BMC Bioinformatics

Identifying spatially similar gene expression patterns in early stage fruit fly embryo images: binary feature versus invariant moment digital representations

BMC Bioinformatics 2004, 5:202 doi:10.1186/1471-2105-5-202

Rajalakshmi Gurunathan (Rajalakshmi.Gurunathan@asu.edu)
Bernard Van Emden (Bernard.VanEmden@asu.edu)
Sethuraman Panchanathan (panch@asu.edu)
Sudhir Kumar (s.kumar@asu.edu)

ISSN 1471-2105

Article type Software

Submission date 30 Apr 2004

Acceptance date 16 Dec 2004

Publication date 16 Dec 2004

Article URL http://www.biomedcentral.com/1471-2105/5/202

Identifying spatially similar gene expression patterns in early stage fruit fly embryo images: binary feature versus invariant moment digital representations

Rajalakshmi Gurunathan[1,2] , Bernard Van Emden[1,3], Sethuraman Panchanathan[2] and Sudhir Kumar[1,3§]

[1]Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University, Tempe, AZ 85287-5301, USA

[2]Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287-8809, USA

[3]School of Life Sciences, Arizona State University, Tempe, AZ 85287-4501, USA

[§]Corresponding author

Email addresses:
SK: s.kumar@asu.edu
SP: panch@asu.edu
RG: rajalakshmi.gurunathan@asu.edu
BVE: bernard.vanemden@asu.edu

[§]Address for correspondence:
Dr. Sudhir Kumar
Biodesign Building A-240
Arizona State University
Tempe, AZ 85287-5301, USA
Tel: (480) 727 6949
Fax: (480) 727 6947
E-mail: s.kumar@asu.edu

Abstract

Background

Modern developmental biology relies heavily on the analysis of embryonic gene expression patterns. Investigators manually inspect hundreds or thousands of expression patterns to identify those that are spatially similar and to ultimately infer potential gene interactions. However, the rapid accumulation of gene expression pattern data over the last two decades, facilitated by high-throughput techniques, has produced a need for the development of efficient approaches for direct comparison of images, rather than their textual descriptions, to identify spatially similar expression patterns.

Results

The effectiveness of the Binary Feature Vector (BFV) and Invariant Moment Vector (IMV) based digital representations of the gene expression patterns in finding biologically meaningful patterns was compared for a small (226 images) and a large (1819 images) dataset. For each dataset, an ordered list of images, with respect to a query image, was generated to identify overlapping and similar gene expression patterns, in a manner comparable to what a developmental biologist might do. The results showed that the BFV representation consistently outperforms the IMV representation in finding biologically meaningful matches when spatial overlap of the gene expression pattern and the genes involved are considered. Furthermore, we explored the value of conducting image-content based searches in a dataset where individual expression components (or domains) of multi-domain expression patterns were also included separately. We found that this technique improves performance of both IMV and BFV based searches.

Conclusions

We conclude that the BFV representation consistently produces a more extensive and better list of biologically useful patterns than the IMV representation. The high quality of results obtained scales well as the search database becomes larger, which encourages efforts to build automated image query and retrieval systems for spatial gene expression patterns.

Background

The complexity of animal body form arises from a single fertilized egg cell in an odyssey of gene expression and regulation that controls the multiplication and differentiation of cells [1-3]. For over two decades, *Drosophila melanogaster* (the fruit fly) has been a canonical model animal for understanding this developmental process in the laboratory. The raw data from experiments consist of photographs (two dimensional images) of the *Drosophila* embryo showing a particular gene expression pattern revealed by a gene-specific probe in wildtype and mutant backgrounds. Manual, visual comparison of these spatial gene expressions is usually carried out to identify overlaps in gene expression and to infer interactions [4-6].

Whole fruit fly embryo and other related gene expression patterns have been published in a wide variety of research journals since late 1980's. These efforts have now entered a high-throughput phase with the systematic determination of patterns of gene expression [e.g., 7]. As a result, the amount of data currently available has doubled leading to the imminent availability of multiple expression patterns of every gene in the *Drosophila* genome [7]. In addition, the use of micro-array technology to study *Drosophila* development has revealed additional and important insights into changes in gene expression levels over time and under different conditions at a genomic scale [8, 9].

With this rapid increase in the amount of available primary gene expression images, searchable textual descriptions of images have become available [7, 10, 11]. However, a direct comparison of the gene expression patterns depicted in the images is also desirable to find biologically similar expression patterns, because textual descriptions (even using a highly structured and controlled vocabulary) cannot fully capture all aspects of an expression pattern. In fact, there is a need for automated identification of images containing overlapping or similar gene expression patterns [6, 12] in order to assist researchers in the evaluation of similarity between a given expression pattern and all other existing (comparable) patterns in the same way that the BLAST [13] technique functions for DNA and protein sequences. Of course, unlike the genomes with four letters and proteomes with 20 letters, all gene expression anatomies cannot be easily reduced to, and thus represented by, a small number of components.

We previously proposed a binary coded bit stream pattern to represent gene expression pattern images [6]. In this digital representation, referred to as the Binary Feature Vector (BFV; BSV in [6]), the unstained pixels in the images (white regions and background) were denoted by a value of 0 and the stained areas (colored and foreground: gene expression) were denoted by a value of 1. Based on the BFV representations of the expression pattern, we proposed a Basic Expression Search Tool for Images (BESTi) [6] with an aim to produce biologically significant gene expression pattern matches using image content alone, without any reference to textual descriptions. We found that the BESTi approach generated biologically meaningful matches to query expression patterns [6].

In this paper, we explore how a more sophisticated Invariant Moment Vectors (IMV, [14]) based digital representation of gene expression patterns performs in generating an ordered list of best-matching images that contain similar/overlapping gene expression patterns to that depicted in a query image. IMV are frequently used in natural image processing (e.g., optical character recognition [15]) and have a number of desirable properties, including the compensation for variations of scale, translation, and rotation. If successful, IMV representations hold the promise of producing significantly shorter computing times for image-to-image matching compared to BFV.

Previously, we had examined the performance of the BFV representation for a limited dataset of early stage images [6]. Here we compare the relative performances of BFV and IMV first using a dataset containing 226 images (from 13 research papers). Then we test for scalability of the BESTi search by using a seven times larger dataset containing 1819 (1593 new + 226 previous) images from 262 additional research papers (list available upon request from the authors). Both datasets contained lateral views of early stage (1-8) embryos.

During these investigations, we also developed another measure of image-to-image similarity for the BFV representation. This measure is aimed at finding images that contain as much of the query image expression pattern as possible, but without penalizing for the presence of any expression outside the overlap region in the target image. In

addition, we examined whether partitioning a multi-domain expression pattern into multiple BFV representations, each containing only one domain, yields a better result set.

Recently, Peng and Myers [16] have proposed a different procedure involving the global and local Gaussian Mixture Model (GMM) of the pixel intensities (of expression) to identify images with similar patterns. This GMM method is expected to find images with intensity and spatial similarities. This is different from the BFV and IMV methods examined here, which are intended to find only spatially similar patterns. This focus is important because, as mentioned in [6], the differences in gene expression intensity among images in published literature can arise simply due to use of different techniques, illumination conditions, or biological reasons. However, Peng and Myers method [16] appears to be promising and we plan to examine its effectiveness in a separate paper.

Results and Discussion

Data set generation

An image database of 226 gene expression pattern images was initially generated using data from the literature [17-29]. All were lateral images and exhibited early stage (1-8) expression patterns. These images were selected because they had some commonality of gene expression (as seen by the human eye), which allowed us to evaluate the performance of the BESTi in finding correct as well as false matches under controlled conditions. BESTi was also tested for scalability on a larger dataset containing 1819 (1593 plus the 226) lateral views of early stage embryos. These 1593 images were obtained from 262 articles.

In order to present comprehensible result sets in this paper, we have primarily discussed the findings from the dataset of 226 and provided information on how those queries scaled when they were conducted for the larger dataset. In general, our focus was to show the retrieval of biologically significant matches based on both the visual overlap of the spatial gene expression pattern and the genes associated with the pattern retrieved.

Each image was standardized and the binary expression pattern extracted following the procedures described previously [6]. These extracted patterns, their invariant moments

(ϕ_1 through ϕ_7), and binary feature representations were stored in a database. We also calculated and stored the expression area (the count of the number of 1's in the binary feature represented image), the X and Y coordinates of the centroid $(\bar{x}, \bar{y})$, and the principal angle (θ) for each extracted pattern.

To quantify the similarity of gene expressions in two images, we computed two measures (S_S, S_C) based on the BFV representation (See equations 2 and 3 in **Methods**). S_S is designed to find gene expression patterns with overall similarity to the query image, whereas S_C is for finding images that contain as much of the query image expression pattern as possible without penalizing for the presence of any expression outside the overlap region in the target image. For a given pair of gene expression patterns (A and B), S_S is the same irrespective of which image in the pair is the query image. That is, $S_S(A,B) = S_S(B,A)$. This is not so for S_C, because S_C measures how much of the query gene expression pattern is contained in the image. Therefore, $S_C(A,B) \neq S_C(B,A)$.

For IMV representation, we computed one dissimilarity measure (D_ϕ, equation 13 in **Methods**). Results from D_ϕ should be compared to that from S_S, as both of these measurements do not depend on the reference image, *i.e.*, $D_\phi(A,B) = D_\phi(B,A)$ and, also they capture overall similarity or dissimilarity.

Matches and their biological significance

The effectiveness of the BESTi in finding biologically similar expression patterns was geared towards determining the biological validity of the results obtained from the image matching procedure. All results were based solely on quantitative similarities between images without using any textual descriptions. All images were lateral views from the early stages of fruit fly embryogenesis and were oriented anterior end to the left and dorsal to the top. We refer to the images retrieved as the BESTi-matches.

Performance of BFV-S_S search: Figure 1A shows the query image with gene expression restricted to the anterior (left) portion of the embryo, except that the expression is absent at the anterior terminus [22]. The query image depicts the expression of the *sloppy paired* (*slp1*) gene in a wildtype embryo. The BESTi-matches based on the S_S measure

for the representations are given in Figure 1A1-A8. BESTi retrieves images showing similar expression patterns, all of which are from same research article as the query image [22]. These images depict the expression patterns of *sloppy paired* genes (*slp1* and *slp2*) in a variety of genetic backgrounds or in combination with a head gap gene *orthodentical* (*otd*); all of these genes are essential for the pattern formation in *Drosophila* head development [30]. In fact, *slp1* and *slp2* are tightly linked genes found in the *slp* locus of the *Drosophila* genome. They are not only closely related in their primary sequence structure, but also significantly similar in their expression pattern (compare Figure 1A7 and 1A8).

A search was conducted using the same query image and same distance measure (S_S) on the larger dataset. Figure 2 shows the top-35 matches, which contain all 8 matches shown in Figure 1A (images with blue colored legends). This allowed us to directly compare the quality of matches between the two datasets. Analysis of larger database of images yields more matches for the same S_S cut-off value, as expected. A visual inspection reveals that these are all relevant images (Figure 2), with the larger dataset yielding more images for *otd* (20 images, Figure 2C). Images with expression patterns from *slp1*, *slp2* and combined *otd* expression are found in Figure 2A, B and D. More importantly, searches in the larger dataset provide images containing expression patterns of additional genes: *Kruppel* (Kr), *hunchback* (*hb*), *bicoid* (*bcd*), *nanos*, *snail*, *hu-li tai shao* (*hts*) and *hairy* (Figure 2 E-K). Since these images did not exist in the smaller dataset, they were not included in the search results in Figure 1A. All are biologically useful matches because combinatorial input from gap genes (Kr, *hb*) along with *slp1* establishes the domains of segment polarity genes in the head [22]. As for the *snail*, *hts* and *hairy* genes, there are no known interaction between them and *slp1* (gene in the query image) in the wildtype embryo, but the images show overlap in gene expression due to the genetic backgrounds used [31-33]. Therefore, they are also biologically relevant matches.

Performance of IMV search: We used the same query image for the IMV method applied to the smaller dataset (D_ϕ, results in Figure 1B) and compared the results to the BFV- S_S search. In this case, we obtain images containing expressions of *hb*, Kr, *tailless* (*tll*),

slp1, *hairy* and *infra-abdominal* (*iab*) (type I transcript). It is clear that IMV search produces some biologically disconnected matches. For example, Figures 1B2, 1B4-B7 exhibit no visual overlap in gene expression pattern with the query. Furthermore, even the biologically significant matches were retrieved out of order (Figure 1B1 before 1B3). This happens because D_ϕ retrieves expression patterns that are of similar shape and/or size, regardless of the translation or rotation with respect to the query image.

A comparison of the results from the smaller and larger dataset for the IMV measure is given in Figure 3. Twenty-six images were retrieved from the larger dataset when we used the same maximum distance value for the same query image. Of these, only two images were with expression pattern from *slp1* (Figure 3 A1-A2). The expression of *bcd* was found in two of the results (Figures 3 B1-B2). 13 images containing gap gene expression patterns of Kr, *hb*, *tll*, *giant* (*gt*) and *knirps* (*kni*) (Figures 3 C1-C4, D1-D3, E1-E2, F1-F2, I1 and J1) were also retrieved. Images with expression patterns of *hairy*, *achaete-scute* complex (AS-C), *iab* (type I transcript), IAB5 enhancer, *ventral nervous system defective* (*vnd*), *short gastrulation* (*sog*) and a combined expression of *bcd*, *nanos* and *cap 'n' collar* (*cnc*) accounted for the remaining nine (Figures 3 G1-G2, H1-H2, K1, L1, M1, N1 and O1). We see that the new results also suffer from the same problems as before. For example, images in Figure 3 C, E, K and L have no common expression pattern with the query image. Hence these are not biologically significant results even though few of them (Figures 3 C1-C4, E1-E2) contain expression patterns of developmentally connected genes (Kr and *tll* with *slp1*).

Since both *Ss* and D_ϕ measures capture the overall similarity or dissimilarity, we can use Figures 2 and 3 to compare the relative effectiveness of the BFV and IMV methods on the larger dataset. We clearly see that the BFV method performs much better in retrieving both overlapping and similar expression patterns that are also biologically significant.

In addition to the Hu moments, one could also compute Zernike moments, which are based on the polar coordinate system. Both Hu moments and Zernike moments are susceptible to the same problem namely expression patterns showing a similar shape but translated to different locations in the embryo would be in the same result set. We chose

to study the Hu Invariant Moment Vectors mainly because the centroid of the image can be used to distinguish between similarly shaped but translated expression patterns. With Zernike moments, the image must be inherently contained within a unit circle anchored at the centroid [34]. Thus, there is no straightforward method to eliminate the translational problem.

Using the Hu moments, the spatial location problem can be corrected by considering the Euclidean difference in the centroid location expressed in pixels (ΔC_{XY}) of the query and results. In the case of BFV-S_S search results in Figure 1 (A1-A8), the maximum ΔC_{XY} is less than or only slightly greater than the minimum ΔC_{XY} for the IMV search results (Figure 1 B1-B8). Therefore, in the present case, the IMV-based BESTi search results need to be pared down using the centroid location difference. For example, if we consider results based on a ΔC_{XY} lesser than or equal to 50 pixels, images shown in Figure 1 B2, B4-B7 would be removed producing a more meaningful result set.

Performance of BFV-S_C search: Figure 1C shows the result for the same query image as used in Figure 1A, but using the newly devised S_C distance for the BFV representation (BFV-S_C search). This is expected to retrieve images with gene expression patterns that contain the largest amount of the overlap with the expression pattern in the query image. The top eight hits shown (Figure 1C1-C8) all contain over 93% of the query expression pattern: five of the matches are to the expression of *hunchback* (*hb*; C1, C3-C6) and the remaining three are from *slp1* under different genetic backgrounds. As mentioned above, the combinatorial input from gap genes (including *hb*) along with *slp1* establishes the domains of segment polarity genes in the head [22]. Therefore, gene expression patterns found by BFV- S_C search are for developmentally connected genes. However, using the same query image, BFV-S_C search yielded only two images in common with the BFV-S_S results (Figure 1; C7 and C8 are the same as A5 and A4, respectively). This difference occurs because S_S is designed to find gene expression patterns with overall similarity to the query image (Figure 1A), whereas S_C is intended for finding images that contain as much of the query image expression pattern as possible and exclusive of the presence of the gene expression in the result image outside the region of overlap with the query

image. Therefore, BFV-S_S and BFV-S_C have the capability of finding gene expression patterns from different biological perspectives.

Using the same minimum similarity value for the BFV-S_C in the larger dataset resulted in 55 images, given in Figure 4. Gene expression patterns of *slp1* and *otd* accounted for 8 of these images (Figure 4A and 4B). 22 images contained expression patterns of the various gap genes *hb*, Kr, *kni* and *tll* (Figure 4C, 4E-F, 4I-L) that were co-expressed with *bcd* and *nanos* (Figure 4E and 4J) or with *en* (Figure 4I). Five other genes, developmentally connected to the gene, *slp1*, in the query image were also retrieved in this result set (*eve*, *twist, dpp* (*decapentaplegic*) [35]; *en* (*engrailed*) [36]; *arm* (*armadillo*) [37]; Figure 4M-Q). These images were not found in the top-35 of S_S result set, which accentuates the different capabilities of the two BFV similarity measures in retrieving biologically relevant matches. The remaining images had expression patterns of AS-C, *sc* (s*cute*), *snail*, *hairy*, *zen* (*zerknullt*), *run*, Hsp83, *nmo* (*nemo*), Tc'hb, *iab*, *hts* and *sog* (Figure 4D, 4G-H, 4R-Z) which are not known to be directly related to the gene *slp1*. All but seven of these images (Figures 4 D3-D4, H1-H2, R1, X1 and Y1) were from a different developmental stage than the query image. Hence, by limiting the results to those from a specific stage, extraneous matches can be removed. The seven images having the same stage as the query image were retrieved because of their significant overlap (more than 94%) with the query gene expression pattern. Thus, we observe that the new distance measure S_C has the potential to identify images containing expression patterns of developmentally connected genes, other than those retrieved by S_S, thus improving the overall performance of the BFV method and the BESTi tool.

Analysis of multi-domain gene expression patterns

Due to the presence of multiple areas of expression, some patterns in the database that appeared to contain much better matches (by eye and biologically) to the query image were not found or ranked very high. Hence, we also analyzed multi-domain expression patterns separately for the smaller dataset. Developmental biologists are also interested in finding such patterns as they contain overlaps with the expression domains in the query image. In fact, a large number of the expression patterns available today contain multiple isolated domains of expressions since more than one topologically distinct region of

expression may be produced by many genes, transgenic constructs, probes or experimental techniques (multiple staining). In such cases, we need to consider each of these regions individually as well as in the context of the composite pattern. Biologically, it is important to consider them separately because different regions of expression may be under the control of distinct *cis*-regulatory sequences [e.g., 28, 38] or may represent the expression of different genes in a multiply-stained embryo.

Separating multi-domain gene expression patterns into individual components was straightforward; we simply generated multiple images from the same initial image and included them in the target dataset. This resulted in 192 additional images (418 total) in the database all of which were components of the initial gene expression patterns. The images were separated into expression regions horizontally and/or vertically depending on the gene expression. For this new set of images, the IMV as well as BFV representations were re-calculated and the BESTi query constructed as above.

Results from BFV-S_S and IMV queries for this data set are given in Figures 1D and 1E, respectively. Now, many images with multiple regions of expression are retrieved in the result set (Figure 1D: D1-D8) and many of them show an even better match with the query pattern than those in Figure 1A for the BFV-based BESTi search. For instance, gene expression patterns are now retrieved (with more than 55% pattern similarity) from embryos with the expression of *tailless* (*tll*), which is known to interact with *slp1* in defining the embryonic head [22], and with a composite expression of *race (related to angiotensin converting enzyme)*, *sog* (*short gastrulation*) and *eve* (*even-skipped*) due to enhanced *race* expression in the anterior domain caused by a transgenic construct causing ectopic expression of *sog*[19]. Therefore, the strategy of dividing multi-domain expression data into individual domains provides additional flexibility to query individual components or sub-sets of complex expression patterns. Results also improved for IMV (Figure 1E), but again the outcome reinforced the need to use the difference in centroid to limit the result set.

Next we examine the performance of S_S, S_C and D_ϕ in finding BESTi matches for a query pattern with multiple regions of expression (Figure 5A). This complex expression pattern

consists of anterior and posterior domains caused by enhanced *race* expression resulting from dosage alteration of *dpp* in a *gastrulation defective* (*gd*) mutant background, and a middle stripe due to misexpressed *sog* using an *eve* stripe-2 enhancer [Figure 2d in 19]. The results from this query are shown in Figure 5A1-A8 (only the original image set (226) was used as the target database in this case). We again find that S_S finds many images from the same paper as well as some images from other research articles with similar expression patterns. The results correctly include expression pattern of *eve* (Figure 5A4), of another pair-rule gene (*ftz*: *fushi tarazu*; Figure 5A6), and of two other developmentally related genes [39, 40].

When D_ϕ is used as a search criterion, it produces some correct matches in the result set (Figure 5B1-B8). However, it generally fails to rank biologically meaningful matches as the best matches. Use of the centroid in this case is also not productive, as most of the matches show very close centroids. The principal angle (θ) value calculated does not show a significant difference in the early stage embryos used in this study. The results using the S_C based search are given in Figure 5C1-C8. They show a number of images in common with the S_S results. However, as expected, there are significant differences between the two searches.

The results in Figures 5D and 5E demonstrate the power of the BESTi-search when the multi-domain expression data are represented in their component patterns (domain database). In this case, all the BESTi searches are based on the use of S_S as the search criterion. These searches are based on the complete expression (Figure 5D) and on one of its components (bottom-left domain, Figure 5E). All, but one, BESTi-matches in Figure 5D contain both domains of expression. In contrast, the use of only the left, anterior, domain (Figure 5E) in the BESTi search produces many other images in which the gene expression pattern is similar to only the anterior-ventral query pattern. Therefore, the use of individual expression components as search arguments increases the potential of directly identifying different overlapping expression patterns.

Conclusions

We have found that it is possible to identify biologically significant gene expression patterns from a dataset by first extracting numeric signature descriptors and then using those descriptors in a computerized search of the database for expression patterns with similar signatures or maximum pattern similarities. We find that the BFV methodologies provide a longer and more biologically meaningful set of expression pattern matches than IMV. Even though IMV representations will produce much faster retrieval speeds for large collections of embryogenesis images, the lack of biological validity of BESTi-matches retrieved makes IMV undesirable for the present problem. Instead, investigations and strategies aimed at improving the real time performance of the BFV representation will better serve the developmental biological research.

Methods

The wide variety of input methodologies, illumination conditions, equipment, and publication venues involved in the acquisition and presentation of gene expression patterns makes the available gene expression pattern data rather diverse. Extracting a gene expression pattern from its background requires the use of a combination of manual and automatic techniques. Each image is first standardized into a binary image as described in [6]. The standardized images are then represented using the Binary Feature Vector (BFV) [6], and the Invariant Moment Vectors (IMV) [14]. Similarity measures S_S and S_C are derived from BFV of which, S_S is the one's complement of the distance metric D_E presented in [6] and S_C is a new measure introduced in this paper. The third metric D_ϕ is deduced from the invariant moment vectors.

Binary Sequence Vector Analysis

The binary coded bit stream pattern, in which the two possible states indicate staining over or under a threshold value, is called as Binary Feature Vector (BFV). This is referred to as the Binary Sequence Vector (BSV) in [6]. In other words, we represent each image as a sequence of 1's and 0's, where the black pixels (stained areas) are denoted by a value of 1 and the white pixels (unstained and background) are denoted by a

value of 0. This BFV holds the gene expression and localization pattern information of each image.

The expression patterns are ordered by evaluating a set of difference values, D_E, between the binary feature vectors of every possible pair of images in the dataset. D_E was introduced in [6] and is formally given as,

$$D_E = Count(\text{A XOR B})/Count(\text{A OR B}) \quad (1)$$

The term *Count*(A XOR B) corresponds to the number of pixels not spatially common to the two images and the term *Count*(A OR B) provides the normalizing factor, as it refers to the total number of stained pixels (expression area) depicted in either of the two images being compared. For simplicity, we use the one's complement of D_E, as a measure of similarity of gene expression patterns between two images, S_S, is given by the equation

$$S_S = (1 - D_E). \quad (2)$$

S_S quantifies the amount of similarity based on the overlap between two expression patterns. S_S is equal to 1 when the two expression patterns are identical ($D_E = 0$).

We introduce a new similarity measure in this paper that does not penalize for any non-overlapping region. The measure S_C quantifies the amount of similarity based on the containment of one expression pattern in the other given by

$$S_C = Count(\text{A AND B})/Count\ (\text{A}) \quad (3)$$

If the entire query image is contained within the result set images found in the database, *i.e.*, there is complete overlap (with respect to the query image) S_C is equal to 1. Note that, $S_C(A,B) \neq S_C(B,A)$, because the denominator corresponds to the gene expression area of the query image.

Invariant Moment Vector (IMV) Analysis

Some methodologies of image analysis produce numeric descriptors that compensate for variations of scale, translation and rotation. In the following section, we describe the

invariant moment analysis of gene expression data. Invariant moment calculations have been used in optical character recognition and other applications for many years [15].

To calculate these invariant moment descriptors the standardized binary image [6] is converted to a binary representation of the same pattern (BFV). From this binary sequence of the image, the invariant moments and other descriptors are extracted using the following method [14, 41]. The continuous scale equation used is

$$M_{pq} = \iint x^p y^q f(x, y) dxdy \,, \tag{4}$$

where M_{pq} is the two-dimensional moment of the function of the gene expression pattern, $f(x, y)$. The order of the moment is defined as $(p+q)$, where both p and q are positive natural numbers. When implemented in a digital or discrete form this equation becomes

$$M_{pq} = \sum_x \sum_y x^p y^q f(x, y) \,. \tag{5}$$

We then normalize for image translation using $\overline{x}$ and $\overline{y}$ which are the coordinates of the center of gravity, centroid, of the area showing expression. They are calculated as

$$\overline{x} = \frac{M_{10}}{M_{00}} \text{ and } \overline{y} = \frac{M_{01}}{M_{00}} \,. \tag{6}$$

Discrete representations of the central moments are then defined as follows:

$$\mu_{pq} = \sum_x \sum_y (x - \overline{x})^p (y - \overline{y})^q f(x, y) \tag{7}$$

A further normalization for variations in scale can be implemented using the formula,

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{8}$$

and $\gamma = \frac{p+q}{2} + 1$ is the normalization factor. From the central moments, the following values are calculated:

$$\phi_1 = \eta_{20} + \eta_{02}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (\eta_{03} - 3\eta_{21})^2$$

$$\phi_4 = (\eta_{30} - \eta_{12})^2 + (\eta_{03} + \eta_{21})^2$$

$$\begin{aligned}\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ &+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned}$$

$$\begin{aligned}\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ &+ 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\end{aligned}$$

$$\begin{aligned}\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ &+ (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{30})^2]\end{aligned} \qquad (9)$$

where ϕ_7 is a skew invariant to distinguish mirror images. In the above, ϕ_1 and ϕ_2 are second order moments and ϕ_3 through ϕ_7 are third order moments. ϕ_1 (the sum of the second order moments) may be thought of as the "spread" of the gene expression pattern; whereas the square root of ϕ_2 (the difference of the second order moments) may be interpreted as the "slenderness" of the pattern. Moments ϕ_3 through ϕ_7 do not have any direct physical meaning, but include the spatial frequencies and ranges of the image.

In order to provide a discriminator for image inversion (and rotation), sometimes called the "6", "9" problem, it has been suggested [14, 42] that the principal angle be used to determine "which way is up". This is extremely important in embryo images because gene expression at the anterior and posterior regions may simply appear to be mirror images of each other to the invariant moments, but biologically they are completely distinct. The principal axis of the gene expression pattern $f(x, y)$ is the angular

displacement of the minimum rotational inertia line that passes through the centroid $(\bar{x}, \bar{y})$ and is given as:

$$\sum\sum[(x-\bar{x})\sin\theta-(y-\bar{y})\cos\theta)]^2 f(x,y)=0. \qquad (10)$$

The slope of the principal axis is called the principal angle θ. It is calculated knowing that the moment of inertia of f around the line $(y-\bar{y})\cos\theta=(x-\bar{x})\sin\theta$ is a line through $(\bar{x}, \bar{y})$ with slope θ. We can find the θ value at which the momentum is minimum by differentiating this equation with respect to θ and setting the results equal to zero. This produces the following equation:

$$\theta=\frac{1}{2}\tan^{-1}\frac{2\eta_{11}}{\eta_{20}-\eta_{02}} \qquad (11)$$

Using the condition $|\theta|<45^0$ one can distinguish the "6" from the "9" and rotationally similar gene expression patterns.

In invariant moment analysis, our initial method of image comparison calculates the Euclidean distance between the images using all moments (ϕ_1 through ϕ_7) and combinations of these moments. For example, if the first two invariant moments are used, then

$$X=\eta_{20}+\eta_{02} \quad \text{and} \quad Y=\sqrt{(\eta_{20}-\eta_{02})^2+4\eta_{11}^2} \qquad (12)$$

and the distance D_{ij}, between a pair of images *i* and *j* where *i*, *j* = 1, 2,…n is given by

$$D_{ij}=\sqrt{(X_i-X_j)^2+(Y_i-Y_j)^2}\,.$$

This can be expanded to use all of the moment variables. Here, the Euclidean distance, D_{ϕ}, between any two images is calculated as

$$D_{\phi} = \left[\sum_{j=1}^{7} \left| x_{ij} - x_{qj} \right|^2 \right]^{1/2} \qquad (13)$$

where i and q designate images whose distance is being calculated and j designates the parameters used in the distance calculation and $j = 1, 2, \ldots, 7$. This assumes that all moments have the same dimensions or that they are dimensionless.

Using this method, it is possible to rank each of the images in order of their similarity based on, for example, the first two invariant moments that have clear-cut physical meanings. Expansion to include additional moments or parameters can be performed in a number of ways. It is possible to add additional parameters to the distance calculation making sure that each of the parameters has the same dimension. For example, ϕ_1 has the dimension of distance squared, while ϕ_2 has the dimension of the fourth power of distance, thus requiring the square root function to equalize dimensions for comparable distance calculation purposes. In general, the greater number of invariant moments used in the distance calculation, the more selective the ranking. We have also allowed for the use of the centroids and principal angle as a means of list limiting.

Authors' contributions

SK originally conceived the project, developed the image distance measures based on the BFV representation, wrote an early version of the manuscript, and edited it until the final version. RG was responsible for writing new and using pre-existing programs to perform the image distance and parameter calculations, helped prepare the figures, searched the literature for gene expression data, maintained the database of gene expression pattern images, and helped in writing the manuscript. BVE provided the IMV method description, managed the day-to-day activities in the project, and did significant editing to produce the manuscript in the desired format for the journal. SP originally proposed the use of invariant moment vectors for biological image analysis, contributed significantly for the image distance and parameter calculations and provided critical feedback during the later stages of revision.

Acknowledgements

We thank Dr. Robert Wisotzkey for biological remarks, Dr. Dana Desonie for editorial comments and Dr. Stuart Newfeld for useful suggestions. This research was supported in part by research grants from National Institutes of Health (S.K.) and the Center for Evolutionary Functional Genomics (S.K.) at the Arizona State University.

References

1. Carroll SB, Grenier JK, Weatherbee SD: **From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design.** Massachusetts, MA: Blackwell Scientific; 2000.
2. Davidson E: **Genomic Regulatory Systems: Development and Evolution**. New York, NY: Academic Press; 2000.
3. Rougvie AE: **Control of developmental timing in animals**. *Nat Rev Genet* 2001, **2**(9):690-701.
4. Gieseler K, Wilder E, Mariol MC, Buratovitch M, Berenger H, Graba Y, Pradel J: **DWnt4 and wingless elicit similar cellular responses during imaginal development**. *Dev Biol* 2001, **232**(2):339-350.
5. Takaesu NT, Johnson AN, Sultani OH, Newfeld SJ: **Combinatorial Signaling by an Unconventional Wg Pathway and the Dpp Pathway Requires Nejire (CBP/p300) to Regulate dpp Expression in Posterior Tracheal Branches**. *Dev Biol* 2002, **247**(2):225-236.
6. Kumar S, Jayaraman K, Panchanathan S, Gurunathan R, Marti-Subirana A, Newfeld SJ: **BEST: A novel computational approach for comparing gene expression patterns from early stages of *Drosophila melanogaster* development**. *Genetics* 2002, **162**(4):2037-2047.
7. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, Rubin GM: **Systematic determination of patterns of gene expression during *Drosophila* embryogenesis**. *Genome Biol* 2002, **3**(12):research0088.0081-0088.0014.
8. Montalta-He H, Reichert H: **Impressive expressions: developing a systematic database of gene-expression patterns in *Drosophila* embryogenesis**. *Genome Biol* 2003, **4**(2):205.
9. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP: **Gene expression during the life cycle of *Drosophila melanogaster***. *Science* 2002, **297**(5590):2270-2275.
10. FlyBase: **The FlyBase database of the Drosophila genome projects and community literature**. *Nucleic Acids Res* 1999, **27**:85-88.
11. Janning W: **FlyView, a Drosophila image database, and other Drosophila databases**. *Seminars in Cell and Developmental Biology* 1997, **8**(5):469-475.
12. Bard JBI: **Introduction: Making and filling gene-expression developmental databases**. *Seminars in Cell and Developmental Biology* 1997, **8**(5):455-458.
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
14. Hu M-K: **Visual pattern recognition by moment invariants**. *IRE Transactions of Information Theory* 1962:179-187.
15. Castleman KR: **Digital Image Processing**. New Jersey: Prentice Hall; 1996.
16. Peng H, Myers EW: **Comparing in situ mRNA expression patterns of Drosphila embryos**. In: *Proceedings of RECOMB: 2004; San Diego, CA*: ACM Journals; 2004.

17. Arnosti DN, Gray S, Barolo S, Zhou J, Levine M: **The gap protein knirps mediates both quenching and direct repression in the *Drosophila* embryo**. *Embo J* 1996, **15**(14):3659-3666.
18. La Rosee-Borggreve A, Hader T, Wainwright D, Sauer F, Jackle H: ***hairy* stripe 7 element mediates activation and repression in response to different domains and levels of Kruppel in the *Drosophila* embryo**. *Mech Dev* 1999, **89**(1-2):133-140.
19. Ashe HL, Levine M: **Local inhibition and long-range enhancement of Dpp signal transduction by Sog**. *Nature* 1999, **398**:427-431.
20. Casares F, Sanchez-Herrero E: **Regulation of the infraabdominal regions of the bithorax complex of *Drosophila* by gap genes**. *Development* 1995, **121**(6):1855-1866.
21. Goldstein RE, Jimenez G, Cook O, Gur D, Paroush Z: **Huckebein repressor activity in *Drosophila* terminal patterning is mediated by Groucho**. *Development* 1999, **126**:3747-3755.
22. Grossniklaus U, Cadigan KM, Gehring WJ: **Three maternal coordinate systems cooperate in the patterning of the *Drosophila* head**. *Development* 1994, **120**(11):3155-3171.
23. Gutjahr T, Frei E, Noll M: **Complex regulation of early *paired* expression: initial activation by gap genes and pattern modulation by pair-rule genes**. *Development* 1993, **117**(2):609-623.
24. Hartmann C, Taubert H, Jackle H, Pankratz MJ: **A two-step mode of stripe formation in the *Drosophila* blastoderm requires interactions among primary pair rule genes**. *Mech Dev* 1994, **45**(1):3-13.
25. Hulskamp M, Pfeifle C, Tautz D: **A morphogenetic gradient of *hunchback* protein organizes the expression of the gap genes *Kruppel* and *knirps* in the early *Drosophila* embryo**. *Nature* 1990, **346**(6284):577-580.
26. Hulskamp M, Tautz D: **Gap genes and gradients - the logic behind the gaps**. *BioEssays* 1991, **13**:261-268.
27. Hulskamp M, Lukowitz W, Beermann A, Glaser G, Tautz D: **Differential regulation of target genes by different alleles of the segmentation gene *hunchback* in Drosophila**. *Genetics* 1994, **138**(1):125-134.
28. Gaul U, Jackle H: **Role of gap genes in early Drosophila development**. *Adv Genet* 1990, **27**:239-275.
29. Gaul U, Jackle H: **Pole region-dependent repression of the *Drosophila* gap gene *kruppel* by maternal gene products**. *Cell* 1987, **51**:549-555.
30. Royet J, Finkelstein R: **Pattern formation in Drosophila head development: the role of the orthodenticle homeobox gene**. *Development* 1995, **121**(11):3561-3572.
31. Stathopoulos A, Levine M: **Linear signaling in the Toll-Dorsal pathway of Drosophila: activated Pelle kinase specifies all threshold outputs of gene expression while the bHLH protein Twist specifies a subset**. *Development* 2002, **129**(14):3411-3419.
32. Brent AE, MacQueen A, Hazelrigg T: **The Drosophila wispy gene is required for RNA localization and other microtubule-based events of meiosis and early embryogenesis**. *Genetics* 2000, **154**(4):1649-1662.

33. Zhang H, Levine M: **Groucho and dCtBP mediate separate pathways of transcriptional repression in the Drosophila embryo**. *Proc Natl Acad Sci U S A* 1999, **96**(2):535-540.
34. Teh C, Chin R: **On Image Analysis by the Methods of Moments**. *IEEE Transactions on Patterns Analysis and Machine Intelligence* 1988, **10**(4):496-513.
35. Riechmann V, Irion U, Wilson R, Grosskortenhaus R, Leptin M: **Control of cell fates and segmentation in the Drosophila mesoderm**. *Development* 1997, **124**(15):2915-2922.
36. Cadigan KM, Grossniklaus U, Gehring WJ: **Localized expression of sloppy paired protein maintains the polarity of Drosophila parasegments**. *Genes Dev* 1994, **8**(8):899-913.
37. Bhat KM, van Beers EH, Bhat P: **Sloppy paired acts as the downstream target of wingless in the Drosophila CNS and interaction between sloppy paired and gooseberry inhibits sloppy paired during neurogenesis**. *Development* 2000, **127**(3):655-665.
38. Sanchez L, Thieffry D: **A logical analysis of the *Drosophila* gap-gene system**. *J Theor Biol* 2001, **211**(2):115-141.
39. Frasch M, Warrior R, Tugwood J, Levine M: **Molecular analysis of even-skipped mutants in Drosophila development**. *Genes Dev* 1988, **2**(12B):1824-1838.
40. Abbott MK, Kaufman TC: **The relationship between the functional complexity and the molecular organization of the Antennapedia locus of Drosophila melanogaster**. *Genetics* 1986, **114**(3):919-942.
41. Jayaraman K, Panchanathan S, Kumar S: **Classification and indexing of gene expression images**. *Proceedings of Society of Photo-optical Instrumentation Engineers* 2001, **4472**:471-481.
42. Rosenfeld A, Kak AC: **Digital Picture Processing**, 2nd edn. New York: Academic Press; 1982.
43. Zhao C, York A, Yang F, Forsthoefel DJ, Dave V, Fu D, Zhang D, Corado MS, Small S, Seeger MA, Ma J: **The activity of the Drosophila morphogenetic protein Bicoid is inhibited by a domain located outside its homeodomain**. *Development* 2002, **129**(7):1669-1680.
44. Gao Q, Finkelstein R: **Targeting gene expression to the head: the Drosophila orthodenticle gene is a direct target of the Bicoid morphogen**. *Development* 1998, **125**(21):4185-4193.
45. Wimmer EA, Cohen SM, Jackle H, Desplan C: **buttonhead does not contribute to a combinatorial code proposed for Drosophila head development**. *Development* 1997, **124**(8):1509-1517.
46. Schulz C, Tautz D: **Autonomous concentration-dependent activation and repression of Kruppel by hunchback in the Drosophila embryo**. *Development* 1994, **120**(10):3043-3049.
47. Tsai C, Gergen JP: **Gap gene properties of the pair-rule gene runt during Drosophila segmentation**. *Development* 1994, **120**(6):1671-1683.
48. Janody F, Reischl J, Dostatni N: **Persistence of Hunchback in the terminal region of the Drosophila blastoderm embryo impairs anterior development**. *Development* 2000, **127**(8):1573-1582.

49. Sauer F, Wassarman DA, Rubin GM, Tjian R: **TAF(II)s mediate activation of transcription in the Drosophila embryo**. *Cell* 1996, **87**(7):1271-1284.
50. Strunk B, Struffi P, Wright K, Pabst B, Thomas J, Qin L, Arnosti DN: **Role of CtBP in transcriptional repression by the Drosophila giant protein**. *Dev Biol* 2001, **239**(2):229-240.
51. Colas JF, Launay JM, Vonesch JL, Hickel P, Maroteaux L: **Serotonin synchronises convergent extension of ectoderm with morphogenetic gastrulation movements in Drosophila**. *Mech Dev* 1999, **87**(1-2):77-91.
52. Wu X, Vasisht V, Kosman D, Reinitz J, Small S: **Thoracic patterning by the Drosophila gap gene hunchback**. *Dev Biol* 2001, **237**(1):79-92.
53. Ghiglione C, Perrimon N, Perkins LA: **Quantitative variations in the level of MAPK activity control patterning of the embryonic termini in Drosophila**. *Dev Biol* 1999, **205**(1):181-193.
54. Pankratz MJ, Busch M, Hoch M, Seifert E, Jackle H: **Spatial control of the gap gene knirps in the Drosophila embryo by posterior morphogen system**. *Science* 1992, **255**(5047):986-989.
55. Melnick MB, Perkins LA, Lee M, Ambrosio L, Perrimon N: **Developmental and molecular characterization of mutations in the Drosophila-raf serine/threonine protein kinase**. *Development* 1993, **118**(1):127-138.
56. Parkhurst SM, Lipshitz HD, Ish-Horowicz D: **achaete-scute feminizing activities and Drosophila sex determination**. *Development* 1993, **117**(2):737-749.
57. Zhou A, Hassel BA, Silverman RH: **Expression cloning of 2-5A-dependent RNAase: A uniquely regulated mediator of interferon action**. *Cell* 1993, **72**:753-765.
58. Niessing D, Dostatni N, Jackle H, Rivera-Pomar R: **Sequence interval within the PEST motif of Bicoid is important for translational repression of caudal mRNA in the anterior region of the Drosophila embryo**. *Embo J* 1999, **18**(7):1966-1973.
59. Yagi Y, Suzuki T, Hayashi S: **Interaction between Drosophila EGF receptor and vnd determines three dorsoventral domains of the neuroectoderm**. *Development* 1998, **125**(18):3625-3633.
60. Cowden J, Levine M: **The Snail repressor positions Notch signaling in the Drosophila embryo**. *Development* 2002, **129**(7):1785-1793.
61. Miskiewicz P, Morrissey D, Lan Y, Raj L, Kessler S, Fujioka M, Goto T, Weir M: **Both the paired domain and homeodomain are required for in vivo function of Drosophila Paired**. *Development* 1996, **122**(9):2709-2718.
62. Schulz C, Tautz D: **Zygotic caudal regulation by hunchback and its role in abdominal segment formation of the Drosophila embryo**. *Development* 1995, **121**(4):1023-1028.
63. Goff DJ, Nilson LA, Morisato D: **Establishment of dorsal-ventral polarity of the Drosophila egg requires capicua action in ovarian follicle cells**. *Development* 2001, **128**(22):4553-4562.
64. Sackerson C, Fujioka M, Goto T: **The even-skipped locus is contained in a 16-kb chromatin domain**. *Dev Biol* 1999, **211**(1):39-52.
65. Rusch J, Levine M: **Regulation of a dpp target gene in the Drosophila embryo**. *Development* 1997, **124**(2):303-311.

66. Steingrimsson E, Pignoni F, Liaw GJ, Lengyel JA: **Dual role of the Drosophila pattern gene tailless in embryonic termini**. *Science* 1991, **254**(5030):418-421.
67. Hamada F, Bienz M: **A Drosophila APC tumour suppressor homologue functions in cellular adhesion**. *Nat Cell Biol* 2002, **4**(3):208-213.
68. Klinger M, Soong J, Butler B, Gergen JP: **Disperse versus compact elements for the regulation of *runt* stripes in *Drosophila***. *Dev Biol* 1996, **177**:73-84.
69. Bashirullah A, Halsell SR, Cooperstock RL, Kloc M, Karaiskakis A, Fisher WW, Fu W, Hamilton JK, Etkin LD, Lipshitz HD: **Joint action of two RNA degradation pathways controls the timing of maternal transcript elimination at the midblastula transition in Drosophila melanogaster**. *Embo J* 1999, **18**(9):2610-2620.
70. Verheyen EM, Mirkovic I, MacLean SJ, Langmann C, Andrews BC, MacKinnon C: **The tissue polarity gene nemo carries out multiple roles in patterning during Drosophila development**. *Mech Dev* 2001, **101**(1-2):119-132.
71. Wolff C, Schroder R, Schulz C, Tautz D, Klingler M: **Regulation of the Tribolium homologues of caudal and hunchback in Drosophila: evidence for maternal gradient systems in a short germ embryo**. *Development* 1998, **125**(18):3645-3654.

Figures

Figure 1 - BESTi search results with smaller dataset

Results from the BESTi-search for the same query image [22] based on (A) BFV [S_S], (B) IMV [D_ϕ] and (C) BFV [S_C] representations in the original dataset (226 images); and based on (D) BFV [S_S] and (E) IMV [D_ϕ] representations in the domain database (in which distinct domains of the multi-domain expression patterns were added to the original dataset as additional data points). The search argument and the results retrieved are shown on the left and right of the arrow, respectively. The original data used to generate these expression patterns are shown above this row. BESTi-matches are arranged in descending order starting with the best hit for the given search image. Values of difference in centroids (ΔC_{XY}) and principal angles ($\Delta\theta$) are also given. Each image is identified by the last name of the first author of the original research article and the figure number with the following abbreviations: Ashe [19]; Casares [20]; Gaul1 [28]; Grossniklaus [22]; Hartmann [24]; Hulskamp1 [27]; Hulskamp3 [26].

Figure 2 - BESTi search results for S_S with larger dataset

Comparison of search results from the small (226 images) and large (1819 images) dataset using the S_S measure for the same query image (Figure 1A) [22]. Panels (A-K) are based on the genes whose expression patterns were retrieved as follows (A) *slp1*, (B) *slp1* and *otd*, (C) *otd*, (D) *slp2*, (E) Kr, (F) *hb*, (G) *hb* and *bcd*, (H) Hb, *bcd* and *nanos*, (I) *snail*, (J) *hts* and (K) *hairy*. Images are referenced with the last name of the first author of the original article and its figure number: Grossniklaus [22]; Zhao [43]; Gao [44]; Wimmer [45]; Schulz1 [46]; Tsai [47]; Janody [48]; Stathopoulos [31]; Brent [32]; Zhang [33]. Common search results between the small and large dataset are indicated with dark blue image names.

Figure 3 - BESTi search results for D_ϕ with larger dataset

Comparison of search results from the small (226 images) and large (1819 images) dataset using the D_ϕ measure for the same query image (Figure 1A) [22]. Panels (A-O) are based on the genes whose expression patterns were retrieved as follows (A) *slp1*, (B)

bcd, (C) Kr, (D) *hb*(D1,D3) and Hb(D2), (E) *tll*, (F) *gt*, (G) *hairy*, (H) AS-C, (I) *hb* and Kr, (J) *kni*, (K) *iab* (type I transcript), (L) IAB5 enhancer, (M) *vnd*, (N) *sog* and (O) *nanos*, *bcd* and *cnc*. Images are referenced with the last name of the first author of the original article and its figure number: Grossniklaus [22]; Sauer[49]; Tsai[47]; Hulskamp1[27]; Gaul1[28]; Strunk[50]; Colas[51]; Wu[52]; Ghiglione[53]; Pankratz[54]; Melnick[55]; Janody[48]; Zhang[33]; Parkhurst[56]; Zhou[57]; Stathopoulos[31]. Common search results between the small and large datasets are indicated with dark blue image names.

Figure 4 - BESTi search results for S_C with larger dataset

Comparison of search results from the small (226 images) and large (1819 images) dataset using the D_ϕ measure for the same query image (Figure 1A) [22]. Panels (A-Z) are based on the genes whose expression patterns were retrieved as follows (A) *slp1*, (B) *otd*, (C) *hb*, (D) AS-C, (E) *nanos*, *bcd* and Hb, (F) Kr, (G) *sc*, (H) *snail*, (I) *en* and *hb*, (J) *bcd* and *hb*, (K) *kni* and *hb*, (L) *tll*, (M) *eve*, (N) *twist*, (O) *dpp*, (P) *en*, (Q) *arm*, (R) *hairy*, (S) *zen*, (T) *run*, (U) Hsp83, (V) *nmo*, (W) Tc'hb, (X) *iab*, (Y) *hts* and (Z) *sog*. Images are referenced with the last name of the first author of the original article and its figure number: Grossniklaus [22]; Gao [44]; Hulskamp1[27]; Hulskamp3 [26]; Zhao [43]; Gaul1[28]; Tsai [47]; Niessing [58]; Sauer[49]; Parkhurst[56]; Janody[48]; Schulz2 [46]; Yagi [59] Cowden [60]; Stathopoulos[31]; Miskiewicz [61]; Schulz1 [62]; Goff [63]; Sackerson [64]; Rusch [65]; Steingrimsson [66]; Hamada [67]; Zhang[33]; Klingler [68]; Bashirullah [69]; Verheyen [70]; Wolff [71]; Casares [20]; Brent [32]. Common search results between the small and large dataset are indicated with dark blue image names.

Figure 5 - BESTi search results with multiple domains of expression using smaller database

Results from BESTi-search for a query image with multiple domains of expression. (A) BFV [S_S], (B) IMV [D_ϕ] and (C) BFV [S_C] searches for the same expression pattern in the original database (226 images). (D) BFV [S_S] search using the complete multi-domain expression in the original database and (E) BFV [S_S] search using only the pattern on the left in the domain database. Search argument and the results retrieved are shown on the left and right of the arrow, respectively. Original data used to generate these expression

patterns are shown above this row. BESTi-matches are arranged in descending order starting with the best hit for the given search statistic. Values of difference in centroids (ΔC_{XY}) and principal angles ($\Delta\theta$) are also given for panels A, B and C. Each image is identified by the last name of the first author of the original research article and the figure number; with the abbreviations as follows: Ashe [19]; Arnosti [17]; Borggreve [18]; Casares [20]; Gaul1 [28]; Gaul2 [29]; Grossniklaus [22]; Hartmann [24]; Hulskamp1 [27]; Hulskamp2 [25]; Hulskamp3 [26].

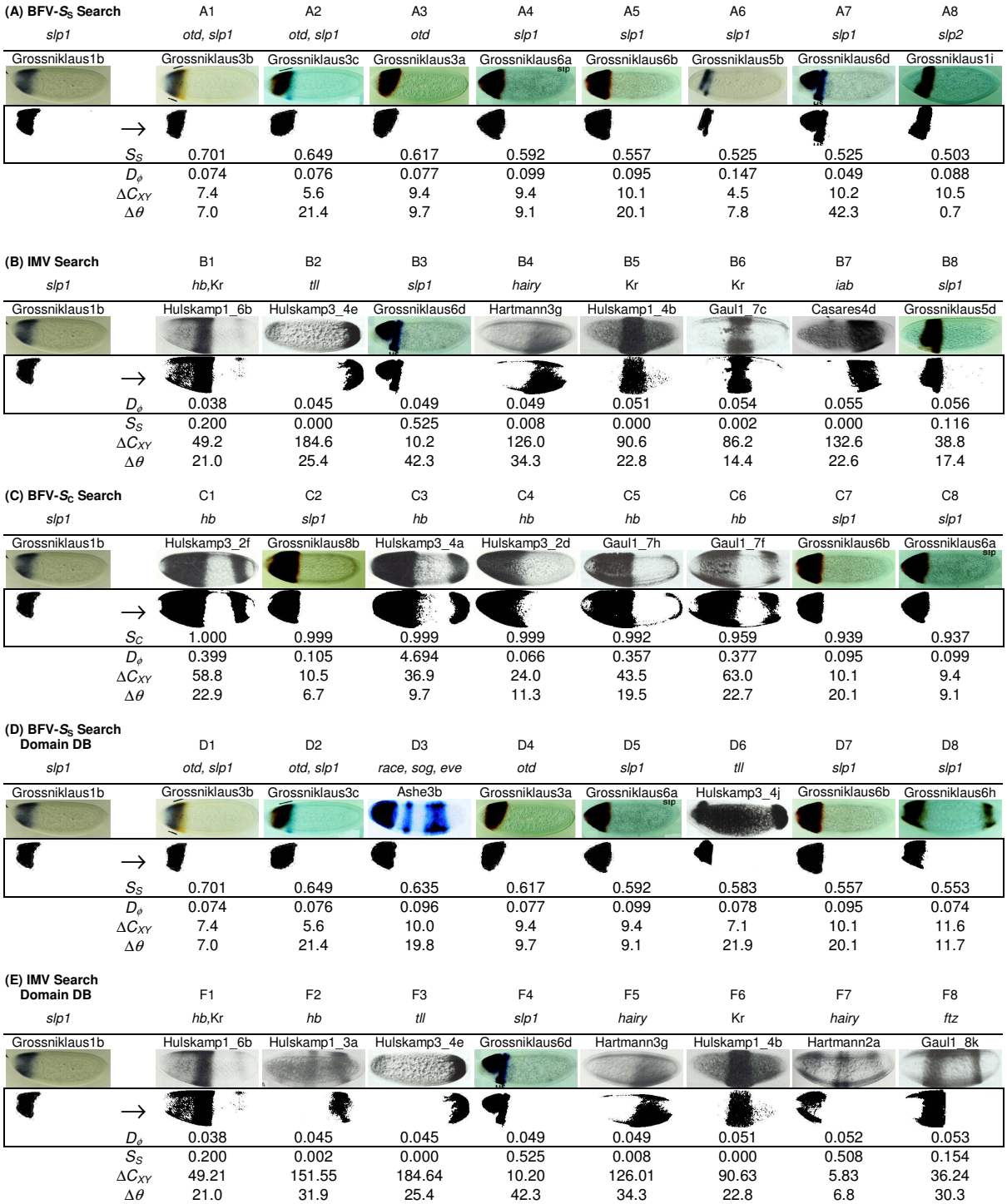
(A) BFV-S_S Search
A1 A2 A3 A4 A5 A6 A7 A8
slp1 otd, slp1 otd, slp1 otd slp1 slp1 slp1 slp2
Grossniklaus1b Grossniklaus3b Grossniklaus3c Grossniklaus3a Grossniklaus6a Grossniklaus6b Grossniklaus5b Grossniklaus6d Grossniklaus1i
S_S 0.701 0.649 0.617 0.592 0.557 0.525 0.525 0.503
D_ϕ 0.074 0.076 0.077 0.099 0.095 0.147 0.049 0.088
ΔC_{XY} 7.4 5.6 9.4 9.4 10.1 4.5 10.2 10.5
$\Delta\theta$ 7.0 21.4 9.7 9.1 20.1 7.8 42.3 0.7
(B) IMV Search
B1 B2 B3 B4 B5 B6 B7 B8
slp1 hb,Kr tll slp1 hairy Kr Kr iab slp1
Grossniklaus1b Hulskamp1_6b Hulskamp3_4e Grossniklaus6d Hartmann3g Hulskamp1_4b Gaul1_7c Casares4d Grossniklaus5d
D_ϕ 0.038 0.045 0.049 0.049 0.051 0.054 0.055 0.056
S_S 0.200 0.000 0.525 0.008 0.000 0.002 0.000 0.116
ΔC_{XY} 49.2 184.6 10.2 126.0 90.6 86.2 132.6 38.8
$\Delta\theta$ 21.0 25.4 42.3 34.3 22.8 14.4 22.6 17.4
(C) BFV-S_C Search
C1 C2 C3 C4 C5 C6 C7 C8
slp1 hb slp1 hb hb hb hb slp1 slp1
Grossniklaus1b Hulskamp3_2f Grossniklaus8b Hulskamp3_4a Hulskamp3_2d Gaul1_7h Gaul1_7f Grossniklaus6b Grossniklaus6a
S_C 1.000 0.999 0.999 0.999 0.992 0.959 0.939 0.937
D_ϕ 0.399 0.105 4.694 0.066 0.357 0.377 0.095 0.099
ΔC_{XY} 58.8 10.5 36.9 24.0 43.5 63.0 10.1 9.4
$\Delta\theta$ 22.9 6.7 9.7 11.3 19.5 22.7 20.1 9.1
(D) BFV-S_S Search Domain DB
D1 D2 D3 D4 D5 D6 D7 D8
slp1 otd, slp1 otd, slp1 race, sog, eve otd slp1 tll slp1 slp1
Grossniklaus1b Grossniklaus3b Grossniklaus3c Ashe3b Grossniklaus3a Grossniklaus6a Hulskamp3_4j Grossniklaus6b Grossniklaus6h
S_S 0.701 0.649 0.635 0.617 0.592 0.583 0.557 0.553
D_ϕ 0.074 0.076 0.096 0.077 0.099 0.078 0.095 0.074
ΔC_{XY} 7.4 5.6 10.0 9.4 9.4 7.1 10.1 11.6
$\Delta\theta$ 7.0 21.4 19.8 9.7 9.1 21.9 20.1 11.7
(E) IMV Search Domain DB
F1 F2 F3 F4 F5 F6 F7 F8
slp1 hb,Kr hb tll slp1 hairy Kr hairy ftz
Grossniklaus1b Hulskamp1_6b Hulskamp1_3a Hulskamp3_4e Grossniklaus6d Hartmann3g Hulskamp1_4b Hartmann2a Gaul1_8k
D_ϕ 0.038 0.045 0.045 0.049 0.049 0.051 0.052 0.053
S_S 0.200 0.002 0.000 0.525 0.008 0.000 0.508 0.154
ΔC_{XY} 49.21 151.55 184.64 10.20 126.01 90.63 5.83 36.24
$\Delta\theta$ 21.0 31.9 25.4 42.3 34.3 22.8 6.8 30.3

Figure 1

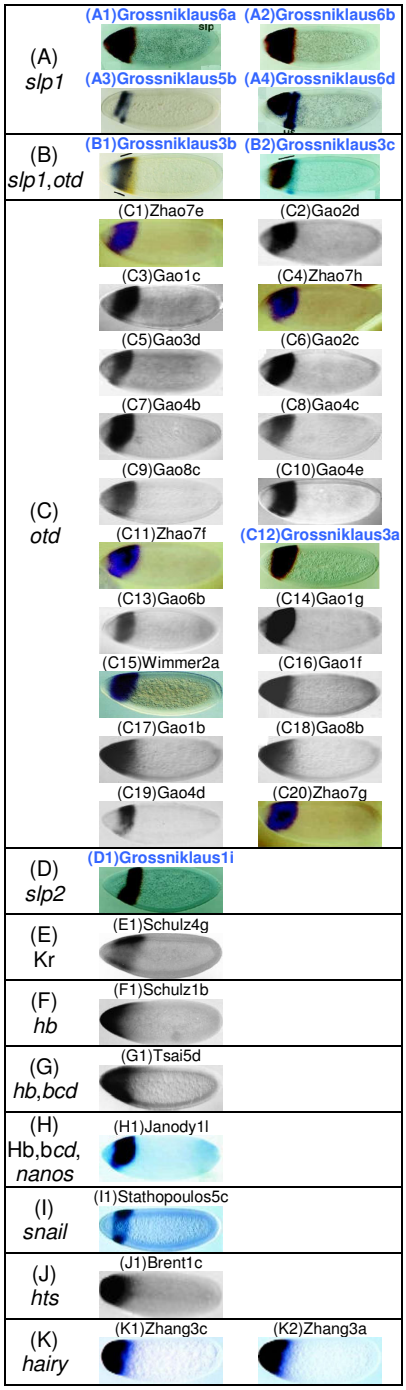
(A) slp1
(A1)Grossniklaus6a
(A2)Grossniklaus6b
(A3)Grossniklaus5b
(A4)Grossniklaus6d
(B) slp1,otd
(B1)Grossniklaus3b
(B2)Grossniklaus3c
(C) otd
(C1)Zhao7e
(C2)Gao2d
(C3)Gao1c
(C4)Zhao7h
(C5)Gao3d
(C6)Gao2c
(C7)Gao4b
(C8)Gao4c
(C9)Gao8c
(C10)Gao4e
(C11)Zhao7f
(C12)Grossniklaus3a
(C13)Gao6b
(C14)Gao1g
(C15)Wimmer2a
(C16)Gao1f
(C17)Gao1b
(C18)Gao8b
(C19)Gao4d
(C20)Zhao7g
(D) slp2
(D1)Grossniklaus1i
(E) Kr
(E1)Schulz4g
(F) hb
(F1)Schulz1b
(G) hb,bcd
(G1)Tsai5d
(H) Hb,bcd, nanos
(H1)Janody1l
(I) snail
(I1)Stathopoulos5c
(J) hts
(J1)Brent1c
(K) hairy
(K1)Zhang3c
(K2)Zhang3a

Figure 2

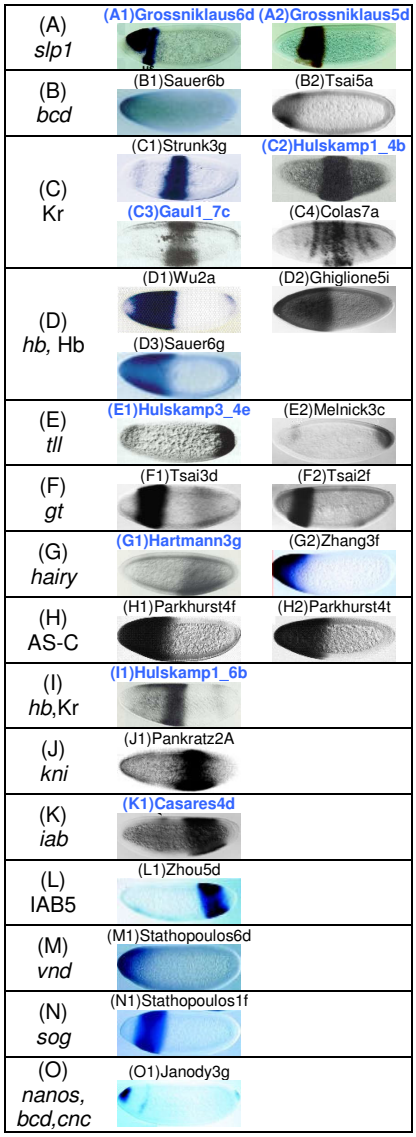
(A) slp1
(A1)Grossniklaus6d
(A2)Grossniklaus5d
(B) bcd
(B1)Sauer6b
(B2)Tsai5a
(C) Kr
(C1)Strunk3g
(C2)Hulskamp1_4b
(C3)Gaul1_7c
(C4)Colas7a
(D) hb, Hb
(D1)Wu2a
(D2)Ghiglione5i
(D3)Sauer6g
(E) tll
(E1)Hulskamp3_4e
(E2)Melnick3c
(F) gt
(F1)Tsai3d
(F2)Tsai2f
(G) hairy
(G1)Hartmann3g
(G2)Zhang3f
(H) AS-C
(H1)Parkhurst4f
(H2)Parkhurst4t
(I) hb,Kr
(I1)Hulskamp1_6b
(J) kni
(J1)Pankratz2A
(K) iab
(K1)Casares4d
(L) IAB5
(L1)Zhou5d
(M) vnd
(M1)Stathopoulos6d
(N) sog
(N1)Stathopoulos1f
(O) nanos, bcd,cnc
(O1)Janody3g

Figure 3

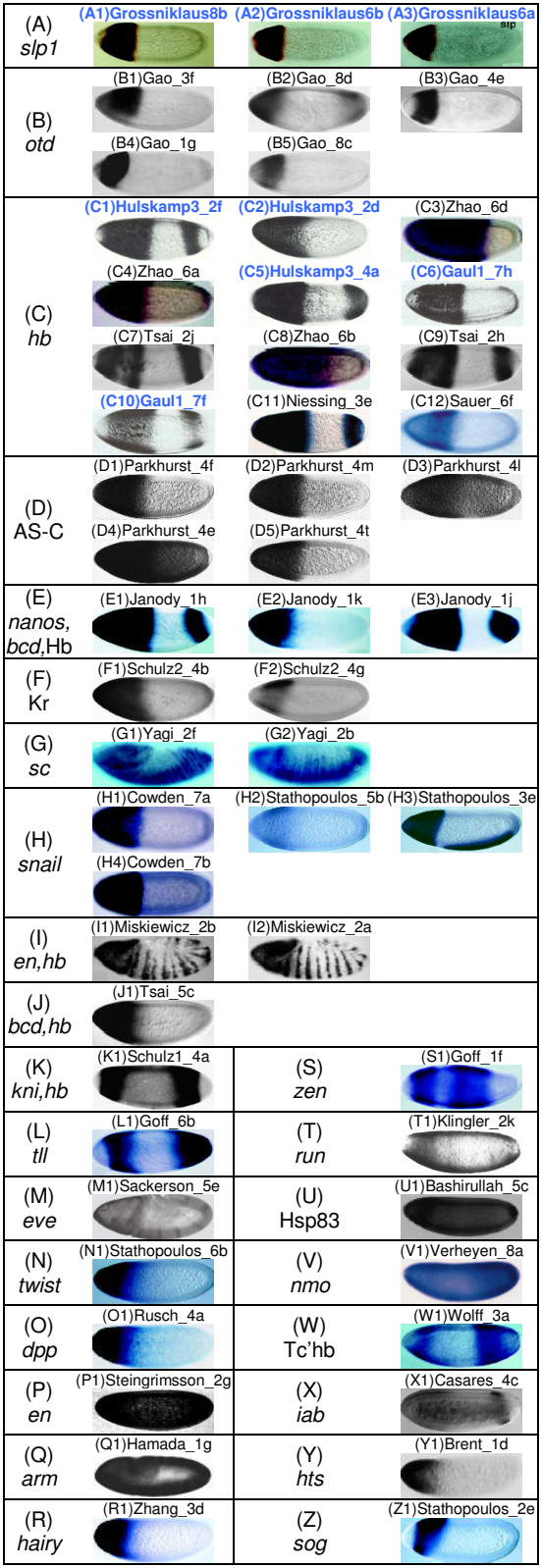
(A) slp1
(A1)Grossniklaus8b
(A2)Grossniklaus6b
(A3)Grossniklaus6a
slp
(B) otd
(B1)Gao_3f
(B2)Gao_8d
(B3)Gao_4e
(B4)Gao_1g
(B5)Gao_8c
(C) hb
(C1)Hulskamp3_2f
(C2)Hulskamp3_2d
(C3)Zhao_6d
(C4)Zhao_6a
(C5)Hulskamp3_4a
(C6)Gaul1_7h
(C7)Tsai_2j
(C8)Zhao_6b
(C9)Tsai_2h
(C10)Gaul1_7f
(C11)Niessing_3e
(C12)Sauer_6f
(D) AS-C
(D1)Parkhurst_4f
(D2)Parkhurst_4m
(D3)Parkhurst_4l
(D4)Parkhurst_4e
(D5)Parkhurst_4t
(E) nanos, bcd,Hb
(E1)Janody_1h
(E2)Janody_1k
(E3)Janody_1j
(F) Kr
(F1)Schulz2_4b
(F2)Schulz2_4g
(G) sc
(G1)Yagi_2f
(G2)Yagi_2b
(H) snail
(H1)Cowden_7a
(H2)Stathopoulos_5b
(H3)Stathopoulos_3e
(H4)Cowden_7b
(I) en,hb
(I1)Miskiewicz_2b
(I2)Miskiewicz_2a
(J) bcd,hb
(J1)Tsai_5c
(K) kni,hb
(K1)Schulz1_4a
(S) zen
(S1)Goff_1f
(L) tll
(L1)Goff_6b
(T) run
(T1)Klingler_2k
(M) eve
(M1)Sackerson_5e
(U) Hsp83
(U1)Bashirullah_5c
(N) twist
(N1)Stathopoulos_6b
(V) nmo
(V1)Verheyen_8a
(O) dpp
(O1)Rusch_4a
(W) Tc'hb
(W1)Wolff_3a
(P) en
(P1)Steingrimsson_2g
(X) iab
(X1)Casares_4c
(Q) arm
(Q1)Hamada_1g
(Y) hts
(Y1)Brent_1d
(R) hairy
(R1)Zhang_3d
(Z) sog
(Z1)Stathopoulos_2e

Figure 4

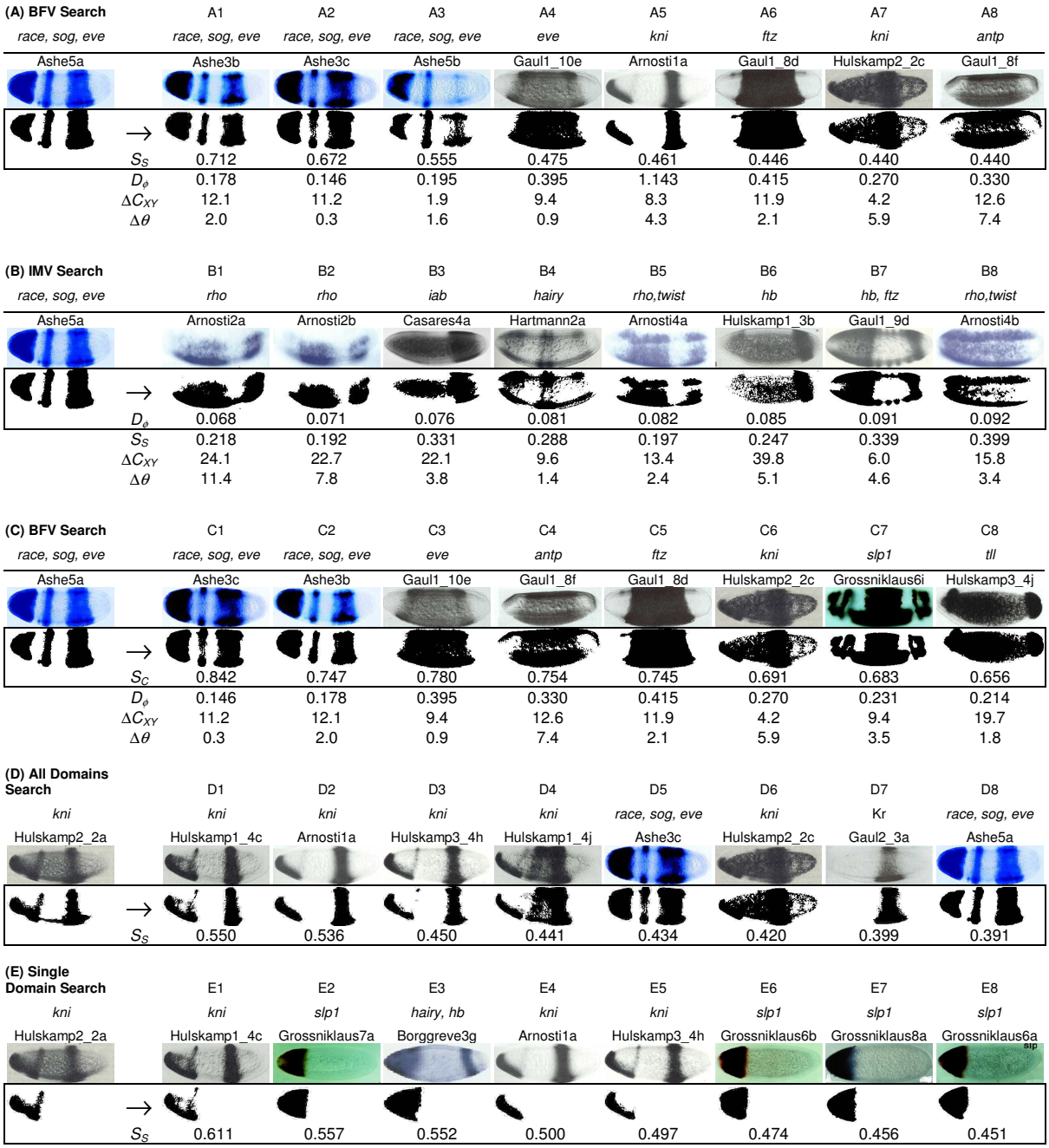
(A) BFV Search
A1 A2 A3 A4 A5 A6 A7 A8
race, sog, eve | race, sog, eve | race, sog, eve | eve | kni | ftz | kni | antp
Ashe5a | Ashe3b | Ashe3c | Ashe5b | Gaul1_10e | Arnosti1a | Gaul1_8d | Hulskamp2_2c | Gaul1_8f
S_S 0.712 0.672 0.555 0.475 0.461 0.446 0.440 0.440
D_ϕ 0.178 0.146 0.195 0.395 1.143 0.415 0.270 0.330
ΔC_{XY} 12.1 11.2 1.9 9.4 8.3 11.9 4.2 12.6
$\Delta\theta$ 2.0 0.3 1.6 0.9 4.3 2.1 5.9 7.4
(B) IMV Search
B1 B2 B3 B4 B5 B6 B7 B8
race, sog, eve | rho | rho | iab | hairy | rho,twist | hb | hb, ftz | rho,twist
Ashe5a | Arnosti2a | Arnosti2b | Casares4a | Hartmann2a | Arnosti4a | Hulskamp1_3b | Gaul1_9d | Arnosti4b
D_ϕ 0.068 0.071 0.076 0.081 0.082 0.085 0.091 0.092
S_S 0.218 0.192 0.331 0.288 0.197 0.247 0.339 0.399
ΔC_{XY} 24.1 22.7 22.1 9.6 13.4 39.8 6.0 15.8
$\Delta\theta$ 11.4 7.8 3.8 1.4 2.4 5.1 4.6 3.4
(C) BFV Search
C1 C2 C3 C4 C5 C6 C7 C8
race, sog, eve | race, sog, eve | race, sog, eve | eve | antp | ftz | kni | slp1 | tll
Ashe5a | Ashe3c | Ashe3b | Gaul1_10e | Gaul1_8f | Gaul1_8d | Hulskamp2_2c | Grossniklaus6i | Hulskamp3_4j
S_C 0.842 0.747 0.780 0.754 0.745 0.691 0.683 0.656
D_ϕ 0.146 0.178 0.395 0.330 0.415 0.270 0.231 0.214
ΔC_{XY} 11.2 12.1 9.4 12.6 11.9 4.2 9.4 19.7
$\Delta\theta$ 0.3 2.0 0.9 7.4 2.1 5.9 3.5 1.8
(D) All Domains Search
D1 D2 D3 D4 D5 D6 D7 D8
kni | kni | kni | kni | race, sog, eve | kni | Kr | race, sog, eve
Hulskamp2_2a | Hulskamp1_4c | Arnosti1a | Hulskamp3_4h | Hulskamp1_4j | Ashe3c | Hulskamp2_2c | Gaul2_3a | Ashe5a
S_S 0.550 0.536 0.450 0.441 0.434 0.420 0.399 0.391
(E) Single Domain Search
E1 E2 E3 E4 E5 E6 E7 E8
kni | kni | slp1 | hairy, hb | kni | kni | slp1 | slp1 | slp1
Hulskamp2_2a | Hulskamp1_4c | Grossniklaus7a | Borggreve3g | Arnosti1a | Hulskamp3_4h | Grossniklaus6b | Grossniklaus8a | Grossniklaus6a
S_S 0.611 0.557 0.552 0.500 0.497 0.474 0.456 0.451

Figure 5