Gene expression atlas of a developing tissue by single cell expression correlation analysis

Josephine Bageritz1,2, Philipp Willnow1,2,3, Erica Valentini1,2, Svenja Leible1,2, Michael Boutros1,2,* and Aurelio A. Teleman1,2,*

The Drosophila wing disc has been a fundamental model system for the discovery of key signaling pathways and for our understanding of developmental processes. However, a complete map of gene expression in this tissue is lacking. To obtain a gene expression atlas in the wing disc, we employed single cell RNA sequencing (scRNA-seq) and developed a method for analyzing scRNA-seq data based on gene expression correlations rather than cell mapping. This enables us to compute expression maps for all detected genes in the wing disc and to discover 824 genes with spatially restricted expression patterns. This approach identifies clusters of genes with similar expression patterns and functional relevance. As proof of concept, we characterize the previously unstudied gene CG5151 and show that it regulates Wnt signaling. Our method will enable the leveraging of scRNA-seq data for generating expression atlases of undifferentiated tissues during development.

The Drosophila wing imaginal disc has been an important model system for studying tissue growth, pattern formation, epithelial morphogenesis, inter-cellular signaling, cell competition and tissue biophysics. Yet, the expression patterns of most genes in the wing disc are not known. In the past, genes with non-uniform expression patterns were identified by in situ hybridization screens or by profiling RNA obtained from dissected disc regions. Methods analogous to those used for differential gene expression analysis across treatments or timepoints are then used for identifying patterns in such spatial transcriptomics data. Recent advances in scRNA-seq allow the clustering of cells into cell types based on their expression profiles, however, clustering per se does not provide spatial information. Single cell sequencing has enabled the generation of genome-wide spatial expression maps by dissociating a tissue, sequencing single cells and then reassembling the tissue in silico by mapping back cells to tissue locations based on the expression patterns of known genes. Finally, data from scRNA-seq and fluorescence in situ hybridization (FISH) can be combined to first cluster cells by cell type and subsequently extract spatial information.

The wing disc presents several challenges. It is composed mainly of pluripotent, undifferentiated stem-like cells; hence, it consists of few cell types. With 50,000 cells, the wing disc has ten times more cells than, for instance, the Drosophila embryo. Most mammalian organs have more cells, representing an even larger challenge. We present here a scRNA-seq analysis method for generating gene expression maps in a large and undifferentiated tissue, based on analyzing gene expression correlations rather than mapping sequenced cells back to their tissue locations. This allowed us to compute expression maps for all genes in the wing disc and discover 824 genes with spatially restricted expression.

Results

scRNA-seq of Drosophila wing disc cells. To construct a gene expression map of the Drosophila wing disc, we dissociated cells from wing disc fragments of third instar female larvae and sequenced their messenger RNA. With DropSeq we sequenced 1,644 cells with median depth of 3,774 transcripts and 1,134 genes per cell (Supplementary Fig. 1a,b). With 10X Genomics we sequenced 2,554 cells with median depth of 10,620 transcripts and 1,998 genes per cell (Supplementary Fig. 2a). These sequencing depths are in line with what others have reported (Supplementary Table 1). We could unambiguously identify true cell barcodes (Supplementary Fig. 1c) indicating low ambient mRNA from cell breakage during sample preparation. Two biological replicates correlated highly with each other (Pearson correlation r = 0.93, Supplementary Fig. 1d) indicating reproducibility. The sum of all single cell reads correlated well with RNA-seq data of nondissociated wing discs (Supplementary Fig. 1e), suggesting that scRNA-seq captured most of the expressed genes and that the procedure did not strongly alter gene expression.

To identify cell types, we clustered cells using the Seurat R package and visualized clusters using a t-distributed stochastic neighbor embedding (t-SNE) plot (Fig. 1a). This revealed two main clusters corresponding to wing disc proper cells and associated adult muscle precursor cells (AMPs) (Fig. 1a,b). Since we focus here on the wing disc proper, we excluded the 520 AMPs from subsequent analyses.

Identification of spatially restricted genes (SRGs). To identify genes with spatially restricted expression patterns, we plotted for every gene the number of cells in which the gene was detected versus its average expression level in the cells where it was detected (Fig. 1c). The stronger a gene is expressed, the higher the chance it will be detected in a cell. Most genes lie on a curve that increases and asymptotes near the total number of sequenced cells. SRGs are genes observed in fewer cells than expected, given their expression level (dots above the curve, Fig. 1c, left panel). We identified SRGs by residuals that are one standard deviation below the mean on the inverse graph (green points, Fig. 1c, right panel) and obtained a set of 824 SRGs (Supplementary Table 2). As a benchmark, we compiled a list of 68 genes from the literature that are expressed in restricted domains of the wing (Supplementary Table 3). The SRGs include 63 of these (magenta dots in Fig. 1c,d). In our DropSeq data set, we did not find as SRGs aos, arr, Dig, fng and Mes2. However, we identified
arr and fng as SRGs in the 10X Genomics data set. In comparison, a similarly sized set of 829 ‘highly variable genes’ (HVGs) obtained using the Seurat package contained 17 benchmark genes (Fig. 1d and Supplementary Fig. 3). This is likely because the HVG concept was designed for a different purpose—to identify genes with variable expression. Genes expressed ubiquitously but at varying levels are classified as HVGs, although they may not be spatially restricted. Our SRG analysis is more closely related to M3Drop/NBDrop, which detects nonubiquitous genes by analyzing drop-out rates. To identify benchmark genes (Supplementary Table 3) with a false-negative rate below 10%, the SRG algorithm generates a list of 824 genes and NBDrop generates a list of 1,000 genes (Supplementary Fig. 3a). In sum, the SRG analysis is well suited for identifying genes with spatially restricted expression domains.

Using the SRGs, the cells formed five clusters along the wing proximal-distal axis, corresponding to all the major wing regions: the wing margin, wing pouch, proximal wing, hinge and notum (Fig. 1e,f). These cell populations could also be identified using Seurat HVGs or NBDrop genes (Supplementary Fig. 4), indicating that cell clustering works with different sets of genes. We found no biases in these clusters in terms of the number of unique molecular identifiers (nUMIs) per cell, read alignment rate, fraction of mitochondrial RNA or representation of the two biological replicates (Supplementary Fig. 5).

**Fig. 1 | Single cell sequencing of wing disc cells identifies SRGs.** a,b, Two-dimensional t-SNE representation of all sequenced cells reveals two main cell clusters (a) corresponding to wing disc cells and AMPs, based on differential expression of AMP genes in the two clusters (b). n = 1,468 high-quality cells. c, Identification of SRGs as genes observed in fewer cells than expected based on their expression level. n = 9,929 genes. Black lines, regression lines. d, The set of 824 SRGs contains most of the 68 benchmark genes known to have spatially restricted expression domains based on literature (Supplementary Table 3). In comparison, a similarly sized set of HVGs contains 17. Analysis was done on the DropSeq data. e,f, Two-dimensional t-SNE representation of all wing disc cells using the 824 SRGs for dimensional reduction identifies five clusters along the proximal-distal axis of the wing disc (e), based on expression of known marker genes (f). n = 615 high-quality wing disc proper cells.
In total, 2.5% of the cells were mutually exclusive.

e, Schematic representation of the method used to generate computed expression maps.

d, Concept for generating expression maps based on gene expression correlations. A positive correlation of gene expression across all sequenced wing disc cells.

c, Top hits from genome-wide correlation analysis of gene expression across all sequenced wing disc cells. \( n = 948 \) cells. Correlation was calculated using Pearson’s correlation coefficient (Correl. coeff.) with one outlier removed. See Methods for details.

To generate a gene expression atlas of the wing disc, we first tested whether we could map the location of the sequenced cells to the wing disc based on the presence or absence of expression of genes with known expression domains, such as engrailed for the posterior of the wing, or \( ci \) for the complementary anterior (Supplementary Fig. 6). However, we could not confidently map cell locations because the transcriptome coverage of current single cell approaches is not sufficient to distinguish whether a gene is not expressed or not detected in any given cell. For instance, although \( \sim 35\% \) of wing disc cells should express engrailed and the other \( \sim 65\% \) should express \( ci \), in our DropSeq library only 14% of cells were \( en^{+} \) (>0 reads) and 28% were \( ci^{+} \) (left edge of graph, Fig. 2a). Setting a minimum nUMI per cell threshold did not solve this: above a threshold of 12,000 UMI per cell, 84% of cells were \( en^{+} \) or \( ci^{+} \), with only 45 of the 948 sequenced cells passing this threshold (Fig. 2a). In total, 2.5% of the cells were \( en^{+}/ci^{+} \) double-positive, suggesting they are doublets. Using our 10X Genomics data set, which has more sequencing depth (Supplementary Table 1) we reached 90% confidence on the cell location in the \( en^{+} \) or \( ci^{+} \) domain with a 45,000 UMI per cell threshold, which leaves only 140 cells in the data set (Supplementary Fig. 2b). Since cells would need to be mapped more precisely along the anterior–posterior, dorsal–ventral and proximal–distal axes using multiple markers, the uncertainties compound and it is not possible to confidently map a cell’s location based on these scRNA-seq data. In line with this, we tested methods that successfully reconstructed spatial gene expression from scRNA-seq data in zebrafish and the \( Drosophila \) embryo17,18, but these methods did not work well for our data set (Supplementary Fig. 7). We therefore searched for an alternative method.

Computation of gene expression maps based on gene expression correlations. We noticed good correlations in gene expression...
between genes across the hundreds of sequenced cells: we calculated correlation coefficients between en and all other genes in the genome, and as expected based on their expression patterns, the top genes correlating to en are inv and hh, and the top anti-correlating gene is ci (Fig. 2b). Likewise, the top genes correlating or anti-correlating to wg or dpp are expressed in overlapping or complementary expression patterns, respectively (Fig. 2b). The underlying data can be visualized using two-dimensional histograms (Fig. 2c): few cells express both wg and frizzled 2 (fz2), which are expressed in complementary domains. In contrast, multiple cells have detectable transcripts for both wg and Wnt6 (Fig. 2c), which are expressed in overlapping domains. Likewise, few cells are en/+ci+, whereas many cells are en+/+inv+ or en/+hh+. This analysis also identifies novel genes that correlate with en, suggesting a similar expression pattern (CR44334, Fig. 2b).

We therefore conceived a method for calculating gene expression maps based on gene correlations, without mapping the location of sequenced cells in the tissue and without first identifying variable genes. This method uses the correlation coefficient between two genes to determine whether their expression domains are overlapping (positive correlation), complementary (negative correlation) or orthogonal (no correlation) (Fig. 2d). For a given cell within the known expression domain of Gene 1 (red dot, Fig. 2d), uncharacterized Gene 2 is likely also expressed in this cell if the two genes correlate and not expressed if they anti-correlate. If the correlation coefficient is close to zero, the expression domain of Gene 1 is not informative with regards to Gene 2. We compiled a map of the wing disc containing the expression domains of 58 ‘mapping genes’ known from the literature to have distinct expression patterns. We overlaid, aligned and thresholded these patterns, yielding binary expression domains (Fig. 2e and Supplementary Table 4). We then calculated a cross-correlation matrix between these 58 mapping genes and all genes in the DropSeq data set. To compute an expression map of a gene, for each position in the wing we added the correlation coefficients between this gene and the mapping genes with a +1 or −1 weighting factor depending on whether the mapping gene is expressed in that position or not (Fig. 2e). We verified the resulting ‘computed expression maps’ by comparing them to FISH (Figs. 3 and 4).

Comparison of computed maps to in situ. This method generates computed expression maps for all genes detected in the sequencing experiment. We illustrate here three approaches to identify genes of interest, based on the similarity of their expression patterns to known genes: (1) clustering genes using a two-dimensional dendrogram, (2) searching for genes that correlate or anti-correlate with one specific gene of interest and (3) generating an expression interaction network. To cluster genes by expression pattern, we calculated a cross-correlation matrix of gene expression for all 824 SRGs from our DropSeq data set against each other and then hierarchically clustered them (Fig. 3a and Supplementary Data 1). Visual inspection of this dendrogram confirmed that neighboring genes have similar expression patterns (for example, en/inv/hh, hh/hh/zfh2 or wg/Wnt4/Wnt6/ct). We selected three clusters containing both characterized and uncharacterized genes and performed FISH on all genes in the clusters. The ‘red’ cluster (Fig. 3b) contains genes expressed in the wing pouch with a pattern along the anterior–posterior axis. This includes the mapping gene ‘kn’ and genes with unknown expression patterns and functions in the wing. FISH confirmed that CG9850 has a mild ‘kn-like’ stripe that is less accentuated than kn, that CG3168 has a broader expression pattern in the pouch that is repressed at the dorsal/ventral boundary (Supplementary Fig. 6a) and that Trim9 (ref. 24) is expressed predominantly in the pouch with an inverse venation pattern and inhibition at the dorsal/ventral boundary. The FISH for CG7201 had some elements of the predicted map, such as higher expression medially and broad repression at the dorsal/ventral boundary, but also differed from the map. Thus, overall, the computed maps are able to predict the main features of the gene expression patterns. FISH for genes in the orange and light-blue clusters analogously confirmed the main characteristics of the computed maps (Fig. 3c,d). These expression patterns implicate a number of genes with previously uncharacterized functions in anterior–posterior patterning and in ptc or dpp signaling.

A second way to identify interesting genes is to select genes with expression patterns that correlate or anti-correlate with genes of interest such as sens, wg or Dpp (Supplementary Tables 5–7). Among these are many genes previously implicated in the respective signaling pathways. Hence, we only performed in situ for the top genes that have not previously been characterized. In situ for genes that correlate with the neurogenic gene senseless confirmed the computed expression patterns (Fig. 4a). This implicates genes in wing neurogenesis such as Fhos, previously known to be involved in actin stress fiber formation25, ImpL3, a metabolic enzyme and Rau and cpo, which are involved in neurogenesis in other organisms26–28. The expression of CG10249/Kank, involved in attaching muscle to epidermis29, correlates with wingless (Fig. 4b) and the uncharacterized gene CG9689 correlates with Dpp (Supplementary Fig. 8).

We selected CG5151 to study in more detail, as it is functionally uncharacterized and has a human ortholog called LDLRAD4/C18orf81. The computed map predicts that CG5151 is expressed weakly along the dorsal/ventral boundary and in a proximal ring coinciding with wingless. FISH and a green fluorescent protein (GFP) transcript trap in the endogenous CG5151 locus confirmed this expression pattern, and also detected expression of CG5151 in AMPs that are not included in our computed maps (Fig. 4b,c). We tested if CG5151 is involved in Wingless or Notch signaling. Knockdown of CG5151 in the posterior wing (GFP+) caused wing notching (typical for Notch/wingless loss-of-function, Fig. 4d) and reduced Wingless expression (Fig. 4e). In sum, our mapping strategy identified a previously uncharacterized gene that has an expression pattern overlapping with wingless, and affects Wingless and Notch signaling. The human ortholog LDLRAD4 is elevated in liver cancer and promotes tumorigenesis26, where Wnt or Notch signaling may be involved.

Downsampling and algorithm performance. To test how this algorithm’s performance depends on the number of cells sequenced, we down-sampled our data by randomly selecting subsets of cells from our DropSeq data set, and used these to generate expression maps. The computed expression maps lose details and become more stochastic when <500 cells are used (Supplementary Figs. 9 and 10). Likewise, we tested the influence of nUMIs per cell by selecting from our data set the 300 cells that were most deeply sequenced (8,552 UMI per cell average), least deeply sequenced (3,169 UMI per cell) or in the middle (Supplementary Fig. 11). Map quality drops when computed from data with <8,000 UMI per cell. This can be compensated, however, by increasing cell number because our full data set of 618 cells has an average sequencing depth of 5,832 UMI per cell.

We applied our algorithm on our 10X Genomics data set. This yielded similar computed maps to the maps computed with the DropSeq data set (Supplementary Fig. 12a), hence the results can be reproduced with an independent data set and a different scRNA-seq technology.

De novo identification of ‘mapping genes’. In the wing disc the expression patterns of many genes are known, which allowed us to choose a suitable set of ‘mapping genes’ from the literature. This may not be the case for other tissues or organisms. It would be helpful to identify de novo from scRNA-seq data a suitable set of ‘mapping genes’ for which in situ need to be performed. The most informative
genes are the ones with high correlations to other genes in the data set. Hence we devised an algorithm that recursively pulls out of the data set the genes that correlate most highly with other genes and identifies them as ‘mapping genes’ (Supplementary Fig. 12b). For the wing, this yielded 65 mapping genes, including many of the ones we used to generate the maps, such as 

\[ \text{wg}, \text{Doc1, salm, hth, brk, Dll, en, ptc or ap} \] (Supplementary Table 8).

**Application to Drosophila embryo data.** We tested if our approach can be applied to other scRNA-seq data sets, such as to a previously published data set of the *Drosophila* embryo\(^1^\). To identify embryo SRGs, we plotted the average nUMI per cell for all genes versus the number of cells in which the gene was detected (Supplementary Fig. 13). We observed the same relationship as in our wing disc data set (Fig. 1c) indicating the SRG approach may be broadly applicable to scRNA-seq data sets. We computed expression maps for the embryo based on the scRNA-seq data and 85 reference genes from ref. \(^1^\) (Supplementary Fig. 14). Since the DistMap algorithm\(^1^\) yields thresholded maps, the two sets of maps look different. Nonetheless, our algorithm predicted many features of the in situ such as the ventral accumulation of CG4500, CG14688, babos and stumps, the ventral exclusion of CG32053 or the anterior–ventral accumulation of gcm. Worth noting is that the 85 reference genes used for the DistMap algorithm do not correspond to the optimal set of mapping genes needed by our algorithm for the embryo (Supplementary Table 9). For instance, gcm is one of the top mapping genes identified ‘de novo’ for which in situ would need to be done. Hence the quality of the computed
Articles

**Discussion**

To our knowledge, current methods that reconstruct tissue expression maps from scRNA-seq data do so by mapping back sequenced cells to their locations within the tissue. Here, we take a different approach by identifying genes that are co-expressed in the wing and hence likely functionally related.

This can be visualized using Cytoscape on the expression cross-correlation matrix (Fig. 4f and Supplementary Fig. 15). Genes with expression patterns linked to signaling pathways of interest such as Dpp, Wnt, Notch or EGF could be of interest for future study.

**Fig. 4 | Discovery of genes in processes of interest based on their expression pattern.** a, b. Computed expression maps and in situ hybridizations for genes correlating with either senseless (a) or wingless (b). Images representative of ten discs and two biological replicates. Scale bars, 100 µm. c. Expression of CG5151 using a GFP transcript trap in the endogenous CG5151 locus reveals expression at the dorsal/ventral boundary and in a more proximal ring, similar to that of wingless. Image representative of three discs and one biological replicate. Scale bar, 100 µm. d, e. Knockdown of CG5151 in the posterior compartment of the wing disc using hedgehog-Gal4 (hhGal4) causes notching of the posterior wing margin (in 7 of 16 wings at 25 °C, with 0 of 19 control wings showing notching), a typical Notch or wingless loss-of-function phenotype (d) and loss of Wingless protein (eight of eight knockdown discs showed reduced Wg expression at 29 °C while zero of ten control discs had reduced Wg, two biological replicates) (e). Scale bars, 500 µm for d and 100 µm for e. f. Gene-network analysis of the most connected genes linked to Dpp, Wnt, Notch and/or EGF signaling pathways. Only edges with a minimum correlation coefficient of 0.1 are shown. Self-correlations are excluded.

expression maps would likely improve by using mapping genes suited to our method.

In addition to generating gene expression maps, this work identifies genes that are co-expressed in the wing and hence likely functionally related. This can be visualized using Cytoscape on the expression cross-correlation matrix (Fig. 4f and Supplementary Fig. 15). Genes with expression patterns linked to signaling pathways of interest such as Dpp, Wnt, Notch or EGF could be of interest for future study.
approach and do not attempt to map back any of the cells in our data set. Instead, we compute tissue expression maps using gene expression correlations.

One approach for dealing with high drop-out rates is to impute gene expression. This approach works well when the number of cells sequenced is many times the number of different cell types in the tissue (that is, the cells are ‘over-sampled’). Hence every cell type is present multiple times in the data set. Clustering cells by similarity will yield clusters representing a single cell type, and missed genes can be imputed. In the wing disc, where many of the 50,000 cells are different from each other, this would require sequencing >400,000 cells. By sequencing <50,000 cells, imputation will blur cell types by averaging together cells that are actually different from each other. Hence, we did not use imputation here.

Although the calculated expression maps capture the main features of the real expression patterns, they are not perfect. For instance, rau has an expression gap in the medial domain that is not predicted, and Impl3 is less expressed ventrally than predicted (Fig. 4a). The quality and resolution of the expression maps depend on parameters that can be further refined: (1) The alignment of the ‘mapping gene’ domains to each other (left side, Fig. 2c) is nontrivial. Each map derives from an in situ on an individual wing disc with unique morphology. Furthermore, every map must be aligned to every other map, which is a problem that scales exponentially with the number of mapping genes. (2) Map qualities increase with the number of single cells sequenced and the number of mapping genes used.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0492-x.

Received: 5 September 2018; Accepted: 13 June 2019; Published online: 29 July 2019

References
1. Buchmann, A., Alber, M. & Zartman, J. J. Sizing it up: the mechanical feedback hypothesis of organ growth regulation. Semin. Cell Dev. Biol. 35C, 73–81 (2014).
2. Restrepo, S., Zartman, J. J. & Basler, K. Coordination of patterning and growth by the morphogen DPP. Curr. Biol. 24, R245–R255 (2014).
3. Worley, M. I., Setiawan, L. & Hariharan, I. K. Regeneration and tissue reconstruction in Drosophila imaginal discs. Annu. Rev. Genet. 46, 289–310 (2012).
4. Garcia-Bellido, A. The cellular and genetic bases of organ size and shape in Drosophila. Int. J. Dev. Biol. 53, 1291–1303 (2009).
5. Kornberg, T. B. & Guha, A. Understanding morphogen gradients: a problem of dispersion and containment. Curr. Opin. Genet. Dev. 17, 264–271 (2007).
6. LeGoff, L. & Lecuit, T. Mechanical forces and growth in animal tissues. Cold Spring Harb. Perspect. Biol. 8, a019232 (2015).
7. Johnston, L. A. Socializing with MYC: cell competition in development and as a model for premalignant cancer. Cold Spring Harb. Perspect. Med. 4, a014274 (2014).
8. Calleja, M., Moreno, E., Pelaz, S. & Morata, G. Visualization of gene expression in living adult Drosophila. Science 274, 252–255 (1996).
9. Ibrahimi, D. M., Beths, B., Kornberg, T. B. & Klebe, A. Microarray comparison of anterior and posterior Drosophila wing imaginal disc cells identifies novel wing genes. G3 J. 3, 1353–1362 (2013).
10. Reeves, N. & Posakony, J. W. Genetic programs activated by proneural genes in the developing Drosophila PNS. Dev. Cell 8, 413–425 (2005).
11. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. Nat. Meth. 15, 343–346 (2018).
12. Edgard, D., Johnson, P. & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. Nat. Meth. 15, 339–342 (2018).
13. D’Avie, K. et al. A single-cell transcriptome atlas of the aging drosophila brain. Cell 174, 982–998 (2018).
14. Han, X. et al. Mapping the mouse cell atlas by microwell-seq. Cell 173, 1307 (2018).

Acknowledgements
We thank the High Throughput Sequencing group of the DKFZ Genomics and Proteomics Core Facility for providing excellent next generation sequencing services. We thank J-P. Mall and the DKFZ Single-Cell Open Laboratory (scOpenLab) for assistance with the 10X Genomics experiment. We thank E. Rempel for implementing the DropSeq computational cookbook on our Galaxy Server. J.B. was supported by a research stipend from the Fritz Thyssen Foundation. P.W. was funded by a fellowship from CellNetworks—Cluster of Excellence (EXC31). Research in the laboratories of M.B. and A.A.T. is supported by ERC Grants of the European Commission.

Author contributions
J.B., P.W., S.L. and A.A.T. performed experiments. J.B., P.W., E.V., M.B. and A.A.T. analyzed data. J.B., P.W. and A.A.T. wrote the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0492-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.B. or A.A.T.

Peer review information: Nina Vogt and Tal Naway were the primary editors on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Methods

_Drosophila_ stocks. The following fly lines were used: w^{118}, CG15130 (BGDR ID 102217), CG151 MiMIC (Bloomington stock 52188). Stocks were maintained at 25 °C with a 12 h light/dark cycle, except for the crosses used in the knockdown experiments with RNAi-transgene and GAL4/UAS expression, for which crosses were maintained at 29 °C. See also the Reporting Summary.

Single cell sample preparation from wing disc tissue. Wing discs of female wandering third instar _w^{118} _larvae were dissected in Schneider's medium in batches of five animals to prevent hypoxia and transferred into a tube containing Schneider's medium (Thermo Fisher Scientific) on ice for a maximum time of 30 min. The isolated wing discs were rinsed once with Schneider's medium and then incubated for 15 min in a water bath at 37 °C in TrypLE Select Enzyme (10x, Thermo Fisher Scientific), with gentle mixing every 5 min. Schneider's medium was then added to the tissue pellets, and the dissociation was performed using a P1000 pipette. The cell suspension was then passed through a 10 μM cell strainer to remove undigested tissue and cell clumps. Cells were manually counted using a plastic hemocytometer (C-Chip N01). The entire cell isolation protocol was done with PBS-Triton (0.1%) coated microcentrifuge tubes and tips to minimize cell loss.

scRNA-seq by Drop-seq technology. Drop-seq experiments were performed as published in ref. 31 following the detailed online protocol (Drop-seq-Protocol v.1.0-May 2015, mccarrollab.org/download/905/). In brief, cells and barcoded beads (ChemeGene) were co-flown in an Agarose coated microfluidics device (Flowjet) and co-assembled in monodisperse droplets with a 0.6 ratio of AMPure beads (Agencourt) and eluted in the necessary amount of water to obtain 400–1,000 pg μM cell strainer to remove undigested tissue and cell clumps. Cells were manually counted using a plastic hemocytometer (C-Chip N01). The entire cell isolation protocol was done with PBS-Triton (0.1%) coated microcentrifuge tubes and tips to minimize cell loss.

scRNA-seq by 10X technology. The 10X experiments were performed using the GemCode Single-Cell Instrument, Single Cell 3’ Library and Gel Bead Kit v2 and Single Cell A Chip Kit (10X Genomics) following the manufacturer’s protocol (CG00052, SingleCell3_ReagentKits2UserGuide_RevD). In brief, the single cell suspension was resuspended in PBS and about 9,000 cells were loaded in one lane of the chip. Nanoliter-scale gel bead-in-emulsions (GEMs) were generated, mRNA reverse transcription was performed using the 10X transcript amplification protocol. The final library was PCR-amplified for 14 cycles and showed an average size of about 500 bp. Paired-end sequencing was carried out with the Illumina HiSeq2500 instruments at the DKFZ Genomics and Proteomics Core Facility (Heidelberg, Germany).

Preprocessing of scRNA data. For Drop-seq data, paired-end sequence reads were processed as described in ref. 31. The available R command lines were implemented in our in-house Galaxy server (http://galaxy.b110.dkfz.de/galaxy/) following the default settings described in detail in the Drop-seq computational cookbook v1.2 (mccarrollab.org/wp-content/uploads/2016/03/Drop-seqAlignmentCookbook1Jan2016.pdf). The reads were aligned to the _Drosophila_ reference genome (BGDP6 v.87 (GCA 000002155.4)) using STAR v2.5.2b-0 with the default parameters. The cell number was estimated by plotting the cumulative fraction of reads per cell against the sorted cell barcodes (decreasing number of reads) and determining the point of inflection. The raw digital gene expression matrices were generated for the two batches. The 10X Genomics data were analyzed using the Cell Ranger Pipeline v2.2. The reads were aligned using STAR to the _Drosophila_ reference genome (BGDP6 v.87 (GCA 000002155.4)). The estimated cell number was derived by plotting the UMIs counts against the barcodes and revealed 2,554 cells used for downstream analysis. Further filtering of the expression matrices was done to ensure high-quality single cell data. By using the Seurat R package, we selected cells with low expression of mitochondrial encoded genes (<20%), high alignment rate (>85%) and a minimum number of detected genes (>200). Note, of Digital gene expression matrix from the 10X Genomics experiment did not contain any mitochondrial encoded genes. For the Drop-seq data, outlier cells (>3,000 detected UMIs), which could be potential cell doublets, were also excluded from further analysis. After subsampling the wing disc cell population, we also applied a reasonable UMI cutoff (>2,000). The UMI cutoff was empirically determined by performing correlation analysis with genes of known expression patterns. Using cells with at least 2,000 detected UMIs showed the expected correlation coefficient values among our reference set. Additionally, we also removed genes that were detected in only one cell. This filtering resulted in 615 high-quality wing disc single cells, which were subsequently merged together in a single Digital gene expression matrix. Before principal component analysis (PCA) and clustering, the data were log transformed (log, ) and re-scaled by multiplying by 10,000.

Identification of SRGs. SRGs were identified by analyzing the nUMIs, as this led to the smallest spread in the data. A scatter plot was generated for all detected genes whereby the x axis is the number of cells in which the gene was detected (nUMI >0) and the y axis is the average nUMI for that gene in the cells in which it was detected (that is, not across the entire cell population, since this also contains cells not expressing the gene). A linear model was then applied and adjusted to fit the data, and residuals were calculated for each gene relative to the linear model. The average and standard deviation of the residuals was calculated, and SRGs were defined as genes with residuals (<mean – 1 s.d.).

Batch correction, PCA and clustering. For cluster identification we applied the Seurat R package v2.3.4 package17 and followed largely the tutorial instructions from the Seurat website at http://satijalab.org/seurat/. To reduce dimensionality, PCA was run on the entire transcriptome after scaling and centering the data and removing technical confounder factors (number of UMIs, number of genes and alignment rate). Wees omninialtions were plotted to detect any remaining confounding effects by (1) plotting the t-SNE plots for evenly distributed batches, number of genes and transcripts, and fraction of mitochondrial RNA among the clusters (Supplementary Fig. 5), (2) analyzing the loading of genes (‘PC loading’) of the different batches for their similarity and (3) comparing the inter-batch correlations (Supplementary Fig. 1d). The jack straw statistical analysis (num.mix = 20, num.replicate = 1,000, prop. freq. = 0.01) and plotting the eigenvalues in decreasing order (‘Elbow plot’), was used to select PCs as input for clustering. We used t-SNE for visual representation of the clusters and highlighting marker gene expression. Two distinct clusters of cells were identified by means of the above described procedure, one of which was strongly defined by expression of AMP marker genes. As this cell type was irrelevant for the present study, we excluded it based on the t-SNE plot from further analysis. After AMI’s were removed, the list of SRGs derived specifically for the wing disc cell population was used for dimensional reduction. Selection of principal components and clustering was again performed as described above. The proportion of wing disc cells within each cluster was found to be similarly represented in both batches underlying the robustness of the identified clusters.

Generating bulk mRNA-seq data. Wing discs from 50 female wandering third instar _w^{118} _larvae were dissected in Schneider's medium, five larvae at a time to prevent hypoxia and transferred to a tube containing Schneider’s medium on ice. The wing discs were then lysed in TRizol (Thermo Fisher) for total RNA isolation following the manufacturer’s protocol. RNA library preparation (TruSeq Stranded mRNA Sample Preparation Kit, Illumina) and sequencing (50 nucleotide single-end reads, HiSeq2500, Illumina) were done at the DKFZ Genomics and Proteomics Core Facility following the manufacturers’ protocols. Fastq sequencing data was processed using the scater R package (v.1.10.1)17 by applying the default settings.

Comparison of single cell and bulk transcriptomic profiles. To compare gene expression data at single cell and bulk levels, we calculated Pearson’s correlation coefficients (R) of gene expression for all possible gene pairs across the cells. Only genes detected in both sets were used for comparison. For single cell data, the average UMI expression for each gene was first calculated and then converted to average transcripts per million (ATPM). Gene counts were converted to transcripts per million (TPM) and isofrom counts averaged. The log-transformed data was plotted (1 + ATPM and 1 + TPM, respectively).

Calculation of gene expression correlations. For gene expression correlation analysis, only cells with UMI >2,000 were considered since, of all the reads per cell reduced the correlation coefficients. Gene expression correlations across cells were calculated using the Pearson’s correlation coefficient, except that the one single cell contributing most strongly to the correlation coefficient was removed to avoid outliers from influencing the correlation. Specifically, for genes _x_ and _y_, where the nUMI for each gene in _x_ are _x_ and _y_ respectively, the mean _x_ and _y_ across cells were calculated. Then, for every cell, the numerator of the Pearson’s coefficient _a_ = ( _x_ − _x_)( _y_ − _y_). The cell with the maximum _a_ was excluded, and the Pearson’s correlation coefficient was calculated for all other cells.

Generation of gene clustering dendrogram. The gene clustering dendrogram was generated by first generating a cell to cell correlation table for all SRGs using the correlation function described in the previous section, and then clustering and plotting using the R hclust() function.

Generation of computed wing disc maps. To generate computed gene expression maps, we first performed in situ hybridizations of reference genes with published
expression patterns, which we term 'mapping genes.' Genes for which we could not confirm the expression pattern were discarded, yielding a list of 58 confirmed mapping genes (for the list see Supplementary Table 4). For each mapping gene, we then selected one representative image, either from our in situ or from the published literature, depending on which had better signal. These images were morphed in Photoshope using the 'Puppet Warp' function to fit one reference wing disc shape, and then the images for all 58 mapping genes were aligned to each other. Images were then thresholded using ImageJ to obtain binary images, thereby defining an expression domain for each mapping gene. Computed expression maps were then calculated as follows. We call \( m_1, m_2, \ldots, m_n \) the 58 mapping genes, \( \epsilon(x, c) \) the gene of unknown expression pattern for which the expression map is being computed and \( \{c_1, c_2, \ldots, c_{50,000}\} \) the 50,000 cells in the wing disc. Following the calculation described in the section above, a modified version of the Pearson's correlation that excludes one outlier was calculated for gene \( x \) relative to each of the 58 mapping genes, yielding 58 correlation coefficients \( \{r_{x,1}, r_{x,2}, \ldots, r_{x,58}\} \). Computation of the expression map consists of determining an expression level \( e \) for gene \( x \) in each cell \( c \):

\[
e(x, c) = \sum_{j=1}^{58} r_{x,j} x a(m_j, c)
\]

where the parameter \( a(m_j, c) \) is equal to +1 if the mapping gene \( m_j \) is expressed in cell \( c \), and it equals -1 if it is not. This essentially sums together all the correlation coefficients of gene \( x \) relative to the 58 mapping genes with a weighting factor of \( \pm 1 \) depending on whether that mapping gene is expressed in that cell or not.

Immunostainings. Immunostainings of wandering third instar wing discs were performed as previously described, using monoclonal mouse anti-Wingless (clone 4D4, 1:50, Developmental Studies Hybridoma Bank). Secondary antibody staining was performed using fluorescently labeled antibodies at a dilution of 1:500, together with Hoechst 33342 (1:2,000, Invitrogen) nucleic acid staining. The specimens were mounted in Vectashield mounting medium (Vector Laboratories) and imaged with a Leica TCS SP8 confocal microscope (Leica). Images were analyzed and processed in ImageJ 2.0.0-rc-59.

FISH. In situ probes with lengths of 250–500 nucleotides were designed to detect all transcript variants of the gene of interest. A DNA template containing a T7 promoter sequence (included in the reverse primer oligonucleotide) was generated by PCR from cDNA and used to generate digoxigenin-labeled RNA probes by in vitro transcription reaction using the DIG RNA labeling Kit (Roche). The DNA template was removed by DNase I digest and the in situ probe purified from the in vitro transcription reaction using the DIG RNA labeling Kit (Roche) and the DIG RNA Clean-up, Macherey–Nagel. The purified in situ probe was performed overnight at 4 °C in the respective blocking solution containing pre-absorbed anti-digoxigenin Fab fragments conjugated to horseradish peroxidase (Roche) (1:1,000). Unbound Fab fragments were removed by rinsing three times with PBST and washed twice with PBST 10 min before staining cell nuclei with DAPI (1:2,000 in PBST) for 15 min. The imaging of the antibody to the in situ probe was performed overnight at 4 °C in the respective blocking solution containing pre-absorbed anti-digoxigenin Fab fragments conjugated to horseradish peroxidase (Roche) (1:1,000). Unbound Fab fragments were removed by rinsing three times with PBST and washed twice with PBST for 10 min. Binding of the antibody to the in situ probe was performed overnight at 4 °C in the respective blocking solution containing pre-absorbed anti-digoxigenin Fab fragments conjugated to horseradish peroxidase (Roche) (1:1,000). Unbound Fab fragments were removed by rinsing three times with PBST and washed twice with PBST for 10 min.

Network analysis. For constructing a network and detecting modules, the cross-correlation matrix between core signaling pathway components in the wing imaginal disc (Dad, sens, aos, dpp, wg) and the set of SRGs was calculated. A correlation coefficient cutoff of >0.1 was applied to the network, self-correlations were excluded. Visualization of the network was done using the Cytoscape software v.3.6.1 (ref. 30).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. Data generated or analyzed during this study are included in this published article (and its Supplementary Information files). Raw sequencing data have been deposited to NCBI GEO with accession number GSE127832.

Code availability. All custom code, together with sample data, are freely available on the Boutros laboratory repository of Github [https://github.com/boutroslab/Supplemental-Material/tree/master/Bageritz_2019]. A markdown document describing how to identify SRGs using R is provided. The software package is composed of four pieces of software: (1) 1_cross_correlation_all_genes calculates the cross-correlation of all genes against all genes provided as input in an expression matrix. (2) 2_identify_best_mapping_genes recursively identifies the gene with the highest correlation score in the cross-correlation matrix generated by software no. 1, and pulls it out as a mapping gene. See Supplementary Fig. 12b for a schematic diagram. (3) 3_cross_correlation_to_mapping_genes calculates the cross-correlation between all genes and the SRGs, taking as input an expression matrix. (4) 4_calculate_expression_maps calculates the expression maps following the algorithm described above in the section Generation of computed wing disc maps.

References. 32. McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scafer: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformetrics 33, 1179–1186 (2017).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Clearly defined error bars
  - State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

The software running the Leica confocal microscope was Leica Application Suite X (LAS X) version 2.0.1.14392. Image acquisition: SPOT software v4.7.0.12. The sequencer runs RTA v2 software.

Data analysis

Custom software is provided on Github. All other software used in this study is publicly or commercially available: Images were analyzed with ImageJ. Network analysis was done with Cytoscape software v3.6.1. Dendograms were made with the R hclust() function. Cluster identification was done with the Seurat R package. Software for sequence analysis: Drop-seq_tools v1.12, M3Drop/NBDrop v3.10.4, DistMap v0.1.0, STAR v2.5.2b-0, R v3.5.2, R Studio v1.1.463, Cell Ranger v2.2, scater R package v1.10.1, bcl2fastq Conversion Software v2.18.0.12, v2.19.0.316, v2.20.0.422. Figure preparation: Adobe Photoshop CS6 version (13.0 x64).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated or analysed during this study are included in this published article (and its supplementary information files).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
No sample size calculations were performed. All in situ studies were performed on at least 10 animals, which from experience is sufficient to assure that the selected examples are representative.

Data exclusions
No exclusion criteria were pre-established. As described in the Methods section, for the Drop-Seq data we selected cells with low expression of mitochondrial encoded genes (<5%), high alignment rate (>85%) and a minimum number of detected genes (>200). Outlier cells (>3,000 detected UMIs), which could be potential cell doublets, were excluded from further analysis. We removed Adult Muscle Precursor cells from the analysis, because they were not the focus of this study. We excluded cells falling below a reasonable UMI cutoff (<2,000) which was empirically determined by performing correlation analysis with genes of known expression patterns. Using cells with at least 2,000 detected UMIs showed the expected correlation coefficient values among our reference set.

Replication
Except where noted in the figure legends, experiments were replicated at least 2 times. Only data that reproduced are included in this manuscript.

Randomization
All animals were randomly allotted.

Blinding
Blinding is not relevant to this study because no treatments were performed that could be biased, and the readouts (e.g. notching in Fig 4d) are objective.

Reporting for specific materials, systems and methods

Materials & experimental systems

Methods

Antibodies

As described in the Methods section: Immunostainings of wandering 3rd instar wing discs were performed as previously described (3), using monoclonal mouse anti-Wingless (clone 4D4, 1:50, Developmental Studies Hybridoma Bank). Secondary antibody staining was performed using fluorescently labeled antibodies at a dilution of 1:500 (Goat anti-Mouse IgG (H+L) Highly Cross-Adsorbed Secondary Antibody, Alexa Fluor 546, catalog number A-11030, lot number 1661231), together with Hoechst 33342 (1:2,000, InvitrogenTM) nucleic acid staining.
Validation
Both the anti-wg and the anti-senseless antibodies have been extensively used in the field and have been validated in the following publications:
- Antagonistic Interactions Between Wingless and Decapentaplegic Responsible for Dorsal-Ventral Pattern in the Drosophila Leg” (William J. Brook, Stephen M. Cohen; Pubmed ID: 8703069)
- Inducible progenitor-derived Wingless regulates adult midgut regeneration in Drosophila” (Cordero et al, 2012; Pubmed ID 22948071)
- Active Wnt proteins are secreted on exosomes” (Gross et al. 2012; Pubmed ID 22983114)

Animals and other organisms
Policy information about studies involving animals: ARRIVE guidelines recommended for reporting animal research

| Laboratory animals                      | Drosophila melanogaster, strain w[1118], female wandering 3rd instar larvae |
|----------------------------------------|--------------------------------------------------------------------------------|
| Wild animals                           | This study did not involve wild animals.                                      |
| Field-collected samples                | This study did not involve samples collected from the field.                |