Gene expression patterns are the lingua franca of developmental biology and Drosophila melanogaster is a canonical model organism for understanding animal development. More than one hundred thousand images containing gene expression patterns have been published in research articles that have used low and high-throughput techniques. The primary data are images that exist as graphical objects. We have developed a unique knowledge base (FlyExpress) to facilitate visual mining of images from Drosophila melanogaster embryogenesis. By clicking on specific locations in pictures of fly embryos from different stages of development and different visual projections, users can produce a list of genes and publications instantly. In FlyExpress, each queryable embryo picture is a heat-map that captures the expression patterns of more than 4500 genes and more than 2600 published articles. In addition, one can view spatial patterns for particular genes over time as well as find other genes with similar expression patterns at a given developmental stage. Therefore, FlyExpress is a unique tool for mining spatiotemporal expression patterns in a format readily accessible to the scientific community.

Availability: http://www.flyexpress.net

Contact: s.kumar@asu.edu

Received on June 25, 2011; revised on September 20, 2011; accepted on October 6, 2011

Spatial gene expression patterns are the lingua franca of developmental biology and Drosophila melanogaster is a canonical model organism for understanding animal development. More than one hundred thousand images containing gene expression patterns have been published in research articles that have used low and high-throughput techniques. Kumar et al. (2002), Lecuyer et al. (2007), and Tomancak et al. (2002). With the increase in the volume of information for spatiotemporal expression patterns, scientists have an unprecedented opportunity to identify co-regulated genes by comparing spatial expression profiles, whose similarity is the first significant clue about potential joint roles in developmental processes. With the increasing size of image datasets, such spatial pattern comparisons have become increasingly more useful, because genes with similar and overlapping patterns can now be found with a much higher probability, for example, to predict functions for new genes or to understand newly discovered patterns for existing genes. However, classical manual inspection of spatial patterns for matches does not scale well for such large datasets, which in turn hampers the discovery of gene interactions and hypotheses generation. Computer-assisted approaches using annotations of images with textual controlled vocabulary, while useful, are subject to serious limitations because it is virtually impossible to describe all aspects of a spatial pattern by words alone. Furthermore, the text of published literature rarely contains controlled vocabulary descriptions of spatial patterns and Google-like text searches produce limited results.

We have developed a unique resource for mining the gene expression patterns visually for a given developmental stage and visual projection. This is intended to enable biologists to interrogate specific coordinates in a developing embryo and to quickly reveal genes expressed at that location (Fig. 1A) as well as a list of published research articles that contain spatiotemporal gene expression at the selected location (Fig. 1B). A user can select multiple locations simultaneously, and the genes and publications are ranked by the number of locations. The outcome is a set of links to individual gene expression pattern images as well as publications.

For each developmental stage and visual projection, the user is presented with a global view of gene activity [Genomewide-Expression-Maps, (GEMs)]. They are constructed by the aggregation and normalization of spatial expression patterns that share the same developmental stage and view (see Methods in Konikoff et al. 2011). The normalized spatial aggregates are then rendered into heat-maps with the darkest regions showing the largest numbers of expressed genes. Using this procedure we generated 18 GEMs from 48 190 images for Tomancak et al. (2002) data (BDGP; 3 views × 6 stage ranges; Fig. 1A), 15 GEMs from 30 095 images for Lecuyer et al. (2007) data (Fly-FISH; 3 views × 5 stage ranges), and 44 GEMs for 5977 images from more than 2600 published articles.
S. Kumar et al.

Fig. 1. Visual discovery tools in FlyExpress. (A) The user selects a lateral GEM for stage range 9–10, clicks on several points on the GEM to generate a list of genes with expression present, selects a gene of interest from the list to generate a list of embryo images for this gene, and (C) uses BESTi search to generate a list of embryo images and their genes with similar patterns of expression. (B) The user selects a lateral PubMed GEM for stage 10, and clicks on several points on the GEM to generate a list of publications with images showing expression at the selected locations.

(PubMed publications, years 1980–2006; Fig. 1B). By the end of this year, we expect to have over 20,000 images in PubMed GEMs. To build these GEMs, we developed a digital library of expression images in which all the images and expression patterns have been size standardized and aligned, a prerequisite for biologically meaningful comparison of spatial expression patterns and synthesizing information across genes and publications. For all the data, we determined the anatomical view of each image, because comparisons are not meaningful for images with different views (e.g. lateral versus ventral). In this work, we have preserved the original biological stage range annotations by the authors of the data, which leads to three types of GEMs (BDGP, Fly-FISH, PubMed). In the future, we plan to assign individual stages to all the images, which will enable users to query all the data through integrated GEMs. Unlike the high-throughput data (BDGP and FlyFISH), images from PubMed data were manually annotated by our team of curators. We are continuing to add images from more recent articles and developing systems for authors to directly submit their data to our resource for inclusion.

In addition to the GEM-based data mining, FlyExpress provides facilities to display expression patterns for any gene and a tool to identify co-expressed and, thus potentially co-regulated, genes by searching for other genes with similar or overlapping expression profiles by using the Basic Expression Search Tool for Images (BESTi), e.g. Kumar et al. (2002). In the output, genes are sorted based on their spatial similarity score with the query profile and the probability of randomly (by chance) finding a match with a given S-score or better (P-value; methods described in Konikoff et al., 2011). In summary, FlyExpress addresses an urgent need to standardize and summarize image data from Drosophila embryogenesis in order to facilitate the generation of novel gene interaction hypotheses and accelerate biological discovery. And, it provides a template for creating similar image searchable databases of gene expression patterns from other developmental model systems.

ACKNOWLEDGEMENTS

We thank Drs Susan Celniker, Erwin Frise, William Gelbart, Thomas Kaufmann, Henry Krause, Eric Lécuyer, Gerry Rubin, Pavel Tomancak and many members of the fly community for their support and invaluable advice.

Funding: US National Institutes of Health grant (HG002516-07 to S.K.) and Arizona State University.

Conflict of Interest: none declared.

REFERENCES


