MulteeSum: A Tool for Comparative Spatial and Temporal Gene Expression Data

Miriah Meyer, *Member, IEEE*, Tamara Munzner, *Member, IEEE*, Angela DePace,
and Hanspeter Pfister, *Senior Member, IEEE*

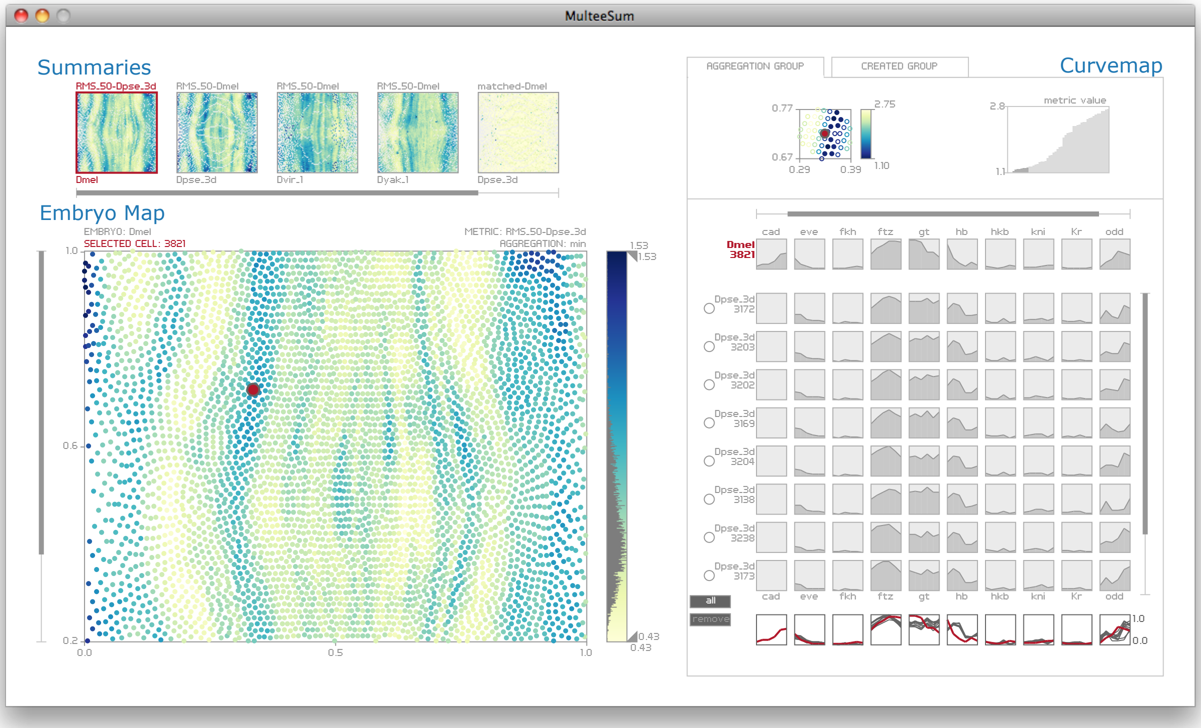


Fig. 1. MulteeSum is a visualization system that supports inspection and curation of data sets showing gene expression over time, in conjunction with the spatial location of the cells where the genes are expressed. It is the first tool to support comparisons across multiple such data sets. This screenshot includes data for four related species of *Drosophila*.

Abstract—Cells in an organism share the same genetic information in their DNA, but have very different forms and behavior because of the selective expression of subsets of their genes. The widely used approach of measuring gene expression over time from a tissue sample using techniques such as microarrays or sequencing do not provide information about the spatial position within the tissue where these genes are expressed. In contrast, we are working with biologists who use techniques that measure gene expression in every individual cell of entire fruitfly embryos over an hour of their development, and do so for multiple closely-related subspecies of *Drosophila*. These scientists are faced with the challenge of integrating temporal gene expression data with the spatial location of cells and, moreover, comparing this data across multiple related species. We have worked with these biologists over the past two years to develop MulteeSum, a visualization system that supports inspection and curation of data sets showing gene expression over time, in conjunction with the spatial location of the cells where the genes are expressed — it is the first tool to support comparisons across multiple such data sets. MulteeSum is part of a general and flexible framework we developed with our collaborators that is built around multiple summaries for each cell, allowing the biologists to explore the results of computations that mix spatial information, gene expression measurements over time, and data from multiple related species or organisms. We justify our design decisions based on specific descriptions of the analysis needs of our collaborators, and provide anecdotal evidence of the efficacy of MulteeSum through a series of case studies.

Index Terms—Spatial data, temporal data, gene expression.

1 Introduction

Even though most cells in the body share identical genetic information in their DNA, the functions and morphologies of the cells vary greatly both spatially, in different organs and tissues, and temporally during development and in response to the environment. This diversity is largely due to the selective expression of specific subsets of genes in different cells over time. These dynamic spatial patterns of gene

- *M. Meyer and H. Pfister are with Harvard University, E-mail: miriah:pfister@seas.harvard.edu.*
- *T. Munzner is with University of British Columbia., E-mail: tmm@cs.ubc.ca.*
- *A. DePace is with Harvard Medical School, E-mail: angela_depace@hms.harvard.edu.*

Manuscript received 31 March 2010; accepted 1 August 2010; posted online 24 October 2010; mailed on 16 October 2010.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org.

expression ultimately determine the morphology and physiology of an organism — understanding the genomic source of these patterns is a grand challenge in biology [3].

To address this challenge, the Berkeley Drosophila Transcription Network Project (BDTNP) [1] developed experimental and computational methods to extract gene expression information from every cell of the fruitfly *Drosophila melanogaster* embryo during a critical hour of development when highly dynamic patterns of gene expression determine the segmented body plan of the adult fly. These methods are being adapted to gather such data for multiple related fruitfly species [5]. Now, the scientists are faced with the challenge of integrating nonspatial gene expression measurement data with the spatial location of cells in an organism and, moreover, comparing this data across multiple related species. We collaborated with a group of these scientists to develop a visualization tool that enables this analysis.

In this work we present the design and implementation of MulteeSum, a multiple summary expression explorer, shown in Figure 1. MulteeSum is a visualization system that supports inspection and curation of data sets showing gene expression over time, in conjunction with the spatial location of the cells where the genes are expressed. Very few previous tools successfully integrate the temporal and nonspatial gene expression data with the spatial locations of cells. Furthermore, MulteeSum is the first tool to support comparisons across multiple such data sets. We do so with a general and flexible framework built around multiple *summaries* for each cell, allowing the biologists to explore the results of computations that mix spatial information, gene expression measurements over time, and data from multiple related species or organisms. The tool enables two specific types of functionality for their research: the ability to quickly view the entire gene expression profile over time for a given cell or group of cells, and to support the comparison of multiple related species. It is generally accepted in the biology community that comparative high spatial resolution expression data, whether derived from imaging or sequencing technologies, will become more prevalent. We therefore anticipate that our strategies to explore this type of data will be broadly useful.

In the rest of this paper we justify our design decisions with specific descriptions of the data and needs of our collaborators (Section 3), discuss our comparison support through the use of *groups* and *summaries* (Section 4), describe the design and evolution of MulteeSum (Sections 5 and 6), and provide anecdotal evidence of the efficacy of the tool through a series of case studies (Section 7).

2 Previous Work

Gene expression data is primarily derived from two experimental sources: microarrays or sequencing where many genes can be measured, but the spatial component is lost, and imaging, where many fewer genes are measured, but their spatial relationship is preserved. In this work, the data sets are generated using imaging techniques. The vast majority of gene expression information, however, is currently derived from microarrays, and the visualization convention is a nearly-universally used heatmap [26, 9]. In this matrix view, samples are mapped to columns and genes are mapped to rows, with each element in the matrix assigned a color based on the gene expression value [7, 22, 23]. These views, however, do not explicitly encode temporal relationships in the data.

An extension to the heatmap, called a *curvemap*, uses a time curve as the base visual unit in a matrix display, supporting the comparison of temporal expression data. This view is implemented in Pathline [17], a tool designed for comparative functional genomics data. The curvemap represents time series gene expression data as filled curves in a heatmap-style matrix layout, and includes overlay plots of curves along the rows and columns to facilitate the discovery of trends. We incorporate the curvemap display in MulteeSum for visualizing the temporal gene expression data from individual cells in the embryos.

Visualization strategies to integrate the spatial component of gene expression data are just emerging as high-throughput imaging based data becomes more prevalent. The few that exist take advantage of a 3D representation to plot the spatial expression pattern of a few genes at a single time point, such as the Allen Brain Atlas viewer [15] and PointCloudXplore [21, 25, 20]. The Allen Brain Atlas viewer maps gene expression values onto a 3D representation of a mouse brain, encoding the expression value of a single gene at a single time using color. PointCloudXplore was developed specifically to visualize the BDNTP data and supports the visualization of data for a single average model of a fruitfly embryo. In this tool, a single gene at a single time point is assigned a color which is then used to colormap each cell in a 3D embryo representation according to the expression value for that cell. The tool also supports a 2D embryo representation, which is the preferred view by our collaborators for analysis tasks because of the lack of data obstruction. Several genes can be rendered at once, however the expression of multiple genes in a cell leads to color blending which makes the expression values of individual genes indistinguishable. The tool thus includes a parallel coordinates view that plots, for every cell in the embryo, the values of the selected genes, as well as clustering capabilities to assign cells to groups with similar expression patterns.

For the current needs of our collaborators, PointCloudXplore has several drawbacks. First, in the tool, selecting multiple genes at the same time point produces an equivalent view as selecting one gene at multiple time points – for the analysis our collaborators are doing these two classes are not equivalent, and they are specifically interested in observing patterns in the genes over time. Second, the tool supports looking at a few genes for many cells, while our collaborators want to analyze the expression levels for the full set of genes in a just a few cells at a time. And third, PointCloudXplore does not support the comparison of multiple related species. We designed MulteeSum to specifically address these limitations.

3 Data and Tasks

This section describes our collaborators' data collection process and the resulting *virtual embryo* data sets as well as their tasks for exploring, understanding, and comparing these data sets.

3.1 Virtual Embryo

The fruitfly is an ideal model system to identify the genetic sources of variation in the dynamic gene expression patterns for several reasons. First, scientists have recently sequenced the complete genomes of 12 related species of *Drosophila* [6]. Second, the interactions of the set of genes that direct the segmentation of the adult body plan in the *D. melanogaster* embryo is well-understood and characterized in the *blastoderm* stage, which is the critical hour of development when the fate of cells is determined [13]. And third, it is now possible to measure the expression levels of many genes at cellular resolution using 2-photon fluorescent microscopy and image processing techniques [12, 8].

3.1.1 Data Collection

Using these imaging techniques, our collaborators measure the expression levels of a subset of genes in the blastoderm embryos of multiple fruitfly species. Briefly, thousands of embryos from a specific species are stained to show their nuclei and two genes, one gene of interest and one to serve as a registration marker. Each stained embryo is imaged using a 2-photon microscope, creating a stack of images at progressive depths. A compilation of an example image stack from the microscope is shown in Figure 2(a). The images in the stack are segmented and processed to produce a computational representation where each cell in the embryo has a unique id, a position, and quantitative measurements representing the amount of expression for the two genes in the nucleus. Finally, this computational representation is mapped to a template representing the average embryo morphology for the species. This process is repeated for all of the stained embryos, resulting in a *virtual embryo* (VE), where the average expression over time for an arbitrary number of genes is recorded per cell (this is referred to as a *virtual point cloud* in previous work [21, 25]). Currently, our collaborators have a complete VE for 1 species (*D. melanogaster*, or `Dmel`), preliminary data for 3 species (*D. pseudoobscura*, or `Dpse`; *D. virilis*,

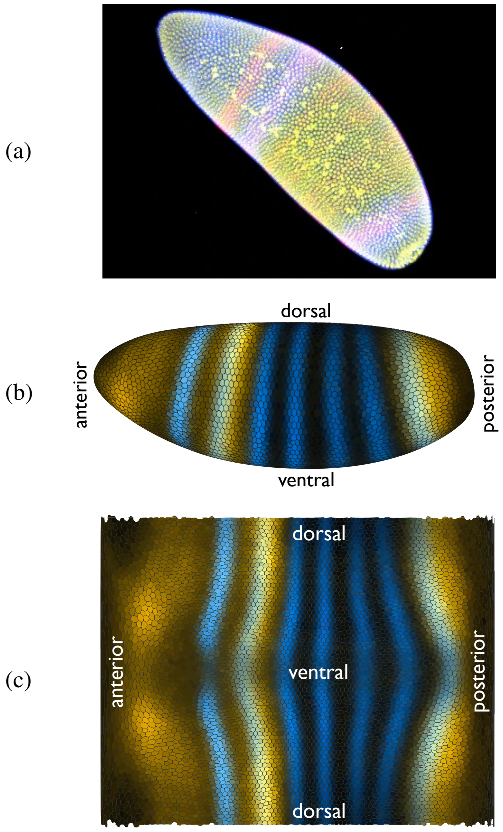


Fig. 2. (a) A compilation of an image stack of a Dmel embryo stained for two genes, eve (expressed as stripes, stained in red) and hb (expressed along the two ends, stained in blue) and nuclei (stained in green). Before channel unmixing, which takes place during image processing, there is substantial overlap in the green and red channels, resulting in the yellow color in the mid-section of the embryo. (b) Thousands of image stacks are processed and combined to build a 3D embryo representation, shown here for Dmel and colormapped with a single time step of eve shown in blue and hb shown in yellow. (c) The unique geometry of the Drosophila blastoderm embryos allows for a lossless parameterization to a 2D *embryo map* representation. The images in (b) and (c) were made using PointCloudXplore [21].

or Dvir; *D. yakuba*, or Dyak), and are in the process of collecting data for another 8 species.

3.1.2 Spatial Layout of Cells

During the blastoderm stage there is no cell division occurring, and all of the cells lie on the surface of the embryo. Thus, in the resulting virtual embryo representation the number of cells does not change over time, and the virtual embryo is hollow inside. This first characteristic allows for a static 3D representation of the VE, shown in Figure 2(b). The second characteristic allows for a lossless parameterization of the cell positions in the VE using a cylindrical projection, shown in Figure 2(c). We call this 2D representation an *embryo map*.

Each VE has around five thousand cells, but that number can vary up or down from species to species by approximately one thousand. For example, the Dmel VE contains 6078 cells, the Dpse VE contains 4970 cells, the Dvir VE contains 5476 cells, and the Dyak VE contains 6162 cells. The templates for each VE used in the data acquisition process are registered to a common 3D coordinate frame using the two major axes of the embryos, which allows the biologists to compare groups cells between species in roughly the same areas, such as cells making up the head or tail. There is not a one-to-one mapping of cells between the species, however, because of variations in the virtual embryo shapes, the numbers of cells, and distribution of cellular densities. These variations add a layer of complexity when comparing VEs, which we discuss further in Section 4.1.

The shape variation is eliminated in the 2D VE representation as the embryo maps share a common coordinate system across all the species, which is important for some comparative analysis tasks. The cell positions within each embryo map, however, are not the same.

3.1.3 Temporal Expression Profiles

When staining a batch of embryos to create a VE, each individual is stained for one gene of interest from approximately 50 different genes. The number of genes is fixed due to the size of the gene set involved with determining the body plan of an adult fly. The full set of genes is available for the Dmel VE, while the preliminary Dpse VE contains 12 genes, the preliminary Dvir VE contains 6 genes, and the preliminary Dyak VE contains 10 genes. Each stained embryo is manually classified as being at one of six time points in the blastoderm stage. The number of time points is also fixed due to the number of visually discernible substages. The process of registering the batch of stained and classified embryos to a species's template results in a matrix of expression values, called an *expression profile*, for each cell in the VE. This expression profile consists of a set of time series, one for each gene. The individual expression values are normalized during the data acquisition process and fall in the range of 0.0 to 1.0, where 0.0 indicates no measured gene expression.

3.1.4 Data Sets

To summarize: our collaborators currently work with 4 VEs and are actively acquiring data for 8 more. Each VE contains roughly between 4000 and 6000 cells. Each cell has a unique identifier, a 3D and 2D spatial position, and an expression profile matrix with dimensions $timepoints \times genes$, or 6×50.

3.2 Tasks

Our collaborators seek to use the VEs to find biologically meaningful differences in the dynamic expression patterns between species. Some of their initial questions include:

- Are there groups of cells in one species that have unique expression profiles compared to cells in the other species?
- Do groups of cells with similar expression patterns in different species exist in the same location in the embryos?
- How do groups of cells with a known morphological function relate to patterns in location and/or gene expression?

Addressing these questions requires the ability to explore, analyze, and compare the VEs. However, any specific question requires comparing the expression patterns of only a subset of the 50 genes. The typical working set size is 10 genes or fewer.

4 Summaries and Groups

The tasks described above all hinge on the idea of looking for the membership of cells in meaningful groups. The membership criteria can be based on spatial positions, expression profiles, some known biology, or any sort of combination thereof. The problems of finding a new group and judging the quality of an existing group can be addressed by using a measure of similarity for the cells. We now describe a mechanism for supporting measures of cell similarity, called *summaries*, and discuss how the workflow of our collaborators involves generating and using summaries to create, curate, and understand groups of cells.

4.1 Summaries

There are myriad ways to define the similarity of cells. One definition could be based on the similarity of spatial locations, while another could be based on the similarity of expression profiles. Quantifying the spatial similarity of cells in VEs is a hard problem for several reasons. First, some biologically meaningful neighborhoods of cells within a VE are anisotropic with respect to their spatial locations because many of the genes our collaborators are studying express in patterns that are not radially symmetric, such as the characteristic stripes of the eve gene shown in Figure 2. Second, there is no one-to-one mapping of cells between the species due to the variations discussed in Section 3.1. And third, the cells that form specific anatomical features in the adult fly can occur in slightly different locations among the species, such as the cells that form the boundaries of the head, thorax and abdomen.

Quantifying the expression profile similarity of cells is equally challenging because of the many biological interpretations of a meaningful expression difference. For example, quantitative differences in some genes could be more important than in others. Or, a time shift in a gene's time series may, or may not, constitute an interesting difference. Also, a large difference in just a single gene may be more relevant than small differences in many genes.

Meaningful measures of similarity are most often a complex combination of both spatial and expression profile similarity for the biological questions our collaborators are pursuing. Given the large search space of similarity options and the thousands of cells in each VE, manually classifying all interesting groups from just the raw data is a daunting task. The biologists thus incorporate computations into their workflow to transform the VE data into quantitative measures of similarity — we call these summaries. More specifically, a summary is a single quantitative value for each cell in a VE that expresses a measure of similarity, summarizing some potentially complicated similarity definition.

4.1.1 Computing Summaries

The types of computations our collaborators use to create summaries range from the simple to the complex. At one end of this spectrum there are computations that consist of a simple select operation that chooses the expression value of a single gene at a single time point for each cell. On the other end of the spectrum there are complex machine learning algorithms that use a multistage mix of spatial and expression similarities across multiple VEs. Creating summaries that compare between VEs requires more complex computations than those that only use spatial and expression data from within a single VE.

The comparative computations the biologists are currently using have a general structure of computing the similarity value of cell through a comparison with a subset of cells in another VE. This structure includes three key components: a *metric*, which quantifies the similarity of a pair of cells; an *aggregation*, which generates a single value from a set; and an *aggregation group*, which is a set of cells from which a single similarity value is aggregated, either from a single VE or from multiple VEs.

For instance, to address the first task in Section 3.2, finding groups of cells in one species that have unique expression profiles compared to cells in the other species, our collaborators created a summary called RMS_50 using a comparative computation. The computation first establishes, for each cell in a VE (c_i), an aggregation group from the 50 spatially nearest cells in another VE. Then, for each cell in the aggregation group, a root-mean-square metric (RMS) calculates the distance between its expression profile and that of c_i. Finally, from the list of 50 distance values, the aggregation operation *min* selects the smallest value to be the similarity value of c_i in the RMS_50 summary. The values in this summary are thus the distance of each cell in expression profile space to its best matching cell in another VE. Cells with a high value can be interpreted as having unique expression profiles, and the biologists would look for groups of cells with these high values.

The computation used to generate the RMS_50 summary is only one choice from the vast number of possible computations that combine spatial data, expression data, and multiple VEs. It is not clear, however, which of these computations will produce summaries that are biologically meaningful. Thus, we designed MulteeSum to allow the biologists to explore and characterize summaries.

4.1.2 Visualizing Summaries

Prior to our involvement, the biologists were relying almost exclusively on computational methods for extracting comparative information from the VEs. It was immediately clear that they needed an interactive visualization tool to explore and characterize their computational summaries by linking these results to the full VE data from which they were created. The broad, and growing, set of scientific and computational questions that they wanted to answer led us to think about how to frame their initial specific questions in a more general context, and how to flexibly combine their computational analysis with visual exploration.

With these goals in mind, we designed MulteeSum to support a triad of data types: spatial locations, temporal gene expression profiles, and summaries. This triad is cell-centric, where each cell in a VE has one position in an embryo map, one expression profile, and a set of similarity values from multiple summaries. The task of the biologist is to explore the triad of data for one or multiple VEs and to determine what aspects of the summaries are biologically interesting.

The variability of the relationships between the elements of this triad required us to use a more complex design pattern [11] than that of most visualization systems. Many visualization systems are designed to exploit hierarchical structure. In fact, the two-level structure of overview and detail is so pervasive that it is invoked as a mantra [24]. Even three-level hierarchies, as for example in the MizBee system [16], are not unusual. Another very common pattern is "peer-to-peer" linked views, where neither view is a parent of the other. The triad of spatial locations, temporal gene expression profiles, and summaries, however, cannot be encompassed by either of these patterns because the summaries are multiple, different overviews that reflect various possibly-interesting subsets of the spatial and the expression data.

Summaries can also be thought of as the output of a data transformation step, for which the underlying VE data is the input. MulteeSum was intended to fill a gap within the existing analysis workflow of our collaborators, rather than be a monolithic solution that would be used for every step in their workflow. It sits downstream of both the data collection and processing that creates the VEs, and also downstream of the data transformation used to computationally create the summaries. MulteeSum is agnostic to the type of computation or system used to generate the summaries, making the complete workflow flexible and generalizable.

MulteeSum was designed to support these specific tasks:

- Characterize the extent to which a summary is biologically meaningful.
- Find an interesting threshold value for the summary.
- Use the summary to pick interesting cells.
- Understand the underlying data for the cells deemed interesting.
- Find which spatial neighbors of an interesting cell share its characteristics.

These tasks are often a feedback loop as well as a source of information to the biologists for generating new summaries. Finally, finding biologically interesting summaries ultimately allows the biologists to determine meaningful groups of cells.

4.2 Groups

As we discussed in the beginning of this section, when we consider the biologists' scientific questions from an abstract point of view, they can be expressed as finding meaningful groups of cells. In MulteeSum, we include summaries as a computational indicator of cell similarity, and thus potential group membership. Furthermore, we explicitly include support for the creation, curation, and understanding of groups in the tool.

We define a *group* to be a set of similar cells that may come from a single, or multiple, VEs. There are three specific types of groups that we use in our tool: an *aggregation group*, as discussed in Section 4.1.1; a *created group*, which is a group of cells created on-the-fly during an interactive visualization session; and an *existing group*, which is predefined group of cells loaded into the tool at runtime. Existing

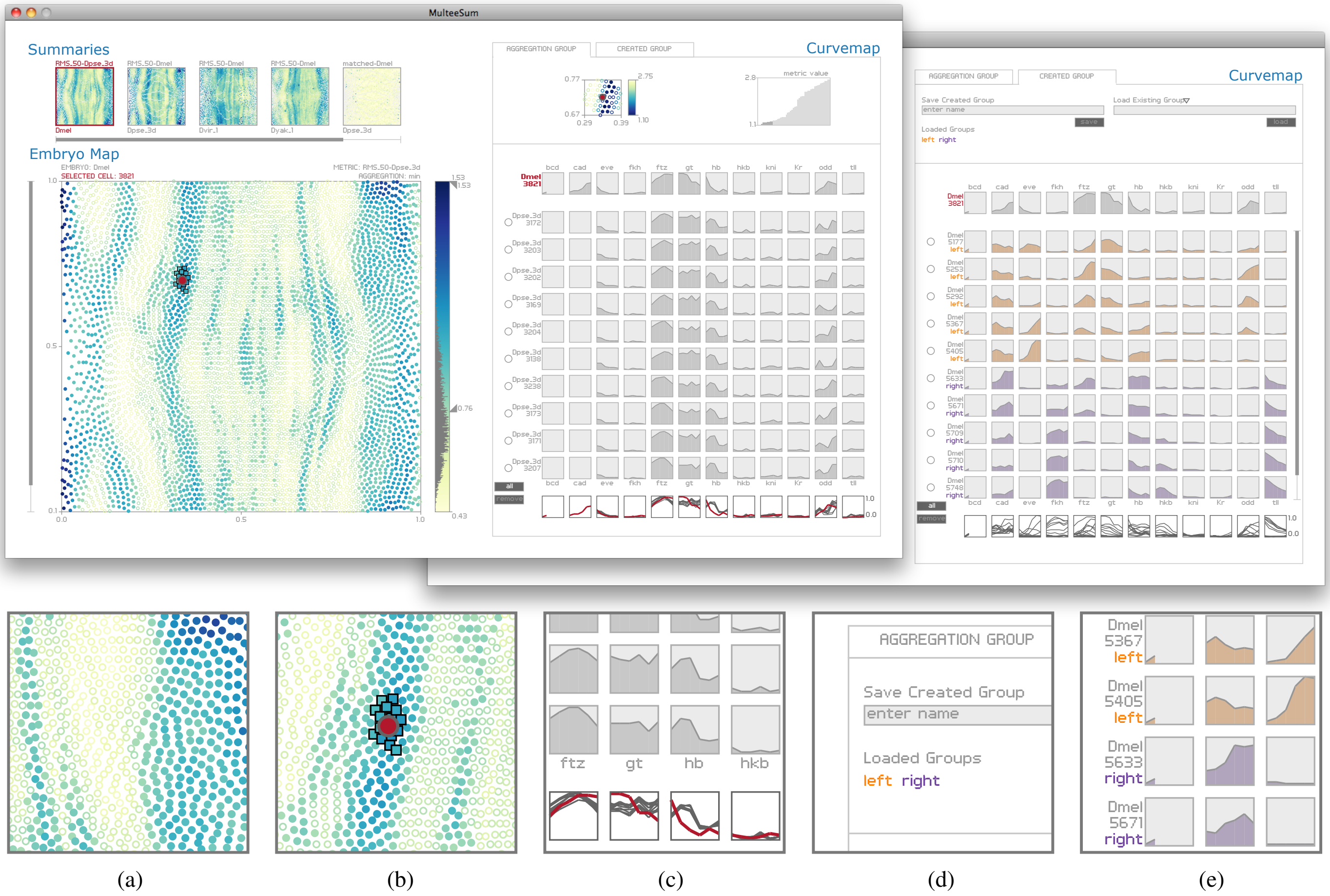


Fig. 3. Snapshots of MulteeSum. The curvemap views in the two screenshots along the top show the AGGREGATION GROUP tab on the left and the CREATED GROUP tab on the right. (a) The summary value of a cell is encoded with color in an embryo map. Filtered cells are rendered as circles with the remaining cells rendered as filled disks. (b) The selected cell is rendered in red, while cells in the created group are rendered as squares. (c) Rolling over cell labels in the curvemap highlights the cell's associated curves in the overlay plots. (d) Existing groups loaded into the curvemap are assigned a color, which is used when rendering the group's label. (e) The existing group's color is also used in the curvemap to indicate the expression profiles of group members.

groups can be created outside the tool to reflect some biological conjecture or to indicate the results of a clustering algorithm, or they can be created in a prior MulteeSum session.

5 MulteeSum

MulteeSum is a visualization system composed of multiple linked views [19], including a Summary View, an Embryo Map View, and a Curvemap View. The tool is designed to support exploration of the intertwined relationships in the spatial locations of cells, expression profiles, and summaries.

The input to MulteeSum is a set of summaries and a set of VEs. Each VE contains a list of cells, and each cell contains a location in the embryo map and an expression profile. Optionally, a set of existing groups can be loaded as well.

5.1 Summary View

In the Summary View, shown in the upper left corner of Figure 3, a thumbnail image appears for each summary loaded into the tool. The thumbnails are small embryo maps of the summary's associated VE, where the summary value at each cell's location is encoded with color — we are using sequential color maps from ColorBrewer [2]. The thumbnails provide an overview of the spatial distribution of summary values and allow for coarse comparisons between summaries. A summary, and the associated VE, is selected for viewing in the Embryo Map View by clicking the thumbnail. The currently selected summary is outlined in red.

5.2 Embryo Map View

Beneath the Summary View in Figure 3 is the Embryo Map View. This view contains an embryo map of the VE associated with the selected summary where the summary value of each cell in the embryo map is encoded using color. The 2D embryo map is the preferred view of our collaborators for analysis tasks as there is no obstruction of data and the common coordinate frame enables easier interspecies comparisons. A slider to the right of the embryo map is used to threshold cells based on their summary value. Filtered cells are rendered as circles with the remaining, potentially interesting cells rendered as filled disks, shown in Figure 3(a). We scent the slider [27] with a histogram showing the distribution of cells over the range of summary values — the biologists expect to see a normal distribution in the histogram for well-defined computations.

To determine a good threshold for the selected summary, as well as to characterize the effectiveness of the summary, the expression profiles of potentially interesting cells can be analyzed in the Curvemap View by selecting individual cells in the embryo map. The selected cell is rendered in red in the embryo map and its associated expression profile is shown in the top row of the curvemap, thus linking the cell's spatial and summary information with its expression profile. Finally, brushing with the right mouse button is used in the embryo map to select a meaningful group of cells, which are added to the current created group in the tool. Cells in the created group are rendered in the embryo map as squares, shown in Figure 3(b).

When switching between summaries, the state of the Embryo Map View is saved for the currently selected summary before switching to a new one.

5.3 Curvemap View

The Curvemap View is shown to the right of the Embryo Map View in Figure 3. This view is designed to compare and analyze the complete temporal gene expression profiles of selected cells and groups of cells, as well as to aid in curating and understanding meaningful groups of cells. This tabbed view has two panes, one for comparing the expression profile of the selected cell with those of its aggregation group (AGGREGATION GROUP tab), and one for analyzing the expression profiles in the created and existing groups (CREATED GROUP tab) — both of these panes are shown in Figure 3.

Each pane is divided into two components, with the bottom component of both containing a curvemap display of the associated cells' temporal expression profiles. A curvemap is a small-multiple matrix of filled line charts that encodes time series data using position [17], which is shown to be a better encoding than color for accurate judgements about quantitative values and time series trends [4, 14]. The curvemap display is designed specifically to enhance the detection of temporal trends, an important characteristic for helping the biologists make sense of the complex and dynamic relationships between the genes and the cells.

In MulteeSum, the rows of the curvemap are the cells while the columns are the genes. The current selected cell occupies the top row. We augment the main curvemap matrix with overlay multiples where all the curves for each column are superimposed in a single shared frame. These plots support the detection of trends for each gene across the set of cells. Rolling over any of the cells' labels highlights its associated curves in the overlay plots, an example of which is shown in Figure 3(c). The y-axis of both the line charts and overlay plots have the range of [0.0, 1.0]. Group cells can be removed from the curvemap display by using the radio and remove buttons.

In the AGGREGATION GROUP Pane, the top component of the pane displays the spatial distribution of the aggregation group in an embryo map with each cell colored according to its metric value. The component also includes a barchart of the metric values for more precise comparisons. Rolling over any of the cells' labels in the curvemap highlights the cell in both the aggregation group embryo map and the barchart, and vice versa, linking the cell's spatial location, expression profiles, and metric value. The aggregation group embryo map and barchart are both additionally used for adding aggregation group cells to the curvemap display. By default, the first 10 cells in the group are displayed.

The top component of the CREATED GROUP Pane, shown on the far right in Figure 3, contains controls for saving the created group, both within the tool and to a file, and loading existing groups. When an existing group is loaded into the curvemap it is assigned a color from a repeating 9-element colormap — we use a qualitative colormap from ColorBrewer [2]. The existing group label is rendered in the assigned color, as are the expression profiles for the existing group members in the curvemap. Figures 3(d) and 3 (e) show examples. This color coding in the curvemap helps to make existing group membership visually salient. If a cell has membership in multiple groups its expression profile is repeated in the curvemap, color coded each time for the specific existing group. Rolling over an existing group label highlights the curves associated with all group members in the overlay plots.

5.4 Implementation

We implemented MulteeSum using the Processing programming language [18]. Executables and source code are freely available at http://multeesum.org.

6 Iterative Refinement

In this section we briefly discuss the details of our collaboration with the biologists that led to the development of MulteeSum — one of the biologists is a co-author on this paper. We began working with the group two years ago and developed two earlier versions of the tool with feedback from two members of the group. MulteeSum was developed with feedback from an additional four group members. We obtained feedback from the biologists through informal interviews and observed their workflow by spending one day a week in their lab. The development of each version of the tool took roughly 2-3 months, with several months in between for collecting feedback and designing the next prototype.

When we first began working with our collaborators they were using static images created in Matlab to explore the results of their initial computational comparisons. Their first goal was to find unique groups of cells between the species. They created RMS_50 summaries, filtered the cells based on their summary values, and then clustered the remaining cells using their expression profiles. Finally, they plotted the resulting groups on an embryo map to see the spatial distribution of group members, and generated heatmaps for each group to analyze the expression profiles. As these views lacked interactivity, there was no easy way to visualize the expression profile for an individual cell at a specific location.

Our first prototype took these summaries and visual encodings and added interactivity, giving the biologists the ability, for the first time, to select an individual cell at a specific spatial location, and to compare the cell's expression profile to that of its aggregation group. Using this prototype the biologists discovered that their filtering method was not producing meaningful groups of cells. They asked that the next version of the tool include support for filtering cells interactively.

Our second prototype encoded the summary value for each cell in an embryo map using color, and included a slider for thresholding the summary values, giving the biologists the ability to filter cells dynamically. Although we still encoded expression profiles using a heatmap, we began experimenting by augmenting the heatmaps with overlay plots for each gene to enhance the perception of temporal trends. Upon completion of the second prototype, our discussions with the group focused on how to include more computational functionality in the tool to support a broader range of questions. These discussions are what led to the idea of a general and flexible workflow where summaries generated in an upstream data transformation step are used in a downstream visualization tool to enable exploration of computational comparisons in the context of the underlying VE data sets. The feedback loop in this workflow is that observations made in MulteeSum can lead to the generation of new, more informative summaries at the data transformation step.

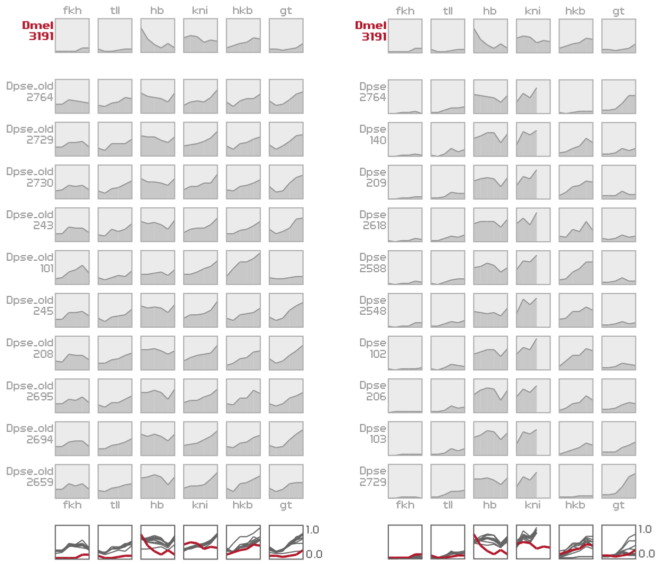
7 Case Studies

We present three case studies to provide anecdotal evidence for the efficacy of MulteeSum. The first case study discusses the biologists' discovery of significant noise in one of their data sets, while the second case study presents a series of findings that indicate the need for more sophisticated summary computations. The last case study uses a simple summary to probe the biological subtleties of gene regulation.

7.1 Wrangling Data

While exploring an early version of the Dpse VE using our second prototype system discussed in Section 6, the biologists discovered that the data set was in fact plagued by significant low-level noise. Using summaries that compared the Dmel and early Dpse data sets, they noticed that the expression profiles of many cells with high summary values, *i.e.*, those cells with a large dissimilarity from their aggregation group cells, were actually quite similar to the expression profiles of their aggregation group in terms of the most highly expressed genes. By analyzing and comparing the expression profiles of many of these high-value cells, they found that the summary values were being dominated by contributions from low-level noise in the Dpse data for a handful of genes. An example of this noisy data is shown for a Dpse aggregation group in Figure 4(a). These high, but uninteresting, summary values were masking the types of expression profile differences the biologists were hoping to find. They made the decision to generate a new Dpse VE, the data for which is shown in 4(b) for the same set of cells.

The discovery of noise in the early data was enabled by the interactive linking of summary values with expression profiles. Being able to quickly explore the complete expression profiles of many different cells was also important for noticing the global trends in the subset of the genes.



(a) (b)

Fig. 4. Comparing a selected cell from Dmel with its 10 most similar cells from two different Dpse VEs. (a) An early version of the Dpse data set was found to be plagued with low-level noise in the expression levels of a handful of genes. (b) A new, cleaner version of the Dpse data.

7.2 Refining Computation

Currently, the biologists are performing computational comparisons of the VE data using pairwise comparison metrics like RMS and Pearson correlation. They are interested in using more sophisticated algorithms to pull out specific, biologically interesting characteristics in the data sets. In this case study we present a series of examples that illustrate the types of characteristics one of the biologists is interested in, which she found using the RMS_50 summary for Dmel in MulteeSum.

Investigating the expression profile for the Dmel cell shown in Figure 4(b), this biologist noticed that the RMS metric did not match the cell with the biologically most similar cell from the aggregation group. In this case, the RMS best match, shown in the second row of the curvemap, has significant differences in the expression levels of both the hkb and gt genes compared to the selected cell shown in the top row. This is in contrast to the other aggregation group cells, which match the selected cell's expression trends more closely for these genes. This mismatch indicates that the RMS metric is sensitive to small variations over the entire expression profiles, at times obscuring the biologically significant differences in just a small set of genes. For this example, a possible next computation would weight the contributions for specific genes of interest more heavily than others, or filter low-level noise.

Another biologically meaningful trend that the RMS_50 summary obscures is shown in Figure 5, where the expression profile of the selected cell significantly varies from those in its aggregation group in just a single gene, prd. The summary value for this selected cell was only moderately high, but the potential biological implications for this variation are very important to this biologist. She noted that a significant variation in a single gene could indicate a novel regulation mechanism, and this specific example provides an interesting direction for follow-up experiments. Similar to the previous example, this observation indicates the need for computations that are sensitive to variations in just a single, or a small set of, genes in the expression profiles.

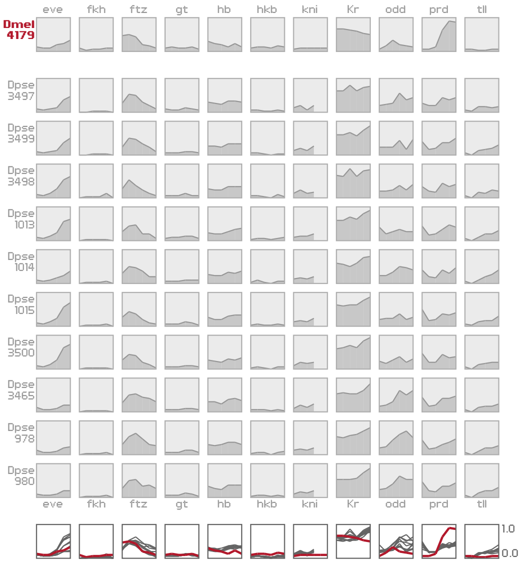
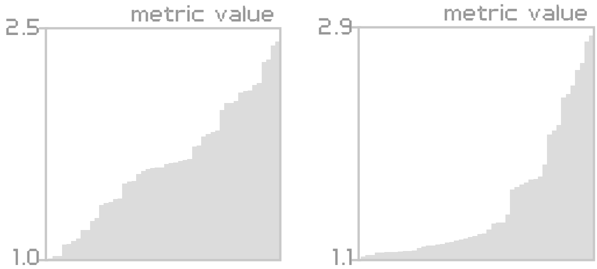


Fig. 5. Single gene variations in the expression profile of a cell to that of its aggregation group are of particular interest to our collaborators, such as the variation in the prd shown in this example.

In a third example shown in Figure 6, two cells with roughly similar summary values were selected by the biologist. One cell resides in the anterior of the embryo, while the other in the posterior. Of interest is the distribution of metric values for the aggregation groups — shown in Figure 6(a) are the values for the anterior cell and in Figure 6(b) are the values for the posterior cell. She noticed for the posterior cell the distribution has a very long, flat tail of low values, indicating a potentially large neighborhood of similar cells in the comparison embryo, compared with the relatively few similar cells for the anterior cell. This finding reiterated to the biologist that the RMS computation can assign similar summary values to cells that have very different comparison trends to their aggregation groups.



(a) (b)

Fig. 6. (a) A cell selected from the anterior of the Dmel VE using the RMS_50 summary shows a relatively rapid increase for the RMS metric values of its aggregation group. (b) A cell selected from the posterior of the Dmel VE with a similar summary value shows a markedly different trend.

The biologists are planning to incorporate more sophisticated comparison algorithms into their workflow, such as manifold learning methods [10], to further enable their discovery of biologically meaningful groups. Their initial explorations using summaries like RMS_50 are helping them determine the types of characteristics they are interested in. This determination is made possible through the linking of the spatial data, expression profiles, and summaries in MulteeSum.

7.3 Understanding Gene Regulation

Another one of the members of the group is developing physically-based statistical models that describe the genomic controls of dynamic

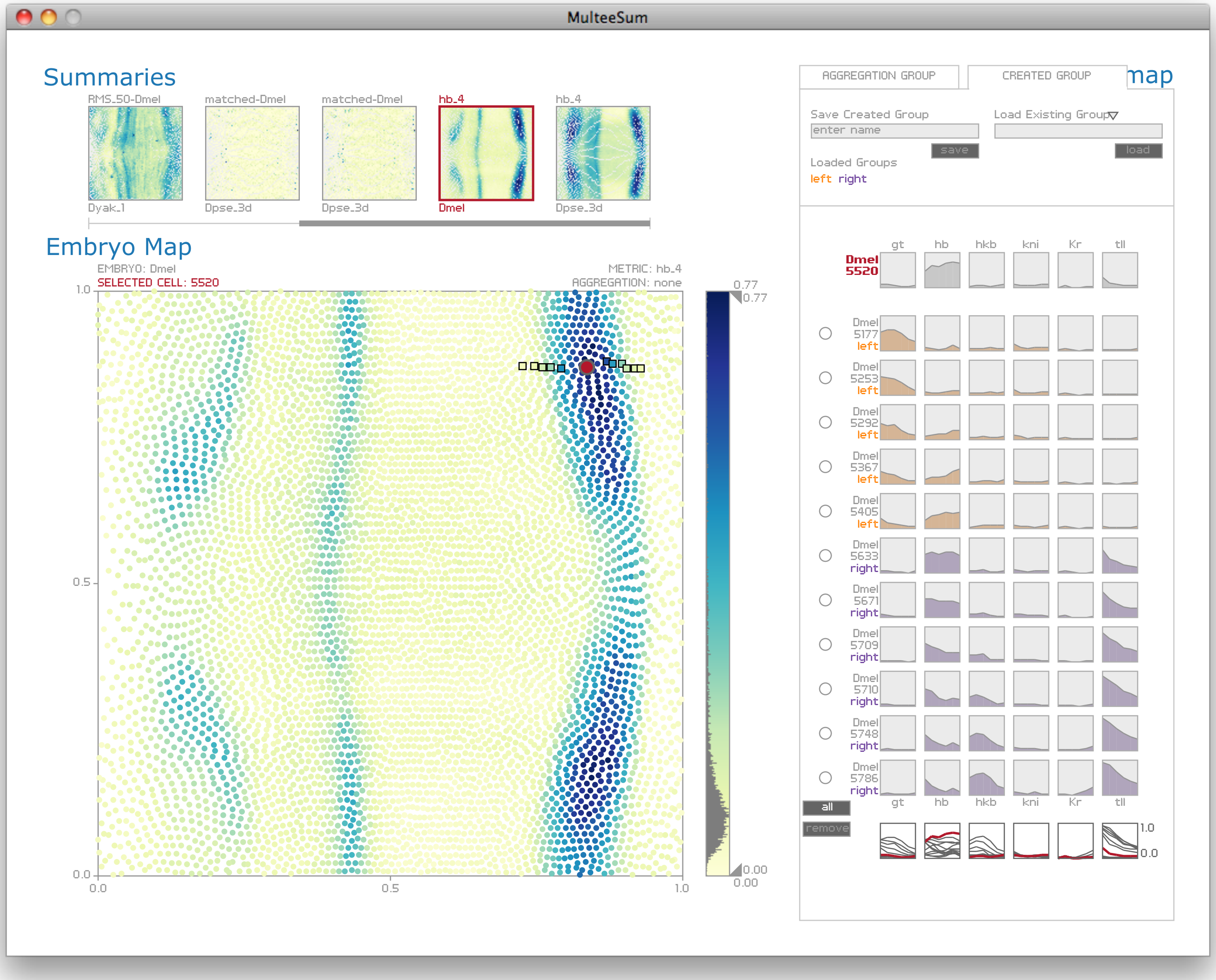


Fig. 7. A summary generated by selecting the expression value of the hb gene at time point 4 at each cell. One of the biologists is interested in understanding the expression patterns in a handful of genes on either side of the region of high expression.

gene expression patterns. Her initial study involves correlating the expression levels of the `hb` gene, the expression levels of genes that are known to regulate `hb` (`gt`, `hkb`, `kni`, `Kr`, and `tll`), and the genomic sequence that dictates how `hb` is regulated. In her statistical model, she aggregates the expression data for these genes from an entire VE, looking for trends in a global manner. She is using MulteeSum to "spot-check" the results of her model to see if individual cells behave the way her model predicates that they will.

She created a summary for `Dmel` from the expression level of the `hb` gene at a single time point at each cell in the VE. This summary is shown in Figure 7. Of particular interest is the band of high values in the posterior of the embryo. On either side of this band two groups of cells were created to analyze how the expression profiles change moving inwards towards the band. These groups are the *left group* and the *right group*, and both were generated by manually selecting cells in MulteeSum.

The groups were loaded into the curvemap, with the left group assigned an orange color and the right group assigned a purple color. A high-value cell from the middle of the band is selected for comparison. The expression profile of the selected cell shows consistently low levels of all regulator genes and high levels of `hb`. In the left group, the repressor gene `gt` is high, while in the right group the repressor gene `tll` is high — the biologist expected to see this from prior knowledge about these genes. Several trends she was not aware of, however, were immediately obvious to her. She noted that even though the expression of the repressor `tll` in the right group is markedly going down over time, the level of `hb` does not exhibit a comparable change. Also of interest is that while all the final levels of `tll` in the right group are the same, the final levels of `hb` in the same set of cells varies. This latter observation is a possible hint that the final levels of `tll` do not matter for the expression of `hb`. Both of these observations point to potentially interesting features in how the `hb` gene is regulated, and are trends she will look to confirm in her statistical model.

This biologist found the curvemap representation particularly useful for observing these subtle differences in the expression profiles. For her, encoding the temporal expression profiles as curves is key because the coarse temporal resolution of these data sets leads her to trust the overall trends more than any single time point. She felt that her new observations would have been nearly impossible to see in PointCloudXplore or in a heatmap display. This scenario also confirmed that the working set size supported by the curvemap display was a good fit for her analysis needs.

8 Conclusions and Future Work

In this paper we discuss our development process and design of MulteeSum, a visualization system for inspecting and curating biological data sets that contain temporal gene expression profiles for cells with known spatial positions. It is the first tool to support the comparison of multiple such data sets. We describe the use of summaries for performing computational comparisons of the data sets, and discuss how MulteeSum incorporates summaries with the underlying data for flexible and extendable support of a broad range of scientific questions. MulteeSum was designed to support tasks that require the ability to explore, analyze, and compare VEs, with three specific examples listed in Section 3.2. It has been deployed to our collaborators, and we present three case studies illustrating its successful use for the first of the three example tasks. We plan to follow up with the biologists as they continue to obtain and analyze new VE data, and hope to report on their usage of the tool for addressing the remaining two tasks in future work.

Although MulteeSum was developed for visualizing the genes involved with the body segmentation of *Drosophila* embryos, we believe that our design is broadly applicable to any data sets that incorporate temporal gene expression data with the spatial location of cells. With the advent of high-throughput imaging techniques, this type of data is becoming more prevalent within the biology community. Our core ideas for linking computational data in the form of summaries with spatial and expression data are general and support a broad range of analysis. The workflow supported by MulteeSum allows users to generate summaries using any combination of computational methods and algorithms, including MATLAB, R, open-source software, and custom tools.

Extending the MulteeSum approach to other model organisms and domains would be very interesting future work. The two obvious issues are scalability, and finding an appropriate spatial map. While the current curvemap display is a good fit for our design goal of exploring trends across tens of cells, tens of genes, and tens of time points at once, it is unlikely to support analysis tasks that require the simultaneous inspection of hundreds or thousands of genes and cells, or higher resolution time series with dozens or hundreds of time points. Extending MulteeSum to domains such as astronomy, where researchers study large collections of light curves from stars, may require a higher-level abstraction beyond curvemaps. User studies to better understand the scalability and limitations of curvemap displays would address many open questions.

Also, the current embryo map view is specific to *Drosophila* embryos and takes advantage of a known lossless 2D parameterization of the 3D embryo shape. At first glance, a 3D view may seem to be required to explore structures like a mouse brain [15]. However, based on our experience with the response of the *Drosophila* researchers to 2D versus 3D views, we argue that developing abstract 2D representations of 3D shapes is likely to offer many advantages to researchers dealing with the combination of summaries, spatial locations, and temporal gene expression. When comparing vastly unrelated species, such as a fly to a mouse, a compact 2D representation may well be essential.

Finally, we are continuing our collaboration with the biologists to develop additional tools that incorporate findings from MulteeSum with sequence-based data. These future tools will focus on helping them in their quest to uncover the genomic source of expression differences between species.

Acknowledgments

The authors wish to thank the members of the DePace Lab, as well as Charless Fowlkes, for their input and support during the development of this work. This work was supported in part by the National Science Foundation under Grant 0937060 to the Computing Research Association for the CIFellows Project, the Helen Hay Whitney postdoctoral fellowship, the Armenise Foundation of Harvard Medical School, and the Natural Sciences and Engineering Research Council of Canada.

References

[1] Berkley Drosophila Transcription Network Project. `http://bdtnp.lbl.gov/`, accessed March 15, 2010.

[2] C. Brewer. `http://colorbrewer.org`, accessed March 15, 2010.

[3] S. B. Carroll, J. K. Grenier, and S. D. Weatherbee. *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design*. Blackwell Science, 2001.

[4] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[5] A. DePace et al. Quantitative comparison of anterior posterior patterning in closely related Drosophila. In progress.

[6] Drosophila 12 Genomes Consortium et al. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–18, Nov 2007.

[7] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. National Academy of Sciences*, 95(25):14863–14868, 1998.

[8] C. Fowlkes, C. Hendriks, S. Keränen, G. Weber, O. Rübel, M. Huang, S. Chatoor, A. DePace, L. Simirenko, C. Henriquez, A. Beaton, R. Weiszmann, S. Celniker, B. Hamann, D. Knowles, M. Biggin, M. Eisen, and J. Malik. A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell*, 133(2):364–374, 2008.

[9] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, and A.-C. Gavin. Visualization of omics data for systems biology. *Nature Methods*, 7:S56–S68, 2010.

[10] S. Gerber, T. Tasdizen, and R. Whitaker. Dimensionality reduction and principal surfaces via kernel map manifolds. *Proc. Intl. Conf. Computer Vision (ICCV)*, 2009.

[11] J. Heer and M. Agrawala. Software design patterns for information visualization. *IEEE Trans. Visualization and Computer Graphics (TVCG)*, 12(5), 2006.

[12] C. L. L. Hendriks, S. V. E. Keranen, C. C. Fowlkes, L. Simirenko, G. H. Weber, A. H. DePace, C. Henriquez, D. W. Kaszuba, B. Hamann, M. B. Eisen, J. Malik, D. Sudar, M. D. Biggin, and D. W. Knowles. Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution 1: data acquisition pipeline. *Genome Biol*, 7(12):R123, Jan 2006.

[13] T. B. Kornberg and T. Tabata. Segmentation of the Drosophila embryo. *Current Opinion in Genetics & Development*, 3(4):585–94, Aug 1993.

[14] H. Lam, T. Munzner, and R. Kincaid. Overview use in multiple visual information resolution interfaces. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2007)*, 13(6):1278–1285, 2007.

[15] C. Lau, L. Ng, C. Thompson, S. Pathak, L. Kuan, A. Jones, and M. Hawrylycz. Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain. *BMC Bioinformatics*, 9(1):153, 2008.

[16] M. Meyer, T. Munzner, and H. Pfister. MizBee: A multiscale synteny browser. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2009)*, 15(6):897–904, 2009.

[17] M. Meyer, B. Wong, M. Styczynski, T. Munzner, and H. Pfister. Pathline: A tool for comparative functional genomics. *Computer Graphics Forum (Proc. EuroVis 10)*, 29(3), 2010.

[18] C. Reas, B. Fry, and J. Maeda. *Processing: A Programming Handbook for Visual Designers and Artists*. MIT Press, 2007.

[19] J. C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proc. Intl. Conf. on Coordinated and Multiple Views in Exploratory Visualization (CMV)*, pages 61–71. IEEE Computer Society, 2007.

[20] O. Rübel, G. H. Weber, M.-Y. Huang, E. W. Bethel, M. D. Biggin, C. C. Fowlkes, C. L. Luengo Hendriks, S. V. E. Keranen, M. B. Eisen, D. W. Knowles, J. Malik, H. Hagen, and B. Hamann. Integrating data clustering and visualization for the analysis of 3D gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 7(1):64–79, 2010.

[21] O. Rübel, G. H. Weber, S. V. E. Keränen, C. C. Fowlkes, C. L. L. Hendriks, L. Simirenko, N. Y. Shah, M. B. Eisen, M. D. Biggin, H. Hagen, D. Sudar, J. Malik, D. W. Knowles, and B. Hamann. PointCloudXplore: Visual analysis of 3D gene expression data using physical views and parallel coordinates. In *Proc. Eurographics/IEEE Symp. Visualization (EuroVis)*, pages 203–210. Eurographics Association, 2006.

[22] A. J. Saldanha. Java Treeview – extensible visualization of microarray

data. *Bioinformatics*, 20(17):3246–3248, 2004.

[23] J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results. *Computer*, 35(7):80–86, 2002.

[24] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. IEEE Symp. Visual Languages (VL)*, page 336, Washington, DC, USA, 1996. IEEE Computer Society.

[25] G. H. Weber, O. Rübel, M.-Y. Huang, A. H. DePace, C. C. Fowlkes, S. V. E. Keranen, C. L. Luengo 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 Trans. Comput. Biol. Bioinformatics*, 6(2):296–309, 2009.

[26] J. N. Weinstein. A postgenomic visual icon. *Science*, 319(5871):1772–1773, March 2008.

[27] W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, 2007.