FlyExpress: visual mining of spatiotemporal patterns for genes and publications in *Drosophila* embryogenesis

Sudhir Kumar[1,5,*], Charlotte Konikoff[1,2], Bernard Van Emden[1], Christopher Busick[1], Kailah T. Davis[1], Shuiwang Ji[1,3], Lin-Wei Wu[1], Hector Ramos[1], Thomas Brody[4], Sethuraman Panchanathan[3], Jieping Ye[1,3], Timothy L. Karr[1], Kristyn Gerold[1], Michael McCutchan[1] and Stuart J. Newfeld[1,5]

[1]Center for Evolutionary Medicine and Informatics, Biodesign Institute, Arizona State University (ASU), Tempe, AZ 85287, [2]Department of Biology, University of Washington, Seattle, WA 98195, [3]School of Computing, Informatics, and Decision Systems Engineering, ASU, Tempe, AZ 85287, [4]National Institutes of Health, Neural Cell-Fate Determinants Section, Bethesda, MD 20892 and [5]School of Life Sciences, ASU, Tempe, AZ 85287, USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: Images containing spatial expression patterns illuminate the roles of different genes during embryogenesis. In order to generate initial clues to regulatory interactions, biologists frequently need to know the set of genes expressed at the same time at specific locations in a developing embryo, as well as related research publications. However, text-based mining of image annotations and research articles cannot produce all relevant results, because 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

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; 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

*To whom correspondence should be addressed.
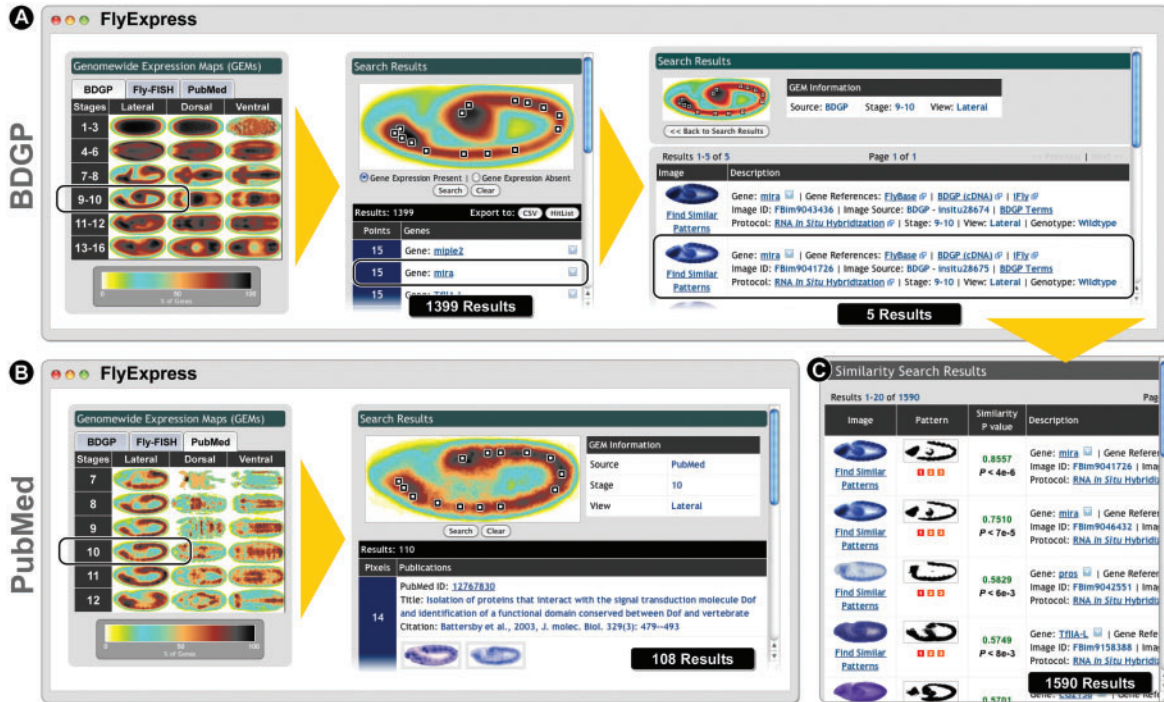
Fig. 1. Visual discovery tools in FlyExpress. (**A**) The user selects a lateral GEM for stage range 9–10, clicks on several points on the GEM to generate a list of genes with expression present, selects a gene of interest from the list to generate a list of embryo images for this gene, and (**C**) uses 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.

REFERENCES

Konikoff,C.*et al.* (2011) Comparison of embryonic expression within multigene families employing the FlyExpress discovery platform reveals significantly more spatial than temporal divergence. *Developmental Dynamics* [Epub ahead of print, doi: 10.1002/dvdy.22749, September 29, 2011].

Kumar,S. *et al.* (2002) BEST. A novel computational approach for comparing gene expression patterns from early stages of *Drosophila melanogaster* development. *Genetics*, **162**, 2037–2047.

Lécuyer,E. *et al.* (2007) Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function, *Cell*, 131, 174–187.

Tomancak,P. *et al.* (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.*, **3**, RESEARCH0088.