FINDING GENOME-TRANSCRIPTOME-PHENOME ASSOCIATION WITH STRUCTURED ASSOCIATION MAPPING AND VISUALIZATION IN GENAMAP

ROSS E CURTIS[1,2], JUNMING YIN[2], PETER KINNAIRD[3], AND ERIC P XING[4]

[1]*Joint Carnegie Mellon-University of Pittsburgh PhD Program in Computational Biology,* [2]*Lane Center for Computational Biology,* [3]*Human Computer Interaction Institute,* [4] *Machine Learning Department*
Carnegie Mellon University, Pittsburgh, PA 15213, USA
Email: {rcurtis, junmingy, kinnaird, epxing}@cs.cmu.edu

Despite the success of genome-wide association studies in detecting novel disease variants, we are still far from a complete understanding of the mechanisms through which variants cause disease. Most of previous studies have considered only genome-phenome associations. However, the integration of transcriptome data may help further elucidate the mechanisms through which genetic mutations lead to disease and uncover potential pathways to target for treatment. We present a novel structured association mapping strategy for finding genome-transcriptome-phenome associations when SNP, gene-expression, and phenotype data are available for the same cohort. We do so via a two-step procedure where genome-transcriptome associations are identified by GFlasso, a sparse regression technique presented previously. Transcriptome-phenome associations are then found by a novel proposed method called gGFlasso, which leverages structure inherent in the genes and phenotypic traits. Due to the complex nature of three-way association results, visualization tools can aid in the discovery of causal SNPs and regulatory mechanisms affecting diseases. Using well-grounded visualization techniques, we have designed new visualizations that filter through large three-way association results to detect interesting SNPs and associated genes and traits. The two-step GFlasso-gGFlasso algorithmic approach and new visualizations are integrated into GenAMap, a visual analytics system for structured association mapping. Results on simulated datasets show that our approach has the potential to increase the sensitivity and specificity of association studies, compared to existing procedures that do not exploit the full structural information of the data. We report results from an analysis on a publically available mouse dataset, showing that identified SNP-gene-trait associations are compatible with known biology.

1. Introduction

Understanding the genetic basis of human diseases has the potential for improved treatment and prevention. Genome-wide association study (GWAS) is a popular approach that uncovers genetic variants associated with complex traits. Typically, GWAS uses machine learning or statistical approaches to search for genetic polymorphisms, often single-nucleotide polymorphisms (SNPs), which are associated with a disease. These studies find mutations in the genome that play a role in disease development and progression. Successful studies over the last few years have uncovered susceptibility loci for many diseases such as Crohn's disease[1], asthma[2], and heart disease[3]. GWAS is commonly regarded as a powerful tool for uncovering the genetic basis of human diseases[4].

Despite its recent success and popularity, fundamental limitations of GWAS impede its ability to explain disease variation and lead to clinical treatments[5]. For example, while many studies have identified SNPs associated with different diseases, these SNPs often explain only a fraction of the heritability of the disease, suggesting that other genetic factors remain undiscovered[4].

Additionally, a majority of identified SNPs do not affect protein sequence; these SNPs may have a yet unknown regulatory role in the cell[6]. Hence, we are far from a complete understanding of how discovered SNPs actually regulate cellular pathways leading to disease[7]. Furthermore, without adequate understanding of how SNPs operate in a biological context, it is often impossible for a study's results to lead to clinical applications[5]. Overcoming these barriers is a significant problem facing the genetics community today. A proposed solution to these barriers is the combination of additional data types to the traditional GWAS data (SNP genotyping and trait measurements)[5,8].

Recently, GWAS analyses have incorporated gene expression data[6-7,9]. Given gene expression data for a given cell type or tissue, SNPs can be mapped to the gene expression levels. SNP-gene associations can provide insight into regulatory patterns in the genome and help elucidate the mechanism by which a SNP leads to disease. Moreover, methods leveraging gene-network structure have been developed to enhance the discovery of associations from SNPs to cellular pathways[10-11]. The integration of expression data into GWAS analyses has led to the identification of new disease genes[12-14], demonstrating the potential of this approach.

To date, the general strategy has been to integrate gene expression data into a GWAS study *after* the main analysis and from individuals *unrelated* to the original study[6-7]. However, in the increasingly common scenario where expression, phenotype, and genomic data are available from the same cohort, expression data can be incorporated directly into the primary analysis. In this case, all the data guides the discovery of genome-transcriptome-phenome associations. This unified three-way association framework has the potential to reveal the functional relationships between associated genomic variations and physical phenotypes, via intermediate phenotypes. This has a direct impact on personalized medicine as different patients may have different regulatory mechanisms by which disease arises, but traditional SNP-trait association studies are insufficient to uncover the hidden mechanisms. Despite the promise of unified three-way association analysis, work done on this problem is limited[15-16].

In this paper, we present a new algorithmic and visualization strategy for three-way association analysis with GWAS and gene expression data. We use *structured association mapping* to find genome-transcriptome-phenome associations. In the framework of structured association mapping, the structure of the data guides the algorithm to find true associations and eliminate false positives. While incorporating gene expression data and its structure into GWAS can enhance discovery statistically, the additional data also increases the difficulty of interpreting the results. The results, with the trait and genetic structures, are a large, complex sea of data that must be explored, suggesting the applicability of a visualization approach[17]. Available visualization tools for general GWAS[18-19] are not built for this type of data, thus new visualization tools are needed to use the structure of genome-transcriptome-phenome association data to guide the analyst to discover the biological mechanisms driving genome-phenome associations.

In Section 2, we present a novel structured association mapping strategy that incorporates genome-wide SNP data, expression data, and phenotype data in a unified three-way association framework. We find genome-transcriptome-phenome associations through a two-step process. First, we find the association of SNPs to gene expression data by leveraging the genetic network structure among genes by applying GFlasso[10]. In the second step, we find associations from genes to phenotypes using a new proposed method called gGFlasso, which incorporates structure information from both the predictors (genes) and responses (traits).

In Section 3 we present novel visualizations we have designed to guide geneticists through the exploration of three-way association results, guided by the structure of the data. We implement these tools in GenAMap, visual analytics software for association analysis[20].

Finally, we demonstrate our approach by simulation (Section 4) and on a NIH heterogeneous stock mice dataset[21] (Section 5). Results on simulated datasets show that our strategy outperforms existing methods in terms of sensitivity and specificity. Using the mouse data, we have uncovered the associations of eight genes in the mouse H2 complex on Chr 17 with mouse immunology traits, compatible with current knowledge.

2. Statistical Model

Let $\boldsymbol{X}$ be an $N \times P$ genotype matrix for N individuals where each row represents the allele states of an individual at P loci; let $\boldsymbol{Y}$ be an $N \times J$ gene expression matrix where expression levels of J genes are measured for the same set of individuals; finally, let $\boldsymbol{Z}$ be an $N \times K$ phenotype matrix where each row records K phenotypic traits of an individual. We assume that there are two *undirected weighted* relevance graphs $G_G = (V_G, E_G)$ and $G_T = (V_T, E_T)$ available for J genes and K traits, respectively. Each node in G_G represents a gene and each node in G_T represents a trait, hence $|V_G| = J,$ $|V_T| = K$. There is a vast literature on network construction algorithms, and any strategy can be applied to create a network where a connection between nodes u and v represents a relationship between nodes. Each edge $\{u, v\} \in E_G$ or E_T is associated with a weighted connection between nodes. The three-way association study is then performed in a two-step procedure: first, we find genome-transcriptome associations by applying GFlasso to $\boldsymbol{X}$ and $\boldsymbol{Y}$; next, we search for transcriptome-phenome associations between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ using a new structured sparse regression method called graph-graph-guided fused lasso (gGFlasso).

2.1. *Finding genome-transcriptome associations using GFlasso*

To find associations between SNPs and gene expression levels, we use GFlasso[10], a structured association mapping approach that estimates the association strengths of predictors jointly for multiple correlated responses by a sparse regression strategy. In our case, SNPs are treated as predictors and gene expression levels are treated as responses. GFlasso assumes that a relatively small number of markers are associated with each gene and that highly correlated genes tend to be influenced by a common subset of SNPs. This assumption is explicitly expressed as two regularization terms in a linear regression model (Eq. 1). Input to GFlasso includes $\boldsymbol{X}$, $\boldsymbol{Y}$, and G_G; the output is a regression coefficient matrix $\boldsymbol{B_1}$, where B_{pv} denotes the strength of pth SNP associated with vth gene. $\boldsymbol{B_1}$ is estimated by solving the following optimization problem[22]:

$$\boldsymbol{B_1} = \underset{\boldsymbol{B} \in \mathbb{R}^{P \times J}}{argmin} ||\boldsymbol{Y} - \boldsymbol{XB}||_F^2 + \lambda \sum_j \sum_p |B_{pj}| + \gamma \sum_{\{u,v\} \in E_G} \sum_p |B_{pu} - sign(\rho_{uv}) B_{pv}|, \tag{1}$$

where $||.||_F$ is the Frobenius norm of the matrix. The second term is the L_1 lasso penalty[23], which has the property of shrinking the strengths of irrelevant SNPs towards zero; the last term is the *graph-guided fused lasso* penalty, which encourages highly correlated genes (connected by an edge in E_G) to be associated with the same SNPs. sign(ρ_{uv}) controls the pattern of fusion applied to the association strengths: if two genes are negatively correlated, their corresponding association strengths are encouraged to have different sign. λ and γ are the regularization parameters and can be estimated by cross-validation on a validation set. For each gene, the set of associated SNPs are those with non-zero association strengths in the estimated coefficient matrix $\boldsymbol{B}_1$.

2.2. *Finding transcriptome-phenome associations using gGFlasso*

To find associations between genes and phenotypic traits, we develop a new method by extending GFlasso, called gGFlasso, which incorporates the correlation structure of genes G_G as well as the dependency graph of traits G_T. In this setting, expression levels are treated as predictors and traits are regarded as responses. The GFlasso framework only uses the dependency structure on responses and assumes that variations of correlated responses (traits) are likely to be explained by a common set of predictors (genes). However, predictors can also be correlated, such as co-expressed genes, and it seems natural to also exploit the predictor's dependency graph. Given G_G, gGFlasso assumes that correlated (connected) genes tend to influence the same subsets of traits. This assumption is encoded as an additional fusion penalty in the linear regression model:

$$\boldsymbol{B_2} = \underset{\boldsymbol{B} \in \mathbb{R}^{J \times K}}{argmin} ||\boldsymbol{Z} - \boldsymbol{YB}||_F^2 + \lambda \sum_j \sum_k |B_{jk}| + \gamma_1 \sum_{\{u,v\} \in E_G} \sum_k |B_{uk} - sign(\rho_{uv}) B_{vk}|$$
$$+ \gamma_2 \sum_{\{m,l\} \in E_T} \sum_j |B_{jm} - sign(\rho_{ml}) B_{jl}|. \tag{2}$$

As in Eq. 1, λ, γ_1, and γ_2 are regularization parameters and B_{jk} is the association strength of jth gene with kth trait. Note that there are two graph-guided fused lasso penalty terms in Eq. 2: if two genes u and v are connected in G_G, the fusion penalty encourages their influence on each trait to be similar; if two traits m and l are connected by an edge in E_T, the association strength B_{jm} and B_{jl} for each gene j is encouraged to have the same absolute value. As we will demonstrate in our experiments, this joint framework that accounts for the information in the correlation structure of predictors (gene expressions) and responses (traits) has the potential to increase the sensitivity and specificity of association studies. Combined with the L_1 lasso penalty, the estimated coefficient matrix $\boldsymbol{B_2}$ has a large fraction of zero entries and the remaining non-zero entries tend to show a block structure (see Figure 4).

2.3. *Implementation notes*

We have automated GFlasso-gGFlasso in GenAMap. We use a coordinate descent approach and the “η-trick”[24] to optimize Eq. 2. The update equations and a C++ implementation of gGFlasso are available for download from sailing.cs.cmu.edu/genamap. GenAMap uses a linear search strategy and hold-out validation to find values for the regularization parameters λ, γ_1, and γ_2. For our simulated data with 100 SNPs, 500 genes, and 20 traits, the full GFlasso-gGFlasso runs in less than one day. Using GenAMap to run the GFlasso-gGFlasso for 2500 SNPs with associations to 2000 genes would return results in less than one week.

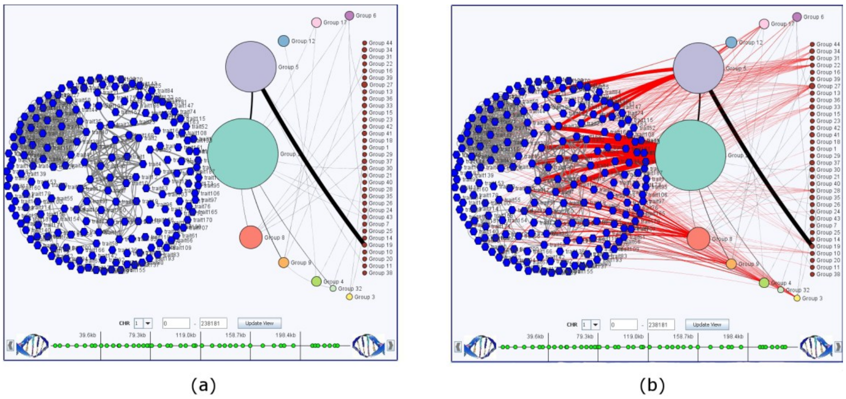

Fig. 1. Overview of three-way association visualization using the KK-layout[a] . (a) The structure (edges) of the traits (blue hexagons) and gene groups (circles) are displayed without showing the association edges. (b) The red association edges from gene groups to traits are shown. SNPs are represented at the bottom of the display as green circles in GenAMap’s genome browser.

3. Visualization Strategies in GenAMap

The GFlasso-gGFlasso analysis of a genome-transcriptome-phenome dataset leaves the genetics analyst with a large, complex sea of data to interpret. In addition to the gene-gene and trait-trait relationships, potentially hundreds of SNP-to-gene and gene-to-trait associations are identified. The analyst must explore this data to pinpoint the associations that lead to insight into disease.

The analyst explores the data with different strategies, depending on what questions he is interested in. In one scenario, the analyst comes into the study with questions about specific traits. The analyst explores the network relationships between these traits and looks for associated gene networks. He then examines the gene networks for genomic associations, leading to the discovery of SNPs that perturb the gene networks associated with the phenotypes of interest. Alternatively,

[a] Large, high resolution figures from this paper are available at http://sailing.cs.cmu.edu/genamap/threeway.html

the analyst starts with a genomic region of interest. In this case, the first considers the genomic region and its associations to genetic pathways. He then identifies the traits associated with the discovered genes. In either scenario, analysts have to filter through the dataset in an exploratory fashion. As visualization methods are particularly adept at guiding the identification of interesting data through exploratory analysis[17], visualization can be a powerful tool in either scenario.

To design a visualization system that would facilitate the exploration of three-way association results with gene and trait structure we follow Shneiderman's mantra[25]: 1) *overview first*, 2) *zoom and filter*, and 3) *details on demand*. In this section, we discuss our design of the visualizations, which are implemented in GenAMap using JAVA and the JUNG toolkit[26]. A video overview of our visualization tools is available online at http://sailing.cs.cmu.edu/genamap/threeway.html.

3.1. *Overview first: an introduction to the visualization*

When the analyst first examines three-way association results in GenAMap, he is presented with an overview of the data (Figure 1, 2). Traits are represented as blue hexagons and edges in the trait network are displayed as weighted (by strength of correlation) gray lines. Genes are grouped according to GFlasso results: genes associated with common SNPs form a group. Visually, gene groups are represented as circles where the size of the circle represents the size of the group. Edges between genes groups are shown as black lines, the thickness of the line representing the number of edges between the genes in the two groups.

GenAMap enables the analyst to explore the three-way results through two different layouts. The KK-layout is designed to present trait-trait and gene-gene structures separately (Figure 1). The positions of traits are determined by the KK-layout algorithm[27] and gene groups are plotted in a half-circle with the ten largest groups placed in an arc. Association edges are shown in red. When the analyst hovers over a gene group, a tool-tip reports the group's gene count and significant gene ontology (GO) enrichment.

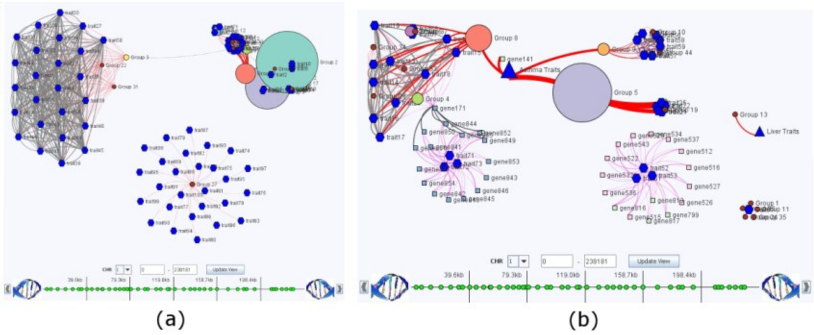

Fig. 2. GenAMap overview of three-way association results using the force-directed layout.

Analysts can also explore three-way association results using a force-directed layout (FDL). In this layout, gene and trait nodes repel each other and edges act as springs that pull nodes together.

The analyst can adjust the repulsion and attraction parameters to adjust the display. In Figure 2, we show the data from Figure 1 now represented in the FDL. In Figure 2a, the association and correlation edge spring tension is high, causing connected nodes to pull into a tightly clustered group. As the spring tension is relaxed (Figure 2b), the structure of the connected gene-trait clusters is visible. In both the KK-layout and FDL, the analyst can customize the display by turning labels on and off, adjusting parameters, or manually repositioning nodes.

GenAMap enables the analyst to simultaneously explore SNP-gene and gene-trait associations. The SNP-gene association strengths are visualized using color-encoding. For example, the analyst selects SNPs of interest and clicks to color all genes by association. The gene view updates so that the brightness of the color of each group represents its strongest association to the selected SNPs. Alternatively, the analyst selects gene groups of interest and colors the SNPs by association. SNPs are then colored: white represents a strong association and black represents no association. In Figure 3e we demonstrate both of these encodings: SNPs are colored by association to the teal gene group, and groups are colored by association to selected SNPs.

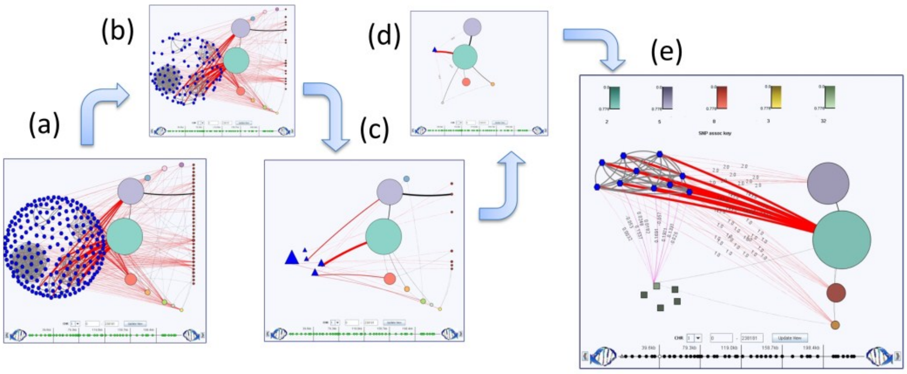

Fig. 3. Demonstration of zoom and filter tools in GenAMap. (a) The analyst starts with all data visible and filters data based on association resulting in (b). He adjusts the association edge threshold and groups traits into trait groups (c). Trait groups and gene groups not connected/associated with a trait group are removed (d). A trait group and gene group of interest are expanded and the analyst considers gene-transcriptome-phenome association simultaneously (e).

3.2. *Zoom and filter: identifying interesting signals in the data*

GenAMap is designed to enhance the analyst's ability to filter through the dataset to identify genome-transcriptome-phenome associations. We demonstrate GenAMap's filter and zoom tools through a series of steps shown in Figure 3. In Figure 3a, the data is loaded in the KK-layout. The analyst uses a filter to removes all genes and traits without associations (Figure 3b). Next, the analyst adjusts the association edge threshold to remove weak associations. Interested in the largest gene group (Group 2, teal color), he removes all groups that do not have a network edge to Group 2. Each connected component in the trait graph is then collapsed into a *trait group* (represented as a triangle) to simplify the display (Figure 3c). To explore only the trait group with the most associations to Group 2 (thickest edge), all trait groups and gene groups not connected to this trait group are manually removed (Figure 3d). Finally, the analyst expands the trait group and a strongly-associated gene group. He colors the SNPs by association to the now-visible genes (squares) in this gene group. After identifying associated SNPs, the analyst colors all genes by association to the identified SNPs (Figure 3e). In summary, GenAMap allows the analyst to filter and explore based on network connectivity, association, edge thresholds, and grouping strategies. These strategies can be employed in any three-way analysis, starting from traits, SNPs, or genes.

3.3. *Details on demand: resources for further exploration*

Once the analyst has found interesting gene-trait associations, GenAMap directly links the analyst to more information. The analyst can directly link to the UniProt database[28], Google search, or to dbSNP[29]. GO information for genes is available through GenAMap's integration with BiNGO[30].

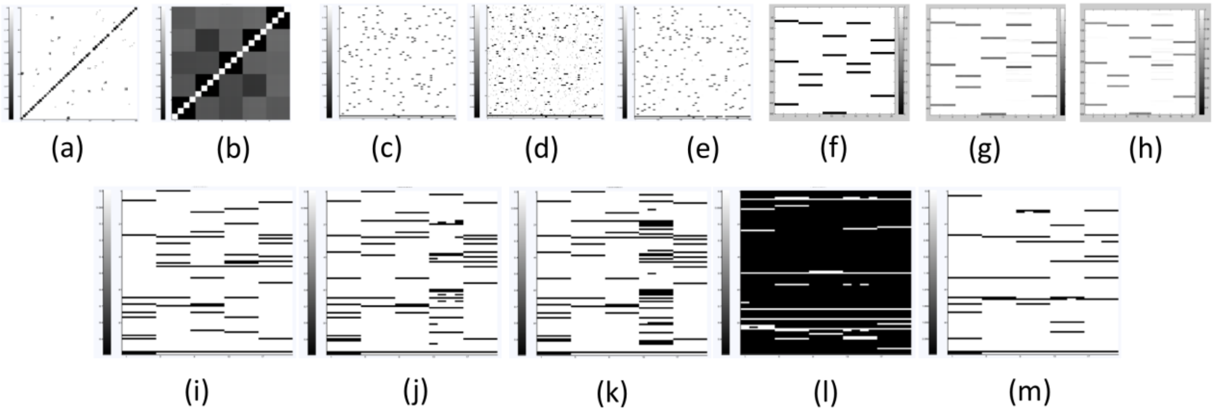

Fig. 4. Results of a simulation study on a dataset simulated with $\sigma_1^2 = \sigma_2^2 = 1$. Dark pixels indicate large values. (a) Correlation matrix of genes thresholded at 0.5 (b) Correlation matrix of traits thresholded at 0.5 (c) True $\boldsymbol{B_1}$ matrix (d) Recovered $\boldsymbol{B_1}$ matrix by lasso (e) Recovered $\boldsymbol{B_1}$ matrix by GFlasso. (f) True $\boldsymbol{B_2}$ matrix (g) Recovered $\boldsymbol{B_2}$ matrix by gGFlasso. (h) Recovered $\boldsymbol{B_2}$ matrix by GFlasso (i) True SNP-trait association matrix $\boldsymbol{B_3}$ (j) Recovered $\boldsymbol{B_3}$ by GFlasso-gGFlasso strategy. (k) Recovered $\boldsymbol{B_3}$ by GFlasso-GFlasso strategy (l) Recovered $\boldsymbol{B_3}$ by lasso (m) Recovered $\boldsymbol{B_3}$ by plink, values plotted are –log10(*p*-value).

4. Simulated Data Analysis

We performed a simulation experiment to validate our GFlasso-gGFlasso strategy for three-way association analysis, comparing the results with a GFlasso-GFlasso strategy, a lasso-lasso strategy and PLINK[31], a popular pairwise GWAS tool. We randomly selected $N = 250$ individuals and $P = 100$ SNPs from the mice dataset[21] to create a genotype matrix $\boldsymbol{X}$. The number of genes J was set to 500; we assumed 50 correlated gene groups of equal size. The true $P \times J$ association matrix $\boldsymbol{B_1}$

was created as follows: for each group, we randomly selected three causal SNPs; one (random) housekeeping SNP was assumed to be associated with all genes. Strengths of association for selected SNP-gene pairs were set to 1 (Figure 4c). Given $\boldsymbol{X}$ and the association matrix $\boldsymbol{B_1}$, the gene expression matrix $\boldsymbol{Y}$ was simulated as $\boldsymbol{Y} = \boldsymbol{XB_1} + \boldsymbol{\varepsilon_1}$, where $\boldsymbol{\varepsilon_1}$ is an $N \times J$ matrix with entries independently generated from $N(0, \sigma_1^2)$. The gene correlation matrix, simulated with noise level $\sigma_1^2 = 1$ and thresholded at 0.5, reveals the inherent structure of the genes (Figure 4a). We set K = 20 traits. To generate the trait matrix $\boldsymbol{Z}$, we first created a $J \times K$ association matrix $\boldsymbol{B_2}$ similar to the procedure for obtaining $\boldsymbol{B_1}$: we assumed five groups of correlated traits of equal size each with three causal gene groups (Figure 4f). Then, $\boldsymbol{Z} = \boldsymbol{YB_2} + \boldsymbol{\varepsilon_2}$, where entries of $\boldsymbol{\varepsilon_2}$ were independently generated from $N(0, \sigma_2^2)$. We also calculated the true SNP-trait association matrix, $\boldsymbol{B_3}$ (Figure 4i), by assuming that if a SNP is associated with a gene and that gene influences a trait, then there is an association between the SNP and the trait.

Table 1. True positive rates (TPR) and false positive rates (FPR) of different methods in recovering the association matrices at different noise levels.

	$\sigma_1^2 = \sigma_2^2 = 1/4$		$\sigma_1^2 = \sigma_2^2 = 1$		$\sigma_1^2 = 4, \sigma_2^2 = 16$	
	TPR	FPR	TPR	FPR	TPR	FPR
$\boldsymbol{B}_1$ by GFLasso	0.9454	0.0060	0.8965	0.0057	0.7884	0.0092
$\boldsymbol{B}_1$ by Lasso	0.9758	0.7632	0.9535	0.0763	0.9081	0.7528
$\boldsymbol{B}_2$ by gGFLasso	1.0000	0.0000	0.9333	0.0016	0.7067	0.0205
$\boldsymbol{B}_2$ by GFLasso	0.9800	0.0004	0.9233	0.0039	0.7000	0.0207
$\boldsymbol{B}_3$ by GFlasso-gGFlasso	0.9200	0.0266	0.8600	0.0305	0.8400	0.2450
$\boldsymbol{B}_3$ by GFlasso-GFlasso	0.9200	0.0266	0.8600	0.0615	0.8400	0.2500
$\boldsymbol{B}_3$ by Lasso-Lasso	1.0000	0.8803	1.0000	0.9030	1.0000	0.8559
$\boldsymbol{B}_3$ by PLINK	0.6300	0.0150	0.5357	0.0234	0.5000	0.0294

In Table 1, we present a summary of performance results evaluated by the true positive rate (TPR) and false positive rate (FPR). As the table and Figure 4e show, in a low noise setting, GFlasso is able to recover SNP-gene pairs with true association with a tiny FPR. We compared the performance of GFlasso and gGFLasso in recovering $\boldsymbol{B_2}$. Although both methods successfully recover true gene-trait pairs, GFlasso tends to produce a number of false positive pairs with weak associations (Figure 4h) and hence has a higher FPR. Both GFlasso-GFlasso and GFlasso-gGFlasso have a much lower FPR than the lasso-lasso strategy in terms of recovering $\boldsymbol{B_3}$ (Figure 4l). We were interested in whether the methods that do not incorporate the gene expression data would similarly recover $\boldsymbol{B_3}$, and so we ran the default association algorithm in PLINK to find SNP-trait associations (Figure 4m). We found that although there is a low FPR, the power of detecting true associations decreases significantly (Figure 4m).

5. Mouse Data Analysis

We ran GFlasso-gGFlasso on the NIH heterogeneous stock mice dataset[21] using GenAMap to demonstrate that: 1) our method scales to real-world problems, 2) the visualization tools can filter and zoom to find association signals in real data, and 3) the method finds genome-transcriptome-phenome associations compatible with established biological knowledge.

The NIH heterogeneous stock mice dataset has data for 259 mice with 12545 SNPs, gene expression data (from the liver) for 46147 genes, and 178 phenotypic traits. We excluded genes that did not have signals in at least 50 mice, leaving 9742 genes, and normalized the expression values using lumi[32]. All traits with more than 30% of the mice having missing values were excluded, leaving 161 traits. Missing values were imputed using *k*-nearest neighbor imputation[33]. The network was created using a soft-threshold method for finding scale-free networks[34].

We compared the results from the GFlasso-gGFlasso strategy to results found from PLINK on the SNP-trait data, where all SNP-trait pairs with a *p*-value < 1*e*-4 were considered significant associations. We found that 1385 SNPs had associations to 131 traits in the PLINK results. In the two-step results, we found that 943 SNPs were associated with 746 genes and that 412 genes were associated with 133 traits. We found 604 SNPs that were associated to a trait in the PLINK analysis and were associated to a gene using GFlasso-gGFLasso. Of the 604 SNPs with an association to a trait using PLINK and to a gene using GFlasso, we found that 546 also had at least one associated gene that was associated with the same trait in the gGFlasso results. These findings suggest that the GFlasso-gGFlasso has found SNP-gene-trait associations that may help to explain associations found in the data by the PLINK analysis, in addition to the new signals that were not found by PLINK.

We wanted to explore this large results dataset to find biologically interesting results. We explored the association results using GenAMap's three-way visualization tools (see supplementary website for video). In our analysis, we found a group of 15 genes that was enriched for the GO process *antigen processing* (*p*-value = 4.54e-6). Using GenAMap, we identified eight genes in the group (*H2-T23*, *H2-Q1*, *H2-Ab1*, *Bat5*, *Tmem63b*, *C4a*, *Psmb9*, *NM_010396*) that were associated with a sub-network of seven immunology traits (%CD8+/CD3+, CD4+/CD8+, %CD4+/CD3+, %CD8+, %B220+, %CD4+, %CD3+). We found that these genes were associated to SNPs on Chr 17 (Figure 5), with the strongest signals to *rs8242408*.

rs8242408 is located on Chr 17 in the intron region of *Tap1* and 6kb upstream from *Psmb9*. All of the identified SNPs are part of the mouse H2, the major histocompatibility complex (MHC). The eight genes we identified are also located in the H2 region; three (*H2-T23, H2-Q1, H2-Ab1*) are known H2 genes and *Psmb9* is known to be involved in antigen processing to create class I binding peptides (http://www.phosphosite.org). The H2 region is the mouse ortholog to the human HLA region and encodes genes involved in the mouse immune response[35]. Two classes of H2 genes are presented to immune cells for identification. Some H2 genes (class I) are expressed in virtually all cells and display "self" antigens, while others (class II) are expressed only in antigen-presenting cells[36]. CD8+, CD4+, etc., refer to proteins on the surface of immune response cells that bind to the antigens on the surface of the other cells in the organism.

Despite the known association of H2 genes and immunology traits, the precise mechanism of how the identified SNPs influence the expression of these associated H2 genes is yet unknown. It is likely that the discovered SNPs act as *cis*-regulatory elements of these genes, which in turn have effects on the immunology traits. Although this investigation merits further study, the discovered SNP-gene-trait associations are compatible with current biological knowledge and could lead to further insight into how SNPs in the H2 region affect the immune system. Of note is the strong association to *rs8242408* in the promoter region of *Psmb9*, which suggests expression levels of *Psmb9* could have a regulatory effect on other identified genes.

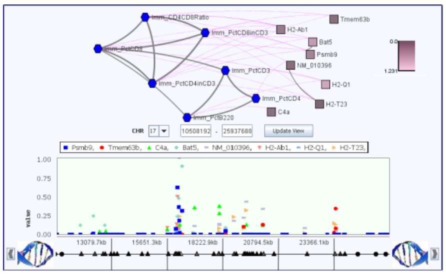

Fig. 5. Results of mouse analysis. In our mouse three-way association analysis we found eight genes associated in *cis* to the mouse H2 region and are associated with immune response phenotypes.

6. Discussion

High-throughput sequencing technology has advanced rapidly over the past few years and it will soon become routine to obtain personal genome information. Personal genomic data will possibly allow us to provide personalized treatment for each individual. While a number of exciting findings have emerged from genome-wide association studies, many results haven't translated to the clinic for improved patient care. In the post-GWAS era, it has been suggested that the GWAS data could be more insightful when integrated with other data types. In this paper we considered the problem of combining GWAS and gene expression data to find genome-transcriptome-phenome associations from an algorithmic and a visualization perspective. We have proposed a novel two-step strategy that employs structured association mapping to find associations from genome to phenome by leveraging gene expression data and its network structure. Using simulated and mouse datasets, we demonstrated that integrating this intermediate data type into a unified framework has the potential to increase the accuracy of studies and offer more insights on the functional relationships between identified GWAS loci and associated phenotypic traits. Although it is challenging to have a full understanding of such relationships and translate them into medical practice, we believe we have taken an important step forward.

We are currently undertaking further steps to demonstrate our approach with biological analysis using the NIH heterogeneous stock mice. Additionally, more sophisticated optimization algorithms could be developed to increase the computational efficiency of gGFlasso.

Acknowledgments

This work is supported by the Defense Advanced Research Projects Agency [Z931302]; National Science Foundation [DBI-0640543, IIS-0713379]; National Institutes of Health [1R01GM087694, 1R01GM093156, 0015699]; an Alfred P. Sloan Research Fellowship to EPX; and a Ray and Stephenie Lane Research Fellowship to JY.

References

1. G. Lettre and J. D. Rioux, *Hum. Mol. Genet.* **17**(R2), R116-R121 (2008).
2. D. S. Postma and G. H. Koppelman, *Proceedings of the American Thoracic Society* **6**, 283-287 (2009).
3. R. McPherson et al., *Science* **316**(5830), 1488-91 (2007).
4. R. A. Manolio et al., *Nature* **461**, 747-753 (2009).
5. E. E. Schadt, *Nature* **461**, 218-223 (2009).
6. E. E. Schadt et al., *PLoS Biol* **6**(5), e107 (2008).
7. M. I. McCarthy and J. N. Hirschorn, *Hum. Mol. Genet.* **17**(R2), R156-R165 (2008).
8. A. Califano, A. Butte, S. Friend, T. Ideker, and E. E. Schadt, *Nature Precedings* **5732**(1), doi:10.1038 (2011).
9. Y. Gilad, S. A. Rifkin, and J. K. Pritchard, *Treds Genet* **24**(8), 408-145 (2008).
10. S. Kim and E. P. Xing, *PLoS Genet* **5**(8), e1000587 (2009).
11. J. Zhu et al., *Nature Genet* **40**(7), 854-861 (2008).
12. W. Cookson, L. Liang, G. Abecasis, M. Moffatt, and M. Lanthrop, *Nature Reviews Genetics* **10**, 184-194 (2009).
13. A. C. Silveira et al., *Vision Research* **50**(7), 698-715 (2010).
14. Y. Hsu et al., *PLoS Genetics* **6**(6), e1000977 (2010).
15. Y. Chen et al., *Nature* **452**, 429-435 (2008).
16. V. Emilsson et al., *Nature* **452**(27), 423-430 (2008).
17. J. D. Fekete, J. J. vanWijk, J. T. Stasko, and C. North, *LNCS* **4950**, 1-18 (2008).
18. R. J. Pruim et al., *Bioinformatics* **26**(18), 2336-2337 (2010).
19. D. Ge et al., *Genome Res* **18**(4), 640-3 (2008).
20. R. E. Curtis and E. P. Xing, in *Proc. ISMB*, Technology Track (2010).
21. M. Johannesson et al., *Genome Res* **19**(1), 150-8 (2009).
22. X. Chen, S. Kim, Q. Lin, J. G. Carbonell, and E. P. Xing, *CoRR* (2010).
23. R. Tibshirani, *Royal Statist Soc B* **58**(1), 267-288 (1996).
24. A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, *JMLR* **9**, 2491-2521 (2008).
25. B. Shneiderman, *IEEE Software* **11**(6), 70-77 (1994).
26. J.O. Madahain., D. Fisher. P. Smyth. S. White. and Y.B. Boey, *J Stat Softw* **VV**(II) (2005).
27. T. Kamada and S. Kawai, *Information Processing Letters* **31**(1), 7-15 (1989).
28. The UniProt Consortium, *Nucleic Acids Res.* **39**, D214-D219 (2011).
29. S. T. Sherry et al., *Nucleic Acids Res.* **29**(1), 308-11 (2001).
30. S. Maere, K. Heymans, and M. Kuiper, *Bioinformatics* **21**, 3448-3449 (2005).
31. S. Purcell, et al., *American Journal of Human Genetics*, 81 (2007).
32. P. Du, W.A. Kibbe, S.M. Lin. *Bioinformatics* **24**(13),1547-8 (2008).
33. O. Troyanskaya, et al., *Bioinformatics* **17**(6):520-525.
34. B. Zhang and S. Horvath. *Stat Appl Genet Molec Bio* 4(1): Article 17 (2005).
35. P.M. Stuart, *eLS* doi:10.1002/9780470015902.a0000921.pub3 (2010).
36. A. Kumanovics, T. Takada, K. F. Lindahl. *Immunology* 21:629-657 (2003).