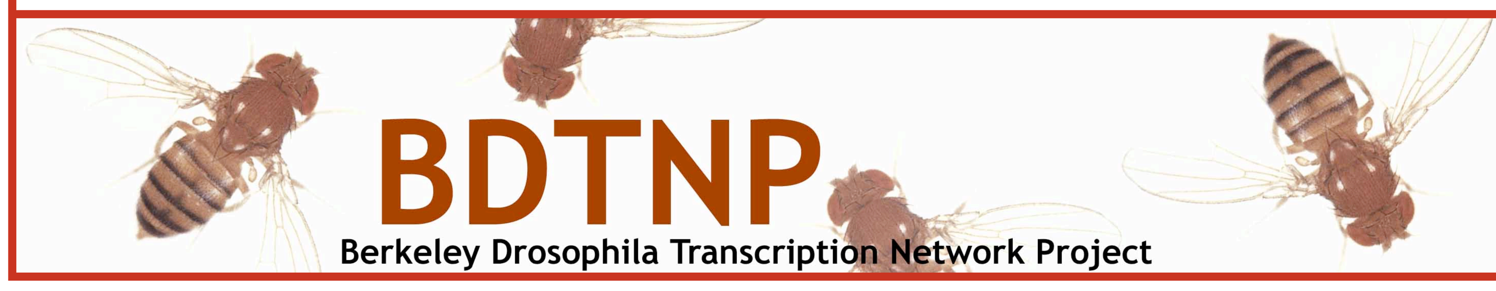
Counting Gene Expressing Nuclei in Whole *Drosophila* Blastoderm Embryos



Soile V. E. Keränen, Cris L. Luengo Hendriks, Damir Sudar, Mark D. Biggin, David W. Knowles

Life Sciences and Genomics Divisions, Lawrence Berkeley National Laboratory

Abstract

Complex temporal and spatial patterns of gene expression specificy morphology and tissue identity. To fully understand pattern formation, we need to characterize the expression of large number of genes during development in three-dimensions with cellular resolution. To make such datasets useful, gene expression patterns must be quantitative and reduced to a computable form. The Berkeley Drosophila Transcription Network Project is developing a set methods to do this. Our goal is to build an expression atlas that will record the expression of 1,000 genes in wild type pregastrula embryos and up to 200 genes in a series of mutant embryos, each mutant for one of 34 early acting transcription factor.

We have adapted fluorescent *in situ* hybridization protocols for this purpose. Stained and mounted blastoderm embryos are imaged whole by multiphoton confocal microscopy. One fluorescent channel is reserved for a DNA-stain to detect nuclei; the other two channels contain different gene expression patterns detected with tyramide-reactions. The confocal image stacks are then analyzed to yield a 3D computer representation of gene expression around each nucleus (posters 358A Luengo Hendriks et al., 350B Fowlkes et al.).

There are many ways to look such data. In this poster, we introduce some computational methods for analysing gene expression using numerical pointcloud data on *eve*, *ftz*, *kni*, *rho*, and *sna* mRNA expression patterns in a set of stage 5 embryos. The examples here show that it is possible to connect abstract numerical data on gene expression with biological events. Eventually, we hope to be able to use similar approaches for mining regulatory information from unknown patterns in a high through-put manner.

From expression patterns to numerical data sets

To collect data on gene expression, fluorescently triple-stained blastoderm embryos (sytox green for DNA and Cy3 and coumarin-tyramides for two mRNAs) are staged according to the percentage of cell membrane development and imaged in confocal microscope to record the 3D expression patterns as image stacks (Figure 1). The images are converted into text-files of local expression levels that are readable to various data analysis scripts and programming languages (see poster 358A Luengo Hendriks et al.). This allows flexible data sharing, such as mapping expression data into a virtual embryo (see posters 350B Fowlkes et al., 374B Weber et al.), as well as more abstract analyses of multiple expression features (Figure 2). The versatility of numerical data means that the direction of research depends on mathematical tools chosen by the researcher.

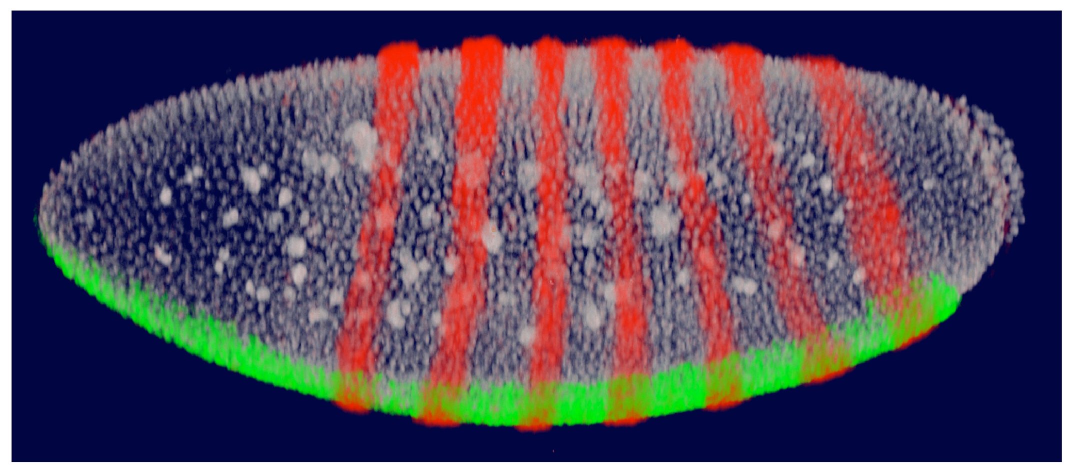


Figure 1

A pseudocolor volume-rendered and thresholded image of a *D. melanogaster* blastoderm embryo triple-stained for *ftz* mRNA (red), *sna* mRNA (green) and DNA (white) (for details, see also poster 374B Weber et al.).

While visual information is intuitively fast to comprehend, computational methods allow quantification of the various patterning features for more critical analyses.

id	x	y	z	Nx	Ny	Nz	Da	Db	Vn	Vc	Sytox	Cy3_n	Cy3_a	Cy3_b	Cy3_g	Cou_n	Cou_a	Cou_b	Cou_g
0	0.45	0.45	1.5	0.45	0.45	1.5	1	1	0.304	0.304	0.071	0.074	0.079	0.095	0.079	0.069	0.076	0.271	0.088
## Using MAKEPOINTCLOUD of 25 February 2005.																			
1009	202.8	183.1	76.68	-0.08	0.938	-0.34	0	0	330.3	871.3	84.04	64.96	39.57	39.99	49.63	52.15	57.76	65.92	57.33
2018	336.1	121.2	165.1	0.336	0.223	0.915	0	0	263.2	660.3	38.7	34.33	16.69	29.17	27.02	14.47	12.09	37.06	15.26
3027	118.2	123.6	169.3	-0.24	0.324	0.916	0	0	154.2	884.4	32.26	30.72	19.55	29.39	24.09	5.275	5.246	25.44	7.203
4036	320.7	37.24	86.3	0.248	-0.93	-0.29	0	0	276.6	626.9	84.54	65.55	37.41	38.57	50.99	14.52	17.23	45	18.52
1	296.7	34	124.9	0.144	-0.94	0.312	0	0	819.1	1834	72.75	57.06	21.53	36.32	41.17	34.51	38.7	46.57	37.9
1010	126.6	163.8	83.28	-0.36	0.871	-0.33	0	0	559.5	1521	73.06	55.02	35.4	35.44	42.37	28.07	35.7	55.05	32.73
2019	164.6	145.1	44.55	-0.26	0.502	-0.83	0	0	208.3	778.4	69.62	55.12	40.79	37.75	42.69	22.29	27.55	54.72	26.55
3028	130.9	167.8	85.74	-0.37	0.876	-0.31	0	0	230.1	625.7	72.91	55.81	37.54	36.73	44.29	24.13	35.16	50.07	32.09
4037	250.1	149.6	38.53	0.063	0.569	-0.82	0	0	250.2	773.3	73.23	64.15	48.57	44.74	51.63	12.65	11.58	47.8	15.03
5046	194	156.6	162.5	-0.09	0.671	0.736	0	0	151.2	417.4	39.32	36.6	22.04	29.29	28.96	9.733	11.33	31.14	12.03
2	255.6	171.6	59.02	0.073	0.796	-0.6	0	0	510.9	1445	78.29	65.85	46.97	43.62	51.64	11.8	11.4	48.98	14.87
1011	365.8	99.71	60.26	0.485	-0.08	-0.87	0	0	226.8	547.1	80.77	62.98	40.35	40.46	49.59	11.9	11.15	45.31	14.47
2020	243.6	40.92	58.34	0.064	-0.81	-0.58	0	0	240.1	535.8	85.99	73.87	45.07	45.6	59.07	48.04	56.77	59.51	55.71
3029	266.2	123.3	29.85	0.131	0.186	-0.97	0	0	301.2	953	73.54	62.13	37.7	42.84	46.24	33.79	32.75	64.58	34.72
4038	164.6	143.5	169.1	-0.2	0.541	0.817	0	0	196.1	623.6	33.25	36.52	22.08	30.46	28.57	5.819	5.336	26.18	7.41
5047	208.8	51.67	160.3	-0.01	-0.72	0.692	0	0	301.8	791.2	46.63	40.02	16.4	32.87	31.19	6.299	3.95	29.19	8.058
3	186.9	158.1	159.9	-0.09	0.678	0.73	0	0	378.6	1203	38.59	39	23.65	34.53	31.34	20.99	25.48	40.95	22.44
1012	357.9	152.1	127.9	0.474	0.821	0.319	0	0	397.7	839.7	70.62	55.27	24.58	33.02	41.77	8.694	7.315	37.7	10.87
2021	334.1	126.3	165.1	0.33	0.307	0.893	0	0	304.5	793.3	38.62	34.5	16.8	30.09	26.99	22.84	23.14	43.41	24.04
3030	388.5	136.2	98.19	0.66	0.713	-0.24	0	0	261.7	639.4	90.4	83.98	65.09	52.85	69.93	10.63	9.24	42.13	12.7
4039	368.6	111.8	63.53	0.512	0.143	-0.85	0	0	261.1	694.9	73.5	59.18	41.76	44.63	48.03	12.02	11.59	45.07	14.7
4	273.6	138.4	35.78	0.189	0.381	-0.9	0	0	294.5	796.6	74.94	61.45	41.55	40.93	48.4	32.02	34.33	59.65	34.52
1013	94.87	58.14	75.34	-0.29	-0.7	-0.65	0	0	196.7	533.7	73.21	53.64	36.01	37.26	42.43	11.08	12.61	45.18	14.54

Table 1 Beginning of a pointcloud file. Each row contains nucleus id, its X,Y,Z-co-ordinates, nuclear and cytoplasmic volumes, the average intensity of the DNA stain in the nucleus (sytox), and the average intensities of the two RNA-stains for apical cytoplasm (a), basal cytoplasm (b), nucleus (n), and the whole region (g). The number of lines below the header lines equals to the number of segmented objects (for details, see poster 358A Luengo Hendriks et al.)

Taking a first peek at numerical data sets

The expression intensities in this poster are normalized from 1 - 100, with 1% top and bottom intensity nuclei set into 1 or 100. As seen below, the distribution of intensities can vary between the genes and developmental stages. This tells what fraction of cells express the gene and at what level. Since the amount of computation often increases exponentially with the size of the data set, such primary intensity profiles might be useful for dividing unknown expression patterns into smaller cohorts to speed up further regulatory analyses.

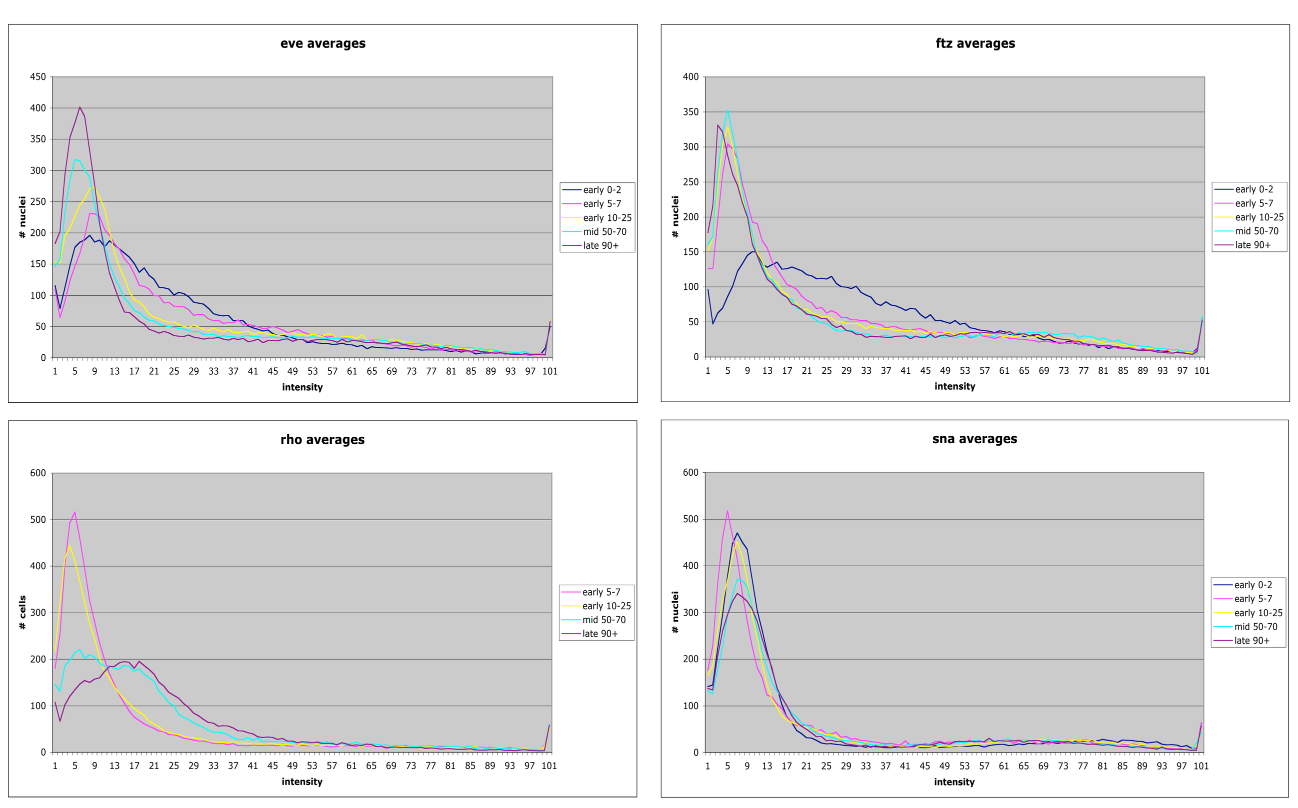


Figure 2 Averaged intensity profiles of four genes at five different stages of blastoderm embryos (early 0-2% cell wall invagination, early 5-7%, early 10-25%, mid 50-70%, and late 90+% cell wall invagination; n = 4-17). In *eve* and *ftz* the intensity differences between cells become sharper as the pattern evolves, whereas in *rho* they decrease and in *sna* the profiles stay quite similar. Such changes might reflect, e.g., overall sharpening or diffusion of the pattern borders (development of binary vs graded expression patterns) or generic up- or downregulation of gene expression.

Quantifying the spatial information in the patterns

Each pointcloud has data complexity equivalent to that of a microarray. However, pointcloud dataset contains spatial information that can be analyzed at various resolutions, data not available from methods using summed total expression such as microarray data. Simple analyses on spatial features can be used to support and verify the results of more complex spatial analyses later. For example, calculating the distribution of minimum distances from expressing to non-expressing cells and the inverse is one way to generate 1st pass quantitative measures on spatial characteristics of expression patterns.

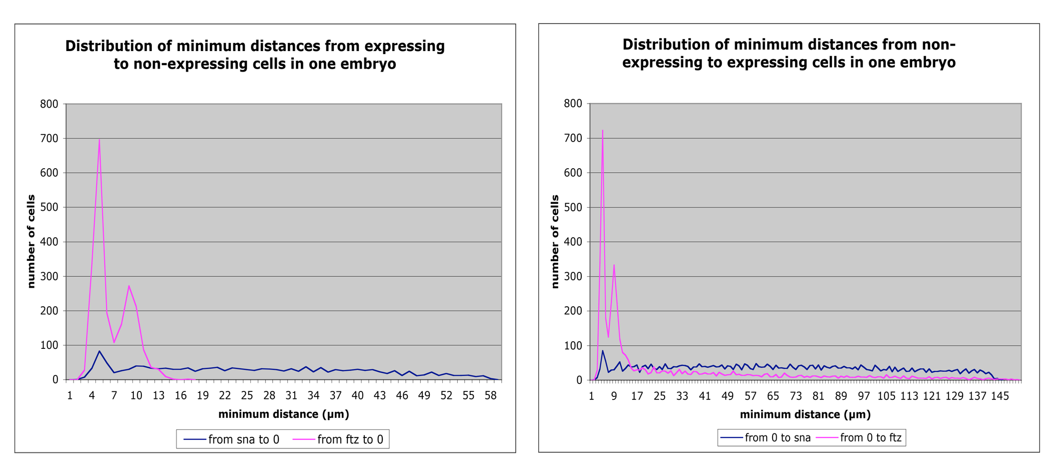


Figure 3 Distribution of minimum distance from an expressing to the nearest non-expressing cell and vice versa in one embryo stained for *ftz* and *sna*. At intensity cut-off 30 (≥ 30 expression, < 30 no expression), most of the minimum distances from ftz to non-expressing are short, reflecting the pattern that consists of narrow stripes, whereas many of the sna expressing cells are within a large band of cells. While the left histogram reflects the shape of the expression pattern, the minimum distances from non-expressing to nearest expressing cell in the right histogram reflect the spatial spread of the pattern, in this case 7 round stripes for *ftz* and one ventral band for *sna* (for picture of the expression, see Figure 1).

Measuring the expression patterns domain by domain…

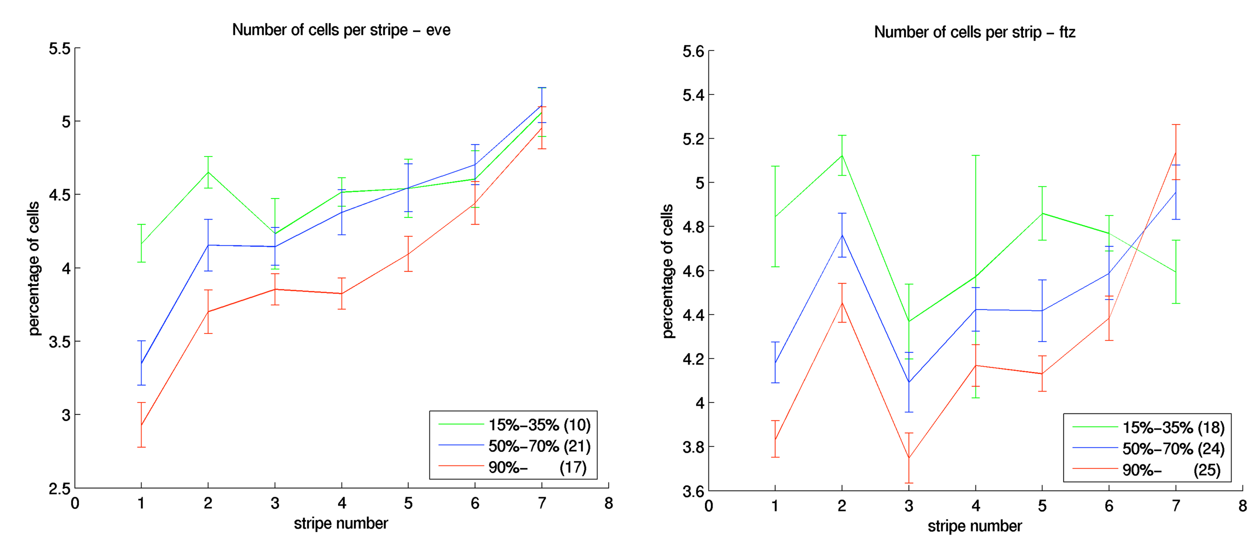


Figure 4. The average percentage of the cells in a stripe compared to the total number of cells at three different stages of blastoderm cellularization. Left *eve*, right *ftz*.

The individual elements of the patterns are easily seen by visual inspection of an image, but extracting them from numerical datasets can be more problematic. Because different expression domains may be controlled by different regulatory modules, spatial subdivisions of the expression data are still essential for dissecting the developmental regulatory network. Smoothing the data and then isolating the areas with sharp intensity jumps (1/x cell whorl-thing) allows automated identification of isolated expression domains with minimal user input.

For example, *eve* and *ftz* are easily seen as having seven stripes at blastoderm stage (as in Figure 6). However, computational analysis of the individual stripes shows that on average, different stripes can contain different numbers of cells, and that the temporal profiles of the stripes can be different, even though they belong to the same gene. Moreover, though both *eve* and *ftz* are pair-rule genes, the profiles of their stripe patterns are different (Figure 4).

Average class distribution for six nearest neighbors for a class of cells

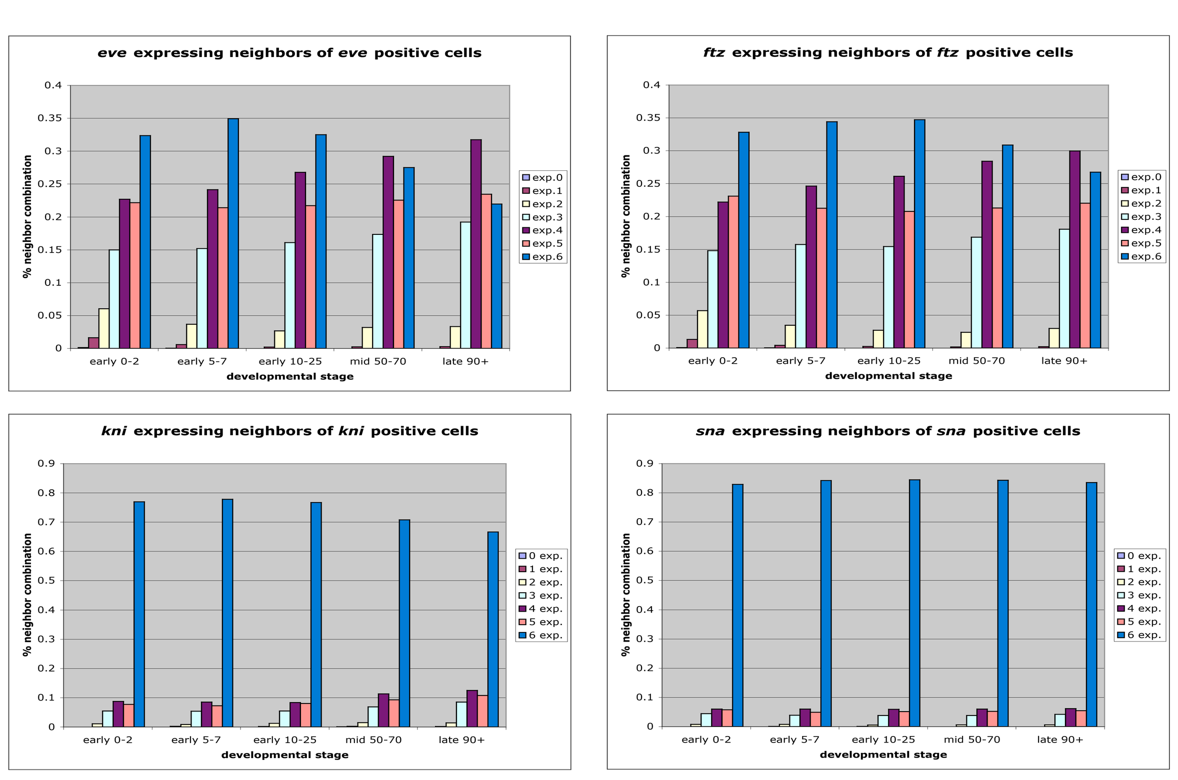
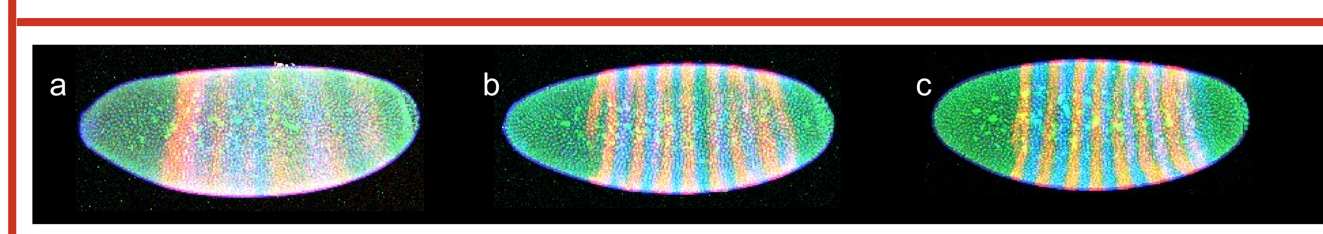


Figure 5. Bar graphs showing the average frequency at which 6 nearest neighbors for a cell expressing a gene also express the same gene. The frequency at which all 6 nearest neighbors (6 exp.) express the gene drops in *ftz*, *eve* and *kni*, but not in *sna*.

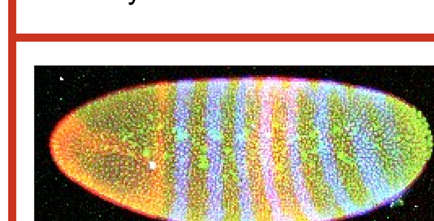
Plotting the fractions of expressing cells by different numbers of neighbors that belong to the same expression class (0/6-6/6) is another method for measuring shape information. The fraction of cells surrounded by only gene expression (6/6) decreases as the relative border length in the pattern increases. This tells how clumped the expression is (e.g. thin *ftz* stripes vs broad *sna* band), and how this pattern shape feature evolves during development (e.g. decrease of 6/6 in later *eve*, *ftz*, and *kni* patterns).

Degree of co-expression cell-by-cell



	e+f	e	f	-	e+f	e	f	-	e+f	e	f	-
obs	405	1123	1512	3617	117	1786	1904	2213	53	1585	1703	2285
exp	440	1088	1477	3652	639	1264	1382	2734	511	1127	1245	2743
diff.	-35	35	35	-35	-522	522	522	-522	-458	458	458	-458

Figure 6. Three embryos (**a** early 0, **b** early 25, **c** mid 65) stained for *eve* mRNA (red), *ftz* mRNA (blue) and DNA (green). Below are shown the observed and expected frequencies of total cells expressing *eve+ftz*, *eve*, *ftz*, or neither in a smoothed pointcloud data set when all the cells in embryo are considered. When the stripes grow narrower, *eve+ftz* overlap reduces.



	kni total vs ftz total				kni total, ftz stripe 4			
	k+f	k	f	-	k+f	k	f	-
obs	170	946	1755	3178	166	950	130	4803
exp	355	761	1570	3363	55	1061	241	4692
diff.	-185	185	185	-185	111	-111	-111	111

Figure 7. An embryo (mid 50) stained for *kni* mRNA (red), *ftz* mRNA (blue) and DNA (green) has *in toto* fewer co-expressing cells in a smoothed pointcloud data set than random expectation (when the total distribution of expressing cells all around the embryo is independent of rules for distribution of other gene expressing cells), but locally, the posterior domain of *kni* expression overlaps with *ftz* stripe 4 (red arrow) showing up as higher than expected co-expression frequencies.

Strong and/or consistent positive and/or negative correlations may be identified by binning the normalized co-expression intensities and then calculating the difference between observed and expected frequencies of nuclei in each co-expression bin (Figure 6). However, more complex analyses based on domain-by-domain comparisons are probably required for discovering specific interactions, since, e.g., summing the all different overlap profiles of individual domains can mask local patterning detail between genes (for example, Figure 7).

Future prospects

Our initial goal is to map in depth the expression patterns of 34 early developmental regulatory transcription factors, on which the biology is well known. This data will be used for:

- Generating a first pass regulatory network for blastoderm pattern formation to guide the high through-put analysis of ~1000-1500 genes
- Further development of novel tools and approaches for mining 3D expression data

Examples of non-spatial methods to be developed further:

* Exploration on the spatiotemporal dynamics of total co-expression data and associating different dynamics with different known regulatory interactions

* Development of statistical tests to detect and discriminate between subtle patterns

Examples of spatial methods to be developed further:

* More methods for finding and describing of intensity peaks in 3D data sets

* Using spatial data to limit the number of pairwise gene and/or domain comparisons

Please see other BDTNP posters for further methodology and other forms of pattern analyses