*i*Circos: Visual Analytics for Translational Bioinformatics

Suresh K. Bhavnani[1], Mamta Abbas[2], Vickie McMicken[3], Numan Oezguen[4], Jeffry Tupa[5]

[1,4] Institute for Translational Sciences, University of Texas Medical Branch; [2]Computer Information Systems, [3]Management and Information Science, [5]School of Communication, University of Houston Clear Lake

[1]skbhavnani@gmail.com; [2]abbasmamta@gmail.com; [3]vmcmicken@teleshare.net; [4]nuoezgue@utmb.edu; [5]jeffry.tupa@gmail.com

ABSTRACT

Translational bioinformatics increasingly involves the discovery of associations between molecular and phenotype information, with the goal of transforming those discoveries into novel methods for diagnosis and treatment. To enable such complex analyses, researchers need approaches that provide the simultaneous representation and interactive analysis of patients, and their molecular and phenotype information. Because few existing visual analytical systems provide appropriate capabilities, we developed a prototypical visual analytical system called *iCircos*, which enables the simultaneous and interactive exploration of molecular and phenotype information. We discuss our overall method for developing the prototype by integrating user needs and design heuristics from visual analytics, with agile programming in HTML5 and SVG. A demonstration of the prototype to explore molecular and phenotype associations in two disease datasets suggests that iCircos has the potential to accelerate translational discoveries in complex disease datasets. We conclude by discussing insights about designing visual analytical systems for translational bioinformatics, and present our future plans for user testing and adding advanced interactivity to the prototype.

Categories and Subject Descriptors

H.5.0 Information Interfaces and Presentation.

General Terms

Design, Human Factors.

Keywords

Visual Analytics, translational bioinformatics, Circos.

1. INTRODUCTION

The exponential increase of molecular (e.g., gene expression), and phenotype (e.g., blood pressure) information far exceeds the cognitive capabilities of most researchers to rapidly analyze such data. This is especially true in translational bioinformatics (TBI), where researchers attempt to discover relationships in molecular and phenotype information with the goal of transforming those discoveries into more effective diagnoses and treatments [4].

To conduct such complex analysis, TBI researchers require methods to: (1) **simultaneously represent** molecular and phenotype information to enable translational insights; and (2) **interactively explore** the representation to formulate hypotheses about the underlying biological mechanisms.

IHI'12, January 28–30, 2012, Miami, Florida, USA.

One promising approach to address such complex analysis is through *visual analytics*, which is the science of analytical reasoning facilitated by interactive visualizations [11]. Interactive visualizations, when based on principles of perception and interaction, leverage the massively parallel architecture of the human visual cortex, enabling rapid detection of complex patterns in graphical data [5].

However, while there are several visual analytical systems that have been developed to assist in the analysis of complex data [11], few enable researchers to simultaneously represent and interactively analyze patients and their molecular and phenotype information. For example, bipartite networks can effectively represent patients and genes as two sets of nodes, with connected weighted edges representing gene expression. However, such networks are limited in the number of phenotype variables that can be graphically represented [2]. For example, patient nodes can represent gender and race by node color and shape respectively, but require additional networks for more variables.

Here we describe how we used an existing visual representation called a *Circos Diagram* [9] to build a new prototypical system called *Interactive Circos (iCircos)*. This prototype enables the *simultaneous* representation, and *interactive* exploration of patients, and their associated molecular and phenotype information. We begin by discussing our motivation for developing iCircos, followed by its **design** (based on an analysis of user needs and interaction design heuristics), **implementation** (based on the selection of a programming platform and design of the system architecture), and a demonstration of its **usefulness** (based on its use in exploring two disease datasets). We conclude with insights about future functionality and usability assessment.

2. PRIOR WORK AND MOTIVATION

Our attempt to build iCircos was based on two motivations: (1) the nature of molecular and phenotype data being analyzed by TBI researchers; and (2) the lack of an appropriate approach to analyze such data.

2.1 Nature of Data and Analysis

Our TBI team has increasingly been approached by biologists and clinicians who wish to simultaneously analyze molecular and phenotype data. For example, we were requested to analyze a dataset of 83 asthma patients, 18 cytokine expressions for each of the patients, and 25 phenotype variables. Our task as TBI researchers was to help the domain experts analyze patterns of cytokine expression across the patients, and how those patterns related to phenotype information. The hope was that such an analysis would lead to a molecular-based classification of asthma patients with insights into the underlying biological pathways, resulting in improved diagnosis and novel therapeutic targets.

To address the above need, we began by using bipartite networks [1] to visualize and quantitatively analyze the relationship of asthma patients and their cytokine (molecules involved in inter-

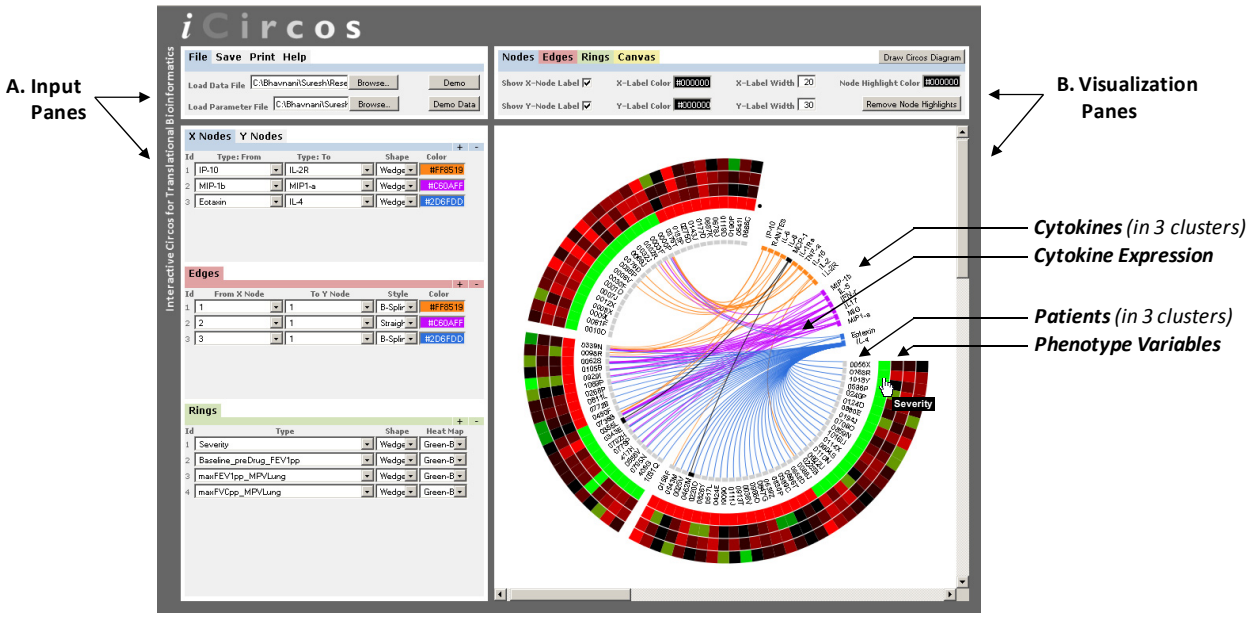

Figure 1. The iCircos interface design and functionality was based on an understanding of user needs, general heuristics for interface design, and specific heuristics for visual analytics. The **input panes** (A) enable the selection of a data file, the selection of rows and columns from the data file, and graphical attributes. The **visualization panes** (B) enable the generation and interaction with the Circos ideogram. Here, edges represent normalized cytokine expression values, and are colored based on the connecting cytokines. A cytokine node (colored black) is double-clicked to reveal its connected edges and nodes. The patients and their respective phenotype variables are sorted (within each patient cluster) based on an outer ring (marked by the dot) representing severity (red=severe, green=not severe).

cellular communication) expressions. The analysis revealed a complex but understandable relationship between three clusters of patients, and three clusters of cytokines. Next, we attempted to add to the network additional nodes representing phenotype variables such as lung function, with the respective connecting edges representing normalized values. However, adding phenotype nodes to the network caused confusion because the meaning of a node and edge were overloaded. We therefore used standard statistical measures (e.g., Kruskal Wallis) to analyze the relationship of the patient clusters to the phenotype variables.

While the above analysis provided important insights, the researchers wished to simultaneously represent molecular and phenotype information in a single visualization to explore the possibly complex multi-way relationships (patients X molecular expression X phenotype X demographics) that could be concealed by the aggregate statistical measures. This led us to explore alternate multivariate visual representations.

2.2 Alternate Visual Representations

We experimented with several other visual analytical systems that were designed to interactively explore multivariate data. Of them the most relevant were: (1) *TreeMap* [8] which enables the analysis of hierarchical data such as populations at different levels of granularities (Countries, Cities, States, etc.) using dynamically created nested rectangles, whose size represented a user-selected variable; (2) *CateRank* [7] which enables the analysis of correlations between sets of categorical variables such as gender and race; (3) *Circos* [9] which enables the visualization of a network using nodes and edges, in addition to node attributes. As shown in the right pane in Figure 1, the Circos visualization can be used to place patient and cytokine nodes around the inner circle, which are connected through weighted edges representing normalized cytokine expression values. Concentric rings can be used to represent through a heatmap, phenotype variables such as lung function corresponding to patients in the inner circle. In addition, the patients and cytokines could be clustered based on the results from the network analysis, and edges colored based on the cytokine clusters to which they were connected.

Based on our analysis, the Circos representation appeared the most appropriate for our purposes because it did not assume a hierarchical structuring of data, was not limited to categorical data, and had a node-edge representation that was intuitive. However, the current Circos system [9] used to generate a Circos diagram requires writing a script to convert the data, and the use of a batch process to execute the script. This motivated us to develop an interactive prototype to help rapidly generate and explore Circos diagrams.

3. DESIGN of iCIRCOS

Our design was based on the intersection of **user needs**, and **design heuristics** from usability engineering and visual analytics.

3.1 User Needs

The intended users of the iCircos system are TBI researchers working in collaboration with domain experts (computational biologists and clinicians) who request different versions of the visualization, and provide interpretation of the patterns. To understand the affordances of the existing Circos system, we used it to analyze asthma molecular and phenotype information in collaboration with asthma domain experts [2]. This case study helped to abstract 4 user needs of TBI researchers:

1. **General Data Format.** The asthma dataset, in addition to several other datesets we have obtained, contained a relatively small number of patients (n<100) due to the difficulty of collecting human samples. The data also contained molecular measures (e.g., cytokine or gene expression), and phenotype information (e.g., lung function and glucose levels). The domain experts wished to first discover the relationship between patients and the molecular information, and then understand how those

discoveries relate to phenotype information. The data were stored in CSV format where the rows represented patients, the columns represented normalized values of molecular expression, or phenotype variables that are either categorical (e.g., severe vs. non-severe), or continuous (e.g., lung function). Our prototype therefore should be designed to use this general data format.

2. **Sufficiency of a Subset of Input Parameters.** The existing Circos system provides more than 100 graphical parameters enabling a huge space of possible diagrams to be generated. While clearly powerful, our case study suggested that a subset of these parameters (size and color of nodes and edges, ring radius, ring thickness, and rotation) was sufficient for enabling domain experts to explore the complex relationship between patients and their molecular and phenotype information.

3. **Interactive Sorting based on Phenotype Variables.** While generating the asthma Circos diagram, we sorted the patients within each cluster based on phenotype values in specific rings. This significantly enhanced the detection of patterns based on changes in color. As shown in Figure 1, the patients are sorted based on the first outer ring representing severe (red) and non-severe patients (green). We had to do this manually in the Circos system, but given its usefulness, decided to automate it in iCircos.

4. **Direct Manipulation and Linking.** While analyzing the Circos diagram, the domain experts asked to interact directly with the diagram to explore specific relationships. For example, they wished to inspect a specific patient by double-clicking on the respective node, and highlighting all its edges and connected nodes. Furthermore, they wished to select specific cytokines of interest, and retrieve from external databases on the web, known biological pathways that implicated those cytokines.

3.2 Interaction Design

The above user needs were operationalized into a graphical user interface (GUI) by using (1) ten general-purpose heuristics [10] that have been extensively used to design and evaluate a wide range of GUIs, and (2) seven heuristics [12] that are specific to the interaction design of visual analytical systems.

Heuristics for Designing General Purpose GUIs. A key heuristic for designing effective GUIs is to create an **aesthetic and minimalist design** so the focus of the user is on the task rather than on the interface. As shown in Figure 1, the GUI was designed to make the clear separation between the input panes on the left, and the visualization panes on the right. The input panes enable the user to load a file, and select which fields are to be visualized. *X Nodes* refer to the molecular information across columns, *Y Nodes* refer the patient information across rows, and the *Subtypes* allow the user to break up the X nodes and Y nodes into clusters. (We chose to use the terms *X nodes* and *Y nodes* to allow for some flexibility for alternate data formats.) Edges allow the user to select the color of subsets of edges, such as the orange colored edges emanating from the first cytokine cluster, and Rings allow the user to specify which phenotype variables to display as rings. Finally, the top visualization pane enables users to modify graphical parameters such as the radius and thickness of the rings.

Another important design heuristic is to provide **help and documentation** for first-time and infrequent users. We therefore added a *Demo* and *Demo Data* buttons in the top left input pane to make clear how an input file maps to the visualization. Similarly, we used the other eight general-purpose usability heuristics including a message to make the system status visible when the visualization is being generated; labels that matched the domain experts' understanding of terms such as nodes, edges, and rings; user control and freedom by enabling the user to change any parameters for the visualization.

Heuristics for Designing Visual Analytics GUIs. We also used seven heuristics that were specific to visual analytics [12]. The heuristics helped to systematically explore use cases as discussed below (the seven heuristics are bolded):

(1) **Select an interesting data point**, was incorporated by enabling users to highlight a node or edge; (2) **Explore the data** was incorporated by enabling users to add or remove rings and change color attributes of edges; (3) **Reconfigure the representation** was incorporated by enabling users to sort the nodes based on a specific ring. As shown, the nodes have been sorted by severity within each cluster, and denoted by a dot close to the respective first ring in the visualization; (4) **Encode a different representation** will be incorporated by enabling users to change the heatmap ring representation into a bar graph; (5) **Abstract/Elaborate the information** was incorporated by allowing users to zoom in to see more details such as the range of edge weights incident to a patient node; (6) **Filter the information** was designed by enabling the user to set a threshold of edge weights to display; (7) **Connect** was incorporated by allowing users to reveal the relationship between elements by double clicking on a node to highlight all connected edges and nodes (as shown by the black nodes and edges in Figure 1). The above heuristics therefore included Shneiderman's well known mantra for visual analytics [12] *Overview first, zoom and filter, then details-on-demand*.

4. iCIRCOS IMPLEMENTATION

The above design (which iteratively evolved throughout the implementation) was implemented by selecting a programming platform, designing the system architecture, and using agile programming for managing the software development.

4.1 Programming Platform

The need for interactive visualization and online linking, suggested we develop a browser-based system. We selected HTML5 with JavaScript (JS) as our programming platform because: (1) the platform provides powerful capabilities for generating Scalable Vector Graphics (SVG) objects (e.g., circles) that can be manipulated in the same way as interface objects (e.g., scroll bars). Furthermore, SVG enables zooming without pixelation which is important to inspect nodes and edges; and (2) unlike Adobe Flash, the HTML5 platform is non-proprietary and is therefore supported by a wide range of delivery devices including the Apple iPad. To reduce lines of code to generate the SVG graphical objects we used the Raphaël JS software library, and to layout and refine the interface, we used Dreamweaver.

4.2 Architecture and Software Development

We developed a system architecture where iCircos could be run on the web or on a local machine, and enable the uploading of a local data file in CSV format. The data is loaded into memory and converted into graphical objects based on user inputs. This conversion is based on three main data structures: (1) The *nodeEdge* array which consisted of patient-cytokine node pairs, each with their connecting edges and attributes; (2) The *outerRing* array which consisted of rings (representing phenotype variables) specified by the user, each with graphical attributes such as thickness; (3) The *generalAttributes* data structure which consisted of attribute values such as inner radius that were required for the overall visualization. The above data structures each had their own JS functions to populate them with values

derived from the interface fields, and with specific functions to draw the nodes and rings in the Circos diagram.

A main function named *generate* was called when the user selected the *Draw Circos Diagram* button (located above the visualization pane). The *generate* function populated the above three data structures based on data extracted from the interface, and passed them to the draw functions to generate the diagram.

Direct manipulation of the visualization such as selecting a node was achieved by retrieving the objects from the Document Object Model (DOM) using a mouse-over event, and executing appropriate functions on the retrieved object.

We used principles from agile software development [6] to iteratively and incrementally learn and code in HTML5, JS, SVG, and Raphaël, with frequent user feedback from a researcher experienced in using Circos. The design and development team had no experience in HTML5/SVG and consisted of a graduate student from computer and information science with experience in Java programming, a graduate student from management and information science with experience in Visual Basic programming, and a graduate student from communications and graphic design experienced in HTML, CSS, and graphic design. The team was led by a faculty member experienced in C programming, interface design, evaluation, and team management. The working prototype shown in Figure 1 was developed and tested in 5 weeks.

5. USEFULNESS OF iCIRCOS

To test the usefulness of iCircos and to provide formative evaluation of the UI, we used it to analyze the asthma dataset described earlier, and to analyze a diabetes dataset which was not used during the design of iCircos.

5.1 Asthma

Asthma is a chronic inflammatory disease of the airways, characterized by hyperactivity to nonspecific stimuli. Asthmatic patients are currently classified as either severe or non-severe based primarily on their response to glucocorticoids [3]. Unfortunately, the current classification of asthma patients has not been sufficiently predictive to guide treatment, leading us to analyze how patients were similar or different based on their molecular profiles. A bipartite network analysis helped to identify three clusters of patients that had a complex but understandable relationship to three clusters of cytokines [1]. Because the network representation was unable to simultaneously represent the molecular and phenotype information, we used the existing Circos system to analyze both types of information simultaneously.

To test if we could regenerate the same diagram we had earlier created using the Circos system, and whether the interactivity enabled any new insights, we repeated the asthma analysis using iCircos. Figure 1 demonstrates that we were able to generate the same diagram that we had generated using the existing Circos system. By selecting a datafile and input parameters, patients and cytokines nodes were placed in the inner circle, each clustered based on the specified subtype information in the data file. To explore the relationship between the cytokines, severity and lung function, we selected those phenotype variables in the Rings panel, and sorted each patient cluster using severity. As shown, we also removed all edges below 0.8 by using *Edge Threshold*.

The diagram generated from iCircos made salient that there was approximately an equal number of severe (red) and non-severe (green) in each patient cluster. Not unexpectedly, the non-severe patients appeared to have higher lung function (as shown by the

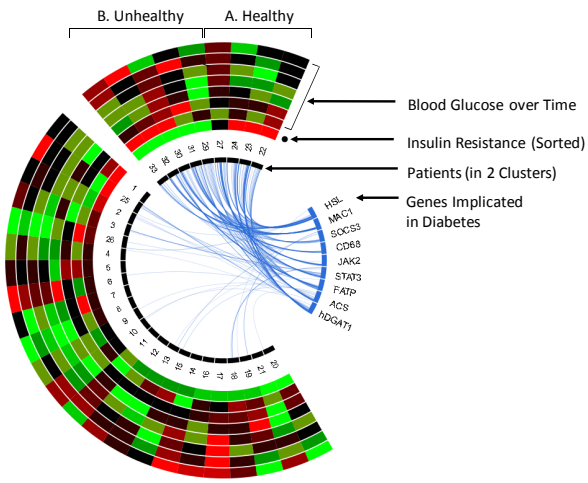

Figure 2. Circos diagram for the diabetes dataset generated through iCircos enabled the identification of molecular differences between healthy and unhealthy patients.

generally brighter colors in rings 3-5 for non-severe patients) compared to severe patients. However, the simultaneous visualization of molecular and phenotypes information, assisted by the interactivity provides a new insight about why the current classification of severe and not severe is not predictive: They are not predictive because the severity phenotype could be caused by *different types* of underlying molecular processes depending on the cytokine clusters to which they are related, and therefore require different therapeutic interventions. The visualization therefore suggests a more complex classification of asthma patients, which combines molecular and phenotype information. We intend to test this hypothesis in our future research.

5.2 Diabetes

While the relationship between obesity and Type 2 Diabetes is well known, much less is known about why lean people develop the same disease. Our molecular and clinical domain experts were therefore interested in discovering how lean adults (fat <30%) with a specific gene mutation were (1) similar or different in their molecular profile to a comparable group of patients who did not have that mutation, and (2) how those patterns related to different phenotype variables. Similar to the asthma project, we first conducted a bipartite network analysis, which helped to identify two clusters of patients based on their molecular profiles. However, it was difficult to use the network representation to visually explore how these patient clusters related to phenotype information.

To address the above limitation, and to understand how useful iCircos was for analysis, we informally observed how a TBI researcher with knowledge of diabetes used iCircos to analyze the data. As shown in Figure 2, he represented the patients (divided into two clusters based on the network analysis) and genes as nodes in the inner circle, and used edge thickness to represent the normalized gene expression values. Next, he used the first outer ring to represent *insulin resistance* (referred to as Rd), and outer rings 2-8 to represent the concentration of glucose in the blood of patients at 0, 30, 60, 90, 120, 150, and 180 minutes respectively, after consuming glucose. Finally, he sorted the patients within each cluster based on Rd (the first outer ring marked with a dot). This sorting resulted in a subset of patients (marked A in the top cluster) who had high Rd values and therefore healthy, and another subset of patients (marked B) who had low Rd values and therefore unhealthy. The researcher confirmed that patient subset

A had a faster absorption of glucose (as shown by their glucose level colors getting progressively darker across the rings representing 0-180 minutes) compared to patient subset B. This distinction between the two subsets enabled the researcher to explore the genes involved more closely. The complex three-way connection between Rd, glucose absorption, and gene expressions suggested to the researcher that the latter two could be a proxy for Rd which is difficult to sample requiring invasive procedures. This insight could lead to a new simpler approach to predict the risk of Type 2 diabetes in lean adults.

While iCircos appeared to help the researcher arrive at a new insight, he had difficulty in remembering which variable each ring represented. He therefore recommended three methods to address this short term memory constraint: (1) transient display of the variable name when the mouse moves over a ring, and persistent display based on a check box, (2) ability to collapse all panels except the Rings panel to enable all the ring selections to be visible, and (3) ability to rearrange the order of the rings and their radius through direct manipulation of the rings, and ability to drag the dot (close to the current sorted ring) to sort on other rings.

6. DISCUSSION

Our attempt to build a visual analytical prototype to enable the simultaneous and interactive analysis of data led to the following preliminary conclusions, which will guide our future research:

1. **Role of Interactivity in Discovery.** The obvious advantage of iCircos over the existing Circos system was the speed at which we could generate many different versions of the diagrams using different phenotype variables. However, we believe another important advantage of interactivity in the discovery process is that we were able to store *feedback from a sequence of changes* in short term memory, which enabled a more complex understanding of the data. For example, by sorting the rings and changing the edge weight thresholds in the asthma dataset, we comprehended the complex interactions because the feedback from multiple changes was available in short term memory. Interactivity therefore in addition to enabling more diagrams, could also be qualitatively changing how the data is being comprehended.

2. **Role of Heuristics in Visual Analytical Design**. Although the seven heuristics for designing visual analytical systems has only recently been proposed [12], they helped us to synergistically explore the space of possible methods to add interactivity to iCircos. However, iCircos depends on clusters derived from network analysis, strongly suggesting the need to understand how multiple representations complement each other to deal with complex analytical problems like the ones we have discussed.

3. **Role of HTML5/SVG in Visual Analytical Design**. Despite our limited experience in HTML5 and SVG, it was possible to develop a working prototype of iCircos in a period of 5 weeks. However, we experienced two important hurdles: (1) **Browser incompatibility** in dealing with the opening of local files from a local running html file. For example, Mozilla FireFox allows opening of local files, while Google Chrome does not. (2) **Slow performance** in the generation of the Circos diagram, which currently takes approximately 4 seconds. Our future experiments with this new programming platform will enable us to explore how to address these issues so we can add direct manipulation to the diagram such as dragging a ring to change its radius.

7. CONCLUSIONS

To enable TBI researchers to simultaneously and interactively explore associations between patients and their molecular and phenotype information, we developed a visual analytical prototype and demonstrated its usefulness by using it to derive insights in two biomedical datasets. Our preliminary results suggest that TBI researchers could benefit by using a combined and interactive representation of molecular and phenotype information to rapidly detect complex patterns. Our results also suggest that developers of visual analytical systems can rapidly develop useful interactive systems guided by visual analytics design heuristics and by using HTML5 and SVG. However, performance issues need to be addressed for adding direct manipulation to the diagram. Accordingly, our future research and development efforts will focus on exploring other programming environments that are more appropriate. Finally, we plan to conduct a rigorous user test to ensure that the prototype is useful and usable for analyzing complex multidisciplinary datasets, with the goal of accelerating discoveries in translational bioinformatics.

8. ACKNOWLEGEMENTS

This research was supported in part by NIH grants 1U54RR02614 UTMB CTSA(ARB), and CDC/NIOSH # R21OH009441-01A2. We thank N. Abate, A. Narain, G. Vallabha, A. Ganesan, M. Vallabhaneni, and S. Dasari, for their contributions.

9. REFERENCES

[1] Bhavnani, S. K., et al. How Cytokines Co-occur across Asthma Patients: From Bipartite Network Analysis to a Molecular-Based Classification. *Journal of Biomedical Informatics* (in press).

[2] Bhavnani, S. K., Pillai, R., Calhoun, W. J., and Brasier, A. R. How Circos Ideograms Complement Networks: A Case Study in Asthma. *Proceedings of AMIA Summit on Translational Bioinformatics* (2011).

[3] Brasier, A.R., Victor, S., et al. Molecular phenotyping of severe asthma using pattern recognition of bronchoalveolar lavage derived cytokines. *J Allergy Clin Immunol* 2008; 121:30–37.

[4] Butte, A. J. Viewpoint Paper: Translational Bioinformatics: Coming of Age. *JAMIA* 15(6): 709-714 (2008).

[5] Card, S., MacKinlay, J., and Shneiderman, B., eds: *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann 1998.

[6] Craig, L. *Agile and Iterative Development: A Manager's Guide*. (2004). Addison-Wesley.

[7] Filippova, D., Shneiderman, B. Interactive Exploration of Multivariate Categorical Data: Exploiting Ranking Criteria to Reveal Patterns and Outliers. *HCIL Tech. Report #38* (2009).

[8] Johnson, B., and Shneiderman, B. Treemaps: a space-filling approach to the visualization of hierarchical information structure. *Proceedings of IEEE Visualization Conference* (1991), 284–291.

[9] Krzywinski, M. et al. Circos: an Information Aesthetic for Comparative Genomes. *Genome Res* (2009) 19:1639-1645.

[10] Nielsen, J., and Mack, R. L. (Eds.), *Usability Inspection Methods*, (1994) John Wiley & Sons, New York, NY.

[11] Thomas, J., and Kristin A. Cook (eds.), *Illuminating the Path: The Research and Development Agenda for Visual Analytics* (Washington, DC: DHS, 2005).

[12] Yi, J. S., Kang, Y. A., Stasko, J., & Jacko, J. A. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, (2007), 13(6).