GenAMap: Visualization Strategies for Structured Association Mapping

Ross E Curtis[*]
Joint Carnegie Mellon-University of Pittsburgh
PhD Program for Computational Biology

Peter Kinnaird[†]
Human Computer Interaction Institute
Carnegie Mellon University

Eric P Xing[‡]
Machine Learning Department
Carnegie Mellon University

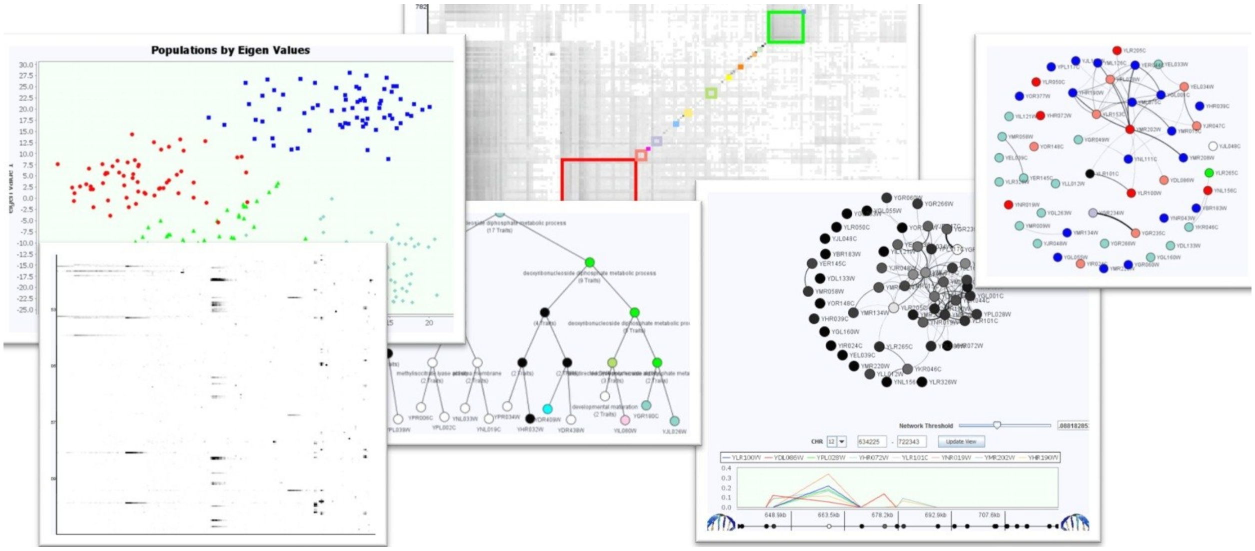


Figure 1. GenAMap is a visual analytics system for structured association mapping. It incorporates several different visualizations in a novel way to lead biologists to relevant SNPs and their associated traits. GenAMap enables biologists to explore the structure of the genome and trait data while exploring association strengths.

Abstract

Association mapping studies promise to link DNA mutations to gene expression data, possibly leading to innovative treatments for diseases. One challenge in large-scale association mapping studies is exploring the results of the computational analysis to find relevant and interesting associations. Although many association mapping studies find associations from a genome-wide collection of genomic data to hundreds or thousands of traits, current visualization software only allow these associations to be explored one trait at a time. The inability to explore the association of a genomic location to multiple traits hides the inherent interaction between traits in the analysis. Additionally, researchers must rely on collections of in-house scripts and multiple tools to perform an analysis, adding time and effort to find interesting associations. In this paper, we present a novel visual analytics system called GenAMap. GenAMap replaces the time-consuming analysis of large-scale association mapping studies with exploratory visualization tools that give geneticists an overview of the data and lead them to relevant information. We present the results of a preliminary evaluation that validated our basic approach.

Keywords: structured association mapping, visual analytics, eQTL analysis, genome-wide association studies

Index Terms: H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

1 Introduction

Understanding the specific interactions between DNA, genes, and traits holds the promise of innovative treatments for many diseases. Association mapping is a popular strategy for examining the complex relationships between DNA and genes. Although machine learning has provided several new, powerful tools for unraveling these connections, the output of today's machine learning algorithms is itself a sea of data, challenging to analyze. We have developed an integrated visual analytics system to aid biological researchers in understanding and analyzing this data.

The motivation for association mapping comes from the central dogma of biology – DNA codes for mRNA, which is expressed as a gene, and then 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. Although much of the human genetic sequence is identical across individuals, there are many places in the genome where the sequence has been mutated, causing a genetic polymorphism between individuals in the population. If the DNA sequence is thought of as a string made up of 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 normal function of a cell,

[*] e-mail: rcurtis@cs.cmu.edu
[†] e-mail: kinnaird@cs.cmu.edu
[‡] e-mail: epxing@cs.cmu.edu

some SNPs can have a much larger effect. The inheritance of SNPs that turn off important genes, or change the coding sequence of important genes, can interact with other genes to lead to disease. Many human diseases or syndromes have a genetic component, and successful association mapping studies have found mutations in the genome associated with human diseases such as cancer [1], diabetes [2], and Alzheimer's disease [3].

As association mapping finds SNPs associated with disease it leads to insight into disease prevention, acquisition, and progression. Association mapping strategies can be divided up into two types, although they are similar in their approach. One type of association study is an expression quantitative trait locus (eQTL) study [4]. In an eQTL study, genome-wide SNP data is collected along with gene expression data for thousands of genes. Gene expression data indicate the expression level, or amount, of each gene in a cell. eQTL studies allow researchers to find associations between SNPs and the expression of genes. In a genome-wide association study (GWAS), genome-wide SNP data is also collected, but the association analysis is performed against one or more clinical traits, such as asthma or diabetes [5]. Although the two strategies are similar, eQTL studies have the promise of linking changes in the genome to the function of individual cells, while GWAS look primarily at genetic polymorphisms that affect the entire organism. Either approach can lead to insight into how biological systems can be affected by small changes in the DNA.

Despite the promise of association studies, however, many SNPs that have been found in association studies only explain a small part of the genetic component of the disease. When associated SNPs only have a small effect, this suggests that other important contributing SNPs have not been found and that the heritability of the disease is not fully explained [5]. This so-called "missing heritability" is a challenge facing the association mapping community today. One approach to overcome the problem is *structured association mapping*. Structured association mapping is a machine learning approach that leverages structure in the data in order to enhance the discovery of weaker signals. Initial studies have suggested that structured association mapping could lead to increased insight and greater statistical power [6,7]; however, there remain several barriers before such algorithms can be widely used.

Structured association mapping methods are more complex than simple statistical methods normally employed by biologists, thus requiring greater specialization to run the method and interpret the results. These methods generally output a raw matrix of values representing associations between the genome and the traits. For a simple organism like yeast, a dataset will generally include a few thousand SNPs and up to six thousand gene measurements. For a human dataset, there might be hundreds of thousands of SNPs and over ten thousand gene expression measurements. These results are typically analyzed using command line tools and in-house scripts, which limits their availability to many geneticists.

Training aside, it remains a difficult task to identify relevant signals from a vast amount of output involving hundreds of traits across a genome-wide scan of SNPs. Analysts must manually identify queries and code up their own implementation to find patterns in the data and explore specific interactions between a trait and the genome. Current visualization software available only allows for the exploration of association data on a trait-by-trait basis, which is not sufficient to explore large datasets with thousands of traits. In this paper, we contribute a new visual analytics system called GenAMap. Our work with GenAMap has been motivated by the promise of structured association mapping strategies and the need for visualizations to explore the results from these machine learning algorithms. We present a comprehensive system for structured association mapping analysis including four visualizations that allow genetics researchers to explore results from structured association mapping algorithms. GenAMap is specifically designed to give geneticists an overview of the results then lead them to specific gene-genome interactions. Once a geneticist has used GenAMap to identify a significant interaction, they can explore the interaction in the tool that leads them to external links to biological databases such as UniProt [8] and dbSNP [9]. GenAMap specifically aids in the exploration of association mapping results through the integration of multiple views so that an analyst can explore the structure of the traits while considering their association to the genome.

We have conducted a preliminary qualitative user study to assess the utility of our visualization techniques and to get feedback on the overall system. The results from this study suggest that GenAMap represents a dramatic improvement over current practice, saving time and effort in analysis.

The outline of the paper is as follows: we first review the related work for association studies, structured association mapping methods, and the promise of visualization in biological problems. We present GenAMap through a discussion of a typical association analysis using the tool, followed by two case studies on real datasets. We report results from a preliminary user study that investigates the utility of the visualizations used in GenAMap and finish with a discussion of our conclusions and future work.

2 RELATED WORK

Today's most advanced computational biology studies commonly employ machine learning to identify patterns in the data. Due to the vast amount of data produced by these algorithms and the sparseness of useful output, new strategies are required to help biology researchers identify the links between DNA and the genes that could eventually produce new treatments for diseases.

Although machine learning has carried us well so far, we believe that the next steps strongly indicate that visual analytics, which combines algorithmic analysis with visualization strategies to take advantage of the analytic capabilities of *both* machines and people, will advance the field still further. Visual analytics are best employed when analysts need to explore their data, understand the overall structure of the data, discover weak patterns most easily recognized by humans, and retain the ability to perform detailed analysis [10].

Our inquiry focuses on structured association mapping analyses for three primary reasons: first, they hold significant promise for impactful outcomes, second, they employ very large amounts of sparse data (classically suggesting the likely fruitfulness of a visualization approach), and finally, the data often conform to easily visualized structures that might support cognition by favoring exploration and investigation over direct query.

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 – immediately suggesting a visualization strategy following Shneiderman's well-known mantra: overview first, zoom and filter, details on demand [11].

2.1 Current methods for association mapping

In a typical association mapping study, hundreds of thousands of SNPs are collected for a number of individuals. Clinical traits or gene expression measurements are also collected. Association algorithms are generally run one gene or trait at a time against genome-wide SNPs. These results are then explored trait-by-trait through a Manhattan plot (a logarithmically scaled scatter plot designed to highlight small variations from a normal range) or a summary table. For studies with only a handful of clinical traits, as is common in many GWAS, this approach can work quite well; there are excellent tools such as WGAViewer [12] and

LocusZoom [13] that facilitate exploring these results. However, these visualization strategies fall quite short in larger studies with thousands of traits or genes, as in eQTL studies or GWAS studies on many related phenotypes.

To date, there are very few tools built to explore the association signals from the genomic data to thousands of traits. eQTL Explorer [14] is one such tool that represents associations as arrows along chromosomes that the analyst can browse through. This approach makes it difficult to understand the overall patterns in association and does not scale well to larger datasets with many genes. Additionally, the relationships between genes, and the relationship between a gene and multiple loci, are obscured in this visualization. eQTL Viewer [15], on the other hand, provides a broad over-view of the associations with a heat map view, however, this display does not allow the analyst to explore interactions between genes, and the strength of eQTL signals is not available to the analyst. Thus, all published tools are limited in their capacity to explore the results from an eQTL study and are not extendable to a GWAS study with clinical traits.

2.2 The promise of structured association mapping

Structured association mapping algorithms are a new strategy to association mapping that can identify signals that could not be discovered with previous, simple methods. There are currently two types of structured association mapping algorithms available: those that leverage structure across traits, and those that leverage population structure.

Mutations in the genome often do not affect just one gene or trait but multiple correlated traits. When looking for these SNPs, we can use the information from the correlation structure of the traits/genes in our analysis. For example, GFlasso [6] is a regression approach that finds associations from SNPs to a cluster of highly related traits or genes. Traits or genes that are highly correlated are likely to be under the regulation of the same genetic loci, and incorporating this information in the analysis can lead to greater statistical power and the detection of weaker signals [6]. A similar approach, the TreeLasso [16] takes advantage of the hierarchical structure of the trait data clustered as a tree.

Another important dimension of the data is the population structure of the individuals in the study. Different subpopulations will respond to SNPs in different ways. For example, an Asian population might respond to the loss of an important gene differently than an American population due to the differences in the genome between the two populations. An algorithmic approach that takes advantage of this structure in the data is the MPGL algorithm [7], which uses the information in the different populations to find associations, while still recognizing their differences.

Whether exploring the results of a population-structured association analysis, or a trait-structured association analysis, the structure of the data can provide further evidence to the geneticist of how a SNP might be involved in the association. For example, seeing the relationship between many correlated genes that are all associated to the same SNP could provide evidence on how that SNP affects all the traits; this information would not be available in a one-by-one visual analysis. This is especially important in eQTL analysis, and has been recognized by the figures in recent papers that show the correlation structure of the traits, colored by association [17,18]. These types of figures show the potential of a visualization system that guides the analyst to relevant signals through the structure of the data itself.

Despite the improvements that we have seen in performing structured association mapping studies with thousands of SNPs and thousands of traits, in order to adequately explore the data, researchers still have to rely on in-house scripts, command line tools, and customized software [19]. The reliance on these types of tools limits the number of individuals who can perform these analyses to only the very specialized, and limits the insight that can be gleaned from exploring the structure of the data itself.

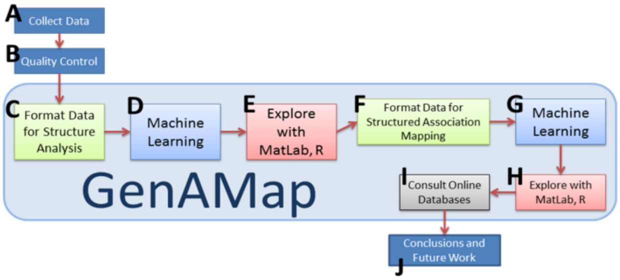


Figure 2. GenAMap is integrated into the workflow of a structured association mapping analysis.

In Figure 2, we show a typical workflow investigators would follow in a structured association mapping study. Data would be collected for SNPs and phenotypes (whether genes or traits). After quality control and preprocessing steps, the data would be formatted in order to run machine learning algorithms on the data to generate structure. Investigators would explore the structure using tools like Matlab or R in order to determine what structured association analyses were appropriate for the data and to initially identify interesting patterns. They would then reformat their data and run machine learning algorithms to find the structured association results. They would then write their own scripts or use Matlab and R in order to explore the data, observer interesting patterns, and find specific associations that were interesting to them. They would investigate the details of these associations using online databases and a handful of other tools in order to understand them. This analysis would lead the investigators to conclusions about the data and lead to other hypotheses to be tested in future studies. GenAMap's integrated system is designed to guide researchers through steps C through I (as shown in Figure 2).

2.3 The promise of visualization in biology

With improved technology, high-throughput methodologies are generating huge biological datasets. The integration of visualization into the analysis of these large datasets is becoming a popular and effective strategy.

The success of visualization strategies can be seen in many areas of biology. For example, Cytoscape [20] has become an extremely popular application for visualizing biological networks and exploring relationships between genes. By integrating information into a visualization that can handle thousands of genes, Cytoscape has become a standard tool for the analysis of biological networks [20]. In other domains, the recent development of ABySS-Explorer [21] has shown that visualization can enhance the analysis of complex biological tasks like genome assembly through visual representation of the contigs. Another recent approach to visualization, MulteeSum, demonstrated the potential for visualization to help researchers identify spatial and temporal patterns in gene expression data [22]. Thus, although the problems in biology are varied, visualization is proving to be a reliable tool to aid in the exploration of the vast amount of data available.

3 AN ASSOCIATION STUDY WORKFLOW IN GENAMAP

In this section, we will give an overview of GenAMap by discussing a typical structured association mapping workflow using GenAMap as shown in Figure 2. This workflow includes several machine learning and visualization steps; GenAMap is intended to integrate all analytic and visualization tools necessary to complete a structured association analysis.

3.1 Data collection and structure analysis

In an association study, two types of data must be collected from the individuals in the study. Genotype data is collected as SNPs for hundreds of thousands to millions of SNPs, and trait data is collected as gene expression data from a microarray or as clinical trait data collected in the office. Once the data has been preprocessed, it is ready to be imported into GenAMap. GenAMap will then provide tools to create and visualize the structure for both of these data types separately.

GenAMap helps the analyst determine which structured association algorithms will be appropriate by providing visualizations to explore these data. We discuss the specific visualizations in Section 4; the purpose of these visualizations is to help the analyst decide on an appropriate analysis. For example, if the analyst notices strong population stratification in the data, he will want to perform a population analysis to find associations. If the analyst notices that the traits form many highly connected clusters, he will want to use an algorithm such as GFlasso to take advantage of this information. Once the analyst has explored the data and determined the appropriate structured analysis for their dataset, he is ready to run the actual association algorithm.

3.2 Exploring structured association results

GenAMap automatically runs all structured association mapping algorithms on an external cluster using parallelization. Once the results from the algorithms are available, the analyst has a number of tools to explore the associations in the data.

GenAMap supports Shneiderman's mantra: *overview first*, *zoom and filter*, and *details on demand* [11]. First, GenAMap allows the analyst to get an overview of the association results. For example, geneticists can quickly identify whether SNPs are generally association with many traits in the same part of the clustered network. Once the geneticist has gotten a feel for the overall network, they can start to zoom and filter to the most interesting patterns in the data. Geneticists will then identify specific trait clusters that are association with genetic loci. They will then zoom in to see the network and correlation structure of the traits of interest. Finally, once the geneticist has zoomed into a particular interaction, GenAMap provides details on demand to specifically characterize the association. Researchers can query online resources directly from the tool in order to find out the function of key genes in the network, and they can look up SNPs online to find out why they might be associated with the traits of interest.

4 Exemplary Use Cases

We now present two case studies with real data to highlight the novel visualization strategies in GenAMap.

4.1 A case study on a yeast eQTL dataset with trait structure

In the first case study, we demonstrate an analysis in GenAMap using a yeast eQTL dataset [23]. This dataset is generated from a cross between the yeast BY4617 strain and the RM11-1a strain. The dataset has 5637 gene expression measurements and 1260 SNP markers from the two parent and 112 progeny strains.

A dataset from a genetic cross in a model organism like yeast allows us to understand how genetic variants affect gene expression globally. These datasets can also lead an analyst to new regulatory genes in specific biological processes.

4.1.1 The network view

An analyst loads the yeast gene expression and SNP data into GenAMap. The analyst believes that an association mapping algorithm leveraging gene structure might make sense for these data. In order to run this algorithm, the analyst needs to build a gene-gene network. The analyst can also use GenAMap to explore this network to ensure that a structured approach makes sense for these data. The analyst uses GenAMap to create a gene-gene network using a previously defined method [18] and then explores the network using GenAMap's *network view* (Figure 3).

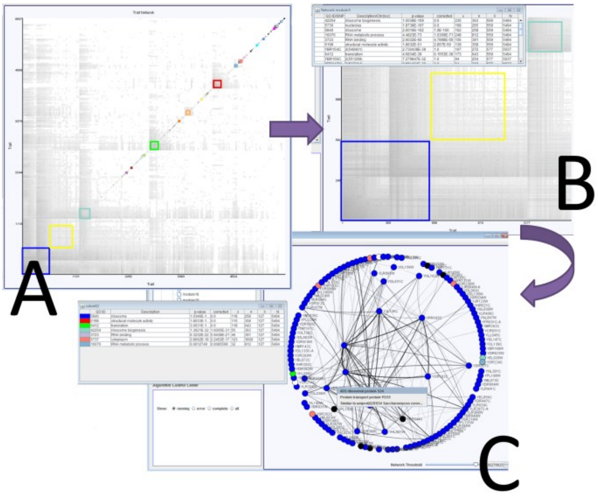


Figure 3. Exploring a gene network in GenAMap. (A) An overview of the entire network, with gene modules identified. (B) Zoomed in regions of the network with GO functional enrichment. (C) Node-edge representation of specific regions in the network colored by GO category.

By default, the network graph is visualized in a hierarchically clustered matrix, showing a heat map of all edges between genes (overview first). For the yeast data it is a 5637x5637 matrix (Figure 3A). GenAMap weights the edges between traits, thus strong edges are shown as dark gray or black in the heat map, weak edges are shown as light gray, and no edge is represented as white. The hierarchical clustering ensures that strongly connected genes are shown next to each other in the heat map. GenAMap also identifies *gene modules* in the network, outlined in color in Figure 3A. Gene modules are regions of the network with many strongly-connected genes. The analyst clicks on the modules in the heat map to view results from a gene-ontology (GO) enrichment test for the module; the GO enrichment test allows the analyst to identify the common function of the different gene modules. Because the gene modules are enriched for common GO categories in this case, the analyst concludes that this gene network is appropriate to use in a structured association analysis to find SNPs associated with functionally coherent groups of genes.

The analyst notices a gene cluster that is enriched for the GO category ribosome (p-value = 2.6e-102). He is intrigued by the low p-value so he zooms into this region of the network in the heat map (Figure 3B). He wants to explore the relationship between these genes, and so he switches to a node-edge view of the first 150 genes in this region. The analyst performs a GO analysis on these genes and finds that this specific set of genes are highly enriched for ribosome (p-value = 8.5e-163). He colors the genes by GO category (Figure 3C) and notices that almost every gene in the module is a ribosome gene. Because these genes are clustered together in the network, the analyst concludes that this module is made up of co-expressed ribosomal genes.

In order to identify the key genes in the network, the analyst uses the dynamic query controls [24] in the network view to adjust the network threshold to add and remove edges. He moves the highest connected genes to the center of the network (Figure

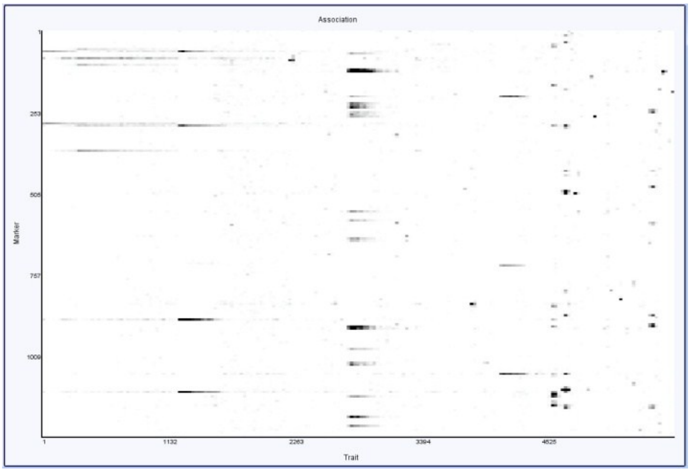


Figure 4. GenAMap gives an overview of the association results through a heat map view where SNPs are plotted on the y axis and genes are plotted on the x axis.

3C). He right-clicks on these genes to look them up in UniProt for further analysis.

4.1.2 The network association view

After exploring the network, the analyst uses GenAMap to run a structured association mapping algorithm, GFlasso, using the SNP, gene expression, and network data. Once the algorithm has completed, the analyst uses the *network association view* to explore the results (Figure 4 and 5).

The network association view is similar to the network view, integrated with the genome view. As with the network view, the analyst can explore the overview of the data, zoom in and filter, and then get details on demand. The network association view incorporates tightly coupled coordinated views [25], allowing the analyst to interactively correlate between SNPs and the network. The analyst will use this view to get a feel for the data and find specific SNP-gene associations for further investigation.

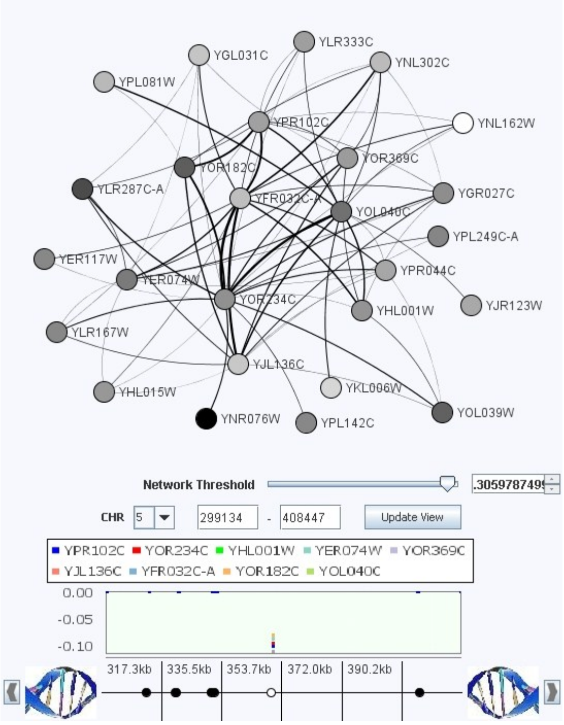
The analyst first considers GenAMap's overview of the association results shown in Figure 4. This heat map shows a matrix of the association values. SNPs are shown along the y-axis, and the genes are shown along the x-axis; the traits have been clustered by hierarchical clustering. In this case, the yeast data associations are represented by a 1260x5637 matrix. Black represents a strong association, and white represents no association. For both this heat map view and the network heat map, the analyst can zoom in and out of the matrix through a series of resolutions. Each resolution is a 200 pixel by 200 pixel matrix; the association results initially displayed to the analyst in Figure 4 is at a resolution where each pixel represents six SNPs and 30 genes. Because the data is inherently sparse, GenAMap colors the pixel by the maximum association value between all SNPs and traits represented. This ensures that the analyst can focus on the signals in the data; the signals are preserved even at lower resolutions.

The analyst looks at the heat map shown in Figure 4, and quickly gains insight into the yeast regulation patterns present in these results. The analyst notices the series of long (and short) horizontal black lines in the matrix. These lines represent associations between a SNP and a cluster of genes. The presence of such patterns indicates to the analyst that gene clusters in the yeast network are associated with a common SNP. Because these lines overlap, the analyst concludes that some gene clusters are associated with multiple SNPs, representing a case where multiple mutations affect the same set of genes.

This view has made it visually obvious which of the gene clusters are associated with multiple genetic locations and approximately where in the genome these association lie. The analyst can now use his knowledge of the gene network he obtained from the network view to zoom into clusters of traits that are associated with different SNPs. Because the analyst has previously identified the 150 ribosome genes, he zooms into the part of the heat map associated with these genes and switches to the node-edge representation of the associations shown in Figure 5. From the heat map view, the analyst knows that these genes are associated with SNPs on two chromosomes and he wants to explore these associations.

In the node-edge representation of the network, the analyst can explore the gene structure of the network while identifying associations. The view is integrated with a simple genome browser where nodes represent SNPs [26] (bottom of Figure 5). The analyst can use this genome browser to switch between chromosomes and zoom into certain chromosomal regions.

In this particular analysis, the analyst colors the nodes in the genome view by their association to the genes in the network. This allows him to identify the white colored SNP on chromosome 5, ignoring the rest of the SNPs in the genome that are not associated with these genes. He adjusts the network threshold on the network to find the highest connected genes. After adjusting the threshold and removing unconnected genes, he finds 25 highly connected genes, shown in Figure 5. The analyst colors these genes by their association to the SNP located on chromosome 5 and also consults the Manhattan plot of these genes. The analyst is able to recognize that all of these genes have some association to this SNP (because they are not colored black), although the signal is much stronger for many of the genes as some are white or light gray and some are colored darker, representing weaker associations. The analyst queries UniProt for information about these genes, and finds that *YER074W* is a gene located near this SNP in the genome. Through this analysis, the analyst gains important information about this SNP and genes



Network Threshold .305978749

CHR 5 ▾ 299134 - 408447 Update View

Figure 5. In the network association view, GenAMap shows interaction between genes, integrated with the association strengths of the genes to SNPs in the genome.

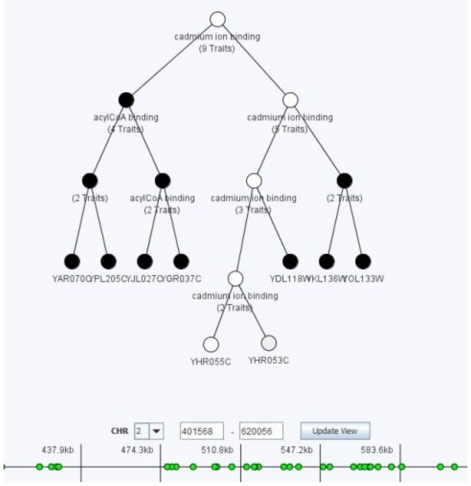


Figure 6. In the association tree view, the analyst explores genes structured as a tree in order to identify functionally relevant branches of the tree that are associated with a genomic region.

associated with this SNP, directing the next steps in his investigation.

4.1.3 Exploring association results through the association tree view

In the previous example, the analyst was interested in associations to gene clusters. Often geneticists will want to explore the associations of a particular SNP or SNP region to find out if the genes associated with a SNP are actually in a gene cluster, or to find the strongest associations from a SNP to genes. GenAMap's *association tree view* allows for these types of explorations (Figure 6).

In the *tree view*, the leaves of the tree represent genes, and other nodes represent the aggregation of genes descending from them. Each non-leaf node is labeled by the number of aggregated genes below the node and by a GO enrichment annotation (if the genes have a significant functional enrichment). By default, the nodes are colored by this GO annotation. The tree view only shows three to eight levels of the tree at a time and allows the analyst to browse through the tree. In the association tree view, the tree view is integrated with the genome view.

An analyst is specifically interested in a genomic region on chromosome 2 (base-pair 560000) in the GFlasso results. From the association tree view, the analyst browses to this genomic location, selects several SNPs in the region, and colors the tree by association to these SNPs. Each node in the tree is then colored by strength of association to these SNPs, white represents a strong association to the genome location and black represents no association. As seen in Figure 6, a non-leaf node is colored by the strength of the strongest association of all the traits it represents.

The analyst is interested to find the genes with the strongest associations to these SNPs. From the root of the tree, the analyst follows the white nodes to browse down the tree until he finds the genes (leaves, shown in Figure 6). Interestingly, this part of the tree only had two genes associated to this genomic locus. The analyst looks these genes up in UniProt through GenAMap's links to find out what they are and why they might be affected by mutations in this genomic location. He also uses the genome browser to link to the SNP location in the *Saccharomyces* Genome Database (SGD) [27]. Further exploration in the tree will allow the analyst to find associations between the genes and SNPs, identify whether other related genes in the tree are also associated, and discover the common GO enrichment of associated branches in the tree.

4.2 A case study on a mouse GWA dataset

The second dataset that we consider is a mouse dataset [28]. This dataset has measurements for 179 clinical traits and 12546 SNPs for 269 mice. Using a dataset with clinical traits and SNPs allows geneticists to identify SNPs that are associated with a particular disease trait of interest. In this example, we will focus on traits related to asthma in mice.

4.2.1 The population structure view

After the analyst loads the data into GenAMap, he is ready to begin his analysis. From the analyst's knowledge of the data source, he believes that there is population structure in the data. He explores the population structure through the *population structure view* (the population structure view is integrated into the top half of Figure 7) after running machine learning to generate population structure [29] and eigenvalues from the data. Individuals are plotted according to their eigenvalues, and colored according to population assignment. The analyst dynamically colors the plot for different numbers of populations. He can see the number of individuals assigned to each population using a pie chart. The analyst finds that the mice split up into four distinct subpopulations across the first five eigenvalues.

4.2.2 The population association view

In the *population association* view (Figure 7), the analyst explores the results from MPGL on the mouse data with four populations. GenAMap integrates the population structure view, the network view, and the genome view to help the analyst explore these associations. The analyst has explored the overall network and identified seven traits related to asthma for further exploration.

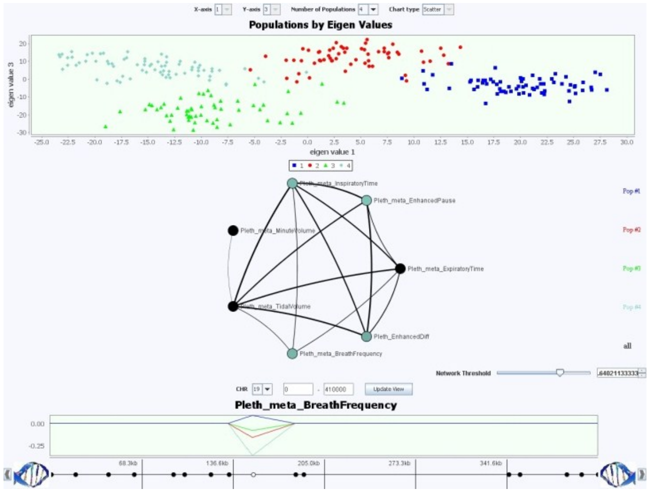
The analyst wants to find the SNPs associated with these traits. He colors the genome by association to these traits, and identifies a SNP on chromosome 19 that is strongly associated to at least one trait. He selects this SNP, able now to ignore the rest of the genome, and colors the traits in the network that are associated to this SNP. Each trait is colored by the color of the population with the largest beta value (association). The analyst finds four asthma traits associated with this SNP, with the strongest association in each case being the association to population #4. The analyst investigates the association of each of these traits one-by-one by adding the Manhattan plot to the genome view. For the trait "breath frequency," the analyst finds that population #4 and #1 are strongly associated with this SNP, more than population #2 and #3. The analyst investigates this association by linking to dbSNP through GenAMap. He suspects that this SNP on chromosome 19, or one close-by, plays a role in asthma in mice.

5 IMPLEMENTATION

GenAMap is implemented using Java SE. GenAMap visualizations are built by the integration and expansion of three visualization toolkits publically available [30, 31, 32].

6 USER STUDY

We performed a preliminary qualitative user study to assess the utility of our visualization techniques and to get feedback on steps we could take to improve the visualizations. We recruited PhD students and post-docs with specific research interests in genetics from two universities. We had eight volunteers participate in the study: seven PhD students and one post-doc. There were four male participants, and four females. All of the participants are involved in genetics research with an emphasis on machine learning development. We assessed the level of expertise in association mapping of the candidates based on three criteria: 1) self-rated expertise in association mapping, 2) self-reported participation in an association mapping project before, and 3) the



X-axis 1 Y-axis 3 Number of Populations 4 Chart type Scatter

Network Threshold .64021133333

CHR 19 0 - 410000 Update View

Figure 7. The population association view is an integrated view enabling the exploration of association strengths across different populations.

ability to explain what an eQTL study entails. Using these criteria, our participants consisted of four experts in association mapping (met two or more criteria) and four non-experts.

Each participant met with an investigator privately in a standard office space. The investigator guided them through five different tutorials. GenAMap was run on a standard desktop-computer with a 22-inch screen. Participants were encouraged to think-aloud as they used GenAMap; they were given semi-structured tasks to explore the tools with guidance and on their own. We asked for verbal feedback at each stage of the study. Sessions lasted about an hour; the investigator took notes of all comments throughout the session, and users filled in a survey upon completing the evaluation.

6.1 Survey results

Our post-survey had twelve questions where the user had to rate the software on a scale from 1-to-5. Overall, the users reported that GenAMap allowed them to explore association results better than other tools (average score 5.0) and that GenAMap allowed them to get an overall feel for the structure in the data (4.75). They all agreed that GenAMap lead to insight that was not available using other tools (4.71), and that they would recommend GenAMap to other researchers (4.75). The lowest scores from the survey were in regards to the usability of the system. While the utility scores just mentioned were high, the participants did not agree as strongly that GenAMap was easy to learn (average score 3.75), or that the visualization strategies were always easy to understand (4.0). However, the lowest rating that we had from any user was a score of 3 for any of these questions.

In the free response part of the survey, the users were prompted for the most useful part of GenAMap that they explored. Four of the users specifically mentioned the incorporation of outside data, including GO category analysis and external databases. In fact, when asked specifically about external links, seven of the participants responded very positively. When asked what views led to the greatest insight, seven participants specifically mentioned a visualization strategy that incorporated multiple views that allowed them to explore the association results between the genome and the traits represented in some structure.

6.2 Think-aloud comments and results

We presented three integrated views to explore the results of association analysis to the users: the association tree view (Section 4.1.3), the population association view (Section 4.2.2), and the network association view (Section 4.1.2). None of the users had seen association results presented in such a way previously.

All the users reported that they liked each of the visualizations. The tree view was met with some reservation, but in the end users found that they could think of different uses for it. The users gave us many ideas of different queries they wanted to be added to explore the tree, and one user specified that the tree view was their favorite visualization technique. All the users mentioned that GenAMap was an improvement over how they would normally do these types of studies. For example, six of the users mentioned that the tree view was a more convenient way to explore the results than MATLAB or another command line interface.

Users also liked the integration of the different views of the structure of the data. Five users specifically mentioned that they would have had to use a combination of tools or done the work by hand in order to complete a similar analysis. One user remarked, “By myself I would have to go back and forth between the human genome browser and a network viewer. It is really nice that it is integrated into the software.” Five of the users specifically mentioned that they liked being able to explore the genome and the gene-gene network in the same integrated tool. One user said, “The ability to interact with the network and the genome is excellent,” and another commented that “it really puts it into perspective.” Five users specifically mentioned that using GenAMap was easier, more convenient, and saved time by allowing them to more systematically explore the data.

Many of the feature requests the users offered were related to documentation and exporting data. Five users wanted better documentation and links incorporated into the software so they could more easily identify what different plots and charts represented. Additionally, five users mentioned that they would have liked the ability to export data from GenAMap for further analysis using more specialized tools.

7 Conclusions and Future Work

Structured association mapping is a powerful machine learning approach to discover weak signals in association mapping datasets. Because the results from these analyses are complex, adequate visualization techniques are necessary to explore the structure of the traits and genome to find important association signals. We have developed visualizations and integrated machine learning into a visual analytics system called GenAMap, which enables analysts to explore the results from structured association mapping studies. GenAMap’s visualizations give analysts an overview of the results and lead them to focus on important signals. We demonstrated these visualizations through case studies, and have provided further support through a user study.

The results from our user study suggest that GenAMap made analyzing results from structured association mapping easier and saved time, while providing additional insight. The users in the study liked how GenAMap incorporated multiple views to provide a feel for the structure of the genome and the traits while exploring the associations. They felt that the coordinated visualization helped to put things into perspective and avoided unnecessary and awkward integration of specialized tools. Users also felt that GenAMap helped them to focus their attention on important associations and that GenAMap was an improvement over the command line scripts they normally use. GenAMap also has several resources to link to outside information such as GO annotation, SNP pages, and UniProt. These proved to be a key feature in GenAMap to aid researchers in the analysis of the data. GenAMap was able to help the users focus their attention on the important signals, and then quickly direct their attention to the outside sources that could explain the signals.

Despite the improvement that GenAMap has over current applications, we feel that there is still room for improvement.

Specifically, we plan to continue to develop GenAMap to add more links to outside information, as requested by the users. Additionally, even though GenAMap incorporates much of the pipeline for association analysis, the participants in our study suggested that they need to export the data for additional specialized analysis. We plan to incorporate this feature into the tool in future releases. Additionally, we plan to work to provide more legends, keys, and consistency to the tool based on user feedback. Although users were able to understand the visualization strategies in GenAMap, many felt that with a few more legends, color bars, and tooltips, the tool would be easier to pick up and use without consulting documentation or tutorials.

Our experience with GenAMap therefore suggests these five general rules for building an adequate biological visualization system: 1) the ability to focus the user's attention on the important information in a complex data sets, especially large arrays of multi-dimensional data, 2) the ability to coordinate multiple views when analyzing connections between different data types, 3) the ability to link out directly from the tool to outside information and biological databases in order to strongly integrate into existing work flows, 4) the ability to export intermediate results for further specialized analysis, and 5) intuitive displays with legends, tooltips, and color bars to enable the user to understand the data as it is presented to them. We hope that these guidelines will prove useful in the development of future biological visualizations.

We further expect that the development of GenAMap will make the analysis of structured association mapping available to more genetics researchers by moving the analysis away from command line scripts into a visual system where users can explore associations, structure, and small-scale interaction.

Acknowledgments

This work is supported by DARPA [Z931302]; NSF [DBI-0640543, IIS-0713379]; NIH [1R01GM087694, 1R01GM093156, 0015699]; and an Alfred P. Sloan Research Fellowship to EPX.

References

[1] M Yeager et al., "Identification of a new prostate cancer susceptibility locus on chromosome 8q24," *Nature Genetics*, vol. 41, pp. 1055-1057, Sep. 2009.

[2] H Yaguchi, K Togawa, M Moritani, and M Itakura, "Identification of candidate genes in the type 2 diabetes modifier locus using expression QTL," *Genomics*, vol. 85, no. 5, pp. 591-599, 2005.

[3] S C Waring and R N Rosenberg, "Genome-Wide Association Studies in Alzheimer Disease," *Arch Neurol*, vol. 65, no. 3, pp. 329-334, 2008.

[4] Y Gilad, S A Rifkin, and J K Pritchard, "Revealing the architecture of gene regulation: the promise of eQTL studies," *Treds Genet*, vol. 24, no. 8, pp. 408-145, 2008.

[5] R A Manolio et al., "Finding the missing heritability of complex disease," *Nature*, vol. 461, pp. 747-753, 2009.

[6] S Kim and E P Xing, "Statistical estimation of correlated genome associations to a quantitative trait network," *PLoS Genet*, vol. 5, no. 8, p. e1000587, 2009.

[7] K Puniyani, S Kim, and E P Xing, "Multi-population GWA mapping via multi-taks regularized regression," *Bioinformatics*, vol. 26, no. 12, pp. i208-i216, 2010.

[8] The UniProt Consortium, "Ongoing and future developments at the Universal Protein Resource," *Nucleic Acids Res.*, vol. 39, pp. D214-D219, 2011.

[9] S T Sherry et al., "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308-11, 2001.

[10] S Card, J Mackinlay, B Shneiderman, and M Kaufmann, *Readings in Information Visualization*., 1999.

[11] B Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," in *Proc 1996 IEEE Visual Languages*, Boulder, CO, pp. 336-343, 1996.

[12] D Ge et al., "WGAViewer: Software for Genomic Annotation of Whole Genome Association Studies," *Genome Res*, vol. 18, no. 4, pp. 640-3, 2008.

[13] R J Pruim et al., "LocusZoom: regional visualization of genome-wide association scan results," *Bioinformatics*, vol. 26, no. 18, pp. 2336-2337, 2010.

[14] M Mueller et al., "eQTL Explorer: integrated mining of combined genetic linkage and expression experiments," *Bioinformatics*, vol. 22, no. 4, pp. 509-511, 2005.

[15] W Zou, D L Aylor, and Z B Zeng, "eQTL Viewer: visualizing how sequence variation affects genome-wide transcription," *BMC Bioinformatics*, vol. 8, no. 7, pp. doi:10.1186/1471-2105-8-7, 2007.

[16] S Kim and E P Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

[17] P C A Dubois et al., "Multiple common variants for celiac disease influencing immune gene expression," *Nature genetics*, vol. 42, pp. 295-302, 2010.

[18] J Zhu et al., "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks," vol. 40, no. 7, pp. 854-861, 2008.

[19] SD Buckingham, "Scientific Software: seeing the SNPs between us.," *Nature Methods*, vol. 5, pp. 903-908, 2008.

[20] M Smoot, K Ono, J Ruscheiriski, P L Wang, and T Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431-432, 2011.

[21] I Birol et al., "De novo transcriptome assembly with ABySS," *Bioinformatics*, vol. 25, no. 21, pp. 2872-2877, 2011.

[22] M Meyer, T Munzner, A DePace, and H Pfister, "MulteeSum: A Tool for Comparative Spatial and Temporal Gene Expression Data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 908-917, 2010.

[23] R B Brem and L Kruglyak, "The landscape of genetic complexity across 5700 gene expression traits in yeast," *Proc Natl Acad Sci USA*, vol. 102, no. 5, pp. 1572-1577, 2005.

[24] B Shneiderman, "Dynamic Queries for Visual Information Seeking," *IEEE Software*, vol. 11, no. 6, pp. 70-77, 1994.

[25] C North and B Shneiderman, "Snap-Together Visualization: Can Users Construct and Operate Coordinated Views," *Intl. Journal of Human-Computer Studies, Academic Press*, vol. 53, no. 5, pp. 715-739, 2000.

[26] A Martin and M Ward, "High Dimensional Brushing for Interactive Exploration of Multivariate Data," in *Proceedings of IEEE Visualization*, pp. 271-278, 1995.

[27] The Saccaromyces Genome Database. [Online]. http://yeastgenome.org

[28] M Johannesson et al., "A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: the NIH heterogeneous stock," *Genome Res*, vol. 19, no. 1, pp. 150-8, 2009.

[29] J K Pritchard, M Stephens, and P Donnelly, "Inference of Population Structure Using Multilocus Genotype Data," *Genetics*, vol. 155, pp. 945-959, 2000.

[30] Madahain JO, Fisher D, Smyth P, White S, and Boey YB, "Analysis and Visualization of Network Data using JUNG," vol. VV, no. II, 2005.

[31] D Gilbert. (2010) JFreeChart open source library. [Online]. http://www.jfree.org/jfreechart/index.html

[32] JHeatChart. [Online]. http://freshmeat.net/projects/jheatchart