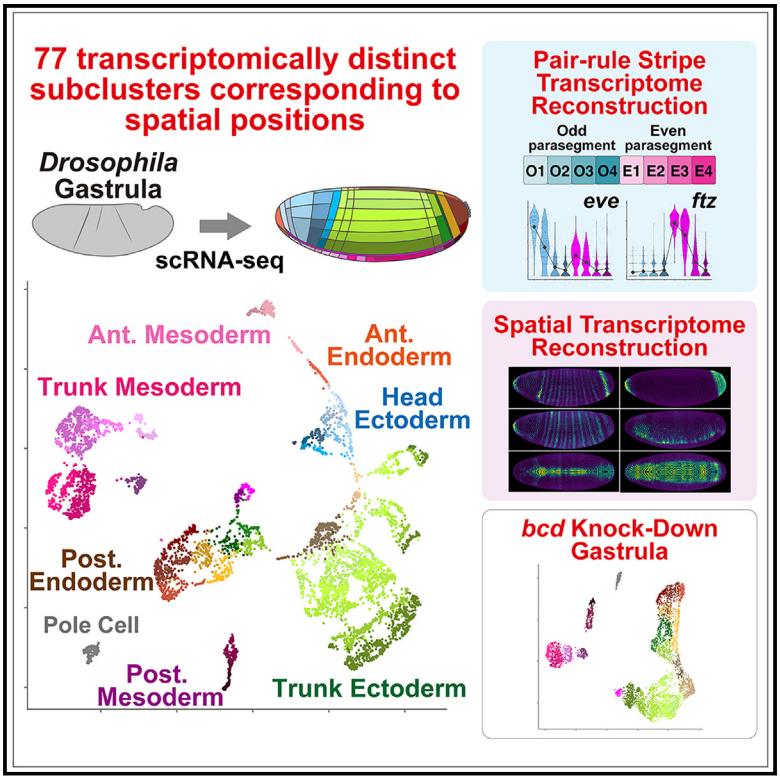
Cell Reports

Resource

Single-cell transcriptome atlas of *Drosophila* gastrula 2.0

Graphical abstract



Highlights

- The single-cell transcriptome analysis of *Drosophila* gastrula identifies 77 subclusters
- The expression profiles of plasma-membrane-related genes distinguished future germ layers
- A limited number of genes showed significant stripe patterns in the lateral ectoderm
- The genome-wide and accurate spatial patterns of gene expression were reconstructed

Authors

Shunta Sakaguchi, Sonoko Mizuno, Yasushi Okochi, ..., Mitsutaka Kadota, Honda Naoki, Takefumi Kondo

Correspondence

takefumi.kondo@riken.jp

In brief

Sakaguchi et al. established the single-cell transcriptome atlas of *Drosophila* gastrula and annotated cells into 77 identities corresponding to spatial positions. The expression profiles of plasma-membrane-related genes reflect future germ layer lineages better than transcription factors. In addition, they computationally reconstructed accurate spatial transcriptome at the single-cell resolution.

Sakaguchi et al., 2023, Cell Reports *42*, 112707
July 25, 2023
https://doi.org/10.1016/j.celrep.2023.112707

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

Single-cell transcriptome atlas of *Drosophila* gastrula 2.0

Shunta Sakaguchi,[1] Sonoko Mizuno,[1] Yasushi Okochi,[2,3] Chiharu Tanegashima,[4,10] Osamu Nishimura,[4,10] Tadashi Uemura,[1,5] Mitsutaka Kadota,[4,10] Honda Naoki,[2,6,7] and Takefumi Kondo[8,9,10,11,*]

[1]Laboratory of Cell Recognition and Pattern Formation, Graduate School of Biostudies, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[2]Laboratory of Theoretical Biology, Graduate School of Biostudies, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[3]Faculty of Medicine, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[4]Laboratory for Phyloinformatics, RIKEN Center for Biosystems Dynamics Research, Chuo-ku, Kobe, Hyogo 650-0047, Japan
[5]Center for Living Systems Information Science, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[6]Laboratory of Data-driven Biology, Graduate School of Integrated Sciences for Life, Hiroshima University, Higashihiroshima, Hiroshima 739-8511, Japan
[7]Theoretical Biology Research Group, Exploratory Research Center on Life and Living Systems (ExCELLS), National Institutes of Natural Sciences, Okazaki, Aichi 444-8585, Japan
[8]Graduate School of Biostudies, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[9]The Keihanshin Consortium for Fostering the Next Generation of Global Leaders in Research (K-CONNEX), Sakyo-ku, Kyoto 606-8501, Japan
[10]Present address: Laboratory for Developmental Genome System, RIKEN Center for Biosystems Dynamics Research, Chuo-ku, Kobe, Hyogo 650-0047, Japan
[11]Lead contact
*Correspondence: takefumi.kondo@riken.jp
https://doi.org/10.1016/j.celrep.2023.112707

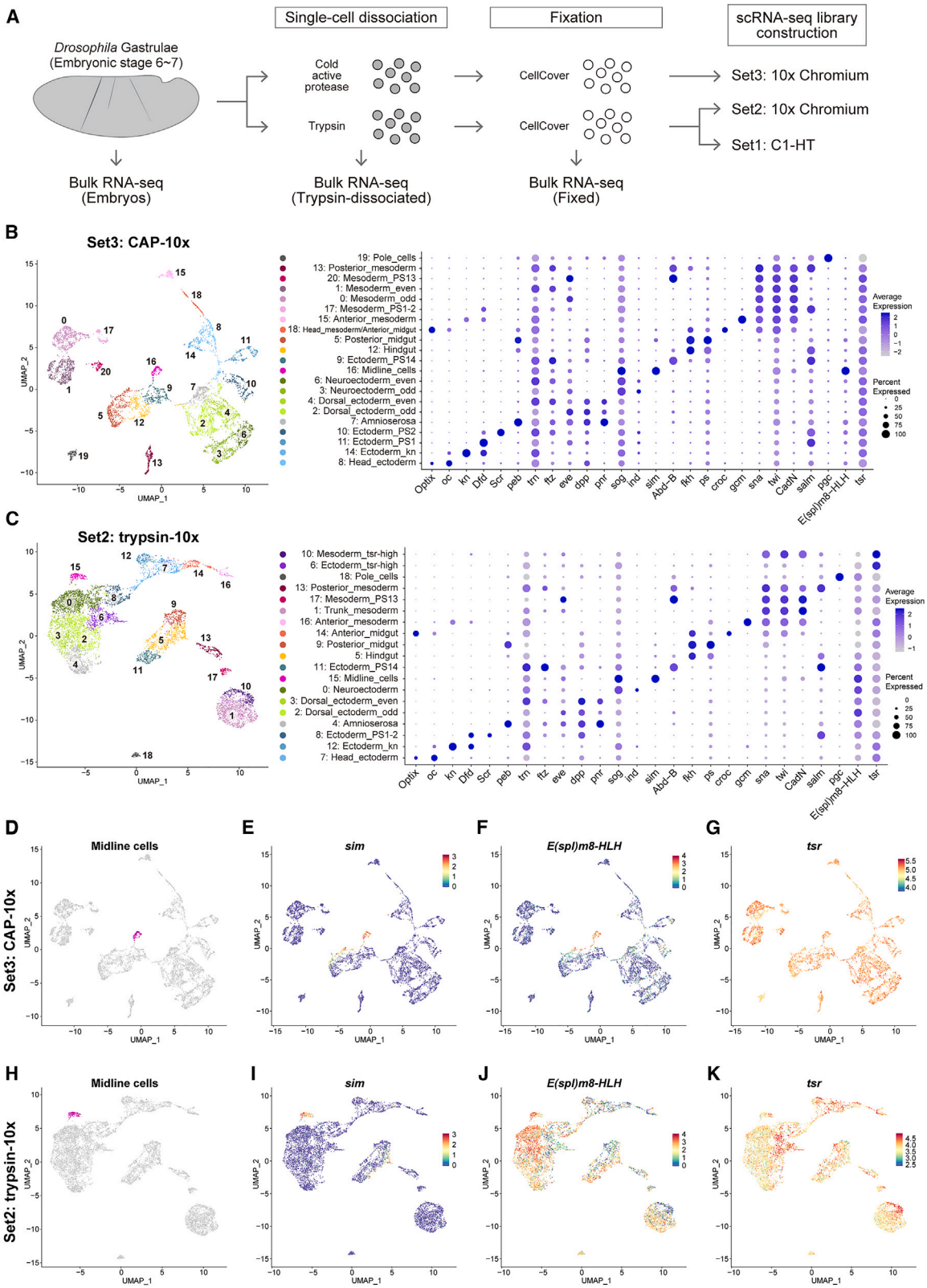
SUMMARY

During development, positional information directs cells to specific fates, leading them to differentiate with their own transcriptomes and express specific behaviors and functions. However, the mechanisms underlying these processes in a genome-wide view remain ambiguous, partly because the single-cell transcriptomic data of early developing embryos containing accurate spatial and lineage information are still lacking. Here, we report a single-cell transcriptome atlas of *Drosophila* gastrulae, divided into 77 transcriptomically distinct clusters. We find that the expression profiles of plasma-membrane-related genes, but not those of transcription-factor genes, represent each germ layer, supporting the nonequivalent contribution of each transcription-factor mRNA level to effector gene expression profiles at the transcriptome level. We also reconstruct the spatial expression patterns of all genes at the single-cell stripe level as the smallest unit. This atlas is an important resource for the genome-wide understanding of the mechanisms by which genes cooperatively orchestrate *Drosophila* gastrulation.

INTRODUCTION

One of the fundamental goals of developmental biology is to understand how genes cooperatively orchestrate morphogenesis and physiological functions at the cellular and tissue levels. In the scheme of programmed control of development, positional information is established by the combination of morphogens and cell-cell interactions. Then each cell is canalized into a specific fate depending on its position in embryos.[1–3] The dynamics of gene regulatory networks is considered to be essential for transforming gradual analog information into a discrete digital pattern of gene expression.[4,5] After cell fate canalization, it is widely considered that the combinatorial action of transcriptional factors (TFs) establishes a transcriptome profile that defines the pattern of cell and tissue morphogenesis.[6,7] Technological advances have led to a genome-wide understanding of gene expression control and cell differentiation processes.[8,9] However, our understanding has yet to sufficiently link the transcriptome with cell and tissue behavior. To understand how TFs orchestrate cell/tissue behavior through establishing characteristic transcriptomes, it is crucial to have a single-cell transcriptome atlas of developing embryos containing both accurate spatial and lineage information.

Drosophila gastrulae have been an excellent model system for studying multicellular morphogenesis for decades.[10–12] The cell-specific expression of genes has been extensively analyzed using *in situ* hybridization (ISH).[13,14] However, it is still difficult to obtain quantitative transcriptome profiles of each cell together with spatial information. For example, the Berkeley Drosophila Transcription Network Project (BDTNP) established a gene-expression database as a virtual embryo by integrating the quantified fluorescence ISH (FISH) data from multiple embryos. However, the number of genes analyzed was less than 100.[15–17]



(legend on next page)

In this decade, single-cell RNA sequencing (scRNA-seq) has become a standard technique that enables the analysis of transcriptomes at the single-cell level.[18,19] Since cells need to be dissociated into the single-cell level from tissues for scRNA-seq, several computational methods have also been developed to restore the spatial information of scRNA-seq data.[20–23] For *Drosophila* gastrulae, scRNA-seq analysis and spatial reconstruction of gene expression were performed.[24] However, there is room for improvement in its quality because, for many genes, the reconstructed spatial pattern from scRNA-seq data did not match the original pattern uncovered by ISH. For example, the famous 14-stripes expression patterns of segment polarity genes (e.g., wingless [*wg*], engrailed [*en*]) have not been fully reconstructed in the *Drosophila* Virtual Expression eXplorer (DVEX; https://shiny.mdc-berlin.de/DVEX/). Recently, we computationally overcame this limitation by developing a machine learning method, Perler, based on generative linear mapping.[25]

Although we and others have made efforts to improve computational methods,[26–28] there still is a fundamental limitation with the scRNA-seq data; the number of high-quality cells sequenced was 1,297, which is much lower than the number of cells in a gastrula (approximately 6,000 cells),[24] so there might be cells not in the scRNA-seq data. In addition, Karaiskos et al. distinguished only 13 clusters in their scRNA-seq data, which may not be sufficient to fully describe the entire *Drosophila* gastrula. Therefore, further acquisition of scRNA-seq data will enable us to perform a more precise transcriptomic characterization and improve the spatial-transcriptome reconstruction.

In this study, we aimed to improve the single-cell transcriptome atlas of *Drosophila* gastrula and update it to version 2.0. To this end, we profiled single-cell transcriptomes that can be annotated into 77 clusters and further recapitulated the transcriptome profiles of each pair-rule stripe. Finally, we cataloged the spatial expression patterns of all genes in *Drosophila* gastrulae via computational integration with reference spatial expression patterns using Perler[25] or NovoSpaRc.[26] Since *Drosophila* gastrula is one of the most well-characterized multicellular systems, this atlas provides an important quantitative resource for a wide range of biological fields as a reference for understanding the principles that link gene regulatory networks and cell differentiation to cell behavior and tissue morphogenesis.

RESULTS

scRNA-seq of fixed cells dissociated from *Drosophila* embryos

To conduct scRNA-seq analysis for *Drosophila* gastrulae, we first reexamined the protocols of the single-cell dissociation step. We tried the mechanical dissociation protocol as previously reported[24] but did not recover enough cells in our hands. Therefore, we next attempted gentle breaking of the vitelline membrane and dissociated the cells enzymatically. It has recently been recognized that enzymatic dissociation at room temperature leads to artificial changes in the transcriptome. To overcome this problem, cell dissociation using cold-active protease (CAP, also known as subtilisin A from *Bacillus licheniformis*) at low temperatures has been shown to be a good solution.[29–32] We examined both enzymes to assess the artificial effect of trypsin and the usefulness of CAP on single-cell dissociation of *Drosophila* embryos (Figure 1A). In addition, a step of non-cross-linking fixation using CellCover is added to avoid the gene expression change after cell dissociation to cell lysis steps (Figure 1A). Bulk RNA sequencing (RNA-seq) analysis indicates CellCover fixation could preserve the transcriptome profile (Figure S1A).

We then performed scRNA-seq analysis using three different protocols: set 1, trypsin dissociation and Fluidigm C1 mRNA Seq HT IFC (Set1 trypsin-C1HT); set 2, trypsin dissociation and 10x Genomics Chromium V3.1 (Set2 trypsin-10x); and set 3, CAP dissociation and 10x Genomics Chromium V3.1 (Set3 CAP-10x) (Figure 1A). In all datasets, gene expression was quantified by counting the different unique molecular identifiers (UMIs), short random nucleotide sequences added to each transcript in the reverse transcription step per gene and per cell.[33] After filtering the data of high-quality cells (see STAR Methods for details), 1,243, 7,314, or 6,180 cells remained, and the expression of 4,480, 3,222, or 4,053 genes per cell in the median was detected for Set1, Set2, or Set3 data, respectively (Table S1). Set1 trypsin-C1HT data showed higher median UMI counts per cell (152,429) than the other 10x data (22,506 [Set2] and 37,610 [Set3]). One of the reasons for higher sensitivity is greater sequencing depth per cell, which is consistent with a previous report that the C1 platform can produce rich information.[34] Unsupervised clustering and the extraction of marker genes for each cluster using Seurat v3[22] revealed that each dataset contains all major cell types, such as mesoderm (*snail* [*sna*], *twist* [*twi*]), trunk dorsal ectoderm (*decapentaplegic* [*dpp*], *pannier* [*pnr*]), trunk neuroectoderm (*short gastrulation* [*sog*]), head ectoderm (*Optix*, *ocelliless* [*oc*]), terminal endoderm (*fork head* [*fkh*]), pole cells (*polar granule component* [*pgc*]), and dorsal amnioserosa cells (*pebbled* [*peb*]), indicating there was no cell type bias in all datasets (Figures 1B, 1C, S1B, and S1C). The Set1 trypsin-C1HT data were composed of four biological replicates, and there was no obvious batch effect among them, indicating the reproducibility of our protocol (Figure S1B).

Figure 1. Comparison between trypsin and CAP dissociation for scRNA-seq

(A) Schematic diagram of the data acquired in this study.

(B and C) Uniform Manifold Approximation and Projection (UMAP) plot of the Set3 CAP-10x scRNA-seq data (B) and the Set2 trypsin-10x scRNA-seq data (C) with Seurat cluster information. Dot plot shows the expression patterns of typical marker genes for each cluster.

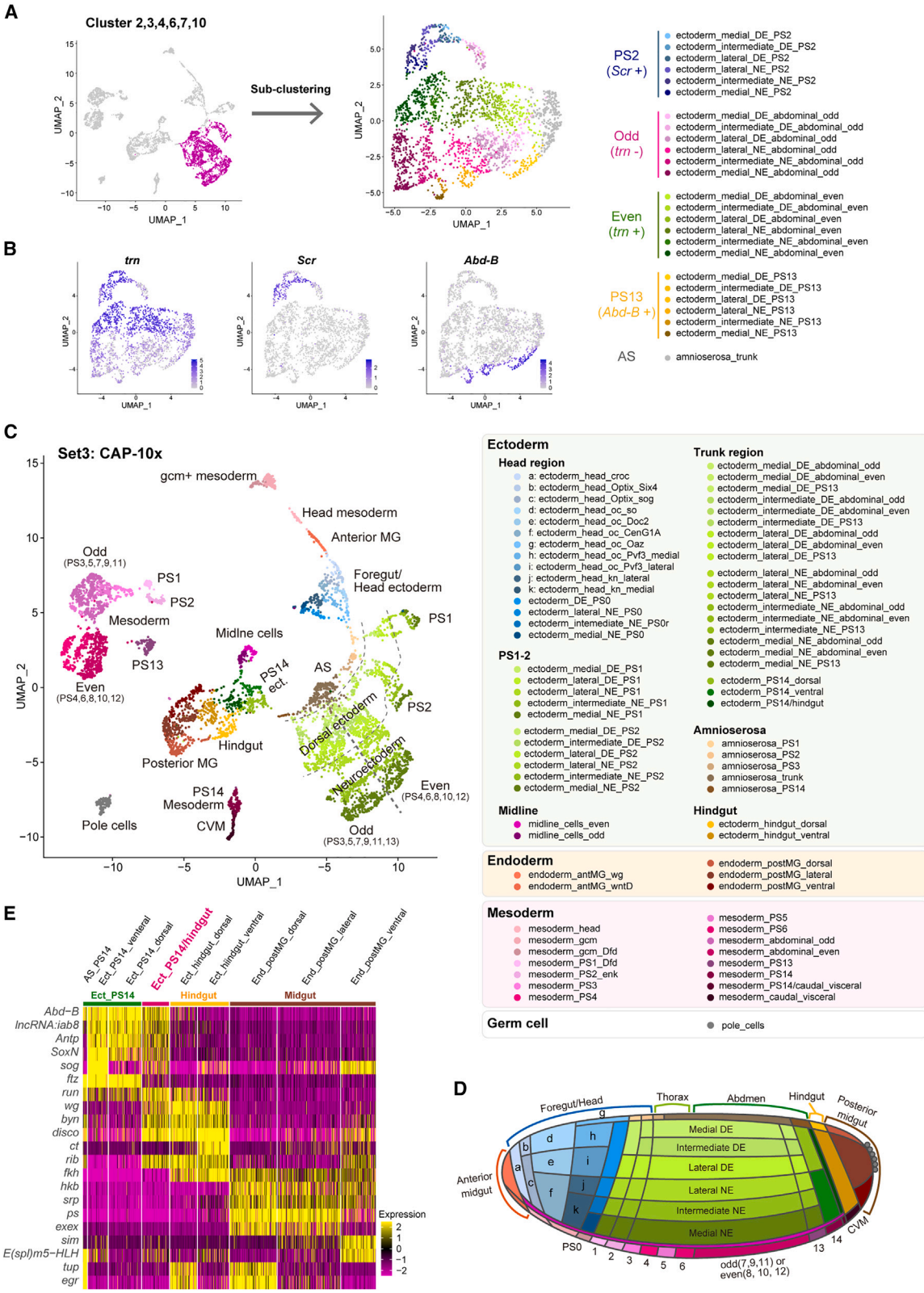
(D) Midline cells are colored magenta in the UMAP plot of the Set3 CAP-10x data.

(E–G) Expression patterns of *sim* (E), *E(spl)m8-HLH* (F), and *tsr* (G) in the Set3 CAP-10x data.

(H) Midline cells are colored magenta in the UMAP plot of the Set2 trypsin-10x data.

(I–K) Expression patterns of *sim* (I), *E(spl)m8-HLH* (J), and *tsr* (K) in the Set2 trypsin-10x data.

The expression patterns in (E–G) and (I–K) represent the log-transformed values after SCTransform normalization.



(legend on next page)

Trypsin dissociation causes artificial upregulation of Notch target genes

To reveal the extent to which different dissociation methods affect the single-cell transcriptome profile, we inspected all scRNA-seq data in depth. Midline cells (mesoectoderm) are known to highly express Notch target genes *single-minded* (*sim*) and some *Enhancer of split* (*E(spl)*) complex genes, such as *E(spl)m5-helix-loop-helix* (*E(spl)m5-HLH*), and *E(spl)m8-HLH*.[35–37] In Set3 CAP-10x data, all these genes were specifically expressed in the midline cell cluster (Figures 1D–1F). On the other hand, we noticed that, in Set2 trypsin 10x data, although the midline cell cluster was identified by specific expression of *sim*, *E(spl)m8-HLH* showed strong and broad expression not only in the midline cell cluster but also in other clusters (Figures 1H–1J). To assess whether the artificial induction of *E(spl)* complex genes is due to trypsin treatment, we performed bulk RNA-seq. Although there was a high correlation between intact embryos and trypsin-dissociated cells, trypsin-dissociated cells showed higher expression of *E(spl)* complex genes, indicating that trypsin treatment artificially upregulates the expression independent of cell type (Figures S1G and S1H).

In addition, Set2 trypsin-10x data had Seurat clusters that showed high *twinstar* (*tsr*) expression (Figure 1K). For example, the trunk mesoderm was divided into two clusters (clusters 1 and 10 in Figure 1C), and cluster 10 showed higher expression of *tsr* than cluster 1 (Figure 1K). Furthermore, cluster 6, which seems to belong to the trunk ectoderm, also showed high *tsr* expression. On the other hand, clusters 6 and 10 showed relatively low expression of *E(spl)* complex genes (Figure 1J). To characterize these cells with high *tsr* expression (*tsr*-high cells), we identified highly expressed genes in cluster 10 compared with cluster 1 and performed gene set enrichment analysis. By Gene Ontology (GO) enrichment analysis, the term ''oxidative phosphorylation'' was enriched in highly expressed genes of the *tsr*-high cell cluster (Figure S1I), suggesting that these cells exhibited some metabolic stress responses.

These results suggest that there are two types of cells in Set2 trypsin-10x data: one increased some of the Notch target genes (*E(spl)* complex genes), and the other showed some kind of stress response upon trypsin treatment. Furthermore, Set1 trypsin-C1HT data also showed strong and broad expression of *E(spl)* complex genes (Figures S1D and S1E) and clusters showing high *tsr* expression (Figure S1F). On the other hand, in Set3 CAP-10x data, there was no such cluster (Figures 1F and 1G), indicating that these cellular responses are specific to trypsin treatment but not CAP treatment.

Identification of 77 transcriptomically distinct subclusters in Set3 data

To investigate the detailed single-cell transcriptome diversity in the gastrulae, we performed subclustering of all scRNA-seq data and manually annotated each subcluster based on known gene expression patterns from databases (Berkeley Drosophila Genome Project [BDGP] *in situ* database [https://insitu.fruitfly.org],[38–40] Fly-FISH [https://fly-fish.ccbr.utoronto.ca])[13,14] and information from the literature. In the trunk region of the gastrula, the amnioserosa, dorsal ectoderm, ventral neuroectoderm, mesoectoderm (midline cells), and mesoderm emerge along the dorsal-ventral (DV) axis.[41] On the other hand, along the anterior-posterior (AP) axis, cells were divided into 14 parasegments (PSs),[12,42–46] and even parasegments express *tartan* (*trn*) and *fushi tarazu* (*ftz*) specifically.[47,48] Therefore, we inferred the origin of the trunk ectodermal cells in Set3 for each of the AP and DV axes separately. We picked up the trunk ectoderm cells (Figure 2A, corresponding to PS2–13) and assigned these cells to seven DV identities by k-means clustering with 35 selected DV genes (see STAR Methods for details). Along the AP axis, we assigned these cells to four AP identities (parasegment 2 [PS2], trunk ectoderm odd [PS3, 5, 7, 8, 9, 11], trunk ectoderm even [PS4, 6, 8, 10, 12], and PS13) (Figure 2A). Combining these AP and DV identities divided the trunk ectoderm cells into 25 subclusters (Figure 2A). We also performed subclustering for other Seurat clusters in Set3 and eventually divided the cells into 77 subclusters for Set3 data (Figures 2C, 2D, and S2; Table S2). Notably, the head region located anterior to parasegment 1 could be divided into 15 subclusters, including future foregut primordium (Figure 2C), indicating that the head region can be subdivided into smaller areas at the transcriptome level. This result may reflect the complexity of later head development. Note that, during the subclustering process, 62 potential doublet cells were identified and discarded from the dataset (see STAR Methods for details). The remaining Set3 data consisted of 6,118 cells.

We also performed subclustering of Set1, Set2, and previously reported data in Nikos Karaiskos et al. (NK-data)[24] similarly. The apparent differences between Set2 and Set3 are that we could not clearly separate even/odd parasegmental identity of ventral neuroectoderm in Set2 data. In Seurat clustering using all cells, the dorsal ectoderm appears to be divided into even and odd parasegments, while the ventral neuroectoderm is not separated (Figures S3A and S3B). Furthermore, even after subclustering, the lateral ectoderm could not be clearly divided into even and odd parasegment along the DV axis. These differences could be due to transcriptome distortion by trypsin treatment. In the case of Set1 and NK-data, only 32 and 28 subclusters could be distinguished, respectively (Figures S3C and S3D),

Figure 2. Seventy-seven subclusters of *Drosophila* gastrulae identified from scRNA-seq data

(A) Subclustering of trunk ectodermal cells (PS2-13). (Left) Trunk ectodermal cells are colored magenta in the UMAP plot of the Set3 CAP-10x data. (Right) UMAP plot of trunk ectodermal cells with 25 subcluster information.

(B) Expression patterns of genes expressed in specific parasegments. *trn*, even parasegment; *Scr*, PS 2; *Abd-B*, PS13.

(C) UMAP plot of the Set3 CAP-10x scRNA-seq data with information on the 77 subclusters.

(D) Schematic diagram showing the inferred spatial location of each subcluster in gastrula.

(E) Heatmap showing the typical marker genes for subclusters of the posterior ectodermal and endodermal cells. AS, amnioserosa; CVM, caudal visceral mesoderm; DE, dorsal ectoderm; MG, midgut; NE, neuroectoderm.

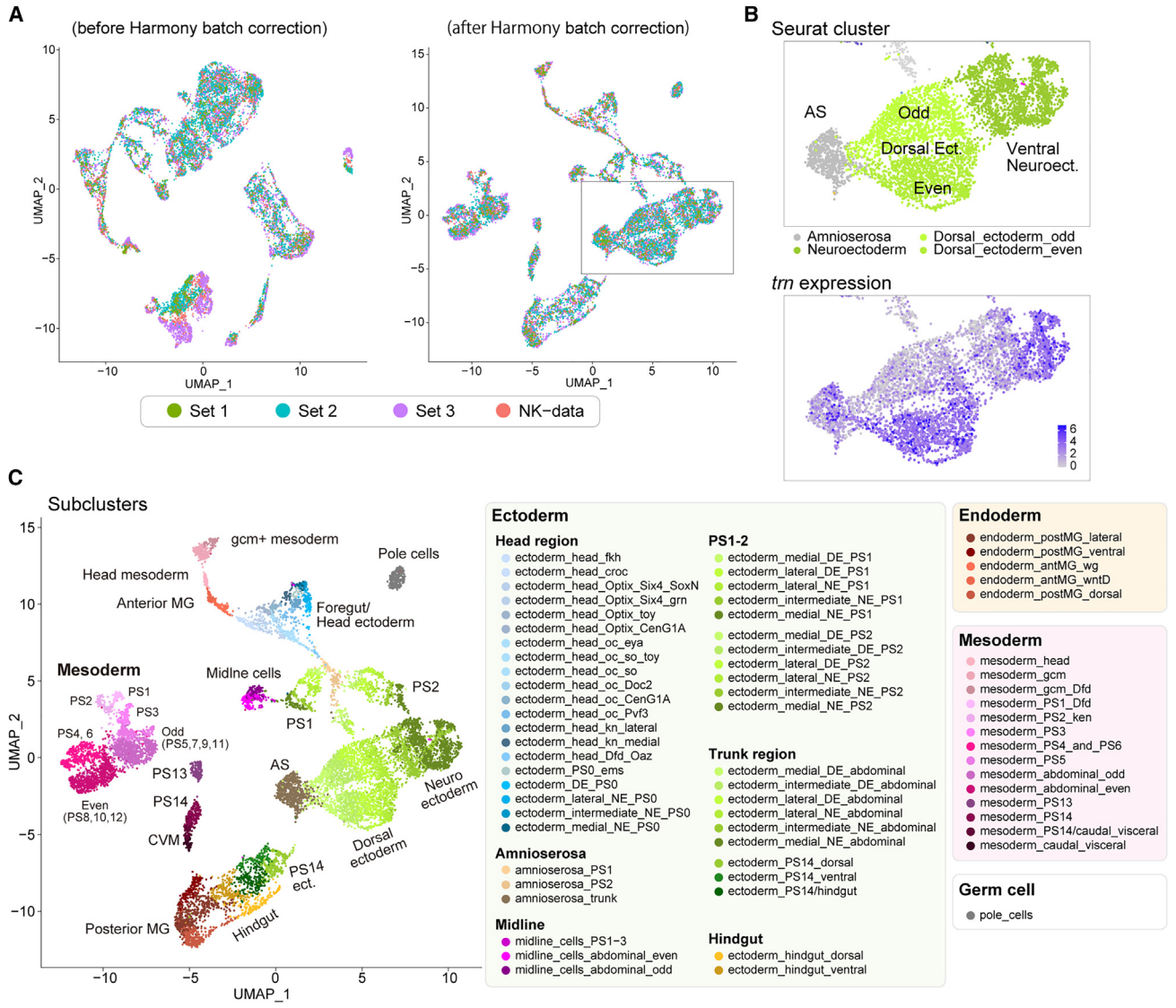


Figure 3. Harmony integration of all datasets

(A) UMAP plot of scRNA-seq data merging Set1, Set2, Set3, and NK-data before Harmony batch correction (left) and after Harmony batch correction (right).

(B) Enlarged view of the integrated UMAP of the boxed area in (A). (Top) Seurat clusters. (Bottom) *trn* expression.

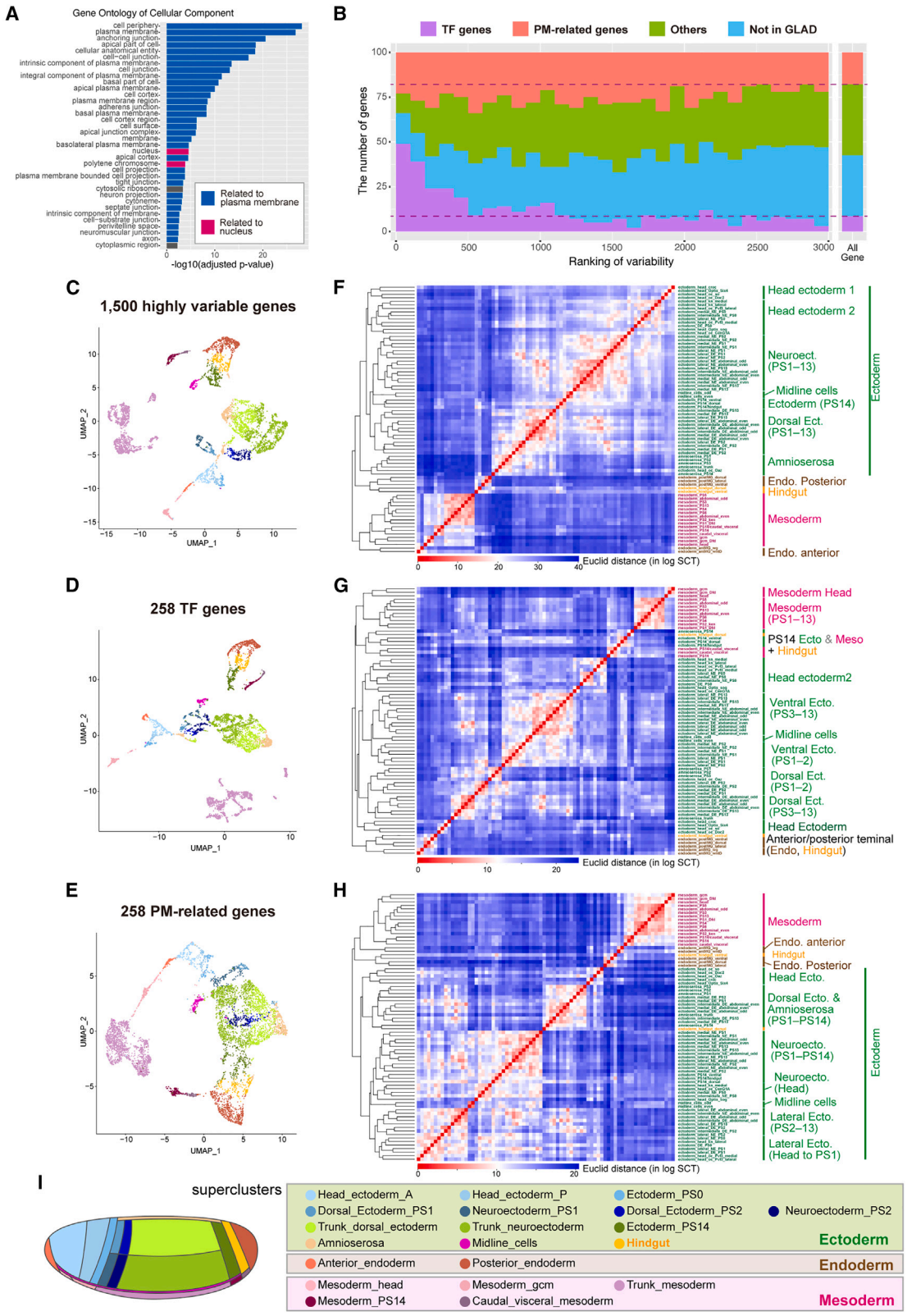
(C) UMAP plot of the Harmony-integrated data with information on the 68 subclusters.AS, amnioserosa; CVM, caudal visceral mesoderm; DE, dorsal ectoderm; MG, midgut; NE, neuroectoderm.

suggesting that sequencing more cells, rather than deeper sequencing of each cell, is more important for identifying minor cell types, as already mentioned.[49,50] The results of the subclustering are summarized in Table S2.

To further test whether more cells allow us to identify more clusters, we integrated all scRNA-seq data using Harmony[51] (Figure 3A) and did clustering. However, we could subdivide the Harmony-integrated data into only 68 distinct subclusters (Figure 3C). Clustering of the integrated data tended to fail to separate parasegmental identities. The integrated data showed a gradual *trn* expression pattern in the ventral ectoderm on the Uniform manifold approximation and projection (UMAP) plot (Figure 3B). However, Seurat clusters did not correspond to even/odd parasegmental identity with high or low *trn* expression. Subclusters that could be identified in Set3 but not in the integrated data were those corresponding to PS13 ectoderm, PS3 and 14 amnioserosa, and PS4 and 6 mesoderm (Figures 2C and 3C). These results suggest that increasing the number of cells does not necessarily increase resolution and could be exacerbated by adding lower-quality data, such as for trypsin-dissociated cells.

The Set1 trypsin C1HT data show a deeper depth of transcript counts per cell and a larger total number of detected genes (14,785 in Set1, 13,214 in Set2, and 13,335 in Set3)

CellPress
OPEN ACCESS



(legend on next page)

(Table S1). Also, 1,374 genes were only detected in Set1 but not in Set2 and Set3. To test the usefulness of this depth, we compared marker gene detection sensitivity between Set1 and Set2, which were based on trypsin dissociation, for each Seurat cluster detected in Harmony-integrated data. As a result, we found that 331 marker genes were only detected in the Set1 trypsin C1HT data, and they showed lower expression levels than Set1/Set2 common or Set2 only marker genes (Figure S3E). Furthermore, the same comparison was performed between Set1 and Set3, and 321 marker genes detected only in Set1 showed lower expression than Set1/Set3 common or Set3 only marker genes. One example of Set1-only marker genes is *Sarcoglycan delta* (*Scgdelta*), detected as a posterior midgut marker. Its expression was rarely detected in Set2 and Set3 (Figure S3F). These results suggest that deeper Set1 data help detect low-expressed genes and characterize each cell more comprehensively.

Because (1) the CAP dissociation could well preserve the original expression patterns of some Notch target genes, and (2) the integration of all available datasets did not improve the quality compared with Set3 only, we mainly focused on the Set3 CAP-10x dataset using CAP for further analysis.

Potential intermediate-state cells

We noticed two of the 77 subclusters in Set3 CAP-10x were difficult to annotate with an equivocal identity using well-known marker genes. Three subclusters were identified from the subclustering of cluster 18 (Figure S2B). One of them was "endoderm_antMG_wg," which specifically expresses anterior endoderm markers, such as *fkh*, *huckebein* (*hkb*), and *serpent* (*srp*), as well as *wg*. Another subcluster named "mesoderm_head" showed the expression of mesoderm markers, such as *sna*, *twi*, and *heartless* (*htl*). The third subcluster, expressing *wnt inhibitor of Dorsal* (*wntD*), was positive for both endoderm and mesoderm markers. Consistent with previous reports that mesoderm genes *sna* and *twi* are also expressed in endodermal cells,[52–54] and *wntD* is known to repress mesoderm differentiation,[55,56] this cluster did not show mesoderm gene expressions other than *sna* and *twi*. Therefore, we annotate this third subcluster as "endoderm_antMG_wntD" at this moment.

Another example of a potential intermediate state is "ectoderm PS14/hindgut," which seems to be between PS14 ectoderm and hindgut (Figure 2E). It expressed both PS14 ectoderm markers (*Abdominal B* [*Abd-B*]) and hindgut markers (*disconnected* [*disco*], *wg*, and *brachyenteron* [*byn*]). Multiplex FISH of *Abd-B* and *byn* revealed the presence of cells co-expressing them at the border between the future epidermis and hindgut (Figure S2C), suggesting that intermediate-state cells exist between the epidermis and hindgut in the gastrula. We annotate this subcluster as ectoderm PS14/hindgut at this moment. We concluded that the scRNA-seq data contain enough information to distinguish the spatial origin at the single-cell level and transient intermediate states that have not been recognized.

Expression profile of plasma-membrane genes better represents the major cell types

Next, we analyzed the features of the transcriptome profile that contributed to the classification of each cell. GO term analysis revealed that genes encoding TF genes, as well as plasma-membrane (PM)-related genes were highly enriched in 3,000 highly variable genes (HVGs) (Figure 4A). Therefore, HVGs were classified into four categories based on the Gene List Annotation for *Drosophila* (GLAD) database[57]; TF genes, PM genes, non-TF, and non-PM genes in GLAD (other genes), and genes not included in GLAD (see STAR Methods for details). Since TF genes were enriched in the top 1,500 HVGs (Figure 4B), the top 1,500 HVGs were used for the hierarchical clustering analysis below to focus on the significance of TF enrichment. Hierarchical clustering of 76 subclusters (pole cells were removed) with the top 1,500 HVGs classified the subclusters into three germ layers (ectoderm, endoderm, and mesoderm), indicating that the differences in the cellular transcriptome at the gastrula stage reflect the differences among future cell lineages (Figures 4C and 4F). To understand how much information each gene set alone holds to characterize the cell type, the hierarchical clustering was performed for each of the four categories, using 258 genes with high variances in each category. The 258 TF genes well segregated each cell along with the original positions on the UMAP plot (Figure 4D). By hierarchical clustering using only TF genes, subclusters tended to be classified by spatial location compared with the case using all 1,500 HVGs (Figure 4G). For example, the subclusters of "ectoderm_PS14," "mesoderm_PS14," and "mesoderm_caudal_visceral" form a single group across the types of mesoderm and ectoderm.

On the other hand, the set of PM-related genes well reproduced the clustering pattern with all 1,500 HVGs, and hierarchical clustering categorized 76 subclusters with their germ layer identities beyond the spatial proximity in the embryo (Figures 4E and 4H). Furthermore, the sets of non-TF and/or non-PM genes also classified the subclusters into three germ layers (Figures S4B–S4E). These clustering analyses revealed that, without any prior functional knowledge about each gene, only the mRNA expression profiles of TF genes were insufficient to distinguish future cell lineages in this gastrula stage. On the

Figure 4. Clustering analysis with GLAD categories

(A) GO enrichment analysis of 3,000 HVGs using g:Profiler. The terms of cellular components are presented.

(B) The number of genes belonging to each category in each bin of 3,000 HVGs divided into 30 bins from the top. The red dotted lines show the expected numbers by randomly sampling TF- and PM-related genes.

(C–E) UMAP plot with 1,500 HVGs (C), 258 TFs in 1,500 HVGs (D), and top 258 PM-related HVGs (E). Cells were colored by super cluster information based on prior annotations (I).

(F–H) Hierarchical clustering analyses of 76 subclusters based on the Euclidean distances in log-transformed gene-expression space with top 1,500 HVGs (F), 258 TFs in 1,500 HVGs (G), and top 258 PM-related HVGs (H). Subclusters were colored based on the future germ layers.

(I) Super-cluster information used to color the cells in the UMAP plot. See Table S2 for details.

CellPress
OPEN ACCESS

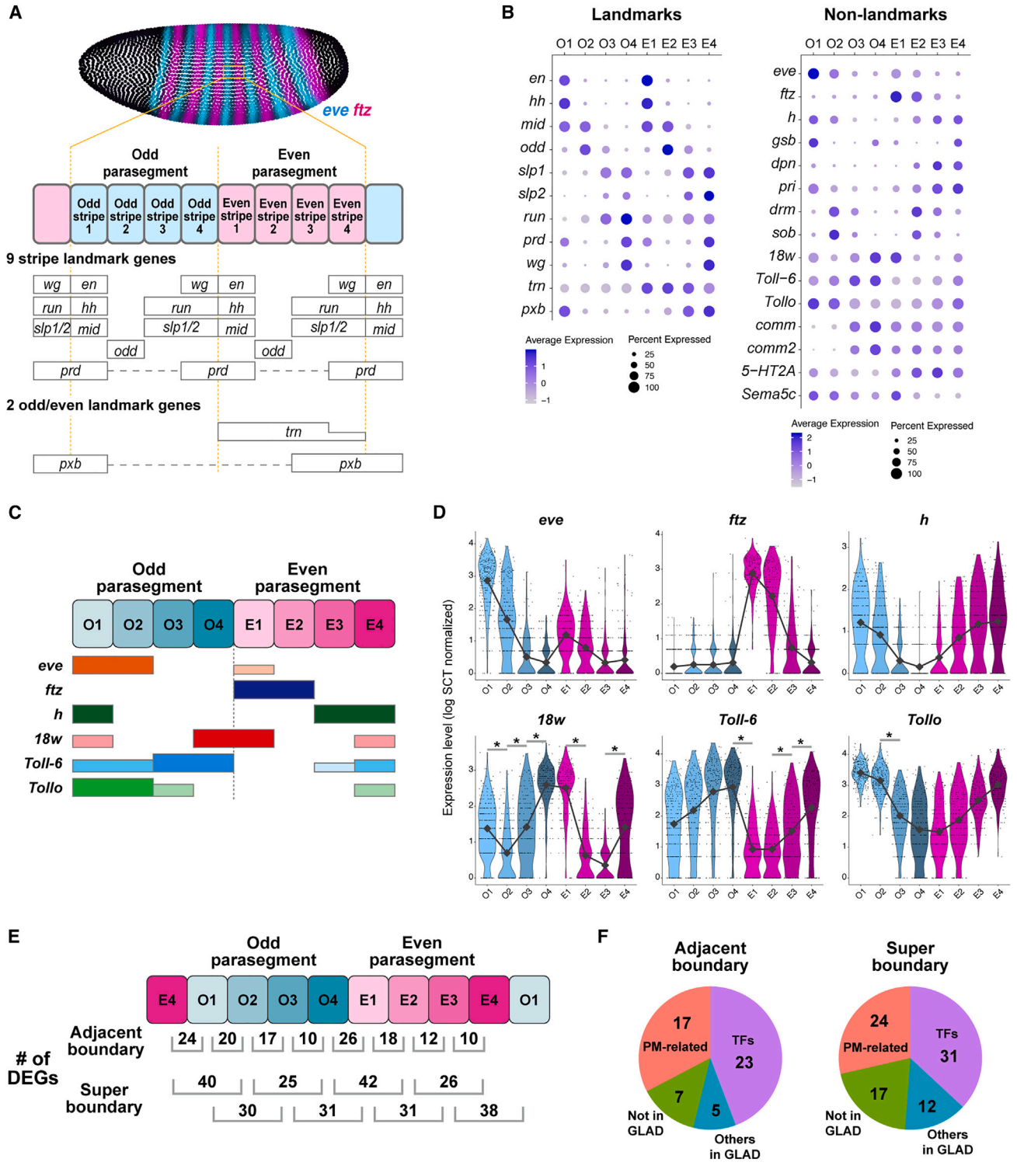


Figure 5. Inference of the pair-rule stripe identities to ectodermal cells

(A) (Top) Examples of the stripe expression patterns of pair-rule genes (*eve*, *ftz*) from the BDTNP ISH database. *eve* is expressed in odd parasegments, and *ftz* is expressed in even parasegments. (Bottom) Expression patterns of nine stripe landmark genes and two odd/even landmark genes.

(B) Dot plots showing the expression patterns of landmark genes used for stripe assignment and non-landmark genes in each stripe.

(C) Reported stripe patterns of *eve*, *ftz*, *h*, *18w*, *Toll-6*, and *Tollo*.[47,62]

(legend continued on next page)

other hand, those of other effector genes better represented the differentiation status of the three germ layers.

Transcriptome-level differences between the single-cell stripes along the AP axis

In *Drosophila* gastrulae, each parasegment of the lateral ectoderm comprises four stripes along the AP axis, and each stripe is a single-cell-wide column (Figure 5A). Each stripe has different identities with combinatorial sets of pair-rule genes directing polarized myosin localization, cell intercalation movement, and germband extension.[58–60] Because pair-rule genes encode TFs, they should control downstream effector genes to drive cell intercalation movement, and some effector genes, such as *18 wheeler* (*18w*, also known as *Toll-2*), *Toll-6*, *Tollo* (also known as *Toll-8*), and *trn* were identified.[61,62] It has been proposed that, from detailed live imaging analyses, the difference between the third and fourth stripes in each parasegment might be difficult to distinguish, suggesting that the strength of cell-cell interaction between them is weak and the difference in gene expression profiles may also be smaller.[63] Tetley et al. also proposed that "super-boundaries" that interface between cells of non-adjacent stripes ("skipped" boundary) are more contractile, implying that these boundaries have more significant differences in receptor expression patterns and stronger cell-cell interaction than boundaries of adjacent identities.[63]

To clarify the whole picture of the genetic basis of germband extension, we need to describe quantitatively how the gene expression profiles differ among each stripe. To do this, by using pair-rule genes and segment polarity genes as stripe landmarks, we categorized the trunk ectodermal cells into eight single-cell stripes that span an odd and even parasegmental unit (Figures 5A, 5C, S5A, and S5B) (see STAR Methods for details). Furthermore, the inference of the stripe pattern showed that, in addition to the landmark genes used for inference, other stripe genes, such as *18w*, *Toll-6*, and *Tollo*, showed gradual changes (Figure 5B). These patterns correlated well with the reported expression patterns.[62] These results indicated that this accurately reconstructed the stripe pattern at the single-cell column level.

This transcriptome information of eight single-cell stripes provides opportunities to quantitatively compare the differences in gene expression profiles between them. First, we conducted the differentially expressed gene (DEG) analyses between all pairs of adjacent identities. Based on an expression difference of $\geq$1.75 fold and a family-wise error rate (FWER) of $\leq$0.01, 10 to 26 genes were identified as DEGs between adjacent pairs (Figure 5E). Similar to the DEG composition of whole scRNA-seq data, most DEGs between adjacent stripes were TF or PM genes, and there was little contribution from other cytoplasmic genes (Figure 5F). As proposed by the "Toll receptor code," all boundaries within parasegment showed at least one Toll receptor gene (*18w*, *Toll-6*, and *Tollo*) as DEGs (Figure 5D). In addition to these known PM genes, our scRNA-seq data revealed that transmembrane genes, such as *commissureless* (*comm*), *comm2*, and *Semaphorin 5c* (*Sema5c*), were quantitatively differentially expressed in a stripe manner (Figure S5C), suggesting that these genes also play a role in cell-cell recognition for cell intercalation. In terms of the number of DEGs, the difference between parasegments (Odd4 vs. Even1 and Even4 vs. Odd1) was larger than that between cell stripes within parasegments, and the difference between the third and fourth stripes (Odd3 vs. Odd4 and Even3 vs. Even4) was the lowest (Figure 5E). In addition, by comparing the differences between super-boundaries, more DEGs were identified than the differences between adjacent pairs (Figure 5E). These results are consistent with the proposed models of super-boundaries and smaller differences between the third and fourth stripes.[63] This dataset will be important for a quantitative understanding of the sufficient genetic mechanisms for germband extension.

scRNA-seq analysis of the *bicoid* mutant

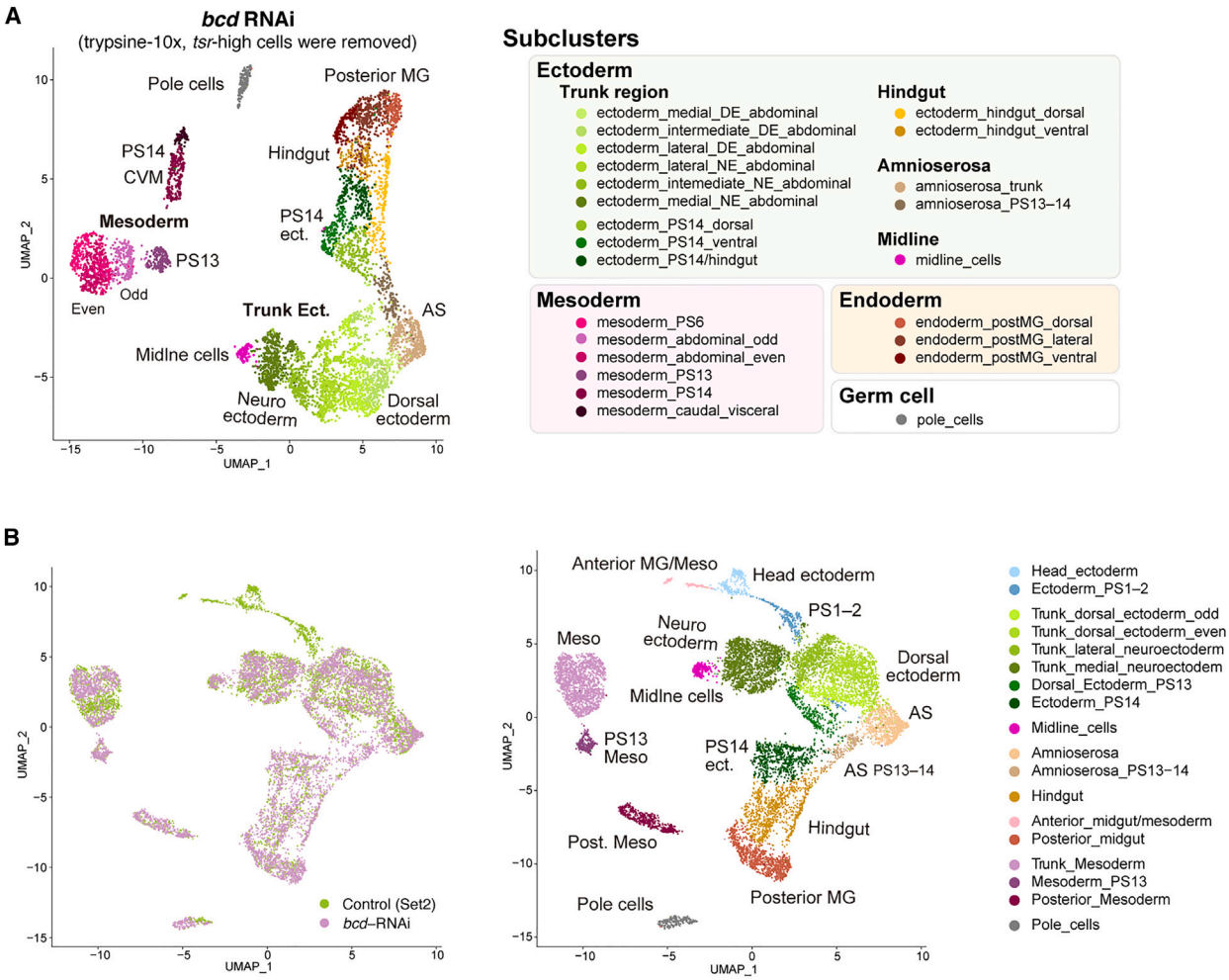
During development, perturbations of axis formation and positional information compromise the process of cell fate determination. However, these cell fate changes have often been assessed by the expression of limited marker genes, most of which encode TFs, and it is not clear whether the cells transformed at the level of the whole transcriptome. To address this issue, we performed scRNA-seq analysis of the *bicoid*-depleted embryos. The AP axis of *Drosophila* is determined by the morphogen gradients of the anterior Bicoid (Bcd) and the posterior Nanos (Nos). The loss of Bcd function results in converting the anterior identity into the posterior one.[2,64] The anterior part of *bcd* mutants eventually shows posterior profiles. However, the developmental history of cells in the anterior region to reach the state differs from that in the original posterior region. For example, the onset of the anterior *hunchback* (*hb*) expression in *bcd* mutants is delayed compared with that in the posterior.[64] Since these historical differences may affect the final state,[1] there may still be transformed cells with a mixed state of both anterior and posterior identities or a state not present in the control gastrula at the transcriptome level.

By acquiring and subclustering *bcd*-RNAi scRNA-seq data, cell types belonging to the anterior region of wild-type embryos, such as the anterior midgut, head/PS1-2 ectoderm, and anterior mesoderm, were not identified (Figure 6A). In addition, consistent with previous reports,[65,66] no clear expression of anterior genes, such as *oc* and *Deformed* (*Dfd*), was detected in these *bcd*-RNAi data (Figure S6A). Since *bcd*-RNAi data were obtained by the trypsin-10x protocol, we compared this with Set2 trypsin-10x data. The *bcd*-RNAi data were merged with the control Set2 data without batch correction tools. The anterior clusters consisted of control cells only (Figure 6B, Head_ectoderm, Ectoderm_PS1-2, anterior midgut/mesoderm), while other clusters were mixed well with cells from both data. The ratio of cells

(D) Violin plots showing the expression patterns of *eve*, *ftz*, *h*, *18w*, *Toll-6*, and *Tollo* in each reconstructed pair-rule stripe. The gray line indicates the median values for each stripe. Expression levels represent the log-transformed values after SCTransform normalization. Asterisks in the bottom panels indicate significant differences in expression ($|FC| \geq 1.75$ and FWER < 0.01).

(E) The number of DEGs between adjacent or super-boundaries.

(F) GLAD category breakdowns for unique DEGs of adjacent or super-boundaries.



(legend on next page)

assigned to posterior clusters in *bcd*-RNAi embryos was significantly larger than that in wild-type embryos and almost double (Figure 6C; Table S6). These results support the complete transformation of the anterior region of *bcd*-RNAi embryos into the posterior identity.

To further investigate *bcd*-RNAi embryos, we did DEG analyses between Set2 and *bcd*-RNAi data for each subcluster. As a result, 161 genes upregulated in *bcd*-RNAi (Figure S6B, left) and 113 genes downregulated in *bcd*-RNAi (Figure S6B, right) were detected as DEGs in at least one subcluster. Among them, 94 upregulated and 61 downregulated genes were detected as DEGs in two or more subclusters, and 68 upregulated and 37 downregulated genes were detected in both the trunk subclusters and the posterior subclusters, which are supposed to contain transformed cells in *bcd*-RNAi embryos. Since most DEGs show no region specificity, they may reflect the difference in genetic background between control and *bcd*-RNAi embryos.

In addition, to clarify whether there is a trace of anterior identity in *bcd*-RNAi data, we searched for genes common to both this upregulated DEG list and the list of genes detected as markers only in the anterior cluster in both Set2 and Set3 (see STAR Methods for details). Using this criterion, we identified a single gene, *Distal-less* (*Dll*). In control embryos, *Dll* was expressed in the dorsal region of the PS1-2 ectoderm and amnioserosa that were absent in *bcd*-RNAi embryos (Figure S6C, left and middle). On the other hand, in *bcd*-RNAi data, *Dll* expression was detected in the trunk amnioserosa cluster (Figure S6C, right). Since *Dll* is expressed in anterior amnioserosa cells but not in trunk amnioserosa in control embryos, *Dll*-positive amnioserosa cells in *bcd*-RNAi might be transformed from PS1-2 amnioserosa to trunk amnioserosa and our data capture a subtle residual feature of the transformed cells. However, this residual *Dll* expression was specific to anterior amnioserosa. It was not detected in other dorsal ectodermal regions, supporting the idea that anterior regions of *bcd*-RNAi embryos are almost completely canalized into the posterior identity at the transcriptome level.

Spatial reconstruction of all gene expression patterns at single-cell resolution

We reconstructed the spatial expression pattern using our Set3 data (6,118 cells) and Perler,[25] and then compared the results with those obtained using the NK-data (1,297 cells). First, by leave-one-gene-out cross-validation (LOOCV), Set-3-based reconstruction showed a higher prediction score (median correlation coefficient = 0.66) than that of NK-data-based reconstruction (median correlation coefficient = 0.61) (Figure 7A). Second, the gene-gene correlations in scRNA-seq data were better conserved in Set-3-based reconstruction than in NK-data-based reconstruction (Figure 7B). Finally, Set-3-based reconstruction maintained the scale of expression values, but NK-data-based reconstruction did not (Figures 7C–7E). For example, Set-3-based reconstruction showed low background signals, and no re-scaling was needed as in NK-data reconstruction. We also found that the reconstructed pattern of some genes was qualitatively improved using our Set3 data. For example, the pattern of segment polarity genes (*wg* and *en*) became much more evident in Set3 reconstruction, and the background between the stripes was almost zero in the Set-3-based reconstruction (Figure 7F). In addition, ISH for *C15* showed expression only in the dorsal amnioserosa, and ISH for *egr* showed a broader expression along the dorsal midline. The reconstructed pattern of *C15* with NK-data and Perler showed a broad expression like *egr*, while Set3 and Perler reconstructed a pattern similar to ISH (Figure 7G). These results indicate that the reconstruction based on Set3 data is more accurate and provides better interpretability for applications in future biological studies.

Recently, in addition to Perler, other computational methods for spatial reconstruction have been proposed. One of them is NovoSpaRc, which adopts a different strategy from Perler and is based on the hypothesis that physically neighboring cells share similar transcriptional profiles and the framework of optimal transport.[26,67] We attempted spatial reconstruction using NovoSpaRc and our Set3 data. Overall, NovoSpaRc showed performance comparable with Perler. First, the spatial reconstruction by Perler and NovoSpaRc showed a high correlation (Figure S7A). Second, the prediction performances of spatial reconstruction by LOOCV were also comparable (Figure S7B). Finally, we also examined the degree to which the spatial reconstruction by Perler and NovoSpaRc conserved the gene-gene correlation in the original scRNA-seq. Perler maintained slightly higher gene-gene correlations than NovoSpaRc (Figure S7C).

On the other hand, from a qualitative point of view, NovoSpaRc showed more spatially uniform patterns than Perler. Although, for example, both methods well reconstructed the ventral expression of mesodermal gene *twi*, the expression within ventral mesodermal in Perler looked more spatially variable than that in NovoSpaRc (Figure S7D). This difference could be because NovoSpaRc takes physical distances between cells into account.

DISCUSSION

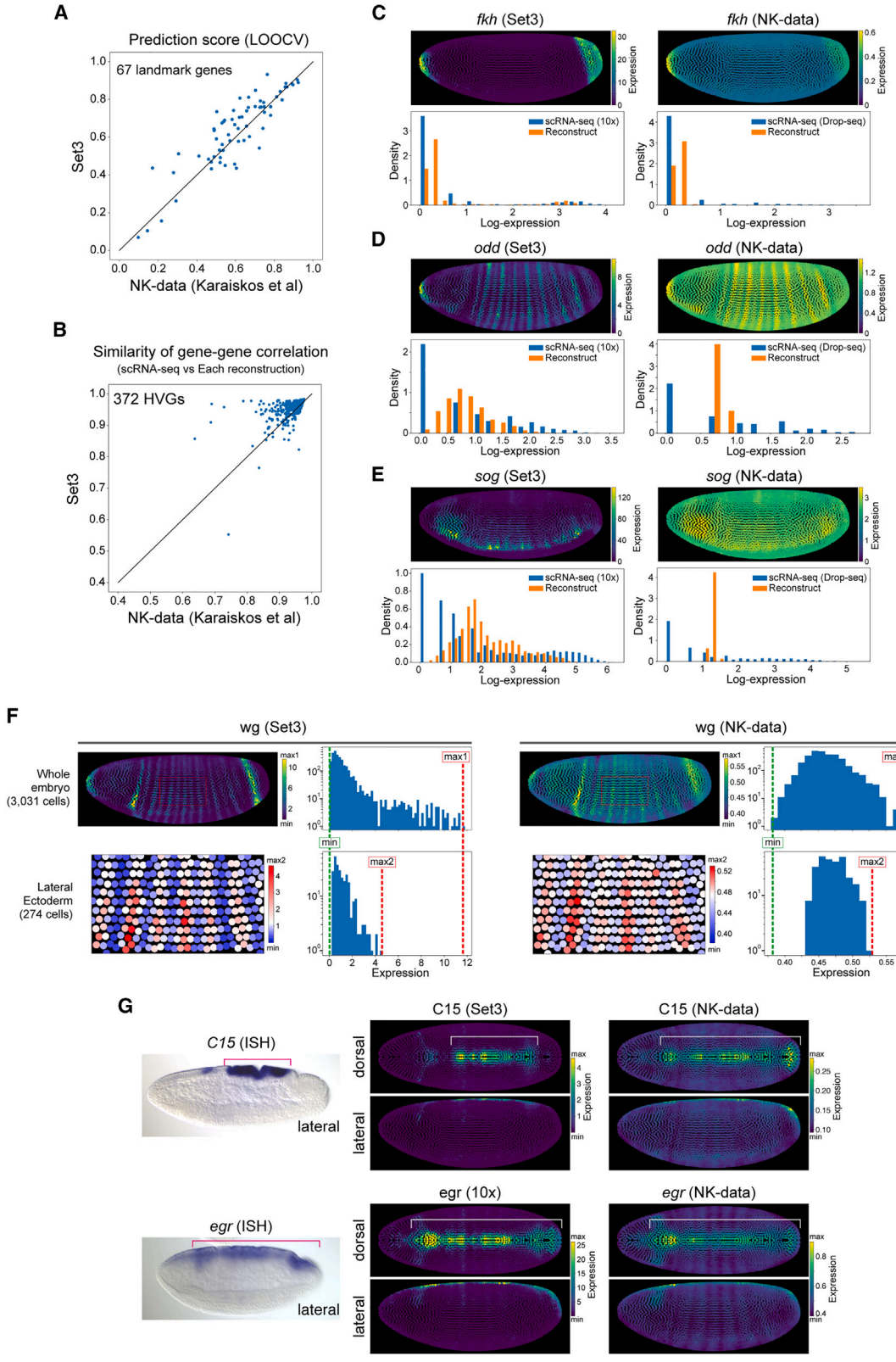
We conducted scRNA-seq analysis to establish the single-cell transcriptome atlas of *Drosophila* gastrulae with higher accuracy and spatial resolution. These data consist of 6,118 cells covering the entire gastrula and allowed us to identify 77 subclusters. We also recapitulated the stripe expression patterns along the AP axis with single-cell-wide column resolution. We found that, at the transcriptome level, rather than the primary TF layer in the regulatory network, the subsequent layer of PM genes or other

Figure 6. Analysis of fate transformation in *bcd* knockdown mutants with scRNA-seq

(A) UMAP plot of the *bcd*-RNAi scRNA-seq data with information on the 24 subclusters. *tsr*-high cells were removed.

(B) UMAP plot of the data merging Set2 and *bcd*-RNAi data without batch correction methods. Cells are colored according to the original dataset (left) and Seurat cluster information (right).

(C) Ratio of the total number of cells to the number of cells in each Seurat cluster for each dataset. The results of the Fisher's exact test are listed in Table S6.



(legend on next page)

cytoplasmic genes showed mRNA expression profiles that better represented the features of the three germ layers. A spatially reconstructed dataset is also established.

Artificial effect of trypsin treatment during cell dissociation

Single-cell dissociation is one of the critical steps for scRNA-seq analysis, and minimizing the artificial effect of dissociation on gene expression is of critical importance. Here, we compared two proteases, trypsin and CAP, and found that only trypsin treatment at 25°C increased the expression of the *E(spl)* complex genes, known targets of Notch signaling, regardless of cell type. This result suggests that trypsin treatment of *Drosophila* gastrula cells induces Notch signal activation. Although the detailed mechanism of Notch activation by trypsin is unclear, these results indicate that cell dissociation methods need to be carefully considered not only for mammalian tissues, as previously reported,[29–32,68] but also for insect tissues.

Several methods other than CAP have also been proposed to minimize the artificial effect of enzymatic treatment on the transcriptome. One is to add transcriptional inhibitors such as actinomycin D.[69] In this study, we attempted the CAP method first rather than transcriptional inhibitors because these inhibitors may not block the degradation of mRNAs. On the other hand, cell dissociation at low temperatures using CAP is expected to minimize transcription and degradation. Although we did not test the use of these inhibitors, they could be useful if target tissues cannot be dissociated using CAP at low temperatures. The other method is single-nucleus RNA-seq. The advantages of single-nucleus RNA-seq are that (1) isolation of nuclei is easier than dissociation of cells from complex tissues, and (2) tissues can be flash-frozen to suppress the gene expression changes. However, the number of transcripts and genes detected per cell tends to be lower than with scRNA-seq.[70,71] In this study, we used scRNA-seq to obtain deeper transcriptome data because of the simple tissue structure of *Drosophila* gastrula. However, since scRNA-seq was applicable only to embryos after cellularization, snRNA-seq should be used to obtain single-cell transcriptome profiles of pre-cellularization embryos, as recently published.[72]

Transient intermediate state during cell differentiation

Detailed subclustering revealed potential intermediate-state cells in *Drosophila* gastrulae, belonging to ectoderm PS14/hindgut, that are likely to be intermediate between PS14 ectoderm and hindgut. Intermediate/hybrid (or multilineage priming) states have also been identified in the embryogenesis of other organisms by scRNA-seq analysis,[73–75] suggesting that the transient intermediate state is a common step during cell differentiation. It is thought that such intermediate states do not persist, and the cells should eventually differentiate into one of the two states, but how the direction of differentiation is determined is still poorly understood. Cell differentiation proceeds in parallel with morphogenesis during development, and we recently found that morphogenesis can modulate cell differentiation.[76] Since the intermediate ectoderm PS14/hindgut region undergoes dynamic changes in tissue shape (hindgut/endoderm invaginations), the cell differentiation paths may depend on the completion of invagination of the hindgut at this time. In the future, it is essential to investigate the lineages of these cells in detail using time-course analyses of cell differentiation and morphogenesis.

PM DEGs

Here, we found that PM genes show higher variability between cells of fly gastrula rather than cytoplasmic genes, and highly variable PM genes are sufficient to classify cell types established in the gastrula. Furthermore, DEGs between adjacent pair-rule stripes were also mainly composed of PM genes and pair-rule TF genes. These results suggest that PM genes, rather than other cytoplasmic genes, contribute more strongly to cellular and tissue-level regulation during gastrulation. This result is consistent with previous reports showing that cellular signaling mediated by transmembrane proteins triggers cellular and tissue behaviors.[61,62,77] Since variable genes other than TF and PM genes also classified cells into three germ layers to some degree (Figure S4), the combinatorial profiles of these cytoplasmic or unclassified genes may be able to define the basic properties of cells and tissues, or the range of behavioral capabilities. Variable expressions of PM genes might then orchestrate the local cell-cell interaction, followed by driving cell and tissue morphogenesis in combination with their capacities.

Figure 7. Spatial reconstruction by Perler

(A) Comparison between the leave-one-gene-out cross-validation (LOOCV) results of Perler reconstructions based on NK-data and Set3 data. Each dot indicates each landmark gene. The x and y axis show the correlations between reference ISH expression patterns and the reconstructed expression pattern of each gene based on NK-data and Set3 data, respectively.

(B) Comparison of gene-gene correlation structure conservation between Set3-based and NK-data-based reconstruction by Perler. Each dot indicates each gene that was commonly included in top 500 HVGs of both datasets (372 genes). The x axis shows gene-gene correlation structure conservation in NK-data-based reconstruction, and the y axis shows gene-gene correlation structure conservation of Set3-based reconstruction. The definition of gene-gene correlation structure conservation is described in STAR Methods.

(C–E) Examples of reconstructed expression by Perler on Set3 (left) and NK-data (right). In each plot, upper panels show reconstructed expression patterns. Color maps are linear and zero-max scaled. Bottom panels show density histograms of the gene expression in the original scRNA-seq and the reconstruction. Expression patterns are log scaled, and each bin size is 0.2 in density histograms.

(F) The reconstructed expression patterns of *wg* by Perler based on Set3 (left) and NK-data (right). In each plot, upper-left panel shows reconstructed expression patterns in whole embryo. Bottom-left panel shows the enlarged views of the region enclosed with a red rectangle in the upper-left panel. Upper-right and bottom-right panels show the histograms of the expression in whole embryo and the region shown in bottom-left panel, respectively. Expression values in plots are linear scaled, and y axes in the histogram are log scaled. The maximum expression patterns in the whole-embryo plots and enlarged plots are indicated by "max1" and "max2" (red dashed lines), respectively. The minimum expression in the whole embryo is indicated by "min" (green dashed line).

(G) (left) ISHs of *C15* and *egr* from the Berkeley Drosophila Genome Project (BDGP; https://insitu.fruitfly.org/).[38–40] Lateral view (middle). The lateral or dorsal views of Set3-based reconstructed patterns of *C15* and *egr* by Perler. (Left) The lateral or dorsal views of NK-data-based reconstructed patterns of *C15* and *egr* by Perler. Red and white lines show the region of expression along the dorsal midline.

Transmembrane DEGs between stripes include not only three Toll receptor genes (*18w*, *Toll-6*, and *Tollo*), *trn*, and *5-HT2A*, which are known to play a role in the regulation of germband extension,[61,62,78,79] but also factors that have not been recognized as regulators of germband extension but are known to be involved in axon guidance, such as *comm*, *comm2*,[80,81] and *Sema5c*.[82] There is growing evidence that many transmembrane proteins identified as factors for neural-network formation are also involved in epithelial morphogenesis and homeostasis[83–86] Furthermore, *Sema5c* was recently reported to be involved in the morphogenesis of follicle epithelia.[87] In this analysis, only a few genes were identified as pair-rule DEGs. Although it remains possible that some essential genes are being missed because the threshold is too high, the limited transmembrane DEGs are expected to be sufficient to organize the dynamics of epithelial morphogenesis in a redundant or cooperative manner.

Non-linear conversion from spatial information to cell-type-specific transcriptome

The mechanisms of cell fate specification and differentiation have been studied extensively in genome-wide research using transcriptomic and epigenetic analyses.[8,9] However, it is still a fundamental biological question how a limited number of TFs and signaling generate various cell types during development. Also, local intercellular communication can affect TF activity in a post-transcriptional manner. To solve this issue, it is necessary to comprehensively clarify the relationship between TFs and downstream gene expression at the single-cell level with spatial information. At least in the case of *Drosophila* gastrula, we revealed that, rather than the expression profiles of TF genes, those of PM genes or other effector genes better represented the cell differentiation status corresponding to the three germ layers (Figures 4 and S4). In other words, initial positional information remains to some extent when viewed from the perspective of only mRNA expression profiles of all TFs. These analyses using our scRNA-seq data support the idea of a non-linear combinatorial scheme of transcriptome establishment by TFs.

It has been proposed that sequential logic can overcome the bottleneck of combinatorial logic.[88] In this theoretical view, there is a limit to the transcriptome pattern that can be established from only the combination state at the time. However, if we take the sequential logic wherein the time ordering of factors informs the outcome, the diversity of target configurations dramatically increases even with the same regulatory network. Although the anterior part of *bcd* mutants eventually got the posterior combination of gap genes, the temporal histories of gap gene expression (e.g., *hb*) and pair-rule gene expression (e.g., *eve* and *ftz*) are slightly different from those of the original posterior region.[64] However, our scRNA-seq analysis of *bcd*-depleted embryos revealed that the anterior part of them acquired transcriptome characteristics for cells in the posterior region (Figure 7), suggesting that the temporal histories of gap genes and pair-rule genes do not significantly affect the formation of transcriptome, and the status at the last moment just before gastrulation starts (50 min after the onset of nuclear cycle 14) determines the cell fate. Therefore, the sequential scheme has less contribution to cell fate control in the *Drosophila* gastrula, possibly because of the short duration of the process. These results support the proposed possibility that "subsequent layers serve to transform the positional information, fully available already at the gap gene layer, into an explicit commitment to repeated but discrete cell types."[2] Furthermore, even though there are some noise and sharp discontinuities along the AP axis, all cells in *bcd* mutants eventually canalize into cell types that are present in the wild type at the level of the transcriptome, suggesting that the robust gene regulatory mechanism is operating not only with a handful of marker genes but also with a multitude of genes across the whole genome. Similar canalization into the defined transcriptomic state existing in wild types upon perturbations has been reported in scRNA-seq of zebrafish mutants.[73,74]

In this study, using our scRNA-seq data with Perler or NovoSpaRc, we reconstructed the spatial transcriptome of *Drosophila* gastrula at single-cell resolution with high accuracy. Gene regulatory network analysis has made progress in recent years with scRNA-seq and assay for transposase-accessible chromatin with sequencing (ATAC-seq) in many species, including flies. Therefore, this gastrula spatial atlas could be used as an essential reference to reveal the gene regulatory network of early embryos and to bring up the whole picture of cell-cell communication. Future integrated analyses of gene regulatory networks with single-cell epigenetic profiling and spatial signaling activity with spatial-transcriptome data will provide us with more detailed insights into the mechanisms by which the gradual positional information is non-linearly converted into discrete patterns of cell differentiation genome-wide and also enable a deeper understanding of the developmental systems that orchestrate tissue morphogenesis and functions.

Limitations of the study

Although both Perler and NovoSpaRc reconstructions seem highly accurate, there is still a limitation. In both methods, the spatial gene expression along the AP axis appears to be well reconstructed, whereas that along the DV axis is insufficient. For example, in the UMAP plot of the original scRNA-seq data, the expression levels of *vnd* and *ind* were mutually exclusive in the *brk*+ medial neuroectoderm, intermediate neuroectoderm, and midline cells. However, these levels were intermingled in both reconstructed data (Figures S7E–S7H). This output is probably because the reference BDTNP ISH data are not sufficient and accurate because of the limited number of genes analyzed and the incomplete computational integration of the imaging data from multiple embryos. The construction of more precise reference data in the future will enable us to perform a more accurate reconstruction of the spatial transcriptome. Taken together, at present, it would be better to use the results of spatial reconstruction with reference to the original scRNA-seq data for future biological applications.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.112707.

ACKNOWLEDGMENTS

We would like to thank the Kyoto Stock Center, the Bloomington Drosophila Stock Center for fly stocks; Platform for Advanced Genome Science (PAGS) and NGS Core Facility of the Graduate Schools of Biostudies, Kyoto University, for supporting the sequencing analysis; Seishi Ogawa, Masahiro Nakagawa, and Toshiko Sato for supporting the 10x Chromium; Masayo Miki for assistance with the experiments; Tadao Usui and Yu-Chiun Wang for critical comments on the manuscript; and members of the Uemura laboratory, Shigehiro Kuraku, and the Laboratory for Phyloinformatics RIKEN BDR for discussion. This work was supported by JSPS KAKENHI (15K14535, 17KT0021, 22H05167 to T.K.; 20J23385 to S.S.; and 16H06279 to PAGS); JST FOREST Program (JPMJFR204V to T.K.); the Naito Foundation (to T.K.); and the Keihanshin Consortium for Fostering the Next Generation of Global Leaders in Research (K-CONNEX) established by the Building of Consortia for the Development of Human Resources in Science and Technology; and MEXT (to T.K.). S.S. was supported by a JSPS Research Fellowship for Young Scientists.

AUTHOR CONTRIBUTIONS

T.K. conceived the project. S.S. and T.K. performed the experiments and analyzed the data. S.M. performed FISH analysis. S.S., Y.O., and H.N. developed and implemented the computational method. C.T., O.N., and M.K. assisted in establishing the scRNA-seq method. T.U. provided feedback on this study. S.S. and T.K. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 4, 2022
Revised: March 27, 2023
Accepted: June 13, 2023
Published: July 10, 2023

REFERENCES

1. Briscoe, J., and Small, S. (2015). Morphogen rules: Design principles of gradient-mediated embryo patterning. Development *142*, 3996–4009. https://doi.org/10.1242/dev.129452.
2. Petkova, M.D., Tkačik, G., Bialek, W., Wieschaus, E.F., and Gregor, T. (2019). Optimal Decoding of Cellular Identities in a Genetic Network. Cell *176*, 844–855.e15. https://doi.org/10.1016/j.cell.2019.01.007.
3. Collinet, C., and Lecuit, T. (2021). Programmed and self-organized flow of information during morphogenesis. Nat. Rev. Mol. Cell Biol. *22*, 245–265. https://doi.org/10.1038/s41580-020-00318-6.
4. Small, S., and Arnosti, D.N. (2020). Transcriptional Enhancers in Drosophila. Genetics *216*, 1–26. https://doi.org/10.1534/genetics.120.301370.
5. Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E.E.M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature *462*, 65–70. https://doi.org/10.1038/nature08531.
6. Heisenberg, C.-P., and Bellaïche, Y. (2013). Forces in Tissue Morphogenesis and Patterning. Cell *153*, 948–962. https://doi.org/10.1016/j.cell.2013.05.008.
7. Kondo, T., and Hayashi, S. (2015). Mechanisms of cell height changes that mediate epithelial invagination. Dev. Growth Differ. *57*, 313–323. https://doi.org/10.1111/dgd.12224.
8. Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170–1187. https://doi.org/10.1016/j.cell.2016.09.018.
9. Stricker, S.H., Köferle, A., and Beck, S. (2017). From profiles to function in epigenomics. Nat. Rev. Genet. *18*, 51–66. https://doi.org/10.1038/nrg.2016.138.
10. Wieschaus, E., and Nüsslein-Volhard, C. (2016). The Heidelberg Screen for Pattern Mutants of Drosophila: A Personal Account. Annu. Rev. Cell Dev. Biol. *32*, 1–46. https://doi.org/10.1146/annurev-cellbio-113015-023138.
11. Gheisari, E., Aakhte, M., and Müller, H.A.J. (2020). Gastrulation in Drosophila melanogaster: Genetic control, cellular basis and biomechanics. Mech. Dev. *163*, 103629. https://doi.org/10.1016/j.mod.2020.103629.
12. Jaeger, J., Manu, and Reinitz, J. (2012). Drosophila blastoderm patterning. Curr. Opin. Genet. Dev. *22*, 533–541. https://doi.org/10.1016/j.gde.2012.10.005.
13. Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T.R., Tomancak, P., and Krause, H.M. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. Cell *131*, 174–187. https://doi.org/10.1016/j.cell.2007.08.003.
14. Wilk, R., Hu, J., Blotsky, D., and Krause, H.M. (2016). Diverse and pervasive subcellular distributions for both coding and long noncoding RNAs. Genes Dev. *30*, 594–609. https://doi.org/10.1101/gad.276931.115.
15. Fowlkes, C.C., Hendriks, C.L.L., Keränen, S.V.E., Weber, G.H., Rübel, O., Huang, M.Y., Chatoor, S., DePace, A.H., Simirenko, L., Henriquez, C., et al. (2008). A Quantitative Spatiotemporal Atlas of Gene Expression in the Drosophila Blastoderm. Cell *133*, 364–374. https://doi.org/10.1016/j.cell.2008.01.053.
16. Keränen, S.V.E., Fowlkes, C.C., Luengo Hendriks, C.L., Sudar, D., Knowles, D.W., Malik, J., and Biggin, M.D. (2006). Three-dimensional

morphology and gene expression in the Drosophila blastoderm at cellular resolution II: dynamics. Genome Biol. *7*, R124. https://doi.org/10.1186/gb-2006-7-12-r124.

17. Luengo Hendriks, C.L., Keränen, S.V.E., Fowlkes, C.C., Simirenko, L., Weber, G.H., DePace, A.H., Henriquez, C., Kaszuba, D.W., Hamann, B., Eisen, M.B., et al. (2006). Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution I: data acquisition pipeline. Genome Biol. *7*, R123. https://doi.org/10.1186/gb-2006-7-12-r123.
18. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. Mol. Cell. *58*, 610–620. https://doi.org/10.1016/j.molcel.2015.04.005.
19. Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. Nature *541*, 331–338. https://doi.org/10.1038/nature21350.
20. Achim, K., Pettit, J.-B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D., and Marioni, J.C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat. Biotechnol. *33*, 503–509. https://doi.org/10.1038/nbt.3209.
21. Satija, R., Farrell, J.a., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502. https://doi.org/10.1038/nbt.3192.
22. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.
23. Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell *177*, 1873–1887.e17. https://doi.org/10.1016/j.cell.2019.05.006.
24. Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N., and Zinzen, R.P. (2017). The Drosophila embryo at single-cell transcriptome resolution. Science *358*, 194–199. https://doi.org/10.1126/science.aan3235.
25. Okochi, Y., Sakaguchi, S., Nakae, K., Kondo, T., and Naoki, H. (2021). Model-based prediction of spatial gene expression via generative linear mapping. Nat. Commun. *12*, 3731. https://doi.org/10.1038/s41467-021-24014-x.
26. Nitzan, M., Karaiskos, N., Friedman, N., and Rajewsky, N. (2019). Gene expression cartography. Nature *576*, 132–137. https://doi.org/10.1038/s41586-019-1773-3.
27. Cang, Z., and Nie, Q. (2020). Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nat. Commun. *11*, 2084. https://doi.org/10.1038/s41467-020-15968-5.
28. Tanevski, J., Nguyen, T., Truong, B., Karaiskos, N., Ahsen, M.E., Zhang, X., Shu, C., Xu, K., Liang, X., Hu, Y., et al. (2020). Gene selection for optimal prediction of cell position in tissues from single-cell transcriptomics data. Life Sci. Alliance *3*, e202000867. https://doi.org/10.26508/LSA.202000867.
29. Adam, M., Potter, A.S., and Potter, S.S. (2017). Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: A molecular atlas of kidney development. Development *144*, 3625–3632. https://doi.org/10.1242/dev.151142.
30. Denisenko, E., Guo, B.B., Jones, M., Hou, R., de Kock, L., Lassmann, T., Poppe, D., Clément, O., Simmons, R.K., Lister, R., and Forrest, A.R.R. (2020). Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. *21*, 130. https://doi.org/10.1186/s13059-020-02048-6.
31. Miyawaki-Kuwakado, A., Wu, Q., Harada, A., Tomimatsu, K., Fujii, T., Maehara, K., and Ohkawa, Y. (2021). Transcriptome analysis of gene expression changes upon enzymatic dissociation in skeletal myoblasts. Gene Cell. *26*, 530–540. https://doi.org/10.1111/GTC.12870.
32. O'Flanagan, C.H., Campbell, K.R., Zhang, A.W., Kabeer, F., Lim, J.L.P., Biele, J., Eirew, P., Lai, D., McPherson, A., Kong, E., et al. (2019). Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. Genome Biol. *20*, 210. https://doi.org/10.1186/s13059-019-1830-0.
33. Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166. https://doi.org/10.1038/nmeth.2772.
34. Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., Kim, J., Murray, J., and Raj, A. (2018). Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH. Cell Syst. *6*, 171–179.e5. https://doi.org/10.1016/j.cels.2018.01.014.
35. Cowden, J., and Levine, M. (2002). The Snail repressor positions Notch signaling in the Drosophila embryo. Development *129*, 1785–1793.
36. Morel, V., and Schweisguth, F. (2000). Repression by Suppressor of Hairless and activation by Notch are required to define a single row of single-minded expressing cells in the Drosophila embryo. Genes Dev. *14*, 377–388. https://doi.org/10.1101/gad.14.3.377.
37. Zinzen, R.P., Cande, J., Ronshaugen, M., Papatsenko, D., and Levine, M. (2006). Evolution of the Ventral Midline in Insect Embryos. Dev. Cell *11*, 895–902. https://doi.org/10.1016/j.devcel.2006.10.012.
38. Hammonds, A.S., Bristow, C.A., Fisher, W.W., Weiszmann, R., Wu, S., Hartenstein, V., Kellis, M., Yu, B., Frise, E., and Celniker, S.E. (2013). Spatial expression of transcription factors in Drosophila embryonic organ development. Genome Biol *14*, R140. https://doi.org/10.1186/gb-2013-14-12-r140.
39. Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S.E., et al. (2002). Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol *3*, RESEARCH0088. https://doi.org/10.1186/gb-2002-3-12-research0088.
40. Tomancak, P., Berman, B.P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., Celniker, S.E., and Rubin, G.M. (2007). Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol *8*, R145. https://doi.org/10.1186/gb-2007-8-7-r145.
41. Reeves, G.T., and Stathopoulos, A. (2009). Graded Dorsal and Differential Gene Regulation in the Drosophila Embryo. Cold Spring Harbor Perspect. Biol. *1*, a000836. https://doi.org/10.1101/cshperspect.a000836.
42. Akam, M. (1987). The molecular basis for metameric pattern in the Drosophila embryo. Development *101*, 1–22.
43. Ingham, P.W. (1988). The molecular genetics of embryonic pattern formation in Drosophila. Nature *335*, 25–34.
44. Jaeger, J. (2011). The gap gene network. Cell. Mol. Life Sci. *68*, 243–274. https://doi.org/10.1007/s00018-010-0536-y.
45. McGinnis, W., and Krumlauf, R. (1992). Homeobox genes and axial patterning. Cell *68*, 283–302. https://doi.org/10.1016/0092-8674(92)90471-N.
46. Lawrence, P.A. (1992). The Making of a Fly: The Genetics of Animal Design (Wiley-Blackwell).
47. Clark, E., and Akam, M. (2016). Odd-paired controls frequency doubling in Drosophila segmentation by altering the pair-rule gene regulatory network. Elife *5*, e18215. https://doi.org/10.7554/eLife.18215.
48. Graham, P.L., Anderson, W.R., Brandt, E.A., Xiang, J., and Pick, L. (2019). Dynamic expression of Drosophila segmental cell surface-encoding genes and their pair-rule regulators. Dev. Biol. *447*, 147–156. https://doi.org/10.1016/j.ydbio.2019.01.015.
49. Heimberg, G., Bhatnagar, R., El-samad, H., Thomson, M., Heimberg, G., Bhatnagar, R., El-samad, H., and Thomson, M. (2016). Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. Cell Syst. *2*, 239–250. https://doi.org/10.1016/j.cels.2016.04.001.

50. Zhang, M.J., Ntranos, V., and Tse, D. (2020). Determining sequencing depth in a single-cell RNA-seq experiment. Nat. Commun. *11*, 774. https://doi.org/10.1038/s41467-020-14482-y.

51. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.

52. Thisse, B., Stoetzel, C., Gorostiza-Thisse, C., and Perrin-Schmitt, F. (1988). Sequence of the twist gene and nuclear localization of its protein in endomesodermal cells of early Drosophila embryos. EMBO J. *7*, 2175–2183. https://doi.org/10.1002/j.1460-2075.1988.tb03056.x.

53. Alberga, A., Boulay, J.L., Kempe, E., Dennefeld, C., and Haenlin, M. (1991). The snail gene required for mesoderm formation in Drosophila is expressed dynamically in derivatives of all three germ layers. Development *111*, 983–992. https://doi.org/10.1242/dev.111.4.983.

54. Reuter, R., Grunewald, B., and Leptin, M. (1993). A role for the mesoderm in endodermal migration and morphogenesis in Drosophila. Development *119*, 1135–1145. https://doi.org/10.1242/DEV.119.4.1135.

55. Ganguly, A., Jiang, J., and Ip, Y.T. (2005). Drosophila WntD is a target and an inhibitor of the Dorsal/Twist/Snail network in the gastrulating embryo. Development *132*, 3419–3429. https://doi.org/10.1242/dev.01903.

56. Rahimi, N., Averbukh, I., Haskel-Ittah, M., Degani, N., Schejter, E.D., Barkai, N., and Shilo, B.-Z. (2016). A WntD-Dependent Integral Feedback Loop Attenuates Variability in Drosophila Toll Signaling. Dev. Cell *36*, 401–414. https://doi.org/10.1016/j.devcel.2016.01.023.

57. Hu, Y., Comjean, A., Perkins, L.A., Perrimon, N., and Mohr, S.E. (2015). GLAD: an Online Database of Gene List Annotation for Drosophila. J. Genomics *3*, 75–81. https://doi.org/10.7150/jgen.12863.

58. Bertet, C., Sulak, L., and Lecuit, T. (2004). Myosin-dependent junction remodelling controls planar cell intercalation and axis elongation. Nature *429*, 667–671. https://doi.org/10.1038/nature02581.1.

59. Zallen, J.a., and Wieschaus, E. (2004). Patterned gene expression directs bipolar planar polarity in Drosophila. Dev. Cell *6*, 343–355.

60. Zallen, J.a., and Blankenship, J.T. (2008). Multicellular dynamics during epithelial elongation. Semin. Cell Dev. Biol. *19*, 263–270. https://doi.org/10.1016/j.semcdb.2008.01.005.

61. Paré, A.C., Vichas, A., Fincher, C.T., Mirman, Z., Farrell, D.L., Mainieri, A., and Zallen, J.A. (2014). A positional Toll receptor code directs convergent extension in Drosophila. Nature *515*, 523–527. https://doi.org/10.1038/nature13953.

62. Paré, A.C., Naik, P., Shi, J., Mirman, Z., Palmquist, K.H., and Zallen, J.A. (2019). An LRR Receptor-Teneurin System Directs Planar Polarity at Compartment Boundaries. Dev. Cell *51*, 208–221.e6. https://doi.org/10.1016/j.devcel.2019.08.003.

63. Tetley, R.J., Blanchard, G.B., Fletcher, A.G., Adams, R.J., and Sanson, B. (2016). Unipolar distributions of junctional Myosin II identify cell stripe boundaries that drive cell intercalation throughout Drosophila axis extension. Elife *5*, e12094. https://doi.org/10.7554/eLife.12094.

64. Staller, M.V., Fowlkes, C.C., Bragdon, M.D.J., Wunderlich, Z., Estrada, J., and DePace, A.H. (2015). A gene expression atlas of a bicoid-depleted Drosophila embryo reveals early canalization of cell fate. Development *142*, 587–596. https://doi.org/10.1242/dev.117796.

65. Finkelstein, R., and Perrimon, N. (1990). The orthodenticle gene is regulated by bicoid and torso and specifies Drosophila head development. Nature *346*, 485–488. https://doi.org/10.1038/346485a0.

66. Jack, T., and McGinnis, W. (1990). Establishment of the Deformed expression stripe requires the combinatorial action of coordinate, gap and pair-rule proteins. EMBO J. *9*, 1187–1198. https://doi.org/10.1002/j.1460-2075.1990.tb08226.x.

67. Moriel, N., Senel, E., Friedman, N., Rajewsky, N., Karaiskos, N., and Nitzan, M. (2021). NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport. Nat. Protoc. *16*, 4177–4200. https://doi.org/10.1038/s41596-021-00573-7.

68. Van Den Brink, S.C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C.S., Robin, C., and Van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat. Methods *14*, 935–936. https://doi.org/10.1038/nmeth.4437.

69. Wu, Y.E., Pan, L., Zuo, Y., Li, X., and Hong, W. (2017). Detecting Activated Cell Populations Using Single-Cell RNA-Seq. Neuron *96*, 313–329.e6. https://doi.org/10.1016/j.neuron.2017.09.026.

70. Bakken, T.E., Hodge, R.D., Miller, J.A., Yao, Z., Nguyen, T.N., Aevermann, B., Barkan, E., Bertagnolli, D., Casper, T., Dee, N., et al. (2018). Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. PLoS One *13*, e0209648. https://doi.org/10.1371/journal.pone.0209648.

71. Basile, G., Kahraman, S., Dirice, E., Pan, H., Dreyfuss, J.M., and Kulkarni, R.N. (2021). Using single-nucleus RNA-sequencing to interrogate transcriptomic profiles of archived human pancreatic islets. Genome Med. *13*, 128. https://doi.org/10.1186/s13073-021-00941-8.

72. Albright, A.R., Stadler, M.R., and Eisen, M.B. (2022). Single-nucleus RNA-sequencing in pre-cellularization Drosophila melanogaster embryos. PLoS One *17*, e0270471. https://doi.org/10.1371/journal.pone.0270471.

73. Briggs, J.A., Weinreb, C., Wagner, D.E., Megason, S., Peshkin, L., Kirschner, M.W., and Klein, A.M. (2018). The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. Science *360*, eaar5780. https://doi.org/10.1126/science.aar5780.

74. Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A., and Schier, A.F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science *360*, eaar3131. https://doi.org/10.1126/science.aar3131.

75. Packer, J.S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., et al. (2019). A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. Science *365*, eaax1971. https://doi.org/10.1126/science.aax1971.

76. Kondo, T., and Hayashi, S. (2019). Two-step regulation of trachealess ensures tight coupling of cell fate with morphogenesis in the Drosophila trachea. Elife *8*, e45145. https://doi.org/10.7554/eLife.45145.

77. Manning, a.J., Peters, K.a., Peifer, M., and Rogers, S.L. (2013). Regulation of Epithelial Morphogenesis by the G Protein-Coupled Receptor Mist and Its Ligand Fog. Sci. Signal. *6*, ra98. https://doi.org/10.1126/scisignal.2004427.

78. Colas, J.F., Launay, J.M., Vonesch, J.L., Hickel, P., and Maroteaux, L. (1999). Serotonin synchronises convergent extension of ectoderm with morphogenetic gastrulation movements in Drosophila. Mech. Dev. *87*, 77–91.

79. Schaerlinger, B., Launay, J.M., Vonesch, J.L., and Maroteaux, L. (2007). Gain of affinity point mutation in the serotonin receptor gene 5-HT2Dro accelerates germband extension movements during Drosophila gastrulation. Dev. Dynam. *236*, 991–999. https://doi.org/10.1002/dvdy.21110.

80. Keleman, K., Rajagopalan, S., Cleppien, D., Teis, D., Paiha, K., Huber, L.A., Technau, G.M., and Dickson, B.J. (2002). Comm Sorts Robo to Control Axon Guidance at the Drosophila Midline. Cell *110*, 415–427. https://doi.org/10.1016/S0092-8674(02)00901-7.

81. Keleman, K., Ribeiro, C., and Dickson, B.J. (2005). Comm function in commissural axon guidance: Cell-autonomous sorting of Robo in vivo. Nat. Neurosci. *8*, 156–163. https://doi.org/10.1038/nn1388.

82. Yazdani, U., and Terman, J.R. (2006). The semaphorins. Genome Biol. *7*, 211. https://doi.org/10.1186/gb-2006-7-3-211.

83. Hinck, L. (2004). The versatile roles of "axon guidance" cues in tissue morphogenesis. Dev. Cell *7*, 783–793. https://doi.org/10.1016/j.devcel.2004.11.002.

84. Vaughen, J., and Igaki, T. (2016). Slit-Robo Repulsive Signaling Extrudes Tumorigenic Cells from Epithelia. Dev. Cell *39*, 683–695. https://doi.org/10.1016/j.devcel.2016.11.015.

85. Yoo, S.K., Pascoe, H.G., Pereira, T., Kondo, S., Jacinto, A., Zhang, X., and Hariharan, I.K. (2016). Plexins function in epithelial repair in both Drosophila and zebrafish. Nat. Commun. *7*, 12282. https://doi.org/10.1038/ncomms12282.

86. Cammarota, C., Finegan, T.M., Wilson, T.J., Yang, S., and Bergstralh, D.T. (2020). An Axon-Pathfinding Mechanism Preserves Epithelial Tissue Integrity. Curr. Biol. *30*, 5049–5057.e3. https://doi.org/10.1016/j.cub.2020.09.061.

87. Stedden, C.G., Menegas, W., Zajac, A.L., Williams, A.M., Cheng, S., Özkan, E., and Horne-Badovinac, S. (2019). Planar-Polarized Semaphorin-5c and Plexin A Promote the Collective Migration of Epithelial Cells in Drosophila. Curr. Biol. *29*, 908–920.e6. https://doi.org/10.1016/j.cub.2019.01.049.

88. Letsou, W., and Cai, L. (2016). Noncommutative Biology: Sequential Regulation of Complex Networks. PLoS Comput. Biol. *12*, e1005089. https://doi.org/10.1371/journal.pcbi.1005089.

89. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002.

90. Kaminow, B., Yunusov, D., and Dobin, A. (2021). STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. Preprint at bioRxiv. https://doi.org/10.1101/2021.05.05.442755.

91. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296. https://doi.org/10.1186/s13059-019-1874-1.

92. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. *47*, W191–W198. https://doi.org/10.1093/nar/gkz369.

93. Saxena, A., Wagatsuma, A., Noro, Y., Kuji, T., Asaka-Oba, A., Watahiki, A., Gurnot, C., Fagiolini, M., Hensch, T.K., and Carninci, P. (2012). Trehalose-enhanced isolation of neuronal sub-types from adult mouse brain. Biotechniques *52*, 381–385. https://doi.org/10.2144/0000113878.

94. Kishi, J.Y., Lapan, S.W., Beliveau, B.J., West, E.R., Zhu, A., Sasaki, H.M., Saka, S.K., Wang, Y., Cepko, C.L., and Yin, P. (2019). SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. Nat. Methods *16*, 533–544. https://doi.org/10.1038/s41592-019-0404-0.

95. Beliveau, B.J., Kishi, J.Y., Nir, G., Sasaki, H.M., Saka, S.K., Nguyen, S.C., Wu, C.T., and Yin, P. (2018). OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. Proc. Natl. Acad. Sci. USA *115*, E2183–E2192. https://doi.org/10.1073/pnas.1714530115.

96. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

97. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York). https://doi.org/10.1007/978-0-387-98141-3.

98. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. *16*, 278–313. https://doi.org/10.1186/s13059-015-0844-5.

99. Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.

100. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

101. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. j. *17*, 10–12. https://doi.org/10.14806/EJ.17.1.200.

102. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. *12*, 323. https://doi.org/10.1186/1471-2105-12-323.

103. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. https://doi.org/10.1093/bioinformatics/btp616.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
RNase quiet	Nacalai	Cat#09147-14
Trehalose	Nacalai	Cat#11667-34
trypsine-EDTA	Sigma	Cat#T3924
trypsin inhibitor	Sigma	Cat#T6522
BSA	Wako	Cat#012-23881
ULTRAPURE BSA	Thermo Fisher	Cat#AM2616
RNasin plus	Promega	Cat#N2611
CellCover	Anacyte Laboratories	Cat#800-125
Bacillus licheniformis protease (CAP)	Sigma	Cat#P5380
ERCC spike-in mix	Thermo Fisher	Cat#4456740
RNeasy Lipid Tissue Mini Kit	Qiagen	Cat#74804
RNeasy Mini Kit	Qiagen	Cat#74104
dextran sulfate 500 kDa	Wako	Cat#193-09981
DAPI	Dojindo	Cat#D212
SlowFade Diamond Antifade Mountant	Invitrogen	Cat#S36963
Critical commercial assays		
Chromium Next GEM Single Cell 3' Reagent Kits v3.1	10x Genomics	Cat#PN-1000121
Fluidigm C1 with the C1 Single-Cell mRNA Seq HT IFC	Fluidigm	Cat#101-4981
SMART-Seq v4 Ultra Low Input RNA Kit	Clonetech	Cat#634888
Nextera XT DNA Library Preparation Kit	Illumina	Cat#FC-131-1024
Deposited data		
fastq files of scRNA-seq	This paper	DRA: DRA009858, DRA011653, and DRA011680
fastq files of scRNA-seq	Karaiskos et al.[24]	SRA: GSM2494783 – GSM2494789
Gene List Annotation for *Drosophila*	Hu et al.[57]	https://www.flyrnai.org/tools/glad/web/
Processed scRNA-seq data file, including UMI count table, Seurat object and loom files.	This paper	Mendeley Data: https://dx.doi.org/10.17632/k8g638cmxv.1
The viewer for browsing the dataset on the browser.	This paper	Mendeley Data: https://dx.doi.org/10.17632/k8g638cmxv.1
UMI count table of NK-data	Karaiskos et al.[24]	https://shiny.mdc-berlin.de/DVEX/
FISH reference table	Karaiskos et al.[24]	https://shiny.mdc-berlin.de/DVEX/
FISH reference table	BDTNP[15–17]	http://bdtnp.lbl.gov
Experimental models: Organisms/strains		
D.melanogaster, *y*[1] *w*[*]	Gift from Shigeo Hayashi lab	N/A
D.melanogaster, *w*[*1118*]	Bloomington Drosophila Stock Center	Stock number 5905
D.melanogaster, *w*; *R14E10-GAL4[attP2] UAS-mCD8.chRFP (III)*	Bloomington Drosophila Stock Center	Stock number 48641
D.melanogaster, *w*; *UAS-mCD8.chRFP (III)*	Bloomington Drosophila Stock Center	Stock number 27392
D.melanogaster, *UAS-bcd RNAi* (TRiP.GL00407)	Bloomington Drosophila Stock Center	Stock number 35478

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
D.melanogaster, *matalpha4-GAL-VP16[15]; matalpha4-GAL-VP16[67]*	Bloomington Drosophila Stock Center	Stock number 80361
Oligonucleotides		
Primers for C1HT, See Table S3	This paper	N/A
Probe for SABER FISH, See Table S3	This paper	N/A
Software and algorithms		
R version 4.0.3	R project	RRID:SCR_001905
Drop-seq tools version 2.5.1	Macosko et al.[89]	RRID:SCR_018142
STARsolo version 2.7.7a	Kaminow et al.[90]	RRID:SCR_021542
Seurat version 3.2.2	Stuart et al.[22]	RRID:SCR_016341
SCTransform version. 0.3.2	Hafemeister and Satija[91]	RRID:SCR_022146
g:Profiler	Raudvere et al.[92]	https://biit.cs.ut.ee/gprofiler; RRID:SCR_006809
Mclust version 5.4.7	https://cran.r-project.org/	https://cran.r-project.org/
Harmony version 0.1.0	Korsunsky et al.[51]	RRID:SCR_022206
Python 3.8.3	Python.org	RRID:SCR_008394
Perler version 0.1.0	Okochi et al.[25]	https://github.com/yasokochi/Perler
NovoSpaRc version 0.4.3	Moriel et al.[67]	https://github.com/rajewsky-lab/novosparc
Other		
Code	This paper	https://github.com/TKondolab/flygastrula2 or https://dx.doi.org/10.5281/zenodo.8012908

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Takefumi Kondo (takefumi.kondo@riken.jp).

Materials availability

Materials generated in this study are available upon request.

Data and code availability

- All raw sequence data were deposited in the DDBJ Sequence Read Archive (DRA) under accession numbers DRA009858, DRA011653, and DRA011680. Processed scRNA-seq datasets (including UMI count table, Seurat object, and loom files to visualize data in SCope (https://scope.aertslab.org/)) and the viewer for browsing the datasets on the browser have been deposited at Mendeley Data. DOI is listed in the key resources table.
- Jupyter Notebooks of code used for data analysis have been deposited at GitHub. DOI was generated through Zenodo and is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Fly strains

All stocks were maintained on standard laboratory food containing corn flour, corn grits, dry yeast, glucose, agar, propionic acid, and butyl p-hydroxybenzoate. The following fly strains were used as a control: *y w* for set 1 C1HT and set 2 trypsin-10x data. *w*; *R14E10-GAL4[attP2] UAS-mCD8.chRFP (III)* for set 3 CAP-10x data and *w* for SABER-FISH. Maternal RNAi knockdown of *bcd* was performed as previously reported.[64] Briefly, *UAS-bcd RNAi* (TRiP.GL00407) females were crossed with *matalpha4-GAL-VP16[67]* and *matalpha4-GAL-VP16[15]* males. The *matalpha4-GAL-VP16[67]/+*; *matalpha4-GAL-VP16[15]/UAS-bcd RNAi* (TRiP.GL00407) females hatched from it were crossed with *UAS-bcd RNAi* (TRiP.GL00407) males, and embryos obtained from this cross were used as *bcd*-RNAi embryos.

METHOD DETAILS

Preparation of single-cell suspensions

All equipment, including the forceps, brushes, and nylon mesh, was treated with RNase quiet (Nacalai) and washed well with RNase-free water. Embryos were collected by egg laying for 20–30 min and kept for 90 min at 25°C. Then, embryos were dechorionated using bleach and washed with RNase-free PBS. The developmental stages of embryos were monitored under a fluorescent stereomicroscope (Nikon SMZ18), and stage 6–7 embryos shortly after initiation of gastrulation were picked and transferred into 10 μL of ice-cold homogenization buffer (1x RNase-free PBS, 5% trehalose) in a 1.5 mL microtube (Watson, PROKEEP protein low binding tube). Trehalose was included in the whole dissociation process as a cell protectant[93]. After collecting 150–300 embryos at the bottom of the microtubes, the vitelline membranes were broken by slowly turning the tip of the pipette tip (Axygen, Maxymum Recovery 200 μL Universal Fit Tip with Filter). The disrupted embryos were suspended in 500 μL of ice-cold homogenization buffer and pelleted by centrifugation at 800 rcf for 2 min at 4°C. After removing the supernatant, the pellet was resuspended in 500 μL of ice-cold homogenization buffer, followed by centrifugation at 800 rcf for 2 min at 4°C.

For trypsin treatment, the pellet was resuspended in 1x trypsin-EDTA (Sigma, T3924) and kept at 25°C for 10 min. 500μL of ice-cold stopping buffer (1x PBS, 5% trehalose, 0.375% BSA (WAKO, 012–23881), 0.1 mg/mL trypsin inhibitor (Sigma, T6522)) was added. After washing with 500μL of ice-cold wash buffer1 (1x PBS, 5% trehalose, 0.375% BSA (WAKO, 012–23881)) twice, the pellet was resuspended with 200 μL of ice-cold loading buffer (1x PBS, 5% trehalose, 0.5 mg/mL ULTRAPURE BSA (Thermo Fisher, AM2616), 1/200 RNasin plus (Promega, N2611)).

For CAP treatment, the pellet after homogenization was resuspended in 500μL CAP solution (5 mg/mL *Bacillus licheniformis* protease (Sigma P5380), 5% trehalose, in 1x PBS), and kept at 6°C for 30 min. Then, 500 μL of ice-cold wash buffer2 (1x PBS, 5% trehalose, 0.5 mg/mL ULTRAPURE BSA (Thermo Fisher, AM2616)) was added. After washing with wash buffer2 four times, the pellet was resuspended in 200 μL of ice-cold loading buffer.

For either trypsin or CAP treatment, the cells suspended in 200 μL of ice-cold loading buffer were filtered through a cell strainer (FLOWMI Cell Strainers for 1000uL Pipette Tip, 40um Porosity) and fixed with 1 mL of CellCover (Anacyte Laboratories) for 1 h at 25°C, and then kept at 4°C overnight. The fixed cells were washed with 500 μL of ice-cold loading buffer and resuspended in 100 μL of ice-cold loading buffer. After counting the density of cells using a hemocytometer, the density was adjusted to approximately 200 or 300 cells/μL for Fluidigm C1HT, or approximately 300 cells/μL for 10x genomics Chromium.

Single-cell RNA-seq using C1HT

scRNA-seq library preparations using Fluidigm C1 with the C1 Single-Cell mRNA Seq HT IFC were performed according to the manufacturer's protocol with some modifications. Before proceeding to the cell lysis step, all 800 capture sites in the IFC were automatically imaged using an Axio Observer.Z.1 (Zeiss) equipped with an Axiocam 105 color (Zeiss) and an electric stage. One modification was custom primers with inserted 8 base UMI for the reverse transcription reaction. Primer sequences are listed in Table S3. We also added the ERCC spike-in mix (Thermo Fisher, 4456740) to the Lysis Mix. Another modification was the concentration of the primers used in the library amplification step. A 10-fold lower concentration of enrichment primer was used. The PCR cycle for library amplification was 12. After quality control and quantification using Bioanalyzer and qPCR, the libraries were sequenced with a NextSeq 500 (Illumina), 75 cycles high-output kit v2 (Read1: 15 cycles, Read2: 69 cycles, Index1: 8 cycles, total 92 cycles).

Single-cell RNA-seq using 10x Chromium

Library preparations using 10x Chromium with the Chromium Next GEM Single Cell 3' Reagent Kits v3.1, were performed according to the manufacturer's protocol. The PCR cycles were 11 for cDNA amplification and 11 for library amplification. After quality control and quantification using Bioanalyzer and qPCR, the libraries were sequenced using NextSeq 500 (Read1: 28 cycles, Read2, 56 cycles), NovaSeq 6000 (Illumina) (Read1: 28 or 151 cycles, Read2, 91, 98, or 151 cycles), or HiSeq X (Illumina) (Read1: 151 cycles, Read2, 151 cycles). For the trypsin dataset (set 2), the libraries were sequenced using the NovaSeq 6000. For the set 3 CAP dataset, the same library was sequenced three times with NextSeq 500, NovaSeq 6000, and HiSeq X. All reads obtained from the three sequencing times were integrated for analysis.

Bulk RNA-seq

For total RNA preparation from embryos, 80 stage 6–7 embryos were harvested, and total RNA was purified using RNeasy Lipid Tissue Mini Kit (Qiagen). For total RNA preparation from dissociated cells, 200–300 embryos at stage 6–7 were dissociated into single-cell suspensions by trypsin-EDTA treatment as described above. After washing, the cells were passed through a 40 μm strainer and pelleted. Total RNA was purified from approximately 40,000 cells using the RNeasy Mini Kit (Qiagen). For total RNA preparation from fixed cells, 200–300 embryos at stage 6–7 were dissociated into single-cell suspensions by trypsin-EDTA treatment and fixed by CellCover as described above. Cells were stored at 4°C for one day and pelleted. Total RNA was purified from approximately 40,000 pelleted cells using an RNeasy Mini Kit (Qiagen).

cDNA was synthesized from 250 ng of each total RNA using the SMART-Seq v4 Ultra Low Input RNA Kit (Clonetech). Then, a library for Illumina sequencers was constructed from 0.0625 ng cDNA using the Illumina Nextera XT DNA Library Preparation Kit. The

libraries were sequenced on an Illumina NextSeq 500 to obtain single-end reads with a length of 76 bases. Each sample was analyzed in duplicates. For each library, 36,577,021–41,844,986 reads were sequenced.

SABER-FISH

SABER-FISH was performed using a protocol described in Kishi et al., 2019[94] with some modifications. The "Balance" list of candidate probe sequences pre-designed using OligoMiner pipeline[95] was downloaded from https://oligopaints.hms.harvard.edu/genome-files. Thirty probe sequences that target exons common to all isoforms of each gene were randomly extracted from the list. SABER-FISH probes were prepared by PER amplification using DNA oligos (Table S3) purchased from IDT.

Embryo fixation was performed by a protocol as previously described.[76] Briefly, embryos were dechorionated in 50% bleach for 2 min and fixed in 1:1 4% PFA containing 1 mM CaCl2 and heptane for 20 min at room temperature. The vitelline membrane was removed by shaking in 1:1 methanol and heptane. Fixed embryos were rinsed in methanol 2 times and collected to DNA LoBind Tubes (Eppendorf). And then, embryos were washed in PBS with 0.2% Tween 20 and 0.2% Triton X-100 (2 × 2 min), washed in PBSTw (PBS with 0.1% Tween 20) (3 × 5 min), and placed in 1:1 PBSTw and Whyb buffer (2× SSC pH 7.0 with 1% Tween 20 and 40% formamide) for 5 min. Before hybridization, embryos were incubated in Whyb for 10 min at 43°C. And then, embryos were incubated with 1 μg/100 μL probes in pre-warmed Hyb1 buffer (Whyb with 2.5% dextran sulfate 500 kDa (FUJIFILM Wako, 193–09981)) for 16–48 h at 43°C. After hybridization, embryos were washed in Whyb (quickly once and 3 × 30 min) and washed in 2× SSCT (2× SSC with 0.1% Tween 20) (3 × 5 min) at 43°C. Before fluorescent oligo hybridization, the tube was returned to room temperature, and embryos were washed in PBSTw (3 × 5 min). The tube was then transferred to 37°C and once the tube was warm, the PBSTw was removed and replaced with 0.2 μM DNA oligos conjugated with Alexa-fluor (synthesized by Thermo Fisher, Table S3) in pre-warmed Hyb2 buffer (PBSTw with 2.5% dextran sulfate 500 kDa). After incubation for 20 min at 37°C in the dark, embryos were washed in PBSTw (quickly once and 3 × 5 min) at 37°C. Then, embryos were incubated with 1 μg/mL DAPI (Dojindo) in PBSTw for 30 min at room temperature in the dark and then washed in PBSTw (3 × 15 min). Embryos were mounted in SlowFade Diamond Antifade Mountant (Invitrogen). Images were taken using a Zeiss LSM800 with a 40× water immersion objective (Objective LD LCI Plan-Apochromat 40x/1.2 Imm Corr DIC M27, Zeiss).

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of 10x Chromium data

Read1, including UMIs and cell barcodes, was trimmed to 28 base lengths using fastx_trimmer (FASTX-toolkit, version 0.0.14, http://hannonlab.cshl.edu/fastx_toolkit). Adapter trimming and quality filtering were performed using fastp (version 0.20.1, for NextSeq 500 data with q 20 –cut_tail -l 28 –max_len1 28 –max_len2 55 –trim_poly_g –trim_poly_x, for 10x NovaSeq data -q 20 –cut_tail -l 28 –max_len1 28 –max_len2 97 –trim_poly_g –trim_poly_x, for HiSeq X data with -q 20 –cut_tail -l 28 –max_len1 28 –max_len2 97 –trim_poly_x).[96] The trimmed reads were mapped to the genome sequence of *Drosophila melanogaster* (BDGP6.22.98) and UMI-counted using STARsolo (version 2.7.7a).[90] In this process, since STARsolo (version 2.7.7a) cannot account for multi-gene reads for UMI counting, a modified gtf annotation file in which genes overlapped in the same direction of the genome were integrated and treated as the same gene (Table S4) was used. For cell filtering, the median of the total UMI per cell in the filtered output of STARsolo was calculated, and cells with a total UMI two times higher than the mean value were filtered as potential doublets. Then, cells in which either the number of genes detected, the UMI proportion of ribosomal RNA genes, or the UMI proportion of mitochondrial genome genes were outside the range of an average value ±2.5 × SD were filtered as low-quality cells. The remaining UMI-count tables were loaded into Seurat (version 3.2.3)[22] and normalized using the SCTransform function with an option (vars.to.regress = c("percent.mt", "percent.rRNA")).[91] "percent.mt" and "percent.rRNA" were labels of metadata which contain the percentages of transcripts from the mitochondrial genome and nuclear rRNA genes to total detected transcripts in each cell respectively. If the residual_variance (Pearson residual) returned by SCTransform is 1, the variance for that gene is considered to be noise. So, we define the (residual_variance −1) as the biological variance. Since cumulative sums plateaued at the top 3,000 genes when values were taken in descending order, the number of HVGs used for dimensionality reduction was set at 3,000. Then PCA analysis was performed using the RunPCA function in Seurat. To determine the number of dimensions to be used in subsequent analyses, an Elbow plot of the standard deviation of each principal component in Set3 was produced using the ElbowPlot function in Seurat. Based on this plot, the number of dimensions used was set to 30. The same number of dimensions was used throughout this paper unless otherwise mentioned. UMAP analysis was performed using RunUMAP (dims = 1:30, n.neighbors = 20L) functions, followed by unsupervised graph-based clustering with FindNeighbors and FindClusters functions in Seurat. In this study, this clustering output by Seurat using all cell data is referred to as "Seurat cluster". Seurat clusters showing high expression of ribosomal protein genes and not expressing the markers corresponding to the embryonic space were filtered out as low-quality cells. Each cluster was manually annotated based on the marker genes identified by the FindAllMarkers function. All plots were generated using Seurat or ggplot2 in R unless otherwise noted.

Analysis of C1HT data

Adapter trimming and quality filtering were performed using fastp (version 0.20.1) with options (-q 20 –cut_tail -l 14 –max_len1 14 –max_len2 68 –trim_poly_g –trim_poly_x), and the trimmed reads were mapped to the genome sequence of *Drosophila melanogaster* (BDGP6.22.98) with a modified gtf annotation file as described above and UMI-counted using STARsolo. Only the data derived from

the cells determined to be a singlet from the image of the capture site were loaded into Seurat. For each batch, cells in which either the number of genes detected, the UMI proportion of ribosomal RNA genes, the UMI proportion of mitochondrial genome genes, or the UMI proportion of ERCC spike-ins were outside the range of an average value ±2.5 × SD were filtered as low-quality cells. Then all four batches were merged and normalized using SCTransform with an option (vars.to.regress = c("percent.mt", "percent.rRNA", "percent.ERCC")). "percent.ERCC" were labels of metadata which contain the percentage of ERCC to total detected transcripts in each cell. Dimensionality reduction, graph-based clustering, and Seurat cluster annotation were performed similarly to the 10x data. All plots were generated using Seurat or ggplot2 (version 3.3.3)[97] in R, unless otherwise noted.

Analysis of scRNA-seq data of NK-data

Fastq files (GSM2494783 – GSM2494789) reported in Karaiskos et al., 2017[24] were obtained from SRA database. Adapter trimming and quality filtering were performed using fastp (version 0.20.1) with options (-q 20 –cut_tail -l 20 –max_len1 20 –max_len2 64 –trim_poly_g –trim_poly_x), and potential SMART adapter were further trimmed from read 2 using Cutadapt (version 3.4) with option (-m 20:20 -G AAGCAGTGGTATCAACGCAGAGTACATGGG). The trimmed reads were mapped to the genome sequence of *D. melanogaster* (BDGP6.22.98) with a modified gtf annotation file as described above or the reference that combines this *D. melanogaster* and *D. virilis* (GCF_003285735.1_DvirRS2) references using STAR. Then, the UMI count tables were generated from the BAM files using in TagReadWithGeneFunction, DetectBeadSubstitutionErrors, DetectBeadSynthesisErrors (with –PRIMER_SEQUENCE AAGCAGTGGTATCAACGCAGAGTAC) and DigitalExpression (with –NUM_CORE_BARCODES 5000) in Drop-seq tools (version 2.5.1).[89] For data containing *D. melanogaster* and *D. virilis* cells, only cells with more than 90% of the total number of UMIs mapped to *D. melanogaster* were considered to be *D. melanogaster* cells and extracted for further analyses. As in the original paper, cells with a total UMI count of 12,500 or more were retained as high-quality cells. For each batch, cells in which either the number of genes detected, the UMI proportion of ribosomal RNA genes, or the UMI proportion of mitochondrial genome genes were outside the range of an average value ±2.5 × SD were filtered as low-quality cells. Then, all seven batches were merged and normalized using SCTransform with an option (vars.to.regress = c("percent.mt", "percent.rRNA")). Dimensionality reduction, graph-based clustering, and Seurat cluster annotation were performed in the same way as the 10x data. All plots were generated using Seurat or ggplot2 (version 3.3.3)[97] in R, unless otherwise noted.

Filtering out *tsr*-high cells from trypsin data

To remove *tsr*-high cells from Set2 trypsin-10x data and *bcd*-RNAi data, the correlation coefficients between all pairs of 2,000 HVGs were calculated with the correlate function in the corrr package (version 0.4.3) of R. Then, genes positively and negatively correlated with *tsr* were extracted to perform principal component analysis (PCA) and model-based clustering by the Mclust function in mclust package of R (version 5.4.7, https://cran.r-project.org/package=mclust) with options (pca = 30, G = 2, modelNames = "VVV"). Finally, the cluster with high *tsr* expression was filtered out as stressed cells for further data integration.

For trypsin-C1HT data (Set1), after removing pole cells, normalization using SCTransform, dimensionality reduction using RunPCA and clustering using the FindNeighbors and FindClusters functions (resolution = 2.0) with 30 dimensions were performed. Of clusters detected, two clusters that showed high expression of *tsr* were removed.

Subclustering of each scRNA-seq data

For subclustering, we further applied unsupervised graph-based clustering using the FindNeighbours and FindClusters functions for each set unit of clusters as shown in Table S2. If the number of cells in the cluster set unit was 500 or more, normalization by the SCTransform function was performed again, and then genes with highly variable features detected by SCTransform were used for dimensionality reduction by the RunPCA function. We noticed that if the number of cells was below 500, SCTransform does not normalize properly. In that cases, additional normalization and HVG selection were not performed. If the number of cells was more than 30, 30 dimensions were used for RunUMAP, FindNeighbours, and FindClusters, and if the number of cells was 30 or less, dimension number −1 was used. The resolution parameter in the FindClusters function was empirically determined. The maximum value at which no subcluster can be annotated based on knowledge of the literature was adopted. One exception is that subclustering along the DV axis of the lateral ectoderm was performed using k-means clustering (n = 7) with only 35 DV genes (*Ama, Ance, Atx-1, bbg, brk, C15, CG13653, cic, cv-2, dap, Doc1, Doc2, Doc3, dpp, Dr, Dtg, Egfr, egr, emc, ind, mirr, peb, pnt, pnt, rho, sog, SoxN, srp, stg, tup, ush, vn, vnd, Z600, zen*). Each subcluster was manually annotated based on the marker genes shown in Table S2. During subclustering, cells showing expression of both ectodermal and mesodermal genes were removed as doublets and filtered. In addition, cells that did not express the markers corresponding to the embryonic space were also removed as low-quality cells. The remaining singlet dataset consisted of 6,118, 1018, 4,855, and 1,476 cells for Set3, Set1, Set2, and NK-data, respectively.

Harmony integration of scRNA-seq datasets

The control integration data was generated using Set1, Set2, Set3, and NK-data after subclustering. Each dataset was individually normalized by SCTransform, and 3,000 genes to be used for integration were determined using the FindIntegrationFeatures function in Seurat. Dimensionality reductions were performed in the same way as the 10x data. Then, batch correction was performed using the Harmony (version 0.1.0)[51] with 30 PCs, followed by dimensionality reduction using the RunUMAP function and clustering using the FindNeighbors and FindClusters functions. Each cluster was annotated as in the case of individual datasets.

Subclustering of the integrated datasets

As for each individual dataset, subclustering analysis was performed for the control integration dataset. First, the cluster set unit was divided into the original four datasets. If the number of cells in the cluster set unit was 500 or more, normalization by the SCTransform function was performed again. If the cell number is less than 500, renormalization was not performed, and the values of the scale.data slot in the SCT assay of Seurat object were centered by the ScaleData function (with do.scale = FALSE). Then, the 3,000 HVGs used for re-integration were selected using the FindIntegrationFeatures function in Seurat, and dimensionality reduction was performed using the RunPCA function. Then, batch correction was performed using the Harmony with 30 PCs, followed by the dimensionality reduction using the RunUMAP function. Clustering and annotation were performed as in the case of individual datasets described above.

Comparison of sensitivity for marker genes

First, the integrated data described above was split again for each derived dataset. Among them, the FindAllMarkers function (method = 'MAST') was applied to Set1 (C1HT) and Set2 (10x Chromium) data to detect marker genes in each cluster at an FWER threshold of 0.01. In each of Set1 and Set2, genes detected as markers in at least one cluster were listed. To calculate the mean expression value of each marker gene, cell data were extracted from all clusters in which the gene was detected as a marker gene, and the average expression level of the gene in the cell group was calculated.

Merge of Set2 and *bcd*-RNAi data

For the Set2 and *bcd*-RNAi data after subclustering, both were re-normalized separately by the SCTransform function and merged using the "merge" function in Seurat. 3,000 HVGs for this merged dataset were determined using the SelectIntegrationFeatures function in Seurat, and the RunPCA function was applied without batch correction. 30 PCs were used for dimensionality reduction by the RunUMAP function and clustering by the FindNeighbors function and the FindClusters Clustering. Since, each Set2 and *bcd*-RNAi data were mixed well in this UMAP plot of this merged data (Figure 6B), further batch correction was not applied. Each cluster was annotated as for individual datasets. The Fisher exact test was used to test whether, for each Seurat cluster, the ratio of cells assigned to the cluster is different between *bcd*-RNAi and control dataset.

DEG analysis between Set2 control *bcd*-RNAi

First, the subcluster information of *bcd*-RNAi and Set2 were manually linked. For cells annotated as "Dorsal_lateral_ectoderm_PS13" and "Amnioserosa_PS13-14" in the Seurat cluster after merging Set2 and *bcd*-RNAi, this cluster information was used instead of the subcluster information. Then, DEGs were determined by the FindMarkers function (method = 'MAST') between *bcd*-RNAi and Set2 for each subcluster present in both *bcd*-RNAi and Set2 and filtered by an FWER threshold of 0.01 and |logFC| threshold of 0.5. DEGs with high expression in *bcd*-RNAi were considered positive DEGs, and those with low expression in *bcd*-RNAi were considered negative DEGs.

To extract marker genes for each subcluster in Set2, the FindAllMarkers function (method = 'MAST') was applied to Set2 to detect marker genes at an FWER threshold of 0.01, and marker genes detected only in anterior subclusters not found in *bcd*-RNAi were defined as anterior markers. The list of these anterior markers was compared with the list of positive DEGs, and the common genes including *Dll* were considered candidates for residual features of transformation from anterior to posterior in *bcd*-RNAi embryos.

Gene Ontology enrichment analysis

For GO enrichment analysis of all high-quality cells, cells in the pole cell cluster were removed, and the list of the top 3,000 highly variable features was extracted using Seurat. The gene list was analyzed using g:Profiler (https://biit.cs.ut.ee/gprofiler) with g:SCS algorithm,[92] and significantly enriched terms in cellular components were identified at an FDR threshold of 0.01 and term_size lower than 4,000. For the analysis of highly expressed genes in the *tsr*-high cluster of Set2 10x trypsin dataset, 328 highly expressed genes in cluster 10 (Mesoderm_tsr-high) compared to cluster 1 (Trunk_mesoderm) were identified using the FindMarkers function with MAST[98] at an FWER threshold of 0.01 and logFC threshold of 0.25. The gene list was analyzed using g:Profiler, and significantly enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) terms were identified at an FDR threshold of 0.01.

Hierarchical clustering analysis with GLAD

The gene list of each GLAD category was downloaded from https://www.flyrnai.org/tools/glad/web/. Since some genes listed in the "Transcription factor/DNA binding" category are also listed in other categories, a modified database was prepared to exclude these duplications. An integrated gene list of "Trans-membrane proteins," "Receptors," "Secreted proteins," and "Matrisome" categories was used as the list of plasma-membrane (PM)-related genes. Genes listed in GLAD categories other than the "Transcription factor/DNA binding" category and plasma-membrane-related genes were considered as other cytoplasmic genes. Since "PM-related genes" includes genes that function in the organelles, the genes having the "intracellular membrane-bounded organelle" (GO:0043231) or its child term (Including "mitochondrion," "Golgi apparatus" and "endoplasmic reticulum") were transferred to other cytoplasmic genes. The modified GLAD list is presented in Table S5. In addition, since TF genes were particularly enriched in the top 1,500 HVGs (Figure 4B), the top 1,500 HVGs were used in this clustering analysis to focus on the significance of this TF enrichment.

For the analysis using only a specific gene set, to align the numbers to the lowest TFs, only 258 genes were selected from the top of variance in all sets. The dimensionality reduction analysis was performed using Seurat, as described above. The cell identity in the UMAP plot was colored using pre-annotated information (Figure 4I). For the hierarchical clustering analysis, the average normalized expression value of each gene for each of the 76 subclusters (pole cells were removed) was calculated using the AverageExpression function of Seurat. Euclidean distances for all pairs of subclusters in log-transformed gene-expression space were calculated using the dist function of R. Then, hierarchical clustering based on the Euclidean distances was performed with the hclust function with the average method. Euclidian distances and the structure of hierarchical clustering were drawn as a heatmap using the heatmap.2 function in the gplot package (version 3.1.1, https://CRAN.R-project.org/package=gplots) of R.

Assignment of stripe identities

scRNA-seq data annotated as "trunk ectoderms 2" (see Table S2 for details) were extracted. It is considered that each parasegment is composed of four stripes along the A-P axis, and each stripe is a single-cell-wide column and has different gene expression profiles.[63] Therefore, to infer the stripe positions in parasegment, "trunk ectoderms 2" cells were analyzed by k-means clustering (n = 4) with nine landmark genes of the stripe position (Figure 5A). Then, data in each stripe were divided into odd or even parasegment by k-means clustering (n = 2) with *trn* for stripes 1 and 2 (or *pxb* for stripes 3 and 4) and genes positively and negatively correlated with it. K-means clustering was performed with the k-means function in the ClusterR package (version 1.2.2, https://CRAN.R-project.org/package=ClusterR) of R. The correlation coefficient between all pairs of 1,000 HVGs was calculated with the correlate function in the corrr package (version 0.4.3) of R. All plots were generated by Seurat or ggplot2 (version 3.3.3) in R. DEGs between each adjacent boundary or super boundary were identified using the FindMarkers function with MAST and filtered by an FWER threshold of 0.01, and FC threshold of 1.75. FC threshold was empirically determined by plotting the FC distribution of DEGs between adjacent boundaries with an FDR of 0.01 or less.

Spatial reconstruction of gene expression

Preprocessing scRNA-seq data

For Set3 data, because the BDTNP FISH data does not contain pole cells, 123 cells in the "pole_cells" cluster were removed from the dataset. As described above, the UMI-count table of the remaining 5,995 cells was renormalized using SCTransform. Then, dimensionality reduction analysis was performed using the RunPCA function of Seurat with default settings.

The raw count table (dge_raw.txt) for NK-data was obtained from *Drosophila* Virtual Expression eXplorer (https://shiny.mdc-berlin.de/DVEX/). This count table was loaded into Seurat and normalized using the SCTransform function without options. Note that, in this data, pole cells were already removed, and the UMI counts for mitochondrial and rRNA genes were omitted.

For both datasets, a log-scaled count ("data" slot in "SCT" assay of the Seurat object) and HVGs detected by SCTransform were used for spatial reconstruction.

Selection of ISH reference landmark genes

ISH spatial references were constructed mainly based on the BDTNP database (D_mel_wt__atlas_r2.vpc from http://bdtnp.lbl.gov) and DVEX (bdtnp.txt). The DVEX reference was forked from the BDTNP reference, but three genes (*bowl*, *ems*, and *exex*) were only in the DVEX reference.

Both scRNA-seq data were derived from stage 6–7 embryos, while ISH reference data were established for stage 5 embryos. Some genes in the ISH data dynamically changed the expression pattern from stage 5 to stage 6–7. Therefore, genes whose expression patterns significantly changed between the two time points and that could worsen the reconstruction were removed from the reference. As a result, 67 genes remained as landmarks for spatial reconstruction (Table S7). In addition, among 3,039 cells in the DVEX reference, eight cells with $y < 0$ were removed.

Spatial reconstruction by Perler

Perler (version 0.1.0) Python package was obtained from https://github.com/yasokochi/Perler. For both set 3 and NK-data, log-scaled counts and reference above were loaded into the PERLER object, and then the EM algorithm was performed using the em_algorithm method with option (optimize_pi = False). Next, the distances between the scRNA-seq data points and reference data points were calculated using the calc_dist method with default parameters. Optimization of hyperparameters was performed by the loocv method with default parameters and the gridsearch method with parameters (grids = ((0,1), (0.01,1))). Finally, spatial gene expression patterns were reconstructed by the spatial reconstruction method with parameters (mirror = False, _3d = True, z_scored = False).

Spatial reconstruction by NovoSpaRc

NovoSpaRc (version 0.4.3) reconstruction was mainly performed according to https://github.com/rajewsky-lab/novosparc. First, log-scaled counts and reference data were loaded into the Tissue object of NovoSpaRc. Cost matrices for the optimal transport framework were calculated by the set_up_smooth_costs method based on 30 principal components (PCs) and the setup_linear_cost method with the reference and default parameters. Then, spatial reconstruction was performed using the reconstruction method with parameters (alpha_linear = 0.3, epsilon = 5e-3). Note that the alpha_linear parameter and the number of used PCs were determined by grid search so that the LOOCV score described below was maximized.

Leave-one-gene-out cross-validation (LOOCV)

Each of the 67 landmark genes in the ISH reference was removed from the reference as the true expression, and spatial reconstruction by Perler or NovoSpaRc was performed using the remaining 66 genes as the reference. Pearson correlations between the reconstructed expression pattern of the removed gene and the truth were then calculated.

For the decision of hyperparameter for NovoSpaRc, we calculated the score:

$$J = -\frac{1}{2}\sum_{i}^{67}\ln\left(1-\rho_i^2\right)$$

Here, ρ_i is the Pearson correlation coefficient between the ISH expression and the reconstructed expression of gene i. The hyperparameter and number of PCs with the highest score were selected.

Comparison of gene correlation conservation

First, 372 common HVGs included in the top 500 HVGs of both set 3 and NK-data scRNA-seq datasets were selected. For each HVG, Pearson correlation coefficients between the gene and other 371 HVGs in the original scRNA-seq data and those in the set-3-based or NK-data-based reconstruction were calculated. Then, the Pearson correlation coefficient between these two correlation scores for each of the 372 HVGs was calculated as gene-gene correlation structure conservation between the original scRNA-seq data and either reconstruction. For the comparison between Perler and NovoSpaRc, the top 500 HVGs of the set 3 data were used.

Plotting reconstructed expression pattern

Plotting was performed using the scatter function in the matplotlib (version 3.3.4) package.[99] The reconstructed expression values were converted from a log-scale to a linear scale. For the lateral view, all the cells in the reference were plotted. For dorsal and ventral views, cells with $z > 0$ and $z \leq 0$ were used, and the cells were mirrored on the x-z plane. The anterior is left in all plots, and the dorsal is up in lateral views.

Density plot of ind *and* vnd *expression*

For the plot of scRNA-seq data, intermediate or medial neuroectoderm cells in the abdomen and PS13 and midline cells were extracted from set 3 data. For the reconstruction data plot, cells with $|x| < 50$ and $-55^\circ < \theta < 0^\circ$ were extracted. θ is the angle between the y axis and the line segment drawn from the center of the embryo to the cell in a cross-section parallel to the y-z plane containing the cell, expressing the position of the cell on the DV-axis (Figure S7G). Density estimation was performed using the Gaussian_kde class in the stats module in the Scipy package (version 1.6.0).[100] Estimation results were plotted using the pcolormesh function in the matplotlib package (version 3.3.4).

Analysis of bulk RNA-seq data

Sequenced reads were quality trimmed using Trim Galore (version 0.6.4, https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and Cutadapt (version 1.18)[101] with the –nextseq 20 option. After removing the 76th base from each read, the remaining reads were mapped to the genome sequence of *Drosophila melanogaster* (BDGP6.22.98) using STAR with a modified gtf annotation file, as described above. Gene expression was calculated using RSEM (version 1.3.3),[102] and differential expression analysis was performed using edgeR (version 3.32.1).[103] After removing the mitochondrial and ribosomal RNA genes, low expression genes with CPM less than 0.1 in all six samples were also filtered out. Normalization was performed using calcNormFactors. Spearman correlation coefficients were calculated using the cor function in R. DEGs were identified using the glmQLFit and glmQLFTest functions in the edgeR package at an FDR threshold of 0.01 and logFC threshold of 2.

ADDITIONAL RESOURCES

All reconstructed stripe patterns, spatial reconstruction results, and a viewer for them were available in https://github.com/TKondolab/flygastrula2 and https://dx.doi.org/10.17632/k8g638cmxv.1.