# Visual Comparison of Multiple Gene Expression Datasets in a Genomic Context

**Krzysztof Borowski[1,2], Jung Soh[1] and Christoph W. Sensen[1]**

[1]Sun Center of Excellence for Visual Genomics, University of Calgary, Faculty of Medicine, 3330 Hospital Drive NW, Calgary, AB, T2N 4N1, Canada

**Summary**

The need for novel methods of visualizing microarray data is growing. New perspectives are beneficial to finding patterns in expression data. The Bluejay genome browser provides an integrative way of visualizing gene expression datasets in a genomic context. We have now developed the functionality to display multiple microarray datasets simultaneously in Bluejay, in order to provide researchers with a comprehensive view of their datasets linked to a graphical representation of gene function. This will enable biologists to obtain valuable insights on expression patterns, by allowing them to analyze the expression values in relation to the gene locations as well as to compare expression profiles of related genomes or of different experiments for the same genome.

## 1 Introduction

The growth of information generated from microarray experiments has occurred at an exponential rate [1]. As microarray experiments involve high-throughput methods, they tend to create a large amount of data. Furthermore, with many genes per genome and many possible experiments designed to measure variations in expression of said genes, the potential amount of data generated from microarray experiments can be immense.

Numerical methods of comparing datasets of microarray experiments are important and clustering algorithms perform a large amount of expression data analysis. While statistical means of dealing with the data exist, statistical support of gene expression values can be cryptic; many genes have multiple functions in the organism, and not all changes at the mRNA production level are distinctly tied to the experiment at hand [2]. Thus, it is important to look further than just clustering algorithms and statistical analyses when contemplating microarray data [3].

Visual presentation of dataset is significant because it provides intuitive insights into patterns and leads to conclusions, which might be missed when using only statistical assessments. Thus, presenting the data in an informative, visually intuitive way should be an essential part of any microarray analysis program. Because of this need, visualization tools such as Bluejay (Browser for Linear Units in Java, `http://bluejay.ucalgary.ca`) gain value as part of the toolset of biological researchers. As a browser of linear biological data, Bluejay is able to display entire genomes in a way that allows easy access to further information about each gene. Through integrated JavaTM packages, Bluejay is able to present microarray experiment data in a genomic context, which makes it a useful, scalable, integrated package for viewing both genomes and their respective expression datasets [4]. Bluejay provides both intuitive

[2]To whom correspondence should be addressed. E-mail: `kkborows@ucalgary.ca`

visualization methods as well as the algorithms often used in mathematical microarray analysis, through the integration of TIGR MeV (TIGR MultiExperiment Viewer, `http://www.tm4.org/mev.html`), an open-source microarray analysis package built using Java [5]. This allows Bluejay to incorporate the analysis abilities of TIGR MeV (e.g., clustering) into its own set of microarray tools.

Bluejay also has the ability to display multiple genomes concurrently and link similar genes using visible lines. When viewing multiple genomes simultaneously, Bluejay displays the information in the following way: there is always the main genome and additional genomes added either to the outside of the current genomes loaded (if the genome shape option is set to `Circular`, as in Figure 4), or above them (if the genome shape option is set to `Linear`, as in Figure 5). Straight lines connect similar genes on all genomes displayed.

Bluejay currently has the ability to present a single instantaneous microarray dataset in a single bar-chart lane. Enabling the user to see expression data in relation to the gene itself places the microarray dataset in a genomic context, which creates a link between expression and function. This bar-chart lane can be changed to load one dataset in `Single` mode, and can be used to iterate through multiple time points in a single experiment in `Player` mode (Figure 1).

However, multiple datasets could not be shown in unison in Bluejay. Because the user was only able to see one instance of data at a time, visual information about patterns in one dataset could be forgotten when another is loaded. While tools exist that display multiple gene expression datasets together, the limitation to singular dataset display is still the case with most microarray visualization tools [6, 4]. To improve effectiveness of Bluejay as a microarray data visualization tool, especially to work with the multiple genome comparison capability, we have now added the ability to visualize multiple datasets simultaneously.

We have enhanced the microarray data visualization capability of Bluejay such that it can: (i) display multiple microarray datasets for one genome simultaneously, with multiple lanes of data display within the loaded genome at the same time; and (ii) display multiple microarray datasets for multiple genomes simultaneously, which is a natural integration of Bluejay's whole genome comparison functionality with multiple microarray dataset visualization for a single genome.

## 2 Methods

The genomic context offered by Bluejay is apparent with expression data corresponding to genes in question for each expression ratio displayed (Figure 1). Bluejay sets two drawing areas for microarray data, one for positive ratios and one for negative ratios. Both the up-regulation and down-regulation areas will be referred to as one lane of microarray data. Bluejay handles lane allocation by a system of lane requests from a static context in Java in order to limit lane creation and allow for a set lane quantity.

The number of datasets that can be displayed simultaneously rose from one to three (an upper bound) per genome. For example, Figure 2 shows a case of displaying three datasets for a genome. Importantly, a new painter class was added to Bluejay to facilitate allocating and painting multiple microarray data lanes instead of just one. With Bluejay's ability to show multiple genomes at a time, multiple microarray lanes were created for each genome loaded. The GUI for dataset selection is located within the `Experiments` tab in Bluejay, and presented

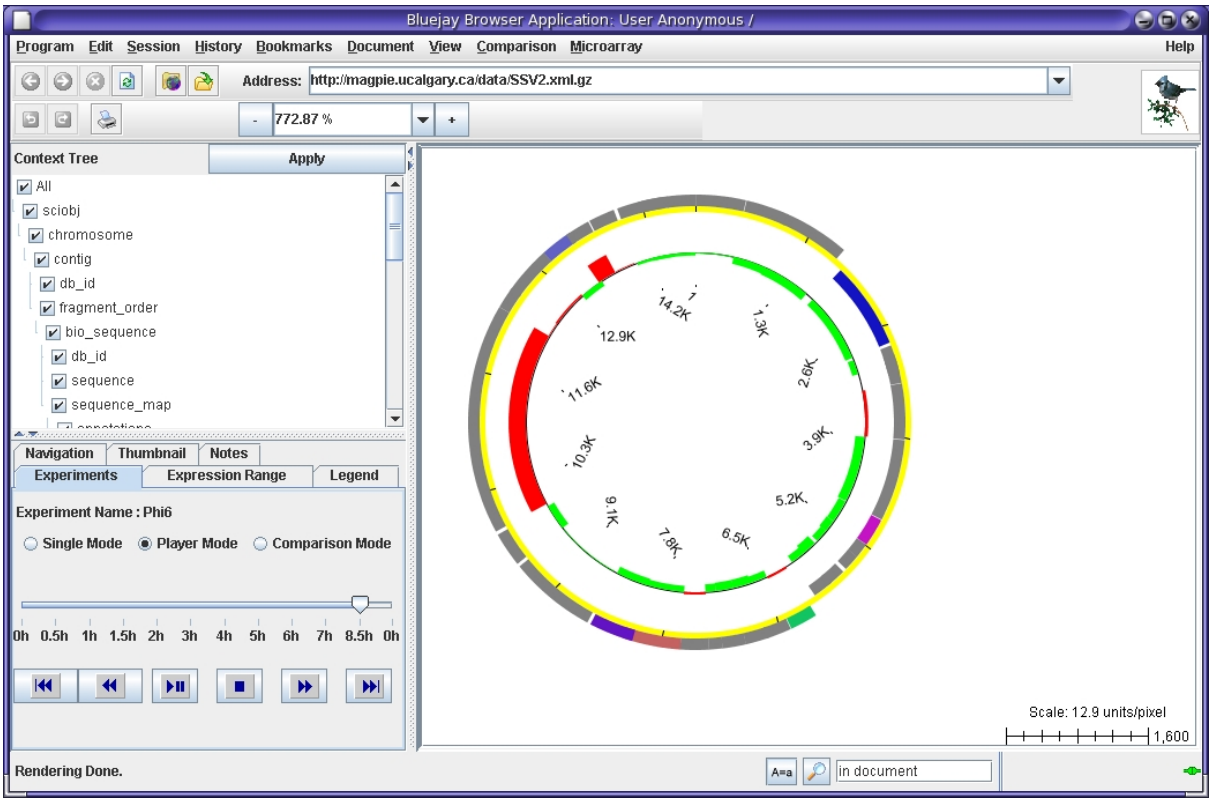


**Figure 1: Bluejay in microarray `Player` viewing mode, with a dataset for *Sulfolobus spindle-shaped virus* 2 (SSV2) at 8.5 hours after infection loaded. Up-regulated genes are shown by red outer bars on the microarray lane, down-regulated genes by green inner bars.**

as a microarray visualization mode named `Comparison`, in addition to the existing `Single` and `Player` modes. The user can click a dataset or multiple datasets from the loaded dataset list, to enable the display of the chosen datasets. The maximum expression datasets that can be displayed per genome is limited to a reasonable number to limit clutter, which can be increased or decreased programmatically if desired.

Due to the changes, analysis of dataset selection takes place before painting. In `Comparison` mode, the painting process is continued through classes until it reaches the specific painter class for the given object undergoing painting. The new painter class is then responsible for lane allocation and painting itself. Bluejay takes the selected datasets into consideration, and allocates a new microarray lane to provide space on the display for each of the selected microarray datasets within the applicable genome. All gene objects with microarray data are then processed by the painter class, which decides which lane within each genome will be the lane the current expression value is painted upon. The display area devoted to this specific expression value currently being painted is then filled with either of the positive or negative expression values, and a rectangular shape directly aligned with the gene on the genome lane is shown.

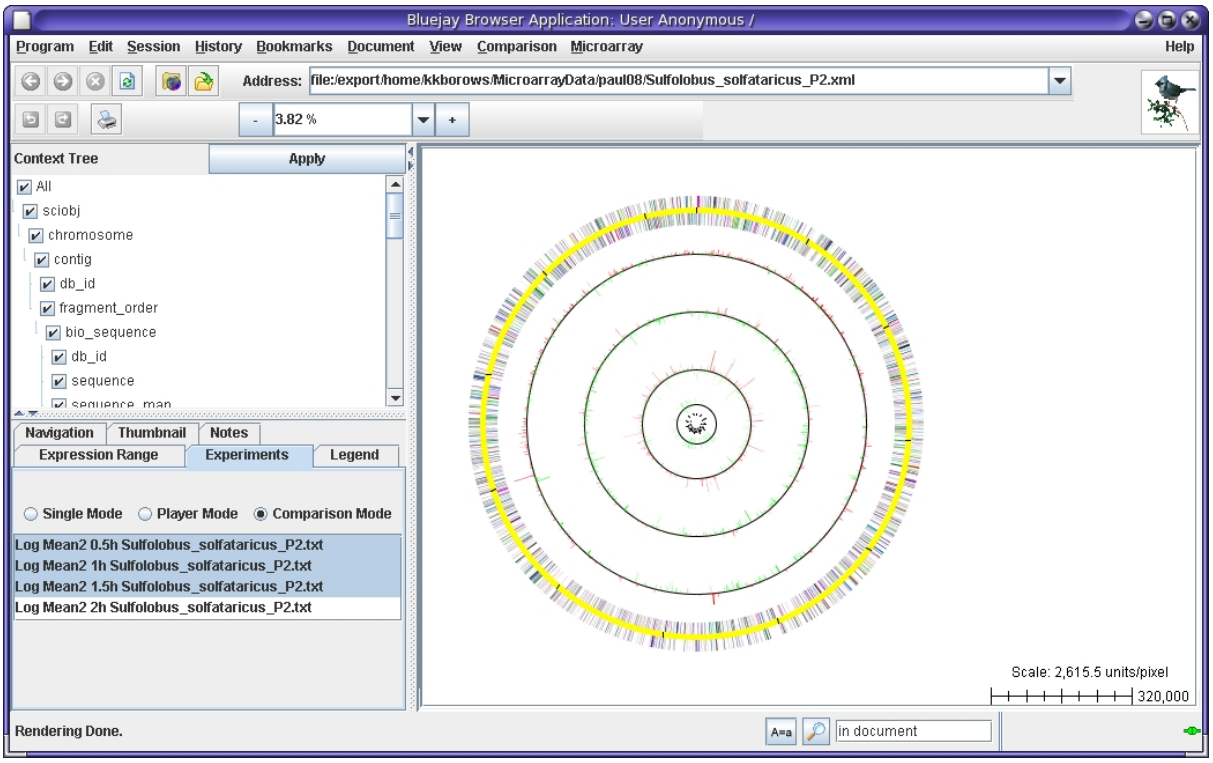


**Figure 2: Multiple microarray data (*Sulfolobus solfataricus* at different times after UV exposure) loaded in Bluejay and displayed simultaneously. Three different expression lanes corresponding to 0.5, 1, and 1.5 hours after UV exposure, respectively, are shown inside the *Sulfolobus solfataricus* genome.**

# 3 Results

## 3.1 Simultaneous display of multiple microarray datasets for one genome

A set of expression data from a *Sulfolobus solfataricus* UV irradiation experiment [7] is used to show the addition of multiple lanes in the `Comparison` mode of Bluejay's microarray view. Figure 2 shows that the user can control the selection of particular microarray datasets. The lanes with red and green expression bars are expression lanes. The `Experiments` tab on the left side of the image has the options of `Single`, `Player` and `Comparison` modes. In `Comparison` mode, a selection list is given from which the user can select the expression dataset(s) they want displayed. The expression data lanes are ordered in the same order as in the selection menu. A clearer view of the same data is seen in Figure 3, where semantic zooming is also presented.

Upon each selection, the canvas is repainted to include the newly selected datasets. Selecting a dataset can be done in three ways: single, multiple, and range selections. This follows the familiar file selection method of a window system. If only one dataset needs to be displayed, simply clicking on a dataset is necessary. If adding another to an already selected dataset, holding down the `Ctrl` key and selecting a dataset will paint the original and additional dataset, which can be repeated. Finally, to select a consecutive number of datasets, selecting the first dataset, holding down the `Shift` key and then selecting the final dataset the user desires to choose will cause the program to paint the consecutive datasets.

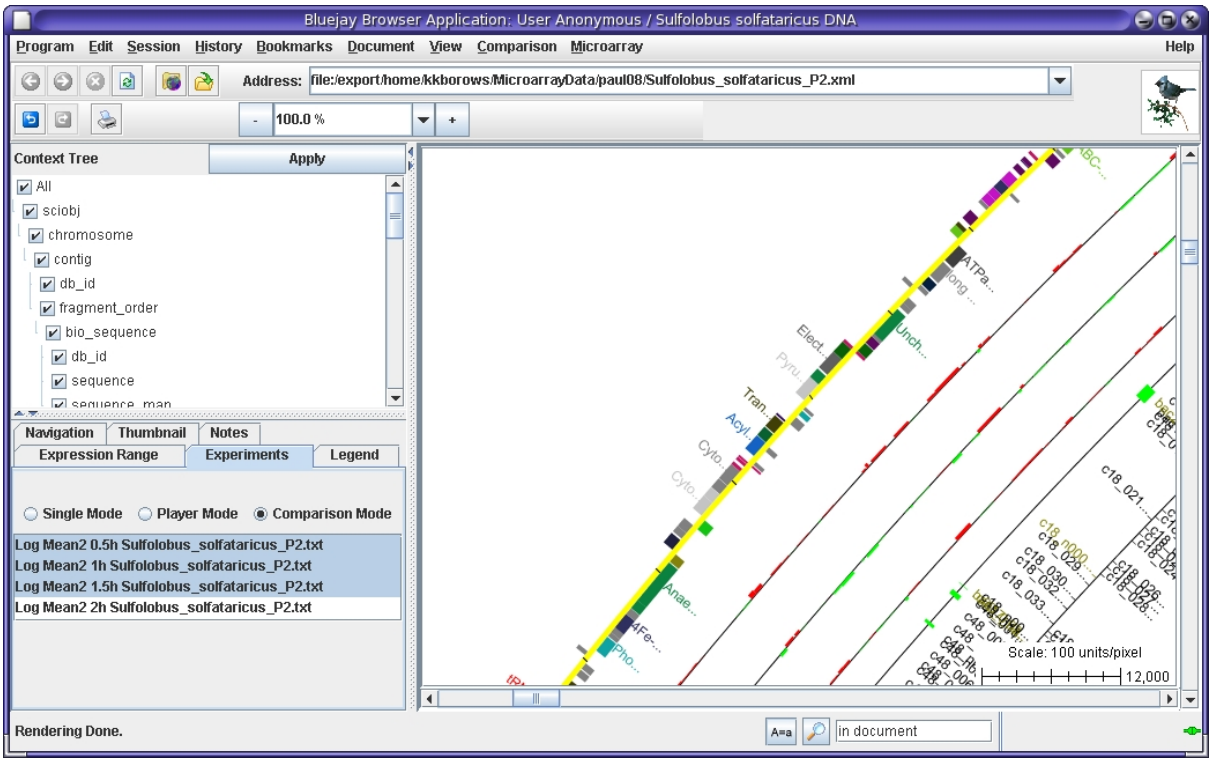


**Figure 3:** ***Sulfolobus solfataricus*** **from Figure 2, with the same datasets selected. This is a zoomed-in view showing detailed expression comparisons, focused on a specific area of the genome.**

### 3.2 Simultaneous display of multiple microarray datasets for multiple genomes

Figure 4 shows the result of adding the *Sulfolobus spindle-shaped virus* 2 (SSV2) genome as the slave genome on top of the *Sulfolobus spindle-shaped virus* 1 (SSV1) master genome, and then loading and selecting the sample microarray data provided with the Bluejay. Selections are presented in the same order as in the dataset list for each separate genome; the selected datasets which correspond to the master genome will appear from the outside to the inside, with the first set on the outermost microarray lane and the third set on the innermost microarray lane. The same will occur with the datasets selected for the slave genomes. The display order is equivalent to the single genome visualization previously described and follows the same rules; a dataset corresponding to a specific displayed genome will be displayed only within that genome, and ordered in the same way as the selections in the dataset list. If the genome in question is not loaded, the data will not be shown.

As is evident from Figure 4, too many lanes with expression information may result in compressed images at some viewing levels. This causes visually relating the sizes of expression bars to be difficult, as shown with the 1 hour dataset microarray lane of SSV1, where the abundance of data creates an obvious inability to paint the data with enough resolution to distinguish expression levels. The solution to this is found by either zooming into the display (for an effect similar to the one seen in Figure 3) or changing the genome shape from `Circular` to `Linear`. Figure 5 presents the linear view of the same genomes and microarray datasets as in Figure 4. Here we see a more pronounced view of the bar-chart presentation of microarray datasets. The values in the bar-chart rise in the positive vertical direction to represent up-regulation of a gene, and fall downwards in the negative vertical direction to represent down-regulation. Similar genes that are connected by lines can be found easily across genomes so that the associated

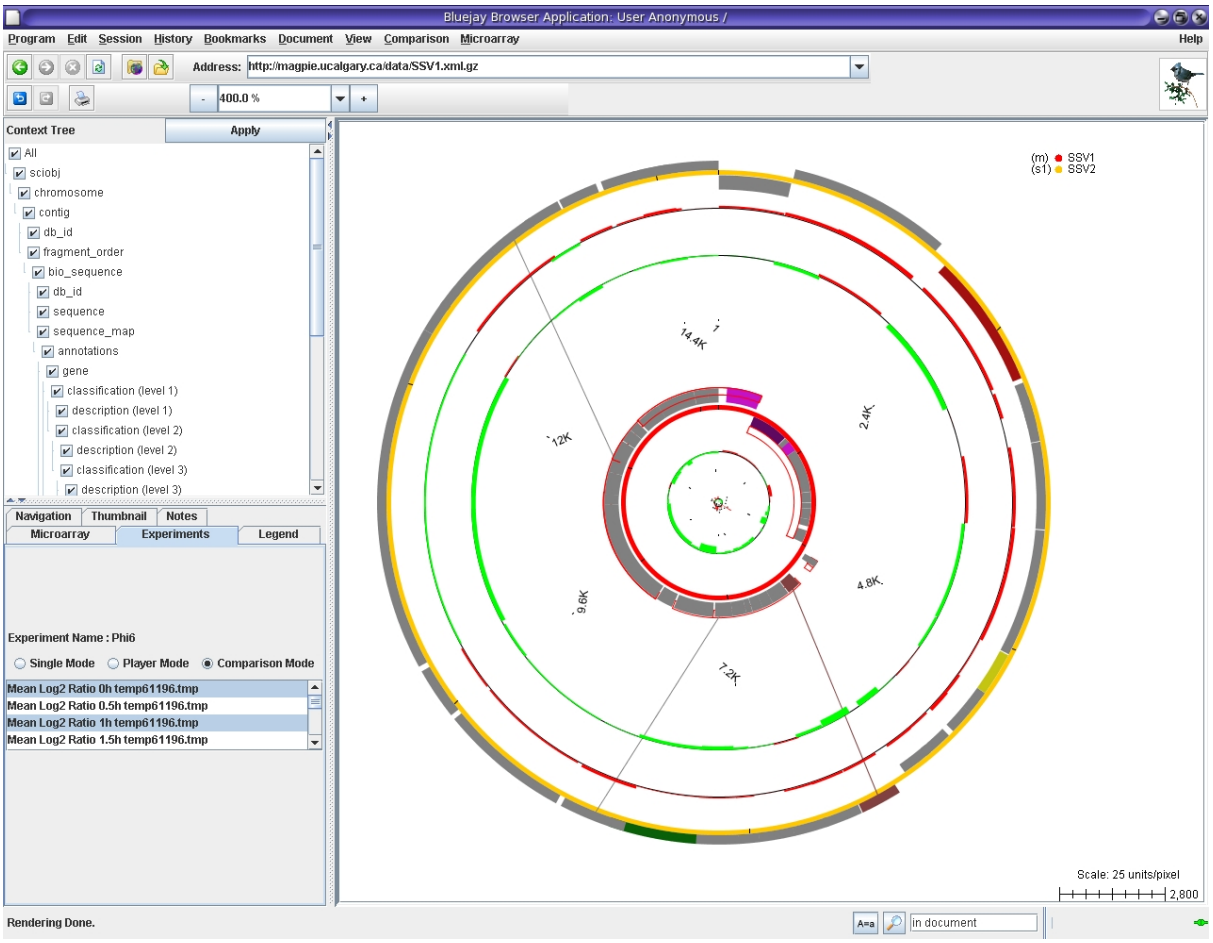


**Figure 4: Multiple microarray datasets for multiple genomes are displayed. SSV1 is loaded as master and SSV2 loaded as slave1. Datasets loaded correspond to 0 and 1 hour for both genomes.**

expression datasets compared quickly and efficiently.

## 4 Discussion

Hibbs et al. [8] uses a heat map format to show similarities of multiple expression datasets. Bluejay's native visualization methods do not deal with the heat map representation that Hibbs et al. [8] and many others use [3]. By creating a visual relation to the gene and its location on the entire genome, Bluejay integrates expression data into a genomic context instead of simply providing more data at a time.

ChARMView [9] shows multiple experiment display in a chromosomal context, aligning expression values to chromosome locations to see differences in expression. This chromosomal context is based on the same essential idea of providing the genomic context as in Bluejay. ChARMView provides a way to see chromosomal breaks in repeated microarray experiments and chromosomal aberrations, by visualizing multiple independent microarray datasets. ChARMView does some statistical analysis, but currently is not able to do the wide range of clustering analyses that Bluejay is capable of through TIGR MeV.

As microarray data can be accurate to a limited extent; repeated experiments are often done to verify results [10, 11]. Bluejay can be used to visually compare two experiments to confirm

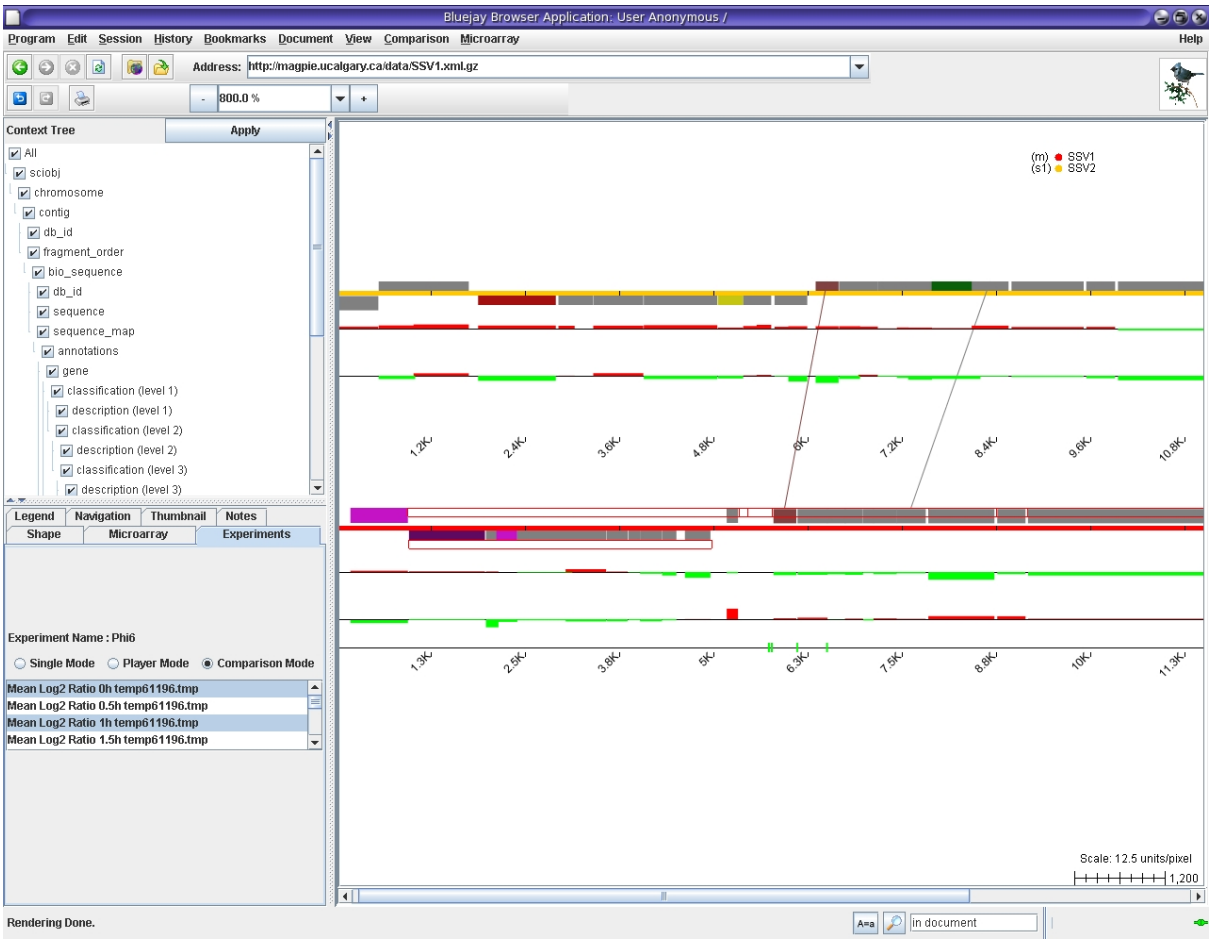


**Figure 5: `Linear` shape view of the same genomes and microarray datasets as in Figure 4**

reliability of the expression data, thanks to the multiple dataset view. Visual comparison is not limited to such data verification purposes, however. Wang et al. [12] describes a method for quick bird virus detection based on microarray data. The method involves comparing microarray results with the "naked eye", and claims to be a rapid way of distinguishing between a non-infected sample and Newcastle disease or avian influenza. Bluejay's multiple microarray experiment display will be useful for diagnosis of infection or disease by such methods. Furthermore, template expression files could be prepared to allow quick comparison and discovery of abnormal expression. Joining this potential for results verification and diagnostics is the potential for further insight development using Bluejays genomic context display.

Zhang et al. [13] suggest that well-designed visualization tools for microarray experiment data can facilitate immediate recognition of unknown patterns and unexpected relationships by letting the user focus on the areas of interest. They also classify a few task types that such tools may enhance: determining characteristics of unknown genes, clarifying diagnostic categories, extracting networks from databases and confirming clinical hypotheses. Bluejay has the ability to focus on genome areas and thus on expression values at a certain area of interest. The intuitive way of creating a link between gene function and gene expression found in Bluejay, together with the current improvement of allowing multiple dataset display in this same genomic context, facilitates the potential discoveries in the task types.

Wilkinson [14] states that avoiding clutter, maintaining simple geometric forms, sorting and organizing, as well as annotating data is essential to proper data presentation. Differences in expression values are presented in Bluejay through the differences in the heights of the bars drawn, direction (inwards and outwards, or upwards and downwards), and color (red and green). The painting of microarray data alongside the genome elements uses simple geometric shapes and is organized in a genome-encapsulated order, so that datasets for a specific genome are only displayed within that genome.

Tufte [15] states that removing irrelevant detail while displaying data is a must, explaining the effectiveness of providing detailed data on demand. In Bluejay, gene annotation is provided by a simple click on either the gene itself or the expression value, and tooltip generation by mouse hovering provides the expression values of all displayed datasets in a numerical fashion. Zooming in and out of the genome and thereby changing data focus is done with ease by the user. Focusing on datasets at various levels is helpful not only in the search for patterns and hypotheses, but also when presenting results of dataset analysis.

Integrating information from two different species is an interesting new area. Homologous genes exist in different organisms, and research into specificity of functional relationships between genes has tried to solidify the knowledge about the relationship by finding it in multiple organisms [16]. Allowing Bluejay to correctly display homologous gene expression data from two different organisms would make contributions to functional gene annotation as well as to the creation of biological network models [11]. Seeing these changes relative to their location on the genome may impact hypothesis creation by researchers.

## 5 Conclusions

Simultaneous display of expression data from multiple experiments is useful in determining patterns which selected statistical algorithms cannot [9]. Being able to view visual information across multiple experiments or at different times within one experiment provides a deeper understanding of changes in the expression rate of specific genes; a gene in a cluster at an hour into an experiment may exist in an entirely different cluster two hours in. Thus, Bluejay employing multiple expression dataset display functionality, together with its clustering analyses derived from TIGR MeV, would provide researchers with a powerful view of visible changes in clustering of genes or gene groups at specific locations on the genome. Visual displays of information can show variances or similarities between experiments and could lead to defining operon areas of the genome, when expression values coincide across a group of neighboring genes.

Bluejay is already an extensive microarray and genome viewing toolset [4], and the newly added multiple microarray dataset display capability enhances its utility even more. Scientists will want to find significant patterns by comparing multiple expression datasets in one comprehensive view. The genomic context of expression values is explicit in Bluejay and allows researchers to gain insight on expression patterns due to the concurrently available information on actual gene location and function. Domain-relevant insight into the data is given by Bluejay's handling of expression data, which may lead to potential discoveries not easily found from line graphs or heat maps.

## Acknowledgements

We would like to thank Andrew Ah-Seng, Paul Gordon, and Anguo Dong for their contributions to the earlier microarray experiment data visualization and genome comparison modules of Bluejay. This work was supported by Genome Canada/Genome Alberta through Integrated and Distributed Bioinformatics Platform for Genome Canada, as well as by the Alberta Science and Research Authority, Western Economic Diversification, National Science and Engineering Research Council, Canada Foundation for Innovation, and the University of Calgary. Christoph Sensen is the iCORE/Sun Microsystems Industrial Chair for Applied Bioinformatics.

## References

[1] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. NCBI GEO: Mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Research*, 35(Database issue):D760–D765, 2007.

[2] W. A. Rensink and S. P. Hazen. Statistical issues in microarray data analysis. *Methods in Molecular Biology*, 323:359–66, 2006.

[3] T. Werner. Bioinformatics applications for pathway analysis of microarray data. *Current Opinion in Biotechnology*, 19(1):50–54, 2008.

[4] A. L. Turinsky, A. C. Ah-Seng, P. M. Gordon, J. N. Stromer, M. L. Taschuk, E. W. Xu, and C. W. Sensen. Bioinformatics visualization and integration with open standards: The Bluejay Genomic Browser. *In Silico Biology*, 5(2):187–198, 2005.

[5] A. I. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, and J. Quackenbush. TM4: A free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2):374–378, 2003.

[6] P. Saraiya, C. North, and K. Duca. An evaluation of microarray visualization tools for biological insight. In *Proceedings of IEEE Symposium on Information Visualization*, pages 1–8, 2004.

[7] D. Gotz, S. Paytubi, S. Munro, M. Lundgren, R. Bernander, and M. F. White. Responses of hyperthermophilic crenarchaea to UV irradiation. *Genome Biology*, 8(10):R220, 2007.

[8] M. A. Hibbs, N. C. Dirksen, K. Li, and O. G. Troyanskaya. Visualization methods for statistical analysis of microarray clusters. *BMC Bioinformatics*, 6:115, 2005.

[9] C. L. Myers, X. Chen, and O. G. Troyanskaya. Visualization-based discovery and analysis of genomic aberrations in microarray data. *BMC Bioinformatics*, 6:146, 2005.

[10] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg. Protein interactions: Two methods for assessment of the reliability of high throughput observations. *Molecular & Cellular Proteomics*, 1(5):349–356, 2002.

[11] O. G. Troyanskaya. Putting microarrays in a context: Integrated analysis of diverse biological data. *Briefings in Bioinformatics*, 6(1):34–43, 2005.

[12] L. C. Wang, C. H. Pan, L. L. Severinghaus, L. Y. Liu, C. T. Chen, C. E. Pu, D. Huang, J. T. Lir, S. C. Chin, M. C. Cheng, S. H. Lee, and C. H. Wang. Simultaneous detection and differentiation of Newcastle disease and avian Influenza viruses using oligonucleotide microarrays. *Veterinary Microbiology*, 127(3–4):217–226, 2008.

[13] L. Zhang, J. Kuljis, and X. Liu. Information visualization for DNA microarray data analysis: A critical review. *IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews*, 38(1):42–54, 2008.

[14] L. Wilkinson. Presentation graphics. In N. J. Smelser and P. B. Baltes, editors, *International Encyclopedia of the Social and Behavioral Sciences*, volume 9, pages 6369–6379. Elsevier, New York, NY, USA, 2001.

[15] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 2 edition, 2001.

[16] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

[17] P. N. Seibel, J. Kruger, S. Hartmeier, K. Schwarzer, K. Lowenthal, H. Mersch, T. Dandekar, and R. Giegerich. XML schemas for common bioinformatic data types and their application in workflow systems. *BMC Bioinformatics*, 7:490, 2006.