Genetic Fine-Mapping With Dense Linkage Disequilibrium Blocks

Chen Mo
University of Maryland School of Medicine
Zhenyao Ye
University of Maryland School of Medicine
Kathryn Hatch
University of Maryland School of Medicine
Yuan Zhang
The Ohio State University
Qiong Wu
University of Maryland, College Park
Song Liu
Qilu University of Technology (Shandong Academy of Sciences)
Qing Lu
University of Florida
Braxton Mitchell
University of Maryland School of Medicine
L. Elliot Hong
University of Maryland School of Medicine
Peter Kochunov
University of Maryland School of Medicine
Tianzhou Ma
University of Maryland, College Park
Shuo Chen (ShuoChen@som.umaryland.edu)
University of Maryland School of Medicine

Research Article

Keywords: ℓ0 graph norm shrinkage, fine-mapping, GWAS, linkage disequilibrium, nicotine addiction, regression shrinkage

Posted Date: August 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-649530/v1

Title: Genetic fine-mapping with dense linkage disequilibrium blocks

Chen Mo[1†], Zhenyao Ye[1†], Kathryn Hatch[1], Yuan Zhang[4], Qiong Wu[5], Song Liu[6], Qing Lu[7], Braxton Mitchell[1], L. Elliot Hong[1], Peter Kochunov[1], Tianzhou Ma[3*†] and Shuo Chen[1,2*†]

[1]Maryland Psychiatric Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, Maryland, United States of America.
[2]Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, Maryland, United States of America.
[3]Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, Maryland, United States of America.
[4]Department of Statistics, College of Arts and Sciences, The Ohio State University, Columbus, Ohio, United States of America.
[5]Department of Mathematics, University of Maryland, College Park, Maryland, United States of America.
[6]School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong, China.
[7]Department of Biostatistics, University of Florida, Gainesville, Florida, United States of America.

[*]Correspondence: ShuoChen@som.umaryland.edu (SC); tma0929@umd.edu (TM)
[†]Equal contributor

Abstract

Background: Fine-mapping is an analytical step for causal prioritization of the polymorphic variants in a trait-associated genomic region observed in genome-wide association studies (GWAS). Prioritization of causal variants can be challenging due to linkage disequilibrium (LD) patterns among hundreds to thousands of polymorphisms associated with a trait. Hence, we propose an ℓ_0 graph norm shrinkage algorithm to disentangle LD patterns by *dense* LD blocks consisting of highly correlated single nucleotide polymorphisms (SNPs). We further incorporate the dense LD structure for fine-mapping. Based on graph theory, the concept of **"dense"** refers to a condition where a block is composed mainly of highly correlated SNPs. We demonstrated the application of our new fine-mapping method using a large UK Biobank (UKBB) sample related to nicotine addiction. We also evaluated and compared its performance with existing fine-mapping algorithms using simulations.
Results: Our results suggested that polymorphic variances in both neighboring and distant variants can be consolidated into dense blocks of highly correlated loci. Dense-LD outperformed comparable fine-mapping methods with increased sensitivity and reduced false-positive error rate for causal variant selection. Applying to a UKBB sample, this method replicated the loci reported in previous findings and suggested a strong association with nicotine addiction.
Conclusion: We found that the dense LD block structure can guide fine-mapping and accurately determine a parsimonious set of potential causal variants. Our approach is computationally efficient and allows fine-mapping of thousands of polymorphisms.

Keywords: ℓ_0 graph norm shrinkage; fine-mapping; GWAS; linkage disequilibrium; nicotine addiction; regression shrinkage

1 Background

1.1 Fine-mapping and existing methods

Genome-wide association studies (GWAS) are crucial to understand the associations between genetic variants and many complex biological traits such as nicotine addiction, asthma, and schizophrenia [1, 2, 3, 4, 5, 6, 7, 8]. Thousands of diseases-associated genetic loci have been discovered and replicated [2, 3, 9, 10]; however, even replicated loci may differ from the causal variants due to linkage disequilibrium (LD) among proximal and distant loci. Specifically, non-causal variants can be associated with a trait because of the high correlation with the causal variants due to LD [4, 11, 12]. Hence, a post-hoc examination of trait associations is required to distinguish likely causal variants from those associated with the trait.

A polygenic trait may be associated with multiple single nucleotide polymorphisms (SNPs), and fine-mapping allows prioritization of causal variants among the genetic variants identified by GWAS [4]. Fine-mapping algorithms are used to prioritize plausible causal variants by accounting for a complex LD structure, thus providing further insight into the underlying biological mechanisms of the trait [4, 12]. From a statistical perspective, genetic fine-mapping is a variable selection procedure that identifies parsimonious sets of variants from large numbers of correlated SNPs.

Recently, a wide range of extension and related applications of regression- shrinkage-based [13, 14, 15, 16] and Bayesian-based [17, 18, 19, 20] fine-mapping procedures have been therefore developed. Fine-mapping based on regression shrinkage methods, such as the least absolute shrinkage and selection operator (LASSO) [21] and elastic net (ENET) [22], can comprehensively and efficiently evaluate the joint effects of multiple SNPs. However, regression shrinkage methods tend to produce sparse models that lead to excessive exclusion of variants. They also face the challenge of the strong correlation among numerous SNPs. For example, selecting an appropriate sparse model using regression shrinkage methods can be difficult when an irrelevant variable is highly correlated with the true predictor. They also may lack the consistency of selection among highly correlated variables [23, 24].

Bayesian fine-mapping methods select variants by estimating the probability of an SNP being included as a causal variant in the model based on the posterior inclusion probability (PIP) [25, 26]. As a result, it is convenient to select putative causal SNPs by ranking the SNPs based on their PIP. It provides the statistical quantification of the likelihood of being a causal SNP. However, the computational cost of Bayesian methods can increase exponentially with the width of the genomic region and, therefore, is impractical for large sets of SNPs (e.g., > 1000) [4, 25, 26]. Bayesian fine-mapping approaches may spread the estimated PIP across highly correlated SNPs, and none of them would be large enough to distinguish causal and non-causal SNPs [4, 25]. Thus, understanding the LD pattern of SNPs is essential for all fine-mapping approaches.

1.2 LD block structure

Understanding the patterns of LD across the human genome has been central for genetics research because it provides insights for disease gene mapping and genetics across populations [27]. The block-wise structure has been widely used to model the patterns of LD [28, 29, 30]. The block-wise LD structure is built based on the observations that: i) the LD score between a pair of SNPs generally decreases with their physical distance, and ii) high LD scores extend through a genomic region until

abruptly discontinued by recombination hotspots or other population genetic factors [31, 32]. Although the haplotype block methods generally group proximal SNPs with high correlation into physically contiguous regions, these approaches may be limited to optimally reveal the network structure of LD [27, 28, 29, 33].

In the current research, we review the foundation for classic LD block methods. We first demonstrate the relationship between the LD score and physical distance between a pair of SNPs in Figure 1. We note that the physical distance and LD score are almost uncorrelated within a close neighborhood (e.g., < 100 Kb) owing to the influences, although the LD score generally decays as the distance increases more than 300 Kb. This finding suggests there may exist finer block structure within a genomic region with < 100 Kb, which inspires our investigation of the dense block structure in this neighborhood.

The concept of the "dense" block originated from graph pattern recognition in the field of computer science. In the current research, we consider SNPs as nodes and LD scores as (undirected) edges of a graph. A dense block refers to a subgraph consisting of a number of edges that approximates the maximum number of possible edges it can have [34]. We illustrate the concept of dense LD blocks by Figure 2. Figure 2A shows the LD matrix from the reference genome for a genomic region of 500 kb. Figure 2B-left shows the block structure detected by conventional haplotype block methods, while Figure 2B-right shows the dense block structure by our proposed approach (the order of SNPs is reorganized to show the block structure). Clearly, the dense block structure can better allocate SNPs with high LD scores into communities (Figure 2D). We also notice that the dense LD block structure can better guide disease gene mapping. Figure 2C shows that the $-\log_{10}(p)$-values of SNPs in dense blocks are more coherent than those in conventional haplotype blocks. Therefore, we are motivated to develop dense-LD-block-based fine-mapping strategies.

1.3 Dense LD block structure and fine mapping

An accurate fine-mapping model requires exact knowledge of the LD pattern (i.e., structure) among the SNPs within a GWAS identified region [4, 11, 12, 35, 36]. Previous research on multivariate statistics also has shown that the accuracy of variable selection relies on accurate knowledge of the network structure of the dependence between predictors [37, 38, 39, 40, 41, 42, 43]. Generally, the LD pattern refers to two aspects: i) the local aspect, based on the pairwise LD scores between SNPs that are directly available, and ii) the global aspect, based on the (latent) network topological structure (e.g., the community structure) of the LD matrix. Existing fine-mapping approaches conservatively focus on the local aspect of the LD pattern (i.e., haplotype structure) and lack of complete incorporation of all highly correlated SNPs (i.e., latent dense structure), especially for a large number of variants [4]. To fill this gap, we propose a new fine-mapping approach, named Dense-LD, first to learn the underlying network of the LD matrix using our data-driven procedure (i.e., ℓ_0 graph norm shrinkage algorithm). Next, Dense-LD provides a regression shrinkage strategy by fully leveraging the dense LD block structure and letting variants borrow strengths from each other based on organized LD patterns. It reduces the number of variants and simultaneously selects a group of variants having high correlations. Notably, the SNPs selected by Dense-LD can come from different genes residing in a large physical distance (see section 3 for details). Moreover, Dense-LD selects a larger number of variants than traditional regression shrinkage fine-mapping methods (e.g., LASSO and ENET) and reduces the possibility of falsely removing true

causal variants. The foundation of this statistical framework is the joint consideration of apparently independent sources of information, LD patterns, and trait associations (p-value).

We evaluated our fine-mapping approach using simulation studies and by replicating findings on the genetics of nicotine addiction using a large and inclusive UK biobank (UKBB) sample [44]. The simulation results showed that fine-mapping with dense LD blocks effectively enhanced the sensitivity and reduced the false-positive error rate in the causal variant selection procedure. Our fine-mapping analysis in the UKBB application prioritized 81 variants associated with nicotine dependence that have both high statistical significance ($p < 10^{-50}$) and large effect sizes (> 0.75). These potential causal variants reside across the genes of IREB2, CHRNA3, CHRNA5, CHRNB4, HYKK, and PSMA4 in a highly correlated LD block (average $r^2 = 0.89$), resulting in a systematic and multi-gene causal variant selection process that aids our understanding of the underlying genetic mechanisms for complex traits.

2 Methods

We propose a new fine-mapping approach, Dense-LD, consisting of two steps: i) detect dense LD blocks via ℓ_0 graph norm shrinkage, and ii) prioritize potential causal variants based on the latent dense LD block structure. Figure 3 provides an overview of the procedure.

2.1 Detect dense LD blocks via ℓ_0 graph norm shrinkage

Our goal is to extract dense LD block structure from the LD matrix $\boldsymbol{R}$ based on a genomic region of interest (e.g., identified from the Manhattan plot of GWAS study). Let each entry $0 \le r_{ij} \le 1$ in $\boldsymbol{R}$ represent the LD score between SNPs $i$and j ($1 \le i < j \le n$). Denote $G = \{V, E\}$as a graph notation for the LD pattern, where the vertex/node set V represents n SNPs (i.e., $|V| = n$) in the genomic region, and the edge set E indicates the levels of LD between the n alleles (i.e., $|E| = n \times (n-1)/2$). $\mathcal{T}(G)$ denotes the latent graph topology of the LD patterns which elucidates the assignment of nodes into subsets V_c, $c = 1, \ldots, C$ and $V = \cup_{c=0}^{C} V_c$. We assume that $\mathcal{T}(G)$ encompasses C dense LD blocks $\{G_c\}$ induced by node sets $\{V_c\}_{c=1,\ldots,C}$ and **"singleton"** nodes in V_0. Specifically, we consider the following graph structure: i) the dense LD blocks $\{G_c\}$ with high within-block edge weights indicate highly correlated within-block SNPs; ii) edges connecting two dense LD blocks G_c and $G_{c'}$, $1 \le c < c' \le C$, show medium edge weights (i.e., LD scores); iii) the **"singleton"** nodes that do not belong to the main structure of the network are contained in V_0. Their internal and external connections are low-weighted and may be irregular. We aim at extracting the dense LD blocks which can further guide the causal variant selection.

2.2 A graph mixture model for *R*.

Given the three classes of subgraphs in G, we let an entry r_{ij} follow a marginal mixture distribution that $r_{ij} \sim \sum \pi_k f_k(\theta_{\mathbf{k}})$, where π_k is the proportion for a mixture component k($k = 0, 1, \cdots, K$), where $K \le C(C-1)/(2+1)$, and $f_k(\theta_{\mathbf{k}})$ is the corresponding distribution (e.g., beta distribution) with parameters $\theta_{\mathbf{k}}$. We assume that edges within dense blocks belong to a mixture component, interactions

between blocks constitute $K-1$ mixture components, and edges of the rest of the graph are from the null component. For any $1 \le i < j \le n$, the mixture distribution of r_{ij} is defined as follows:

$$r_{ij} \sim f_k(\boldsymbol{\theta}_{\mathbf{k}})$$

$$\delta_{ij} = \begin{cases} 1, \text{ if } \{i,j\} \subseteq V_c; \\ k, \text{ if } i \in V_{c_1} \text{ and } j \in V_{c_2} \text{ for some } 1 \le c_1 < c_2 \le C; \\ 0, \text{ otherwise}; \end{cases} \quad (1)$$

where δ_{ij} is an indicator variable assigning an edge to a mixture component which is determined by the latent graph structure of G (i.e., the graph categories). We also have $E(r_{ij}|\delta_{ij}=1) > E(r_{ij}|\delta_{ij}=k, k>1) > E(r_{ij}|\delta_{ij}=0)$. The LD graph-structure-based mixture model is distinct from the conventional mixture model because it cannot freely assign an edge r_{ij} to any mixture component due to the graph constraint. For example, given a dense LD block G_c, we have $\left\{e_{ij} \in G_c \text{ and } e_{ij'} \in G_c\right\} \Rightarrow e_{jj'} \in G_c$. Therefore, the estimating methods, including expectation-maximization (EM) and nonparametric Bayes methods, may not be directly applicable.

We consider the latent graph topological structure $\mathcal{T}(G)$ as the key parameter of the LD graph-structure-based mixture model because estimating $\boldsymbol{\theta}_{\mathbf{k}}$ is straightforward given $\mathcal{T}(G)$. Although K is unknown in our infinite mixture model, it can also be estimated by the likelihood principle with a given $\mathcal{T}(G)$. For the purpose of fine-mapping, our primary goal is to detect dense LD blocks while prohibiting false-positive edges (with medium LD scores) from being included in the blocks.

To link the underlying dense LD blocks with the input data $\boldsymbol{R}_{\mathbf{n}\times\mathbf{n}}$, we introduce a matrix $\mathbf{U} = \{u_{ij}\}_{1 \le i,j \le n}$ obtained by thresholding $\boldsymbol{R}_{\mathbf{n}\times\mathbf{n}}$:

$$u_{ij} = \begin{cases} r_{ij}, \text{ if } \delta_{ij} = 1; \\ 0, \text{ otherwise} \end{cases} \quad (2)$$

Estimating dense LD blocks $\{G_c\}$ through directly optimizing the likelihood of the mixture model, however, is rather computationally costly and even intractable [45]. Community detection methods are often used to estimate the graph structure, and then the mixture model parameters are straightforward with estimated community structure [46]. Since the commonly used community detection methods are not suited for the dense block structure, we resort to a new objective function for dense block extraction:

$$\widehat{U} = \underset{C,\ \{V_c\}}{argmax} \log\|\mathbf{U}\|_1 - \tau_0 \log\|\mathbf{U}\|_0, \quad (3)$$

where $\|\mathbf{U}\|_0 = \sum_{i,j} I(|u_{ij}| > 0)$ is the element-wise ℓ_0 matrix norm (i.e., graph (size) norm), $\|\mathbf{U}\|_1 = \sum_{i,j} |u_{ij}|$ is the element-wise ℓ_1 matrix norm, and $0 < \tau_0 < 1$ is a tuning parameter. We maximize $\|\mathbf{U}\|_1$ to assign a maximal number of high LD score edges into dense blocks with high

sensitivity. In the meanwhile, we penalize the ℓ_0 graph norm $\|\mathbf{U}\|_0$ to prohibit including false-positive and medium LD score edges into blocks. For example, assigning one false-positive SNP into a dense block comes with a high cost of greatly increasing the $\|\mathbf{U}\|_0$ term, which is against our objective function. Note that $\delta_{ij}=1$, only if $e_{ij}\in G_c$ and G_c is a dense LD block. Therefore, we regulate the size of each dense block to ensure the low false-positive rate. The tuning parameter τ_0 controls the level of parsimony. In general, larger τ_0 leads to denser yet smaller sized LD blocks. In practice, we can optimize τ_0 based on the likelihood function of the mixture model as follows:

$$\ell(\boldsymbol{\theta}_{\mathbf{k}}, \widehat{\mathcal{T}}_{\tau_0}(G)) = \sum_{i<j} \log\{\sum_{k=0}^{K} f_k(w_{ij};\boldsymbol{\theta}_{\mathbf{k}})\, I(h(e_{ij};\widehat{\mathcal{T}}_{\tau_0}(G)) = k)\}. \quad (4)$$

By implementing the ℓ_0 graph norm shrinkage, we can accurately recognize the underlying dense LD blocks despite the false-positive noise outside dense LD blocks. In practice, the direct optimization of Eq 3 is NP (nondeterministic polynomial time) complex, and we develop computationally efficient algorithms to implement the optimization (see Additional file 2 A1 for details). The extracted dense block structure can further guide the variable selection for fine-mapping.

2.3 Select causal variants based on a dense LD block structure

We perform causal variant selection using genotypic data $\boldsymbol{X}_{\mathrm{S}\times \mathrm{n}}$ (n SNPs) and phenotypic data $\boldsymbol{Y}_{\mathrm{S}\times 1}$, a scalar trait, collected from S participants. Our goal is to integrate the estimated dense LD block structure $\widehat{\mathcal{T}}_{\tau_0}(G)$ into the variable selection procedure. Penalized regression has been a popular tool for variable selection as it enjoys a number of theoretical and numerical advantages. However, commonly used variable selection models like LASSO and ENET tend to randomly select a few genetic variants from a set of highly correlated SNPs that are associated with the phenotype [4, 21, 22]. Recently, fused-LASSO, a structured regularization method, has been developed to encourage similarity within a group of variables while retaining the overall sparsity [47, 48, 49, 50]. This method is well suited for our fine-mapping because the estimated dense LD block structure can naturally provide proximity indices between the coefficients of the multivariate predictors. Our newly developed fine-mapping method, Dense-LD, is built on top of the advances in penalized regression shrinkage, accounting for the structural patterns found in the LD matrix.

Based on the estimated dense LD blocks according to Eqs 3 and 4, we propose the graph guided fused-LASSO, integrating the objective function as:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\sum_{s=1}^{S}(y_s - \mathbf{x}_s^{\mathrm{T}}\boldsymbol{\beta})^2 + \lambda \sum_{i,j:\delta_{ij}=1} |\beta_i - \beta_j| + \gamma\cdot\lambda\sum_{i=1}^{n}|\beta_i|, \quad (5)$$

where $\widehat{\delta}_{ij}$ is determined by the estimated dense LD blocks (i.e., $\{\widehat{V}_c\}$) by optimizing Eq 3, $\boldsymbol{Y}_{\mathrm{S}\times 1} = \{y_1, y_2, \ldots, y_S\}$, $s = 1, 2, \ldots, S$ are the phenotype for S subjects in our data, $x_s = \{x_{s,1}, x_{s,2}, \ldots, x_{s,n}\}^T \in (0, 1, 2)$, $i<j = 1, 2, \ldots, n$ are the n genotypes of the s^{th} subject, and $\boldsymbol{\beta} = \{\beta_1, \beta_2, \cdots, \beta_n\}$ represents the phenotype-genotype associations. The graph object was converted from the adjacency matrix that represented the block structure of SNPs based on the dense LD patterns

and passed with term $|\beta_i - \beta_j|$ for regularization. Guided by the graph object, fused-LASSO not only provides shrinkage on coefficients of single SNPs but also their differences between the highly correlated SNPs within the same dense LD block [47, 48, 49, 50]. Here, we only penalize the difference between coefficients of SNPs within a dense block because they tend to have more homogeneous effect sizes. In Eq 5, λ and γ are the regularization parameters that control the weight of penalty terms. The regularization becomes a pure fusion of the coefficient vector beta when $\gamma = 0$; otherwise, a non-zero γ introduces a ratio of sparsity to terms $|\beta_i|$ and $|\beta_i - \beta_j|$, corresponding to the sparsity of coefficients and the sparsity of their differences, respectively. Given the dense LD blocks, the objective function Eq 5 is convex, and thus can be effectively implemented by various versions of the alternating direction method of multipliers (ADMM) algorithms [51]. The shrinkage pathway is given by a fixed γ value along with various λs ranging from zero to the value that all coefficients are shrunken to zero. We implement a grid search step for optimal results by changing the values of λ and γ according to the application needs. Moreover, our method can be extended to binary traits based on generalized fused LASSO algorithms (see Additional file 2 A2) [52, 53, 54, 55]. We can further evaluate the uncertainty of the selected SNPs by bootstrapping [56].

In summary, we propose Dense-LD as a new fine-mapping method by integrating two independent sources of information: LD matrix from the reference genome and sample genotypic and phenotypic data. We select causal variants based on the extracted dense LD block structure. Since the ℓ_0 graph norm shrinkage is applicable for a broad genomic region, Dense-LD can evaluate thousands of SNPs by treating them as a single fine-mapping region. This advantage provides universal consideration for all SNPs in the entire genomic region, explicitly accounting for their correlations. Accordingly, this approach can improve the accuracy of fine-mapping in many applications. In addition, our approach is computationally efficient with a complexity of $O(n^4)$. We demonstrate the performance of our method using a data example and extensive simulations.

3 Application to Nicotine Addiction from the UK Biobank

We applied our method to the UKBB data sample for the identification of genetic variants associated with nicotine addiction [44]. UKBB is a large prospective study that collects abundant health information, including both genotypic and phenotypic data of 500,000 volunteer participants aged between 40-69 years in the UK [44]. We considered cigarettes per day (CPD) as the phenotype of interest because of its relevance to nicotine dependence, focusing on Caucasians to ensure a homogeneous LD pattern [8, 57]. A total of 142,752 participants, including both previous and current smokers with available genotypic data, were used for our analysis (see Additional file 1 Appendix S1 for a detailed description).

GWAS was performed to identify the genomic regions of interest for further investigation using our fine-mapping method by PLINK [58]. The inclusive criteria, involving minor allele frequency (MAF) > 0.01, Hardy-Weinberg equilibrium (HWE) > 0.001, missingness per marker (GENO) > 0.05, and missingness per individual (MIND) > 0.02, were used for quality control and filtering out individuals or SNPs in GWAS. The Manhattan plot shows genomic regions on chromosomes 8, 15, and 19 that consist of SNPs highly associated with CPD (i.e., strong marginal associations) (see Additional file 1 Figure S1). A set of loci in the genomic region of interest on chromosome 15 (15:78,700,000 to 79,230,000) showed extremely significant associations. We focused on this genomic region on chromosome 15, which consisted of 1733 SNPs, for genetic fine-mapping.

We estimated the LD matrix ($\boldsymbol{R}$) of 1733 SNPs using PLINK [58] and applied the optimization of Dense-LD (Eqs 3 and 4) to identify the dense LD block structure of the matrix $\boldsymbol{R}_{1733\times1733}$. Dense-LD detected 68 dense blocks as shown in Figure 4 A-B. This dense block-wise structure is distinct from the haplotype block-wise structure detected by other methods (e.g., BigLD [30] and *ldetect* algorithm [59]). We provide the haplotype blocks identified by BigLD and *ldetect* in Additional file 1 Figure S2 and Figure S3, respectively. The distributions of LD scores were more differentiated between intra-block and inter-block detected by Dense-LD (see Figure 4B and Additional file 1 Table S1). Using different values ranging from 0.5 to 0.9 for tuning parameter (τ_0), the block structure was almost unchanged, indicating that the block detection of Dense-LD was stable. To support the applicability of our approach, we also analyzed the dense LD block structure of the CPD-associated genomic region in chromosome 19 (see Additional file 1 Figure S4). Next, we implemented dense-LD-block-guided regression shrinkage (Eq 5) and selected 81 SNPs from the same dense LD block (see Figures 4A-B). This dense LD block structure beneficially assists in evaluating the joint effect of highly correlated SNPs in the same block. As shown in Figure 4C, the 81 SNPs showed both highly significant *p*-values and large effect sizes as compared to the rest of the SNPs. This pattern demonstrated the strong capability of leveraging strengths from SNPs with strong LD scores and similar association properties.

We further explored the properties and functions of the 81 SNPs using the SNP annotation portal such as the Functional Annotation of Variant–Online Resource (FAVOR) [60, 61, 62]. We found that the 81 SNPs are located in the intronic regions of the genes IREB2, CHRNA3, CHRNA5, CHRNB4, and HYKK (see details in Additional file 4). Also, some of these regions were predicted to have potential enhancer elements. Previous studies discovered that these genetic regions were associated with smoking-related behaviors and diseases [63, 64, 65, 66, 67].

For comparison, we applied other penalization methods such as LASSO and ENET to this dataset. Due to the lack of LD structure information, both methods selected 14 SNPs (see Additional file 1 Figure S5), missing important and highly significant variants located within the same dense LD block. We also conducted fused-LASSO-based variant selection using haplotype blocks instead of dense blocks as in Dense- LD [30]. Since the haplotype blocks were sparser and wider than dense blocks, almost all SNPs in the region were selected, contradicting the goal of fine-mapping (see Additional file 1 Figure S2). We did not apply Bayesian models because the computational time for a genomic region of 1733 SNPs and 80 causal variants is more than four weeks.

4 Simulation study

To evaluate the performance of Dense-LD, we first generated genotypic data $\boldsymbol{X}$ from 500 participants using the SNP simulation tool HAPGEN2 [68] and estimated the LD matrix using PLINK [58]. We focused on a genomic region of chromosome 15, including 1000 SNPs with physical positions ranging from 78,700,000 to 79,230,000 bp.

To reveal the latent Dense-LD structure, we implemented the objective function Eqs 3 on this genomic region. The estimated Dense-LD structure consists of 26 dense LD blocks of highly correlated SNPs (average $r^2 > 0.80$). We let 35 SNPs within a dense LD block be causal variants for a trait. We varied the effect sizes by setting $\beta = 0.6$, 0.8, and 1. The remaining 965 SNPs were set as non-causal with $\beta = 0$. We then simulated the phenotype variable from a normal distribution by $y_i \sim N(\beta^T X_i, 1)$. We repeated this procedure 100 times for each setting.

We implemented the Dense-LD algorithm and competing methods, including penalized regression shrinkage methods (LASSO [21] and ENET [22]) and Bayesian methods (CAVIAR [18, 19], PAINTOR [17, 69, 70], and JAM [20] on the simulated data sets. We evaluated the performance of fine-mapping models by comparing the selected causal variants with the ground truth (SNPs of $\beta \neq 0$). We summarized the results in Table 1. Our model outperformed the comparable methods regarding the sensitivity and specificity. Dense-LD performs better than the regression shrinkage methods because it allows predictors to effectively borrow strengths from each other based on dense LD structure. Due to the limitation of the size of genomic regions for fine-mapping by Bayesian methods, we split the 1000 SNPs into 25 subregions. We then performed fine-mapping on individual subregions and at last integrated the results. In general, the sensitivity of Bayesian methods was relatively small in part because these methods often only allow a limited maximal number of causal variants, and splitting a genomic region may lead to inaccurate estimation of the dependence structure. The specificity of all approaches is high, reflecting well-controlled false positive error rates. In general, the more causal variants a fine-mapping method selects, the higher probability we can obtain to cover the true causal variants. Figure 5 demonstrates the trend of increased sensitivity with the elevating number of selected causal variants. Dense-LD revealed all true causal variants with a small number of selected causal variants more efficiently than competing methods. Besides the above evaluation metrics, we also provided the receiver operating characteristic (ROC) curves with similar results (see Additional file 1 Figure S6). In summary, the simulation results show that the accuracy of fine-mapping by Dense-LD is improved by leveraging the estimated Dense-LD structure.

We further evaluated the uncertainty of Dense-LD selected causal variants by estimating the confidence interval of $\hat{\beta}$ and inclusion probability using a bootstrap procedure. The inclusion probabilities of true causal variants are high for all settings, and the confidence intervals are sound (see Additional file 3). The computational cost of Dense-LD is affordable. The average computational time for Dense-LD on each data set was 0.16 seconds on a high-performance computer cluster with 27 nodes and 72 Intel(R) Xeon(R) Gold 6150 CPUs (2.70GHz).

5 Discussion

We have developed a novel fine-mapping method that identifies causal variants while incorporating the latent dense LD block structure in a genomic region. This method innovatively characterizes the LD structure by recognizing blocks with condensed high correlations, assisting in selecting causal variants. When high LD occurs between more distant SNPs in a genomic region (e.g., highly correlated SNPs in distinct conventional haplotype blocks), our method becomes advantageous as it can better capture the dependence structure between SNPs and improves the accuracy of fine-mapping. Our method is also computationally efficient and thus is capable of performing fine-mapping in a genomic region with thousands of SNPs and a large maximal number of causal variants.

Our fine-mapping approach is the first to explore the latent structure of the input LD matrix using dense blocks to the extent of our knowledge. The ℓ_0 graph norm shrinkage was employed to discover the dense LD block structure to ensure that dense blocks exclusively consist of SNPs with high correlations. Therefore, the extracted dense block structure is different from the conventional haplotype block detection. Specifically, we can allocate distant SNPs with high correlations from different haplotype blocks into a dense block. In consequence, the LD scores be- tween SNPs in dense blocks are much

higher than those in haplotype blocks, which provides an overall accurate estimate of the dependence structure between SNPs. Our simulation and application studies showed that the dense LD block structure could better guide causal variants selection, consistent with the recent statistical literature [36]. The dense-block structure can be further incorporated into summary-statistics-based analysis (e.g., lassosum), and extended to fine-mapping models [71]. In our application analysis, Dense-LD identified causal variants from different haplotype blocks in an extended genomic region of 1733 SNPs on chromosome 15. The genes of these causal variants (e.g., IREB2, CHRNA3, CHRNA5, CHRNB4, and HYKK) are related to smoking-related diseases and functionally correlated [8, 11, 72, 73]. Our method showed that the selected SNPs from these genes are strongly correlated, although many are distant. By leveraging the extracted dependence structure, Dense-LD captured these features in fine-mapping. Hence, the causal variants selected by Dense-LD can deepen our understanding of the genetics of nicotine ad- diction by providing a comprehensive and systematic fine-mapping analysis. We compared our results with competing methods using regression shrinkage. The results from regression shrinkage seem over-conservative by selecting a few causal variants while missing SNPs correlated with those selected. The direct application of Bayesian fine-mapping approaches to a genomic region of 1733 SNPs is computationally formidable, and thus they were not used in the data example. In simulations, our method outperformed both regression shrinkage and Bayesian methods with higher accuracy of causal variant selection.

In summary, we develop a new fine-mapping toolkit with improved accuracy and computational efficiency. Both our simulation studies and the UKBB nicotine addiction study show how dense-block structure can reveal the latent organized LD patterns and guide fine-mapping. In light of enhanced computational efficiency, Dense-LD is applicable to fine-mapping for a variety of complex traits in wide-range genomic regions and numerous causal variant candidates.

Appendix

Abbreviations

CPD: cigarettes per day; ENET: elastic net; GENO: missingness per marker; GWAS: genome-wide association study; HWE: Hardy-Weinberg equilibrium; Kb: kilo-basepairs; LASSO: least absolute shrinkage and selection operator; LD: linkage disequilibrium; MAF: minor allele frequency; MIND: missingness per individual; PIP: posterior inclusion probability; ROC: receiver operating characteristic; SNP: single nucleotide polymorphism; UKBB: UK Biobank.

Acknowledgments

Not Applicable

Authors' contributions

CM, ZY, SC, and TM developed the method and wrote the manuscript. CM and ZY performed the analysis and results visualization. SC and TM supervised the project and provided conceptualization. KH, YZ, QW, SL, QL, BM, LEH, and PK contributed to manuscript editing, provided critical feedback, and help to shape the research, analysis, and manuscript. All authors have read and approved the manuscript.

Funding

This work was supported by the National Institute on Drug Abuse of the National Institutes of Health under Award Number 1DP1DA04896801. Additional support for computer cluster was provided by NIH R01 grants EB008432 and EB008281.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the UK Biobank (UKB), https://www.ukbiobank.ac.uk/. We provide the GWAS summary statistics, linkage disequilibrium scores, simulated sample data, code and software used in this study in the GitHub repository, https://github.com/cmomo/DenseLD.

Declarations

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]Maryland Psychiatric Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, Maryland, United States of America. [2]Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, Maryland, United States of America. [3]Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, Maryland, United States of America. [4]Department of Statistics, College of Arts and Sciences, The Ohio State University, Columbus, Ohio, United States of America. [5]Department of Mathematics, University of Maryland, College Park, Maryland, United States of America. [6]School of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong, China. [7]Department of Biostatistics, University of Florida, Gainesville, Florida, United States of America.

Supporting Information
Additional file 1 The information of cigarettes per day (CPD) from UK Biobank, and additional figures and tables.
Additional file 2 Mathematical equations for optimization of the objective function in l_0 graph norm shrinkage and fused-LASSO algorithm for binary data.
Additional file 3 Bootstrapping results of Dense-LD in simulation.
Additional file 4 Table of annotations for the 81 SNPs identified by Dense-LD.

References
1. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., *et al.*: Complement factor h polymorphism in age-related macular degeneration. Science **308**(5720), 385–389 (2005)
2. Mills, M.C., Rahal, C.: A scientometric review of genome-wide association studies. Communications biology **2**(1), 1–11 (2019)
3. Tam, V., Patel, N., Turcotte, M., Boss'e, Y., Par'e, G., Meyre, D.: Benefits and limitations of genome-wide association studies. Nature Reviews Genetics **20**(8), 467–484 (2019)
4. Schaid, D.J., Chen, W., Larson, N.B.: From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics **19**(8), 491–504 (2018)
5. Torgerson, D.G., Ampleford, E.J., Chiu, G.Y., Gauderman, W.J., Gignoux, C.R., Graves, P.E., Himes, B.E., Levin, A.M., Mathias, R.A., Hancock, D.B., *et al.*: Meta-analysis of genome-wide association studies of asthma in ethnically diverse north american populations. Nature genetics **43**(9), 887 (2011)
6. Schizophrenia Working Group of the Psychiatric Genomics Consortium and others: Biological insights from 108 schizophrenia-associated genetic loci. Nature **511**(7510), 421–427 (2014)
7. Pickrell, J.K., Berisa, T., Liu, J.Z., S'egurel, L., Tung, J.Y., Hinds, D.A.: Detection and interpretation of shared genetic influences on 42 human traits. Nature genetics **48**(7), 709 (2016)
8. Berrettini, W., Doyle, G.: The chrna5–a3–b4 gene cluster in nicotine addiction. Molecular psychiatry **17**(9), 856–866 (2012)
9. Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J.: Five years of gwas discovery. The American Journal of Human Genetics **90**(1), 7–24 (2012)
10. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., Yang, J.: 10 years of gwas discovery: biology, function, and translation. The American Journal of Human Genetics **101**(1), 5–22 (2017)
11. Icick, R., Forget, B., Clo¨ez-Tayarani, I., Pons, S., Maskos, U., Besson, M.: Genetic susceptibility to nicotine addiction: Advances and shortcomings in our understanding of the chrna5/a3/b4 gene cluster contribution. Neuropharmacology **177**, 108234 (2020)
12. Broekema, R., Bakker, O., Jonkers, I.: A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. Open biology **10**(1), 190221 (2020)
13. Arbet, J., McGue, M., Chatterjee, S., Basu, S.: Resampling-based tests for lasso in genome-wide association studies. BMC genetics **18**(1), 1–15 (2017)
14. Ayers, K.L., Cordell, H.J.: Snp selection in genome-wide and candidate gene studies via penalized logistic regression. Genetic epidemiology **34**(8), 879–891 (2010)
15. He, Q., Cai, T., Liu, Y., Zhao, N., Harmon, Q.E., Almli, L.M., Binder, E.B., Engel, S.M., Ressler, K.J., Conneely, K.N., *et al.*: Prioritizing individual genetic variants after kernel machine testing using variable selection. Genetic epidemiology **40**(8), 722–731 (2016)
16. Kim, S., Xing, E.P.: Statistical estimation of correlated genome associations to a quantitative trait network. PLoS Genet **5**(8), 1000587 (2009)
17. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., Pasaniuc, B.: Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet **10**(10), 1004722 (2014)
18. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., Eskin, E.: Identifying causal variants at loci with multiple signals of association. Genetics **198**(2), 497–508 (2014)
19. Hormozdiari, F., Van De Bunt, M., Segre, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., Eskin, E.: Colocalization of gwas and eqtl signals detects target genes. The American Journal of Human Genetics **99**(6), 1245–1260 (2016)

20. Newcombe, P.J., Conti, D.V., Richardson, S.: Jam: a scalable bayesian framework for joint analysis of marginal snp effects. Genetic epidemiology **40**(3), 188–201 (2016)
21. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
22. Cho, S., Kim, H., Oh, S., Kim, K., Park, T.: Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. In: BMC Proceedings, vol. 3, pp. 1–6 (2009). BioMed Central
23. Zhao, P., Yu, B.: On model selection consistency of lasso. The Journal of Machine Learning Research **7**, 2541–2563 (2006)
24. Xue, F., Qu, A.: Variable selection for highly correlated predictors. arXiv preprint arXiv:1709.04840 (2017)
25. Guan, Y., Stephens, M.: Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics, 1780–1815 (2011)
26. Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M., Auton, A., Myers, S., Morris, A., *et al.*: Bayesian refinement of association signals for 14 loci in 3 common diseases. Nature genetics **44**(12), 1294 (2012)
27. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., *et al.*: The structure of haplotype blocks in the human genome. Science **296**(5576), 2225–2229 (2002)
28. Wang, N., Akey, J.M., Zhang, K., Chakraborty, R., Jin, L.: Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. The American Journal of Human Genetics **71**(5), 1227–1234 (2002)
29. Barrett, J.C., Fry, B., Maller, J., Daly, M.J.: Haploview: analysis and visualization of ld and haplotype maps. Bioinformatics **21**(2), 263–265 (2005)
30. Kim, S.A., Cho, C.-S., Kim, S.-R., Bull, S.B., Yoo, Y.J.: A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated snps. Bioinformatics **34**(3), 388–397 (2018)
31. Jeffreys, A.J., Kauppi, L., Neumann, R.: Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. Nature genetics **29**(2), 217–222 (2001)
32. Stephens, M., Scheet, P.: Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. The American Journal of Human Genetics **76**(3), 449–462 (2005)
33. Wei, P., Pan, W.: Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. Bioinformatics **24**(3), 404–411 (2008)
34. Lov'asz, L.: Large Networks and Graph Limits vol. 60. American Mathematical Soc., Budapest, Hungary (2012)
35. Dadaev, T., Saunders, E.J., Newcombe, P.J., Anokian, E., Leongamornlert, D.A., Brook, M.N., Cieza-Borrella, C., Mijuskovic, M., Wakerell, S., Al Olama, A.A., *et al.*: Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. Nature communications **9**(1), 1–19 (2018)
36. He, K., Kang, J., Hong, H.G., Zhu, J., Li, Y., Lin, H., Xu, H., Li, Y.: Covariance-insured screening. Computational statistics & data analysis **132**, 100–114 (2019)
37. Nicora, G., Vitali, F., Dagliati, A., Geifman, N., Bellazzi, R.: Integrated multi-omics analyses in oncology: A review of machine learning methods and tools. Frontiers in Oncology **10**, 1030 (2020)
38. Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., Ma, S.: A selective review of multi-level omics data integration using variable selection. High-throughput **8**(1), 4 (2019)
39. Li, C., Li, H.: Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics **24**(9), 1175–1182 (2008)
40. Jin, J., Zhang, C.-H., Zhang, Q.: Optimality of graphlet screening in high dimensional variable selection. The Journal of Machine Learning Research **15**(1), 2723–2772 (2014)
41. Wang, X., Leng, C.: High-dimensional ordinary least-squares projection for screening variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **78**(3), 589–611 (2015). doi:10.1111/rssb.12127
42. Wu, C., Pan, W.: A powerful fine-mapping method for transcriptome-wide association studies. Human genetics **139**(2), 199–213 (2020)
43. Pan, W., Shen, X.: Penalized model-based clustering with application to variable selection. Journal of Machine Learning Research **8**(5) (2007)
44. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., *et al.*: Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. Plos med **12**(3), 1001779 (2015)
45. Bickel, P.J., Chen, A.: A nonparametric view of network models and newman–girvan and other modularities. Proceedings of the National Academy of Sciences **106**(50), 21068–21073 (2009)
46. Zhao, Y., Levina, E., Zhu, J., *et al.*: Consistency of community detection in networks under degree-corrected stochastic block models. The Annals of Statistics **40**(4), 2266–2292 (2012)
47. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(1), 91–108 (2005)
48. Tibshirani, R.J., Taylor, J., *et al.*: The solution path of the generalized lasso. The Annals of Statistics **39**(3), 1335–1371 (2011)
49. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of statistical software **33**(1), 1 (2010)
50. Friedman, J., Hastie, T., H¨ofling, H., Tibshirani, R., *et al.*: Pathwise coordinate optimization. The annals of applied statistics **1**(2), 302–332 (2007)
51. Zhu, Y.: An augmented admm algorithm with application to the generalized lasso problem. Journal of Computational and Graphical Statistics **26**(1), 195–204 (2017)

52. Lee, S.H., Yu, D., Bachman, A.H., Lim, J., Ardekani, B.A.: Application of fused lasso logistic regression to the study of corpus callosum thickness in early alzheimer's disease. Journal of neuroscience methods **221**, 78–84 (2014)
53. Lee, S.-I., Lee, H., Abbeel, P., Ng, A.Y.: Efficient l˜ 1 regularized logistic regression. In: Aaai, vol. 6, pp. 401–408 (2006)
54. Yu, D., Lee, S.J., Lee, W.J., Kim, S.C., Lim, J., Kwon, S.W.: Classification of spectral data using fused lasso logistic regression. Chemometrics and Intelligent Laboratory Systems **142**, 70–77 (2015)
55. Lin, T., Ma, S., Zhang, S.: An extragradient-based alternating direction method for convex minimization. Foundations of Computational Mathematics **17**(1), 35–59 (2017)
56. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: the Lasso and Generalizations. Chapman and Hall/CRC, Boca Raton, FL (2015)
57. Erzurumluoglu, A.M., Liu, M., Jackson, V.E., Barnes, D.R., Datta, G., Melbourne, C.A., Young, R., Batini, C., Surendran, P., Jiang, T., et al.: Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. Molecular psychiatry, 1–18 (2019)
58. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., *et al.*: Plink: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics **81**(3), 559–575 (2007)
59. Berisa, T., Pickrell, J.K.: Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics **32**(2), 283 (2016)
60. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., Aslibekyan, S., *et al.*: Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nature genetics **52**(9), 969–983 (2020)
61. National Center for Biotechnology Information (NCBI) [Internet]: Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/. Accessed: 2020-8-27 (1988)
62. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., *et al.*: The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research **47**(D1), 1005–1012 (2019)
63. Amos, C.I., Wu, X., Broderick, P., Gorlov, I.P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., *et al.*: Genome-wide association scan of tag snps identifies a susceptibility locus for lung cancer at 15q25. 1. Nature genetics **40**(5), 616–622 (2008)
64. DeMeo, D.L., Mariani, T., Bhattacharya, S., Srisuma, S., Lange, C., Litonjua, A., Bueno, R., Pillai, S.G., Lomas, D.A., Sparrow, D., *et al.*: Integration of genomic and genetic approaches implicates ireb2 as a copd susceptibility gene. The American Journal of Human Genetics **85**(4), 493–502 (2009)
65. Minic˜a, C.C., Mbarek, H., Pool, R., Dolan, C.V., Boomsma, D.I., Vink, J.M.: Pathways to smoking behaviours: biological insights from the tobacco and genetics consortium meta-analysis. Molecular psychiatry **22**(1), 82–88 (2017)
66. Conlon, M.S., Bewick, M.A.: Single nucleotide polymorphisms in chrna5 rs16969968, chrna3 rs578776, and loc123688 rs8034191 are associated with heaviness of smoking in women in northeastern ontario, canada. Nicotine & Tobacco Research **13**(11), 1076–1083 (2011)
67. Furberg, H., Kim, Y., Dackor, J., Boerwinkle, E., Franceschini, N., Ardissino, D., Bernardinelli, L., Mannucci, P.M., Mauri, F., Merlini, P.A., *et al.*: Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nature genetics **42**(5), 441 (2010)
68. Su, Z., Marchini, J., Donnelly, P.: Hapgen2: simulation of multiple disease snps. Bioinformatics **27**(16), 2304–2305 (2011)
69. Kichaev, G., Pasaniuc, B.: Leveraging functional-annotation data in trans-ethnic fine-mapping studies. The American Journal of Human Genetics **97**(2), 260–271 (2015)
70. Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindstroem, S., Kraft, P., Pasaniuc, B.: Improved methods for multi-trait fine mapping of pleiotropic risk loci. Bioinformatics **33**(2), 248–255 (2017)
71. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., Sham, P.C.: Polygenic scores via penalized regression on summary statistics. Genetic epidemiology **41**(6), 469–480 (2017)
72. Hancock, D.B., Wang, J.-C., Gaddis, N.C., Levy, J.L., Saccone, N.L., Stitzel, J.A., Goate, A., Bierut, L.J., Johnson, E.O.: A multiancestry study identifies novel genetic associations with chrna5 methylation in human brain and risk of nicotine dependence. Human molecular genetics **24**(20), 5940–5954 (2015)
73. Saccone, N.L., Wang, J.C., Breslau, N., Johnson, E.O., Hatsukami, D., Saccone, S.F., Grucza, R.A., Sun, L., Duan, W., Budde, J., *et al.*: The chrna5-chrna3-chrnb4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in african-americans and in european-americans. Cancer research **69**(17), 6848–6856 (2009)
74. Goldberg, A.V.: Finding a Maximum Density Subgraph. University of California, Berkeley, CA (1984)
75. Charikar, M.: Greedy approximation algorithms for finding dense components in a graph. In: International Workshop on Approximation Algorithms for Combinatorial Optimization, pp. 84–95 (2000). Springer
76. Tsourakakis, C.E.: Mathematical and algorithmic analysis of network and biological data. arXiv preprint arXiv:1407.0375 (2014)
77. Stella, X.Y., Shi, J.: Multiclass spectral clustering. In: Null, p. 313 (2003). IEEE
78. Bolla, M.: Spectral Clustering and Biclustering: Learning Large Graphs and Contingency Tables. John Wiley & Sons, Hoboken, NJ (2013)
79. Chen, S., Kang, J., Xing, Y., Zhao, Y., Milton, D.K.: Estimating large covariance matrix with network topology for high-dimensional biomedical data. Computational Statistics & Data Analysis **127**, 82–95 (2018)

Tables

Table 1 Sensitivity and specificity in simulation.

Beta		Dense-LD	LASSO	ENET	CAVIAR	PAINTOR	JAM
0.6	Sensitivity	1.0000(0.0000)	0.7607(0.0058)	0.8128(0.0055)	0.2551(0.0055)	0.1149(0.0004)	0.2571(0.0031)
	Specificity	0.9563(0.0009)	0.9167(0.0029)	0.9152(0.0022)	0.8199(00016)	0.9077(0.0007)	0.9694(0.0006)
0.8	Sensitivity	1.0000(0.0000)	0.6826(0.0065)	0.7340(0.0053)	0.2437(0.0053)	0.1369(0.0049)	0.2191(0.0035)
	Specificity	0.9561(0.0003)	0.8693(0.0040)	0.8576(0.0029)	0.8413(0.0011)	0.8978(0.0008)	0.9694(0.0005)
1	Sensitivity	1.0000(0.0000)	0.7991(0.0051)	0.8388(0.0047)	0.2251(0.0050)	0.1337(0.0034)	0.2183(0.0021)
	Specificity	0.9485(0.0003)	0.8891(0.0028)	0.8689(0.0023)	0.8511(0.0009)	0.9027(0.0007)	0.9709(0.0004)

Mean (standard error) of 100 simulations.

Figures

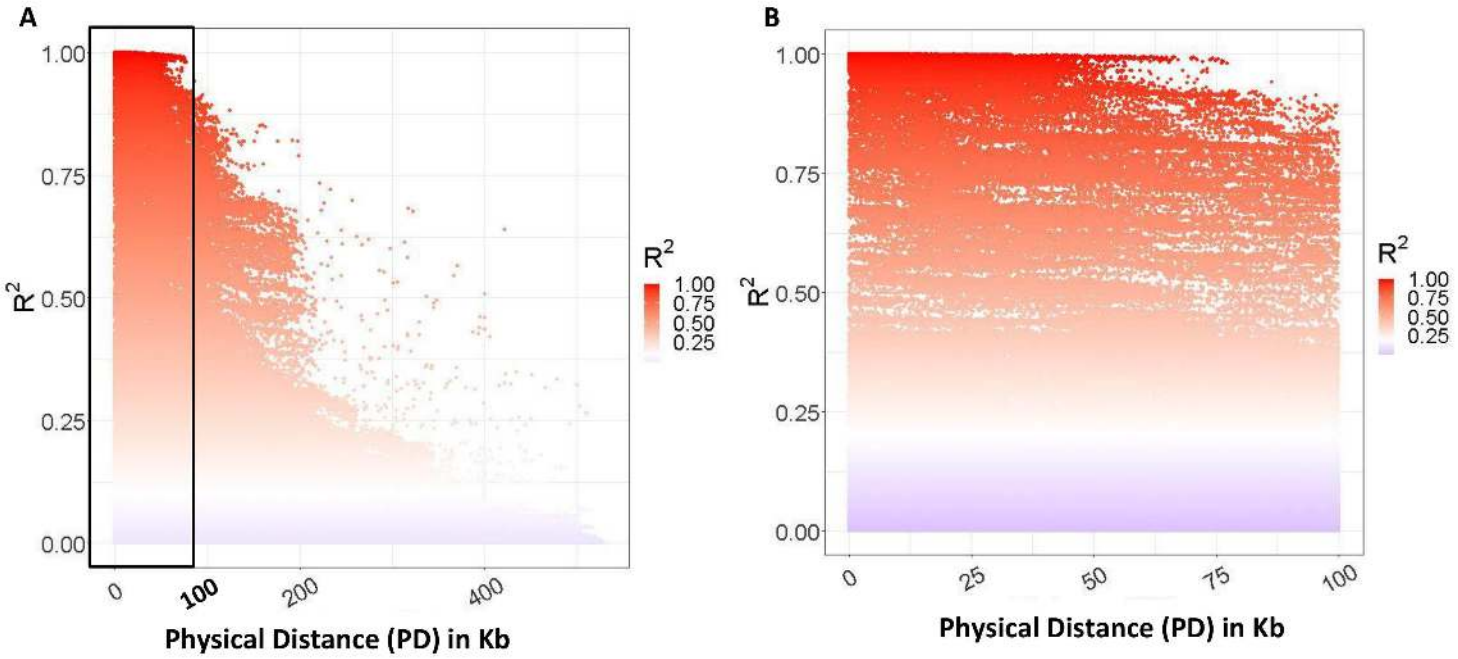

Figure 1

LD decay (in Kb) plot. A) displays the pairwise correlations (r2) of all SNPs pairs in the selected region from CPD data. The zoom plot in B) shows the relationship between r2 and the physical distance of pairs of SNPs within PD = 100 Kb and shows a non-monotonic and nonlinear pattern.

Figure 2

Manhattan plots and heatmaps before and after LD pattern detection via Dense-LD. (A) shows the raw heatmap of the genomic region. (B) shows the heatmaps of the haplotype block (left) and the dense block (right) learned from (A). (C) shows the Manhattan plots with SNPs ordered according to haplotype block (left) and dense block (right). The vertical lines in the Manhattan plots indicate the block boundaries. The plots on the left represent the SNPs in their natural physical position. On the contrary, the plots on the right reorder the SNPs based on rankings based on LD pattern detection via graph norm shrinkage. SNPs are colored along with their block ID. The blue rectangle highlights an example of SNPs with distinct levels of trait associations, having moderate correlations (r2) ranging from 0.5 to 0.6, which are assigned to two different blocks. (D) shows the histograms of r2 inside (red) and outside (blue) blocks detected by different LD block detection methods. The results of haplotype blocks detected by ldetect algorithm is provided in Additional file 1 Figure S3)

Figure 3

Overview of Dense-LD procedure. A genomic region selected based on genome-wide analysis results is passed to Dense-LD fine-mapping following two steps: i) detection of the LD structure of SNPs, and ii) selection of SNPs based on the SNPs data and the LD structure.

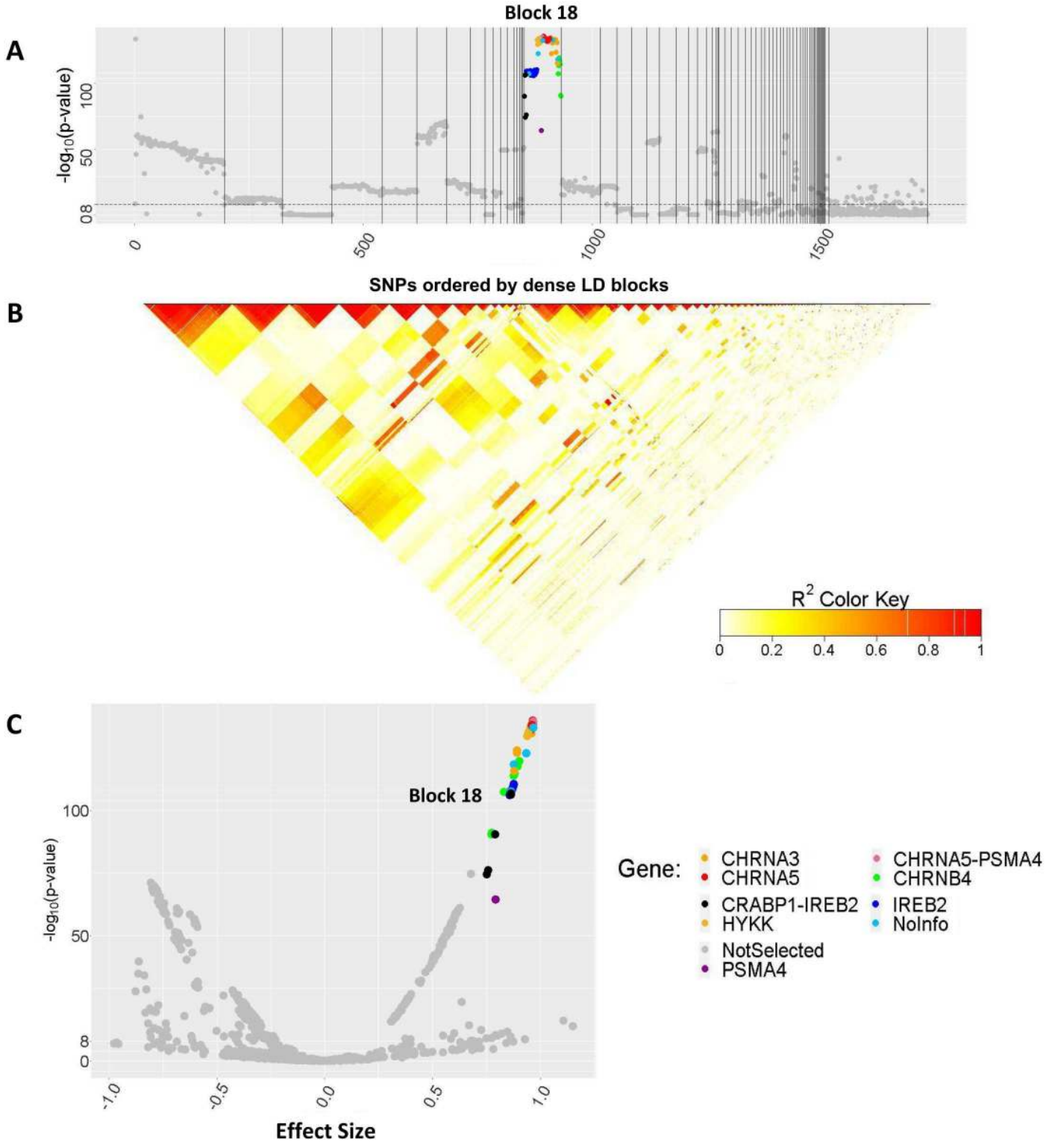

Figure 4

Results of Dense-LD for CPD. (A) and (B) are the Manhattan plot and heatmap for SNPs ordered according to Dense-LD, respectively. The Manhattan plot shows the non-selected SNPs in grey and

highlights the selected SNPs in other colors. The SNPs are colored according to the genes they are located in. The horizontal dashed line (at − log10 (p-value) = 8) corresponds to the commonly used genome-wide significance level (p-value = 5 × 10−8). The block ID of selected SNPs is shown on top of the plot. The volcano plot (C) shows that the SNPs in the 18th block gathered relatively close to each other, indicating similar trait associations within the same block in terms of p-values and effect sizes.

Figure 5

Comparison of different methods in the simulation. The x-axis is the number of selected SNPs, and the y-axis is the proportion of causal variants included across the simulations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.docx
- AdditionalFile2.pdf
- AdditionalFile3.xlsx
- AdditionalFile4.xlsx