The Role of Complementary Bipartite Visual Analytical Representations in the Analysis of SNPs: A Case Study in Ancestral Informative Markers

Suresh K. Bhavnani, PhD1, Gowtham Bellala, MS2, Sundar Victor, MS1, Mamta Abbas3, Vickie McMicken4, Jeffry Tupa5, Shyam Visweswaran, MD, PhD6

1Inst. for Translational Sciences, University of Texas Medical Branch, Galveston, TX; 2Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI; 3Computer Information Systems, 4Management and Information Science, 5School of Communication, University of Houston Clearlake, Houston TX; 6Department of Biomedical Informatics, and the Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA

Abstract

Several studies have shown how sets of single nucleotide polymorphisms (SNPs) can help to classify subjects based on their continental origins, with applications to case-control studies, and population genetics. However, most of these studies use dimensionality-reduction methods such as principal component analysis or clustering methods that result in unipartite (either subjects or SNPs) representations of the data. Such analyses conceal important bipartite relationships such as how subject and SNP clusters relate to each other, and the genotypes that determine their cluster memberships. Here we explore the use of three bipartite visual analytical representations (bipartite network, heatmap with dendrograms, and Circos ideogram) that enable the simultaneous visualization of subjects, SNPs, and subject attributes. The results demonstrate the role that the bipartite representations each play in revealing novel insights into SNP data, which are difficult to derive from purely unipartite views of the data. We conclude by discussing the implications of the results for the analysis of SNPs in genomic studies associated with disease, and the need for a general framework that guides the use of complementary bipartite visual analytics to analyze complex relationships in SNP data.

Introduction

Because more than 99% of the 3 billion base pairs in the human genome are identical across all humans [1], the remaining less than 1% contains crucial information about how humans vary. This variation, resulting from millennia of natural selection and random drift, is coded in approximately 20-30 million locations on the human genome, commonly referred to as single nucleotide polymorphisms (SNPs). By exploiting the development of inexpensive high throughput genotyping technologies, recent studies have identified SNPs that are associated with the risk of developing specific diseases [2], and SNPs that are highly associated with continental origins. For example, several studies have identified SNPs that have large differences in genotype frequencies between two or more continental populations such as Africans and Europeans [3]. Identification of such SNPs (referred to ancestry informative markers or AIMs), have important implications for research in risk assessment, diagnosis, prognosis, and therapy of common diseases. For example, because African Americans have around 10-15% European admixture [4], AIMs can be used to select or assign subjects to subpopulations in case-control studies, with the potential of reducing confounding based on continental origins [5]. AIMs can also shed light on which genes are implicated in diseases that are over expressed in certain populations, such as the high incidence of diabetes mellitus in African Americans [6].

While the above studies have begun to identify important differences in genotype frequencies between subpopulations, to the best of our knowledge the methods used have focused on generating a unipartite view (either SNPs, or subjects) of the data. For example, studies that attempt to identify AIMs typically use dimensionality reduction methods such as principal component analysis (PCA), or clustering methods such as k-means. Such methods aim to identify a parsimonious set of SNPs that can separate the data into distinct population clusters, in addition to revealing admixture. The output of these methods typically includes a plot of subjects (e.g., scatter plot, dendrogram) to show how they relate to each other based on appropriate distance measures.

Although these dimensionality reduction and clustering methods are powerful, they do not provide a bipartite view of the data, potentially concealing important relationships. For example, they cannot directly reveal which clusters of subjects are related to which clusters of SNPs, nor can they reveal the nature of their membership based on the proportion of genotypes. To address these limitations, we explored the use of bipartite visual analytical
representations to analyze SNP data. Such representations enable the simultaneous view of (1) subjects and SNPs, and (2) the type and frequency of genotype associations between the subjects and SNPs.

We begin by describing the conceptual underpinnings of the methods used to analyze AIMs and other SNPs, which led us to explore bipartite visual analytical methods. We then describe how we extracted from the International HapMap Project’s data repository, a dataset of subjects and SNPs related to ancestry, why and how we represented and analyzed the data using bipartite visual analytical representations with associated quantitative measures. Next, we discuss how the bipartite representations revealed complex associations between subjects and SNPs. We conclude by discussing the implications of the results for the analysis of SNPs in genomic studies associated with disease, and the need for a general framework that guides the use of complementary bipartite visual analytics to analyze complex relationships in SNP data.

**Conceptual Underpinnings of Current Methods Used to Analyze SNP Data**

We briefly describe the nature of SNP data because it is central to understanding the methods used to analyze such data. The human genome consists of a sequence of approximately 3 billion locations, where each location has one of the four possible molecules A, C, G or T (referred to as nucleotides). While the majority of these locations on the genome are identical across all humans, a minority of the locations (referred to as SNPs) can vary across humans. A SNP typically has only two possible molecules (e.g., A or G), of which one is less common in the population (minor allele), and the other is more common (the major allele). However, because humans carry two copies of the genome, SNPs that have molecules A and G can have three combinations across both copies of the genome: AA, AG and GG. These three combinations are referred to as the genotypes of the SNP.

Several detailed reviews of methods used to analyze SNP data exist including distance-based measures used for exploratory analysis of genetic data [7, 8], and model-based algorithms for estimating ancestral information in subjects [9, 10]. Here we focus on the conceptual underpinnings of the typical methods used to analyze SNPs, and why they motivated us to explore the use of bipartite representations. The current methods can be broadly classified as univariate and multivariate analysis of SNPs.

**Univariate Analysis of SNPs and Subjects.** The three genotypes described above form the basic input of most methods to analyze SNPs. Although the methods for analyzing SNP data are still evolving, many studies begin with the use of the univariate chi-square test to identify which SNPs are the most significant across the populations being studied (e.g., subjects from different ancestries, or between diseased and healthy populations). This method compares for each SNP the proportion of the three genotypes between the two or more groups being studied, and outputs the significance for each SNP. Because of the large number of SNPs being tested, some could be significant just by chance. Therefore the results are adjusted for false discovery using methods such as the *Bonferroni* correction. Researchers then use the most significant SNPs for further analysis. While this method is powerful, it is limited because it treats each SNP independently, when SNPs could in fact be working in groups.

**Multivariate Analysis of SNPs and Subjects.** Univariate tests such as chi-square are complemented with multivariate methods which consider the variance across all the data. These methods can be broadly classified into two categories: (1) distance-based, and (2) model-based.

(a) **Distance-Based Methods.** The distance-based methods typically consist of two steps [3, 7]. The first step is the use of dimensionality reduction methods to project the data into a lower dimensional space. For example, PCA is often used to identify the dimensions (referred to as components) which describe maximum variability in the data, and model them as linear combinations of the SNPs. PCA also outputs each subject’s (or SNP’s) coordinates along the identified components [7], and plots them in two dimensions using pairs of components as axes. This plot visually suggests how many clusters exist in the data. The second step attempts to find boundaries for clusters in the data. For example, *k*-means takes as input the coordinates generated from PCA, along with the inferred number of clusters, and generates the cluster boundaries for the subjects (or SNPs). This two-step approach is especially suitable for exploratory data analysis, where the (unknown) number of clusters is determined through visual inspection of the data in low dimensions from the PCA output, which is then used to generate the cluster boundaries in *k*-means. Finally, to test the significance of the ancestry informative SNPs identified using the above steps, AIMs researchers often use Wright’s F-statistic (FST) [11], which measures the diversity of randomly chosen alleles within the same sub-population relative to that found in the entire population. FST is often expressed as the proportion of genetic diversity due to genotype differences among populations [12].

(b) **Model-Based Methods.** The model-based algorithms such as STRUCTURE [13] and ADMIXMAP [14] assume an underlying probabilistic model, and estimate the parameters in the model from the data using standard statistical
methods such as maximum-likelihood or Bayesian methods. In addition to clustering subjects in the data, these algorithms can estimate ancestral information in admixed subjects [9, 10]. Given that the data consists of $K$ different populations, these algorithms output a vector of $K$ values for each subject in the data, where the vector values correspond to the ancestry proportion in a subject’s genome content that is derived from each population.

Although the above methods are powerful in separating subjects based on continental origins and disease subtypes, or in identifying the important SNPs, they are based on a unipartite view of the data: they can be used either to analyze cluster SNP clusters based on subjects, or subject clusters based on SNPs. In this paper, we explore the use of bipartite representations of the data to help reveal important insights into SNP data such as inter-cluster relationships between SNPs and subjects, in addition to the nature of the cluster memberships.

Method

Research Question. Our goal was to demonstrate the value of bipartite representations to analyze SNP data. We therefore posed the research question:

*What is the bipartite relationship between subjects (from different continental origins), and SNPs (known to code for ancestry information)?*

Data Selection. To address the research question, we selected SNP data from Phase 2 HapMap (release 23) database [15]. Because prior research has identified [16] and verified [17] that 128 AIM SNPs contain strong signal related to ancestry, we extracted that set from 60 unrelated subjects from Ibadan, Nigeria (henceforth referred to as **Yoruba Africans**), and 60 unrelated subjects from Utah America with ancestry from northern and western Europe (henceforth referred to as **Utah Americans**). Of the 128 SNPs, 78 SNPs had complete data, and the remaining 30 SNPs had missing data for <5% of the subjects, which we excluded from the analysis. The final dataset contained genotype data for 78 SNPs and 120 subjects and had no missing genotypes. Because prior research [18] has shown no significant differences in AIMs SNPs between Africans and African Americans, the two ancestry groups in this data set therefore represent the major race classes typically used in US case-control studies.

Data Encoding. For each SNP, the three genotypes (e.g., AA, AG, GG) were coded as 0, 1, or 2 denoting whether a subject was a major homozygotic (having two copies of the major allele), a heterozygote (having one copy of each allele) or a minor homozygote (having two copies of the minor allele) respectively. The minor allele of a SNP was determined to be the one that had the lower frequency in the data. This encoding, referred to as the additive genetic model [2], represents an increased disease risk (or probability of an ancestry) of $r$ for the heterozygote when compared to the major homozygote, and $2r$ for the minor homozygotic when compared to the major homozygote. This additive model is typically used in many genetic analyses, including in the analysis of AIMs SNPs.

Data Analysis. Our analysis consisted of two steps: (1) exploratory visual analysis through the use of three bipartite visual representations chosen to identify emergent bipartite relationships between subjects and SNPs; and (2) quantitative analysis through the use of methods suggested by the emergent visual patterns. This two-step method was motivated by our earlier studies [19-21] that used a similar approach, and which have revealed that bipartite relationships can exhibit in different patterns (e.g., nested clusters, disjoint clusters), each prompting the use of quantitative methods that make the appropriate assumptions about the underlying data.

1. Exploratory Visual Analysis. To analyze the data, we used the following three bipartite representations.

(a) Bipartite Networks. Networks provide a powerful approach for representing and analyzing complex relationships. They are increasingly being used to analyze a wide range of molecular phenomena such as gene regulation [22], disease-gene associations [23], and disease-protein associations [24]. A network (also referred to as a graph in mathematics) consists of a set of points or nodes, joined in pairs by lines or edges; nodes represent one or more types of entities (e.g., subjects and SNPs). Edges between the nodes represent a specific relationship between the entities (e.g., a homozygote relationship between a subject and a SNP). Figure 1A shows a bipartite network (where edges exist only between different types of entities) [25] of subjects and SNPs, which was created using Pajek [26] (version 1.23).

Node diameter was used to represent the sum of the edge weights connected to that node. This enabled a rapid visual inspection to determine for example, which patients have overall high aggregate genotype values, and how such subjects relate to the rest of the network. Edge weights in the network were used to represent the genotype (0, 1, or
2. Global patterns in the network were visualized and analyzed using the Kamada-Kawai\(^1\) layout algorithm [27]. The algorithm results in nodes that are connected by high edge weights to be pulled together, and those with low edge weights to be pushed apart. This algorithm is fast but approximate, and well-suited for small to medium-sized networks consisting of between 50-1000 nodes [28]. As shown in Figure 1A, the result is that nodes with a similar pattern of edge weights (e.g., black nodes in the upper left-hand side) are placed close to each other.

Network analyses provide two advantages for analyzing complex relationships. (1) They do not require \textit{a priori} assumptions about the relationship of nodes within the data, such as the hierarchical assumption of hierarchical clustering, or disjoint clusters of \textsc{k}-means. Instead, by using a simple pair-wise representation of nodes and edges, network layouts enable the identification of multiple structures (e.g., hierarchical, disjoint, overlapping, nested) in a single representation [28], but frequently do reveal disjoint clusters. Therefore, while layout algorithms such as Kamada-Kawai depend on the force-directed assumption and its implementation, such algorithms are viewed as less biased for data exploration because they do not impose a particular cluster structure on the data, often leading to the identification of more complex structures in the data [19]. (2) Networks also enable the simultaneous visualization of \textit{multiple raw values} (e.g., subject-SNP associations, subject attributes), \textit{aggregated values} (e.g., sum of edge weights), and \textit{emergent global patterns} (e.g., clusters) in a uniform visual representation. The overall network representation therefore enables the rapid generation of hypotheses based on complex multivariate relationships, and enables a more informed approach for selecting quantitative methods to verify the patterns in the data.

\textbf{(b) Heatmaps}. While networks provide a powerful method for visualizing data, the edges can often get very dense, making it difficult to analyze the edges and their weights connected to specific nodes. We therefore used a second bipartite representation called a bipartite heatmap [29]. Here, instead of using Euclidean distances and edge weights to represent relationships between nodes (such as in the bipartite network), a heatmap uses a color grid to represent such relationships. As shown in Figure 2B, the rows represent subjects, and the columns represent SNPs, and the cells represent the genotypes: white = 0, gray = 1, and black = 2.

\textbf{(c) Circos Ideograms}. While heatmaps enable inspection of subjects and their relationship to each SNP, they cannot simultaneously represent attributes of the entities such as the sex of specific subjects, nor do they allow interactive exploration of the relationship between subsets of the data such as subjects which have high admixture (resulting from mating between subjects from reproductively isolated ancestral populations [3] of the two races). We therefore used a third bipartite representation called a Circos Ideogram [30]. As shown in Figure 3B, the Circos Ideogram represents the bipartite network of subjects and SNPs in an inner circle, and attributes of subjects such as sex in the outer rings.

2. **Quantitative Analysis.** The insights derived from the three bipartite visualizations were analyzed using three quantitative methods, which were selected based on their appropriateness to the emergent patterns in the network.

\textbf{(a) Agglomerative Hierarchical Clustering}. Because the network layout suggested the presence of distinct clusters for subjects and for SNPs, we used the agglomerative hierarchical clustering method to verify the number of clusters, and to identify the boundaries of the clusters. The clustering was done using the Manhattan dissimilarity measure (to handle the 0, 1, and 2 edge weights representing the genotype) with the Ward linkage function [30]. The number of clusters and their boundaries were determined based on natural breaks in the SNP, and the subject dendrograms. The dendrograms were also combined with the heatmaps to aid in the visual analysis of the results.

\textbf{(b) Betweenness Centrality}. To identify the subjects that had high admixture of SNPs from the two ancestries, we calculated the \textit{betweenness centrality} [26] for each node in the network. This measure calculates the proportion of the shortest paths between every pair of nodes in the network that have to pass through a given node. In a highly clustered network, nodes that have a high betweenness centrality value tend to be those that are between clusters because they act as “bottlenecks” or “bridges” for the shortest paths that start from one cluster, and end in another cluster. The measure is given by the expression shown on the right where \(\sigma_s\) denotes the total number of shortest paths between nodes \(s\) and \(t\), and \(\sigma_v\) denotes the number of those paths that pass through node \(v\).

\textbf{(c) Clusteredness and Bipartite Modularity}. To test whether the clusters in the network could have occurred by chance, we compared the variance, skewness, and kurtosis of the dissimilarities in the HapMap data, to 1000 random

---

\(^1\) The Kamada-Kawai layout algorithm is approximate because it does not guarantee a globally optimal layout. The method is therefore used to explore the data using different starting conditions, and the observed topology verified using appropriate quantitative methods.
permutations of this data. For each network permutation, we preserved the size of the network, in addition to the edge weight distribution of each SNP when analyzing the SNP dendrogram, and the edge weight distribution for each subject when analyzing the subject dendrogram. Significant breaks in the HapMap’s subject or SNP dendrograms would result in a significantly larger variance, skewness, and kurtosis of the dissimilarity measures, compared to the same measures generated from the random networks.

We also used the RGraph algorithm [31] to calculate the modularity of the bipartite network. This algorithm attempts to partition a bipartite network into clusters (or modules) by optimizing modularity. The modularity of a particular clustering of the nodes is defined as the number of edges falling within the clusters, minus the expected number of such edges in a network of the same size with randomly reassigned edges. Modularity values range from -1 to +1, where high values (>0.3) represent significantly more edges within clusters compared to random networks of the same size, zero represents no difference compared to random networks, and negative values represent fewer edges within clusters compared to random networks [25]. Because the modularity algorithm for weighted bipartite networks is an active area of research (personal communication Guimera), we instead calculated the modularity for a binarized-version of the network using the dominant model (an edge represents the presence of one or more copies of the minor allele), in addition to the recessive model (an edge represents the presence of two copies of the minor allele).

Results

The bipartite visualizations and quantitative analysis revealed distinct SNP and subject clusters, in addition to a subset of subjects that represents an admixed population. For each result, we describe the results of the visual analysis, followed by their quantitative verification.

**SNP and Subject Clusters.** The bipartite network visualization of 120 subjects and 78 SNPs revealed a complex but understandable clustered pattern. As shown in Figure 1A, there are two major clusters of SNPs and subjects, one to the left, and one to the right. The SNPs are connected to subjects via black, gray and white edges that denote genotype values 2, 1, and 0 (see section on Data Encoding for details). The left cluster consists of SNPs that have predominantly the minor homozygote genotype (genotype 2) for the Yoruba African subjects (black nodes), whereas the right cluster (shown in pink) consists of SNPs that predominantly have the minor homozygote genotype for the Utah American subjects (white nodes). Henceforth, for brevity, we refer to the left SNP cluster as “Utah American SNPs” and to the right SNP cluster as “Yoruba African SNPs”. In addition, there appear to be additional SNPs that circle mainly the Yoruba African SNPs, and which have weaker connections to both the Yoruba African subjects, and the Utah American subjects, as shown by the relatively fewer black edges.

To quantitatively verify the above visual result, we used agglomerative hierarchical clustering. Figure 1B shows the resulting dendrogram of the SNPs. As shown, the dendrogram shows a significant break at 3 as well as 2 clusters. Because the 3 clusters corresponded well with the three distinct topologies of the SNPs, we colored the SNP nodes in the bipartite network based on the boundaries of the three clusters identified by the SNP dendrogram. A dendrogram of the subjects showed that the subject clusters perfectly match the two ancestries defined a priori. It is important to note that the dendrogram alone did not provide an explanation for the nature of the third less dominant SNP cluster (colored red). The nature of the third cluster was more apparent in the network based on its ring topology, and its genotype pattern with the other two dominant SNP clusters. The two methods therefore together provided an explanation for the overall topology of the bipartite network and its subparts, in addition to the quantitative verification of its boundaries.

To generate a network based on a parsimonious subset of the SNPs, and to examine the admixture based on the dominant SNP clusters (blue and pink), we removed the middle SNP cluster (red nodes) from the network, and re-laid out the network using the Kamada Kawai algorithm. Figure 2A shows the result of this transformation on the network. As shown, the original Yoruba African and Utah American subjects continue to be strongly clustered around their respective SNPs. The network also revealed a subset of subjects between the two clusters which appeared to have admixture of SNPs because the members of this subset genotype 2 associations to some Utah American SNPs, and to some Yoruba African SNPs.

The clusteredness of the subjects in the HapMap data was significant when compared to 1000 random networks based on variance of the dissimilarities (HapMap = 74822.5, Random Mean = 1023.6, p<.001 two-tailed test), skewness of the distribution of dissimilarities (HapMap = 10.56, Random Mean = 4.3, p<.001 two-tailed test), and kurtosis of the distribution of dissimilarities (HapMap = 114.01, Random Mean = 24.28, p<.001 two-tailed test).
The advantage of the above method is that it considers the weights of the edges, and therefore there is no loss of genotype information. However, because modularity algorithms currently do not support weighted bipartite networks, we binarized the edge weights by removing all edges representing genotypes 0 and 1, and created a network where edges represented only genotype 2 (recessive genetic model shown in Figure 3A). For this network, the modularity was above 0.3 (patients = 0.46, SNPs = 0.48) indicating that the clustering of the subjects, and of the SNPs were significant compared to random chance [26]. In both cases, the highest modularity was achieved with the same clusters identified by the hierarchical clustering. We also removed all edges with weight 0, and created a network where an edge represented either genotype 1 or 2 (dominant genetic model, not shown). For this network, the modularity was lower (patients = 0.24, SNPs = 0.25) suggesting that the recessive model separates the two continentally population more strongly compared to the dominant model for this dataset.

**Subjects with Ancestry Admixture.** To analyze the admixed subjects who are located in the center of the network, we used the betweenness centrality measure. Because genotype 2 appeared to be the main determinant of the clusters, we used the recessive model to conduct this analysis. Figure 3A shows the results of the betweenness centrality measure, which sizes nodes based on the proportion of shortest paths that pass through them. As shown,
the measure correctly identified 12 Utah Americans, and 7 Yoruba Africans that have genotype 2 relationships for both clusters. These are the 19 subjects in the center of the network enclosed by the yellow shape in Figure 3A.

The measure also identified SNPs that had strong connections to the admixed subjects, and therefore were implicated in the admixture. However, due to the density of black edges in the network, it was difficult to determine which SNPs from each cluster were connected to subjects from the opposite cluster. Furthermore, the admixed subjects were scattered across the heatmap (rows containing dark cells representing genotype 2 in the upper right and lower left areas of Figure 2B) because they do not form a cluster based on their membership to the same SNPs.

To address this difficulty, we used the Circos representation which enabled a closer inspection of this subset of subjects across all the SNPs. Figure 3B shows the Circos ideogram where the edges and SNP nodes can be highlighted based on one or the other cluster of subjects to which they are connected. Therefore, as shown in Figure 3B, we highlighted all edges that were connected to the Utah American nodes (white nodes), to explore which Yoruba African nodes are, and are not, connected to them. As shown, the representation quickly revealed 8 SNPs (colored white on the right hand side of the diagram) that account for the admixture of the Utah Americans, and the remaining 10 SNPs (colored black) are not involved in that admixture. Similarly, the Circos ideogram enabled the identification of SNPs that were involved in the admixture of the Yoruba African set, and those that were not. The Circos ideogram therefore helped to closely examine the admixture based on the inter-cluster relationship, which was not directly possible through the network or the heatmap representations. Additionally, the outer ring of the ideogram shows the sex of the subjects (red=male, green=female), revealing how their proportion varies across the admixed subjects in the two continental groups. The Circos ideogram therefore enabled the exploration of the three-way association between SNPs, subjects, and their attributes.

Discussion

Our goal was to explore the role of bipartite visual analytical representations in the analysis of SNP data. While the results matched many of the results from earlier AIMs studies [16,17], they provided a richer understanding of the associations in the data. First, while the larger set of 128 SNPs we used to seed our study are clearly discriminatory for subjects from a large number of continental origins, the network analysis helped to identify a smaller set of 40 SNPs that is possibly sufficient to form strong clusters of Utah Americans, and Yoruba Africans. Furthermore, this smaller set also enabled us to closely examine the admixed subjects and which SNPs were involved in that admixture. The results also provided a more complex understanding of the association between the SNP and subject clusters, in addition to the nature of the cluster memberships. These in turn enabled us to understand the complementary role that each bipartite representation played in revealing the associations as discussed below:

**Relationship between Clusters.** The bipartite network of 78 SNPs in Figure 1A revealed (1) SNPs that were highly discriminatory of the two ancestries, (2) SNPs that were not discriminatory of the two ancestries, and (3) the relationship of the SNP and subject clusters. Because the weakly connected SNPs formed a ring-like structure around the Yoruba African SNP cluster, it suggested a weak but none-the-less relatively stronger relationship to that
cluster compared to the Utah American cluster. This pattern was difficult to discover from the heatmap because of a
fundamental difference in the two representations: Spring layouts like Kamada-Kawai for 2D layouts, have two
degrees of freedom in laying out the nodes, and therefore can show multiple adjacency relationships; in contrast,
dendrograms have only one degree of freedom because nodes can be located either along the x or y axis restricting
the number of adjacencies that can be represented simultaneously. These adjacencies have to be inferred by
inspecting the color gradations in the heatmap, which is perceptually more difficult to comprehend, compared to
layouts that use distance to show similarity. However, while these relationships are difficult to discover from the
heatmap and dendrograms, we used them to confirm those patterns after the fact. In contrast, the network layout
while suggesting distinct SNP and subject clusters, cannot on its own discover the boundaries of the clusters, and
therefore we used the dendrograms and modularity to discover those boundaries, which we then confirmed by
overlaying them as similarly colored nodes in the network. The two representations therefore together
enabled the
discovery and confirmation of the relationship between the clusters.

**Nature of Cluster Memberships.** In addition to the identification of the cluster boundaries, and the relationship
between the clusters, the bipartite representations also revealed the nature of the cluster memberships. Unlike
unipartite representations produced by methods like PCA and k-means (and for that matter even unipartite
networks), bipartite networks through weighted edges explicitly show the nature of the relationship between a
subject and a SNP. The overall topology of the network, revealed insights such as what kind of relationship is
responsible for the formation of the clusters. For example, the two dominant clusters in our dataset were mainly held
together because of the genotype 2 relationships from their respective subject populations. This might not always be
the case in different datasets. For example, clusters could be held together with very few genotype 2 relationships,
and be dominated instead with many genotype 1 relationships. Of course one might argue that such information can
anytime be extracted directly from the raw data, but the power of the bipartite visual analytical representations lies
in the fact that they can suggest patterns that the researcher might not otherwise think about analyzing.

Similar to the inadequacy of any single representation to enable the comprehension of the clusters and their
relationship to each other, networks and heatmaps were also unable to provide the full picture of the admixed
population. While the network helped to identify the existence of the admixed population, the density of the edges
did not allow a direct inspection of the nature of their admixture, and which SNPs in both clusters were responsible
for that admixture. Furthermore, these admixed subjects were spread out in the heatmap as their discovery is based
on a network-based relationship which is not the basis of the clustering algorithm. In contrast, the Circos
representation enabled the selection of edges based on the subject cluster to which they were connected, which
helped to quickly identify which nodes were, or were not, implicated in the admixture. Therefore, while the Circos
representation is not designed to identify clusters, it can enable the inspection of the admixture in a much more
effective way compared to networks and heatmaps.

**Methodological and Theoretical Implications**

The results have methodological and theoretical implications. From a methodological perspective, the bipartite
representations could intuitively show a researcher studying the data from a case-control study, not only which
Subjects have high admixture, but also the reason for that admixture based on the type and nature of SNP cluster membership. For example, if the SNPs involved are the topic of the case-control study, then the bipartite representation can reveal important information to make critical decisions to prevent confounding experimental results. Furthermore, when studying SNP data beyond AIMs, researchers can use the identity of the SNP membership to rapidly derive data-driven hypotheses for the disease causation. For example, we used the Genetic Association Database (GAD) to analyze the known association of the 18 Yoruba African SNPs and the 22 Utah American SNPs to diseases. We found that the Yoruba African SNPs were associated with hypertension, schizophrenia, prostate cancer, Parkinson's disease and autoimmune inflammatory diseases. In contrast, the Utah American SNPs were associated with osteoporosis, schizophrenia, bipolar disorder, and multiple sclerosis. These associations demonstrate a differential prevalence of diseases in the two populations based on the SNPs to which they were associated. In case-control SNP data, such differential prevalence of SNP-disease associations provides a starting point for analyses for elucidating the associations between the disease under study, and to other diseases.

From a theoretical perspective, we have demonstrated that the network representation enabled us to rapidly explore the effect of different models (e.g., recessive, dominant) on the SNP and subject clustering, and how the emergent patterns could be rapidly detected and quantitatively verified through network measures such as modularity and betweenness. Future research should explore how these methods complement current methods used to analyze AIMs SNP data. We have also elucidated the limitations of networks, and how to overcome them through the use of multiple bipartite visual representations. The results showed that each representation provides different affordances, and therefore play the role of enabling discovery, confirmation, explanation, and inspection for different tasks. This understanding has inspired us to explore the development of a complimentary visual analytical framework, which could explain and guide the use of multiple visual analytical representations, to rapidly enable discoveries in complex SNP data.

**Conclusion and Future Work**

Although there exist many powerful methods to analyze SNP data, to the best of our knowledge they rely mostly on unipartite representations of the data. Here we explored the use of three bipartite visual analytical representations and associated quantitative methods with the goal of enabling a deeper understanding of SNP-subject data. The results suggest that bipartite representations of AIMs SNP data can provide not only an understanding of the SNP and subject clusters based on different models, but also how they are related to each other, and the nature of the membership of the subjects to different SNP clusters. Therefore, while the development of new methods holds a high premium in the informatics field, we believe there is much to understand in how to strategically combine existing visual analytical methods to reveal new insights in a domain.

While we have demonstrated the value of bipartite representations in only one dataset related to AIMs, our current research suggests that the approach is more general. For example, we have begun to use the same approach to represent Alzheimer’s patients and controls, along with the SNPs known to be implicated in the disease. The results are revealing complex patterns of bipartite clustering which have the potential of leading to a deeper understanding of the underlying genetics in Alzheimer’s disease. Therefore we believe that our approach has implications beyond the analysis of AIMs SNPs as it can be used for the analysis of a wide range of SNP data.

Finally, we believe we have only scratched the surface in understanding the complimentary role of multiple bipartite visual analytic representations. Our future research hopes to develop a comprehensive framework which integrates current methods with bipartite visual analytical representations with the goal of helping researchers to rapidly identify complex SNP-related phenomena, and unravel the mysteries related to the genetic causes of complex diseases.

**Acknowledgements**

This work was supported by NIH grant 1U54RR02614 UTMB CTSA (ARB). We thank G. Vallabha, and A. Narain for their contributions.

**References**