Original Paper

Human Heredity

Hum Hered 2017/2018;83:117–129
DOI: 10.1159/000489761

Published online: January 22, 2019

Inferring Gene-Disease Association by an Integrative Analysis of eQTL Genome-Wide Association Study and Protein-Protein Interaction Data

Jun Wang[a] Jiashun Zheng[b] Zengmiao Wang[c] Hao Li[a, b] Minghua Deng[a, d, e]

[a]Center for Quantitative Biology, Peking University, Beijing, China; [b]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA; [c]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA; [d]School of Mathematical Sciences, Peking University, Beijing, China; [e]Center for Statistical Sciences, Peking University, Beijing, China

Keywords
Data integration · Disease-associated gene · Hidden Markov random field

Abstract

Objectives: Genome-wide association studies (GWASs) have revealed many candidate SNPs, but the mechanisms by which these SNPs influence diseases are largely unknown. In order to decipher the underlying mechanisms, several methods have been developed to predict disease-associated genes based on the integration of GWAS and eQTL data (e.g., Sherlock and COLOC). A number of studies have also incorporated information from gene networks into GWAS analysis to reprioritize candidate genes. ***Methods:*** Motivated by these two different approaches, we have developed a statistical framework to integrate information from GWAS, eQTL, and protein-protein interaction (PPI) data to predict disease-associated genes. Our approach is based on a hidden Markov random field (HMRF) model, and we called the resulting computational algorithm GeP-HMRF (a GWAS-eQTL-PPI-based HMRF). ***Results:*** We compared the performance of GeP-HMRF with Sherlock, COLOC, and NetWAS methods on 9 GWAS datasets, using the disease-related genes in the MalaCards database as the standard, and found that GeP-HMRF significantly improves the prediction accuracy. We also applied GeP-HMRF to an age-related macular degeneration disease (AMD) dataset. Among the top 50 genes predicted by GeP-HMRF, 7 are reported by the MalaCards database to be AMD-related with an enrichment p value of 3.61 $\times 10^{-119}$. Among the top 20 genes predicted by GeP-HMRF, CFHR1, CGHR3, HTRA1, and CFH are AMD-related in the MalaCards database, and another 9 genes are supported by the literature. ***Conclusions:*** We built a unified statistical model to predict disease-related genes by integrating GWAS, eQTL, and PPI data. Our approach outperforms Sherlock, COLOC, and NetWAS in simulation studies and 9 GWAS datasets. Our approach can be generalized to incorporate other molecular trait data beyond eQTL and other interaction data beyond PPI.

Background

Genome-wide association studies (GWASs) have become a powerful method to identify genetic variants associated with a complex disease. However, most of the significant SNPs identified by GWASs are located in the non-coding regions of the genome, making it difficult to

KARGER

E-Mail karger@karger.com
www.karger.com/hhe

Hao Li and Minghua Deng
Center for Quantitative Biology
Peking University, No. 5 Yiheyuan Road, Haidian District
Beijing, 100871 (China)
E-Mail haoli@genome.ucsf.edu and dengmh@pku.edu.cn

interpret the results [1]. In the post-GWAS era, it remains a challenge to identify disease-associated genes based on statistically significant SNPs. In the case where no SNP reaches a genome-wide significance threshold, it is even more challenging to delineate gene-disease associations from the weak GWAS signal. A number of previous studies attempted to address these issues, and the approaches can be classified into the following two general categories.

Studies in the first category try to integrate information from GWASs of intermediate molecular traits. Several algorithms have been developed to make use of the colocalization information of GWAS candidate SNPs and eQTL SNPs to infer an association between diseases and genes [2–9]. The underlying hypothesis is that if an allele appears more frequently in patients than in healthy controls, and at the same time this allele is associated with the expression of a gene, it is likely that this gene is associated with the disease (or influences the disease risk through changed expression). Initially, a common approach was to map a significant SNP to the nearby gene in the genome, and to check whether the SNP influences the expression of the gene (eSNP) using eQTL data. Several GWAS analyses employed this idea and found interesting candidate genes of type 2 diabetes, for example [10–12]. These studies focused on the cis-eSNPs and ignored the trans-eSNPs.

He et al. [13] developed a Bayesian inference model called Sherlock, trying to colocalize both cis- and trans-eQTL signal with the GWAS signal. The underlying hypothesis of Sherlock is that if a gene is the driver of a disease, then an SNP that influences the expression of the gene (i.e., eSNP) is also likely to influence the disease phenotype; thus, the eQTL signal of the gene will overlap with the GWAS signal. Specifically, by comparing the similarity between a gene's eQTL profile (eQTL *p* values across all SNPs) and the GWAS profile (GWAS *p* values across all SNPs), they calculate the likelihood ratio (LR) as the evidence of the association between the gene and the disease.

Similar to Sherlock, other algorithms such as COLOC [4], eCAVIAR [5], and ENLOC [9] have developed subsequently which estimate the posterior probability that the same SNP is causal in both the GWAS and the eQTL study to get disease-associated genes. Sherlock is slightly different from COLOC, eCAVIAR, and ENLOC in the way it utilizes SNPs. Sherlock takes all eQTL SNPs passing a soft threshold across the genome no matter whether they are cis or trans, while the other 3 methods consider only cis-effect SNPs in a continuous region on the chromosome.

One challenge of the colocalization method is that a GWAS-significant region may contain several SNPs which are highly correlated in a linkage disequilibrium (LD) block. The causal SNPs might be surrounded by other significant SNPs in the LD block. Facing the uncertainty brought by LD, Sherlock and COLOC assume at most 1 causal SNP in an LD block. Sherlock selects the most significant eQTL SNP in an LD block and discards all other SNPs in the same LD block (see "Considering the LD Blocks" below for details on how Sherlock deals with the LD issue). COLOC incorporates the at-most-one-causal assumption into the prior of hidden states of SNPs. eCAVIAR has no restriction on the number of causal SNPs in an LD and utilizes the information from the 1000 Genomes Project [14] or the HapMap Project [15] to represent the correlation of SNPs in the LD. ENLOC builds a hierarchical Bayesian model and uses an enrichment parameter to account for the LD. Even though eCAVIAR, COLOC, and ENLOC utilize all SNPs in the LD block, they ignore SNPs from other distant regions in the genome.

Studies in the second category try to prioritize candidate disease-associated genes by integrating information from gene networks into GWASs. This has been motivated by the recognition that the genes associated with a disease tend to be functionally or physically coupled. Several different approaches have been developed: Chen et al. [16] proposed a Markov random field (MRF) model to incorporate pathway topology into association analysis; GWAB [17] performs network propagation on a gene co-functional network; NetWAS [18] trains a support vector machine classifier to prioritize genes with edges in a tissue-specific network as features; and REGENT [19] utilizes a hierarchical model to integrate the embedding of genes, which are related based on multiple networks, into GWAS data. However, the gene-level association scores in all these methods are directly derived from the nearby SNPs' GWAS *p* value, without considering the eQTL information.

These two general approaches motivated us to integrate the eQTL information and gene-gene interaction information in one unified statistical framework to infer disease-associated genes. Based on the approach used by Sherlock [13], we can compute the LR of each gene, integrating the evidence from GWAS and eQTL data. But instead of testing each gene independently, as Sherlock does, we combine the LR from Sherlock and the protein-protein interaction (PPI) information to build a unified statistical model. The basic assumption underlying the use of the PPI information is "guilty by association" [20].

If there are many disease-associated genes in a gene's interaction partners, the gene's potential to be associated with the disease will increase. Chen et al. [16] showed that the suspect genes are more likely to be neighbors in a gene network.

We built a hidden MRF (HMRF) model [21] to integrate GWAS eQTL and PPI data. The resulting algorithm is called GeP-HMRF (a GWAS-eQTL-PPI-based HMRF). GeP-HMRF is composed of the MRF [22, 23] to model the interaction of genes and emission function to integrate the observed GWAS and eQTL profile. To demonstrate the utility of GeP-HMRF, we compared GeP-HMRF with Sherlock, COLOC, and NetWAS methods on 9 GWAS datasets and found that GeP-HMRF significantly outperforms the other 3 methods. We took age-related macular degeneration (AMD) as an example to illustrate the performance of GeP-HMRF in detail. GeP-HMRF discovered 2 more genes than Sherlock in the top 50 predictions. Among the top 20 genes predicted by GeP-HMRF, 4 are AMD-related reported by the MalaCards database [24], and 9 additional genes are supported by the literature but not included in the MalaCards database.

Methods

Modeling PPIs by MRF

A PPI network can be represented by an undirected graph $G = (V, E)$, in which $V = \{1, ..., n\}$ is the set of genes and $E = \{(i, j)$: i and j are directly connected$\}$ is the set of edges [16]. And we denote $c_i = \{j: (i, j) \in E\}$ as the neighborhood of gene i. Let Z_i be the status of gene i which is a binary indicator variable,

$$Z_i = \begin{cases} 1, & \text{if gene } i \text{ is associated with the disease;} \\ 0, & \text{if gene } i \text{ is not associated with the disease.} \end{cases}$$

Then, we denote $\boldsymbol{Z} = \{Z_1, ..., Z_n\}$ as the status of V. Thus, $\boldsymbol{Z}$ is a spatial random vector whose elements may be correlated with each other. Since each gene has 2 statuses, there are 2^n configurations of the nodes' statuses in the PPI network. Our goal is to estimate the value of $\boldsymbol{Z}$ based on the topology of the PPI network and the observed eQTL and GWAS data.

A previous study [16] has shown that a pair of interacting genes tends to have the same status. Thus, in the PPI network, the probability of two directly connected genes having the same status is higher than having a different status. Similar to Deng et al.'s work [22], we utilized the nearest neighbor Gibbs measure to model the probability of a PPI network, which has the following form:

$$P\left(\boldsymbol{Z}|\lambda_0\right) = \frac{1}{M\left(\lambda_0\right)} \exp\left\{\alpha N_1 + \beta_1 N_{11} + \beta_2 N_{01} + \beta_3 N_{00}\right\}. \qquad (1)$$

In equation 1,

$$N_1 = \sum\nolimits_{i=1}^{n} Z_i$$

is the number of status 1 genes, N_{ll}' *is* the number of (l, l') interacting pairs in E,

$$N_{11} = \sum_{(i,j)\in E} Z_i Z_j$$

$$N_{00} = \sum_{(i,j)\in E} \left(1 - Z_i\right)\left(1 - Z_j\right)$$

$$N_{01} = \sum_{(i,j)\in E} \left(1 - Z_i\right) Z_j + Z_i \left(1 - Z_j\right)$$

and $\lambda_0 = (\alpha, \beta_1, \beta_2, \beta_3)$ are the prior parameters. β_i represent the weights of 3 different kinds of edges connecting non-associated or associated genes. And $M(\lambda_0)$ is a normalizing equation which is the summation of all 2^n configurations:

$$M\left(\lambda_0\right) = \sum_{\boldsymbol{Z}} \exp\left\{\alpha N_1 + \beta_1 N_{11} + \beta_2 N_{01} + \beta_3 N_{00}\right\}.$$

When n is large in the practical problem, it is prohibitive to calculate $M(\lambda_0)$ directly. The Gibbs measure in equation 1 has the Markov property:

$$P\ (Z_i|Z_{-i}) = P\ (Z_i|Z_{c_i}).$$

Thus, a MRF model has been constructed for describing the relations between directly interacting genes.

Emission Function of the HMRF

Since $\boldsymbol{Z}$ is not observable, the above MRF model describing the PPI can be treated as the hidden layer in the HMRF model. We also need an emission function to model the observed GWAS and eQTL data given the genes' status $\boldsymbol{Z}$. Our data consists of the p values of SNPs related to the gene expression trait (eQTL profile), denoted as vector $\boldsymbol{X}$, and the p values of the SNPs related to the phenotypic trait (GWAS profile), denoted as vector $\boldsymbol{Y}$. For a fixed gene i, we select the putative eSNPs passing a low significance threshold (say 10^{-5}) in the eQTL data. The eQTL profile of a gene is denoted as vector $\boldsymbol{Xi} = [X_{i_1}, ..., X_{i_m}]$, and the GWAS profile of these corresponding SNPs is denoted as $\boldsymbol{Yi} = [Y_{i_1}, ..., Y_{i_m}]$. The emission function is to compute $P(\boldsymbol{Xi}, \boldsymbol{Yi}|Z_i)$. Assuming all SNPs in vector $\boldsymbol{Xi}$ are conditionally independent (see also "Considering the LD Blocks" below), then

$$P\left(\boldsymbol{X}_i, \boldsymbol{Y}_i \,|\, Z_i\right) = \prod_{j=1}^{m} P\left(X_{i_j}, Y_{i_j} \,|\, Z_i\right).$$

According to Sherlock, the likelihood function at each SNP at a given Z_i is computed by summing over the hidden variables U_{i_j} and V_{i_j}:

$$P\left(X_{i_j}, Y_{i_j}|Z_i\right) = \sum_{U_{i_j}\in\{0,1\}, V_{i_j}\in\{0,1\}} P\left(U_{i_j}\right) P\left(V_{i_j}|Z_i, U_{i_j}\right) P\left(X_{i_j}|U_{i_j}\right) P\left(Y_{i_j}|V_{i_j}\right), \qquad (2)$$

where U_{i_j} is a binary random variable indicating whether SNP j of gene i is an eSNP: $P(U_{i_j} = 1) = \pi_1$, SNP j is an eSNP. And V_{i_j} is a binary random variable indicating whether SNP j is a disease-related SNP. When gene i is not a disease-related gene, SNP j is a disease-related SNP with probability: $P(V_{i_j} = 1|Z_i = 0, U_{i_j}) = \pi_2$. When gene i is a disease-related gene and SNP j is an eSNP of gene i, SNP j should be disease-related with probability 1: $P(V_{i_j} = 1|Z_i = 1, U_{i_j} = 1) = 1$. As for the calculation of $P(X_{i_j}|U_{i_j})$ and $P(Y_{i_j}|V_{i_j})$, please refer to the online supplementary material of Sherlock [13].

```
Algorithm 1: Gibbs sampler

Initialize $z^{(0)}$ according to Sherlock score ($z_j^{(0)} = 1$ if Sherlock score >0, $z_j^{(0)} = 0$ otherwise)
For iteration $i = 1, 2, \ldots, 5{,}000$,
    $z_1^{(i)} \sim P(Z_1^{(i)} = z_1 | Z_2^{(i-1)} = z_2^{(i-1)}, Z_3^{(i-1)} = z_3^{(i-1)}, \ldots, Z_n^{(i-1)} = z_n^{(i-1)}, \boldsymbol{X}, \boldsymbol{Y}, \lambda_0)$
    $z_2^{(i)} \sim P(Z_2^{(i)} = z_2 | Z_1^{(i)} = z_1^{(i)}, Z_3^{(i-1)} = z_3^{(i-1)}, \ldots, Z_n^{(i-1)} = z_n^{(i-1)}, \boldsymbol{X}, \boldsymbol{Y}, \lambda_0)$
    …
    $z_n^{(i)} \sim P(Z_n^{(i)} = z_n | Z_1^{(i)} = z_1^{(i)}, Z_2^{(i)} = z_2^{(i)}, \ldots, Z_{n-1}^{(i)} = z_{n-1}^{(i)}, \boldsymbol{X}, \boldsymbol{Y}, \lambda_0)$
Permute the updating order of genes
End for
Discard the first 1,000 samples as burn-in, and choose 1 sample in every 10 samples
```

Assuming the MRF model (equ. 1) is a network prior for hidden layer $\boldsymbol{Z}$, and the likelihood function (equ. 2) is an emission function, we have built a HMRF model. Given the observed eQTL profile and GWAS profile, the posterior distribution of $\boldsymbol{Z}$ follows:

$$P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y}, \lambda_0) \propto P(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z})\, P(\boldsymbol{Z}|\lambda_0). \tag{3}$$

The posterior distribution in equation 3 can define a MRF, too, refer to online supplementary section 1.1 "Proof of the posterior distribution is also an MRF" for proof (for all online supplementary material, see www.karger.com/doi/10.1159/000489761). Ideally, we would like to find out the maximum posterior probability to infer the configuration of $\boldsymbol{Z}$, but it is impossible to do so when n is large. Thus, we utilize the Gibbs sampling method [25] to get the posterior mean of each gene's status Z_i.

Making an Inference Based on the Gibbs Sampling

The posterior distribution of a gene's status $\boldsymbol{Z}$ can be divided into two parts:

$$P(\mathbf{Z}|\boldsymbol{X}, \boldsymbol{Y}, \lambda_0) = P(Z_i|\mathbf{Z}\boldsymbol{-i}, \boldsymbol{X}, \boldsymbol{Y}, \lambda_0)\; P(\mathbf{Z}\boldsymbol{-i}|\boldsymbol{X}, \boldsymbol{Y}, \lambda_0).$$

And for a specific gene i, the log odds of posterior probability is:

$$\begin{aligned}\log\frac{P\left(Z_i = 1|\boldsymbol{Z}_{-i}, \boldsymbol{X}, \boldsymbol{Y}, \lambda_0\right)}{P\left(Z_i = 0|\boldsymbol{Z}_{-i}, \boldsymbol{X}, \boldsymbol{Y}\ \lambda_0\right)} &= \log\frac{P\left(\boldsymbol{X}_i, \boldsymbol{Y}_i|Z_i = 1\right)P\left(Z_i = 1|\boldsymbol{Z}_{-i}, \lambda_0\right)}{P\left(\boldsymbol{X}_i, \boldsymbol{Y}_i|Z_i = 0\right)P\left(Z_i = 0|\boldsymbol{Z}_{-i}, \lambda_0\right)}\\ &= \log\frac{P\left(\boldsymbol{X}_i, \boldsymbol{Y}_i|Z_i = 1\right)}{P\left(\boldsymbol{X}_i, \boldsymbol{Y}_i|Z_i = 0\right)} + \alpha\\ &\quad + \left(\beta_1 - \beta_2\right)K_{i1} + \left(\beta_2 - \beta_3\right)K_{i0}\,.\end{aligned}$$

The first part

$$\log\frac{P\left(\boldsymbol{X}_i, \boldsymbol{Y}_i \,|\, Z_i = 1\right)}{P\left(\boldsymbol{X}_i, \boldsymbol{Y}_i \,|\, Z_i = 0\right)}$$

is the log-LR, which is the measure employed by Sherlock and can be treated as the evidence from the observed eQTL and GWAS profiles. The second part is the network prior parameter α which generally influences the proportion of positive genes. A smaller α corresponds to a smaller chance for a gene to be positive. The third part is the contribution from positive neighboring genes, while $\beta_1 - \beta_2$ is the weight and K_{i1} is the number of positive genes in the neighbors of gene i. When there is 1 more positive gene among the neighbors of gene i, the log odds of posterior probability will be increased by $\beta_1 - \beta_2$. The fourth part is the contribution from the neighboring genes with $z_i = 0$. Similar to the third part, $\beta_2 - \beta_3$ is the weight, and K_{i0} is the number of genes with $z_i = 0$ among the neighbors of gene i.

We sample the gene's status Z_i according to its posterior probability described in equation 4 with fixed prior parameter $\lambda_0 = (-5, 0.8, -0.001)$ in real data analysis (see online supplementary section 1.3 "The details of Gibbs sampling in the GeP-HMRF method" for detail about the Gibbs sampler). After the convergence of an MCMC chain, we get the posterior mean of each gene $P(Z_i = 1|\mathrm{Z}_{-i}, \boldsymbol{X}, \boldsymbol{Y}, \lambda_0)$ by counting the frequency of $Z_i = 1$. We found that the posterior mean is positively correlated with the gene's degree in the network (see online supplementary section 1.4 "The posterior mean has a positive correlation with the gene's degree in network" for details). In order to decrease the bias brought by the degree and assess the significance of the posterior means, we run different MCMC chains 1,600 times using randomly permutated log-LR for genes and keeping their neighbors unchanged. Thus, we get 1,600 randomized posterior means for each gene and treat them as the background. Finally, the p value for each gene is computed by comparing the original posterior mean (based on the original log-LR) to the background. GeP-HMRF takes the p values as the measure of gene-disease associations.

Considering the LD Blocks

When aligning the eQTL eSNPs and GWAS SNPs, we face the problem that different eSNPs may be located in the same LD blocks. After applying a soft cutoff of 10^{-5} to the eQTL profile, most eSNPs naturally fall into different LD blocks. When there are still several eSNPs in the same LD block, we only chose the eSNP with the most significant eQTL p value in this block. Thus, we can make sure the SNPs used in the Sherlock and GeP-HMRF model are in different LD blocks. For a fixed eSNP in vector Xi, we chose the corresponding SNP in the GWAS data to construct Yi. If there is no corresponding SNP in the GWAS data, we chose the most adjacent GWAS SNP within the same LD block of the eSNP. Most of the alignment comes from the exact match, a small fraction of <10% is from adjacency within the same LD block. SNPs without any alignment will be ignored.

Results

Simulation Studies

We did a simulation using synthetic data to study the performance of GeP-HMRF. The simulation data was constructed in two steps. First, we chose 8 independent SNPs on chromosome 21 and simulated the genotype of

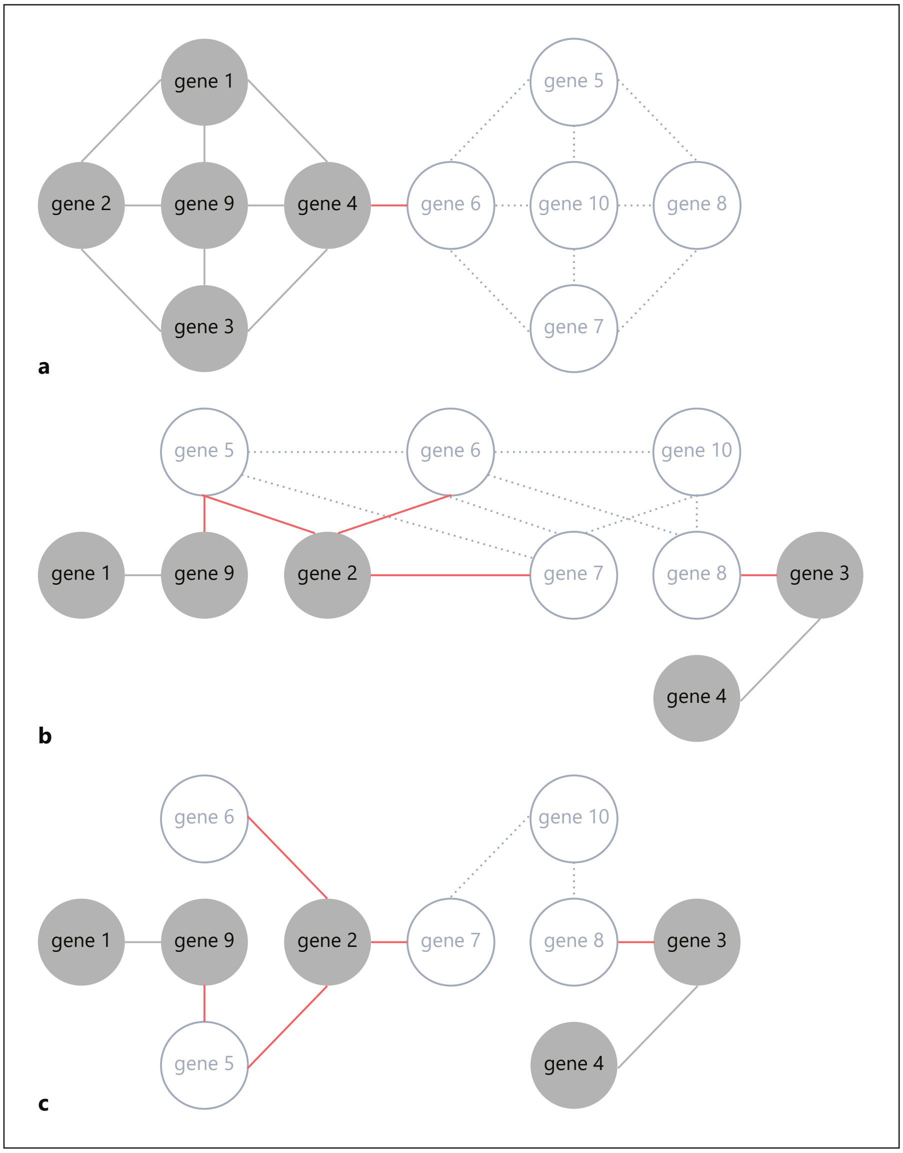

Fig. 1. The 3 artificial networks used in the simulation. The genes in the grey circle are causal to phenotype, while the genes in white are non-causal. The edge in red links a causal and a non-causal gene. Network (a) with a high clustering coefficient of 0.571 has a high proportion of (1, 1) edges (47.1%); network (b) with a high clustering coefficient of 0.600 has 14.2% (1, 1) edges; network (c) with the lowest clustering coefficient of 0 has 22.2% (1, 1) edges.

2,000 samples using HAPGEN2 developed by Su et al. [26]. Then, according to the genotype of the 8 SNPs, we simulated the expression data of 8 genes corresponding to the 2,000 samples. For each gene, there is one SNP that affects its expression:

$$g_i = \mu + \theta_i \,\mathrm{SNP}_i + \varepsilon_i,\ i = 1, 2, \ldots, 8,$$

where $\mu = 5$, $\theta_i = 0.5$, and $\varepsilon_i \sim N(0, 1)$. We constructed an artificial interaction network (a) with 10 genes (Fig. 1). The expressions of the 8 genes are described above, and we added two additional genes whose expressions are set as random variables following normal distributions $N(6, 2^2)$, $i = 9, 10$; i.e., there is no SNP associated with genes 9 and 10. The disease risk of sample j is determined by genes 1–4 and gene 9:

$$\log \frac{P\left(D_j = 1\right)}{P\left(D_j = 0\right)} = \omega_0 + \sum_{i \in \{1,\,2,\,3,\,4,\,9\}} \omega_i g_i + \xi,$$

where $\omega_0 = -26$, $\omega_i = 1$, and $\xi \sim N(0, 0.5^2)$. Cases are obtained from Bernoulli distribution with probability $P(D_j = 1)$; the left samples are treated as controls. Then, we have the genotypes and gene expression profiles of cases and controls. Standard linear regression analysis (lm function in R) was used to get the association between SNPs and genes (eQTL analysis), and standard χ^2 test

Table 1. eQTL *p* values in synthetic data

ID	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8
Gene 1	**0.000009**	0.610091	0.290353	0.947109	0.321585	0.734279	0.316854	0.168246
Gene 2	0.638634	**0.000001**	0.467930	0.969769	0.304657	0.208840	0.782695	0.482802
Gene 3	0.003592	0.194502	**0.000002**	0.206140	0.964279	0.865404	0.973628	0.788787
Gene 4	0.093823	0.822662	0.804729	**0.000006**	0.875819	0.668831	0.764818	0.445460
Gene 5	0.518030	0.317525	0.149372	0.976911	**0.000183**	0.192892	0.881140	0.825146
Gene 6	0.631310	0.817056	0.637363	0.345251	0.734619	**0.000001**	0.042815	0.351485
Gene 7	0.205045	0.858568	0.445119	0.511781	0.511964	0.665488	**0.000001**	0.910775
Gene 8	0.804331	0.018984	0.883694	0.944684	0.656816	0.394546	0.802157	**0.000007**
Gene 9	0.013282	0.407290	0.739397	0.453253	0.965608	0.266826	0.055263	0.647740
Gene 10	0.898361	0.064430	0.038205	0.249196	0.372424	0.023148	0.363086	0.546274

Genes 1–8 each have a significant eSNP. Since we added noise to the expression data, the *p* values of the eSNPs are different. Genes 9 and 10 do not have eSNPs.

Table 2. GWAS *p* values in synthetic data

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8
p value	**0.009447**	**0.000018**	**0.000354**	**0.001101**	0.701814	0.988640	0.641454	0.459382

We first simulated the genotype of 8 SNPs in 2,000 samples. Based on the genotype, we simulated 10 genes' expression. Then, we chose 5 genes' expression to simulate the phenotype. Standard χ^2 test was used to get the GWAS *p* values. Since we added noise to the expression and phenotype, SNP1 is not very significant in this GWAS result.

(chisq.test function in R) was applied to get the association between SNPs and phenotype (GWAS). The *p* values of the association tests are shown in Tables 1 and 2. Based on the eQTL *p* value and GWAS *p* value, we applied the Sherlock method to get the log-LR of each gene:

$$\log \frac{P\left(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid Z_i = 1\right)}{P\left(\boldsymbol{X}_i, \boldsymbol{Y}_i \mid Z_i = 0\right)},$$

which is shown in Table 3.

In Tables 1 and 2, four of the causal genes (genes 1–4) have an individual eSNP that is significantly associated with the disease, while the non-causal genes (genes 5–8) each have a significant eSNP but this is not correlated with the disease phenotype. Thus, genes 1–4 have high log-LR values, while genes 5–8 have low log-LR values as shown in Table 3. Genes 9 and 10 do not have any eSNPs associated with them; thus, we cannot be sure whether these two genes are associated with the disease based on the colocalization of eQTL and the GWAS profiles. But when utilizing the information from the PPI network, the posterior for these two genes to be related to the disease differentiates. The posterior mean of gene 9 becomes larger than that of gene 10 with the help of its 4 positive neighbors. This simulation demonstrates that when the guilty-by-association assumption is satisfied, our method can prioritize those genes which are densely connected with genes owning high log-LR values.

The performance of Sherlock, COLOC, and GeP-HMRF are assessed under 4 different effect sizes and 4 sets of prior parameter λ_0, as well as 3 networks. Since the NetWAS method does not accept the user-customized networks, we cannot implement and compare the NetWAS method in this simulation. Under each setting of effect size, we repeated the above simulation 1,000 times, i.e., we simulate 1,000 eQTL datasets and 1,000 GWAS datasets. Sherlock, COLOC, and GeP-HMRF are applied to these datasets under each effect size to get the average performance. In addition, GeP-HMRF is applied using 4 different prior parameters λ_0 and 3 networks, as shown in Table 4.

In 14 out of 16 settings (4 different effect sizes and 4 different sets of prior parameters), GeP-HMRF(a) corresponding to the network (a) has a higher area under the curve (AUC) than Sherlock and the COLOC method. The

Table 3. log-likelihood ratio (log-LR) and posterior means of each gene

ID	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7	Gene 8	Gene 9	Gene 10
log-LR	0.097	4.830	3.933	1.044	–0.001	–2.130	–1.361	–0.037	0.002	0.009
Posterior mean	0.401	0.986	0.970	0.620	0.297	0.056	0.099	0.307	0.445	0.317
p value	0.314	0.037	0.094	0.253	0.766	0.994	0.907	0.721	0.308	0.764

The log-LR cannot distinguish Genes 9 and 10, since they do not own eSNPs. However, adding network information will help improve the rank of the posterior mean of Gene 9, because Gene 9 has positive neighbors.

Table 4. Average performance of Sherlock, COLOC, and GeP-HMRF under different prior parameters and effect sizes, as well as different networks

Effect size (θ_i, ω_i)	Method	(–1, 0.5, –0.01)			(–1, 0.25, –0.01)			(–1, 0.1, –0.01)			(–2, 0.25, –0.01)		
		TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
(0.5, 1.0)	Sherlock	0.378	0.035	0.918	0.378	0.035	0.918	0.378	0.035	0.918	0.375	0.034	0.917
	COLOC	0.382	0.036	0.882	0.382	0.036	0.882	0.382	0.036	0.882	0.382	0.036	0.882
	GeP-HMRF(a)	0.497	0.001	**0.989**	0.452	0.001	**0.981**	0.445	0.001	**0.980**	0.426	0.001	**0.973**
	GeP-HMRF(b)	0.394	0.002	0.975	0.386	0.004	0.969	0.394	0.004	0.958	0.390	0.002	0.962
	GeP-HMRF(c)	0.394	0.004	0.958	0.404	0.004	0.959	0.418	0.004	0.948	0.408	0.004	0.954
(0.4, 1.0)	Sherlock	0.375	0.032	0.892	0.375	0.032	0.892	0.375	0.032	0.892	0.375	0.032	0.892
	COLOC	0.371	0.037	0.808	0.371	0.037	0.808	0.371	0.037	0.808	0.371	0.037	0.808
	GeP-HMRF(a)	0.504	0.001	**0.959**	0.462	0.003	**0.939**	0.446	0.004	**0.933**	0.437	0.005	**0.924**
	GeP-HMRF(b)	0.366	0.002	0.924	0.370	0.004	0.912	0.402	0.006	0.903	0.390	0.008	0.907
	GeP-HMRF(c)	0.390	0.008	0.893	0.392	0.010	0.896	0.406	0.008	0.885	0.416	0.006	0.894
(0.3, 1.0)	Sherlock	0.330	0.047	0.805	0.330	0.047	0.805	0.330	0.047	**0.805**	0.330	0.047	**0.805**
	COLOC	0.343	0.053	0.697	0.343	0.053	0.697	0.343	0.053	0.697	0.343	0.053	0.697
	GeP-HMRF(a)	0.451	0.016	**0.843**	0.421	0.024	**0.815**	0.384	0.029	0.800	0.380	0.031	0.790
	GeP-HMRF(b)	0.322	0.028	0.776	0.346	0.038	0.772	0.342	0.054	0.752	0.356	0.032	0.747
	GeP-HMRF(c)	0.356	0.074	0.718	0.348	0.072	0.724	0.354	0.068	0.723	0.342	0.062	0.728
(0.4, 0.5)	Sherlock	0.329	0.053	0.814	0.329	0.053	0.814	0.329	0.053	0.814	0.329	0.053	0.814
	COLOC	0.338	0.054	0.742	0.338	0.054	0.742	0.338	0.054	0.742	0.338	0.054	0.742
	GeP-HMRF(a)	0.457	0.012	**0.886**	0.417	0.016	**0.857**	0.395	0.020	**0.844**	0.384	0.028	**0.840**
	GeP-HMRF(b)	0.346	0.014	0.839	0.350	0.030	0.829	0.344	0.046	0.817	0.334	0.044	0.820
	GeP-HMRF(c)	0.354	0.044	0.801	0.348	0.044	0.803	0.348	0.070	0.805	0.342	0.058	0.797

GeP-HMRF based on network (a) outperforms Sherlock and COLOC in 14 of 16 settings. As θ_i increases from 0.3 to 0.5, the effect size of SNPs also increases, along with the improvement of AUCs of all 3 methods.

TPR, true-positive rate; FPR, false-positive rate; AUC, area under the curve.

Bold indicates that this algorithm performed the best. (a), (b), (c) stand for different networks, see text for more details.

average true-positive rate and false-positive rate are the results given a cutoff of 0.1 on the *p* value. With increasing effect size, the AUCs of Sherlock, COLOC, and GeP-HMRF also increase. Also, GeP-HMRF applied to network a shows better results than those for network (b) which are in turn better than those for network (c). Network (a) has a high proportion of (1, 1) edges and (0, 0) edges and, thus, is an ideal example for the guilty-by-association assumption. When applied to network (b), with 2 (1, 1) edges, 5 (1, 0) edges, and 7 (0, 0) edges, GeP-HMRF still performs better than Sherlock and COLOC in 12 circumstances, especially in the cases with larger effect sizes. Network (c) has 2 (1, 1) edges, 5 (1, 0) edges, and 2 (0, 0) edges, and thus has the highest proportion of (1, 0) edges, which is not consistent with the guilty-by-association assumption. Since Sherlock and COLOC do not use

Table 5. The distribution of edges under different prior parameters and different networks

Prior parameters	Network (a)				Network (b)				Network (c)			
	*N*11	*N*01	*N*00	*N*1	*N*11	*N*01	*N*00	*N*1	*N*11	*N*01	*N*00	*N*1
(–1, 0.50, –0.01)	0.229	0.419	0.352	0.435	0.254	0.417	0.329	0.424	0.144	0.399	0.457	0.336
(–1, 0.25, –0.01)	0.117	0.408	0.475	0.320	0.121	0.415	0.464	0.317	0.100	0.392	0.507	0.293
(–1, 0.10, –0.01)	0.086	0.397	0.517	0.285	0.087	0.395	0.519	0.281	0.084	0.392	0.524	0.278
(–2, 0.25, –0.01)	0.020	0.219	0.761	0.129	0.021	0.221	0.758	0.129	0.018	0.213	0.769	0.125

From equation 4, we know that α controls the general proportion of positive nodes (*N*1) in a network. $\beta_1 - \beta_2$ affects the proportion of (1, 1) edges, while $\beta_2 - \beta_3$ affects the proportion of (0, 0) edges. With a decreasing $\beta_1 - \beta_2$, there are less positive genes and (1, 1) edges.

Table 6. Comparison of Sherlock, COLOC, NetWAS, and GeP-HMRF model on 9 GWAS datasets [27–32]

Datasets	Source	AUC				Top 50 prediction			
		Sherlock	COLOC	GeP-HMRF	NetWAS	Sherlock	COLOC	GeP-HMRF	NetWAS
AMD	Fritsche et al. [27]	0.619	0.567	**0.682**	0.570	5	2	**7**	1
Crohn disease	Barrett et al. [28]	0.666	0.556	**0.701**	0.637	3	**5**	3	1
Crohn disease	Franke et al. [29]	0.664	0.537	**0.680**	0.593	**3**	**3**	**3**	0
Crohn disease	Liu et al. [30]	0.669	0.553	**0.724**	0.628	3	3	**4**	0
HDL cholesterol	Do et al. [31]	0.776	0.508	**0.807**	0.527	2	1	**3**	1
HDL cholesterol	Teslovich et al. [32]	0.679	0.546	**0.780**	0.537	**2**	1	**2**	0
LDL cholesterol	Do et al. [31]	0.627	0.506	**0.704**	0.644	0	1	**2**	0
Total cholesterol	Do et al. [31]	0.739	0.501	**0.779**	0.677	0	1	**3**	0
Total cholesterol	Teslovich et al. [32]	0.695	0.465	**0.739**	0.666	**2**	1	0	1

We applied the 4 methods on 9 GWAS datasets. Column 1: GWAS names, Column 2: sources of the GWAS datasets, Columns 3–6: areas under the curve from the four algorithms, Columns 7–10: overlap with the MalaCards database in the top 50 predictions of the four algorithms. Numbers in bold indicate that this algorithm performed the best.

network information, this simulation shows that GeP-HMRF can improve the inference when the guilty-by-association assumption is satisfied.

Prior parameters can influence the edges' distributions, see Table 5. From equation 4, we know that $\beta_1 - \beta_2$ is the weight added to the neighbor genes when the gene is positive. A larger $\beta_1 - \beta_2$ leads to a higher proportion of positive genes (see column *N*1) and a higher proportion of (1, 1) edges (see column *N*11), while a smaller $\beta_1 - \beta_2$ leads to more negative genes and a lower proportion of (1, 1) edges. In addition, the influence of $\beta_1 - \beta_2$ depends on the network topology, too. For example, networks (a) and (b) are relatively denser than network (c). Even with the same prior parameter (–1, 0.5, –0.01), the distribution of the 3 kinds of edges are different in the 3 networks. The parameter $\beta_1 - \beta_2$ tends to have a bigger influence in dense networks, since the fold changes of *N*1 in networks (a) and (b) are larger than those in network (c), when $\beta_1 - \beta_2$ drops from 0.5 to 0.25. In network (a), the disease-related genes are densely connected. So, the performance of GeP-HMRF achieves the best results when $\beta_1 - \beta_2 = 0.5$, although the difference is small. In the sparse network (c), $\beta_1 - \beta_2$ has a smaller influence. The performance of GeP-HMRF does not show a trend with a decreasing $\beta_1 - \beta_2$.

Application of GeP-HMRF to GWASs of Complex Human Phenotypes

We applied GeP-HMRF, Sherlock, COLOC as well as NetWAS methods to analyze 9 GWAS datasets (see Table 6) and 1 merged eQTL dataset from GTEx version 6 [33]. See online supplementary section 1.5 ("Sources for the 9 GWAS datasets and eQTL dataset used in real data analysis") for the links to the GWAS dataset and details about the eQTL dataset. We utilized the PPI information from

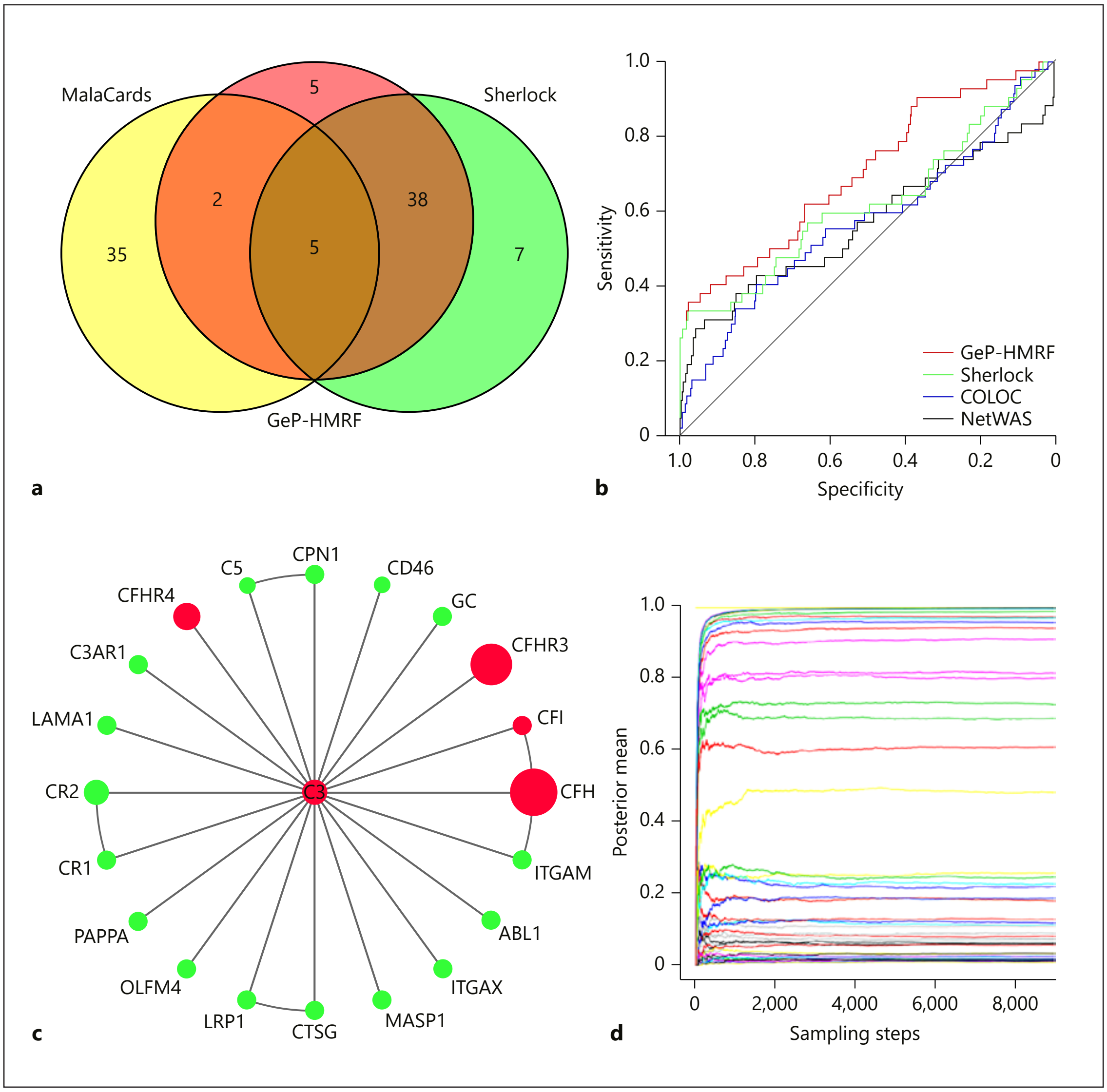

Fig. 2. Comparison of GeP-HMRF with Sherlock, COLOC, and NetWAS on the age-related macular degeneration (AMD) dataset. **a** Among the top 50 predicted genes, Sherlock has 5 genes overlapping with known AMD-related genes, while GeP-HMRF has 2 more overlaps. **b** Receiver-operating characteristic curve comparison of Sherlock, COLOC, GeP-HMRF, and NetWAS. **c** Among the neighboring genes of C3 in the protein-protein interaction network, red nodes are known AMD-related genes in MalaCards, and the node size represents the value of log-likelihood ratio (LR). C3 has 3 highly confident genes (CFH, CFHR3, and CFHR4) with large log-LR. **d** The posterior means of 50 randomly selected genes become stabilized after sampling 2,000 steps in the Markov chain.

the HPRD (human protein reference database [34]). We used the prior parameters (–5, 0.8, –0.001) for GeP-HMRF in the real-data analysis. The reason for selecting the prior parameter is explained in online supplementary section 1.6 ("Choice of prior parameters in real data analysis"). We took the MalaCards database [24] as the standard. The MalaCards database is a well-accepted database on disease-gene relations cited by more than 150 papers. Each disease in MalaCards is associated with a prioritized list of genes, obtained from 9 sources. The sources for gene-disease relations are from both manually curated (e.g., OMIM, Orphanet, SwissProtKB, ClinVar, and COSMIC) and text-mined resources (e.g., DISEASES, GeneTests, Novoseek, and GeneCards). MalaCards defines an overall disease-gene score, computed as a weighted sum of individual scores derived from the 9 sources. The individual scores depend on the level of manual curation of the information source, and on the confidence score as-

Table 7. Top 20 genes predicted to be associated with AMD

Gene	Sherlock		Posterior mean	GeP-HMRF	COLOC		NetWAS		Supporting evidence
	score	rank			score	rank	score	rank	
CFHR1	27.04	2	1	6.25E–04	6.49E–05	27312	6.46E–03	5962	MalaCards
CFHR3	22.73	7	1	6.25E–04	8.61E–09	27322	3.92E–02	4232	MalaCards
HTRA1	17.97	13	1	6.25E–04	7.41E–01	252	1.84E–01	576	MalaCards
TRIM31	17.26	14	1	6.25E–04	4.81E–02	9369	6.04E–02	3280	
PILRA	10.86	33	1	6.25E–04	9.82E–01	45	5.67E–02	3414	26691988, 24439028
ZFP57	26.45	3	1	6.25E–04	8.57E–01	165	–8.13E–02	10773	29739930
BTBD16	24.87	4	0.99999	6.25E–04	8.31E–32	27326	–6.04E–02	9722	23577725, 19405847
HLA-G	23.30	5	0.99999	6.25E–04	9.11E–01	136	3.73E–01	28	28442288
ZBTB41	22.93	6	0.99999	6.25E–04	6.38E–01	320	–5.23E–05	6342	21855625
HCG4P8	20.39	8	0.99999	6.25E–04	8.82E–01	152	–1.42E+00	22565	
HLA-F-AS1	19.43	10	0.99999	6.25E–04	9.03E–01	145	–1.31E+00	21532	
HCG23	16.56	16	0.99999	6.25E–04	9.90E–01	34	–1.35E+00	21897	
TMEM45A	14.51	20	0.99992	6.25E–04	1.14E–02	25601	–1.56E–01	13769	
PMS2P1	12.38	26	0.99937	6.25E–04	9.93E–01	25	–1.54E+00	23543	
DMBT1	12.03	27	0.99925	6.25E–04	4.71E–05	27314	–2.51E–02	7818	16642439
RDH5	11.70	30	0.99850	6.25E–04	9.99E–01	8	–3.17E–01	16030	11448328, 12967826
GPR108	6.00	77	0.73070	6.25E–04	4.57E–36	27327	4.80E–02	3807	29739930
CFH	29.77	1	1	9.38E–04	4.29E–01	531	1.36E–01	1169	MalaCards
C4B	16.36	17	1	9.38E–04	9.93E–01	26	8.16E–02	2462	26742632, 7090374
RNF5	13.09	23	1	9.38E–04	9.94E–01	21	–2.13E–01	15038	

AMD, age-related macular degeneration.

Four are well-known AMD-related genes in MalaCards, and 9 genes are supported by literature (the PubMed ID of the supporting literature is shown).

signed by the text-mining source. We compared the performance of the 4 methods in two different ways: (1) we drew the receiver-operating characteristic curve and calculated the AUC for each algorithm; (2) for the top 50 predictions of each algorithm, we counted the number of genes known to be associated with the phenotype in MalaCards.

GeP-HMRF has the largest AUC (larger than Sherlock, COLOC, and NetWAS) on all 9 datasets. The average difference between the GeP-HMRF and Sherlock methods is 0.0514 with a *p* value 0.0374 (using a two-sided *t* test). GeP-HMRF also significantly outperforms COLOC and NetWAS with the average increase of AUCs of 0.2063 (*p* value 2.2335×10^{-8}) and 0.1204 (*p* value 9.6620×10^{-5}), respectively. Looking at the aspect of the top 50 predictions, GeP-HMRF has the highest overlap with the MalaCards database in 5 out of the 9 datasets. In another 2 datasets, GeP-HMRF achieved the equal best performance with Sherlock or COLOC. There are two cases in which Sherlock or COLOC found more genes in MalaCards than GeP-HMRF. The scores of the 4 methods in each GWAS datasets are supplied in the online supplementary files.

In the following, we describe in more details the results and the comparison between different algorithms using AMD as an example. Late AMD is the leading cause of blindness of elder people in Western countries. It has a prevalence of 0.05% before the age of 50 years and increases to 11.8% after the age of 80 years [35]. As shown in Figure 2a, the overlap of the top 50 genes predicted by Sherlock and the MalaCards database is 5, while GeP-HMRF hits two more genes. The details of SNPs supporting these 7 genes are listed in online supplementary section 1.7 (“Details of SNPs supporting the 7 positives gene in the top 50 predictions of GeP-HMRF”). The overlap between the top 50 GeP-HMRF-predicted genes and those in MalaCards is highly significant with a *p* value of 3.61×10^{-119}(Fisher exact test). Also, GeP-HMRF has a larger AUC than Sherlock, COLOC, and NetWAS (Fig. 2b). Table 7 shows the top 20 genes and their associated statistics predicted by GeP-HMRF. The top 3 genes (CFHR1, CGHR3, and HTRA1) are known to be associated with AMD in MalaCards, as well as the gene CFH (ranked 18). In addition, 9 other genes (PILRA [27, 36], ZFP57 [37], BTBD16 [38–40], HLA-G [41], ZBTB41 [42, 43], DMBT1 [44], RDH5 [45, 46], GPR108 [37, 47], C4B [48, 49]) are reported as AMD-related in the literature. For example, the RDH5 gene encodes the protein retinol dehydrogenase 5, which is an enzyme catalyzing the biosynthesis of 11-cis retinaldehyde. 11-cis retinaldehyde constitutes the universal chromophore of visual pigments. Two independent studies from Wada et al. [45] and Yamamoto et al. [46] revealed that mutations of RDH5 are associated with a degenerative macula and scattered white dots in the retina.

To illustrate how the PPI network helps GeP-HMRF to identify disease-associated genes, we took gene C3 as an example. C3 is a well-known AMD-related gene, replicated in many independent studies [50–52]. The gene was ranked 40 by GeP-HMRF (not in Table 7). Figure 2c shows the direct neighbors of C3, with known AMD-related genes labeled in red. The size of nodes in Figure 2c represents the value of log-LR. It shows that C3 interacts with CFH (log-LR 29.77), CFHR3 (log-LR 22.73), CFHR4 (log-LR 6.90), and CR2 (log-LR 4.77) in PPI network, while C3 is located far away from those 4 genes (C3 is located on chr19, while the other 4 genes are located on chr1). These neighboring genes with a high log-LR have a high confidence to be associated with AMD, and increase the C3 gene's posterior probability of also being associated with AMD. The average log-LR of the neighbors of the C3 gene is significantly higher than that of all genes (*p* value 0.041, *t* test). When computing the log-LR based on the eQTL and GWAS profiles only, the log-LR of C3 is 5.97, and its rank is 78. After combining the PPI information with the eQTL and GWAS profile, the rank of C3 in GeP-HMRF increases to 40. The rank of C3 is promoted by the help from its neighbors (details on gene C3 and its neighbors can be found in online supplementary section 1.8 "Neighbor genes of C3").

For the computation of the posterior means by Gibbs sampling, we sampled 5,000 steps of each Markov chain and 1,000 steps for burn-in. Figure 2d shows the fluctuation of the posterior means of 50 randomly selected genes. In the figure, we sample 9,000 steps to get as many steps as possible from the Markov chain to demonstrate the long-term behavior. After 2,000 steps, the posterior mean of each gene became stabilized, which is a sign of convergence.

Discussion

GeP-HMRF is based on the hypothesis of "guilty by association," which was used by the previously developed algorithms that integrate GWAS with gene network information. Chen et al. [16] showed that the disease-associated genes tend to be connected. We also observed that the known disease-related genes connect with each other more densely than the average connectivity between genes not associated with the disease in all 3 diseases we analyzed. As was shown for C3 in the example of AMD, GeP-HMRF can help find some hot spots with high log-LR genes clustering together. These hotspot genes in the cluster share functional similarities or have physical interactions. These clusters can provide a new perspective for understanding the mechanism of disease. However, caution is needed as such analysis may also generate additional false positives, since the hypothesis of guilty by association may not always be true.

One challenge facing GeP-HMRF or similar approaches is that a gene interaction network is not static but context dependent. The network can vary among different tissues or be influenced by the disease state. For many diseases, the tissue of origin is not known. We have tried to apply GeP-HMRF on 9 GWAS datasets using 33 tissue-specific eQTL datasets as input. GeP-HMRF achieved good performance on Crohn disease using the eQTL from whole blood, esophagus mucosa, and adrenal gland tissue. Crohn disease is a complex chronic immune disorder that primarily affects the digestive system. Esophagus mucosa is one of the affected tissues, but the small intestine and colon are more often affected. Lage et al. [53] derived disease-tissue association scores based on the co-occurrence of disease-related and tissue-related terms in PubMed abstracts. The top 6 tissues/cell types for Crohn disease are adrenal cortex, liver, appendix, CD4 T cells, skin, and monocytes. Whole blood contains many immune cells like the CD4 T cells and monocytes, which might be the reason it performs well. For the cholesterol, the eQTL from adipose tissue gives a good result. Adipose tissue is a major site for cholesterol storage [54], which might explain the reason for the good performance. More details on the results from tissue-specific eQTLs are listed in online supplementary section 1.9 ("Apply GeP-HMRF based on 33 tissue-specific eQTL datasets from GTEx"). The merged eQTL dataset can achieve moderate performance. Merged eQTLs can be an option when the disease-related tissue is unknown. In addition, our knowledge of the network is far from complete with many missing nodes and links. Thus, future work will greatly benefit from better understanding of the tissue origin of the diseases and more comprehensive knowledge of gene networks in different tissues, such that may be provided by the human cell atlas program.

We have developed a unified statistical framework to integrate GWAS, eQTL, and PPI data to infer disease-related genes. Our approach combines the strengths of previously developed methods that integrate GWAS and eQTL data and those that utilize both GWAS and gene network information. We have implemented our method in a computational algorithm called GeP-HMRF. We tested the performance of GeP-HMRF across a number of GWAS datasets and showed that GeP-HMRF has significantly improved the ability to identify disease-associated genes. Our approach can be generalized to incorporate other types of molecular trait information such as epigenomic or metabolomics data.

Availability of Data and Materials

The datasets generated and analyzed during the current study are available from https://github.com/JunWangmath/GeP-HMRF.

Acknowledgments

This work has been partly supported by the National Key Basic Research project of China (No. 2015CB910303), the National Key Research and Development program of China (No. 2016YFA0502303), the National Natural Science Foundation of China (No. 31471246), and the China Scholarship Council. It is also supported by NIH grant AG043080, and a grant from the California Institute of Quantitative Biology.

References

1 Visscher PM, Brown MA, McCarthy MI, Yang J: Five years of GWAS discovery. Am J Hum Genet 2012;90:7–24.

2 Nica AC, Dermitzakis ET: Using gene expression to investigate the genetic basis of complex disorders. Hum Mol Genet 2008; 17:R129–R134.

3 Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, Dermitzakis ET: Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet 2010;6: 1000895.

4 Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V: Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet 2014;10:1004383.

5 Hormozdiari F, van de Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E: Colocalization of GWAS and eQTL signals detects target genes. Am J Hum Genet 2016;99:1245–1260.

6 Gamazon ER, Wheeler HE, Shah KP, et al: A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 2015;47:1091–1098.

7 Guo H, Fortune MD, Burren OS, Schofield E, Todd JA, Wallace C: Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. Hum Mol Genet 2015;24:3305–3313.

8 Fortune MD, Guo H, Burren O, et al: Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. Nat Genet 2015;47:839–846.

9 Wen X, Pique-Regi R, Luca F: Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. PLoS Genet 2017;13:1006646.

10 Schadt EE, Molony C, Chudin E, et al: Mapping the genetic architecture of gene expression in human liver. PLoS Biol 2008;6:e107.

11 Kang HP, Yang X, Chen R, Zhang B, Corona E, Schadt EE, Butte AJ: Integration of disease-specific single nucleotide polymorphisms, expression quantitative trait loci and coexpression networks reveal novel candidate genes for type 2 diabetes. Diabetologia 2012;55: 2205–2213.

12 Emilsson V, Thorleifsson G, Zhang B, et al: Genetics of gene expression and its effect on disease. Nature 2008;452:423–428.

13 He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, Li H: Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. Am J Hum Genet 2013;92:667–680.

14 1000 Genomes Project Consortium: A global reference for human genetic variation. Nature 2015;526:68–74.

15 International HapMap Consortium: The International HapMap project. Nature 2003; 426:789–796.

16 Chen M, Cho J, Zhao H: Incorporating biological pathways via a Markov random field model in genome-wide association studies. PLoS Genet 2011;7:1001353.

17 Shim JE, Bang C, Yang S, Lee T, Hwang S, Kim CY, Singh-Blom UM, Marcotte EM, Lee I: GWAB: a web server for the network-based boosting of human genome-wide association data. Nucleic Acids Res 2017;45:W154–W161.

18 Greene CS, Krishnan A, Wong AK, et al: Understanding multicellular function and disease with human tissue-specific networks. Nat Genet 2015;47:569–576.

19 Wu M, Zeng W, Liu W, Zhang Y, Chen T, Jiang R: Integrating embeddings of multiple gene networks to prioritize complex disease-associated genes; in 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp 208–215, IEEE.

20 Oliver S: Guilt-by-association goes global. Nature 2000;403:601–603.

21 Zhang Y, Brady M, Smith S: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 2001;20:45–57.

22 Deng M, Zhang K, Mehta S, Chen T, Sun F: Prediction of protein function using protein-protein interaction data. J Comput Biol 2003; 10:947–960.

23 Kunsch H, Geman S, Kehagias A, et al: Hidden Markov random fields. Ann Appl Probab 1995;5:577–602.

24 Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Stein TI, Bahir I, Belinky F, Morrey CP, Safran M, Lancet D: MalaCards: an integrated compendium for diseases and their annotation. Database 2013;12:bat018.

25 Andrieu C, De Freitas N, Doucet A, Jordan MI: An introduction to MCMC for machine learning. Machine Learn 2003;50:5–43.

26 Su Z, Marchini J, Donnelly P: Hapgen2: simulation of multiple disease SNPs. Bioinformatics 2011;27:2304–2305.

27 Lonsdale J, Thomas J, Salvatore M, et al: The genotype-tissue expression (GTEX) project. Nat Genet 2013;45:580–585.

28 Prasad TK, Goel R, Kandasamy K, et al: Human protein reference database. Nucl Acids Res 2009;37(suppl 1):767–772.

29 Fritsche LG, Igl W, Bailey JN, et al: A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat Genet 2016;48:134–143.

30 Barrett JC, Hansoul S, Nicolae DL, et al: Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 2008;40:955–962.

31 Franke A, McGovern DP, Barrett JC, et al: Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010;42:1118–1125.

32 Liu JZ, van Sommeren S, Huang H, et al: Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet 2015;47:979–986.

33 Do R, Willer CJ, Schmidt EM, et al: Common variants associated with plasma triglycerides and risk for coronary artery disease. Nat Genet 2013;45:1345–1352.

34 Teslovich TM, Musunuru K, Smith AV, et al: Biological, clinical and population relevance of 95 loci for blood lipids. Nature 2010;466: 707–713.

35 De Jong PT: Age-related macular degeneration. N Engl J Med 2006;355:1474–1485.

36 Logue MW, Schu M, Vardarajan BN, Farrell J, Lunetta KL, Jun G, Baldwin CT, DeAngelis MM, Farrer LA: A search for age-related macular degeneration risk variants in Alzheimer disease genes and pathways. Neurobiol Aging 2014;35:1510–1517.

37 Barbeira AN, Dickinson SP, Bonazzola R, et al: Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun 2018;9:1825.

38 Naj AC, Scott WK, Courtenay MD, Cade WH, Schwartz SG, Kovach JL, Agarwal A, Wang G, Haines JL, Pericak-Vance MA: Genetic factors in nonsmokers with age-related macular degeneration revealed through genome-wide gene-environment interaction analysis. Ann Hum Genet 2013;77:215–231.

39 Courtenay MD: Gene-Environment Interaction in Age-Related Macular Degeneration: Exogenous Estrogen, Cigarette Smoking, and VEGF Pathway Polymorphisms. PhD thesis, University of Miami, 2014.

40 Swaroop A, Chew EY, Rickman CB, Abecasis GR: Unraveling a multifactorial late-onset disease: from genetic susceptibility to disease mechanisms for age-related macular degeneration. Ann Rev Genom Hum Genet 2009; 10:19–43.

41 Svendsen SG, Wu CL, Juel HB, Faber C, Carosella ED, LeMaoult J, Nissen MH: Expression of HLA-G in the retinal pigment epithelial cell line, ARPE-19. Invest Ophthalmol Visual Sci 2014;55:2950–2950.

42 Williamson JF, McLure CA, Guymer RH, Baird PN, Millman J, Cantsilieris S, Dawkins RL: Almost total protection from age-related macular degeneration by haplotypes of the regulators of complement activation. Genomics 2011;98:412–421.

43 Scheetz TE, Fingert JH, Wang K, et al: A genome-wide association study for primary open angle glaucoma and macular degeneration reveals novel loci. PLoS One 2013;8: 58657.

44 Schmidt S, Hauser MA, Scott WK, et al: Cigarette smoking strongly modifies the association of loc387715 and age-related macular degeneration. Am J Hum Genet 2006;78:852–864.

45 Wada Y, Abe T, Sato H, Tamai M: A novel Gly35Ser mutation in the RDH5 gene in a Japanese family with fundus albipunctatus associated with cone dystrophy. Arch Ophthalmol 2001;119:1059–1063.

46 Yamamoto H, Yakushijin K, Kusuhara S, Escaño MFT, Nagai A, Negi A: A novel RDH5 gene mutation in a patient with fundus albipunctatus presenting with macular atrophy and fading white dots. Am J Ophthalmol 2003;136:572–574.

47 Morris N, Pulagam VL, Haines J, Iyengar SK: Examining AMD GWAS signals in light of regulatory eQTL variants. Invest Ophthalmol Visual Sci 2014;55:2216–2216.

48 van Lookeren Campagne M, Strauss EC, Yaspan BL: Age-related macular degeneration: complement in action. Immunobiology 2016; 221:733–739.

49 Grassmann F, Cantsilieris S, Schulz-Kuhnt AS, et al: Multiallelic copy number variation in the complement component 4A (C4A) gene is associated with late-stage age-related macular degeneration (AMD). J Neuroinflamm 2016;13:81.

50 Seddon JM, Yu Y, Miller EC, et al: Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration. Nat Genet 2013;45:1366–1370.

51 Yates JR, Sepp T, Matharu BK, et al: Complement C3 variant and the risk of age-related macular degeneration. N Engl J Med 2007; 357:553–561.

52 Spencer KL, Olson LM, Anderson BM, Schnetz-Boutaud N, Scott WK, GallinsP, Agarwal A, Postel EA, Pericak-Vance MA, Haines JL: C3 R102G polymorphism increases risk of age-related macular degeneration. Hum Mol Genet 2008;17:1821–1824.

53 Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S: A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc Natl Acad Sci USA 2008;105:20870–20875.

54 Iwayama T, Nitobe J, Watanabe T, et al: The role of epicardial adipose tissue in coronary artery disease in non-obese patients. J Cardiol 2014;63:344–349.