Wang *et al. BMC Medical Genomics* 2020, **13**(Suppl 3):19
https://doi.org/10.1186/s12920-020-0667-4

BMC Medical Genomics

TECHNICAL ADVANCE Open Access

Discovering weaker genetic associations guided by known associations

Haohan Wang[1], Michael M. Vanyukov[2], Eric P. Xing[1,3] and Wei Wu[4*]

From Joint 30th International Conference on Genome Informatics (GIW) & Australian Bioinformatics and Computational Biology Society (ABACBS) Annual Conference
Sydney, Australia. 9–11 December 2019

Abstract

Background: The current understanding of the genetic basis of complex human diseases is that they are caused and affected by many common and rare genetic variants. A considerable number of the disease-associated variants have been identified by Genome Wide Association Studies, however, they can explain only a small proportion of heritability. One of the possible reasons for the missing heritability is that many undiscovered disease-causing variants are weakly associated with the disease. This can pose serious challenges to many statistical methods, which seems to be only capable of identifying disease-associated variants with relatively stronger coefficients.

Results: In order to help identify weaker variants, we propose a novel statistical method, Constrained Sparse multi-locus Linear Mixed Model (CS-LMM) that aims to uncover genetic variants of weaker associations by incorporating known associations as a prior knowledge in the model. Moreover, CS-LMM accounts for polygenic effects as well as corrects for complex relatednesses. Our simulation experiments show that CS-LMM outperforms other competing existing methods in various settings when the combinations of MAFs and coefficients reflect different scenarios in complex human diseases.

Conclusions: We also apply our method to the GWAS data of alcoholism and Alzheimer's disease and exploratively discover several SNPs. Many of these discoveries are supported through literature survey. Furthermore, our association results strengthen the belief in genetic links between alcoholism and Alzheimer's disease.

Keywords: Weak association, Linear mixed model, GWAS

Background

Genome Wide Association Studies (GWAS) have allowed people to address one of the most fundamental tasks in genetic research, which is to uncover associations between genetic variants and complex traits. Many efforts have been made which employ traditional statistical testing methods such as the Wald test to test the association of each individual SNP with a certain human disease, yet there are still a large amount of missing heritability to be discovered [1], which is due to the relatively low statistical power of these methods. In order to increase the power of the association mapping, many statistical approaches have been proposed.

For example, linear regression and the Lasso variants have been introduced to account for polygenic effects commonly seen in complex human diseases [2, 3]. Following the success of Lasso methods, the Adaptive Lasso with the oracle property under some regularity conditions [4], and the Precision Lasso that works with correlated and linearly dependent variables [3] were proposed.

However, a natural limitation of the Lasso-based approaches is that they do not account for confounding effects raised by population structure and other complex relatedness in the GWAS data. In order to correct such effects, linear mixed models (LMMs) have been developed and received much attention in the recent years

*Correspondence: weiwu2@cs.cmu.edu
[4]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
Full list of author information is available at the end of the article

[5, 6]. Recently, Segural *et al* introduced a multi-locus LMM that utilizes step-wise selection to model polygenetic effects [7]. Further Liu *et al* extended the multi-locus LMM by dividing the model into fixed effect model and random effect model and use them iteratively [8]. On an alternative approach, recent studies also proposed a multi-locus extension to the standard LMM to account for polygenic effects with the introduction of priors on coefficients [9, 10].

Despite the success of the aforementioned methods achieved, these methods are not effective in identifying genetic variants with weaker coefficients. Considering the current notion that many complex human diseases are likely to be caused and affected by many–rather than a few–genetic variants with small coefficients on a certain disease [11] and yet only a limited number of significant disease-associated variants have been identified from GWAS, we conjecture that the variants with small coefficients are difficult to identify given the presence of the variants with much larger coefficients, and that they will become easier to detect when conditioning on frequently reported SNPs which usually have larger coefficients. Following this belief, we propose a novel statistical method, Constrained Sparse Multi-locus Linear Mixed Model (CS-LMM), [12, 13] to uncover novel genetic variants of smaller coefficients by: 1) incorporating those frequently reported or known variants as a prior knowledge to the model, 2) accounting for polygenic association with a multivariate sparse regularized regression, and 3) correcting for population structure and complex relatedness (including family structure and other cypticx relatedness).

The performance of the CS-LMM model is evaluated using extensive simulation experiments. We also apply our CS-LMM model to an alcoholism and an Alzheimer's Disease GWAS data, with the prior knowledge of the reported SNPs associated with each disease. We identify a set of SNPs having weak associations with each disease. Most of our findings are consistent with previously published results.

Methods

We formally introduce our model named Constrained Sparse Multi-locus Linear Mixed Model (CS-LMM) that aims to uncover genetic variants with weaker associations of a disease by incorporating variants of known associations as a prior knowledge.

Model

Given frequently reported or known variants (will be called known variants later for simplicity) with relatively larger coefficients, our model CS-LMM aims to uncover novel variants of smaller coefficients. In order to achieve this, let $\mathbf{X}$ denote genotype data, $\mathbf{Z}$ denote population identification, $\mathbf{y}$ denote phenotype data (we first assume quantitative traits here, and discuss the case-control data or binary traits later), and let $\mathcal{K}$ denote the set of the variants that are known or frequently reported. The "coefficient" is mathematically defined as the coefficient of linear regression [14]. With these settings, we have our CS-LMM model formally presented as:

$$\begin{aligned}
&\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon \\
&\mathbf{u} \sim N(0, \mathbf{I}\sigma_u) \\
&\epsilon \sim N(0, \mathbf{I}\sigma_\epsilon) \\
&\text{subject to} \quad ||\beta||_1 \leq c, \\
&|\beta_i| > 0, \quad \forall i \in \mathcal{K}, \\
&|\beta_j| < |\beta_i|, \quad \forall i \in \mathcal{K}, j \notin \mathcal{K}
\end{aligned}$$

where β is the fixed genetic effects; u denotes the random population effects; ϵ is natural noise. We also introduce a constraint term $||\beta||_1 \leq c$ with the belief that only a subset of the SNPs are associated with the phenotype, where c is a constant.

Algorithm

We proceed to introduce a three-phase algorithm to estimate the parameter β, σ_u, and σ_ϵ in the CS-LMM model.

- **Step I. Fitting known variants of larger coefficients**: We first fit a linear regression model to determine the coefficients (magnitude of β_i) for the known SNPs, by solving the following equation:

$$\hat{\beta}_i = \arg\min_{\beta_i} ||\mathbf{y} - \sum_i \mathbf{X}^i \beta_i||_2^2, \quad \forall i \in \mathcal{K} \tag{1}$$

- **Step II. Correcting for population stratification and complex relatedness**: Then, we consider to estimate σ_u and σ_ϵ for population stratification. Since $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon$ ($\mathbf{u} \sim N(0, \sigma_u)$ and $\epsilon \sim N(0, \sigma_\epsilon)$) is equivalent to $\mathbf{y} \sim N(\mathbf{X}\beta, \mathbf{Z}\mathbf{Z}^T\sigma_u^2 + I\sigma_\epsilon^2)$, we can estimate the variance term with a maximum likelihood estimation of Gaussian distribution by maximizing the following:

$$l(\sigma_u, \sigma_\epsilon | \mathbf{y}', G) \propto N(\mathbf{y}' - \bar{\mathbf{y}'} | 0, \sigma_u^2 \mathbf{Z}\mathbf{Z}^T + \sigma_\epsilon^2 \mathbf{I}) \tag{2}$$

where $\bar{\mathbf{y}'}$ is the empirical mean of y' that is calculated by

$$\mathbf{y}' = \mathbf{y} - \sum_i \mathbf{X}^i \hat{\beta}_i \tag{3}$$

and $\mathbf{Z}\mathbf{Z}^T$ is the genomic relationship matrix that is estimated as $\mathbf{Z}\mathbf{Z}^T = (\mathbf{X}^j)(\mathbf{X}^j)^T$, following the convention [15].
We then solve Eq. 2 for σ_u and σ_ϵ, where we can adopt the trick of introducing $\delta = \frac{\sigma_\epsilon^2}{\sigma_u^2}$ to replace σ_u^2 for more efficient optimization [16].

Finally, we can correct the population stratification by rotating the original data:

$$\tilde{\mathbf{X}}^j = (\text{diag}(\Gamma) + \delta\mathbf{I})^{-\frac{1}{2}}\mathbf{V}^T\mathbf{X}^j$$
$$\tilde{\mathbf{y}}' = (\text{diag}(\Gamma) + \delta\mathbf{I})^{-\frac{1}{2}}\mathbf{V}^T\mathbf{y}'$$

where $\mathbf{Z}\mathbf{Z}^T = \mathbf{U}\Gamma\mathbf{V}^T$ is the singular value decomposition.

- **Step III. Fitting variants with smaller coefficients**: Finally, we try to use the rest SNPs to explain the residual phenotypes, with solving the following:

$$\hat{\beta}_j = \arg\min_{\beta_j} ||\tilde{\mathbf{y}}' - \sum_j \tilde{\mathbf{X}}^j\beta_j||_2^2$$
$$\text{subject to} \quad |\beta_j| < \min|\beta_i|, \quad \forall j \quad \forall i$$

To solve this problem efficiently, we relax this constrain to a Lasso constrain as follows:

$$\hat{\beta}_j = \arg\min_{\beta_j} ||\tilde{\mathbf{y}}' - \sum_j \tilde{\mathbf{X}}^j\beta_j||_2^2 + \sum_j \lambda||\beta_j||_1 \quad (4)$$

This new Lasso problem is solved via proximal gradient descent [17].

Stability Selection In Step III, to achieve a stable variable selection, we follow the regime of stability selection [18]: we run the algorithm 100 times, each time with half of the data points sampled without replacement from the original data. The final selected variables are the ones that are chosen more than 75% of chances over 100 runs.

Implementation

The implementation of CS-LMM is available as a python software. Without installation, one can run the software with a single command line. It takes the Plink binary data as input. An extra file containing the known association variants is recommended. If this extra file is not available, CS-LMM will first employ standard testing methods such as Wald test to select variants with the strongest signals. In order to identify a specific number (denoted as K) of SNPs associated with the disease, users can inquire the model with the number K or with a specific weight of the regularization term (λ in Eq. 4). If neither the number of SNPs nor the regularization weight is specified, the software will estimate the parameters using cross validation. The detailed instruction on how to use the software can be found in the Additional file 1. The implementation is available as a standalone software[1]. The computational complexity and scalability scales linearly with the number of samples and SNPs.

[1] https://github.com/HaohanWang/CS-LMM

Results

Simulations

In order to evaluate the performance of CS-LMM, we compare it with several existing association methods regarding their ability to uncover weaker associations. In particular, we compare CS-LMM to the following methods:

- Standard Wald test with the standard FDR control using the Benjamini–Hochberg (BH) procedure [19]: the most popular test used in GWA studies;
- L1-regularized linear regression (i.e. the Lasso);
- Adaptive Lasso: an extension of Lasso that weighs the regularization term [4] (enabled by the method introduced in [20] for high-dimensional data);
- Precision Lasso: a novel improvement of Lasso that is more stable and consistent than Lasso [3];
- Linear mixed model: the most popular method of population stratification;
- Sparse linear mixed model (sparse LMM): a combination of sparse variable selection and population stratification [9, 21].
- Multi-locus linear mixed model (MLMM): an improvement of linear mixed model with step-wise selection to enable polygenetic modelling [7].
- Fixed and random model Circulating Probability Unification (FarmCPU): a novel extension of MLMM that iteratively uses fixed effect model and random effect model [8]

Data generation

We generate the simulation data comprehensively to reflect real world scenarios of genetic data with population structure under different minor allele frequencies (MAFs) and coefficients. We use the SimuPop [22] software to simulate the real world genomic data with population structure. We simulate p SNPs for n individuals, denoted as $\mathbf{X}$, and let $\mathbf{X}^j$ denote the j^{th} SNP. These individuals are from g populations and each population has f subpopulation.

In our simulation experiments, the SNPs come from two sets with two different MAFs: 20% of these SNPs are from one set (denoted as Set v) which has an MAF as m_v while the rest of the 80% SNPs are from the other set (denoted as Set u) which has a MAF as m_u. We assume there are k SNPs associated with the phenotype, of which, 20% are from set v and the rest are from set u.

In addition, the known SNPs in our simulation have higher MAFs and larger coefficients than the SNPs to be discovered. More specifically, for a SNP j, if $j \in k$ and $j \in v$, it simulates the SNP that is already known to be associated with the trait and it has coefficient $\beta_j = e_v c_j$. On the other hand, if $j \in k$ and $j \in u$, SNP j simulates the undiscovered associated SNP that has coefficient $\beta_j = e_u c_j$. If

$j \notin k$, SNP j simulates a SNP that is not associated with the phenotype and has the coefficient $\beta_j = 0 c_j = 0$. c_j is the base coefficient, sampled from a uniform distribution $U(0, 1)$. This simulation process is showed in Fig. 1.

We generate the associated phenotype $\mathbf{y}$ as $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where $\epsilon \sim N(0, 1)$ is the natural noise. We further transform $\mathbf{y}$ into a binary phenotype with a Binomial sampling procedure with the probability of success achieved through feeding $\mathbf{y}$ into the inverse logit function.

Following [1], we conduct experiments with a variety of the settings with different combinations of MAFs ($m_u = 0.005, 0.01$), coefficients ($e_u = 5, 10, 25$) of the SNPs to be discovered, and heritability ($0.1, 0.3, 0.5, 0.7$) of the phenotype. For the known SNPs, we keep $m_v = 0.1$ and $e_v = 50$. We choose $n = 500$, $p = 500000$, and $k = 10$ for the following experiments. For each configuration of the data, we repeat the experiments 10 times with different random seeds, and the reported result is based on the union of the results from all runs.

Evaluation

To conduct a fair comparison, we evaluate these models only regarding their ability to uncover the associated SNPs that are not already known to CS-LMM, as CS-LMM takes the known SNPs as a prior knowledge. For each method, we follow the convention to select the parameter λ (the weight of regularizer), which leads to the desired number of the selected variables (denoted as K) [3, 23]. This helps to avoid overly complex models, which tend to be selected by automatic measures such as cross validation, the Akaike information criterion (AIC), and the Bayesian information criterion (BIC) [24]. Moreover, it is known that the performance of parameter estimation and prediction are not directly coupled, e.g., as mentioned in [25] and the hyperparameter selected through cross-validation tend to report more false positives [3]. In our experiments, we select exactly $K = k$ variables.

Results

Figure 2 shows the precision-recall curve of CS-LMM compared to the Wald test, Lasso, Adaptive Lasso, Precision Lasso, LMM, sparse LMM, MLMM, and FarmCPU. The figure shows 24 experiments with three choices of coefficients (e_u) across two choices of MAFs m_u of the SNPs to be discovered, and four choices of heritability. In particular, plots in Figure 2 represent MAFs and coefficients correspond to heritability 0.1 (a), 0.3 (b), 0.5(c), and 0.7(d).

Figure 2a represents the most challenging case since the heratibility is as small as 0.1. All the methods do not behave well in this setting, and MLMM seems to have tiny advantages over other methods. Figure 2b and c illustrate the more realistic cases with heritabilities set as 0.3 and 0.5. Within this set-up, we can see CS-LMM has clear advantages over other methods. Sparse LMM and vanilla LMM are also behaving well, but still inferior to CS-LMM. Figure 2d represents a simple scenario where the heritability is 0.7. In this setting, simpler univeriate testing methods, such as Wald and LMM, can also perform well, and CS-LMM behave roughly slightly shy of these univariate testing methods. In general, CS-LMM behave better than the competing methods in most settings of the experiments.

Other experiments Other than the main experiment shown in Fig. 2, we have tested our methods in a larger range of choices of coefficients and MAF, tested the methods when we have different choices of k, and tested the methods under a larger number of samples.

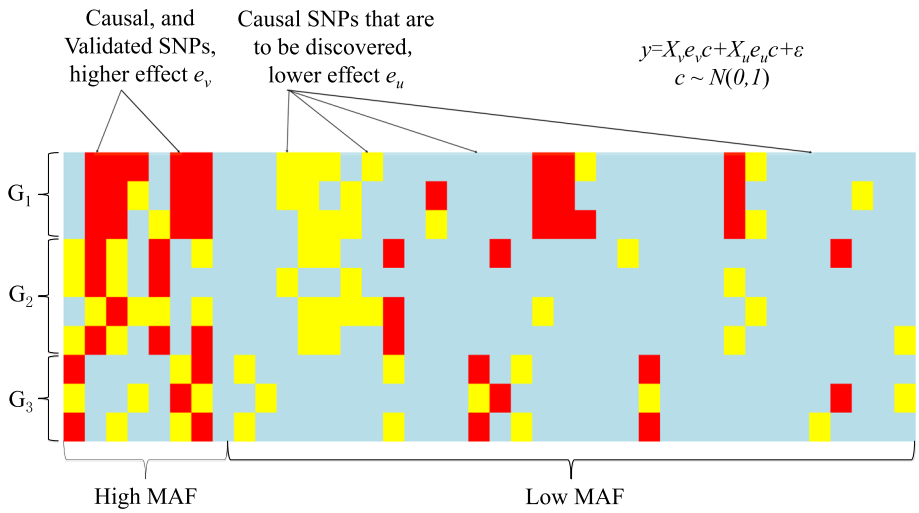

Fig. 1 An illustration of the generation process of SNP array data. This figure shows the data is generated with three populations as an example

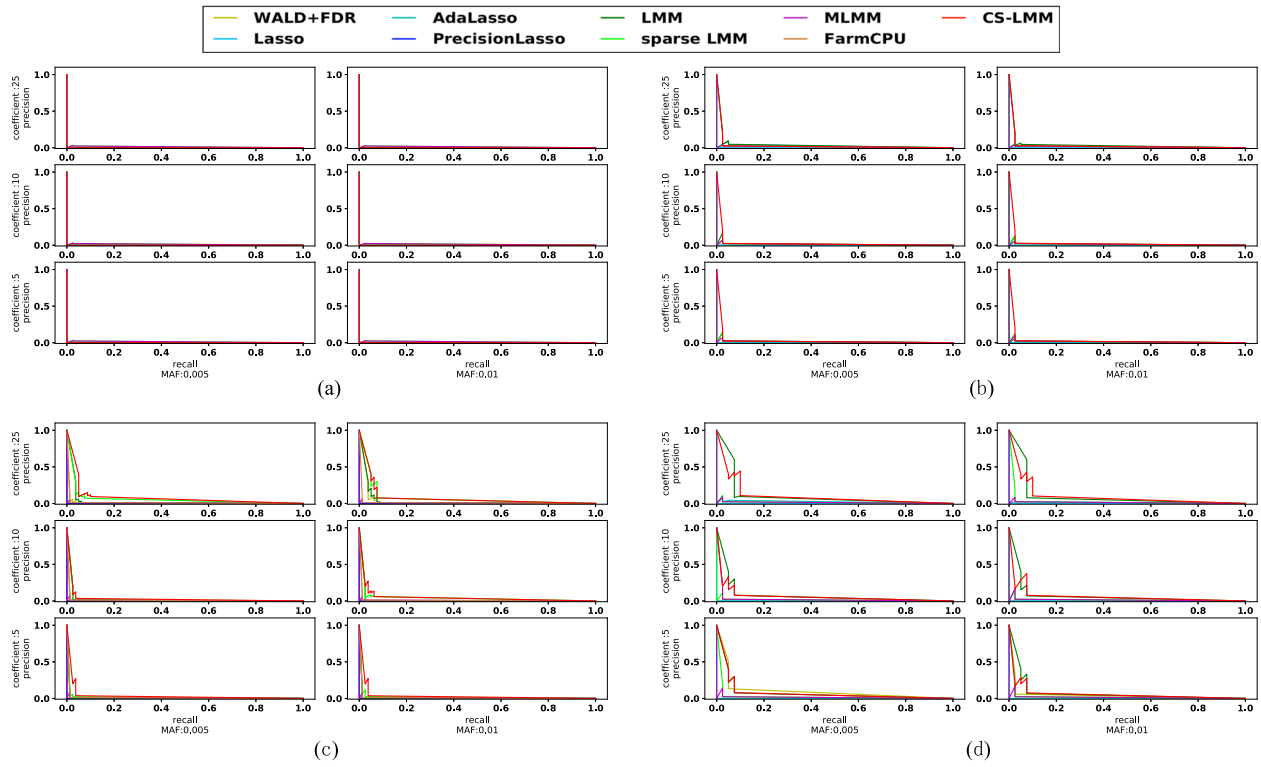

Fig. 2 Simulation results of CS-LMM compared to other models in terms of the precision-recall curve. The x-axis is recall and y-axis is precision. This figure is split into four components based on the heritability. **a**: heritability is 0.1; **b** heritability is 0.3; **c** heritability is 0.5; **d** heritability is 0.7;

We also reported other other evaluation criteria including true positives, false positives and area under ROC (auROC) under a broader setting of the experiment. There more thorough tests are included in Additional file 1: Section 4–7.

Taken together, these results show that CS-LMM outperforms other competing existing approaches in most cases, in particular, in the settings when the heratibility is at an intermediate level. Notably, these are also the settings that resemble real life scenarios for complex human diseases, and thus demonstrating the necessity and promising usages of CS-LMM in the real life.

Application to real data

Alcoholism study

We apply our method CS-LMM to the case-control GWAS data collected from subjects with and without alcoholism by The Center for Education and Drug Abuse Research (CEDAR) at the University of Pittsburgh. The data set consists of 383 individuals that include 305 subjects reported to be addicted to the consumption of alcohol through their lifetime. The data consists of 234 male subjects and 149 female subjects. The ages of these subjects range from 21 to 31. There are 519,138 genotyped SNPs in the data. The missing values are imputed as the mode of corresponding SNPs. To take the full advantage of our method, we collect the SNPs associated with alcoholism that are reported in GWAS Catalog [26] with p-values smaller than 1e-8 as the known SNPs to build in the CS-LMM model. The four SNPs we collect include: *rs1789891*, *rs7590720*, *rs2835872*, and *rs4478858*. With these known alcoholism-associated SNPs fed into CS-LMM, we run the model to uncover additional SNPs that have weaker associations with alcoholism.

We inquire 20 SNPs from the model, and CS-LMM returns 21 predicted SNPs when converges, including the 4 known SNPs we feed into the model as a prior knowledge, and thus the model discovers 17 alcoholism-associated SNPs. Table 1 lists the SNPs associated with alcoholism that are identified by CS-LMM. Since it is challenging to verify the reliability of these findings experimentally, we instead conduct a literature survey to find out whether the genes where these SNPs reside are linked to alcoholism or related disorders. Even though this type of "verification" may not provide conclusive evidence about the association between the identified SNPs and the disease, it can provide clues about whether the findings are worth further investigation.

Encouragingly, all the SNPs we discovered are linked to alcoholism, through the gene these SNPs reside in, in previously published results (shown in Table 1). For example, the 5^{th}, the 6^{th}, and the 17^{th} SNPs are within

Table 1 The top SNPs that CS-LMM identifies in an alcoholism study with four known associations

Rank	SNP	Chr	Chr Position	Est. Coe.	MAF	Gene	Disease [Literature]
1	rs1789891	4	99329262	4.2E3	0.15	*ADH1B*	ALC [27]
2	rs7590720	2	216033935	1.7E3	0.29	*PECR*	ALC [28]; AD [29]
3	rs2835872	21	37654970	1.5E3	0.25	*KCNJ6*	ALC [30]; DS [31]
4	rs4478858	1	31411078	1.4E3	0.44	*SERINC2*	ALC [32]
5	rs1789924	4	99353129	-2.2E-4	0.33	*ADH1C*	ALC [33]
6	rs698	4	99339632	-2.2E-4	0.33	*ADH1C*	ALC [33]
7	rs2851300	4	99358667	-2.2E-4	0.33		
8	rs10483038	21	37652469	-1.6E-4	0.25	*KCNJ6*	ALC [30]; DS [31]
9	rs1344694	2	216028914	-1.3E-4	0.32	*PECR*	ALC [28]; AD [29]
10	rs4147536	4	99317955	-7.6E-5	0.30	*ADH1B*	ALC [27]
11	rs12482570	21	37705475	-5.9E-5	0.28	*KCNJ6*	ALC [30]; DS [31]
12	rs857975	21	37629311	-5.8E-5	0.28	*KCNJ6*	ALC [30]; DS [31]
13	rs4147544	4	99213357	-5.7E-5	0.45	*ADH6*	ALC [34]
14	rs702860	21	37636327	-5.6E-5	0.26	*KCNJ6*	ALC [30]; DS [31]
15	rs2835853	21	37642590	-5.6E-5	0.26	*KCNJ6*	ALC [30]; DS [31]
16	rs717859	21	37640500	-5.6E-5	0.26	*KCNJ6*	ALC [30]; DS [31]
17	rs11499823	4	99353592	-5.6E-5	0.12	*ADH1C*	ALC [33]
18	rs2835910	21	37713604	-5.5E-5	0.30	*KCNJ6*	ALC [30]; DS [31]
19	rs4355398	4	99237168	-3.8E-5	0.25		
20	rs2187483	4	99212946	-9.7E-7	0.38	*ADH6*	ALC [34]
21	rs2835831	21	37614931	-6.9E-7	0.30		

The SNPs are ranked by the absolute values of their estimated coefficients. The first four SNPs with the largest coefficients in the upper panel are known SNPs that our model CS-LMM takes as prior knowledge. The rest SNPs in the lower panel are ones predicted by the model. The MAFs reported in the table are calculated using the case-control alcoholism GWAS 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) [35], and listed in the 'Gene' column. Abbreviations: ALC, Alcoholism; AD, Alzheimer's Disease; DS, Down Syndrome; Est. Coe.: Estimated Coefficients. Note that the literature support may refer to how the genes that the corresponding SNPs reside in are related to the phenotype, instead of the SNPs themselves. See discussions in Section *Alcoholism Study* for details

the region of the gene *ADH1C*, which encodes class I alcohol dehydrogenase, gamma subunit, a member of the alcohol dehydrogenase family. *ADH1C* has been shown to be associated with alcoholism in different populations [33]. Also, there are seven different SNPs residing within the region of *KCNJ6*, which encodes a member of the G protein-coupled inwardly-rectifying potassium channel. *KCNJ6* is also reported to be associated with alcoholism previously [30]. The 9th SNP resides within the region of *PECR*. Interestingly, previous evidence shows that *PECR* is not only associated with alcoholism [28], but also plays some role in Alzheimer's disease [29]. A previous study reported that the protein level of *PECR* is significantly altered in the cortical lipid rafts of the murine model of AD, compared to the control mice [29]. This result is consistent with a previous study suggesting associations between daily alcohol users and Alzheimer's patients [36].

The 10th SNP is within the region of *ADH1B*, which is also known to be related with alcoholism. The 13th SNP and the 20th SNP are within in the region of gene *ADH6*, which is also known as an alcohol dependence gene [34].

Alzheimer's disease study

Encouraged by our results from the alcoholism association mapping, we take a step further to investigate whether there is a genetic link between alcoholism and AD. We apply our method to a late-onset AD dataset provided by Harvard Brain Tissue Resource Center and Merck Research Laboratories [37]. The genotype data was generated from 540 subjects, and consists of the measurements for about 500,000 SNPs. There are 82 male subjects and 87 female subjects. The gender of the rest patients are unidentified. There are 366 subjects diagnosed with AD. The average age of these subjects is 56. The missing values are imputed as the mode of the corresponding SNPs. We use the two SNPs, *rs2075650* (gene *APOE*) and *rs157580* (gene *TOMM40*) as a prior knowledge to build into CS-LMM. These two SNPs are reported to be associated with AD with p-value less than 1e-20 in

GWAS Catalog [26]. We inquire the model for 20 SNPs that are associated with AD, and 22 SNPs are reported. The results are shown in Table 2. The reason that we use different thresholds (1e-20 for Alzheimer's disease and 1e-8 for Alcoholism) to choose SNPs are prior knowledge is mainly due to the fact that Alzheimer's disease is studied much more extensively than alcoholism in GWAS catalog, and p-values for SNPs that are reported to be associated with Alzheimer's disease tend to be smaller than those for alcoholism. We verify our findings following the same logic presented in the previous section.

Among the 19 SNPs associated with AD in Table 2, we found that the 6th SNP within gene *ABCA9* is previously reported associated with AD [41], confirming again that our method CS-LMM can identify biologically meaningful variants. Also noticeably, the 15th SNP resides within gene *ESRRG*, which encodes estrogen related receptor γ. Interestingly, evidence suggests that ERRγ plays key an role in alcohol-induced oxidative stress [42, 43]. This result also potentially verifies the existence of the pleiotropic effects between alcoholism and AD.

Since this short list of SNPs shows a promising application of CS-LMM, we also apply CS-LMM to identify a longer list of 200 SNPs for further studies. The longer list is reported in Additional file 1 (Section S2 and S3).

We also apply the competing existing methods to these two data sets, none of these methods identify a list of SNPs that are consistent with published results to the extent that CS-LMM achieves.

Discussion

We developed a novel method: Constrained Sparse multi-locus Linear Mixed Model (CS-LMM) that conditions on the associations that have already been discovered to identify disease-associated SNPs with weaker signals. Our CS-LMM model accounts for polygenic effects as well as corrects for complex relatedness such as population structure, family structure and cryptic relatedness. Our simulation experiments show that CS-LMM outperforms other competing existing methods in terms of uncovering the variants with weaker signals in various settings which reflect real life scenarios for common and rare diseases.

Table 2 The top SNPs that CS-LMM identifies in an AD study with two known associations

Rank	SNP	Chr	Chr Position	Est. Coe.	MAF	Gene	Disease [Literature]
1	rs2075650	19	44892362	0.21	0.18	*APOE*	AD [38]
2	rs157580	19	44892009	0.02	0.27	*TOMM40*	AD [39]
3	rs10027926	4	3412927	-8.3E-11	0.14	*RGS12*	SCZ [40]
4	rs12641989	4	3418113	-7.8E-11	0.14	*RGS12*	SCZ [40]
5	rs3088231	4	3420484	-7.5E-11	0.13	*RGS12*	SCZ [40]
6	rs10512523	17	69044919	5.2E-11	0.28	*ABCA9*	AD [41]
7	rs4076949	1	234066399	4.2E-11	0.18	*SLC35F3*	
8	rs874418	4	3440342	-3.9E-11	0.19	*HGFAC*	
9	rs6842419	4	3475572	-3.2E-11	0.16	*DOK7*	
10	rs16844383	4	3445516	-2.9E-11	0.21	*HGFAC*	
11	rs12131508	1	234017193	1.7E-11	0.17	*SLC35F3*	
12	rs12506821	4	3282833	-1.6E-11	0.16		
13	rs11485175	1	222437868	1.4E-11	0.23		
14	rs584507	10	6489788	1.2E-11	0.24	*PRKCQ*	
15	rs12563692	1	216818264	-1.2E-11	0.30	*ESRRG*	ALC [42, 43]
16	rs6446731	4	3283024	-1.1E-11	0.26		
17	rs7984051	13	70233817	-8.1E-12	0.25		
18	rs2327771	20	13295734	3.0E-12	0.29	*ISM1*	
19	rs7548651	1	234012812	2.4E-12	0.20	*SLC35F3*	
20	rs4330674	8	133209259	-1.2E-12	0.24	*WISP1*	
21	rs16885750	5	56578982	-8.1E-13	0.12	*C5orf67*	
22	rs938412	3	188571269	3.6E-13	0.31	*LPP*	

The SNPs are ranked by the absolute values of their estimated coefficients. The first two SNPs with largest coefficients are known SNPs the model takes as a prior knowledge. The rest are SNPs predicted by the model. The MAFs reported in the table are calculated using the AD GWAS data. The information of whether a SNP is located within a region of a gene is taken from the dbSNP. Abbreviations: ALC, Alchoholism; AD, Alzheimer's Disease; SCZ, Schizophrenia; Est. Coe.: Estimated Coefficients. Note that the literature support may refer to how the genes that the corresponding SNPs reside in are related to the phenotype, instead of the SNPs themselves. See discussions in Section *Alzheimer's Disease Study* for details.

Interestingly, in the case of 'rare variants with weak coefficients', which is categorized as the most challenging case in [1, 44], CS-LMM is superior to other competing methods. Our simulations also show that CS-LMM can particularly outperforms other methods consistently in terms of controlling false positives.

Furthermore, we apply CS-LMM to alcoholism and AD studies. For about top 20 SNPs associated with either alcoholism or AD that CS-LMM identifies, many of the SNPs reside within genes that were previously implicated in the corresponding diseases. Interestingly, our results further verify the pleiotropic effects between alcoholism and AD. The results indicate that two alcoholism-associated SNPs, *rs7590720* (previously known) and *rs1344694* (newly discovered), reside in PECR. The protein level of PECR was shown to be abnormally altered in a murine model of AD compared to the control mice, suggesting the involvement of PECR in the disease mechanism of AD. Similarly, our results also show that a novel AD-associated SNP, *rs12563692*, resides in ESRRG which encodes estrogen related receptor γ. Notably, ERRγ plays key an role in alcohol-induced oxidative stress and liver injury.

One interesting aspect regarding CS-LMM is about the three-phase learning algorithm we develop for estimating the parameters of the model. Two alternative strategies of learning the parameters are: 1) directly solving it as a convex optimization problem with explicit constrains; and 2) solving it as a standard Lasso with relaxation on the regularization on known associations. We tested these two algorithms in simulations, and our three-phase learning algorithm outperforms these two alternative strategies.

To tailor CS-LMM for case-control data or binary traits, a simple extension can be made that replaces the linear regression cost function with logistic regression cost function. Interestingly, our results indicates that CS-LMM works well with case-control data as it is (data not shown), without any extensions required. In fact, extending CS-LMM to logistic regression (or any other generalized linear models with a nontrivial link function) will affect the results adversely. For a generalized linear model, we believe CS-LMM will only function as desire when the link function is identity.

Conclusions

In summary, we have proposed and developed a novel software tool, CS-LMM, for disease association mapping which takes into account genetic variants of known associations, polygenic effects, as well as population structure and complex relatedness. The results from our simulation experiments and real data analysis demonstrate that CS-LMM can be served as an effective tool for association studies for complex human diseases.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12920-020-0667-4.

Additional file 1: Supplementary of *Discovering Weaker Genetic Associations Guided by Known Associations, with Application to Alcoholism and Alzheimer's Disease Studies*. The file has instructions of using the software and extra experimental results.

Abbreviations

AD: Alzheimer's disease; CS-LMM: Constrained sparse multi-locus linear mixed model; GWAS: Genome wide association studies; LMM: Linear mixed model; MAF: Minor allele frequency; SNP: Single nucleotide polymorphism

Acknowledgements

The authors would like to thank Steven Knopf from University of Pittsburgh for instructions in using the Alcoholism data. The authors would also like to thank Dr. Bryon Aragam from Carnegie Mellon University for early stage discussion in the designing the method and experiments.

About this supplement

This article has been published as part of BMC Medical Genomics Volume 13 Supplement 3, 2020: Proceedings of the Joint International GIW & ABACBS-2019 Conference: medical genomics (part 2). The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-13-supplement-3.

Authors' contributions

HW proposed and the idea, conducted the experiment and wrote the manuscript. MMV prepared the Alcoholism data. EPX read and wrote the manuscript. WW designed the experiment, read and wrote the manuscript. All authors read and approved the final manuscript.

Funding

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.

Availability of data and materials

The programs CS-LMM is available at https://github.com/HaohanWang/CS-LMM. The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. [2]Department of Pharmaceutical Sciences, Departments of Psychiatry, and Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA. [3]Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. [4]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.

Received: 16 November 2019 Accepted: 20 January 2020
Published: 24 February 2020

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–53.

2. Ogutu JO, Schulz-Streeck T, Piepho H-P. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proc. 2012;6(S2):. https://doi.org/10.1186/1753-6561-6-s2-s10.
3. Wang H, Lengerich BJ, Aragam B, Xing EP, Stegle O. Precision Lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. Bioinformatics. 2019;35(7):1181–7.
4. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–29.
5. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. Genetics. 2008;178(3):1709–23.
6. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010;42(4):355–60.
7. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet. 2012;44(7):825.
8. Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet. 2016;12(2):1005767.
9. Rakitsch B, Lippert C, Stegle O, Borgwardt K. A lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics. 2012;29(2):206–14.
10. Wang H, Aragam B, Xing EP. Variable selection in heterogeneous datasets: A truncated-rank sparse linear mixed model with applications to genome-wide association studies. IEEE; 2017. https://doi.org/10.1109/bibm.2017.8217687.
11. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11(6):446–50.
12. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JNP, Mott R, Flint J. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet. 2006;38(8):879–87.
13. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ, et al. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44(4):369–75.
14. Park J-H, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF, Chatterjee N. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Nat Acad Sci. 2011;108(44):18026–31.
15. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-y, Freimer NB, Sabatti C, Eskin E, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348–54.
16. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. Fast linear mixed models for genome-wide association studies. Nat Methods. 2011;8(10):833–5.
17. Parikh N, Boyd S, et al. Proximal algorithms. Found Trends® Optim. 2014;1(3):127–239.
18. Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc Ser B Stat Methodol. 2010;72(4):417–73.
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
20. Huang J, Ma S, Zhang C-H. Adaptive Lasso for sparse high-dimensional regression models. Stat Sin. 2008;Oct 1:1603–18.
21. Wang H, Yang J. Multiple confounders correction with regularized linear mixed effect models, with application in biological processes. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2016. https://doi.org/10.1109/bibm.2016.7822753.
22. Peng B, Kimmel M. simupop: a forward-time population genetics simulation environment. Bioinformatics. 2005;21(18):3686–3687.
23. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009;25(6):714–21.
24. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. Ann Stat. 2006;34(3):1436–62. https://doi.org/10.1214/009053606000000281.
25. de los Campos G, Sorensen D, Gianola D. Genomic heritability: what is it? PLoS Genet. 2015;11(5):1005048.
26. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). Nucleic Acids Res. 2017;45(D1):896–901.
27. Frank J, Cichon S, Treutlein J, Ridinger M, Mattheisen M, Hoffmann P, Herms S, Wodarz N, Soyka M, Zill P, et al. Genome-wide significant association between alcohol dependence and a variant in the adh gene cluster. Addict Biol. 2012;17(1):171–80.
28. Treutlein J, Cichon S, Ridinger M, Wodarz N, Soyka M, Zill P, Maier W, Moessner R, Gaebel W, Dahmen N, et al. Genome-wide association study of alcohol dependence. Arch Gen Psychiatr. 2009;66(7):773–84.
29. Chadwick W, Brenneman R, Martin B, Maudsley S. Complex and multidimensional lipid raft alterations in a murine model of alzheimer's disease. Int J Alzheimers Dis. 2010;2010:1–56. https://doi.org/10.4061/2010/604792.
30. Kang SJ, Rangaswamy M, Manz N, Wang J-C, Wetherill L, Hinrichs T, Almasy L, Brooks A, Chorlian DB, Dick D, et al. Family-based genome-wide association study of frontal theta oscillations identifies potassium channel gene kcnj6. Genes Brain Behav. 2012;11(6):712–9.
31. Cooper A, Grigoryan G, Guy-David L, Tsoory MM, Chen A, Reuveny E. Trisomy of the g protein-coupled k+ channel gene, kcnj6, affects reward mechanisms, cognitive functions, and synaptic plasticity in mice. Proc Nat Acad Sci. 2012;109(7):2642–7.
32. Zuo L, Wang K, Zhang X-Y, Krystal JH, Li C-SR, Zhang F, Zhang H, Luo X. Nkain1–serinc2 is a functional, replicable and genome-wide significant risk gene region specific for alcohol dependence in subjects of european descent. Drug Alcohol Depend. 2013;129(3):254–64.
33. Peng Q, Gizer IR, Wilhelmsen K, Ehlers C. Associations between genomic variants in alcohol dehydrogenase (adh) genes and alcohol symptomatology in american indians and european americans: Distinctions and convergence. Alcohol Clin Exp Res. 2017;41(10):1695–704. https://doi.org/10.1111/acer.13480.
34. Park BL, Kim JW, Cheong HS, Kim LH, Lee BC, Seo CH, Kang T-C, Nam Y-W, Kim G-B, Shin HD, et al. Extended genetic effects of adh cluster genes on the risk of alcohol dependence: from gwas to replication. Hum Genet. 2013;132(6):657–68.
35. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbsnp: the ncbi database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.
36. Zhou S, Zhou R, Zhong T, Li R, Tan J, Zhou H. Association of smoking and alcohol drinking with dementia risk among elderly men in china. Curr Alzheimer Res. 2014;11(9):899–907.
37. Zhang B, Gaiteri C, Bodea L-G, Wang Z, McElwee J, Podtelezhnikov AA, Zhang C, Xie T, Tran L, Dobrin R, et al. Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. Cell. 2013;153(3):707–20.
38. Li H, Wetten S, Li L, Jean PLS, Upmanyu R, Surh L, Hosford D, Barnes MR, Briley JD, Borrie M, et al. Candidate single-nucleotide polymorphisms from a genomewide association study of alzheimer disease. Arch Neurol. 2008;65(1):45–53.
39. Naj AC, Beecham GW, Martin ER, Gallins PJ, Powell EH, Konidari I, Whitehead PL, Cai G, Haroutunian V, Scott WK, et al. Dementia revealed: novel chromosome 6 locus for late-onset alzheimer disease provides genetic evidence for folate-pathway abnormalities. PLoS Genet. 2010;6(9):1001130.
40. Guipponi M, Santoni FA, Setola V, Gehrig C, Rotharmel M, Cuenca M, Guillin O, Dikeos D, Georgantopoulos G, Papadimitriou G, et al. Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes. PloS One. 2014;9(11):112745.
41. Piehler AP, Özcürümez M, Kaminski WE. A-subclass ATP-binding cassette proteins in brain lipid homeostasis and neurodegeneration. Front Psychiatry. 2012;3:17. https://doi.org/10.3389/fpsyt.2012.00017.
42. Kim D-K, Kim Y-H, Jang H-H, Park J, Kim JR, Koh M, Jeong W-I, Koo S-H, Park T-S, Yun C-H, et al. Estrogen-related receptor γ controls hepatic cb1 receptor-mediated cyp2e1 expression and oxidative liver injury by alcohol. Gut. 2013;62(7):1044–54. https://doi.org/10.1136/gutjnl-2012-303347.
43. Han Y-H, Kim D-K, Na T-Y, Ka N-L, Choi H-S, Lee M-O. Rorα switches transcriptional mode of errγ that results in transcriptional repression of cyp2e1 under ethanol-exposure. Nucleic Acids Res. 2016;44(3):1095–104.

44. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Rev Genet. 2008;9(5): 356–69.

Publisher's Note