

Comparison of Embryonic Expression Within Multigene Families Using the FlyExpress Discovery Platform Reveals More Spatial Than Temporal Divergence

Charlotte E. Konikoff,[†] Timothy L. Karr, Michael McCutchan, Stuart J. Newfeld, and Sudhir Kumar*

Background: Overlaps in spatial patterns of gene expression are frequently an initial clue to genetic interactions during embryonic development. However, manual inspection of images requires considerable time and resources impeding the discovery of important interactions because tens of thousands of images exist. The FlyExpress discovery platform was developed to facilitate data-driven comparative analysis of expression pattern images from *Drosophila* embryos. **Results:** An image-based search of the BDGP and Fly-FISH datasets conducted in FlyExpress yields fewer but more precise results than text-based searching when the specific goal is to find genes with overlapping expression patterns. We also provide an example of a FlyExpress contribution to scientific discovery: an analysis of gene expression patterns for multigene family members revealed that spatial divergence is far more frequent than temporal divergence, especially after the maternal to zygotic transition. This discovery provides a new clue to molecular mechanisms whereby duplicated genes acquire novel functions. **Conclusions:** The application of FlyExpress to understanding the process by which new genes acquire novel functions is just one of a myriad of ways in which it can contribute to our understanding of developmental and evolutionary biology. This resource has many other potential applications, limited only by the investigator's imagination. *Developmental Dynamics* 241:150–160, 2012. © 2011 Wiley Periodicals, Inc.

Key words: *Drosophila*; FlyExpress; gene expression patterns; image analysis; multigene family

Key findings:

- FlyExpress resource facilitates analysis of expression pattern images from *Drosophila* embryos.
- Image-based pattern searching yields fewer but more precise results than text-based searching.
- Gene family expression more frequently displays spatial divergence than temporal divergence.

Accepted 30 August 2011

INTRODUCTION

High throughput and individual laboratory efforts have made available a vast collection of spatial patterns for a large number of developmentally rele-

vant genes (Tomancak et al., 2002; Lécuyer et al., 2007). These data can be a key to the discovery of previously unknown links between genes and new components within developmental networks. A common first step in

discovering gene interactions is to identify genes with overlapping expression patterns. However, the standard practice of manual inspection of images is not efficient given the extraordinary number of images

Additional Supporting Information may be found in the online version of this article.

School of Life Sciences and Center for Evolutionary Medicine and Informatics in the Biodesign Institute, Arizona State University, Tempe, Arizona

Grant sponsor: NIH; Grant number: 2R01HG002516-07.

[†]Dr. Konikoff's present address is Department of Biology, University of Washington, Seattle, WA 98195

*Correspondence to: Sudhir Kumar, School of Life Sciences and Center for Evolutionary Medicine and Informatics in the Biodesign Institute, Arizona State University, Tempe, AZ 85287. E-mail: s.kumar@asu.edu

DOI 10.1002/dvdy.22749

Published online 29 September 2011 in Wiley Online Library (wileyonlinelibrary.com).

available today (Gurunathan et al., 2004; Peng et al., 2007; Walter et al., 2010). This problem has been addressed using textual descriptions of gene expression images using controlled vocabularies (CVs; Janning, 1997; Brody, 1999; FlyBase, 1999; Drysdale, 2001; Tomancak et al., 2002; Matthews et al., 2005; Grumblin and Strelets, 2006).

Textual descriptions are always limited by the size and nature of the vocabulary and the current state of scientific knowledge. Furthermore, no textual annotations exist for a plethora of spatial expression patterns due to the tremendous resources required, as assigning such annotations must be done manually (Gurunathan et al., 2004). Furthermore, under many circumstances it is not feasible to fully describe an expression pattern in words alone. There is a great need for computational approaches that directly compare the primary information—images containing expression patterns (Kumar et al., 2002; Khatri and Draghici, 2005; Walter et al., 2010). These tools have the unique potential to facilitate large-scale synthesis of all available developmental image data and to integrate information across experiments and laboratories.

For instance, computational frameworks for sequence analysis have already revolutionized the analysis of DNA and proteins (Altschul et al., 1990). However, image analysis of gene expression patterns poses unique challenges. Sequence data can be easily expressed in four or twenty building blocks (for DNA and proteins, respectively), but images containing spatial expression patterns show remarkable variation in orientation, size, lighting conditions, and color-spectrum depending on the laboratory techniques used. Thus, biological image analysis is more challenging and a new frontier of computational biology (Gurunathan et al., 2004; Peng et al., 2007; Walter et al., 2010).

A current issue is how to facilitate the large-scale analysis of gene expression patterns from *D. melanogaster* embryos. Over 100,000 images capturing spatial aspects of gene expression are now available (Tomancak et al., 2002; Lécuyer et al., 2007). These images are a key

resource for understanding the early development of fruit flies and other metazoans because a large number of genes are shared (Koonin et al., 2004; Bier, 2005). Thus, we developed the FlyExpress platform, which contains a unique digital library of standardized expression patterns and tools that facilitate comparative expression analysis (<http://www.flyexpress.net>).

Here, we show that image-based searching in FlyExpress often yields fewer but more high-quality results compared with text-based searching when the specific goal is to find co-expressed genes. We then demonstrate how the FlyExpress resource may be used for scientific discovery by conducting an analysis of multigene family expression divergence from spatial and temporal perspectives. Overall, we found more instances of spatial than temporal divergence among paralogous genes with comparable data. Looking more closely at this diversity, the data revealed most of this expression divergence occurs after the maternal to zygotic transition.

RESULTS

By virtue of having a unique digital library of expression patterns that are uniformly oriented, aligned and scaled, FlyExpress makes it possible for investigators to visually and computationally compare gene expression patterns within the Berkeley *Drosophila* Genome Project (BDGP; Tomancak et al., 2002) and Fly-FISH (Lécuyer et al., 2007) image collections. These high-throughput in situ hybridization datasets contain images depicting the expression patterns of various *D. melanogaster* genes across development (Fig. 1A). There are 1,091 genes shared by both datasets, 2,263 unique to BDGP and 1,342 unique to Fly-FISH.

There are a total of 57,045 BDGP images (of individual whole-mount embryos) in FlyExpress. This dataset captures expression throughout embryogenesis. There are 43,065 images of whole-mount embryos from Fly-FISH available for searching in FlyExpress. The majority of these images capture embryonic development through stage 9. Also note that the in situ hybridization protocols

used by each are different—BDGP visualizes gene expression in whole-mount embryos using the single color alkaline phosphatase reaction and light microscopy while Fly-FISH uses two color fluorescence in situ hybridization and confocal microscopy. Thus, BDGP images are valuable for assessing expression presence or absence in various parts of the embryo across development, whereas Fly-FISH images are valuable for a high-resolution examination of gene expression during the earliest stages of development. Currently, searching for Fly-FISH images using a BDGP query (and vice versa) cannot be conducted in FlyExpress, as the images from each dataset are sufficiently different as to be incompatible. However, both BDGP and Fly-FISH curators assign images to stage ranges and textually annotate them with CV terms. Thus, FlyExpress allows one to compare image and text-based search techniques within either dataset.

Image-based analyses are facilitated in FlyExpress by the Basic Expression Search Tool for Images (BESTi; Kumar et al., 2002). For each gene, BESTi uses spatial profiles derived from images capturing gene expression by means of a series of computational filtering and adaptive-thresholding techniques (see the Experimental Procedures section). This accommodates images captured under varying microscopy and experimental conditions. Three binary spatial profiles per standardized image are included to develop discrete representations that accommodate expression intensity differences across the embryo (black and white spatial profiles). FlyExpress users may choose one of these three spatial profiles (*a*, *b*, or *c*) to conduct a search for overlapping expression. For each standardized image, the *b* spatial profile represents expression most closely depicted in the image itself, while the *a* and *c* profiles represent under or over-staining, respectively.

Just as each standardized image is assigned a unique identification number in FlyExpress, their respective spatial profiles are also assigned a unique identifier. Inclusion of spatial profiles allows FlyExpress users to further customize searches. To discover genes that show expression

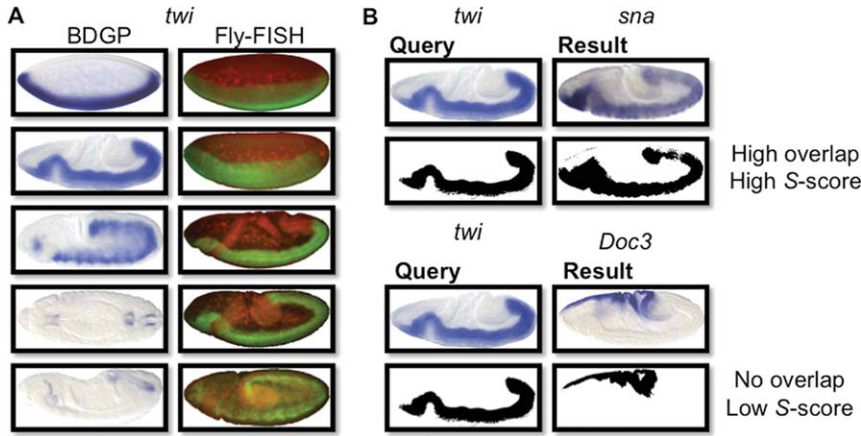


Fig. 1.

Fig. 1. *twist* gene expression patterns and spatial similarity. **A:** Standardized BDGP (left) and Fly-FISH images (right) viewable in FlyExpress depicting expression of *twist* (*twi*) are shown at progressively older developmental stages from top to bottom in both columns. In BDGP images gene expression is blue and in Fly-FISH images gene expression is green/yellow. **B:** Spatial similarity is expressed as an S-score and measured based on the common presence and absence of expression at corresponding coordinates in the spatial profiles of query and database images. Using the *twi* image presented in (A), we show a *sna* spatial profile and corresponding standardized image that display considerable spatial overlap and FlyExpress calculates the pair has a high S-score ($S = 0.57$ reflecting 57% overlap). For images with no overlap in comparison to *twi*, such as *Doc3*, FlyExpress calculates an S-score for the pair of zero ($S = 0.0$).

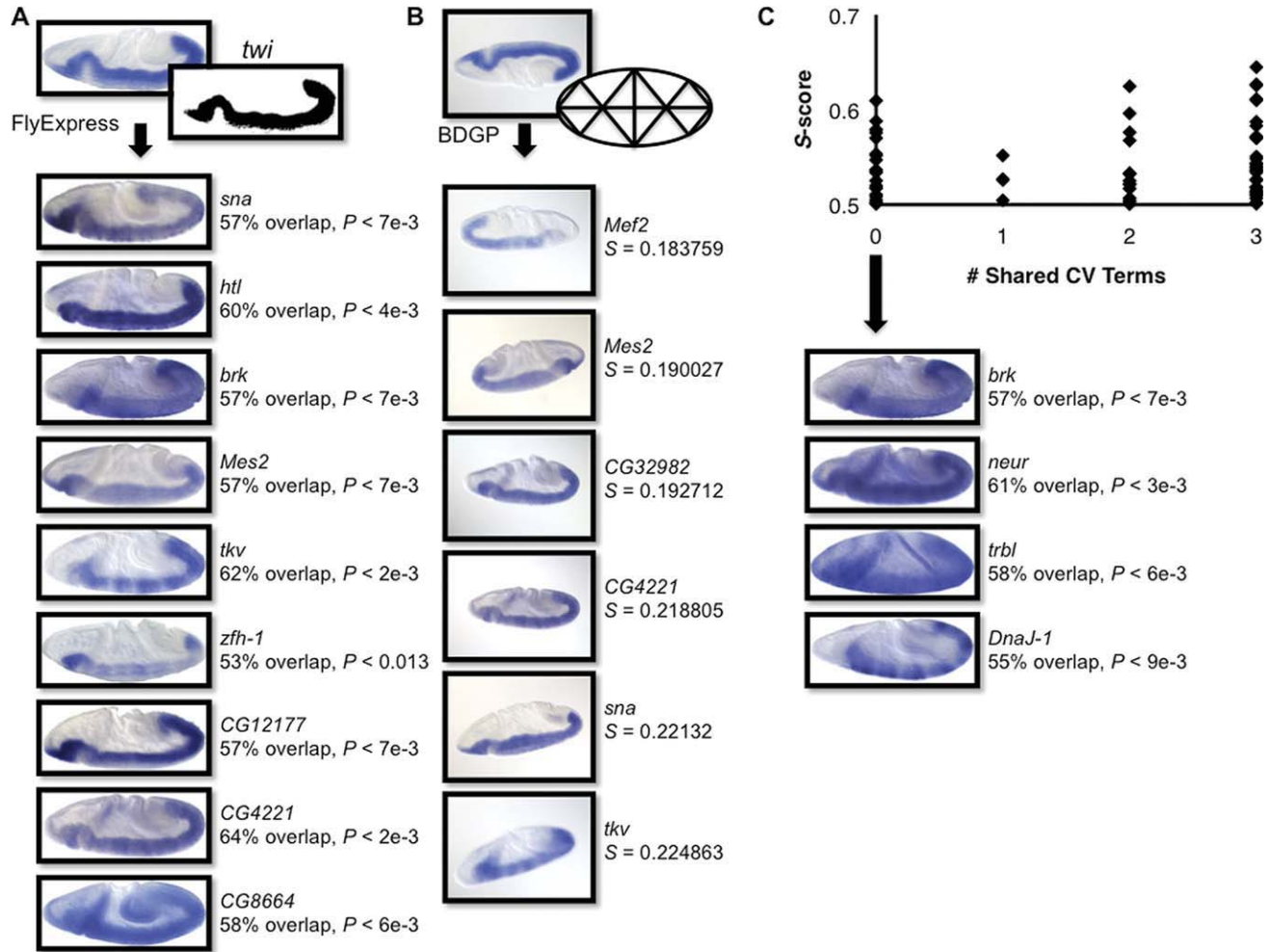


Fig. 2. Image searching method comparison for BDGP data. **A,B:** Images searches for overlapping expression patterns for the same query image were performed for *twi* using FlyExpress (A; FBim9035159 FlyExpress profile c) and BDGP's image search tool that disregards embryo orientation (B). The top nine images of 86 are shown for FlyExpress while all six images are shown for the BDGP method. **C:** The number of CV terms a given FlyExpress identified gene shares with the query gene are graphed. Examples of four images with $\geq 50\%$ overlap but sharing 0 CV terms with the query gene are shown below the graph.

patterns similar to that of a query gene, FlyExpress compares the spatial profiles of that gene with all the black and white spatial profiles in the database. These comparisons are restricted to images from the same data source, anatomical view, and stage range as the query pattern.

Spatial similarity is measured based on the common presence and absence of expression at corresponding coordinates in the spatial profiles of query and database images. For patterns with larger overlaps, the similarity value (*S*-score) will be higher (Fig. 1B). The statistical significance of the observed match is assessed using the probability of randomly finding a match occurring with the given similarity or better (*P*-value). The distribution used to derive the *P*-value for a given *S*-score is generated from pairwise comparisons of all patterns in the query database in a stage and view specific manner (see the Experimental Procedures section and Supp. Fig. S2, which is available online, for details). This method is analogous to that used in the BLAST search for homologous sequences (Altschul et al., 1990).

Image vs. Text-Based Searching

We began our analysis of the relative value of data-driven (image-based) vs. text-based searching by using the well-known *twist* (*twi*) expression pattern depicted in Figure 1A. *twist* expression is activated by the dorsal–ventral axis organizing gene *Dorsal*; its expression defines the ventral domain of the embryo and its function dictates that these cells become mesoderm (reviewed in Stathopoulos and Levine, 2002). FlyExpress searching (FBim9035159, profile *c*) produces 86 genes with a *P*-value < 0.02 and *S*-score ≥ 0.5 . The six of the nine most similar expression patterns found by searching with the query image shown in Figure 2A are biologically meaningful. They include known *twi* target genes such as *sna* (Ip et al., 1992), *htl* (Stathopoulos et al., 2004) and *brk* (Markstein et al., 2004) and additional genes with roles in dorsal–ventral axis or mesoderm specification such as *Mes2* (Zimmermann et al., 2006), *tkv* (Dorfman and Shilo, 2001) and *zfh-1*

(de Velasco et al., 2006). Importantly, genes that are not currently characterized, such as *CG12177*, *CG4221*, and *CG8664*, also exhibit significant overlap. These genes are now candidates for testing to determine their involvement, if any, in the *twi* developmental network or dorsal–ventral axis formation.

We next compared FlyExpress image-based search results for *twi* with those of BDGP's more complex image-based search tool (Frise et al., 2010). In contrast to FlyExpress data-driven searching that discovers similar and overlapping expression patterns simply based on the presence or absence of expression at a given point, BDGP data-driven searching uses a more sophisticated method that also considers expression intensity (Frise et al., 2010). Using the same query image, BDGP yields a set of six genes, all of which exhibit nearly congruent gene expression with the query pattern (Fig. 2B). This is in contrast to FlyExpress, which returns a longer list of genes (including the six found by BDGP) that exhibit a wider range of spatial overlap with the query and each other.

We then conducted a text-based search using CV terms associated with the *twi* query image (trunk mesoderm primordium P2, head mesoderm primordium P4 and anterior endoderm anlage). A purely text-based search will yield 116, 122, and 246 genes sharing 1, 2, or 3 CV terms with the query gene, respectively. The gene counts produced by text searching are significantly larger than those obtained from either type of image-based search and they have the widest range of expression overlap with the query (Fig. 2C). Of the 86 FlyExpress identified genes having $\geq 50\%$ overlap with the *twi* query 25, 4, 18, and 39 genes shared 0, 1, 2, or 3 CV terms respectively, with the query. The ability of FlyExpress to identify 25 genes with patterns that overlap $\geq 50\%$ with the *twi* pattern but share no common CV terms with *twi* (29% of the FlyExpress list) is strong evidence of the value of image-based search techniques for biological discovery. Four of these 25 are shown in Figure 2C.

We then performed a FlyExpress image-based search of a Fly-FISH *twi* image (Fbim9497472, profile *a*) from a

similar developmental stage and anatomical view. Note that represented genes and embryonic stage ranges for the BDGP and Fly-FISH datasets are different, and thus *P*-values will vary, even if similar query images are chosen to search within each image collection. FlyExpress searching identifies 132 genes with a *P*-value < 0.05 and *S*-score ≥ 0.5 (Fig. 3A). Only one gene from the top nine of the FlyExpress *twi* search of BDGP (Fig. 2A; *tkv*) was among the top nine in the Fly-FISH search. To gain insight into this discrepancy, we first examined the Fly-FISH dataset for the presence of the other eight top genes identified in the FlyExpress BDGP *twi* search. All but one (*htl*) is absent from the Fly-FISH dataset. Upon closer inspection, we found *htl* among our initial Fly-FISH results, but the image displayed slightly < 50% overlap to the *twi* query image and thus was not counted in the final tally. Note that many genes showing ubiquitous or near-ubiquitous expression had slightly > 50% overlap with the query spatial profile, thus providing a possible explanation for this discrepancy. Note that our 50% overlap cutoff is also relatively arbitrary. If the cutoff was lowered to 45% then the *htl* image would be included. As for BDGP above, a FlyExpress search of the Fly-FISH dataset returned genes with significant expression overlap to the query and to each other. At this time there is no other image searching method for analyzing Fly-FISH data.

The Fly-FISH *twi* query image is associated with four CV terms (blastoderm nuclei, expression in mesoderm, subset blastoderm nuclei and zygotic). A purely text-based search will yield 151, 98, 393, and 93 genes sharing 1, 2, 3, and 4 CV terms with the query image. Text-based searching within FlyExpress identified set of 132 genes revealed that 103, 8, 1, 13, and 7 share 0, 1, 2, 3, and 4 CV terms respectively, with the query (Fig. 3B). The ability of FlyExpress to identify 103 genes with patterns that overlap $\geq 50\%$ with the *twi* pattern but share no common CV terms with *twi* (78% of the FlyExpress list) is again strong evidence of the value of image-based search techniques for biological discovery. Four of the 103 images belonging to genes associated with 0 shared

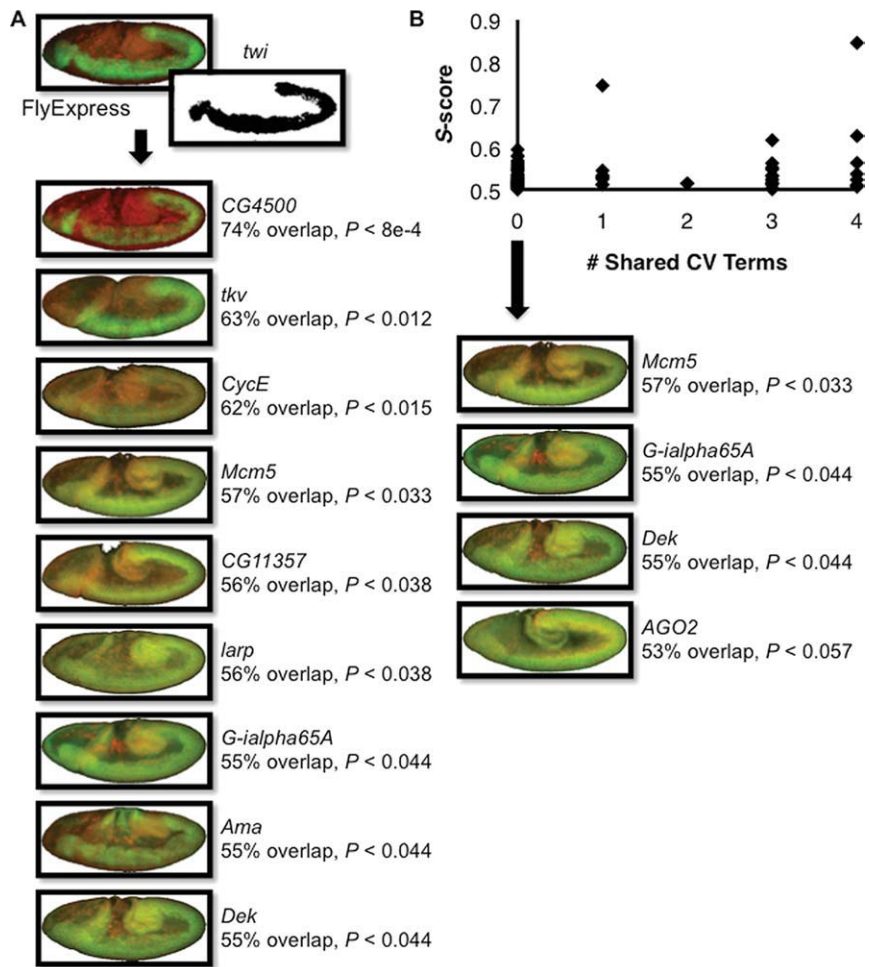


Fig. 3. Image vs. text searching method comparison for Fly-FISH data. **A:** A FlyExpress search for overlapping expression of *twi* using a similar image (Fbim9497472 FlyExpress profile a). The top nine images of 132 genes identified are shown. **B:** The number of CV terms a given FlyExpress identified gene shares with the query gene are graphed. Examples of four images with $\geq 50\%$ overlap but sharing 0 CV terms with the query gene are shown below the graph.

CV terms are shown in Figure 3B. Note that many genes displayed ubiquitous or near-ubiquitous expression and are textually annotated as such. Thus, when looking for genes showing significant overlap with an expression pattern covering a large area of the embryo an abundance of significant overlap with ubiquitously expressed genes is expected.

Three more analyses of FlyExpress identified images (derived from a variety of query images) with regard to the number of CV terms shared by the FlyExpress list and the query were conducted to determine the robustness of this result beyond a *twi* query image. Independently selected exemplar images chosen from Interactive Fly (Brody, 1999), as well as randomly selected from the BDGP and Fly-FISH datasets were examined. These included embryos of various stages and different views (Supp. Table S1; Supp. Fig. S1). In each analysis 2–82% of the FlyExpress identified genes with $> 50\%$ overlap with the query did not share a CV term with the query and would have been missed in a text search. Examination of the images missed by text-based searching revealed that if the query embryo has an expression pattern covering a large area of the embryo, then ubiquitous expression patterns were often missed.

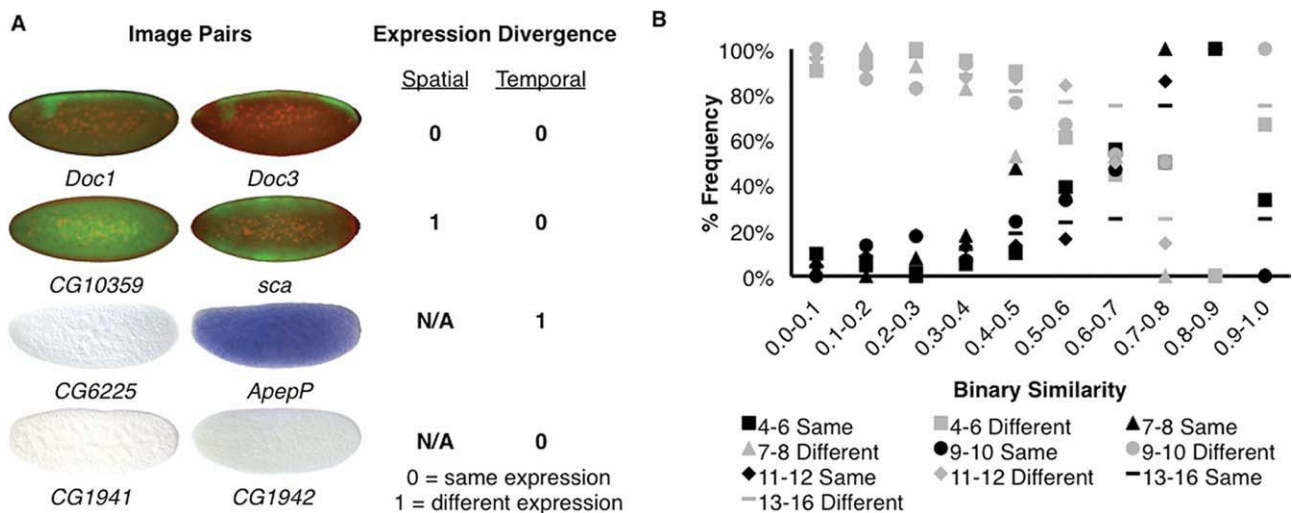


Fig. 4. Paralogous gene pair classification with analysis of spatially and temporally different gene pairs over developmental time. **A:** Example image pairs classified as having the same (0) or different (1) spatial or temporal expression patterns. BDGP gene expression is in blue. Fly-FISH gene expression is in green/yellow. N/A - no divergence was visible. **B:** For the BDGP dataset the frequencies of spatially different and temporally different gene pairs within all gene pairs were graphed according to developmental stage and S-score. Images exhibiting ubiquitous or no expression are not included. Legend showing characters used on the graph for each developmental stage and its two expression pattern comparisons (similar or different).

Paralogous Gene Expression Analysis Using FlyExpress

Using our data-driven approach, we next demonstrate how the FlyExpress resource may be used to analyze gene co-expression. Standardized *in situ* images from both datasets for 251 multigene families ranging in size from 2–20 members (801 genes total; Supp. Table S2) were manually compared, in a pairwise manner, to determine the presence or absence of spatial and/or temporal expression pattern divergence. Standardized images enable this type of quantitative image-based analysis, as embryos and their expression patterns are comparable if they are of the same data source, stage and anatomical view, due to uniform size and shape (Fig. 4A). Two family members (evolutionarily known as paralogs because we are examining genes in the same species) were judged to have the same spatial expression if their images from the same anatomical view, developmental stage range, and data source exhibited expression in the same embryonic regions. Two paralogs were judged to have the same temporal expression if their images depicted the presence or absence of expression in the same developmental stage range. Corroborating microarray data was obtained to improve confidence (Arbeitman et al., 2002).

Image pairs were also compared computationally (post hoc) using the Basic Expression Search Tool for Images (BESTi; (Kumar et al., 2002)). As expected, image pairs showing the highest levels of divergent expression in the manual inspection assay also had the lowest *S*-scores (smallest overlap) and image pairs with the lowest levels of divergent expression had the highest *S*-scores (highest overlap; Figure 4B shows the graph for spatial expression). For an *S*-score between 0.6 and 0.7 the similar and different expression pairs are at roughly equal frequency. However, for an *S*-score of 0.9 to 1.0 the clean correlation between *S*-score and manual assessment of expression overlap ceases.

To determine the source of the discrepancy between manual and computational estimates of similarity we reexamined the datasets and soon realized that computers, unlike humans,

cannot recognize and discard noise automatically. In such cases, BESTi may underestimate the amount of expression similarity for a particular gene pair. For example, if one image has artifactual staining in addition to normal gene expression while another does not, or if different artificial staining is present in two images that otherwise depict the same expression pattern true similarity will be underestimated. A second example is when one image is over-stained while another is under-stained for the same gene expression pattern. A third example applies particularly to older embryos during organogenesis where it results from naturally occurring morphological movements. For instance, Malpighian tubules and peripheral neurons migrate in all directions, elongate, twist, extend and retract almost continuously during late stages of development. As a result, a comparison of images with expression of the same gene in two stage 16 embryos that are 15 min apart in age may have a similarity score less than 1.0 due to morphological events that occurred during those 15 min. The skill necessary to ascertain the age of an embryo, within 15 min, takes considerable effort to develop.

Alternatively, BESTi may overestimate expression similarity between two images. The most common instance of this is when expression appears in the same two-dimensional space but is, in reality, present in different regions of three-dimensional space (e.g., above and below the plane of focus). For example, visceral musculature and fat body expression may be impossible to distinguish two-dimensionally in images of laterally oriented stage 13 embryos. Comparison of expression by a gene present in the visceral musculature and another in fat body will lead to significant similarity overestimation. See Supp. Fig. S3 for examples illustrating how binary similarity may be over or underestimated. Thus, while the FlyExpress resource simplifies the identification and quantification of similar expression patterns, the biological importance of these results must be evaluated by a biologist, as is true for all computational systems.

Of 1,068 gene pairs in our analysis, 807 (75.6%) exhibited spatial gene

expression divergence during at least one stage of development. To determine if spatial pattern divergence is more frequent in genes with particular functions, we tested the most divergent gene pairs for GO term enrichment. In the analysis, members of gene pairs showing a predominance of spatial divergence across development (the pair had divergent comparisons over a majority of stages) were compared with the larger group of gene pairs that simply had enough image data to be assessed over the majority of developmental stages. We found three Biological Process terms at a statistically significant frequency (Fig. 5A; Supp. Table S3).

To determine if spatial divergence varied across developmental stages, the proportion of gene pairs exhibiting spatial differences was determined for each developmental stage range. These proportions for each dataset, individually and combined, were then graphed according to median stage range developmental times (Campos-Ortega and Hartenstein, 1985). The graph (Fig. 5B) reveals the fraction of spatially divergent gene pairs increases significantly in all three cases (P -value < 0.05) as embryogenesis progresses. Just 2.9% of the pairs show spatial differences in the youngest embryos (stages 1–3) that increased to 93.2% of pairs in the oldest (stages 13–16). Specifically, in the earliest stages there is an excess of spatial similarity because ubiquitous expression is abundant. The proportion of divergent pairs then expands swiftly with stages 4–6 showing roughly equal proportions of similar and different pair expression. By stages 10–12 (just halfway through development based on time), nearly 90% of the gene pairs show spatial divergence and spatial divergence reaches a maximum shortly thereafter.

In contrast, there were just 168 (15.7%) gene pairs exhibiting temporal divergence over at least one stage range. Note that only 291 of the gene pairs had comparable image data for at least one stage range in the Fly-FISH dataset, compared with 869 in BDGP (91 gene pairs had images in both datasets). This is almost fivefold less than the number of gene pairs exhibiting spatial divergence in at least one stage range. This distinction

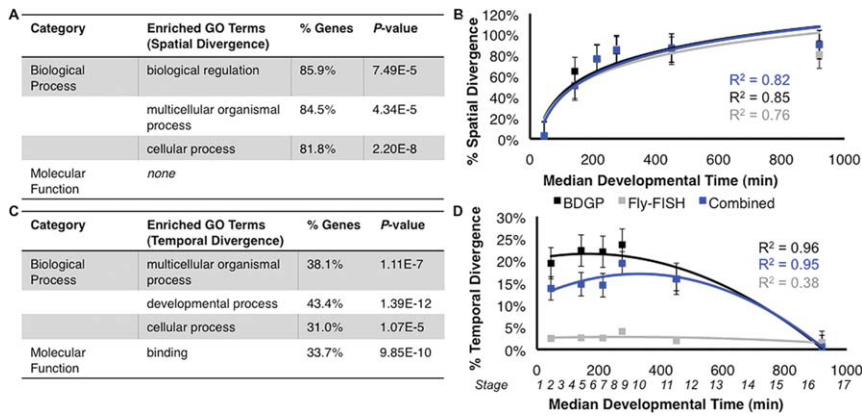


Fig. 5. Paralogous gene pair expression analysis. **A:** Statistically significant GO term enrichment is present for three terms in members of gene pairs with spatial expression divergence across a majority of developmental stages. **B:** An average spatial divergence was calculated for all pairwise comparisons with spatial differences, and noted as percent divergence, for the BDGP, Fly-FISH, and combined datasets. These averages are graphed according to developmental stage after each stage was converted to time in minutes since fertilization based on the established timeline of embryonic development in *Drosophila*. Standard deviations are shown and statistical significance (P -value < 0.05) was determined by the two-tailed P -value for a comparison of the mean divergence values of the first and last developmental time points. The median times of the seventeen Campos-Ortega and Hartenstein (1985) stages are also included below the minute intervals. **C:** Statistically significant GO term enrichment is present for four terms in members of gene pairs with temporal expression divergence across a majority of developmental stages. **D:** The average divergence of gene pairs showing temporal differences was calculated and graphed as above.

was truly unexpected as a priori there was no reason to think that either spatial or temporal divergence would predominate. To determine if temporal divergence was associated with gene pairs of specific function, GO term enrichment was again examined. As above, members of gene pairs showing a predominance of temporal divergence across development (the pair had divergent comparisons over a majority of stages) were compared with the larger group of gene pairs that simply had enough image data to be assessed over the majority of developmental stages. We found three Biological Process terms at a statistically significant frequency (Fig. 5C; Supp. Table S4). Two of these terms were also found among spatially divergent pairs. The term “biological regulation” is enriched only in spatially divergent genes, while the term “developmental process” is enriched only in temporally divergent ones. Furthermore, temporally divergent pairs are enriched in the Molecular Function term “binding.”

To determine if temporal divergence varied across developmental stages, the proportion of gene pairs exhibiting temporal differences was determined for each developmental

stage range. In the BDGP dataset, the fraction of gene pairs displaying temporal expression differences is roughly stable during the first third of development at 20–24% (stages 1–9 are not significantly different) but decreases significantly (P -value < 0.05) during the last two-thirds of development (Figure 5D). The Fly-FISH dataset, with its emphasis on early embryos (the vast majority derive from stages 1–9), showed few temporally different pairs, as well as relatively fewer comparable pairs overall compared with BDGP, and thus was unchanged over time. The decrease in temporal pairs for BDGP (which as the larger dataset is responsible for the combined results) is in sharp contrast to the trajectory for spatial divergence (continuously increasing in frequency). For BDGP, the fraction of temporally different pairs ranges from a high of 24% (stages 8–9) to a low of 0.6% in the oldest embryos (stages 13–16).

Taken together the spatial and temporal divergence data reveals that gene pairs within a multigene family are more likely to diverge in expression temporally rather than spatially before the maternal to zygotic expression transition (stages 4–5; 20% tem-

porally vs. 2.9% spatially for BDGP). In contrast, during late stages of development gene pairs within a multigene family are more likely to diverge in expression spatially than temporally (stages 13–16; 90.3% spatially vs. 0.6% temporally for BDGP).

DISCUSSION

Overall, our results for image vs. text-based searching illustrate that image-based searching for overlapping expression patterns using the FlyExpress resource presents unique advantages compared to text-based searching. Importantly, the comparison of image and text-based searching suggests that co-expressed genes (even when they display substantial overlap) do not always share common text annotations. Therefore, if one’s aim is to find genes with similar or overlapping patterns of expression then direct searching of standardized expression patterns is more powerful than the text-based searching. However, as our study of spatial divergence suggests, users of image searching must remember that, as with any computational method, decisions about biological significance are theirs.

In addition, textual descriptions remain an important feature of discovery in developmental biology because they are crucial for communicating expression details across different embryonic stages, anatomical views and even species boundaries. Textual annotations also allow incorporation of the three-dimensional nature of the developing embryo, and keen annotators may easily distinguish between expression and background or artifact staining. Therefore, FlyExpress incorporates BDGP and Fly-FISH produced text information alongside image data so users can more easily discern the biological relevance of search results.

Our analysis of expression divergence among paralogous genes in multigene families suggests spatial rather than temporal divergence in gene expression patterns is the predominant contributor to the development of new morphological features, especially after the maternal to zygotic transition. One example of a multigene family with significant spatial and temporal divergence is the

innexin family. This family is highly conserved and plays a variety of roles in cell migration and in organogenesis (Bauer et al., 2002; Lechner et al., 2007). Genes associated primarily with spatial divergence are also enriched in the GO term “biological regulation.” One example of a multi-gene family with significant spatial but not temporal divergence is the *Minichromosome maintenance* family. This family is highly conserved across phyla and functions in genome maintenance and DNA replication (Forsburg, 2004). Genes primarily associated with temporal divergence are enriched in two different GO terms: “developmental process” and “binding.” One example of a multigene family with significant temporal but not spatial divergence is the *Imaginal Disc Growth Factor (IDGF)* family. This family is also conserved (evolved from chitinases) and influences cell proliferation (Kawamura et al., 1999; Zurovcová and Ayala, 2002).

Looking toward the future, increasingly sophisticated and interconnected computational tools for image analysis of gene expression patterns will speed the discovery of new relationships between genes, genomic segments, and ultimately species. To foster the discovery process, FlyExpress contains automated portals into FlyBase and FlyMine. Looking more broadly, the search method used by FlyExpress could potentially be applied to other two-dimensional expression datasets, provided that images are capable of being effectively standardized. Ultimately, effective standardization will result in significant overlap of morphological features, and thus allow for expression pattern comparisons.

Our usage of FlyExpress to identify clues to molecular mechanisms underlying the process whereby new genes acquire novel functions is just one of a myriad ways in which this discovery platform can contribute to our understanding of developmental and evolutionary biology. This resource also has many other potential uses, limited only by the investigator’s imagination. One example is to examine expression overlap in gene networks, particularly threshold activation of genes in response to important signaling gradients such as the one produced by *Dorsal* (Stathopoulos and Levine,

2002; Papatsenko and Levine, 2005). Alternatively, FlyExpress could be used to predict interactions between embryonic genes with specific spatial expression profiles, as has been done in oogenesis (Yakoby et al., 2008).

EXPERIMENTAL PROCEDURES

FlyExpress Discovery Platform: Gene Expression Images and Image-Based Searching

Digital images revealing patterns of *D. melanogaster* gene expression, captured by RNA in situ hybridization, were retrieved from BDGP Release 2 (Tomancak et al., 2002) and Fly-FISH September 2007 release (Lécuyer et al., 2007). All images were processed using an in-house, semi-automated pipeline to standardize and align embryos, where multi-embryo images were manually divided into separate images and partial embryo images were discarded. Image processing was carried out using Matlab and our own image processing routines. In this pipeline, images were extracted by means of the following procedure (Matlab functions are shown in italics): read original image with *imread*, convert to grayscale with *rgb2gray*, apply windowed low-pass Gaussian filter (*imfilter* and *fspecial*) to blur the image a little so that edge detection would detect only the main outside edges of the embryo and not edges caused by expression patterns, shadows, etc., delineate embryo edges using *edge* Canny edge detection, expand points with *imdilate*, fill holes in the image with *imfill*, shrink dilations, blur again and sharpen the image edges with *imerode*, *strel*, and *medfilt2*. The *bwlabel* function is used to identify individual embryo boundaries. Finally all pixels outside the selection region (outside the embryo) are set to pure white. The resulting image is saved in RGB color as a bitmap file.

The next standardization step is embryo alignment, which is done by rotating embryos using *imrotate*, such that the major axis of the embryo is parallel to the horizontal, drawing a bounding box to enclose the embryo, and cropping the embryo using *imcrop* to automatically remove back-

ground external to the smallest embryo bounding box. For consistent orientation, we used anterior-on-the-left and dorsal-toward-the-top format for lateral images and anterior-on-the-left for all other views (e.g., dorsal and ventral; see also Kumar et al., 2002). During quality control, experienced biologists corrected orientation and alignment images using *flipdim*, as necessary.

To size standardize and align all images, we chose a cellular aspect ratio of 2.5 (320 × 128 pixels) based on natural aspect ratios (Markow et al., 2009), on the need to avoid pixel padding in image representation and to make sure that each line of pixels and all images end on byte, word and long word boundaries. Using the *imresize* function, all embryo images were resized and a standardized collection created.

Developmental stages for Fly-FISH and BDGP embryos are available from the image source and were thus assigned to embryos. BDGP embryos are annotated with developmental stage ranges (1–3, 4–6, 7–8, 9–10, 11–12, or 13–16) using the Bownes system (16 stages; Roberts, 1986) whereas Fly-FISH embryos are classified into stage ranges (1–3, 4–5, 6–7, 8–9, and later) using the Campos-Ortega and Hartenstein (1985; 17 stages) system. We found that a large fraction of Fly-FISH images contained multiple embryos from different developmental stages. We carefully reviewed and assigned appropriate stage ranges to individual embryos for these cases. We also assigned anatomical embryo views (e.g., lateral, dorsal, and ventral) for all images, because computational comparison of spatial expression patterns is only biologically meaningful if conducted within each view. These stage and view assignments were added during our quality control process, where all embryos were examined by at least one developmental biologist and image standardization and expression extraction were carried out manually, when needed. This produced a total of 99,148 standardized embryo images: 42,065 Fly-FISH and 57,083 BDGP. Within BDGP, there are 14,257 dorsal, 37,825 lateral, and 5,964 ventral views, and within Fly-FISH, there are 3,494 dorsal, 37,211 lateral, and 1,360 ventral views.

Comparative expression analysis to identify images (and thus genes) with overlapping expression patterns requires digital descriptions of spatial expression patterns (Kumar et al., 2002). We used adaptive intensity and color thresh-holding to automatically delineate the expression profile from the embryo background. For Fly-FISH images, expression patterns are captured in green and yellow colors, and in BDGP images, blue color captures the spatial profile. In the process of image extraction there are four parameters: the color layer (red, green, or blue), whether to reverse intensity, an adjustment parameter for shift threshold and the amount of variance between the three expression patterns produced. The RGB image is loaded with *imread*. Based on the color channel (layer) parameter, a histogram of intensities for just that color layer is created with *imhist*. The histogram is “pre-cut” to only include intensities between 5 and 250. The sum of the intensities in the pre-cut histogram is calculated with *sum*. Upper and lower limits of the histogram are computed for the range between 10% and 90% of this sum. An area (“integration”) vector is calculated for the color layer (specified by the color parameter) of the original image. Using this area vector and the original intensities, we calculate the first moment and the centroid of the area under the curve about the y-axis and the first moment and the centroid of the area under the curve about the x-axis. A threshold value is calculated as the centroid about the y-axis + the lower limit calculated above. The minimum distance from the x–y center is calculated for each point in the intensity histogram. The *plus_da* (the differential best area + variance) and *minus_da* (the differential best area - variance) are calculated. *plus_da* and *minus_da* thresholds are calculated from these. The image background is cleared using *imdilate* (Image, SE).

The three binary extraction patterns are extracted from this image with cleared background based on the adjustment parameter and the threshold (for the best extraction), the *minus_da* threshold (for the minus extraction) and the *plus_da* threshold (for the plus extraction). The code

for this algorithm is available upon request. Three binary (black and white) patterns enable searches of the database using patterns at different levels of expression intensity (*a*, *b*, and *c* patterns). Embryos with ubiquitous expression in the earliest stages of development as well as those with no expression were noted and marked for exclusion from comparative image analysis.

To identify the degree of spatial overlap between patterns, we used the low-level bitmap Jaccard similarity index (Kumar et al., 2002), which traces its roots to the Tanimoto measure (Tanimoto, 1958) and is a member of a family of similarity measures that include Taversky, Euclidian, Hamming, and Ochia measures (Bradshaw, 2001). In this case, the similarity score (*S*) between two images (*Q* and *D*) is given by $S_{QD} = |Q \cap D| / |Q \cup D|$, where $|Q \cap D|$ is the size of the intersection of expression (count of black pixels) between images *Q* and *D* and $|Q \cup D|$ is the size of the union between images *Q* and *D*. We have previously shown that this approach emphasizes spatial overlap, which is biologically more meaningful than shape matching and invariant moment based features (Kumar et al., 2002; Gurunathan et al., 2004). We have also found that it performs with an effectiveness similar to the computationally more intensive Gaussian Mixture Model method (Peng et al., 2007), which in our hands is very sensitive to shifts in image properties, such as the color and contrast (Gargsha et al., 2008, 2009a,b; Roy et al., 2009). Thus, we used multiple binary feature vectors to represent the expression information in our analysis and provide the greatest flexibility. For each *S*-score, we also computed the probability that any pair of images will show an equal or higher value by chance alone (*P*-value). Empirical *S*-score distributions derived from image pairs from the same data source (BDGP or Fly-FISH), developmental stage and anatomical view are used to determine *P*-values (Supp. Fig. S2).

For images from PubMed publications, gene expression patterns were manually extracted from the PDF file by our pipeline team. Extracted images were then standardized and expression pattern representations

created using the same methods described above for BDGP and Fly-FISH. Images were manually annotated, for example for developmental stage and anatomical view (lateral, dorsal, ventral), by our team of curators at Harvard (FlyBase) and by biologists on our pipeline team. Images from publications present special challenges due to their varying quality. For images of low resolution or poor quality, it is not possible to make specific developmental stage determination. In this case the best possible stage range (from-stage/to-stage) is determined for the annotation. Publication information (for example author, year, and title) for the papers was obtained from the FlyBase publically available database.

A detailed description of all computational aspects of FlyExpress and of the extracted gene expression pattern database can be found in Kumar et al. (2011) and the source code for all algorithms is available upon request.

Finding and Aligning Multigene Families

All genes present in the BDGP (Tomancak et al., 2002) and Fly-FISH (Lécuyer et al., 2007) datasets were analyzed by *tblastn*, using the longest protein isoform available from NCBI and the *D. melanogaster* genome sequence. Genes were considered paralogous if reciprocal hits containing $\geq 50\%$ amino acid positives with *E*-value $< 10^{-20}$ were returned. Additional hits meeting these criteria were analyzed by *tblastn* until no more paralogous genes were retrieved completing that particular multigene family. The subset of multigene families with expression data for at least two paralogs was retained for further analysis. Multigene family sequences were aligned with default parameters in MAFFT (Katoh et al., 2005; Nuin et al., 2006) and sequences re-gapped with GeneDoc (<http://www.nrbsc.org/gfx/genedoc/index.html>). Alignments were corrected by eye and modified (tails cut off) if necessary.

Expression Assessment

For each gene pair, the standardized BDGP or Fly-FISH expression patterns in FlyExpress were manually

compared to determine if the same or different expression patterns were present, either spatially or temporally. Manual comparison was necessary because we could not identify an *S*-score cut-off that adequately represented spatial and temporal divergence in comparisons involving embryos with artifact staining, high backgrounds, or embryo pairs with three-dimensional expression considerations. Even when using manual inspection, posterior spiracles and other openings that are frequently a source of artifacts must show expression in at least two images of the same gene at the same stage and view to be considered legitimate.

Image comparisons were performed without any additional information, but comparisons within a multigene family were verified to be free of logical inconsistencies. Pattern similarity or difference was determined on the basis of the presence or absence of expression within the same region of the embryo. Although only images of the same data source, developmental stage and view were compared, gene pairs were binned into BDGP stage ranges and each pair given an overall assessment. The exception was stage 1–3 BDGP embryos where the exact stage cannot be assigned due to inability to visualize the number of nuclei. Gene pairs were assigned an overall spatial and temporal assessment of Same (0), Different (1), or Ambiguous/No Data (null).

Spatially, only images of the same data source, developmental stage, and anatomical view were compared. A gene pair was classified as having the same expression in a stage if all image pairs showed expression in the same embryonic regions. A gene pair was classified as having divergent expression if image pairs within a stage showed expression in different embryonic regions. Gene pairs with no images meeting the above criteria (including those with no image data) were classified as Ambiguous/No Data.

Temporally, images of the same data source and stage were examined. Two genes were classified as having the same expression in a stage if all image pairs showed expression presence or absence during that stage, accompanied by corroborating microarray data (Arbeitman et al., 2002).

Two genes were classified as having different expression in a stage if at least one image pair showing expression presence for one gene and absence for the other was present, again accompanied by corroborating microarray data. Gene pairs not falling into one of the above categories were classified as Ambiguous/No Data. Gene pairs exhibiting temporal expression differences in a stage were not assessed spatially for that stage.

GO Term Analysis

FlyBase (release FB2009_09) GO term annotations (Tweedie et al., 2009) were retrieved for paralogous BDGP and Fly-FISH genes (http://flybase.bio.indiana.edu/static_pages/term/term/term/term.html). Overall enrichment was determined in the context of total FlyBase term annotations by means of hypergeometric distribution. Significantly enriched terms were those with P -value < 0.0001 . For spatial analysis, there were 211 genes from 158 gene pairs where spatial patterns were assessed over at least 2/3 of BDGP stages. Of these 167 genes from 119 gene pairs showed a spatial expression divergence over a majority of BDGP stages. For temporal analysis, there were 392 genes from 285 pairs assessed over at least 2/3 of BDGP stage ranges. Of these 92 genes (from 57 gene pairs) exhibited temporal expression divergence over a majority of BDGP stages.

Developmental Time Course Analysis

Individual pairwise comparisons were binned according to developmental stage range. The fraction of genes exhibiting spatial or temporal divergence was calculated for each bin (1/0+1) and plotted against the median BDGP stage range times (Campos-Ortega and Hartenstein, 1985). Statistical significance (P -value < 0.05) was determined by the two-tailed P -value.

ACKNOWLEDGMENTS

We thank Toni Marco and Rachel Sipes for assistance with paralogous gene analysis. S.K. was funded by a NIH grant.

REFERENCES

- Altschul S, Gish W, Miller W, Myers E, Lipman D. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410.
- Arbeitman M, Furlong E, Imam F, Johnson E, Null B, Baker B, Krasnow M, Scott M, Davis R, White K. 2002. Gene expression during the life cycle of *Drosophila melanogaster*. *Science* 297:2270–2275.
- Bauer R, Lehmann C, Fuss B, Eckardt F, Hoch M. 2002. The *Drosophila* gap junction channel gene *innexin 2* controls foregut development in response to Wingless signalling. *J Cell Sci* 115:1859–1867.
- Bier E. 2005. *Drosophila*, the golden bug, emerges as a tool for human genetics. *Nat Rev Genet* 6:9–23.
- Bradshaw J. 2001. YAMS - Yet another measure of similarity. In: Euromug01, 13th–14th September 2001, Cambridge (UK).
- Brody T. 1999. The interactive fly: gene networks, development and the Internet. *Trends Genet* 15:333–334.
- Campos-Ortega JA, Hartenstein V. 1985. The embryonic development of *Drosophila melanogaster*. New York: Springer-Verlag. xi, 227 p.
- de Velasco B, Mandal L, Mkrtchyan M, Hartenstein V. 2006. Subdivision and developmental fate of the head mesoderm in *Drosophila melanogaster*. *Dev Genes Evol* 216:39–51.
- Dorfman R, Shilo BZ. 2001. Biphasic activation of the BMP pathway patterns the *Drosophila* embryonic dorsal region. *Development* 128:965–972.
- Drysdale R. 2001. Phenotypic data in FlyBase. Briefings in Bioinformatics 2:68–80.
- FlyBase. 1999. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res* 27:85–88.
- Forsburg SL. 2004. Eukaryotic MCM proteins: beyond replication initiation. *Microbiol Mol Biol Rev* 68:109–131.
- Frise E, Hammonds AS, Celniker SE. 2010. Systematic image-driven analysis of the spatial *Drosophila* embryonic expression landscape. *Mol Syst Biol* 6:345.
- Garghesha M, Jenkins M, Rollins A, Wilson D. 2008. Denoising and 4D visualization of OCT images. *Opt Express* 16:12313–12333.
- Garghesha M, Jenkins M, Wilson D, Rollins A. 2009a. High temporal resolution OCT using image-based retrospective gating. *Opt Express* 17:10786–10799.
- Garghesha M, Qutaish M, Roy D, Steyer G, Bartsch H, Wilson D. 2009b. Enhanced volume rendering techniques for high-resolution color cryo-imaging data. *Proc Soc Photo Opt Instrum Eng* 7262:72655V.
- Grumbling G, Strelets V. 2006. FlyBase: anatomical data, images and queries. *Nucleic Acids Res* 34:D484–D488.
- Gurunathan R, Van Emden B, Panchanathan S, Kumar S. 2004. Identifying spatially similar gene expression patterns

- in early stage fruit fly embryo images: binary feature vs. invariant moment digital representations. *BMC Bioinformatics* 5:202.
- Ip YT, Park RE, Kosman D, Yazdanbakhsh K, Levine M. 1992. dorsal-twist interactions establish snail expression in the presumptive mesoderm of the *Drosophila* embryo. *Genes Dev* 6:1518–1530.
- Janning W. 1997. FlyView, a *Drosophila* image database, and other *Drosophila* databases. *Semin Cell Dev Biol* 8:469–475.
- Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511–518.
- Kawamura K, Shibata T, Saget O, Peel D, Bryant P. 1999. A new family of growth factors produced by the fat body and active on *Drosophila* imaginal disc cells. *Development* 126:211–219.
- Khatri P, Draghici S. 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21:3587–3595.
- Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* 5:R7.
- Kumar S, Jayaraman K, Panchanathan S, Gurunathan R, Marti-Subirana A, Newfeld S. 2002. BEST: a novel computational approach for comparing gene expression patterns from early stages of *Drosophila melanogaster* development. *Genetics* 162:2037–2047.
- Kumar S, Konikoff C, Van Emden B, Busick C, Davis K, Ji S, Wu L, Ramos H, Brody T, Panchanathan S, Ye J, Karr T, Gerold K, McCutchan M, Newfeld SJ. 2011. FlyExpress: visual mining of spatiotemporal patterns for genes and publications in *Drosophila* embryogenesis. *Bioinformatics* (in press; Aug. 15)
- Lechner H, Josten F, Fuss B, Bauer R, Hoch M. 2007. Cross regulation of intercellular gap junction communication and paracrine signaling pathways during organogenesis in *Drosophila*. *Dev Biol* 310:23–34.
- Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, Hughes T, Tomancak P, Krause H. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* 131:174–187.
- Markstein M, Zinzen R, Markstein P, Yee KP, Erives A, Stathopoulos A, Levine M. 2004. A regulatory code for neurogenic gene expression in the *Drosophila* embryo. *Development* 131:2387–2394.
- Markow T, Beall S, Matzkin L. 2009. Egg size, embryonic development time and ovoviviparity in *Drosophila* species. *J Evol Biol* 22:430–434.
- Matthews KA, Kaufman TC, Gelbart WM. 2005. Research resources for *Drosophila*: the expanding universe. *Nat Rev Genet* 6:179–193.
- Nuin P, Wang Z, Tillier E. 2006. The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics* 7:471.
- Papatsenko D, Levine M. 2005. Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the *Drosophila* embryo. *Proc Natl Acad Sci U S A* 102:4966–4971.
- Peng H, Long F, Zhou J, Leung G, Eisen M, Myers E. 2007. Automatic image analysis for gene expression patterns of fly embryos. *BMC Cell Biol* 8(suppl 1):S7.
- Roberts DB. 1986. *Drosophila*: a practical approach. Washington DC: IRL Press. xix, 295 p.
- Roy D, Steyer G, Gargasha M, Stone M, Wilson D. 2009. 3D cryo-imaging: a very high-resolution view of the whole mouse. *Anat Rec* 292:342–351.
- Stathopoulos A, Levine M. 2002. Linear signaling in the Toll-Dorsal pathway of *Drosophila*: activated Pelle kinase specifies all threshold outputs of gene expression while the bHLH protein Twist specifies a subset. *Development* 129:3411–3419.
- Stathopoulos A, Tam B, Ronshaugen M, Frasch M, Levine M. 2004. pyramus and thisbe: FGF genes that pattern the mesoderm of *Drosophila* embryos. *Genes Dev* 18:687–699.
- Tanimoto T. 1958. An elementary mathematical theory of classifications and prediction. In: IBM Corp. Internal Report 17th November.
- Tomancak P, Beaton A, Weiszmam R, Kwan E, Shu S, Lewis S, Richards S, Ashburner M, Hartenstein V, Celniker S, Rubin G. 2002. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol* 3:RESEARCH0088.
- Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, Zhang H. 2009. FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res* 37:D555–D559.
- Walter T, Shattuck DW, Baldock R, Bastin ME, Carpenter AE, Duce S, Ellenberg J, Fraser A, Hamilton N, Pieper S, Ragan MA, Schneider JE, Tomancak P, Heriche JK. 2010. Visualization of image data from cells to organisms. *Nat Methods* 7:S26–S41.
- Yakoby N, Bristow CA, Gong D, Schafer X, Lembong J, Zartman JJ, Halfon MS, Schupbach T, Shvartsman SY. 2008. A combinatorial code for pattern formation in *Drosophila* oogenesis. *Dev Cell* 15:725–737.
- Zimmermann G, Furlong EE, Suyama K, Scott MP. 2006. Mes2, a MADF-containing transcription factor essential for *Drosophila* development. *Dev Dyn* 235:3387–3395.
- Zurovcová M, Ayala F. 2002. Polymorphism patterns in two tightly linked developmental genes, *Idgf1* and *Idgf3*, of *Drosophila melanogaster*. *Genetics* 162:177–188.