

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, Jeffrey Tupa⁵, Shyam Visweswaran, MD, PhD⁶

¹Institute 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-Clear Lake, 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 individuals 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 that result in unipartite (either individuals or SNPs) representations of the data. Such analyses conceal important bipartite relationships such as how individual and SNP clusters relate to each other, and the genotypes that determine their cluster membership. 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 individuals, SNPs, and individual attributes. The results demonstrate the role that bipartite representations play in revealing novel insights into SNP data, and the need for a general framework that guides the use of complementary bipartite visual analytical representations to analyze complex relationships in SNP data.

Introduction

Recent studies have identified single nucleotide polymorphisms (SNPs) that are associated with the risk of developing specific diseases, and SNPs that are highly associated with continental origins [e.g., 1] for use in case-control studies. However, such studies typically use methods such as principal component analysis or clustering that generate a *unipartite* view (either SNPs, or subjects) of the data, which could potentially conceal important relationships in molecular data [e.g., 2]. Here we explore the use of three *bipartite* visual analytical representations to analyze SNPs that code for continental origins, commonly referred to as ancestry informative markers (AIMs).

Methods

We analyzed 128 previously identified AIM SNPs [1] for a group of 60 Yoruba Africans and 60 Utah Americans extracted from the HapMap Phase II database. The analyses consisted of: (1) **exploratory visual analysis** using *bipartite networks, heatmaps, and Circos ideograms*; and (2) **quantitative analysis** using *agglomerative hierarchical clustering, betweenness centrality, clusteredness, and bipartite modularity*.

Results and Conclusion

In addition to matching the results from earlier studies on these AIM SNPs [1], the bipartite analysis provided two additional insights: (1) **relationship between SNP and subject clusters** which helped to identify a subset of 40 SNPs that were highly discriminatory between Utah Americans, and Yoruba Africans; and (2) **nature of cluster memberships** which helped to explain the reason underlying the SNP and subject clusters, in addition to identifying which SNPs were strongly associated with the admixed subjects. These results suggest the need for a complementary visual analytical framework that combines multiple methods to help analyze a wide range of SNP data including those related to complex diseases such as Alzheimer's and breast cancer.

Acknowledgements: Supported by NIH grants 1U54RR02614 UTMB CTSA (ARB), and HHSN276201000030C.

References

1. Kosoy R, Nassir R, Tian C, White PA, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat.* 2009;30(1):69-78.
2. Bhavnani SK, Victor S, Calhoun WJ, Busse WW, Bleecker E, Castro M, Ju H, Pillai R, Oezguen N, Bellala G, Brasier AR. How Cytokines Co-occur across Asthma Patients: From Bipartite Network Analysis to a Molecular-Based Classification. *Journal of Biomedical Informatics* (in press).