GenAMap: Visual Analytics Software for Structured Association Mapping

Ross E Curtis[1,2], Sally Wenzel[3], Peter Kinnaird[4], Eugene Bleecker[5], Deborah A Meyers[5], Eric P Xing[6§]

[1] Joint Carnegie Mellon – University of Pittsburgh PhD Program in Computational Biology, Pittsburgh, PA

[2] Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA

[3] Pulmonary Allergy Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA

[4] Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA

[5] Center for Human Genomics, Wake Forest University School of Medicine, Winston-Salem, NC

[6] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA

[§]Corresponding author

Email addresses:

REC: rcurtis@cs.cmu.edu
SW: wenzelse@upmc.edu
PK: kinnaird@cs.cmu.edu
EB: ebleeck@wfubmc.edu
DAM: dmeyers@wfubmc.edu
EPX: epxing@cs.cmu.edu

Abstract

The incorporation of additional data into GWAS, such as eQTL and gene expression data, helps to elucidate the underlying biological mechanisms behind SNP-trait associations. The addition of these new data types also adds another dimension to the study, as the additional data has inherent structure that can be leveraged in a systematic way to enhance the discovery of signals in the data and to eliminate false positives. We present a new paradigm for GWAS analysis, termed *structured association mapping*, which leverages data structures *algorithmically* and *visually* in a visual analytics software platform, GenAMap. GenAMap is available from http://sailing.cs.cmu.edu/genamap.

Key words: genetic association mapping, genome-wide association study, eQTL studies, visual analytics, automation

Genome-Wide Association Studies

As genome sequencing technology has improved over the last decade, it has led to an increased understanding of how mutations in the genome lead to complex phenotypes and heritable human diseases. A popular strategy to determine how sequence variation affects the inheritance of complex traits is the genome-wide association study (GWAS) [1,2]. GWAS has led to the successful identification of many so-called *disease genes* and *susceptibility loci* for a variety of diseases such as cancer [3], diabetes [4], and Alzheimer's disease [5]. In a traditional GWAS, individuals are sequenced for genetic polymorphisms across the genome. The individuals are divided into case and control groups, and machine learning or statistical techniques identify mutations that are associated with the disease of interest.

The motivation for GWAS comes from the central dogma of biology, which is that certain parts of the DNA (genes) code for mRNA, which is translated into the proteins that run the cell and the organism. Thus, mutations in the genome at the DNA level can directly affect the entire organism by altering the creation or function of proteins in the cell. Although most of the human genetic sequence is identical across individuals, there are places in the genome where the sequence has been mutated, causing a genetic polymorphism between individuals. If we think of DNA as a string made up of four nucleotides (characters), then a genetic polymorphism is a difference in the sequence between two individuals. The most common type of genetic polymorphism is a single-nucleotide polymorphism (SNP), an instance where one nucleotide is different between individuals. For example, some individuals may inherit a G at a particular location instead of the A that is common in the population. Although many SNPs make little or no difference to gene expression levels and the normal function of a cell, some SNPs can have a much larger effect. SNPs that turn off important genes, or change the coding sequence of regulator genes, can interact with other expressed genes and lead to a diseased phenotype or greater susceptibility to disease.

The goal of GWAS, then, is to identify genetic polymorphisms associated with disease, which ideally lead to greater insight about disease prevention, acquisition, and progression. However, the success of many GWAS in explaining SNP-phenotype associations and leading to clinical treatments has been limited [6], in part, because traditional GWAS only considers the association from the SNPs to the phenotype and ignores the underlying biological system. Thus for many studied diseases, discovered SNPs only explain a fraction of the disease heritability [1] or identify SNPs that do not affect protein sequence and thus have an unknown regulatory role in the cell [7].

Recent developments in genetics use a variety of strategies to improve discovery in GWAS. For example, the result that SNPs associated with complex traits are likely to be expression quantitative trait loci (eQTLs) [8] has led some studies to use eQTL information to select SNPs for GWAS or to perform GWAS using gene expression data instead of SNPs. Other strategies have added data into GWAS analysis from other "-omes," creating a meta-dimensional analysis with multiple data types. By incorporating established knowledge about the biological system, discovery in association studies is enhanced. In fact, as another strategy, gene expression data is now commonly used to integrate transcriptome information into association studies [7,9,10]. Successful integration of eQTL analysis into GWAS has led to the identification of new disease genes in humans and mice [11,12,13,14].

The subject of this paper concerns the growing complexity of GWAS when gene expression or other data is added to the analysis. While the additional data can be overwhelming, it can also be leveraged *algorithmically* and *visually* to improve and enhance discovery in GWAS. In this paper, we will propose how this can be done, demonstrating this new paradigm of thinking about GWAS and eQTL studies through a visual analytics software platform called GenAMap.

Algorithmically, traditional, simple methods that look for pairwise associations between SNPs and genes or multiple phenotypic traits do not take advantage of all the information in today's GWAS data. Structure inherent in the genome, transcriptome, and phenome provide information about the underlying biological system and thus can guide analytic methods to discover associations that are hidden to simple methods. The recent development of a new generation of GWAS algorithms, termed *structured association mapping* algorithms, utilizes structural and other known information inherent to GWAS data to discover genome-transcriptome-phenome associations.

While initial studies have suggested that structured association mapping leads to increased insight and greater statistical power in association study [15,16], there are barriers to the wide-spread use of these algorithms. For example, the power of structured association mapping comes with more sophisticated machine learning techniques that require greater specialization to run and interpret. Also, due to the data complexity, results from these algorithms become a sea of data that can be challenging to explore.

In response to the challenges of today's association studies, we present recent advancements in structured association mapping. We review the development of structured association mapping algorithms and report on the development of new visualization strategies to explore the results from these data. We use a visual analytics software tool, GenAMap, to showcase these algorithms and visualizations. GenAMap is a comprehensive system for structured association mapping to 1) automate the execution of structured association mapping algorithms and 2) provide new visualizations specifically designed to aid in the exploration of association mapping results.

We begin this paper with an overview of structured association mapping. We then present a discussion of the motivation, design, and implementation of GenAMap. Finally, we demonstrate this new approach to GWAS and eQTL analysis through two cases studies using yeast and human data. Finally, we use GenAMap to analyze the NIH heterogeneous stock mice data [17]. By using structured association mapping and visualization in GenAMap, we find an eQTL hotspot on the mouse chromosome 14 that is associated with axon genes. We further investigate this eQTL hotspot and find specific genes associated with anxiety traits in mouse. The three-way analysis using structured association mapping provides additional mechanistic insight into the SNP-trait association that has not been possible using other state-of-the art methods.

Structured Association Mapping

In this section we will briefly review structured association mapping, an emerging algorithmic paradigm for GWAS. Structured association mapping is advantageous over simple, pairwise methods because it takes advantage of structure in the genome, transcriptome, and phenome in the discovery of association signals. We will consider each of these "-omes" in turn to describe the available structural information that is available and how structured association mapping leverages that structure in a systematic framework.

Structured association mapping is built using sparse-regression techniques, built off of the lasso [18]. The lasso is advantageous in association mapping as it selects the most informative predictors (SNPs) for each response (genes or traits) and eliminates false positives. As we incorporate further information from the data into this statistical paradigm, we not only select the most informative SNPs, but we also leverage the structure through the addition of further optimization penalties to enhance discovery.

Let us define the problem of association mapping as follows. Let $\boldsymbol{X}$ be an $N \times P$ genotype matrix for N individuals and P SNPs and let $\boldsymbol{Y}$ be an $N \times J$ gene expression matrix where expression levels of J genes are measured for the same individuals. Finally, let $\boldsymbol{Z}$ be an $N \times K$ phenotype matrix where each row records K phenotypic traits of an individual. Then, using lasso regression to find associations between $\boldsymbol{X}$ and $\boldsymbol{Z}$, we optimize the following equation:

$$\boldsymbol{B} = \underset{\boldsymbol{B} \in \mathbb{R}^{P \times J}}{argmin} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 + \lambda\, |\boldsymbol{B}| \qquad (1)$$

$||.||_F$ is the Frobenius norm of the matrix. The first term in the equation penalizes based on prediction error, and the second term is the L_1 lasso penalty, which has the property of shrinking the strengths of irrelevant SNPs towards zero. In this scenario, $\boldsymbol{B}$ is a $P \times K$ matrix representing associations between SNPs and phenotypes.

Genome structure

Structure and information from the genome can provide insight in association analysis. For example, consider population structure. While many SNPs may be population-specific, some SNPs may have similar effects across populations. The multi-population group lasso (MPGL) is a sparse-regression method that allows associations to be discovered in different populations independently, while incorporating information across all populations [16]. This is done by building on the multi-task learning research in machine learning. More specifically, MPGL does this through the introduction of a L_1/L_2 penalty instead of the lasso penalty, as shown in Eq. 2:

$$\boldsymbol{B} = \underset{\boldsymbol{B} \in \mathbb{R}^{P \times J}}{argmin} ||\boldsymbol{Y} - \boldsymbol{XB}||_F^2 + \lambda \, ||\boldsymbol{B}||_{L_1/L_2} \text{ where } ||\boldsymbol{B}||_{L_1/L_2} = \sum\nolimits_{j=1}^{p} \left\|B_j\right\|_F^2 \quad (2)$$

In this case, $\boldsymbol{B}$ is calculated trait-by-trait. If we assume C populations, then $\boldsymbol{B}$ is a $P \times C$ matrix, with one column for each population. Thus, rows represent SNPs and columns represent associations from SNPs to populations for one trait.

Alternatively, we can consider known features about the SNPs in our dataset. For example, certain SNPs are conserved across species, in genetic promoter regions, or in non-coding DNA. Using this information, we can use the Adaptive Multi-task lasso (AMTL) to find genome-transcriptome or genome-phenome associations [19]. AMTL uses multi-task learning to optimize a lasso-type equation as well as select weights on the SNP features. The result is a $P \times K$ $\boldsymbol{B}$ matrix representing genome-phenome associations.

Transcriptome and phenome structure

Now we consider leveraging structural information in the transcriptome in association analysis to detect eQTLs. The methods that we consider can also be used for genome-phenome association analysis with multiple-correlated traits.

We will review two structured association mapping algorithms that leverage information in the traits or genes. These methods assume that related traits or genes tend to be influenced by a common, small subset of SNPs. Biologically, this might be the case when a mutation in a genetic regulator affects the expression levels of multiple genes in a common pathway.

Graph-guided fused lasso (GFlasso) [15], extends the lasso such that a network structure is used to guide the discovery of associations. This network can be constructed using simple techniques based on correlation, or it can reflect known gene-gene or protein-protein interactions that are experimentally tested. We define $G_G = (V_G, E_G)$ as a relevance graph where each node represents a gene in $\boldsymbol{Y}$ and each edge represents a weighted relationship between two nodes in the network graph. GFlasso is then described by the following optimization problem:

$$\begin{aligned} \boldsymbol{B} &= \underset{\boldsymbol{B} \in \mathbb{R}^{P \times J}}{argmin} \|\boldsymbol{Y} - \boldsymbol{XB}\|_F^2 + \lambda \sum_j \sum_p |B_{pj}| \\ &\quad + \gamma \sum_{\{u,v\} \in E_G} \sum_p \left| B_{pu} - sign(\rho_{uv}) B_{pv} \right| \qquad (3) \end{aligned}$$

In Eq. 3, $\boldsymbol{B}$ is a $P \times J$ matrix representing genome-transcriptome associations. Similarly, we can create a network graph $G_T = (V_T, E_T)$ for the traits and substitute $\boldsymbol{Z}$ for $\boldsymbol{Y}$ and G_T for G_G to find a $P \times K$ matrix representing genome-phenome associations. GFlasso thus leverages the network structure of the genes or traits to find associations to SNPs that affect networks or pathways of genes.

A similar approach to GFlasso is TreeLasso [20]. TreeLasso builds a hierarchical clustering tree from the genetic network and uses the tree as information about the relationships between genes or traits to guide the association discovery.

Joint three-way analysis

Finally, we consider a structured association mapping approach that uses combined genome, transcriptome, and phenome data to perform a joint-three way association analysis, GFlasso-gGFlasso [21]. This is done through a two-stage process. First, we find genome-transcriptome associations using GFlasso as just described. Next, we find transcriptome-phenome associations using the graph-Graph-guided fused lasso (gGFlasso):

$$\boldsymbol{B}_2 = \underset{\boldsymbol{B} \in \mathbb{R}^{J \times K}}{argmin} \|\boldsymbol{Z} - \boldsymbol{YB}\|_F^2 + \lambda \sum_j \sum_k |B_{jk}| \\ + \gamma_1 \sum_{\{u,v\} \in E_G} \sum_k |B_{uk} - sign(\rho_{uv}) B_{vk}| \\ + \gamma_2 \sum_{\{m,l\} \in E_T} \sum_j |B_{jm} - sign(\rho_{ml}) B_{jl}|. \qquad (4)$$

In gGFlasso, we add a second fusion penalty to the GFlasso framework to encourage related genes in the network to influence related traits in the trait network. This model assumes that the effects of genes in the same pathways might be similar on multiple-related traits.

Each of the five structured association mapping algorithms that we have reviewed here (MPGL, AMTL, GFlasso, TreeLasso, and gGFlasso) is available online and is automated in GenAMap.

Motivation, Implementation, and Design of GenAMap

Advances in machine learning hold significant potential to facilitate discovery in association studies. However, two key obstacles hinder these advances from being widely accepted in practice. The first obstacle is the expertise required to run structured association mapping algorithms. The state-of-the-art practice is to distribute machine learning algorithms via command line implementations, which must be extensively customized before they can be used in practice. The second obstacle is the exploration of the results after the algorithms complete. In a structured association analysis, the analyst is no longer considering a small list of SNP-trait associations, but rather a sea of data with complex structure. Current visualization strategies for exploring structure are not customizable to association studies [22], and association visualization strategies for GWAS are limited in the types of analyses they can perform; most can only look at one trait or gene expression level at a time [23,24,25].

We present new visualization strategies to explore the results of structured association analysis as well as an automated processing system through a *visual analytics* software tool called GenAMap. Visual analytics is a rapidly emerging field that describes the combination of advanced information visualization techniques with statistics and machine learning. Visual analytics tools combine the power of automated information extraction with the intuition and cognitive strengths of human decision making [26]. By combining the strength of cutting-edge machine learning technology for structured association mapping with novel visualizations built for the exploration of structured association data, we propose a powerful system for structured association mapping.

Automation: GenAMap is a software suite of algorithmic tools for association mapping

The integration of structured association mapping analysis into GWAS has been slowed, in part, by the state of the art practice in the development of these algorithms. Structured association mapping algorithms are generally made available as crude, command-line implementations (if they are made available at all). Thus, for a geneticist to use a structured association mapping algorithm, he/she must download a rough implementation of the algorithm and customize the code to fit his/her study. As part of the GenAMap system, we incorporated an end-user-friendly strategy for the deployment of new machine learning algorithms to increase their accessibility for geneticists and biologists. We argue that the wide-spread acceptance of more accessible approaches could potentially accelerate biological discovery by facilitating the incorporation of cutting-edge machine learning techniques. More specifically, GenAMap runs structured association mapping algorithms through an automatic processing system called Auto-SAM<CITE…if accepted for publication>. In contrast to the general strategy of posting a raw implementation on the web, we systematically develop each algorithm so it will automatically run in a distributed parallel-computing environment. Thus, little technical specialization is required for a genetics analyst to pick up GenAMap and run the algorithms.

GenAMap runs a variety of algorithms through Auto-SAM including structure-generating algorithms, association algorithms, and structured association mapping algorithms, listed in Table 1. To generate structure, Auto-SAM provides algorithms to build networks and find population structure. GenAMap runs baseline association methods through Auto-SAM including PLINK's chi-square and Wald tests [27]. Most notably, GenAMap automates five structured association mapping algorithms: GFlasso, TreeLasso, AMLT, MPGL, and gGFlasso. Analysts can also load in their own structures and results into GenAMap, bypassing Auto-SAM and using GenAMap's visualizations to analyze the association results from any algorithm.

Our approach of distributing structured association mapping algorithms through Auto-SAM has several advantages over other distribution methods such as CRAN-R [28] (for examples see glasso [29] or bioconductor [30]). 1) By running algorithms on a distributed system with access to a cluster-computing system, Auto-SAM is able to handle much larger datasets and run algorithms in parallel; 2) through the use of a database, analyses are made available to entire teams of analysts; 3) the integration of Auto-SAM with GenAMap provides state-of-the-art visual analytic tools that enable the analyst to explore and analyze the data and results, including links to external databases and integration with gene-ontology resources.

Visualization: GenAMap is visualization software for association mapping

In structured association mapping, millions of SNPs, tens of thousands of genes, and hundreds of phenotypic traits are analyzed. Although machine learning has carried association mapping well thus far, we believe that the next steps strongly indicate that visualization, combined with machine learning, will advance the field still further by taking advantage of the analytic capabilities of *both* machines and people. The vast amount of input and output to these algorithms, and the sparseness of useful output, classically suggests that a visualization strategy will aid analysts in the exploration of this data to identify the links between DNA, genes, and traits to eventually produce new treatments for disease.

Analyzing structured association mapping results is a near perfect fit for visualization for many reasons. First, once an analyst has run structured association mapping algorithms, the focus of the investigation becomes more exploratory than query driven [31]. Information visualization, “the use of computer-supported, interactive visual representations of data to amplify cognition,” as a field, touts its strengths as generating exploration-based insights, explanatory and persuasive interaction, and aesthetic representations [32]. Visualization techniques, therefore, excel when providing an explanation of the overall structure of the data or to find weak or unexpected patterns most easily recognized by humans [33], critical requirements for structured association analysis.

Indeed, the success of visualization strategies has emerged already in many areas of biology. For example, Cytoscape [22] has become an extremely popular application for visualizing biological networks and exploring relationships between genes. In other domains, the recent development of ABySS-Explorer [34] has shown that visualization can enhance the analysis of complex biological tasks like genome assembly through a visual representation of the contigs. Another recent approach to visualization in biology, MulteeSum, demonstrated the potential for visualization to aid in the identification of spatial and temporal patterns in gene expression data [35]. For simple GWAS with one trait, excellent visualization tools have been built to explore linkage disequilibrium (LD), strength of association, and surrounding genes in the association results [23,24]. Thus, visualization is proving to be reliable in its application to biology.

Given the promise of visualization in genetics research, we present visualizations in GenAMap built to explore structured association mapping results. In GenAMap, we use multiple-coordinated views to enable analysts to explore the structures of the genome, transcriptome, and phenome simultaneously when performing association analysis. In our experience, in a structured association study, researchers need to get an overall picture of the patterns of association in the data, and then they need to focus their attention on specific, important signals in the data. This immediately suggests a visualization strategy following Shneiderman's well-known mantra: overview first, zoom and filter, details on demand [36]. As we will show, this mantra has provided an excellent strategy for the development of the visualizations that guide discovery in association studies. In this paper, we present these visualizations through two cases studies using GenAMap.

Implementation

GenAMap is implemented in Java SE. To facilitate the rapid development of high-quality visualizations, we have integrated and customized open-source visualization Java toolkits into GenAMap, including JUNG [37], JHeatChart [38], and JFreeChart [39]. GenAMap communicates with Auto-SAM, the automatic processing system, through an Apache web-interface. Data storage is done via MySQL, and algorithms are run on an automatic processing system implemented in Java SE, C++, R [28], and MATLAB. Algorithms are parallelized and run using Condor [40].

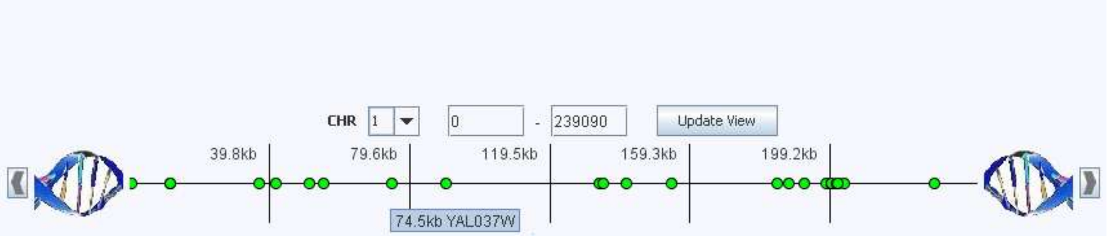
Case Studies

In this section, we demonstrate the visualization tools available in GenAMap with two cases studies using real data. Through these studies, we highlight the structured association mapping methods available to run in GenAMap, and we also describe the visualizations available to explore the results from these analyses.

Case Study 1: Exploring gene networks and eQTLs using yeast data

We first introduce GenAMap through a demonstrative analysis using a dataset from budding yeast, *Saccharomyces cerevisiae* [41]. Because this dataset has been extensively studied [42,41,43,44], it serves as an excellent dataset to highlight the capabilities of structured association mapping and GenAMap in a scenario where much is already known about the associations in the data. This dataset was generated by crossing a laboratory strain (BY4716) of yeast with a wild-type vineyard strain (RM11-1a) to create 112 progeny yeast strains. Each of the 114 strains were sequenced for 1260 unique SNP markers. Gene expression data was also collected from each strain for over 6000 traits. After preprocessing the gene expression data, we used 5637 gene expression measurements for each yeast strain. The data collection and preprocessing steps were completed independently outside of the GenAMap software system.

Figure 1 - GenAMap's genome browser
GenAMap provides a simple genome browser that allows analysts to explore the mutation marker data that they load into GenAMap. SNPs are represented by green circles across the genome. Analysts can use these SNPs to directly link to external databases, such as SGD or dbSNP. SNP labels are displayed as the analyst hovers over the SNPs.



Exploring the gene network

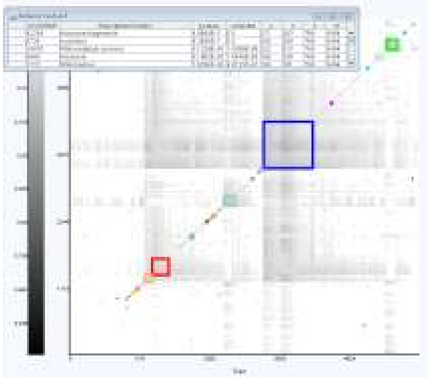
Once the data has been preprocessed, it is ready to import into GenAMap to begin association analysis. We import the SNP data as a tab-delimited file into GenAMap using the import wizard. When the import finishes, we can explore the data using GenAMap's genome browser (Figure 1). GenAMap's genome browser is a simple chromosome-by-chromosome browser that displays each SNP as a green circle. We use the genome browser to check the distribution of SNPs on each chromosome and to directly link to the *Saccharomyces* Genome Database (SGD) [45] for more information about the SNPs. For future analyses, we download and standardize twelve features from the SGD for each SNP and add these features to the dataset in GenAMap. These features include eleven discrete variables describing the location of the SNP (intron region, binding site, exon, etc.) and one continuous feature (conservation score) [19]. As we browse through the SNPs, we can request to see these features by right-clicking on a selected SNP (or many selected SNPs) in the genome browser.

Similarly, we load the gene expression data into GenAMap using the import wizard. Once the gene expression data have loaded into GenAMap, we use GenAMap to automatically build a gene-gene network using the soft-thresholding method to create a scale-free topological overlap matrix [46]. GenAMap can also create a network using the pairwise correlation coefficient or glasso [29]. When the network is created, we want to get an overall picture of the gene interactions to understand the network structure. GenAMap supports this type of analysis through the discovery of *gene modules* within the network. To find gene modules, we first use GenAMap to automatically run a simple pairwise association test, the Wilcoxon-Sum Rank test with false discovery rate (FDR) correction [42], to find SNP-gene associations. Simultaneously, we use GenAMap to run hierarchical clustering to cluster highly connected genes in the network. Once these two algorithms finish running, we run a job in GenAMap to discover the top twenty connected gene modules in the clustered network. GenAMap identifies these modules automatically on the parallel computing cluster using a previously defined algorithm [42] and also calculates eQTL enrichment (using the pairwise associations) and gene ontology (GO) enrichment (using BiNGO [47]) for each discovered module.

At this point, we have used GenAMap to prepare a gene expression data network; the next step is to use GenAMap's visualization tools to explore the gene network to find interesting interactions. First, we explore the patterns in the entire network to get an overview. In Figure 2 we show a screen-shot of GenAMap's overview visualization of gene networks. This overview is presented as a heat map, where darker pixels represent a weighted relationship between genes. The genes in the heat map have been clustered, and 20 identified modules are outlined in color. As we select different gene modules in the network, GenAMap displays the module's eQTL and GO enrichments. We find that the modules are significantly enriched for GO category and eQTL association, consistent with previous reports [42].

Figure 2 - GenAMap trait overview exploration

GenAMap provides an overview of gene and trait networks to aid analysts in their exploration of the networks. Here, we present a genetic network generated from the yeast data. The network has been clustered by hierarchical clustering, and twenty highly connected gene modules have been automatically identified by GenAMap (outlined in color). As the analyst clicks in these different modules, an information display appears to report the GO and eQTL enrichment of the genes that belong to the particular module.

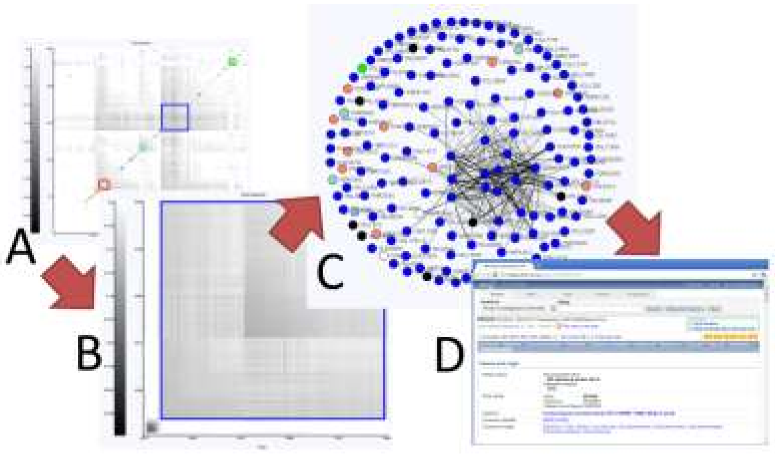


As we referenced in the design section, all of GenAMap's visualization tools are developed to give an overview first, provide tools to zoom and filter, and then link to details on demand. The network view follows this pattern. Once we have an overview picture of the gene network, we use GenAMap to drill down into the data to explore interesting sub-networks. We provide one simple example of this type of top-down exploration.

From the network overview, we observe that the largest of all the modules in the network is the blue module, made up of 788 genes. This module is enriched for many GO categories including “ribosome biogenesis” (*p*-value = 4.0604e-169) and has eQTL enrichment to many SNPs including a SNP on chromosome II:548401 (*p*-value = 2.769e-48). To explore this sub-network further, we manually zoom into this region of the heat chart display of the network. GenAMap displays gene and trait networks at a series of resolutions, and so as we zoom into this region of the network we see the finer detail of the gene-gene relationships (Figure 3). We select the most highly connected part of the network and switch to the *JUNG view*, which displays sub-networks of up to 200 traits/genes in a ball and stick representation. We summarize the process we follow to zoom into this region in Figure 3.

In the JUNG view, genes are represented as circle nodes, and relationships between genes in the network are represented as weighted lines. Thicker lines imply a strong weight or degree of connection or correlation between genes. There are several different layouts available for us in this view, including a simple circle layout and the KK-layout [48] shown in Figure 3C. Now that we have zoomed into this region, we use GenAMap to get details about these genes. We perform a GO enrichment analysis, which finds that the genes we selected are enriched for the GO category ribosome (*p*-value = 4.89e-169). We adjust the edge threshold manually to remove edges with lower weights; this allows us to find the most highly connected genes in the network. Because the top-connected genes in this network may be important players in the sub-network, we right-click on these gene’s labels to link directly to Google search and to the gene’s UniProt webpage [49]. These details on demand help us to understand what genes are particularly active in this sub-network, for example *RPS24A*, a ribosomal protein from chromosome V is the gene with the strongest connections in the network.

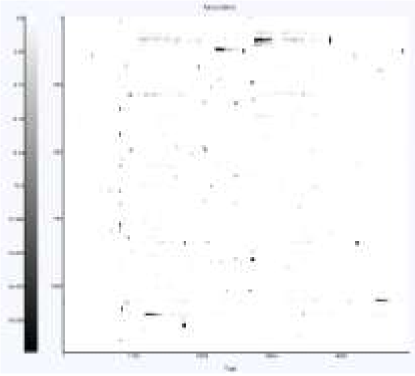
Figure 3 - Using GenAMap to explore genetic networks
We demonstrate using GenAMap visualizations to explore a genetic network. A) From the overview of the network, the analyst can see the different gene modules in the network. B) The analyst zooms into a module of interest in the network. C) The analyst switches to a node-edge representation of this sub-network and adjusts the edge threshold, layout, and labels. D) The analyst uses GenAMap to link directly to external data sources for more information.



Finding eQTLs using GenAMap

Given the high modularity of the gene network, we determine to run TreeLasso [20] to find SNPs associated with genes. TreeLasso uses the network structure to cluster the genes into a tree structure, and the structure guides the algorithm to find associations from SNPs to related genes. In Figure 4, we present an overview of the results from running TreeLasso automatically in GenAMap. This view shows a heat map where SNPs are plotted along the x-axis and the genes plotted along the y-axis. Genes are clustered according to hierarchical clustering as before. Dark pixels represent an association between the SNPs and the genes, and white pixels represent no associations.

Figure 4 - GenAMap overview of association results
GenAMap provides a heat chart visualization to explore the results from an eQTL association analysis. SNPs are plotted along the *x*-axis and genes are clustered along the *y*-axis. This view allows the analyst to explore the overview of the results. For example, in these results from running TreeLasso on the yeast data, many SNPs are associated with all the genes in a gene module, and some gene modules are associated with many different SNPs in different genomic locations.

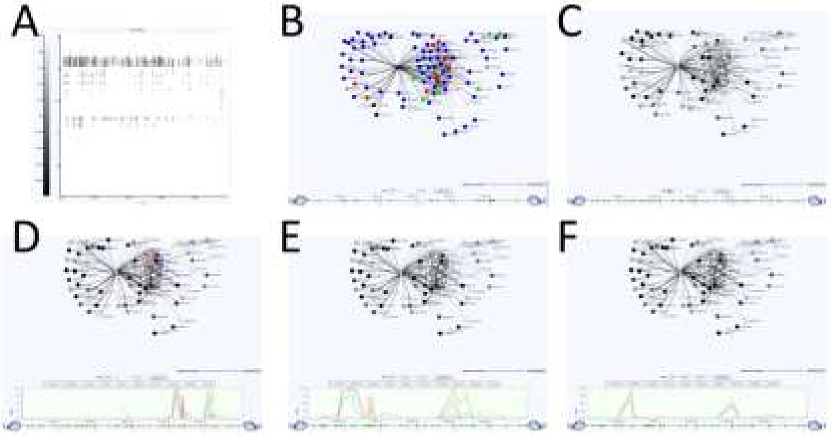


From the results shown in Figure 4, we observe that many SNPs are associated with clusters of genes, meaning that the associations follow the modular structure of the data. We also observe that many of the gene modules are associated with more than one SNP, suggesting some kind of interaction between the SNPs to regulate or affect the gene expression of the module. We zoom into the heat chart to see the finer structure of the associations in the largest cluster (Figure 5A). We notice that there are ten SNPs in the same genomic region that are associated with these genes. To explore these associations, we select the 131 genes in the cluster with strong associations and switch to the JUNG view.

When exploring associations, the JUNG view shows the network view as before, but GenAMap also provides access to the genome browser. By using these coordinated views, the analyst can explore structure in the genome and the traits, while simultaneously querying to understand the associations between the two. Once in the JUNG view, we perform a GO enrichment test to see if our selected genes have a common function. Indeed, the genes are enriched for the GO annotations “nucleolus” (*p*-value = 2.091e-107), “ribosome biogenesis” (*p*-value = 2.623e-99) and “RNA metabolic process” (*p*-value = 7.081e-66). In Figure 5B, we show these genes colored by GO category. All genes with the GO annotation for “nucleus” are shown in blue. Genes without this annotation, but annotated as “ribosome biogenesis” are shown in red. Genes not annotated as either, but annotated as “RNA metabolic process” are shown in green. We can change this coloring based on the GO categories we are interested in. We conclude that this group of genes appears to be a functionally cohesive group of genes involved in ribosome biogenesis that co-locate to the nucleolus.

From the overview, we know that these functionally coherent genes have strong associations to at least ten SNPs on chromosome II. We select half of chromosome II and color the genes by association to the selected SNPs (Figure 5C). Genes with a strong association to these SNPs are shown in white, genes shown in black are not associated, and gray genes represent varying levels of association. In this view, the SNPs being considered are shown as yellow triangles for our reference. We discover that most of the genes in this module are associated to this region on chromosome II. To further explore this association, we select ten genes in the module (highlighted in salmon in Figure 5D) and view the Manhattan plot of the associations strengths of these genes across chromosome II. We zoom into the region with the strongest associations (Figure 5E) and note that there are many SNPs with associations to these genes.

Figure 5 - Using GenAMap to find eQTLs in yeast data
GenAMap provides many tools for analysts to explore association results while using the structure of the data to guide the discovery of associations. We demonstrate some of these tools. A) The analyst can zoom into certain regions to see finer detail of the SNP-trait associations. This panel is a zoomed-in region from Figure 4. B) The analyst switches to the JUNG view to explore the genes associated with the region and perform a GO enrichment test. C) The analyst colors the genes by strength of association to the genomic region. D) The analyst selects up to ten interesting genes (salmon colored) and views the Manhattan plot of associations from these genes across the genome. E) The analyst zooms into interesting regions in the genome view. F) The analyst can switch between association tests for further insight into the associations.



The number of SNPs associated to these genes complicates our analysis. We want to find the SNPs that are the most likely to be associated with the genes in this module. Because we have already added feature data to our SNPs, we can run AMTL to find associations in the yeast dataset. AMTL, unlike TreeLasso, takes into account SNP features instead of genetic structure. Thus, AMTL selects SNP-gene associations based on features that predict the likelihood of the SNP to cause a change in gene expression. Once the AMTL analysis is complete, GenAMap allows us to switch between the TreeLasso results and the AMTL results easily in the same view of the data (Figure 5E and 5F). Indeed, the AMTL results find associations to far fewer SNPs on chromosome II for this set of ten genes. We inspect the two SNPs on chromosome II that have associations in the AMTL results to these genes. We use GenAMap to link to the SGD for more information about these SNPs and find that one SNP is in *RPB5*, a component of RNA polymerase, and the other SNP is in *PYC2*, a gene near *SDS24*.

In this demonstration using the yeast data, we have shown how GenAMap enables an analyst to survey a gene expression network to find modules and then to drill down for further detail about these modules. We have also demonstrated how GenAMap enables the exploration of association results to discover gene modules under the regulation of eQTL hotspots. Finally, we show how GenAMap allows analysts to compare the results from different structured association mapping tests to better understand association signals.

Case Study 2: Using GenAMap to explore population structure and associations in a human asthma dataset

We now demonstrate a different set of analysis and visualization tools available in GenAMap using a human dataset collected for the study of asthma. We use the data collected through the Severe Asthma Research Program (SARP) and the Cooperative Study for the Genetics of Asthma (CGSA). This data set is an excellent dataset for demonstration because it complements the yeast dataset, showing the versatility of GenAMap to perform analyses appropriate for different datasets from different species. Additionally, this dataset has been collected from individuals in two distinct populations, a non-Hispanic white and an African American population, allowing us to show analyses in GenAMap that use population structure.

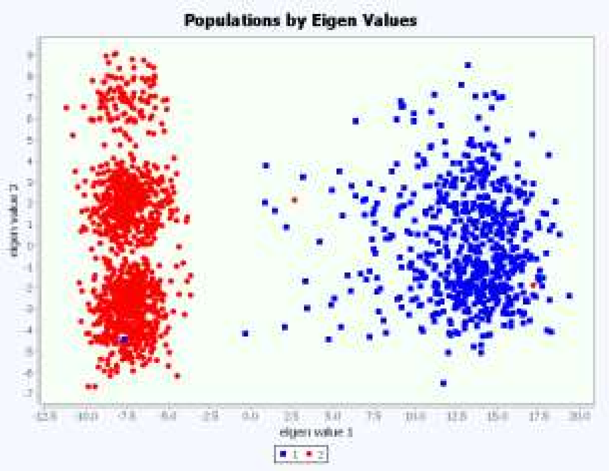
The combination of the SARP and CGSA datasets creates a dataset with 1745 individuals genotyped for 752256 SNPs. In addition to case/control assignment, 18 other clinical traits are available. Due to implementation constraints, GenAMap cannot import datasets of larger than 5000 SNPs for 2000 individuals. As a preprocessing step to find the most interesting SNPs, we use PLINK [27] to run the chi-square test on each of the SNPs against the asthma phenotype (cases/controls) in each population. We choose a significance cut-off at 2.5e-3 to select 3785 SNPs that were associated with asthma in one of the two populations (three SNPs were associated with asthma in both populations at this significance level). We import these 3785 SNPs and the 19 traits into GenAMap for analysis.

Assessing population structure

The first step we take upon importing the data is to explore the population structure of the data. We compare race assignments for the individuals found in two ways: 1) self-reported race and 2) the results from running Structure [50] (using GenAMap) on the data assuming 2-10 populations. In the population view, GenAMap plots the individuals by Eigenvalue; the analyst can explore 2D plots for the first five Eigenvalues to compare different numbers of populations in the data. In this dataset, we find that the populations separate into two distinct populations (Figure 6). We also find that the population assignments made by running Structure strongly agree with self-reported race.

Figure 6 - Analyzing population structure in GenAMap

GenAMap provides an interactive view for analysts to explore population structure. Population assignments are plotted by individual by Eigenvalue. The analyst can adjust the 2D plot to adjust between the first five Eigenvalues. Here, we present the results from a population analysis on the asthma data. Self-reported race is plotted according to the first two Eigenvalues. The plot shows clear separation between the populations.

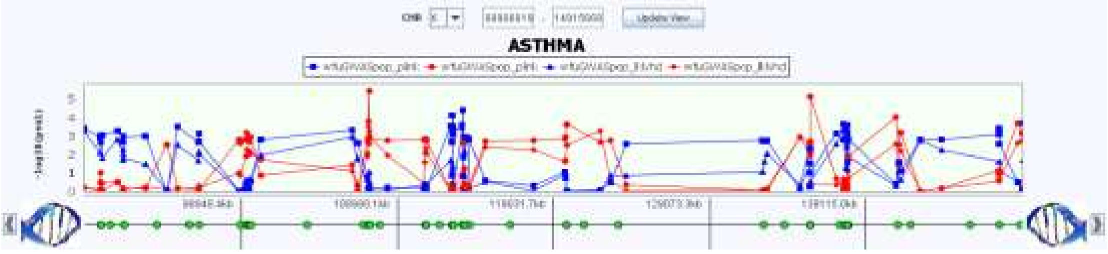


Exploring association by population

Given the strong separation between populations, we choose to perform simple, baseline association analyses that find associations for each population one at a time. GenAMap provides four simple statistics to explore associations by population [51]. These four analyses are automated and run in parallel in GenAMap. The four analyses are 1) the Wald (qualitative traits) or chi-squared likelihood (binary traits) test as implemented by PLINK [27], 2) a two-sided t-test on the phenotype distribution by genotype, 3) a likelihood test [52], and 4) a cross-ten validation score by linear regression.

Once this analysis completes, GenAMap provides visualization tools to explore the differences in association by population and by test. Additionally, GenAMap provides an analysis tool to explore the similarities and differences in the results.

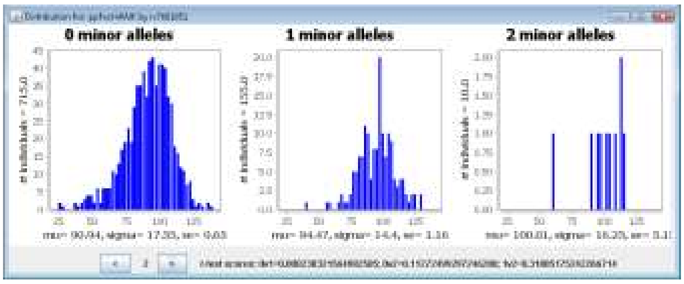
Figure 7 – Interactive Manhattan plot for population data
GenAMap provides an interactive Manhattan plot for exploring associations in population data as well as results between different tests. For example, here we show the results of two tests looking for genetic associations to asthma. The blue lines represent population 1 (African American population) and the red lines represent population 2 (non-Hispanic white population). Different tests are represented by different shapes in the plot.



In Figure 7, we present GenAMap's visualizations built to explore association by population and test. Figure 7 shows a region on chromosome 6 where many associations were found to the trait "Asthma" in the data. In this interactive Manhattan plot, we can add and remove tests, and we can also add and remove populations. For example, in Figure 7 we show the results from the PLINK and likelihood tests for both populations. From the plot, it is readily observable that the two tests find the same associations at approximately the same level of significance. We also recognize the interesting pattern that the SNPs associated with asthma in the African American population are not associated with asthma in the non-Hispanic white population (and vice-versa).

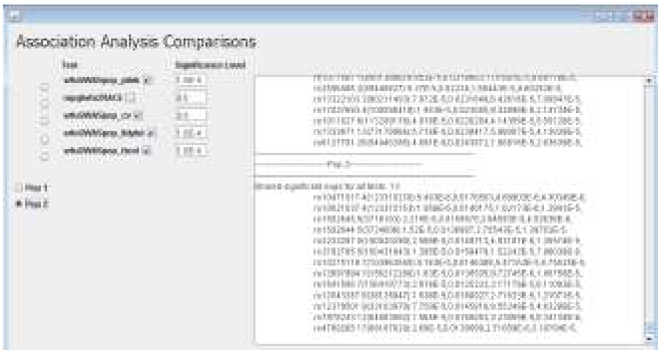
From this view, GenAMap provides other tools for the analyst that link to further details about the SNPs and associations. Directly from this view, the analyst can query for and link to the dbSNP [53] page of any SNP. For binary traits, the analyst can select a SNP and request to view the frequency table of the trait by genotype. For continuous traits, the analyst can compare the distributions of the trait by genotype (Figure 8).

Figure 8 – Frequency distribution of asthma trait by genotype
When exploring SNP-gene associations, GenAMap provides links to tools that allow the analyst to explore the discovered association. For example, consider a case where the analyst considers a discovered SNP-trait association. The analyst can query dbSNP to find out information about the SNP, and the analyst can use GenAMap to visualize the frequency distribution of the trait by genotype.


0 minor alleles
1 minor alleles
2 minor alleles
25
50
75
100
125
mu= 90.94, sigma= 17.35, se= 0.65
mu= 94.47, sigma= 14.4, se= 1.16
mu= 100.01, sigma= 16.25, se= 5.1
&lt;
2
>

Finally, we have implemented a dynamic query system in GenAMap to compare results from different association tests. The capabilities of this system are highlighted in Figure 9. Figure 9 shows the results from a query that compares the four different tests in population 2. In the query dialog, we select the tests that we want to compare, set the significance level of the tests, and choose what populations we want to consider. By performing this analysis, we find 13 SNPs that are associated with asthma across the genome in population 2 (non-Hispanic whites), in all four tests, at the significance levels we set. In addition to the simple statistical tests, we can also use GenAMap to run MPGL on the asthma data to find associations by population.

Figure 9 – Comparing association tests
GenAMap allows analysts to compare association results across tests. In this dynamic query tool, the analyst can select which tests to include in the comparison and the significance level of each test. They can see which SNP-trait associations are significant across all tests, and also which associations are unique to any given test.



Mouse Analysis

Thus far, we have introduced GenAMap as a visual analytics software system for structured association analysis. By combining the strengths of human intuition and structured association mapping, GenAMap is a powerful software system for association analysis.

One resource that is available for association studies is the NIH heterogeneous stock mice dataset [17]. The dataset consists of 460 mice that have been genotyped for 12,545 markers and phenotyped for 97 traits [54]. Additionally, expression profiling was recently added to the dataset from the liver, lung, and brain [55]. The expression profiling was done in the liver and lung for 260 genotyped mice, and in the brain for 460 mice. This dataset has been studied for SNPs associated with the mice phenotypes and for eQTLs. Cutting-edge resources are available to explore these associations and to investigate the strength of association on a trait-by-trait or a genomic-location basis. Thus, this dataset provides an excellent test bed to demonstrate a structured association analysis.

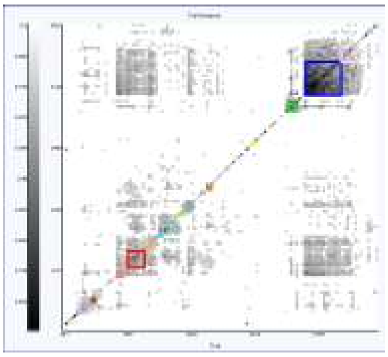
However, to date, the effect of an eQTL on a genome-phenome association is assumed based on the location of the eQTL in the genome. While this offers some insight into the mechanisms behind the association, the discovery of three-way genome-transcriptome-phenome associations can be enhanced using structured association mapping. As we will show in this section, by using the GFlasso-gGFlasso strategy [21] to find three-way associations, we not only uncover SNP-trait associations, but also find SNP-gene-trait associations that uncover some of the biological mechanisms behind the associations.

To prepare the data for analysis, we preprocessed the expression data from each tissue (hippocampus, liver, and lung) using lumi [56]. In each tissue, we retained all probes that had a significant signal ($d < .05$) for at least 95% of the mice. We limit our study to 218 mice that have gene expression measurements across all three tissues. We imputed missing phenotypic traits using k-nearest neighbor imputation [57], and we excluded all phenotypes missing values for more than 30% of the mice. In summary, our dataset included 218 mice. Each mouse is genotyped for 12545 SNPs, has measurements for 173 phenotypic traits, and has gene expression level measurements from the liver (7102 probes), lung (9698 probes) and hippocampus (9733 probes).

Genetic network analysis

We imported SNP data, gene expression data from three tissues, and phenotypic trait data into GenAMap. We analyzed the data using GenAMap in three steps: genetic network analysis, eQTL analysis, and three-way genome-transcriptome-phenome analysis. In this section, we discuss the results from the network analysis of the three tissues in GenAMap.

Figure 10 – Mouse gene network analysis
We used GenAMap to create gene-gene networks from the expression data for each tissue. GenAMap finds the top 20 connected modules and GO and eQTL enrichment for each module. Here, we show the gene-gene network generated using the hippocampus gene expression data.



After importing the data into GenAMap, we used GenAMap to construct genetic networks for each tissue using the soft-thresholding method described by Zhang and Horvath [46]. We also used GenAMap to run PLINK [27] to find SNP-gene associations. GenAMap automatically considers all *p*-values less than 1e-3 to be significant, which, although naïve in its approach, is a sufficient cutoff that allows us to get an overall idea of the associations in the dataset. We used GenAMap to cluster each of the three gene networks by hierarchical clustering and to run a dynamic programming algorithm [42] to find the top 20 connected gene modules in each network. For each of the 20 connected gene modules, GenAMap also performs enrichment analyses for the modules in terms of eQTL enrichment and GO category. We used the GO slim annotation and the associations found by PLINK for this analysis. In Figure 10, we show the annotated network generated from the gene expression data from the brain. The top connected modules identified by GenAMap are outlined in color. Tables showing the top GO and eQTL enrichments for each module for all three tissues are available in the Supplementary Worksheet.

We found that the gene networks were quite dissimilar across the three tissues. For each tissue, we found the number of unique genes (some genes are represented by multiple probes) and the number of unique edges between genes. We compare the three tissues in Table 2. While many genes (78% of the genes in the liver dataset) are shared between the three networks, few edges are common across all three networks (only 14% of the edges in the liver gene network are common across all three tissues). Because the set of genes included in each network are similar, we suggest that the differences between the networks are due to a difference in regulatory patterns of expression across the three tissue types.

Similarly, we found the gene modules found in each network to be distinct. Specifically, we found little overlap between the genes in each of the top 20 modules identified across tissues (see Supplementary Worksheet). Also, we noticed a difference in the GO and eQTL enrichments for the modules across tissues. In the liver, we found nine gene modules that were enriched for a GO category; these are listed in Table 3 and include enrichments for *mitochondrion*, *catalytic activity*, and *generation of metabolites and energy*. While the hippocampus network had eight modules enriched for a GO category, only two matched the liver enrichments and different GO categories were represented including *ribosome*, *calcium ion binding*, and *transport*.

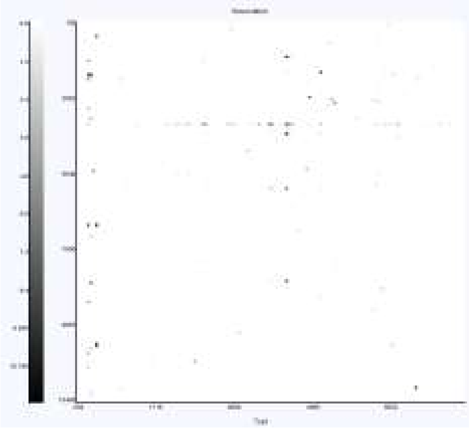
The eQTL enrichments for the modules were also different across the three tissues. Of note, we found five modules in the lung gene expression network that were significantly associated with enrichment for association to the SNP rs3023797. No modules in the liver or brain were significantly enriched for association with this SNP. Interestingly, rs3023797 is located in the exon region of the gene *Ttf1*, transcription termination factor, RNA polymerase I [53]. *Ttf1* has previously been shown to have important regulatory roles in lung function and development in mice [58]. These results, therefore, suggest that mutations in *Ttf1* affect the expression patterns in the lung, but not in the other tissue types. Similarly, six gene modules in the lung network also had an enrichment for association to chromosome 12 (26000000), which was not seen in the other two tissues. This suggests that there is a second mutation that is affecting lung expression patterns, but not hippocampus or liver expression.

eQTL analysis using GFlasso

Given the modularity of the gene expression networks, we used GenAMap to run GFlasso [15] to identify eQTLs for each tissue type. GenAMap uses cross-ten-validation to find optimal values for λ and γ using a linear search strategy (documented online: http://sailing.cs.cmu.edu/genamap). We downloaded all results from GenAMap and found all SNP-gene associations. SNPs within 2MB of each other and associated with the same gene are counted as the same association. We found the genomic locations of all genes [59] to classify associations as *cis*- or *trans*-associations. We define an association as a *cis* association when the SNP and gene are located on the same chromosome and located within 10MB of each other.

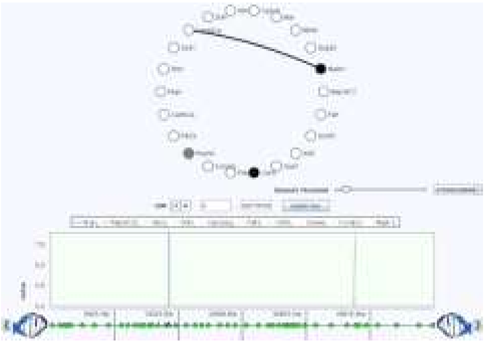
Our results were quite different in the different tissues. In the liver, GFlasso identified six SNP-gene associations; all six associations were *cis* associationsThe results from the lung were similarly sparse, with 25 SNP-gene associations discovered. GFlasso found one *trans* association and 24 *cis* associations. Overall, GFlasso found two *cis* SNP-gene associations that were common across all three tissues (*Gps2* and *Psmb6*), one *cis* association common between liver and lung (*C4b*), and four *cis* associations common between lung and hippocampus (*Mrpl15*, *Hsd17b11*, *Rpl21*, *Hbb-b1*).

Figure 11 – eQTLs found in hippocampus tissue
We used GenAMap to find SNP-gene associations in the hippocampus gene expression data using GFlasso. In this figure, we show the overview of the results in GenAMap. This is a heat chart representation of the associations, where SNPs are represented along the *y* axis and the clustered genes are represented along the *x* axis. We have zoomed into the section of the gene graph where there are the most associations. We note an eQTL hotspot (represented by a horizontal line of associations).



In contrast to the sparsity of the GFlasso results in liver and lung, GFlasso found many eQTLs using the hippocampus data. Specifically, GFlasso identified 467 SNP-gene associations for 103 SNPs and 268 genes. GFlasso identified 138 *cis* associations and 329 *trans* associations in the dataset. 79 genes were associated with more than one SNP, and 6 SNPs were associated with more than 20 genes. Although the sparsity of the results for liver and lung is surprising, our results are consistent with previous reports [55] that found that "*trans*-eQTLs are twice as common as *cis*-" in the brain, and that *trans*-eQTLs are much more common in the brain than in the other two tissues. Because the strength of GFlasso is to identify SNPs that affect multiple-correlated genes, it is no surprise that it was able to identify many *trans*-eQTLs in the hippocampus gene expression data as compared to the other two tissue samples. Because the results from the hippocampus are the most interesting, we will focus on these signals in the remainder of this section.

Figure 12 – Association of axon genes to chromosome 14
We found that rs8244120 on chromosome 14 was associated with 140 genes enriched for *cell projection*, implying function in neuronal axons. Here, we show 22 of these genes in GenAMap's node-link view, colored by the strength of association to rs8244120. White genes are strongly associated and black genes are weakly associated (gray is intermediate). We found that some of the genes were also associated with another SNP on chromosome 14 (shown) and some of the genes were associated with a SNP on chromosome 18 (not shown).



We present the overview of the SNP-gene association results for the brain gene measurements in Figure 11. From the overview, we note in particular one long horizontal line, suggesting the discovery of an eQTL hotspot that regulates many genes in *trans*. We also note the presence of other, shorter horizontal lines, including some short lines that overlap with some of the genes of the largest eQTL hotspot. We used GenAMap to discover the location of the SNP that was associated with these genes, rs8244120 on chromosome 14. We used GenAMap to find rs8244120 in dbSNP and found that it is located in the exon coding region of two genes: *Tmem55b* and *Apex1*. *Apex1* has been annotated for GO categories such as *DNA binding* and *DNA demethylation*, suggesting that *Apex1* potentially regulates the expression of other genes through its interaction with the genome.

To better understand the genes associated with this genomic region, we used GenAMap to create a subset of all genes associated with rs8244120. GenAMap identified 140 genes associated with rs8244120 in the GFlasso results. We performed a GO enrichment analysis using GenAMap on these genes to see if they shared common annotations. In fact, GenAMap found that the associated genes were enriched for the GO category *cell projection* (*p*-value = 2.65e-5). Cell projection is defined as "A prolongation or process extending from the cell, e.g. a flagellum or axon" [60]. Indeed, many of the genes in this subset are annotated to GO categories indicating involvement in brain function (e.g. *Gas7*, *Nrp1*, *Stx1a* are annotated to the GO category *neuron projection development*). We selected the 22 genes annotated with the *cell projection* annotation and saved them as a subset for further analysis. These 22 genes were enriched for many GO annotations including *cell projection* (*p*-value = 1.4911e-27), *neuron projection* (*p*-value = 5.3239e-17), *axon* (*p*-value = 3.3027e-9) and *dendrite* (*p*-value = 2.6457e-7).

We were interested to consider the associations of the identified cell projections genes to the SNPs (Figure 12). We plotted the Manhattan plot of the associations for the genes across GenAMap's genome browser. We noticed that all of the genes were associated with rs8244120, as expected, but that many genes had other associations as well. Two of the genes were also associated to rs13482353 (also on chromosome 14), and three of the genes were associated with rs3722205 on chromosome 18. We looked into these two SNPs in more detail and found that there are 27 genes associated with rs13482353, 25 of which are also associated with rs8244120. We also found that 25 of the 27 genes associated with rs3722205 are also associated with rs8244120. These results suggest that these SNPs may interact in some way to regulate gene expression in the mouse hippocampus.

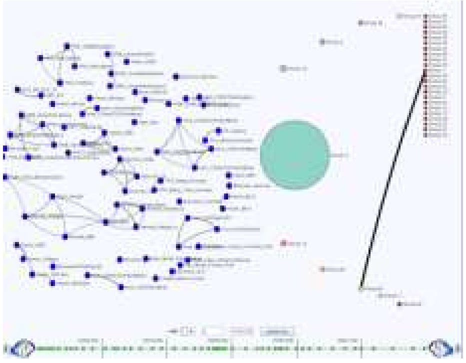
We also investigated an unrelated set of 41 genes that are associated with rs1348069 on chromosome 10 in the GFlasso results. We found that these 41 genes are enriched for several GO categories including *extracellular ligandgated ion channel activity* (*p*-value = 2.2018e-4), *membrane depolarization* (*p*-value = 2.5393e-4), and *synaptic transmission* (*p*-value = 5.1675e-4). rs1348069 is in the intron region of *Slc5a8*, a gene that has been annotated to the GO category *ion transport*, suggesting that this SNP may play a role in cell ion signaling by affecting these genes through altering the function or expression of *Slc5a8*.

Joint three-way genome-transcriptome-phenome analysis using GFlasso-gGFlasso

Given the results of our eQTL analysis, we determined to run gGFlasso to find associations from the brain tissue genes to the clinical trait phenotypes. We ignored all traits that were marked as “Covariates,” since these were largely dates, experimenter ids, and other variables such as gender and litter. Overall, GFlasso-gGFlasso found 759 SNP-gene-trait associations. These associations included 138 associations to the X chromosome, which we ignored due to possible gender effects. The results of GFlasso-gGFlasso thus consist of 621 associations between 98 SNPs on 18 chromosomes to 156 genes that are associated with 94 phenotypic traits.

We compared the GFlasso-gGFlasso results with the top 29 results reported using a SNP-trait association method [54]. We found nine matches where GFlasso-gGFlasso found a SNP-gene-trait association that matched the previously reported SNP-trait associations. We list these matches in Table 4, and complete results from our analysis are available in the Supplementary Worksheet. The GFlasso-gGFlasso results suggest associated genes that help to explain the SNP-trait associations that were previously discovered.

Figure 13 – Overview of three way GFlasso-gGFlasso association analysis
We show the overview of the trait-network and gene-network from GenAMap for the GFlasso-gGFlasso analysis; associations are not shown. In this visualization, circles represent groups of genes, associated to the same regions in the genome. Hexagons represent traits. The edges between genes or between traits represent the connections in the gene or trait network. In this data, we note that there are very few edges between gene groups. The largest gene group is the teal group, representing genes associated with the eQTL hotspot on chromosome 14. The trait network consists of small sub-groups of related traits.



We used GenAMap to drill down to explore the specific associations in the results. First, we considered the overall structure of the gene and trait data (Figure

13), noting that the largest gene group was associated with the eQTL hotspot on chromosome 14 as discovered in the previous section. To better understand the associations of these genes to the phenotypic traits, we used GenAMap to zoom into these genes and the associated traits, filtering out all other genes, traits, and associations (Figure 14). After exploring the results, we were especially interested in six genes that were found to be associated with sub-networks of anxiety traits (Elevated plus maze open arm time, distance, latency, etc.) due to the probable link between the brain and the traits themselves. In Figure 15, we show these traits, the correlations between traits (represented as gray lines between traits), and the gene-trait associations (pink lines between genes and traits). We also considered the associations of these genes to the genome and found that the genes were associated with two regions on chromosome 14. These results are consistent with previous findings that found two peaks on chromosome 14 associated with these traits [54]. Furthermore, the results also suggest potential mechanisms for these associations. For example, consider *Calb1*, a gene associated with the two eQTL hotspots and the anxiety traits. *Calb1* has been annotated to the axon, and knockout mice are known to show severe impairment in motor coordination [59]. Similarly, *Gabrd* is also associated with one eQTL hotspot and these traits. *Gabrd* knockouts have increased postpartum depression and anxiety, along with other disorders, and *Gabrd* is annotated to be involved in ion transport [59]. Thus, current knowledge supports the model that the GFlasso-gGFlasso results uncover: mutations on chromosome 14 affect the expression levels of *Calb1* and *Gabrd* in the hippocampus to affect anxiety traits such as *Elevated plus maze open arm time.*

Figure 14 – Gene-trait associations for genes associated with chromosome 14
We used GenAMap to explore the joint three-way associations corresponding to the eQTL hotspot on chromosome 14. We removed all other gene groups and then expanded the group to see the individual genes (squares). We filtered out all traits without associations to these genes.

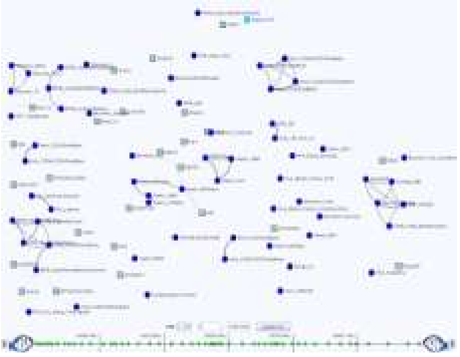
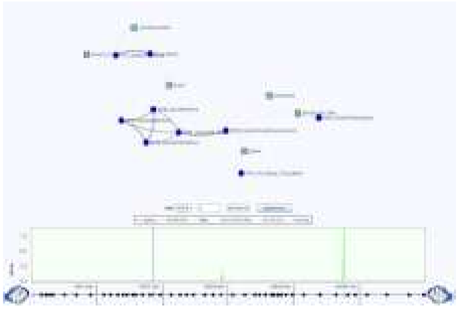
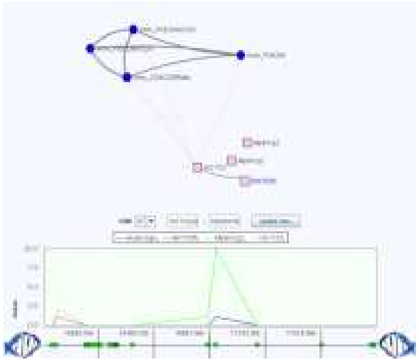


Figure 15 – Joint SNP-gene-trait associations from chromosome 14
We found a subnetwork of traits and associated genes involved in brain function. These genes were also associated with the overlapping eQTL hotspots on chromosome 14.



We also considered other associations that GFlasso-gGFlasso uncovered. For example, the GFlasso-gGFlasso results found an association between chromosome 17 and immunology traits (CD4+/CD8+, %CD4+/CD3+, and %CD8+), which was also reported as a strong signal using the simple SNP-trait association method [54]. We were interested to see if GFlasso-gGFlasso provided further mechanistic insight into this association. We used GenAMap to drill down to this association (Figure 16). We found a gene group consisting of four genes that were associated with these three immunology traits. One gene, *H2-T22*, was associated with all three correlated immunology traits. *H2-T22* was found to be associated with rs13482952 on chromosome 17, which is 3.2 mega-bases away from the *H2-T22* coding region. Given that the resolution for this cross is about 2MB, this SNP likely affects expression of *H2-T22* in *cis*. In fact, this region on chromosome 17 is part of the mouse H2 region, the major histocompatibility complex (MHC). The H2 region is the mouse ortholog to the human HLA region and encodes genes involved in the mouse immune response [61]. To summarize the H2 genes in mice, there are two classes of H2 genes. 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 [62]. The immunology traits associated with *H2-T22*: CD8+, CD4+, and CD3+, refer to proteins on the surface of immune response cells that bind to the antigens on the surface of other cells in the organism. *H2-T22* has been annotated as a membrane protein [59], and it likely participates in this immune response pathway. As the immune response is common across all cell types, we would expect to find this association in all cells, including the brain tissues.

Figure 16 – Immunity associations from chromosome 17
We found a small group of genes associated with the H2 region on chromosome 17. We also found that these genes were associated with a subset of immunology traits.



Discussion

In the post-GWAS era, the challenges facing geneticists are varied and difficult. While there is great potential for discovery and advancement with the ever-growing data available to geneticists, it is easy to drown in the multi-dimensional complexity and to slide by without taking advantage of the rich treasure-trove of information. Given the sophistication of methods needed to analyze and interpret results from the combination of genome, transcriptome, and phenome data, the inter-reliance between the fields of human genetics, molecular biology, machine learning, and information visualization is paramount. It is the frequent and successful collaboration between researchers from biology, visualization, and machine learning that will allow geneticists to fully capture the potential of the vast amount of data available.

In this project, we have described the development of GenAMap, a visual analytics software platform for GWAS and eQTL studies. GenAMap is a suite of algorithmic tools that provide ready-to-use access to cutting edge machine learning research in GWAS and eQTL analysis. Not only have we built GenAMap to provide access to state-of-the art analytic methods, we have designed visualizations to enable analysts to explore the sea of data that results from these types of algorithms. By building on tried-and-tested visualization principles, we have developed visualization strategies that will enable analysts to explore association results from any analysis. Through multiple-coordinated views, we provide analysts with the ability to explore the structure in the genome, transcriptome, and phenome simultaneously, while considering associations between the data types. We provide instant access to online databases, GO annotations, and association strengths. These tools enable the analyst to explore the data in ways that would not be possible using command-line query tools.

Furthermore, we have shown that GenAMap enables biological discovery through two case studies and an analysis on the mouse data. By using GFlasso-gGFlasso, we have not only uncovered SNP-trait associations, but have identified specific genes that are associated with the eQTL hotspots and the clinical traits themselves. Indeed, using additional data and more sophisticated techniques allows us to understand the biological mechanisms behind SNP-trait associations. Understanding the biological mechanisms behind SNP-trait associations brings us one step closer to the prevention and treatment of complex diseases.

To combat the increasing complexity of genetics analysis, we believe that research must follow a pattern of collaboration and cooperation between disciplines, even those as vastly different as genetics, information visualization, and machine learning. We believe that GenAMap serves as an exemplary foray into this type of multi-disciplinary collaboration to build a suite of tools and visualizations based on cutting-edge machine learning technology. The problems facing geneticists today are a near perfect-fit for visualization and machine learning. As these fields come together with solid collaboration, the potential for discovery will continue to accelerate.

List of abbreviations

AMTL adaptive multi-task lasso

eQTL expression quantitative trait locus

FDR false discovery rate

GFlasso graphical fused lasso

GO gene ontology

GWAS genome-wide association study

JUNG Java Universal Network/Graph Framework

MPGL multi-population group lasso

TreeLasso tree-guided group lasso

SNP single nucleotide polymorphism

SGD Saccharomyces Genome Database

Competing Interests

REC and EPX have applied for a provision patent application for GenAMap.

Authors' contributions

REC developed the software and drafted the manuscript. EPX directed the project and drafted the manuscript. PK provided a visualization perspective in drafting the manuscript. SW helped give a human genetics perspective to the drafted manuscript, and JLW gave a molecular biology perspective to the drafted manuscript. EB, DAM, and SW collected the human genetics data. AG helped to implement parts of GenAMap and drafted the manuscript.

Acknowledgements

We would like to thank Seunghak Lee for help us acquire the SNP feature data for the yeast dataset.

Funding: 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]; and an Alfred P. Sloan Research Fellowship to EPX.

References

1. Manolio RA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ,

McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler E, Gibson G, Haines JL, Mackay TF C, McCarroll SA, Visscher PM: **Finding the missing heritability of complex disease.** *Nature* 2009, **461**:747-753.

2. Hindorff LA, MacArthur J, (European Bioinformatics Institute), Wise A, Junkins HA, Hall PN, Klemm AK, Manolio TA: **A Catalog of Published Genome-Wide Associations Studies**. www.genome.gov/gwastudies
3. Yeager M, Chatterjee N, Ciampa J, Jacobs KB, Gonzalez-Bosquet JG, Hayes RB, Kraft P, Wacholder S, Orr N, Berndt S, Yu K, Hutchinson A, Wang Z, Amundadottir L, Feigelson HS, Thun MJ, Diver WR, Albanes D, Virtamo J, Weinstein S, eta: **Identification of a new prostate cancer susceptibility locus on chromosome 8q24.** *Nature Genetics* 2009, **41**:1055-1057.
4. Yaguchi H, Togawa K, Moritani M, Itakura M: **Identification of candidate genes in the type 2 diabetes modifier locus using expression QTL.** *Genomics* 2005, **85**(5):591-599.
5. Waring SC, Rosenberg RN: **Genome-Wide Association Studies in Alzheimer Disease.** *Arch Neurol* 2008, **65**(3):329-334.
6. Schadt EE: **Molecular networks as sensors and drivers of common human diseases.** *Nature* 2009, **461**:218-223.
7. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S,Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC,

Schuetz E, Rushmore TH, Ulrich R: **Mapping the genetic architecture of gene expression in human liver.** *PLoS Biol* 2008, **6**(5):e107.

8. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ: **Trait-Associated SNPs are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS.** *PLoS Genetics* 2010, **6**(4):e1000888.
9. McCarthy MI, Hirschorn JN: **Genome-wide association studies: potential next steps on a genetic journey.** *Hum. Mol. Genet.* 2008, **17**(R2):R156-R165.
10. Gilad Y, Rifkin SA, Pritchard JK: **Revealing the architecture of gene regulation: the promise of eQTL studies.** *Treds Genet* 2008, **24**(8):408-145.
11. Cookson W, Liang L, Abecasis G, Moffatt M, Lanthrop M: **Mapping complex disease traits with global gene expression.** *Nature Reviews Genetics* 2009, **10**:184-194.
12. Hsu Y, Zillikens MC, Wilson SG, Farber CR, Demissie S, Soranzo N, Bianchi EN, Grundberg E, Liang L, Richards JB, Estrada K, Zhou Y, van Nas A, Moffatt MF, Zhai G, Hofman A, van Meurs JB, Pols H A P, Price RI, Nilsson O, Pastinen T, Cupples LA, Lusis AJ, Schadt EE, Ferrari S, Uitterlinden AG, Rivadeneira F, Spector TD, Karasik D, Kiel DP: **An integration of genome-wdie association study and gene expression profiling to prioritize the discovery of novel susceptibility loci for osteoporosis-related traits.** *PLoS Genetics* 2010, **6**(6):e1000977.
13. Silveira AC, Morrison MA, Ji F, Xu H, Reinecke JB, Adams SM, Arneberg TM, Janssian M, Lee J, Yuan Y, Schaumberg DA, Kotoula MG, Tsironi EE, Tsiloulis AN, Chatzoulis DZ, Miller JW, Kim IK, Hageman GS, Farrer LA, Haider NB, DeAngelis MM: **Convergence of linkage, gene expression and association data**

demonstrates the influence of the RAR-related orphan receptor alpha (RORA) gene on neovascular AMD: A systems biology based approach. *Vision Research* 2010, **50**(7):698-715.

14. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE: **Variations in DNA elucidate molecular networks that cause disease.** *Nature* 2008, **452**:429-435.
15. Kim S, Xing EP: **Statistical estimation of correlated genome associations to a quantitative trait network.** *PLoS Genet* 2009, **5**(8):e1000587.
16. Puniyani K, Kim S, Xing EP: **Multi-population GWA mapping via multi-taks regularized regression.** *Bioinformatics* 2010, **26**(12):i208-i216.
17. Johannesson M, R Lopez-Aumatell, Stridh P, Diez M, Tuncel J, Blazquez G, Martinez-Membrives E, Canete T, Vicens-Costa E, Graham D, Copley RR, Hernandez-Pliego P, Beyeen AD, Ockinger J, Fernandez-Santamaria C, Gulko PS, Brenner M, Tobena A, Guitart-Masip M, Gimenez-Llort L, Dominiczak A, Holmdahl R, Gauguier D, Olsson T, Mott R, Valdar W, Redei EE, Fernandez-Teruel A, Flint J: **A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: the NIH heterogeneous stock.** *Genome Res* 2009, **19**(1):150-8.
18. Tibshirani R: **Regression shrinkage and selection via the lasso.** *Royal Statist Soc B* 1996, **58**(1):267-288.
19. Lee S, Zhu J, Xing EP: **Adaptive Multi-Task Lasso: with Application to eQTL Detection.** In *Advances in Neural Information Processing Systems 23 (NIPS),*

2010.

20. Kim S, Xing EP: **Tree-guided group lasso for multi-task regression with structured sparsity.** In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

21. Curtis RE, Yin J, Kinnaird P, Xing EP: **Finding Genome-Transcriptome-Phenome Associations with Structured Association Mapping and Visualization in GenAMap.** *Pacific Symposium on Biocomputing* 2012, **17**:327-338.

22. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**(11):2498-504.

23. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnkie M, Abecasis GR, Willer CJ: **LocusZoom: regional visualization of genome-wide association scan results.** *Bioinformatics* 2010, **26**(18):2336-2337.

24. Ge D, Zhang D, Need AC, Marin O, Fellay J, Telenti A, Goldstein DB: **WGAViewer: Software for Genomic Annotation of Whole Genome Association Studies.** *Genome Res* 2008, **18**(4):640-3.

25. Mueller M, Goel A, Thimma M, Dickens NJ, Aitman TJ, Mangion J: **eQTL Explorer: integrated mining of combined genetic linkage and expression experiments.** *Bioinformatics* 2005, **22**(4):509-511.

26. Keim DA, Mansmann F, Schneidewind J, Thomas J, Ziegler H: **Visual Analytics: Scope and Challenges**. In *Visual Data Mining*: Springer-Verlag Berlin; 2008:10.1007/978-3-540-71080-6 6.

27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA R, Bender D, Maller J, Skalr P,de Bakker, P I W, Daly MF, Sham PC: **PLINK: a toolset for whole-genome association and population-based linkage analysis.** *American Journal of Human Genetics* 2007, **81**(3):559-75.
28. R Development Core Team: **R Foundation for Staistical Computing**. http://www.R-project.org
29. Friedman J, Hastie T, Tibshirani R: **Sparse inverse covariance estimation with the graphical lasso.** *Biostatistics* 2007, **9**(3):432-441.
30. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettlin M, Dudoit S, Ellis B, Gautier L, Ge Y: **Bioconductor: Open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5**:R80.
31. Fekete JD, vanWijk JJ, Stasko JT, North C: **The Value of Information Visualization.** *LNCS* 2008, **4950**:1-18.
32. Card SK, Mackinlay JD, Shneiderman B: **Information Visualization: Using Vision to Think.** *Morgan-Kaufmann* 1998:San Francisco, California.
33. Card S, Mackinlay J, Shneiderman B, Kaufmann M, *Readings in Information Visualization*. 1999.
34. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horseman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ M: **De novo transcriptome assembly with ABySS.** *Bioinformatics* 2011, **25**(21):2872-2877.
35. Meyer M, Munzner T, DePace A, Pfister H: **MulteeSum: A Tool for Comparative Spatial and Temporal Gene Expression Data.** *IEEE Transactions on Visualization and Computer Graphics* 2010, **16**(6):908-917.

36. Shneiderman B: **The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations.** In *Proc 1996 IEEE Visual Languages*, Boulder, CO, 1996, 336-343.

37. Madahain JO, Fisher D, Smyth P, White S, Boey YB: **Analysis and Visualization of Network Data using JUNG.** 2005, **VV**(II).

38. Castle T: **JHeatChart**. http://freshmeat.net/projects/jheatchart

39. Gilbert D: **JFreeChart open source library**. http://www.jfree.org/jfreechart/index.html

40. Thain D, Tannenbaum T, Livny M: **Distributed computing in practice: the Condor experience.** *Concurrency - Practice and Experience* 2005, **17**(2-4):323-356.

41. Brem RB, Kruglyak L: **The landscape of genetic complexity across 5700 gene expression traits in yeast.** *Proc Natl Acad Sci USA* 2005, **102**(5):1572-1577.

42. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE: **Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks.** 2008, **40**(7):854-861.

43. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L.: **Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors.** *Nat Genet* 2003, **35**:57-64.

44. Lee SI, Dudley AM, Drubin D, Silver PA, Krogan NJ, Pe'er D, Koller D: **Learning a prior on regulatory potential from eQTL data.** *PLoS Genet* 2009, **5**(1):e1000358.

45. **The Saccaromyces Genome Database**. http://yeastgenome.org

46. Zhang B, Horvath S: **A General Framework for Weighted Gene Co-**

Expression Newtork Analysis. *Stat Appl Genet Molec Biol* 2005, **4**(1):Article 17.

47. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks.** *Bioinformatics* 2005, **21**:3448-3449.
48. Kamada T, Kawai S: **An algorithm for drawing general indirect graphs.** *Information Processing Letters* 1989, **31**(1):7-15.
49. The UniProt Consortium: **Ongoing and future developments at the Universal Protein Resource.** *Nucleic Acids Res.* 2011, **39**:D214-D219.
50. Pritchard JK, Stephens M, Donnelly P: **Inference of Population Structure Using Multilocus Genotype Data.** *Genetics* 2000, **155**:945-959.
51. Curtis RE, Wenzel S, Myers DA, Bleecker E, Xing EP: **Population analysis of asthma genome-wide association data using GenAMap.** *Presented at the 61st Annual Meeting of the American Society of Human Genetics* 2011.
52. Wu T T, Chen Y F, Hastie T, Sobel E, Lange K: **Genome-wide association analysis by lasso penalized regression.** *Bioinformatics* 2009, **25**(6):714-721.
53. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res.* 2001, **29**(1):308-11.
54. Valdar W, Solberg LC, Gauguler D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JN, Mott R, Flint J: **Genome-wide genetic association of complex traits in heterogeneous stock mice.** *Nat Genet* 2006, **172**:1783-1797.
55. Huang GJ, Shifman S, Valdar W, Johannesson M, Yalcin B, Taylor MS, Taylor JM, Mott R, Flint J: **High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues.** *Genome Research* 2009, **19**:1133-

1140.

56. Du P, Kibbe WA, Lin SM: **lumi: a pipeline for processing Illumina microarray.** *Bioinformatics* 2008, **24**(13):1547-1548.

57. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17**(6):520-525.

58. Reynolds PR, Allison CH, Willnauer CP: **TTF-1 regulates α5 nicotinic acetylcholine receptor (nAChR) subunits in proximal and distal lung epithelium.** *Respiratory Research* 2010, **11**(175):doi: 10.1186/1465-9921-11-175.

59. Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT, Mouse Genome Database Group: **The Mouse Genome Database (MGD): a premier model organism resource for mammalian genomics and genetics.** *Nucleic Acids Res* 2011, **39**(suppl 1):D842-D848.

60. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R: **QuickGO: a web-based tool for Gene Ontology searching.** *Bioinformatics* 2009, **25**(22):3046-6.

61. Stuart PM: **Major Histocompatibility Complex (MHC): Mouse.** *eLS* 2010.

62. Kumanovics A, Takada T, Lindahl KF: **Genomic Organization of the Mammalian MHC.** *Annual Review of Immunology* 2002, **21**:629-657.

63. Chen X, Kim S, Lin Q, Carbonell JG, Xing EP: **Graph-Structured Multi-task Regression and an Efficient Optimization Method for General Fused Lasso.** *CoRR* 2010.

64. Friedman J, Hastie T, Tibshirani R: **Regularization paths for generalized linear models via coordinate descent.** *J Stat Softw* 2010, **33**(1):1-22.

Figures

Figure 1 - GenAMap's genome browser

GenAMap provides a simple genome browser that allows analysts to explore the mutation marker data that they load into GenAMap. SNPs are represented by green circles across the genome. Analysts can use these SNPs to directly link to external databases, such as SGD or dbSNP. SNP labels are displayed as the analyst hovers over the SNPs.

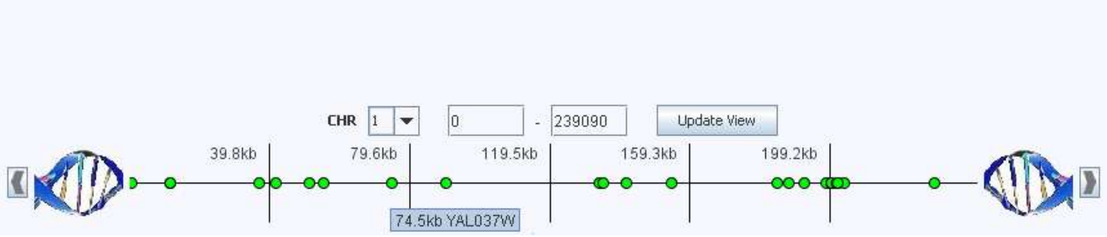


Figure 2 - GenAMap trait overview exploration

GenAMap provides an overview of gene and trait networks to aid analysts in their exploration of the networks. Here, we present a genetic network generated from the yeast data. The network has been clustered by hierarchical clustering, and twenty highly connected gene modules have been automatically identified by GenAMap (outlined in color). As the analyst clicks in these different modules, an information display appears to report the GO and eQTL enrichment of the genes that belong to the particular module.

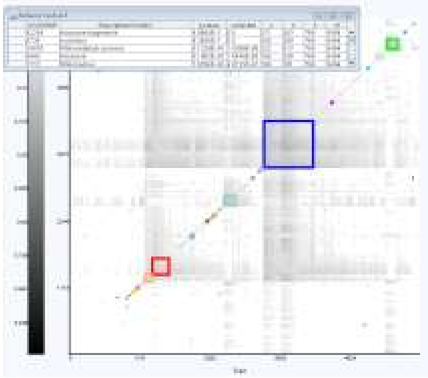


Figure 3 - Using GenAMap to explore genetic networks
We demonstrate using GenAMap visualizations to explore a genetic network. A) From the overview of the network, the analyst can see the different gene modules in the network. B) The analyst zooms into a module of interest in the network. C) The analyst switches to a node-edge representation of this sub-network and adjusts the edge threshold, layout, and labels. D) The analyst uses GenAMap to link directly to external data sources for more information.

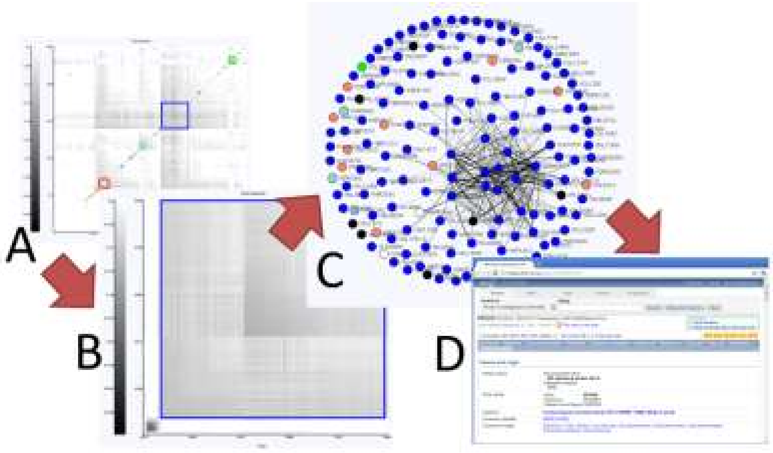


Figure 4 - GenAMap overview of association results
GenAMap provides a heat chart visualization to explore the results from an eQTL association analysis. SNPs are plotted along the *x*-axis and genes are clustered along the *y*-axis. This view allows the analyst to explore the overview of the results. For example, in these results from running TreeLasso on the yeast data, many SNPs are associated with all the genes in a gene module, and some gene modules are associated with many different SNPs in different genomic locations.

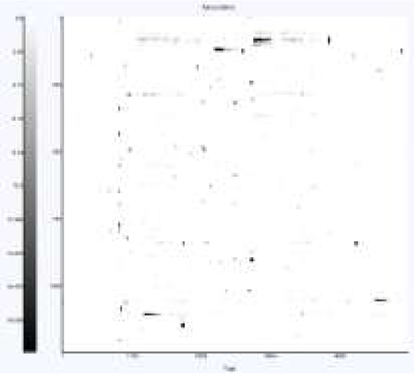


Figure 5 - Using GenAMap to find eQTLs in yeast data
GenAMap provides many tools for analysts to explore association results while using the structure of the data to guide the discovery of associations. We demonstrate some of these tools. A) The analyst can zoom into certain regions to see finer detail of the SNP-trait associations. This panel is a zoomed-in region from Figure 4. B) The analyst switches to the JUNG view to explore the genes associated with the region and perform a GO enrichment test. C) The analyst colors the genes by strength of association to the genomic region. D) The analyst selects up to ten interesting genes (salmon colored) and views the Manhattan plot of associations from these genes across the genome. E) The analyst zooms into interesting regions in the genome view. F) The analyst can switch between association tests for further insight into the associations.

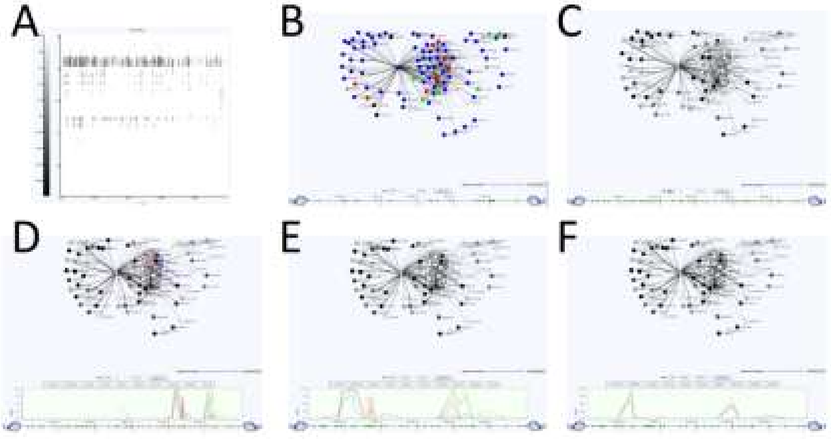


Figure 6 - Analyzing population structure in GenAMap

GenAMap provides an interactive view for analysts to explore population structure. Population assignments are plotted by individual by Eigenvalue. The analyst can adjust the 2D plot to adjust between the first five Eigenvalues. Here, we present the results from a population analysis on the asthma data. Self-reported race is plotted according to the first two Eigenvalues. The plot shows clear separation between the populations.

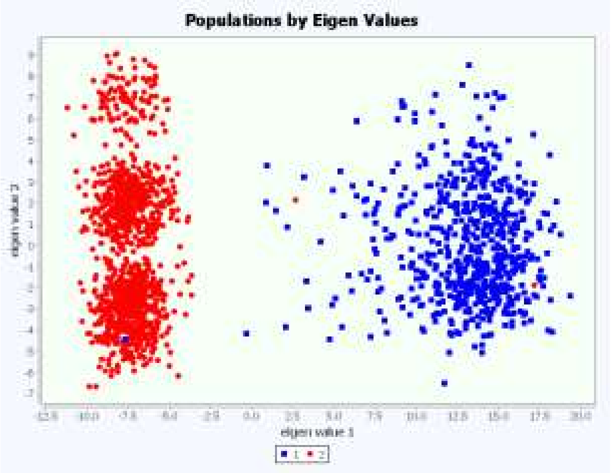


Figure 7 – Interactive Manhattan plot for population data

GenAMap provides an interactive Manhattan plot for exploring associations in population data as well as results between different tests. For example, here we show the results of two tests looking for genetic associations to asthma. The blue lines represent population 1 (African American population) and the red lines represent population 2 (non-Hispanic white population). Different tests are represented by different shapes in the plot.

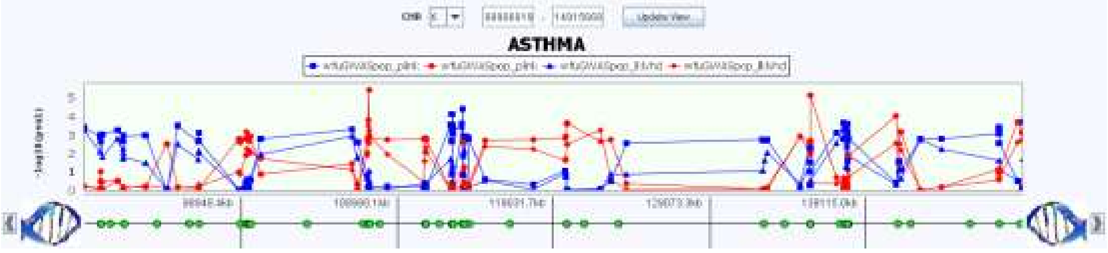


Figure 8 – Frequency distribution of asthma trait by genotype

When exploring SNP-gene associations, GenAMap provides links to tools that allow the analyst to explore the discovered association. For example, consider a case where the analyst considers a discovered SNP-trait association. The analyst can query dbSNP to find out information about the SNP, and the analyst can use GenAMap to visualize the frequency distribution of the trait by genotype.

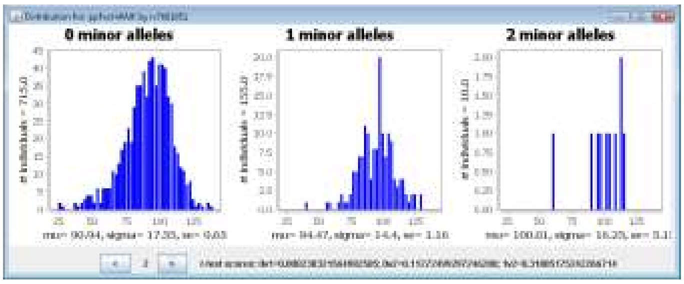
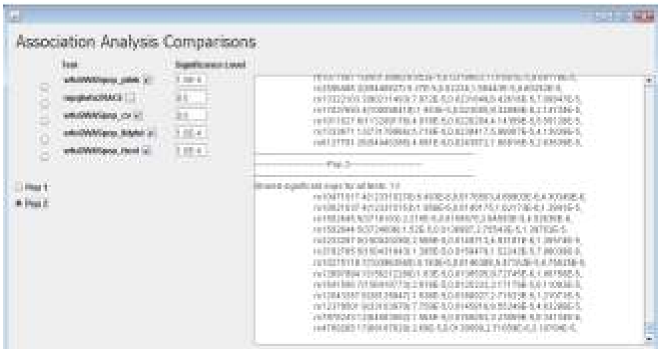


Figure 9 – Comparing association tests

GenAMap allows analysts to compare association results across tests. In this dynamic query tool, the analyst can select which tests to include in the comparison and the significance level of each test. They can see which SNP-trait associations are significant across all tests, and also which associations are unique to any given test.


Association Analysis Comparisons

Test

Significance Level

Figure 10 – Mouse gene network analysis

We used GenAMap to create gene-gene networks from the expression data for each tissue. GenAMap finds the top 20 connected modules and GO and eQTL enrichment for each module. Here, we show the gene-gene network generated using the hippocampus gene expression data.

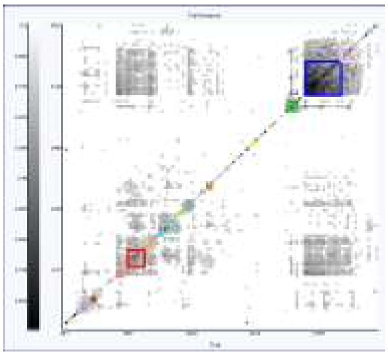


Figure 11 – eQTLs found in hippocampus tissue
We used GenAMap to find SNP-gene associations in the hippocampus gene expression data using GFlasso. In this figure, we show the overview of the results in GenAMap. This is a heat chart representation of the associations, where SNPs are represented along the *y* axis and the clustered genes are represented along the *x* axis. We have zoomed into the section of the gene graph where there are the most associations. We note an eQTL hotspot (represented by a horizontal line of associations).

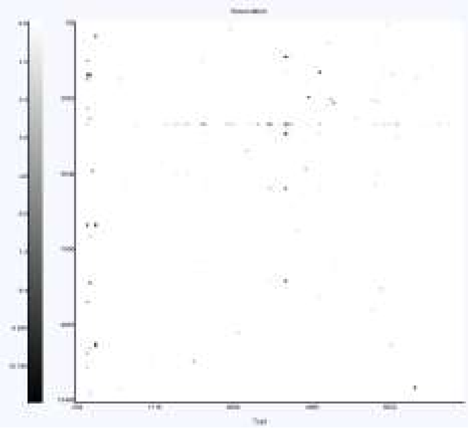


Figure 12 – Association of axon genes to chromosome 14
We found that rs8244120 on chromosome 14 was associated with 140 genes enriched for *cell projection*, implying function in neuronal axons. Here, we show 22 of these

genes in GenAMap's node-link view, colored by the strength of association to rs8244120. White genes are strongly associated and black genes are weakly associated (gray is intermediate). We found that some of the genes were also associated with another SNP on chromosome 14 (shown) and some of the genes were associated with a SNP on chromosome 18 (not shown).

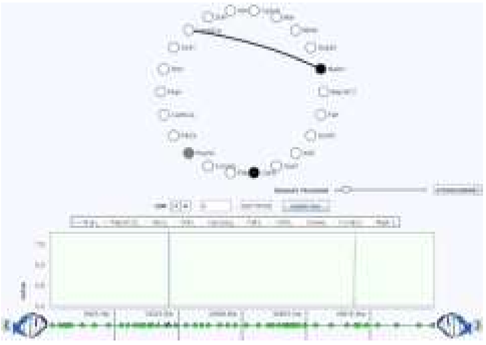


Figure 13 – Overview of three way GFlasso-gGFlasso association analysis
We show the overview of the trait-network and gene-network from GenAMap for the GFlasso-gGFlasso analysis; associations are not shown. In this visualization, circles represent groups of genes, associated to the same regions in the genome. Hexagons represent traits. The edges between genes or between traits represent the connections in the gene or trait network. In this data, we note that there are very few edges between gene groups. The largest gene group is the teal group, representing genes associated with the eQTL hotspot on chromosome 14. The trait network consists of small sub-groups of related traits.

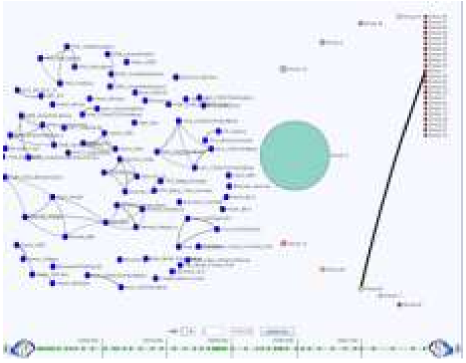


Figure 14 – Gene-trait associations for genes associated with chromosome 14
We used GenAMap to explore the joint three-way associations corresponding to the eQTL hotspot on chromosome 14. We removed all other gene groups and then expanded the group to see the individual genes (squares). We filtered out all traits without associations to these genes.

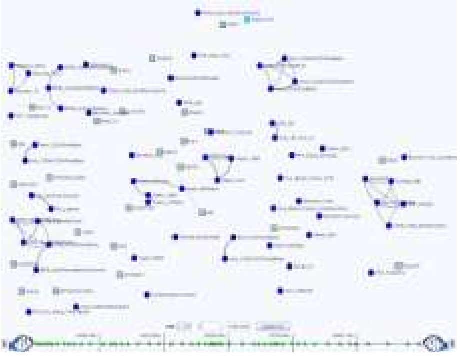


Figure 15 – Joint SNP-gene-trait associations from chromosome 14
We found a subnetwork of traits and associated genes involved in brain function. These genes were also associated with the overlapping eQTL hotspots on chromosome 14.

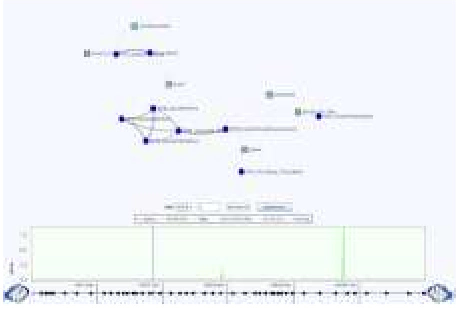
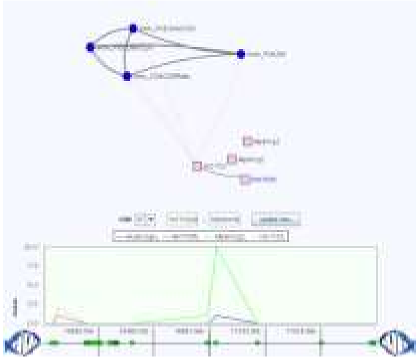


Figure 16 – Immunity associations from chromosome 17
We found a small group of genes associated with the H2 region on chromosome 17. We also found that these genes were associated with a subset of immunology traits.



Tables

Table 1 - Algorithms available to run in GenAMap

Type	Algorithm
Structured Association Mapping	**GFlasso** [63]
	MPGL [16]
	TreeLasso [20]
	AMTL [19]
	gGFlasso [21]
Pairwise Association	**Wald Test** [27]
	Wilcoxon Sum-rank test [42]
	Lasso [64]
	association by population [51]
Network Generation	**Correlation**
	Glasso [29]
	Scale-free network [46]
Tree Generation	**Hierarchical clustering**
Population Assignment	**Structure** [50]
Gene network analysis	**Gene module discovery** [42]

Table 2 – Comparison of gene networks across mouse tissues

Tissue	# genes	% genes shared w/ brain	% genes shared w/ liver	%genes shared w/ lung	% genes shared across tissues	# network edges	% edges shared w/ brain	% edges shared w/ liver	% edges shared w/ lung	% edges shared by all tissues
Brain	7960	100	59.6	78.1	57.6	170982	100	8.9	16.6	4.2
Liver	5879	80.8	100	86.8	78.0	48768	31.1	100	22.9	14.8
Lung	7968	78.1	64.0	100	57.6	105933	26.8	10.5	100	6.8

Table 3 – Gene modules with GO enrichment in the liver network

Module number	# genes in module	eQTL location	eQTL *p*-value	GO Category	GO *p*-value
1	446	11 (4877160)	1.47E-57	mitochondrion	3.80E-04
2	104	17 (61151939)	6.10E-07	catalytic activity	1.96E-04
4	201	14 (9353843)	7.42E-114	ion channel activity	2.02E-04
5	97	19 (20354841)	3.38E-31	mitochondrion	1.11E-13
8	89	17 (61151939)	1.81E-07	cytoplasm	3.73E-04
12	45	13 (56818025)	2.56E-10	regulation of gene expression epigenetic	5.59E-05
14	22	1 (76152963)	8.61E-07	generation of metabolites and energy	6.28E-04
15	34	19 (21138174)	4.18E-10	ER	7.08E-04
20	20	6 (42868138)	1.31E-11	nucleic acid binding	2.34E-04

Table 4 – GFlasso-gGFlasso associations matching previous results [54]

SNP	Chr	Gene	Trait
rs13459079	4	*C1qb*	Alkaline phosphatase
rs4226889	7	*Nsmce1*	Weight at 6 weeks
rs3718803	11	*Pcdh20*	Aspartate Transaminase
rs3023277	11	*Psmb6*	Mean corpuscular haemglobin
rs6326787	11	*Gabrd*	Startle response
rs6380524	11	*Ube2g1*	Startle response
rs4229111	11	*Mpp3*	Startle response
rs1348295	17	*H2-T22*	CD4+/CD8+
rs1348295	17	*H2-T22*	%CD4+/CD3+
rs1348295	17	*H2-T22*	%C8+ cells