# Systematic image driven analysis of the spatial Drosophila embryonic expression landscape

Erwin Frise, Ann S. Hammonds, and Susan E. Celniker[1]

[1]Berkeley Drosophila Genome Project, Department of Genome Dynamics, Lawrence Berkeley National Laboratory, MS 64R0121, 1 Cyclotron Road, Berkeley, CA, 94720 United States

Correspondence: Erwin Frise, E-mail erwin@fruitfly.org

## Supplementary Information

**Contents:**

## Supplementary Methods

### Separating touching embryos

We created a heuristic algorithm to detect the presence of more than one embryo in an image enclosed by the segmentation boundary, and to determine the positions in the boundary where the embryos join each other. The embryo of interest was always located in the center of the image. Thus, we choose the point in the center of the image, laid lines in one degree intervals l=1, ..., 360 from this center point and intersected the lines with the boundary. We then determined the number of boundary instances b intersecting each line. As illustrated in Supplementary Figure 2A, there is usually only one boundary intersecting the line from the center (line a in the figure) unless the boundary encloses a second (or third) embryo (line c). However, small dents on the boundary or errors in the initial segmentation produce more than one line (see line b in figure for example) and mimic a touching embryo. To detect such cases we also calculated the standard deviation of boundary distances from the image center intersecting each of the 360 lines. Neither the number of boundary instances nor the normalized standard deviation alone was sufficient, but the combination of both permitted us to determine both the presence and location of a touching embryo (Supplementary Figure 2B). We evaluated each line l for the values of b and s (Supplementary Figure 2B).

### Computing staining intensity and removal of Differential interference contrast (DIC) microscopy shadows

Previous efforts with this image dataset converted the images to grayscale and used the resulting gray values between 0 and 255 as staining intensities. Shadows are a known artifact of capturing images using Normarski or Differential interference contrast (DIC) microscopy. These shadows are indistinguishable from weak staining in these grayscale images (Supplementary Figure 2J, note the essentially unstained embryo). Morphological landmarks such as the cephalic furrow and the amnioproctodeal invagination also produce shadows that confound the image processing (Supplementary Figure 2C and 2I, note the morphological structures clearly visible in the grayscale image). Examination of the RGB color channels of the images revealed that the red (R) and blue (B) channels were highly correlated for the entire image including the DIC shadows but the red channel was stronger at pixels with staining. Thus, we determined the staining intensity

$$I = \frac{256R}{B+1}$$

The resulting image I virtually eliminated DIC shadows and resulted in a high correlation of the grayscale staining intensities to the color image. However, areas with intense staining resulted in low scale gray values (Supplementary Figure 2D). This was illustrated by plotting the staining intensities R, B and I at vertical line 600 where I increased between the horizontal coordinate positions between 950 and 1200 (Supplementary Figure 2F). We calculated the relationship for all possible R and B values (Supplementary Figure 2G), computed the gradient $\nabla I$, and at R and B combinations where the gradient $\nabla I$ exceeded 0.005 we linearly lowered the R and B values to the first point where $\nabla I$ was below that threshold. The result of this dynamic adaptation was an accurate representation of the staining intensities (Figure 1E and Figure 1K). In the resulting image, 0 was defined as the darkest staining intensity, 255 as no staining.

**Reducing the complexity of the expression landscape**

To filter the datasets we created an automated pipeline, *TIfilter* (written for Matlab 7.2 or higher), with following processing steps (Supplementary Figure 1, Table I, Supplementary Table I). All staining intensities were based on the numerical values of the TIs produced by TIgen.

First, we eliminated the homogeneously stained or unstained embryos by removing all TIs with nearly identical staining intensities across all triangles. We calculated the vector norm for all 311 triangles and eliminated all TIs with a value below 50.

Second for each gene, we eliminated multiple redundant expression patterns by combining redundant TIs to a single TI and ordered the TIs in a developmental time-line (Supplementary Figure 7). To eliminate duplicates we selected all TIs for each gene and calculated their pair-wise correlation distance $d_{i,j}$ to factor for any experimental staining intensity differences. If the distance was below an empirically determined value of 0.4, TIs were assumed to be identical. Otherwise, we clustered them using Affinity Propagation Clustering (APC) with the distance input $s(i,j) = -d_{i,j}$ (Supplementary Figure 7B). We then created one representative TI for each cluster using the darkest staining intensity for each of the 311 triangles for all TIs in the cluster. The darkest staining intensity was the lowest numerical value for each identical triangle across all TIs. This procedure had the additional benefit of collapsing all focal planes into one. To order them relative to each other, we used graph theory. We computed the pairwise Euclidean distance of each representative image to each other. Staining intensities were already largely equalized in the representative TIs so that euclidean distance performed better than correlation distance. Using the distances as weights and the filtered TIs as nodes in a graph we determined the minimal spanning tree of this graph (Prim, 1957). Observations from image capture, microarray data and previous analysis suggested that the area of the embryo with gene expression and/or the level of gene expression increases as development progresses. We used this fact to identify the first and final TI in the graph. We calculated the total staining $S$ for each TI with the triangle staining intensities $d_i$:

$$S = \sum_{i=1}^{311} |d_i - 255|$$

We selected the TI with *min(S)* as starting TI and the TI with *max(S)* as the end TI. We then sorted the intermediate TI with a shortest path algorithm (Dijkstra, 1959) (Supplementary Figure 7C). We verified results from randomly selected sorted TIs by comparing the latest TI at the selected stage range to TIs in the next stage range and by evaluating morphological features of the embryos. Examples for sorted patterns are shown in Supplementary Figure 7D-F.

Finally, we discriminated against uncertain or faint expression patterns by selecting only TIs with a pronounced difference between the expressed regions and regions with no expression. To select the most distinct patterns we clustered staining intensities of each TI into three clusters using the consensus of three randomly initialized k-means clusterings. We then compared the median intensities of the triangles in the cluster with the lowest value with the triangles in the cluster with the middle value. If the difference between the intensities exceeded 30, we selected the TI as a distinct pattern.

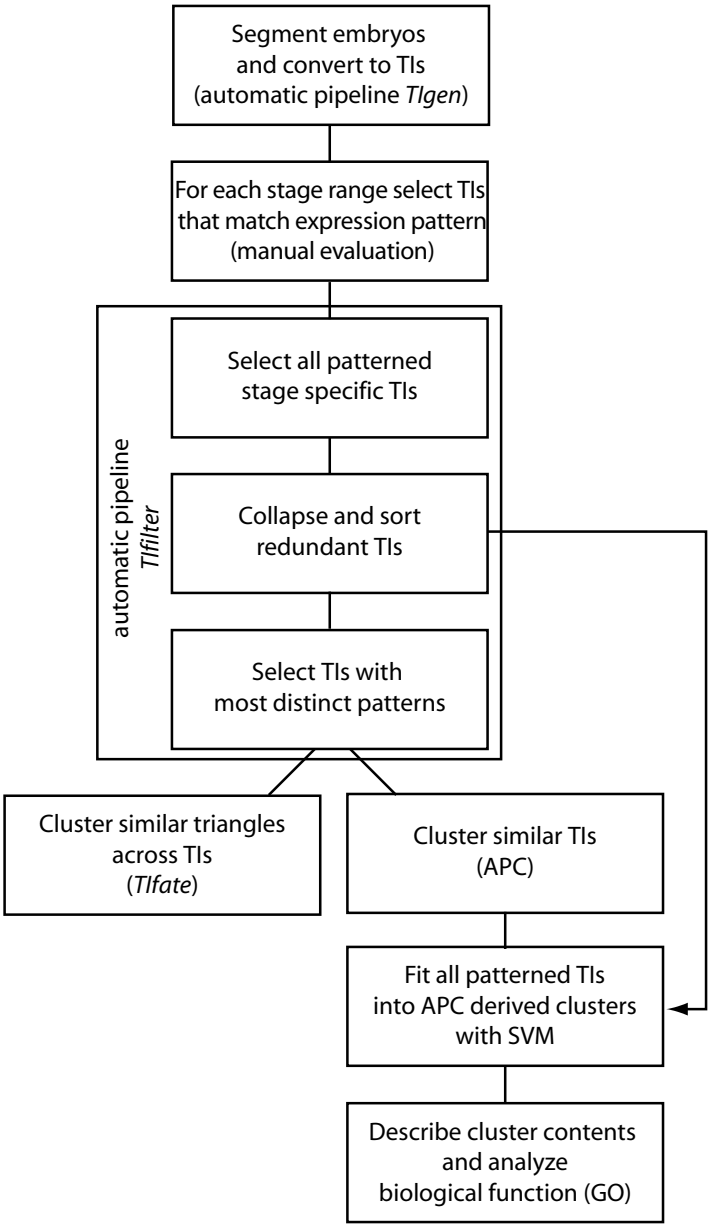
**Support vector machine classification of the TIs**

To fit all patterned TIs into the clusters generated by APC, we used the clustered TIs as training set for a support vector machine (SVM). We normalized the dataset so that each triangle across all embryos was independently normalized between 0 and 1. We then used a SVM from the Matlab libsvm package (version 2.86, (Fan *et al*, 2005)) for a binary classification of the TI into one or more clusters. As training sets we used the TIs of each cluster as a positive set and random samples in the size of four times the number of the positive training set from other clusters as a negative set and used five-fold cross validation (four parts for training and one parts for cross-validation). For each of the clusters we optimized the SVM parameters for a radial basis function (RBF) kernel with a systematic grid search as described (Chang and Lin, 2001), trained the SVM with the contents of each cluster and performed a binary classification of the entire dataset with the trained SVM for each cluster (Supplementary Figure 8).

**Associations between biological functions**

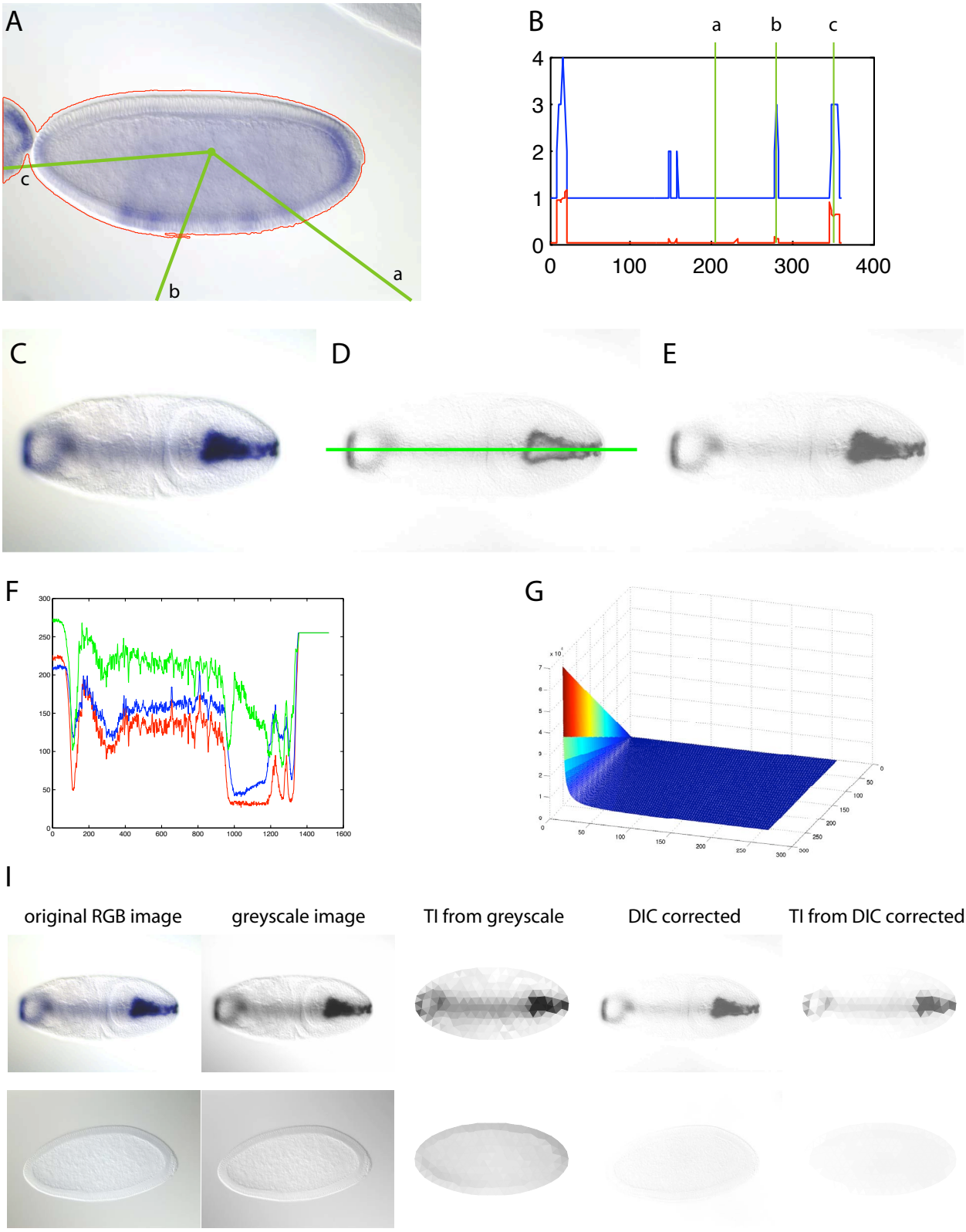
To find the most significant associations between two or more biological functions we filtered for GO-terms that co-occurred in multiple clusters (Figure 7B). We selected for all terms that were enriched in at least three clusters resulting in 869 unique term pairs composed of 199 unique GO-terms (Dataset 9). For visualization, we clustered closely related terms together as a node and connected the nodes resulting in the graph shown in Supplementary Figure 16. In detail, we computed the Jaccard distance matrix among all selected GO-terms based on their co-occurrence in multiple clusters and condensed the dataset further by applying APC to the negative value of the Jaccard distance. The cluster centers given by the APC procedure were selected as terms for the nodes. We combined all terms for each cluster into a binary vector (1=term present in the cluster, 0=term absent) and computed the Jaccard distance between the clusters. If the distance exceeded a threshold of 0.3, we linked the terms together and imported linked terms into Cytoscape to created the graph (Shannon *et al*, 2003). We used the options in Cytoscape to graph the size of the nodes based on the number of connections.

## Supplementary Figures



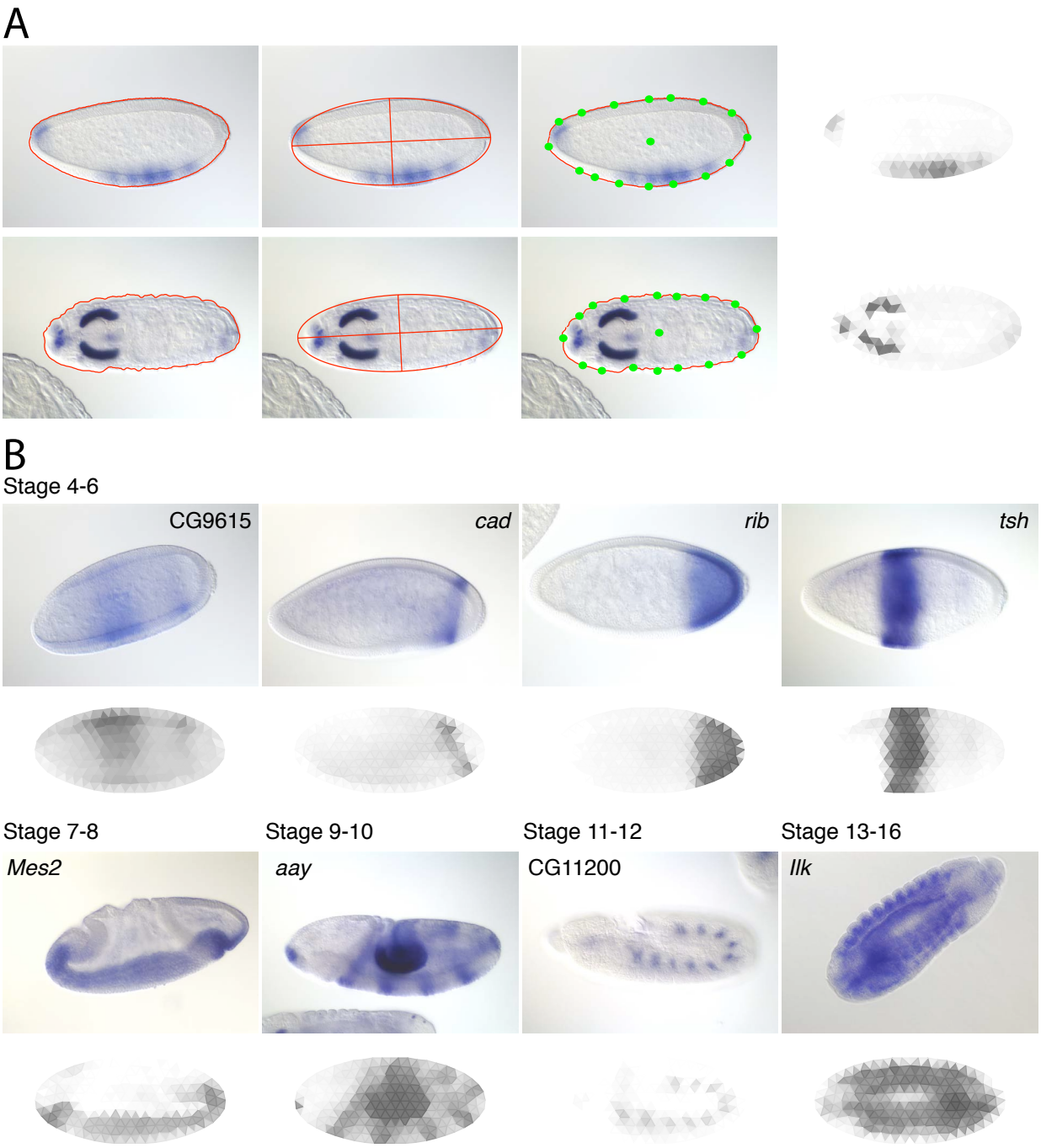

**Supplementary Figure 1** Flowchart for the analysis of the expression landscape. Shown is the process flow and the contribution of the programs *TIgen, TIfilter* and *TIfate*. Filtered distinct patterns were subjected to clustering with *TIfate* and APC. The GO analysis of the clusters included all patterned TIs. APC = affinity propagation clustering.

**Supplementary Figure 2**



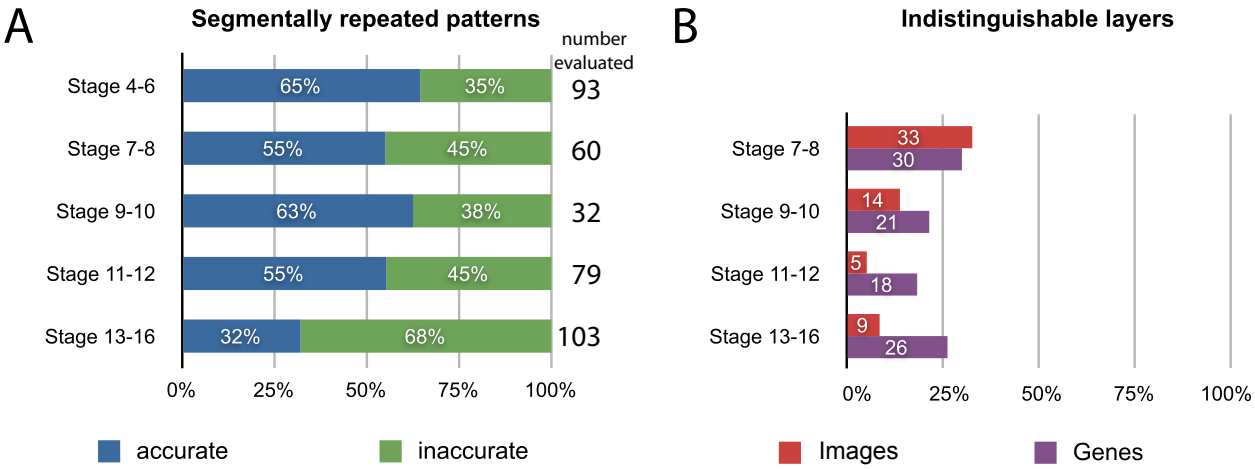
**Supplementary Figure 2** Separating touching embryos and converting color images to grayscale intensities (**A**) Initial segmentation result for the embryo from Figure 1A with another embryo touching the main embryo both enclosed by the boundary (red). Lines (a), (b) and (c) originating from the center of the image illustrate the different scenarios for detecting touching embryos. Line a crosses the boundary once. Line b crosses the boundary three times due to a minor segmentation problem. Line c crosses the boundary three times as a result of two touching embryos in the image. (**B**) Graph for all possible lines in (A), plotted from 0 to 360 degrees in one-degree intervals (x-axis). The number times the lines intersect the boundary is shown in blue, the normalized standard deviation in red (y axis). Line a crosses the outline once and line b and c cross the outline three times. Line b and c distinguish themselves by a low standard deviation in b and a high standard deviation in c. (**C**) Color image (insitu35798.jpe) at stage 7-8 with heavy DIC shadows. (**D**) Conversion to grayscale without dark staining corrections. The green line corresponds to the plots in F. (**E**) Conversion to grayscale after dark staining corrections were applied. (**F**) Plot of all pixels of line 600 from D of the red channel (red), the blue channel (blue) and the computed grayscale intensity (green). Note the increase of the grayscale intensity values in areas of low red/blue values. (**G**) Plot of all possible red/blue combinations and the computed grayscale values. (**I**) Effect of the correction of the DIC shadows on the color image of (C) and an unstained embryo at stage 4-6 (insitu13172.jpg). The images converted to grayscale and resulting TIs (2nd and 3rd panels from the left) show the DIC shadows and, in case of the unstained embryo, mimic weak expression at the posterior. DIC corrected images and TIs (two right panels) do not show any shadows.

**Supplementary Figure 3**



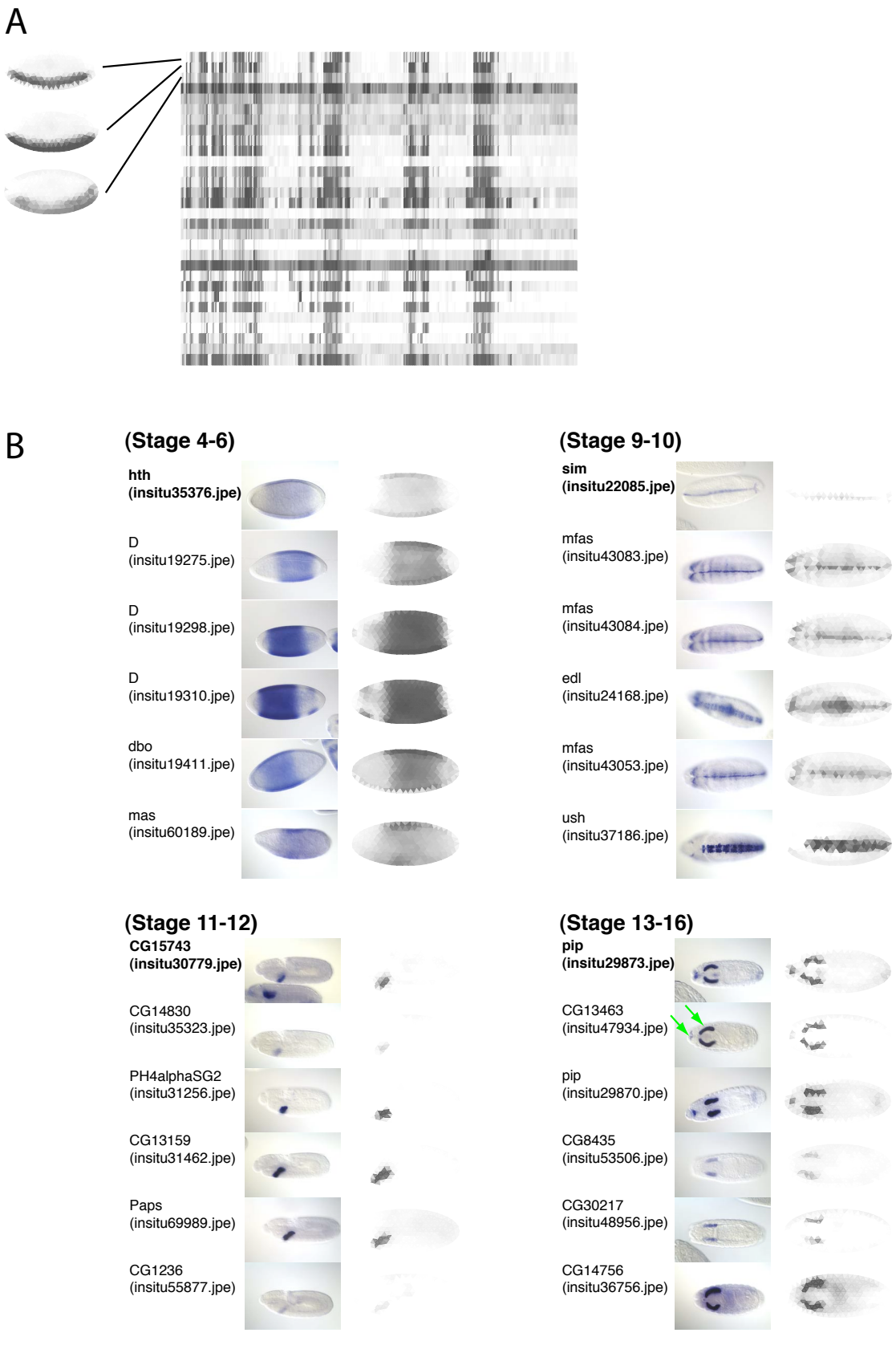

Examples of embryo segmentation and TIs. (**A**) Segmentation for embryos in images insitu19144.jpe (top, same as Figure 1K) for *tin* and insitu29873.jpe (bottom) for *pipe* (*pip*). From left to right, shown are the segmentation results (corresponding to Figure 1B), the fitted ellipse (as in Figure 1C), the selected 16 points on the circumference (as in Figure 1D) and the resulting TIs. (**B**) Examples of staining patterns converted to TI at all stage ranges. Note that *CG9615*, *cad*, *Mes2* and *Ilk* TIs were reoriented so that the final TIs have the conventional anterior at the left and ventral at the bottom orientation. The TI for *Ilk* at stage 13-16 does not resolve all structural details but still provides a reasonable approximation of the expression.

**Supplementary Figure 4**



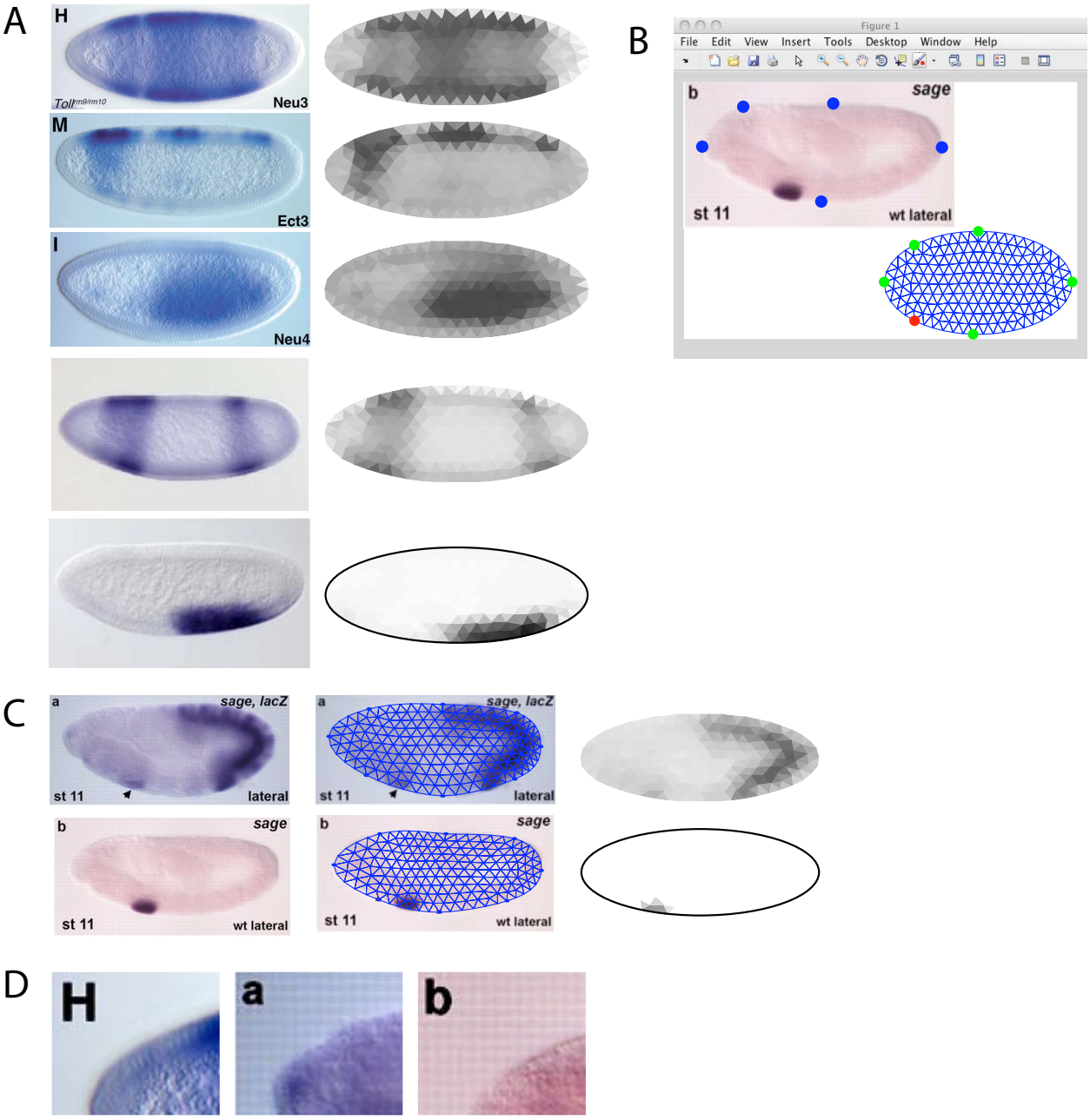

Quality assessment for segmentally repeated patterns and images with indistinguishable layers. (**A**) Percent of segmentally repeated patterns that were accurately represented in the TIs at different stages. Shown to the right is the number of TIs with segmentally repeated patterns that were scored for each stage range. Of the segmentally repeated patterns, expressed in stages 4-12, more than 50% of the TIs match the digital images. (**B**) Percent of images and genes with images that cannot be resolved to a particular tissue. Stage 7-8 contains the largest fraction of images where it is not possible for a curator to assign the expression to a particular tissue layer. At later stage ranges (11-12 and 13-16) with terminally differentiated tissues, expression patterns are easier to annotate and the number of images that are accurately represented is high. However, a single inaccurate image causes the gene to be labeled as inaccurate and complex expression patterns affecting multiple tissues are represented by high numbers of images.

**Supplementary Figure 5**



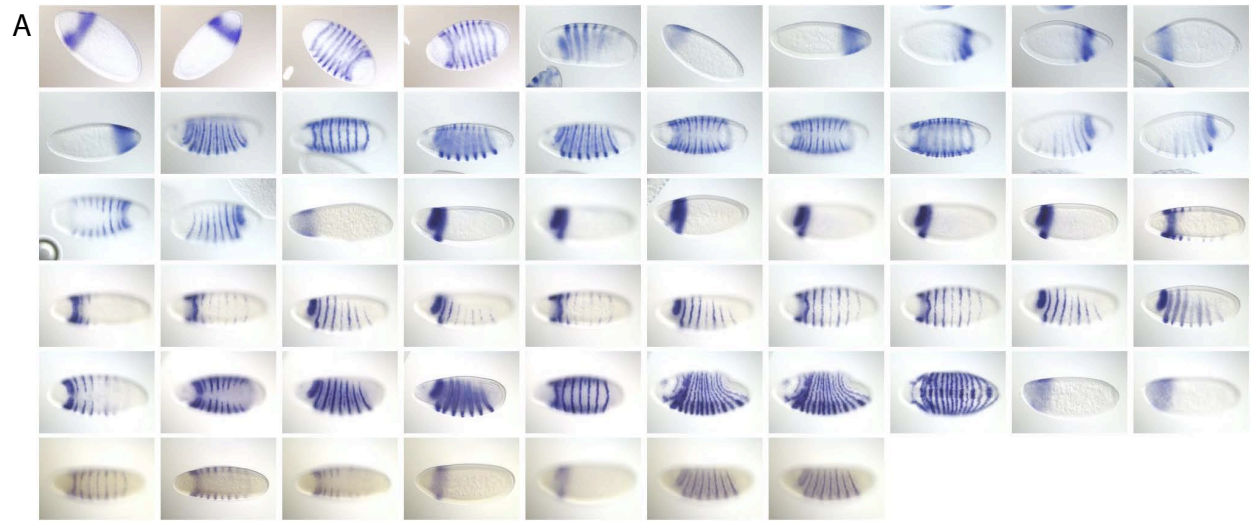
**Supplementary Figure 5** Examples of how TIs can be used to find related patterns. (**A**) Alignment of vectorized TIs shown for the overall similar patterns in Cluster #14. For the first three data vectors, the corresponding TIs are shown. (**B**) TIs at different stage ranges for the genes *homothorax* (*hth)* (stage 4-6)**,** *single-minded* (*sim)* (stage 9-10)*,* CG15743 (stage 11-12) and *pipe* (*pip)* (stage 13-16) were used to identify genes with similar expression patterns in the entire (unfiltered) dataset. The first row shows the original image and the TIs used for the search, the subsequent five rows show the search results in descending order of similarity. Hits were ranked by their distance score. Top hits for *hth* include the related gene *D* and for *sim* the related gene *mfas*. Note the patterns are highly correlated but their intensities are not. The query with *pip* revealed a missing annotation. *pip* was annotated with the terms “salivary gland” and “spiracle” but the first hit for *pip*, CG13463 was only annotated with the term “salivary gland”. CG13463 shows expression in both tissues (green arrows).

**Supplementary Figure 6**



**Supplementary Figure 6** Examples of TIs generated from images in the literature. (**A**) Images of varied quality taken from Stathopoulos et al. (Stathopoulos *et al*, 2002) (top three) and Segal et al. (Segal *et al*, 2008) (bottom two) and the TIs generated by the fully automatic *TIgen* pipeline. The top three images had an approximate 1000x400 pixel resolution, the bottom two images 200x120 pixels. Note that the relative intensities are represented correctly but the background is stronger than the BDGP TIs as a result of the color reproduction and bluish tint. (**B**) Screenshot of the interactive Matlab segmentation option from *TIgen*. When *TIgen* is invoked with an option for manual anchor point selection, *TIgen* displays the image and the outline for the standardized ellipse. The user selects a point in the *in situ* image corresponding to the red dot on the standardized ellipse. Once selected program prompts the user for selecting the next point. Points that are already done are shown in blue on the *in situ* image and in green on the ellipse. For this Figure, a saved high-resolution image for the interactive tool was merged with a screenshot of an empty window. (**C**) Images from Abrams et al. (Abrams *et al*, 2006). 8 anchor points on the circumference and one point in the center were selected manually. Shown are the images (left, 400x200 pixels resolution), the ellipses aligned over the images after manual anchor point selection (middle) and the resulting TIs (right). Note that both images have rough texture (see enlargements in (D)) and the top image has text and an arrow touching the embryo boundary. (**D**) Color enhanced and enlarged upper left corner of the top image in (A) (left) and the two images in (C). Note the stripes in the background for the images in (C), which necessitated interactive anchor point selection.

**Supplementary Figure 7**



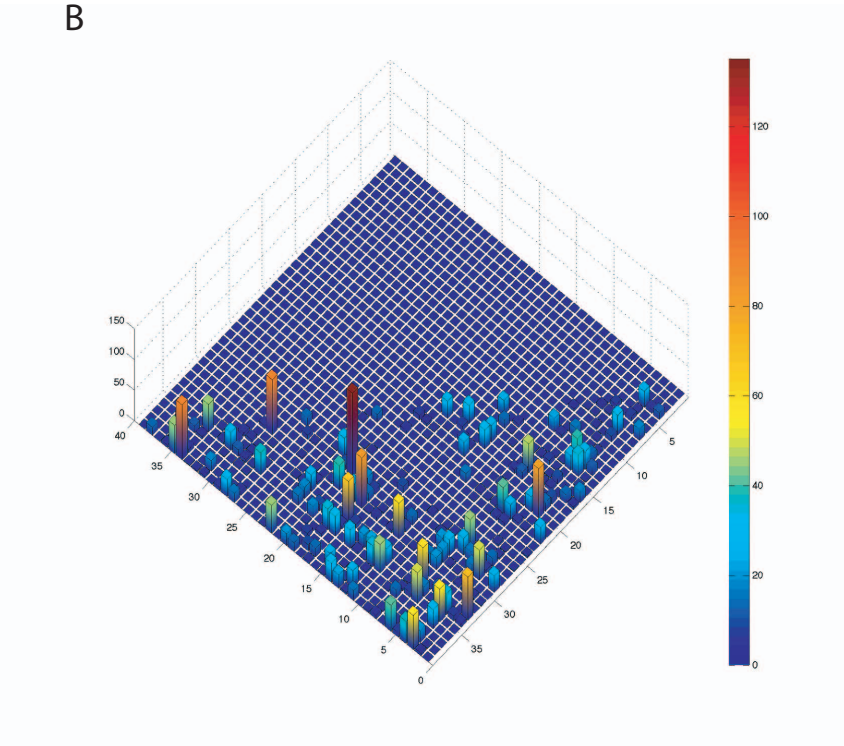
**Supplementary Figure 7** Reduction of multiple related images to principal TIs. (**A**) Database images for *slp1* at stages 4-6. Note the almost identical images in the third row. (**B**) Results of clustering the images with all images for each cluster shown. Five distinct clusters were identified. (**C** and **D**) Output for sorting the clusters with the shortest path algorithm. Cluster #5 was computationally identified as starting point and Cluster #3 as the endpoint (see Materials and Methods). The sorted clusters in D provide a progression of the pattern from the anterior to the entire embryo. The composite TIs are represented with the minimum values for each triangle in each cluster. (**E**) Examples of two genes at stage 4-6 and (**F**) two genes at stage 9-11 that were sorted and collapsed like *slp1*. CG5888 is contrast enhanced for better visibility of the pattern.

## Supplementary Figure 8

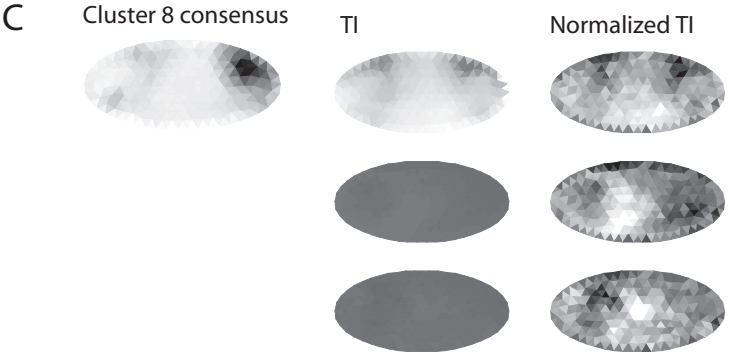
A

| Cluster | Cross validation accuracy |
|---|---|
| 1 | 94.55 |
| 2 | 90.00 |
| 3 | 97.65 |
| 4 | 88.33 |
| 5 | 92.94 |
| 6 | 96.00 |
| 7 | 91.43 |
| 8 | 96.36 |
| 9 | 95.38 |
| 10 | 89.09 |
| 11 | 97.65 |
| 12 | 97.33 |
| 13 | 96.67 |
| 14 | 98.00 |
| 15 | 94.12 |
| 16 | 95.00 |
| 17 | 95.56 |
| 18 | 94.00 |
| 19 | 96.00 |
| 20 | 95.79 |
| 21 | 93.64 |
| 22 | 86.15 |
| 23 | 92.50 |
| 24 | 95.38 |
| 25 | 97.78 |
| 26 | 91.82 |
| 27 | 93.33 |
| 28 | 92.86 |
| 29 | 93.00 |
| 30 | 93.33 |
| 31 | 97.50 |
| 32 | 100.00 |
| 33 | 96.92 |
| 34 | 91.43 |
| 35 | 93.33 |
| 36 | 91.11 |
| 37 | 96.36 |
| 38 | 97.14 |
| 39 | 96.67 |

B

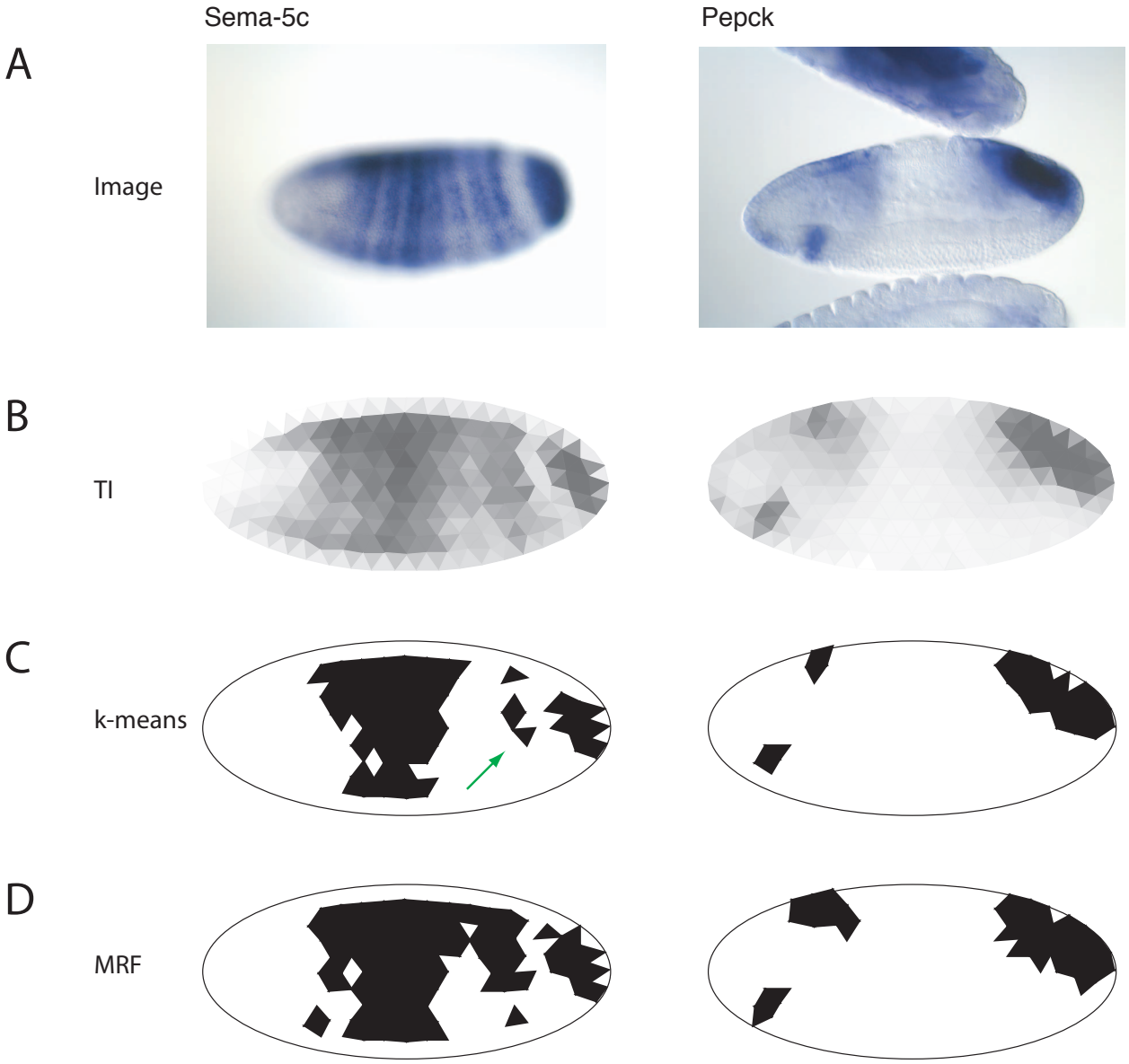



C



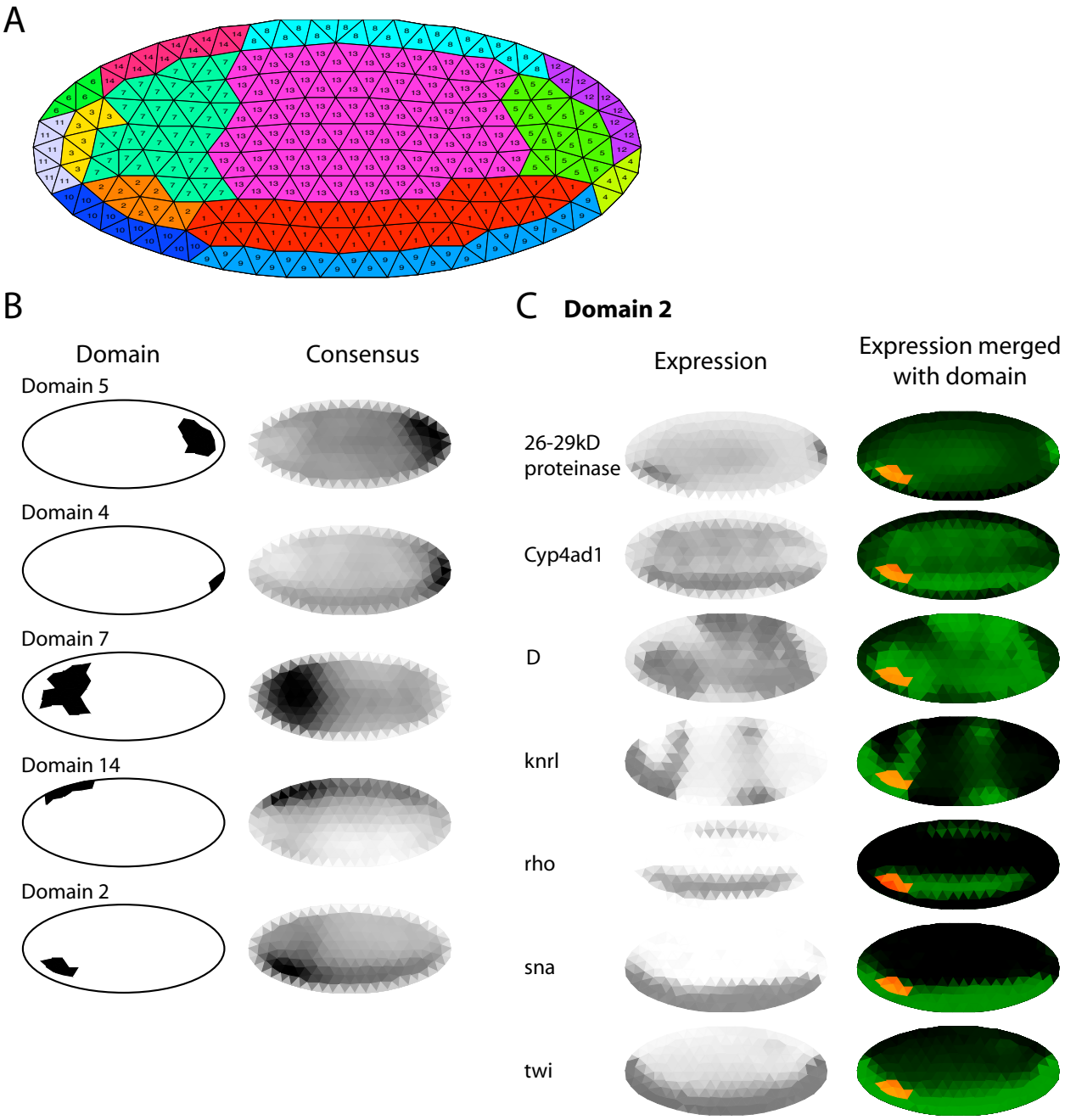
**Supplementary Figure 8** SVM classification training and results. (**A**) Cross validation results for each cluster for SVM training from the libsvm svmtrain output. Optimizing the parameters for each of the 39 clusters resulted in cross validation values above 90% with the exception of Cluster #4, #10 and #22. (**B**) Number of gene expression patterns that were classified into more than one cluster. The X and Y axes show the cluster number and the bar heights and color, from blue (lowest) to red (highest), indicate the number of patterns. We classified the entire dataset of 2,693 TIs with the optimized parameters from (A), including the 553 TIs from the training set, allowing classification into multiple groups. Only a small number of patterns were classified more than once, with the exception of clusters #20 and #29 that shared 135 TIs. (**C**) Example of poorly defined expression patterns that are associated with Cluster #8 because of their similarity to the training set. The consensus has a posterior and anterior dorsal pattern that is also seen in one of the members in the original training set. The raw TIs of two other cluster members appear almost ubiquitous but show a stronger expression around the position of the consensus when contrast enhanced by normalization (right).

**Supplementary Figure 9**

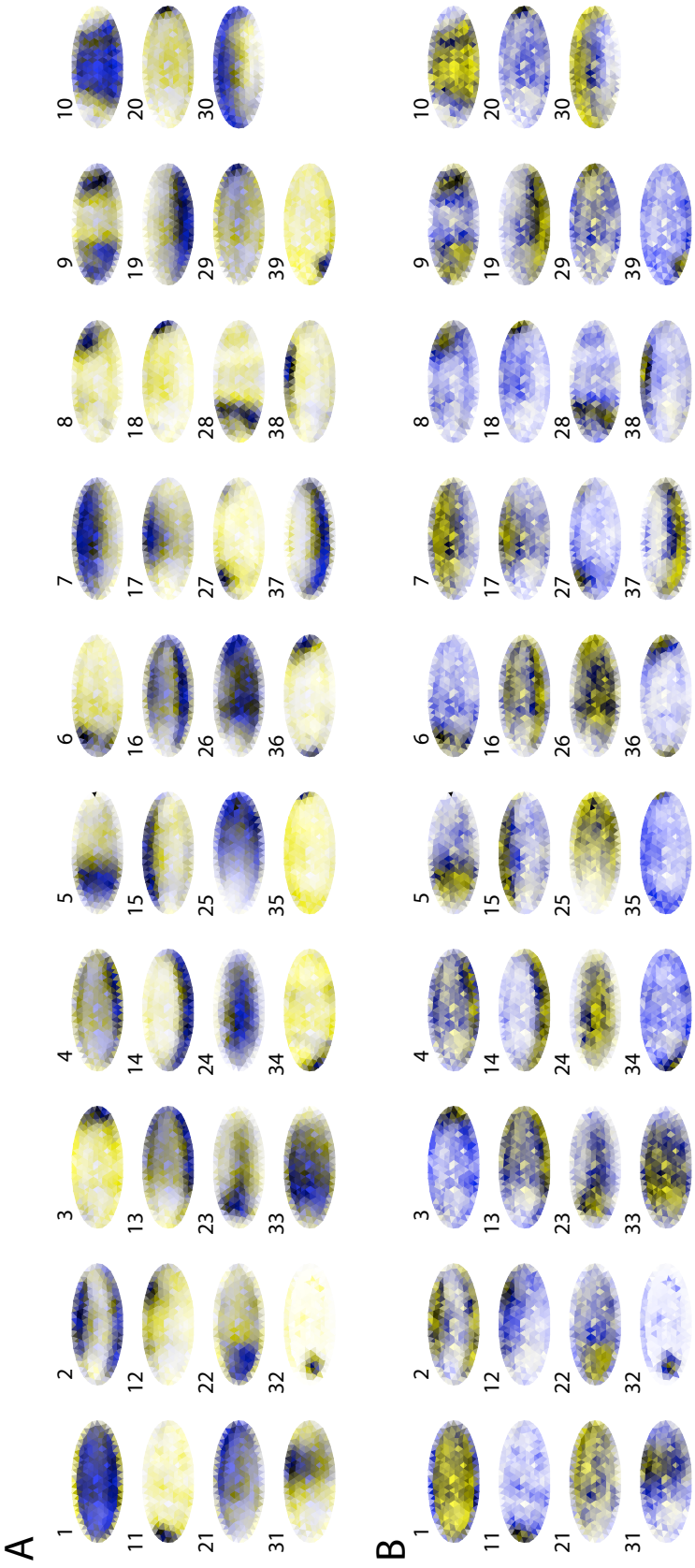



Result of extracting expression patterns with poorly defined boundaries. (**A**) Original images for *Sema-5c* (insitu38935.jpe) and *Pepck* (insitu42925.jpe). (**B**) TIs for the two images. The striped expression pattern for *Sema-5c* is only approximated due to the narrow gaps between the stripes. The three components of the *Pepck* expression pattern have different signal intensities and weak boundaries. Note the successful automatic separation of the two touching embryos for the *Pepck* image to create the TI. (**C**) Extraction of the expression patterns with k-means clustering. For both examples, significant parts of the patterns are missing. Note the completely absent part of the pattern at the posterior of the *Sema-5c* embryo (arrow). (**D**) Extraction of the expression patterns with MRF-algorithm. The resulting segmentation represents these complex patterns sufficiently accurate for analysis.

**Supplementary Figure 10**

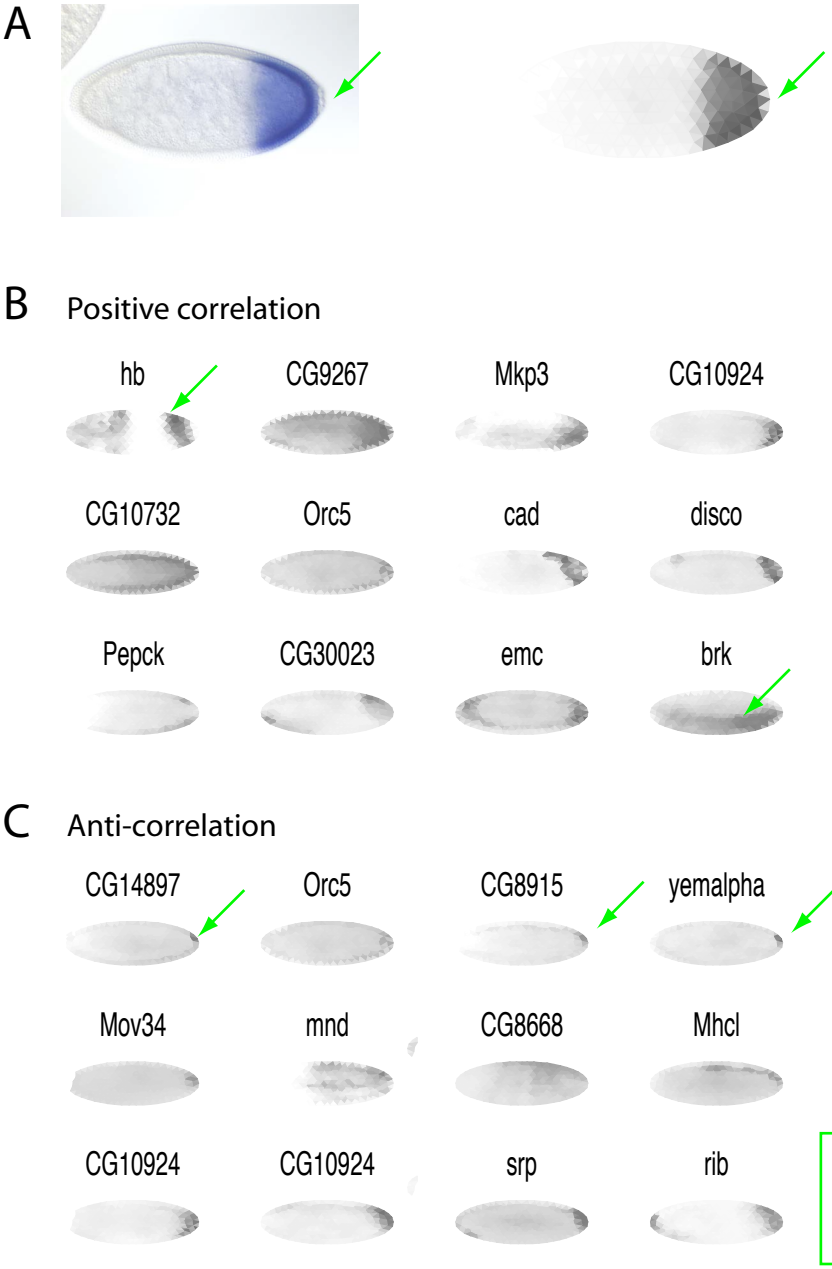



Genes defining expression domains. (**A**) Copy of Figure 3B showing 14 expression domains. (**B**) Normalized consensus patterns for 5 domains. Shown are domains 2, 4, 5, 7 and 14 from (A) and the consensus pattern for all genes with expression overlapping to each of the shown domains. The consensus patterns correspond remarkably well to each domain. (**C**) Examples for genes overlapping with domain 2 (as shown in (B) bottom). Gene expression is shown in grayscale (left) and green (right) with the overlapping domain 2 merged in yellow on the right. Note that only the *26-29kD proteinase* pattern corresponds to domain 2 but has another domain of expression at the posterior pole. All other patterns are only partially overlapping or larger than the domain (e.g. *twi*).




**Supplementary Figure 11**
Diversity of the patterns in each cluster. The consensus expression patterns from Figure 4A are shown in two yellow/blue color schemes with the consensus pattern in one color and the standard deviation in the other color. The standard deviation values are based on all 2,693 TIs grouped into the 39 clusters. In both cases, white indicates no expression. In (**A**) blue indicates uniform expression, yellow the normalized standard deviation. In (**B**) colors are reversed with the normalized standard deviation in blue. Triangles with greater variation are darker yellow or darker blue. Due to the normalization the darker standard deviation colored triangles primarily highlight variable regions and not poorly fitting patterns.

**Supplementary Figure 12**

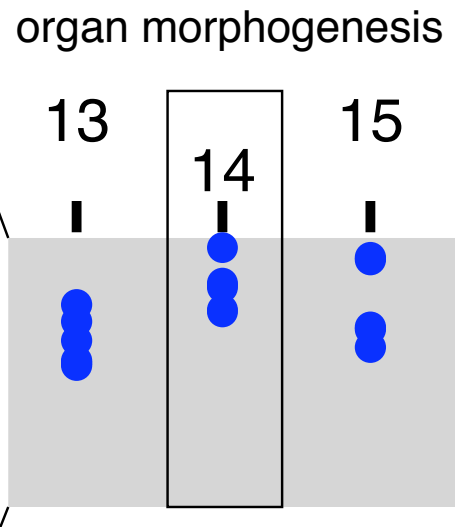



Pairwise MRF identifies partially overlapping and anti-correlated gene expression patterns (**A**) An image (insitu21116.jpe) and the corresponding TI for the gene *ribbon* (*rib*). Note that *rib* is excluded from the pole cells (arrow). (**B**) Top positive correlated patterns. Results show embryos with overlapping rib gene expression boundaries (e.g. *hunchback* (*hb*) and *Ptx1, rib* boundary indicated with green arrows) or embryos with the same rib boundary marked by changing intensities (e.g. *brinker* (*brk)*, arrow indicates the intensity change) (**C**) Top anti-correlated patterns. A search for anti-correlated patterns identifies several genes with expression in the pole cells (arrows) including *oskar* (*osk)*, a well-described gene that is known to control pole cell development.

## Supplementary Figure 13

A Cluster 14

GO:0001746 Bolwig's organ morphogenesis
GO:0007443 Malpighian tubule morphogenesis
GO:0045467 R7 cell development
GO:0045466 R7 cell differentiation
GO:0007465 R7 cell fate commitment
GO:0048800 antennal morphogenesis
GO:0007375 anterior midgut invagination
GO:0042675 compound eye cone cell differentiation
GO:0042676 compound eye cone cell fate commitment
GO:0042679 compound eye cone cell fate specification
GO:0001745 compound eye morphogenesis
GO:0001751 compound eye photoreceptor cell differentiation
GO:0042051 compound eye photoreceptor development
GO:0001752 compound eye photoreceptor fate commitment
GO:0046667 compound eye retinal cell programmed cell death
GO:0048546 digestive tract morphogenesis
GO:0007391 dorsal closure
GO:0007394 dorsal closure, elongation of leading edge cells
GO:0048567 ectodermal gut morphogenesis
GO:0048619 embryonic hindgut morphogenesis
GO:0042067 establishment of ommatidial polarity
GO:0001736 establishment of planar polarity
GO:0016334 establishment or maintenance of polarity of follicular epithelium
GO:0048592 eye morphogenesis
GO:0001754 eye photoreceptor cell differentiation
GO:0042706 eye photoreceptor cell fate commitment
GO:0007455 eye-antennal disc morphogenesis
GO:0007440 foregut morphogenesis
GO:0010160 formation of organ boundary
GO:0048547 gut morphogenesis
GO:0007442 hindgut morphogenesis
GO:0007560 imaginal disc morphogenesis
GO:0016348 imaginal disc-derived leg joint morphogenesis
GO:0007480 imaginal disc-derived leg morphogenesis
GO:0008587 imaginal disc-derived wing margin morphogenesis
GO:0007476 imaginal disc-derived wing morphogenesis
GO:0008586 imaginal disc-derived wing vein morphogenesis
GO:0007474 imaginal disc-derived wing vein specification
GO:0007392 initiation of dorsal closure
GO:0007478 leg disc morphogenesis
GO:0001707 mesoderm formation
GO:0007509 mesoderm migration
GO:0048332 mesoderm morphogenesis
GO:0048333 mesodermal cell differentiation
GO:0001710 mesodermal cell fate commitment
GO:0007500 mesodermal cell fate determination
GO:0007501 mesodermal cell fate specification
GO:0008078 mesodermal cell migration
GO:0001738 morphogenesis of a polarized epithelium
GO:0002009 morphogenesis of an epithelium
GO:0016331 morphogenesis of embryonic epithelium
GO:0016333 morphogenesis of follicular epithelium
GO:0042683 negative regulation of compound eye cone cell fate specification
GO:0016318 ommatidial rotation
GO:0048645 organ formation
GO:0009887 organ morphogenesis
GO:0007428 primary branching, open tracheal system
GO:0007458 progression of morphogenetic furrow during compound eye morphogenesis
GO:0042682 regulation of compound eye cone cell fate specification
GO:0046669 regulation of compound eye retinal cell programmed cell death
GO:0046668 regulation of retinal cell programmed cell death
GO:0046666 retinal cell programmed cell death
GO:0007432 salivary gland boundary specification
GO:0008052 sensory organ boundary specification
GO:0016360 sensory organ precursor cell fate determination
GO:0048729 tissue morphogenesis
GO:0007472 wing disc morphogenesis



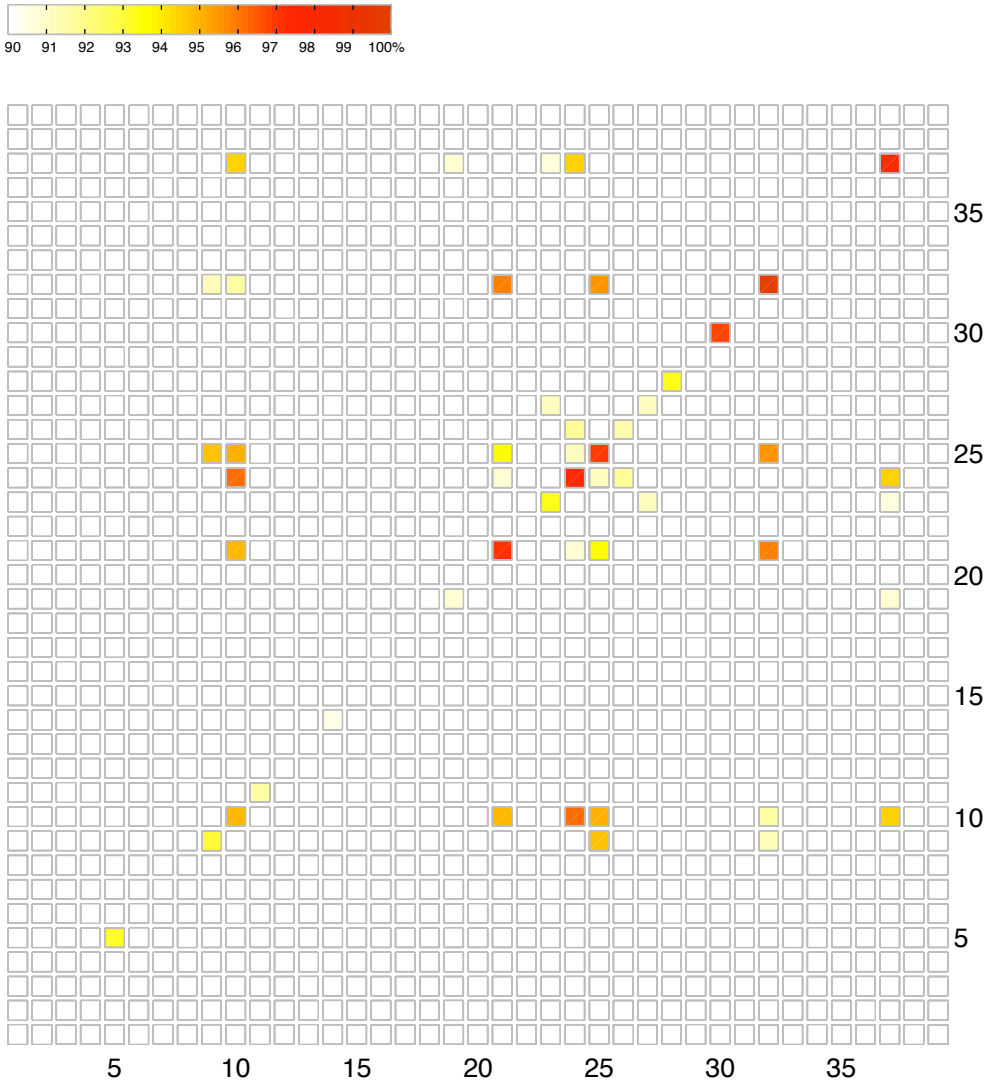

B

**Cluster 6**
GO:0055034 Bolwig's organ development
GO:0001746 Bolwig's organ morphogenesis
GO:0045466 R7 cell differentiation
GO:0007465 R7 cell fate commitment
GO:0042675 compound eye cone cell differentiation
GO:0042676 compound eye cone cell fate commitment
GO:0001751 compound eye photoreceptor cell differentiation
GO:0001752 compound eye photoreceptor fate commitment
GO:0046667 compound eye retinal cell programmed cell death
GO:0042067 establishment of ommatidial polarity
GO:0042706 eye photoreceptor cell fate commitment
GO:0046669 regulation of compound eye retinal cell programmed cell death
GO:0046668 regulation of retinal cell programmed cell death
GO:0008052 sensory organ boundary specification
GO:0016360 sensory organ precursor cell fate determination

**Cluster 27**
GO:0007465 R7 cell fate commitment
GO:0001752 compound eye photoreceptor fate commitment
GO:0046667 compound eye retinal cell programmed cell death
GO:0042067 establishment of ommatidial polarity
GO:0042706 eye photoreceptor cell fate commitment
GO:0016318 ommatidial rotation
GO:0046666 retinal cell programmed cell death

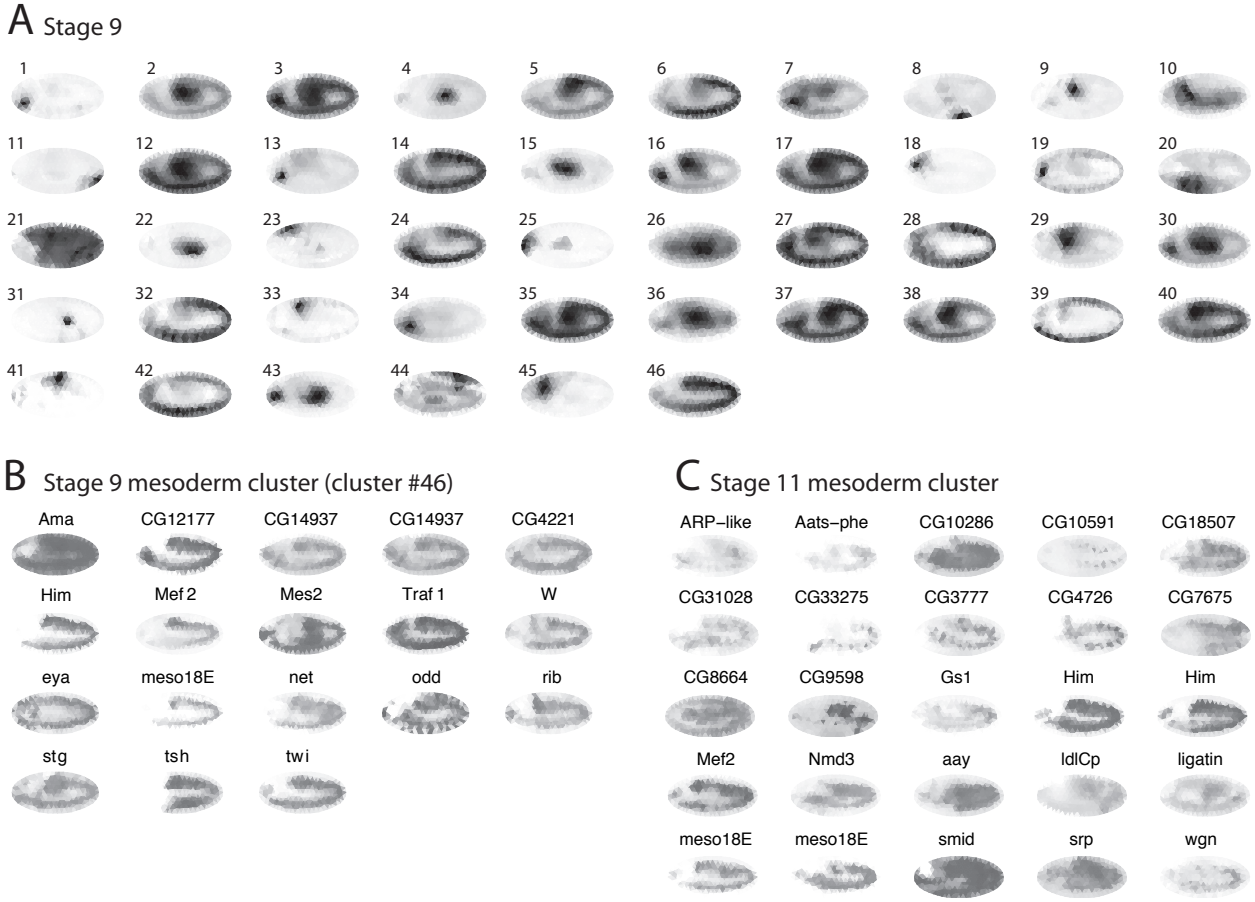
**Supplementary Figure 13** Additional details from Figure 7B. (**A**) Magnification of the category "organ morphogenesis" around cluster #14 (framed). The contents for the entire category are shown on the left and the GO terms corresponding to the blue dots in cluster #14 are highlighted in blue. The magnification picture is rotated 180 degrees from Figure 7B so that the order of the dots correspond to the order of the GO terms. The large gray space at the bottom was originally reserved for printing the category and is not indicative of further terms not shown. (**B**) Excerpts of the contents for the category "sensory organ development" for clusters #6 and #27. Only enriched terms are shown. The contents for the two clusters highlight the variable terms for eye development that required the broad categorization in Figure 7B.

**Supplementary Figure 14**




Genetic interactions within and between the 39 clusters from Figure 4A. Enriched genetic interactions are displayed as level of significance between white (90% and below) and red (100%). Shown are interactions between the 39 clusters on the horizontal and vertical axis and the interactions within each cluster on the diagonal. 611 genes of the genetic interaction dataset (Droid v4.0) match our 1881 genes. Note that all of the highly enriched interactions are on the diagonal.

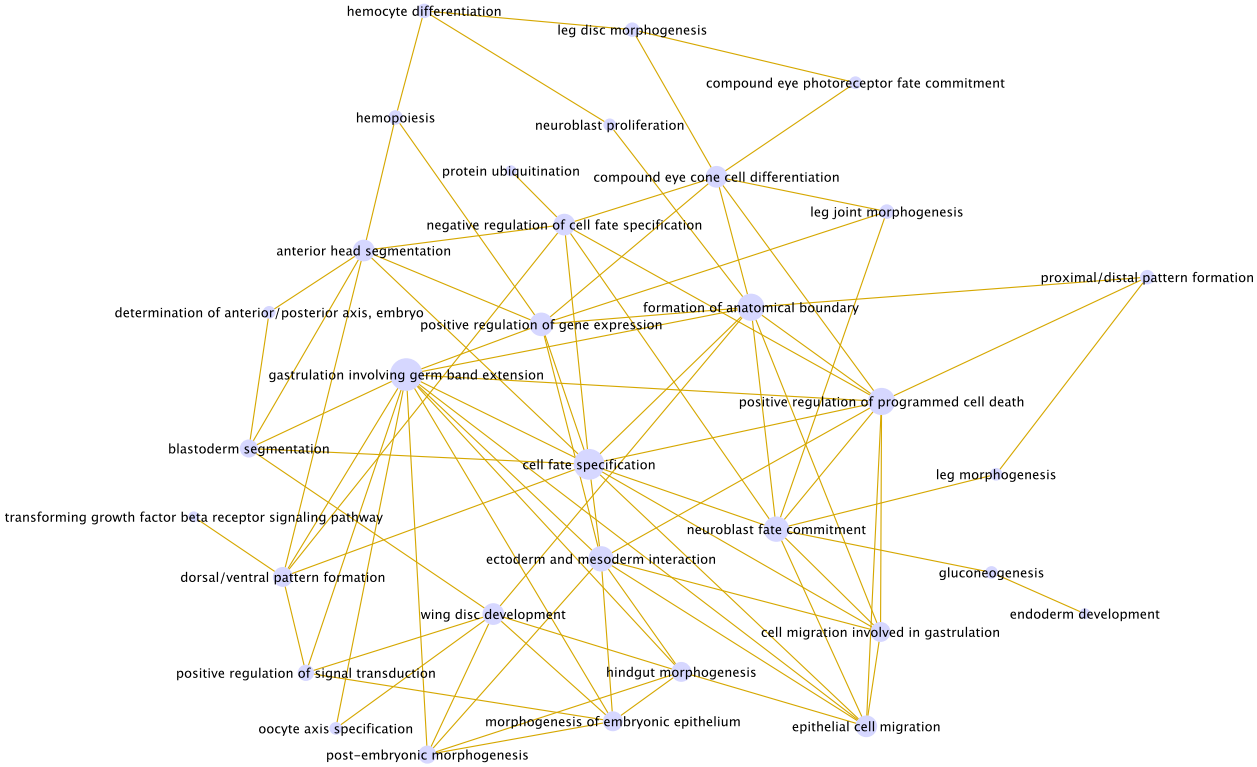
**Supplementary Figure 15**




Clustering genes with similar patterns at later stages. (**A**) Normalized consensus patterns for 46 clusters identified by APC using the 609 distinct stage 9-10 TIs from the *TIfilter* pipeline (Steps are summarized in Supplementary Figure 1). (**B**) Cluster #46 containing many genes with previously identified roles in mesoderm development. (**C**) Similar mesoderm cluster identified by APC clustering of stage 11-12 TIs.

## Supplementary Figure 16

A



# Supplementary Figure 16

B

**positive regulation of gene expression**
positive regulation of macromolecule biosynthetic process
positive regulation of nucleobase nucleoside nucleotide and nucleic acid metabolic process
positive regulation of macromolecule metabolic process
positive regulation of transcription DNA-dependent
positive regulation of RNA metabolic process
positive regulation of transcription

**post-embryonic morphogenesis**
regulation of transcription from RNA polymerase II promoter
instar larval or pupal development
post-embryonic development

**wing disc development**
cell fate commitment
appendage morphogenesis
appendage development
imaginal disc-derived wing morphogenesis
imaginal disc-derived appendage morphogenesis
wing disc morphogenesis
imaginal disc-derived appendage development
imaginal disc development
imaginal disc morphogenesis
gland development

**dorsal/ventral pattern formation**
imaginal disc pattern formation
apical protein localization
ectoderm development
imaginal disc-derived wing vein specification

**neuroblast proliferation**
development of primary sexual characteristics
neuroblast division
asymmetric neuroblast division
somatic stem cell division
antennal development

**blastoderm segmentation**
segmentation
central nervous system development
embryonic pattern specification
periodic partitioning
stem cell differentiation
segment specification

**oocyte axis specification**
tissue development
anterior/posterior axis specification
anatomical structure arrangement
oocyte construction
oocyte morphogenesis
oocyte development

**proximal/distal pattern formation**
proximal/distal pattern formation
imaginal disc
salivary gland boundary specification
branched duct epithelial cell fate determination
open tracheal system

**positive regulation of signal transduction**
positive regulation of cellular process
imaginal disc-derived wing vein morphogenesis
positive regulation of cell communication
regulation of signal transduction
regulation of immune system process
regulation of cell communication
dorsal/ventral axis specification

**compound eye photoreceptor fate commitment**
eye photoreceptor cell fate commitment
photoreceptor cell fate commitment
neuron fate commitment
compound eye retinal cell programmed cell death

**formation of anatomical boundary**
peripheral nervous system development
organ formation
formation of organ boundary
sensory organ precursor cell fate determination
sensory organ boundary specification

**endoderm development**
endoderm formation
negative regulation of oskar mRNA translation

**positive regulation of programmed cell death**
genital disc development
positive regulation of apoptosis
apoptotic program
cardiac cell differentiation
induction of apoptosis
regulation of phosphorylation
asymmetric cell division
induction of programmed cell death

**epithelial cell migration**
epithelial cell migration
open tracheal system
lymph gland development
regulation of mitosis
germ cell migration

**determination of anterior/posterior axis**
embryo
tripartite regional subdivision
embryonic axis specification
eye-antennal disc morphogenesis

**hindgut morphogenesis**
tube development
gut development
ectodermal gut morphogenesis
ectodermal gut development
gut morphogenesis
digestive tract morphogenesis

**ectoderm and mesoderm interaction**
formation of primary germ layer
mesoderm morphogenesis
mesoderm formation
embryonic hindgut morphogenesis
Malpighian tubule morphogenesis
tissue morphogenesis
anterior midgut invagination
mesodermal cell fate commitment
mesodermal cell differentiation
mesodermal cell fate determination
mesodermal cell fate specification

**gluconeogenesis**
primary branching
open tracheal system
glucose metabolic process
pyruvate metabolic process
zygotic determination of anterior/posterior axis embryo
regulation of oskar mRNA translation
hexose metabolic process
monosaccharide metabolic process
analia development
monosaccharide biosynthetic process
hexose biosynthetic process
alcohol biosynthetic process

**morphogenesis of embryonic epithelium**
exocrine system development
salivary gland development
open tracheal system development
embryonic morphogenesis
tube morphogenesis
morphogenesis of a branching structure
morphogenesis of an epithelium
ventral furrow formation

**cell fate specification**
cell fate determination
heart development
ventral cord development

**gastrulation involving germ band extension**
gastrulation
mesoderm development
gastrulation with mouth forming first
negative regulation of transcription from RNA polymerase II promoter

**cell migration involved in gastrulation**
ganglion mother cell fate determination
mesodermal cell migration
pericardial cell differentiation
mesoderm migration
visceral muscle development

**leg morphogenesis**
limb morphogenesis
limb development
imaginal disc-derived leg joint morphogenesis
regulation of peptidase activity
regulation of caspase activity
regulation of endopeptidase activity

**neuroblast fate commitment**
stem cell fate determination
stem cell fate commitment
neuroblast fate determination
neuroblast differentiation

**negative regulation of cell fate specification**
regulation of cell fate commitment
regulation of cell fate specification
negative regulation of cell fate commitment
initiation of dorsal closure
dorsal/ventral pattern formation
imaginal disc

**anterior head segmentation**
posterior head segmentation
head segmentation
negative regulation of cell differentiation
R7 cell fate commitment
specification of segmental identity
head
establishment of ommatidial polarity

**compound eye cone cell differentiation**
compound eye cone cell fate commitment
positive regulation of transcription from RNA polymerase II promoter
progression of morphogenetic furrow during compound eye morphogenesis

**transforming growth factor beta receptor signaling pathway**
amnioserosa formation
transmembrane receptor protein serine/threonine kinase signaling pathway

**hemopoiesis**
immune system development
hemopoietic or lymphoid organ development

**leg disc morphogenesis**
imaginal disc-derived leg morphogenesis
imaginal disc-derived limb morphogenesis

**hemocyte differentiation**
crystal cell differentiation

**protein ubiquitination**
regulation of Notch signaling pathway
morphogenesis of follicular epithelium
protein modification by small protein conjugation

**Supplementary Figure 16** Associations between biological functions. (**A**) Connections between GO terms. Each node is a cluster of GO terms and the GO term at the center as identified by the clustering algorithm of the cluster is shown. Terms condensed to nodes are overwhelmingly functionally related. The size of the blue circle corresponds to the number of connections. Shown are only the nodes connections with a Jaccard distance equal or greater than 0.3. Distances between nodes are nearly identical except for a weaker link between gastrulation and mesoderm development. (**B**) Node contents of (A). Each node corresponding to the node labels (A) is shown in bold and the terms condensed in the node below. As previously demonstrated (Bonini and Fortini, 1999), we found nodes centered on both a cell death term and an eye development term linked together. Likewise, the cell death node is also connected to a node centered on neural development terms, which was an association that was known previously (Buss and Oppenheim, 2004). The most interconnected nodes are cell death, gastrulation and cell fate specification which are important developmental processes at stages 4-6.

# Supplementary Tables

## Supplementary Table I

| Step | # of TI | | | |
|---|---|---|---|---|
| | Stage 4-6 | Stage 7-9 | Stage 9-10 | Stage 11-13 |
| Start with all patterns (lateral view) | 5745 | 3227 | 3247 | 8446 |
| Remove meshes with no/ubiquitous expression | 4454 | 2499 | 2405 | 6834 |
| Remove redundant patterns | 2693 | 1820 | 1764 | 3660 |
| Select most distinct patterns | 553 | 636 | 609 | 1440 |

Results of filtering TIs at different stage ranges.

## Supplementary Table II

| Domain 1 | Domain 2 | Domain 3 | Domain 4 | Domain 5 | Domain 6 | Domain 7 | Domain 8 | Domain 9 | Domain 10 | Domain 11 | Domain 12 | Domain 13 | Domain 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aats-lys | 26-29kD-prote | Aats-lys | CycA | Aats-lys | Adgf-A | Aats-lys | Ama | Ama | BicD | Adgf-A | Calx | Aats-lys | Ance |
| Adam | Aats-lys | Adgf-A | Cyp310a1 | Cctgamma | Alhambra | Fer2LCH | CycA | Cctgamma | Cyp310a1 | Alhambra | CycA | Alhambra | BicD |
| Alhambra | Ama | Alhambra | Eip71CD | CycA | BicD | Hsp83 | Cys | Cyp306a1 | Mdr49 | Cad74A | Doc2 | D | Dl |
| Ama | Arf51F | Cad74A | Gapdh1 | Cyp306a1 | Cad74A | Map60 | D | Cyp310a1 | Mes2 | Hsp83 | Eip71CD | DnaJ-1 | Doc3 |
| Cctgamma | Cyp4ad1 | Cyp307a1 | Gfat2 | Eip71CD | Cyp307a1 | Nek2 | Dl | D | Ndae1 | ImpE2 | MESR3 | Fer2LCH | Nek2 |
| Cyp4ad1 | D | D | Grp1 | Fer2LCH | ImpE2 | Sec61beta | Doc2 | Mdr49 | NetA | Mdr49 | Mkp3 | Jafrac1 | Pepck |
| D | Dg | Fer2LCH | Hsp27 | Grp1 | Mdr49 | SoxN | Doc3 | Mes2 | Pka-C3 | Mkp3 | Mlf | Map60 | bun |
| Dg | DnaJ-1 | Hsp83 | ImpE2 | Hsp27 | Optix | danr | Eip71CD | Ndae1 | PyK | Ndae1 | Nek2 | Mhcl | eiger |
| Fer2LCH | Fer2LCH | ImpE2 | Jafrac1 | Jafrac1 | PyK | fj | Mipp1 | NetA | Traf1 | Optix | Orc5 | PpD3 | fj |
| Jafrac1 | Gapdh1 | Map60 | MESR3 | MESR3 | RhoGEF3 | gukh | Myo31DF | Pepck | gukh | Pka-C3 | Pepck | Sema-5c | hb |
| Map60 | Grip75 | Mdr49 | Mcm2 | Map60 | Sec61beta | hb | Orc5 | Pka-C3 | knrl | PyK | PyK | aay | lok |
| Mcm3 | GstS1 | Ndae1 | Mes2 | Mcm2 | apt | loner | PyK | RhoGAP71E | mus210 | Sec61beta | aay | bin3 | mus210 |
| Mes2 | Hsp83 | Optix | Mkp3 | Mkp3 | bowl | mus210 | bun | Traf1 | noc | Smg5 | apt | danr | neur |
| Mo25 | Jafrac1 | Pepck | Myo31DF | Mlf | cnc | slp1 | dbo | Wnt8 | ry | apt | bun | en | nuf |
| Ndae1 | Map60 | RhoGEF3 | Ndae1 | Mov34 | eiger | yellow-e3 | eiger | brat | sna | bowl | dm | feo | spdo |
| NetA | Mcm3 | Sec61beta | Nek2 | Nek2 | lok | | exex | fj | spdo | dm | eiger | l(2)08717 | tld |
| Pka-C3 | Mdr49 | Sep5 | Orc5 | Orc5 | mus210 | | hth | fu2 | stumps | fkh | gk | odd | toc |
| PpD3 | Mes2 | Tif-IA | Ote | Ote | nAcRalpha-96Ab | | mew | gk | trx | hkb | l(3)00305 | pgc | tok |
| Sep5 | Ndae1 | apt | Pepck | Pepck | noc | | net | gukh | twi | knrl | mspo | prd | zen |
| SoxN | NetA | bowl | Pros54 | PpD3 | ry | | pgc | msk | | lok | nAcRalpha-96 | run | |
| TER94 | NetB | cnc | Ptx1 | Ptx1 | slp1 | | pyd3 | neur | | mus210 | pgc | shu | |
| Traf1 | Osi19 | danr | Wnt8 | PyK | tkv | | rho | ry | | noc | pxt | sob | |
| Wnt8 | Pka-C3 | dm | apt | RhoGAP18B | toc | | slmb | sna | | ry | tipE | yellow-e3 | |
| aay | PyK | eiger | cad | Sep5 | zen | | toc | stg | | slp1 | zen | | |
| bin3 | RhoGEF3 | emc | charybde | Set | | | ush | stumps | | toc | | | |
| brk | RnrS | en | disco | Src64B | | | zen | tin | | twi | | | |
| bs | Sec61beta | fj | dm | TER94 | | | | tkv | | | | | |
| chif | Sep5 | halo | emc | Vinc | | | | trx | | | | | |
| danr | SoxN | hb | fkh | aay | | | | twi | | | | | |
| dm | TER94 | lok | fu2 | apt | | | | yemalpha | | | | | |
| en | bowl | mus210 | gk | bin3 | | | | | | | | | |
| feo | btd | noc | gukh | charybde | | | | | | | | | |
| fu2 | danr | numb | gwl | cib | | | | | | | | | |
| gcl | dm | ry | hb | disco | | | | | | | | | |
| gk | en | slp1 | hkb | dm | | | | | | | | | |
| gukh | gukh | smg | l(2)08717 | exex | | | | | | | | | |
| ial | hb | sog | l(3)00305 | feo | | | | | | | | | |
| l(1)sc | ial | stg | mspo | fu2 | | | | | | | | | |
| l(2)08717 | knrl | toc | osk | gk | | | | | | | | | |
| odd | l(2)08717 | twi | pxt | halo | | | | | | | | | |
| phyl | loner | | ry | hb | | | | | | | | | |
| prd | mus210 | | shu | l(3)00305 | | | | | | | | | |
| pxb | neur | | sim | mnd | | | | | | | | | |
| run | noc | | spdo | mspo | | | | | | | | | |
| ry | phyl | | srp | osk | | | | | | | | | |
| sala | pxt | | trx | pcs | | | | | | | | | |
| shu | rho | | twi | sala | | | | | | | | | |
| slp2 | ry | | yemalpha | scu | | | | | | | | | |
| smg | sala | | | shu | | | | | | | | | |
| sna | slp2 | | | toc | | | | | | | | | |
| sog | smg | | | zen | | | | | | | | | |
| tkv | sna | | | | | | | | | | | | |
| twi | sog | | | | | | | | | | | | |
| | spdo | | | | | | | | | | | | |
| | srp | | | | | | | | | | | | |
| | stg | | | | | | | | | | | | |
| | twi | | | | | | | | | | | | |
| | zfh1 | | | | | | | | | | | | |

Genes represented by the TIs expressed in each of the 14 domains shown in Figure 3B/ Supplementary Figure 10A. Only named genes are shown; a complete list of genes, including unnamed genes, can be found in the online materials (http://www.fruitfly.org/insitu/FriseMSB/domains). Note that domains contain genes expected at their location. For example, domain 2 (anterior) contains nervous system specific genes *rhomboid* (*rho*), *sanpodo* (*spdo*) and *SoxNeuro* (*SoxN*); domain 4 (posterior terminal) contains germ-line specific and anterior terminal specific genes *caudal* (*cad*), *huckebein* (*hkb*) and *oskar* (*osk*); domain 9 (ventral) contains mesoderm specific genes *Mes2, sna, tin, twi*.

## Supplementary Datasets

**Dataset 1**
Compressed (Zip) file of the software code for *TIgen*, *TIfilter*, *TIfate, TIbin2*, *TImrf* and *TImrf2* and the triangle coordinates and assignments for the virtual representation (Figure 1F). See included README-MSB.txt for details.

**Dataset 2**
CSV file containing the results of *TIfilter* at stage 4-6. The columns are: "reference" is a unique identifier for the TI; "order" is the temporal sort order of the pattern; "patterned" is 1 if the TI was included in the 553 distinctly patterned TIs, 0 otherwise; "symbol" is the gene symbol and "FBgn" the Flybase FBgn number.

**Dataset 3**
CSV file with the TI assignments to each of the 39 clusters. The first column, "reference" refers to the unique "reference" identifier in Dataset 1, the other columns are the 39 clusters. TIs are in the clusters if their column entry is set to 1.

**Dataset 4**
Compressed (Zip) file with the gene symbols in each cluster.

**Dataset 5**
List of all GO-terms on the vertical axis of Figure 7B, sorted by the categories in Figure 7B.

**Dataset 6**
CSV file detailing the GO-terms associated to each cluster in Figure 7B. The columns are: "cluster" is the cluster number (1-39), "parent_goid" is the GO-id for the categories on the vertical axis in Figure 7B; "parent_goterm" is the corresponding term to parent_goid; "goid" is the GO-id in each cluster and "goterm" is the corresponding term to goid.

**Dataset 7**
List of GO-terms in each cluster in Figure 7B, sorted first by category, then by cluster. This dataset is a human readable version of Dataset 6.

**Dataset 8**
List of clusters with significant correlation between the categorized GO-terms.

**Dataset 9**
List of the 869 unique term pairs (199 GO-terms) that are correlated to each other and summarized in Supplementary Figure 16. Each line contains one correlated term pair and the correlated GO-terms (GO-id in parenthesis) are separated by a hypen ("-").

## References

Abrams EW, Mihoulides WK, Andrew DJ (2006) Fork head and Sage maintain a uniform and patent salivary gland lumen through regulation of two downstream target genes, PH4alphaSG1 and PH4alphaSG2. *Development* **133:** 3517-3527.

Bonini NM, Fortini ME (1999) Surviving Drosophila eye development: integrating cell death with differentiation during formation of a neural structure. *Bioessays* **21**: 991-1003.

Buss RR, Oppenheim RW (2004) Role of programmed cell death in normal neuronal development and function. *Anat Sci Int* **79**: 191-197.

Chang C-C, Lin C-J (2001) LIBSVM: a library for support vector machines.

Dijkstra EW (1959) A note on two problems in connexion with graphs. v. *Numerische Mathematik* **1:** 269-271.

Fan R-E, Chen P-H, Lin C-J (2005) Working set selection using the second order information for training SVM. *Journal of Machine Learning Research* **6:** 1889-1918.

Prim R (1957) Shortest Connection Networks and Some Generalizations. *Bell System Technical Journal* **36:** 1389-1401.

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **451:** 535-540.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498-2504.

Stathopoulos A, Van Drenth M, Erives A, Markstein M, Levine M (2002) Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo. *Cell* **111:** 687-701.