Lee et al. BioData Mining (2020) 13:6
https://doi.org/10.1186/s13040-020-00216-9

BioData Mining

METHODOLOGY Open Access

Identification of therapeutic targets from genetic association studies using hierarchical component analysis

Hao-Chih Lee[1,2], Osamu Ichikawa[1,2,3], Benjamin S. Glicksberg[1,2,4], Aparna A. Divaraniya[1,2], Christine E. Becker[1,2], Pankaj Agarwal[5] and Joel T. Dudley[1,2*]

* Correspondence: joel.dudley@gmail.com
[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[2]Institute for Next Generation Healthcare, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
Full list of author information is available at the end of the article

Abstract

Background: Mapping disease-associated genetic variants to complex disease pathophysiology is a major challenge in translating findings from genome-wide association studies into novel therapeutic opportunities. The difficulty lies in our limited understanding of how phenotypic traits arise from non-coding genetic variants in highly organized biological systems with heterogeneous gene expression across cells and tissues.

Results: We present a novel strategy, called GWAS component analysis, for transferring disease associations from single-nucleotide polymorphisms to co-expression modules by stacking models trained using reference genome and tissue-specific gene expression data. Application of this method to genome-wide association studies of blood cell counts confirmed that it could detect gene sets enriched in expected cell types. In addition, coupling of our method with Bayesian networks enables GWAS components to be used to discover drug targets.

Conclusions: We tested genome-wide associations of four disease phenotypes, including age-related macular degeneration, Crohn's disease, ulcerative colitis and rheumatoid arthritis, and demonstrated the proposed method could select more functional genes than S-PrediXcan, the previous single-step model for predicting gene-level associations from SNP-level associations.

Keywords: Genome-wide association study, Network biology, Gene candidate discovery

Introduction

Genome-wide association studies (GWAS) seek to identify how genetic variations, typically represented as single-nucleotide polymorphisms (SNPs), contribute to variability in expression of phenotypic traits or diseases across the population. GWAS, which is made possible by the availability of the reference human genome [1, 2], represents contemporary efforts to map collective genetic architecture to common diseases. Since the first GWAS in 2005, applications of this technique have facilitated identification of risk

variants for various diseases, including age-related macular degeneration [3], inflammatory bowel disease (IBD) [4–6], type 2 diabetes [7, 8] and many others. For example, GWAS have discovered over 200 risk loci for IBD that encompass a wide range of biological processes, including innate and adaptive immunity, autophagy, and epithelial permeability [9].

Currently, identification of therapeutic targets from GWAS remains difficult and relatively inefficient, largely because SNP associations often do not directly indicate optimal therapeutic targets nor the complex mechanism underlying disease pathogenesis [10]. The presence of non-coding causal SNPs is one of the major obstacles to functional implications of the mechanisms of disease [11]. Studies have demonstrated the widely-spread SNP associations with tiny effect sizes can collectively contribute to a large portion of heritability for complex traits such as schizophrenia [12] and height [13]. These ubiquitous genetic signals across genome, acting directly on any genes, may propagate through interconnected gene regulatory network to affect functions of disease-related genes [14]. Studies have also shown that hub genes, genes interacting with many other genes, are subject to negative evolutionary selection [15–17], hinting the potential of combing network topology with genetic signals in search of therapeutic targets. This "omnigenic" point of view thus make us wonder how to distill the ubiquitous genetic signals into effects on the gene network.

To this end, we developed a hierarchical approach that maps disease associations from SNPs to genes, and eventually to transcriptomic modulation. We first developed tissue-specific co-expression networks to determine co-expression modules, a collection of genes that are modulated concurringly, and used it to demonstrate that genetic associations can be hierarchically mapped to these gene modules. We demonstrate that this approach, requiring only GWAS summary data, determines module associations as accurate as those computed directly from individual-level data. We then applied this technique to GWAS of four complex disorder to demonstrate the applicability of GWAS component analysis and gene candidate discovery.

Methods

Overview of the proposed method

We took a two-stage approach to discover disease-associated gene components (Fig. 1a). First, we mapped SNP associations to gene associations using S-PrediXcan [18], which utilizes a linear model that maps SNP dosage to gene expression to predict gene associations Z_g^G from SNP associations Z_i^X (Fig. 1b). Both associations are linked by

$$Z_g^G \approx \Sigma_{i \in Model_g} W_{gi} \frac{\sigma_i^X}{\sigma_g^G} Z_i^X \tag{1}$$

where W is the weight matrix of the linear model fitted using individual-level data from the Genotype–Tissue Expression project (GTEx) [19], and σ_g^G and σ_i^X are the standard deviations of a gene g and a SNP i, respectively.

In the second stage, we estimated the disease association of an eigen-gene component L_l that represents the activity of a co-expression module. A co-expression module represents a group of genes whose expression is collectively modulated, while the eigen-gene component summarizes the overall expression of this gene group by the largest

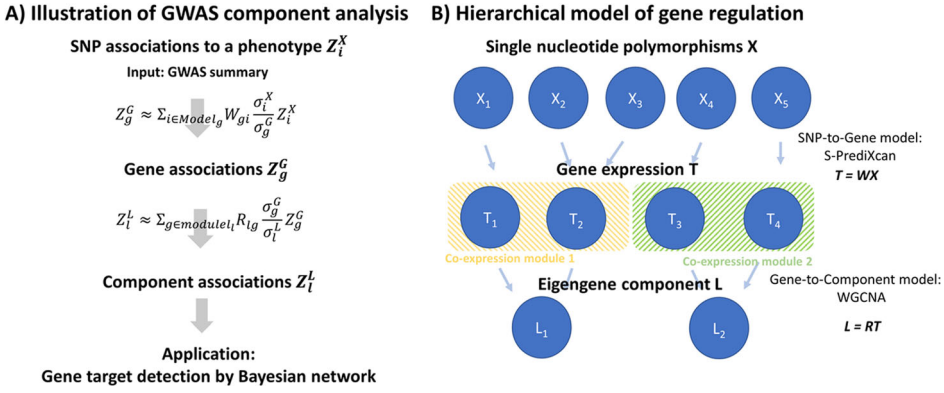

Fig. 1 Schematic diagram of GWAS component analysis. **a** We designed a two-stage method to map SNP associations to component associations. Associated GWAS components were integrated with Bayesian networks to facilitate therapeutic discovery. GWAS component is a component with a score Z_l^L significantly deviating from 0. **b** In the first stage, S-PrediXcan was used to infer gene associations from SNP associations. Gene-to-Component model were fitted by WGCNA, and the corresponding associations were computed as in Eq. (2) (see text). Models in B) were built using GTEx reference data

variation. Specifically, the eigengene of a co-expression module is defined as the first principal component of the measured gene expression profile in the module [20]. Given the linearity of principal component analysis, we can compute the eigengene by multiplying a weight matrix R to the gene expression matrix. We note that this is analogous to the way S-PrediXcan computes gene expressions from SNP dosages and thus the statistical theory of S-PrediXcan can be carried over, as we summarize in the following. Given the weight matrix R, the disease association Z_l^L of an eigen-gene component L_l can be estimated by

$$Z_l^L \approx \Sigma_{g \in Model_l} R_{lg} \frac{\sigma_g^G}{\sigma_l^L} Z_g^G, \tag{2}$$

where σ_l^L is the standard deviation of the eigen-gene component L_l. Equations (1) and (2) thus transfer the disease associations from genomic space to transcriptomic space, and ultimately to co-expression modules. Under the null hypothesis, Z_l^L is a standard normal random variable. We thus refer to a component, with a score Z_l^L significantly deviating from 0 as a genome-wide significant (GWAS) component.

We propose using Bayesian networks (BN) to discover putative causal genes of GWAS components. We sought to discover functional genes by determining the overlap between a GWAS component and a tissue-specific BN. The "functionality" of a gene candidate g_0 is evaluated based on the odds ratio of the overlap between its children on the BN and the GWAS component. Specifically, we selected putative causal genes by testing whether the set $S_1 = \{g \in B |\, g$ is in a GWAS component$\}$ is overrepresented by $S_2 = \{g \in B | g$ is downstream of g_0 in the BN$\}$, where B is the set of background genes defined by the intersection of genes used in constructing S – PrediXcan models and the Bayesian networks. The Bayesian networks were constructed using RIMBANet [21].

Computation of gene-level associations by S-PrediXcan

To map SNP associations to gene associations, we used the recently proposed method S-PrediXcan to predict tissue-specific gene associations. We briefly summarize S-PrediXcan as follows: given X_i, the allelic dosage for SNP i, T_g, the predicted expression of gene g, and Y, the phenotype of interest, S-PrediXcan assumes a pre-trained model that maps allelic dosages to the predicted expression by

$$T_g = \Sigma_{i \in Model_g} W_{gi} X_i + \epsilon \tag{3}$$

where W is the weight matrix of the linear model fitted using individual-level genotype data [18]. On top of this linear model, S-PrediXcan estimate the gene association $Z_g^G = \gamma_g / se(\gamma_g)$ from the SNP associations $Z_i^X = \beta_i / se(\beta_i)$, where β_i and γ_g are estimators of effect sizes and $se(\beta_i)$ and $se(\gamma_g)$ are the variances of the estimators of gene g and SNP i, respectively. Barbeira et al. [18] demonstrated that both associations are approximately related by

$$Z_g^G \approx \Sigma_{i \in Model_g} W_{gi} \frac{\sigma_i^X}{\sigma_g^G} Z_i^X \tag{4}$$

where σ_g^G and σ_i^X are the standard deviations of gene g and SNP l. Similar results were obtained via a different derivation [22]. We summarized their approximation as follows: Given random variables X_i whose covariance Γ is known, the association of its linear transformation $T_g = \Sigma_{i \in Model_g} W_{gi} X_i$ to trait Y can be approximated by Eq. (4), where $\sigma_g^G = \sum_{ij} W_{ig} \Gamma_{ij} W_{gj}$.

Similar methods exist for mapping SNP associations to gene associations. Several methods infer gene-level associations as aggregated effects of a group of SNPs by modeling the linkage disequilibrium (LD) structure using, for example, chi-squared statistics [23, 24] or hypothesis testing [25]. Another class of methods attempt to integrate expression quantitative trait loci (eQTLs) with GWAS signals. For example, COLOC seeks to determine whether eQTL and GWAS signals are consistent with a shared causal variant [26]. Summary mendelian randomization (SMR) includes instrumental variables to determine the causative effects of gene expressions on a phenotype from eQTLs [27]. TWAS [22] and S-PrediXcan [18] combine information of the LD structure and eQTLs into multivariate analysis to infer trait-associated genes. Theoretical and empirical comparison on COLOC, SMR, TWAS and S-PrediXcan can be found in [18].

Computation of GWAS component associations

Our proposed method further assumes that overall activity of a co-expression module, termed eigen-gene component L, can be represented by a mixture of gene expression T, i.e.,

$$L_l = \Sigma_{g \in Module_l} R_{lg} T_g.$$

The matrix R consists of the weighted contributions of genes to an eigen-gene component. Applying the relation in Eq. (4) to L, trait association Z_l^L can be approximated by

$$Z_l^L \approx \Sigma_{g \in Module_l} R_{lg} \frac{\sigma_g^G}{\sigma_l^L} Z_g^G$$

Building the eigen-gene component models

To determine the weight matrix R, we applied weighted correlation network analysis (WGCNA) to the GTEx RNA-seq data. Covariates were first removed following the procedure used in building S-PrediXcan models. For consistency, we confined the analysis to the same genes used in building S-PrediXcan models. Co-expression modules were estimated from each tissue independently. We tuned the minimum of module size to 5 to allow detection smaller modules. The eigen-gene component was then computed as the first principal component of the expression matrix of co-expressed genes.

Construction of Bayesian networks

The Bayesian networks (BN) were constructed using RIMBANet [21, 28, 29]. The estimation and validation of BNs are reported in previous studies [30, 31]. Briefly, GTEx data were first normalized to ensure a normal distribution, and then discretized into three clusters using the k-means approach. The number of clusters was adjusted to two if any of the three clusters contained only a few samples. Each gene was limited to having no more than three parent nodes. The final network was pooled into a consensus network from 1000 repeated runs. Cycles and weak edges were then pruned to ensure that the final network was a directed acyclic graph.

Simulation test

We simulated a scenario to test Eq. (1) using genotype data of 2504 individuals from the 1000 Genomes Project [32]. We first used S-PrediXcan to compute the predicted gene expression of these 2504 individuals. Eigen-gene components were then computed as weighted averages of these predicted gene expression using WGCNA models fitted from the GTEx Whole Blood data. We then simulated a trait caused by a single component as $Y = L_1 + \alpha\,\epsilon$ with a randomly selected eigen-gene component L_1. We tested the GWAS component method under various signal-to-noise ratios (SNRs) $std(L_1)/(std(L_1) + \alpha)$, which represents heritability in a broad sense. The selected component is referred to as the causal component, whereas the other components are non-causal. In this scenario, we expect to see a strong z-score from the selected component and minor signals from the other components. The associations to the eigen-gene components were then tested using 1) predicted eigen-gene components from genotypes of 2504 individuals and 2) the proposed GWAS component analysis.

In silico validations of putative targets

To evaluate these gene candidates, we conducted two in silico validations. First, we evaluated whether mutations in gene candidates could result in disease phenotypes in mouse models. The Phenotype/Alleles project of Mouse Genotype Informatics (MGI) is a database that provides rich information about mutant alleles and their resultant phenotypes in various mouse models [33]. We extracted the phenotypes of disease mouse models (Supplementary Table 6) from MGI. Phenotypes associated with gene candidates were also obtained from MGI. We considered a gene candidate to be a hit if its associated phenotypes were significantly enriched in the phenotypes of at least one disease mouse model.

Second, we evaluated whether the perturbation of a gene candidate could result in disease signatures in cell lines. For this purpose, we queried the characteristic direction of a single-gene perturbation, including shRNA knockdown, overexpression and ligand binding, from L1000CDS [34]. Disease characteristic direction signatures were constructed using crowd curated data CREED [35], including one AMD, five CD, 22 UC and seven RA case–control studies (Supplementary Table 7). The cosine distance was used to evaluate the relevance of two characteristic direction signatures. Specifically, a gene candidate was considered a LINCS hit if its characteristic direction was significantly correlated (cosine distance ~ 1) or anticorrelated (cosine distance ~ − 1) with at least one disease characteristic direction.

Statistical overrepresentation test

The overrepresentation test is a statistical test for determining whether the level of overlap between two sets is due to chance. The test requires three inputs: two sets S_1, S_2 to be compared and a background set B. It is assumed that the elements in S_1 and S_2 are all drawn from the background set B. The chance that the observed data were generated by random overlap can be evaluated by the hypergeometric distribution

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

where

$$a = \#(S_1 \cap S_2), b = \#(S_1 \setminus S_2), c = \#(S_2 \setminus S_1), d = \#(B \setminus (S_1 \cup S_2))$$

are elements of the contingency matrix. For significant associations between S_1 and S_2, we define enrichment odds ratio $OR = (a/b)/(c/d)$. Two sets are said to be enriched if OR is greater than 1 and p is less than a given threshold.

We summarize the overrepresentation tests used in our study below:

1. Bayesian network: in the results section, we sought to discover functional genes by determining the overlap between an associated component and a tissue-specific BN. The "functionality" of a gene candidate g_0 is evaluated based on the OR of the overlap between its children and the associated components. Specifically, the inputs for the overrepresentation test are as follows:

   $B = \{\text{genes used in S-PrediXcan}\} \cap \{\text{genes in the Bayesian Network}\}$

   $S_1 = \{g \in B | g \text{ is in an GWAS component}\}$

   $S_2 = \{g \in B | g \text{ is downstream of } g_0 \text{ in the Bayesian Network}\}$

2. Mouse genome informatics: in the results section, we evaluated in silico whether the mutation of a gene has been associated with relevant disease phenotypes in a mouse model. We set up the overrepresentation test as follows

   $B = \{\text{all mouse phenotypes in MGI}\}$

   $S_1 = \{p \in B | p \text{ is a phenotype associated with a gene candidate}\}$

$$S_2 = \{p \in B | p \text{ is a phenotype of the mouse model}\}$$

Matching characteristic directions

The characteristic direction is a computational method for finding a high-dimensional vector that best differentiates gene expression between cases and controls [36]. The characteristic direction, generally unit-normalized, is determined as the maximizer of the ratio of posteriors of two Gaussians with a shared covariance:

$$log\frac{P(G=0|X=x)}{P(G=1|X=x)} = \log\frac{\pi_0}{\pi_1} - \frac{1}{2}(\mu_0 - \mu_1)^T\Sigma^{-1}(\mu_0 - \mu_1) + x^T\Sigma^{-1}(\mu_0 - \mu_1).$$

G indicates the cases and controls. We aim to compare how similar the characteristic direction of a gene perturbation experiment is to the characteristic direction of a disease. We followed Clark. et al. to compute the similarity of two characteristic directions by the cosine distance:

$$d(v_1, v_2) = \langle v_1, v_2 \rangle / ||v_1|| \|v_2\|$$

To estimate the null distribution, we randomly sampled 10,000 characteristic directions and computed the cosine distance between sampled characteristic directions and the targeted disease characteristic direction. We found the empirical distribution is roughly bell-shape but slightly skewed. We thus used the average and standard deviation of this empirical distribution to normalize cosine distance. We call a LINCS hit if the absolute value of the normalized cosine distance from a given gene-perturbed characteristic direction to a target disease characteristic direction is larger than 1.96.

Multiple testing correction

We used the Holm-Sidak method to correct the family-wise error rate when required. Specifically:

1. When testing the association of disease and component we corrected the number of components tested in each tissue.
2. The LINCS database contains replicates of single-gene perturbation across cell type and time points. The CREEDS database also contains replicates of disease signatures. We thus corrected for the number of combinations of disease signatures and gene perturbation signatures when testing whether perturbation of a gene candidate could result in producing disease signatures in cell lines.
3. The MGI database may contain several mouse models of the same disease. We corrected the number of mouse models of each disease when testing whether the mutation of a gene has been associated with relevant disease in a mouse model.

Results

To determine the weight matrix R, we applied weighted correlation network analysis (WGCNA) to RNA-seq data from GTEx to infer co-expression modules (Methods). Among 44 tissues analyzed, we generally detected 213 ± 89 co-expression modules in one tissue. On average, each component contains 19 genes. In general, the co-expression modules determined by WGCNA are likely to reflect biological pathways

and gene functions [20], and we sought to probe if these co-expression modules were linked to genetics. We compared the co-expression modules to a multi-species co-expression network, in which the gene-gene interactions are present in multiple species and assumed to be genetically conserved [37]. Specifically, we formed a network by enumerating all gene-gene combinations within WGCNA modules and compare it to the multi-species co-expression network. We found that, for a single tissue, ~ 2% WGCNA edges are overlapped with the multi-species co-expression network despite the overlap is very significant (Odds ratios range from 1.45 to 39.87, Supplementary Table 1). In addition, 23% WGCNA edges, if detected in more than 2 tissues, can be found in the multi-species co-expression network (Table 1). We also found that connected components formed by WGCNA edges detected in more than 2 tissues carry clearly defined biological functions, such as ribosomal protein synthesis, ATP synthesis and structural maintenance of chromosomes (Fig. 2a-d). Overall, these results show that co-expression models estimated by WGCNA are consistent with biological knowledge.

We first validated that the associations estimated by Eq. (2) using summary-level data is consistent with those estimated using individual-level data. To this end, we simulated gene expressions and the eigen-gene activity from individual genotype data using PrediXcan [38]. 2504 samples were simulated using the genotype data collected in the 1000 Genomes Project [32]. We then randomly selected an eigen-gene component and used its activity, injected with different level of random noises, as a trait to conduct a Genome-wide association study. The summary statistics of these SNP correlations was used as input to compute component associations using Eqs. (1) and (2) and benchmark against the associations estimated directly using simulated eigen-gene activity. The results showed a linear correspondence between associations estimated by individual-level and summary data (Fig. 3). We confirmed by Kolmogorov–Smirnov test

Table 1 Gene pairs co-expressed in multiple tissues and genetically conserved gene pairs. P-values report the difference in ratios compared to the one estimated from gene pairs found in one tissue (bottom row)

# tissues detected	# WGCNA edge (n_1)	# genetically conserved edges (n_2)	ratio (n_2 / n_1)	*p*-value
14	1	0	0	0.548
13	1	1	1	8.78×10^{-23}
12	7	2	0.286	3.58×10^{-13}
11	9	5	0.556	1.68×10^{-58}
10	12	4	0.333	1.75×10^{-28}
9	19	8	0.420	2.06×10^{-69}
8	28	15	0.536	6.96×10^{-163}
7	33	14	0.424	6.65×10^{-120}
6	43	14	0.326	3.58×10^{-91}
5	75	20	0.267	4.62×10^{-104}
4	180	45	0.25	1.18×10^{-209}
3	495	87	0.176	2.21×10^{-261}
2	3344	218	0.0652	1.96×10^{-170}
1	78,656	817	0.0104	N/A

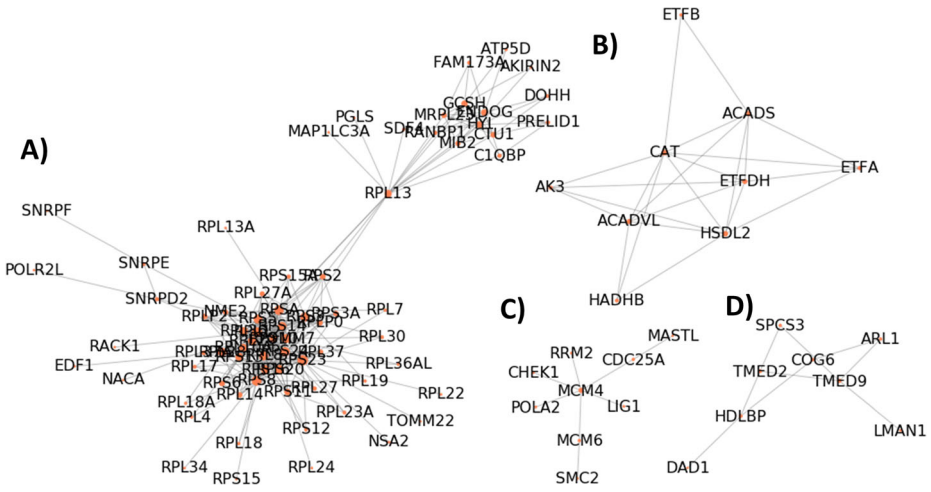

Fig. 2 Connected components formed by gene-pairs co-expressed in more than 2 tissues have well-defined biological functions, such as **a** ribosome complex, **b** electron transfer chain, **c** mini-chromosome maintenance and **d** Golgi vesicle transport

that the z-scores from non-causal components followed a normal distribution. These results show that the proposed method conforms the statistical theory of S-PrediXcan.

Next, we investigated whether the associated components capture biological information. We applied our method to GWAS summary statistics of six traits of blood cell counts [39], including neutrophils, eosinophils, basophils, lymphocytes, monocytes and red blood cells, to obtain GWAS components of these six traits in the whole blood tissue. The whole blood tissue was selected since all 6 GWAS traits are measured from blood samples. We then performed cell type enrichment analysis [40] to determine the relevant cell types using genes from GWAS components. Figure 4 shows the p-values of the enriched cell types. Genes of GWAS components associated with lymphocyte counts, basophil counts, and neutrophil counts are enriched in B cells, Myeloid cells, and the neutrophil-like HL-60 cell line respectively. These results confirmed that our method could capture gene sets enriched in expected cell types, though we did not observe a perfect one-to-one correspondence.

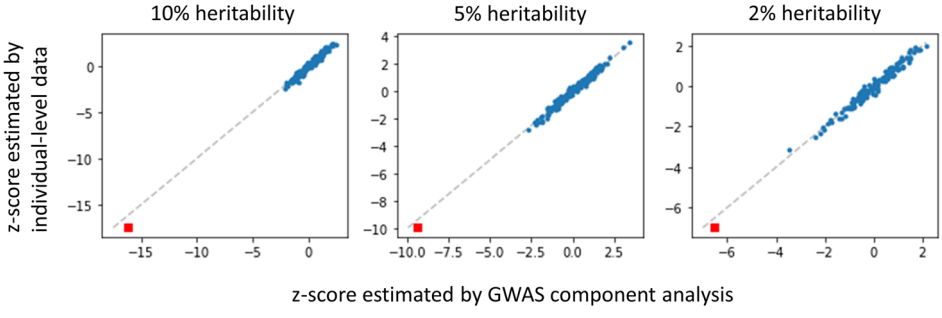

Fig. 3 Component associations estimated using GWAS summary (x-axis) are consistent with estimated using individual-level data (y-axis). Data were simulated as a trait that can be attributed to a randomly chosen eigen-gene component with 10% (left), 5% (middle) and 2% (right) heritability. Red squares indicate the z-score of the causal component. We confirmed by Kolmogorov–Smirnov test that the z-scores from non-causal components (blue points) followed a normal distribution ($p = 0.1691$, 0.5393 and 0.2542 for 10, 5 and 2% (right) heritability, respectively)

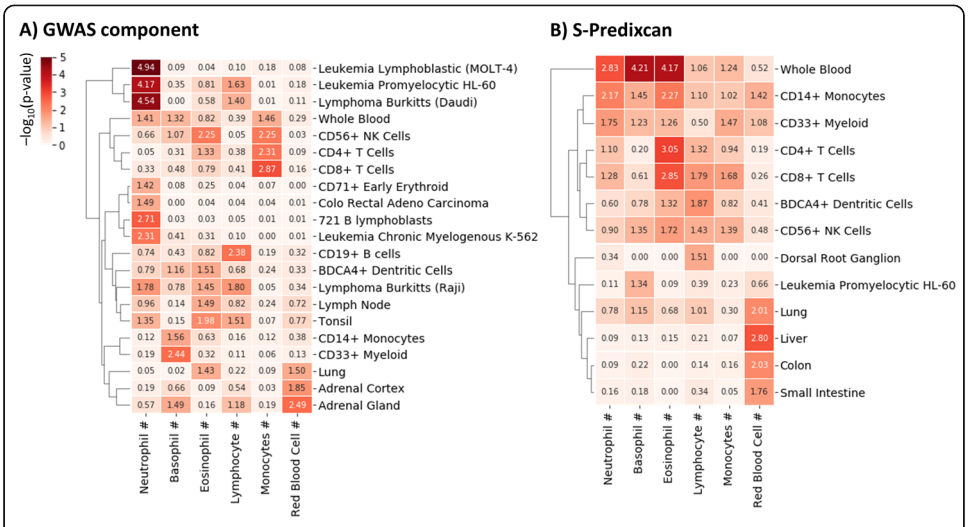

Fig. 4 GWAS components select cell-specific signatures. Cell type enrichment were used to determine cell types from genes selected by GWAS component analysis (**a**) and S-Predixcan (**b**). Colors indicate $-\log_{10}$(p-value). Results were generated by CTen web server

Last, we investigated the potential of our approach to discover putative therapeutic gene targets. To this end, it would be valuable to discover targets that might specifically impact the gene component. We projected the associated components onto Bayesian networks (BNs) constructed from GTEx [31] data. We ranked BN genes by the odds ratio of overlap between a node's downstream genes and genes in the GWAS component. Significance was determined by testing whether the odds ratio is statistically greater than 1. We applied this approach to four disease phenotypes and discovered 147, 47, 103 and 158 putative gene targets for age-related macular degeneration (AMD), Crohn's disease (CD), ulcerative colitis (UC) and rheumatoid arthritis (RA) respectively. The full list of significant gene targets is provided in Supplementary Tables 2, 3, 4, 5.

With the "omnigenic" point of view [14], we wonder if "core genes" on the gene regulatory network are better therapeutic targets than "peripheral genes" that directly carry genetic variations. Core genes are defined as functional genes that give rise to phenotypes but are not necessarily carrying genetic variants. Our approach attempts to capture this subset of genes while we used S-PrediXcan's results to represent "peripheral genes" related directly to genetic variations. MGI hits and LINCS hits are used to measure the possibility of a gene being a therapeutic target. In Table 2, we reported the ratio of functional genes among all gene candidates available in each database. Overall, we demonstrated the proposed methods could select more functional genes than S-PrediXcan that selects genes directly influenced by SNP-level associations. Especially, we observed a higher ratio of MGI hits, but a comparable rate in LINCS hits, using GWAS components. Among these gene candidates, 5 of them were targets of known medications (Table 3) listed on DrugBank. TNF, selected by S-PrediXcan from both UC and AMD GWAS summary, is a target of infliximab, Chloroquine and Etanercept. ALOX5, selected by our approach as a gene candidate for UC, is a target of Mesalazine. SLOC1A2, FCGR3A and C1QA, selected by our approach as gene candidates for RA, are also targets of

Table 2 In silico validation of gene candidates for four disease phenotypes

	AMD	CD	UC	RA
GWAS components + BNs (p_1)				
LINCS hits	0.059 (3/51)	0.143 (2/14)	0.146 (6/41)	0.326 (17/52)
MGI hits	0.311 (42/135)	0.326 (15/46)	0.361 (35/97)	0.493 (73/148)
S-Predixcan (p_2)				
LINCS hits	0.093 (5/54)	0.243 (18/74)	0.121 (7/58)	0.237 (22/93)
MGI hits	0.227 (29/128)	0.367 (50/136)	0.269 (28/104)	0.277 (57/206)
$p_1 - p_2$				
LINCS hits	−0.034	−0.1	0.025	0.089
MGI hits	0.084	− 0.047	0.092	0.216***

Numerators represent hits, and denominators represent the number of genes retrieved by GWAS components + BNs or S-PrediXcan. *** indicates $p < 0.001$

medications for RA such as Etanercept, Hydrocortisone and Ibuprofen. These results further support the idea that our approach can improve selection of functional gene candidates from a GWAS summary.

Discussion

We investigated a few genes captured by our approach and are both MGI and LINCS hits. HCK was found to be a gene candidate driving a GWAS component associated with UC in artery aorta. HCK was previously found to be genetically associated with inflammatory bowel disease and predicted as a causal factor that regulates NOD2, IL10 and ALOX5 [41]. Consistent with this, the BN suggests that HCK regulates ALOX5 (Fig. 5a), whose absence has a protective role in an experimental mouse model of colitis [42].

C1QA was found to be a gene candidate from an AMD-associated component in adrenal gland. C1q and the classical complement pathway has been suspected to play a role in the disease progression induced in retinal degeneration, potentially through local expression of C1q from subretinal microglia/macrophages that instigates inflammasome activation and inflammation [43]. Inspection of the neighborhood on the BN suggests that C1QA regulates MS4A4A (Fig. 5b), a membrane-spanning protein that is expressed on macrophage-lineage cells [44, 45].

We also identified PDGFRA as a gene candidate from a RA-associated component in stomach. PDGFR has been found to be upregulated in RA synoviocytes and synovial tissues and may play a role in synoviocyte-driven extracellular matrix degradation in RA [46]. PDGFR signaling has been shown to be one of potential

Table 3 Gene candidates with known indications. Results are queried from DrugBank

Target	Indication	Drug
GWAS components + BNs		
ALOX5	UC	Mesalazine
SLOC1A2	RA	Hydrocortisone, Ibuprofen, Indomethacin
FCGR3A	RA	Etanercept
C1QA	RA	Etanercept
S-PrediXcan		
TNF	UC	Infliximab
TNF	RA	Chloroquine, Etanercept, Infliximab

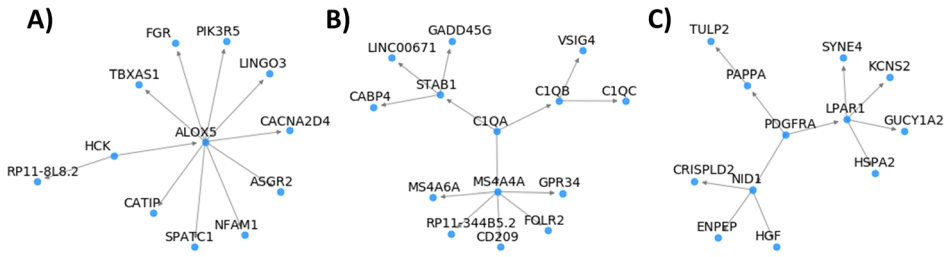

Fig. 5 Downstream genes of selected gene candidates on the Bayesian networks: **a** HCK in artery aorta, **b** C1QA in adrenal gland, and **c** PDGFRA in adrenal gland. For simplicity, genes more than three steps away from a gene candidate were excluded

mechanisms of immatinib mesylate, a tyrosine kinase inhibitor that reduces activation of RA synoviocytes [47]. Inspection of the neighborhood on the BN (Fig. 5c) suggests that PDGFRA regulate LPAR1 which may contribute to development of arthritis via cellular infiltration [48].

GWAS component analysis provides a complementary viewpoint to genetic mapping. Instead of locating risk variants, this approach looks for transcriptomic modulation that is influenced by genetic variants. This added dimension allows interpretation of GWAS results on pathways more relevant to phenotypes. In contrast to previously developed techniques, our method detects novel disease-associated components rather than enriched pathways from databases [49]. In this study, we applied WGCNA to single-tissue gene expression independently. As our results showed that genes co-expressed in multiple tissues usually carry well-defined functions, integrating multiple tissues may improve the construction of co-expression networks, as has been done previously [50]. However, such joint modeling often operates on shared genes across tissues, limiting its applicability when integrating with S-Predixcan models. Currently our method utilizes WGCNA to estimate co-expression modules in an unsupervised manner. Further work is required to integrate WGCNA with GWAS summary to construct a disease centric co-expression network.

The key to GWAS component analysis is its ability to utilize and stack models estimated from the reference genome and tissue-specific gene expression in a principled way. Combining models is crucial to obtaining a holistic picture of the complex biological systems underlying diseases [51, 52]. Although comprehensive measurements of every aspect of these systems would in theory offer a direct solution, such data are generally lacking. Instead, reference data focused on specific features of systems are accumulating at unprecedented speed. In this study, we combined two models in sequential order, demonstrating the feasibility of combining co-expression networks with GWAS associations. In the future, we expect to integrate additional types of functional data into this framework, and we envision that general approach of combining local models estimated from various data sources will enable comprehensive characterization of complex diseases.

Conclusions

Here we describe a hierarchical approach, GWAS component analysis, for detecting disease-associated components from GWAS summary data. GWAS component analysis

utilizes correlations of gene expression to further summarize SNP associations into associations of eigen-gene components. We evaluated GWAS component analysis on synthetic data and confirmed its consistency with respect to associations estimated using individual-level data. The application to GWAS of six blood cell counts revealed enriched cell types that coincide with current knowledge. We further demonstrated that GWAS component analysis can be used for therapeutics discovery by coupling it with Bayesian networks. Investigation of selected gene candidates suggests that our integrated framework can discover functional gene candidates from a GWAS summary.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13040-020-00216-9.

Additional file 1.
Additional file 2.
Additional file 3.
Additional file 4.
Additional file 5.
Additional file 6.
Additional file 7.

Acknowledgements
We thank Roman Kosoy and Rachel Hodos for helpful discussions of this manuscript.

Authors' contributions
JD, PA and H-CL conceived and designed the study. H-CL carried out the implementation. H-CL, OI, BG, AD and CB analyzed and interpreted the results. JD, PA, OI, BG, CB. and H-CL wrote the manuscript. The author(s) read and approved the final manuscript.

Funding
This work was supported by a Postdoctoral Fellowship from GlaxoSmithKline.

Availability of data and materials
Python scripts for reproducing results are available on https://github.com/howchihlee/gwas_component_analysis.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
P.A. is an employee of BioInfi. O.I. is an employee of Sumitomo Dainippon Pharma Co., Ltd.

Author details
[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [2]Institute for Next Generation Healthcare, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. [3]Drug Research Division, Sumitomo Dainippon Pharma. Co. Ltd., 3-1-98 Kasugade-naka, Konohana-ku, Osaka 554-0022, Japan. [4]Hasso Plattner Institute of Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY 10032, USA. [5]BioInfi, 1150 First Avenue, Ste 501, King of Prussia, PA 19406, USA.

Received: 18 March 2020 Accepted: 29 May 2020
Published online: 17 June 2020

References
1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science (80-). 2001;291(5507):1304–51.
2. Consortium†The International HapMap. The International HapMap Project. Nature. 2003;426:789.
3. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, et al. Complement factor H variant increases the risk of age-related macular degeneration. Science (80-). 2005;308(5720):419–21.
4. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science (80-). 2006;314(5804):1461–3.
5. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, et al. Genome-wide association study identifies five novel susceptibility loci for Crohn's disease and implicates a role for autophagy in disease pathogenesis. Nat Genet. 2007;39(5):596.

6. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47(9):979–86.
7. Demenais F, Kanninen T, Lindgren CM, Wiltshire S, Gaget S, Dandrieux C, et al. A meta-analysis of four European genome screens (GIFT consortium) shows evidence for a novel region on chromosome 17p11. 2–q22 linked to type 2 diabetes. Hum Mol Genet. 2003;12(15):1865–73.
8. Guan W, Pluzhnikov A, Cox NJ, Boehnke M. Meta-analysis of 23 type 2 diabetes linkage studies from the international type 2 diabetes linkage analysis consortium. Hum Hered. 2008;66(1):35–49.
9. Van Limbergen J, Wilson DC, Satsangi J. The genetics of Crohn's disease. Annu Rev Genomics Hum Genet. 2009;10:89–116.
10. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat Rev Genet. 2018;19(2):110.
11. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and Epigenetic Fine-Mapping of Causal Autoimmune Disease Variants. Nature. 2015;518(7539):337.
12. Loh P-R, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet. 2015;47(12):1385.
13. Shi H, Kichaev G, Pasaniuc B. Contrasting the genetic architecture of 30 complex traits from summary association data. Am J Hum Genet. 2016;99(1):139–53.
14. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169(7):1177–86.
15. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24(1):14–24.
16. Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. Gene co-expression network connectivity is an important determinant of selective constraint. PLoS Genet. 2017;13(4):e1006402.
17. Josephs EB, Wright SI, Stinchcombe JR, Schoen DJ. The relationship between selection, network connectivity, and regulatory variation within a population of Capsella grandiflora. Genome Biol Evol. 2017;9(4):1099–109.
18. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9(1):1–20.
19. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45(6):580–5.
20. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1):559.
21. Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, et al. An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenet Genome Res. 2004;105(2–4):363–74.
22. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48(3):245–52.
23. Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A versatile gene-based test for genome-wide association studies. Am J Hum Genet. 2010;87(1):139–45.
24. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS Comput Biol. 2016;12(1):e1004714.
25. Zhu X, Stephens M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nat Commun. 2018;9(1):1–14.
26. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383.
27. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48(5):481–7.
28. Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. Nat Methods. 2016;13(4):310.
29. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet. 2008;40(7):854–61.
30. Divaraniya AA. Mapping the Shared Molecular Architecture of Complex Inflammatory Diseases. Mount Sinai: Icahn School of Medicine; 2017.
31. Cohain A, Divaraniya AA, Zhu K, Scarpa JR, Kasarskis A, Zhu J, et al. Exploring the reproducibility of probabilistic causal molecular network models. In: Pacific Symposium on Biocomputing. World Scientific; 2017. p. 120–31.
32. ConsortiumT 1000 GP. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
33. Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT. MGD: the mouse genome database. Nucleic Acids Res. 2003;31(1):193–5.
34. Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, et al. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. NPJ Syst Biol Appl. 2016;2:16015.
35. Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, Rouillard AD, et al. Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. Nat Commun. 2016;7:12846.
36. Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, et al. The characteristic direction: a geometrical approach to identify differentially expressed genes. BMC Bioinformatics. 2014;15(1):79.
37. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science (80-). 2003;302(5643):249–55.
38. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091–8.
39. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell. 2016;167(5):1415–29.
40. Shoemaker JE, Lopes TJS, Ghosh S, Matsuoka Y, Kawaoka Y, Kitano H. CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data. BMC Genomics. 2012;13(1):460.
41. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119–24.
42. Cuzzocrea S, Rossi A, Mazzon E, Di Paola R, Genovese T, Muia C, et al. 5-Lipoxygenase modulates colitis through the regulation of adhesion molecule expression and neutrophil migration. Lab Investig. 2005;85(6):808–22.

43. Jiao H, Rutar M, Fernando N, Yednock T, Sankaranarayanan S, Aggio-Bruce R, et al. Subretinal macrophages produce classical complement activator C1q leading to the progression of focal retinal degeneration. Mol Neurodegener. 2018; 13(1):45.
44. Sanyal R, Polyak MJ, Zuccolo J, Puri M, Deng L, Roberts L, et al. MS4A4A: a novel cell surface marker for M2 macrophages and plasma cells. Immunol Cell Biol. 2017;95(7):611–9.
45. Mattiola I, Tomay F, De Pizzol M, Silva-Gomes R, Savino B, Gulic T, et al. The macrophage tetraspan MS4A4A enhances dectin-1-dependent NK cell–mediated resistance to metastasis. Nat Immunol. 2019;20(8):1012.
46. Charbonneau M, Lavoie RR, Lauzier A, Harper K, McDonald PP, Dubois CM. Platelet-derived growth factor receptor activation promotes the prodestructive invadosome-forming phenotype of synoviocytes from patients with rheumatoid arthritis. J Immunol. 2016;196(8):3264–75.
47. Sandler C, Joutsiniemi S, Lindstedt KA, Juutilainen T, Kovanen PT, Eklund KK. Imatinib mesylate inhibits platelet derived growth factor stimulated proliferation of rheumatoid synovial fibroblasts. Biochem Biophys Res Commun. 2006;347(1):31–5.
48. Miyabe Y, Miyabe C, Iwai Y, Takayasu A, Fukuda S, Yokoyama W, et al. Necessity of lysophosphatidic acid receptor 1 for development of arthritis. Arthritis Rheum. 2013;65(8):2037–47.
49. Jin L, Zuo X-Y, Su W-Y, Zhao X-L, Yuan M-Q, Han L-Z, et al. Pathway-based analysis tools for complex diseases: a review. Genomics Proteomics Bioinformatics. 2014;12(5):210–20.
50. Pierson E, Koller D, Battle A, Mostafavi S, ConsortiumGte. Sharing and specificity of co-expression networks across 35 human tissues. PLoS Comput Biol. 2015;11(5):e1004220.
51. Califano A, Butte AJ, Friend S, Ideker T, Schadt E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet. 2012;44(8):841.
52. Yugi K, Kubota H, Hatano A, Kuroda S. Trans-omics: how to reconstruct biochemical networks across multiple 'omic'layers. Trends Biotechnol. 2016;34(4):276–90.

Publisher's Note