RESEARCH

Refining multivariate disease phenotypes for high chip heritability

Jiangwen Sun[1], Henry R. Kranzler[2] and Jinbo Bi[1*]

[*]Correspondence: jinbo@engr.uconn.edu
[1]Department of Computer Science and Engineering, University of Connecticut, 371 Fairfield Way, U-4155, Storrs, CT, 06269 USA
Full list of author information is available at the end of the article

Abstract

Background: Statistical genetics shows that the success of both genetic association studies and genomic prediction methods is positively associated with the heritability of the trait used in the analysis. Identifying highly heritable components of a complex disease can thus enhance genetic studies of the disease. Existing heritable component analysis methods use data from related individuals to compute linearly-combined traits to maximize heritability. Recent advances in acquiring genome-wide markers have enhanced heritability estimation using genotypic data from apparently unrelated individuals, which is referred to as the chip heritability. Novel statistical models are thus needed to identify disease components (subtypes) with high chip heritability.

Methods: We propose an optimization approach to identify highly heritable components of a complex disease as a function of multiple clinical variables. The heritability of the components is estimated directly from unrelated individuals using their genome-wide single nucleotide polymorphisms. The proposed approach can also model the fixed effects due to covariates, such as age and race, so that the derived traits have high chip heritability after correcting for fixed effects. A new sequential quadratic programming algorithm is developed to efficiently solve the proposed optimization problem.

Results: The proposed algorithm was validated both in simulations and the analysis of a real-world dataset that was aggregated from genetic studies of cocaine, opoid, and alcohol dependence. Simulation studies demonstrated that the proposed approach could identify the hypothesized component from multiple synthesized features. A case study on cocaine dependence (CD) identified a quantitative trait that achieved chip heritability of 0.86 estimated using a cross-validation process. This quantitative trait corresponded to the likelihood of an individual's membership in a CD subtype. Clinical analysis showed that the subtype enclosed individuals who reported heavy use of cocaine but few withdrawal symptoms.

Conclusions: Extensive experiments on both synthetic and real-world data demonstrate the effectiveness of the proposed approach as a means to find meaningful disease components with high chip heritability.

Keywords: phenotype-genotype association analysis; chip heritability; quadratic optimization; heritable component analysis

Introduction

Identifying genetic variation that underlies complex diseases has important implications in medicine. To date, genome-wide association studies (GWAS) have had limited success in dissecting the genetic etiology of complex diseases. For instance, very few associations identified for substance use disorders at a genome-wide significant level have been replicated [1, 2, 3]. Complex disorders are often characterized by multiple disease indicators. For example, to diagnose whether a patient has a lifetime drug dependence disorder, clinicians interview the patient to understand his or her drug use behaviors, the negative consequences of the drug use, the treatment history and other co-occuring medical conditions. All of these clinical variables are used to arrive at a diagnosis of dependence on a certain drug [4]. There is substantial variation in these variables in the disease population, and these variables also present different levels of heritability, i.e., some are more genetically influenced than others. This phenotypic heterogeneity diminishes evidence of genetic association. Statistical genetics also shows that the success of most gene discovery studies is positively associated with the heritability of the trait used in the association analysis [5]. Hence, identifying more homogeneous and highly heritable components of a complex disease could enhance the association analysis.

The ability to translate genotype information into a quantitative prediction of disease phenotypes is important for precision medicine [6]. Genomic prediction methods that predict a phenotype based on genome-wide single nucleotide polymorphisms (SNPs) may provide a suitable analytic tool [7, 8]. These methods expand the traditional single-marker-regression-based GWAS model for detecting few variants of large effect to multi-marker predictive models with many variants of small effect. The predictive ability of genomic prediction methods relies on several factors, especially trait heritability [9]. If we identify higly heritable components of a complex disease, it could also improve the utility of genomic prediction methods to predict subtypes (defined by the components) of the disease.

Because the success of both association analysis and genomic prediction is dependent on the trait heritability, heritability can be a valid target for refining multivariate disease phenotypes. The narrow-sense heritability h^2 is defined by the percentage of phenotypic variance that is due to additive genetic effects [10]. The broad-sense heritability H^2 is defined as the overall genetic contribution to the phenotypic variation. The heritability of a quantitative trait is commonly estimated from related individuals in pedigrees. Recent advances in acquiring dense genome-wide genetic markers have enhanced heritability estimation from apparently unrelated individuals using their genome-wide SNPs. The SNP-based heritability, often referred to as the chip heritability, is defined as the portion of the phenotypic variation that can be explained by the genotyped genetic markers [11]. It has been argued that estimating h^2 from unrelated individuals has an advantage over traditional pedigree-based methods because the estimated chip h^2 corresponds only to the causal-variant heritability that is tagged by the genotyped SNPs [8, 12].

Phenotype refinement is an important but underdeveloped genetics research area. Unsupervised cluster analysis or latent class analysis has been commonly used to partition a study population into subgroups based on clinical variables [13, 14, 15, 16, 17, 18]. This approach can create subgroups of individuals that differ in clinical symptoms and features, but may have limited utility in genetic analysis. Because genetic data are not used during the creation of the subgroups, the resultant subtypes (subgroups) are not guaranteed to have high heritability, and hence may not be informative for genetic association.

More relevant to this present work, a number of prior methods identify the principal components of clinical data that are heritable, and characterize the components by linear combinations of clinical variables [19, 20, 21, 22, 23]. Thus, these methods are often called heritable component analysis. All existing methods decompose the variance of clinical data into two components: the variance due to additive genetic effects estimated from pedigrees; and the variance due to other effects (residuals). Then, they solve a generalized eigen-decomposition problem to identify the linear combination of the clinical variables that maximizes the ratio of additive-genetic variance versus the residual variance, thus leading to high heritability of the resultant linearly combined trait. Nearly all of these methods use pedigree-based heritability estimation (an exception is [23]), and all assume a genetic model that is based on a single causal variant, an assumption that is commonly violated for complex diseases.

Although the latest heritable component analysis method [23] is effective and computationally efficient, a fundamental question is how much heritability of the derived trait can be explained by the genotyped SNPs. Because GWAS and genomic predictions mainly utilize the genotyped SNPs, the utility of the derived trait may be limited by a low chip heritability. Thus, novel statistical models are needed to directly target high chip heritability. In this paper, we propose an approach to identify the components of a multivariate disease phenotype that maximizes the chip h^2. To estimate the chip heritability of a given trait, the latest methods use the restricted maximum likelihood (REML) method, which assumes that the trait follows a mixed effect model with random genetic effects, and fixed effects due to covariates, such as age, sex and race [8, 12]. To identify a trait of high chip h^2, we need to solve the inverse problem of (chip) heritability estimation. In other words, we now search for a trait (e.g., a linearly-combined trait) so that its chip heritability is high when estimated using the REML method. Directly solving the inverse problem leads to a quadratic optimization problem that can be optimized efficiently via a sequential quadratic programming algorithm. We validated the proposed approach in simulations as well as in the analysis of a real-world dataset that was aggregated from genetic studies of cocaine, opioid, and alcohol dependence. Our experimental results demonstrated the effectiveness and generalizability of the proposed approach.

Methods

The proposed statistical model

Given a set of n subjects, we denote their trait values of a quantitative trait y by a vector $\mathbf{y}$ of length n. We use a matrix $\mathbf{Z}_{n\times m}$ to represent their standardized genotypic data at m genetic markers, and $\mathbf{C}_{n\times p}$ to represent their data on p covariates. The matrix $\mathbf{Z}$ is calculated from the genotypic data as follows. Let f_j be the frequency of reference allele at the j-th genetic variant, r_{ij} be the number of copies of reference allele that the i-th subject has at the j-th locus. The standardized genotype z_{ij} is calculated as $(r_{ij}-2f_j)/\sqrt{2f_j(1-f_j)}$ [8]. The chip heritability estimation method assumes the following mixed-effect linear model [8, 12] that characterizes how a phenotype is related to genotypes and covariates:

$$\mathbf{y} = \mathbf{C}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \varepsilon, \qquad (1)$$

where $\boldsymbol{\varepsilon}$ is a vector of length n, which specifies residual effects. In Eq.(1), all covariates create fixed effects (fixed $\boldsymbol{\beta}$) on the phenotype whereas genetic effects are random (random $\mathbf{u}$). Assume that $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ follow Gaussian distributions: $\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}\sigma_u^2)$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$. Then, the covariance of $\mathbf{y}$ between individuals, denoted by $\boldsymbol{\Omega}_{n\times n}$, can be calculated as:

$$\boldsymbol{\Omega} = \mathbf{Z}\mathbf{Z}^T\sigma_u^2 + \mathbf{I}\sigma_e^2. \tag{2}$$

Let σ_g^2 be the phenotypic variance attributable to all of the m genetic causal variants. Then, we have $\sigma_g^2 = m\sigma_u^2$. Let $\mathbf{G} = \mathbf{Z}\mathbf{Z}^T/m$, which is referred to as the genetic relationship matrix (GRM) among subjects determined by the causal variants. Then Eq. (2) can be re-written as:

$$\boldsymbol{\Omega} = \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_e^2 \tag{3}$$

where σ_g^2 and σ_e^2 can be estimated by the REML method [24, 11]. The chip heritability estimated on the m causal variants is computed as $h^2 = \sigma_g^2/\sigma_p^2$, where $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$ is the total phenotypic variance. Because the causal variants of y are usually unknown for a trait, recent research has proposed to estimate a GRM using genome-wide SNPs [8, 12].

The main idea of REML for estimating the variance components is to first eliminate the fixed effect due to covariates from the observed values of y and then estimate the variance components from the random effect part. The REML finds n basis vectors represented by columns of a matrix $\mathbf{L}_{n\times n}$. This matrix has two sub-matrices $\mathbf{L} = [\mathbf{L}_1\ \mathbf{L}_2]$ with $\mathbf{L}_1$ of size $n \times p$ and $\mathbf{L}_2$ of size $n \times (n-p)$. The two sub-matrices satisfy $\mathbf{L}_1^T\mathbf{C} = \mathbf{I}_{p\times p}$, and $\mathbf{L}_2^T\mathbf{C} = 0$. Let $\tilde{\mathbf{y}} = \mathbf{L}^T\mathbf{y}$, $\tilde{\mathbf{y}}_1 = \mathbf{L}_1^T\mathbf{y}$ and $\tilde{\mathbf{y}}_2 = \mathbf{L}_2^T\mathbf{y}$. It can be derived that $\tilde{\mathbf{y}}$ follows the following multivariate Gaussian distribution given the multivariate Gaussian assumption of $\mathbf{y}$:

$$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{L}_1^T\mathbf{y} \\ \mathbf{L}_2^T\mathbf{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\beta} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{L}_1^T\boldsymbol{\Omega}\mathbf{L}_1 & \mathbf{L}_1^T\boldsymbol{\Omega}\mathbf{L}_2 \\ \mathbf{L}_2^T\boldsymbol{\Omega}\mathbf{L}_1 & \mathbf{L}_2^T\boldsymbol{\Omega}\mathbf{L}_2 \end{bmatrix} \right).$$

We have $\tilde{\mathbf{y}}_2 \sim N(0, \mathbf{L}_2^T\boldsymbol{\Omega}\mathbf{L}_2)$ and the conditional distribution:

$$\tilde{\mathbf{y}}_1|\tilde{\mathbf{y}}_2 \sim N\big(\boldsymbol{\beta} + \mathbf{L}_1^T\boldsymbol{\Omega}\mathbf{L}_2(\mathbf{L}_2^T\boldsymbol{\Omega}\mathbf{L}_2)^{-1}\tilde{\mathbf{y}}_2, (\mathbf{C}^T\boldsymbol{\Omega}\mathbf{C})^{-1}\big).$$

Then, the log likelihood of $\tilde{\mathbf{y}}$ can be decomposed into:

$$\ell(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}) = \ell_1(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}_1|\tilde{\mathbf{y}}_2) + \ell_2(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}_2),$$

where ℓ_2 is not a function of the fixed-effect parameter $\boldsymbol{\beta}$. The two variance components, i.e., σ_g^2 and σ_e^2 can be estimated by maximizing ℓ_2, and there is no additional information in ℓ_1 for estimating the variance components. Once σ_g^2 and σ_e^2 are estimated, a generalized least squares estimate of $\boldsymbol{\beta}$ can be obtained as:

$$\hat{\boldsymbol{\beta}} = \tilde{\mathbf{y}}_1 - \mathbf{L}_1^T\boldsymbol{\Omega}\mathbf{L}_2(\mathbf{L}_2^T\boldsymbol{\Omega}\mathbf{L}_2)^{-1}\tilde{\mathbf{y}}_2.$$

The second log likelihood component ℓ_2 is calculated as (after removing constant):

$$\ell_2(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}_2) = -\frac{1}{2}(\ln |\mathbf{L}_2^T \boldsymbol{\Omega} \mathbf{L}_2| + \tilde{\mathbf{y}}_2^T (\mathbf{L}_2^T \boldsymbol{\Omega} \mathbf{L}_2)^{-1} \tilde{\mathbf{y}}_2).$$

It has been shown in an early work [25] that when $\mathbf{L}_1^T \mathbf{C} = \mathbf{I}_{p \times p}$ and $\mathbf{L}_2^T \mathbf{C} = 0$, we have

$$\boldsymbol{\Omega} - \boldsymbol{\Omega} \mathbf{L}_2 (\mathbf{L}_2^T \boldsymbol{\Omega} \mathbf{L}_2)^{-1} \mathbf{L}_2^T \boldsymbol{\Omega} = \mathbf{C}(\mathbf{C}^T \boldsymbol{\Omega}^{-1} \mathbf{C})^{-1} \mathbf{C}^T.$$

Substituting these equations into the calculation of ℓ_2 yields:

$$\ell_2(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}_2) = -\frac{1}{2}(\ln |\boldsymbol{\Omega}| + \ln |\mathbf{C}^T \boldsymbol{\Omega}^{-1} \mathbf{C}| + \mathbf{y}^T \mathbf{P} \mathbf{y}), \tag{4}$$

where $\mathbf{P} = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1} \mathbf{C}(\mathbf{C}^T \boldsymbol{\Omega}^{-1} \mathbf{C})^{-1} \mathbf{C}^T \boldsymbol{\Omega}^{-1}$ and $\boldsymbol{\beta}$ can be obtained by:

$$\boldsymbol{\beta} = (\mathbf{C}^T \boldsymbol{\Omega}^{-1} \mathbf{C})^{-1} \mathbf{C}^T \boldsymbol{\Omega}^{-1} \mathbf{y}. \tag{5}$$

Given data on $\mathbf{y}$, $\mathbf{C}$ and $\mathbf{Z}$, σ_g^2 and σ_e^2 are obtained by maximizing the log likelihood of observing the trait values $\ell(\sigma_g^2, \sigma_e^2; \mathbf{y})$ which corresponds to maximizing $\ell_2(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}_2)$ [11]. The chip heritability of a trait y is computed using the resultant optimal ${\sigma_g}^2$ and ${\sigma_e}^2$.

In our study, however, we solve the inverse problem of the above estimation model. A definitive quantitative trait y is not known beforehand but needs to be derived from a set of known clinical variables. Let $\mathbf{X}_{n \times d}$ be the data matrix of d clinical variables $\mathbf{x}$ for the same n subjects as in $\mathbf{Z}$. A trait y is defined by a linear function of $y = \mathbf{w}^\top \mathbf{x}$ where $\mathbf{w}$ is the vector of combination coefficients. Correspondingly, the trait values $\mathbf{y} = \mathbf{X}\mathbf{w}$. Unlike the heritability estimation process that finds the best values of σ_g^2 and σ_e^2 to maximize the likelihood of observing the values of y, the inverse problem searches for the best $\mathbf{w}$ so to form a trait $\mathbf{y}$ that maximizes the likelihood, (or equivalently the log likelihood $\ell(\sigma_g^2, \sigma_e^2; \mathbf{y}, \mathbf{C}, \mathbf{Z})$), of observing a large heritability, i.e., a large σ_g^2 but small σ_e^2. For simplicity and easy interpretation of the resultant model, here we only consider linear models, but the proposed method can be easily extended to construct non-linear models through kernel mapping [26].

Notice that the highest possible heritability of a trait y is 1 when $\sigma_g^2 = 1$ and $\sigma_e^2 = 0$. We hence propose to formulate an optimization problem, in which we search for the optimal $\mathbf{w}$ that maximizes the log likelihood $\ell(\sigma_g^2, \sigma_e^2; \mathbf{y}, \mathbf{C}, \mathbf{Z})$ (or equivalently, $\ell_2(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}_2)$) of observing $\sigma_g^2 = 1$ and $\sigma_e^2 = 0$. According to Eq.(3), the covariance matrix $\boldsymbol{\Omega} = \mathbf{G}$ when $\sigma_g^2 = 1$ and $\sigma_e^2 = 0$. We substitute the values of these parameters into the log likelihood Eq.(4), and remove any constant terms. The resultant maximization problem is equivalent to the following minimization problem:

$$\min_{\mathbf{w}} \quad \mathbf{w}^T (\mathbf{X}^T \mathbf{P} \mathbf{X}) \mathbf{w} \tag{6}$$

where $\mathbf{P}$ is calculated as:

$$\mathbf{P} = \mathbf{G}^{-1} - \mathbf{G}^{-1} \mathbf{C}(\mathbf{C}^T \mathbf{G}^{-1} \mathbf{C})^{-1} \mathbf{C}^T \mathbf{G}^{-1}. \tag{7}$$

When $\sigma_g^2 = 1$ and $\sigma_e^2 = 0$, we have $\sigma_p^2 = 1$ because $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$. This requires to impose a constraint to the optimization problem so that the total phenotypic variance that is due to either genetic or environmental effect should be scaled to 1. An estimate of σ_p^2 can be obtained by calculating the sample variance after correcting for the covariate effects as $\hat{\sigma}_p^2 = \frac{1}{n}(\mathbf{X}\mathbf{w} - \mathbf{C}\boldsymbol{\beta})^T(\mathbf{X}\mathbf{w} - \mathbf{C}\boldsymbol{\beta})$. Since $\boldsymbol{\beta}$ can be estimated according to Eq.(5), by substituting the $\boldsymbol{\beta}$ value, $\hat{\sigma}_p^2$ can be computed by

$$\hat{\sigma}_p^2 = \frac{1}{n}\mathbf{w}^T\mathbf{X}^T(\mathbf{J}^T\mathbf{J})\mathbf{X}\mathbf{w}$$

where $\mathbf{J} = \mathbf{I} - \mathbf{C}(\mathbf{C}^T\boldsymbol{\Omega}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\boldsymbol{\Omega}^{-1}$. To further simplify the notation, denote

$$\mathbf{Q} = \frac{\mathbf{J}^T\mathbf{J}}{n}, \tag{8}$$

then

$$\hat{\sigma}_p^2 = \mathbf{w}^T(\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{w}.$$

Combining the objective function and the constraint together, the proposed optimization problem is formulated as:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w}, \\ \text{subject to} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{w} = 1. \end{aligned} \tag{9}$$

According to statistical learning theory [26], only maximizing the training heritability (by minimizing Eq.(9)), the resultant model may overfit the training data $\mathbf{X}$. If overfitting occurs, the optimal $\mathbf{w}$ of Eq.(9) may correspond to a trait that has high heritability on the data that is used to train the linear model, but when the model is applied to a new sample, the trait has low heritability. In order to prevent overfitting and identify a trait with high heritability that can generalize, we incorporate a regularizer $R(\mathbf{w})$ in our formulation (9). The optimization problem becomes:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{n}\mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w} + \frac{\lambda}{d}R(\mathbf{w}) \\ \text{subject to} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{w} = 1, \end{aligned} \tag{10}$$

where λ is a hyper-parameter and needs to be tuned, and $\frac{1}{n}$ and $\frac{1}{d}$ are included to pre-balance the two items in the objective function. The value of λ can either be chosen by users according to domain knowledge or determined using a cross-validation process as done in our experiments. According to learning theory [26], minimizing $\frac{1}{n}\mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w}$ corresponds to empirical risk minimization, whereas minimizing the objective in Eq.(10) corresponds to structural risk minimization that improves the generalizability of the resultant model. There are many different ways to define $R(\mathbf{w})$ [23]. The L_2 vector norm defined by $\|\mathbf{w}\|_2^2 = \sum_i w_i^2$ is a common choice. The L_1 vector norm defined by $\|\mathbf{w}\|_1 = \sum_i |w_i|$ can be a better choice when model sparsity is required to select variables for use in the model. In more complicated applications where variables may be grouped and feature selection among groups is expected, a structured regularizer, such as the group lasso $||\mathbf{w}||_{2,1} = \sum_{k=1}^{L} \sqrt{\sum_{i\in\mathcal{G}_k} w_i^2}$, can be used where $\mathcal{G}_k$ contains the indices of variables belonging to a group k.

Optimization algorithm

In this paper, we use the L_1 norm penalty $\|\mathbf{w}\|_1$ to be $R(\mathbf{w})$, and develop an efficient algorithm to solve the resultant optimization problem as follows:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{n}\mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w} + \frac{\lambda}{d}\|\mathbf{w}\|_1 \\ \text{subject to} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{w} = 1. \end{aligned} \tag{11}$$

The algorithm we will describe next, although is designed for Problem (11), can be modified to solve Problem (10) that may take another form of the regularizers.

Due to the use of the $\|\mathbf{w}\|_1$ norm, the objective function in Problem (11) is not continuously differentiable and a gradient decent type of approach cannot be applied directly. A well known strategy to overcome this obstacle is to decompose $\mathbf{w}$ into two parts: $\mathbf{w} = \mathbf{u} - \mathbf{v}$, where both $\mathbf{u}$ and $\mathbf{v}$ are vectors of the same size as that of $\mathbf{w}$, and all the components in $\mathbf{u}$ and $\mathbf{v}$ are required to be non-negative (i.e., $\mathbf{u} \geq 0$, and $\mathbf{v} \geq 0$). Because $\mathbf{Xw} = \mathbf{Xu} - \mathbf{Xv}$, we denote $\boldsymbol{\gamma} = [\mathbf{u}^T, \mathbf{v}^T]^T$, $\mathbf{H} = [\mathbf{X}, -\mathbf{X}]$, and then we have $\mathbf{Xw} = \mathbf{H}\boldsymbol{\gamma}$. By the change of variables, Problem (11) can be equivalently re-written as:

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \quad & f : \frac{1}{n}\boldsymbol{\gamma}^T(\mathbf{H}^T\mathbf{P}\mathbf{H})\boldsymbol{\gamma} + \frac{\lambda}{d}\sum_{i=1}^{2d}\gamma_i \\ \text{subject to} \quad & g_1 : \boldsymbol{\gamma}^T(\mathbf{H}^T\mathbf{Q}\mathbf{H})\boldsymbol{\gamma} - 1 = 0 \\ & g_{2:e} : \boldsymbol{\gamma} \succeq 0, \end{aligned} \tag{12}$$

where f denotes the objective function, g's denote the constraints, and $e = 2d + 1$, indicating the number of constraints in that group. It is straightforward to show that Eq.(12) is equivalent to Eq.(11) in the sense that at optimality $\mathbf{w} = \mathbf{u} - \mathbf{v} = \boldsymbol{\gamma}(1 : d) - \boldsymbol{\gamma}(d + 1 : 2d)$. When Eq.(12) reaches optimality, at least one of the two components u_i and v_i at any i-th position of the two vectors will be 0. Otherwise, by setting $\tilde{u}_i = u_i - v_i$ and $\tilde{v}_i = 0$ if $u_i \geq v_i$, or $\tilde{u}_i = 0$ and $\tilde{v}_i = v_i - u_i$ if $u_i < v_i$, we obtain a better solution with $\tilde{u}_i$ and $\tilde{v}_i$ than $(\mathbf{u}, \mathbf{v})$. Therefore, at optimality, $\sum_{i=1}^{2d}\gamma_i = \sum_{i=1}^{d} u_i + v_i = \sum_{i=1}^{d}|w_i| = \|\mathbf{w}\|_1$. Then, Eq.(12) becomes exactly the same as Eq.(11).

Eq.(12) is not a convex problem because of the quadratic equality constraint. However, it can be efficiently solved using a sequential quadratic programming (SQP) algorithm [27] because both of the objective and constraints are either in a quadratic or a linear form. The gradient of the objective and constraint functions with respect to γ can be calculated as:

$$\begin{aligned}\triangledown f &= \frac{2}{n}(\mathbf{H}^T\mathbf{PH})\gamma + \frac{\lambda}{d}\mathbf{1},\\ \triangledown g_1 &= 2(\mathbf{H}^T\mathbf{QH})\gamma,\\ \triangledown g_{2:e} &= \mathbf{I}.\end{aligned}$$

Let $\boldsymbol{\alpha}$ be the Lagrange multipliers, the Lagrangian function of this problem can be written as:

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = f(\boldsymbol{\gamma}) + \sum_i \alpha_i g_i(\boldsymbol{\gamma});$$

and the Hessian of the Lagrangian function with respect to $\boldsymbol{\gamma}$ is computed as:

$$\triangledown^2_{\mathcal{L}} = 2\mathbf{H}^T(\frac{\mathbf{P}}{n} + \alpha_1\mathbf{Q})\mathbf{H}.$$

We iteratively search for the optimal solution to Eq.(12). In the t-th iteration, we have the iterates $\boldsymbol{\gamma}_t$ and $\boldsymbol{\alpha}_t$, and we first solve the following quadratic program to find the moving direction for $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$,

$$\begin{aligned}\min_{\mathbf{p}} \quad & f(\boldsymbol{\gamma}_t) + \triangledown f(\boldsymbol{\gamma}_t)^\top \mathbf{p} + \frac{1}{2}\mathbf{p}^\top \triangledown^2 \mathcal{L}(\boldsymbol{\gamma}_t, \boldsymbol{\alpha}_t)\mathbf{p}\\ \text{subject to} \quad & \triangledown g_1(\boldsymbol{\gamma}_t)^\top \mathbf{p} + g_1(\boldsymbol{\gamma}_t) = 0,\\ & \triangledown g_i(\boldsymbol{\gamma}_t)^\top \mathbf{p} + g_i(\boldsymbol{\gamma}_t) \succeq 0, i \in [2:e].\end{aligned} \tag{13}$$

The optimal solution $\hat{\mathbf{p}}$ to the problem (13) will give the next moving direction for $\boldsymbol{\gamma}$, along which the objective of Problem (12) can be decreased. Let $\hat{\mathbf{q}}$ be the optimal Lagrange multipliers of the problem (13) corresponding to $\hat{\mathbf{p}}$. The next moving direction of $\boldsymbol{\alpha}$ is calculated as $\hat{\mathbf{q}} - \boldsymbol{\alpha}_t$. After the moving directions are computed, we then employ a line search method described in [27] to find the optimal searching step size s and update $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ as follows:

$$\boldsymbol{\gamma}_{t+1} = \boldsymbol{\gamma}_t + s\hat{\mathbf{p}}_t, \qquad \boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + s(\hat{\mathbf{q}}_t - \boldsymbol{\alpha}_t). \tag{14}$$

We summarize the proposed algorithm in Algorithm 1. It has been proved that a SQP based algorithm can converge to a local minimizer $\hat{\boldsymbol{\gamma}}$ of the optimization problem (12) [28].

Data sets

We validated the proposed approach in both simulations and the analysis of a real-world data set that was aggregated from multiple genetic studies of cocaine dependence (CD).

Cocaine use and related behaviors data

We used the *Semi-Structured Assessment for Drug Dependence and Alcoholism* (SSADDA) dataset aggregated from genetic studies of drug dependence to evaluate

Algorithm 1 A sequential quadratic programming approach to solving Problem (11)

```
Input: $\mathbf{Z}$, $\mathbf{C}$, $\mathbf{X}$, $\lambda$
Output: $\gamma$
1. Calculate $\mathbf{P}$ according to Eq.(7), and $\mathbf{Q}$ according to Eq.(8).
2. Initialize $\gamma$ with $\mathbf{u} = \mathbf{1}$, $\mathbf{v} = \mathbf{0}$.
3. Initialize the Lagrange multipliers $\boldsymbol{\alpha} = 1$.
4. Evaluate $f$, $\triangledown f$, $\triangledown g_i$ and $\triangledown^2 \mathcal{L}$ with the current $\gamma$ and $\boldsymbol{\alpha}$.
5. Solve Problem (13) to obtain $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$.
6. Perform a line search to find the searching step size $s$.
7. Update $\gamma$ and $\boldsymbol{\alpha}$ as in Eq.(14).
Repeat 4-7 until $\gamma$ reaches a fixed point.
```

the proposed algorithm. The SSADDA subjects were recruited from multiple sites, including the University of Connecticut Health Center, Yale University School of Medicine, the University of Pennsylvania School of Medicine, McLean Hospital and the Medical University of South Carolina. All subjects participated using procedures approved by the institutional review board at each participating site. There were 6,621 subjects genotyped with a total of 1,140,420 SNPs genome-wide. Among the subjects, 2,674 were stratified into the African American population using STRUCTURE software v2.3 [29], and only these subjects were used in our experiments to avoid spurious findings due to population structure. We removed 537 subjects who had other family members in the data so the GRM was computed for unrelated individuals.

A series of data cleaning steps were performed to ensure the quality of genotypic markers. Markers that meet any of the following conditions were excluded: low call rate ($< 98\%$ subjects received values for the marker), G/C and A/T markers (to avoid strand issues), deviation from Hardy-Weinberg equilibrium at $p < 10^{-8}$, significant cohort calling discrepancy and being monomorphic. We also removed non-autosomal markers, so that only markers from the 22 autosomal chromosomes were used in the analysis. After these data cleaning steps, 690,864 SNPs remained. Genetic relationship was estimated for each pair of subjects by the genome-wide complex trait analysis (GCTA) software [11] using all 690,864 SNPs. We then excluded 385 subjects whose relatedness to some subjects was greater than 0.025 (corresponding to the relatedness of second cousins). The remaining sample, 1,752 subjects, was used in our analysis.

All subjects were interviewed with a computer-assisted assessment system called the SSADDA [4], which consists of survey questions designed for cocaine use and related behaviors. All subjects were reported to have used cocaine in their lifetime. The responses to those questions in the SSADDA led to the definition of thirteen important cocaine use related variables, based on which a diagnosis of CD was determined. There were seven binary variables as listed below, which represent the seven cocaine dependence criteria in DSM-IV.

- *F1* - tolerance to cocaine;
- *F2* - withdrawal from cocaine;
- *F3* - using cocaine in larger amounts or over longer period than intended;
- *F4* - persistent desire or unsuccessful efforts to cut down or control cocaine use;
- *F5* - great amount of time spent in activities necessary to obtain, use or recover from the effects of cocaine;

- *F6* - gave up or reduced important social, occupational, or recreational activities because of cocaine use;
- *F7* - cocaine use despite knowledge of persistent or recurrent physical or psychological problems likely to have been caused or exacerbated by cocaine.

In our experiments, positive responses to the seven variables were coded by 1 and negative responses were coded by 0. We also included six continuous variables in the analysis as listed below:

- *F8* - number of cocaine symptom endorsed;
- *F9* - age when first used cocaine;
- *F10* - age when last used cocaine;
- *F11* - age when first diagnosed with DSM-IV cocaine dependence;
- *F12* - age when last diagnosed with DSM-IV cocaine dependence;
- *F13* - transition time in years between the first cocaine use and the first cocaine dependence diagnosis.

All these variables were normalized to the range of $[0, 1]$ in the analysis.

Synthetic data

Following the same design principle used in the simulations for testing chip heritability in [8], we used the real-life genotypic data in the CD study but synthesized phenotypic data. We simulated quantitative traits based on the mixed-effect linear model shown in Eq.(1). We first synthesized a dataset that contained 5 phenotypic features, all of which were created with moderate to high heritability, and were used to form a quantitative trait of very high heritability reaching 0.8. We then added irrelevant features, which varied mainly due to covariates, to create five other simulated datasets. These datasets consisted of 10, 20, 30, 40 and 50 features where only the first 5 of them were used in the model of the final trait. These datasets were used to determine whether the proposed algorithm could identify the right features for use in the model.

To synthesize features with genetic effects, we randomly picked 2,000 of the 690,864 SNPs in the cocaine use data set and used them as the causal variants of these features. The random effect coefficient u_j associated with each of the 2,000 markers was generated independently by sampling from the standard normal distribution $N(0, 1)$. The residual component ε_i for each individual was drawn from the normal distribution of mean 0 and variance $\text{var}(\mathbf{z}_i\mathbf{u})(1/h^2 - 1)$ where $\mathbf{z}_i$ is the i-th row of $\mathbf{Z}$, $\text{var}(\cdot)$ is the sample variance of a random variable and h^2 is the heritability of the feature. To synthesize features with no genetic effects, we ignored the term $\mathbf{Zu}$ in Eq.(1) and created ε by randomly sampling from the standard normal distribution. To further synthesize features with fixed covariate effects, we used sex and age of the individuals in the CD study as the covariates, and arbitrarily set their effects, i.e., the β coefficients, to 0.2 and 0.5.

We evaluated the proposed method in two different experimental settings:

Setting 1: This setting assumed that there were no covariate effects in the quantitative trait. The five relevant features were simulated as follows. We used the procedure described in the above paragraph to create four features with h^2 equal to 0.2: $\mathbf{x}_1, \cdots, \mathbf{x}_4$. Then we simulated the final quantitative trait $\mathbf{y}_1$ with $h^2 = 0.8$ using the same procedure. A five-entry weight vector was created with arbitrary values, such as $\mathbf{w} = [0.22, 0.67, 0.60, 0.30, 0.22]$, used in our experiments. Then, the fifth feature was directly computed as $\mathbf{x}_5 = (\mathbf{y}_1 - \sum_{i=1}^{4} w_i \mathbf{x}_i)/w_5$. By simulating the data in this way, we knew that there was at least one linear combination of the five features in the data that would result in a composite trait (i.e., $\mathbf{y}_1$) with h^2 of 0.8. Hence, if our approach worked, it should at least find this linear combination if there was no any other one that gave even higher h^2. Note that the heritability of the fifth feature had to depend on the empirical estimation, but given how it was created, there were genetic effects in this feature.

We then created 45 other features that had no genetic effects, and added a certain number of these features to the original 5 features to create 5 other datasets. Hence, there were in total 6 datasets for 1,752 subjects with 5, 10, 20, 30, 40 and 50 features. We used this set of data (i.e., the discovery set) in training to retrieve the combination models. Then we repeated the above procedure to create another independent set of data (i.e., the validation set) to validate the resultant models.

Setting 2: This setting assumed that the two covariates, sex and age, had fixed effects to the features and the final trait. We generated 5 features by adding fixed effects to the 5 useful features created in Setting 1. Because fixed effects do not change h^2 of a trait, we computed a composite trait $\mathbf{y}_2$ using the same pre-specified weight vector $\mathbf{w}$ that was used in Setting 1. We then created 45 other features with only covariate effects using the procedure described early on. Five other datasets were generated consisting of 10, 20, 30, 40 and 50 features. Note that the optimal weight vector for these datasets should have zero entries for all features except the first 5 features that were synthesized. Similarly, a discovery suite of the six datasets and another suite of them were synthesized using the same procedure for training and validation, respectively.

We estimated the chip h^2 of the features created in the synthetic datasets using GCTA software. The four features synthesized with a pre-specified $h^2 = 0.2$ had empirical chip h^2 values 0.2 ± 0.01 in these datasets. The chip h^2 estimate of the fifth feature was 0.57 in the discovery set and 0.6 in the validation set. Because fixed effects do not affect trait heritability, the five relevant features and the final quantitative traits had the same empirical chip h^2 in Settings 1 and 2. The features simulated with no genetic effects had h^2 estimates that ranged from 0 to 0.05, and most of these features had estimates less than 10^{-5}.

The proposed analyses

We first validated the proposed approach in a variety of experiments with the synthetic data. Then we applied our approach to the real-life cocaine use data to identify important components or subtypes of the disease defined by linear combinations of clinical features. Such a combination can be used to define a disease subtype because it produces a quantitative trait for each individual, which amounts to the membership likelihood of the individual in a subtype. Because the actual causal variants were known for synthetic data, we calculated the GRM of the individuals directly using the causal variants. In the case study for CD, because the real causal variants were unknown, we followed the commonly-used procedure in the literature on chip heritability estimation [30] and computed the GRM using all 690,864 SNPs that remained in the data. All of the reported chip heritability was estimated using GCTA software.

Tuning of the hyperparameter: For both the simulation and the CD case study, we performed 10 times three-fold cross validation (CV) to help determine a proper value of λ. At each fold of the CV, a linear model was derived by running the proposed method on 2/3 of the data in the dataset, and then tested on the remaining 1/3 of the data. The cross validated h^2 of the derived trait was estimated using only subjects in the remaining 1/3 of the data which was not used to train the model. We ran the same CV process for each pre-specified choice of λ (the choices we used are reported in the results section) and chose the λ value that gave a trait of the highest cross validated h^2 for each experimental setting.

Evaluation metrics: We reported and investigated the CV performance (including the mean values and standard deviations of the validation h^2 obtained in the CV process described above) for each λ choice in each experiment. Once the best value of λ was chosen through the cross validation for a dataset, we applied the proposed approach with the best λ to the entire data in the dataset to derive a quantitative trait. The chip heritability of this trait was estimated using the separately-synthesized validation datasets in simulations and by another cross validation process in the case study. In other words, in simulations, we estimated the valiation chip h^2 using the trait values computed by the linear model on the newly-synthesized validation samples. In the CD case study, we computed the trait h^2 using SNPs of 2/3 of the subjects randomly sampled from the dataset and repeated the random sampling 10 times to report the averaged h^2 value. We named this process the evaluation CV. Moreover, besides the heritability as a major evaluation metric, we also measured the effectiveness of our approach by comparing the derived trait models and the linear model implanted in the simulated data. We calculated the squared difference between the learned weights $\hat{\mathbf{w}}$ and the true weights $\mathbf{w}$, i.e., $\mathrm{SE}(\mathbf{w}) = \|\mathbf{w} - \hat{\mathbf{w}}\|_2^2$ and the mean of squared residuals $\mathrm{SE}(\mathbf{y}) = (1/n)\sum_{i=1}^{n}(y_i - \mathbf{x}_i\hat{\mathbf{w}})^2$, and reported the values in plots. Additional evaluation steps were conducted in the case study to clinically interpret the resultant quantitative trait (see the later paragraph).

Comparison: The validated chip h^2 of our derived traits was compared with that of all quantitative features in the data in both simulations and the CD case study. In each of the experiments, our derived trait was also compared with the commonly-used disease phenotype, often referred to as a symptom count, which was the quantitative trait created by equal weighted aggregation of all features in the data. Given that no prior method existed to identify heritable disease components using the genome-wide SNPs, on the real-life cocaine use data, we compared our approach with a recently published method [23] that aimed to derive linearly-combined traits using pedigrees of related individuals. This comparison considered whether a pedigree-based heritable component analysis method can identify a disease component with a chip heritability comparable to that found by our approach. As multi-member families were included in the original cocaine use dataset, i.e., a superset of the sample used by our approach, it was feasible to apply the method in [23] to derive a trait. Then we computed the trait values on the unrelated individuals used by our approach to compare the chip h^2 of the two approaches using the evaluation CV. Note that the prior pedigree-based approach was actually given a favor because it used the superset of 2,674 subjects (unrelated individuals were treated as one-member pedigrees) to derive the trait in comparison with our approach that used only the 1,752 unrelated individuals.

Clinical interpretation: It is very important to understand the clinical implications of the quantitative trait (or an empirical subtype) derived by our approach from the aggregated CD study data. From prior work [17, 31], we identified three key steps to ensure the clinical validity of an empirical subtype. We first examined the specific features selected by our approach for use in the model. Second, we studied the distribution (or histogram) of the quantitative scores among the 1,752 subjects. From the distribution plot, we examined whether there were obvious subgroups of the scores. Third, the subgroups of subjects were characterized and compared on 11 of the most important clinical variables reflecting cocaine use and related behaviors including both the features selected and those not selected for use in the linear model. The individuals receiving very high or very low values of the quantitative trait may show the most representative features of the subtype.

Results

Simulations

We pre-specified 21 different λ values ranging from 0 to 0.04 with step size 0.002 for use in the cross-validation tuning process. The validation or test h^2 for each λ choice was plotted for each of the six datasets in Figure 1 (for Setting 1 where data were generated without covariate effects) and Figure 2 (for Setting 2 where data were generated with covariate effects). The mean, median, and the standard deviations of the test h^2 values in the cross validation were plotted for each tested λ. These two figures show that our approach could identify components (quantitative traits) with test h^2 estimate of ~ 0.8, which was the heritability of the implanted heritable component (the simulated true model), for all datasets even with many irrelevant features in some of the datasets for both settings. This result demonstrates that our approach identified highly heritable disease components and could successfully correct for fixed covariate effects.

In addition, both Figures 1 and 2 show that there was overfitting of the learned models to the training data when λ was too small, especially when the number of irrelevant features grew to 25, 35 and 45. The larger the number of irrelevant features in the data, the more severe was the overfitting seen when λ was small. On the other hand, when λ was too large, underfitting could occur, so the test h^2 showed a peak in most of the plotted curves. The best choice of λ for the six datasets (sorted from the smallest number of features in the data to the largest number of features) were 0, 0, 0, 0.002, 0.004 and 0.004 for Setting 1, and 0, 0, 0, 0,002, 0.004 and 0.006 for Setting 2.

Then λ was set to the optimal value for each of the six datasets, and we re-ran the algorithm to generate the final quantitative trait from each dataset. We then compared the chip h^2 between these derived traits and the commonly used disease traits, such as individual features and the equally weighted aggregation of individual features. The results are shown in Figure 3 where all h^2 values were estimated using only the *validation* datasets. Results from both settings are shown. For all the datasets and for both settings, our approach could recover the quantitative traits of h^2 close to 0.8. When the number of irrelevant features in the data increased, the h^2 values of the derived traits decreased as expected. Typically, when more irrelevant features were included in the data, the learning problem became more challenging. Because covariates do not affect h^2 estimate when their effect is properly corrected during the estimation, we did not differentiate the two settings when discussing and comparing the h^2 values of individual features and the aggregated traits. Recall that in our simulation, the highest chip h^2 of individual features was 0.6 in validation. Figure 3 also included this most heritable feature for comparison. Among the traits derived by equally weighted aggregation, the one developed from the 5-feature dataset reached the highest h^2 $(= 0.66)$. As expected, chip h^2 of these traits decreased along with the increasing number of irrelevant features. The trait developed from the 50-feature dataset had the lowest h^2 $(= 0.28)$. These results demonstrate that our approach could identify quantitative traits that are more heritable (i.e., with high chip heritability) than those commonly used.

The squared difference (error) between learned weights and the optimal (implanted) weights: SE($\mathbf{w}$) together with the squared error between the derived traits and simulated traits: SE($\mathbf{y}$) are presented in Figure 4. The results from both settings are provided. We observed that clearly when the number of irrelevant features increased, the noisier data made the learning problem more difficult, and SE($\mathbf{y}$) and SE($\mathbf{w}$) increased in both of the settings.

A case study of cocaine dependence

In this study, we used the same pre-specified λ values in the simulations. In all expriments, we used age, sex and the first three principal components of the GRM as covariates. The test h^2s of all traits derived for each λ choice in the CV process are plotted in Figure 5. It shows that there was overfitting when λ was too small as well. The trait derived with $\lambda = 0.004$ achieved the highest cross-validated h^2 on average. We thus derived a model by running our approach with $\lambda = 0.004$ and all the 1,752 subjects in the data. We examined this model and the resultant quantitative trait as discussed in the proposed analyses section.

The weights that each variable received in the model are shown in Figure 6. Of the 13 clinical variables, five (F8 - F12) received a zero coefficient, and were completely ruled out from the model. Variable F13 had an coefficient close to 0 (¡ 10^{-5}), thus its impact on the resultant trait was minimal. Variables with significant coefficients in the model are the seven cocaine criteria: F3 - using cocaine in larger amounts or over longer period than intended, F2 - withdrawal from cocaine, and F5 - great amount of time spent in activities necessary to obtain, use or recover from the effects of cocaine, had the highest impact on the trait. Both F3 and F5 had negative weights, which indicated that positive response to these two variables would lower the score or value of this trait. In contrast, F2 had a positive coefficient, which indicated that a positive response to this variable would increase the score. Variables F4 - persistent desire or unsuccessful efforts to cut down or control cocaine use, and F1 - tolerance to cocaine, had limited impact on the trait.

The cross validated h^2 estimate of the trait derived by the proposed approach is 0.87 (with a standard error of 0.13). For comparison, we ran the approach proposed in [23], which identifies heritable components with pedigrees as genetic inputs and its formulation also had a hyper-parameter λ. For fair comparison, we first ran 10 times cross validation to choose a proper value for λ. With this λ, we developed a trait using the entire African American sample set. We then estimated the trait h^2

using the exact same setting (i.e., GRM and covariates) as for the traits derived by the proposed approach. We estimated the h^2 for all six continuous variables in the data using the same setting. All of the h^2 values are plotted in Figure 7 together with the trait derived by the proposed approach. The figure clearly shows that the trait derived by the proposed approach had the highest h^2 among all compared traits, and was significantly higher than that of the trait derived using the prior approach [23]. Note that one of the continuous variables in the cocaine use data was the counting of CD criteria. It was defined as the number of positive responses to the seven CD criteria, i.e., a quantitative trait resulting from a linear combination of F1-F7 with equal weights. This trait was reported to be a better trait for genome-wide association analysis than the binary CD diagnosis [1]. The CD diagnosis had a value of h^2 close to zero when estimated using our data. These results demonstrate the effectiveness of our approach in identifying disease components with high chip heritability from complex multivariate phenotypes. It is worth noting that the trait with the second highest h^2 estimate was the one derived using the prior approach [23]. This implies that (1) both of these two methods (i.e., the proposed method and the one previously reported [23]) can identify heritable components with high chip heritability; but (2) the proposed method outperforms as it directly maximizes heritability using genetic variants.

Figure 8 shows the distribution of the trait values (i.e., the membership scores) of the subjects. It shows that based on the scores, the samples can be partitioned into four subgroups. There were 250 subjects (14.27% of total) in Group 1, which had a mean score of -2.22. Group 2 consisted of 323 subjects and comprised 18.44% of the entire sample set. Its mean score was -0.8. Group 3 was the largest one and consisted of 821 subjects (46.86% of the sample). The mean score of this group was -0.2. Group 4 was the smallest, comprised of 237 individuals (13.53% of the sample), with a mean score of 1.22.

To understand the clinical implications of the derived trait, we characterized the four groups using 11 important clinical variables, including the 7 CD criteria (F1-F7), the total number of CD criteria endorsed (F8), age of first cocaine use (F9), age when first diagnosed with DSM-IV CD (F11), and the transition time in years from first cocaine use to first DSM-IV CD diagnosis (F13). The results are summarized in Table 1. Only 5.6% of the subjects in Group 1 had experienced cocaine withdrawal symptoms (F2), despite the face that this group contained heavy cocaine users (as shown by other variables). Most of the subjects (99.6%) in this group reported using cocaine in larger amounts or over a longer period than indended (F3). Moreover, this group had the highest percentage of subjects (96%) who spent a great amount of time in activities related to cocaine (F5). Group 2 had the lowest percentage of subjects (23.6%) with tolerance to cocaine (F1), but the highest percentage of subjects (92.92%) who used cocaine despite knowledge that problems were likely caused by cocaine (F7). Subjects in Group 3 endorsed all of the seven CD criteria with similar percentages (87.04% was the lowest and 90.60% was the highest among the seven criteria). Group 4 had the lowest percentage of subjects who endorsed F3, F5 and F7. Group 4 also had the highest percentage of subjects with persistent desire or unsuccessful attempts to cut down their cocaine use (F4). Group 1 and Group 2 had a similar transition time from first cocaine use to first CD diagnosis

(F13): 8.01 years for Group 1 and 8.07 years for Group 2. These were significantly shorter than the transition time in Groups 3 and 4, which were 12.15 years and 15.19 years, respectively.

Conclusion

We developed an approach to identify composite traits from multivariate phenotypes that are highly heritable, as estimated using genome-wide SNPs. The trait we derived is in the form of a linear combination of variables related to the phenotype, that is $\mathbf{y} = \mathbf{X}\mathbf{w}$. A quadratic optimization problem was formulated, in which optimal $\mathbf{w}$ was sought to optimize the log likelihood for estimating variance components in REML. In this formulation, variance components are set to their ideal values with the additive genetic variance component σ_g^2 equal to 1 and other components equal to 0. To avoid overfitting, we incorporated a regularization term in our formulation. An efficient algorithm based on the sequential quadratic programming framework was developed to solve the proposed optimization problem. We evaluated the proposed approach on both synthetic and real world data. The empirical results demonstrate the effectiveness of our approach as a means to identify traits with much higher chip h^2 than commonly-used disease phenotypes.

In this paper, the pairwise genetic relationship among subjects was estimated from genome-wide SNPs. However, it can also be estimated from SNPs restricted to a specific region, such as on a particular chromosome or in genes related to a pathway, to explore the genetic architecture of a trait. When SNPs within a specific region are used, the trait resulting from the proposed approach will achieve the maximized genetic variance component corresponding to this region. In an application, such as substance dependence, there are known pathways involved, so it may be of utility to determine whether there is a composite trait, the variance of which can be largely explained by the variants within the pathways. This will be a future application of our approach.

Competing interests

The authors Jiangwen Sun and Jinbo Bi declare that they have no competing interests. Henry R Kranzler has served as a consultant or advisory board member for the following companies: Alkermes, Lilly, Lundbeck, Otsuka, and Pfizer. He is a member of the Alcohol Clinical Trials Group of the American Society of Clinical Psychopharmacology, which is supported by Abbvie, Alkermes, Ethypharm, Lilly, Lundbeck, and Pfizer.

Author's contributions

JB, HRK and JS designed the overall study of finding heritable disease subtypes. JS and JB designed the new algorithm discussed in this paper. JS implemented the algorithm and performed the data analyses. HRK provided the aggregated drug dependence data and helped in interpreting the results. JS and JB wrote the first draft of this manuscript.

Acknowledgements

This work was supported by NIH grant R01DA037349, NSF grant DBI-1356655, and IIS-1447711. Jinbo Bi was also supported by NSF grants IIS-1320586 and IIS-1407205. Henry R. Kranzler was also supported by NIH grants R01DA030976, R01AA021164, R01AA023192, and R01CA184315.
We thank Joel Gelernter, M.D. from Yale University who was instrumental in recruiting, characterizing, and genotyping the subjects in the SSADDA dataset used here. Kathleen Brady, M.D., Ph.D. of the Medical University of South Carolina, Roger Weiss, M.D., of McLean Hospital and Harvard Medical School, and David Oslin, M.D., of the University of Pennsylvania Perelman School of Medicine oversaw study recruitment at their respective sites. That work was supported by NIH grants R01AA011330, R01AA017535, R01DA012690, R01DA018432, and R01DA012849.
Publication costs for this article were funded by NIH grant R01DA037349.

Author details

[1]Department of Computer Science and Engineering, University of Connecticut, 371 Fairfield Way, U-4155, Storrs, CT, 06269 USA. [2]Treatment Research Center, University of Pennsylvania Perelman School of Medicine, 3900 Chestnut Street Philadelphia, PA, 19104 USA.

References

1. Gelernter, J., Sherva, R., Koesterer, R., Almasy, L., Zhao, H., Kranzler, H.R., Farrer, L.: Genome-wide association study of cocaine dependence and related traits: Fam53b identified as a risk gene. Mol Psychiatry (2013)
2. Gelernter, J., Kranzler, H.R., Sherva, R., Koesterer, R., Almasy, L., Zhao, H., Farrer, L.A.: Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways. Biol Psychiatry **76**(1), 66–74 (2014)
3. Treutlein, J., Rietschel, M.: Genome-wide association studies of alcohol dependence and substance use disorders. Curr Psychiatry Rep **13**(2), 147–55 (2011)
4. Pierucci-Lagha, A., Gelernter, J., Chan, G., Arias, A., Cubells, J.F., Farrer, L., Kranzler, H.R.: Reliability of dsm-iv diagnostic criteria using the semi-structured assessment for drug dependence and alcoholism (ssadda). Drug Alcohol Depend **91**(1), 85–90 (2007)
5. Balding, D.J., Bishop, M.J., Cannings, C.: Handbook of Statistical Genetics, 3rd edn., p. . John Wiley & Sons, Chichester, England ; Hoboken, NJ (2007)
6. de los Campos, G., Gianola, D., Allison, D.B.: Predicting genetic predisposition in humans: the promise of whole-genome markers. Nature Reviews Genetics **11**(12), 880–886 (2010)
7. Meuwissen, T.H., Hayes, B.J., Goddard, M.E.: Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**(4), 1819–1829 (2001)
8. Yang, J., Montgomery, G.W., Goddard, M.E., Visscher, P.M., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G.: Common SNPs explain a large proportion of the heritability for human height. Nature Genetics **42**(7), 565 (2010)
9. Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B., de los Campos, G.: Beyond missing heritability: prediction of complex traits. PLoS Genetics **7**(4), 1002051 (2011)
10. Hill, W.G., Wray, N.R.: Heritability in the genomics era - concepts and misconceptions. Nature Reviews Genetics **9**(4), 255–266 (2008)
11. Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M.: Gcta: a tool for genome-wide complex trait analysis. Am J Hum Genet **88**(1), 76–82 (2011)
12. Speed, D., Hemani, G., Johnson, M.R., Balding, D.J.: Improved heritability estimation from genome-wide SNPs. American Journal of Human Genetics **91**(6), 1011–1021 (2012)
13. Kranzler, H.R., Wilcox, M., Weiss, R.D., Brady, K., Hesselbrock, V., Rounsaville, B., Farrer, L., Gelernter, J.: The validity of cocaine dependence subtypes. Addict Behav **33**(1), 41–53 (2008)
14. Bi, J., Gelernter, J., Sun, J., Kranzler, H.R.: Comparing the utility of homogeneous subtypes of cocaine use and related behaviors with dsm-iv cocaine dependence as traits for genetic association analysis. Am J Med Genet B Neuropsychiatr Genet **165B**(2), 148–56 (2014)
15. Sun, J., Bi, J., Chan, G., Oslin, D., Farrer, L., Gelernter, J., Kranzler, H.R.: Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors. Addictive Behaviors (2012)
16. Gelernter, J., Panhuysen, C., Wilcox, M., Hesselbrock, V., Rounsaville, B., Poling, J., Weiss, R., Sonne, S., Zhao, H., Farrer, L., Kranzler, H.R.: Genomewide linkage scan for opioid dependence and related traits. American Journal of Human Genetics. **78**(5), 759–769 (2006)
17. Babor, T.F., Caetano, R.: Subtypes of substance dependence and abuse: implications for diagnostic classification and empirical research. Addiction (Abingdon, England) **101**, 104–10 (2006)
18. Hu, V.W., Addington, A., Hyman, A.: Novel autism subtype-dependent genetic variants are revealed by quantitative trait and subphenotype association analyses of published gwas data. PloS ONE **6**(4), 19067 (2011)
19. Ott, J., Rabinowitz, D.: A principal-components approach based on heritability for combining phenotype information. Hum Hered **49**(2), 106–11 (1999)
20. Wang, Y., Fang, Y., Jin, M.: A ridge penalized principal-components approach based on heritability for high-dimensional data. Hum Hered **64**(3), 182–91 (2007)
21. Klei, L., Luca, D., Devlin, B., Roeder, K.: Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet Epidemiol **32**(1), 9–19 (2008)
22. Oualkacha, K., Labbe, A., Ciampi, A., Roy, M.A., Maziade, M.: Principal components of heritability for high dimension quantitative traits and general pedigrees. Statistical Applications in Genetics and Molecular Biology **11**(2) (2012)
23. Sun, J., Bi, J., Kranzler, H.R.: Quadratic optimization to identify highly heritable quantitative traits from complex phenotypic features. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '13, pp. 811–819. ACM, New York, NY, USA (2013)
24. Patterson, H.D., Thompson, R.: Recovery of inter-block information when block sizes are unequal. Biometrika **58**(3), 545–554 (1971)
25. Verbyla, A.P.: A conditional derivation of residual maximum likelihood. Australian Journal of Statistics **32**(2), 227–230 (1990). doi:10.1111/j.1467-842X.1990.tb01015.x
26. Vapnik, V.N.: An overview of statistical learning theory. IEEE Transactions on Neural Networks **10**(5), 988–999 (1999)
27. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, New York (2006)
28. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York (2006)
29. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of population structure using multilocus genotype data. Genetics **155**(2), 945–59 (2000)
30. Yang, J., Benyamin, B., ..., Visscher, P.M.: Common snps explain a large proportion of the heritability for human height. Nat Genet **42**(7), 565–9 (2010)
31. Hesselbrock, V.M., Hesselbrock, M.N.: Are there empirically supported and clinically useful subtypes of alcohol dependence? Addiction **101**(Suppl 1), 97–103 (2006)

Figures

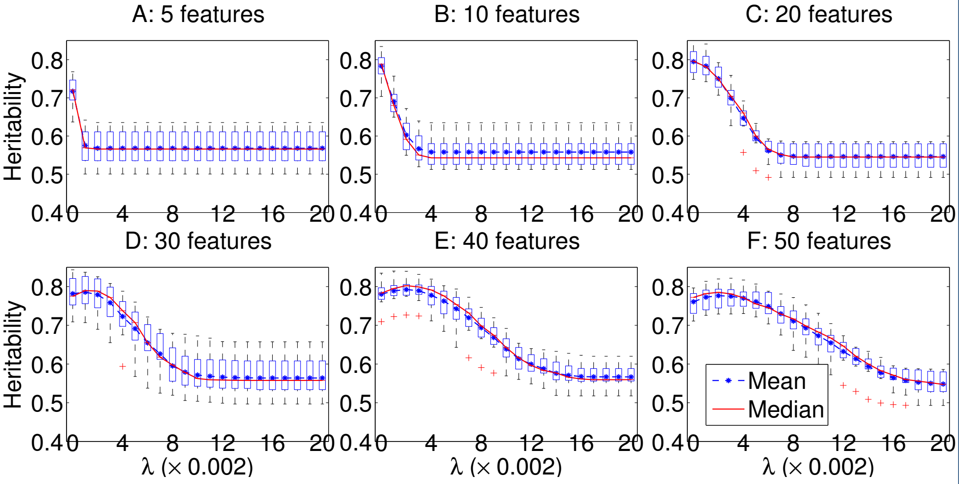

Figure 1 Simulation study: the **testing** h^2 of the quantitative traits developed in three-fold cross validation with varying λ and total number of phenotypic features in setting 1, in which datasets are simulated without covariate effect on the phenotypic features.

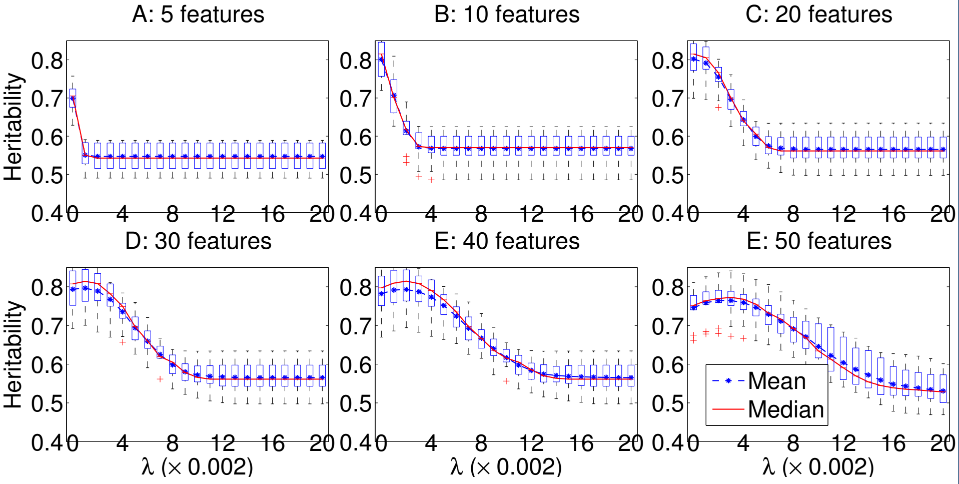

Figure 2 Simulation study: the **testing** h^2 of the quantitative traits developed in three-fold cross validation with varying λ and total number of phenotypic features in setting 2, in which datasets are simulated with covariate effect on the phenotypic features.

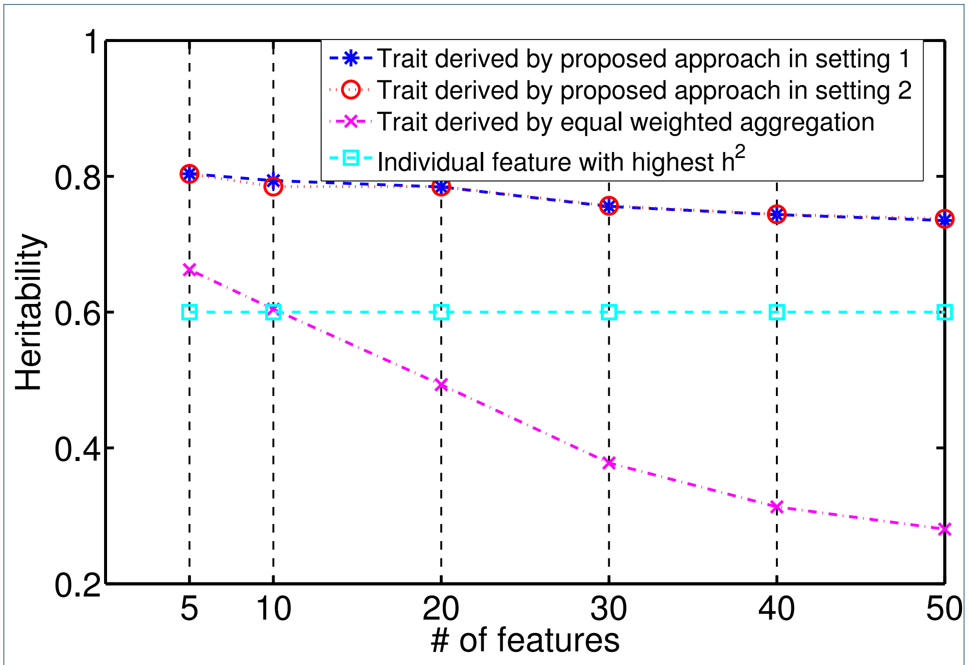

Figure 3 Simulation study: comparison of h^2 of commonly used traits with that of the quantitative traits derived by the proposed approach with chosen λ's. Results from both settings (setting 1 - without covariates, setting 2 - with covariates) are shown. For h^2 of individual features in each dataset, we show the highest. Because in our simulation the feature that has the highest h^2 estimate is shared across all the datasets with varying number of features, the corresponding h^2 curve is a straight line. Since covariates do not affect the h^2 estimate when their effect is properly corrected for during the estimation, we do not differentiate the two settings and show one curve for traits derived by equal weighted aggregation of individual features in the data and one for individual feature that has the highest h^2 estimate.

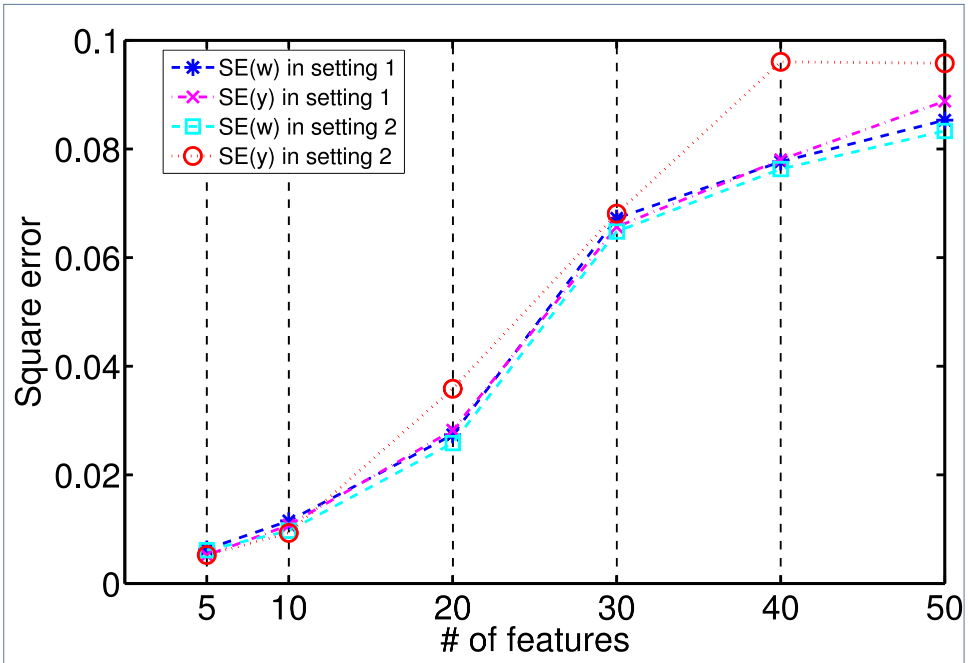

Figure 4 Simulation study: the square error of feature weights $(\mathbf{w})$ in models derived by the proposed approach with chosen λ's: SE$(\mathbf{w})$ and that of resulted quantitative traits $(\mathbf{y})$: SE$(\mathbf{y})$, comparing to the optimal (implanted) model coefficients $(\hat{\mathbf{w}})$ and traits $(\hat{\mathbf{y}})$. SE$(\mathbf{w})$ is calculated as $\|\mathbf{w}-\hat{\mathbf{w}}\|_2^2$; and SE$(\mathbf{y})$ is calculated as $\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2/n$, where n is the total number of subjects in the data. Results from both settings (setting 1 - without covariates, setting 2 - with covariates) are shown.

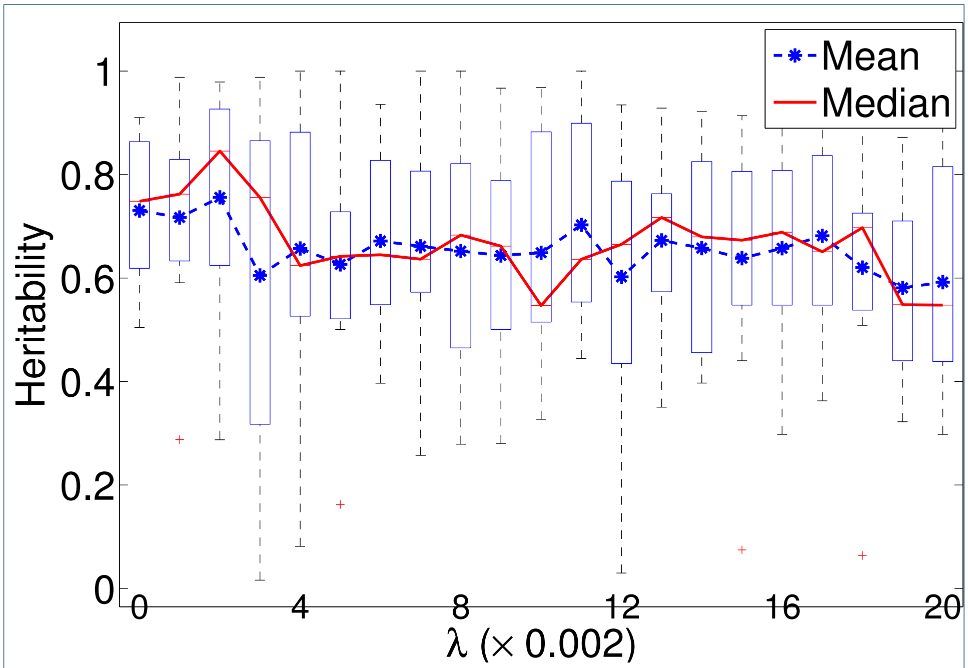

Figure 5 Case study on cocaine dataset: the testing h^2 of the composite traits derived in three-fold cross validation with varying λ.

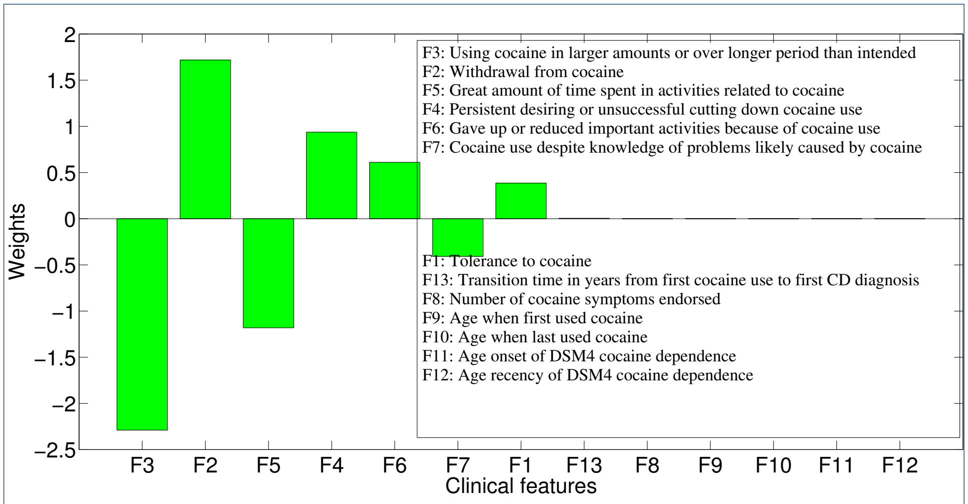

Figure 6 Case study on cocaine dataset: the weights of variables in the linear combination learned by the proposed method with $\lambda = 4$ and the entire sample set.

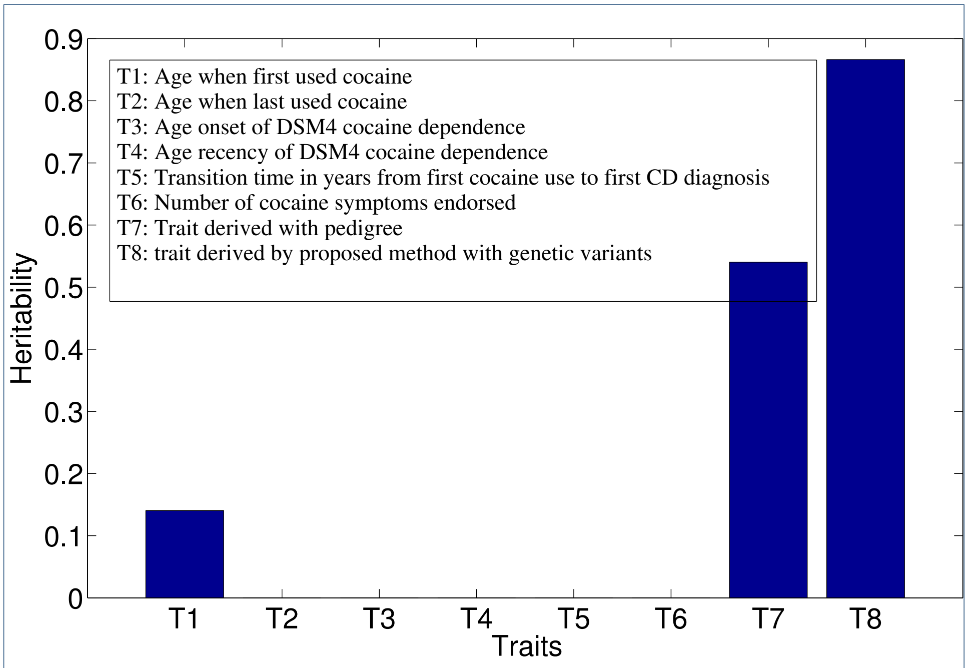

Figure 7 Case study on cocaine dataset: comparison of h^2's of six individual continuous variables in the data, composite traits derived by the approach in [23] and that derived by the proposed approach with $\lambda = 2$.

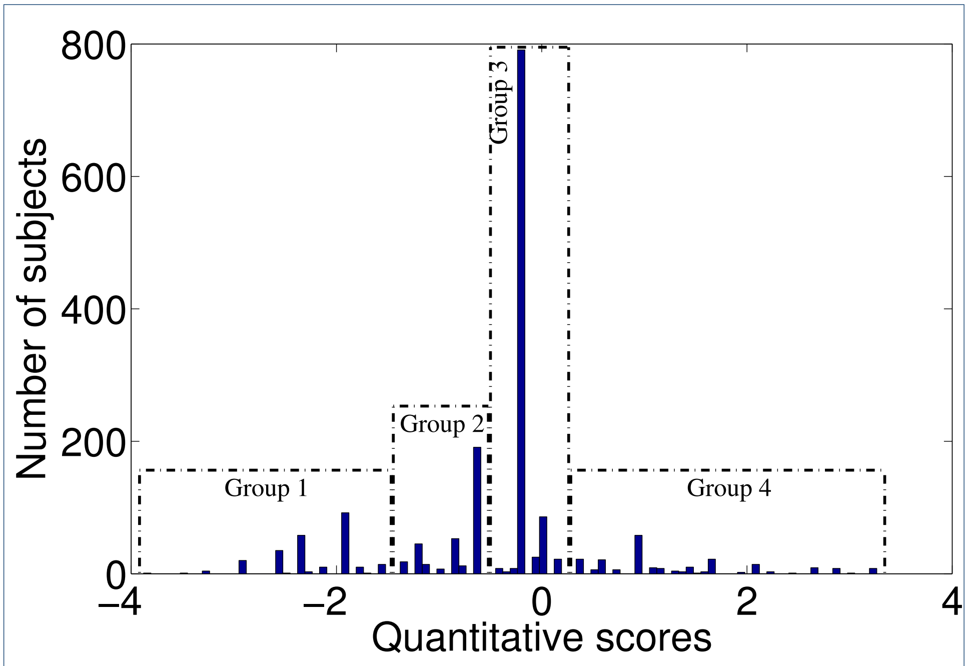

Figure 8 Case study on cocaine dataset: the distribution of scores for the composite trait derived by the proposed method with $\lambda = 2$ and the entire sample set.

Table 1 Characteristic of the three subject groups on important clinical variables related to cocaine use.

Variable	Group1 250(14.27)	Group2 339(19.35)	Group3 926(52.85)	Group4 237(13.53)
Tolerance to cocaine	124(49.60)	80(23.60)	807(87.15)	123(51.90)
Withdrawal from cocaine	14(5.60)	275(81.12)	813(87.80)	192(81.01)
Using cocaine in larger amounts or over longer period than intended	249(99.60)	323(95.28)	816(88.12)	103(43.46)
Persistent desiring or unsuccessful cutting down cocaine use	223(89.20)	326(96.17)	839(90.60)	233(98.31)
Great amount of time spent in activities related to cocaine	240(96.00)	290(85.55)	823(88.88)	82(34.60)
Gave up or reduced important activities because of cocaine use	170(68.00)	212(62.54)	806(87.04)	156(65.82)
Cocaine use despite knowledge of problems likely caused by cocaine	209(83.60)	315(92.92)	817(88.23)	170(71.73)
Number of CD criteria endorsed	4.92(1.07)	5.37(1.08)	6.18(2.10)	4.47(1.50)
Age when first used cocaine	21.93(6.34)	22.32(6.51)	21.44(5.75)	22.97(8.42)
Age onset of DSM4 cocaine dependence	28.24(7.88)	28.31(7.95)	26.63(6.70)	28.32(8.06)
Transition time in years from first cocaine use to first CD diagnosis	8.01(12.05)	8.07(12.91)	12.15(21.30)	15.19(24.65)

$N(\%)$ is shown for the first seven binary variables, where N is the number of subjects who are positive on the corresponding variable within a group and % is the percentage of N in the group.

$\mu(\sigma^2)$ is shown for the last four continuous variable, where μ is the group mean and σ^2 the standard deviation.