and cannot replace each other, although joint analysis may better fit the biology of complex diseases. For recent developments in joint G-E interaction analysis, we refer to [4, 5].

Our literature review suggests that, in most of the existing joint G-E interaction analyses, attention has been on "simple" data, for example, continuously distributed gene expressions [6] and single nucleotide polymorphisms (SNPs) with moderate to high MAFs (minor allele frequencies). Comparatively, attention to rare features, for example, SNPs with low MAFs (often defined as MAF < 5%) and certain methylation data, has been limited. Rare features are not uncommon in practice. In Figure 1 (Data S1), for the NFBC1996 data to be analyzed in Section 4, we show the genotype distributions of the rare features (post screening). Published studies have established that "ordinary" statistical methods lose power with rare features [7, 8], and that as features get increasingly rare, an unreasonably large sample size will be needed to detect their effects. Here, it is noted that such conclusions have been drawn for main-effect-only methods, most of which conduct marginal analysis. However, it is sensible to expect similar conclusions for interaction analysis. Some early studies inappropriately drop rare features from analysis [9]. With the development of personalized medicine, the significance of rare features for complex human diseases has been firmly recognized [10–12]. Its theoretical basis is that features that strongly predispose to diseases are likely to be deleterious and thus kept at low frequencies by purifying selection [13, 14]. Examining rare features can assist identifying subpopulations that may benefit from targeted treatment.

In main-effect-only analysis, it has been recognized that the most effective and possibly the only feasible strategy for identifying rare features is pooling. That is, as opposed to identifying the individual effects of rare features, the combined effects of "related" rare features, for example those in the same genetic region, are identified. Popular data pooling/collapsing strategies include gene-based bins [15, 16], windows of a fixed length [17], windows of a fixed number of variants [18, 19], and others. A common limitation of these approaches is that they do not take into account the directions of features' effects on a response variable. Generically, methods for analyzing rare features can be classified into four main categories: burden tests using linear statistics [20, 21], variance-components-type tests using quadratic statistics [22, 23], hybrid methods combining burden and variance-components-type tests [15, 24], and other dimension-reduction-based approaches. Examples in the last category include [25, 26], which conduct unsupervised clustering to create denser features. Another example is a penalization method called ConvexConcave Rare variant Selection (CCRS) [27]. However, it has been found that, even after applying the aforementioned aggregation methods, a large portion of aggregated rare features may still be too sparse, and they may still have to be discarded. Here, we note that the aforementioned and many other approaches are limited to marginal analysis in the hypothesis testing framework and are not directly applicable to joint analysis. Recognizing limitations of the existing data aggregation techniques, in a recent study, Yan and Bien [28] develop a more effective strategy for aggregating and selecting rare features, which leverage side information (additional prior information) in the form of a tree. A tree-based parameterization strategy is introduced to translate the feature aggregation problem into a sparse modeling one. Statistical and numerical investigations show that this approach can significantly improve over the existing ones. This flexible, data-adaptive, and tree-based aggregation approach is integrated into a log-contrast regression model in Reference [29]. It is noted that this approach has only been applied to main-effect-only analysis.

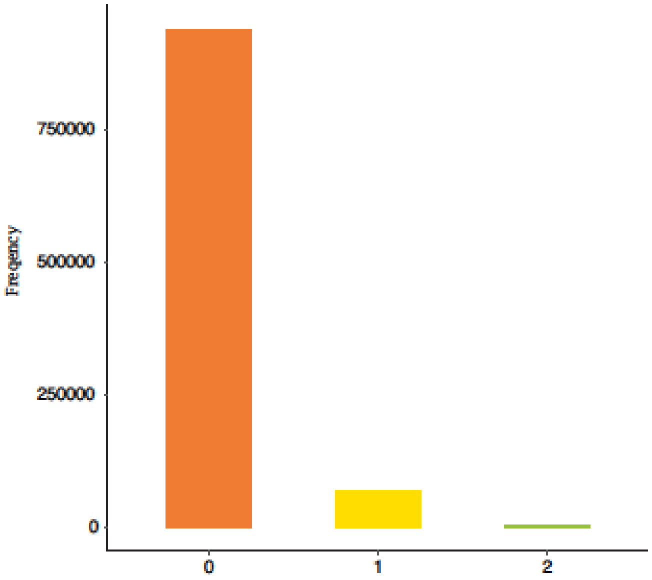

FIGURE 1 A small example of aggregating features within branches

With the high significance of G-E interaction analysis, there has been some effort on detecting interactions between rare features and E variables. For example, Lu and others [30] propose an aggregated statistic, which is derived from the MAF-based logistic principal component analysis (MLPCA). A limitation of this approach is that the adopted unsupervised technique is not ideal to indicate how genetic variants are modified by environment factors to affect disease risk and traits. Zhao and others [31] aggregate genetic and G-E interaction information across markers and construct score tests to identify important G-E interactions. Yang and others [32] develop a family of data-adaptive G-E interaction tests in the framework of adaptive powered score testing. It is noted that these works mostly belong to the marginal analysis paradigm. For joint analysis, Lin and others [33] develop a variance component score test within the induced generalized linear mixed model (GLMM) framework and apply ridge regression to estimate the nuisance main effects. Lim and others [34] adopt a kernel-based method to leverage joint information across rare variants under the GLMM framework. However, in these studies, there has been no attention to the "main effects, interactions" variable selection hierarchy [35, 36].

In this article, we consider joint G-E interaction analysis where a significant number of candidate G features are rare. Although certain individual components of this analysis share some common ground with the existing studies, overall, this study complements and advances published literature in the following aspects. Unlike most of the existing G-E interaction studies, there is special attention to rare features. It differs from most of the existing rare feature studies by conducting joint analysis (which differs significantly from "marginal analysis + hypothesis testing") and by accommodating interactions (and the accompanying unique challenges in particular the "main effects, interactions" variable selection hierarchy). It also advances from many existing pooling studies for rare features by adopting the cutting-edge tree-based aggregation technique [32] and from Reference [32] by conducting joint interaction analysis. In addition, the proposed approach can directly go beyond rare features and be applied to other types of data that also have individual weak effects, and hence data integration is needed.

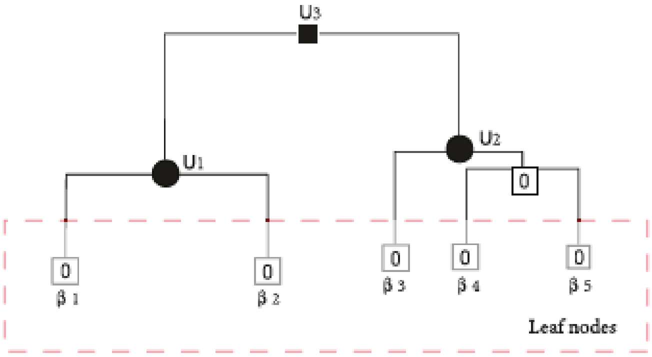

FIGURE 2 Aggregating $\boldsymbol{\beta}$ (left) and $\boldsymbol{\xi}_{\mathbf{k}}(k = 1, \dots, 3$; right) in $\mathcal{T}$

2 | METHODS

2.1 | Data and model

Y is denoted as the disease outcome/phenotype. In what follows, we consider a continuously distributed outcome and corresponding linear regression. The proposed approach can be directly applied to other

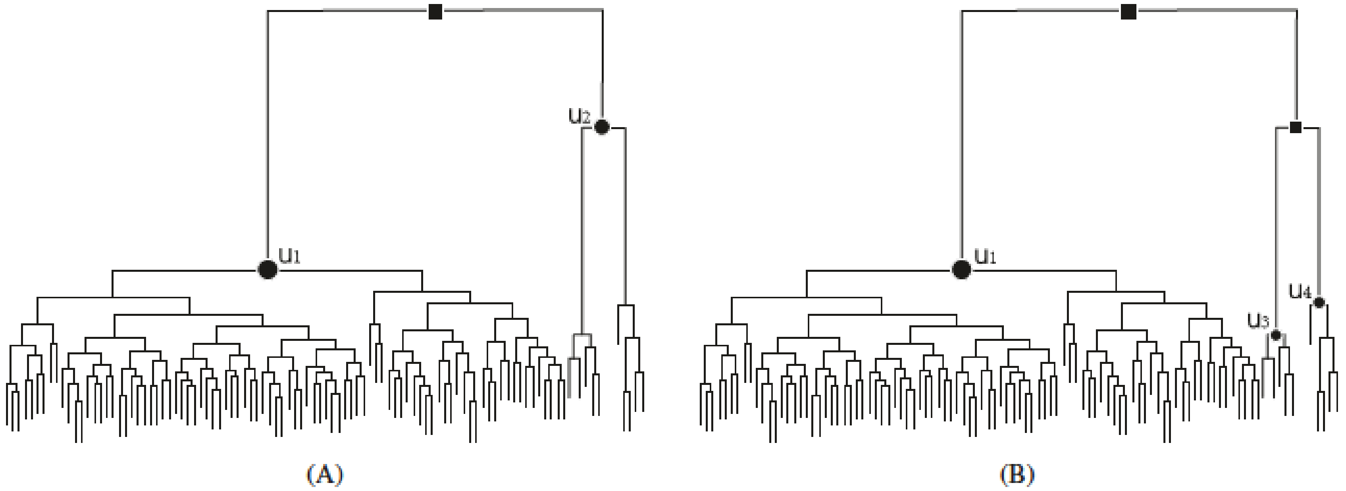

FIGURE 3 The tree structure $\mathcal{T}$ of p leaves with $(p, m, s) = (200{,}20{,}0.4)$. Gray leaves have zero effects, leaves with the other colors have nonzero effects, and leaves with the same color have the same effects

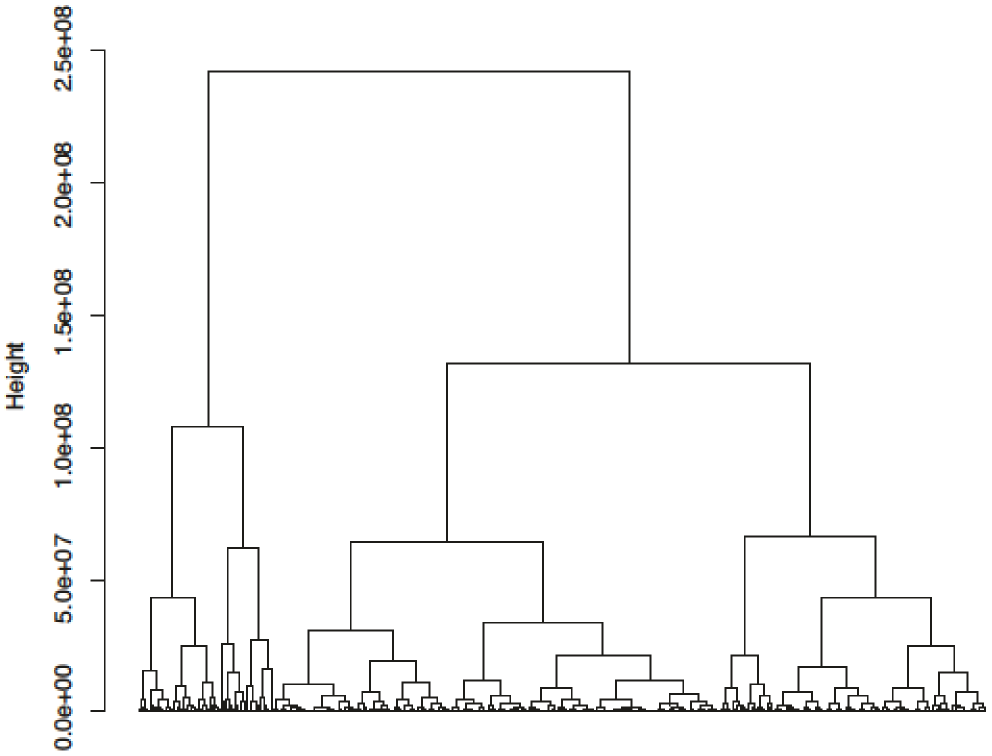

FIGURE 4 NFBC1996 data: distribution of genotypes for the rare single nucleotide polymorphisms (SNPs) after screening

types of outcomes/phenotypes by adopting corresponding regression models and likelihood functions. $Z = (Z_1, \dots, Z_p)'$ is denoted as the p rare features. In our data analysis, we consider SNPs with low MAFs. Further, $X = (X_1, \dots, X_q)'$ is denoted as the q clinical/environmental risk factors. Following strong advocate in the recent literature, we also consider the interactions with demographic and clinical variables. It is also possible to limit interactions to narrowly defined E factors. Consider the joint regression model:

$$Y = \sum_{k=1}^{q} \alpha_k X_k + \sum_{j=1}^{p} \left(\beta_j Z_j + \sum_{k=1}^{q} \eta_{kj} X_k Z_j \right) + \varepsilon, \qquad (1)$$

where α_k's, β_j's, and η_{kj}'s are the regression coefficients for the main E effects, main G effects, and their interactions, respectively. ε is the random error. With proper normalization, the intercept term has been omitted. There are multiple ways of respecting the "main effects, interactions" variable selection hierarchy. Here, we adopt the decomposition strategy [37], where $\eta_{kj} = \beta_j \xi_{kj}$. Then, model (1) can be rewritten as:

$$Y = \sum_{k=1}^{q} \alpha_k X_k + \sum_{j=1}^{p} \left(\beta_j Z_j + \sum_{k=1}^{q} \beta_j \xi_{kj} X_k Z_j \right) + \varepsilon.$$

Denote $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)'$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$, and $\xi_k = (\xi_{k1}, \dots, \xi_{kp})'$. Assume n iid observations $\{(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i), i = 1, \dots, n\}$. Denote $\boldsymbol{y}$ as the n-vector composed of y_i's, $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}^{(k)}$ as the matrices composed of x_i's, z_i's, and $\boldsymbol{w}_i^{(k)} = (x_{ik} z_{i1}, \dots, x_{ik} z_{ip})'$s, respectively. In the matrix form, the least squares objective function is $L(\theta) = \frac{1}{2n} \left\| \boldsymbol{y} - \boldsymbol{X}\alpha - \boldsymbol{Z}\beta - \sum_{k=1}^{q} \boldsymbol{W}^{(k)} (\beta \odot \xi_k) \right\|_2^2$, where $\theta = (\alpha', \beta', \xi_1', \dots, \xi_q')'$, $\| \cdot \|_2$ is the l_2 norm, and $\odot$ is the component-wise product.

We note that the data and model settings have been extensively adopted in the literature, with the difference that Z represents rare features. It is expected that other loss functions, for example, the robust ones, can also be adopted.

2.2 | Estimation

With data aggregation, one of the most critical steps is to determine the regions within which rare features are pooled. Quite a few approaches have been developed for this purpose. Some utilize biological information, for example, functionalities of SNPs. However, this may be not sufficiently effective as the functions of many SNPs, especially those in noncoding regions, are unknown. Another family of approaches utilizes information on features' physical locations, which is usually known. When SNPs are densely measured, those physically close can be in high linkage disequilibrium (LD) and have similar biological functions and/or statistical effects [38]. In our numerical study, for SNP data, we follow [28] and conduct hierarchical clustering analysis of the physical locations of SNPs

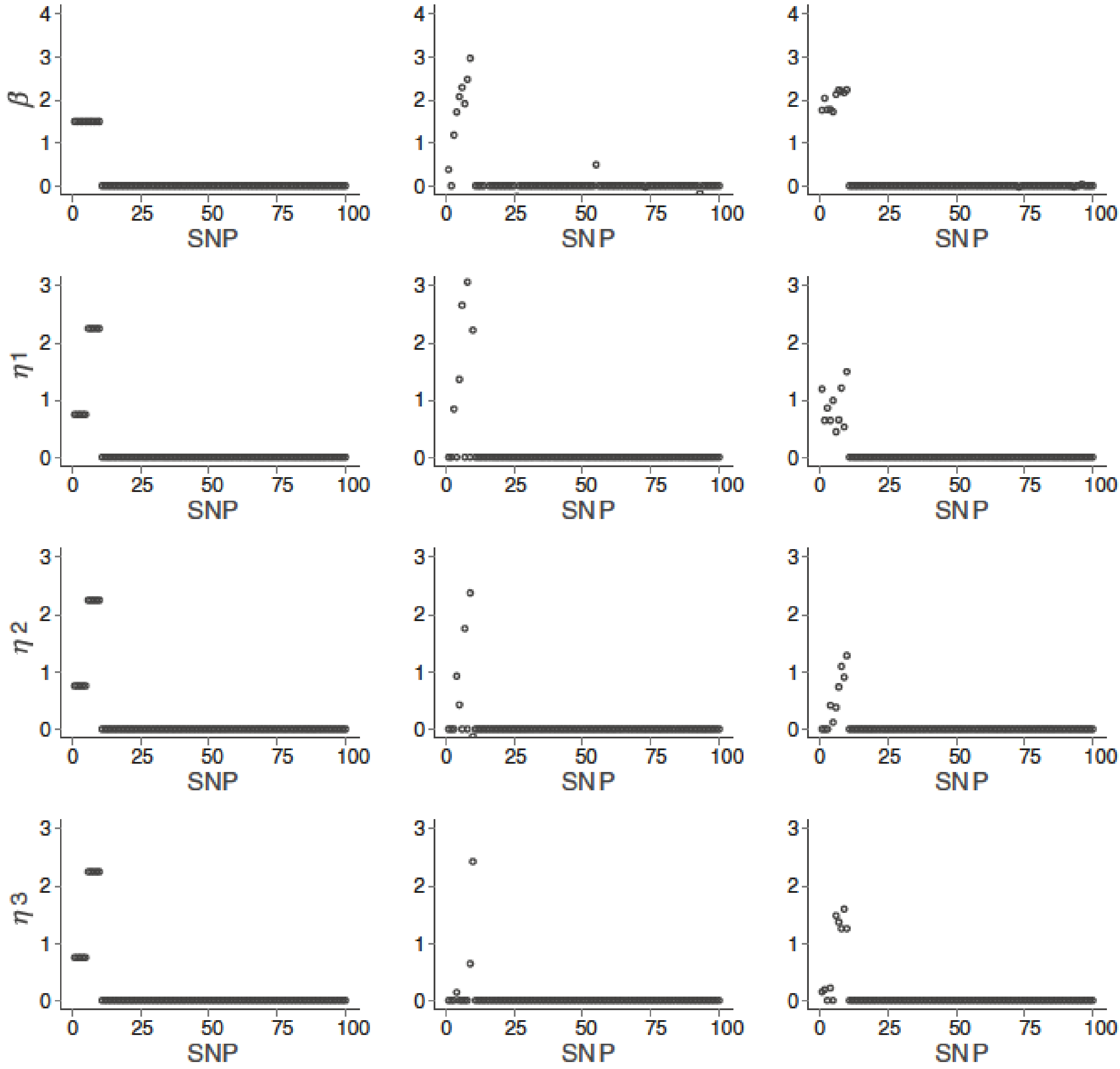

FIGURE 5 Toy example: true (left) and estimated (center: Lasso; right: proposed) main G effects and interactions

to form a tree $\mathcal{T}$, as showcased in Figures 2, 3, and 4. The consideration is that features physically close to each other tend to have related biological functions, which has been established for SNP and some other types of data. We refer to Reference [28] for more discussions on the tree construction. Advancing from Reference [28], we also incorporate interactions and propose densifying β and $\xi_{\mathbf{k}}$ using the same tree structure.

Let u be a node, which is a branching point in a tree. A node is called a leaf node, if it has no additional nodes coming out of it. For example, in Figure 2, those in the red box are leaf nodes. The ancestor(u) and descendant(u) are denoted as the ancestors and descendants of node u in $\mathcal{T}$, respectively. The set of nodes in the path from the root of $\mathcal{T}$ to the jth leaf can be written as ancestor$(j) \cup \{j\}$. Assign a parameter γ_{0u} (γ_{ku}) to each node u in $\mathcal{T}$. Similar to [28], we can conduct a tree-based parameterization to associate β_j and ξ_{kj} with $\mathcal{T}$. Specifically, β_j and $\xi_{kj}(k = 1, \dots, q)$ are decomposed into the sum of all the parameters on the path:

$$\beta_j = \sum_{u \in \text{ancestor}(j)\cup\{j\}} \gamma_{0u}, \quad \xi_{kj} = \sum_{u \in \text{ancestor}(j)\cup\{j\}} \gamma_{ku}. \qquad (2)$$

When $\gamma_{0\,\text{descendant}(u)} = 0$ $(\gamma_{k\,\text{descendant}(u)} = 0)$, β_j's (ξ_{kj}'s) associated with the leaves lying beneath node u are equal. For example, with the tree in Figure 2, coefficients of all the nodes beneath nodes u_1 and u_2 are zero. According to (2), β_j's are aggregated into two groups: $\beta_1 = \beta_2 = \gamma_{0u_1} + \gamma_{0u_3}$

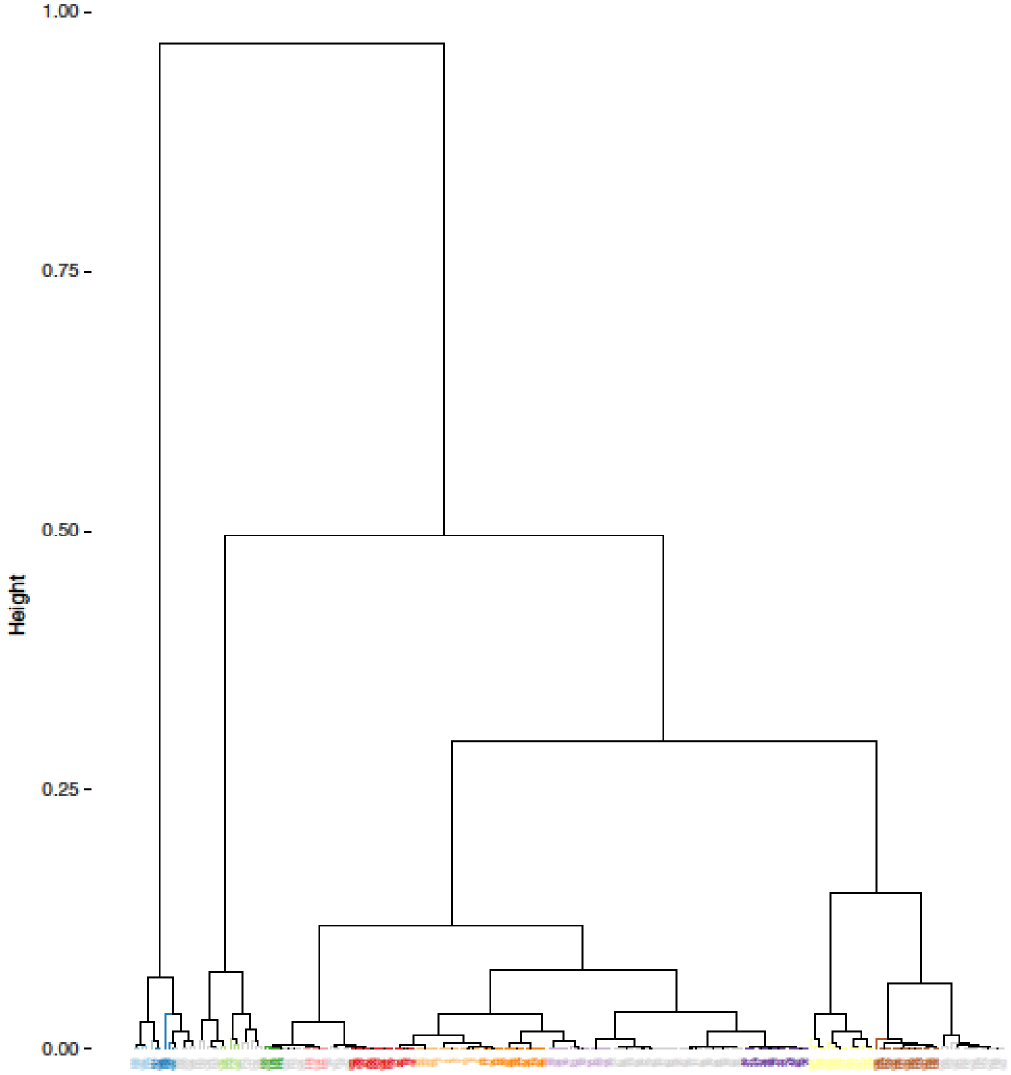

FIGURE 6 Simulation: estimation error with $(n, p, q, s) = (200,200,5,0.4)$

and $\beta_3 = \beta_4 = \beta_5 = \gamma_{0u_2} + \gamma_{0u_3}$. As such, feature aggregation can be achieved by introducing sparsity to $\gamma_0(\gamma_k)$.

For regularized estimation and selection of important interactions and main effects, we propose the penalized objective function:

$$Q_n(\theta, \Gamma) = L(\theta) + a\lambda \sum_{\ell=1}^{|\mathcal{T}|} \left[\omega_{0\ell} \, |\gamma_{0\ell}| + \sum_{k=1}^{q} \omega_{k\ell} \, |\gamma_{k\ell}| \right] + (1-a)\lambda \sum_{j=1}^{p} \left[\widetilde{\omega}_{0j} \, |\beta_j| + \sum_{k=1}^{q} \widetilde{\omega}_{kj} \, |\xi_{kj}| \right],$$
$$s.t. \;\; \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\gamma}_0, \boldsymbol{\xi}_k = \boldsymbol{A}\boldsymbol{\gamma}_k (k = 1, \dots, q), \qquad (3)$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q)' \in \mathbb{R}^{(q+1)\times|\mathcal{T}|}$, $\lambda \geq 0$ and $a \in [0, 1]$ are tuning parameters, $\omega_{0\ell}, \omega_{k\ell}, \widetilde{\omega}_{0j}, \widetilde{\omega}_{kj}$ are covariate-specific weights (more details below), $|\mathcal{T}|$ denotes the number of nodes in $\mathcal{T}$, $\boldsymbol{A} \in \{0, 1\}^{p\times|\mathcal{T}|}$ is a matrix with elements $\boldsymbol{A}_{jr} := 1_{\{u_r \in \text{ancestor}(j)\cup\{j\}\}} = 1_{\{j\in \text{descendant}(u_r)\cup\{u_r\}\}}$, and $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\gamma}_0$ and $\boldsymbol{\xi}_k = \boldsymbol{A}\boldsymbol{\gamma}_k$ are the compact forms of (2). Similar to other penalized interaction analyses, interactions, and main effects with nonzero coefficients are identified as being important for the response.

Rationale The overall strategy is similar to other penalizations, with the first term quantifies lack-of-fit—it can be revised to accommodate other data types/models. The two penalty terms induce different types of sparsity, which are controlled by λ and balanced by a. The second penalty is relatively "simple" and has been considered in the existing penalized G-E interaction studies. In particular, the Lasso penalty is directly imposed to $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_k$, identifying important main effects and interactions. With the decomposition strategy, the variable selection hierarchy is guaranteed. The weights $\widetilde{\omega}_{0j}, \widetilde{\omega}_{kj}$ lead to weighted (adaptive) penalization. For choosing weights, we refer to Reference [28] and many other publications. The most straightforward choice, which is adopted in our numerical study, is to set the weights equal to 1. The most significant advancement over the existing G-E interaction analysis is

TABLE 1 Simulation Scenario 1

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	0.97(0.07)	0.03(0.01)	1.00(0.00)	0.21(0.13)	4.63(0.54)	0.21(0.08)
L1_dense	0.62(0.08)	0.04(0.01)	0.78(0.07)	0.22(0.04)	6.43(0.77)	3.12(1.08)
L1_ag_h	0.84(0.09)	0.13(0.05)	1.00(0.04)	0.68(0.16)	5.68(1.06)	3.69(0.72)
Lasso	0.74(0.06)	0.02(0.00)	0.95(0.04)	0.19(0.03))	6.43(0.8)	2.64(1.00)
Proposed	0.82(0.13)	0.01(0.01)	1.00(0.00)	0.08(0.07)	4.85(0.58)	0.82(0.47)
$m = 40$						
Oracle Lasso	0.98(0.03)	0.03(0.01)	1.00(0.00)	0.09(0.06)	5.62(0.86)	0.39(0.16)
L1_dense	0.61(0.09)	0.05(0.02)	0.77(0.08)	0.24(0.05)	7.12(0.82)	3.72(1.56)
L1_ag_h	0.75(0.11)	0.15(0.08)	0.99(0.06)	0.84(0.05)	7.03(1.68)	5.66(1.15)
Lasso	0.74(0.07)	0.03(0.01)	0.94(0.06)	0.19(0.03)	7.05(0.76)	2.79(1.21)
Proposed	0.89(0.09)	0.01(0.01)	1.00(0.01)	0.09(0.04)	5.80(0.56)	1.06(0.42)
$m = 60$						
Oracle Lasso	0.94(0.06)	0.02(0.01)	1.00(0.01)	0.04(0.03)	6.42(0.98)	0.59(0.34)
L1_dense	0.61(0.09)	0.04(0.01)	0.75(0.10)	0.22(0.03)	6.74(0.66)	3.47(1.20)
L1_ag_h	0.56(0.18)	0.14(0.12)	1.00(0.00)	0.9(0.08)	8.04(1.82)	6.22(1.09)
Lasso	0.73(0.06)	0.03(0.00)	0.93(0.05)	0.19(0.03)	6.54(0.64)	2.66(1.15)
Proposed	0.88(0.09)	0.02(0.01)	1.00(0.01)	0.08(0.03)	6.48(0.48)	1.13(0.39)
$m = 80$						
Oracle Lasso	0.91(0.05)	0.02(0.01)	1.00(0.02)	0.02(0.02)	6.88(0.74)	0.81(0.44)
L1_dense	0.59(0.08)	0.05(0.01)	0.74(0.09)	0.23(0.04)	7.38(0.70)	4.35(1.29)
L1_ag_h	0.57(0.16)	0.15(0.18)	0.98(0.05)	0.93(0.06)	7.85(2.03)	6.48(1.76)
Lasso	0.71(0.07)	0.03(0.01)	0.93(0.05)	0.19(0.03)	7.24(0.59)	3.53(1.52)
Proposed	0.89(0.07)	0.03(0.01)	1.00(0.02)	0.12(0.04)	7.07(1.84)	1.56(0.62)
$m = 100$						
Oracle Lasso	0.86(0.05)	0.02(0.01)	0.97(0.03)	0.02(0.02)	7.05(0.87)	1.23(0.55)
L1_dense	0.60(0.07)	0.05(0.01)	0.76(0.08)	0.23(0.04)	7.20(0.60)	4.56(1.57)
L1_ag_h	0.20(0.16)	0.15(0.16)	0.95(0.09)	0.85(0.13)	8.62(1.66)	7.71(1.22)
Lasso	0.73(0.07)	0.03(0.01)	0.94(0.04)	0.18(0.03)	7.16(0.64)	3.53(1.43)
Proposed	0.88(0.06)	0.03(0.01)	0.97(0.03)	0.13(0.04)	7.09(0.57)	1.84(0.87)

Note: In each cell, mean (SD) based on 500 replicates.

the first term. Penalty is imposed to $\gamma_{0\ell}$ and $\gamma_{k\ell}$, which, with the constraint defined in (2), induces *fusion* to the coefficients in $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_k$. This fusion is built on the tree structure (as showcased in Figures 2 and 3). In particular, following [28], we leave the root ($\gamma_{k|\mathcal{T}|}$ for $k = 0, \ldots , q$) unpenalized with $\left\{\omega_{k|\mathcal{T}|} = 0\right\}_{\{k=0,1,\ldots,q\}}$. This allows all features to be aggregated into one single group with coefficients fused to a nonzero value. Under $\mathcal{T}$, nearby features, which are expected to have similar effects, are put into the same data aggregating sets. Their effects are fused to be similar, which allows nearby rare features to borrow strength from their neighbors. The aggregated effects can be considerably larger than the individual ones, making them more likely to be identified. It is noted that, with the proposed penalty, $\gamma_{0\,\mathrm{descendant}(u)}$ ($\gamma_{k\,\mathrm{descendant}(u)}$) is encouraged but not forced to be zero. As such, with this fusion/data aggregation technique, features in the same aggregating sets not necessarily have the same coefficients, making this approach more flexible than, for example, those directly adding up rare features.

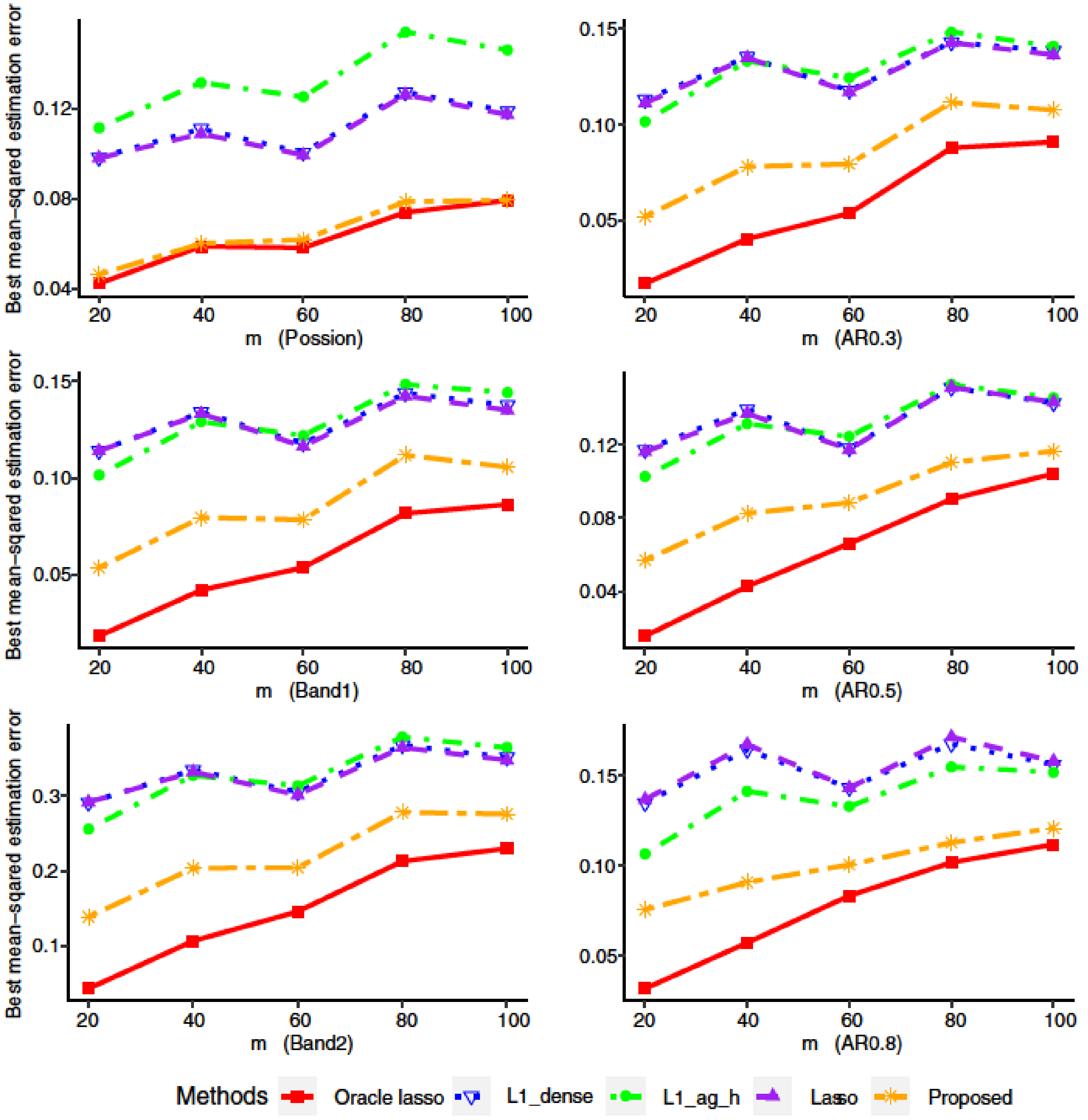

FIGURE 7 Pairwise LD analysis of rare single nucleotide polymorphisms (SNPs) (post screening). Top: LD decay plot; bottom: LD heatmap

A toy example To better appreciate working characteristics of the proposed method, we simulate a small dataset with $n = 100$ and $p = 100$. The tree structures for the main G effects and (components of) G-E interactions are shown in Figure 3. The true aggregating sets are determined based on Figure 3. In particular, the main G effects β_j' are aggregated into two groups, corresponding to nodes u_1 and u_2. All the leaves under u_1 have coefficients zero. β_j' under node u_2 are set to be 1.5. ξ_{kj}'s are aggregated into three groups, corresponding to nodes u_1, u_3, u_4. ξ_{kj}'s under node u_1 are set as 0, and those under nodes u_3 and u_4 are set as 0.75 and 2.25, respectively. Finally, the G-E interactions are calculated as $\eta_{kj} = \beta_j \xi_{kj}$. There are in total 10 main G effects and 30 G-E interactions with nonzero coefficients, and they satisfy the variable selection hierarchy. We graphically show the true regression coefficients in the left column of Figure 5. The SNP measurements are simulated from a Poisson(0.02) distribution and truncated at 2 if

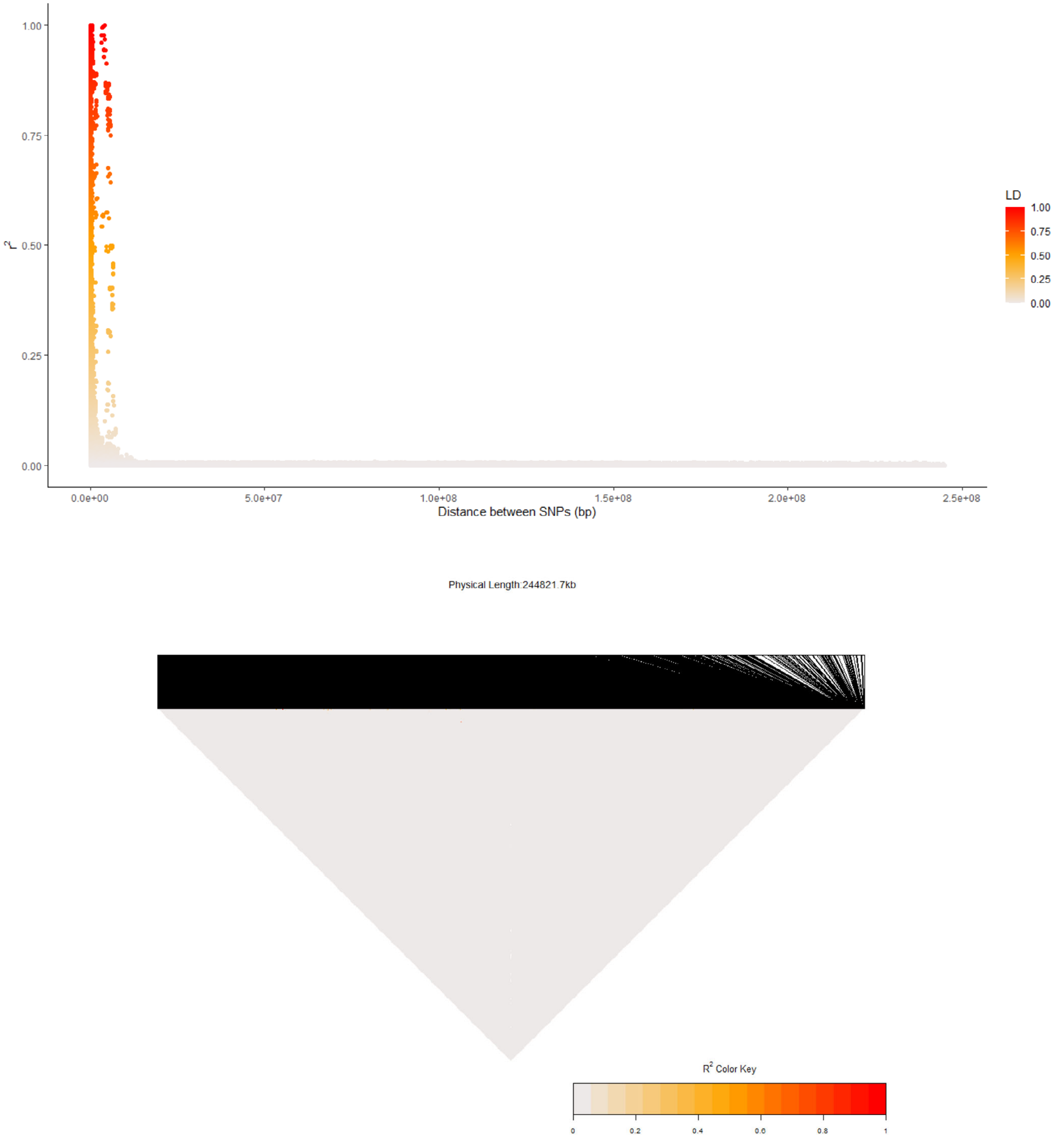

FIGURE 8 Pairwise LD analysis of all single nucleotide polymorphisms (SNPs) (post screening). Top: LD decay plot; bottom: LD heatmap

needed. We then simulate three E variables as having a Bernoulli distribution with probability of success 0.7. The response variable is generated from a linear regression model with a standard normally distributed random error. Beyond the proposed approach, we also consider the Lasso approach as a benchmark, which shares the same penalization framework as the proposed approach but does not conduct data aggregation. The estimation results using the proposed and Lasso approaches are graphically presented in Figure 5. By borrowing strength and effectively aggregating data, the proposed approach is observed to have significantly better identification and estimation

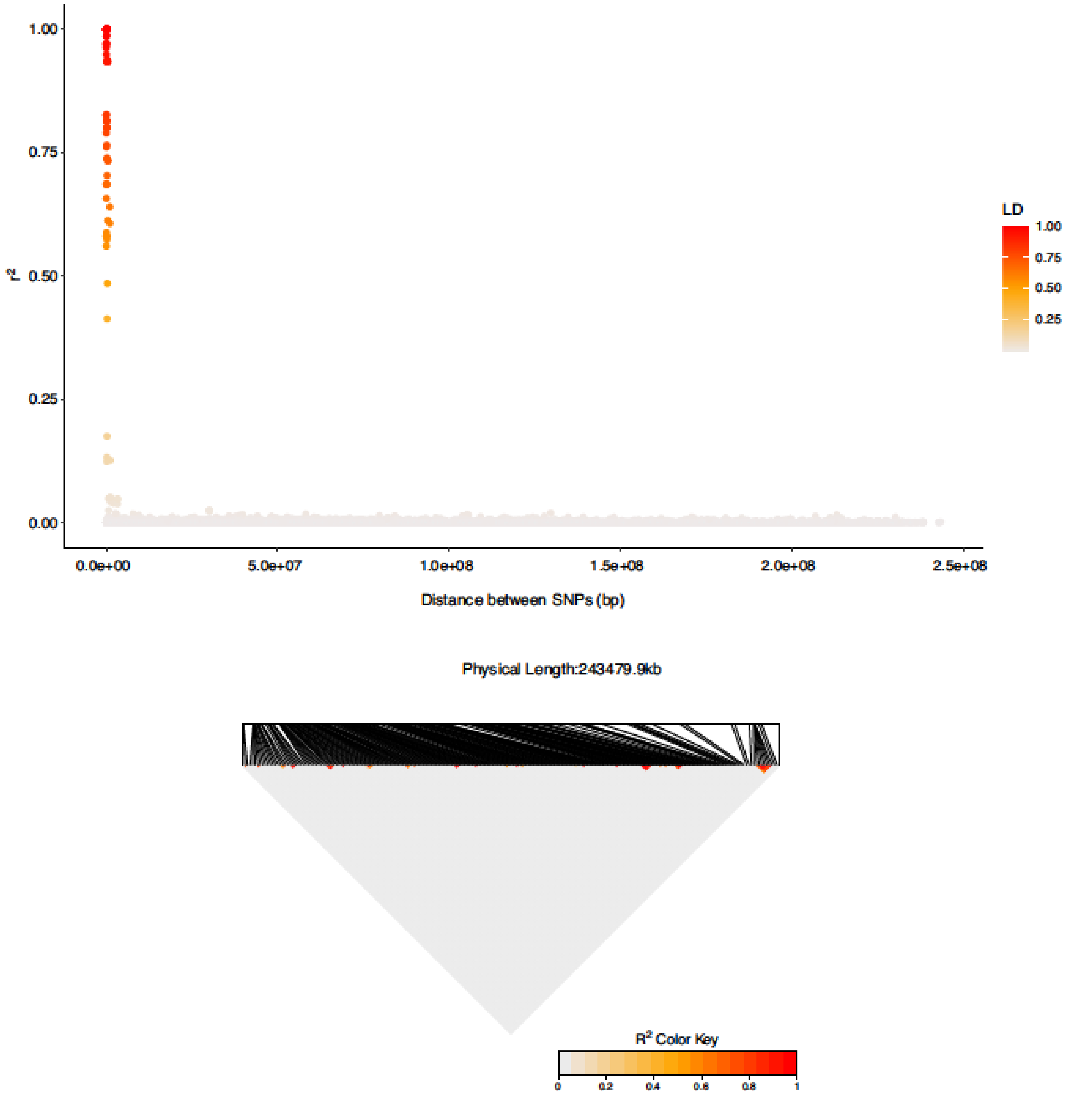

FIGURE 9 NFBC1996 data analysis: tree $\mathcal{T}$ of single nucleotide polymorphisms' (SNPs') physical positions (post screening)

accuracy. More definitive results based on larger-scale simulations are presented below in Section 3.

2.3 | Computation

With fixed tuning parameters, the optimization of (3) can be conducted using an iterative coordinate descent (CD) algorithm, which optimizes the objective function with respect to one of the three (sets of) vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{\xi}_k$'s at a time and iteratively cycles through all of the parameters until convergence. Let $\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}$, and $\boldsymbol{\xi}_k^{(t)}$ denote the estimates of $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{\xi}_k$ at iteration t, respectively. The proposed algorithm proceeds as follows:

Step 1 Initialize $t = 0$, $\boldsymbol{\beta}^{(t)} = \mathbf{0}$, $\boldsymbol{\xi}_k^{(t)} = \mathbf{0}$, and $\boldsymbol{\alpha}^{(t)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

Step 2 Update $t = t + 1$. With $\boldsymbol{\xi}_k$ and $\boldsymbol{\alpha}$ fixed at $\boldsymbol{\xi}_k^{(t-1)}$ and $\boldsymbol{\alpha}^{(t-1)}$, optimize (3) with respect to $\boldsymbol{\beta}$. Let $\widetilde{\boldsymbol{y}}^{(t)} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}^{(t-1)}$ and $\widetilde{\boldsymbol{Z}}^{(t)} = \boldsymbol{Z} + \sum_{k=1}^{q} \boldsymbol{W}^{(k)} \odot \left(\mathbf{1}_{n\times 1}\left(\boldsymbol{\xi}_k^{(t-1)}\right)' \right)$ with

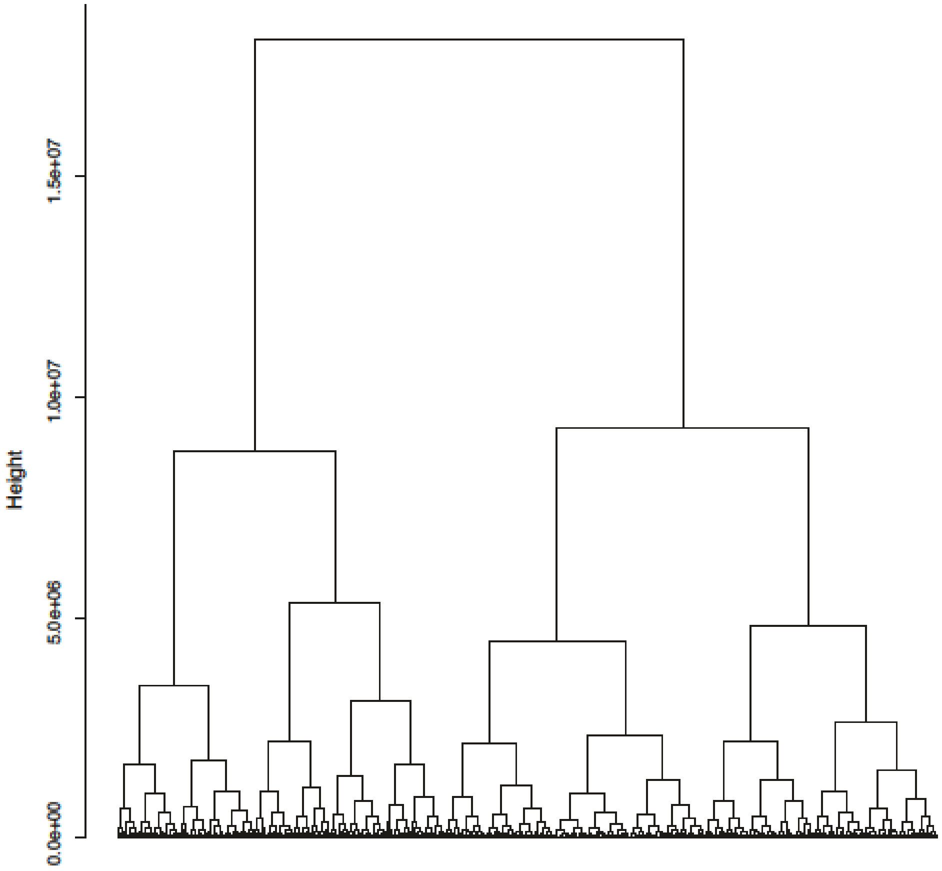

FIGURE 10 NFBC1996 data analysis: tree $\mathcal{T}$ of single nucleotide polymorphisms' (SNPs') physical positions (post screening), under the alternative screening approach

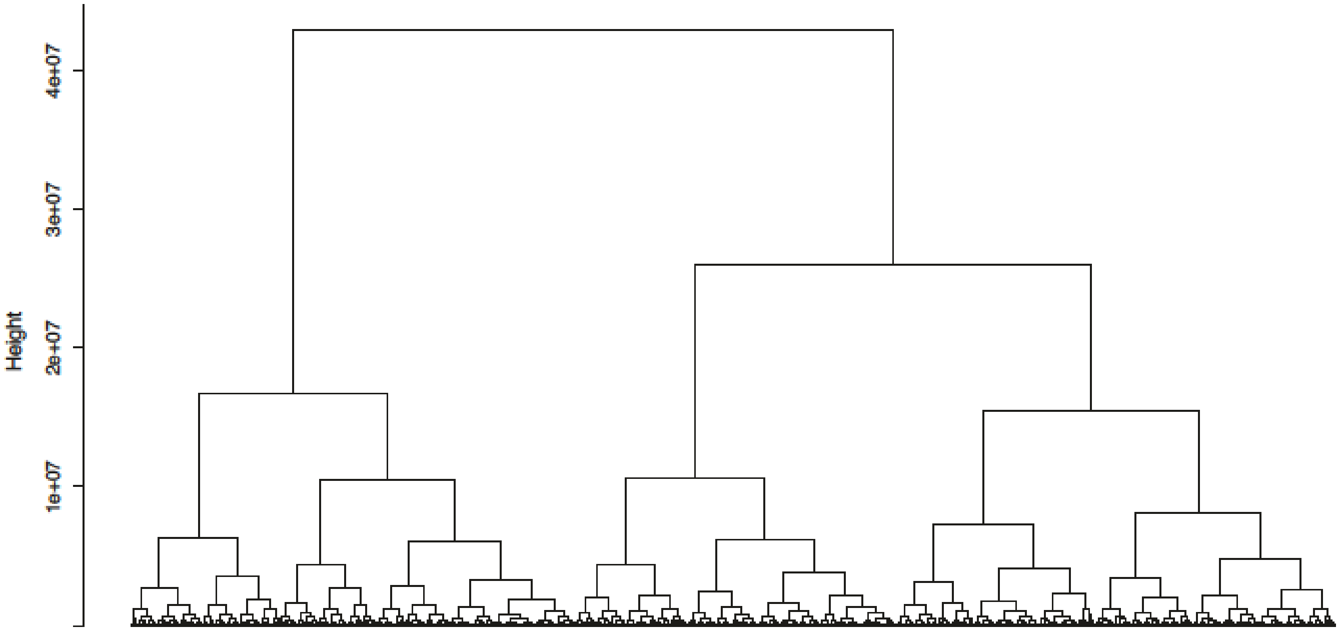

FIGURE 11 NFBC1996 data analysis: tree $\mathcal{T}$ of single nucleotide polymorphisms' (SNPs') physical positions (post screening), with glucose as the response variable

TABLE 2 Analysis of the NFBC1996 data using the proposed approach: identified main effects and interactions

SNP	Main effect	Gender	CRP	Glucose	TC	HDL
rs12208417	0.015	0.045	0.014	—	0.005	0.046
rs6488338	−0.017	−0.037	0.074	—	−0.014	—
rs2453244	−0.015	−0.044	−0.007	0.022	−0.067	—
rs11042023	−0.008	−0.057	—	—	—	0.027
rs2511841	−0.012	−0.081	0.014	0	−0.028	0.076
rs4965685	0.016	0.037	—	0.005	—	0.029
rs489487	0.01	0.05	—	−0.011	—	—
rs7306908	−0.016	−0.069	—	−0.015	—	—
rs10949732	0.017	0.039	−0.095	—	—	0.013
rs4575188	0.011	0.036	0.016	—	0.017	—
rs4720078	−0.016	−0.048	—	−0.012	−0.003	0.013
rs7039156	−0.018	−0.034	—	—	−0.022	−0.011
rs1676996	0.017	0.053	—	0.042	0.026	0.01
rs1386894	0.009	0.057	—	—	−0.014	—
rs1180819	−0.015	−0.051	0.005	—	−0.008	−0.053
rs10512052	0.014	0.042	−0.041	—	0.011	—
rs1237044	−0.016	−0.03	0.007	—	—	−0.109
rs10773484	−0.013	−0.065	—	−0.01	—	—
rs1284412	0.012	0.029	0.055	0.051	—	—
rs2571249	0.016	0.052	0.05	—	−0.027	−0.001
rs4149570	0.009	0.031	—	—	—	−0.005
rs1372555	−0.013	−0.049	−0.011	0.001	—	−0.035
rs3782631	0.013	0.06	0.01	—	—	—
rs2121671	0.015	0.048	−0.004	−0.011	−0.016	—
rs1870591	−0.011	−0.084	—	—	—	0.008
rs10508924	0.015	0.043	0.018	−0.018	—	−0.011
rs2834889	−0.011	−0.025	0.019	—	0.019	−0.003
rs7186722	−0.022	−0.069	—	0.006	0.046	0.094
rs3092379	0.013	0.038	—	—	0.008	0.006
rs2150855	−0.017	−0.051	−0.005	—	0.017	0.046
rs516783	0.013	0.058	—	0.009	—	−0.009
rs7506974	0.012	0.022	—	−0.049	—	0.009
rs3898586	0.014	0.061	0.04	—	0.003	—
rs344386	−0.024	−0.045	−0.02	—	—	0.038

$\mathbf{1}_{n\times 1} = (1, \dots, 1)_{n\times 1}$. Then

$$\beta^{(t)} = \arg\min_{\beta} \frac{1}{2n} \left\| \widetilde{\boldsymbol{y}}^{(t)} - \widetilde{\boldsymbol{Z}}^{(t)} \boldsymbol{\beta} \right\|_2^2 + \lambda \left\{ a \sum_{\ell=1}^{|\mathcal{T}|} w_{0\ell} \left|\gamma_{0\ell}\right| + (1-a) \sum_{j=1}^{p} \widetilde{w}_{0j} \left|\beta_j\right| \right\}$$
$$s.t. \ \ \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\gamma}_0. \tag{4}$$

To simplify notation, we consider the representative setting with $w_{0|\mathcal{T}|} = 0$ and $\left\{w_{0\ell} = 1, \widetilde{w}_{0j} = 1\right\}_{\{l \neq |\mathcal{T}| \, | j \in \{1, \dots, p\}\}}$. Problem (4) can be efficiently solved with the consensus ADMM algorithm [28]. Taking the form of a decomposition-coordination procedure, it combines the benefit of dual decomposition and augmented Lagrangian methods for constrained optimization.

Step 3 With $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ fixed at $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\alpha}^{(t-1)}$, optimize (3) with respect to $\boldsymbol{\xi} = \left(\xi_1, \dots, \xi_q\right)$. Let $\breve{\boldsymbol{y}}^{(t)} = \boldsymbol{y} -$

TABLE 3 Simulation Scenario 2

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	1.00(0.02)	0.05(0.03)	1.00(0.00)	0.65(0.13)	3.01(1.08)	0.31(0.17)
L1_dense	0.14(0.07)	0.00(0.00)	0.42(0.19)	0.02(0.01)	7.03(1.00)	6.64(3.60)
L1_ag_h	0.89(0.09)	0.08(0.04)	0.98(0.08)	0.61(0.16)	6.18(0.94)	3.84(1.08)
Lasso	0.14(0.07)	0.00(0.00)	0.42(0.19)	0.02(0.01)	6.94(0.99)	6.46(3.36)
Proposed	0.67(0.20)	0.01(0.01)	0.94(0.09)	0.18(0.08)	5.03(1.00)	3.11(2.38)
$m = 40$						
Oracle Lasso	0.95(0.06)	0.05(0.02)	0.99(0.04)	0.57(0.11)	4.58(1.11)	0.96(0.70)
L1_dense	0.12(0.07)	0.00(0.00)	0.36(0.19)	0.02(0.02)	7.47(0.95)	7.39(3.07)
L1_ag_h	0.85(0.13)	0.24(0.08)	1.00(0.00)	0.76(0.15)	7.19(0.94)	5.61(1.93)
Lasso	0.12(0.07)	0.00(0.00)	0.36(0.19)	0.02(0.02)	7.4(0.89)	7.37(2.96)
Proposed	0.56(0.22)	0.01(0.01)	0.90(0.09)	0.18(0.09)	5.99(1.01)	4.41(1.99)
$m = 60$						
Oracle Lasso	0.88(0.08)	0.03(0.02)	0.98(0.04)	0.48(0.11)	5.33(1.38)	1.93(1.18)
L1_dense	0.14(0.07)	0.00(0.00)	0.42(0.19)	0.02(0.02)	7.19(0.77)	6.86(2.41)
L1_ag_h	0.60(0.21)	0.26(0.15)	0.94(0.12)	0.83(0.11)	7.76(1.32)	7.13(4.07)
Lasso	0.14(0.07)	0.00(0.00)	0.42(0.19)	0.02(0.02)	7.22(0.76)	6.75(2.38)
Proposed	0.50(0.19)	0.02(0.01)	0.87(0.11)	0.22(0.08)	6.21(0.68)	4.92(2.09)
$m = 80$						
Oracle Lasso	0.75(0.10)	0.04(0.03)	0.96(0.06)	0.41(0.11)	6.25(1.03)	2.97(1.27)
L1_dense	0.09(0.06)	0.00(0.00)	0.30(0.18)	0.02(0.01)	7.79(0.52)	7.98(2.60)
L1_ag_h	0.42(0.18)	0.30(0.17)	0.86(0.12)	0.82(0.15)	8.39(0.91)	7.98(2.48)
Lasso	0.09(0.06)	0.00(0.00)	0.30(0.18)	0.02(0.01)	7.76(0.50)	7.85(2.53)
Proposed	0.34(0.17)	0.01(0.01)	0.76(0.15)	0.25(0.10)	7.03(0.66)	6.63(2.21)
$m = 100$						
Oracle Lasso	0.69(0.11)	0.04(0.03)	0.93(0.06)	0.40(0.12)	6.31(1.13)	3.06(1.38)
L1_dense	0.11(0.08)	0.00(0.00)	0.33(0.19)	0.02(0.02)	7.61(0.69)	7.86(3.35)
L1_ag_h	0.41(0.13)	0.26(0.13)	0.84(0.13)	0.76(0.18)	8.42(1.04)	8.34(2.78)
Lasso	0.11(0.08)	0.00(0.00)	0.33(0.19)	0.02(0.02)	7.60(0.80)	7.91(3.40)
Proposed	0.38(0.17)	0.02(0.01)	0.78(0.13)	0.25(0.09)	6.73(0.70)	5.75(2.59)

Note: In each cell, mean (SD) based on 500 replicates.

$\boldsymbol{X}\boldsymbol{\alpha}^{(t-1)} - \boldsymbol{Z}\boldsymbol{\beta}^{(t)}$ and $\left(\widetilde{\boldsymbol{W}}^{(k)}\right)^{(t)} = \boldsymbol{W}^{(k)} \odot \left(\mathbf{1}_{n\times 1}\left(\boldsymbol{\beta}^{(t)}\right)'\right)$. Then

$$\boldsymbol{\xi}^{(t)} = \arg\min \frac{1}{2n}\left\|\breve{\boldsymbol{y}}^{(t)} - \sum_{k=1}^{q}\left(\widetilde{\boldsymbol{W}}^{(k)}\right)^{(t)}\xi_k\right\|_2^2 + \lambda\left(a\sum_{\ell=1}^{|\mathcal{T}|}\sum_{k=1}^{q} w_{k\ell}\,|\gamma_{k\ell}| + (1-a)\sum_{j=1}^{p}\sum_{k=1}^{q}\widetilde{w}_{kj}\,|\xi_{kj}|\right)$$

$$s.t.\ \xi_k = \boldsymbol{A}\boldsymbol{\gamma}_k,\ k = 1, \dots, q.$$

The algorithm is similar to that in **Step 2**.

Step 4 Compute $\boldsymbol{\alpha}^{(t)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\left(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}^{(t)} - \sum_{k=1}^{q}\mathbf{W}^{(k)}\left(\boldsymbol{\beta}^{(t)} \odot \xi_k^{(t)}\right)\right)$.

Step 5 Repeat **Steps 2–4** until convergence. In our numerical study, convergence is concluded if $\frac{|Q_n(\boldsymbol{\theta}^{(t)},\boldsymbol{\Gamma}^{(t)}) - Q_n(\boldsymbol{\theta}^{(t-1)},\boldsymbol{\Gamma}^{(t-1)})|}{|Q_n(\boldsymbol{\theta}^{(t-1)},\boldsymbol{\Gamma}^{(t-1)})|} < 10^{-4}$.

TABLE 4 Simulation Scenario 3

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	1.00(0.01)	0.06(0.05)	1.00(0.00)	0.70(0.14)	3.11(0.98)	0.36(0.24)
L1_dense	0.16(0.07)	0.00(0.00)	0.43(0.17)	0.02(0.01)	7.03(1.02)	7.92(3.93)
L1_ag_h	0.91(0.09)	0.09(0.04)	1.00(0.04)	0.67(0.17)	5.73(1.27)	3.51(1.59)
Lasso	0.16(0.07)	0.00(0.00)	0.43(0.17)	0.02(0.01)	6.95(0.99)	7.67(3.91)
Proposed	0.68(0.21)	0.01(0.01)	0.94(0.09)	0.18(0.07)	5.24(1.09)	3.49(2.02)
$m = 40$						
Oracle Lasso	0.95(0.06)	0.05(0.05)	0.99(0.03)	0.59(0.12)	4.65(1.38)	1.29(0.77)
L1_dense	0.14(0.06)	0.00(0.00)	0.39(0.15)	0.02(0.01)	7.69(0.92)	8.36(3.42)
L1_ag_h	0.79(0.20)	0.23(0.10)	0.99(0.06)	0.81(0.10)	7.03(1.23)	6.18(2.52)
Lasso	0.14(0.06)	0.00(0.00)	0.39(0.15)	0.02(0.01)	7.66(1.15)	8.55(5.08)
Proposed	0.55(0.21)	0.02(0.01)	0.87(0.14)	0.20(0.08)	6.25(0.85)	5.00(1.74)
$m = 60$						
Oracle Lasso	0.84(0.10)	0.03(0.02)	0.96(0.06)	0.50(0.10)	5.18(0.89)	1.89(1.00)
L1_dense	0.14(0.06)	0.00(0.00)	0.38(0.16)	0.02(0.01)	7.14(0.64)	6.73(2.76)
L1_ag_h	0.57(0.17)	0.24(0.13)	0.96(0.11)	0.83(0.14)	7.16(0.85)	6.54(3.20)
Lasso	0.14(0.06)	0.00(0.00)	0.38(0.16)	0.02(0.01)	7.13(0.68)	6.70(2.78)
Proposed	0.48(0.18)	0.02(0.01)	0.85(0.12)	0.23(0.09)	5.98(0.83)	4.82(2.10)
$m = 80$						
Oracle Lasso	0.77(0.09)	0.03(0.02)	0.96(0.05)	0.42(0.10)	6.65(1.16)	3.58(2.22)
L1_dense	0.11(0.06)	0.00(0.00)	0.30(0.16)	0.02(0.01)	7.93(0.68)	9.09(3.69)
L1_ag_h	0.48(0.16)	0.38(0.18)	0.87(0.13)	0.81(0.17)	8.90(1.11)	10.33(5.29)
Lasso	0.11(0.06)	0.00(0.00)	0.30(0.16)	0.02(0.01)	7.99(0.89)	8.87(3.45)
Proposed	0.37(0.16)	0.02(0.01)	0.78(0.12)	0.25(0.09)	6.97(0.71)	6.99(3.32)
$m = 100$						
Oracle Lasso	0.68(0.10)	0.04(0.02)	0.91(0.08)	0.41(0.10)	6.30(0.98)	3.18(1.38)
L1_dense	0.11(0.08)	0.00(0.00)	0.32(0.20)	0.02(0.02)	7.72(0.66)	8.29(3.28)
L1_ag_h	0.44(0.13)	0.30(0.14)	0.89(0.12)	0.81(0.17)	8.35(1.21)	9.28(5.37)
Lasso	0.11(0.08)	0.00(0.00)	0.32(0.20)	0.02(0.02)	7.66(0.63)	8.00(3.06)
Proposed	0.36(0.18)	0.02(0.01)	0.76(0.15)	0.25(0.10)	6.79(0.64)	6.36(2.63)

Note: In each cell, mean (SD) based on 500 replicates.

The proposed objective function is bounded from below. In each iteration step, its value decreases. As such, convergence is guaranteed. It is satisfactorily achieved with a moderate number of iterations in all of our numerical studies. The tuning parameters (λ, a) are chosen using a modified BIC criterion with the degree of freedom defined as the effective number of parameters [39]. With simple updates, the proposed computational algorithm is affordable. For one simulation replicate (details described below), computation can be accomplished within 3 min on a regular desktop. To facilitate numerical analysis within and beyond this study, we have developed R code and made it publicly available at http://github.com/shuanggema/.

3 | SIMULATION

We consider a total of six scenarios to examine the dependence of performance on distributional properties,

TABLE 5 Simulation Scenario 4

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	0.99(0.02)	0.05(0.04)	1.00(0.00)	0.69(0.14)	3.41(1.90)	1.02(1.19)
L1_dense	0.17(0.07)	0.01(0.00)	0.41(0.10)	0.02(0.01)	8.57(2.02)	9.95(4.83)
L1_ag_h	0.91(0.13)	0.10(0.05)	0.98(0.12)	0.65(0.16)	6.99(1.89)	6.14(3.6)
Lasso	0.17(0.07)	0.01(0.00)	0.41(0.10)	0.02(0.01)	8.46(1.86)	9.63(4.38)
Proposed	0.64(0.16)	0.01(0.01)	0.90(0.12)	0.19(0.06)	6.86(2.82)	5.83(4.37)
$m = 40$						
Oracle Lasso	0.90(0.09)	0.06(0.04)	0.99(0.03)	0.65(0.10)	6.12(1.96)	3.57(3.17)
L1_dense	0.16(0.06)	0.01(0.00)	0.38(0.13)	0.02(0.01)	8.92(1.75)	12.11(6.68)
L1_ag_h	0.88(0.12)	0.29(0.08)	0.99(0.06)	0.83(0.11)	7.43(1.90)	11.15(6.73)
Lasso	0.16(0.06)	0.01(0.00)	0.38(0.13)	0.02(0.01)	9.07(1.89)	12.71(7.06)
Proposed	0.55(0.16)	0.02(0.01)	0.89(0.12)	0.22(0.08)	7.04(1.52)	9.00(7.59)
$m = 60$						
Oracle Lasso	0.82(0.09)	0.04(0.03)	0.97(0.06)	0.54(0.09)	6.00(1.46)	4.43(3.64)
L1_dense	0.15(0.06)	0.01(0.00)	0.39(0.10)	0.03(0.01)	7.66(0.80)	9.24(3.39)
L1_ag_h	0.72(0.18)	0.35(0.10)	0.98(0.06)	0.84(0.12)	7.48(1.00)	9.96(4.20)
Lasso	0.15(0.06)	0.01(0.00)	0.39(0.1)	0.03(0.01)	7.86(1.23)	9.59(4.18)
Proposed	0.53(0.12)	0.02(0.01)	0.90(0.09)	0.22(0.09)	6.61(0.92)	6.69(3.21)
$m = 80$						
Oracle Lasso	0.69(0.11)	0.05(0.03)	0.94(0.08)	0.46(0.11)	7.32(1.46)	6.43(2.90)
L1_dense	0.14(0.07)	0.01(0.00)	0.35(0.14)	0.03(0.01)	8.80(1.36)	12.12(5.52)
L1_ag_h	0.60(0.16)	0.50(0.16)	0.93(0.1)	0.90(0.10)	8.78(1.23)	13.18(6.26)
Lasso	0.14(0.07)	0.01(0.00)	0.35(0.14)	0.03(0.01)	8.67(1.11)	11.97(5.70)
Proposed	0.46(0.14)	0.03(0.01)	0.84(0.10)	0.24(0.09)	7.33(0.96)	9.55(4.85)
$m = 100$						
Oracle Lasso	0.63(0.11)	0.04(0.02)	0.89(0.08)	0.43(0.08)	7.50(1.53)	5.79(3.04)
L1_dense	0.13(0.05)	0.01(0.00)	0.36(0.12)	0.03(0.02)	8.68(1.58)	10.78(5.38)
L1_ag_h	0.56(0.14)	0.38(0.15)	0.94(0.10)	0.88(0.10)	9.01(1.72)	14.04(10.54)
Lasso	0.13(0.05)	0.01(0.00)	0.36(0.12)	0.03(0.02)	8.69(1.39)	10.26(3.83)
Proposed	0.41(0.13)	0.03(0.01)	0.80(0.12)	0.26(0.09)	7.54(1.22)	8.53(3.28)

Note: In each cell, mean (SD) based on 500 replicates.

especially correlation. To mimic data analyzed in the next section, we simulate G variables with properties similar to SNPs. Under Scenario 1, the G variables are independently generated from a Poisson(0.02) distribution and truncated at 2 if needed. The five E variables are generated from a Bernoulli distribution with probability of success 0.7. Under Scenarios 2–6, we first generate p continuous variables from multivariate normal distributions, and then dichotomize the continuous variables at the 0.98 and 0.995 percentiles to generate the three-level G measurements. The multivariate normal distributions have marginal means 0 and variances 1. Two correlation structures with different parameters are considered and referred to as Band1, Band2, AR(0.3), AR(0.5), and AR(0.8). Here, Band1 and Band2, the two banded correlation structures, have correlation coefficients of variables j and k as $0.3^{|j-k|}I(j-k| < 2)$ and $0.3^{|j-k|}I(j-k| = 2) + 0.5^{|j-k|}I(j-k| < 2)$, respectively. The three auto-regressive structures correspond to weak, moderate, and strong correlations, respectively. We note that such correlation

TABLE 6 Simulation Scenario 5

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	1.00(0.01)	0.05(0.03)	1.00(0.00)	0.68(0.14)	2.76(0.82)	0.33(0.32)
L1_dense	0.11(0.09)	0.00(0.00)	0.34(0.22)	0.02(0.01)	6.67(0.97)	5.54(2.47)
L1_ag_h	0.91(0.07)	0.07(0.03)	1.00(0.00)	0.59(0.18)	6.28(0.82)	3.73(0.84)
Lasso	0.11(0.09)	0.00(0.00)	0.34(0.22)	0.02(0.01)	6.64(1.03)	5.57(2.45)
Proposed	0.65(0.24)	0.01(0.01)	0.95(0.09)	0.18(0.11)	4.73(0.82)	2.49(1.13)
$m = 40$						
Oracle Lasso	0.93(0.08)	0.05(0.03)	0.98(0.05)	0.55(0.13)	4.26(0.81)	1.07(0.76)
L1_dense	0.10(0.08)	0.00(0.00)	0.29(0.19)	0.02(0.02)	7.3(0.7)	6.89(2.40)
L1_ag_h	0.79(0.13)	0.22(0.07)	0.99(0.06)	0.77(0.12)	6.74(1.00)	4.95(1.61)
Lasso	0.10(0.08)	0.00(0.00)	0.29(0.19)	0.02(0.02)	7.21(0.67)	6.79(2.32)
Proposed	0.54(0.24)	0.02(0.01)	0.88(0.14)	0.25(0.12)	5.68(0.84)	4.04(1.82)
$m = 60$						
Oracle Lasso	0.89(0.08)	0.04(0.02)	0.99(0.04)	0.49(0.13)	5.17(1.26)	1.95(1.39)
L1_dense	0.14(0.07)	0.00(0.00)	0.41(0.17)	0.02(0.01)	7.15(0.73)	7.03(3.15)
L1_ag_h	0.57(0.18)	0.24(0.12)	0.92(0.13)	0.78(0.12)	7.45(1.15)	6.36(2.38)
Lasso	0.14(0.07)	0.00(0.00)	0.41(0.17)	0.02(0.01)	7.11(0.68)	6.84(3.00)
Proposed	0.57(0.16)	0.02(0.01)	0.87(0.09)	0.29(0.11)	5.99(0.64)	4.77(2.14)
$m = 80$						
Oracle Lasso	0.75(0.09)	0.03(0.01)	0.96(0.07)	0.4(0.09)	6.57(1.37)	2.85(1.36)
L1_dense	0.09(0.08)	0.00(0.00)	0.28(0.18)	0.02(0.02)	7.79(0.69)	7.64(2.67)
L1_ag_h	0.42(0.19)	0.29(0.18)	0.9(0.13)	0.85(0.12)	8.62(1.09)	8.43(2.50)
Lasso	0.09(0.08)	0.00(0.00)	0.28(0.18)	0.02(0.02)	7.74(0.70)	7.41(2.61)
Proposed	0.39(0.20)	0.02(0.01)	0.83(0.14)	0.27(0.11)	6.87(0.73)	5.87(2.35)
$m = 100$						
Oracle Lasso	0.69(0.10)	0.04(0.03)	0.92(0.06)	0.41(0.11)	6.20(1.20)	3.20(1.56)
L1_dense	0.10(0.07)	0.00(0.00)	0.30(0.18)	0.02(0.02)	7.56(0.65)	7.48(2.09)
L1_ag_h	0.40(0.15)	0.24(0.13)	0.88(0.12)	0.82(0.13)	8.41(1.25)	8.25(2.52)
Lasso	0.10(0.07)	0.00(0.00)	0.30(0.18)	0.02(0.02)	7.52(0.65)	7.43(2.20)
Proposed	0.38(0.17)	0.02(0.01)	0.81(0.11)	0.28(0.10)	6.84(0.69)	6.20(2.21)

Note: In each cell, mean (SD) based on 500 replicates.

structures have been considered in quite a few 3 studies. For the E variables, we first generate five continuous variables from a multivariate normal distribution with marginal means 0, marginal variances 1, and an AR(0.3) correlation structure. Then, two variables are dichotomized at 0 to create two binary variables, leading to three continuous and two binary E variables. Under all scenarios, the G variables have low (MAF 1%–5%) and very low (MAF < 1%) frequencies. In practical data analysis, more common variants are expected. Here, we focus on rare variants whose effects are more difficult to quantify. The proposed approach is expected to have better performance for variants that are less rare.

The nonzero main effects and interactions are generated as follows. For the SNPs, based on their adjacency (correlation) information, the true tree structure $\mathcal{T}$ of the p leaves is shown in Figure 6. These leaves form m aggregating sets (clusters) with varying sizes, which are indexed by $\boldsymbol{B}^{*}$. This construction is similar to that in [28]. To generate the main G and G-E interaction effects, we first

TABLE 7 Simulation Scenario 6

	I:TPR	I:FPR	M:TPR	M:FPR	RSSE	PMSE
$m = 20$						
Oracle Lasso	1.00(0.02)	0.05(0.04)	1.00(0.00)	0.67(0.15)	2.96(1.00)	0.36(0.22)
L1_dense	0.16(0.07)	0.01(0.00)	0.48(0.15)	0.02(0.01)	7.00(1.34)	7.57(3.81)
L1_ag_h	0.91(0.09)	0.08(0.04)	1.00(0.00)	0.63(0.18)	5.31(0.77)	3.91(0.91)
Lasso	0.16(0.07)	0.01(0.00)	0.48(0.15)	0.02(0.01)	7.02(1.38)	7.43(3.82)
Proposed	0.65(0.16)	0.01(0.01)	0.95(0.06)	0.13(0.07)	4.95(1.01)	2.95(1.90)
$m = 40$						
Oracle Lasso	0.96(0.05)	0.05(0.03)	0.99(0.04)	0.60(0.13)	4.44(1.22)	1.29(0.82)
L1_dense	0.15(0.07)	0.00(0.00)	0.40(0.16)	0.02(0.01)	7.52(0.68)	8.04(2.91)
L1_ag_h	0.83(0.14)	0.25(0.11)	0.99(0.04)	0.78(0.14)	7.02(1.16)	6.28(2.71)
Lasso	0.15(0.07)	0.00(0.00)	0.40(0.16)	0.02(0.01)	7.66(1.04)	8.15(3.24)
Proposed	0.49(0.14)	0.01(0.01)	0.87(0.11)	0.16(0.08)	6.24(0.90)	4.85(2.04)
$m = 60$						
Oracle Lasso	0.86(0.12)	0.04(0.03)	0.98(0.06)	0.52(0.13)	5.61(1.53)	2.37(1.75)
L1_dense	0.13(0.07)	0.00(0.00)	0.39(0.18)	0.02(0.01)	7.16(0.65)	7.82(3.50)
L1_ag_h	0.60(0.21)	0.27(0.13)	0.95(0.12)	0.84(0.13)	7.52(0.97)	8.22(3.82)
Lasso	0.13(0.07)	0.00(0.00)	0.39(0.18)	0.02(0.01)	7.18(0.67)	7.79(3.45)
Proposed	0.46(0.17)	0.01(0.01)	0.85(0.11)	0.18(0.07)	6.09(0.89)	5.57(3.10)
$m = 80$						
Oracle Lasso	0.77(0.11)	0.04(0.02)	0.96(0.07)	0.42(0.10)	7.01(1.82)	4.39(3.22)
L1_dense	0.11(0.07)	0.01(0.00)	0.37(0.17)	0.02(0.02)	8.08(0.70)	9.37(3.90)
L1_ag_h	0.48(0.15)	0.38(0.17)	0.89(0.14)	0.84(0.17)	8.95(1.29)	9.81(3.44)
Lasso	0.11(0.07)	0.01(0.00)	0.37(0.17)	0.02(0.02)	8.05(0.76)	9.32(3.96)
Proposed	0.40(0.17)	0.02(0.01)	0.80(0.14)	0.19(0.08)	7.08(0.70)	7.12(3.37)
$m = 100$						
Oracle Lasso	0.71(0.10)	0.04(0.02)	0.91(0.06)	0.44(0.12)	6.66(1.29)	3.85(2.50)
L1_dense	0.10(0.07)	0.00(0.00)	0.28(0.19)	0.02(0.02)	7.67(0.57)	7.87(3.17)
L1_ag_h	0.44(0.14)	0.27(0.16)	0.89(0.12)	0.81(0.16)	8.25(1.06)	8.14(2.77)
Lasso	0.10(0.07)	0.00(0.00)	0.28(0.19)	0.02(0.02)	7.58(0.59)	7.73(3.14)
Proposed	0.31(0.17)	0.01(0.01)	0.74(0.15)	0.18(0.09)	6.93(0.69)	6.08(2.59)

Note: In each cell, mean (SD) based on 500 replicates.

generate a matrix $\boldsymbol{A}_{B^*} \in \mathbb{R}^{p\times m}$ with binary components $\boldsymbol{A}_{B^*jl} := 1_{\{j\in l\, cluster\}}$. Then, the coefficient vectors are generated via these aggregating sets as: $\boldsymbol{\beta}^* = \boldsymbol{A}_{\boldsymbol{B}^*}\widetilde{\boldsymbol{\beta}}^*$, $\boldsymbol{\xi}_k^* = \boldsymbol{A}_{\boldsymbol{B}^*}\widetilde{\boldsymbol{\xi}}^*_{(k)}$, where $\widetilde{\boldsymbol{\beta}}^*, \widetilde{\boldsymbol{\xi}}^*_k \in \mathbb{R}^m$ have $m \times s$ elements zeroed out, and the remaining elements are independently drawn from a Uniform(0.8, 1.5) distribution. Here, s controls the true level of sparsity. For the main E effects, their nonzero coefficients α_k^*'s are generated from Uniform (0.8,1.2). The response $\mathbf{y} \in \mathbb{R}^n$ values are simulated from (1) with independent Gaussian errors and variances $\sigma^2 = \sum_{i=1}^n \left(\boldsymbol{x}_{i'}\boldsymbol{\alpha}^* + \boldsymbol{z}_{i'}\boldsymbol{\beta}^* + \sum_{k=1}^q x_{ik}\boldsymbol{z}_{i'}\left(\boldsymbol{\beta}^* \odot \boldsymbol{\xi}_k^*\right)\right)^2/(5n)$. The above data generation satisfies the "main effects, interactions" hierarchical structure and aggregative effects of the nearby G features.

We set $n = 200$, $p = 200$, $q = 5$, $s = 0.4$. It is noted that the combined number of unknown parameters is much larger than the sample size. We consider a sequence of m values up to $p/2$. The proposed approach is applied based

TABLE 8 NFBC1996 data analysis: numbers of main G effects and interactions identified by different approaches (diagonal) and their overlaps (off-diagonal)

		Lasso	L1_dense	L1_a_h	Proposed
Main G effects	Lasso	31	22	20	27
	L1_dense	–	29	20	24
	L1_a_h	–	–	23	21
	Proposed	–	–	–	34
Interactions	Lasso	84	50	34	68
	L1_dense	–	65	33	54
	L1_a_h	–	–	40	36
	Proposed	–	–	–	107

on the tree $\mathcal{T}$. To gauge its performance, we further consider the following alternatives. The first is Oracle Lasso, under which the true aggregation structure $\boldsymbol{XA}_{B^*}$ is known, and Lasso (which is the proposed approach with $a = 0$) is applied for regularized selection and estimation. The second is L1_dense, which applies Lasso after first discarding all features with MAF < 1%. It represents approaches that focus on dense features. The third is L1_ag_h, which applies Lasso to features aggregated in the same clusters after the tree is cut at a certain height. This approach conducts feature aggregation based on $\mathcal{T}$, however, in a relatively “brutal” manner. It represents approaches that first conduct clustering, then group features in an unsupervised way, and finally conduct modeling and estimation based on the postaggregation features. Lastly, we also consider the Lasso approach as for the toy example. For each setting, we simulate 500 replicates.

We evaluate identification performance using the true-positive rate (TPR) and false-positive rate (FPR) for main G effects (M:TPR and M:FPR) and interactions (I:TPR and I:FPR) separately. Here, it is noted that the proposed approach conducts penalized variable selection as opposed to hypothesis testing. As such, it does not have a direct false discovery rate control. Nevertheless, TPR and FPR are still highly informative performance measures. We further evaluate estimation performance using the root sum of squared errors (RSSE) defined as $\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2$. In addition, we calculate the best mean-squared estimation error as a function of m, that is, $\min_\Lambda \left\|\hat{\boldsymbol{\theta}}(\Lambda) - \boldsymbol{\theta}^*\right\|_2^2/(p + q + pq)$, where Λ represents a method’s tuning parameter(s). For evaluating prediction performance, for each simulation replicate, we generate an independent testing dataset with size 200 and compute the prediction mean squared error (PMSE).

Summary results are presented in Tables 1 and 3–7 (Data S1). Across all of the simulation scenarios, the proposed approach is observed to perform similarly to the Oracle Lasso and has superior performance compared to the other alternatives. Specifically, it can more accurately identify both the true main effects and interactions while having a small number of false positives. For example, in Table 1, under Scenario 1 and $m = 40$, the proposed approach has (M:TPR, M:FPR, I:TPR, I:FPR) = (0.89, 0.01, 1, 0.09), compared to (0.61, 0.05, 0.77, 0.24) for L1_dense, (0.75, 0.15, 0.99, 0.84) for L1_ag_h, and (0.74, 0.03, 0.94, 0.19) for Lasso. This result and those alike can establish the effectiveness of accommodating rare features (when compared to L1_dense), data aggregation (when compared to Lasso), and more effective data aggregation (when compared to L1_ag_h). We also observe the superiority of the proposed approach in estimation. For example, in Table 6 (Data S1), under Scenario 5 and $m = 60$, the proposed approach has RSSE = 5.99, compared to 7.15(L1_dense), 7.45(L1_ag_h), and 7.11(Lasso). This superiority is further shown in Figure 7 (Data S1). In particular, under Scenario 1 with $\boldsymbol{Z}$ simulated from a Poisson(0.02) distribution, the proposed approach performs nearly as well as the oracle. As m increases, borrowing strength from neighbors decreases, and so estimation performance deteriorates. L1_dense performs very similarly to Lasso. Last but not least, the proposed approach also has satisfactory prediction performance. For example, in Table 1, under Scenario 1, the PMSEs are 3.12 (L1_dense), 3.69 (L1_ag_h), 2.64 (Lasso), and 0.82 (proposed) (Figure 8).

4 | DATA ANALYSIS

To demonstrate the practical applicability of the proposed approach, we analyze the individual-level data from the NFBC (Northern Finland Birth Cohorts) study [40]. This study was conducted to very broadly examine risk factors involved in preterm birth and intrauterine growth retardation, as well as the consequences of these early adverse

TABLE 9 Analysis of the NFBC1996 data using the proposed approach under an alternative marginal screening: identified main effects and interactions

SNP	Main effect	Gender	CRP	Glucose	TC	HDL
rs12222221	−0.253	0.001	−0.001	0.031	−0.002	—
rs4585672	−0.270	−0.001	0.003	0.001	0.006	—
rs6743144	0.230	0.008	0.001	0.028	−0.010	—
rs12548107	0.285	0.009	0.002	−0.021	0.037	—
rs1470829	−0.329	−0.008	—	0.016	0.007	—
rs6127943	0.216	0.007	−0.002	0.002	0.006	—
rs4735825	0.291	0.009	−0.004	−0.007	−0.028	−0.001
rs1882681	−0.462	−0.007	0.001	0.061	−0.030	−0.001
rs4870024	0.300	0.008	—	−0.008	0.039	—
rs937557	−0.236	0.001	—	−0.019	−0.018	—
rs177195	−0.245	0.011	—	−0.008	0.006	—
rs17552964	0.251	0.005	—	—	0.016	0.001
rs3771327	−0.292	0.012	0.001	−0.027	0.024	—
rs2833383	−0.421	0.007	−0.004	−0.005	0.016	−0.001
rs4077636	−0.285	0.002	0.003	0.035	−0.016	—
rs4512398	−0.244	0.003	0.004	−0.004	−0.006	—
rs1025404	−0.288	0.008	−0.005	−0.002	−0.024	—
rs2961725	−0.244	−0.003	0.001	−0.011	−0.011	−0.001
rs1934127	−0.272	0.002	0.002	−0.030	0.028	−0.001
rs1293770	0.223	0.009	—	0.023	0.008	—
rs1407593	0.239	0.010	0.003	−0.007	−0.010	—
rs10906021	0.225	0.004	−0.003	0.013	0.031	0.001
rs12475063	−0.231	0.004	0.004	0.032	0.032	0.001
rs2306970	−0.226	0.008	0.001	−0.015	−0.047	−0.001
rs2868975	0.213	0.008	0.001	0.010	−0.014	—
rs6737978	−0.305	0.005	−0.001	0.017	−0.043	—
rs881204	−0.337	−0.006	−0.001	−0.009	−0.005	—
rs1886434	−0.244	0.009	−0.001	0.014	0.034	0.001
rs7962035	0.282	0.005	0.002	0.054	0.024	0.001
rs11854565	0.354	0.005	0.001	−0.003	0.033	−0.001
rs7209713	−0.307	0.002	−0.001	−0.027	0.010	—
rs2016327	−0.259	0.003	−0.002	−0.003	0.009	0.001
rs4422244	0.225	0.007	−0.002	−0.034	0.002	0.002
rs11812486	−0.300	0.012	0.002	−0.006	0.075	—
rs1345981	−0.293	0.001	0.001	0.023	−0.009	—
rs6122682	−0.233	—	−0.002	−0.003	—	0.001
rs1920083	−0.234	—	0.001	0.008	0.012	—
rs987648	−0.292	—	0.002	−0.016	0.016	−0.001
rs7989689	−0.264	—	0.001	−0.008	0.005	0.001
rs3887251	−0.218	—	—	0.017	−0.009	—
rs1202657	−0.066	—	—	—	—	—

TABLE 10 NFBC1996 data analysis under an alternative marginal screening: numbers of main G effects and interactions identified by different approaches (diagonal) and their overlaps (off-diagonal)

		Lasso	L1_dense	L1_a_h	Proposed
Main G effects	Lasso	47	17	17	23
	L1_dense	–	24	10	10
	L1_a_h	–	–	24	12
	Proposed	–	–	–	41
Interactions	Lasso	102	25	14	28
	L1_dense	–	34	15	16
	L1_a_h	–	–	26	12
	Proposed	–	–	–	163

outcomes on subsequent morbidity. The data collected from Northern Finland forms a unique resource, allowing to study the emergence of diseases, which can be caused by genetic, biological, social, and behavioral risk factors. The NFBC1966 dataset contains 10 traits and 364,590 SNPs for 5402 individuals whose expected year of birth is 1966. In our data analysis, the response variable of interest is BMI (body mass index), which is an important phenotype and critical biomarker for many illness conditions. For the G factors, we consider SNPs. And for the E factors, we consider gender, C-reactive protein (CRP), glucose, total cholesterol (TC), and high-density cholesterol (HDL). We note that these factors are not environmental in the narrow sense. Rather, they are clinical biomarkers. In the recent literature [41], the interactions between clinical/demographic variables and G factors analyzed under the G-E interaction analysis framework and have attracted strong interest. Such analysis can reveal the interplay between clinical/demographic variables and G factors on disease outcomes and other biomarkers.

Data processing is first conducted, following similar procedures as in published studies [42]. In particular, we exclude individuals that have discrepancies between reported sex and sex determined using the X chromosome. Further, individuals with missingness in the response and E variables or with genotype missing call-rates $> 5\%$ are excluded. A SNP is removed from analysis if its MAF $< 1\%$ or missing call-rate $> 1\%$, or it fails the Hardy–Weinberg equilibrium test. The SNP data quality control is conducted using PLINK [43]. These processing procedures lead to data on 5123 individuals and 319,147 SNPs. In principle, the proposed approach can be directly applied. Considering the limited sample size, we further conduct a prescreening to improve estimation. In particular, we split the data into two parts with sizes 2:3. Marginal regression analysis is conducted in smaller part, under which one SNP is analyzed at a time using regression. The 5000 SNPs with the smallest marginal p-values are selected for downstream analysis. In Figure 9 (for the rare SNPs) and 8 (for all of the SNPs) in the Data S1, we examine the LD structures for the SNPs that have passed screening. It is observed that a relatively small number of rare SNPs have high LD values, which may limit the power of information borrowing (Figures 10 and 11).

The larger part of the data is analyzed using the proposed and alternative approaches. It is recognized that this may lead to a smaller sample size and loss of power, compared to the analysis of the whole data. However, separating the screening and analysis data can lead to more objective analysis and comparison and has been adopted in many published studies. The tree $\mathcal{T}$ is constructed using hierarchical clustering and the physical locations of SNPs and shown in Figure 4 (Data S1). For all approaches, tuning parameters are selected using the modified BIC criterion [39]. The proposed approach identifies 34 main G effects and 107 interactions. The detailed estimation results are provided in Table 2. The summary comparison results are presented in Table 8 (Data S1). It is observed that the proposed approach identifies more effects. This is sensible as, with fusion, it can pull some SNPs that otherwise may not be identified. It is also noted that, with the complexity of BMI, more main G effects and interactions (than identified by the proposed and alternative methods) may be involved. In general, penalization approaches, including the proposed, can only identify the relatively strong effects. Alternative techniques will be needed for the identification of weaker effects (Tables 9 to 14).

The alternative methods miss the rare SNP rs6488338, which has MAF = 0.046 and belongs to gene CD163. Published studies have suggested that gene CD163 is associated with pregravid obesity. The adipose tissue expression of gene CD163 is elevated in obesity and type 2 diabetes, and this gene is a novel immune marker for metabolic inflammation [44]. Many other findings are also biologically meaningful. For example, LINGO2 is a protein coding

TABLE 11 Analysis of the NFBC1996 data with Glucose as the response variable using the proposed approach: identified main effects and interactions

SNP	Main effect	Gender	CRP	TC	HDL	LDL
rs10504197	0.023	0.041	−0.017	—	−0.020	—
rs1551547	0.022	0.061	—	—	−0.011	—
rs7460495	−0.031	−0.033	−0.032	—	—	0.010
rs2290526	−0.030	−0.064	—	—	—	—
rs10108007	0.017	0.026	—	—	−0.005	—
rs11998308	0.023	0.049	—	—	—	—
rs9692725	−0.027	−0.052	—	—	—	0.033
rs10091115	0.032	0.042	0.023	0.031	—	—
rs979843	−0.026	−0.047	—	—	−0.043	−0.020
rs7836768	0.025	0.024	0.050	—	—	—
rs9643401	−0.044	−0.058	−0.017	—	0.009	−0.008
rs1896135	0.019	0.045	—	—	—	—
rs2380540	−0.032	−0.033	0.056	—	—	0.028
rs2380607	0.018	0.020	—	—	0.005	—
rs12678469	0.015	0.018	—	0.003	—	—
rs1383978	−0.021	−0.032	—	0.007	—	—
rs1031177	0.018	0.030	0.005	—	—	−0.007
rs959974	−0.023	−0.033	—	—	—	—
rs12334848	0.028	0.027	−0.016	−0.053	—	−0.036
rs2941456	0.013	0.018	—	—	—	—
rs998731	−0.022	−0.016	—	—	—	—
rs6473219	−0.018	−0.018	—	—	—	—
rs272610	0.014	0.013	—	—	—	—
rs4961056	−0.024	−0.048	—	0.009	0.004	—
rs7818882	0.039	0.017	−0.019	−0.007	0.044	−0.083
rs1507883	−0.024	−0.021	0.024	—	—	—
rs1382101	−0.016	−0.016	—	—	—	—
rs1487796	0.038	0.083	0.076	—	—	—
rs13262606	0.021	0.038	−0.009	—	—	—
rs551496	0.015	0.013	—	—	—	—
rs1374633	−0.031	−0.072	0.028	—	-	0.039
rs2890805	0.031	0.053	0.022	−0.032	0.062	—
rs3104966	−0.016	−0.017	—	—	—	—
rs4263730	−0.024	−0.006	−0.038	—	0.032	—
rs2587000	−0.036	−0.045	0.059	0.037	0.017	—
rs2513399	−0.036	−0.036	−0.118	−0.033	—	—
rs2513402	0.038	—	−0.071	0.036	0.010	0.023
rs998980	0.002	—	—	—	—	—

TABLE 12 NFBC1996 data analysis with Glucose as the response variable: numbers of main G effects and interactions identified by different approaches (diagonal) and their overlaps (off—diagonal)

		Lasso	L1_dense	L1_a_h	Proposed
Main G effects	Lasso	42	12	18	31
	L1_dense	–	31	21	11
	L1_a_h	–	–	33	16
	Proposed	–	–	–	38
Interactions	Lasso	55	11	15	51
	L1_dense	–	31	17	15
	L1_a_h	–	–	30	18
	Proposed	–	–	–	86

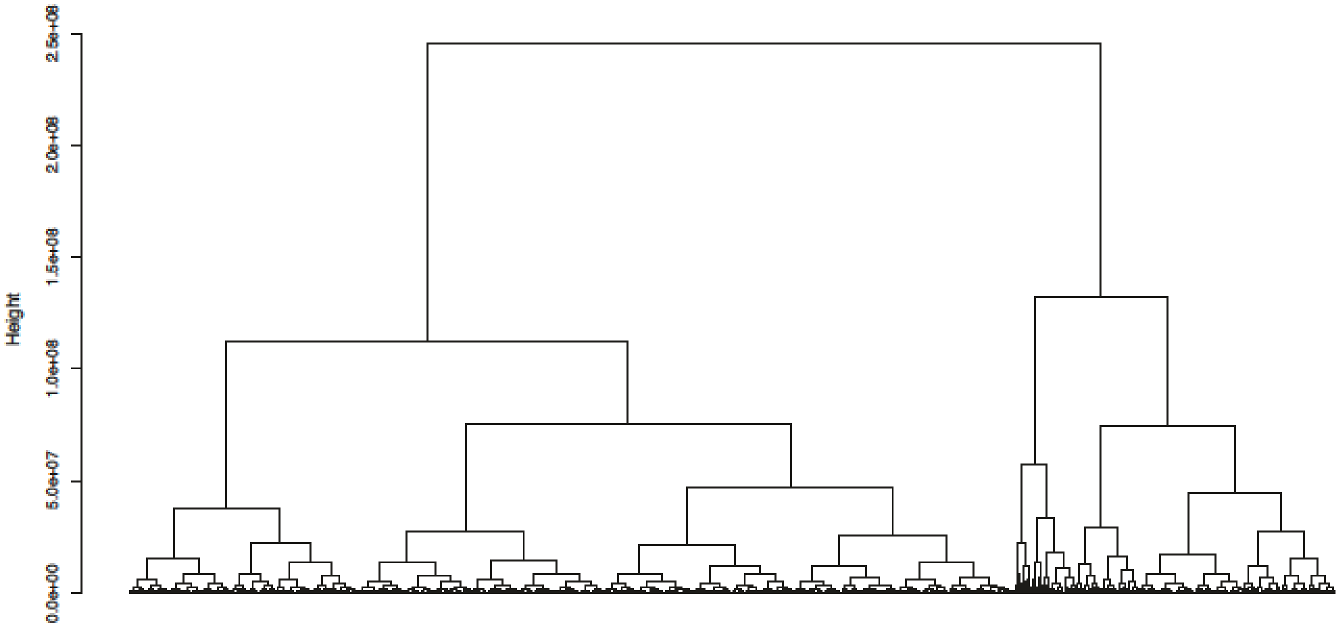

FIGURE 12 NFBC1996 data analysis: tree $\mathcal{T}$ of single nucleotide polymorphisms' (SNPs)' physical positions (post screening), with insulin as the response variable

gene that is an important part of the cellular membrane. It has been linked to obesity in GWAS [45]. Additionally, the variants of LINGO2 have been linked to essential tremor in Parkinson's disease [46]. Thus, its linkage to obesity via interactions with DA signaling seems possible [47]. RBFOX1 is an important RNA-binding protein mediating the incorporation of microexons into many transcripts associated with neurological patterning and tissue development. Its association with obesity has been suggested [48]. Gene *ANO2* has been suggested as playing a role in the pathophysiology of childhood obesity [49]. Studies [50] have revealed that ANO2 is a Ca^{2+}−activated chloride channel in vagal afferents of nodose neurons and a major determinant of CCK-induced satiety, body weight control, and energy expenditure, making it a potential therapeutic target in obesity. TNFRSF1 genotypes have been identified as significantly associated with sTNFR1 plasma levels in obese women [51]. It has been suggested that TNFRSF1A polymorphism can have functional significance in obesity. In addition, genes *TAF4B*, *PCSK5*, *LDLRAD4*, and *TENM4* have also been associated with obesity by GWAS [52].

With real data, it is hard to objectively evaluate the accuracy of identification. To provide further insight and "indirect" support, we apply a resampling-based approach and evaluate prediction performance and stability. Specifically, the dataset is randomly divided into a training and testing set, with sizes 9:1. The model/parameters are estimated using only the training set and then used to

TABLE 13 Analysis of the NFBC1996 data with insulin as the response variable using the proposed approach: identified main effects and interactions

SNP	Main effect	Gender	CRP	TC	HDL	LDL
rs2220326	0.021	0.021	—	—	—	—
rs9598811	−0.023	−0.018	—	——	—	—
rs359379	0.049	0.092	—	−0.020	0.002	−0.055
rs359361	−0.035	−0.032	—	—	—	0.056
rs396985	0.060	0.120	—	—	−0.111	—
rs967958	−0.041	−0.051	0.027	—	—	−0.023
rs1928955	0.020	0.017	0.000	—	−0.013	—
rs1036995	−0.037	0.003	0.072	0.018	−0.026	0.015
rs7324254	−0.019	−0.008	—	—	—	—
rs11148750	−0.023	−0.019	—	—	−0.003	0.003
rs1384607	−0.071	−0.182	−0.009	0.064	0.064	—
rs9541273	−0.029	−0.041	—	—	—	—
rs9571979	0.035	0.084	—	—	—	—
rs7993187	0.055	0.096	—	—	−0.044	—
rs1341525	−0.040	−0.006	—	0.118	—	0.006
rs9572442	−0.025	−0.031	—	—	—	—
rs9542369	0.040	0.087	—	—	—	—
rs1114564	0.040	0.025	—	−0.040	—	—
rs9599903	0.031	0.063	—	—	—	—
rs936457	−0.046	−0.016	—	−0.067	—	−0.065
rs7333339	−0.047	−0.052	0.093	—	—	—
rs287553	−0.028	−0.066	—	—	—	—
rs1324061	0.041	0.045	—	—	−0.034	0.033
rs4597197	0.049	0.065	0.048	—	—	—
rs1505149	0.056	0.121	—	—	0.041	—
rs9574389	−0.021	−0.040	—	—	—	-
rs2274554	0.043	0.054	−0.016	—	−0.018	—
rs1998452	−0.023	−0.014	—	—	—	—
rs7336627	0.035	0.058	—	—	—	——
rs1335852	−0.046	−0.122	—	−0.075	—	—
rs9602002	−0.019	−0.028	—	—	—	—
rs985035	0.030	0.033	—	—	—	—
rs988474	−0.089	−0.085	0.122	—	0.087	−0.117
rs1334166	−0.039	−0.091	—	—	—	—
rs7333936	0.032	0.021	—	0.002	—	0.050
rs184385	0.040	0.031	—	−0.061	−0.090	—

(Continues)

TABLE 13 (Continued)

SNP	Main effect	Gender	CRP	TC	HDL	LDL
rs12855484	−0.052	−0.018	—	−0.041	—	—
rs9516496	0.022	0.040	—	—	—	—
rs7321486	−0.042	−0.023	−0.030	—	0.041	—
rs913427	0.029	0.044	—	—	—	−0.014
rs9556889	0.041	0.032	−0.096	—	—	—
rs679363	−0.039	−0.060	−0.002	0.041	—	—
rs1556799	−0.034	−0.045	—	—	—	—
rs1998550	−0.048	−0.060	−0.003	0.045	−0.025	0.005
rs701556	0.045	0.033	—	—	—	−0.073
rs1571513	−0.063	−0.025	−0.100	−0.083	0.020	—
rs1730649	0.054	0.067	0.031	—	0.112	−0.014
rs2067741	0.042	—	−0.085	0.040	−0.044	0.022
rs937872	0.051	—	0.216	—	0.002	—
rs9317872	−0.055	—	−0.013	—	0.093	0.020
rs9572541	0.052	—	−0.091	0.005	0.045	0.005
rs9300342	−0.020	—	−0.025	—	—	—
rs1998535	0.040	—	—	−0.031	—	−0.046
rs516872	−0.059	—	—	—	0.194	—
rs9572146	−0.011	—	—	—	—	—
rs4405440	−0.003	—	—	—	—	—

TABLE 14 NFBC1996 data analysis with insulin as the response variable: numbers of main G effects and interactions identified by different approaches (diagonal) and their overlaps (off-diagonal)

		Lasso	L1_dense	L1_a_h	Proposed
Main G effects	Lasso	50	18	13	35
	L1_dense	–	43	21	9
	L1_a_h	–	–	52	17
	Proposed	–	–	–	56
Interactions	Lasso	102	24	30	64
	L1_dense	–	90	35	16
	L1_a_h	–	–	99	29
	Proposed	–	–	–	120

make prediction for samples in the testing set, where prediction performance is evaluated using prediction mean squared error (PMSE). This procedure is repeated 1000 times. The training set estimates are also used to evaluate stability. This approach has been extensively adopted in the literature. The squared roots of the average PMSEs are 1.057 (L1_ag_h), 1.058 (L1_dense), 1.052 (Lasso), and 1.046 (Proposed). In the stability evaluation, we compute the OOI (observed occurrence index) for each effect. Briefly, the OOI is the probability of a specific effect being identified across replicates and measures the stability of finding. For the identified main G effects, the mean OOI values are 0.599 (L1_ag_h), 0.636 (L1_dense), 0.604 (Lasso), and 0.638 (Proposed). And for the identified interaction effects, the mean OOI values are 0.544 (L1_ag_h), 0.572 (L1_dense), 0.552

(Lasso), and 0.553 (Proposed). The proposed approach has competitive prediction performance and selection stability (Figure 12).

5 | ADDITIONAL ANALYSIS

To complement the above analysis, additional analysis is conducted and reported in the Data S1. In the first set of analysis, we repeat the above analysis under a different marginal screening approach, with which we select a block of consecutive SNPs. In the second set of analysis, we consider two alternative response variables. The findings have the same patterns as above. The proposed approach is able to make biologically sensible findings with satisfactory prediction and stability performance.

6 | DISCUSSION

In this article, we have developed a new G-E interaction analysis approach, taking advantage of the most recent development in data aggregation. The proposed approach can complement and advance from the existing approaches by effectively accommodating rare features, conducting joint analysis, more effectively aggregating nearby features, and others. It is built on the existing penalized joint G-E interaction analysis and state-of-the-art data integration [28] and has a sensible formulation. Simulation has demonstrated its competitive performance. In the NFBC data analysis, it has generated findings different from the alternatives and with satisfactory prediction and stability performance.

This study can be extended in multiple directions. As briefly described above, it can be (almost) directly applied to other data types/models. A closer examination of the proposed estimation suggests that it may not be specific to SNPs, physical locations (for tree construction), or rare features. When it is expected that certain features may share similar effects, and when a similarity measure can be defined statistically or functionally, the proposed approach may be applicable. In some genetic studies, multiple responses that share related genetic basis are jointly analyzed. In the NFBC1966 study, there are some traits that may share main G effects and interactions. It will be of interest to extend the proposed method to the collective analysis of multiple response variables. It may also be of interest for future research to establish theoretical properties, which may follow from Reference [28] and the existing theoretical studies on penalized G-E interaction analysis. In data analysis, the prediction and stability evaluation can provide some indirect support to the validity of our analysis. It is of interest further examine and validate the findings.

ACKNOWLEDGMENTS

We thank the editor and reviewers for their careful review and insightful comments, which have led to a significant improvement of this article.

DATA AVAILABILITY STATEMENT

Data analyzed in this study is publicly available.

ORCID

Shuangge Ma https://orcid.org/0000-0001-9001-4999

REFERENCES

1. W. T. Boyce, M. B. Sokolowski, and G. E. Robinson, *Genes and environments, development and time*, Proc. Natl. Acad. Sci. 117 (2020), no. 38, 23235–23241.
2. D. Thomas, *Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies*, Annu. Rev. Public Health 31 (2010), 21–36.
3. F. Zhou, J. Ren, X. Lu, S. Ma, and C. Wu, *Gene–environment interaction: Avariable selection perspective*, Methods and Protocols, Epistasis, 2021, 191–223.
4. M. Wu and S. Ma, *Robust genetic interaction analysis*, Brief. Bioinform. 20 (2019), no. 2, 624–637.
5. F. Zhou, J. Ren, G. Li, Y. Jiang, X. Li, W. Wang, and C. Wu, *Penalized variable selection for lipid–environment interactions in a longitudinal lipidomics study*, Genes 10 (2019), no. 12, 1002.
6. C. Wu, Y. Jiang, J. Ren, Y. Cui, and S. Ma, *Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures*, Stat. Med. 37 (2018), no. 3, 437–456.
7. L. Bomba, K. Walter, and N. Soranzo, *The impact of rare and low-frequency genetic variants in common disease*, Genome Biol. 18 (2017), no. 1, 1–17.
8. R. Mukherjee, N. S. Pillai, and X. Lin, *Hypothesis testing for high-dimensional sparse binary regression*, Ann. Stat. 43 (2015), no. 1, 352–381.
9. M. E. Tabangin, J. G. Woo, and L. J. Martin, *The effect of minor allele frequency on the likelihood of obtaining false positives*, BMC Proc. 3 (2009), no. S7, 1–4.
10. Y. Li, N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, T. Jiang, H. Jiang, A. Albrechtsen, G. Andersen, H. Cao, T. Korneliussen, N. Grarup, Y. Guo, I. Hellman, X. Jin, Q. Li, J. Liu, X. Liu, T. Sparsø, M. Tang, H. Wu, R. Wu, C. Yu, H. Zheng, A. Astrup, L. Bolund, J. Holmkvist, T. Jørgensen, K. Kristiansen, O. Schmitz, T. W. Schwartz, X. Zhang, R. Li, H. Yang, J. Wang, T. Hansen, O. Pedersen, R. Nielsen, and J. Wang, *Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants*, Nat. Genet. 42 (2010), no. 11, 969–972.
11. S. Nejentsev, N. Walker, D. Riches, M. Egholm, and J. A. Todd, *Rare variants of ifih1, a gene implicated in antiviral responses, protect against type 1 diabetes*, Science 324 (2009), no. 5925, 387–389.

12. J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, et al., *Evolution and functional impact of rare coding variation from deep sequencing of human exomes*, Science 337 (2012), no. 6090, 64–69.
13. D. B. Goldstein, A. Allen, J. Keebler, E. H. Margulies, S. Petrou, S. Petrovski, and S. Sunyaev, *Sequencing studies in human genetics: Design and interpretation*, Nat. Rev. Genet. 14 (2013), no. 7, 460–470.
14. G. V. Kryukov, L. A. Pennacchio, and S. R. Sunyaev, *Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies*, Am. J. Hum. Genet. 80 (2007), no. 4, 727–739.
15. A. Derkach, J. F. Lawless, D. Merico, A. D. Paterson, and L. Sun, *Evaluation of gene-based association tests for analyzing rare variants using genetic analysis workshop 18 data*, BMC Proc. 8 (2014), no. 1, 1–6.
16. C. Mallaney and Y. J. Sung, *Rare variant analysis of blood pressure phenotypes in the genetic analysis workshop 18 whole genome sequencing data using sequence kernel association test*, BMC Proc. 8 (2014), no. 1, 1–6.
17. J. Xuan, L. Yang, and Z. Wu, *Higher criticism approach to detect rare variants using whole genome sequencing data*, BMC Proceed. BioMed. Central 8 (2014), 1–6.
18. M. Agne, C. H. Huang, I. Hu, H. Wang, T. Zheng, and S. H. Lo, *Considering interactive effects in the identification of influential regions with extremely rare variants via fixed bin approach*, BMC Proc. 8 (2014), no. 1, 1–6.
19. H. C. Yang and H. W. Li, *Analysis of homozygosity disequilibrium using whole-genome sequencing data*, BMC Proceed. BioMed. Central 8 (2014), 1–5.
20. A. P. Morris and E. Zeggini, *An evaluation of statistical approaches to rare variant analysis in genetic association studies*, Genet. Epidemiol. 34 (2010), no. 2, 188–193.
21. Y. J. Sung, J. Basson, and D. C. Rao, *Whole genome sequence analysis of the simulated systolic blood pressure in genetic analysis workshop 18 family data: Long-term average and collapsing methods*, BMC Proceed. BioMed. Central 8 (2014), 1–5.
22. B. M. Neale, M. A. Rivas, B. F. Voight, A. David, D. Bernie, O. M. Marju, K. Sekar, S. M. Purcell, R. Kathryn, and M. J. a. Daly, *Testing for an unusual distribution of rare variants*, PLoS Genet. 7 (2011), no. 3, e1001322.
23. M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, *Rare-variant association testing for sequencing data with the sequence kernel association test*, Am. J. Hum. Genet. 89 (2011), no. 1, 82–93.
24. A. Derkach, J. F. Lawless, and L. Sun, *Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests*, Genet. Epidemiol. 37 (2013), no. 1, 110–121.
25. L. Luo, E. Boerwinkle, and M. Xiong, *Association studies for next-generation sequencing*, Genome Res. 21 (2011), no. 7, 1099–1108.
26. T. Wang and H. Zhao, *A dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms*, Biometrics 73 (2017), no. 3, 792–801.
27. A. Yazdani, A. Yazdani, and E. Boerwinkle, *Rare variants analysis using penalization methods for whole genome sequence data*, BMC Bioinform. 16 (2015), no. 1, 1–8.
28. X. Yan and J. Bien, *Rare feature selection in high dimensions*, J. Am. Stat. Assoc. 534 (2021), 887–900.
29. J. Bien, X. Yan, L. Simpson, and C. L. Müller, *Tree-aggregated predictive modeling of microbiome data*, Sci. Rep. 11 (2021), no. 1, 1–13.
30. M. Lu, H. S. Lee, D. Hadley, J. Z. Huang, and X. Qian, *Logistic principal component analysis for rare variants in gene-environment interaction analysis*, IEEE/ACM Trans. Comput. Biol. Bioinform. 11 (2014), no. 6, 1020–1028.
31. G. Zhao, R. Marceau, D. Zhang, and J.-Y. Tzeng, *Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression*, Genetics 199 (2015), no. 3, 695–710.
32. T. Yang, H. Chen, H. Tang, D. Li, and P. Wei, *A powerful and data-adaptive test for rare-variant–based geneenvironment interaction analysis*, Stat. Med. 38 (2019), no. 7, 1230–1244.
33. X. Lin, S. Lee, D. C. Christiani, and X. Lin, *Test for interactions between a genetic marker set and environment in generalized linear models*, Biostatistics 14 (2013), no. 4, 667–681.
34. E. Lim, H. Chen, J. Dupuis, and C. T. Liu, *A unified method for rare variant analysis of gene-environment interactions*, Stat. Med. 39 (2020), no. 6, 801–813.
35. J. Bien, J. Taylor, and R. Tibshirani, *A lasso for hierarchical interactions*, Ann. Stat. 41 (2013), no. 3, 1111–1141.
36. N. Hao, Y. Feng, and H. H. Zhang, *Model selection for high-dimensional quadratic regression via regularization*, J. Am. Stat. Assoc. 113 (2018), no. 522, 615–625.
37. M. Wu, Q. Zhang, and S. Ma, *Structured gene-environment interaction analysis*, Biometrics 76 (2020), no. 1, 23–35.
38. D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander, *Linkage disequilibrium in the human genome*, Nature 411 (2001), no. 6834, 199–204.
39. W. Pan, and X. Shen, *Penalized model-based clustering with application to variable selection*. J. Mach. Learn. Res. 8 (2007), no. 5, 1145–1164.
40. C. Sabatti, S. K. Service, A. L. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, et al., *Genome-wide association analysis of metabolic traits in a birth cohort from a founder population*, Nat. Genet. 41 (2009), no. 1, 35–46.
41. N. P. Torres-Aguila, C. Carrera, E. Muiño, N. Cullell, J. Cárcel-Márquez, C. Gallego-Fabrega, J. González-Sánchez, A. Bustamante, P. Delgado, L. Ibañez, L. Heitsch, J. Krupinski, J. Montaner, J. Martí-Fàbregas, C. Cruchaga, J. M. Lee, I. Fernandez-Cadenas, and Acute Endophenotypes Group of the International Stroke Genetics Consortium (ISGC), *Clinical variables and genetic risk factors associated with the acute outcome of ischemic stroke: A systematic review*, J. Stroke 21 (2019), no. 3, 276–289.
42. X. Shi, Y. Jiao, Y. Yang, C. Y. Cheng, C. Yang, X. Lin, and J. Liu, *Vimco: Variational inference for multiple correlated outcomes in genome-wide association studies*, Bioinformatics 35 (2019), no. 19, 3693–3700.
43. M. E. Rentería, A. Cortes, and S. E. Medland, "*Using plink for genome-wide association studies (gwas) and data analysis*," *Genome-wide association studies and genomic prediction*, Springer, 2013, pp. 193–213.
44. S. Sindhu, R. Thomas, P. Shihab, E. Al Shawaf, A. Hasan, M. Alghanim, K. Behbehani, and R. Ahmad, *Changes in the*

adipose tissue expression of cd86 costimulatory ligand and cd163 scavenger receptor in obesity and type-2 diabetes: Implication for metabolic disease, J. Glycom. Lipidom. 5 (2015), no. 134, 2153-0637.

45. S. Homma, T. Shimada, T. Hikake, and H. Yaginuma, *Expression pattern of lrr and ig domain-containing protein (lrrig protein) in the early mouse embryo*, Gene Expr. Patterns 9 (2009), no. 1, 1–26.
46. Y. Wu, K. Prakash, T. Rong, H. Li, Q. Xiao, L. Tan, W. Au, J. Ding, S. Chen, and E. Tan, *Lingo2 variants associated with essential tremor and parkinson's disease*, Hum. Genet. 129 (2011), no. 6, 611–615.
47. J. R. Speakman, *Functional analysis of seven genes linked to body mass index and adiposity by genome-wide association studies: A review*, Hum. Hered. 75 (2013), no. 2–4, 57–79.
48. N. Fernàndez-Castillo, G. Gan, M. M. van Donkelaar, M. Vaht, H. Weber, W. Retz, A. Meyer-Lindenberg, B. Franke, J. Harro, A. Reif, et al., *Rbfox1, encoding a splicing regulator, is a candidate gene for aggressive behavior*, Eur. Neuropsychopharmacol. 30 (2020), 44–55.
49. A. G. Comuzzie, S. A. Cole, S. L. Laston, V. S. Voruganti, K. Haack, R. A. Gibbs, and N. F. Butte, *Novel genetic loci identified for the pathophysiology of childhood obesity in the hispanic population*, PLoS One 7 (2012), no. 12, e51954.
50. R. Wang, Y. Lu, M. Z. Cicha, M. V. Singh, C. J. Benson, C. J. Madden, M. W. Chapleau, and F. M. Abboud, *Tmem16b determines cholecystokinin sensitivity of intestinal vagal afferents of nodose neurons*, JCI Insight 4 (2019), no. 5, e122058.
51. A. Mavri, D. Bastelica, M. Poggi, P. Morange, F. Peiretti, M. Verdier, I. Juhan Vague, and M. C. Alessi, *Polymorphism a36g of the tumor necrosis factor receptor 1 gene is associated with pai-1 levels in obese women*, Thromb. Haemost. 97 (2007), no. 01, 62–66.
52. W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant, H. Parkinson, and L. M. Schriml, *Disease ontology 2015 update: An expanded and updated database of human diseases for linking biomedical knowledge through disease data*, Nucleic Acids Res. 43 (2015), no. D1, D1071–D1078.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: M. Liu, Q. Zhang, and S. Ma, *A tree-based gene–environment interaction analysis with rare features*, Stat. Anal. Data Min.: ASA Data Sci. J. (2022), 1–27. https://doi.org/10.1002/sam.11578