Quantitative Visualization of ChIP-chip Data by Using Linked Views

Min-Yu Huang*, Gunther H. Weber†, Xiao-Yong Li‡, Mark D. Biggin‡ and Bernd Hamann*
*Institute for Data Analysis and Visualization (IDAV), Department of Computer Science
University of California, Davis, One Shields Avenue, Davis, CA 95616, USA
Email: {myhuang,bhamann}@ucdavis.edu
†Visualization Group, ‡Genomics Divisions
Lawrence Berkeley National Laboratory (LBNL), One Cyclotron Road, Berkeley, CA 94720, USA
Email: {ghweber,xyli,mdbiggin}@lbl.gov

***Abstract*—Most analyses of ChIP-chip *in vivo* DNA binding have focused on qualitative descriptions of whether genomic regions are bound or not. There is increasing evidence, however, that factors bind in a highly overlapping manner to the same genomic regions and that it is quantitative differences in occupancy on these commonly bound regions that are the critical determinants of the different biological specificity of factors. As a result, it is critical to have a tool to facilitate the quantitative visualization of differences between transcription factors and the genomic regions they bind to understand each factor's unique roles in the network. We have developed a framework which combines several visualizations via *brushing-and-linking* to allow the user to interactively analyze and explore *in vivo* DNA binding data of multiple transcription factors. We describe these visualization types and also provide a discussion of biological examples in this paper.**

***Keywords*-ChIP-chip Analysis; Visualization;**

I. Introduction

Chromatin immunoprecipitation followed by microarray analysis (ChIP-chip) [1] has been widely used to investigate interactions between transcription factors and DNA *in vivo* on a genome-wide scale. These experiments generate massive heterogeneous data sets, and many software tools have been developed to analyze them. These tools typically identify the locations of genomic regions bound by specific transcription factors. However, especially those that are web-based applications, provide only limited interactive operations. Furthermore, they lack the ability to help biologists compare and analyze the behavior of different transcription factors quantitatively. This is a serious limitation as recent studies [2] [3] in *Drosophila melanogaster* show that many factors bind quantitatively to overlapping sets of thousands of genomic regions *in vivo*. Regions bound at high levels are quite different in character from those more poorly bound, with only the more highly bound regions being functional. To better understand how transcriptional regulators behave in cells, it is thus important to have an analysis tool that can quantitatively analyze higher-order differences in DNA binding patterns between factors.

Here we present a framework to facilitate such analyses. It combines the track views present in a traditional genome browser along with a correlation table, scatter plots and parallel coordinates. The design is based on two important visualization principles: multiple views and *brushing-and-linking*. Multiple views allow the user to focus on different aspects of interest in their individual visualizations. While *brushing-and-linking* makes it possible to compare interactively the same selected data in alternate views along with their different contexts.

Section II summarizes other previous ChIP-chip analysis and visualization tools. Section III explains the input and data processing required for our tool. Section IV introduces the different visual components used in our framework. Section V demonstrates the use of our tool with biological examples. Finally, we present ideas for possible future research directions and conclude our paper in Section VI.

II. Previous Work

There are a number of integrated analysis tools for analyzing ChIP-chip data. For example, CisGenome [4] can perform basic analysis tasks, such as peak detection, false discovery rate computation, motif analysis and so on. Most tools are designed to do computationally intensive tasks rather than user-interactive analysis, or they analyze data for only one specific transcription factor at a time. They do not allow quantitative comparison of results for many factors at once directly within the tool.

Popular genome browsers, such as [5] [6] [7], use track views to display genomic data associated with each base pair position of a chromosome. Each track usually displays only properties such as sequence data, annotations from different gene models and experimental data from microarrays, ChIP-chip, etc. We use track views for the same purpose in our framework. Scatter plots are commonly used to illustrate the correlation and other relationships between two variables. Our visualization tool was inspired by GeneBox [8], which uses scatter plots to visualize the results of microarray experiments. Parallel coordinates were developed by Inselberg [9] [10] and Wegman [11] and are a common information visualization technique for high-dimensional data sets.

In the field of information visualization, linking multiple views to assist the user to explore and analyze high-dimensional or complex data is an established concept.

Examples include the work of Henze [12] in computational fluid dynamics data and the WEAVE system [13] that uses Physical Views and Information Visualization Views to explore cardiac data. Both methods use linked views to define features by combining selected subsets of the data in individual views. Our framework was also inspired by PointCloudXplore (PCX) [14], which links various physical views and abstract views to help scientists discover new relationships in 3D, cellular resolution gene expression data from *Drosophila* embryos.

III. Input and Data Processing

The ChIP-chip data obtained from experiments are noisy and cannot be directly used as input for analysis tools. In our implementation, we smooth the data by using a sliding window of 675 bp ("base pair") and compute the average of all data within this window. Other noise-reduction techniques can also be applied instead.

To understand the relationships among transcription factors and their roles in the genetic network, we start by calculating the correlations for all pair-wise comparisons among transcription factors, *e.g.,* f_a and f_b, f_a and f_c,f_b and f_c, etc. We focus on genomic regions for one of each pair of factors that have been identified as significant above some defined false discovery rate (e.g. 1% FDR). Any existing peak detection tool can be applied to determine these "primary peaks" of binding along with the relative ChIP-chip scores at each peak (*i.e.,* the implied level of transcription factor occupancy). Thus, for each transcription factor, the input for our framework should contain at least the genome wide ChIP-chip scores, their locations in base pair, and also the locations of detected primary peaks.

Because the primary peak of f_a does not imply f_b will also have a primary peak at the same location, we compute correlation coefficients between pairs of factors as follows.

i A transcription factor, f_a, is first chosen as the base.
ii A subset of its primary peaks are selected for a given analysis, for example, the top 100.
iii ChIP-chip scores of these two transcription factors, f_a and f_b, are looked up at the peak locations of f_a to form pairs of values. These pairs of ChIP-chip scores are used as coordinates for dots in a scatter plot and to calculate the correlation coefficient. A similar look-up operation for score pairs is also used for parallel coordinates.
iv This process is repeated by choosing other transcription factors in turn as the base.

IV. Visualizations

Our current framework consists of a correlation table, track views in the genome browser, scatter plots and parallel coordinates. All these views are coupled together via *brushing-and-linking*. Detailed descriptions of these components are provided in the following sub-sections.

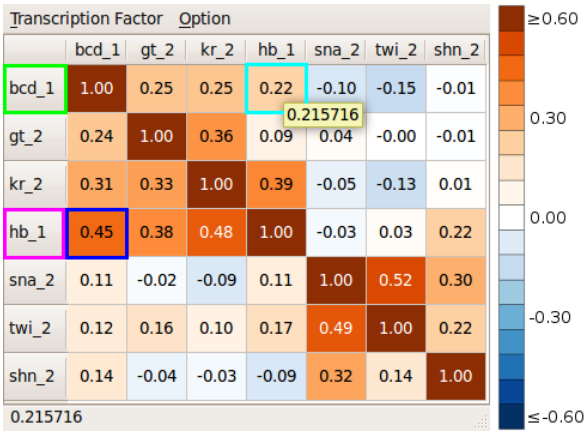


Figure 1. An example of the correlation table, showing correlation coefficients between all possible pair-wise comparisons among the transcription factors, *bicoid* (*bcd*), *giant* (*gt*), *krüppel* (*kr*), *hunchback* (*hb*), *snail* (*sna*), *twist* (*twi*) and *schnurri*(*shn*). The coefficients were calculated using ChIP-chip data from experiments using *Drosophila melanogaster* embryos. For simplicity, we only show two digits after the decimal point. The original value is shown in the tool-tip, or in the status bar when the cell is clicked for selection. The background color of each cell is interpolated using the color bar to the right (suggested by ColorBrewer, http://colorbrewer2.org) according to the correlation coefficient. Corresponding visualizations of cyan, blue, green and magenta boxes are shown in Figure 3(a), 3(b), 4(b) and 4(c), respectively.

A. Correlation Table

The correlation table window is the central graphical user interface (GUI) and the starting point of our analysis tools. Using this table, it is possible to load data sets of transcription factors and adjust parameters such as the number of primary peaks used in the correlation computation. The correlation table also triggers the creation of new views and hence serves as a view factory. Clicking on a table cell (*i.e.,*, a correlation coefficient) creates a new view with the scatter plot of that pair of transcription factors; clicking on a transcription factor name in the vertical table header shows a corresponding parallel coordinates view.

Figure 1 shows an example of the correlation table. In this table, the selected cell indicated by the cyan box represents the Pearson correlation coefficient of *bicoid* (*bcd*) versus *hunchback* (*hb*). Although other kinds of correlation coefficient, such as rank correlation, also could be used, a Pearson correlation places more emphases on the quantitative aspects of the data and thus is more appropriate. The coefficient was computed by first locating the top 100 primary peaks for *bcd*, and then finding the corresponding ChIP-chip scores for *hb* at those locations to form 100 pairs of data. Thus, each row in the table is computed based on the same locations of top primary peaks detected in the corresponding transcription factor shown in the vertical header. Because the primary peaks for each factor are not coincident in the genome, the table is not quantitatively symmetrical, *i.e.,*, *bcd* versus *hb* and *hb* versus *bcd* have different values.

B. Track View

The track view is used to display annotations, sequences, ChIP-chip scores and other quantitative data using nucleic base pairs as abscissa like those in a typical genome browser. The user can switch the chromosome they would like to explore through the graphics user interface. Multiple tracks can be added into the view to display different properties at once. Besides displaying ChIP-chip scores, the track view allows the user to discover relationships between a transcription factor and DNA, such as where binding regions are located. Other visualization tools in our framework focus instead on relationships among transcription factors.

Figure 2 shows the track view window of the example we discussed in Figure 1. We hide transcription factors other than *bcd* and *hb* in this window for simplicity. The user can generate multiple sets of selections, with the same color or different colors, in the track view window, and they will be shown in other views via *brushing-and-linking*. The three regions selected in Figure 2 are also shown in Figure 3 and Figure 4.

C. 2D Scatter Plot

The scatter plot is conceptually the simplest way to interpret each correlation coefficient in the correlation coefficient table. The relationship between two transcription factors is visualized in ChIP-chip score space. The X-axis represents the ChIP-chip score space of the reference transcription factor, while the Y-axis represents the ChIP-chip score space of the second factor. For example, Figure 3(a) shows the corresponding scatter plot of *bcd* versus *hb* selected in Figure 1; Figure 3(b) shows the scatter plot for *hb* versus *bcd*. Besides showing the correlation, scatter plots can easily show any non-linear relationship, support discovery of clusters of dots, or suggest more complicated relationship between two transcription factors in the plot.

Several components in the scatter plot window assist the user in navigating and interacting with the data. A gray bounding box represents the value range in ChIP-chip score space. It might be shown in an anisotropic coordinates for better visual results. Grids and scales are adjusted automatically when the user zooms in or out, or translates the view. Furthermore, an overview window at the corner provides a global context. In the overview window, the gray box indicates the value range in the ChIP-chip score space. A green box represents the viewing area of the primary scatter plot window in isotropic coordinates. The intersection point of the vertical line and the horizontal line in the overview indicates the center of the primary scatter plot window. This overview is particularly useful when the user zooms into the plot to explore only a small subsection. Chromosome selectors also provide means to explore different parts of the data. The user can also obtain each dot's basic information, such as its rank, base pair position, chromosome name and both transcription factors' ChIP-chip scores, by clicking on the dot.

D. Parallel Coordinates

Although one can use multiple 2D scatter plots to discover relationships among more than two transcription factors, it is not easy to work with multiple scatter plots at once. Parallel coordinates provide an alternative way to visualize or analyze high-dimensional or multi-variate data sets. Several examples are shown in Figure 4. Each vertical axis in the plots represents the ChIP-chip score space of an individual transcription factor, the scores being normalized by each factor's maximum score respectively. The red thick vertical part of one axis shows the score range of the base transcription factor used to compute correlation coefficients for the row in the correlation table. The blue thick vertical portions on the other axes represent the corresponding score ranges for the other factors. Corresponding ChIP-chip scores for transcription factors at the same base pair location are connected by the same polyline.

Since there are many polylines usually overlapping in the parallel coordinates, it can be difficult to see important information. We therefore implemented three enhancements to make it easier to discover relationships among factors. First, we added a *highlight mode* where the polyline under the cursor is enhanced while the other polylines are dimmed (Figure 4(c)). Detailed information about this highlighted polyline, such as rank and ChIP-chip scores, is displayed in the status bar. Second, we color polylines based on their rank for the base transcription factor (Figure 4(a)). This coloring scheme can help the user observe when peaks with different binding strength have different relationships with other transcription factors. Third, the user can re-arrange the order of axes by dragging them or the transcription factor's label in the horizontal header of the correlation coefficient table, allowing better visualization of relationships between neighboring axes. Our *brushing-and-linking* mechanism updates all parallel coordinates windows to reflect this changed order of axes.

E. Brushing-and-Linking

ChIP-chip scores of different transcription factors along with related sequence data and annotation information form a complicated high-dimensional data set. Although each of the different visualizations introduced in the previous subsections can help the user focus on selected dimensions to explore and define some important features, many other features are usually more complicated and cannot be analyzed or understood well by a single visualization. One intuitive way to understand a high-dimensional feature is by combining information projected onto its sub-spaces. In our framework, each visualization type conceptually represents a sub-space of the original data set. The user can select a subset of data in any visualization (except the correlation

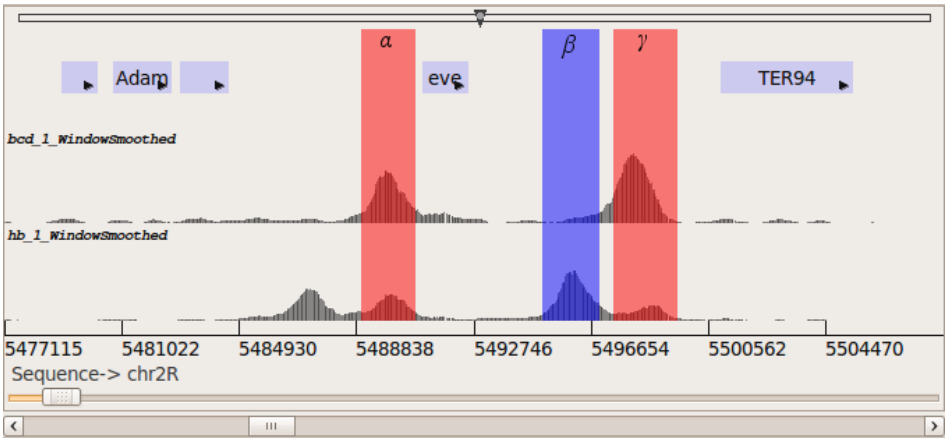
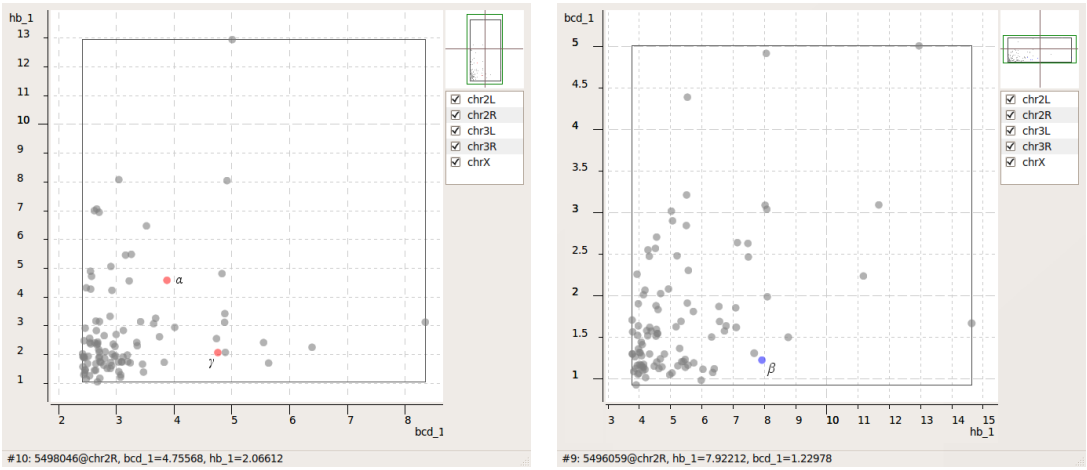


Figure 2. An example of the track view window. The three tracks shown represent gene annotations and ChIP-chip scores for *bcd* and *hb*, respectively. There are two regions brushed in red (α and γ) and one in blue (β) around *eve*.



(a) *bcd* versus *hb* (b) *hb* versus *bcd*

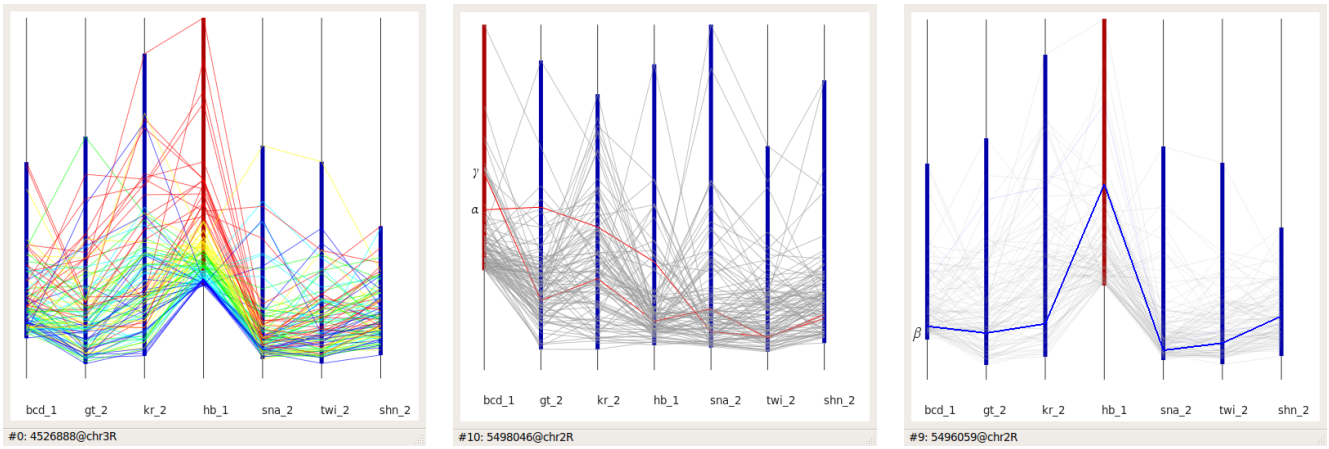
Figure 3. Shown are the corresponding scatter plots when a user clicks on the cells indicated by the cyan box (a) and the blue box (b) in the correlation table shown in Figure 1, respectively. In panel (a), the X coordinate of each dot represents the ChIP-chip score for each of the top 100 primary peaks in *bcd* and each dot's Y coordinate represents the corresponding ChIP-chip score in *hb* at the same base pair location. (b) shows the equivalent coordinates for *hb* on the X axis and *bcd* on the Y axis. Normal dots are shown in gray. Selections made in Figure 2 are also shown in these plots. The status bar shows the detailed information of the dot clicked by the user in the scatter plot.

table) to define a feature or area of interest. *Brushing-and-linking* allows the user to observe the same set of selected data in other visualizations, *i.e.,* in other sub-spaces. Our framework supports multiple brushes, *i.e.,* multiple sets of selection, in different colors to assist in the comparison of multiple features at the same time. Figure 2, Figure 3 and Figure 4 together provide an example of the use of *brushing-and-linking*.

V. Case Study

The tools described in this paper use the output of ChIP-chip experiments, files that contain the normalized probe-level enrichment scores genome wide, and the information of the bound regions identified above a defined FDR including the genomic location and the ChIP-chip score of each region. Our tool allows users without bioinformatics background to explore the functional relationships between factors by examining how the levels of binding by one factor or a group of factors at each genomic region correlates with those of other factors quantitatively.

We use transcription factors involved in *Drosophila* embryogenesis to demonstrate the usefulness of these tools. *Drosophila* embryo development is governed by several groups of transcription factors that coordinate different aspects of patterning. Along the anterior-posterior (A-P) body axes, pattern formation is initiated by the A-P early factors, *bcd*, *cad*, *hb*, *kr*, *gt*, *kni*, *tll* and *hkb*, which act in concert to regulate the latter patterning of the A-P pair-rule factors. Along the dorsal-ventral (D-V) axes, a separate group of D-V factors, including *dl*, *sna*, *shn* and *twi* regulate patterning.



(a) Rank-coloring based on *hb* (b) based on *bcd* (c) based on *hb*, in highlight mode

Figure 4. Examples of parallel coordinates. In (a), polylines are colored by their ranks in *hb* ChIP-chip scores. From high ranks to low ranks, red, yellow, green, cyan and blue represents 20 polylines, respectively. Panel (b) and (c) correspond to the parallel coordinates plots when a user clicks on the green box, and the magenta box in Figure 1, respectively. Color polylines in (b) and (c) represent selections made in Figure 2 via *brushing-and-linking*.

Factors that belong to the same functional group tend to regulate transcription via the same genomic regions, and as a result it is expected that their binding shows a preferential correlation. Figure 1 shows that indeed this is the case. When the binding levels for the top 100 most highly bound regions of each factor are compared, the binding by factors among the A-P early group (*bcd*, *gt*, *kr* and *hb*) tend to correlate more highly with each other than they do with members of the D-V group (*twi*, *sna* and *shn*). As has been shown previously, when additional factors are included in such analysis, more functional groups can be revealed [3]. Thus the correlation in binding levels can be useful guide to whether transcription factors are functionally related.

While such a correlation analysis can reveal functional relatedness between factors, the relationship between any two factors is usually more complex. Factors from different functional groups may regulate the same genes. For example, while the expression of most A-P and D-V factors are primarily regulated by members of the functional group they belong to, they are each to a small degree also regulated by factors of the other groups [15] [3]. Also, while members of the same class often share the same targets, there are important quantitative differences in the degree to which each factor controls these targets. Such complexities can be explored by either scatter plot or parallel coordinates plots. For example, as the scatter plots in Figure 3 show, while there are genomic regions that are bound strongly by both *bcd* and *hb* (*e.g.,* region α), other regions are bound strongly by *bcd* but weakly by *hb* (*e.g.,* region γ), and vice versa (*e.g.,* region β). Figure 2 also shows the same *bcd* and *hb* binding patterns to these three example regions using our genome browser view. These regions are each well studied enhancer elements of the *eve* gene: α is the *eve* stripe 2 enhancer through which *bcd* and *hb* synergistically activate *eve* transcription [16]; γ is the *eve* stripe 1 enhancer for which *bcd* is the chief activator [17]; and β, the *eve* stripe 4/6 enhancer which *hb* represses, while *bcd* has no known functional role [17].

Parallel coordinates plots also support exploration of complex aspects of binding by multiple factors. Figure 4(a) shows that *hb* binding overall has a higher correlation with the other A-P factors, *bcd*, *gt*, and *kr*, than it does with the D-V factors, *sna*, *twi*, and *shn*. However, some regions that are strongly bound by *hb* are at least modestly bound by D-V factors, suggesting that the associated genes may be regulated by both sets of factors. Figure 4(b) shows the example genomic regions α and γ highlighted, allowing one to quickly see that the *eve* stripe 2 enhancer (α) is bound strongly not only by *bcd* and *hb*, but also by *kr* and *gt* (two factors known to be important for defining the expression boundaries of *eve* stripe 2), but not by the other factors included in this analysis, whereas the *eve* stripe 1 enhancer (γ) is only strongly bound by *bcd*. Similarly, Figure 4(c) shows that site β (*eve* stripe 4/6) is bound only by *hb*.

VI. Possibilities for Future Research and Conclusions

Traditional ChIP-chip tools focus on fundamental analyses such as locating binding regions. We have presented a framework which combines several visualizations using to analyze higher-order relationships among different transcription factors in ChIP-chip data. We also have demonstrated using biological examples how easy-to-use and valuable this framework is for discovering functional relationships among factors.

There are still several challenges that remain to be tackled. One is to integrate more low-level analysis tools and also more visualization types into this framework to make it more

powerful and useful. Furthermore, although our framework works perfectly on a local computer, it is desirable to be able to access remote databases directly. When analyzing up to several hundred transcription factors, memory usage becomes a critical problem due to the massive size of each transcription factor's ChIP-chip data set. One could also further improve the efficiency of data structures and algorithms.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health through grant GM70444, by the Director, Office of Science, and Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

We thank Aaron Heckmer, author of the Berkeley Quantitative Genome Browser (BQGB, http://bdtnp.lbl.gov/bqgb/) which our system utilizes. The latest version can be found at http://sanitysewer.com/kelp/. We also thank the members of IDAV at UC Davis, and the members of the Berkeley Drosophila Transcription Network Project (BDTNP) and Visualization Group at LBNL.

REFERENCES

[1] O. Aparicio, J. V. Geisberg, and K. Struhl, "Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo," in *Current Protocols in Cell Biology*, 2004, ch. 17.7.

[2] X.-Y. Li, S. MacArthur, R. Bourgon, D. Nix, D. A. Pollard, V. N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, C. L. L. Hendriks, H. C. Chu, N. Ogawa, W. Inwood, V. Sementchenko, A. Beaton, R. Weiszmann, S. E. Celniker, D. W. Knowles, T. Gingeras, T. P. Speed, M. B. Eisen, and M. D. Biggin, "Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm," *PLoS Biology*, vol. 6, no. 2, February 2008.

[3] S. MacArthur, X.-Y. Li, J. Li, J. B. Brown, H. C. Chu, L. Zeng, B. P. Grondona, A. Hechmer, L. Simirenko, S. V. Keränen, D. W. Knowles, M. Stapleton, P. Bickel, M. D. Biggin, and M. B. Eisen, "Developmental roles of 21 drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions," *Genome Biology*, vol. 10, no. 7, July 2009.

[4] H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong, "An integrated software system for analyzing chip-chip and chip-seq data," *Nature Biotechnology*, vol. 26, no. 11, pp. 1293–1300, November 2008.

[5] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The human genome browser at ucsc," *Genome Research*, vol. 12, no. 6, pp. 996–1006, June 2002.

[6] J. W. Nicol, G. A. Helt, J. Steven G. Blanchard, A. Raja, and A. E. Loraine, "The integrated genome browser: free software for distribution and exploration of genome-scale datasets," *Bioinformatics*, vol. 25, no. 20, pp. 2730–2731, October 2009.

[7] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp, "The ensembl genome database project," *Nucleic Acids Research*, vol. 30, no. 1, pp. 38–41, January 2002.

[8] N. Shah, V. Filkov, B. Hamann, and K. I. Joy, "Genebox: Interactive visualization of microarray data sets," in *International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences*, 2003, pp. 10–16.

[9] A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1, no. 2, pp. 69–91, August 1985.

[10] ——, *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. Springer, 2009.

[11] E. J. Wegman, "Hyperdimensional data analysis using parallel coordinates," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 664–675, September 1990.

[12] C. Henze, "Feature detection in linked derived spaces," in *Proceedings IEEE Visualization*, 1998, pp. 87–94.

[13] D. L. Gresh, B. E. Rogowitz, R. L. Winslow, D. F. Scollan, and C. K. Yung, "Weave: A system for visually linking 3-d and statistical visualizations, applied to cardiac simulation and measurement data," in *Proceedings IEEE Visualization*, 2000, pp. 489–492.

[14] G. H. Weber, O. Rübel, M.-Y. Huang, A. H. DePace, C. C. Fowlkes, S. V. E. Keränen, C. L. L. Hendriks, H. Hagen, D. W. Knowles, J. Malik, M. D. Biggin, and B. Hamann, "Visual exploration of three-dimensional gene expression using physical views and linked abstract views," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 2, pp. 296–309, April 2009.

[15] J. Zeitlinger, R. Zinzen, A. Stark, M. Kellis, H. Zhang, R. Young, and M. Levine, "Whole-genome chip-chip analysis of dorsal, twist, and snail suggests integration of diverse patterning processes in the drosophila embryo," *Genes Developement*, vol. 21, no. 4, pp. 385–390, February 2007.

[16] S. Small, A. Balir, and M. Levine, "Regulation of even-skipped stripe 2 in the drosophila embryo," *The European Molecular Biology Organization Journal*, vol. 11, no. 11, pp. 4047–4057, November 1992.

[17] M. Fujioka, Y. Emi-Sarker, G. Yusibova, T. Goto, and J. Jaynes, "Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients," *Development*, vol. 126, no. 11, pp. 2527–2538, January 1999.