Coupled Mixed Model for joint genetic analysis of complex disorders from independently collected data sets: application to Alzheimer's disease and substance use disorder

Haohan Wang[1], Miaofeng Liu[2], Seunghak Lee[3], Michael M. Vanyukov[4], Wei Wu[5], and Eric P. Xing[6]

1 Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
2 Montreal Institute for Learning Algorithms, Montréal, QC, Canada
3 Human Longevity Inc., San Diego, CA
4 Department of Pharmaceutical Sciences, Departments of Psychiatry, and Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA
5 Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
6 Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

* {weiwu2, epxing}@cs.cmu.edu

Abstract

In the last decade, Genome-wide Association studies (GWASs) have contributed to decoding the human genome by uncovering genetic variations associated with diseases. "Joint analysis", which involves analyzing multiple independently generated GWAS data sets, has been explored extensively as follow-ups. However, most of the analyses remain in the preliminary stage. In this paper, we propose a method, called Coupled Mixed Model (CMM), that allows performing the joint analysis of GWAS on two independently collected data sets with different phenotypes, using a set of multivariate sparse mixed models. The CMM method does not require the data sets to have the same phenotypes as it aims to infer the uncollected phenotypes using statistical learning modeling. Moreover, CMM takes account of the confounders due to population stratification, family structures, and cryptic relatedness, as well as the confounders arising during data collection that frequently appear in joint genetic studies. In this work, we verify the performance of our method using simulation experiments. We also conduct a real data joint analysis of two independent data sets which were generated for investigation of genetic association for Alzheimer's disease and substance use disorder, respectively. Our results reveal new insights into these diseases. A Python implementation of the software is available at: https://github.com/HaohanWang/CMM

Author summary

We present a method, namely Coupled Mixed Model, that allows user to conduct joint GWAS analysis using independently collected data sets, even when there are no shared phenotypes collected among subjects in the two data sets. Our method can automatically infer the missing phenotypes in the joint analysis using likelihood estimation given that the phenotypes are related. More importantly, joint GWAS

analyses across data sets are usually considered challenging due to the existence of confounding factors including population stratification, family structure, confounders due to data collection, or cryptic relatedness. After extensive simulation experiments, we verify the performance of our method. We further apply our method to jointly analyze common genetic factors underlying both Alzheimer's disease and substance use disorder.

Introduction

Genome-wide Association Studies (GWAS) have helped revealed about 10,000 of genetic associations between human genome and diseases [1]. With the success of GWAS on single data sets, a natural follow-up is to investigate the results by integrating multiple data sets [2], which we refer to as "joint study or analysis". A joint study may unveil genetic mechanisms which cannot be found in a single analysis [3]. For example, several recent studies have revealed overlapping genetic factors that influence multiple psychiatric disorders [4], genetic correlations between schizophrenia, ADHD, depression, and use of cigarettes and cannabis [5], and association between schizophrenia and illicit drug use [6]. Co-occurrence of substance use disorder (SUD) and psychopathology has been observed in national epidemiologic surveys [7, 8] which suggests further joint study of SUD and diseases involving cognitive dysfunction should be conducted.

However, joint genetic analysis using independently collected data sets can be extremely challenging. In addition to the issues commonly expected when analyzing single data sets such as population stratification [9], a naïve application of conventional methods for the joint analysis will almost inevitably result in false discoveries which can be caused by other confounding factors such as the confounders due to different data collection procedures. Moreover, independently collected data sets often do not share the phenotypes of interest. For example, there are barely any genetic studies that collect enough samples with both SUD and Alzheimer's disease as phenotypes, which pose challenging situations for studies that aim to find common genetic factors underlying these diseases. These problems are illustrated in detail in Fig. 1. For two data sets that are originally collected for an independent study of the red and the blue phenotype, respectively, a joint analysis aims to discover the common genetic associations for these phenotypes. However, key problems arise as to how to deal with the challenges due to i) missing information about the phenotypes, *e.g.*, as illustrated in Fig. 1, the corresponding phenotypes of the other data set (*i.e.* red phenotype for blue data set, or blue phenotype for red data set), and ii) confounding factors present in different data sets (including population stratification, family structures, cryptic relatedness, and data collection confounders). In other words, all the information that is enclosed in dashed lines in Fig. 1 needs to be inferred.

Several methodological attempts have been made towards joint analysis using different data sets with different phenotypes. The proposed methods are built on summary statistics calculated using single data sets. For example, McGeachie *et al.* conducted a post analysis after independent univariate GWAS testings on different phenotypes [10]. Giambartolomei *et al.* proposed a method which assesses whether the common genetic variants associated with different phenotypes identified using single data sets are consistent [11]. Kang *et al.* jointly analyzed multiple studies with varying environmental conditions using a meta-analysis based on a random-effect model [12]. Zhu *et al.* proposed a approach for meta-analysis using summary statistics that can account for heterogeneity from a variety of different sources. Nieuwboer *et al.* introduced the genome-wide inferred study, which takes GWAS summary statistics, phenotypic means, and covariances as inputs, and outputs an approximation result of GWAS for multiple phenotypes [13]. Turlet *et al.* presented a Multi-Trait Analysis of GWAS (MTAG), which is a generalization of standard, inverse-variance-weighted

Fig 1. Illustration of the existing challenges when conducting joint analysis on two independently collected data sets for two phenotypes. For example, two independent data sets (blue data and red data) are collected with case/control phenotypes for blue symptom and case/control phenotypes for red symptom, respectively. The goal is to estimate the coefficients of SNPs ($\beta^{(1)}$ and $\beta^{(2)}$) for blue symptom and red symptom respectively. The only prior knowledge we have is that there are some common genetic factors between blue symptom and red symptom, and therefore, the blue symptom and red symptom might be correlated in a weak manner. However, we do not have blue symptoms collected for red data set, or red symptoms collected for blue data set. More importantly, there will be population confounding factors and data collection confounding factors intervening the estimation of coefficients and these confounding factors must be estimated and accounted. This figure shows all the variables that are related to the joint analysis. The variables enclosed in dashed lines must be inferred.

meta-analysis [14]. He *et al.* introduced PleioPred, which jointly models multiple genetically correlated diseases and a variety of external information including linkage disequilibrium and diverse functional annotations [15]. Finally, Wen *et al.* proposed a Bayesian hierarchical model for co-localization analysis [16].

In this paper, we propose a method for joint association analysis of two GWAS data sets, namely Coupled Mixed Model (CMM), that aims to address all the aforementioned challenges and provide a reliable joint analysis of the data sets by inferring the missing information as illustrated in Fig. 1. In particular, CMM infers the missing phenotypes and various confounding factors with the maximum likelihood estimation derived from multiple coupled multivariate sparse mixed models.

Different from the summary statistics-based methods described above, our method models the genotypes and phenotypes of the samples directly from the original data sets to estimate the effect sizes of genetic variants. Our method is also different from the approaches for missing phenotype imputation such as [17, 18] in that our method aims to address the challenges when there are no empirical data which allows the correlation between different phenotypes to be measured, which is a common situation for independently collected data sets usually not generated for joint analysis purposes.

Using extensively simulation experiments, we compare our method with previous approaches for conducting joint analysis. Furthermore, we apply our methods to real GWAS data sets previously generated for studying substance use disorder (SUD) and Alzheimer's diseases (AD), respectively, for joint analysis.

Coupled Mixed Model

We now introduce the CMM method for joint genetic analysis of two independently collected GWAS data sets. We first present a model that deals with missing phenotypes and the confounders as discussed above and also shown in Fig. 1, and then in the next section, we present an algorithm that estimates the effect sizes of genetic variants by optimizing the objective function derived from the model.

The following are the notations we use in this work: The *subscript* denotes the *identifier of data set*, and the *superscript in parentheses* denotes the *identifier of phenotypes*. Genotypes and phenotypes are denoted as X and y, respectively. Also, n denotes the sample size, and p denotes the number of SNPs. Specifically, consider a scenario as illustrated in Fig. 1, X_1 and X_2 represent the genotypes of the samples in data sets 1 and 2 with the dimension of $n_1 \times p$ and $n_2 \times p$, respectively. $y_1^{(1)}$ and $y_1^{(2)}$ denote the vectors of phenotypes 1 and 2, respectively, of the dimension $n_1 \times 1$ for the samples in data set 1. Note that $y_1^{(2)}$ is not observed. Similarly, $y_2^{(1)}$ and $y_2^{(2)}$ denote the

vectors of phenotypes 1 and 2, respectively, of the dimension $n_2 \times 1$ for the samples in data set 2. $y_2^{(1)}$ is not observed.

Our method does not require $n_1 = n_2$. However, for the convenience of the presentation of our method, we will assume $n_1 = n_2 = n$. The case of $n_1 \neq n_2$ can be easily generalized by weighing the corresponding cost function components with $1/n_1$ and $1/n_2$, respectively. Following the similar logic, although our method can be extended to the case of generalized linear models (*e.g.* for case-control data), we introduce our method with the simplest linear models for the clearness of discussion.

Straightforwardly, we have the following relations for the scenario shown in Fig. 1:

$$\begin{aligned} y_1^{(1)} &= X_1\beta^{(1)} + u_1^{(1)} + v_1 + \epsilon_1^{(1)} \\ y_1^{(2)} &= X_1\beta^{(2)} + u_1^{(2)} + v_1 + \epsilon_1^{(2)} \\ y_2^{(1)} &= X_2\beta^{(1)} + u_2^{(1)} + v_2 + \epsilon_2^{(1)} \\ y_2^{(2)} &= X_2\beta^{(2)} + u_2^{(2)} + v_2 + \epsilon_2^{(2)} \end{aligned} \tag{1}$$

where $\epsilon_i^{(j)}$ stands for residual noises for data set i with phenotype j, and $\epsilon_i^{(j)} \sim N(0, I\sigma_\epsilon^2)$, where I is an identity matrix with the shape of $n \times n$; $u_i^{(j)}$ accounts for the confounding effects due to population stratification, family structures and cryptic relatedness in data set i with phenotype j; and v_i accounts for the confounding effects due to data collection in data set i.

We have $u_i^{(j)} \sim N(0, K_i\sigma^2_{u^{(j)}})$ for data set i with phenotype j. As observed by Devlin *et al.* [9], population stratification can cause false discoveries because there exist real associations between a phenotype and untyped SNPs that have similar allele frequencies with some typed SNPs not actually associated with the phenotype, which, as a result, can lead to false associations between the phenotype and the typed SNPs. Because these false associations due to confounders from population stratification are phenotype specific, we model $\sigma^2_{u^{(j)}}$ as phenotype-specific. Hence, although we have four different variance terms for population confounders (*i.e.* $u_1^{(1)}$, $u_1^{(2)}$, $u_2^{(1)}$, and $u_2^{(2)}$), they are only parameterized by two scalars (*i.e.,* $\sigma^2_{u^{(1)}}$ and $\sigma^2_{u^{(2)}}$). $K_i = X_iX_i^T$ is the kinship matrix, constructed following standard genetics convention. A more sophisticated construction of the kinship matrix may be used to improve detection of the signals, but these details are beyond the scope of this paper. One can refer to examples in [19–22] for more details.

On the other hand, we have $v_i \sim N(0, I\sigma^2_{v_i})$ for data set i. Because the confounders due to data collection are only related to the data collection procedure, we model $\sigma^2_{v_i}$ as data set specific.

For the independently collected data sets, we only observe $< X_1, y_1^{(1)} >$ and $< X_2, y_2^{(2)} >$. We are interested in estimating $\beta^{(1)}$ and $\beta^{(2)}$, and to achieve this goal, we also need to estimate $y_1^{(2)}$, $y_2^{(1)}$, $\sigma^2_{u^j}$, and $\sigma^2_{v_i}$ in Eq. 1.

In order to estimate $\beta^{(1)}$ and $\beta^{(2)}$, we minimize the joint negative log-likelihood function with the Laplace distribution as a prior distribution over $\beta^{(1)}$ and $\beta^{(2)}$ to account for the fact that there are only a subset of SNPs that contribute to the phenotype. Additionally, as our goal is to identify the SNPs jointly associated with phenotypes 1 and 2, we encourage the method to find common SNPs responsible for both phenotypes. Therefore, we constrain our method to find a non-empty intersection (i.e., the cardinality of the intersection is at least c, where c is a scalar, and $c > 0$) of the support of β_1 and β_2. As a result, the main cost function, which is the negative log-likelihood function of the joint distribution described in Equation 1, with ℓ_1

regularization of β and common support constraint, can be represented as follows:

$$\begin{aligned}\underset{\beta^{(1)},\beta^{(2)},t}{\operatorname{argmin}} \frac{\hat{\sigma_{22}}}{2t}||y_1^{(1)} - X_1\beta^{(1)}||_2^2 + \frac{\hat{\sigma_{11}}}{2t}||y_2^{(2)} - X_2\beta^{(2)}||_2^2 + \frac{1}{2}\log t + \lambda_1||\beta^{(1)}||_1^1 + \lambda_2||\beta^{(2)}||_1^1 \\ \text{s.t.} \quad |\text{supp}(\beta^{(1)}) \cap \text{supp}(\beta^{(2)})| \geq c\end{aligned} \tag{2}$$

where

$$\begin{aligned}\hat{\sigma_{11}} =& (y_1^{(1)})^T(y_1^{(1)}) + (\beta^{(1)})^T(X_2)^T(X_2)(\beta^{(1)}) + 2\,\text{tr}(K_1\sigma^2_{u^{(1)}}) + \sigma^2_{v_1} + \sigma^2_{v_2} \\ \hat{\sigma_{22}} =& (y_2^{(2)})^T(y_2^{(2)}) + (\beta^{(2)})^T(X_1)^T(X_1)(\beta^{(2)}) + 2\,\text{tr}(K_2\sigma^2_{u^{(2)}}) + \sigma^2_{v_1} + \sigma^2_{v_2} \\ t =& |\Sigma|\end{aligned} \tag{3}$$

where Σ is the covariance matrix defined as:

$$\Sigma = \begin{bmatrix} \hat{\sigma_{11}} & \hat{\sigma_{12}} \\ \hat{\sigma_{21}} & \hat{\sigma_{22}} \end{bmatrix}$$

and we have:

$$\begin{aligned}\hat{\sigma_{12}} = \hat{\sigma_{21}} =& (y_1^{(1)})^T(X_1)(\beta^{(2)}) + (\beta^{(1)})^T(X_2)(y_2^{(2)}) + \\ & \text{tr}(K_2\sigma^2_{u^{(1)}}) + \text{tr}(K_2\sigma^2_{u^{(2)}}) + \sigma^2_{v_1} + \sigma^2_{v_2},\end{aligned}$$

and supp$(\cdot)$ stands for the support of a vector and $|\cdot|$ stands for the cardinality.

The detailed derivation of the optimization function 2 is described in S1 File. The key steps involve replacing $y_1^{(2)}$ with $X_1\beta^{(2)}$, and replacing $y_2^{(1)}$ with $X_2\beta^{(1)}$, and then writing out the joint likelihood function of Equation 1.

Algorithm

To solve the optimization function 2, we propose a strategy as follows. We first estimate the parameters $\{\sigma^2_{u^{(1)}}, \sigma^2_{u^{(2)}}, \sigma^2_{v_1}, \sigma^2_{v_2}\}$ following existing techniques [23], then we propose an iterative updating algorithm that decouples the dependency between $\{\beta^{(1)}, \beta^{(2)}\}$ and t in the optimization function 2 and solves for $\{\beta^{(1)}, \beta^{(2)}\}$ and t. We also offer a proof to show that our iterative updating algorithm will converge (see S1 File).

Estimating variances of the population confounders and data collection confounders

Despite the complicated cost function, solving for

$$\{\sigma^2_{u^{(1)}}, \sigma^2_{u^{(2)}}, \sigma^2_{v_1}, \sigma^2_{v_2}\}$$

is relatively straightforward because we have:

$$\begin{aligned} y_1^{(1)} &= X_1\beta^{(1)} + u_1^{(1)} + v_1 \\ y_2^{(2)} &= X_2\beta^{(1)} + u_2^{(2)} + v_2 \end{aligned}$$

where $y_1^{(1)}$, X_1, $y_2^{(2)}$, and X_2 are all fully observed.

Equivalently, we have:

$$\begin{aligned} y_1^{(1)} &\sim N(X_1\beta^{(1)}, K_1\sigma^2_{u^{(1)}} + I\sigma^2_{v_1}) \\ y_2^{(2)} &\sim N(X_2\beta^{(2)}, K_2\sigma^2_{u^{(2)}} + I\sigma^2_{v_2}) \end{aligned}$$

Thus, we can use conventional methods for solving linear mixed models to estimate $\{\sigma^2_{u^{(1)}}, \sigma^2_{u^{(2)}}, \sigma^2_{v_1}, \sigma^2_{v_2}\}$ (e.g. [23,24]). K_i is usually constructed from $X_i X_i^T$ in genomic applications [21], which may introduce an "over-representing problem", since $X_i X_i^T$ has a full rank and hence represents the relationship between every pair of the samples in the data set [25]. To solve this problem, we use a truncated-rank approach proposed in [26] to reduce the rank of K. Specifically, we set the non-dominant eigenvalues of K to be zero with a simple inspection of the slope of the eigenvalues as follows: if $\frac{S_i - S_{i+1}}{S_0} \leq \frac{1}{n}$, where n is the number of samples, then we set the eigenvalue S_i to be zero.

An iterative updating algorithm for estimating effect sizes and the covariance matrix of the joint likelihood

We only have $\{\beta^{(1)}, \beta^{(2)}, t\}$ left to be estimated, however, directly estimating t is difficult since it involves in four coupled terms in Function 2. To simplify the problem, we introduce an approximation to decouple the dependencies among t, β_i and σ_{ii}, leading to a neat solution involving two steps that can be conducted iteratively until convergence. The proof of the convergence of the algorithm is in the S1 File.

Calculating t and σ_{ii} given β_i

Calculating t and σ_{ii} is straightforward because both calculations can be done analytically. The analytical form for σ_{ii} is shown in Equation 3. The analytical form of t can be derived by equating the derivative of this following convex function to zero:

$$\frac{\hat{\sigma_{22}}}{2t}||y_1^{(1)} - X_1\beta^{(1)}||_2^2 + \frac{\hat{\sigma_{11}}}{2t}||y_2^{(2)} - X_2\beta^{(2)}||_2^2 + \frac{1}{2}\log t$$

Estimating $\beta^{(i)}$ given t and σ_{ii}

Numeric methods are needed for estimating $\{\beta^{(1)}, \beta^{(2)}\}$, which we achieve by alternating the direction method of multipliers (ADMM) [27]. ADMM introduces another constraint forcing $\beta^{(1)} = \beta^{(2)}$. In general, constraining two vectors to be equal is stronger than constraining two vectors to have common support. Hence, we remove the common support constraint in Function 2. With ADMM, the optimization involves solving the following three functions iteratively:

$$\begin{aligned}\beta^{(1)} = \underset{\beta^{(1)}}{\operatorname{argmin}} \frac{\hat{\sigma_{22}}}{2t}||y_1^{(1)} - X_1\beta^{(1)}||_2^2 + \lambda_1||\beta^{(1)}||_1^1 \\ + \rho||\beta^{(1)} - \beta^{(2)}||_2^2 + \Lambda^T(\beta^{(1)} - \beta^{(2)})\end{aligned} \tag{4}$$

$$\begin{aligned}\beta^{(2)} = \underset{\beta^{(2)}}{\operatorname{argmin}} \frac{\hat{\sigma_{11}}}{2t}||y_2^{(2)} - X_2\beta^{(2)}||_2^2 + \lambda_2|\beta^{(2)}||_1^1 \\ + \rho||\beta^{(1)} - \beta^{(2)}||_2^2 + \Lambda^T(\beta^{(1)} - \beta^{(2)})\end{aligned} \tag{5}$$

$$\Lambda = \Lambda + \rho(\beta^{(1)} - \beta^{(2)}) \tag{6}$$

Function 4 and Function 5 can be solved via proximal gradient descent [28] where the ℓ_1 regularization term is regarded as the proximal operator.

Discussion

Although we have introduced the CMM method in the context of analyzing two independently collected data sets with different phenotypes, it is noteworthy that CMM can still be used for analyzing multiple data sets with two different phenotypes.

However, analyzing multiple data sets with multiple different phenotypes requires a more sophisticated design of the method which is beyond the scope of this paper.

Implementation

The implementation of Coupled Mixed Model (CMM) is available as a python software. Without installation, one can run the software with a single command line. It takes binary data in a standard Plink format for each of the two data sets as input. If there are mismatched SNPs between the data sets, CMM will use the intersection of these SNPs. We recommend the users to query CMM to identify a specific number SNPs for each data set and CMM can tune the hyperparameters according to this specified number. However, to achieve the same goal, users can also choose to specify regularization parameters (i.e., λ_1, λ_2 in Equation 2 and ρ in Equation 6). If none of these information is specified, CMM will automatically conduct cross-validation to tune parameters. The implementation is available as a standalone software[1] and will be available via the GenAMap platform [29][2].

In theory, the first step of algorithm scales as the cubic of the number of samples, and the iterative updating algorithm scales linearly as the number of samples and the number of SNPs. In practice, as we observe on two data sets with hundreds of samples and 200k SNPs, it takes CMM around a minute to converge given a set of hyperparameters on a modern server (2.30GHz CPU and 128G RAM, Linux OS), and it takes CMM up to an hour to finish the entire hyperparameter tuning process.

Simulation Experiments

To test the performance of our method, we compare CMM to several baseline approaches using simulated data sets. These approaches include:

- HG(W): Meta-analysis conducted using hypergeometric tests [10] when the two independent problems are solved by standard Wald testing with the Benjamini–Hochberg (BH) procedure [30] (The most popular approach in GWAS for a single data set).
- HG(L): Meta-analysis conducted by hypergeometric tests [10] when the two independent problems are solved by standard linear mixed model with the Benjamini–Hochberg (BH) procedure [30].
- CD: Combining data-set approach. CD directly merges two data sets into one, and use the case samples for either phenotype as the case samples for the combined data, which implicitly assumes there is only one "joint" phenotype in the combined data, and then apply the Wald test with the FDR control to identify significant SNPs. This approach can only identify the SNPs that are associated with the "joint" phenotype.
- LR: ℓ_1-regularized logistic regression. Directly apply this approach to the two independent data sets. Select the joint SNPs with the intersection of the support of two coefficients.
- AL: Adaptive Lasso, which is an extension of the Lasso that weighs the regularization term [31] (enabled by the method introduced in [32] for high-dimensional data). AL is applied to the independent data sets with the same

[1] https://github.com/HaohanWang/CMM
[2] http://genamap.org/

manner as LR. We use the logistic regression-equivalent as the phenotypes are binary.

- PL: Precision Lasso, a novel variant of the Lasso, that is designed for correlated and linearly dependent features, mostly used in genomic studies [33]. PL is applied to the independent data sets with the same manner as LR.

- JL: Joint Lasso. A method we implement in this paper for a fair comparison of our proposed CMM method. JL solves the lasso problems jointly with the constraint $\beta^{(1)} = \beta^{(2)}$ with ADMM. This step can be roughly seen as an intermediate step between the popular CD approach and our proposed method. We use the logistic regression-equivalent as the phenotypes are binary.

- CMM: Coupled Mixed Model. Our proposed method.

Data Generation

We use SimPop [34] to simulate the SNPs of the two independently collected data sets. For each data set, we simulate 10000 SNPs for 500 samples from five different populations with migration behaviors. Each population also unevenly splits into five sub-populations. Therefore, it can be seen as these 500 samples are from 25 regions (denoted as G) out of five continents. These two SNP data sets are denoted as X_1 and X_2, following the notation described in Section . X_1 and X_2 are both 500×10000 matrices.

Then we generate two sparse vectors (i.e. $\beta^{(1)}$ and $\beta^{(2)}$) as effect sizes of the association between SNPs (X_1, X_2, respectively) and phenotypes which are sampled from a Gaussian distribution $N(0, 25)$. The sparsity of these two vectors are determined by the percentage of coefficients that are non-zeros, varying in $\{0.1\%, 0.5\%, 1\%\}$. Additionally, we make sure that these two sparse vectors have an overlap of support (*i.e.,* SNPs with non-zero coefficients). The percentages of the overlapping supports range in $\{0.25, 0.5, 0.75, 0.95\}$. Hence, there are 12 sets of simulation experiments for each data set.

To introduce confounders such as population stratification and family structure, we use the aforementioned regions G to affect the phenotypes differently (the effects of these regions are denoted as γ, sampled from a Gaussian distribution $N(0, 100)$). To simulate data collection confounders, we introduce a scalar e, sampled from a Gaussian distribution $N(0, 1)$.

Finally, we have the responses as:

$$r_1^{(1)} = X_1\beta^{(1)} + G_1\gamma_1 + e_1$$
$$r_2^{(2)} = X_2\beta^{(2)} + G_2\gamma_2 + e_2$$

Then, we transform these continuous phenotypes into binary phenotypes via Bernoulli sampling with the outcome of the inverse logit function ($g^{-1}(\cdot)$) over current responses. Therefore, we have:

$$y_1^{(1)} = \mathrm{Ber}(g^{-1}(r_1^{(1)}))$$
$$y_2^{(2)} = \mathrm{Ber}(g^{-1}(r_2^{(2)}))$$

Evaluation

We run the experiments with 10 different random seeds and evaluate these methods with focuses on finding the SNPs associated with both phenotypes, as well as the SNPs

Fig 2. The ROC curves of the compared methods in terms of identifying the SNPs that are jointly associated with both phenotypes.

Fig 3. Roc curves of comparing methods in terms of identifying the SNPs that are responsible for each phenotype.

associated with one phenotype from the corresponding data set. We use the ROC curve for the visualization of the performance comparison. (We also plot the area under ROC curves in S1 Fig. - S3 Fig).

Results

Fig. 2 shows the ROC curves of the compared methods in terms of their ability to find the SNPs associated with both phenotypes. The results favor our CMM method significantly overall. After examining different rows of the plots, we can see that in comparison with other methods, the superiority of the proposed CMM is more evident when there are fewer associated SNPs in each data set, and also when there are fewer SNPs associated with both phenotypes. For example, in the first row of Fig. 2, when only 0.1% of the SNPs are associated with a phenotype, CMM's advantage is clear. However, in the last row of Fig. 2, when 1% of SNPs are associated with a phenotype, CMM barely outperforms traditional LMM methods. Fortunately, in the real world genetic analysis, such a scenario is rare.

By comparing the performances of the compared methods in different columns in Fig. 2, we can see how the overlapping SNPs (i.e., those associated with both phenotypes) affect the results. Interestingly, as the percentage of the overlapping SNPs increases, in general, the performances of all the compared methods increase.

With the clear advantage of CMM, we now proceed to discuss more about the competing methods. Comparing the multivariate methods (LR, AL, JL) and the univariate methods (Wald, LMM, CD), we notice that multivariate methods tend to perform well when there are less associated SNPs as well as less overlapping SNPs, while univariate methods favor the opposite scenarios (i.e., when there are more associated SNPs and more overlapping SNPs). In addition, JL, which can be considered as a multivariate version of CD, barely outperforms CD. Also, notably, as the overlapping SNPs increase, the performance of CD increases clearly, while that of JL does not. This result can be explained as follows: CD only aims to recover the overlapping SNPs, while JL balances between minimizing the two logistic regression cost functions and minimizing the differentiation between coefficients which may not result in a more effective recovery of the overlapping SNPs.

We also notice that LMM performs surprisingly well when there are many associated SNPs. For example, when there are 1% of associated SNPs as shown in the last row of Fig. 2, LMM performs as the second best method; however, it does not perform well with fewer associated SNPs, as shown in the first two rows. In contrast, all the other methods show a stable performance.

Furthermore, we plot the results of the ROC curves of the compared methods in term of their ability to find the associated SNPs separately for each data set, which are shown in Fig. 3, where "Phe 1" and "Phe 2" stands for the two phenotypes, respectively. Notably, it is not surprising to see that the two curves for the two phenotypes almost overlap with each other. This happens because i) the two data sets were generated by using the same data generation protocol with two different sets of parameters, and ii) the results reported are averaged results across 10 different random seeds. Moreover, we notice that, these curves exhibit the same patterns as those in Fig. 2, which is also not surprising for most competing methods when the overlapping set is only selected as the

intersection of the support of the two estimated coefficients.

The results shown in these curves, along with those in Fig. 2, indicate that CMM can identify the overlapping SNPs while also maintaining strong performance when finding associated SNPs from the two data sets independently. In this figure, we only plot the overlapping SNPs CD identifies, since CD is unable to identify any SNPs for the two phenotypes separately.

Together, these simulation results demonstrate that CMM outperforms the seven competing methods in terms of finding overlapping SNPs associated with the two phenotypes, as well as finding associated SNPs from individual data sets.

Real Data Analysis: Joint Genetic Analysis for Alzheimer's Disease and Substance Use Disorder

In the read data analysis, we apply our proposed CMM method to two GWAS data sets generated previously to investigate genetic association in Alzheimer's disease (AD) and substance use disorder (SUD). The AD data set was collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI)[3] and the SUD data set collected by the CEDAR Center at the University of Pittsburgh [4]. For the AD data set, we only used the data generated from the individuals diagnosed with AD or normal controls. There are 477 individuals in the final AD data set. For the SUD data set, we consider the samples with drug abuse history as the case group and the samples with neither drug addiction nor alcohol addiction behavior as the control group, excluding the samples with only alcohol addiction behavior (but not drug abuse history), because alcoholism is usually believed to be related with drug addiction. There are 359 patients in the final SUD data set. We also exclude the SNPs on X-chromosome following suggestions of previous studies [35]. There are 257361 SNPs in these two data sets left to be examined.

Due to the statistical limitation of selecting variables using cross-validation and information criteria in high dimensional data [33], we inquire the method for a fixed number of SNPs to be selected, following previous studies (*i.e.* [36, 37]) and the hyperparameters of our model will be tuned automatically through binary search for the set of parameters to report the number of SNPs we require. To mitigate the computation load, the algorithm will terminate the searching of hyperparameters when the number of the reported SNPs lies within 50% to 200% of the number we inquire.

We inquire for 30 SNPs selected in each data set, and CMM identifies 20 SNPs associated with SUD and 40 SNPs associated with AD when the algorithm converges. There are five SNPs associated with both SUD and AD. The identified SNPs are listed in Tables 1 and 2 for AD and SUD, respectively. Although our method identified 40 SNPs for AD, we only show the top 20 as the primary interest of the current work is to identify common SNPs associated with both of the phenotypes.

In particular, *rs2131691*, which resides in the *ANO3* gene is identified as the SNP with the most significant effect size. *ANO3* is reported to be associated with AD *in silico* [39]. To the best of our knowledge, there is no previous literature that associates *ANO3* with SUD. However, evidence has shown that mutations in *ANO3* produced functional changes that affected striatal signal transduction pathways [59, 60] and also that dysfunction of striatal signal transduction pathways is implicated in SUD [61]. Therefore, these previous observations support our finding that *rs2131691* is connected to SUD.

Also, *rs224534* which resides in *TRPV1* is identified as the 3rd SNP, and 12th SNP to be associated with AD and SUD, respectively. Previous evidence has shown that the

[3] http://adni.loni.usc.edu/
[4] http://www.pitt.edu/ cedar/

Table 1. The top SNPs that our method identifies in Alzheimer's disease study with joint study in substance use disorder The SNPs are ranked by the absolute values of their estimated effect sizes. SNPs in bold are the ones that are shared with the results in substance use disorder disorder study. The MAFs reported in the table are calculated using the case-control AD experimental data. The information of whether a SNP is located within a region of a gene is taken from the Database for Single Nucleotide Polymorphisms (dbSNP) [38], and listed in the 'Gene' column. The literature support showed supports the association between the corresponding gene and the disease either *in vivo* or *in silico* from other independent data sets. Abbreviations: AD: Alzheimer's Disease, ALC: Alcoholism, HD: Huntington's disease, SCZ: Schizophrenia, SMK: Cigarette smoking, SUD: Substance use disorder

Rank	SNP	Chr	Chr. Position	MAF	Gene	Associated Disease (Literature support)
1	**rs2131691**	**11**	**26574855**	**0.47**	***ANO3***	**AD [39]**
2	rs2999899	1	13502964	0.47	*LRRC38*	AD [40]
3	**rs224534**	**17**	**3583408**	**0.34**	***TRPV1***	**AD [41], ALC [42], SUD [43]**
4	rs2944529	10	131160663	0.41	*TCERG1L*	
5	rs2358126	2	11790153	0.47	*LPIN1*	
6	rs975731	12	66801694	0.45	*GRIP1*	SUD [44]
7	rs6063725	20	51865839	0.41		
8	**rs1709317**	**2**	**23536638**	**0.37**	***KLHL29***	**SMK [45]**
9	rs1700106	8	4243970	0.48	CSMD1	AD [46], SCZ [47, 48]
10	**rs4713797**	**6**	**34490756**	**0.48**	***PACSIN1***	**AD [49], HD [50], SUD [51]**
11	**rs1057744**	**14**	**105150705**	**0.50**	***JAG2***	**AD [52]**
12	rs1223552	6	13173601	0.49	*PHACTR1*	
13	rs473015	10	17942519	0.45		
14	rs2546078	5	127332991	0.49	*MEGF10*	AD [53, 54]
15	rs6088244	20	33522398	0.48	*CBFA2T2*	
16	rs12611086	19	34136027	0.43		
17	rs4766335	12	5059683	0.46		
18	rs11134717	5	171698478	0.46		
19	rs7308687	12	5249169	0.50		
20	rs317726	7	29081595	0.48	*CPVL*	AD [55]

positive modulation of the *TRPV1* channels can be a potential target for mitigation of AD [41], suggesting an important involvement of *TRPV1* in AD. On the other hand, [43] have also shown that *TRPV1* plays a key role in morphine addiction. Also, [42] showed that the deletion of *TRPV1* in mice altered behavioral effects of ethanol which indicates a connection between *TRPV1* and alcoholism.

rs1709317, which resides in *KLHL29*, is identified as the 8th SNP and 5th SNP associated with AD and SUD respectively. We do not find strong evidences that support the association between *KLHL29* and either AD or SUD. However, we notice that *KLHL29* has been reported as one of the top 25 small airway epithelium hypomethylated genes of smokers compared with nonsmokers [62], also reported to be associated with smoking cessation [45].

rs4713797, which resides in *PACSIN1*, is identified as the 10th SNP and 6th SNP associated with AD and SUD, respectively. *PACSIN1* has been implicated in various neurodegenerative disorders, including Huntington and Alzheimer's diseases [49, 50, 63]. Additionally, the association between *PACSIN1* and SUD has been reported in a study which investigated whether DNA methylation patterns in early life are prospectively associated with SUD in adolescence [51].

rs1057744, which resides in *JAG2*, is identified as the 11th and 16th SNP associated with AD and SUD, respectively. [52] reported the association of *JAG2* with late-onset Alzheimer's disease. However, it is not previously known that *JAG2* is associated with SUD.

Table 2. The top SNPs that our method identifies in substance use disorder study with joint study in Alzheimer's disease The SNPs are ranked by the absolute values of their estimated effect sizes. SNPs in bold are the ones that are shared with the results in Alzheimer's disease study. The MAFs reported in the table are calculated using the case-control substance use disorder experimental data. The information of whether a SNP is located within a region of a gene is taken from the Database for Single Nucleotide Polymorphisms (dbSNP) [38], and listed in the 'Gene' column. The literature support showed supports the association between the corresponding gene and the disease either *in vivo* or *in silico* from other independent data sets. Abbreviations: AD: Alzheimer's Disease, ALC: Alcoholism, HD: Huntington's disease, SCZ: Schizophrenia, SMK: Cigarette smoking, SUD: Substance use disorder

Rank	SNP	Chr	Chr. Position	MAF	Gene	Associated disease (Literature support)
1	**rs2131691**	**11**	**26574855**	**0.47**	***ANO3***	**AD [39]**
2	rs11167137	8	142312847	0.47	*TSNARE1*	SCZ [56]
3	rs7267819	20	53888742	0.48		
4	rs10759638	9	113280212	0.36	*PRPF4*	SUD [57]
5	**rs1709317**	**2**	**23536638**	**0.36**	***KLHL29***	**SMK [45]**
6	**rs4713797**	**6**	**34490756**	**0.48**	***PACSIN1***	**AD [49], HD [50], SUD [51]**
7	rs13155209	5	10788807	0.38		
8	rs755598	12	103666012	0.42	*STAB2*	SUD [58]
9	rs1997858	6	115688219	0.46		
10	rs17356935	7	106357500	0.44		
11	rs7586009	2	46871829	0.40		
12	**rs224534**	**17**	**3583408**	**0.33**	***TRPV1***	**AD [41], ALC [42], SUD [43]**
13	rs3750534	9	113297844	0.36	*RNF183*	
14	rs2790453	10	28481089	0.47		
15	rs1415640	13	44768660	0.37		
16	**rs1057744**	**14**	**105150705**	**0.49**	***JAG2***	**AD [52]**
17	rs2736100	5	1286401	0.48	*TERT*	
18	rs1398800	6	66487703	0.33		
19	rs2337158	5	168249731	0.38	*TENM2*	
20	rs10865088	2	31304987	0.32		

Together, these results suggest that our findings may reveal some novel understandings of both SUD and AD, and also that our CMM method is able to discover promising genetic variants that are associated with different phenotypes using individually collected GWAS data sets.

Conclusion

In this paper, we propose a method that can enable a joint genetic association analysis of different phenotypes using two independently collected data sets. Typically, conducting joint analysis of independently collected data sets with naïve approaches will inevitably lead to many false discoveries, because these data sets usually do not have the same phenotypes, and also many confounding effects such as population stratification, family structures, cryptic relatedness, and data collection confounders are present in the data sets.

To address the challenges in the joint analysis, we propose a novel method, Coupled Mixed Model (CMM), that can estimate the effect sizes of genetic variants and accounting for both population confounding factors and data collection confounding factors. We further present an algorithm that allows an efficient parameter estimation of the objective function derived from our model. As an essential step of our method involves decoupling of the dependency of parameters and updating them iteratively, we also present a convergence proof of our proposed iterative updating algorithm.

With extensive simulation experiments, we showed the superior performance of our methods in comparison with seven competing methods in terms of both finding the common genetic variants associated with two different phenotypes, as well as finding the genetic variants associated with each phenotype. Further, we apply our CMM method to identify the common SNPs associated with both Alzheimer's disease and SUD. CMM identified five SNPs associated with both Alzhimer's disease and SUD. Literature survey provide strong evidences to support these findings.

To the best of our knowledge, our proposed method is a novel computational tool that enables the joint analysis of two independently collected data sets while accounting for various confounders simultaneously. In the future, possible extensions include development of more sophisticated methods that allow joint association analysis of multiple phenotypes by using multiple individually collected data sets simultaneously.

Supporting information

S1 Fig. Box plot of the auROC of identifying the SNPs that are jointly responsible for both phenotypes The box plot of area under ROC curves of identifying the SNPs that are jointly responsible for both phenotypes are shown. CMM shows a superior performance over these competing methods in all these cases. The trend of these results follow the same trend we observed in the main manuscript.

S2 Fig. Box plot of the auROC of identifying the SNPs that are responsible for Phenotype 1 The box plot of area under ROC curves of identifying the SNPs that are jointly responsible for Phenotype 1 are shown. CMM shows a superior performance over these competing methods in all these cases. The trend of these results follow the same trend we observed in the main manuscript.

S3 Fig. Box plot of the auROC of identifying the SNPs that are responsible for Phenotype 2 The box plot of area under ROC curves of identifying the SNPs that are jointly responsible for Phenotype 2 are shown. CMM shows a superior performance over these competing methods in all these cases. The trend of these results follow the same trend we observed in the main manuscript.

S1 File Details in Methods The file describes 1) Instructions of Using the Software CMM, 2) Derivation of CMM, and 3) Psroof of Convergence .

Acknowledgments

The authors would like to thank Guanghui Min for the early stage discussion of the method. This work is funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. This work is also supported by the National Institutes of Health grants R01-GM093156 and P30-DA035778.

Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). The ADNI is funded by the National Institute on Aging and the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: Abbott; the Alzheimer's Association; the Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences, Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec, Inc.; Bristol-Myers Squibb Co.; Eisai, Inc.; Elan Pharmaceuticals,

Inc.; Eli Lilly and Co.; F. Hoffmann-La Roche, Ltd., and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO, Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development, LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corp.; Pfizer, Inc.; Servier; Synarc, Inc.; Takeda Pharmaceutical Co. The Canadian Institutes of Health Research is providing funds to support the ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles.

References

1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics. 2017;101(1):5–22.

2. Wu C, Wang Z, Song X, Feng XS, Abnet CC, He J, et al. Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. Nature genetics. 2014;46(9):1001–1006.

3. Mukherjee S, Thornton T, Naj A, Kim S, Kauwe J, Fardo D, et al. GWAS of the joint ADGC data set identifies novel common variants associated with late-onset Alzheimer's disease. Alzheimer's & Dementia: The Journal of the Alzheimer's Association. 2013;9(4):P550.

4. Pain O, Dudbridge F, Cardno AG, Freeman D, Lu Y, Lundstrom S, et al. Genome-wide analysis of adolescent psychotic-like experiences shows genetic overlap with psychiatric disorders. bioRxiv. 2018; p. 265512.

5. Walters RK, Adams MJ, Adkins AE, Aliev F, Bacanu SA, Batzler A, et al. Trans-ancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. bioRxiv. 2018; p. 257311.

6. Mallard TT, Harden KP, Fromme K. Genetic risk for schizophrenia influences substance use in emerging adulthood: An event-level polygenic prediction model. bioRxiv. 2018; p. 157636.

7. Grant BF, Goldstein RB, Saha TD, Chou SP, Jung J, Zhang H, et al. Epidemiology of DSM-5 alcohol use disorder: results from the National Epidemiologic Survey on Alcohol and Related Conditions III. JAMA psychiatry. 2015;72(8):757–766.

8. Grant BF, Saha TD, Ruan WJ, Goldstein RB, Chou SP, Jung J, et al. Epidemiology of DSM-5 drug use disorder: results from the National Epidemiologic Survey on Alcohol and Related Conditions–III. JAMA psychiatry. 2016;73(1):39–47.

9. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55(4):997–1004.

10. McGeachie MJ, Clemmer GL, Lasky-Su J, Dahlin A, Raby BA, Weiss ST. Joint GWAS Analysis: comparing similar GWAS at different genomic resolutions

identifies novel pathway associations with six complex diseases. Genomics data. 2014;2:202–211.

11. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS genetics. 2014;10(5):e1004383.

12. Kang EY, Han B, Furlotte N, Joo JWJ, Shih D, Davis RC, et al. Meta-analysis identifies gene-by-environment interactions as demonstrated in a study of 4,965 mice. PLoS genetics. 2014;10(1):e1004022.

13. Nieuwboer HA, Pool R, Dolan CV, Boomsma DI, Nivard MG. GWIS: genome-wide inferred statistics for functions of multiple phenotypes. The American Journal of Human Genetics. 2016;99(4):917–927.

14. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. MTAG: Multi-Trait Analysis of GWAS. bioRxiv. 2017; p. 118810.

15. Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. PLoS genetics. 2017;13(6):e1006836.

16. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS genetics. 2017;13(3):e1006646.

17. Dahl A, Iotchkova V, Baud A, Johansson Å, Gyllensten U, Soranzo N, et al. A multiple-phenotype imputation method for genetic studies. Nature genetics. 2015;47(3):466.

18. Hormozdiari F, Kang EY, Bilow M, Ben-David E, Vulpe C, McLachlan S, et al. Imputing phenotypes for genome-wide association studies. The American Journal of Human Genetics. 2016;99(1):89–103.

19. Listgarten J, Lippert C, Heckerman D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. Nature genetics. 2013;45(5):470.

20. Tucker G, Price AL, Berger B. Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. Genetics. 2014;197(3):1045–1049.

21. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nature genetics. 2014;46(2):100–106.

22. Wang H, Liu X, Xiao Y, Xu M, Xing EP; IEEE. Multiplex confounding factor correction for genomic association mapping with squared sparse linear mixed model. Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on. 2017;.

23. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nature methods. 2011;8(10):833–835.

24. Kang HM, Sul JH, Zaitlen NA, Kong Sy, Freimer NB, Sabatti C, et al. Variance component model to account for sample structure in genome-wide association studies. Nature genetics. 2010;42(4):348–354.

25. Hoffman GE. Correcting for population structure and kinship using the linear mixed model: theory and extensions. PLoS One. 2013;8(10):e75707.

26. Wang H, Aragam B, Xing EP; IEEE. Variable Selection in Heterogeneous Datasets: A Truncated-rank Sparse Linear Mixed Model with Applications to Genome-wide Association Studies. Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on. 2017;.

27. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning. 2011;3(1):1–122.

28. Parikh N, Boyd S, et al. Proximal algorithms. Foundations and Trends® in Optimization. 2014;1(3):127–239.

29. Wang H, Lengerich BJ, Lee MK, Xing EP. GenAMap on Web: Visual Machine Learning for Next-Generation Genome Wide Association Studies. submitted. 2018;.

30. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995; p. 289–300.

31. Zou H. The adaptive lasso and its oracle properties. Journal of the American statistical association. 2006;101(476):1418–1429.

32. Huang J, Ma S, Zhang CH. Adaptive Lasso for sparse high-dimensional regression models. Statistica Sinica. 2008; p. 1603–1618.

33. Wang H, J LB, Aragam B, Xing EP. Precision lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. submitted. 2018;.

34. Peng B, Kimmel M. simuPOP: a forward-time population genetics simulation environment. Bioinformatics. 2005;21(18):3686–3687.

35. Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, Hogan MF, et al. Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. The American Journal of Human Genetics. 2008;83(5):623–632.

36. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009;25(6):714–721.

37. Marchetti-Bowick M, Yin J, Howrylak JA, Xing EP. A time-varying group sparse additive model for genome-wide association studies of dynamic complex traits. Bioinformatics. 2016;32(19):2903–2910.

38. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001;29(1):308–311.

39. Briones N, Dinu V. Data mining of high density genomic variant data for prediction of Alzheimer's disease risk. BMC medical genetics. 2012;13(1):7.

40. Ramanan VK, Nho K, Shen L, Risacher SL, Kim S, McDonald BC, et al. FASTKD2 is associated with memory and hippocampal structure in older adults. Molecular psychiatry. 2015;20(10):1197–1204.

41. Jayant S, Sharma B, Sharma B. Protective effect of transient receptor potential vanilloid subtype 1 (TRPV1) modulator, against behavioral, biochemical and structural damage in experimental models of Alzheimer's disease. Brain research. 2016;1642:397–408.

42. Blednov Y, Harris R. Deletion of vanilloid receptor (TRPV1) in mice alters behavioral effects of ethanol. Neuropharmacology. 2009;56(4):814–820.

43. Nguyen TL, Kwon SH, Hong SI, Ma SX, Jung YH, Hwang JY, et al. Transient receptor potential vanilloid type 1 channel may modulate opioid reward. Neuropsychopharmacology. 2014;39(10):2414–2422.

44. Briand LA, Kimmey BA, Ortinski PI, Huganir RL, Pierce RC. Disruption of glutamate receptor-interacting protein in nucleus accumbens enhances vulnerability to cocaine relapse. Neuropsychopharmacology. 2014;39(3):759–769.

45. Murea M, Lu L, Ma L, Hicks PJ, Divers J, McDonough CW, et al. Genome-wide association scan for survival on dialysis in African-Americans with type 2 diabetes. American journal of nephrology. 2011;33(6):502–509.

46. Swaminathan S, Kim S, Shen L, Risacher SL, Foroud T, Pankratz N, et al. Genomic copy number analysis in Alzheimer's disease and mild cognitive impairment: an ADNI study. International Journal of Alzheimer's Disease. 2011;2011.

47. Håvik B, Le Hellard S, Rietschel M, Lybæk H, Djurovic S, Mattheisen M, et al. The complement control-related genes CSMD1 and CSMD2 associate to schizophrenia. Biological psychiatry. 2011;70(1):35–42.

48. Consortium SPGWASG, et al. Genome-wide association study identifies five new schizophrenia loci. Nature genetics. 2011;43(10):969–976.

49. Takano M, Yamashita T, Nagano K, Otani M, Maekura K, Kamada H, et al. Proteomic analysis of the hippocampus in Alzheimer's disease model mice by using two-dimensional fluorescence difference in gel electrophoresis. Neuroscience letters. 2013;534:85–89.

50. Modregger J, DiProspero NA, Charles V, Tagle DA, Plomann M. PACSIN 1 interacts with huntingtin and is absent from synaptic varicosities in presymptomatic Huntington's disease brains. Human molecular genetics. 2002;11(21):2547–2558.

51. Cecil C, Walton E, Smith R, Viding E, McCrory E, Relton C, et al. DNA methylation and substance-use risk: a prospective, genome-wide study spanning gestation to adolescence. Translational psychiatry. 2016;6(12):e976.

52. Zheng X, Demirci F, Barmada M, Richardson G, Lopez O, Sweet R, et al. Genome-wide copy-number variation study of psychosis in Alzheimer's disease. Translational psychiatry. 2015;5(6):e574.

53. Singh TD, Park SY, Bae Js, Yun Y, Bae YC, Park RW, et al. MEGF10 functions as a receptor for the uptake of amyloid-β. FEBS letters. 2010;584(18):3936–3942.

54. Ray A, Speese SD, Logan MA. Glial Draper rescues Aβ toxicity in a Drosophila model of Alzheimer's Disease. Journal of Neuroscience. 2017;37(49):11881–11893.

55. Shirotani K, Matsuo K, Ohtsuki S, Masuda T, Asai M, Kutoku Y, et al. A simplified and sensitive method to identify Alzheimer's disease biomarker candidates using patient-derived induced pluripotent stem cells (iPSCs). The Journal of Biochemistry. 2017;162(6):391–394.

56. Sleiman P, Wang D, Glessner J, Hadley D, Gur RE, Cohen N, et al. GWAS meta analysis identifies TSNARE1 as a novel Schizophrenia/Bipolar susceptibility locus. Scientific reports. 2013;3:srep03075.

57. Johnson C, Drgon T, Liu QR, Zhang PW, Walther D, Li CY, et al. Genome wide association for substance dependence: convergent results from epidemiologic and research volunteer samples. BMC medical genetics. 2008;9(1):113.

58. Nelson EC, Agrawal A, Heath AC, Bogdan R, Sherva R, Zhang B, et al. Evidence of CNIH3 involvement in opioid dependence. Molecular psychiatry. 2016;21(5):608–614.

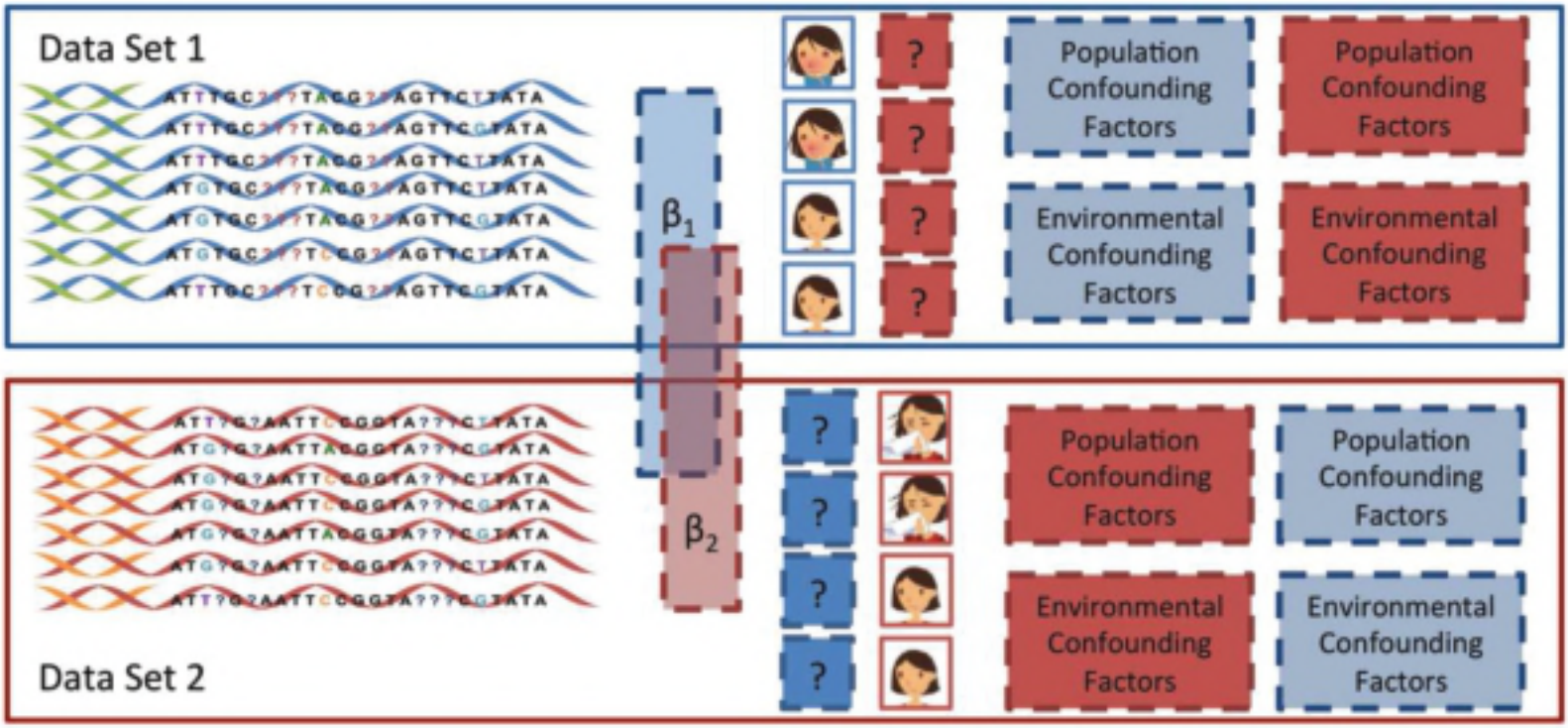
59. Charlesworth G, Plagnol V, Holmström KM, Bras J, Sheerin UM, Preza E, et al. Mutations in ANO3 cause dominant craniocervical dystonia: ion channel implicated in pathogenesis. The American Journal of Human Genetics. 2012;91(6):1041–1050.

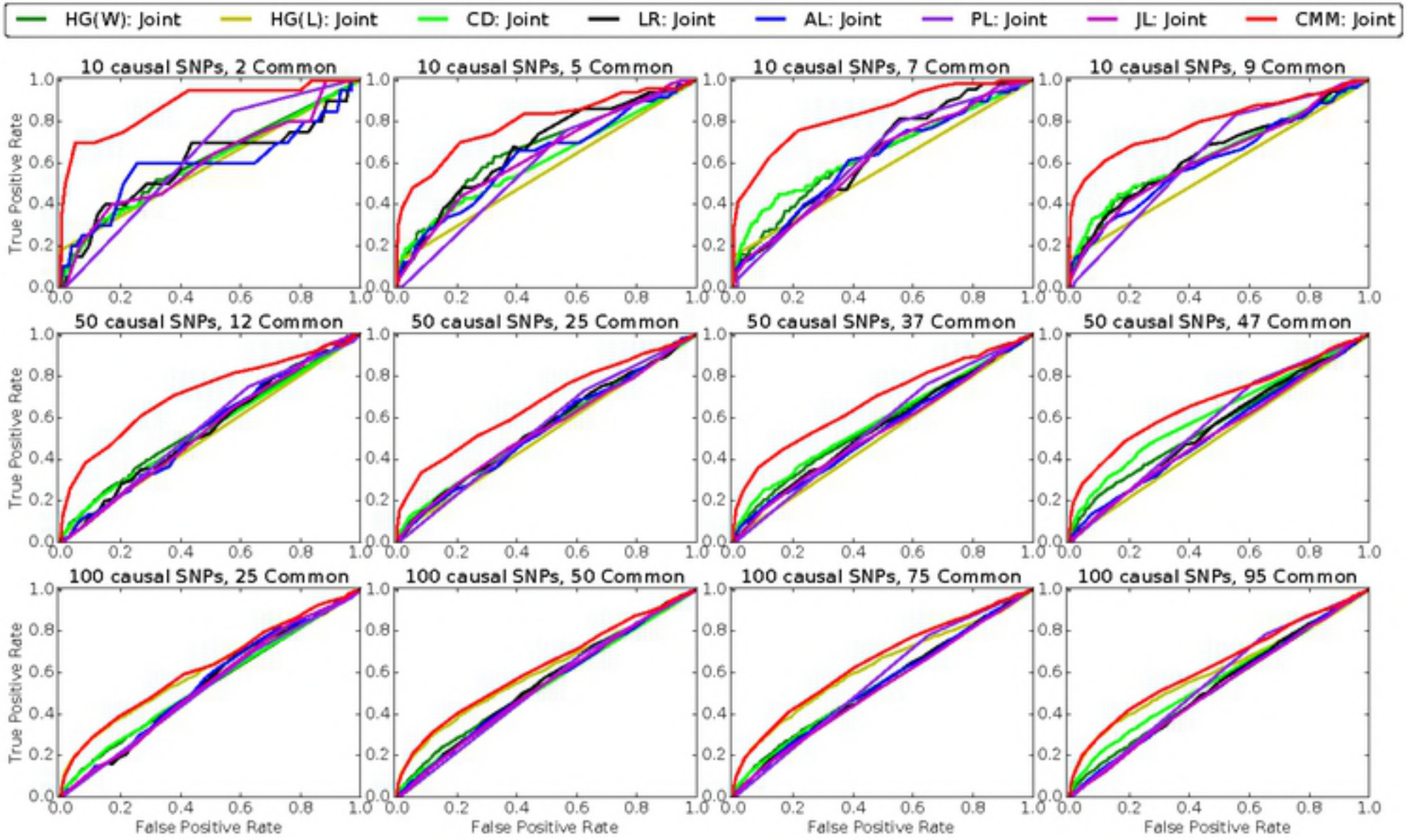
60. Karimi M, Perlmutter JS. The role of dopamine and dopaminergic pathways in dystonia: Insights from neuroimaging. Tremor and Other Hyperkinetic Movements. 2015;5.

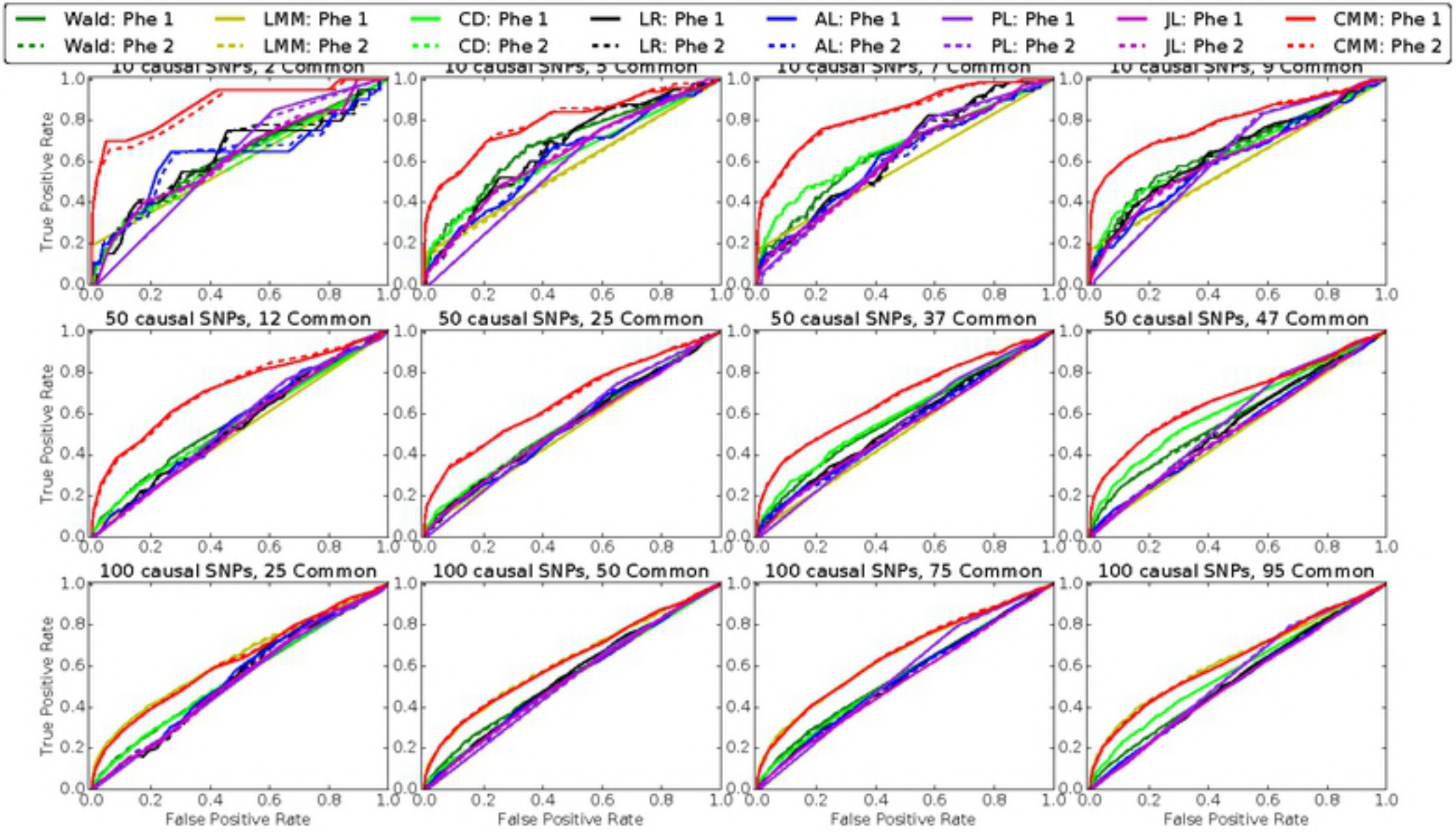
61. Philibin SD, Hernandez A, Self DW, Bibb JA. Striatal signal transduction and drug addiction. Frontiers in neuroanatomy. 2011;5.

62. Buro-Auriemma LJ, Salit J, Hackett NR, Walters MS, Strulovici-Barel Y, Staudt MR, et al. Cigarette smoking induces small airway epithelial epigenetic changes with corresponding modulation of gene expression. Human molecular genetics. 2013;22(23):4726–4738.

63. Anggono V, Koç-Schmitz Y, Widagdo J, Kormann J, Quan A, Chen CM, et al. PICK1 interacts with PACSIN to regulate AMPA receptor internalization and cerebellar long-term depression. Proceedings of the National Academy of Sciences. 2013;110(34):13976–13981.


Data Set 1
β1
β2
Population Confounding Factors
Population Confounding Factors
Environmental Confounding Factors
Environmental Confounding Factors
Population Confounding Factors
Population Confounding Factors
Environmental Confounding Factors
Environmental Confounding Factors
Data Set 2


HG(W): Joint
HG(L): Joint
CD: Joint
LR: Joint
AL: Joint
PL: Joint
JL: Joint
CMM: Joint
10 causal SNPs, 2 Common
10 causal SNPs, 5 Common
10 causal SNPs, 7 Common
10 causal SNPs, 9 Common
50 causal SNPs, 12 Common
50 causal SNPs, 25 Common
50 causal SNPs, 37 Common
50 causal SNPs, 47 Common
100 causal SNPs, 25 Common
100 causal SNPs, 50 Common
100 causal SNPs, 75 Common
100 causal SNPs, 95 Common
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Wald: Phe 1
Wald: Phe 2
LMM: Phe 1
LMM: Phe 2
CD: Phe 1
CD: Phe 2
LR: Phe 1
LR: Phe 2
AL: Phe 1
AL: Phe 2
PL: Phe 1
PL: Phe 2
JL: Phe 1
JL: Phe 2
CMM: Phe 1
CMM: Phe 2
10 causal SNPs, 2 Common
10 causal SNPs, 5 Common
10 causal SNPs, 7 Common
10 causal SNPs, 9 Common
50 causal SNPs, 12 Common
50 causal SNPs, 25 Common
50 causal SNPs, 37 Common
50 causal SNPs, 47 Common
100 causal SNPs, 25 Common
100 causal SNPs, 50 Common
100 causal SNPs, 75 Common
100 causal SNPs, 95 Common
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0