The Drosophila embryo at single-cell transcriptome resolution

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N., Zinzen, R.P.

Science
2017 OCT 13 ; 358(6360): 194-199
2017 AUG 31 (first published online: final publication)
doi: 10.1126/science.aan3235

Publisher: American Association for the Advancement of Science (AAAS)

The *Drosophila* Embryo at Single Cell Transcriptome Resolution

Nikos Karaiskos*[,1], Philipp Wahle*[,2],
Jonathan Alles[1], Anastasiya Boltengagen[1], Salah Ayoub[1], Claudia Kipar[2], Christine Kocks[1],
Nikolaus Rajewsky[$,1], Robert P. Zinzen[$,2]

* contributed equally
[1] Systems Biology of Gene Regulatory Elements,
[2] Systems Biology of Neural Tissue Differentiation,
Berlin Institute for Medical Systems Biology (BIMSB),
Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC),
13125 Berlin, Germany
[$] to whom correspondence should be addressed
nikolaus.rajewsky@mdc-berlin.de
robert.zinzen@mdc-berlin.de

ABSTRACT

By the onset of morphogenesis, *Drosophila* embryos consist of ~6000 cells that express distinct gene combinations. Here, we used single-cell sequencing of precisely staged embryos and devised *DistMap,* a computational mapping strategy to reconstruct the embryo and to predict spatial gene expression approaching single cell resolution. We produce a virtual embryo with ~8000 expressed genes per cell. Our interactive "*Drosophila*-Virtual-Expression-eXplorer" database generates 3D virtual *in situ* hybridizations and computes gene expression gradients (http://dvex.org). We used DVEX to uncover patterned expression of new transcription factors and long non-coding RNAs, as well as signaling pathway components. Spatial regulation of Hippo signaling during early embryogenesis suggests a mechanism for establishing asynchronous cell proliferation. Our approach is suitable to generate *transcriptomic blueprints* for other complex tissues.

INTRODUCTION

Intricate gene regulatory networks produce and maintain complex assemblies of specialized cells such as tissues and organs. To unravel the underlying gene expression dynamics, significant efforts have been made to compare tissue specific materials (*1-4*); however, cell culture often constitutes a poor proxy for *in vivo* complexity and heterogeneity within dissected tissues poses problems (*3-10*). An alternative is isolation of specific cell types via cell sorting (*11-14*); yet, cells are examined in pools, thus obscuring heterogeneity, expression in small subsets may not be detectable and relationships such as expression exclusivity and concomitancy cannot be distilled. This restricts our ability to infer gene regulatory relationships, to predict what functional roles individual cells play and how they integrate with their spatial environment. With the advent of single-cell expression profiling, it has become possible to assess the transcriptomic landscape of complex cell mixtures with single-cell resolution, thereby allowing insights into differentiation trajectories, cell fate decisions, spatial relationships and rare cell types (*15-21*).

The *Drosophila melanogaster* embryo has been an exquisite model for the patterning principles that shape cellular identities. The fertilized egg undergoes 13 rapid nuclear divisions resulting in a syncytial embryo of ~6000 nuclei. By developmental stage 5, nuclei have moved to the surface, become surrounded by cell membranes, and spatial gene expression patterns emerge as cells translate anteroposterior and dorsoventral positional information into transcriptional responses. Stage 6 is marked by the first morphogenetic movements after cellularization completes and gene expression around this stage has been extensively assayed in whole embryos (e.g. (*22*), in mutants converting entire embryos to germ layers (*4*) and in dissected slices (*23*). Available *in situ* databases present systematically annotated spatial gene expression (*24, 25*), but often stop short of single-cell resolution, direct comparison of several genes per cell, assaying transcription of the entire genome including non-coding RNAs, and quantitative assessment.

To overcome these problems, we optimized massively parallel droplet-based single-cell sequencing (Drop-seq) (*19*) and quantified gene expression across >10000 fixed cells from dissociated embryos (*26*) at a median depth of thousands of genes per cell. Computational analysis of the high-resolution *in situ* patterns of 84 genes (*27*) indicated that most if not all cells of the fly embryo have a unique transcriptional identity, highlighting the need to resolve the embryo at single-cell resolution. Previous efforts to map sequenced embryonic cells back to their origin (e.g. (*21*)) did so by reducing mapping complexity (e.g. by binning the entire zebra fish embryo composed of thousands of cells into 128 expression *regions*) and these methods could not correctly map our data at the required resolution. Therefore, we devised a new mapping strategy

and algorithm (*DistMap*) based on spatially distributed scores. The resulting virtual embryo gives access to single-cell transcriptome information at unprecedented spatial resolution (87% of cells in the embryo are confidently resolved) and depth (>8,000 genes/cell).

RESULTS

Deconstructing the transcriptomic state of the embryo

In order to assess transcriptome diversity genome- and embryo-wide with single-cell resolution, we hand-selected embryos at the onset of gastrulation (stage 6, Fig. 1A). Cells were extracted from >5000 precisely staged embryos, methanol fixed (*26*) and sequenced in seven Drop-seq runs corresponding to five biological replicates (Table S2) resulting in a total of ~7975 sequenced *D. melanogaster* cells (Table S3, Fig. 1A, left). The vast majority of the sequenced cells (>90%) represented single-cell transcriptomes, as assessed by mixing cells from stage-matched *D. melanogaster* and *virilis* embryos (Fig. S1A, Table S2, S3). Gene expression correlated well across Drop-seq replicates (R>0.94) and between Drop-seq and stage-matched, unfixed, whole embryos (R>0.88) (Fig. S1B), and was consistent with absolute quantification in individual stage-matched embryos (*28*) (Fig. S1C).

To concentrate on cells with unambiguous patterning information, we excluded pole cells, yolk nuclei and cell doublets from further analysis. Pole cells constitute a discrete lineage and contribute only to the germ line (*29*), while yolk nuclei function primarily in energy metabolism (*30*). Pole cells and yolk nuclei readily separated in a principal component analysis (Fig. S1E). Similarly, t-SNE representation (*31*) correctly groups cells according to cell type, while a central cluster of doublets (cells expressing markers of distinct dorsoventral territories, i.e. mesoderm; neurectoderm; dorsal ectoderm; see Table S4) emerges centrally (Fig. S1E inset, Fig. S1D). Furthermore, we considered only cells with ≥12500 unique transcripts and expressing >5 genes of the BDTNP reference atlas (see below). The remaining ~1300 high quality cells had a median unique transcript number of >20800 mapping to >3100 genes (Fig. S2C). These cells separated along the first two principal components by dorsoventral identities (Fig. S2D), but not by biological replicates (Fig. S2E). The embryos contained a *DsRed* reporter transgene under control of a ventral neurogenic ectodermal *vnd* enhancer (*32*) and *DsRed* transcripts were primarily detected in a subset of cells that also score highly for broad neurectodermal markers (Fig.S2D). We *'in silico* dissected' the embryo by merging the transcriptomes of cells expressing specific markers for various dorsoventral territories and found that the merged transcriptomes accurately

reflected gene expression in the respective domains (Fig. S2A,B). This *dissection* procedure is versatile and represents a distinct advantage of single-cell analysis over traditional bulk analyses (see Supplementary Note 1: '*In silico* dissection').

We concluded that individual cells were sequenced without significant batch effects or bias for particular cellular identities, and that our Drop-seq data accurately reflect the transcriptomic state of individual embryonic cells.

Spatial reconstruction of the embryo

High quality cells could be grouped into 9 prominent cell clusters in the stage 6 embryo (Fig. S2F). Genes well-known for their roles in embryonic patterning and tissue specification were readily identified (Table S5; see Supplementary Note 2: *'Clustering Analysis'*). Spatial expression of cluster-specific driver genes as assessed by *in situ* hybridization (*24, 25*) was similar within clusters and spatial coherence within clusters was further supported by gene ontology (GO) term enrichments (Fig. S3 and Supplementary Note 3: '*GO terms*').

As might be expected, transcription factors were prevalent among the genes that drive cluster identity (Table S5; e.g., *fkh*, *kni, grn*, *hb*, *Abd-B, ken oc,* and *toy*). More surprisingly, several of the most variable genes remain un- or understudied in early development (Table S5; e.g. *DNaseII, Z600, Meltrin, mtt, Atx-1,* and several genes without names (*'CG'*s), though their highly specific expression suggests functional roles in early embryogenesis. Furthermore, numerous long non-coding RNAs (lncRNAs) and a microRNA precursor are among the most variable genes (Table S5; e.g. *CR44683, CR43302, CR43279, CR41257, CR45185*), which indicates a hitherto unrecognized role for lncRNAs in early embryonic patterning and development.

We sought to map cells back to their position of origin to produce a virtual embryo with single-cell transcriptome resolution by using a database of known *in situ* markers (Fig. 1B). The Berkeley *Drosophila* Transcription Network Project (BDTNP) generated *in situ* hybridization data for 84 genes, resulting in a quantitative high-resolution gene expression reference atlas with substantial combinatorial complexity (*27*). To correlate these 84 marker genes with our single-cell transcriptomes, we binarized the BDTNP atlas by manually choosing thresholds for each gene (*24, 25*) (Fig. 2A.I). The combinatorial expression of these 84 binarized BDTNP markers sufficed to uniquely classify almost every position within the embryo (Fig.S5A). Attempts to map our single-cell data using previously published algorithms (*20, 21*) was unsuccessful. We therefore designed a new mapping strategy based on distributed mapping scores (*DistMap*, see Supplement). We binarized the Drop-seq data (Fig. 2A.II, see Supplement), then compared the

profiles of each cell against each bin, collected the (mis)matches into confusion matrices and computed Matthews's correlation coefficients (MCC) for every cell-bin combination (Fig. 2A.III). The result is a distributed mapping score for any sequenced cell across all embryonic positions (Fig. 2A.IV). Fig. 2B explores DistMap's efficacy and demonstrates that sequenced cells mapped to a few cell diameters with high confidence (red) approaching the accuracy of the binarized bins themselves (green), whereas random mapping positions are spread throughout the whole embryo (blue). Due to the high transcriptional complexity, we reasoned that mapping a cell to multiple *likely* positions would be more meaningful than assigning it to a single-location. DistMap covers most of the embryo (Fig. 2C), assigns cells confidently (Fig. S5C) and allows the quantification of many more genes per bin than initially detected in individual sequenced cells, so that most bins exhibit 6500-8500 expressed genes (Fig. 2D).

To compute the spatial expression of a gene we combined normalized gene expression per cell with the MCC scores for every cell-bin pair (see Supplement). This allows querying any given gene across all cells of the virtual embryo and produces a "virtual *in situ* hybridization" (vISH). To assess prediction quality, we computed vISHs of the 84 BDTNP-mapped genes using all high quality cells, as well as from subsamples and compared the resulting vISHs against the BDTNP database. The discrepancies saturate by ~750 cells (Fig. S5B), so that the mapping would only be marginally improved by including more than our full set of ~1300 high quality cells.

To uncover the spatial signature of the nine single-cell clusters described above (Fig. 3A), we calculated the average mapping scores per bin across all cells per cluster (Fig. 3B). The concordance of the spatial signatures of these 9 principle clusters with regionally confined developmental fates (e.g. (*33, 34*)) is striking. While cluster 4 largely encompasses the primordium of the mesoderm, cluster 3 corresponds to the future neurectoderm and ventral epidermis, and cluster 6 corresponds to the dorsal epidermis and extra-embryonic tissues. Cluster 9 in anteroventral regions corresponds the future oesophagus and pharynx, while cluster 2 and 7 will give rise to the anterior- and posterior midgut upon invagination. Furthermore, this spatial cluster mapping is in agreement with GO term enrichments (see Supplementary Note '*GO terms*').

The virtual embryo predicts spatial gene expression

With single-cell transcriptomes confidently mapped onto the embryo, each of the positional bins can be individually queried for gene expression. Our online ***D****rosophila* **V**irtual **E**xpression e**X**plorer (DVEX, www.dvex.org) allows generation of vISHs for single genes and combinations.

Predictions are displayed on a virtual embryo in multiple orientations (Fig. 4A) and expression gradients can be estimated along the anteroposterior and dorsoventral axes (e.g. Fig. 4B). Furthermore, DVEX provides an interactive environment to explore the t-SNE representation, gene expression in clustered cells, and genes driving clustering (e.g. Fig. 4C).

We observed close concordance between vISH predictions and expression detected by RNA *in situ* hybridization for genes expressed in a wide variety of patterns (see Fig. 5, S4), including *vnd*::*DsRed* reporter expression in the ventral neurectoderm. Many of the genes shown were not previously known to be patterned at stage 6. Especially striking were predictions restricted to small patches of cells – expression limited to a few cells is often undetectable in traditional transcriptome studies, but is accurately resolved by vISH (Fig. 5, S4).

vISH can be used to identify genes with distinct spatial patterns. We predicted the spatial expression of the 476 most highly variable genes, clustered their correlation matrix and identified 10 parental branches, which generate *archetypal* expression patterns when averaged (Fig. S5E). These archetypes reflect the predominant transcriptional patterning responses, identify gene sets that respond to similar regulatory cues, and allow the discovery of unstudied and unusual gene expression patterns.

Identification of potential developmental regulators

We generated vISHs for >150 DNA-binding transcription factors that are detectably expressed in the stage 6 embryo (Fig. S6), many of which are un- or understudied with respect to early development. Additionally, we predicted patterned expression of 16 genes that contain DNA binding domains (*35*). These are likely novel transcription factors (Fig. S6A) and we experimentally validated the vISH predictions for two out of two patterned candidates, *CG34224* and *CG10553* (Fig. 5B). This comprehensive overview of transcription factor expression allows spatial assessment of regulator combinations that may activate or restrict target genes locally (Fig. S6B).

Several long non-coding RNAs (lncRNAs) have also been shown to be potent regulators (e.g. (*36*)), but have not been assayed globally and systematically in early *D. melanogaster* development. By screening DVEX, we identified dozens of expressed and patterned lncRNAs (Fig. S6C). The lncRNAs *CR44317*, *CR44691* and *CR45693* for example are weakly expressed, rendering them barely detectable in whole embryo sequencing data (*22*); however, RNA *in situ* hybridization showed reliable transcript signals in the predicted spatial domains (Fig. 5C, S4). Additionally, vISH predictions for *CR45559* and *CR44917* were partially confirmed (Fig. 5C,

S4). The expression patterns of these lncRNAs range from dorsoventral modulation to gap-, terminal- and pair-rule patterns.

CR43432 expression prediction was particularly unusual, as it combines ventral, posterior and dorsal aspects. CR43432 appears to 'wrap around' lateral regions of the embryo and it appeared specifically excluded from the neurectoderm by vISH (Fig. 5D). In fact, expression is strongly anti-correlated with the neurectoderm marker *SoxN* at the single-cell transcriptome level and double ISH confirms mutually exclusive expression (Fig. 5D). Additionally, *CR43432* is highly expressed in yolk nuclei (Fig. 5D, right). The complementary non-neurogenic expression of *CR43432* suggests it might act to delimit neurogenic genes or to promote non-neurogenic fates. In total, we discovered ~40 lncRNAs predicted in a multitude of patterns (Fig. S6C). Taken together, vISH is a powerful tool to discover novel putative regulators of embryonic patterning.

Cell communication by spatially regulated signaling

The Hippo signaling pathway is a major regulator of organ size, cell cycle and proliferation (*37, 38*), but has to our knowledge not been implicated in the early embryo. By querying where transcripts of ligands, ligand modulators, receptors and signal transducers are expressed, we identified patterned expression of major Hippo signaling components along the anteroposterior axis, with overlapping expression primarily in an anterior domain (Fig. 6A). Co-expression of these molecules may promote Hippo signaling, which culminates in the phosphorylation of the transcription factor Yorkie, thereby diminishing Yorkie's nuclear localization (*37, 38*). After mitotic arrest at stage 5, cell cycle re-entry is delayed in an anterior region (*39*) and it is conceivable that active Hippo signaling in that domain delays mitotic onset. Using antibodies against Yorkie and a mitosis marker (phosphorylated histone H3), we detected higher nuclear-cytoplasmic ratios of Yorkie in cells undergoing mitosis in anterior patches at ~stage 7 (Fig. 6B), suggesting active Hippo signaling in intervening regions. To our knowledge Hippo signaling has previously not been implicated in cell-cycle regulation in early *Drosophila* development.

Additionally, we predict that components of other signaling pathways are expressed in a spatially restricted fashion, including alternate ligands, receptors, and antagonists of Dpp/TGFβ (Fig. S7A). Our experimental data suggests anterior repression of the TGFβ signaling cascade (Fig. S7B) (see Supplementary Note 4: '*TGFβ signaling modulation*'). Hence, by analyzing the expression of signaling molecules in the embryo with spatial resolution, we are able to predict where signals originate and where they can be transduced.

Detection of evolutionary gene expression changes

Several *cis*- and *trans*-regulatory circuits have diverged between *D. melanogaster* and *D. virilis* (e.g. (*40, 41*)). We asked whether changes in gene expression patterns over the course of speciation might be detectable using DVEX. Clustering obtained from 673 stage 6 *D. virilis* high quality cells (Fig. S8A) bore a striking similarity *D. melanogaster* with respect to cluster number, proportional cluster size and cluster mapping by vISH (compare Fig. 3, S8A). Gene expression correlation between merged transcriptome data of the two species was high (R=0.77). We used the virtual embryos of *D. melanogaster* and *virilis* to systematically compute vISHs, compare orthologs and identify divergences.

The genes *CG6660* and *GJ14350* are homologous by protein conservation and genomic synteny (Table S8), with *CG6660* predicted not to be expressed in *D. melanogaster* (Fig. S8B, left), while *GJ14350* was predicted to be expressed in an anteroposterior stripe-modulated pattern in *D. virilis* (Fig. S8C, left). By RNA *in situ* hybridization *CG6660* was not detectable, whereas *GJ14350* was expressed in stripes similar to the prediction (Fig. S8C). For the homologous pair *fok*/*GJ17890* (Table S8), *fok* was predicted and verified by *in situ* hybridization to be expressed in an anterior ventral patch in *D. melanogaster* (Fig. S8D), but vISH predicted absence of the anterior patch and weak posterior expression of *GJ17880* in *D. virilis*. *In situ* hybridization in *D. virilis* showed that, while there is a tendency for low posterior expression of *GJ17880* as early as stage 6/7 by RNA *in situ* hybridization, robust posterior staining was not seen until stage 8; however, the absence of anterior expression in *D. virilis* was confirmed (Fig. S8E). These examples illustrate that DVEX can serve as a sensitive tool for the identification of gene expression changes.

Discussion

Here, we resolved a metazoan embryo composed of ~6000 (or ~3000 when considering bilateral symmetry) individual cells. While the *Drosophila* embryo may be an extreme example where each cell has a unique transcriptional profile, the transcriptomes of neighboring cells can be very similar to each other. To successfully map dissociated and sequenced cells to their correct spatial position based on combinatorial expression of marker genes, we required a suitable set of marker genes, deep capture of gene expression in each cell, and powerful computational mapping to be able to confidently score differences between an enormous number of possible mapping

possibilities. To illustrate the latter point, if considering only 1000 sequenced cells across only 1000 locations, one already would have to calculate *one million* possibilities.

We were able to overcome these challenges and to produce a '*virtual embryo*' with ~8000 genes per cell due to three main reasons. Firstly, the 84 *in situ* markers used captured sufficient spatial transcriptional complexity to allow us to guide mapping of each sequenced cell to its positions. Secondly, we optimized our Drop-seq approach to reliably capture thousands of genes per cell. Thirdly, and perhaps most importantly, we devised a novel mapping strategy, '*DistMap*', which reliably maps single-cell transcriptomes back to their origin. *DistMap* is scalable and extendable to other 3D tissues at single-cell resolution. This is because *DistMap* uses measured gene expression and does not require transcript-level imputation and its scoring scheme is suitable for sparse datasets. Additionally, distributed mapping limits the effect of outliers and populates positions with transcript information beyond the base sequencing level; in this way, from an original depth of ~3500 genes captured/cell, we were able to assign on average ~8000 genes/cell. Nevertheless, *DistMap* clearly can be improved in several respects; for example, it currently uses binarized rather than continuous data and maps each cell independently, rather then allowing mapped cells to improve subsequent scoring.

Once a virtual embryo has been produced, what kind of biology can be learned? We first built a computational platform (www.dvex.org) that allows interactive interrogation of single-cell transcriptome data in spatial context, including the computation of gradients. We then leveraged DVEX to compute thousands of virtual *in situs* and to select genes that had 'interesting' expression patterns. For example, we identified patterned transcription factors never implicated in early development before, as well as dozens of lncRNAs with intriguing and sometimes novel expression patterns. Since we had used a second fly species to control for cell doublet frequency, we *en passant* acquired a virtual embryo for *Drosophila virilis*. Even though these species are separated by at least 40 million years of evolution and have clearly diverged *cis*-regulatory DNA sequences (e.g. (*42, 43*)), we found only a few cases with clear expression divergence, which highlights strong selection pressure on maintaining gene expression patterns at this early stage. It also suggests a large extent of gene regulatory plasticity where *cis*-regulatory sequences may diverge, while the overall expression patterns remain largely unchanged.

We uncovered a substantial amount of transcriptional modulation of components of major signaling pathways. Local expression of ligands sets up 'signal sources', but the ability to respond to these signals appears to be heavily regulated at the transcriptional level, even early in development – from patterned expression of specific receptor molecules, to modulators of signal

transduction. One such case is Hippo signaling, which has not been described to play a role in early *Drosophila* development to date. Active Hippo signaling has been connected to cell cycle delay and diminished proliferation (*38*). Thus, the prediction of expression of major Hippo pathway components in an anterior subdomain (Fig. 6A) was of interest. Indeed, we detected evidence of productive Hippo signaling by showing that the transcriptional effector Yki is diminished in anterior nuclei that do not undergo mitosis (Fig. 6B). More than 30 years ago, Volker Hartenstein and Jose Campos-Ortega employed fuchsin staining to show that mitotic reentry after stage 6 occurs asynchronously (*39*). Our data shows that localized Hippo signaling constitutes a mechanism that breaks synchronicity of cell cycle re-entry in early fly embryogenesis.

In general, how many guide *in situs* are needed to reconstruct tissues after dissociation and single cell sequencing? The answer depends, apart from sequencing depth, clearly on the transcriptional complexity and developmental stage of the tissue. In early metazoan development, most decisions about spatial identity are carried out by a temporal cascade of combinations of transcription factors (*44*). In our case, 84 *in situs* (mostly transcription factors) sufficed to uniquely and individually label most of the ~6000 cells. However, it may be possible to assemble complex tissues from sequenced cells *without* using *in situ* markers as guides, somewhat akin to solving a puzzle. Clearly, we need a better understanding of the "design principles" of gene regulation to achieve this or to test ideas about these principles. For example, in early development, the expression of most genes generally does not change in a discontinuous fashion from cell to cell. This feature could be implemented in future versions of *DistMap* to reduce the number of guide expression patterns needed.

ACKNOWLEDGEMENTS

We thank Mark Biggin and Soile Keranen (LBNL) for discussions and unpublished BDTNP data, Rahul Satija for initial DropSeq help, Sarah Ugowski (MDC) for experimental assistance, Steve Small (NYU) for a transgenic *Drosophila* line, Angelike Stathopoulos (Caltech, NIH R35GM118146) for sharing unpublished results. Julia Zeitlinger (Stowers) for Yki antibody, Ed Laufer (Columbia) for pMad antibody. Nir Friedman (Hebrew University) and members of the Rajewsky and Zinzen labs for constructive discussions, Dan Munteanu (BIMSB/MDC) for IT support and the DFG (SPP 1738, RA 838/8-1, RA 838/5-1) for funding. Raw and processed data sets are available from the GEO repository (GSE95025). The *DistMap* R-package is available at https://github.com/rajewsky-lab/distmap.

NR, RZ defined strategy, supervised, procured funding; NK, PW, JA, CKo, NR, RZ designed experimental strategy; PW did fly genetics, embryo collections; JA set-up, CKo supervised and JA, SA, AB, CKo performed Drop-seq; AB, SA, PW prepared sequencing libraries; NK developed and implemented computational analyses/tools including *DistMap* and DVEX; PW, CKi validated predictions experimentally; NK, PW, CKo, NR, RZ analyzed data and wrote manuscript.

REFERENCES

1. S. K. Bowman *et al.*, *Elife* **3**, e02833 (2014).
2. L. Christiaen *et al.*, *Science* **320**, 1349-1352 (2008).
3. N. Soshnikova, D. Duboule, *Science* **324**, 1320-1323 (2009).
4. A. Stathopoulos *et al.*, *Cell* **111**, 687-701 (2002).
5. L. Cherbas *et al.*, *Genome Res* **21**, 301-314 (2011).
6. T. S. Mikkelsen *et al.*, *Nature* **448**, 553-560 (2007).
7. N. C. Riddle *et al.*, *Genome Res* **21**, 147-163 (2011).
8. M. Stoeckius *et al.*, *EMBO J* **33**, 1751-1766 (2014).
9. M. Stoeckius, D. Grun, N. Rajewsky, *EMBO J* **33**, 1740-1750 (2014).
10. T. Schauer *et al.*, *Cell Rep* **5**, 271-282 (2013).
11. S. Bonn *et al.*, *Nat Protoc* **7**, 978-994 (2012).
12. A. Handley, T. Schauer, A. G. Ladurner, C. E. Margulies, *Mol Cell* **58**, 621-631 (2015).
13. V. M. Weake *et al.*, *Genes Dev* **25**, 1499-1509 (2011).
14. F. A. Steiner *et al.*, *Genome Res* **22**, 766-777 (2012).
15. D. A. Jaitin *et al.*, *Science* **343**, 776-779 (2014).
16. K. Shekhar *et al.*, *Cell* **166**, 1308-1323 e1330 (2016).
17. D. Grun *et al.*, *Nature* **525**, 251-255 (2015).
18. A. M. Klein *et al.*, *Cell* **161**, 1187-1201 (2015).
19. E. Z. Macosko *et al.*, *Cell* **161**, 1202-1214 (2015).
20. K. Achim *et al.*, *Nat Biotechnol* **33**, 503-509 (2015).
21. R. Satija *et al.*, *Nat Biotechnol* **33**, 495-502 (2015).
22. B. R. Graveley *et al.*, *Nature* **471**, 473-479 (2011).
23. P. A. Combs, M. B. Eisen, *PLoS One* **8**, e71820 (2013).
24. E. Lecuyer *et al.*, *Cell* **131**, 174-187 (2007).
25. P. Tomancak *et al.*, *Genome Biol* **8**, R145 (2007).
26. J. Alles *et al.*, *BMC Biol* **15**, 44 (2017).
27. C. C. Fowlkes *et al.*, *Cell* **133**, 364-374 (2008).
28. J. E. Sandler, A. Stathopoulos, *Genetics* **202**, 1575-1584 (2016).
29. E. M. Underwood, J. H. Caulton, C. D. Allis, A. P. Mahowald, *Dev Biol* **77**, 303-314 (1980).
30. M. G. Riparbelli, G. Callaini, *Mech Dev* **120**, 441-454 (2003).
31. L. Van der Maaten, G. Hinton, *Journal of Machine Learning Research* **9**, 2579-2605 (2008).
32. M. Markstein *et al.*, *Development* **131**, 2387-2394 (2004).
33. G. M. Technau, J. A. Campos-Ortega, *Roux Arch Dev Biol.* **194**, 196-212 (1985).
34. V. Hartenstein, *Atlas of Drosophila Development*. (Cold Spring Harbor Laboratory Press, 1993), vol. 1, pp. 57.
35. L. S. Gramates *et al.*, *Nucleic Acids Res* **45**, D663-D671 (2017).
36. L. A. Goff, J. L. Rinn, *Genome Res* **25**, 1456-1465 (2015).
37. H. Oh, K. D. Irvine, *Development* **135**, 1081-1088 (2008).
38. J. Huang *et al.*, *Cell* **122**, 421-434 (2005).
39. V. Hartenstein, J. A. Campos-Ortega, *Roux Arch Dev Biol.* **194**, 181-195 (1985).
40. M. Treier, C. Pfeifle, D. Tautz, *EMBO J* **8**, 1517-1525 (1989).
41. R. P. Zinzen *et al.*, *Dev Cell* **11**, 895-902 (2006).
42. E. Emberly, N. Rajewsky, E. D. Siggia, *BMC Bioinformatics* **4**, 57 (2003).
43. M. Z. Ludwig, C. Bergman, N. H. Patel, M. Kreitman, *Nature* **403**, 564-567 (2000).
44. E. Davidson, *The Regulatory Genome. Gene Regulatory Networks In Development And Evolution.*, (Academic Press, ed. 1, 2006), pp. 304.

Figure legends

Figure 1: De- and reconstructing the embryo by single-cell transcriptomics combined with spatial mapping.

(**A**) ~1000 hand-picked stage 6 fly embryos are dissociated per Drop-seq replicate → cells are fixed and counted → single cells are combined with barcoded capture beads, libraries are prepared and sequenced. → Single-cell transcriptomes are deconvolved, resulting in a digital gene expression matrix for further analysis.

(**B**) Single-cell transcriptomes are correlated with high-resolution gene expression patterns across 84 marker genes → cells are mapped to positions within a virtual embryo → virtual *in situ* hybridization of individual genes predicts spatial gene expression.

Figure 2: Reconstructing the embryo by spatial mapping based on distributed scores.

(**A**) DistMap. The 84 BDTNP gene expression patterns (I) and the single-cell expression profiles (II) were binarized. (III) Confusion matrices are calculated scoring expression (dis-)agreement between the transcriptomes and the ~3000 positional bins of the reference atlas. Matthews's correlation coefficients (MCCs) are calculated for every cell/bin combination. (IV) Positional assignment for each cell is distributed based on MCCs across all bins.

(**B**) Density plot showing mapping confidence (mean Euclidean distance) between a cell's highest scoring location and the following six. Single-cell transcriptomes (red) map to embryo positions with similar confidence as cells of the reference atlas (green).

(**C**) Bin coverage across the embryo. >87% of all locations in the embryo are confidently covered (p-value < 0.05, see Methods for details).

(**D**) The virtual fly embryo has a resolution of 6000-8000 genes/cell.

Figure 3: Sequenced cells cluster by spatial identity.

(**A**) 2D t-SNE representation of the high quality cells shows 9 major clusters grouped by transcriptome similarity.

(**B**) Mapping of clusters reveals that cells within each cluster share a contiguous spatial domain.

Figure 4: DVEX accurately predicts spatial gene expression patterns.

DVEX (www.dvex.org) is the online resource for the virtual embryo.

(**A**) Virtual *in situ* hybridization (vISH) for, the pair rule gene *ftz* (red) and the mesodermal gene *sna* (green) in five orientations. Stippled box indicates cells analyzed in **(B)**. EL, egg length; DV, dorsoventral; AP, anteroposterior.

(**B**) Quantification of relative expression per cell mapped along an axis (here dorsoventral) for *stumps* (expressed in the ventral mesoderm, left) and the *vnd::DsRed* reporter (primarily expressed in the ventral neurectoderm, right). Relative expression in log space, thresholds were 0.85, embryos are oriented anterior left.

(**C**) Examples of marker genes and their expression in t-SNE clustered cells. Expression indicated, grey (low) to red (high).

Figure 5: Prediction accuracy and detection of new regulators.

(**A**) vISH predictions are accurate across a wide variety of expression patterns. Expression of CGs had not been reported previously.

(**B**) Patterned expression of putative transcription factors.

(**C**) Patterned expression of lncRNAs.

(**D**) *CR43432* and pan-neurogenic genes are expressed in complimentary patterns. Dual vISH of *SoxN* and *CR43432* (top-left)), double *in situ* hybridization validates the predicted expression. *CR43432* is additionally expressed in yolk nuclei (not shown in vISH).

Figure 6: Spatial regulation of Hippo signaling in the embryo.

(**A**) vISHs predict patterned expression of Hippo signaling components in stage 6 embryos; most are expressed in anterior regions.

(**B**) Shown is the anterior of a stage 7 embryo, cephalic furrow indicated by stippled white line, anterior left. Staining with antibodies against phosphorylated histone H3 (H3S10-P) marks cells undergoing mitosis, nuclear Yorkie (Yki) is depleted in cells not marked by H3S10-P.

Figure 1, Karaiskos, Wahle et al. “Single Cell Embryo”

double column width

page height

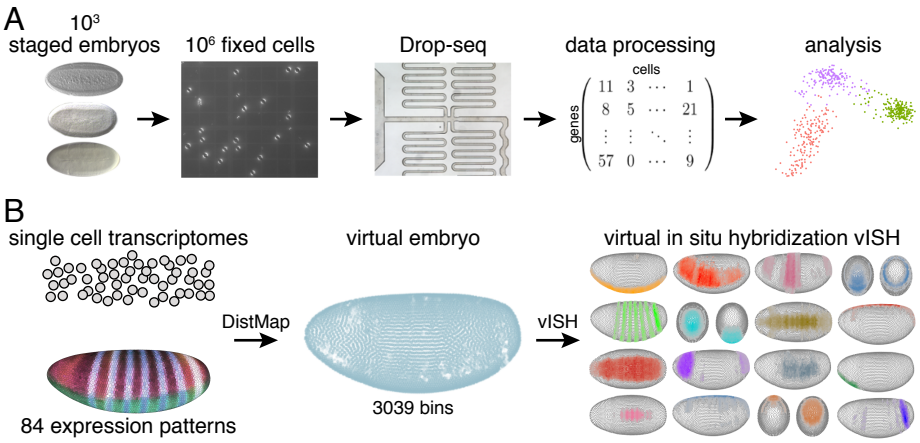


Figure 2, Karaiskos, Wahle et al. “Single Cell Embryo”

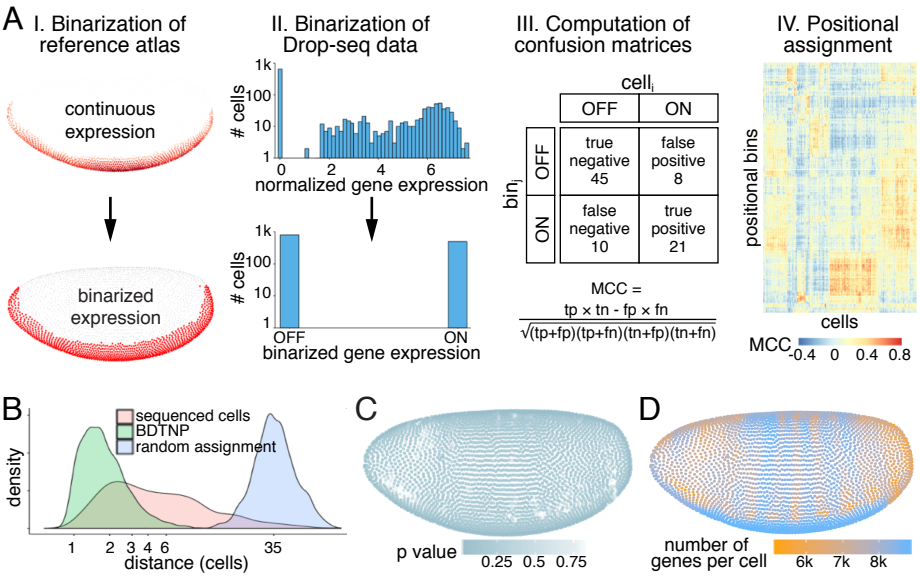


Figure 3, Karaiskos, Wahle et al. “Single Cell Embryo”

single column width

page height

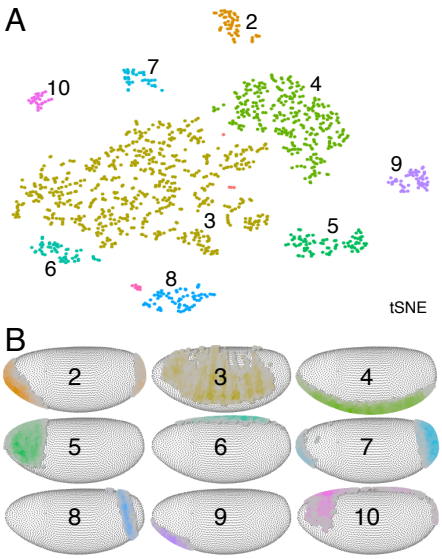


Figure 4, Karaiskos, Wahle et al. “Single Cell Embryo”

double column width

page height

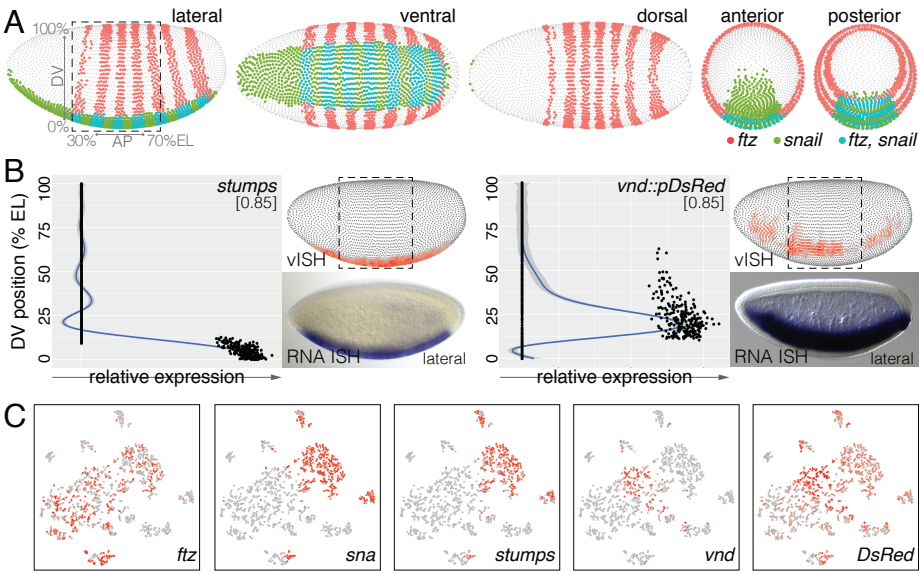


Figure 5, Karaiskos, Wahle et al. “Single Cell Embryo”

double column width

page height

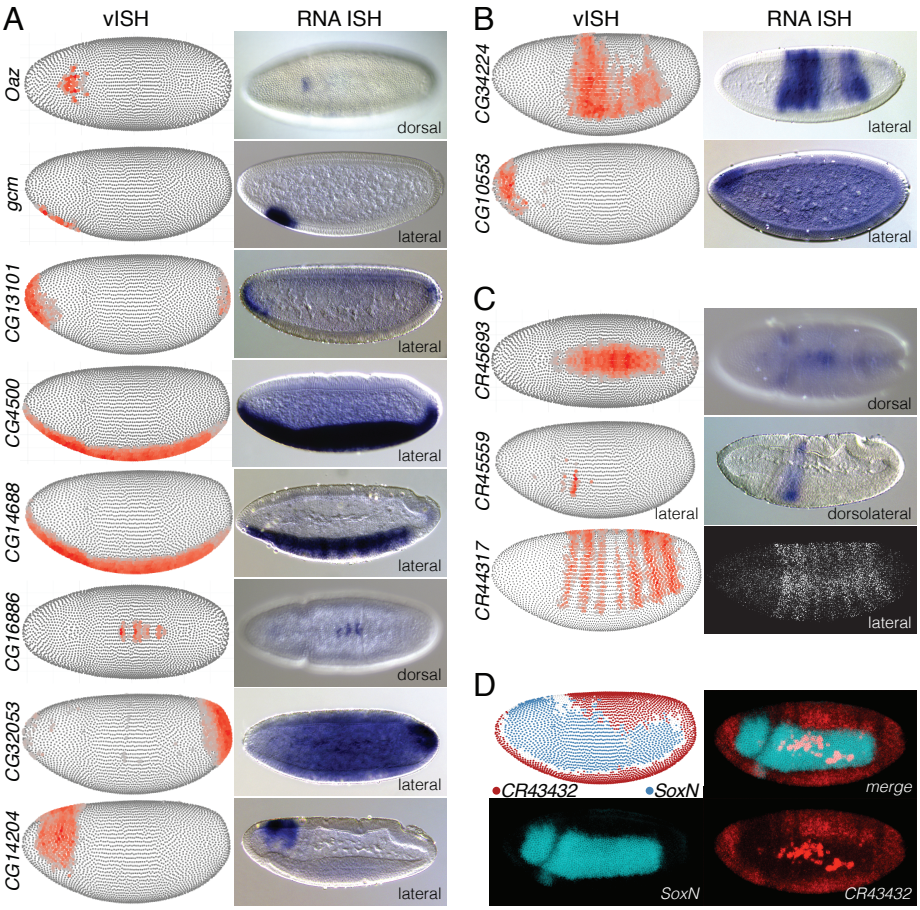
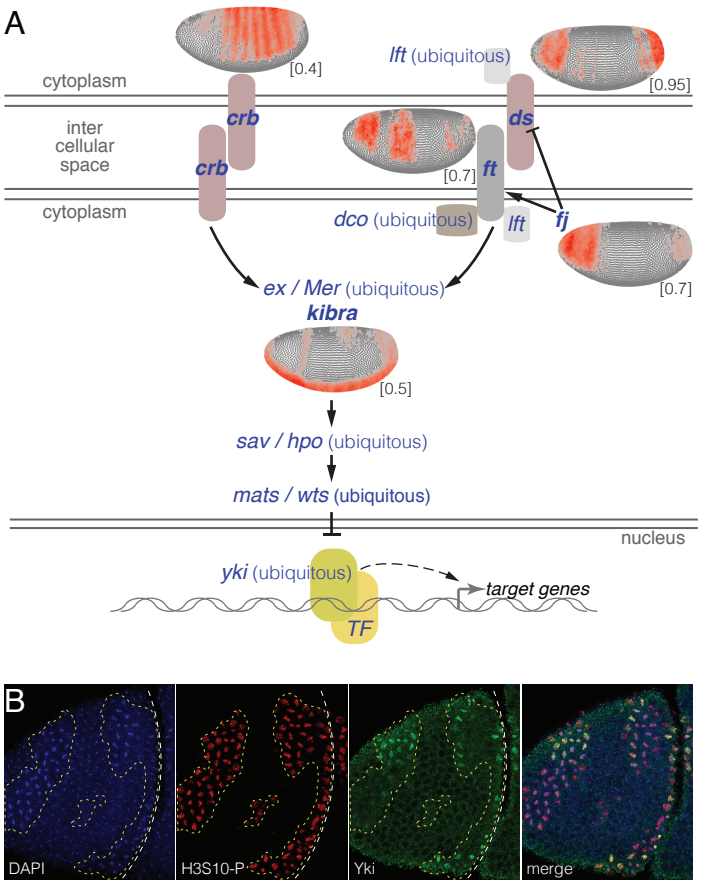


Figure 6, Karaiskos, Wahle et al. “Single Cell Embryo”

1.5 column width

page height



Supplemental

MATERIALS AND METHODS

to

The *Drosophila* Embryo at Single Cell Transcriptome Resolution

Nikos Karaiskos*[,1], Philipp Wahle*[,2],
Jonathan Alles[1], Anastasiya Boltengagen[1], Salah Ayoub[1], Claudia Kipar[2], Christine Kocks[1],
Nikolaus Rajewsky[$,1], Robert P. Zinzen[$,2]

* contributed equally
[1] Systems Biology of Gene Regulatory Elements,
[2] Systems Biology of Neural Tissue Differentiation, Berlin Institute for Medical Systems Biology (BIMSB), Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), 13125 Berlin, Germany
[$] to whom correspondence should be addressed
robert.zinzen@mdc-berlin.de
nikolaus.rajewsky@mdc-berlin.de

Table of Contents

Supplementary Methods 3
Fly Strains, *in situ* hybridizations and immunochemistry 3
Embryo collection and cell isolation for sequencing 4
Drop-seq procedure, single cell library generation and sequencing. 5
Single cell RNA-seq: data processing, alignment and gene quantification. 6
Species separation, cell filtering and data normalization. 6
Correlation of gene expression measurements. 7
Marker genes and primordium assignment. 7
Identification of highly variable genes, principal component analysis clustering and t-SNE. 7
Binarization of the BDTNP reference atlas. 8
Binarization of Drop-seq data. 8
Spatial mapping algorithm: *DistMap*. 9
Simulations to assess bin coverage & cell mapping confidence. 9
vISH computation. 10
vISH visualization. 10
Noise simulations to assess vISH threshold values. 10
Clustering and discovery of archetypal *in situ* patterns. 11
Discovery of new markers with localized expression patterns. 11
Correlation of Drop-Seq and nCounter gene expression analysis. 11
Gene Ontology (GO) Term Enrichment analysis 12
Supplementary Notes 13
Supplementary Note 1: *In silico* dissection of the embryo 13
Supplementary Note 2: Single cell clustering resolves spatial identities 13
Supplementary Note 3: Gene ontology (GO) term analysis 14
Supplementary Note 4: Modulation of TGFβ/Dpp signaling 15
Supplementary References 16
Supplementary Tables 17
Supplementary Table S1: RNA in situ hybridization probes 17
Supplementary Table S2: Drop-seq run information. 18
Supplementary Table S3: Drop-Seq statistics 18
Supplementary Table S4: *Drosophila* embryo marker genes. 19
Supplementary Table S5: Highly variable genes in tSNE clusters. 19
Supplementary Table S6: Pole cell marker genes (available separately online). 19
Supplementary Table S7: BDTNP reference atlas – thresholds and expression matrix (available separately online). 19
Supplementary Table S8: Homology analysis (available separately online). 19
Supplementary Figures and Figure Legends 20
Supplementary Figure S1: Drop-seq data quality and reproducibility. 20
Legend to Supplementary Figure S1 21
Supplementary Figure S2: High quality cells encode spatial patterning information. 22
Legend to Supplementary Figure S2 23
Supplementary Figure S3: GO term analysis of single cell clusters 24
Legend to Supplementary Figure S3 26
Supplementary Figure S4: vISH predictions & experimental validations 27
Legend to Supplementary Figure S4 28
Supplementary Figure S5: Mapping efficacy, vISH thresholds & expression archetypes. 29
Legend to Supplementary Figure S5 30
Supplementary Figure S6: vISH predictions of 155 transcription factors and 42 lncRNAs. 31
Supplementary Figure S7: *Dad* expression inhibits anterior Dpp signaling. 35
Supplementary Figure S8: vISH detects evolutionary changes in expression patterns 36

SUPPLEMENTARY METHODS

Fly Strains, *in situ* hybridizations and immunochemistry

The *Drosophila* strains used were *D. virilis* w^e (*Drosophila* Species Stock Centre, # 15010-1051.17) *and D. melanogaster* y^1 w^{1118} *; P{st.2::Gal4} ; P{vnd::dsRED},* where *st.2::Gal4* was crossed in from a kind gift by Steve Small (NYU) and *vnd::dsRED* was created by standard P-element transgenesis (*1*) after placing the early *vnd* enhancer (*2*) into *pRED-HStinger* (*3*).

RNA *in situ* hybridizations (ISH) were done according to standard procedures (*4*). Briefly, embryos of the genotypes used for cell isolation were collected for ca. 1.5hrs, aged for another 2.5hrs, then dechorionated and chemically cross-linked in 4% formaldehyde for 25 minutes before washing and dechorionation; fixed embryos were stored at -20°C until use. Antisense RNA *in situ* probes were generated by in vitro transcription using either ESTs obtained from the *Drosophila* Gene Collection (DGRC and BDGP), from PCR products using T7 priming sequences encoded in the reverse primers, or after T/A-cloning into pCRII (TOPO-Dual, Invitrogen). In all cases, inserts and orientations were sequence verified. Anti-sense probes were transcribed with DIG-UTP or FitC-UTP (Roche), hybridized overnight, and visualized using anti-DIG antibodies covalently linked to alkaline phosphatase (Roche). NBT+BCIP staining in whole mount embryos was monitored in a staining dish under a dissection microscope, stopped by washing in PBT, embryos were equilibrated to ethanol, washed with xylene and mounted in Permount (Fisher). Fluorescent *in situ* hybridization was done as described above, except 1° antibody was anti-hapten-HRP conjugates (Roche) followed by blocking with Roche western block reagent (stock used as 5x) in PBT, tyramide signal amplification with FitC or Cy3 (Perkin Elmer), washing in PBT and mounting in ProLong Gold with DAPI (ThermoFisher). Probes and primers are listed In Table S1. Immunohistochemistry was done according to standard procedures, using over-night incubation with 1° antibodies Rabbit-α-pSmad (from Ed Laufer, Columbia (*5*)) at 1:500, Rabbit-α-Yki (from Julia Zeitlinger, Stowers (*6*)) at 1:1000, and Mouse-anti-HistoneH3-phospho-S10 (Abcam #14955) at 1:1000. 2° antibodies were Alexa555 or Alexa 488 conjugates (LifeTechnologies). Imaging was done using Nomarski contrast on a Leica DMi8 microscope for colorimetric stains, or on a Leica SP8 scanning confocal microscope for fluorescence. Confocal Z-stacks were acquired being careful not to saturate signal, Confocal stacks were intensity projected using FIJI(ImageJ) (*7*).

Embryo collection and cell isolation for sequencing

For bulk sequencing, ~40 *D. melanogaster* or *D. virilis* stage 6 embryos were collected as described below. PBT was removed, embryos were shock frozen in liquid nitrogen and broken up using pestles. RNA was extracted by standard TRIZOL® extraction. Quality of extracted RNA was assessed using a BioAnalyzer (Agilent Technologies, Wilmington, DE, USA). Libraries were generated from 200ng extracted RNA with the NEBNext® Ultra™RNA Library Prep Kit for Illumina® using poly(A) mRNA magnetic isolation. Samples were sequenced on the Illumina NextSeq500 platform using 75bp single end sequencing chemistry. Bulk sequencing libraries are deposited in GEO under accession numbers GSM 2494790 (*D. melanogaster*, mel_bulk) and GSM 2494791 (*D. virilis*, vir_bulk).

For single cell sequencing, *D. melanogaster* or *D. virilis* were allowed to lay eggs on apple juice-agar plates for 1 hr. intervals and aged for ~2:30hrs (*D. melanogaster*) or ~3:30hrs (*D.virilis*) at room temperature. Embryos were dechorionated for 1 min in ~4% NaOCl and extensively washed with deionized water and rinsed with PBT (1xPBS, 0.1% Triton X100). Excess liquid was removed and embryos were transferred to 5% agarose gel slices stained light blue with Coomassie Brilliant Blue for contrast. Embryos were monitored under a stereomicroscope and stage 6 embryos were hand-picked by morphological markers (beginning ventral invagination, start of germ band extension, no visible transverse furrow) and immediately transferred to ice-cold PBT (1xPBS, 0.1% Triton X100). Approximately 100 - 200 stage 6 embryos were collected prior to dissociation.

Embryos were washed thoroughly using ice cold cell culture grade PBS to remove any residual detergent, resuspended in 1ml ice cold dissociation buffer (cell culture grade PBS, 0.01% molecular biology grade BSA) and dissociated in a Dounce homogenizer (Wheaton #357544) with gentle, short strokes of the loose pestle on ice until all embryos were disrupted. The suspension was transferred into a 1.5 ml microfuge tube, cells were pelleted for 3' at 1 000*g* at 4°C. The supernatant was exchanged for 1 ml fresh dissociation buffer. Cells were further dissociated using 20 gentle passes through a 22G x 2" needle mounted on a 5 ml syringe. The cell suspension was then gently passed through a 20 µm cell strainer (Merck, NY2002500) into a fresh 1.5ml microfuge tube, residual cells were washed from strainer using a small amount of dissociation buffer. Cells were pelleted again for 3' at 1 000g at 4°C and resuspended in 100µl fresh dissociation buffer. The cell concentration was determined using a Neubauer counting chamber. Samples were fixed (*8*) by adding 4 volumes of ice-cold 100% methanol (final

concentration of 80% methanol in PBS) and thoroughly mixing with a micropipette. Cells were stored at -20°C (up to several weeks without any noticeable decline in sequencing quality).

For Drop-SEQ runs, samples from 8 to 10 handpicking sessions were combined into batches. *D. melanogaster* and *D. virilis* samples were prepared separately and combined just before a Drop-seq run. For rehydration, cells in methanol-fixative were moved to 4°C and kept in the cold throughout the procedure. Cells were pelleted at 3 000g (1 000g for rep.1) for 5 minutes, resuspended in PBS + 0.01% BSA, centrifuged again, passed through a 35 µm cell strainer, counted and diluted for Drop-SEQ as described below. Single cell sequencing digital gene expression (DGE) matrixes are deposited in GEO under accession IDs GSM 2494783 – GSM 2494789 as indicated in Suppl. Table S2. An additional DGE of only the HQCs described in the paper (including their t-SNE cluster assignment) is available in GEO under accession ID GSE 95025.

Drop-seq procedure, single cell library generation and sequencing.

Monodisperse droplets of about 1 nl in size were generated using microfluidic PDMS devices (Drop-SEQ chips, FlowJEM, Toronto, Canada; either self-coated or pre-coated with Aquapel). Barcoded microparticles (Barcoded Beads SeqB; ChemGenes Corp., Wilmington, MA, USA) were prepared and flowed in using a self-built Drop-seq set up according to Macosko et al. 2015 (*9*) (Online-Dropseq-Protocol-v.- 3.1, http://mccarrolllab.com/dropseq/). Cells were loaded according to Suppl. Tables S2 and S3 in 1x PBS + 0.01% BSA. Droplets were collected in 50 ml Falcon tubes; typically, the collection tube was exchanged every ~12.5 minutes, corresponding to ~1 ml of combined aqueous flow volume (1 ml cells and 1 ml of beads). Droplets were broken promptly after collection and barcoded beads with captured transcriptomes were reverse transcribed, exonuclease-treated and further processed as described (*9*). The 1st strand cDNA was amplified by equally distributing beads from one run to 48 PCR reactions (50 µl volume; 4 + 9 cycles). 10 µl fractions of each PCR reaction were pooled (total = 480 µl), then double-purified with 0.6x volumes of Agencourt AMPure XP beads (Beckman Coulter, A63881) and eluted in 12 µl. 1 µl of the amplified cDNA libraries were quantified on a BioAnalyzer High Sensitivity Chip (Agilent). 600 pg cDNA library were fragmented and amplified (12 cycles) for sequencing with the Nextera XT v2 DNA sample preparation kit (Illumina) using custom primers enabling 3'-targeted amplification as described (*9*). The libraries were double-purified with 0.6x volumes of AMPure XP Beads, quantified and sequenced (paired end) on Illumina Nextseq500 sequencers

(library concentration 1.8 p*M*; Nextseq 500/550 High Output v2 kit (75 cycles) in paired-end mode; read1 = 20 bp using the custom primer Read1CustSeqB (*9*) , read 2 = 64 bp).

Single cell RNA-seq: data processing, alignment and gene quantification.

The prepared libraries were sequenced in paired-end mode. We chose read 1 to be 20bp long, which sequences the cell barcode at positions 1-12 and the UMI at positions 13-20 (*9*), while avoiding reading into the poly(A) tail. The remaining 64 sequencing cycles were used for read 2. Sequencing quality was assessed by FastQC, while special attention was paid to the base qualities of read 1 to assure accurate cell and UMI calling. The last base of the UMIs generally showed a higher than expected T-content; we used the Drop-seq tools v. 1.12 to trim poly(A) stretches and potential SMART adapter contaminants from read 2, to add the cell and molecular barcodes to the sequences and to filter out barcodes with low quality bases. The reads were then aligned to a combined FASTA file of the *Drosophila* reference genomes (dm6; GCA_000005245) using STAR v. 2.4.0j with default parameters (*10*). Typically, around 85% of the reads were found to uniquely map to either of the species' genomes; reads not uniquely mapping to a given genome were discarded. The Drop-seq toolkit was further used to add gene annotation tags to the aligned reads and to identify and correct bead synthesis errors, in particular base missing cases in the cell barcode. The Ensembl annotations BDGP6.86 and GCA_000005245.1.33 were used for *melanogaster* and *virilis* respectively. Cell numbers were estimated by plotting the cumulative fraction of reads per cell against the cell barcodes and calculating the inflection point. The DigitalExpression tool was used to obtain the digital gene expression matrix (DGE) for each sample.

Species separation, cell filtering and data normalization.

For efficient handling of mixed species samples and a first view of gene quantification, basic statistics and the doublet rates we used dropbead (https://github.com/rajewsky-lab/dropbead) (*8*). A threshold of 90% (UMIs mapping to one species) was selected to confidently declare reads corresponding to a single-cell transcriptome as not representing a mixed-species doublet. *In silico* separation of the species under this threshold resulted in 7 individual DGEs for *D. melanogaster* and 2 for *D. virilis (*see Suppl. Tables 2 and 3), which were subsequently pooled together by species, giving 7 975 and 2 847 cells, respectively. Within the pooled DGEs, cells containing more than 5 UMIs for mesodermal and dorsal ectodermal marker genes (see Suppl. Table 4) were removed as potential doublets. Pole cells were identified by containing at least 3 UMIs of *pgc*

(*GJ22404* in *D. virilis*) and were removed prior to clustering analysis and spatial mapping. We compiled a list of pole cells markers, by comparing them against the yolk cells and the cells characterized as mesoderm/neurectoderm/dorsal ectoderm (Suppl. Table S6). Similarly yolk nuclei were identified by having at least 10 UMIs of yolk specific markers (see Suppl. Table 4). We further discarded cells expressing less than any 5 of the reference atlas genes used for mapping (threshold: 1 UMI), as our confidence in their mapping accuracy would be accordingly low. Genes mapping to the mitochondrial genome or with biotypes other than *protein coding*, *lncRNA* or *pseudogene* were also removed from the DGE. We normalized the UMI counts for every gene per cell by dividing its UMI count by the sum-total UMIs in that cell, and multiplying it by the number of UMIs that the deepest cell contained. Downstream analysis was performed in log space.

Correlation of gene expression measurements.

Correlations of gene expression levels between single-cell samples were computed by first sub-setting the DGEs of the two samples to the intersection of the genes captured in both libraries (typically ~10,000) and then computing the sum of gene counts across all cells in each library. Plotting of correlations is shown in log-space. For the correlation of our single-cell libraries against mRNA-seq, we converted the gene counts of the latter one into RPKMs and used the mean value of all isoforms for a given gene.

Marker genes and primordium assignment.

We compiled sets of genes, which are well-known to be expressed specifically in the presumptive mesoderm (ME, ventral), neurectoderm (NE, lateral), or dorsal ectoderm (DE) (see Suppl. Table 4). Cells were then classified with respect to gross dorsoventral origin by computing a column-specific score per cell reflecting the average expression of any column's markers present in that cell. The scores were scaled and centered across every cell and subsequently across the three dorsoventral regions. Cells scoring higher than 1.3 in one of the columns were assigned to that column; otherwise they were designated 'undetermined'.

Identification of highly variable genes, principal component analysis clustering and t-SNE.

We used Seurat (*11*) for identifying the highly variable genes, i.e. genes with relatively high average expression and variability, in each of the pooled DGEs. Those sets were subsequently

projected along the first few principal components to encompass more genes (see Satija et al., 2015 for more details on this procedure). Principal component analysis was performed over these extended sets. The first 16 or 18 principal components were identified as statistically significant and were used as input for clustering of *D. melanogaster* or *D. virilis* cells, respectively. We used Seurat to identify the marker genes for each of the clusters in the t-SNE representation.

Binarization of the BDTNP reference atlas.

The BDTNP reference atlas (*12*) quantifies the relative mRNA levels of 84 genes in the context of a virtual stage 5 *D. melanogaster* embryo at pseudo-nuclear resolution. We selected the latest of the 6 time points, i.e. stage 5 with ~100% cellularization, for mapping, as this is the closest match to our data – the earliest we were able collect embryos was beginning developmental stage 6, because cellularization in the *Drosophila* embryo does not complete until the end of stage 5, which means our embryos are developmentally within ~10 minutes of the high resolution BDTNP marker gene atlas used. Due to bilateral symmetry only half of the embryo is relevant for spatial mapping; hence we selected the first 3 039 nuclei. We converted the continuous gene expression levels into binarized on/off states that would guide spatial mapping. We chose a binarization threshold for each gene individually, by inspecting a range of thresholding values and comparing the resulting pattern with published *in situ* hybridizations (*13, 14*). In cases where the BDTNP atlas thresholding resulted in clear discrepancy with published *in situ* data, the reported expression was given primacy and the atlas was manually adjusted. Out of 3 039 locations of the gene atlas, 2 937 exhibit unique gene combination patterns upon binarization. Sup. Fig. S5A shows the logarithmic growth of the number of unique bins as a function of the number of genes considered; all 84 genes were included for maximum bin heterogeneity. The final position dependent expression matrix of the binarized *in situs* used for spatial mapping is included in Suppl. Table S7.

Binarization of Drop-seq data.

We binarized the gene expression levels for the 84 genes that guided the mapping as follows. First, we computed the gene correlation matrix of the 84 genes with the binarized version of the BDTNP atlas. Next, for a given gene and only considering the Drop-seq cells expressing it we computed a quantile value above (below) which the gene would be designated ON (OFF). We sampled a series of quantile values and each time the gene correlation matrix based on this binarized version of our data versus the binarized BDTNP atlas was computed and compared by

calculating the mean square root error between the elements of the lower triangular matrices. Eventually, we selected the quantile value 0.23, as it was found to minimize the distance between the two correlation matrices.

Spatial mapping algorithm: *DistMap*.

Having binarized the BDTNP atlas and our data as described above, we aimed at assigning a distributed score for each of our cells to each of the 3 039 bins of the database as follows. We compared a given cell to all bins by counting the number of genes found ON (OFF) in both the cell and a given bin, as well as the number of genes, which disagreed between the cell and the given bin. This led to a calculation of 3 039 confusion matrices for each cell. In a given cell-bin combination, we interpreted the OFF-OFF cases as true negatives (tn), the ON-OFF cases as false positives (fp), the OFF-ON cases as false negatives (fn) and the ON-ON cases as true positives (tp). We employed the Mathews correlation coefficient (MCC) to weight the confusion matrices and assign a cell-bin score. The MCC scores were subsequently exponentiated.

Simulations to assess bin coverage & cell mapping confidence.

In general a given cell exhibited high scores with more than one positional bins (see also heat map of MCC scores in Fig 4A). We performed simulations to assess whether the obtained high scores could have been generated by chance. First, we permuted the values of the 84 genes for every cell of our data 100 times, generating thus 129 700 simulated cells. We mapped the simulated cells onto the reference atlas via the algorithm described above and calculated then the maximum MCC scores per bin. Their distribution was significantly lower than the corresponding distribution arising from the mapping of real cells (data not shown). Particularly, more than 87% of the bins had at least one MCC score with a p-value smaller than 0.05, leading to the conclusion that we effectively covered more than 87% of the embryo with high confidence (Fig. 2A). Similarly, we permuted the 84 genes 40 times for every bin of the reference atlas, generating thus 121 560 simulated bins, upon which we mapped our Drop-seq data. Comparing the distribution of the maximum MCC scores with the one corresponding to the real bins, we assessed that 878 cells were assigned to the embryo with very high confidence (p-value < 0.05, Suppl. Fig. S5C).

To assess the optimal number of cells needed for mapping, we downsampled the HQCs and mapped them via DistMap. vISHs for the 84 reference genes were computed and compared to the original expression patterns by computing the Hamming distance of their binarized expressions (Fig. 5B). We observed a plateau starting at around ~750 cells, implying that more sequenced

cells would not improve the mapping drastically.

vISH computation.

For the majority of the cells, several positional bins scored highly. Therefore, we adopted a probabilistic mapping strategy, instead of fixing every cell to only its highest probability location. For a given gene, we multiply the vector of its normalized expression across cells with the MCC scores per bin. This multiplication is element-wise and penalizes cases for which the gene is either very lowly expressed (even though the scores for that cell could be very high) or the score for a bin is very low (indicating that the cell expresses the gene but maps better to another bin). We sum the products row-wise, obtaining thus a number for each bin. We then binarize the vector of normalized gene expression and repeat the above process, resulting in a second 3039 dimensional vector. We divide the first vector by the second, element-wise, to account for the different number of cells that could express the gene. The resulting vector, q, is further normalized through q/(1+q) to reduce outlier effects. A threshold is then imposed that sets the values of q to zero that lie below the quantile corresponding to that threshold.

vISH visualization.

We used the coordinate system of the sixth time point of the BDTNP database to depict our computed expression patterns. As the embryo is bilaterally symmetric and we mapped our cells to the first 3 039 bins, we mirrored the final scores for the remaining 3 039 bins across the A/P-D/V plane to obtain a complete picture of the embryo. For the lateral view, we display the first 3 039 bins plotted in x and z. For the ventral and dorsal views, the bins with $z<0$ or $z>0$ were selected and their x/y coordinates were plotted. For the anterior and posterior view, the bins with $x<0$ or $x>0$ were selected and their y/z coordinates were plotted. The red to grey color scale shows high to low expression per positional bin.

Noise simulations to assess vISH threshold values.

We performed simulations in order to identify meaningful threshold values when computing a vISH. Using the raw data of the 1 297 cells, we generated for every gene and for every cell a random Poisson number, with a mean and variance equal to the number of UMIs for that cell and gene; we repeated this process 50 times. We verified that the correlations of gene expression levels were high across the samples (Pearson $R > 0.97$, data not shown), but the cell-to-cell correlations were much lower (Pearson $R \sim 0.7$, data not shown). We spatially mapped these 50

simulated samples. For a given gene we computed a vISH for thresholds in the range [0.5, 0.98] with steps of 0.01. We then computed the per bin standard deviations across the 50 calculated vISH and their sum. As higher thresholds imply fewer bins, this sum is expected to minimize for threshold=0.98. However, for some genes, especially the well-expressed ones, the distribution of these sums against the threshold values showed local minima, which in some cases were even global (Sup. Fig. 5D), revealing robustness of these threshold values against Poisson noise simulated data. Visualizing the vISHs with these obtained values and comparing the patterns with in vivo spatial expression data (*13, 14*) showed good agreement. For the lowly expressed genes our simulations could not provide useful insights on threshold values, even when modeling noise with a negative binomial distribution, allowing for different variances.

Clustering and discovery of archetypal *in situ* patterns.

Having computed a set of gene expression patterns, we computed their distance matrix in order to discover gene expression archetypes. The gene set was selected as the union of the highly variable genes, the three column markers and the genes used for the spatial mapping. Agglomerative clustering was performed to identify the number of parent clusters. We averaged the expression patterns of the genes belonging to each of the parent clusters, thus producing 10 representative *in situ*s of the corresponding identified archetypal classes (Sup. Fig. 5E).

Discovery of new markers with localized expression patterns.

We restricted our survey for new markers of expression patterns to genes either not reported in, or annotated as 'no staining' in the BDGP *in situ* database. We computed the expression patterns for this large set of genes and visually inspected these virtual *in situ* hybridizations. For validation, we chose genes with distinct localized expression patterns.

Correlation of Drop-Seq and nCounter gene expression analysis.

To assess the quantitativeness of the Drop-Seq data gene expression data with respect to NanoString nCounter data from individual stage-matched embryos (beginning gastrulation) (*15*), the DGE was subsetted to the genes included in the nCounter analysis. For each gene a total count across all cells was calculated. The gene *Fdy* was excluded because it resides on the Y-chromosome and the nCounter value based on 3 embryos cannot be representative of a random male/female mixture of cells. For genes which were counted using two primer pairs in the nCounter experiment, the higher value was used in each case, assuming that this reflects

expression values obtained by poly-A selection more closely due to inclusion of more transcript isoforms. For each gene, the total DGE counts and the total nCounter (gastrulation stage) counts were converted to relative expressions by dividing each gene count by the sum of all genes in the respective dataset.

Gene Ontology (GO) Term Enrichment analysis

For the GO term analysis the Comprehensive R Archive Network (CRAN) package "gProfileR" was used. All genes detected in the Drop-seq experiment were used as the background model. For the cluster-specific GO term enrichments, a list of genes identified as significantly overrepresented in a cluster was assembled for each cluster using the "FindAllMarkers" function from Seurat (*11*) and requiring an average difference (avg_diff) > 0 and a p-value (p_val) < 0.05. Cluster specific enrichment of terms is displayed on a blue scale where p-value > 0.05 and on a red scale were p-value < 0.05 (Suppl. Fig. S3). Rows are clustered by complete linkage clustering. The inclusive heat map for biological process terms (Suppl. Fig. S3C) was subsetted to exclude less meaningful terms and is shown in Suppl. Fig. S3B.

SUPPLEMENTARY NOTES

Supplementary Note 1: *In silico* dissection of the embryo

An immediate benefit of the single cell sequencing data is that it enables '*in silico* dissection' of the sample, in this case of the embryo. Based on marker gene sets, cells can be identified and their average gene expression profiles can be generated by merging sequencing reads. We scored cells for selected sets of marker genes corresponding to presumptive mesoderm (ME, ventral), neurectoderm (NE, lateral), and dorsal ectoderm (DE) (see Materials and Methods and Table S4). Ternary analysis of these marker gene sets shows efficient cell separation (Fig. S2A). The aggregated transcriptome tracks from these sets of cells accurately recapitulate the exclusivity of known marker gene expression, such as *snail*, *Dichaete* and *zerknullt* in the mesoderm, neurectoderm or dorsal ectoderm, respectively (Fig. S2B).

In effect, we conducted a marker-based dissection of the intact wild type embryo and obtained tissue specific transcriptomes – something not possible with conventional methods. A major benefit is that the cell population of interest could be refined by any combination of marker genes and even by thresholding on expression levels. We included the function "inSilicoDissect" for this in the *DistMap* package.

Supplementary Note 2: Single cell clustering resolves spatial identities

Clustering analysis of the HQCs identifies at least 9 prominent cell clusters in the stage 6 embryo (Fig. S2F). Many of the most specific genes that drive clustering (e.g. Fig. S2G, see Table S5) are well-known for their roles in embryonic patterning and tissue specification, such as mesodermal determinants (*twist, tin*, *snail*) in cluster 4, DE determinants (*zen, zen2, doc3*) in cluster 6, neurectodermal determinants (*SoxN*, *sog, ind*, *Dr/msh*) in cluster 3, the midgut driver *fkh* in clusters 2 and 7, the terminal determinant *hkb* in cluster 2 and the sub-terminal posterior identity gene *Abd-B* in cluster 8. Examination of cluster-specific driver genes (see Table S5) in *in situ* databases (*14, 16*) reveals that spatial expression is similar within clusters. This demonstrates that clustering reveals basic patterning information within the embryo. While this alone places cells within the respective clusters ventrally, dorsally, laterally and posteriorly, other more complex relationships are also apparent (see Supplementary Note 3: 'Gene ontology (*GO) term analysis*').

Supplementary Note 3: Gene ontology (GO) term analysis

Among the most specific genes that drive clustering (Fig. 2E, Suppl. Table S5), many are transcription factors that are well-known for their role in embryonic patterning and tissue specification. However, a number of these 'driver' genes were not known to play roles in early patterning and development and several unstudied lncRNAs were identified to be highly cluster specific. While specific transcription factor enrichments give a general indication of the origins of cluster cells, this spatial identity is reinforced by gene ontology (GO) term enrichment. Overrepresented molecular function GO-terms across *all* enriched genes primarily relate to transcriptional regulation (Fig. S3A), reflecting the central role transcription regulation plays for cluster identity at stage 6. The only enriched non-transcription terms were found in cluster 3 (Fig. S3A). In addition to neurogenic transcription factors (Table S5), cluster 3 was enriched in *growth factor signaling* molecular function GO-terms. The neurogenic ectoderm is known to be a source of FGF and EGF signaling molecules (*17, 18*), therefore supporting a neurogenic origin of cluster 3 cells.

Spatially more informative is biological function GO-term enrichment. For example, nervous system terms (*ventral cord*, *neuroblasts*, *neurogenesis*, *CNS)* further support the neurogenic identity of cluster 3 cells. Similarly, mesodermal terms (e.g. mesodermal cell differentiation, muscle cell development) place cluster 4 cells ventrally, cluster 6 enriched terms (*amnioserosa, dorsal closure*) indicate dorsal origin, while hindgut terms (e.g. hindgut morphogenesis) place cluster 8 cells posteriorly. However, enriched terms often do not lend themselves to simple spatial assignment. While a cluster may be associated with terms indicative of distinct regions, other clusters may be associated with terms indicative of the same spatial domain. For example, cluster 3, 6, and 8 are all enriched for "*generation of neurons"*, a likely lateral term, though cluster 6 is also enriched for genes associated with "heart development" (likely mesodermal) and "amnioserosa" (dorsal ectoderm), while cluster 8 is enriched for genes annotated to be involved in "*head development*" (anterior) and "*hindgut*" (posterior) (Fig. S3B,C).

Such spatial inferences push the information value of GO terms beyond their intended limits. Enriched terms may, for example, stem from cell populations regulating each other's developmental fate. For example, dorsal ectodermal cells (i.e. cluster 6) do play known roles in regulating heart formation in the underlying dorsal mesoderm in later stages (derived from cluster 4), as well as in helping to pattern the lateral neurectoderm (cluster 3). Nonetheless, GO term enrichment demonstrates that clustering captures a substantial amount of spatial information in the stage 6 embryo.

Supplementary Note 4: Modulation of TGFβ/Dpp signaling

TGFβ signaling establishes dorsal cell identities in the early embryo (e.g. (*19*). vISH identified the expected expression of the *Dpp* ligand dorsally, the Dpp antagonist *sog* laterally and predicted the expression of transcriptional modulators in lateral (*brk*) and dorsal regions (*shn*) (Fig. S7A). Transduction of TGFβ signaling requires phosphorylation of a ubiquitous Smad protein, Mad. Phosphorylated Mad (pMad) is observed specifically in a narrow dorsal gradient at stage 6 to 8.

Dad is the only inhibitory Smad in *Drosophila*; it prevents Mad phosphorylation and thereby inhibits formation of the Mad/Medea complex and its translocation into the nucleus (*20*). Hence, if *Dad* is locally expressed in the early embryo, it would prevent active Dpp signal transduction once translated. DVEX predicts *Dad* expression in anterior and posterior regions, but with low confidence as this prediction is based on detected expression in few cells. In agreement with the sparsity of *Dad* transcripts detected by single cell sequencing (see DVEX tSNE mapping of Dad expressing cells), we could not detect *Dad* expression at stage 6 by *in situ* hybridization. However, *Dad* was detected slightly later in an anterior pattern at stage 8. Immunochemistry for phosphorylated Mad (using anti-pMad Ab (*5*)) shows diminished pMad signal where Dad expression was detected (see arrow in Fig. S7B). To our knowledge, an anterior inhibition of TGFβ signaling by Dad has not been described previously.

More generally, Fig. S7A illustrates striking expression divergence for other components of the Dpp/TGFβ-pathway at the level of ligands, ligand modulators, receptors and transcriptional modulators. This is of particular interest, as it has been demonstrated that specific receptor dimer and ligand-receptor combinations can produce differential signaling effects (e.g. (*21*). Even in the early embryo, the patterned expression of ligands and receptors may lead to regionally distinct signaling outputs.

SUPPLEMENTARY REFERENCES

1. G. M. Rubin, A. C. Spradling, *Science* **218**, 348-353 (1982).
2. M. Markstein *et al.*, *Development* **131**, 2387-2394 (2004).
3. S. Barolo, B. Castro, J. W. Posakony, *Biotechniques* **36**, 436-440, 442 (2004).
4. D. Kosman *et al.*, *Science* **305**, 846 (2004).
5. Z. Guo, I. Driver, B. Ohlstein, *J Cell Biol* **201**, 945-961 (2013).
6. A. Ikmi *et al.*, *Mol Biol Evol* **31**, 1375-1390 (2014).
7. J. Schindelin *et al.*, *Nat Methods* **9**, 676-682 (2012).
8. J. Alles *et al.*, *BMC Biol* **15**, 44 (2017).
9. E. Z. Macosko *et al.*, *Cell* **161**, 1202-1214 (2015).
10. A. Dobin *et al.*, *Bioinformatics* **29**, 15-21 (2013).
11. R. Satija *et al.*, *Nat Biotechnol* **33**, 495-502 (2015).
12. C. C. Fowlkes *et al.*, *Cell* **133**, 364-374 (2008).
13. P. Tomancak *et al.*, *Genome Biol* **3**, RESEARCH0088 (2002).
14. P. Tomancak *et al.*, *Genome Biol* **8**, R145 (2007).
15. J. E. Sandler, A. Stathopoulos, *Genetics* **202**, 1575-1584 (2016).
16. E. Lecuyer *et al.*, *Cell* **131**, 174-187 (2007).
17. A. Stathopoulos *et al.*, *Genes Dev* **18**, 687-699 (2004).
18. R. P. Zinzen *et al.*, *Dev Cell* **11**, 895-902 (2006).
19. F. F. Esteves *et al.*, *PLoS Genet* **10**, e1004625 (2014).
20. H. Inoue *et al.*, *Mol Biol Cell* **9**, 2145-2156 (1998).
21. T. D. Mueller, J. Nickel, *FEBS Lett* **586**, 1846-1859 (2012).

SUPPLEMENTARY TABLES

Supplementary Table S1: RNA in situ hybridization probes

Target gene	*probe generated by*	*primer sequences (5' → 3')* *or EST ID*
dsRED	PCR, subcloned	5'-CTGTTTAATTCGCCCTTCACCG-3' 5'-TACAGGAACAGGTGGTGGCG-3'
CR45693	PCR, direct	5'-TGACAGCCTTCTGACAGGTTTT-3' 5'-TAATACGACTCACTATAGGGAGCAACGTCCTTCAATGGTTTG-3'
gcm	EST	RT01048
SoxN	EST	RH18247
CG4500	EST	RE63419
ana	EST	RH40649
CG16886	EST	RH73259
CG6660	PCR, direct	5'-TAGACTGTGCTTTCGAGCAGTT-3' 5'-ACGTATAATACGACTCACTATAGGGATTTTGTGGCGATGAAACGGAG-3'
fok	PCR, direct	5'-TTCACACGTAGGCACATGAGAA-3' 5'-ACGTATAATACGACTCACTATAGGGACAGCTGAGTCTGAACACCAAA-3'
GJ14350	PCR, direct	5'-ACGAAAGGAAAGCCAGAATCCT-3' 5'-ACGTATAATACGACTCACTATAGGGGGCTTAGTGTTTGGCAAGGTCAG-3'
GJ17890	PCR, direct	5'-CAAAGCTACGCATCGCGAAA-3' 5'-ACGTATAATACGACTCACTATAGGGTCCCGCAAGCAAAATGTCTG-3'
stumps	EST	PC00427
Oaz	EST	AT08673
Kr	EST	RE30918
eve	PCR, subcloned	5'-CAGTCTTGTAGGGCTTGAAGAGC-3' 5'-CTTTGAATCACAAGACGCATACC-3'
babos	EST	GH11432
m4	EST	FI14216
CR44691	PCR, subcloned	5'-GCTGGTTTCCAAAACGTCATGT-3' 5'-TCTCCCTCTCTCTCTCACACTG-3'
CR44917	PCR, subcloned	5'-ATAAAACCATCAACTGCGTGGC-3' 5'-TACCTTTGGGCAATGGTCACTT-3'
CG32053	PCR, direct	5'-GTTTTGACAGTTCGCTTTCCGA-3' 5'-NNTAATACGACTCACTATAGGGCCTACTTTCAGGCTGGTGGAAT-3'
CG14204	PCR, direct	5'-GCACCATTCCATCGGGATATCT-3' 5'-NNTAATACGACTCACTATAGGGATTCTTTCCAAGGCCTCCAGAG-3'
CG14688	PCR, direct	5'-GGCTACATTGTCATCGAGGACT-3' 5'-NNTAATACGACTCACTATAGGGACCGAATAACGATCTCCACCAG-3'
CG34224	PCR, direct	5'-AATTTAGCAATCGGAGCCAGGA-3' 5'-ACTGCTAATACGACTCACTATAGAGGTCTTTAGCTTGGCCCTAAC-3'
CG10553	PCR, direct	5'-AAAATTTCCGGACGCACCTTTC-3' 5'-ACTGCTAATACGACTCACTATAGCGGGAAAACAAAGAGAGTGCTGG-3'
CR44317	PCR, direct	5'-AGCAGGTCCAAATTTTTCCACG-3' 5'-ACTGCTAATACGACTCACTATAGGGCGATGTCGACTAAAGCATTC-3'
CR43432	PCR, direct	5'-TCCGTCCGCGAGTTGTATAAAA-3' 5'-ACTGCTAATACGACTCACTATAGGTTCAGATAGCAGCAGCAAATG-3'
CR45559	PCR, direct	5'-GCAGCATTCCACCAATACACTG-3' 5'-ACTGCTAATACGACTCACTATTAGCAAGTCCTGTTGTCTGGCAAAA-3'
dad	EST	LD47465

Probes were either generated from ESTs (DGRC clone identifiers listed), from PCR products subcloned into pCRII, or directly from PCR products incorporating a T7 promoters in the reverse primer. Images for *stumps*, *Oaz,* and *Kr in situ* hybridizations from the BDGP ISH database (*14*).

Supplementary Table S2: Drop-seq run information.

run ID	Drop-seq Run ID	Drop-seq Date	species mix	CR_Hyd	CR_CAF	final conc. (cells/µl)	GEO accession ID
rep.1	ds015	2016-04-19	yes (1:2)	175 000	118 456	68	GSM 2494783
rep.2	ds022_mix	2016-07-05	yes (2:1)	100 000	72 360	40	GSM 2494784
rep.3	ds022_mel	2016-07-05	no	82 500	59 697	33	GSM 2494785
rep.4	ds024a	2016-09-02	no	sample 1: 320 000	83 750	50	GSM 2494786
rep.5	ds024b	2016-09-02	no		83 750	50	GSM 2494787
rep.6	ds024c	2016-09-02	no	sample 2: 212 500	58 625	35	GSM 2494788
rep.7	ds024d	2016-09-02	no		58 625	35	GSM 2494789
			Total:	890 000	535 263		

Replicate runs for Drop-seq. CR_Hyd, number of cells recovered after rehydration; CR_CAF, number of cells recovered in combined aqueous flow.

Supplementary Table S3: Drop-Seq statistics

run ID	Species	Cell number	Genes	UMIs	Reads	GEO accession ID
rep.1	*D.mel.*	1119	1170	4290	36254	GSM 2494783
	D.vir.	1857	1042	3445	30524	
	mixed	198	—	—	—	
rep.2	*D.mel.*	1104	1464	6253	21300	GSM 2494784
	D.vir.	990	1393	6115	22199	
	mixed	84	—	—	—	
rep.3	*D.mel.*	1563	1319	5814	36729	GSM 2494785
rep.4	*D.mel.*	1257	1286	5490	25908	GSM 2494786
rep.5	*D.mel.*	1044	1132	4340	26527	GSM 2494787
rep.6	*D.mel.*	826	705	1943	18674	GSM 2494788
rep.7	*D.mel.*	1062	1041	4007	25020	GSM 2494789
sum	*D.mel.*	7975		31575	190412	
	D.vir.	2847		9560	30524	
	mixed	282		—	—	

Drop-Seq statistics for cells below the 'knee' and with >1000 UMIs. Genes, median number of genes/cell; UMIs, median number of UMIs/cell; Reads, median number of reads/cell.

Supplementary Table S4: *Drosophila* embryo marker genes.

Cell population	Genes
Dorsal ectoderm	*Ance, CG2162, Doc1, Doc2, egr, peb, tok, ush, zen*
Neurectoderm	*ac, brk, CG8312, l(1)sc, mfas, Ptp4E, sog, SoxN, vnd*
Mesoderm	*CG9005, Cyp310a1, GEFmeso, ltl, Mdr49, Mes2, NetA, ry, sna, stumps, twi, wgn, zfh1*
Yolk cells	*beat-IIIc, CG8129, CG8195, Corp, CNT1, sisA, ZnT77C*
Pole cells	*Pgc*

Supplementary Table S5: Highly variable genes in tSNE clusters.

tSNE Cluster	Genes
1	*spict, Ada1-1, CG18130, Klp10A, mbl, srl, sktl, mrn, CG13300, CR41257, CG7546, CG32243, phyl, CG2837, CG3184*
2	*Aldh, ImpE2, mtt, hkb, ImpL3, CG31038, CG5656, croc, CG13101, fkh, CG8468, CG17119, CG12541, Hsp83, kni*
3	*ind, CG34224, ths, sad, ac, gsb, ImpL2, pyr, Atx-1, pDsRed, SoxN, Dr, sca, Meltrin, sog*
4	*tin, Act87E, twi, CG1673, sprt, CG14687, Mef2, Ilp4, CG14688, sna, CR45361, stumps, CG12177, Mes2, Nplp2*
5	*hbn, fd102C, CG14204, Optix, nrm, fj, eya, oc, Bili, yellow-e, amd, toy, CG3502, Gasp, Adgf-A*
6	*peb, zen2, dap, zen, C15, CG13653, ana, Alk, CG16886, CG43725, Z600, kay, Doc3, CG6753, CG42666*
7	*Pdp1, CG2930, DNaseII, ps, CG32053, CG18754, Gmap, CG7191, Ptx1, Fas2, fkh, mnd, MRE23, CG31431, a*
8	*klg, byn, ken, cad, disco, hb, dpn, CG31871, salm, Blimp-1, rib, D, Abd-B, CG42762, apt*
9	*gcm, ham, ttk, CrebA, shep, RhoL, fok, knrl, kni, zfh1, CG33099, CR44683, srp, btd, NetB*
10	*CR45185, grn, so, CR43302, Oaz, lov, SP2353, toy, Hmx, tll, oc, CG15236, CG42342, Ance, Dll*
11	*rau, Abd-B, grn, CR43617, cic, run, ken, CG32483, EcR, CG34232, veil, hb, CR43279, CG11966, CG10176*

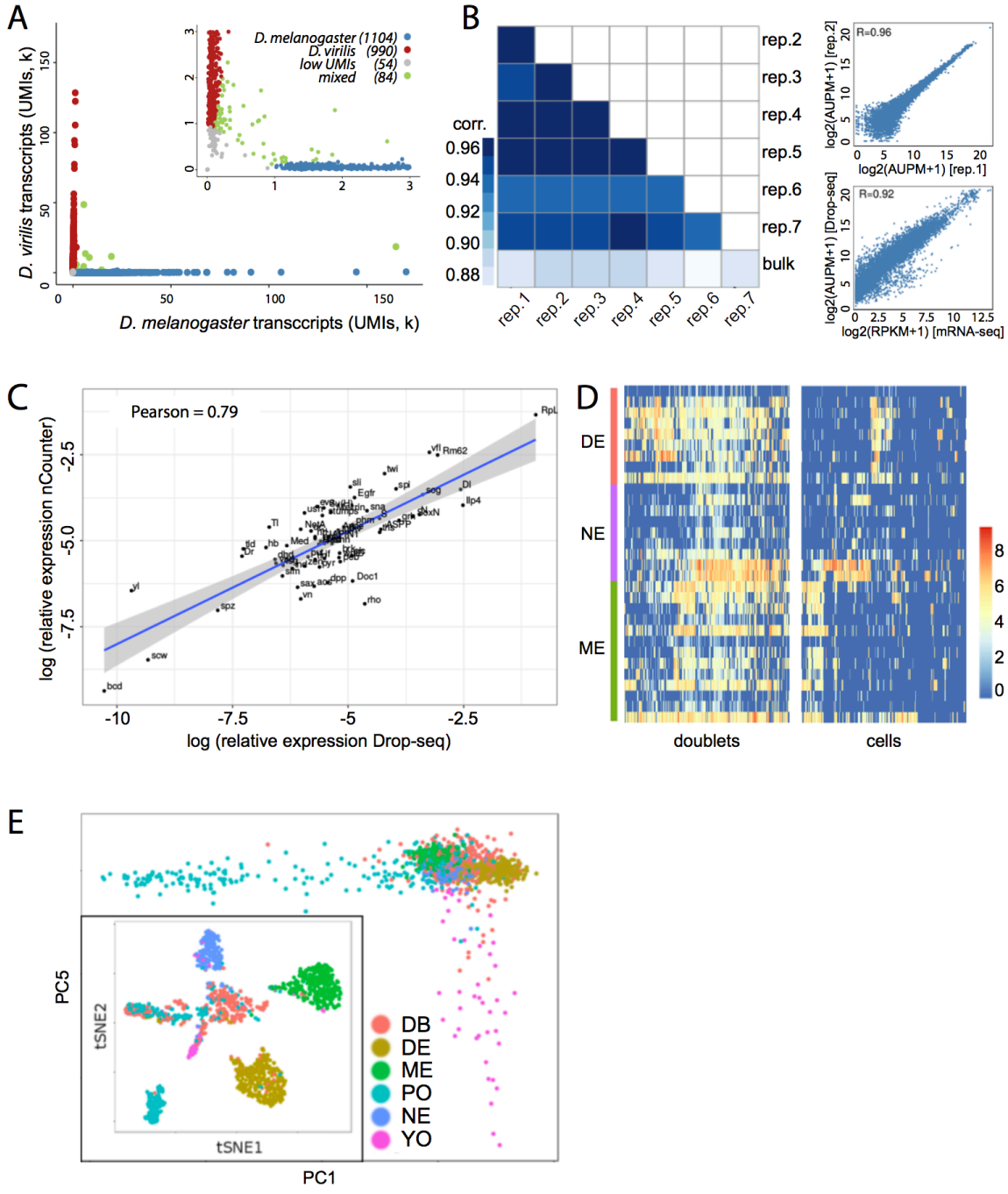
Supplementary Table S6: Pole cell marker genes (available separately online).

Supplementary Table S7: BDTNP reference atlas – thresholds and expression matrix (available separately online).

Supplementary Table S8: Homology analysis (available separately online).

SUPPLEMENTARY FIGURES AND FIGURE LEGENDS

Supplementary Figure S1: Drop-seq data quality and reproducibility.



Legend to Supplementary Figure S1

(**A**) Species separation plot for one of the *D. melanogaster* / *D. virilis* Drop-seq replicates; inset shows zoomed-in intervals to better expose doublet distribution. The mixed species doublet ratio is < 4%. Same species doublets would be expected at a similar rate. Grey colored cells contain less than 1 000 UMIs.

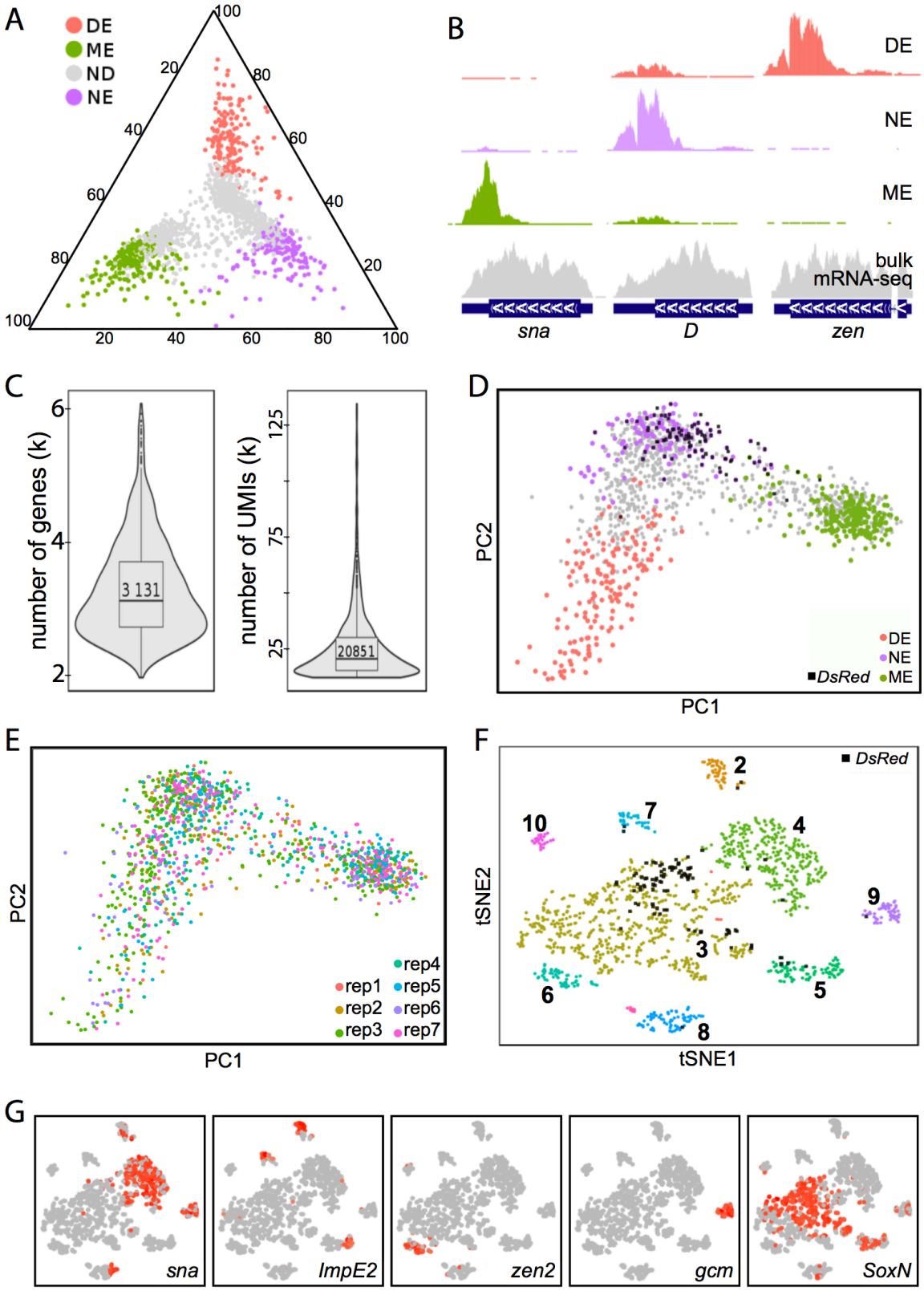
(**B**) Left panel: Correlation matrix comparing aggregated UMIs in Drop-seq replicates and whole embryo mRNA sequencing (bulk) for *D. melanogaster*. Right panel: Correlation of aggregate gene expression between independent Drop-seq replicates (top) and between Drop-seq replicates and bulk mRNA sequencing from unfixed, whole embryos; Pearson correlation shown on the top.

(**C**) Correlation of relative gene expression between aggregated UMIs in Drop-seq and nCounter measurements (*15*).

(**D**) Doublet identification and exclusion. Heat maps of cells (columns) clustered by expression of DE, NE and ME marker genes (rows). Doublets (left) show expression of several marker classes whereas a random subsample of single cells (right, doublets excluded) shows much more exclusive marker class expression. Coloring indicates normalized expression levels (log2(ATPM+1)).

(**E**) PCA of 1369 cells separates yolk (YO) and pole cells (PO); cells colored by marker expression (mesoderm, ME; neurectoderm, NE; dorsoectoderm, DE; doublet, DB). Inset: t-SNE representation of the same cells shows clustering of cell types; doublets cluster centrally.

Supplementary Figure S2: High quality cells encode spatial patterning information.



Legend to Supplementary Figure S2

(**A**) Ternary plot indicating scores for marker expression. Grey cells could not be unambiguously assigned to any of the mesoderm (ME), neurectoderm (NE), or dorsal ectoderm (DE) populations.

(**B**) Genome browser tracks after in silico dissection at three genes regulated along the dorsoventral axis; compare whole embryo mRNA sequencing (bulk, grey) with aggregate transcriptomes of cells expressing DE, NE, or ME markers.

(**C**) Violin plots of gene and UMI numbers detected per cell for highest quality cell.

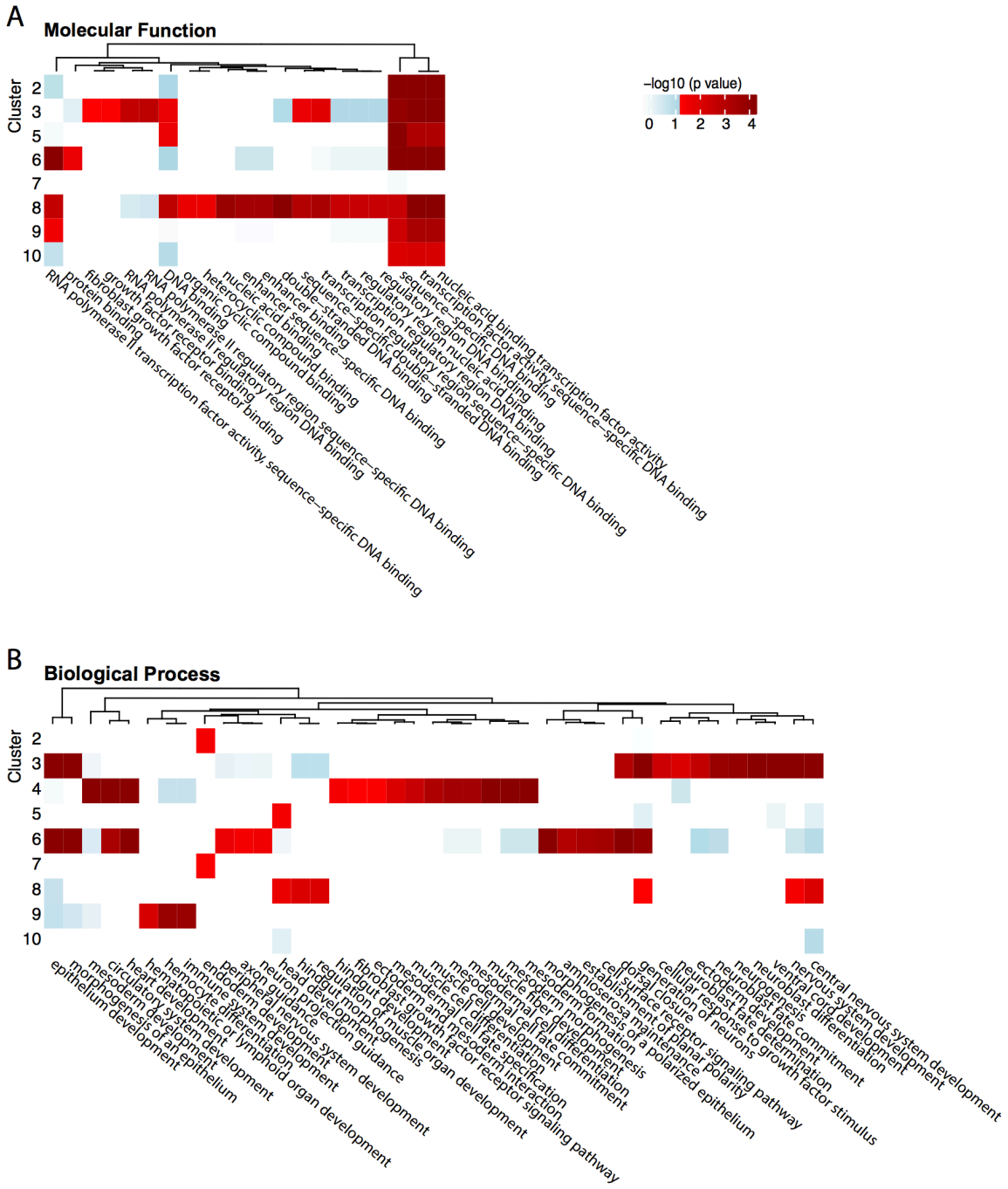
(**D**) Principal component analysis (PCA) of high quality cells; principal components 1 and 2 separate DE, NE, and ME; colors by marker expression, grey cells are unassigned, black boxes indicate strong expression of the *vnd::DsRed* reporter in an NE subset.

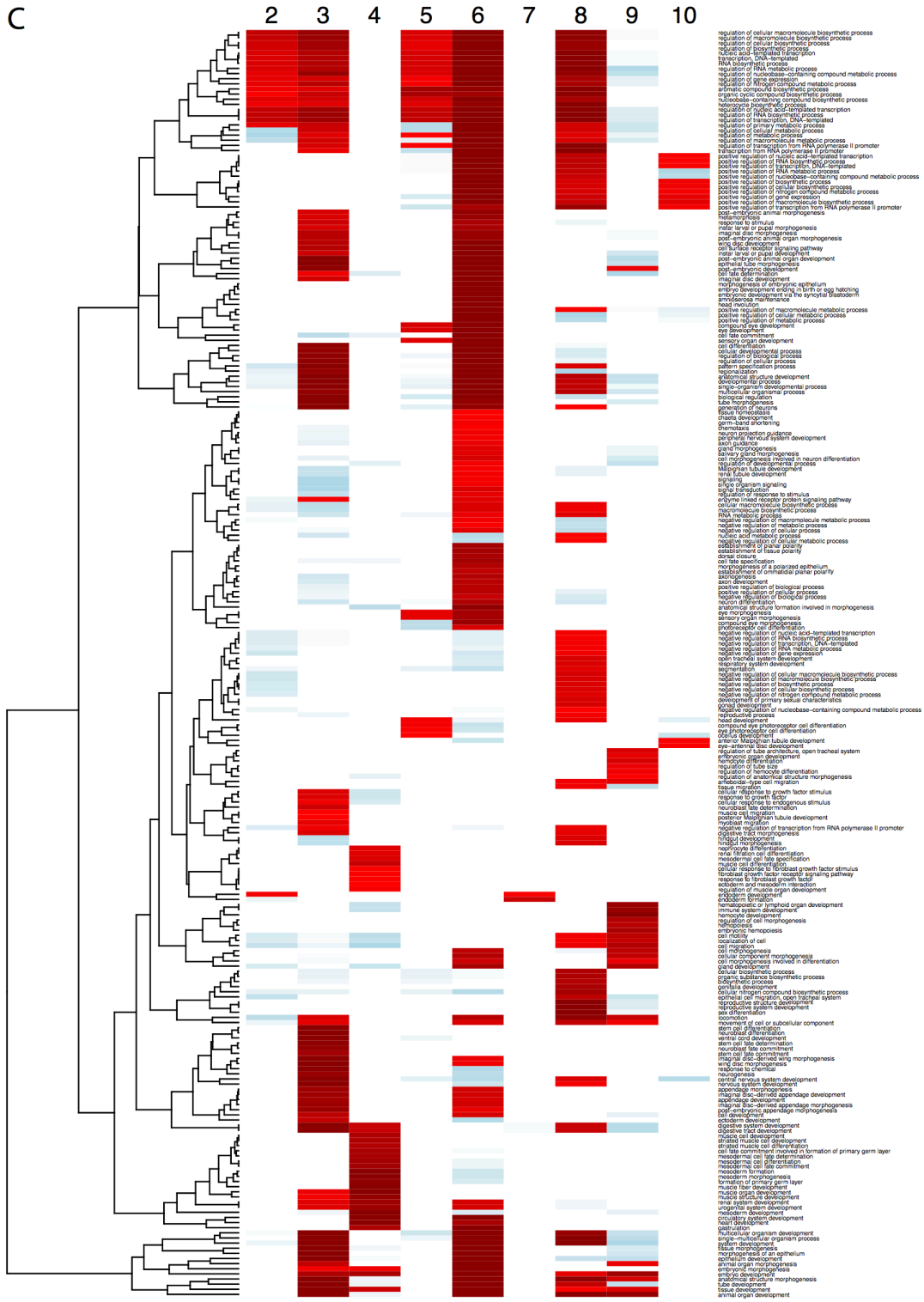
(**E**) Absence of batch effects. PCA of 1297 mapped *melanogaster* cells; cells are colored by Drop-seq runs.

(**F**) t-SNE representation of the high quality cells shows 9 major clusters, black boxes indicate strong expression of the *vnd::DsRed* reporter in a cluster 3 subset

(**G**) Examples of highly variable genes in t-SNE clusters; these genes are known to be patterned, indicating that clusters reflect positional information. Red indicates high expression.

Supplementary Figure S3: GO term analysis of single cell clusters

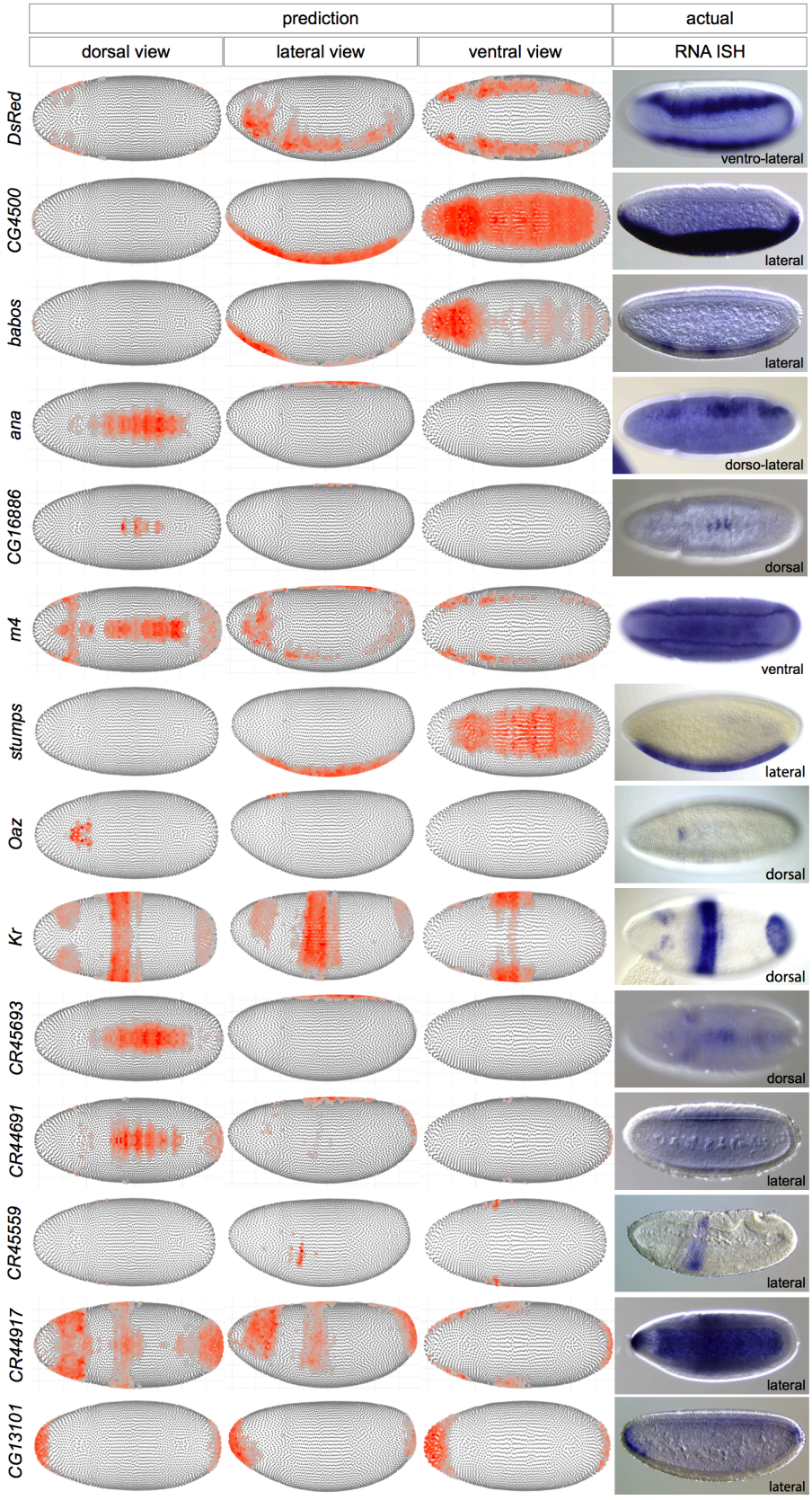



C
2
3
4
5
6
7
8
9
10
–log10
(p value)
4
3
2
1
0

Legend to Supplementary Figure S3

GO term enrichment indicates distinct molecular and biological functions associated with different t-SNE clusters. Molecular Function (**A**) and Biological Process (**B, C**) GO term enrichment among t-SNE cluster-enriched genes. Cluster numbers correspond to t-SNE clusters as indicated in Fig. S2F. Color Scale indicates adjusted p-value. The heat map in (**B**) has been subset for clarity from (**C**)

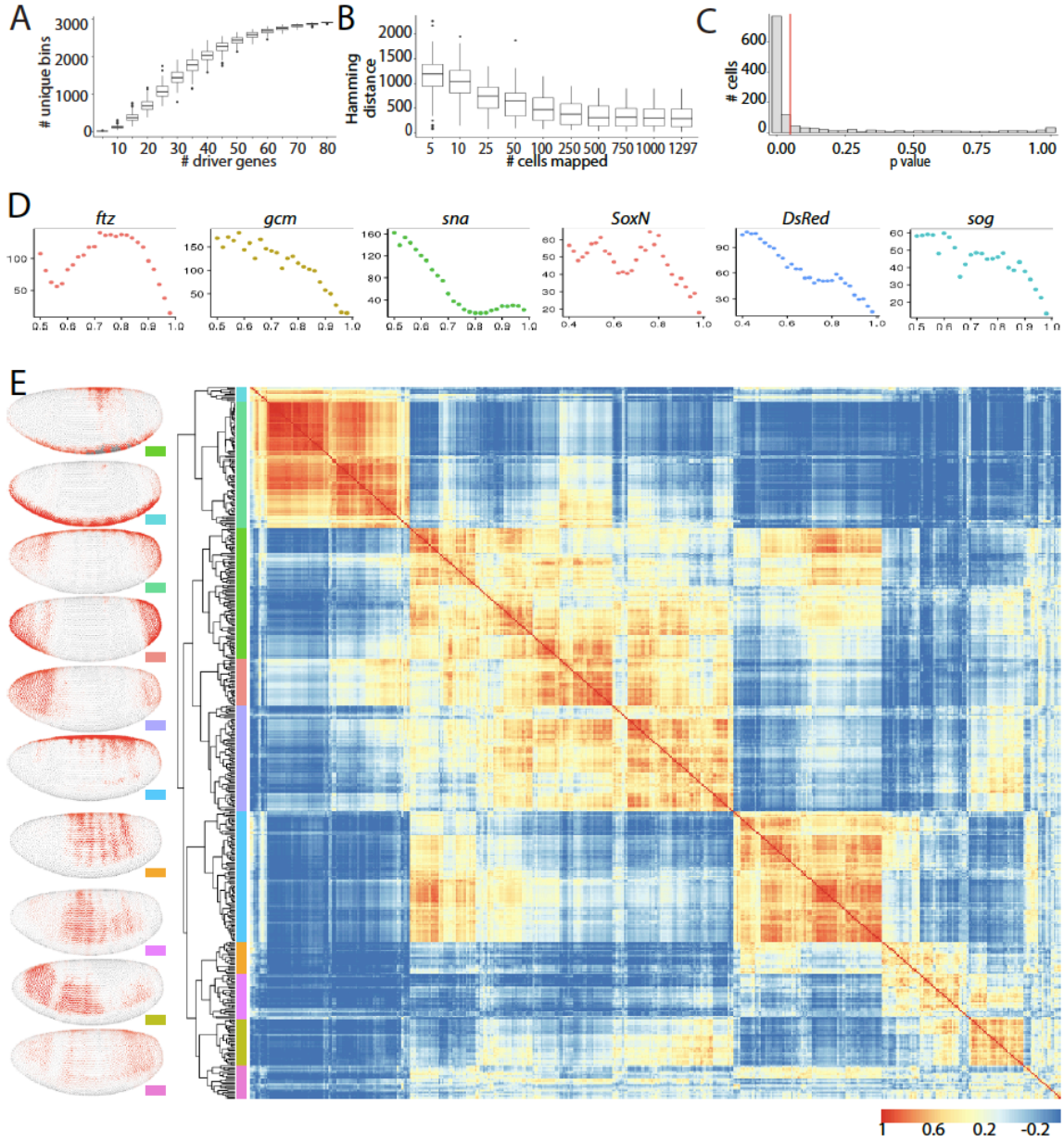
Supplementary Figure S4: vISH predictions & experimental validations



Legend to Supplementary Figure S4

vISH predictions of known and novel marker genes are depicted anterior left in three dorsoventral rotations as indicated (dorsal, lateral, and ventral vISH views). Actual *in vivo* expression shown by RNA *in situ* hybridization (right), Embryo orientation is anterior left, dorsoventral rotation as indicated, shown are developmental stage 5 or 6.

Supplementary Figure S5: Mapping efficacy, vISH thresholds & expression archetypes.



Legend to Supplementary Figure S5

(**A**) Uniqueness of bins as a function of the number of gene expression profiles. Increasing numbers of genes from the binarized dataset (x-axis) were assessed to determine the resulting number of unique bins (y-axis) in the reference atlas.

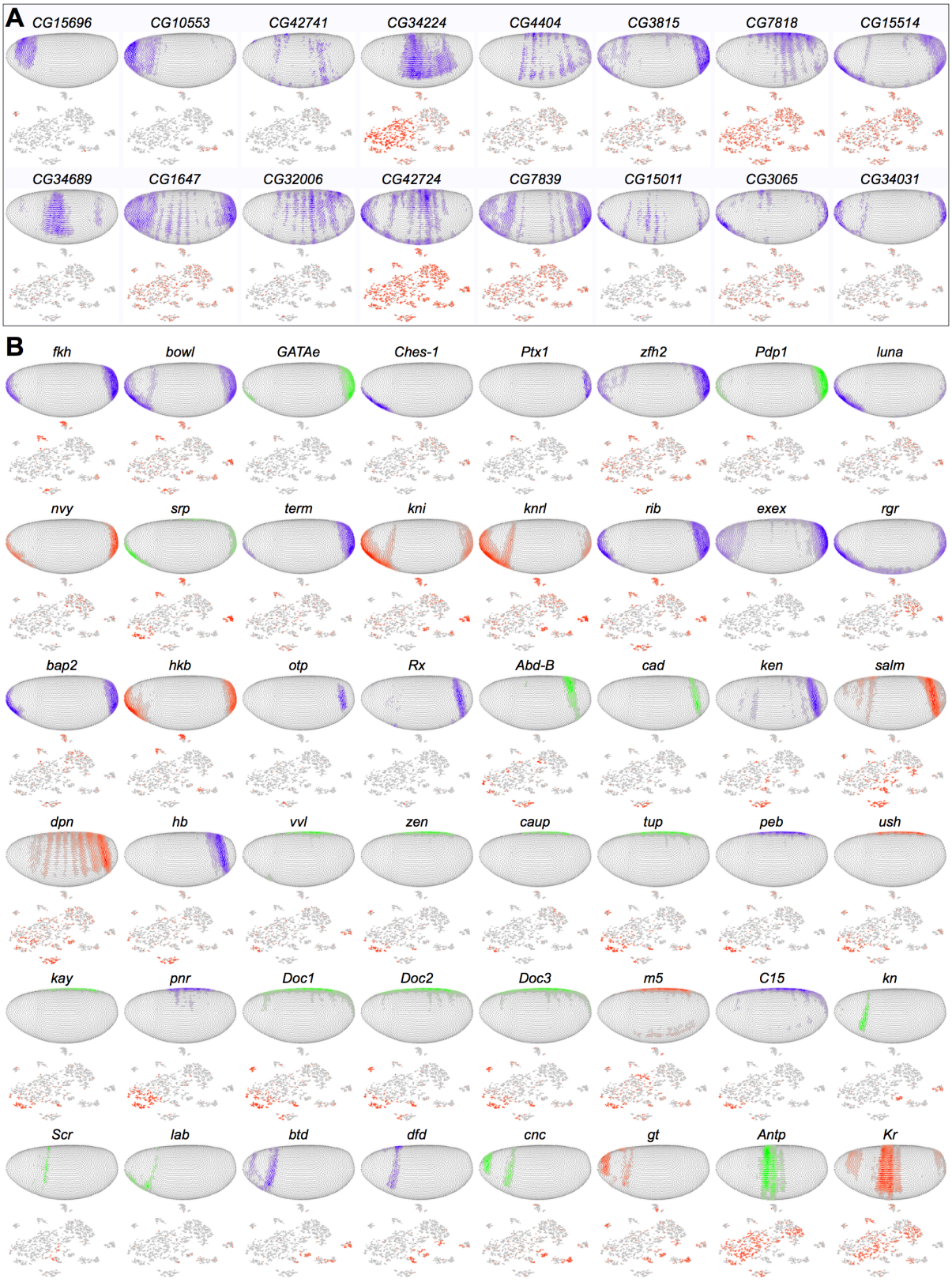
(**B**) Boxplots of discrepancies between computed vISHs and the BDTNP expression patterns, measured by Hamming distances, as a function of the number of cells used for the virtual embryo. 50 virtual embryos were computed for random cell subsets.

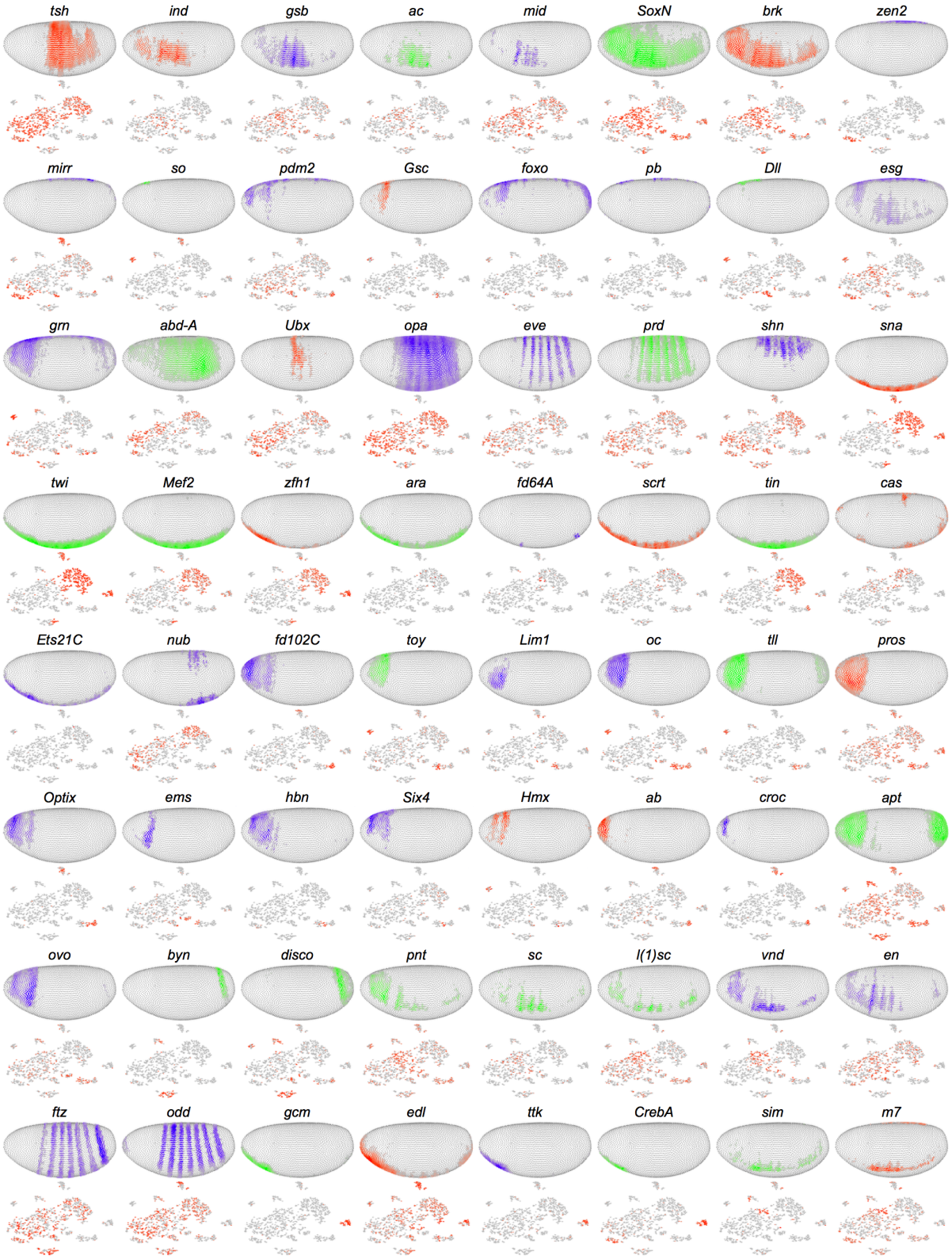
(**C**) Mapping confidence across all cells. At least 878 high quality cells were mapped with a very high confidence on the embryo (p-value < 0.05, see Supp. Materials and Methods for details), see Fig. 2C for spatial map of p-values.

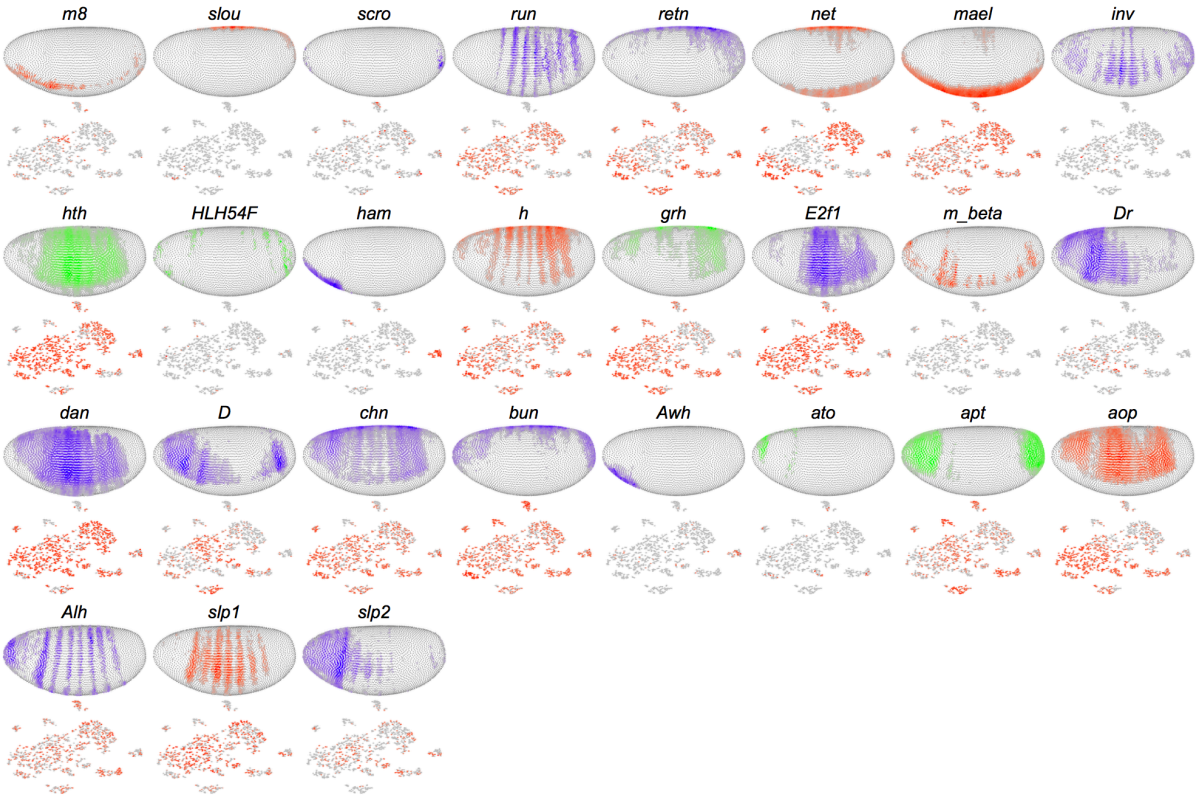
(**D**) Assessment of threshold values via Poisson noise simulations (see also Suppl. Materials and Methods for details). Thresholds (x-axis) are plotted against the sum of standard deviations across all bins (y-axis). As more stringent thresholds imply less bins, the above sum is expected to decrease monotonically. The existence of additional local minima in the cases of ftz, sna, SoxN and sog reveals threshold values around which vISHs are meaningful.

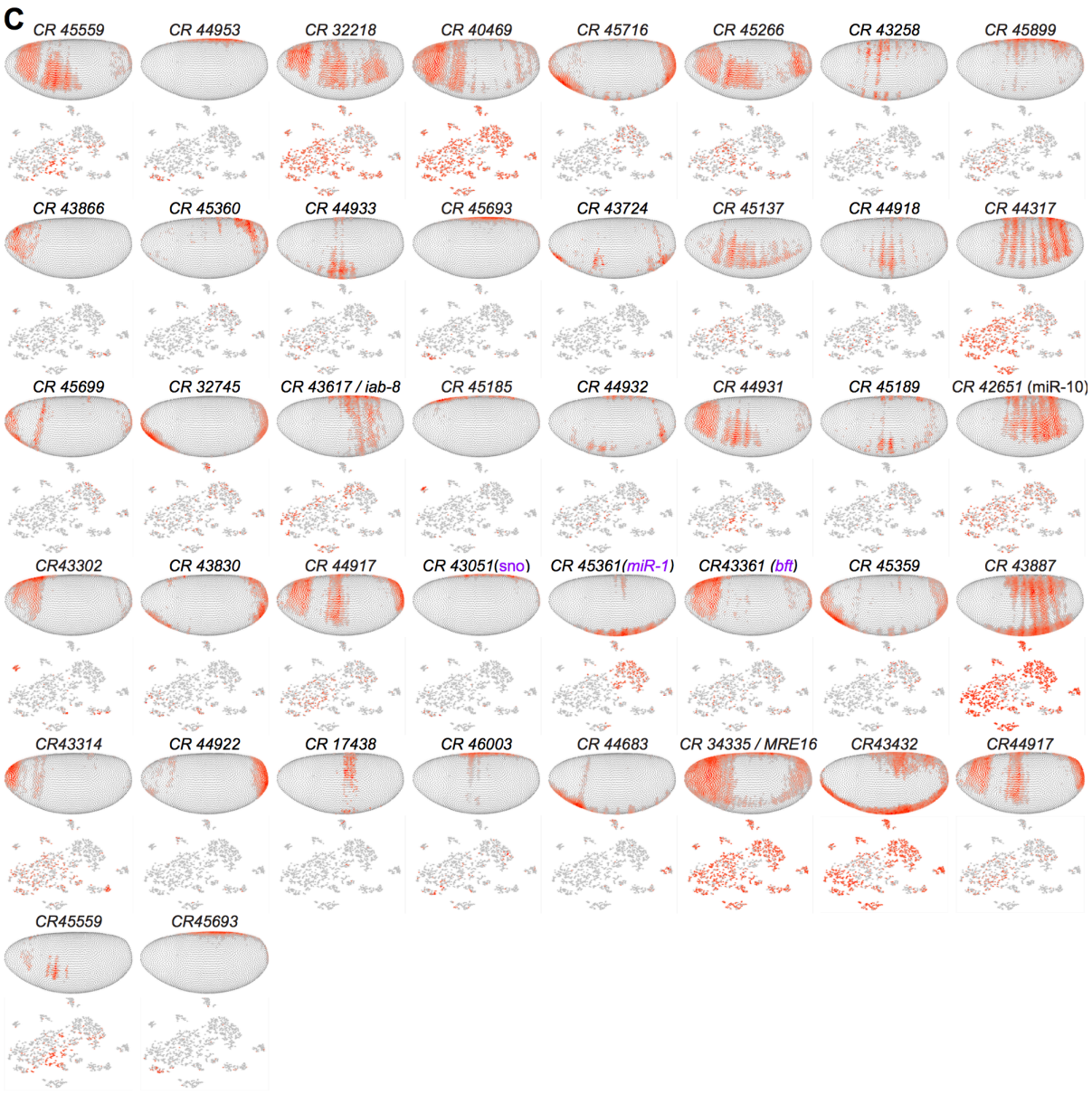
(**E**) The 476 most highly varied genes were clustered according to embryo-wide expression similarity. Heat map shows clustering of the correlation matrices of the computed vISHs with a global threshold value equal to 0.75. Average expression of genes within parent clusters reveals archetypal patterns shown as vISHs to the left. Color scale indicates correlation of Euclidean distances.

Supplementary Figure S6: vISH predictions of 155 transcription factors and 42 lncRNAs.




tsh
ind
gsb
ac
mid
SoxN
brk
zen2
mirr
so
pdm2
Gsc
foxo
pb
Dll
esg
grn
abd-A
Ubx
opa
eve
prd
shn
sna
twi
Mef2
zfh1
ara
fd64A
scrt
tin
cas
Ets21C
nub
fd102C
toy
Lim1
oc
tll
pros
Optix
ems
hbn
Six4
Hmx
ab
croc
apt
ovo
byn
disco
pnt
sc
l(1)sc
vnd
en
ftz
odd
gcm
edl
ttk
CrebA
sim
m7


m8
slou
scro
run
retn
net
mael
inv
hth
HLH54F
ham
h
grh
E2f1
m_beta
Dr
dan
D
chn
bun
Awh
ato
apt
aop
Alh
slp1
slp2

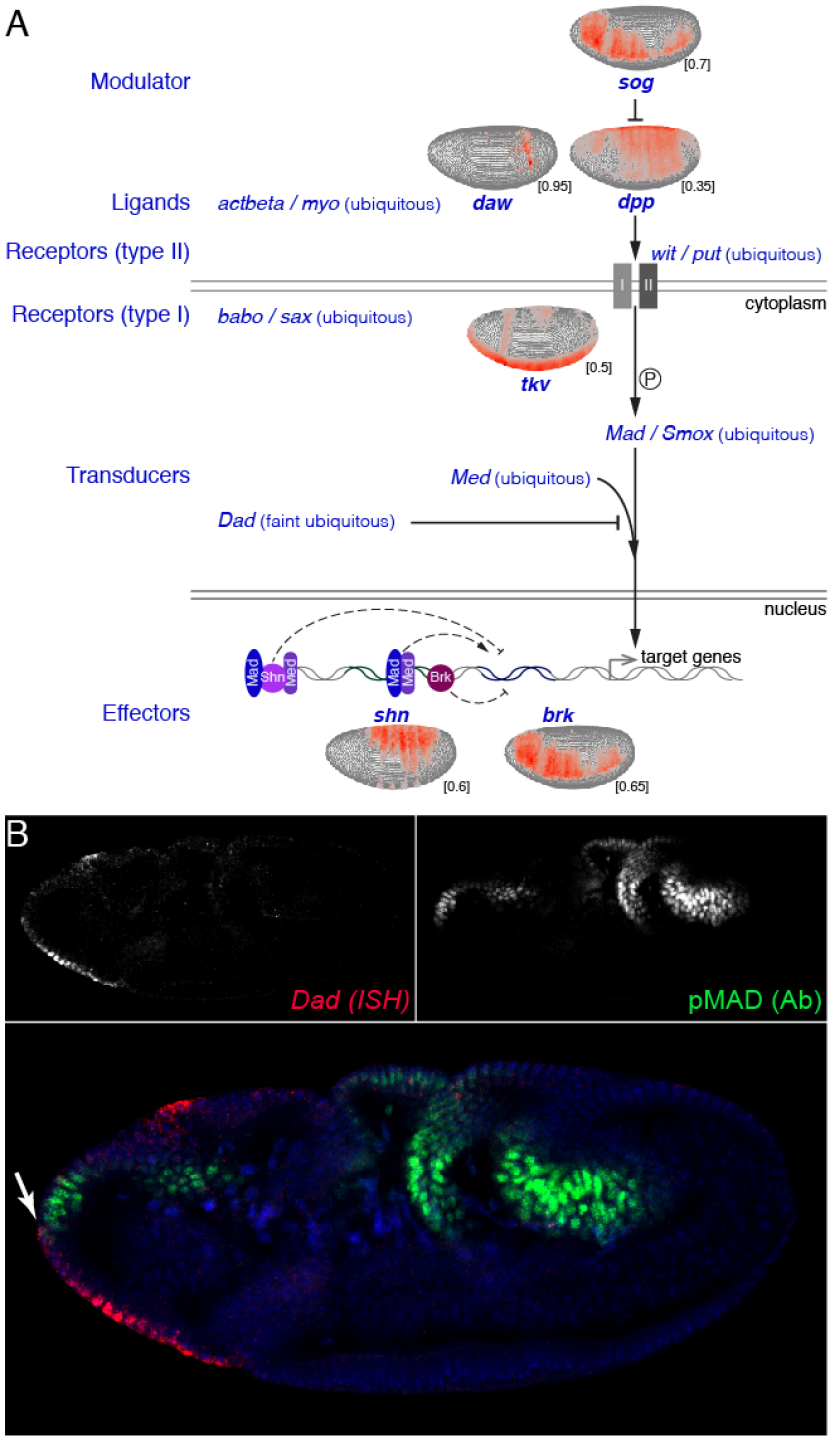


(**A**) vISH prediction and cluster expression of 16 putative transcription factors.

(**B**) vISH prediction and cluster expression of 139 known transcription factors. Those reported primarily as activators are indicated in green, repressors in red, blue where unclear.

(**C**) vISH prediction and cluster expression of 42 lncRNAs

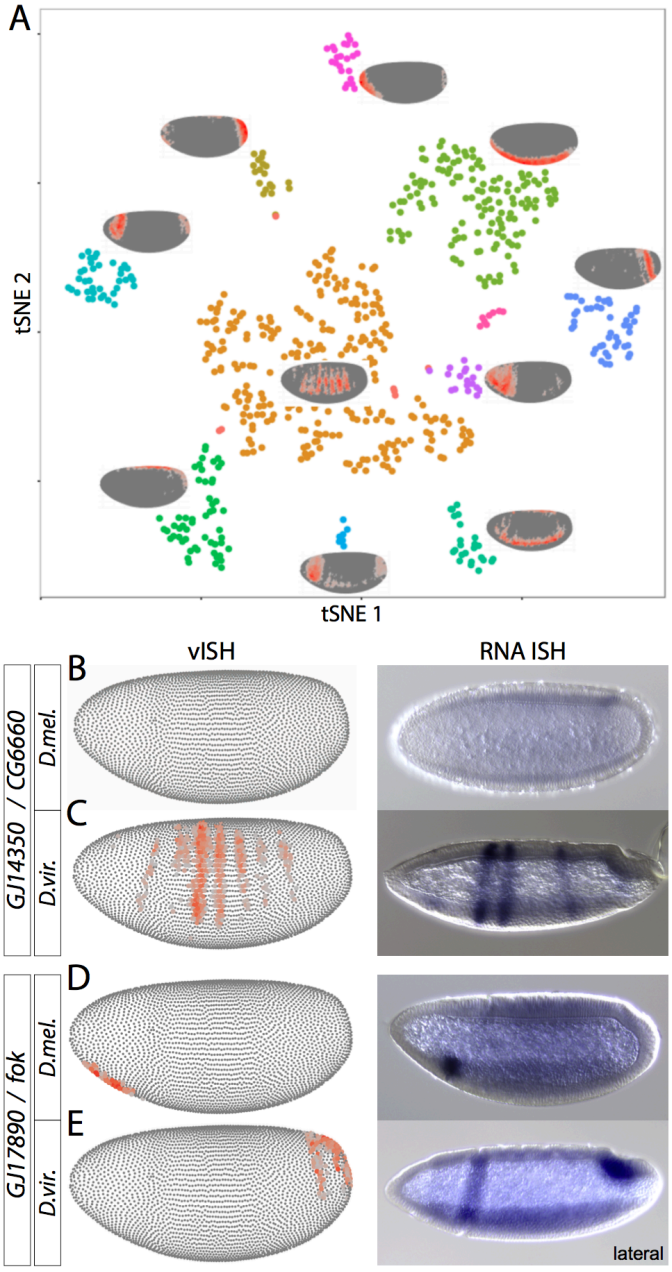
Supplementary Figure S7: *Dad* expression inhibits anterior Dpp signaling.



(**A**) Predicted patterned expression of major Dpp signaling components suggests spatial modulation of Dpp signaling.

(**B**) Co-staining for pSMAD (Ab) and *Dad* (RNA *in situ* hybridization) in a stage 8 embryo. *Dad* expression is detected in anterior regions. Exclusivity of signal suggests repression of Dpp signaling by Dad in anterior regions (white arrow).

Supplementary Figure S8: vISH detects evolutionary changes in expression patterns



(**A**) t-SNE representation of cells from stage 6 *D. virilis* embryos; spatial localization of each cell population on the embryo is indicated as in Fig. 4D.

(**B–E**) Gene expression divergence between gene pairs predicted by vISH (left) and experimentally validated by RNA *in situ* hybridization (right) for the *D. melanogaster* genes *CG6660* (**B**) and *fok* (**D**) and their *D. virilis* homologs *GJ14350* (**C**) and *GJ17890* (**E**).

Note: *CG6660* and *fok* are located within introns of encompassing genes. Drop-seq has a 3' transcript signature which often allows for unambiguous gene assignment. For aligned reads, we found that there are virtually no reads mapping to *CG6660* in *melanogaster, while fok* is expressed – reads clearly belong to *fok* and not the encompassing gene. Thus, transcript overlap is not a confounding factor in this analysis.