Supplementary Material

An Intuitive Graphical Visualization Technique for the interrogation of Transcriptome Data

Natascha Bushati, James Smith, James Briscoe, Christopher Watkins

Intuitive description of t-SNE algorithm

The algorithm takes as parameters the set of N H-points and a positive integer K, which is known as the 'perplexity'. The first stage is to identify the K nearest neighbours of each H-point. This is achieved by assigning nearest neighbour scores that are between 0 and 1: the neighbour score of a point h_j from h_i is proportional to $\exp(\frac{-d(h_i,h_j)^2}{\sigma_i^2})$, where $d(h_i,h_j)$ is the distance between h_i and h_j, and σ_i is a scale parameter chosen individually for each point h_i, in proportion to the distances to its neighbours. A key property of these nearest neighbour scores is that they decline rapidly with distance in H-space. (To accomplish this σ_i is adjusted for each h_i so that the entropy of the neighbor scores for h_i, normalized to be a probability distribution, is K. In effect, this is a 'soft' version of choosing K nearest neighbors.)

Each pair of H-points, h_i and h_j, is then given a neighbour score by averaging the neighbour scores of h_j from h_i, and h_i from h_j. (Note that σ_i and $\sigma_j$may be different, in which case the neighbour scores from h_i to h_j and from h_j to h_i are different.) Averaging the scores in both directions gives a symmetric weight matrix. Importantly, this improves the visualization because a point that is outlying or isolated in H-space will have a large σ, giving high neighbour scores for quite distant points in H-space. In V-space, the isolated point will typically be placed close to one of its neighbours in H-space; this avoids the visualization being dominated by widely spaced isolated or outlying points. Isolated points and outliers can be identified from the neighbour plot, instead of by their spatial position in the visualization.

The second stage of the algorithm is to arrange the V-points. Starting from a random arrangement of V-points, their positions are then optimized by gradient descent, to minimize a matching penalty for the neighbour-scores of the V-points with the neighbour-scores of the H-points.

The neighbour scores of the V-points are calculated in a subtly different way: the score for v_i and v_j is proportional to $\frac{1}{1+d(v_i,v_j)^2}$, where $d(v_i,v_j)$ is the current distance between the V-points v_i and v_j in the scatter plot. The point of this definition is that these V-neighbour scores decline gradually with distance, so that the scatter plot may be expanded to spread out the high-dimensional neighbour relationships more faithfully. Previous visualization methods such as SNE (1) suffered from a 'crowding problem', in which many points were crowded together in the centre of the visualization, surrounded by outliers. t-SNE avoids this by having neighbour scores scale differently with distance in V-space and H-space.

The matching score that is optimized is the Kullback-Leibler divergence between the H and V neighbour scores, when both sets of scores are normalized to sum to 1. Key intuitive properties of this measure of divergence are that points that are close in H-space are dragged towards each other in V-space; points that are close in V-space but far from each other in H-space are dragged apart, but less strongly; and finally that there is negligible penalty for mismatches between distances that are large in both H-space and in V-space. If two points that are close in H-space are separated in the optimization, the strength of the attractive 'force' between them declines with distance, so that, for example, a spherical shell of points is 'ripped' and mapped into a single sheet in the plane, rather than being 'squashed' as in a linear projection.

The optimization is non-convex: many local optima are possible, and different runs give different results, but the overall matching score may be used to select the best of a number of runs. We have found that this is usually unnecessary: most runs are similar.

The non-convexity of the optimization, together with the symmetrisation of neighbour scores in H-space, enables outlying points to be visualized in a neat way. An outlying point in H-space can have a high neighbour score with several well-separated points, in different places in the visualization. The non-convexity of the optimization criterion causes the outlier to be put close to one of its neighbouring points – an outlier thus tends to 'attach itself' to the nearest cluster, even though it may not strictly be part of the cluster.

This phenomenon of 'attachment of outliers' is good for the visualization in that outliers place themselves close to clusters, instead of spacing themselves out and distorting or dominating the whole visualization. Display techniques such as the neighbour plot enable the user to identify the outliers in the visualization, and distinguish them from points that are genuinely well placed relative to their neighbours. If there is a cloud of outliers, there may be no good way to place them in the visualization in a way that preserves neighbour relationships: outliers may distort and dominate PCA, for example. t-SNE tends to place the outliers compactly close to their nearest cluster: this enables the cluster relationships to be displayed correctly, while the outliers can be identified using alternative display modalities, such as neighbour plots.

Sensitivity of t-SNE to noise

A problem encountered with some non-linear dimension reduction methods is that they can be sensitive to noise within a dataset. To examine the robustness of t-SNE to noise we investigated the performance of t-SNE on a synthetic dataset. This dataset consisted of 300 points in 6-dimensional space, 30 points from each of ten Gaussian clusters, giving 300 points in all. The cluster centres were generated from an isotropic Gaussian distribution with variance 9 in each dimension; each cluster had variance 2 in each dimension. Most clusters were therefore well separated in six-dimensional space.

A scatter plot of the first and second principal components, with points coloured according to the clusters from which they were generated indicated that some clusters could be discerned, but not reliably delineated by eye (Supplementary Figure 1a). By contrast, a t-SNE visualisation of the same dataset (using the standard parameter settings, perplexity=30) separated the clusters much more widely (Supplementary Figure 1b). Moreover a neighbour plot (first and second closest neighbours) further aided the

interpretation of this visualisation: it merged two of the generative clusters but all other distinctions were clearly visible (Supplementary Figure 1c). For comparison Supplementary Figure 1d shows the t-SNE scatter plot coloured according to classes discovered by k-means clustering, with k set to 10, the correct number of classes. This method merged two pairs of the generative classes and incorrectly split two of the generative classes.

To test the effect of noise on the visualisation produced by t-SNE we generated a second synthetic dataset of 600 points, consisting of the same 300 points used above, together with an additional 300 'noise points' generated from an isotropic Gaussian distribution with variance 11 on each dimension - this distribution is the mixture distribution for the generative model of clusters with internal variance 2 and variance of the means being 9. The principal component visualisation of these data (Supplementary Figure 1e) would be uninterpretable were it not coloured with the correct generative classes (the noise points are brown). In the case of the t-SNE visualisation of this dataset (Supplementary Figure 1f), points from the same generative classes are grouped together within the noise. Importantly, a neighbour plot for this visualisation (Supplementary Figure 1g) allows seven or eight clusterings to be distinguished by eye, and these correspond to generative classes. By contrast, colouring the t-SNE visualisation according to classes found by k-means with k=10 indicates that some of the generative classes are more-or-less correctly found, but others are not (Supplementary Figure 1h). Overall the neighbour plot gives a more interpretable picture than k-means clustering.

These results show that t-SNE does not break down with noise, as one might fear that a non-linear method would. Indeed, k-means clustering fares worse in this example, even though the data is of a type that conforms to the statistical model implicit in the k-means algorithm (2). This suggests that t-SNE and neighbour plots might provide an interpretable visualisation of transcriptome data that can be used to generate hypotheses, recognise some structure, and inspire further investigation.

Application of t-SNE to additional datasets

To evaluate the utility of t-SNE maps we applied the technique to five datasets. Two representative datasets are described in the main text; here we describe three additional datasets (see Methods for a full description of the data).

t-SNE maps representing the behaviour of the genes in each dataset were generated (Supplementary Figures 3a, 4a, 5a). Examination of the expression behaviour of small groups of neighbouring data points in each of the plots indicated that the t-SNE algorithm had effectively grouped genes with similar behaviours (Supplementary Figures 3a, 4a, 5a). Neighbour plots in which lines joined each data point to its two nearest D-space neighbours confirmed that the positioning of data points reflected the structure of the high dimensional data (Supplementary Figures 3b, 4b, 5b). Moreover, exploring the t-SNE plots in conjunction with the neighbour plots revealed any metric distortions in V-space introduced by the non-linear dimension-reduction. This highlighted which groups of points are really similar and the underlying logic of each mapping.

Dataset 3 describes the transcriptome analysis of mouse serotonergic (5HT) neurons (3). The data contained four conditions consisting of 5HT and non-5HT neurons taken from

stage E12.5 rostral and caudal hindbrain, respectively. Examination of the t-SNE mapping, readily identified regions of the mapping that harboured genes induced (orange in Supplementary Figure 3a) or repressed (magenta) in all 5HT neurons, and regions in which the expression of the genes was restricted to caudal or rostral cell populations (Supplementary Figure 3a). The continuous nature of gene expression patterns in the dataset was apparent. We overlaid the 5 clusters of genes described by Wylie et al. (3) in their original study on to the t-SNE map. The spatial integrity of the clusters confirmed that the validity of the t-SNE mapping and in combination with the neighbour plot provided insight into the logic of the underlying structure. This identified clusters of co-regulated genes independent of additional information and was sufficient to allow the detection of additional genes affiliated with each cluster that were not recognized in the original analysis (Supplementary Figure 3d and data not shown).

Dataset 4 was generated to identify genes induced or repressed by Sonic Hedgehog signalling in the neural tube (4). Neural progenitors of the developing chick neural tube were manipulated by *in ovo* electroporation to repress ('Ptc') or induce ('GliHigh') Shh signalling in a cell autonomous manner. Five conditions were assayed, three ('Ptc', control, 'GliHigh') at an early time point (14h), two (control, 'GliHigh') at a later time point (36h). A t-SNE mapping of differentially expressed genes from this dataset resulted in a ring-like structure, underscoring the continuum of gene expression patterns in this data (Supplementary Figure 4a). Using the corresponding nearest neighbour plot (Supplementary Figure 4b), regions of tightly grouped genes that corresponded to genes situated close to one another in expression space were evident. This identified groups of genes induced or repressed by Shh at all times (magenta, grey), induced or repressed only after short periods (red, light green, pink, yellow) and genes repressed only after longer periods of Shh signalling (orange). Moreover, groups of genes that were independent of Shh signalling but induced or repressed over time in these cells (light blue, dark green) were also identifiable with this method.

The final dataset (Dataset 5) analyzed with t-SNE comprised the transcriptomes of *Drosophila melanogaster* wild type eye, antennal and leg imaginal discs, and antennal and leg discs in which the gene eyeless was misexpressed, in the presence or absence of a null allele of the transcription factor atonal (5). Eyeless is a transcription factor that induces the entire cascade of eye development in imaginal discs and the experiment was designed to identify genes involved in eye development that are regulated by atonal. A t-SNE map and corresponding neighbour plot of the differentially expressed genes revealed clearly separated groups of gene expression patterns. Most notable were genes expressed in both wild type leg discs and eyeless expressing leg discs that lacked atonal (Supplementary Figure 5a, red). Another clearly separated group contained genes upregulated only in eyeless expressing, atonal mutant leg discs (magenta). In addition, a group of genes induced or repressed by eyeless misexpression in all discs, irrespective of atonal (light blue, dark green) was evident, as was a group of genes induced by misexpression of eyeless, which was dependent on atonal (pink). Most strikingly, we identified groups of genes that were induced in the wildtype eye disc and in eyeless expressing antennal and leg discs, which were not expressed in wild type antennal or leg discs (dark blue, light green). Conversely, genes that were expressed in the eye but not in other discs expressing eyeless suggesting the presence of eyeless independent genes within the eye disc (grey).

To further investigate the utility of the t-SNE maps, we performed hierarchical clustering of Datasets 4 and 5 and overlaid the resulting clusters onto the corresponding t-SNE maps (Supplementary Figures 4d, 5d). In both cases the clusters were less spatially coherent than in the case of Datasets 1-3, and the logic underlying the partitioning of the clusters was not immediately clear. This is presumably due to the higher complexity of these datasets, in which changes in time, genotype and tissue type combine to account for the behaviour of the transcriptomes. This highlights a limitation of conventional clustering approaches and indicates how a visualisation technique such as t-SNE can provide a more satisfactory method to capture and assess the behaviour of genes in complex datasets. Moreover, it emphasizes the power of unbiased exploration of the expression data independent of the expectations and assumptions of the initial experimental design.

Finally, we compared the t-SNE maps to plots of the first two PCs of each dataset. This confirmed the superiority of t-SNE (Supplementary Figures 3c, 4c, 5c). It was particularly noteworthy that in the case of Dataset 5 the first 4 PC contributed significantly to explain the bulk of the variation in the data (Supplementary Figure 6). Consequently two-dimensional plots of pairs of PCs significantly underrepresented the data and provided only a partial visualisation of the gene expression behaviours within the experiment. By contrast the two-dimensional t-SNE map was sufficient to visualize the data and identify the key gene expression behaviours (Supplementary Figure 5a).

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Effect of noise on principal component and t-SNE mappings of a synthetic dataset. (**a-d**) Mappings of a synthetic dataset consisting of 300 points in 6-dimensional space clustered into ten Gaussian clusters of 30 points; each cluster had variance 2 in each dimension and a centre generated from an isotropic Gaussian distribution with variance 9 in each dimension. (**a**) First and second principal component projection of the synthetic dataset coloured according to the generative classes of the dataset. (**b**) A t-SNE visualisation of the same dataset coloured in the same manner. (**c**) A neighbour plot, revealing the first and second closest neighbours of each point, displayed on the t-SNE mapping from b. (**d**) The t-SNE mapping from b coloured by membership of the 10 clusters identified from the dataset using k-means (k=10). (**e-h**). Mappings of a 'noisy' synthetic dataset consisting of the same 300 points as above to which were added an additional 300 'noise points', generated from an isotropic Gaussian distribution with variance 11 on each dimension. (**e**) First and second principal component projection of the noisy synthetic dataset coloured according to the 10 clustered generative classes of the dataset. (**f**) A t-SNE visualisation of the same dataset coloured in the same manner. (**g**) A neighbour plot, identifying the first and second closest neighbours of each point, displayed on the t-SNE mapping from f. (**h**) The t-SNE mapping from f coloured by membership of the 10 clusters identified from the dataset by k-means clustering (k=10).

Supplementary Figure 2. Six independent t-SNE mappings of the datasets 1 and 2. (**a**) Six independently generated t-SNE maps of the 2148 probe sets identified as differentially expressed between six stages of human embryogenesis (6). (**b**) Six independently t-SNE maps of 3656 probe sets with periodic behaviour over 36 cycles in the yeast metabolic cycle described by Tu et al. (ref. (7)) In each case although the orientation and topological form of the maps differ the local relationships between individual data points is maintained.

Supplementary Figure 3. t-SNE mapping, PCA and clustering of control and serotonin (5HT) neuron transcriptomes from mouse hindbrain. (**a**) t-SNE map of 3079 probe sets identified as differentially expressed between 4 samples (3). Selected groups of neighbouring data points are highlighted and the expression behaviour (plotted as z-scores) of the selected genes over all conditions shown in the corresponding colours. Cctrl: Caudal, control; C5HT: Caudal, 5HT; Rctrl: Rostral, control; R5HT: Rostral, 5HT. (**b**) Nearest neighbour plots of the t-SNE mapping in (**a**). Each data point in the t-SNE map was connected to its two nearest neighbours in high-dimensional (4D) space and the connectors coloured according to the distance between these data points in high-dimensional space. Red indicates short and blue long distances in the higher dimensional space. Thus short red lines indicate faithful projection of distances. (**c**) Plots of the values of the first and second principal components of the same probe sets used to produce the t-SNE map in (**a**). The lower panel shows nearest neighbour plots of the PC plot as described in (**b**). (**d**) Overlay of clusters 1-5 produced using hierarchical clustering from the original study (3) onto the t-SNE map in (**a**). Data points are coloured according to cluster membership. Note that the five clusters available from the supplementary data of the original paper were not generated from the identical set of probe sets we used to generate the t-SNE mapping, hence the large number of black points not belonging to a cluster.

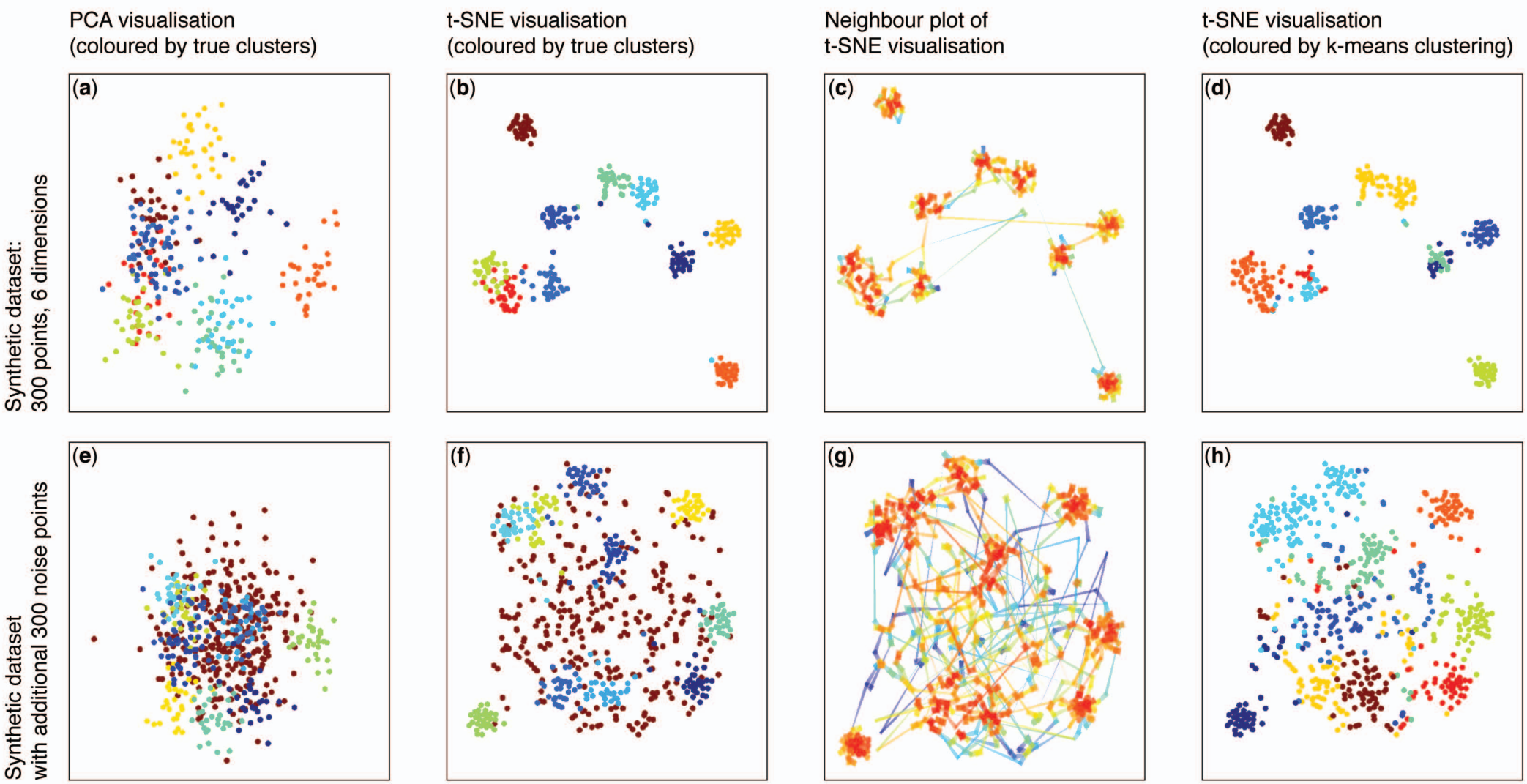
Supplementary Figure 4. t-SNE mapping, PCA and clustering of transcriptomes from chick neural tube cells in which Shh signalling has been manipulated. (**a**) t-SNE map of 2828 probe sets identified as differentially expressed between 5 samples (4). Selected groups of neighbouring data points are highlighted and the expression behaviour (plotted as z-scores) of the selected genes over all conditions shown in the corresponding colours. Ptc: Repression of Shh signalling by $Ptc1^{DLoop2}$, GFP: control, GliHigh: high levels of Shh signalling. (**b**) Nearest neighbour plots of the t-SNE mapping in (**a**). Each data point in the t-SNE map was connected to its two nearest neighbours in high-dimensional (5D) space and the connectors coloured according to the distance between these data points in high-dimensional space. Red indicates short, and blue long distances in the higher dimensional space. Thus short red lines indicate faithful projection of distances. (**c**) Plots of the values of the first and second principal components of the same probe sets used to produce the t-SNE map in (**a**). The lower panel shows nearest neighbour plots of the PC plot as described in (**b**). (**d**) Overlay of 10 clusters produced using hierarchical clustering onto the t-SNE map in (**a**). Data points are coloured according to cluster membership.

Supplementary Figure 5. t-SNE mapping, PCA and clustering of transcriptomes from wild type and eyeless over-expressing drosophila imaginal discs. (**a**) t-SNE map of 1917 probe sets identified as differentially expressed between 6 samples (5). Selected groups of neighbouring data points are highlighted and the expression behaviour (plotted as z-scores) of the selected genes over all conditions shown in the corresponding colours. ate: antennal disc, leg: leg disc, eye: eye disc, UAS-ey: misexpression of eyeless, ato: atonal mutant. (**b**) Nearest neighbour plots of the t-SNE mapping in (**a**). Each data point in the t-SNE map was connected to its two nearest neighbours in high-dimensional (6D) space and the connectors coloured according to the distance between these data points in high-dimensional space. Red indicates short, and blue long distances in the higher dimensional space. Thus short red lines indicate faithful projection of distances. (**c**) Plots of the values of the first and second principal components of the same probe sets used to produce the t-SNE map in (**a**). The lower panel shows nearest neighbour plots of the PC plot as described in (**b**). (**d**) Overlay of 10 clusters produced using hierarchical clustering onto the t-SNE map in (**a**). Data points are coloured according to cluster membership.

Supplementary Figure 6. Principal component values and matrix plots of principal components for the datasets used in this study. (*i*) For each dataset (**a**-**e**) the values (ordered by magnitude) of the eigenvalues (PCs) are shown. (*ii*) A matrix of pairwise plots of the data points projected onto the first five (4 in the case of (**c**)) PCs illustrating that for most of the datasets significant structure is contained in >2 PCs. The main diagonal is a histogram of data point values for the indicated PCs.

SUPPLEMENTARY REFERENCES

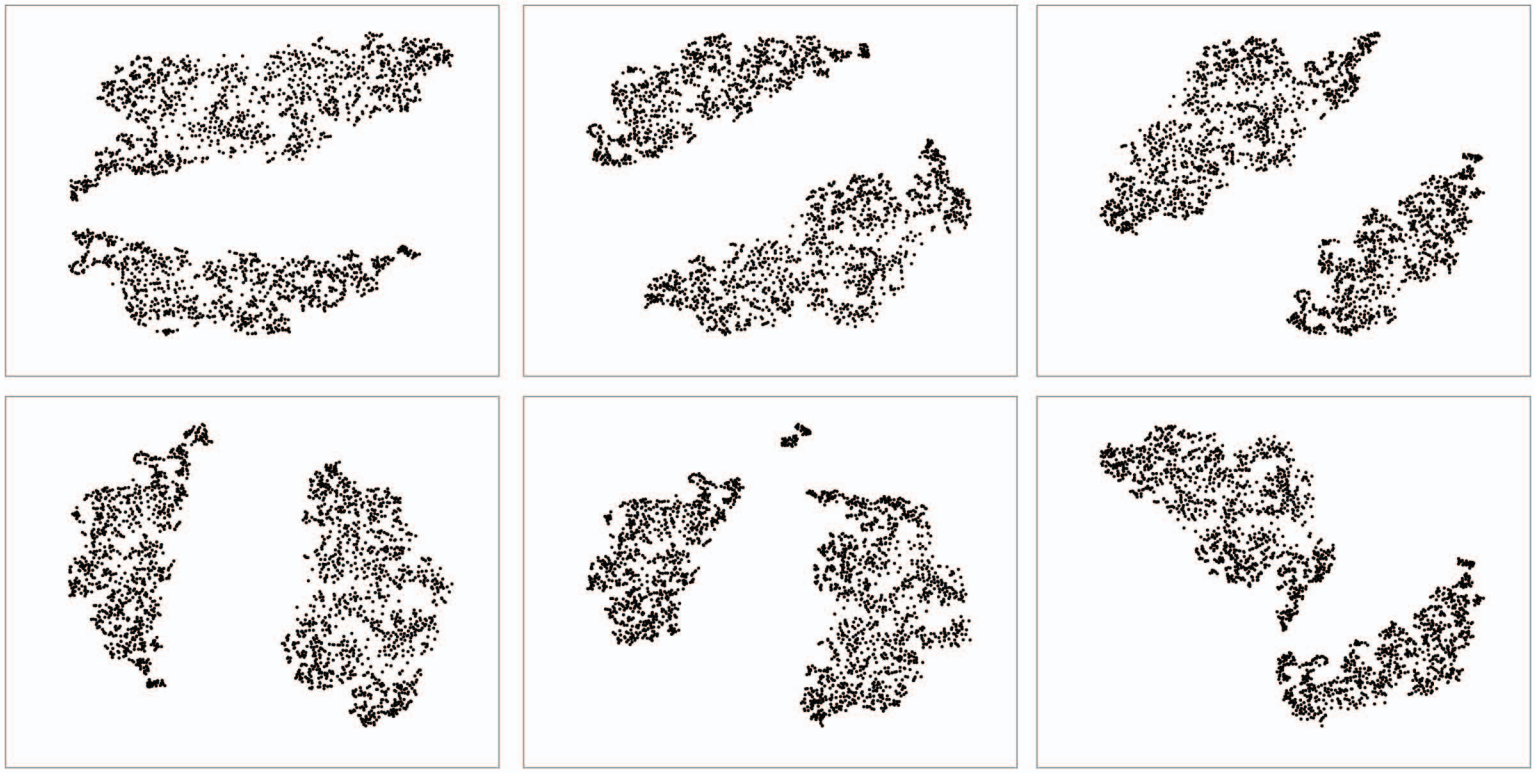
1. Hinton, G. and Roweis, S. (2003) Stochastic Neighbor Embedding. *Neural Information Processing Systems 15 (NIPS'02)*, 857-864.
2. Duda, R. and Hart, P. (1973) Pattern Classification and Scene Analysis. *Wiley-Interscience*.
3. Wylie, C.J., Hendricks, T.J., Zhang, B., Wang, L., Lu, P., Leahy, P., Fox, S., Maeno, H. and Deneris, E.S. (2010) Distinct transcriptomes define rostral and caudal serotonin neurons. *J Neurosci*, **30**, 670-684.
4. Cruz, C., Ribes, V., Kutejova, E., Cayuso, J., Lawson, V., Norris, D., Stevens, J., Davey, M., Blight, K., Bangs, F. *et al.* (2010) Foxj1 regulates floor plate cilia architecture and modifies the response of cells to sonic hedgehog signalling. *Development*, **137**, 4271-4282.
5. Ostrin, E.J., Li, Y., Hoffman, K., Liu, J., Wang, K., Zhang, L., Mardon, G. and Chen, R. (2006) Genome-wide identification of direct targets of the Drosophila retinal determination protein Eyeless. *Genome Res*, **16**, 466-476.
6. Fang, H., Yang, Y., Li, C., Fu, S., Yang, Z., Jin, G., Wang, K., Zhang, J. and Jin, Y. (2010) Transcriptome analysis of early organogenesis in human embryos. *Dev Cell*, **19**, 174-184.
7. Tu, B.P., Kudlicki, A., Rowicka, M. and McKnight, S.L. (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, **310**, 1152-1158.



Bushati_Supplementary Figure 1

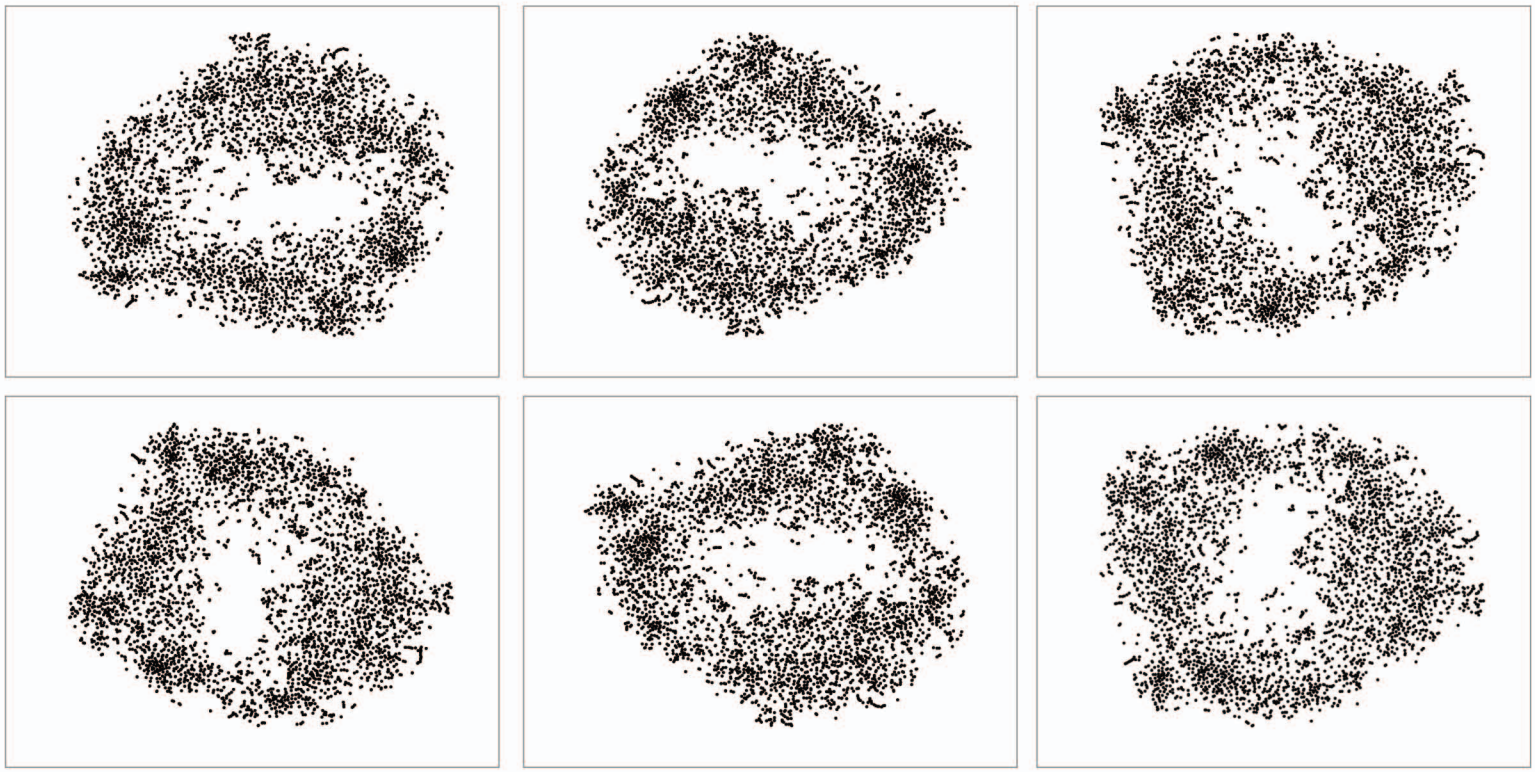
a

Human embryogenesis (Fang et al.)



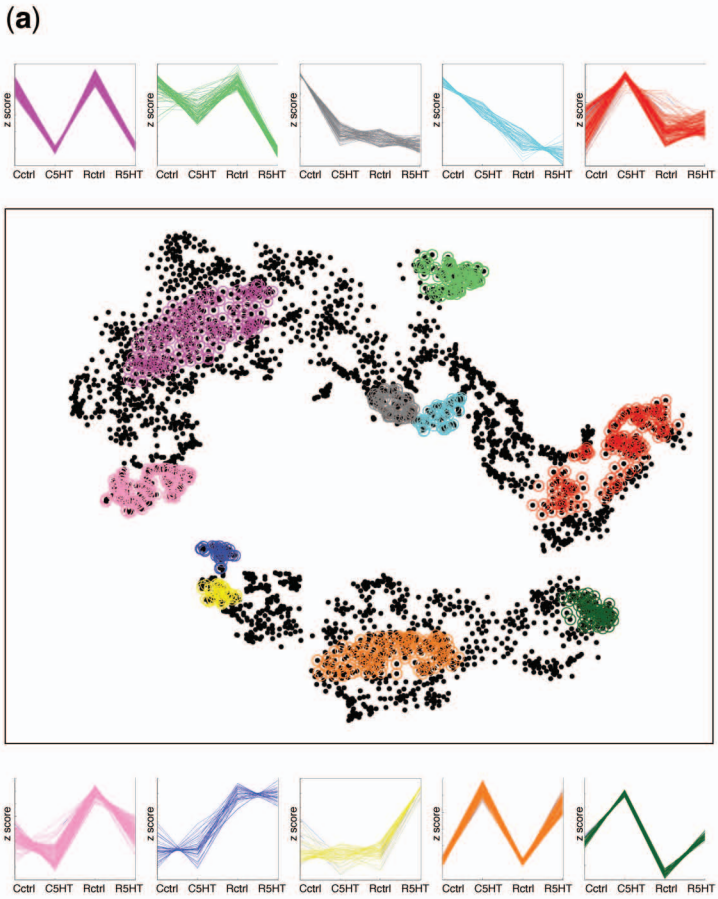
b

Yeast metabolic cycle (Tu et al.)



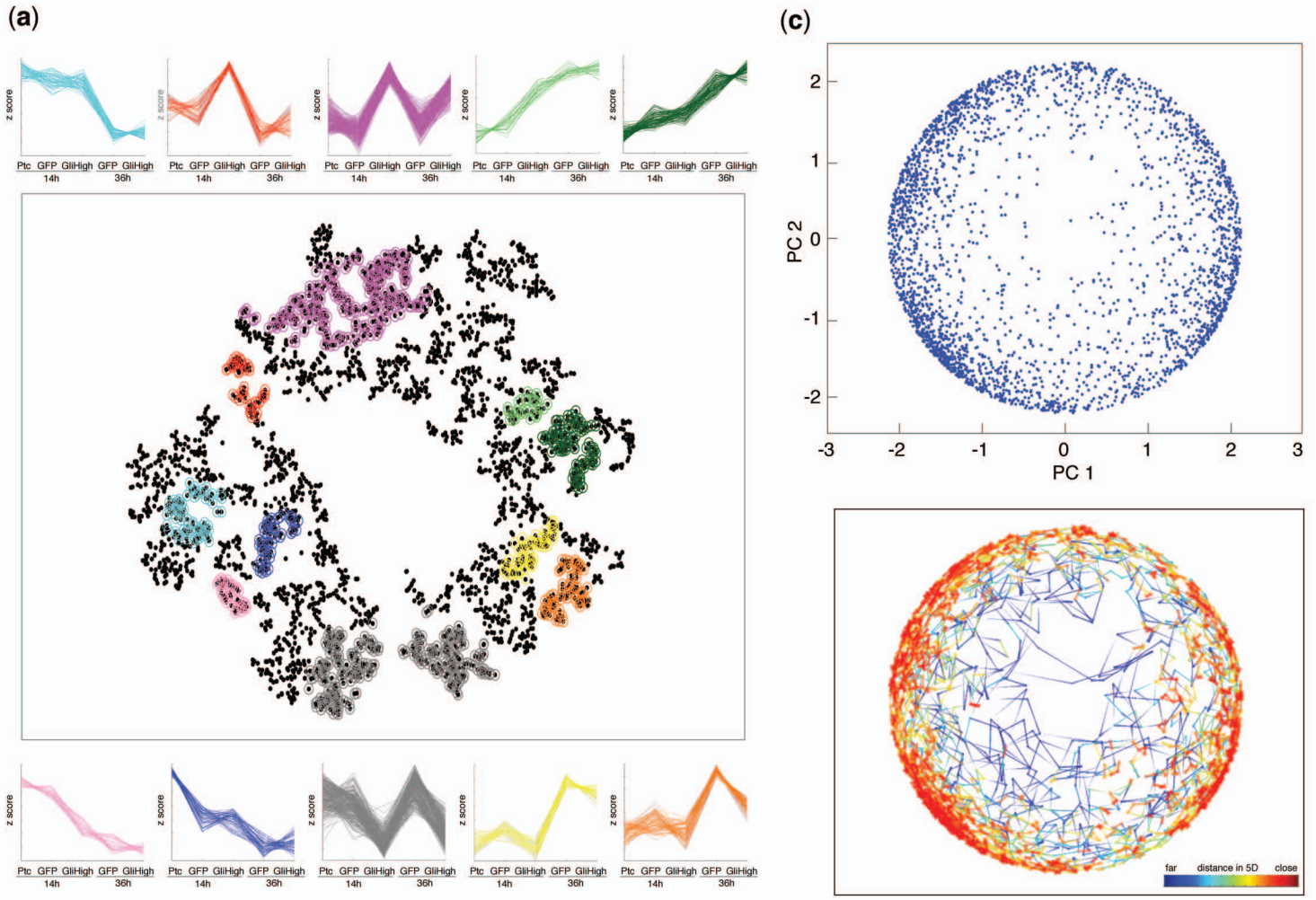
Bushati_Supplementary Figure 2

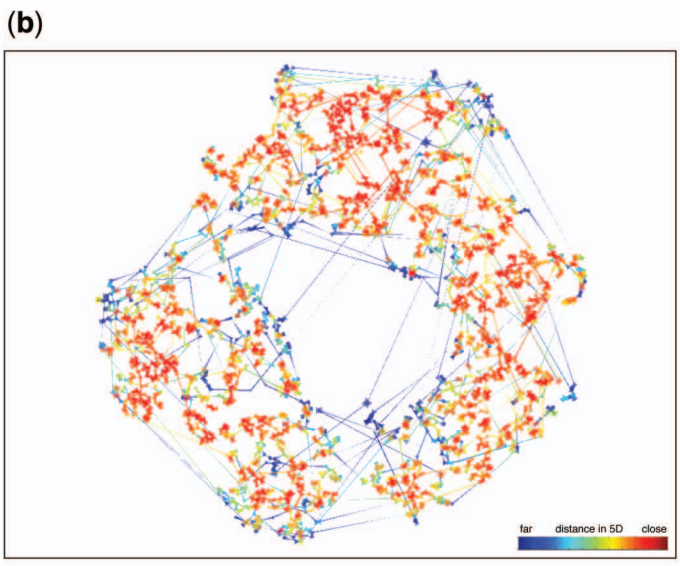
Mouse serotonin neurons (Wylie et al.)

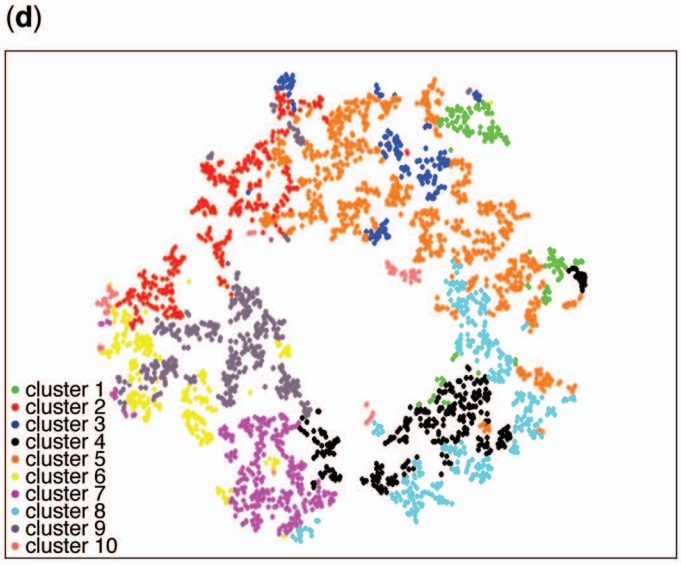


Bushati_Supplementary Figure 3

Chick neural tube cells - Shh signalling (Cruz et al.)

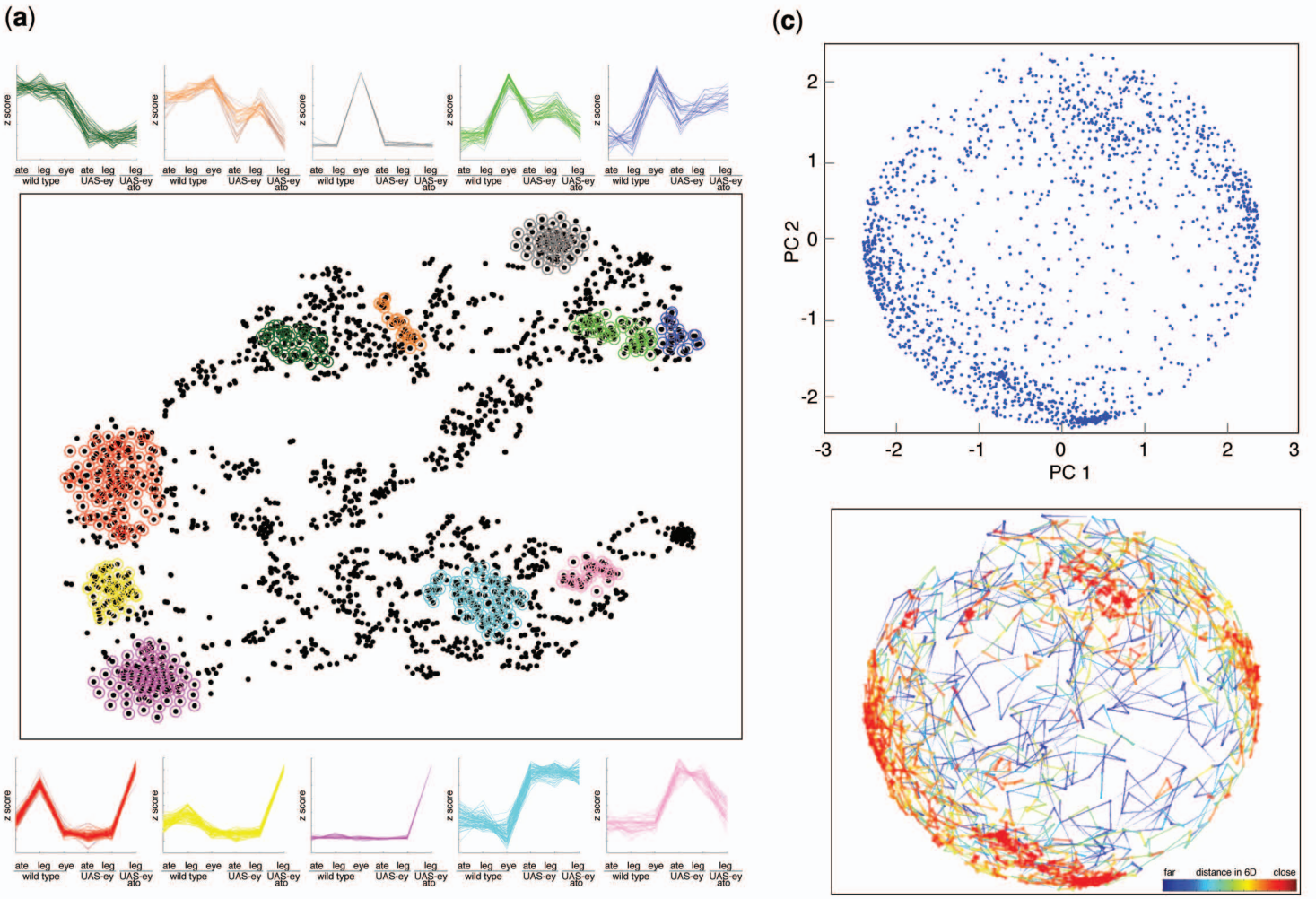


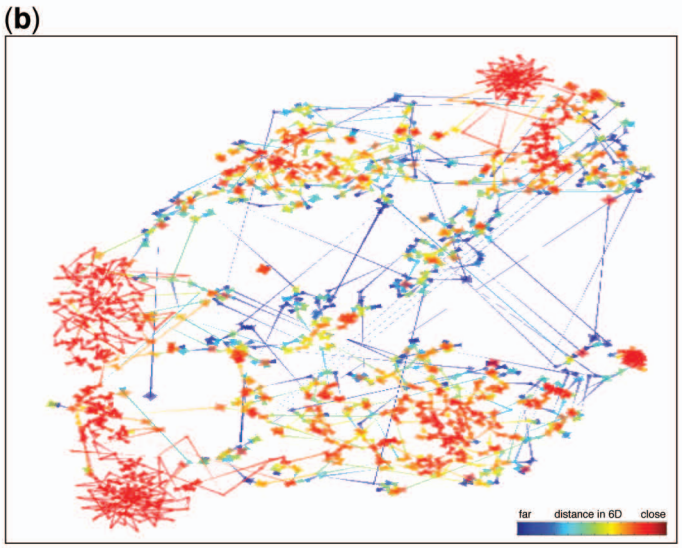


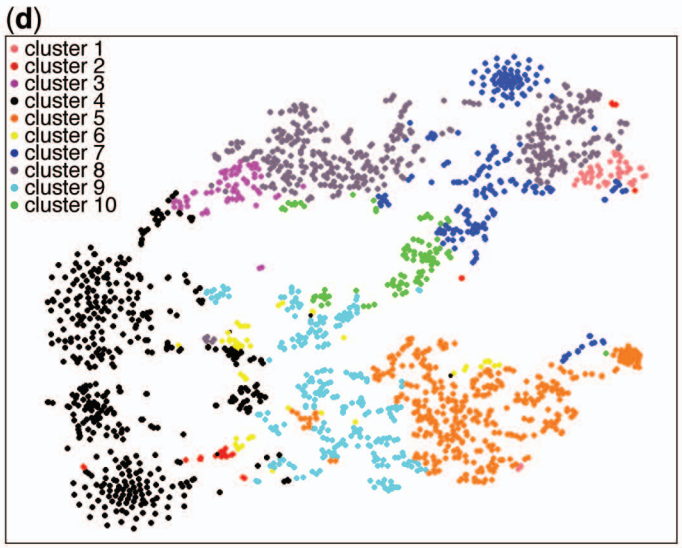


Bushati_Supplementary Figure 4

Drosophila imaginal discs - eyeless misexpression (Ostrin et al.)



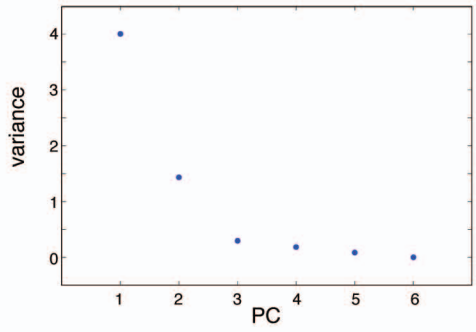


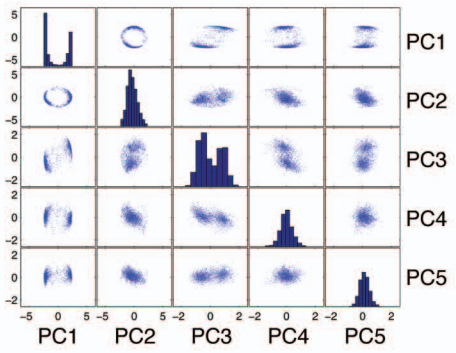


Bushati_Supplementary Figure 5

(**a**)

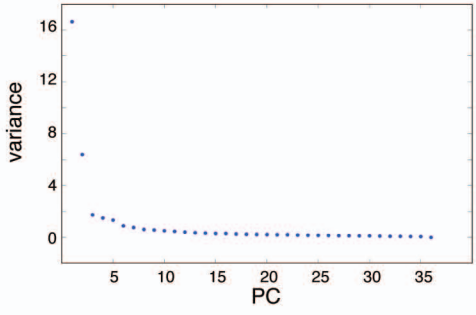
Dataset 1 - Fang et al.
Human embryos
2148 probes
6 conditions

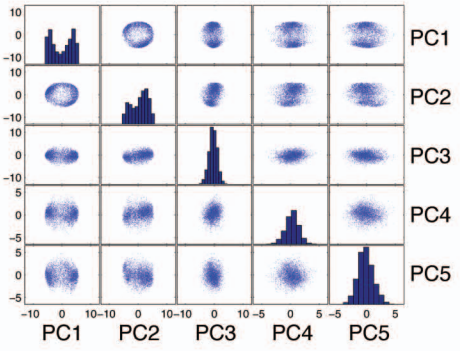




(**b**)

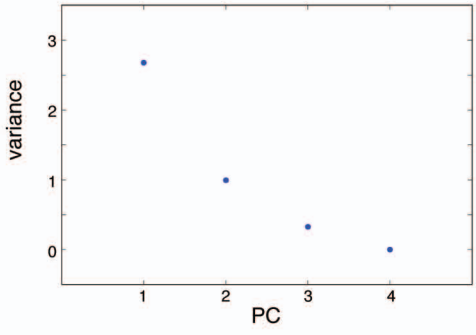
Dataset 2 - Tu et al.
Yeast metabolic cycle
3656 probes
36 conditions

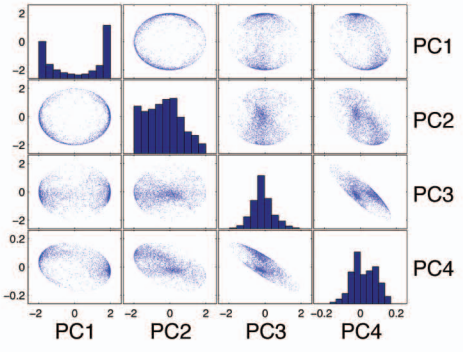




(**c**)

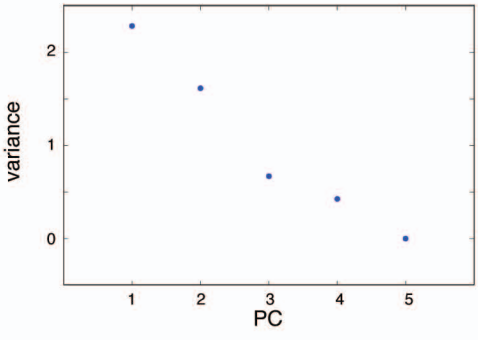
Dataset 3 - Wylie et al.
Mouse 5HT
3079 probes
4 conditions

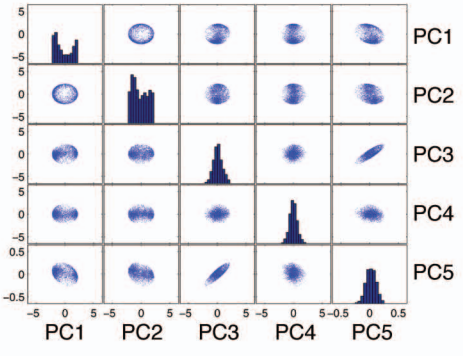




(**d**)

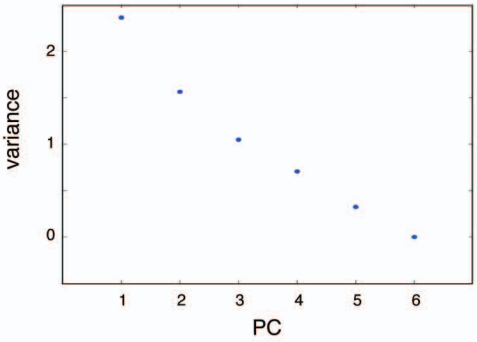
Dataset 4 - Cruz et al.
Chick neural tube
2828 probes
5 conditions

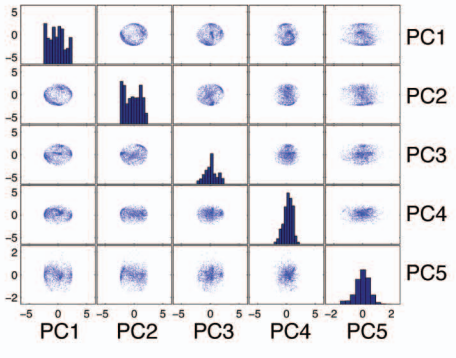




(**e**)

Dataset 5 - Ostrin et al.
Drosophila imaginal discs
1917 probes
6 conditions





Bushati_Supplementary Figure 6