Published in final edited form as:
Nat Methods. 2019 June 17; 16(8): 750–756. doi:10.1038/s41592-019-0492-x.

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

Josephine Bageritz[1,2], **Philipp Willnow**[3,1,2], **Erica Valentini**[1,2], **Svenja Leible**[1,2], **Michael Boutros**[1,2,*], **Aurelio A. Teleman**[1,2,*]

[1]German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

[2]Heidelberg University, 69120 Heidelberg, Germany

[3]CellNetworks - Cluster of Excellence, 69120 Heidelberg University

Abstract

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 complete gene expression atlas in the wing disc, we employed single-cell sequencing (scRNA-seq) and developed a new 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 both known and new 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. This novel method will enable the leveraging of scRNA-seq data for generating expression atlases of undifferentiated tissues during development.

*correspondence: m.boutros@dkfz.de & a.teleman@dkfz.de, tel: +49 6221 42-1620, fax: +49 6221 42-1629.

Data Availability

Data generated or analysed 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 lab repository of Github. 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 #1, and pulls it out as a mapping gene. Please see Suppl. 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”.

Author Contributions

JB, PW, EV, SL and AAT performed experiments. JB, PW, EV, MB and AAT analyzed data. JB, PW, MB and AAT wrote the manuscript.

Competing Financial Interests Statement

The authors declare they have no competing interests.

Europe PMC Funders Author Manuscripts

Keywords

Drosophila; development; single cell RNA sequencing

Introduction

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 biophysics1–8. Despite this, the expression patterns of the vast majority of 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 regions9, 10. Methods analogous to those used for differential gene expression analysis across treatments or timepoints are then used for identifying significant patterns in such spatial transcriptomics data11, 12. Recent advances in single-cell sequencing (scSeq) allow the clustering of cells into cell types based on their expression profiles13–16, however clustering per se does not have 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 genes17–20. Finally, data from scSeq and fluorescence *in situ* hybridization (FISH) can be combined to first cluster cells by cell type and subsequently extract spatial information 21. The wing disc presents several challenges. Firstly, it is composed mainly of pluripotent, undifferentiated stem-like cells, hence it consists of few cell types. Secondly, with 50,000 cells, the wing disc has 10x more cells than, for instance, the *Drosophila* embryo18. Most mammalian organs have even more cells, representing an even larger challenge. We present here a scSeq approach 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

To construct a gene expression map of the *Drosophila* wing disc, we dissociated cells from wing discs of 3rd instar female larvae and sequenced their mRNA. With DropSeq22 we sequenced 1,644 cells with median depth of 3,774 transcripts and 1,134 genes per cell (Suppl. Fig. 1a-b). With 10X Genomics we sequenced 2,554 cells with median depth of 10,620 transcripts and 1,998 genes per cell (Suppl. Fig. 2a). Since we will discuss limitations of these datasets, we point out that our sequencing depths are in line with what others have also reported (Suppl. Table 1)13, 15, 18. We could unambiguously identify true cell barcodes (Suppl. Fig. 1c) indicating low ambient mRNA from cell breakage during sample preparation. Two biological replicates correlated highly to each other (r=0.93, Suppl. Fig. 1d) indicating reproducibility. The sum of all single-cell reads correlated well with RNA-seq data of non-dissociated wing discs (Suppl. Fig. 1e), suggesting scSeq captured most of the genes and did not strongly alter gene expression.

Europe PMC Funders Author Manuscripts

To identify cell types, we clustered cells using the Seurat R package17 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.

To identify genes with spatially restricted expression patterns (“Spatially Restricted Genes”, SRGs), we plotted for every gene the number of cells in which it was detected versus the average expression level in those expressing cells (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, left panel), identified by residuals >1 standard deviation below the mean on the inverse graph (green points, right panel), yielding a set of 824 SRGs (Suppl. Table 2). As a benchmark, we compiled a list of 68 genes from literature that are expressed in restricted domains of the wing (Suppl. Table 3). The SRGs include 63 of these (magenta dots in Fig. 1c-d). Missing are *aos*, *arr*, *Dtg*, *fng* and *Mes2*. Two (*arr* and *fng)* are identified as SRGs in the 10X Genomics dataset. In comparison, a similarly sized set of 829 “Highly Variable Genes” (HVGs) identified using the Seurat package17 contained 17 benchmark genes (Fig. 1d, Suppl. Fig. 3). This is likely because HVGs were designed for a different purpose - to identify genes with variable expression. Genes expressed ubiquitously but at varying levels are identified as HVGs, although they may not be spatially restricted. Our SRG analysis is more related to M3Drop/NBDrop23, which identifies non-ubiquitous genes by analyzing drop-out rates. To identify benchmark genes (Suppl. Table 3) with a false-negative rate <10%, the SRG algorithm generates a list of 824 genes and NBDrop generates a list of 1000 genes (Suppl. 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 (Suppl. 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)/cell, read alignment rate, fraction of mitochondrial RNA or representation of the two biological replicates (Suppl. Fig. 5).

We first tested whether we could determine the location in the wing of the sequenced cells 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 (Suppl. 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 ~35% of wing disc cells should express *engrailed* and the other ~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/cell threshold did not solve this: with 12,000 UMIs/cell, 84% of cells were en^+ or ci^+, with only 45 of the 948 sequenced cells passing this threshold (Fig.

2a). (2.5% of the cells were *en*$^{+}$/*ci*$^{+}$ double-positive, suggesting they are doublets). Using our 10X Genomics dataset, which has more sequencing depth (Suppl. Table 1) we reached 90% confidence on the cell location in the *en*$^{+}$ or *ci*$^{+}$ domain with a 45,000 UMI/cell threshold, which leaves only 140 cells in the dataset (Suppl. 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 scSeq data. In line with this, we tested methods which successfully reconstructed spatial gene expression from scSeq data in zebrafish and the *Drosophila* embryo17, 18, but these methods did not work well for our dataset (Suppl. Fig. 7). We therefore searched for an alternate method.

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 2-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 which 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 expression domain of Gene 1 with known expression pattern (red dot, Fig. 2d), uncharacterized Gene 2 is likely also expressed 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 literature to have distinct expression patterns which we overlayed, aligned, and thresholded, yielding binary expression domains (Fig. 2e, Suppl. Table 4). We then calculated a cross-correlation matrix between these 58 mapping genes and all genes in the genome. 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 tested these resulting 'computed expression maps' by comparing them to fluorescent *in situ* hybridizations (FISH) (Fig. 3-4).

This method generates computed expression maps for all genes detected in the sequencing. We present 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 against each other, and then hierarchically clustered them (Fig. 3a, Suppl. Data 1). Visual inspection of this dendrogram confirmed that neighboring genes have similar expression patterns (e.g. *en*/*inv*/*hh*, *hth*/*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 (Suppl. Fig. 6a), and that Trim924 is expressed predominantly in the pouch with an inverse venation pattern and inhibition at the D/V boundary. The FISH for *CG7201* had some elements of the predicted map, such as higher expression medially and broad repression at the D/V boundary, but also differed somewhat 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). Interestingly, 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* (Suppl. Tables 5-7). Amongst these are many genes previously implicated in the respective signaling pathways. Hence, we only performed *in situs* for the top genes that have not previously been characterized. *In situs* for genes that correlate with the neurogenic gene senseless confirmed the computed expression patterns (Fig. 4a). This implicates novel genes in wing neurogenesis, such as *Fhos,* involved in actin stress fiber formation25, *ImpL3,* a metabolic enzyme, and *Rau* and *cpo,* which are involved in neurogenesis in other organs26–28. The expression of *CG10249*/*Kank*, involved in attaching muscle to epidermis29, correlates with *wingless* (Fig. 4b) and the uncharacterized gene *CG9689* correlates with *Dpp* (Suppl. Fig. 8).

We selected *CG5151* to study in more detail, as it is functionally uncharacterized and has a human ortholog *LDLRAD4*/*C18ORF1*. The computed map predicts *CG5151* is expressed weakly along the dorsal/ventral boundary and in a proximal ring coinciding with wingless. FISH and a GFP transcript trap in the endogenous *CG5151* locus confirmed this expression pattern, and also detected expression of *CG5151* in AMPs which 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 novel uncharacterized gene that has an expression pattern overlapping with wingless, and affects wingless/notch signaling. Interestingly, the human ortholog *LDLRAD4* is elevated in liver cancer and promotes tumorigenesis30, where Wnt or Notch signaling may be involved.

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 dataset, and used these to generate expression maps. The computed expression maps lose details and become more stochastic when <500 cells are used (Suppl. Figs. 9-10). Likewise, we tested the influence of nUMIs/cell by selecting from our dataset the 300 cells that were most deeply sequenced (8,552 UMIs/cell average), least deeply sequenced (3,169 UMIs/cell), or in the middle (Suppl. Fig. 11). Map quality drops when computed from data with <8,000 UMIs/cell. This can be compensated, however, by increasing cell number because our full dataset of 618 cells has an average sequencing depth of 5,832 UMIs/cell.

We next tested the algorithm on our 10X Genomics dataset. This yielded very similar computed maps (Suppl. Fig. 12a), hence the results can be reproduced with an independent dataset and a different scSeq technology.

In the wing disc the expression patterns of many genes are known, hence we could 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 scSeq data a suitable set of 'mapping genes' for which *in situs* need to be performed. The most informative genes are the ones with high correlations to other genes in the dataset. Hence we devised an algorithm which recursively pulls out of the dataset the genes which correlate most highly with other genes and identifies them as 'mapping genes' (Suppl. Fig. 12b). For the wing this yielded 65 mapping genes, including many of the ones we used to generate the maps, such as *wg, Doc1, salm, hth, brk, Dll, en, ptc*, or *ap* (Suppl. Table 8).

We tested if our approach can be applied to other scSeq datasets, such as of the *Drosophila* embryo18. To identify embryo SRGs, we plotted the average nUMI/cell for all genes versus the number of cells in which the gene was detected (Suppl. Fig. 13). We observed the same relationship as in our dataset (Fig. 1c) indicating this may be broadly applicable to scSeq datasets. We computed expression maps for the embryo based on the scSeq data and 85 reference genes from 18 (Suppl. Fig. 14). Since the DistMap algorithm18 yields thresholded maps, the two sets of maps look somewhat different. Nonetheless, our algorithm predicted many features of the in situs such as the ventral accumulation of *CG4500*, *CG14688*, *babos* and *stumps*, the ventral exclusion of *CG34224*, the posterior accumulation 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 (Suppl. Table 9). For instance, *gcm* is one of the top mapping genes identified 'de novo' for which in situs would need to be done. Hence the quality of the computed expression maps would 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 Cytoscape31 on the expression cross-correlation matrix (Fig. 4f, Suppl. 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.

Discussion

To our knowledge, current methods that reconstruct tissue expression maps from scSeq data do so by mapping back sequenced cells to their locations within the tissue. Here we take a different approach, and do not attempt to map back any of the cells in our dataset. 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 (ie the cells are 'over-sampled'). Hence every cell type is present multiple times in the dataset. 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 which is not predicted, and *ImpL3* is less expressed ventrally than predicted (Fig. 4a). The quality and resolution of the expression maps depend on parameters which can be further refined: 1) The alignment of the 'mapping gene' maps to each other (left side, Fig. 2e) is non-trivial. 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 Methods

Drosophila stocks

The following fly lines were used: w^{1118}, $CG5151^{RNAi}$ (VDRC ID 102217), CG5151 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 and GAL4/UAS expression, for which crosses were maintained at 29 °C. Also see "Life Sciences Reporting Summary".

Single cell sample preparation from wing disc tissue

Wing discs of female wandering 3rd instar w^{1118} larvae were dissected in Schneider's medium in batches of 5 animals (to prevent hypoxia) and transferred into a tube containing Schneider's medium on ice for a maximum time of 30 minutes. The isolated wing discs were rinsed once with Schneider's medium and then incubated for 15 minutes in a water bath at 37°C in TrypLE (10x), with gentle mixing every 5 minutes. Schneider's medium was then added to the loosened tissue pellets, followed by gentle mechanical dissociation 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 22 following the detailed online protocol (Drop-seq-Protocol-v1.0-May-2015). In brief, cells and barcoded beads (ChemeGene) were co-flown in an Aquapel coated microfluidics device (FlowJem) and co-encapsulated in aqueous droplets for a maximum period of 15 minutes. Isolated wing disc cells were loaded without further dilution at a concentration found by species mixing experiments to contain a maximum of 3% cell doublets. The aqueous flow rates were adjusted to ensure stable production of monodispers droplets. The size of the droplets was controlled by the oil flow rate. For this project, settings were chosen to generate about 120 μM droplets for batch 1 and 85 μM droplets for batch 2. While the standard barcoded beads were used for batch 1, batch 2 was performed with filtered beads (< 40 μm in diameter) to account for the smaller droplet size. High quality emulsions were broken by perfluorooctanol and reverse transcription of captured mRNA was started immediately after. Subsequently, barcoded beads were incubated with Exonuclease I to remove excess primers, and cDNA was then amplified from 2000 beads per reaction (12-14 PCR cycles). Up to 10 reactions were pooled, purified with a 0.6 ratio of AMPure beads (Agencourt) and eluted in the necessary amount of water to obtain 400-1,000pg/ul of cDNA. Final libraries were prepared using the Illumina Nextera XT kit and 1 ng of amplified cDNA as input. The average size of sequenced libraries was between 700 and 800 bp. Paired-end sequencing was carried out with the Illumina HiSeq2500 instruments at the DKFZ Genomics and Proteomics Core Facility (Heidelberg, Germany).

scRNA-seq by 10x technology

10x experiments were performed using the GemCode Single-Cell Instrument, Single Cell 3' Library & Gel Bead Kit v2 and Single Cell A Chip Kit (10x Genomics, Pleasanton, CA, USA) following the manufacturer's protocol (CG00052_SingleCell3_ReagentKitv2UserGuide_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 transcribed and cDNA amplified using 10 PCR cycles. 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 NextSeq 550 instruments at the DKFZ Genomics and Proteomics Core Facility (Heidelberg, Germany).

Preprocessing of scRNA data

For DropSeq data, paired-end sequence reads were processed as described 22. The available R command lines were implemented in our in-house Galaxy1 server (http://galaxy-b110.dkfz.de/galaxy/) following the default settings described in detail in the Drop-seq computational cookbook v1.2. The reads were aligned to the *Drosophila* reference genome (BDGP6 version 87 (GCA 000001215.4)) using STAR 2.5.2b-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. 10x Genomics data were analyzed using the Cell Ranger Pipeline v2.2. The reads were aligned using STAR to the *Drosophila* reference genome (BDGP6 version 87 (GCA 000001215.4)). The estimated cell number was derived by plotting the UMI 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 package2, we selected cells with low expression of mitochondrial encoded genes (<5%), high alignment rate (>85%) and a minimum number of detected genes (>200). Of note, the DGE 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 subsetting 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 1 cell. This filtering resulted in 615 high-quality wing disc single cells, which were subsequently merged together in a single DGE matrix. Prior to Principle component analysis and clustering, the data were log transformed (log +1) and re-scaled by multiplying by 10,000.

Identification of spatially-restricted genes

Spatially Restricted Genes (SRGs) were identified by analyzing the nUMIs, as this led 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 (i.e. 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 std dev).

Batch Correction, Principle component analysis and clustering

For cell/cluster identification we applied the Seurat R package v 2.3.4 package 17 and followed largely the tutorial instructions from the Seurat website http://satijalab.org/seurat/. In order to reduce dimensionality, principle components analysis (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). We examined the effectiveness of removing confounder effects by *i.)* inspecting the t-SNE plots for evenly distributed batches/libraries, number of genes and transcripts, and fraction of mitochondrial RNA among the clusters (Suppl. Fig. 5), *ii.)* analyzing the loading of genes (“PC loading”) of the different batches/libraries for their similarity and *iii.)* comparing the inter-batch correlations (Suppl. Fig. 1D). The jack straw statistical analysis [num.pc = 20, num.replicate = 1,000, prop.freq = 0.01] and plotting the eigenvalues in decreasing orders (‘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 adult muscle precursor (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 AMPs were removed, the list of spatially-restricted genes (SRGs) derived specifically for the wing disc cell population was used for dimensional reduction. Selection of principle 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 3rd instar w^{1118} larvae were dissected in Schneider's medium, 5 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 Fischer) for total RNA isolation following manufacturer's protocol. RNA library preparation (TruSeq Stranded mRNA Sample Preparation Kit, Illumina) and sequencing (50 nt single-end reads, HiSeq2500, Illumina) were done at the DKFZ Genomics and Proteomics Core Facility (Heidelberg, Germany) following the manufactures' protocol. Fastq sequencing data was processed using the scater R package (v1.10.1)32 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 TPM (Transcripts per million) and isoform counts averaged. Log-transformed data was plotted (1 + ATPM/TPM).

Calculation of gene expression correlations

For gene expression correlation analysis, only cells with nUMI>2,000 were considered, since cells with fewer 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 cell i are x_i and y_i respectively, the means $\bar{x}$ and $\bar{y}$ across cells were calculated. Then for every cell, the numerator of the Pearson's coefficient $a_i = (x_i - \bar{x}) \cdot (y_i - \bar{y})$ was calculated. The cell with the maximum a_i 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 computing a cross-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 list see Suppl. Table 4). For each mapping gene, we then selected one representative image, either from our *in situs* or from the published literature, depending on which had better signal. These images were morphed in Photoshop 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 $\{m_1, m_2, \ldots, m_{58}\}$ the 58 mapping genes, x 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 which 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_i:

$$e(x, c_i) = \sum_{j=1}^{58} r_{x,j} \cdot a(m_j, c_i)$$

where the parameter $a(m_j,c_i)$ is equal to +1 if the mapping gene m_j is expressed in cell c_i, 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 ±1 depending on whether that mapping gene is expressed in that cell or not.

Immunostainings

Immunostainings of wandering 3rd instar wing discs were performed as previously described3, 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.

Fluorescent in situ hybridization

In situ probes with lengths of 250 to 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 reaction solution (RNA Clean-up, Macherey-Nagel). The purified *in situ* probe was stored in 50% formamide at -20°C until further use. Sequences of oligos used to generate the probes are provided in Suppl. Table 10.

For fluorescent *in situ* hybridization, wandering 3rd instar w^{1118} larvae were dissected and fixed in 4% paraformaldehyde for 30 min. Fixed larvae were washed 3 times in phosphate-buffered saline (PBS) containing 0.1% Tween 20 (PBT) for 10 min before dehydration in methanol/PBT (1:1) for 5 min and rinsing in methanol. Next, the larvae were incubated in methanol/PBT (1:1) for 5 min and fixed again in 4% paraformaldehyde containing 0.1% Tween 20 for 20 min. Following three washing steps with PBT for 5 min each, the sample buffer was changed to hybridization solution (HS) (50% formamide, 5x SSC (0.75 M sodium chloride and 75 mM sodium citrate dehydrate), 50 μg/ml heparin, and 0.1% Tween 20) by serial washing steps of 5 min each in HS/PBT dilutions of 30/70, 50/50, and 70/30 (vol/vol). Before hybridization, the larvae were washed for 5 min and 10 min in HS, followed by blocking for 2 h at 65°C in HS supplemented with 100 μg/ml salmon sperm DNA (AppliChem). Hybridization was performed overnight at 65°C with an *in situ* probe concentration of 1.5 ng/ml. The probe was denatured at 80°C for 5 min and cooled on ice prior to adding to the tissues. The following day, the larvae were washed with HS for 5 min and 15 min before changing the washing buffer to PBT through serial washing steps of 5 min each in HS/PBT dilutions of 70/30, 50/50, and 30/70 (vol/vol). Next, the larvae were rinsed and washed three times with PBT for 15 min and blocked in either PBT containing 5% (w/vol) bovine serum albumin or in maleic acid buffer (1M maleic acid, 1.5M NaCl; pH 7.5) supplemented with 0.5% (w/vol) blocking reagent for nucleic acid hybridization and detection (Roche) for 30 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 PBT and washing with PBT for 10 min before staining cell nuclei with DAPI (1:2,000 in PBT) for 15 min. After removal of residual DAPI by washing with PBT for 10 min, localization of the *in situ* probe was visualized using the TSA Plus Fluorescein Kit (PerkinElmer) according to the manufacturer's protocol. For this, the larvae were incubated with the TSA working solution for 3 min or 7 min when blocked with bovine serum albumin or with blocking reagent for nucleic acid hybridization and detection, respectively. Lastly, larvae were rinsed and washed twice with PBT for 10 min. Wing imaginal discs were mounted and analyzed by confocal microscopy. Projections of wing discs shown in this study were generated using the 'sum projection' function in ImageJ.

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 v3.6.1 31.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the High Throughput Sequencing group of the DKFZ Genomics & Proteomics Core Facility for providing excellent next-generation sequencing services. We thank Jan-Philipp Mallm and the DKFZ Single-Cell Open Lab (scOpenLab) for assistance with the 10x Genomics experiment. We thank Eugen 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. as funded by a fellowship from CellNetworks - Cluster of Excellence (EXC81). Research in the labs of M.B. and A.A.T. is supported by ERC Grants of the European Commission.

References

1. Buchmann A, Alber M, Zartman JJ. Sizing it up: The mechanical feedback hypothesis of organ growth regulation. Seminars in cell & developmental biology. 2014; 35C:73–81.
2. Restrepo S, Zartman JJ, Basler K. Coordination of patterning and growth by the morphogen DPP. Current biology : CB. 2014; 24:R245–255. [PubMed: 24650915]
3. Worley MI, Setiawan L, Hariharan IK. Regeneration and transdetermination in Drosophila imaginal discs. Annual review of genetics. 2012; 46:289–310.
4. Garcia-Bellido A. The cellular and genetic bases of organ size and shape in Drosophila. The International journal of developmental biology. 2009; 53:1291–1303. [PubMed: 19924628]
5. Kornberg TB, Guha A. Understanding morphogen gradients: a problem of dispersion and containment. Current opinion in genetics & development. 2007; 17:264–271. [PubMed: 17643982]
6. LeGoff L, Lecuit T. Mechanical Forces and Growth in Animal Tissues. Cold Spring Harbor perspectives in biology. 2015; 8
7. Johnston LA. Socializing with MYC: cell competition in development and as a model for premalignant cancer. Cold Spring Harbor perspectives in medicine. 2014; 4
8. Calleja M, Moreno E, Pelaz S, Morata G. Visualization of gene expression in living adult Drosophila. Science. 1996; 274:252–255. [PubMed: 8824191]
9. Ibrahim DM, Biehs B, Kornberg TB, Klebes A. Microarray comparison of anterior and posterior Drosophila wing imaginal disc cells identifies novel wing genes. G3. 2013; 3:1353–1362. [PubMed: 23749451]
10. Reeves N, Posakony JW. Genetic programs activated by proneural proteins in the developing Drosophila PNS. Developmental cell. 2005; 8:413–425. [PubMed: 15737936]
11. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nature methods. 2018; 15:343–346. [PubMed: 29553579]
12. Edsgard D, Johnsson P, Sandberg R. Identification of spatial expression trends in single-cell gene expression data. Nature methods. 2018; 15:339–342. [PubMed: 29553578]
13. Davie K, et al. A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. Cell. 2018
14. Han X, et al. Mapping the Mouse Cell Atlas by Microwell-Seq. Cell. 2018; 173:1307. [PubMed: 29775597]
15. Croset V, Treiber CD, Waddell S. Cellular diversity in the Drosophila midbrain revealed by single-cell transcriptomics. eLife. 2018; 7
16. Haber AL, et al. A single-cell survey of the small intestinal epithelium. Nature. 2017; 551:333–339. [PubMed: 29144463]
17. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nature biotechnology. 2015; 33:495–502.
18. Karaiskos N, et al. The Drosophila embryo at single-cell transcriptome resolution. Science. 2017; 358:194–199. [PubMed: 28860209]
19. Achim K, et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nature biotechnology. 2015; 33:503–509.
20. Halpern KB, et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature. 2017; 542:352–356. [PubMed: 28166538]
21. Zhu Q, Shah S, Dries R, Cai L, Yuan GC. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nature biotechnology. 2018

22. Macosko EZ, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015; 161:1202–1214. [PubMed: 26000488]
23. Andrews TS, Hemberg M. M3Drop: Dropout-based feature selection for scRNASeq. Bioinformatics. 2018
24. Morikawa RK, Kanamori T, Yasunaga K, Emoto K. Different levels of the Tripartite motif protein, Anomalies in sensory axon patterning (Asap), regulate distinct axonal projections of Drosophila sensory neurons. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108:19389–19394. [PubMed: 22084112]
25. Lammel U, et al. The Drosophila FHOD1-like formin Knittrig acts through Rok to promote stress fiber formation and directed macrophage migration during the cellular immune response. Development. 2014; 141:1366–1380. [PubMed: 24553290]
26. Sieglitz F, et al. Antagonistic feedback loops involving Rau and Sprouty in the Drosophila eye control neuronal and glial differentiation. Science signaling. 2013; 6:ra96. [PubMed: 24194583]
27. Glasscock E, Tanouye MA. Drosophila couch potato mutants exhibit complex neurological abnormalities including epilepsy phenotypes. Genetics. 2005; 169:2137–2149. [PubMed: 15687283]
28. Bellen HJ, Kooyer S, D'Evelyn D, Pearlman J. The Drosophila couch potato protein is expressed in nuclei of peripheral neuronal precursors and shows homology to RNA-binding proteins. Genes Dev. 1992; 6:2125–2136. [PubMed: 1427076]
29. Clohisey SM, Dzhindzhev NS, Ohkura H. Kank Is an EB1 interacting protein that localises to muscle-tendon attachment sites in Drosophila. PloS one. 2014; 9:e106112. [PubMed: 25203404]
30. Liu Z, et al. Low density lipoprotein receptor class A domain containing 4 (LDLRAD4) promotes tumorigenesis of hepatic cancer cells. Experimental cell research. 2017; 360:189–198. [PubMed: 28888937]
31. Su G, Morris JH, Demchak B, Bader GD. Biological network exploration with Cytoscape 3. Curr Protoc Bioinformatics. 2014; 47:8 13 11–24. [PubMed: 25199793]
32. McCarthy DJ, Campbell KR, Lun AT, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017; 33:1179–1186. [PubMed: 28088763]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

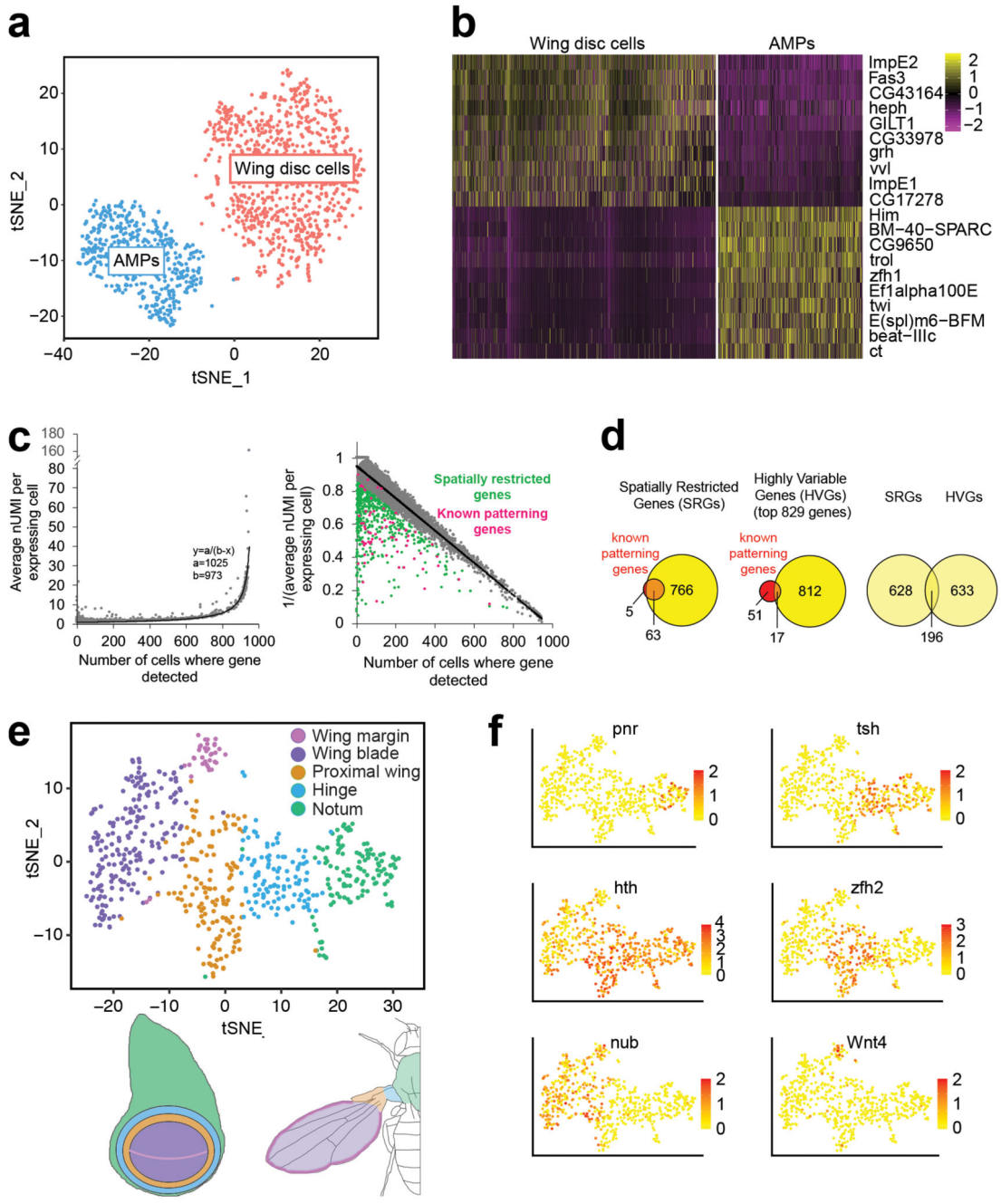

Fig. 1. Single-cell sequencing of wing disc cells identifies Spatially Restricted Genes.

a, b, Two-dimensional t-SNE representation of all sequenced cells reveals two main cell clusters (a) corresponding to wing disc cells and adult muscle precursors (AMPs), based on differential expression of AMP genes in the two clusters (b). n=1,468 high quality cells.

c, Identification of Spatially Restricted Genes (SRGs) as genes observed in fewer cells than expected based on their expression level. n=9929 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 (Suppl. Table 3). In comparison, a

similarly sized set of ‘Highly Variable Genes’ 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 5 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.

Europe PMC Funders Author Manuscripts

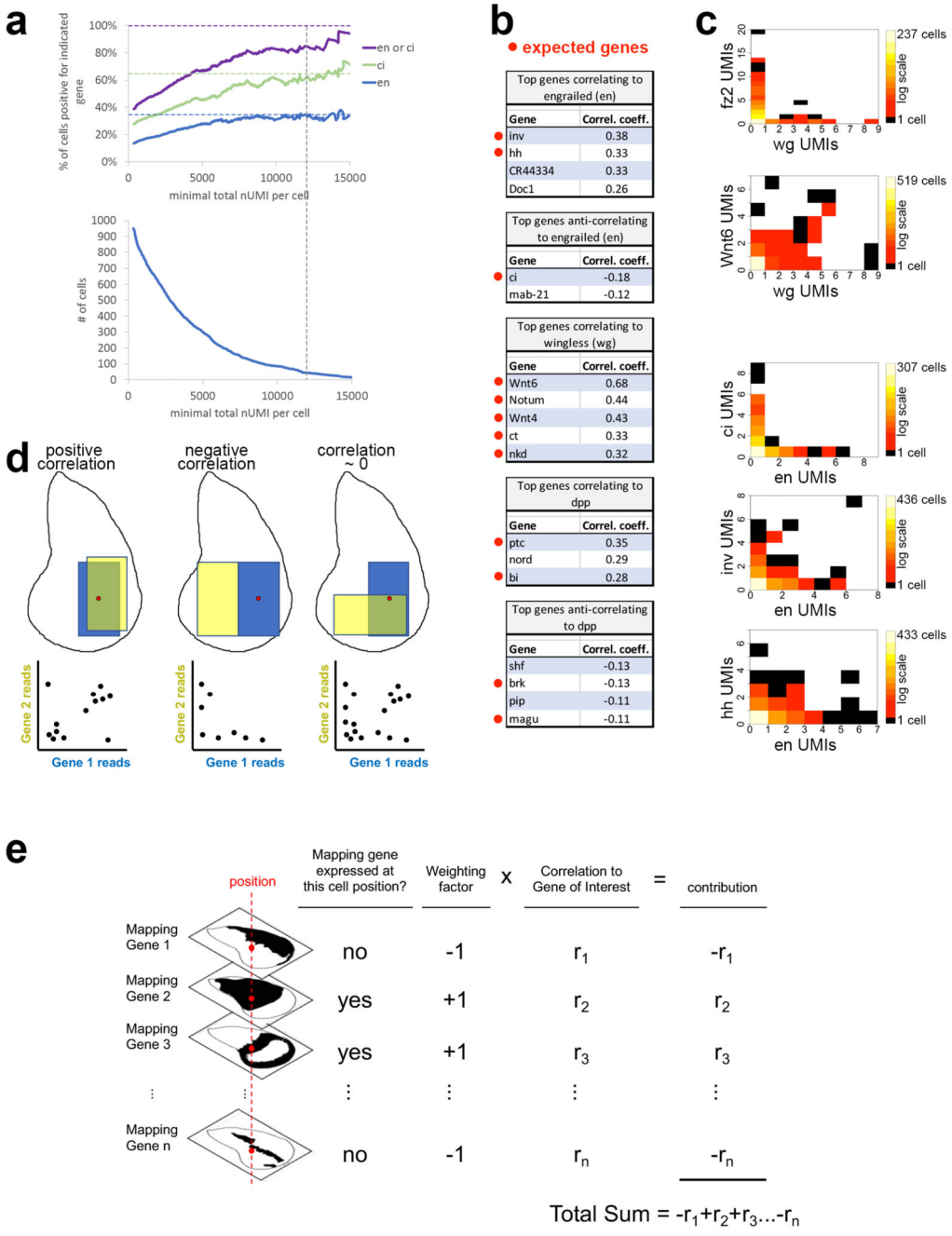

Fig. 2. Generation of gene expression maps based on gene expression correlations

a, The high false-negative rate of scSeq makes it difficult to confidently conclude that a gene is not expressed in any one sequenced cell, and hence to confidently locate its original position in the wing disc. Shown is the number of cells that are positive for expression of *en* or *ci*, which have largely mutually exclusive expression patterns in the wing disc, at different sequencing depth thresholds. nUMI= number of unique molecular identifiers.

Europe PMC Funders Author Manuscripts

b, 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 with one outlier removed. See Methods for details.
c, Two-dimensional histograms showing the distribution of all sequenced cells according to the level of expression of the two indicated genes.
d, Concept for generating expression maps based on gene expression correlations. Positive correlation between two genes indicates they have overlapping expression domains, whereas a negative correlation indicates expression domains that are more mutually exclusive.
e, Schematic representation of the method used to generate computed expression maps.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

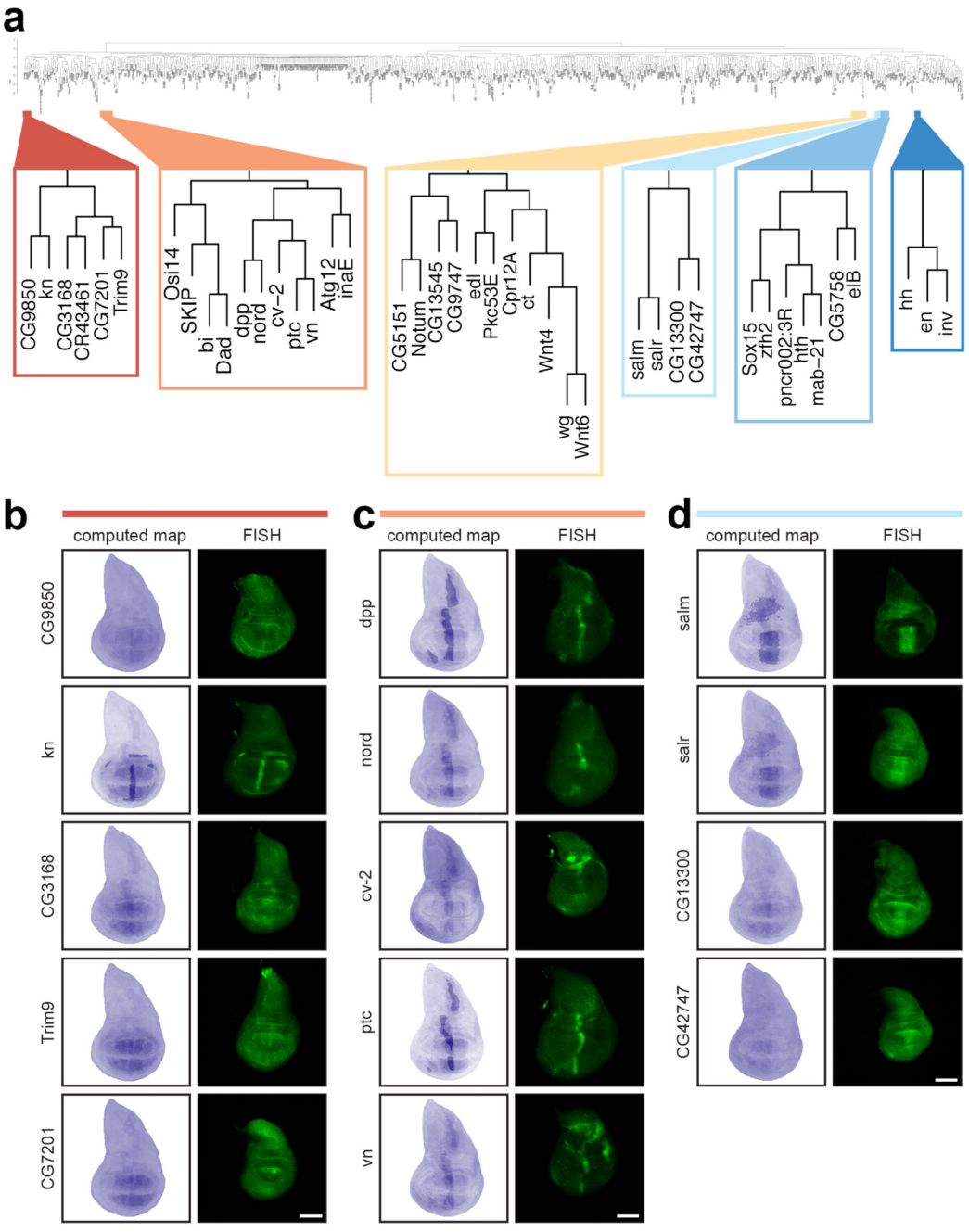

Fig. 3. Computed expression maps for genes of unknown expression patterns agree well with their actual expression detected by *in situ* hybridization.

a, Hierarchical clustering of SRGs by expression correlation identifies clusters of genes with related expression patterns containing both genes of known and unknown function.

b-d, Expression patterns of genes detected by *in situ* hybridization largely confirm the expression patterns predicted by the computed maps. For testing, all genes in the ‘red’ (b), ‘orange’ (c) and ‘light blue’ clusters (d) shown in panel (a) were selected. Images representative of 10 discs and 2 biological replicates. Scale bars = 100μm.

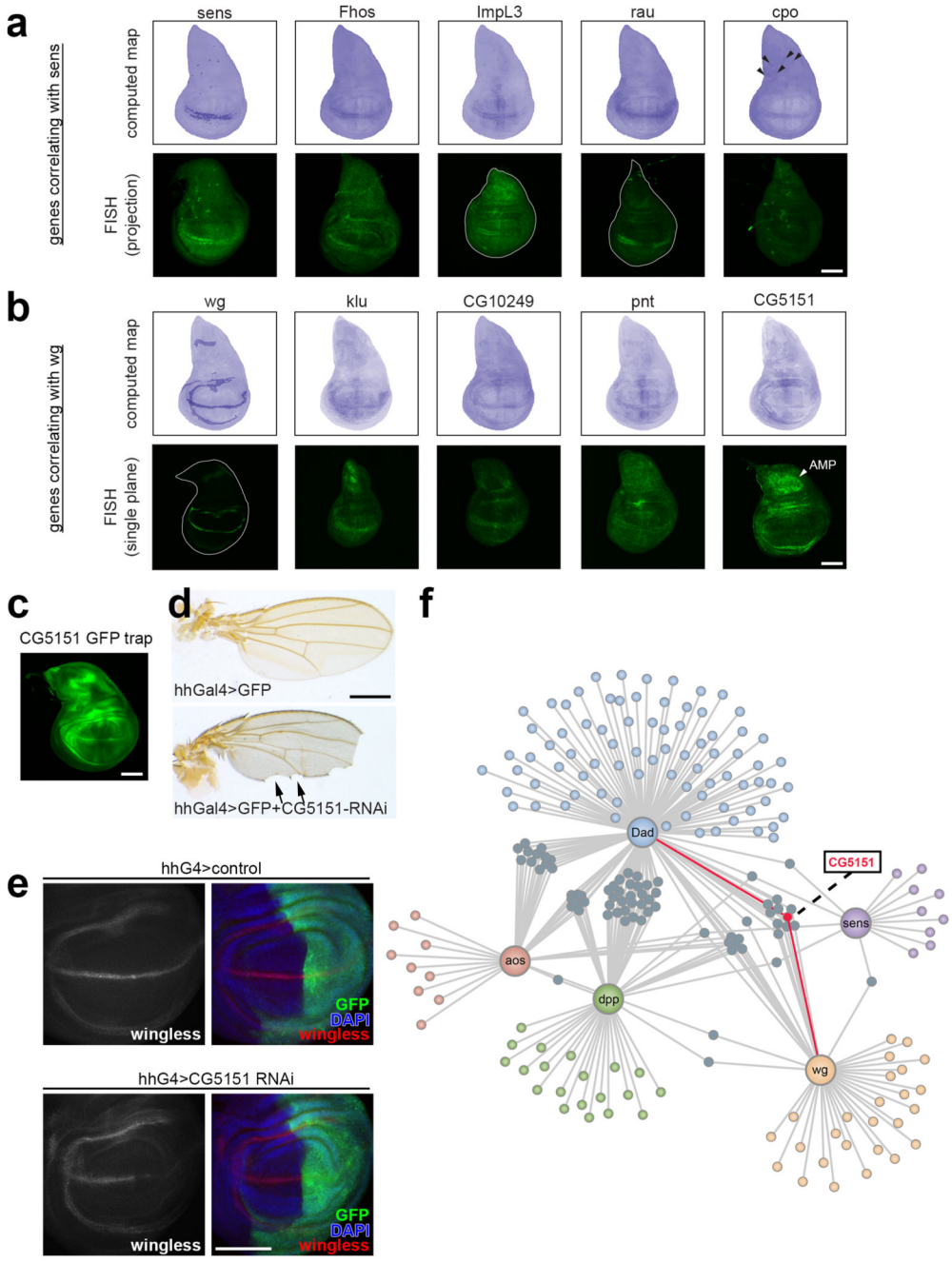

Fig. 4. Discovery of genes in processes of interest based on their expression pattern

a, b, Computed expression maps and *in situ* hybridizations for genes correlating with either *senseless* (a) or *wingless* (b). Images representative of 10 discs and 2 biological replicates. Scale bars = 100μm.

c, Expression of *CG5151* using a GFP transcript trap in the endogenous *CG5151* locus reveals expression at the D/V boundary and in a more proximal ring, similar to that of wingless. Image representative of 3 discs and 1 biological replicate. Scale bar = 100μm.

Europe PMC Funders Author Manuscripts

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 (8 of 8 knockdown discs showed reduced wg expression at 29°C while 0 of 10 control discs had reduced wg. Two biological replicates.) (e). Scale bars = 500μm for (d) and 100μm for (e).

f, Gene-network analysis of the most connected genes linked to Dpp, Wnt, Notch and/or EGRF signaling pathways. Only edges with a minimum correlation coefficient of 0.1 are shown. Self-correlations are excluded.

Europe PMC Funders Author Manuscripts