Vol. 26 no. 6 2010, pages 761–769
doi:10.1093/bioinformatics/btp658

Gene expression

Advance Access publication November 26, 2009

Extraction and comparison of gene expression patterns from 2D RNA *in situ* hybridization images

Daniel L. Mace[1,2], Nicole Varnado[2], Weiping Zhang[2], Erwin Frise[3] and Uwe Ohler[2,*]

[1]Computational Biology and Bioinformatics Graduate Program, [2]Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708 and [3]Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Recent advancements in high-throughput imaging have created new large datasets with tens of thousands of gene expression images. Methods for capturing these spatial and/or temporal expression patterns include *in situ* hybridization or fluorescent reporter constructs or tags, and results are still frequently assessed by subjective qualitative comparisons. In order to deal with available large datasets, fully automated analysis methods must be developed to properly normalize and model spatial expression patterns.

Results: We have developed image segmentation and registration methods to identify and extract spatial gene expression patterns from RNA *in situ* hybridization experiments of *Drosophila* embryos. These methods allow us to normalize and extract expression information for 78 621 images from 3724 genes across six time stages. The similarity between gene expression patterns is computed using four scoring metrics: mean squared error, Haar wavelet distance, mutual information and spatial mutual information (SMI). We additionally propose a strategy to calculate the significance of the similarity between two expression images, by generating surrogate datasets with similar spatial expression patterns using a Monte Carlo swap sampler. On data from an early development time stage, we show that SMI provides the most biologically relevant metric of comparison, and that our significance testing generalizes metrics to achieve similar performance. We exemplify the application of spatial metrics on the well-known *Drosophila* segmentation network.

Availability: A Java webstart application to register and compare patterns, as well as all source code, are available from: http://tools.genome.duke.edu/generegulation/image_analysis/insitu

Contact: uwe.ohler@duke.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 29, 2009; revised on November 18, 2009; accepted on November 19, 2009

1 INTRODUCTION

Advances in high-throughput microscopy have led to a rapid increase of digital image data in biology. New methods to image biological specimens at high resolution, and to visualize expression of genes of interest, have lead to a high interest in imaging in developmental and molecular biology, including the creation of virtual embryos to map expression profiles of important regulatory genes (Fowlkes *et al.*, 2008; Keller *et al.*, 2008). In the past, these images were analyzed in a manual fashion, e.g. comparing two expression patterns by qualitative visual inspection. In order to deal with these datasets more appropriately, it is necessary to develop automated methods for extracting and analyzing images. Methods to quantitatively describe spatial/temporal expression patterns is a relatively new area of research that has begun to be explored in model organisms (Megason and Fraser, 2007). In sea urchins, a comprehensive analysis of normalization methods was used for determining and quantifying spatial expression data (Damle *et al.*, 2006). Recently, large databases of images for *Caenorhabditis elegans* have become available (Hunt-Newbury *et al.*, 2007), accompanied by methods to analyze the data (Bao *et al.*, 2006; Murray *et al.*, 2008). For *Arabidopsis* root images, image registration techniques have been used to quantify tissue-specific expression from green fluorescent protein (GFP) reporter lines (Lee *et al.*, 2006; Mace *et al.*, 2006).

A particularly interesting and fundamental problem that arises with the availability of image data, and which we address in this study, is to compare two samples on the level of their expression profiles: for instance, the same gene under different conditions or across different species, or different genes with the goal to cluster them akin to approaches developed for microarray data. Several general problems arise when comparing image expression data: (i) we need to develop methods to process the raw input images, to eliminate noise under a typical large range of imaging conditions (e.g. different viewpoints, different locations, multiple specimens per image) and to perform normalizations to decouple the variability in morphology from the variability in expression; (ii) we need to represent the expression patterns and to specify appropriate similarity metrics capable of assessing spatial/temporal similarity; and (iii) we need to assess the significance of observed similarities.

The majority of image analysis work in the context of development of model organisms has been carried out for *Drosophila*, and can be broadly grouped into two categories: quantitative high-resolution analysis of a relatively small selected subset of genes (Fowlkes *et al.*, 2008; Janssens *et al.*, 2005; Keranen *et al.*, 2006), and higher throughput pattern analysis of thousands of genes (Harmon *et al.*, 2007; Kumar *et al.*, 2002; Peng *et al.*, 2007). Fine-grained high-resolution analysis often extracts quantitative expression values from image data, e.g. for numerical simulations of specific regulatory pathways. This study follows

[*]To whom correspondence should be addressed.

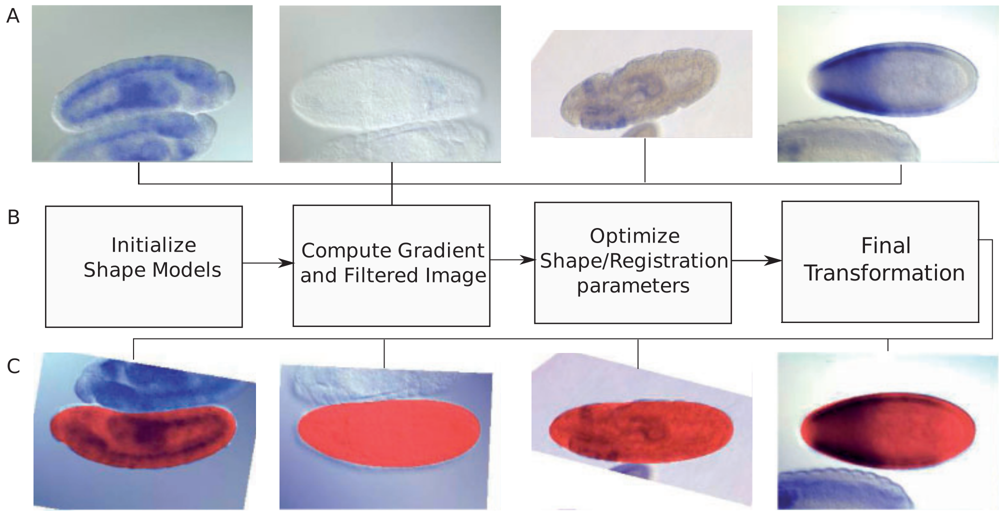
A
B
Initialize Shape Models
Compute Gradient and Filtered Image
Optimize Shape/Registration parameters
Final Transformation
C

Fig. 1. Automated image registration. (**A**) Examples of input RNA *in situ* images. (**B**) Registration flowchart: during initialization, gradient and filtered images are calculated, and the parameters of the registration are randomly initialized. The parameters are then iteratively evaluated and optimized for a fixed number of iterations, using a numerical optimizer. The final parameters are used to transform the image and create the shape model as shown in (**C**). The optimized shape models for the embryos in the example images are shown in red, superimposed on the normalized input images.

previous high-throughput approaches which uses large datasets to quantitatively address expression in the context of spatial patterns: prediction of annotation terms, or clustering/classification of genes. We use a dataset of *Drosophila* embryonic expression patterns (Tomancak *et al.*, 2007) to illustrate how we address the three questions outlined above.

Our work distinguishes itself from previous work on this *Drosophila* data by three main contributions. First, robust and fully automated image analysis techniques are used to process and register the raw images. Through the use of statistical shape models and partial mapping methods, these techniques are capable of handling sources of phenotypic and imaging variability that limited previous approaches. Second, we comprehensively compare different similarity metrics, implement similarity measures that incorporate spatial dependencies to distinguish complex spatial patterns, and validate the different measures against visual annotation terms provided by experts. Third, we develop a new significance testing framework for spatial similarity scores through constrained realization Monte Carlo simulations and demonstrate how it can generalize similarity measures to achieve similar performance to spatial metrics. Finally, we illustrate this method of significance testing on known biological examples, emphasizing the importance for fully automated image registration/comparison models in the context of regulatory interactions. While we use a fly embryo dataset as an example, the general registration and comparison approach is adaptable and thus of interest to the study of spatial expression patterns in a wide range of model organisms.

2 APPROACH

2.1 Image registration

Prior to any quantitative analysis of expression image data, it is necessary to normalize and register the images to a common frame of reference. By first mapping different specimens such as fly embryos to a common reference, we can subsequently apply a large variety of methods for univariate and multivariate data analysis for cross-subject comparisons (between replicates, genes, time stages or different strains or species).

We have developed a fully automated registration method based on statistical shape models and improved numerical optimizers. This approach estimates both the average shape of an embryo, as well as the main components of variation of embryo shape, including orientation, from labeled data. It then uses this model to segment new images into foreground (a single complete embryo) and background (Fig. 1). We applied this registration method to the complete set of 78 621 images in the latest release of the Berkeley *Drosophila in situ* database. Overall, the model-based registration addresses problems that limited the successful application of previous methods on the whole database, highlighting its ability of extracting embryos under a large variety of imaging conditions: multiple embryos or impeding boundaries, changes in lighting/microscope settings, and out of focus boundary regions.

To formally analyze registration accuracy, we uniformly sampled 200 images from 200 uniformly sampled genes from the dataset for quantitative assessment. Each image was manually segmented, registered and compared with the automated segmentation and registration. We found that the most common inconsistencies in registration resulted from incorrectly orienting the axis, and were observed in 20 cases (10%) for the anterior/posterior axis, and in 6 cases (3%) for the dorsal/ventral axis. While our anterior/posterior axis alignment of 90% is consistent with an 85% accuracy reported earlier (Gargesha *et al.*, 2005), our approach differs from Gargesha's approach as we encapsulate the alignment problem into the registration process by allowing the axis-corrected shape model to more properly fit to the embryos in the image. The total registration accuracy was assessed using a test point error (TPE) measure (Zitova and Flusser, 2003) between the manually and automatically registered images. The average TPE was 0.94, with 190 images (95%) having a TPE within a reasonable accuracy rate of

>0.90, as compared with the leading reported registration accuracy (78%; Pan *et al.*, 2006).

After normalization and registration of the images, expression patterns or features useful for classification can be extracted. Here, we are interested in comparing the global 2D expression pattern, and a choice of representation of the patterns has to be made: instead of working with the complete pixel-based 2D patterns, it is common to map them to a smaller set of features representing expression in small subregions of the specimen. As the registration maintains morphological variabilities (size and stretch), it prevents a one-to-one mapping on the pixel level. To be able to evaluate the effect of different metrics in a simple scenario, we therefore chose to project the staining intensities along the anterior/posterior x-axis and dorsal/ventral y-axis as described in Section 3.3. Projections have been used used in a variety of recent applications in *Drosophila* (Janssens *et al.*, 2005; Segal *et al.*, 2008)—as well as Mouse brain images (Liu *et al.*, 2007)—and are particularly suited for the analysis of early embryonic expression as discussed below.

2.2 Correspondence of expression similarities to expert annotations

Using the extracted expression patterns, we assess the importance of using spatial metrics [Haar wavelets (HWs) and spatial mutual information (SMI)] by comparing their performance with two previously used non-spatial metrics [mean squared error (MSE) and mutual information (MI)]. We determined how the similarity values computed by each metric corresponded to manually annotated expression terms in the *in situ* database. For this purpose, we focused on the subset of 27 157 images covering 3127 genes acquired during the time window of developmental stages 4–6. Images in this stage window have been annotated with information on the view from which the images were taken (lateral or dorsal/ventral), and this crucial information is not yet provided for later stages. Furthermore, the images at this stage balance the frequency of spatially diverse expression patterns with the resolution and coverage in the database, and are not subject to the additional complexity that expression in later stages is conditional on earlier expression, which is also reflected in the annotation terms.

Genes in the selected window were annotated with 38 unique terms describing the spatial expression patterns. We removed all genes that had annotation terms which indicated the lack of spatial variability (e.g. ubiquitous, maternal) and non-lateral views, leaving us with 209 genes, 1231 images and 29 annotation terms. For each scoring metric, we calculated an enrichment significance for each annotation term describing how often genes annotated with a particular ontology term show the strongest similarity to genes annotated with the same term. Using a P-value cutoff of 0.05, SMI performed the best, with 22 of the 29 annotation terms being significantly enriched, while MI led to the second best result with 21. Both MSE and HWs led to 19 and 18 enriched terms, respectively (Table 1).

While SMI and HWs are capable of incorporating the spatial structure of expression pattern, they do not fully incorporate the significance of the spatial pattern: genes with complex spatial patterning (e.g. *even-skipped*) will be scored similarly to genes with simpler spatial patterning (e.g. *bicoid*). To better account for this, and to provide actual significance values between expression patterns, we have developed methods that account for the complexity of the spatial patterning. The significance values are computed by comparing the observed similarity value with a null distribution that preserves spatial dependencies between the two gene expression patterns, as described in Section 3.5.1 and shown in Figure 2. To accommodate for low signal/strong lighting effects, the significance values were converted into reweighted significance scores (RSS) as described in Section 3.6. The enrichment significance calculation was repeated for the RSS values and are shown in Table 1. By incorporating the spatial structure using RSS for each score, the scoring metrics performed similarly, with all the metrics resulting in 19–22 enriched significance scores.

While consistent differences between metrics are observed, and the significance estimates produce more stable results across metrics, the overall performance is not dramatically different. The advantage of using appropriate similarity metrics and significance estimates becomes more apparent in noisier scenarios, or when fewer features are used. Many, but not all, terms exhibit distinct patterns along the AP axis; however, when using projections onto the AP axis only, MSE results in only 15 terms compared with the 18 terms for both axis. In comparison, MI, SMI and the RSS scores remain largely consistent regardless of whether the DV axis is incorporated into the score. The full results analogous to Table 1 are given in the Supplementary Material.

2.3 Expression similarity of known co-regulated genes

As application of the registration and expression comparison pipeline, we use SMI and significance tests to validate known biological interactions, suggesting their usefulness for inference on biological data. Gene regulation and spatial patterning are a tightly coupled process: transcription factors acting as activators for a gene are often co-expressed in similar spatial regions, while repressors are often expressed inversely to the targeted gene. We address how such regulatory relationships are reflected in spatial expression profiles in the context of the segmentation network, using a set of the gap, pair rule and segmentation genes adapted from Schroeder *et al.* (2004). This network consists of a set of genes with many direct regulatory interactions and shared functional roles; in many cases, similarities in function/interaction are reflected in noticeable similarities of spatial expression patterns. Since the subset we are using does not contain the irregularities/noise of the full dataset, we here calculated the unweighted significance scores of the anterior/posterior projections described in Section 3.5.1.

The significance values for the similarity scores from all pairwise comparisons are shown in Figure 2. Many of the genes which share functional roles are identified as being significant; for instance, *pdm2* and *nubbin* (also known as *pdm1*) are paralogs with highly similar functional roles and interactions (Yeo *et al.*, 1995). Significant similarity scores can also reflect regulatory relationships between genes with overlapping expression domains in the transcriptional network: *ocelliless* (also known as *orthodenticle*) is positively regulated by *bicoid* (Finklstein and Perrimon, 1990); *Ubx* indirectly regulates *dichaete* through the intermediate activation of *dpp* (Capovilla *et al.*, 1994; Sanchez-Soriano and Russell, 2000). In addition to activators, we observe significant similarities for repressors, a property of MI which scores correlation and anti-correlation equally: *hunchback* represses the expression of *nubbin*, *pdm2* (Kambadur *et al.*, 1998)

Table 1. Enrichment of annotation terms for all genes within the second time window (stages 4–6) for all four similarity values: MSE, Haar, MI and SMI, given as actual similarity scores and RSS

Annotation term	Genes	Actual				RSS			
		MSE	Haar	MI	SMI	MSE	Haar	MI	SMI
Amnioserosa anlage in statu nascendi	19	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Anlage in statu nascendi	30	0.207	0.268	0.039	0.015	0.159	0.152	0.014	0.012
Anterior endoderm anlage in statu nascendi	30	0.015	0.033	0.001	0.004	0.006	0.026	0.001	0.001
Cellular blastoderm	27	0.126	0.142	0.312	0.368	0.265	0.193	0.463	0.355
Clypeolabrum anlage in statu nascendi	10	0.004	0.001	0.006	0.006	0.005	0.001	0.002	0.003
Dorsal ectoderm anlage	15	0.008	0.009	0.002	0.010	0.010	0.006	0.003	0.007
Dorsal ectoderm anlage in statu nascendi	72	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Ectoderm anlage in statu nascendi	13	0.017	0.030	0.017	0.029	0.017	0.035	0.015	0.010
Endoderm anlage in statu nascendi	9	0.216	0.166	0.014	0.023	0.147	0.093	0.011	0.015
Foregut anlage in statu nascendi	25	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001
Gap	27	0.101	0.053	0.020	0.027	0.046	0.049	0.036	0.030
Head epidermis anlage in statu nascendi	4	0.005	0.009	0.005	0.006	0.004	0.005	0.004	0.004
Head epidermis dorsal anlage in statu nascendi	10	0.007	0.003	0.001	0.001	0.018	0.004	0.002	0.002
Head mesoderm anlage	5	0.207	0.181	0.303	0.322	0.154	0.167	0.232	0.215
Head mesoderm anlage in statu nascendi	15	0.075	0.169	0.095	0.029	0.075	0.135	0.112	0.118
hindgut anlage in statu nascendi	20	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Mesectoderm anlage in statu nascendi	12	0.007	0.021	0.032	0.008	0.007	0.012	0.025	0.016
Mesoderm anlage in statu nascendi	18	0.180	0.166	0.161	0.097	0.219	0.101	0.196	0.189
Pair rule	5	0.141	0.071	0.097	0.067	0.137	0.044	0.046	0.047
Pole cell	13	0.971	0.968	0.963	0.984	0.952	0.969	0.968	0.976
Posterior endoderm anlage in statu nascendi	38	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Procephalic ectoderm anlage in statu nascendi	67	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Segmentally repeated	9	0.001	0.001	0.002	0.003	0.001	0.001	0.003	0.002
Trunk mesoderm anlage	5	0.030	0.236	0.006	0.002	0.012	0.160	0.010	0.013
Trunk mesoderm anlage in statu nascendi	17	0.017	0.045	0.162	0.086	0.008	0.066	0.269	0.208
Ventral ectoderm anlage	11	0.008	0.047	0.004	0.004	0.023	0.043	0.003	0.001
Ventral ectoderm anlage in statu nascendi	69	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Visual anlage in statu nascendi	17	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Yolk nuclei	35	0.419	0.485	0.823	0.818	0.639	0.592	0.841	0.805

Each row represents an annotation term, light gray shading represents 0.05 significance, while dark gray represents 0.01 significance values.

and *Ubx* (Pirrotta *et al.*, 1995); *giant* and *Krueppel* mutually repress each other (Kraut and Levine, 1991). Additional significant spatial localization can be a result of conditions in which genes act either in concert, or independently, to regulate other downstream genes. *Nubbin* shows significance with *dichaete*, and double knockout studies have shown that these two genes are essential for the proper formation of *even-skipped* stripes 1, 4, 5 and 6 (Ma *et al.*, 1998).

Not all known interactions are detected as significant, nor is this to be expected given the data. Often, a spatial expression pattern of a gene is the result of complex interactions between many genes across several time stages. For example, the proper development of all *even-skipped* stripes requires the interaction of many genes; the stripes are encoded by distinct *cis*-regulatory regions; and some factors function to control only a subset of the stripes. In such cases, assessing similarity on the level of global gene expression pattern, as pursued here as illustrative example, may therefore be modified to deal with parts of patterns, such as disjoint expression domains or even the boundary of expression domains. We anticipate that the integration of quantitative spatial expression information with other 'traditional' high-throughput data (e.g. binding, expression), will be useful to infer complex interactions in the regulatory networks of multicellular eukaryotic organisms (Fowlkes *et al.*, 2008).

3 METHODS

3.1 Berkeley *Drosophila in situ* database

The Berkeley *Drosophila in situ* database consists of 78 621 images of 3724 genes expressed in *Drosophila* embryos across six time windows (covering the developmental stages 1–3, 4–6, 7–8, 9–10, 11–12, 13–15). An established RNA *in situ* hybridization staining protocol was used to visualize spatial expression patterns as described in Tomancak *et al.* (2002). Annotations are based on an ontology describing embryonic expression patterns, consisting of 314 terms, and were obtained from the latest release of the database (Tomancak *et al.*, 2007). The annotation set was curated by the BDGP group by manually inspecting the *in situ* images and providing ontology terms for each gene at every time stage. Additionally, information on the orientation of the embryos (i.e. the viewpoint) was manually curated for the stage window 4–6.

3.2 Image registration

The approach for segmenting and registering images is based on statistical shape models using signed distance maps to describe object contours (Leventon *et al.*, 2000). Signed distance maps are a representation of contours which contain the distance from the contour of the object for every point in the image: negative distance values depict regions that are inside the object, positive distances are outside and the magnitude represents the

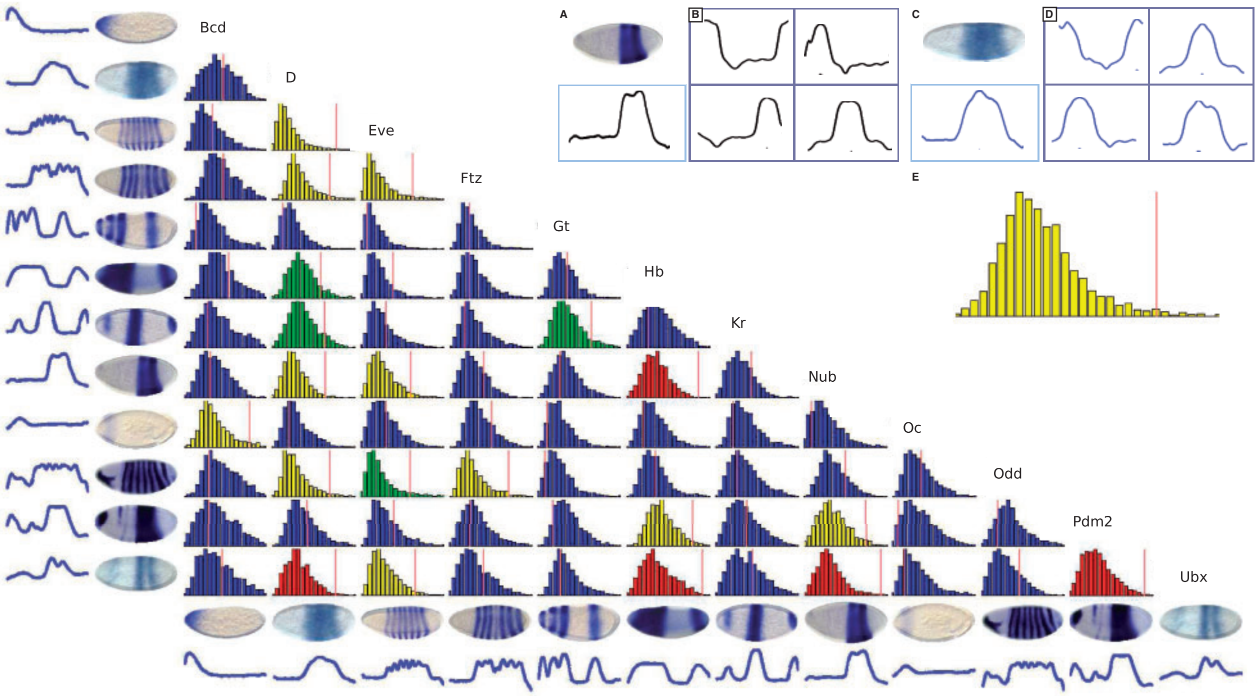

Fig. 2. Generation of background datasets and pairwise significance tests for 2D expression patterns. Upper right panel: computation of significance values, as exemplified on *in situ* images for the genes *nubbin* (**A**) and *dichaete* (**C**). The extracted expression vectors [bottom of (A, C)] of *nubbin* and *dichaete* are used to calculate the background distribution specific for the comparison of these two expression patterns. For each gene, a set of random realizations with constraints on the correlation between spatially adjacent expression values is created as described in Section 3.5 (**B**, **D**). The constrained realizations are used to compute a background distribution of similarity values [histogram in (**E**)]. The observed similarity on the expression patterns is indicated with an orange line, and is used to calculate an empirical P-value for each comparison. Lower left panel: a set of previously described gap, pair rule and segmentation genes (Schroeder *et al.*, 2004) was used to evaluate the significance testing. Images for each gene were registered, and their respective column expression vectors were calculated. Using the constrained realizations (histograms) and observed similarity values (orange lines), significance values were calculated for each similarity score. The color of the histogram represents the significance of the pairwise score (blue: >0.1, green: $(0.1, 0.05]$, yellow: $(0.05, 0.01]$, red: <0.01). The results in this example explore anterior/posterior patterning and are based on the x-axis projections of the expression patterns; for the comparisons in Table 1, projections to both x- and y-axes were used.

actual distance. Signed distance maps are an attractive choice for shape modeling as they provide a continuous representation of a discrete space which is easily interchangeable (the signed distance map can be directly converted from the contour, and the contour can be determined from the signed distance map by calculating the zero crossing of the distance map).

A *Drosophila* shape model was automatically created from a manually curated set of 120 embryo images (Fig. 3). First, the contours of the embryo were manually segmented and transformed into signed distances maps. The objects were then automatically normalized in size by minimizing the distance of each individual signed distance map to the mean signed distance map. The resulting normalized maps were analyzed using a hierarchical principal component analysis (PCA) decomposition (Westerhuis *et al.*, 1998). Let X be the set of all training images, and $X^b \subset X$ be the subset of training images that belong to time stage b, where $b \in B, B = [1,6]$, and $X^i \hat{X}^j = \emptyset \mid i,j \in B, i \neq j$. In standard PCA, a new set of bases $w_t \in W$ is selected such that $w_t = \arg \max \operatorname{var}\left\{w_t \hat{x}_{t-1}\right\}$ where $\hat{x}_{t-1}$ is the deflated matrix from the previous iteration. Each vector w_t is then converted into a 2D signed distance principal component image. Hierarchical PCA extends upon PCA by normalizing the contribution of each basis to each individual block.

In addition to providing characteristic priors on the shape through hierarchical PCA, we also model the filtered intensity values around the contour of the embryos. We create a set of histograms of the intensity values by binning observed intensity values by their respective signed distance. We concentrate on the intensity values close to the contour, i.e. we bin the intensities observed in distances from -25 to 25 in 1 U increments, while remaining bins are not included.

The task of image registration is to find the optimal set of parameters, θ, such that the images are accurately aligned to a common frame of reference. The parameters we optimize over can be separated into two categories: rigid transformation parameters of the image θ_r and the principle shape components of our shape prior θ_s. The principle component parameters provide a non-rigid component to the registration, allowing the underlying shape model to take on a variety of possible shapes which are defined by the signed distance maps. The rigid transformation parameters are used to rotate, translate, scale and flip (horizontally and vertically) each image so that it maps onto the evolving shape model. These two sets of parameters are simultaneously optimized using an in-house implementation of a particle swarm optimizer (Kennedy and Eberhart, 1995).

By using the shape parameters as well as the empirical histogram values, we assess how well the image is registered with the following metric:

$$f\left(\theta_r, \theta_s\right) = \alpha_1 g(s,g) + \alpha_2 h(s,i) \tag{1}$$

$s = \theta_s W$ is the signed distance map created from the linear combination of signed distance principal component images. $g(s,g) = \sum_{j=1}^{n} \left(\frac{1}{1+s_j^2} \left(s_j - g_j\right) \right)$ is an extension to Leventon's original scoring function, and $h(s,i) =$

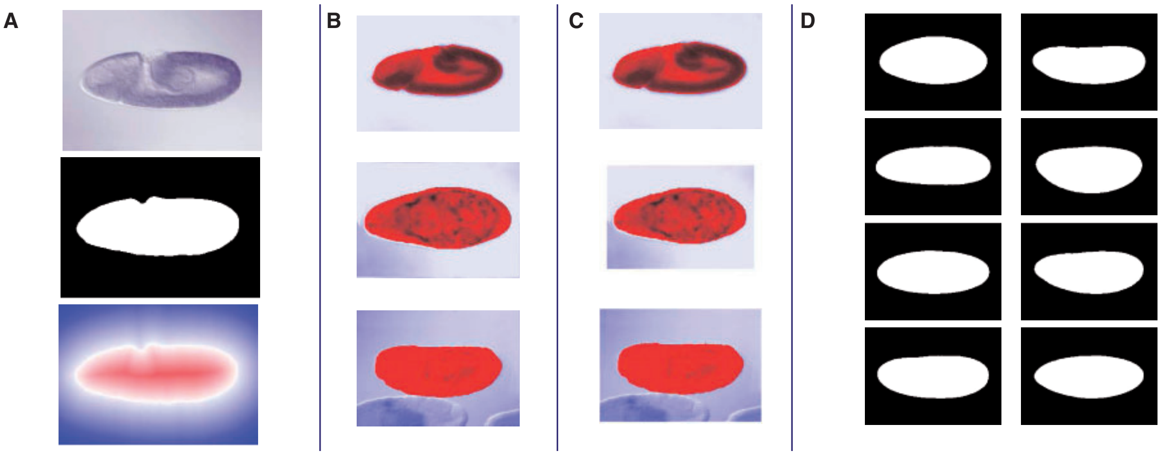

Fig. 3. Shape model. (**A**) Example image showing the creation of the training set. Prior image datasets are split and annotated in two components: the filtered pixel valued images, and the manually segmented contour of the object. The signed distance map is calculated directly from the external contour. Areas in red denote increasingly negative values (internal), while blue depicts increasingly positive values. (**B**) A subset of the 120 images used for the training of the shape model. The external contour is overlayed on the original image. (**C**) The training set is normalized in size. (**D**) The resulting contours are converted into signed distance maps and processed using a hierarchical PCA. Four of the principal shapes of the embryo are shown. These images depict 2 standard deviations of the principal component from the mean of the signed distance map.

$\sum_j p(s_j, i_j)$ is the empirical probability of observing a pixel value of i_j at the signed distance location s_j over all $s_j \in (-25, 25)$.

To summarize, our method differs from Leventon's method by three main distinctions. First, the effect of the intensity distribution is limited to the area around the zero crossing of the evolving shape. This is necessary to limit the contribution of internal staining and multiple/impeding embryos to the score. It also allows for a narrow band approach to be used with minimal effect on the overall score which increases the runtime performance. Second, the variability in size of the objects is normalized prior to creating the shape model. This is required to provide accurate representations of the actual differences in the shape of the *Drosophila* embryo which are not simply related to size. The size of the embryo then becomes an additional parameter in our optimization step. Lastly, we provide an additional term that is based on the direct values of the filtered image. This term is not decomposed into individual components, but rather describes the distribution of the filtered image in relation to the signed distance location. This allows the registration to accurately detect and align the shape model to the correct position of the image.

3.3 Representation of expression patterns

After registration, the transformed image T and shape model S are used to calculate column and row vectors of expression data. To allow a dyadic decomposition for the HWs, 64 columns and 32 rows element vectors are used. The row and column vectors are created by dividing the bounded shape image into equally spaced rows and columns, and computing the mean pixel intensity value of the second channel of the masked image for each column and row entry r_i and c_i. Where $c_i = 1/n_{\mathbf{c_i}} \sum_j^{n_{c_i}} t_{o_{\mathbf{c_i}}(j)}$, and $o_{\mathbf{c_i}}(j)$ is a mapping for column region $\mathbf{c_i}$ of all pixels within the shape model ($s_{\mathbf{c_i}} = 1$) and $n_{\mathbf{c_i}} = \sum s_{\mathbf{c_i}}$. The same representation is used for the rows, resulting in two vectors of expression denoted as $\mathbf{c}$ and $\mathbf{r}$.

3.4 Metrics and evaluation measures

Four metrics were evaluated to determine the similarity between expression patterns: MSE, MI, HWs and SMI. MSE and MI are two frequently used measures for comparing vectors of data. However, both assume that the samples within each vector are independent, which is not the case for spatial and time series data. Interaction terms between individual elements of a sample are capable of describing higher order structures, such as the formation of gradients, or the alternating striped pattern of *odd-* and *even skipped*. These dependencies are relevant not only in terms of kinetics and molecular diffusion, but also regarding interactions between genes. It is for this reason that similarity metrics that account for spatial dependency can be expected to provide more biologically relevant measures for comparing images. We implement two such measures, HWs and SMI, which are the spatial counterparts of MSE and MI, respectively. For all scoring metrics, we compute similarities independently on row and column vectors, and then combine them in a sum weighted by the vector size (i.e. in our case, with 1/3 and 2/3, respectively).

3.4.1 Mean squared error MSE scoring metrics have been previously used to compare gene expression patterns a and b (e.g. in Liu *et al.*, 2007, on RNA in situ brain images). For each row and column, we sum up the difference between the elements as defined:

$$d_{ab}^{\mathrm{MSE}} = \frac{1}{n} \sum_{j}^{n} \left(C_j^a - C_j^b \right)^2 \tag{2}$$

3.4.2 Mutual information We use the standard MI, e.g. as defined in Steuer *et al.* (2002):

$$d_{ab}^{\mathrm{MI}} = H(A) + H(B) - H(A,B), \tag{3}$$

where $H(A)$ and $H(B)$ define the entropy of each variable (or in this case, the column and row vectors for each gene) defined as:

$$H(A) = -\sum_{i=1}^{M_A} p\left(a_i\right) \log p\left(a_i\right) \tag{4}$$

And $H(A,B)$ is the joint entropy:

$$H(A,B) = -\sum_{i=1}^{M_A} \sum_{j=1}^{M_B} p\left(a_i, b_j\right) \log p\left(a_i, b_j\right) \tag{5}$$

3.4.3 Haar wavelets Wavelet analysis allows one to simultaneously examine the frequency and resolution components of a signal. This is accomplished by iteratively decomposing the signal with high and low bandpass filters. The low-frequency filter serves the purpose of downscaling the image into progressively smoother dyadic scales. Let $s \in (1,6)$ denote the dyadic scaling factor of the low-frequency filter. For the HWs, this low pass filter can then be written as: $L_i^s = 0.5 * \left(L_{2i-1}^{s+1} + L_{2i}^{s+1}\right)$, where i represents a specific spatial location. The high pass filter is responsible for creating the coefficients of the wavelet which will be ultimately used to represent the similarity between patterns. The high pass filter is written as: $H_i^s = 0.5 * \left(L_{2i-1}^{s+1} - L_{2i}^{s+1}\right)$. The Haar distance can then be calculated using the MSE: $d_{ab}^{\mathrm{HW}} = \sum_s \sum_{i \in s} \left(H_{a,i}^s - H_{b,i}^s\right)$.

3.4.4 Spatial mutual information SMI is an extension of MI to include neighboring dependencies and has been used in different image analysis applications (e.g. Rodriguez-Carranza and Loew, 1998). Instead of defining the entropy as the 1D probability of observing an event (or in this case, an expression value), we instead define it as the joint probability of observing a value i while its neighboring expression value is j. The entropy values for each variable (gene) then become a 2D entropy.

$$H(A,\hat{A}) = -\sum_{i=1}^{M_A} \sum_{j \in N_i}^{|N_i|} p\left(a_i, \hat{a}_j\right) \log p\left(a_i, \hat{a}_j\right)$$

where N_i are the neighbors of i. and $\hat{a}_j = |a_i - a_j|$ is the difference between neighboring elements.

We chose to use neighboring elements of distance 2. This value was selected as it preserves biological significant spatial patterns (local gradients; patterns such as pair rule or gap), while still being computationally tractable.

The cross-image comparison then becomes a 4D joint entropy:

$$H(A,\hat{A},B,\hat{B}) = -\sum_{i=1}^{M_A} \sum_{j \in N_i}^{|N_i|} \sum_{k=1}^{M_B} \sum_{l \in N_k}^{|N_k|}$$

$$p\left(a_i, \hat{a}_j, b_k, \hat{b}_l\right) \log p\left(a_i, \hat{a}_j, b_k, \hat{b}_l\right)$$

with the SMI being:

$$d_{ab}^{\mathrm{SMI}} = H(A,\hat{A}) + H(B,\hat{B}) - H(A,\hat{A},B,\hat{B})$$

3.4.5 Parzen window kernel density estimate To provide smoother calculations with (spatial) MI, we used a Parzen window kernel density estimate.

$$p\,(x) = \frac{1}{N} \frac{1}{h\sqrt{2\pi}} \sum_{i=1}^{N} \exp\left(-\frac{\left(x - x_i\right)^2}{2h^2}\right), \qquad (6)$$

where h is the bandwidth parameter and controls the smoothness of the estimate. This window allows us to provide a smoother and more robust MI calculation when dealing with sparse data.

3.4.6 Ontology annotation term assessment The fly embryo database frequently contains more than one image per gene. A pairwise blocked matrix was created to represent the similarities between genes where $d_{i,j}^{s,t}$ represents the distance from the s-th image of the i-th gene to the t-th image of the j-th gene. Let D_{ij} represent the distance between genes i and j where $D_{ij} = \min(d_{i,j}^{s,t})$ for the MSE and Haar metrics and $D_{i,j} = \max(d_{i,j}^{s,t})$ for MI and SMI. For each gene i and each annotation term k present for that gene $a_{i,k} = 1$, we perform a Mann–Whitney–Wilcoxon test. To accommodate for multiple disjoint annotations, we performed the rank test on all genes j with the same annotation $a_{j,k} = 1$, or genes with different annotations, but did not share any additional annotations of the original gene $a_{j,l} = 0$ if $a_{i,l} = 1$. Let U_i^k represent the Wilcoxon signed rank value for gene i and annotation term k. The significance for each annotation term, k, was calculated by taking the expectation of the U statistic $E(U^k) = \frac{1}{n} \sum_i^n U_i^k$, and calculating its resulting z-score $z^k = \left(E\left(U_i^k\right) - m_{U^k}\right) / \sigma_{U^k}$. The P-value for each annotation is calculated directly from the z-scores.

3.5 Biological significance testing

The significance of a scoring metric can be computed based on a series of surrogate datasets used for hypothesis testing. We create appropriate surrogate data by drawing constrained realizations (Theiler and Prichard, 1996) which account for both the marginal distributions of the intensity values as well as the joint spatial dependencies between neighboring variables. This strategy is based on surrogate data whose spatial complexity is representative of the underlying spatial processes; genes with high spatial dependencies (e.g. smooth gradients) will result in a surrogate dataset with similar gradient patterns, whereas those having low spatial dependencies (e.g. hard gradients) will have dissimilar patterns. This approach requires the use of a sampler capable of drawing values from both the marginal as well as the spatial dependencies.

To meet these requirements, we sampled as follows: for each column vector, we initialize every element by drawing a sample from the marginal distribution. This initialization provides us with column vectors that have no spatial dependency. To account for the spatial dependencies, an iterative swap sampler is used on this random initialization. Let $p^1\left(a_i, a_j\right)$, $p^2\left(a_i, a_k\right)$, $p^3\left(a_i, a_l\right)$ be the probability of observing an intensity a_i while its first, second and third neighbor elements are a_j, a_k and a_l, respectively, where $j \in N_i^1, k \in N_i^2, l \in N_i^3$ are neighborhood association sets. To account for edge effects, neighbor relationships are considered on a torus. Let $\mathbf{p}\left(a_i, .\right)$ be the cumulative probability of observing a_i. For each iteration, four random locations are chosen: $K = k_c, k_d, k_e, k_f$ with observed values $L = o(k_c), o(k_d), o(k_e), o(k_f)$. Let $\mathbf{h}: K \leftarrow K$ be a permutation of the locations with $\mathbf{h} = h_1, \ldots, h_n$ being the set of all permutations of K with $h_i(k_c, k_d, k_e, k_f) = H^i = \hat{k}_c, \hat{k}_d, \hat{k}_e, \hat{k}_f$ being the permuted locations with observed values $M_i = o(\hat{k}_c), o(\hat{k}_d), o(\hat{k}_e), o(\hat{k}_f)$. The probability of a swap is calculated for all permutations (including the identity permutation) and the most probable swap is chosen. By iteratively resampling from this distribution, surrogate datasets are created that account for both the marginal as well as the spatial dependency structure.

3.5.1 Significance testing For each gene, 40 constrained realizations are created. The background similarity value is constructed by calculating all 1600 pairwise similarity values between the constrained realizations for each gene denoted by $b_{i,j}^s$ for $s \in (1, 1600)$. The empirical P-value of the expression between genes i and j, $p_{i,j}$, is calculated by counting the number of background scores greater than the observed score $|b_{i,j}^s > d_{i,j}|$, and dividing it by the number of elements $p_{i,j} = |b_{i,j}^s > d_{i,j}^{\mathrm{SMI}}| / 1600$.

3.6 Reweighted significance scores

To account for low-staining and highlighting variability commonly observed in the data, we compute a RSS between images i,j: $\mathrm{RSS}_{i,j} = p_{i,j} + (1 - p_{i,j})/X$, where $X = \exp(\sqrt{\mathrm{var}(i) * \mathrm{var}(j)})$. The RSS score is a smooth alternative to thresholding: as the variability in the images increases (stronger signal), the RSS score approaches its actual P-value. Images with low variability (weak signal/stain), have their scores reweighted closer to 1.

4 DISCUSSION

To address questions that arise with the analysis of spatial gene expression patterns, we have implemented a complete pipeline for 2D *Drosophila* RNA *in situ* staining images that is fully automated and highly robust to variable conditions in the input data. In addition to accurate methods for the registration of images to extract expression patterns, we implemented similarity metrics and significance scores that are appropriate for spatial patterns. Such methods are critical for a proper interpretation of spatial expression similarity with dependencies among neighboring areas. We have also demonstrated how our significance tests can be used to generalize metrics that do not include spatial structure directly, we implied their general use for biological inference by validating biologically known relationships. We focused on genes from the developmental stages 4–6, as the database contained viewpoint information for this stage window only. The comparatively simple expression patterns observed at this stage furthermore allowed us to represent them as projections to the *x*- and *y*-axes. While this approach worked well in this scenario, a 2D-grid representation is likely to be more appropriate to analyze more complex patterns later in development, once viewpoint information is available. To this end, the similarity scores and significance estimates can be extended to work with a larger neighborhood (e.g. four neighbors rather than two), albeit at larger computational cost.

Imaging *in situ* hybridizations under bright field microscopy often introduces undesired variability and artifacts. This creates problems for quantitative comparisons: depending on probe affinity, staining protocol, the overall position, lighting conditions and focal plane of the image, genes with near identical spatial expression patterns may exhibit low similarity values. Many of these quantification issues can be addressed by prospective experimental planning, including standardization of microscopy settings, number of replicates, time stages/conditions and staining protocols. Unfortunately, this type of planning is not available in a retrospective study as this; as a result, quantification will be affected by these issues until this information is provided and methods are developed to adequately model these sources of noise. Another critical limitation is the frequent lack of biological replicates, which reduces the ability to filter input noise and model the actual variance of expression patterns.

The goal of our current work was to extract information from large-scale image data, which can then be used as, or combined with, other data on gene expression. For instance, it provides metrics to cluster expression profiles [a task which is currently based on qualitative descriptors obtained by manual annotation (Tomancak *et al.*, 2007)], or to complement other high-throughput data such as obtained from ChIP-chip or microarray experiments (Costa *et al.*, 2007). Our work is certainly not the first computational study based on *Drosophila* digital microscopy images. Some other fly datasets (Fowlkes *et al.*, 2008; Janssens *et al.*, 2005; Keranen *et al.*, 2006) are obtained with the purpose to study the expression of few genes but at much higher quality in terms of resolution and quantification of expression. Each registration method on these complex images is in general tailored to a specific scenario, often with well-controlled imaging protocols/environments (e.g. single embryos, consistent staining/lighting) and small datasets ($<$500 images).

Some recent studies have also addressed *Drosophila* 2D RNA *in situ* images, on the same or similar data that we used here (Kumar *et al.*, 2002; Peng *et al.*, 2007). Our goal was to compare and identify pairwise spatial expression patterns, with an example in gene regulatory interactions, and is thus most similar to the work of (Kumar *et al.*, 2002), a system for database retrieval of RNA *in situ* images based on the global expression pattern. Other studies focused on classifying or clustering patterns, e.g. with the purpose of classifying images into Gene Ontology annotation terms (Ye *et al.*, 2008; Zhou and Peng, 2007). These different goals show the breadth of possible applications on this new type of high-throughput data; however, this also means that direct comparative assessments between them are difficult, even for shared subtasks such as image registration. Our shape-based registration method is fully automated and applied on all 78 621 images in the current release of the expression pattern database. In contrast, other approaches typically selected small application-oriented subsets, or used a registration process that was manually curated, and these sets do not show enough overlap to create a common gold standard set at this point. We did however compare the performance of several distance metrics used within other systems [MSE (Liu *et al.*, 2007), MI (Peng and Myers, 2004) and wavelets (Peng *et al.*, 2007)], with the result that the proposed SMI is more adequate, and that many of the previous metrics can achieve similar performance by incorporating the spatial structure in a significance score.

While our method for the automated registration of images is more robust and complete than many previous approaches, it cannot be an off-the-shelf solution for all image registration problems. Differences in sample preparation and imaging technology, as well as differences in the morphology of specimens which are imaged, pose restrictions on the possible choices of analysis methods. For instance, our previous work on *Arabidopsis* root images (Mace *et al.*, 2006) deals with confocal images of GFP reporter constructs. In addition to the different morphology, the fly dataset poses additional challenges of dealing with multiple objects, and a high variability in imaging conditions; conversely, we needed to apply non-rigid registration for the plant roots. Further, while our method for significance testing and surrogate dataset generation are well suited to the syncytium formation of the *Drosophila* embryo, this is generally not applicable across all organisms. For example, in *C.elegans*, gene expression is heavily influenced by cell lineage, and a more appropriate representation for neighboring dependencies would be between neighboring lineage elements, and not direct spatial location. Additionally, different sampling methods may need to be applied to deal with differences in the structure of the organism. While general platforms for cell-based image analysis have been published (Carpenter *et al.*, 2006), a universal framework for complete multicellular organs or organisms has to be placed at a higher level and poses a challenge for the automated analysis of image expression datasets pertaining to developmental biology.

5 CONCLUSIONS

As images are increasingly becoming a source of high-throughput genomics data, theoretically sound approaches to evaluate them become necessary. Whatever the specific context, appropriate quantitative methods along the lines proposed in this study will allow us to move image expression data from qualitative descriptions to the quantification of gene expression, and to its use as phenotype for which we can assess the significance of changes under perturbations of the genotype or the environment.

Funding: National Science Foundation (grant # MCB 0618304); the Human Frontier Science Program.

Conflict of Interest: none declared.

REFERENCES

Bao,Z. *et al.* (2006) Automated cell lineage tracing in Caenorhabditis elegans. *Proc. Natl Acad. Sci. USA*, **103**, 2707–2712.

Capovilla,M. *et al.* (1994) Direct regulation of decapentaplegic by Ultrabithorax and its role in Drosophila midgut morphogenesis. *Cell*, **76**, 461–475.

Carpenter,A.E. *et al.* (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, **7**, R100.

Costa,I. *et al.* (2007) Semi-supervised learning for the identification of syn-expressed genes from fused microarray and in situ image data. *BMC Bioinformatics*, **8** (Suppl. 10), S3.

Damle,S. *et al.* (2006) Confocal quantification of cis-regulatory reporter gene expression in living sea urchin. *Dev. Biol.*, **299**, 543–550.

Finklstein,R. and Perrimon,N. (1990) The orthodenticle gene is regulated by bicoid and torso and specifies Drosophila head development. *Nature*, **346**, 485–488.

Fowlkes,C.C. *et al*. (2008) A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell*, **133**, 364–374.

Gargesha,M. *et al.* (2005) Automatic annotation techniques for gene expression images of the fruit fly embryo. In *Visual Communications and Image Processing*, Society of Photo-Optical Instrumentation Engineers, Bellingham, WA, pp. 576–583.

Harmon,C. *et al.* (2007) Comparative analysis of spatial patterns of gene expression in Drosophila melanogaster imaginal discs. *Res. Comput. Mol. Biol.*, **4453**, 533–547.

Hunt-Newbury,R. *et al.* (2007) High-throughput in vivo analysis of gene expression in Caenorhabditis elegans. *PLoS Biol.*, **5**, e237.

Janssens,H. *et al.* (2005) A high-throughput method for quantifying gene expression data from early Drosophila embryos. *Dev. Genes Evol.*, **215**, 374–381.

Kambadur,R. *et al.* (1998) Regulation of POU genes by castor and hunchback establishes layered compartments in the Drosophila CNS. *Genes Dev.*, **12**, 246–260.

Keller,P.J. *et al.* (2008) Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *Science*, **322**, 1065–1069.

Kennedy,J. and Eberhart,R. (1995) Particle swarm optimization. *IEEE Intl Conf. Neural Netw.*, **4**, 1942–1948.

Keranen,S.V.E. *et al.* (2006) 3D morphology and gene expression in the Drosophila blastoderm at cellular resolution ii: dynamics. *Genome Biol.*, **7**, R124.

Kraut,R. and Levine,M. (1991) Mutually repressive interactions between the gap genes giant and Kruppel define middle body regions of the Drosophila embryo. *Development*, **111**, 611–621.

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.

Lee,J.-Y.Y. *et al.* (2006) Transcriptional and posttranscriptional regulation of transcription factor expression in Arabidopsis roots. *Proc. Natl Acad. Sci. USA*, **103**, 6055–6060.

Leventon,M.E. *et al.* (2000) Statistical shape influence in geodesic active contours. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.*, **1**, 316–323.

Liu,Z. *et al.* (2007) Study of gene function based on spatial co-expression in a high-resolution mouse brain atlas. *BMC Syst. Biol.*, **1**, 1–19.

Mace,D.L. *et al.* (2006) Quantification of transcription factor expression from Arabidopsis images. *Bioinformatics*, **22**, e323–e331.

Ma,Y. *et al.* (1998) Gene regulatory functions of Drosophila Fish-hook, a high mobility group domain Sox protein. *Mech. Dev.*, **73**, 169–182.

Megason,S.G. and Fraser,S.E. (2007) Imaging in systems biology. *Cell*, **130**, 784–795.

Murray,J.I. *et al.* (2008) Automated analysis of embryonic gene expression with cellular resolution in C. elegans. *Nat. Methods*, **5**, 703–709.

Pan,J.-Y. *et al.* (2006) *Automatic Mining of Fruit Fly Embryo Images*. ACM, New York, NY, pp. 693–698.

Peng,H. and Myers,E.W. (2004) Comparing in situ mRNA expression patterns of Drosophila embryos. In *Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*. Association for Computing Machinery, San Diego, CA, pp. 157–166.

Peng,H. *et al.* (2007) Automatic image analysis for gene expression patterns of fly embryos. *BMC Cell Biol.*, **8**, 693–698.

Pirrotta,V. *et al.* (1995) Distinct parasegmental and imaginal enhancers and the establishment of the expression pattern of the Ubx gene. *Genetics*, **141**, 1439–1450.

Rodriguez-Carranza,C.E. and Loew,M.H. (1998) Weighted and deterministic entropy measure for image registration using mutual information. In *Medical Imaging 1998: Image Processing*, Vol. 3338, 155–166.

Sanchez-Soriano,N. and Russell,S. (2000) Regulatory mutations of the Drosophila Sox gene Dichaete reveal new functions in embryonic brain and hindgut development. *Dev. Biol.*, **220**, 307–321.

Schroeder,M.D. *et al.* (2004) Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol.*, **2**, e271.

Segal,E. *et al.* (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, **451**, 535–540.

Steuer,R. *et al.* (2002) The mutual information detecting and evaluating dependencies between variables. *Bioinformatics*, **18**, 231–240.

Theiler,J. and Prichard,D. (1996) Constrained-realization Monte-Carlo method for hypothesis testing. *Physica D*, **94**, 221–235.

Tomancak,P. *et al.* (2002) Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.*, **3**, research0088.1–research0088.14.

Tomancak,P. *et al.* (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.*, **8**, R145.

Westerhuis,J.A. *et al.* (1998) Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.*, **12**, 301–321.

Yeo,S. *et al.* (1995) On the functional overlap between two Drosophila POU homeo domain genes and the cell fate specification of a CNS neural precursor. *Genes Dev.*, **9**, 1223–1236.

Ye,J. *et al.* (2008) Developmental stage annotation of Drosophila gene expression pattern images via an entire solution path for LDA. *ACM Trans. Knowl. Dis. Data*, **2**, 1–21.

Zhou,J. and Peng,H. (2007) Automatic recognition and annotation of gene expression patterns of fly embryos. *Bioinformatics*, **23**, 589–596.

Zitova,B. and Flusser,J. (2003) Image registration methods: a survey. *Image Vis. Comput.*, **21**, 977–1000.