An Image-Based Data-Driven Analysis of Cellular Architecture in a Developing Tissue

Jonas Hartmann[1,*], Mie Wong[2], Elisa Gallo[1,2,3], Darren Gilmour[2,*]

[1]Cell Biology and Biophysics Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany

[2]Institute of Molecular Life Sciences, University of Zurich (UZH), 8057 Zurich, Switzerland

[3]Collaboration for joint PhD degree between EMBL and Heidelberg University, Faculty of Biosciences, 69120 Heidelberg, Germany

*Correspondence to: Jonas Hartmann, jonas.m.hartmann@protonmail.com

Darren Gilmour, darren.gilmour@imls.uzh.ch

ABSTRACT

Quantitative microscopy is becoming increasingly crucial in efforts to disentangle the complexity of organogenesis, yet adoption of the potent and ever-growing toolbox provided by modern data science has been slow, primarily because data-driven tools focus on "big data" whereas microscopy produces "rich data". We tackle this issue using a newly developed algorithm for point cloud-based morphometry to unpack the rich information encoded in high-resolution 3D image data into a straightforward numerical representation. This allowed us to employ machine learning for multi-modal data integration of cell morphology, intracellular organization, gene expression and annotated contextual knowledge that facilitates data exploration. We apply these techniques to construct and explore a quantitative atlas of cellular architecture for the zebrafish posterior lateral line primordium, an experimentally tractable model of complex self-organized organogenesis. In doing so, we are able to retrieve both previously established and novel biologically relevant patterns, demonstrating the potential of our data-driven approach.

INTRODUCTION

Organogenesis proceeds as a complex multi-scale process. Cells utilize a wide range of molecular machinery in a coordinated fashion to give rise to tissue-scale collective behavior, which in turn feeds back on individual cells' architectural and transcriptional dynamics [Chan et al., 2017]. Uncovering the principles that govern these systems is a long-standing but elusive goal of developmental biology, in part because it is often challenging if not impossible to reduce such complex phenomena to the action of single genes or simple mechanisms [Bizarri et al., 2013]. Thus, there is a persistent need for new techniques that enable integrative analysis of developmental systems.

In recent years, *data science* has arisen as a new interdisciplinary paradigm integrating statistics, computer science and machine learning with the aim of generating knowledge and predictions in a data-driven rather than hypothesis-driven fashion [Dhar et al. 2013; Blei & Smyth, 2017; Baker et al., 2018]. The application of such data-driven approaches to biology promises a new way of extracting information from large and otherwise inscrutable datasets, spurring progress toward a holistic and quantitative picture of biological systems. However, whilst this promise is already being realized to great effect in some fields, for instance in high-throughput cell biology [Roukos & Misteli, 2014; Gut et al., 2018; Chessel & Salas, 2019] and in (multi-)omics analysis [Libbrecht & Noble, 2015; Angerer et al., 2017; Huang et al., 2017; Ching et al., 2018], developmental biology has seen little adoption of data science techniques to date.

This is primarily because the field's main source of data, *in vivo* microscopy, does not readily lend itself to the production of "big data", upon which much of the recent progress in data science is founded. Although imaging datasets of *in vivo* biological systems are often large in terms of computer memory, they generally do not comprise the key element that makes "big data" so useful, namely a very large number of samples (on the order of thousands or more). In addition, the limited number of components that can be labeled and observed simultaneously reduces the number of different measurements associated with each sample. However, imaging data has the crucial advantage that it contains information on both abundance and localization of measured components and thus indirectly encodes rich higher-order information such as patterns, textures, object shapes, object locations, and overall sample structure. In other words, quantitative imaging generates "rich data" rather than "big data".

Employing the power of data science to study development thus entails three challenges: unpacking the rich information encoded in images into a more accessible format (*data extraction*), integrating data across multiple experiments to overcome the limited number of simultaneous measurements (*data integration*), and finally analyzing and visualizing the resulting multi-layered dataset to enable the discovery of biologically meaningful patterns (*data interpretation*).

Here, we address each of these challenges in the context of a comprehensive data-driven analysis of cellular architecture in an intricate developing tissue. We first engineered a data extraction pipeline combining high-resolution live microscopy, 3D single-cell segmentation and a novel point cloud-based algorithm for automated extraction of unbiased quantitative cell descriptors (figure 1A-C). We then co-opted machine learning techniques to perform data integration across experiments, constructing a multi-modal atlas of cellular architecture. Building on this approach, we also mapped contextual knowledge into the atlas and used it to facilitate biological data interpretation through context-guided visualization (figure 1D-F).

As a case study to develop and apply these tools we made use of the zebrafish posterior Lateral Line Primordium (pLLP), a model tissue that migrates collectively along the flank of the developing zebrafish embryo, periodically assembling and depositing rosette-shaped clusters of cells that cease migration and differentiate to form sensory organs [Haas & Gilmour, 2006; Ghysen & Dambly-Chaudière, 2007]. The pLLP is patterned into a leader zone of cells that are highly polarized in the direction of migration and a follower zone where rosettes are being assembled through apical constriction (see figure 2A) [Nechiporuk & Raible, 2008; Lecaudey et al., 2008]. This tight integration of collective migration, patterning and morphogenesis gives rise to an intricate tissue architecture that, although easy to image, is challenging to study at the single-cell level as its cells exhibit a wide variety of shapes and behaviors in addition to being comparably small and densely packed [Galanternik et al., 2016; Nogare et al., 2017]. We show that our data science-inspired analysis retrieves both known and novel patterns of how cells are organized within this complex developing tissue, demonstrating how a data-driven approach can lay a quantitative foundation for the systems-level study of organogenesis.

RESULTS

High-Resolution Live Imaging and 3D Single-Cell Segmentation of the Zebrafish Posterior Lateral Line Primordium

Data-driven approaches are fundamentally reliant on high-quality input datasets. A data-driven analysis of cellular architecture thus requires high-resolution imaging followed by automated segmentation of individual cells. To this aim, we employed AiryScan FAST mode confocal microscopy [Huff, 2016] to achieve high signal-to-noise ratios and high axial resolution – both of which greatly facilitate 3D segmentation – whilst maintaining a sufficiently high acquisition speed for live imaging of a migrating tissue (~20s/channel). In this way, we acquired a large set of single and dual channel volumes of wild-type primordia during migration (N=173 samples in total), all of which expressed the bright cell membrane label *cldnb:lyn-EGFP* [Haas & Gilmour, 2006] (figure 2A, movie 1).

To segment cells, we combined commonplace image processing algorithms into a specialized automated pipeline that uses labeled membranes as its sole input (see materials and methods for details). We found that our pipeline reliably produces high-quality segmentations (figure 2B, movie 2) and that erroneously split or fused cells are rare. To further ensure consistent segmentation quality, each segmented stack was manually double-checked and rare cases of stacks exhibiting substantial segmentation issues were excluded (8 of 173; 4.62%). An expanded visualization of a representative segmentation (figure 2C, movie 3) allows the diversity of cell shapes in the lateral line and their relationship with overall tissue architecture to be appreciated in a qualitative fashion.

Overall, we collected n=15'347 segmented cells from N=165 wild-type primordia as the basis for our quantitative analysis. Unless otherwise specified, the results presented in this study are based on this dataset.

Point Cloud-Based Morphometry for Unbiased Quantification of Cellular Architecture

Even when segmented, confocal volumes of cells are not readily amenable to most types of statistical data analysis and machine learning, as they do not form a canonical feature space (a 2D samples-by-features array). Features can be extracted by measuring specific aspects of each cell such as volume or eccentricity (*feature engineering*) but for exploratory analysis it is preferable to derive an unbiased encoding of 3D image information into a 1D feature vector (*feature embedding*) – a non-trivial problem. While this has previously been tackled in a number of ways [Pincus & Theriot, 2007; Peng & Murphy, 2011; Rajaram et al., 2012; Tweedy et al., 2013; Kalinin et al., 2018], no readily applicable solution for feature embedding of both cell morphologies and subcellular protein distributions from segmented 3D images has been described to our knowledge.

We therefore developed a novel method that allows feature embedding of arbitrary fluorescence intensity distributions. As a starting point, we took a workflow from classical geometric morphometrics consisting of four steps [Adams et al., 2013] (figure 3A): (1) conversion of image data into a point cloud of landmarks, (2) alignment of landmarks by registration to remove rotational variance, (3) re-representation of landmark coordinates based on their deviation from a consensus reference common to all samples, and (4) dimensionality reduction by Principal Component Analysis (PCA). To adapt this classical workflow to make it applicable to 3D images of cells (figure 3B, suppl. figure 1), it was necessary to solve three key problems.

First, a heterogeneous population of cells does not display common reference points shared across individuals, i.e. the equivalent to the nose or eyes in facial shape analysis, which would be required for consistent landmark assignment. We instead implemented a sampling strategy termed *Intensity-Biased Stochastic Landmark Assignment* (ISLA) [inspired by Chan et al., 2018], which treats normalized voxel intensity distributions as multinomial probability distributions and samples them to obtain a predefined number of landmark coordinates. The resulting point cloud is a sparse encoding of the original image wherein the local density of points represents local fluorescence intensity (figure 3C). By adjusting the number of landmarks being sampled, the trade-off between precision and computational performance can be controlled.

Second, a solution was needed to prevent differences in cell rotation and size from obscuring shape information. We first removed size variance by rescaling point clouds such that cell volumes are normalized. The most common solution for achieving rotational invariance, spatial registration of cells [Pincus & Theriot, 2007], is an ill-defined problem for a highly heterogeneous population of cells. We therefore instead chose to re-represent the 3D point cloud of each cell in terms of the points' pairwise distances from each other, a representation that is rotationally invariant. We use the term *Cell Frame Of Reference* (CFOR) to refer to feature spaces where size and rotation have been normalized in this fashion, so only pure shape information is retained. However, size and rotation can have biological significance in the larger context of a biological tissue. Thus, we also pursued a

second approach wherein we register entire tissues (rather than individual cells) prior to feature embedding. In this case, rotational variance of individual cells is not removed but instead becomes biologically meaningful as it now reflects the actual orientation of cells within the tissue. Feature spaces resulting from this approach are referred to as *Tissue Frame Of Reference* (TFOR) and retain size and relevant rotational information (figure 3B).

Third and finally, given that point clouds extracted by ISLA are not matched across multiple cells, the classical method for defining a consensus reference across samples, which is to use the average position of matched landmarks, is not applicable. To address this, we developed *Cluster-Based Embedding* (CBE) [inspired by Qiu et al., 2011]. CBE proceeds by first performing k-means clustering on an overlay of a representative subset of all cells' point clouds. Using the resulting cluster centers as consensus reference points, CBE then re-represents each individual cell's point cloud in terms of a simple proximity measure relative to said cluster centers (figure 3D). We validated CBE using a synthetically generated point cloud dataset, finding that it outperforms an alternative embedding strategy based on point cloud moments and that normalization of size and rotation (i.e. the cell frame of reference, CFOR) is beneficial in the detection of other shape features (suppl. figure 2). For further details on feature embedding with ISLA and CBE as well as our evaluation using synthetic point cloud data see the materials and methods section.

In summary, Cluster-Based Embedding (CBE) of point clouds obtained from 3D images by Intensity-biased Stochastic Landmark Assignment (ISLA) is an expressive and versatile embedding strategy for casting arbitrary 3D fluorescence distributions into an embedded feature space. It thus provides a means of unpacking the rich information encoded in image stacks into an accessible format for further quantitative analysis.

The Cellular Shape Space of a Model Tissue: The Lateral Line Primordium

We applied the ISLA-CBE workflow to the cell boundaries of our segmented dataset to derive an embedded feature space representing cell morphology across the pLLP tissue, which we term the pLLP's *cellular shape space*.

We found that the resulting Principal Components (PCs) describe meaningful shape variation (figure 4A-B) and that a small number of PCs is sufficient to capture most of the shape heterogeneity across the tissue (figure 4C). Interestingly, the cells of the pLLP do not cluster into discrete morphological groups (figure 4A-B, D), implying a continuous shape spectrum between biologically distinct cells such as those at the tissue's leading edge and those at the center of assembled rosettes.

To annotate the dimensions of the shape space with interpretable biological properties, we correlated them against a curated set of simple engineered features (see materials and methods, and suppl. table 2). Bigraph visualizations of these correlations reveal that the highest contributions to pLLP shape heterogeneity in the tissue frame of reference (TFOR) result from rotational orientation along different axes (PCs 1 and 2) as well as from absolute cell length in all three spatial dimensions (PCs 3, 5 and 6) (figure 4E). Some of the PCs do not strongly correlate with any engineered features, implying that this information would have been lost without the use of unbiased feature embedding. Further manual inspection showed PC 4 to relate to additional rotational information for flat cells, PC 7 to the curvature of cells along the tissue's dorso-ventral axis, and PC 8 to cell sphericity. Overall, the dominance of orientation and size in the TFOR shape space is consistent with our previous observations on synthetic data (suppl. figure 2A).

In the cell frame of reference (CFOR) (figure 4F), which is invariant to size and rotation, cell shape heterogeneity is instead dominated by cell sphericity (PC 1), which accounts for nearly 50% of cell shape variation, with lesser contributions from cell surface smoothness (PC 2) and various minor factors that could not be annotated with a clear corresponding engineered feature. Cell sphericity and smoothness are thus important components of the pLLP's cellular architecture, a result that would have gone unnoticed without size- and rotation-invariant shape analysis. The minor factors that remain unidentified are likely mostly noise but PCs 3-5, which still explain a few percentage points of variance each (figure 4C), may encode some meaningful shape information that is not obvious to the human eye and could only be retrieved using a computational approach.

One advantage of image-based data is that it allows features of single cells to be interpreted in the original spatial context of the tissue. To this end, we generated consensus maps of feature variation across primordia based on registered cell centroid positions (figure 4G). As expected from our correlation analysis, features such as TFOR-PC1 (dorso-ventral orientation) are patterned along the corresponding axis of the tissue (top left panel). For TFOR-PC3 (cell height), the consensus map reveals an increase in cell height immediately behind the tissue's 'leading edge' (top right panel), which is consistent with the established notion that leader cells are flatter and follower cells

transition into a more packed and columnar state [Lecaudey et al., 2008]. This leader-follower progression is also reflected in CFOR-PC2 (cell surface smoothness), which displays a very similar pattern to TFOR-PC3 (figure 4G, bottom right panel), possibly due to greater protrusive activity in leaders causing distortions in the cell's shape and thus reducing smoothness. These results show that our shape space analysis allows known key features of pLLP organization to be recovered in a data-driven fashion.

As established above, we found that CFOR-PC1 (sphericity) is a key component of the pLLP's cellular architecture. The corresponding consensus pattern is more intricate than the others (figure 4G, bottom left panel), implying that different processes affect cell sphericity in the primordium. Cells at the front of the tissue register as more spherical due to reduced cell height, again as a consequence of the aforementioned leader-follower transition. The more intriguing pattern is that amongst followers there is a tendency for cells at the center of the tissue to be more spherical than those at the periphery, which manifests as a horizontal stripe in the consensus map. This pattern has not been described previously and may point toward an unknown aspect of pLLP self-organization.

Taken together, these results illustrate how data-driven exploratory analysis of the cellular shape space can reveal interesting patterns of shape heterogeneity.

Data Integration via Machine Learning on Embedded Features

A key strength of the ISLA-CBE pipeline is that it is not limited to embedding object surfaces such as those resulting from single-cell segmentation. Instead, it is capable of embedding arbitrary intensity distributions, including those of the cytoskeleton, organelles, or any other labeled protein. Thus, ISLA-CBE can be used to analyze, and integrate, many aspects of cellular organization beyond morphology.

Our dataset of membrane-labeled tissues encompasses many dual-color stacks containing not only labeled membranes but also one of several other cellular structures, including nuclei (*NLS-tdTomato*, figure 5A), F-actin (*tagRFPt-UtrophinCH*, figure 5C), and the Golgi apparatus (*mKate2-GM130*, figure 5E) (see suppl. table 1 for a complete overview). For each of these structures we computed specific ISLA-CBE feature spaces (figure 5B,D,F).

We sought to combine this intracellular information into an integrated quantitative atlas of cellular architecture. Such atlas integration of image data is usually performed in image space by spatial registration of samples [Peng et al., 2011; Vergara et al., 2017; McDole et al., 2018; Cai et al., 2018]. However, such registration requires stereotypical shapes that can be meaningfully overlaid via spatial transformations, so it is not applicable to the non-stereotypical cells in a developing tissue. Machine learning can be used as an alternative strategy for atlas mapping that side-steps this problem. This works by training a machine learning model to predict a feature measured in only a few samples (e.g. a specific protein distribution) based on reference features that are available for all samples (e.g. cell shape). This strategy can be applied directly to microscopy images by means of deep convolutional neural networks [Johnson et al., 2017; Christiansen et al, 2018]. However, such deep learning typically requires very large datasets. Here, we instead chose to work in the embedded feature space, predicting ISLA-CBE embeddings of protein distributions based on ISLA-CBE embeddings of cell shape. This simplifies the problem such that smaller-scale machine learning algorithms become applicable.

We trained different classical (non-deep) machine learning models to predict the embedded feature spaces of subcellular structures based on the shape space. We optimized and validated this approach using grid search and cross validation (see materials and methods). Focusing on feature spaces in the tissue frame of reference (TFOR), we found that multi-output Support Vector Regression (SVR) yields good predictions across different conditions (suppl. figure 4). We therefore used SVR prediction to generate a complete atlas spanning the embedded TFOR feature spaces of all available markers.

This atlas can be explored using the same visualizations shown previously for the shape space. As an example, we correlated the embedded features representing the Golgi apparatus (*mKate2-GM130*) with engineered features describing cell shape (figure 5G). This shows that Golgi PCs 1 and 2 in the tissue frame of reference again relate to cellular orientation, whereas Golgi PC 3 correlates with cell height and a more columnar cell shape. Interestingly, the point cloud visualization reveals that Golgi

PC 3 represents apical enrichment (figure 5H,I) and the corresponding consensus tissue map shows that PC 3 is high in follower but not in leader cells (figure 5J), indicating increased apical localization of the Golgi apparatus in followers. As such localization is a hallmark of many epithelia, this finding provides unbiased support for a model where follower cells display increased epithelial character.

Besides subcellular protein distributions, a key data modality for the study of developing tissues is gene expression. Indeed, there is great interest in ongoing efforts to integrate gene expression data with cell and tissue morphology and behavior [Battich et al., 2013; Lee et al., 2014; Satija et al., 2015; Karaiskos et al, 2017; Stuart et al., 2019]. The machine learning approach to atlas construction we introduced here can be applied to any quantitative measurement that allows cell shape to be acquired simultaneously, including in-situ measurements of gene expression.

To demonstrate this, we performed single-molecule Fluorescence In-Situ Hybridization (smFISH) of *pea3* RNA (figure 6A), a marker of FGFR signaling activity associated with the progression of follower cell development [Aman & Piotrowski, 2008; Durdu et al., 2014]. We acquired 2-color stacks of smFISH probes and cell membranes from fixed samples (N=31, n=3149), employed automated spot detection to identify and count RNA molecules (suppl. figure 4A-E) and embedded cell shapes with ISLA-CBE. We then used SVR to predict smFISH spot counts based on cell shape and location, finding that by combining all available information (the tissue frame of reference shape space, the cell frame of reference shape space and cell centroid coordinates) the trained model is able to account for 38.2±1.9% of *pea3* expression variance (figure 6B). The residual variance that could not be accounted for is due to high cell-to-cell heterogeneity in *pea3* expression among follower cells that does not seem to follow a clear spatial or cell shape-related pattern (figure 6C). Nevertheless, we were able to recover the graded overall leader-follower expression pattern of *pea3* when running predictions across the entire atlas (figure 6D). Our approach could therefore be used to superimpose at least the key patterns of gene expression in a tissue based on a set of in-situ labeling experiments, which in turn could potentially serve as a reference set to incorporate scRNA-seq data [Stuart et al., 2019].

Data integration based on common reference measurements such as cell shape counters one of the major shortcomings of current-day fluorescence microscopy, which is the limited number of channels and thus of molecular components that can be imaged simultaneously. The approach demonstrated here shows how embedded feature spaces and machine learning can be utilized to perform such data integration even in non-stereotypically shaped samples.

Mapping of Morphological Archetypes Facilitates Biological Interpretation

Feature embedding and atlas mapping enable the conversion of images into rich multi-dimensional numerical datasets. To take full advantage of this, biologically relevant patterns such as the relationships between different cell types, cell shapes and tissue context need to be distilled from this data and presented in a human-interpretable form. Accomplishing this in an automated fashion rather than by laborious manual data exploration remains a challenging problem in data science.

Classically, data interpretation involves relating data to established knowledge. This prompted us to look for ways to encode contextual knowledge quantitatively and use it to probe our atlas in a context-guided fashion. When asking biological questions about pLLP development, it is useful to distinguish different cell populations: the leader cells that are focused on migration, the cells at the center of nascent rosettes that go on to form sensory hair cells later in development, the cells in the periphery of rosettes that will differentiate into so-called support and mantle cells and the cells in between rosettes (inter-organ cells) that will be deposited as a chain of cells between the maturing lateral line organs [Grant et al., 2005; Hernández et al., 2007; Nogare et al., 2017] (figure 7A). These four populations constitute simple conceptual archetypes that facilitate reasoning about the primordium's organization.

To map archetypes into the tissue's cellular shape space, we manually annotated cells that constitute unambiguous examples for each archetype in a subset of samples (N=26, n=624) (figure 7A). We then used a Support Vector Classifier (SVC) with either tissue or cell frame of reference shape features as input to predict the class of all unlabeled cells, finding that for both sets of input features the classifier was readily able to distinguish leader cells, central cells and peripheral cells, confirming that these manually chosen archetypes are indeed morphologically distinct (suppl. figure 5). The inter-organ cells, however, were frequently misclassified as peripheral or central cells, indicating that at this developmental stage they are not substantially different in terms of shape from other follower cells (suppl. figure 5). This illustrates that mapping of manually annotated information into the atlas intrinsically provides an unbiased quality control of biological pre-conceptions, as classification fails for archetypes that are not distinguishable from others.

Importantly, the SVC not only predicts categorical labels but also the probabilities with which each cell belongs to each class, a measure of how much a given cell resembles each archetype. This provides an entry point for human-interpretable visualization of atlas data. Here, we performed PCA on the prediction probabilities to arrive at an intuitive visualization wherein cells are distributed according to their similarity to each archetype (figure 7B,C). In contrast to unsupervised dimensionality reduction of the shape space (see figure 4A,B,D), this new visualization is readily interpretable for anyone with basic prior knowledge of pLLP organization.

One directly noticeable pattern is in the number of cells found in intermediate states between the leader, central and peripheral archetypes (figure 7C). Intermediates between leaders and peripheral

cells are common, as would be expected given the leader-follower axis along the length of the primordium. Similarly, intermediates between peripheral and central rosette cells are common, reflecting the continuous nature of rosette structure along the inside-outside axis. However, cells in an intermediate state between the leader and central rosette archetype are rare, possibly indicating that a direct transition between the two states does not frequently occur.

Any other data available at the single-cell level can be overlaid onto the archetype visualization to take full advantage of its intuitive interpretability (figure 7D). When doing so, aspects of cell shape that have little relevance in the context of the selected archetypes now appear unpatterned (e.g. dorso-ventral orientation). By contrast, the increased cell height in followers – particularly in central cells – is clearly visible. The same pattern can be seen for PC 3 of the embedded protein distribution of F-actin, which is part of the atlas. This reflects apical enrichment of F-actin in follower cells, an important feature of rosette morphogenesis [Lecaudey et al., 2008]. As a final example, plotting CFOR-PC1 (cell sphericity) shows that it is elevated in central rosette cells, especially in a population furthest from the peripheral archetype. This is consistent with the spatial distribution seen in figure 4G and reinforces the finding that cell sphericity exhibits a non-trivial pattern related to rosette organization.

Using the predicted archetype labels, it is now also readily possible to test observations with classical statistical hypothesis testing, which substantiates both the previously known leader-follower difference in cell height (figure 7E) and the novel finding that inner rosette cells are significantly more spherical than outer rosette cells (figure 7F).

Despite its relative simplicity, our archetype-based approach showcases how the mapping of contextual knowledge onto a large and complicated dataset can facilitate its intuitive visualization and biological interpretation.

DISCUSSION

Data-driven approaches represent a major opportunity for the advancement of developmental biology, particularly in the search for holistic system-level explanations. However, seizing this opportunity requires the analysis of image-based multi-layered datasets spanning space, time, several scales and different modalities – a non-trivial task for which data science does not currently provide standardized solutions. Our data-driven analysis of cellular architecture (figure 1) addresses the key challenges entailed in this task, which can be grouped into three central aspects: data extraction, data integration, and data interpretation.

During data extraction, the rich information encoded in images must be pinpointed and transformed into a format more amenable to data science tools. This can be achieved by performing single-cell segmentation (figure 2) and subsequently extracting numerical features describing each cell. Features can be generated through feature engineering, which has been employed successfully in a number of cases [Wang et al., 2017; Viader-Llargues et al., 2018]. However, to gain a more complete and unbiased picture, feature embedding is preferable. We developed a novel feature embedding strategy inspired by classical geometric morphometrics, ISLA-CBE (figure 3). ISLA allows arbitrary fluorescence intensity distributions to be converted to point clouds which in turn are embedded into a feature space with CBE. Previously described embedding strategies focus separately on shape [Tweedy et al., 2013; Kalinin et al., 2018] or on protein distributions [Rajaram et al., 2012; Tweedy et al., 2013; Gut et al., 2018], or require subcellular structures to be segmented into objects [Peng & Murphy, 2011; Johnson et al., 2015]. Furthermore, shape-oriented methods usually require somewhat stereotypical objects that can be registered in order to achieve rotational invariance [Pincus & Theriot, 2007], a problem ISLA-CBE can solve through a simple pairwise distance transform (i.e. the cell frame of reference, CFOR). Recent work established neural network-based autoencoders as another highly promising approach for feature embedding [Johnson et al., 2017] but such deep learning methods tend to require large datasets for training and their internal workings are intractable [Marcus, 2018]. ISLA-CBE has its own shortcomings, most notably that it is not easily reversible, meaning it is not readily possible to reconstruct a point cloud or a microscopy image from any point in the embedded feature space. Nevertheless, ISLA-CBE serves as a simple and broadly applicable tool for feature embedding of segmented cells.

Data integration across experiments is crucial to data-driven developmental biology because live microscopy is limited in the number of components that can be measured simultaneously. The value of generating such integrated data atlases has been demonstrated in several examples [Peng et al., 2011; Vergara et al., 2017], including a dynamic protein atlas of cell division [Cai et al., 2018] and an atlas of cell fate and tissue motion in the mouse post-implantation embryo [McDole et al., 2018]. However, these examples employed registration in image space and thus relied on the stereotypical shape of their samples, a prerequisite that cells in developing tissues do not necessarily conform to.

Deep learning once again represents a promising alternative [Johnson et al., 2017; Christiansen et al, 2018; Ounkomol et al., 2018], though it remains to be demonstrated that it is applicable to the relatively small-scale datasets that can feasibly be produced in a developmental biology context. Here, we opted to use classical machine learning tools to perform multivariate-multivariable regression in order to map embedded feature spaces from different channels onto each other with cell shape as a common reference (figure 5), an approach that could also easily be extended to gene expression data via smFISH (figure 6). This simple and general solution comes with the downside that the resulting atlas cannot easily be viewed as an image overlay. Crucially, however, quantitative data analysis and visualization remain possible.

Finally, data interpretation is perhaps the most important and most challenging aspect. Even in fields leading in the adoption of data-driven methods, mining big datasets for human-readable mechanistic explanations remains a fundamental challenge [Holzinger et al., 2014; Baker et al., 2018]. When aiming to make progress on this front, it is helpful to remind oneself that data on its own is not inherently meaningful; it can only be meaningfully interpreted in the context of an external conceptual framework [Callebaut, 2012; Leonelli, 2019]. Guided by this notion, we sought a way of mapping contextual information onto the cellular shape space, allowing the entire dataset to be analyzed and visualized through a conceptual lens. Similar to recent work on the classification of cellular protrusions [Driscoll et al., 2019], we employed a machine learning classifier to map biologically meaningful cell archetype classes from a manually curated training set onto the entire dataset. We then utilized the prediction probabilities to generate a context-guided visualization that enables the biological interpretation of patterns across the entire atlas (figure 7). Despite its relative simplicity, this approach represents a general strategy for bringing together data and context to facilitate interpretation.

Throughout our analysis, we found evidence that cells in the developing pLLP, from a morphological standpoint, do not fall into distinct groups but rather onto shape spectra between different states (figure 4A,B,D). One such shape spectrum exists along the length of the primordium: leader cells are more flat and less defined by apicobasal polarity in their architectural features, whereas follower cells take on a more columnar architecture with the apical side as a focal point for cellular organization (figures 4G, 5G-J and 7D,E). This provides data-driven support for previous observations that have led to a model where leaders exhibit mesenchyme-like features focused on driving migration and followers assume epithel-like features geared toward rosette morphogenesis [Nechiporuk & Raible, 2008; Lecaudey et al., 2008; Nogare et al., 2017]. A second shape spectrum runs from the inside to the outside of assembling rosettes. As development proceeds, this will eventually be discretized into distinct cell types: the hair cells, support cells and mantle cells [Hernández et al., 2007; Nogare et al., 2017]. Interestingly, we found evidence that this inside-outside pattern is associated with cell sphericity, a key shape factor of the pLLP's CFOR shape space (figures 4F,G and 7D,F). Given the well-established link between cell sphericity and effective cell

surface tension [Matzke, 1946; Lecuit & Lenne, 2007], this leads us to the hypothesis that variations in adhesion and/or contractility could be the driving mechanism behind this pattern, as they would naturally lead to inside-outside sorting of cells in accordance with the Differential Interfacial Tension Hypothesis (DITH) [Brodland, 2002]. Following this discovery in our exploratory analysis, further investigation will be required to test this idea and to dissect its biological implementation and function.

Collectively, the computational strategies presented in this study illustrate the potential of data-driven developmental biology to serve as a means to quantify, integrate and explore image-derived information. For this potential to be fully realized, future work will need to address three key points. First, a versatile open source software framework is needed that simplifies and standardizes handling of multi-layered biological datasets and thereby improves the accessibility of data science tools for developmental biologists. Second, further efforts are needed to facilitate biological interpretation of large multi-dimensional datasets, for instance by extending archetype mapping to other forms of contextual knowledge. Third, atlas datasets such as that of the pLLP need to be extended with additional data and new modalities, including temporal dynamics (based on cell tracking), biophysical properties (such as cell surface tension) and the status of gene regulatory networks (derived e.g. from transcriptomics data). In the long term, we envision a comprehensive tissue atlas, a "digital primordium", which can be mined for patterns and relationships across a wide range of experiments and modalities.

406 # MATERIALS AND METHODS

407 ## Animal Handling

408 Zebrafish (*Danio rerio*) were grown, maintained and bred according to standard procedures
409 described previously [Westerfield, 2000]. All experiments were preformed on embryos younger than
410 3dpf, as is stipulated by the EMBL internal policy 65 (IP65) and European Union Directive
411 2010/63/EU. Live embryos were kept in E3 buffer at 27-30°C. For experiments, pigmentation of
412 embryos was prevented by treating them with 0.002% N-phenylthiourea (PTU) (Sigma-Aldrich, St.
413 Louis, US-MO) starting at 25hpf. For mounting and during live imaging, embryos were anaesthetized
414 using 0.01% Tricaine (Sigma-Aldrich, St. Louis, US-MO).

415 ## Transgenic Fish Lines

416 All embryos imaged carried the membrane marker *cldnb:lyn-EGFP* [Haas & Gilmour, 2006], which
417 was used for single cell segmentation. In addition, different subsets of the main dataset carried one
418 of the following red secondary markers: *cxcr4b:NLS-tdTomato* (nuclei) [Donà et al., 2013],
419 *Actb2:mKate2-Rab11a* (recycling endosomes), *LexOP:CDMPR-tagRFPt* (trans-Golgi network and late
420 endosomes), *LexOP: B4GalT1(1-55Q)-tagRFPt* (*trans*-Golgi), 6x*UAS:tagRFPt-UtrCH* (F-actin) or
421 *atoh1a:dtomato* (transcriptional marker for hair cell specification) [Wada et al., 2010]. Furthermore,
422 in two subsets of the data a red marker was injected as mRNA, namely *mKate2-GM130(rat)* (*cis*-
423 Golgi) [Pouthas et al., 2008] and *mKate2-Rab5a* (early endosomes). Finally, one subset of samples
424 was treated with 1µM LysoTracker™ Deep Red (Thermo Fisher Scientific, Waltham, US-MA) in E3
425 medium with 1% DMSO for 90 minutes prior to imaging. The exact composition of the main dataset
426 is summarized in supplementary table 1.

427 To drive expression of the UAS construct, those fish additionally carried the Gal4 enhancer trap ETL
428 GA346 [Distel et al., 2009]. To drive expression of LexOP constructs [Emelyanov and Parinov, 2008],
429 those fish carried *cxcr4b:LexPR* [Durdu et al., 2014] and were treated with 10µM of the progesterone
430 analogue RU486 (Sigma-Aldrich, St. Louis, US-MO) from 25hpf.

431 Capped mRNA was produced by IVT using the mMESSAGEmMACHINE™ SP6 Transcription Kit
432 (Thermo Fisher Scientific, Waltham, US-MA) according to the manufacturers' instructions and was
433 injected at 250ng/µl into embryos at the 1-cell stage.

434 The following plasmids were generated by MultiSite Gateway Cloning (Invitrogen, Waltham, US-MA)
435 based on the Tol2kit [Kwan et al., 2007]: *LexOP:CDMPR-tagRFPt* (with *cry:mKate2* transgenic
436 marker), *LexOP:B4GalT1(1-55Q)-tagRFPt* (with *cry:mKate2* transgenic marker), *Actb2:mKate2-*
437 *Rab11a* (with *clmc2:GFP* transgenic marker), 6x*UAS:tagFRPt-UtrCH* (with *cry:mKate2* transgenic
438 marker), *sp6:mKate2-GM130(rat)* and *sp6:mKate2-Rab5a*. tagRFPt [Shaner et al., 2008] and UtrCH
439 [Burkel et al., 2007] were kindly provided by Jan Ellenberg and Péter Lénárt, respectively. Zebrafish
440 CDMPR, B4GalT1, Rab5a and Rab11a were cloned by extraction of total RNA from dechorionated

48hpf zebrafish embryos using the RNeasy mini kit (Qiagen, Hilden, Germany) according to manufacturer's instructions, followed by reverse transcription with SuperScript™ III Reverse Transcriptase (Thermo Fisher Scientific, Waltham, US-MA) using both random hexamers and oligo-dT simultaneously, according to the manufacturer's instructions, and finally amplification of genes of interest from cDNA with the following oligonucleotides (Sigma-Aldrich, St. Louis, US-MO) (template-specific region underlined):

CDMPR FOR: GGGGACAAGTTTGTACAAAAAAGCAGGCTGGATGTTGCTGTCTGTGAGAATAATCACT

CDMPR REV: GGGGACCACTTTGTACAAGAAAGCTGGGTCCATGGGAAGTAAATGGTCATCTCTTTCCTC

B4GalT1 FOR: GGGGACAAGTTTGTACAAAAAAGCAGGCTGGATGTCGGAGTCGGTGGGATTCTTC

B4GalT1 REV: GGGGACCACTTTGTACAAGAAAGCTGGGTCTTGTGAATTAACCATATCAGAGATAAATGAAATGTGTCG

Rab5a FOR: GGGGACAGCTTTCTTGTACAAAGTGGCTATGGCCAATAGGGGAGGAGCAACAC

Rab5a REV: GGGGACAACTTTGTATAATAAAGTTGCTTAGTTGCTGCAGCAGGGGGCT

Rab11a FOR: GGGGACAGCTTTCTTGTACAAAGTGGCTATGGGGACACGAGACGACGAATACG

Rab11a REV: GGGGACAACTTTGTATAATAAAGTTGCCTAGATGCTCTGGCAGCACTGC

High-Resolution Live Imaging

Embryos were manually dechorionated with forceps at 30-34hpf and anaesthetized with 0.01% Tricaine (Sigma-Aldrich, St. Louis, US-MO), then transferred into 1% peqGOLD Low Melt Agarose (Peqlab, Erlangen, Germany) in E3 containing 0.01% Tricaine and immediately deposited onto a MatTek Glass Bottom Microwell Dish (35mm Petri dish, 10mm microwell, 0.16-0.19mm coverglass) (MatTek Corporation, Ashland, US-MA). No more than 10 embryos were mounted in a single dish. A weighted needle tool was used to gently arrange the embryos such that they rest flatly with their lateral side directly on the glass slide. After solidification of the agarose, E3 containing 0.01% Tricaine was added to the dish.

The microscope used for imaging was the Zeiss LSM880 with AiryScan technology (Carl Zeiss AG, Oberkochen, Germany), henceforth LSM880. High-resolution 3D stacks (voxel size: 0.099µm in xy, 0.225µm in z) were acquired with a 40X 1.2NA water objective with Immersol W immersion fluid (Carl Zeiss, Oberkochen, Germany). Imaging in AiryScan FAST mode [Huff, 2016] with a piezo stage for z-motion and bi-directional scanning allowed acquisition times for an entire volume to be lowered to approximately 20 seconds (40 seconds for dual-color stacks using line switching). Deconvolution was performed using the LSM880's built-in 3D AiryScan deconvolution with 'auto' settings.

Note that optimal image quality could only be achieved by adjustment of the stage to ensure that the cover glass is exactly normal to the excitation beam. For each dish we imaged, we used 633nm reflected light and line scanning to get a live view of the cover glass interface, which allowed us to

manually adjust the pitch of the stage to be completely horizontal. This process was repeated for both zx and zy line scans.

smFISH: Fixation, Staining and Imaging

Single molecule Fluorescence In-Situ Hybridization (smFISH) was performed according to standard protocols [Durdu et al., 2014; Raj et al., 2008] using previously published Quasar 670-conjugated Stellaris smFISH probes (LGC, Biosearch Technologies, Hoddesdon, UK) designed to target *pea3* mRNA, listed below.

Briefly, embryos were fixed overnight in 4% PFA in PBS-T (PBS with 0.1% Tween-20) at 4°C, then rinsed 3 times in PBS-T and subsequently permeabilized with 100% methanol overnight at -20°C. Embryos were rehydrated with a methanol series (75%, 50%, 25% Methanol in PBS-T, 5min per step) and rinsed 3 times with PBS-T. The yolk was manually removed using forceps. Next, samples were pre-incubated with hybridization buffer (0.1g/ml deyxtrane sulfate, 0.02g/ml RNase-free BSA, 1mg/ml *E. coli* tRNA, 10% formamide, 5x SSC, 0.1% Tween-20 in ddH2O) at 30°C for 30min and subsequently hybridized with *pea3* probe solution (0.1µM in hybridization buffer) at 30°C overnight in the dark. After probe removal, embryos were stained with DAPI (1:1000) in washing buffer (10% formamide, 5x SSC, 0.1% Tween-20 in ddH2O) for 15minutes at 30°C and finally kept in washing buffer for 45min at 30°C

Stained embryos were mounted on glass slides using VECTASHIELD® HardSet™ Antifade Mounting Medium (Vector Laboratories, Burlingame, US-CA) and imaged immediately to prevent loss of signal due to photobleaching. Imaging was performed with a 63x 1.4NA oil immersion objective on the LSM880 in FAST mode with 488nm and 639nm excitation lasers. Stacks were acquired using 8x averaging with 0.187µm z-spacing and a pixel size of 0.085µm, then deconvolved with the built-in 3D AiryScan deconvolution on 'auto' settings.

The following smFISH probes were used:

1: aaggaagacggacagaggca, 2: ctgtgttttaatgagctcca, 3: cttaaccgtttgtggtcatt,
4: ccatccatcttataatccat, 5: agtataaggcacttgctggt, 6: atttccttgcgacctattag,
7: tcaacagtctatttaggggc, 8: atgtatttccttttttgtcgc, 9: aagaggtcttcagattcctg,
10: cctgaagttggcttaaatcc, 11: ggaacttgagcttcggtgag, 12: aacaaactgctcatcgctgt,
13: cactgagttctctgagtgaa, 14: ttcttaatcttcacaggcgg, 15: tagctgaagctttgcttgtg,
16: tcataggcactggcgtaaag, 17: ctggacatgagctcttagat, 18: ttgggggaataatgctgcat,
19: tgagggtggattcatatacc, 20: cggaagggaacctggaactg, 21: agagtgttgccgatggaaac,
22: tgctgaggaggataaggcaa, 23: ccatgtactcctgcttaaag, 24: tcctgtttgaccatcatatg,
25: caggttcgtaagtgtagtcg, 26: tgtgatggtacatggatggg, 27: aaacatgtagccttcactgt,
28: tggcacaacacgggaatcat, 29: tcacctcaccttcaaatttc, 30:accttcacgaaacacactgc,
31: tagttgaagtgagccacgac, 32: gaagggcaaccaagaactgc, 33: atgcgatgaagtgggcattg,

34: atgagtttgaattccatgcc, 35: ttgtcatagttcatggctgg, 36: gtaacgcaaagagcgactca,
37: ttttgcataattcccttctc, 38: aggttatcaaagcttctggc, 39: cgctgattgtcgggaaaagc,
40: gttgacgtagcgctcaaatt, 41: aagaaactccctcatcgagg, 42: tacatgtagcctttggagta,
43: aaaggagaatgtcggtggca, 44: gtggtaaactgggatgggaa, 45: atacaagaggatggggtggg,
46: gaatgcagagtccctaatga, 47: agataggcctcagaagtgag, 48: gcaatctcttgaaccacagt.

Software Development Stack

The software for this study was developed using the Anaconda distribution (Anaconda, Inc., Austin, US-TX) of python 2.7.13 (64-bit) (Python Software Foundation, Beaverton, US-OR) [Van Rossum, 1995].

The following scientific libraries and modules were used: numpy 1.11.3 [Travis & Oliphant, 2006] and pandas 0.19.2 [McKinney, 2010] for numerical computation, scikit-image 0.13.0 [Van der Walt et al., 2014] and scipy.ndimage 2.0 [Jones et al., 2001] for image processing, scikit-learn 0.19.1 [Pedregosa et al., 2011] for machine learning, matplotlib 1.5.1 [Hunter, 2007] and seaborn 0.7.1 [Waskom et al., 2016] for plotting, networkx 1.11 [Hagberg et al., 2008] for graph-based work, tifffile 0.11.1 [Gohlke, 2016] for loading of TIFF images, and various scipy 1.0.0 [Jones et al., 2001] modules for different purposes. Parallelization was implemented using dask 0.15.4 [Dask Development Team, 2016].

Jupyter Notebooks (jupyter 1.0.0, notebook 5.3.1) [Kluyver et al., 2016] were utilized extensively for prototyping, workflow management and exploratory data analysis, whereas refactoring and other software engineering was performed in the Spyder IDE (spyder 3.2.4) [Raybaut et al., 2018]. Version control was managed with Git 2.12.2.windows.2 [Torvalds et al., 2018] and an internally hosted GitLab instance (GitLab, San Francisco, US-CA).

Image Preprocessing

Following AiryScan 3D deconvolution with 'auto' settings on the LSM880, images were converted to 8bit TIFF files using a custom macro for the Fiji distribution [Schindelin et al., 2012] of ImageJ 1.52g [Schneider et al., 2012]. The minimum and maximum values determining the intensity range prior to 8bit conversion were selected manually such that intensity clipping is avoided. Care was taken to apply the same values to all samples of a given marker to ensure consistency.

Samples with the *cxcr4b:NLS-tdTomato* nuclear label exhibited a degree of bleed-through into the *lyn-EGFP* membrane label channel. To prevent this from interfering with single-cell segmentation, we employed a linear unmixing scheme in which the contribution of *NLS-tdTomato* (C, the contaminant image) is removed from the green channel (M, mixed image), resulting in the cleaned membrane channel (U, unmixed image). Our approach assumes that the signal in M is composed according to equation 1, implying that U can be retrieved by subtraction of an appropriate contamination term (eq. 2).

$$M = U + a \cdot C \quad \text{(eq. 1)}$$

$$U = M - a \cdot C \quad \text{(eq. 2)}$$

To compute the optimal bleed-through factor a we minimized a custom loss function (eq. 3), which is essentially simply the Pearson Correlation Coefficient (PCC) of the contaminant image C and the cleaned image U given a particular candidate factor a_i. To ensure that unreasonably high values of a are punished, we centered the values of the cleaned image onto their mean and converted the result to absolute values, causing overly unmixed regions to start correlating with C again.

$$loss = PCC(C,\ |M - a_i \cdot C - mean(M - a_i \cdot C)|) \quad \text{(eq. 3)}$$

We found that this approach robustly removes *NLS-tdTomato* bleed-through, producing unmixed images that could be segmented successfully.

Single-Cell Segmentation

3D single-cell segmentation was performed on membrane-labeled stacks acquired, deconvolved and preprocessed as detailed in the sections above.

The pipeline for segmentation consists of the following steps, applied sequentially:

1. 3D median smoothing with a cuboid *3x3x3vxl* structural element to reduce shot noise.
2. 3D Gaussian smoothing with σ=*3pxl* to further reduce noise and smoothen structures.
3. Thresholding to retrieve a binary mask of foreground objects (i.e. the membranes).
   To automatically determine the appropriate threshold, we use a custom function inspired by a semi-manual approach for spot detection [Raj et al., 2008]. Starting from the most frequent value in the image histogram as a base threshold, we iteratively scan a limited range of positive offsets (usually 0 to 10 in steps of 1, for slightly lower-quality images 0 to 40 in steps of 2) and count the number of connected components in the inverse of the binary mask resultant from applying each threshold. This roughly represents the number of cell bodies in the stack that are fully enclosed in cell membranes at a given threshold. We consider the threshold producing the largest number the best option and use it to generate the final membrane mask.
4. Removal of disconnected components by morphological hole filling.
5. Labeling of connected components on the inverted membrane mask. This ideally yields one connected component per cell, i.e. the cytoplasm.
6. Removal of connected components smaller than 1'000 voxels (artifacts) and re-labeling of connected components larger than 1'000'000 voxels as background objects.
7. Watershed expansion using the labeled connected components as seeds and the smoothed input image as topography (with additional 3D Gaussian smoothing on top of steps 1 and 2, with *σ=3pxl*). The background objects surrounding the primordium are also expanded.
8. Assignment of the zero label to background objects and removal of any objects disconnected from the primordium by retaining only the single largest foreground object.

We manually optimized the parameters of this pipeline for our data by inspecting the output during an extensive set of test runs.

Finally, we also manually double-checked all segmentations and discarded rare cases where many cells had been missed or where several instances of under- or oversegmentation could be observed.

Point Cloud Sampling with ISLA

Intensity-biased Stochastic Landmark Assignment (ISLA) (figure 3A-C, suppl. figure 1A) was applied to cropped-out bounding boxes of single segmented cells. To capture cell shape, the 6-connected inner hull of the binary segmentation mask was used as input image for ISLA. To capture intensity distributions, voxels outside the segmentation mask were set to zero and a simple background subtraction was performed to prevent landmarks from being assigned spuriously due to background signal. The background level was determined as the mean intensity within the masked cell and was subtracted from each voxel's intensity value, with resulting negative values set to zero.

With the inputs so prepared, voxel intensities were normalized such that their sum equals 1 by dividing each by the sum of all. Then, landmarks were assigned by considering the normalized voxel intensities as the probabilities of a multinomial distribution from which 2000 points were sampled (with replacement). This number was determined through test runs in which we found that, for volumes of our size, diminishing returns set in around 500 points and using more than 2000 points no longer improved the performance of various downstream analyses.

Following ISLA sampling, landmark coordinates were scaled from pixels to microns to account for anisotropic image resolution.

Note we recently published another study that utilizes a simplified version of ISLA for some of the data analysis, albeit in a very different way from how it is used here [Wong et al., 2020].

Point Cloud Transformation into TFOR and CFOR

To place cells in a matched *Tissue Frame of Reference* (TFOR) prior to feature embedding (suppl. figure 1B, left route), primordia were aligned using a simple PCA-based approach that does not require full image registration. To this end, 3000 landmarks were sampled from a given primordium's binary overall segmentation mask using ISLA and a modified PCA was applied to the resulting point cloud. Given that the pLLP's longest axis is always its front-rear axis and the shortest axis is always the apicobasal axis, PCA transformation snaps primordia that had been acquired at a slight angle into a consistent frame of reference. Our modification ensured that 180° flipping could not occur in this procedure. To complete the alignment, the primordial point clouds were translated such that the frontal-most point becomes the coordinate system's origin. Finally, the ISLA point clouds extracted from individual cells as described in the previous paragraph were transformed using the same PCA, thus matching their orientation to the common tissue frame of reference.

To create a *Cell Frame of Reference* (CFOR) that is invariant to size and rotation (suppl. figure 1B, right route), point cloud volumes were first normalized such that the sum of the magnitudes of all centroid-to-landmark vectors is 1. Second, cellular point clouds were cast into a pairwise distance (PD) representation. In the PD space, each point of the cloud is no longer characterized by three spatial coordinates but instead by the distances to every other point of the cloud. This representation is rotationally invariant but also extremely high-dimensional (an LxL array, where L is the number of landmarks). To reduce dimensionality, only the 10th, 50th and 90th percentiles of all pairwise distances for each point were chosen to represent that point (resulting in an Lx3 array), which we reasoned would encode both local and global relative spatial location.

Feature Embedding with CBE

To determine reference cluster centers for *Cluster-Based Embedding* (CBE) (figure 3D, suppl. figure 1B), point clouds from multiple samples were centered on their respective centroids and overlaid. K-means clustering was performed on this overlaid cloud (using scikit-learn's *MiniBatchKMeans* implementation) with *k=20*. The resulting cluster centers were used as common reference points for the next step.

Several measures were taken to improve the robustness and performance of this cluster detection. First, individual cellular point clouds were downsampled from 2000 points to 500 points prior to being overlaid, using k-means clustering with *k=500* clusters, the centers of which were used as the new landmarks. Second, not all available cells were used in the overlay. Instead, a representative random subset of primordia (at least 10, at most 25) was selected and only their cells were used in the overlay. The resulting cluster centers were used as reference points across all available samples. Third, the entire overlaid point cloud was downsampled using a density-dependent downsampling approach inspired by Qiu et al., 2011 (see below), yielding a final overlaid cloud of at most 200'000 points, which allowed reference cluster centers to be computed reasonably efficiently.

Density-dependent downsampling was performed using a simplified version of the algorithm described by Qiu and colleagues [Qiu et al., 2011]. First, the local density (LD) at each point is found, which is defined as the number of points in the local neighborhood, i.e. a sphere whose radius is the median pairwise distance between all points multiplied by an empirically determined factor (here 5). Next, a target density (TD) is determined, which in accordance with Qiu et al. was set to be the third percentile of all local densities. Now, points are downsampled such that the probability of keeping each point is given by equation 4. If necessary, the resulting downsampled distribution is further reduced by random sampling in order to reach the maximum of 200'000 points.

$$p(keep\_cell\_i) = \begin{cases} 1, & if\ LD_i < TD \\ \frac{TD}{LD_i}, & otherwise \end{cases} \quad \text{(eq. 4)}$$

Density-dependent downsampling was chosen to avoid cases where high-density agglomerations of landmarks in a particular region accumulate multiple clusters and thus deplete lower-density regions of local reference points; density-dependent downsampling preserves the overall shape of the overlaid point cloud whilst reducing local density peaks.

Following the determination of common reference points, CBE proceeds by extracting features describing the local landmark distribution around said reference points for each separate cellular point cloud. Several such features were implemented, including the number of landmarks in the local neighborhood of each reference point, the number of landmarks assigned to each reference point by the k-means clustering itself, the local density of landmarks at each reference point determined by a Gaussian Kernel Density Estimate (KDE), and either the magnitude or the components of the vector connecting each reference point to the centroid of its 25 nearest neighbors. The results were similar across all of these approaches but we ultimately chose to proceed with the last entry in the above list (the vector components; see figure 3D for a simple 2D example) based on its inclusion of some additional directional information.

The feature extraction described above yields an *n-by-3k* latent feature space, where *n* is the number of cells and *k* the number of shared reference clusters (here *k=20*). As a final step, this space was transformed by Principal Component Analysis (PCA) to bring relevant variation into focus and reduce dimensionality.

In addition to CBE, an alternative embedding based on the moments of the TFOR or CFOR point clouds was also generated as a comparably simple baseline. We computed the 1st raw moments (eq. 5), the 2nd centralized moments (eq. 6) and the 3rd to 5th normalized moments (eq. 7) (55 features in total) and once again used PCA to arrive at a compact and expressive feature space.

$$rM_{1[ijk]} = mean\left(C_z \cdot i + C_y \cdot j + C_x \cdot k\right) \quad \text{(eq. 5)}$$

$$cM_{2[ijk]} = mean\left(\left(C_z - rM_{1[100]}\right)^i \cdot \left(C_y - rM_{1[010]}\right)^j \cdot \left(C_x - rM_{1[001]}\right)^k\right) \quad \text{(eq. 6)}$$

$$nM_{m[ijk]} = \frac{cM_{m[ijk]}}{std\left(C_z - rM_{1[100]}\right)^i \cdot std\left(C_y - rM_{1[010]}\right)^j \cdot std\left(C_x - rM_{1[001]}\right)^k} \quad \text{(eq. 7)}$$

In equations 5-7, C_d is the array of all point cloud coordinates along the spatial dimension d, M_m is the set of raw (rM_m), centralized (cM_m) or normalized (nM_m) moments of the m-th order, and $[i, j, k]$ includes all combinations of length 3 drawn from the integer range $[0, \dots, m]$ that satisfy $i + j + k = m$. All operations are element-wise except $mean(\dots)$ and $std(\dots)$, which compute the mean and standard deviation across a given array.

Synthetic Point Cloud Generator

In order to benchmark the capability of CBE for latent feature extraction, we required a gold standard dataset to test whether CBE recovers known latent parameters underlying a population of

shape objects. We thus wrote a generator to automatically create a large synthetic dataset of cell-like point clouds. This generator functions in 4 stages:

1. Generate a spine (5 parameters)
   The spine is defined by its height along the z dimension (i.e. the apicobasal axis), by an offset, which is the distance and angle by which the end point of the spine is shifted in xy compared to the starting point, and by a connecting line between endpoints, which is a 2^{nd} degree polynomial.
2. Generate the cell surface (9 parameters)
   The surface of the cell is determined based on its distance *d* to the spine as a function of *z*. We used the logit-normal as this distance function on the grounds that it has only two parameters, can be asymmetric, monomodal or bimodal, crosses *(0, 0)* and *(1, 0)*, and has a range *d>0* within the domain *0>z>1*. For each cell, three different logit-normals were used to describe the surface at three uniformly spaced angles around the spine. Additionally, each of those functions was independently scaled by multiplication with another parameter.
3. Sampling of point clouds (0 parameters)
   The actual surface point cloud to be used in feature embedding was generated by sampling 2000 points with the following scheme: first randomly sample a position along *z* and an angle, then determine the corresponding radius by linear interpolation between the logit-normal values of the two adjacent angles at which they are defined. Finally, angle and radius are converted to the Cartesian coordinate system.
4. Centering, scaling and rotation (3 parameters)
   Finally, point clouds are centered on their centroid and cell size is adjusted by multiplication with a scaling parameter, followed by rotation using a homogeneous transformation matrix with two angle parameters.

All in all, this generator requires 17 parameters of which 1 modifies only size (not shape) and 2 modify only rotation. For each synthesized cell, the specific parameter values for the generator were sampled from normal or uniform distributions with hyperparameters set based on empirical experimentation such that a varied population of cell shapes is generated and unreasonably deformed shapes are avoided.

Evaluation of CBE on Synthetic Point Clouds

Using the generator described above, we synthesized a set of *n=20'000* cellular point clouds and embedded them using both CBE and the moment-based alternative embedding strategy. We then tested how well the values of the 17 known generative parameters could be retrieved from the embedded feature spaces using different scikit-learn implementations of multivariate-multivariable regression, specifically k-Nearest Neighbor (kNN) regression, multi-output linear Support Vector Regression (l-SVR) and multi-output radial basis function kernel Support Vector Regression (r-SVR). Optimal hyperparameters for the two SVRs were determined using cross-validated grid search with

GridSearchCV, scanning 5 orders of magnitude surrounding the defaults of C (penalty) and epsilon as well as gamma (RBF kernel coefficient) for r-SVR.

This synthetic experiment showed that cell size and orientation largely obscure other shape features if CFOR-normalization is not performed (suppl. figure 2A). Furthermore, features generated by CBE enable similar or better predictive performance than moments-based features regardless of the method used for prediction (suppl. figure 2B). Interestingly, kNN performs markedly better on CBE-embedded spaces, indicating that CBE more meaningfully encodes point cloud similarity as local neighborhood in the embedded space.

Feature Engineering: Extraction of Simple Shape Measures

We extracted various explicitly engineered features from entire tissues, segmented cell volumes, and segmented cell point clouds. The features used in this study are listed and briefly described in supplementary table 2.

Multi-Channel Atlas Prediction

To construct the multi-channel atlas, we used the secondary markers present in many of our samples (see supplementary table 1), embedded them with ISLA and CBE (see the corresponding sections above) and then trained multivariate-multivariable regressors to predict those embeddings based on the corresponding cell shape embeddings as input features. All embeddings were standardized and PCA-transformed prior to machine learning and only the first 20 PCs were considered. Predictions were performed from shape TFOR to secondary marker TFOR and from shape CFOR to secondary marker CFOR. Note that, because expression of the secondary markers was sometimes heterogeneous across the primordium, only cells with a secondary marker intensity above the 33rd percentile were used as training data.

To select the best machine learning model, the following regressors were tested using 5-fold cross-validation: k-nearest neighbors regression (*sklearn.neighbors.KNeighborsRegressor*), random forest regression (*sklearn.ensemble.RandomForestRegressor*), elastic net regression (*sklearn.linear_model.ElasticNet*), Lasso regression (*sklearn.linear_model.Lasso*), a multi-layer perceptron (*sklearn.neural_network.MLPRegressor*), and a support vector regressor with an RBF-kernel (*sklearn.svm.SVR*). Relevant hyperparameters were optimized on the *NLS-tdTomato* nuclear marker using a 5-fold cross-validated grid search (*sklearn.model_selection.GridSearchCV*) of 5 orders of magnitude surrounding the scikit-learn defaults. Performance was evaluated across different secondary channels and latent feature embeddings (suppl. figure 3A), with the primary aim being high explained variance but also giving some consideration to computational efficiency (training and prediction time). After finding that TFOR predictions perform substantially better than CFOR predictions, possibly because the TFOR shape space incorporates more relevant information for intracellelular marker prediction and/or because CFOR spaces contain more noise and more specific

information that is challenging to predict, we focused further analysis on TFOR for the time being. We selected the RBF SVR as the regressor of choice because of its consistently high performance.

For atlas prediction across the entire dataset, the SVR model was trained for each secondary channel (using the corresponding best hyperparameter set) on all available data for that channel and then applied to predict that channel's embedded space for all other cells.

smFISH: Spot Detection and Analysis

Single-cell segmentation and cell shape feature embedding was performed on the *pea3* smFISH dataset in the same fashion as for the live imaging dataset, resulting in 3'149 cells from 31 samples.

The *blob_log* spot detector from scikit-image was used (*skimage.blob.blob_log*) to identify smFISH spots (suppl. figure 4A-D). Parameters were optimized by manual testing and by scanning different values for the *threshold* and *min_sigma* parameters, arriving ultimately at *min_sigma=1*, *max_sigma=4*, *num_sigma=10*, *threshold=0.22*, *overlap=0.5*, and *log_scale=False*, which produced average counts per cell that are reasonably consistent across different primordia (suppl. figure 4E) and closely approximate previously reported *pea3* smFISH spot counts in the pLLP [Durdu et al., 2014]. However, we also found that optimal settings can vary between different smFISH experiments and thus recommend parameter optimization and careful evaluation of the results for every experiment. Furthermore, in this dataset the spot detector failed to detect a reasonable number of spots in some exceptional samples, so we excluded primordia with a mean count per cell of less than 2 spots from further analysis, leaving 2906 cells from 29 primordia in the dataset.

Because smFISH must be performed on fixed samples, we checked for fixation effects on cell shape, which would make cross-predictions with the live sample atlas more challenging. In bulk comparisons of key shape features (suppl. figure 4F-H) we found no significant differences between live and fixed samples. However, such bulk comparisons may miss subtle fixation effects, making the development of methods to quantify, pinpoint and correct such effects an important future goal.

We trained an RBF-kernel SVR to perform multivariate regression of *pea3* smFISH counts based on as much other information available about each cell as possible (figure 6B, suppl. figure 4E), namely a combined feature space incorporating the first 10 shape TFOR PCs, the first 10 shape CFOR PCs, and the x, y and z coordinates of cell centroids in TFOR. Hyperparameters C (penalty), epsilon, and gamma (RBF kernel coefficient) were again optimized using *GridSearchCV*.

Prediction and Visualization of Morphological Archetypes

The four archetypes were manually annotated in 26 primordia (figure 7A) using Fiji's multi-point selection tool, yielding 93 leader cells, 241 outer rosette cells, 182 inner rosette cells and 108 between-rosette cells (624 cells in total). Only the most clear examples of the respective archetypes were labeled.

Cell archetype prediction was performed using scikit-learn's Support Vector Classifier (SVC) with a Radial Basis Function (RBF) kernel, using either TFOR- or CFOR-embedded cell shapes as input features. We optimized the SVC hyperparameters for each embedding using scikit-learn's *GridSearchCV* with 5-fold cross-validation, testing whether or not to use feature standardization, whether or not to use PCA and keep the first 15, 30 or 50 PCs, as well as screening 5 orders of magnitude surrounding the scikit-learn default values for C (penalty) and gamma (RBF kernel coefficient). The resulting best estimator was used for all further training and prediction.

The confusion matrices in supplementary figure 5 were produced by randomly splitting the annotated cells into a training set (436 cells) and a test set (188 cells). However, predictions for the entire atlas dataset were generated following training with all 624 manually annotated cells.

The archetype space was constructed by inferring the classification probabilities for each class (using *sklearn.svm.SVC.predict_proba*) and performing a PCA on them. The 3D and 2D visualizations in figure 7 were then generated by plotting the first three or two principal components, respectively.

Image Rendering and Expanded View of Segmentation

Fiji's *Straighten* tool was used to align angled samples with the main image axes for the purpose of illustration in figures 1, 2, 5, 6, 7 and in movies 1-3 but never as part of an image analysis pipeline. Maximum projections were created with Fiji whereas 3D movies were rendered with Imaris 7.7.2 (Bitplane, Belfast, UK).

The expanded view of the segmented pLLP (figures 1B, 2C) was generated by first determining the centroids of each segmented cell and then shifting them apart by scaling of their x and y coordinates by a single user-specified factor. The cells were then pasted into an appropriately scaled empty image stack at the new centroid locations, leaving them shifted apart uniformly but not individually rescaled or otherwise transformed. A python implementation of this approach called *tissueRipper* is available under the MIT open source license on GitHub at *github.com/WhoIsJack/tissueRipper*.

Correlation Heatmaps and Bigraphs

The corresponding bigraphs (figures 4E,F and 5G) were generated using a custom plotting function based on the networkx module. The edges were colored according to the signed value of the Pearson correlation coefficient and sized according to its magnitude. Edges with an absolute correlation coefficient smaller than 0.3 were omitted. The nodes of the engineered features were sorted to reduce edge crossings and group similar nodes, which was achieved by minimizing the following custom loss function:

$$loss = \sum_{i=0}^{f_E} \sum_{j=0}^{f_L} \left| \frac{C_i}{f_E} - \frac{j}{f_E} \right| \cdot \left| pcc\left(E_i, L_j\right) \right| \qquad \text{(eq. 8)}$$

where f_E and f_L are the number of engineered and latent features, respectively. C is the current sort order of the engineered features, i.e. a permutation of the integer interval $[0, f_E]$. Finally,

$\left|pcc\left(E_i, L_j\right)\right|$ is the absolute Pearson correlation coefficient of the values of the i-th engineered feature and the j-th latent feature. In essence, this loss function is the sum of all Euclidean rank distances between engineered and latent features, weighted by their corresponding absolute Pearson correlation coefficients. Minimization was performed by random shuffling of the sort order and retaining only shuffles that reduced the loss until no change was observed for 2000 consecutive shuffles.

Note that since the sign of principal components is not inherently meaningful, we flipped it for shape TFOR-PCs 1, 3, 5, 6 and shape CFOR-PC 1 across all analyses presented in this study to ensure that PCs positively correlate with their most defining engineered feature(s), facilitating discussion of the results.

Tissue Consensus Maps

Consensus maps of feature variation (figures 4G, 5J, 6C,D) were based on an overlay of TFOR centroid positions of cells across all relevant samples.

The cut-off for the consensus tissue outline was determined by computing the local density of these overlaid centroids using scipy's Gaussian kernel density estimation with default settings. Regions with densities below 10% of the range between the minimum and maximum density were considered outside of the primordium and are shown as white in the plot.

The local consensus feature values were computed by applying a point cloud-based Gaussian smooth across the individual cells' feature values, using the 0.5th percentile of all pairwise distances between centroids as σ. This smoothed distribution was then plotted as a *tricontourf* plot with matplotlib using up to 21 automatically determined contour levels.

Statistical Analysis

Generally, we use *N* to refer to the number of embryos/primordia and *n* to the number of cells.

Statistical significance for comparisons between two conditions was estimated without parametric assumptions using a two-tailed Mann-Whitney U test (*scipy.stats.mannwhitneyu* with keyword argument *alternative='two-sided'*) with Bonferroni multiple testing correction [Haynes, 2013] where appropriate. We considered $p>0.01$ as not statistically significant.

Significance tests with large sample sizes such as those encountered during single-cell analysis tend to indicate high significance regardless of whether the difference between populations is substantive or technical [Sullivan & Feinn, 2012], which is why we also report estimates of effect size for any comparison that is statistically significant. Effect sizes were estimated using Cohen's d [Cohen, 1977]. The resulting values can be described as *no effect* (d≈0.0), a *small effect* (d≈0.2), a *medium effect* (d≈0.5) or a *large effect* (d≈0.8) [Cohen, 1977].

Materials Availability

Requests for experimental resources and reagents should be directed to and will be fulfilled by Darren Gilmour (darren.gilmour@uzh.ch)

Data and Code Availability

All data and code produced in this study will be made openly available following peer review. We also aim to update the core algorithms to python 3 and make them available as a readily reusable module. Inquiries regarding data and code should be directed to Jonas Hartmann (jonas.m.hartmann @protonmail.com).

ACKNOWLEDGEMENTS

We thank Sabine Görgens and Andreas Kunze for their support with fish and lab maintenance, respectively. We thank the EMBL Advanced Light Microscopy Facility (ALMF) and the UZH Center for Microscopy and Image Analysis (ZMB) for maintenance of and assistance with microscopes. We thank Alejandra Guzman Herrera for generating the *Act2b:mKate2-Rab11a* line. We thank Francesca Peri and Stefano De Renzis for kindly providing temporary lab space. We thank Christian Tischer and Marvin Albert for helpful discussion on image analysis and numerical computation. We thank Stefano De Renzis, Daniel Krueger and Marvin Albert for critical reading of the manuscript. J.H. and E.G. were supported by the EMBL International PhD Programme (EIPP), M.W. was supported by an EMBO Long-Term Fellowship and the EMBL Interdisciplinary Postdoc (EIPOD) Program under Marie Curie COFUNDII Actions. The Gilmour lab was supported by the European Molecular Biology Laboratory (EMBL), the University of Zurich (UZH) and Swiss National Science Funds Grant 31003A_176235.

AUTHOR CONTRIBUTIONS

J.H. and D.G. conceptualized the study. J.H. and M.W. performed molecular biology and live imaging experiments. E.G. performed smFISH experiments. J.H. wrote the code, performed the data analysis, interpreted the results with inputs from D.G. and E.G., and wrote the manuscript with feedback from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

MAIN TEXT REFERENCES

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2013): *A field comes of age: geometric morphometrics in the 21st century*. Hysterix, 24(1), 7-14. doi:10.4404/hystrix-24.1-6283

Aman, A., & Piotrowski, T. (2008). *Wnt/β-Catenin and Fgf Signaling Control Collective Cell Migration by Restricting Chemokine Receptor Expression*. Developmental Cell, 15(5), 749–761. doi:10.1016/j.devcel.2008.10.002

Angerer, P., Simon, L., Tritschler, S., Wolf, F. A., Fischer, D., & Theis, F. J. (2017). *Single cells make big data: New challenges and opportunities in transcriptomics*. Current Opinion in Systems Biology, 4, 85–91. doi:10.1016/j.coisb.2017.07.004

Baker, R. E., Peña, J.-M., Jayamohan, J., & Jérusalem, A. (2018). *Mechanistic models versus machine learning, a fight worth fighting for the biological community?*. Biology Letters, 14(5), 20170660. doi:10.1098/rsbl.2017.0660

Battich, N., Stoeger, T., & Pelkmans, L. (2013). *Image-based transcriptomics in thousands of single human cells at single-molecule resolution*. Nature Methods, 10(11), 1127–1133. doi:10.1038/nmeth.2657

Bizzarri, M., Palombo, A., & Cucina, A. (2013). *Theoretical aspects of Systems Biology*. Progress in Biophysics and Molecular Biology, 112(1-2), 33–43. doi:10.1016/j.pbiomolbio.2013.03.019

Blei, D. M., & Smyth, P. (2017). *Science and data science*. Proceedings of the National Academy of Sciences, 114(33), 8689–8692. doi:10.1073/pnas.1702076114

Brodland, G. W. (2002). *The Differential Interfacial Tension Hypothesis (DITH): A Comprehensive Theory for the Self-Rearrangement of Embryonic Cells and Tissues*. Journal of Biomechanical Engineering, 124(2), 188. doi:10.1115/1.1449491

Cai, Y., Hossain, M. J., Hériché, J.-K., Politi, A. Z., Walther, N., Koch, B., Wachsmuth, M., Nijmeijer, B., Kueblbeck, M., Martinic-Kavur, M., Ladurner, R., Alexander, S., Peters, J.-M., & Ellenberg, J. (2018). Experimental and computational framework for a dynamic protein atlas of human cell division. Nature. doi:10.1038/s41586-018-0518-z

Callebaut, W. (2012). *Scientific perspectivism: A philosopher of science's response to the challenge of big data biology*. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43(1), 69–80. doi:10.1016/j.shpsc.2011.10.007

Chan, C. J., Heisenberg, C.-P., & Hiiragi, T. (2017). *Coordination of Morphogenesis and Cell-Fate Specification in Development*. Current Biology 27(18), R1024–35. doi:10.1016/j.cub.2017.07.010

Chan, C., Theriot, J., Dunn, A. R., Rohatgi, R., & Straight, A. (2018): *Flexible, comprehensive frameworks for quantitative analysis of cell shape and subcellular organization in the context of cell motility*. Stanford University, California. purl.stanford.edu/zw934wr1862

Chessel, A., & Salas, R. E. C. (2019). *From observing to predicting single-cell structure and function with high-throughput/high-content microscopy*. Essays In Biochemistry, EBC20180044. doi:10.1042/EBC20180044

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., Cofer, E. M., Lavender, C. A., Turaga, S. C., Alexandari, A. M., Lu, Z., Harris, D. J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L. K., Segler, M. H. S., Boca, S. M., Swamidass, S. J., Huang, A., Gitter, A., & Greene, C. S. (2018). *Opportunities and obstacles for deep learning in biology and medicine*. Journal of The Royal Society Interface, 15(141), 20170387. doi:10.1098/rsif.2017.0387

Christiansen, E. M., Yang, S. J., Ando, D. M., Javaherian, A., Skibinski, G., Lipnick, S., Mount, E., O'Neil, A., Shah, K., Lee, A. K., Goyal, P., Fedus, W., Poplin, R., Esteva, A., Berndl, M., Rubin L. L., Nelson, P., & Finkbeiner, S. (2018). *In Silico Labeling: Predicting Fluorescent Labels in Unlabeled Images*. Cell, 173(3), 792–803.e19. doi:10.1016/j.cell.2018.03.040

Dhar, V. (2013). *Data science and prediction*. Communications of the ACM, 56(12), 64–73. doi:10.1145/2500499

Driscoll, M. K., Welf, E. S., Jamieson, A. R., Dean, K. M., Isogai, T., Fiolka, R., & Danuser, G. (2019). *Robust and automated detection of subcellular morphological motifs in 3D microscopy images*. Nature Methods. doi:10.1038/s41592-019-0539-z

Durdu, S., Iskar, M., Revenu, C., Schieber, N., Kunze, A., Bork, P., Schwab, Y., & Gilmour, D. (2014). *Luminal signalling links cell communication to tissue architecture during organogenesis*. Nature, 515(7525), 120–124. doi:10.1038/nature13852

Galanternik, M. V., Acedo, J. N., Romero-Carvajal, A., & Piotrowski, T. (2016). *Imaging collective cell migration and hair cell regeneration in the sensory lateral line*. The Zebrafish - Cellular and Developmental Biology, Part B Developmental Biology, 211–256. doi:10.1016/bs.mcb.2016.01.004

Ghysen, A., & Dambly-Chaudiere, C. (2007). *The lateral line microcosmos*. Genes & Development, 21(17), 2118–2130. doi:10.1101/gad.1568407

Gilmour, D., Rembold, M., & Leptin, M. (2017). *From morphogen to morphogenesis and back*. Nature, *541*(7637), 311. doi:10.1038/nature21348

Grant, K. A., Raible, D. W., & Piotrowski, T. (2005). *Regulation of Latent Sensory Hair Cell Precursors by Glia in the Zebrafish Lateral Line*. Neuron, 45(1), 69–80. doi:10.1016/j.neuron.2004.12.020

Gut, G., Herrmann, M. D., & Pelkmans, L. (2018). *Multiplexed protein maps link subcellular organization to cellular states. Science*, 361(6401), eaar7042. doi:10.1126/science.aar7042

Haas, P., & Gilmour, D. (2006). *Chemokine Signaling Mediates Self-Organizing Tissue Migration in the Zebrafish Lateral Line*. Developmental Cell, 10(5), 673–680. doi:10.1016/j.devcel.2006.02.019

Hernández, P. P., Olivari, F. A., Sarrazin, A. F., Sandoval, P. C., & Allende, M. L. (2007). *Regeneration in zebrafish lateral line neuromasts: Expression of the neural progenitor cell marker sox2 and proliferation-dependent and-independent mechanisms of hair cell renewal*. Developmental Neurobiology, 67(5), 637–654. doi:10.1002/dneu.20386

Holzinger, A., Dehmer, M., & Jurisica, I. (2014). *Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions*. BMC Bioinformatics, 15(Suppl 6), I1. doi:10.1186/1471-2105-15-s6-i1

Huang, S., Chaudhary, K., & Garmire, L. X. (2017). *More Is Better: Recent Progress in Multi-Omics Data Integration Methods*. Frontiers in Genetics, 8. doi:10.3389/fgene.2017.00084

Huff, J. (2016): *The Fast mode for ZEISS LSM 880 with Airyscan: high-speed confocal imaging with super-resolution and improved signal-to-noise ratio*. Carl Zeiss Microscopy Application Note, Nature Methods 13(11), i-ii. doi:10.1038/nmeth.f.398

Johnson, G. R., Buck, T. E., Sullivan, D. P., Rohde, G. K., & Murphy, R. F. (2015). *Joint modeling of cell and nuclear shape variation*. Molecular Biology of the Cell, 26(22), 4046–4056. doi:10.1091/mbc.e15-06-0370

Johnson, G. R., Donovan-Maiye, R. M., & Maleckar, M. M. (2017): *Building a 3D Integrated Cell*, preprint on bioRxiv, www.biorxiv.org/content/10.1101/238378v1. doi:10.1101/238378

Kalinin, A. A., Allyn-Feuer, A., Ade, A., Fon, G.-V., Meixner, W., Dilworth, D., Husain, S. S., de Wet, J. R., Higgins, G. A., Zheng, G., Creekmore, A., Wiley, J. W., Verdone, J. E., Veltri, R. W., Pienta, K. J., Coffey, D. S., Athey, B. D., & Dinov, I. D. (2018). *3D Shape Modeling for Cell Nuclear Morphological Analysis and Classification*. Scientific Reports, 8(1). doi:10.1038/s41598-018-31924-2

Lecaudey, V., Cakan-Akdogan, G., Norton, W. H. J., & Gilmour, D. (2008). *Dynamic Fgf signaling couples morphogenesis and migration in the zebrafish lateral line primordium*. Development, 135(16), 2695–2705. doi:10.1242/dev.025981

Lecuit, T., & Lenne, P.-F. (2007). *Cell surface mechanics and the control of cell shape, tissue patterns and morphogenesis*. Nature Reviews Molecular Cell Biology, 8(8), 633–644. doi:10.1038/nrm2222

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Yang, J. L., Ferrante, T. C., Terry, R., Jeanty, S. S. F., Li, C., Amamoto, R., Peters, D. T., Turczyk, B. M., Marblestone, A. H., Inverso, S. A., Bernard, A., Mali, P., Rios, X., Aach, J., & Church, G. M. (2014). *Highly Multiplexed Subcellular RNA Sequencing in Situ*. Science, 343(6177), 1360–1363. doi:10.1126/science.1250212

Leonelli, S. (2019). *The challenges of big data biology*. eLife, 8, e47381. 10.7554/eLife.47381

Libbrecht, M. W., & Noble, W. S. (2015). *Machine learning applications in genetics and genomics*. Nature Reviews Genetics, 16(6), 321–332. doi:10.1038/nrg3920

Marcus, G. (2018): *Deep Learning: A Critical Appraisal*, preprint on arXiv, arxiv.org/abs/1801.00631v1. arXiv:1801.00631v1

Matzke, E. B. (1946). *The Three-Dimensional Shape of Bubbles in Foam-An Analysis of the Role of Surface Forces in Three-Dimensional Cell Shape Determination*. American Journal of Botany, 33(1), 58. doi:10.2307/2437492

McDole, K., Guignard, L., Amat, F., Berger, A., Malandain, G., Royer, L. A., Turuga, S. C., Branson, K., & Keller, P. J. (2018). *In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level*. Cell 175, 859–876. doi:10.1016/j.cell.2018.09.031

Nechiporuk, A., & Raible, D. W. (2008). *FGF-dependent mechanosensory organ patterning in zebrafish*. Science, 320(5884), 1774-1777. doi:10.1126/science.1156547

Nogare, D. D., Nikaido, M., Somers, K., Head, J., Piotrowski, T., & Chitnis, A. B. (2017). *In toto imaging of the migrating Zebrafish lateral line primordium at single cell resolution*. Developmental Biology, 422(1), 14–23. doi:10.1016/j.ydbio.2016.12.015

Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F., & Johnson, G. R. (2018). *Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy*. Nature Methods. doi:10.1038/s41592-018-0111-2

Pasqualato, A., Palombo, A., Cucina, A., Mariggiò, M. A., Galli, L., Passaro, D., Dinicola, S., Proietti, S., D'Anselmi, F., Coluccia, P., Bizzarri, M. (2012). *Quantitative shape analysis of chemoresistant colon cancer cells: Correlation between morphotype and phenotype*. Experimental Cell Research, 318(7), 835–846. doi:10.1016/j.yexcr.2012.01.022

Peng, H., Chung, P., Long, F., Qu, L., Jenett, A., Seeds, A. M., Myers, E. W., & Simpson, J. H. (2011). *BrainAligner: 3D registration atlases of Drosophila brains*. Nature Methods, 8(6), 493–498. doi:10.1038/nmeth.1602

Peng, T., & Murphy, R. F. (2011). *Image-derived, three-dimensional generative models of cellular organization.* Cytometry Part A, 79A(5), 383–391. doi:10.1002/cyto.a.21066

Pincus, Z., & Theriot, J. A. (2007). *Comparison of quantitative methods for cell-shape analysis*. Journal of Microscopy, 227(2), 140–156. doi:10.1111/j.1365-2818.2007.01799.x

Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., & Plevritis, S. K. (2011). *Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE.* Nature Biotechnology, 29(10), 886–891. doi:10.1038/nbt.1991

Rajaram, S., Pavie, B., Wu, L. F., & Altschuler, S. J. (2012). *PhenoRipper: software for rapidly profiling microscopy images*. Nature Methods, 9(7), 635–637. doi:10.1038/nmeth.2097

Roukos, V., & Misteli, T. (2014). *Deep Imaging: the next frontier in microscopy*. Histochemistry and Cell Biology, 142(2), 125–131. doi:10.1007/s00418-014-1239-5

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). *Spatial reconstruction of single-cell gene expression data*. Nature Biotechnology, 33(5), 495–502. doi:10.1038/nbt.3192

Smutny, M., Behrndt, M., Campinho, P., Ruprecht, V., & Heisenberg, C.-P. (2014). *UV Laser Ablation to Measure Cell and Tissue-Generated Forces in the Zebrafish Embryo In Vivo and Ex Vivo*. Tissue Morphogenesis, 219–235. doi:10.1007/978-1-4939-1164-6_15

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). *Comprehensive Integration of Single-Cell Data*. Cell, 177(7), 1888–1902.e21. doi:10.1016/j.cell.2019.05.031

Tweedy, L., Meier, B., Stephan, J., Heinrich, D., & Endres, R. G. (2013). *Distinct cell shapes determine accurate chemotaxis*. Scientific Reports, 3(1). doi:10.1038/srep02606

Veldhuis, J. H., Ehsandar, A., Maître, J.-L., Hiiragi, T., Cox, S., & Brodland, G. W. (2017). *Inferring cellular forces from image stacks*. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1720), 20160261. doi:10.1098/rstb.2016.0261

Vergara, H. M., Bertucci, P. Y., Hantz, P., Tosches, M. A., Achim, K., Vopalensky, P., & Arendt, D. (2017). *Whole-organism cellular gene-expression atlas reveals conserved cell types in the ventral nerve cord of Platynereis dumerilii*. Proceedings of the National Academy of Sciences, 114(23), 5878–5885. doi:10.1073/pnas.1610602114

Viader-Llargués, O., Lupperger, V., Pola-Morell, L., Marr, C., & López-Schier, H. (2018). *Live cell-lineage tracing and machine learning reveal patterns of organ regeneration*. eLife, 7. doi:10.7554/elife.30823

Wang, M. F. Z., Hunter, M. V., Wang, G., McFaul, C., Yip, C. M., & Fernandez-Gonzalez, R. (2017). *Automated cell tracking identifies mechanically oriented cell divisions during Drosophila axis elongation*. Development, 144(7), 1350–1361. doi:10.1242/dev.141473

MATERIALS AND METHODS REFERENCES

Burkel, B. M., Von Dassow, G., & Bement, W. M. (2007*). Versatile fluorescent probes for actin filaments based on the actin-binding domain of utrophin*. Cell motility and the cytoskeleton, 64(11), 822-832. doi:10.1002/cm.20226

Cohen, J. (1977). *CHAPTER 2 - The t Test for Means*. in *Statistical power analysis for the behavioral sciences*, edited by J. Cohen, 19-74, Academic Press, USA. doi:10.1016/B978-0-12-179060-8.50007-4

Dask Development Team (2016). *Dask: Library for dynamic task scheduling*. dask.org

Distel, M., Wullimann, M. F., & Köster, R. W. (2009). *Optimized Gal4 genetics for permanent gene expression mapping in zebrafish*. Proceedings of the National Academy of Sciences, 106(32), 13365-13370. doi:10.1073/pnas.0903060106

Donà, E., Barry, J. D., Valentin, G., Quirin, C., Khmelinskii, A., Kunze, A., Durdu, S., Netwon, L. R., Fernandez-Minan, A., Huber, W., Knop, M., & Gilmour, D. (2013). *Directional tissue migration through a self-generated chemokine gradient*. Nature, 503(7475), 285–289. doi:10.1038/nature12635

Durdu, S., Iskar, M., Revenu, C., Schieber, N., Kunze, A., Bork, P., Schwab, Y., & Gilmour, D. (2014). *Luminal signalling links cell communication to tissue architecture during organogenesis*. Nature, 515(7525), 120–124. doi:10.1038/nature13852

Emelyanov, A., & Parinov, S. (2008). *Mifepristone-inducible LexPR system to drive and control gene expression in transgenic zebrafish*. Developmental biology, 320(1), 113-121. doi:10.1016/ j.ydbio.2008.04.042

Gohlke, C. (2016). *tifffile version 0.11.1*. pypi.org/project/tifffile

Huff, J. (2016): *The Fast mode for ZEISS LSM 880 with Airyscan: high-speed confocal imaging with super-resolution and improved signal-to-noise ratio*. Carl Zeiss Microscopy Application Note, Nature Methods 13(11), i-ii. doi:10.1038/nmeth.f.398

Haas, P., & Gilmour, D. (2006). *Chemokine Signaling Mediates Self-Organizing Tissue Migration in the Zebrafish Lateral Line*. Developmental Cell, 10(5), 673–680. doi:10.1016/j.devcel.2006.02.019

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX*. Los Alamos National Lab., Los Alamos, United States, osti.gov/servlets/purl/960616.

Haynes, W. (2013). *Bonferroni Correction*. in *Encyclopedia of Systems Biology*, 2013 edition, edited by W. Dubitzky, O. Wolkenhauer, K.-H. Cho, H. Yokota, 154–154. doi:10.1007/978-1-4419-9863-7_1213

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in science & engineering, 9(3), 90-95. doi:10.1109/MCSE.2007.55

Jones, E., Oliphant, T., & Peterson, P., and the SciPy Development Team (2001). *SciPy: Open Source Scientific Tools for Python*. www.scipy.org.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C., and the Jupyter Development Team (2016). *Jupyter Notebooks – a publishing format for reproducible computational workflows*. in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, edited by F. Loizides and B. Schmidt, 87-90. doi:10.3233/978-1-61499-649-1-87

Kwan, K. M., Fujimoto, E., Grabher, C., Mangum, B. D., Hardy, M. E., Campbell, D. S., Parant, J. M., Yost, H. J., Kanki, J. P., & Chien, C. B. (2007). *The Tol2kit: a multisite gateway-based construction kit for Tol2 transposon transgenesis constructs*. Developmental dynamics: an official publication of the American Association of Anatomists, 236(11), 3088-3099. doi:10.1002/dvdy.21343

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference, 51-56, conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research 12, 2825-2830, jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Pouthas, F., Girard, P., Lecaudey, V., Ly, T. B. N., Gilmour, D., Boulin, C., Pepperkok, R., & Reynaud, E. G. (2008). In migrating cells, the Golgi complex and the position of the centrosome depend on geometrical constraints of the substratum. Journal of Cell Science, 121(14), 2406–2414. doi:10.1242/jcs.026849

Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., & Plevritis, S. K. (2011). *Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE.* Nature Biotechnology, 29(10), 886–891. doi:10.1038/nbt.1991

Raj, A., Van Den Bogaard, P., Rifkin, S. A., Van Oudenaarden, A., & Tyagi, S. (2008). *Imaging individual mRNA molecules using multiple singly labeled probes*. Nature methods, 5(10), 877-879. doi: 10.1038/NMETH.1253

Raybaut, P., & Cordoba, C., and the Spyder project contributors (2018). *Spyder-IDE GitHub Repository*, github.com/spyder-ide/spyder

Shaner, N. C., Lin, M. Z., McKeown, M. R., Steinbach, P. A., Hazelwood, K. L., Davidson, M. W., & Tsien, R. Y. (2008). *Improving the photostability of bright monomeric orange and red fluorescent proteins*. Nature methods, 5(6), 545. doi:10.1038/NMETH.1209

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., &

Cordona, A. (2012). *Fiji: an open-source platform for biological-image analysis*. Nature methods, 9(7), 676-682. doi:10.1038/nmeth.2019

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). *NIH Image to ImageJ: 25 years of image analysis*. Nature methods, 9(7), 671-675. doi:10.1038/nmeth.2089

Sullivan, G. M., & Feinn, R. (2012). *Using effect size—or why the P value is not enough*. Journal of graduate medical education, 4(3), 279-282. doi:10.4300/JGME-D-12-00156.1

Torvalds, L., and the Git contributors (2018). *Git GitHub Repository*. github.com/git/git

Travis, E., & Oliphant, E. (2006). *A guide to NumPy*. Trelgol Publishing, USA, www.numpy.org.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). *scikit-image: image processing in Python*. PeerJ, 2, e453. doi:10.7717/peerj.453

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial (Vol. 620)*. CWI Report CS-R9526, Amsterdam, Netherlands, msekce.karlin.mff.cuni.cz/~halas/IT/tutorial.pdf.

Wada, H., Ghysen, A., Satou, C., Higashijima, S. I., Kawakami, K., Hamaguchi, S., & Sakaizumi, M. (2010). *Dermal morphogenesis controls lateral line patterning during postembryonic development of teleost fish*. Developmental biology, 340(2), 583-594. doi:10.1016/j.ydbio.2010.02.017

Waskom, M., Botvinnik, O., drewokane, Hobson, P., David, Halchenko, Y., Lukauskas, S., Cole, J. B., Warmenhoven, J., de Ruiter, J., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Martin, M., Miles, A., Meyer, K., Augspurger, T., Yarkoni, T., Bachant, P., Williams, M., Evans, C., Fitzgerald, C., Brian, Wehner, D., Hitz, G., Ziegler, E., Qalieh, A, & Lee, A. (2016): *seaborn: v0.7.1 (june 2016)*. zenodo, zenodo.org/record/54844. doi:10.5281/zenodo.54844

Westerfield, M. (2000). *The zebrafish book. A guide for the laboratory use of zebrafish (Danio rerio)*. 4th ed., University of Oregon Press, Eugene, zfin.org/zf_info/zfbook/zfbk.html.

Wong, M., Newton, L. R., Hartmann, J., Hennrich, M. L., Wachsmuth, M., Ronchi, P., Guzmán-Herrera, A., Schwab, Y., Gavin, A.-C., & Gilmour, D. (2020). *Dynamic buffering of extracellular chemokine enables robust adaptation during directed tissue migration*. Developmental Cell, 52, 1–17. doi:10.1016/j.devcel.2020.01.013

MAIN FIGURE TITLES AND LEGENDS

Figure 1: Overview of key steps in our data-driven analysis workflow

(A) Image data of the tissue of interest is acquired using 3D confocal fluorescence microscopy. Each sample is labeled with a membrane marker to delineate cell boundaries (top) and samples can additionally be labeled with various other markers of interest (bottom, colored). (B) Using an automated image analysis pipeline, single cells are automatically segmented based on the membrane marker to prepare them for analysis, illustrated here by shifting them apart. (C) Next, data extraction takes place to arrive at numerical features representing the cell shapes (yellow) and the various fluorescent protein distributions of additional markers (other colors). (D) Such well-structured data simplifies the application of machine learning techniques for data integration, which here is performed based on cell shape as a common reference measurement. (E) A similar strategy can be used to map manually annotated contextual knowledge (top) into the dataset (bottom), in this case specific cell archetypes chosen based on prior knowledge of the tissue's biology. (F) Finally, all of the resulting data is explored and interpreted through various visualizations and statistics.

Figure 2: Imaging and automated 3D single-cell segmentation of the pLLP

(A) Maximum z-projection of a deconvolved 3D volume of the pLLP acquired using the LSM880 AiryScan FAST mode. (B) The same primordium shown with a semi-transparent color overlay of the corresponding single-cell segmentation. (C) Expanded view of the same primordium; individual segmented cells have been shifted apart without being rescaled or deformed, revealing their individual shapes within the collective. Note that the segmentation faithfully recapitulates the diversity of cell shapes within the pLLP, with the exception of fine protrusions. Since the protrusions of follower cells are often impossible to detect against the membranes of the cells ahead of them, we decided not to include fine protrusions in our analysis. All scale bars: 10µm.

Figure 3: CBE and ISLA for Point Cloud-Based Cell Morphometry

(A) A classical workflow in landmark-based geometric morphometrics. (B) Adapted workflow for morphometrics of arbitrary fluorescence intensity distributions. See suppl. figure 1 for a more detailed version. (C) Illustration of ISLA, our algorithm for conversion of voxel-based 3D images to representative point clouds. Shown are a slice of an input image (left), here a membrane-labeled cell in the pLLP (scale bar: 2µm), the landmarks sampled from this image (middle), here oversampled compared to the standard pipeline for illustration purposes, and the resulting 3D point cloud (right). (D) Illustration of CBE, our algorithm for embedding point clouds into a feature space. In this 2D mock example, two cells are being embedded based on point clouds of their outlines (left). CBE proceeds by performing clustering on both clouds combined (middle) and then extracting the distances along each axis from each cluster center to the centroid of its ten nearest neighbors (right). Note that the most distinguishing morphological feature of the two example cells, namely the

outcropping of cell *a* at the bottom, is reflected in a large difference in the corresponding cluster's distance values (cluster 4, blue).

Figure 4: Analysis of the pLLP's Cellular Shape Space

(A-B) PCA plots of the tissue frame of reference (TFOR) and an cell frame of reference (CFOR) shape spaces of the pLLP. Each point represents a cell and each color represents a different primordium. Selected example cells are shown as point clouds, illustrating that meaningful properties are encoded in PCs, namely cell orientations (A) and cell sphericity and surface smoothness (B). (C) Explained variance ratios of principal components. (D) t-SNE embedding of the shape space showing the absence of obvious clusters as already seen with PCA in (A-B). Colors indicate different primordia as in (A-B). (E-F) Bigraph visualizations of correlations between principal components of the embedded space (bottom nodes) and a set of engineered features (top nodes). Any edge between two nodes indicates a correlation with `Pearson's r > abs(0.3)` and stronger edges indicate stronger correlations. A blue hue implies a positive and a red hue a negative correlation. These correlations together with manual inspection as shown in (A-B) allow the biological meaning of embedded features to be determined. (G) Consensus tissue maps of shape space PCs. The contour map represents the local average of PC values across all registered primordia. The small circles show the centroid positions of cells from a single example tissue to aid orientation. The gray bars indicate, from left to right, the deposition zone, follower zone, transition zone, and leader zone. pLLP shape features show varied patterns, including orientation along the D-V axis (TFOR-PC1), characteristic differences between leader and follower cells (TFOR-PC3, CFOR-PC2), and complicated patterns likely arising from the superimposition of different processes (TFOR-PC1).

Figure 5: Multi-Channel Imaging, Embedding and Data Integration

(A, C, E) Maximum z-projections of two-color stacks showing the membrane in magenta and one of three subcellular structures in yellow. (B, D, F) Tissue frame of reference (TFOR) CBE embeddings corresponding to the three structures shown in A, C and E. The different colors of points indicate different primordia. The three structures are nuclei (N=20, n=2528) (A-B), F-actin (N=19, n=1876) (C-D) and the Golgi apparatus (N=11, n=866) (E-F). (G) Bigraph showing correlations between the Golgi's embedded features and our engineered cells shape features (see supplementary table 2). The first two Golgi TFOR PCs match those found in the cell shape TFOR space (see figure 4E) whereas PCs 3 and 4 are specific to the Golgi. For technical details see the legend of figure 4E. (H-I) Point cloud renderings showing the distribution of Golgi signal (blue, membranes in red) in two example cells, one with a high value in the Golgi's TFOR PC 3 (H) and one with a low value (I), illustrating that PC 3 captures apical enrichment of the Golgi. (J) Consensus tissue map for Golgi PC 3 (apical enrichment), showing increased values behind the leader zone. For technical details see the legend of figure 4G.

Figure 6: *pea3* smFISH as an Example of Data Integration Across Imaging Modalities

(A) Maximum z-projection of a two-color stack of *pea3* smFISH (yellow) and the *lyn-EGFP* membrane marker (magenta). Scale bar: 10µm. (B) Results of SVR regression on *pea3* spot counts using TFOR and CFOR shape features as well as cell centroid coordinates of registered primordia as input. Each blue dot is a cell, the diagonal gray line reflects perfect prediction and blue arrows at the border point to outliers with very high spot counts. On training data, the regressor's explained variance ratio is 0.462±0.011, on previously unseen test data it achieves 0.382±0.019. (C-D) Consensus tissue maps of *pea3* expression generated directly from the *pea3* smFISH dataset (C) or from the full atlas dataset based on SVR predictions of spot counts (D). Note that the prediction for the entire atlas preserves the most prominent pattern – the front-rear gradient across the tissue – but does not capture the noisy heterogeneity among follower cells observed in direct measurements.

Figure 7: Context-Guided Visualization using Morphological Archetypes

(A) A maximum z-projected example stack with colors highlighting different conceptual archetypes in the pLLP that have been manually annotated. (B) A low-dimensional *archetype space* resulting from a PCA of the SVC prediction probabilities (with the SVC having been trained on CFOR shape features). Cells are placed according to how similar they are to each archetype, with those at the corners of the tetrahedron belonging strictly to the corresponding archetype and those in between exhibiting an intermediate morphology. (C) Since inter-organ cells are not morphologically distinct enough at this stage (see suppl. figure 5), the archetype space can be reduced to 2D without much loss of information. (D) Scatter plots of the 2D archetype space with additional information from the cellular shape space and from the protein distribution atlas superimposed in color. (E-F) Boxplots showing data grouped by predicted archetype labels. This form of grouping allows statistical analysis, showing that leader cells are flatter than any other class of follower cells (E) and that central rosette cells are more spherical than peripheral rosette cells (F). Whiskers are 5th/95th percentiles, p-values are computed with a two-sided `Mann-Whitney U-test`, and `Cohen's d` is given as an estimate of effect size.

MAIN FIGURES

Figure 1

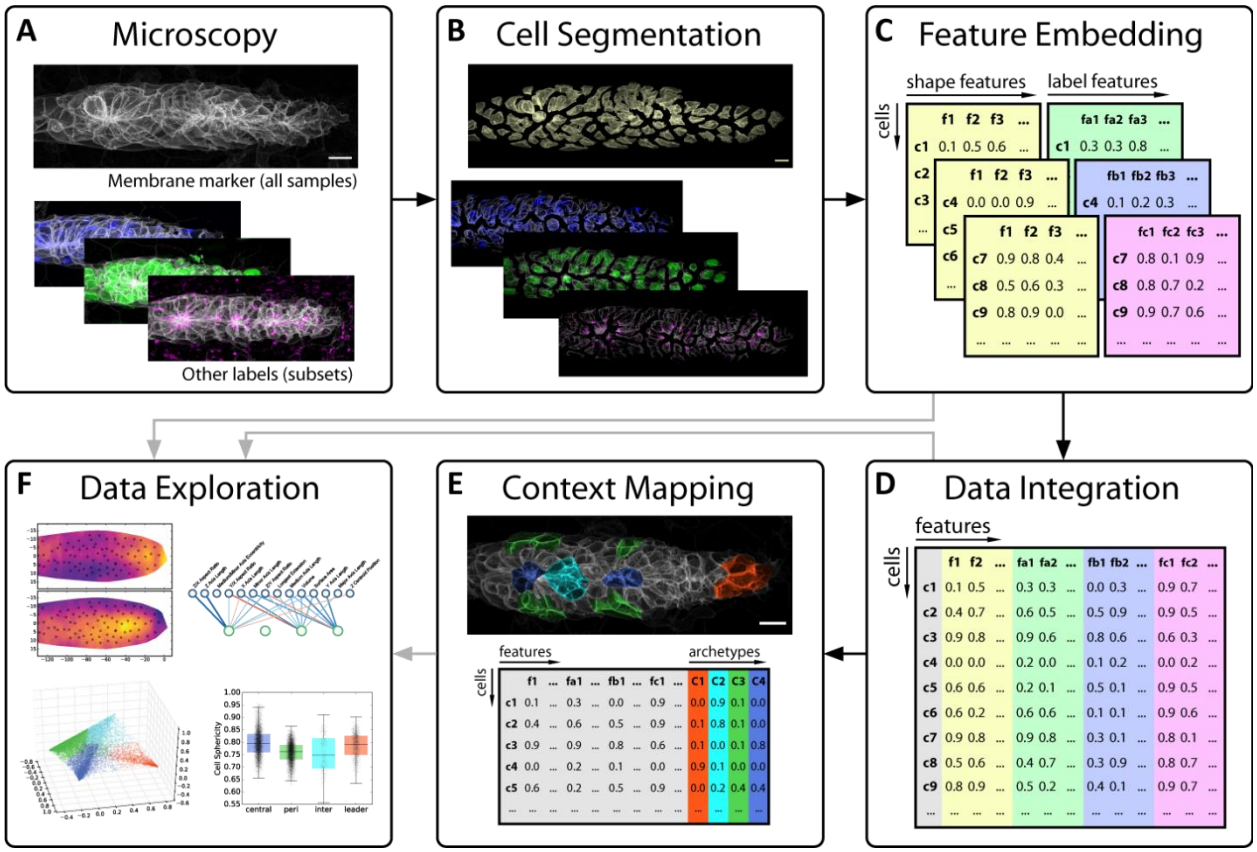


Figure 2

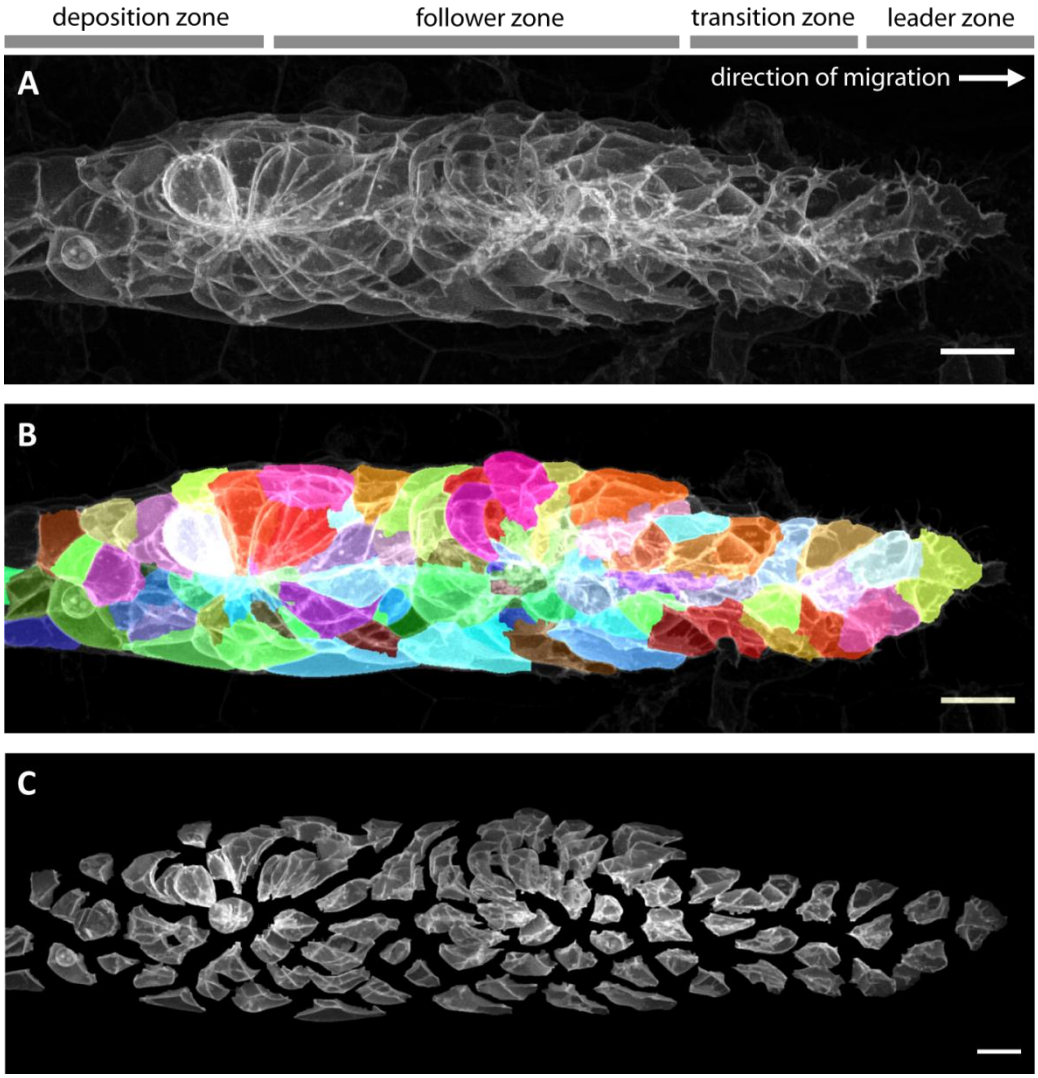


Figure 3

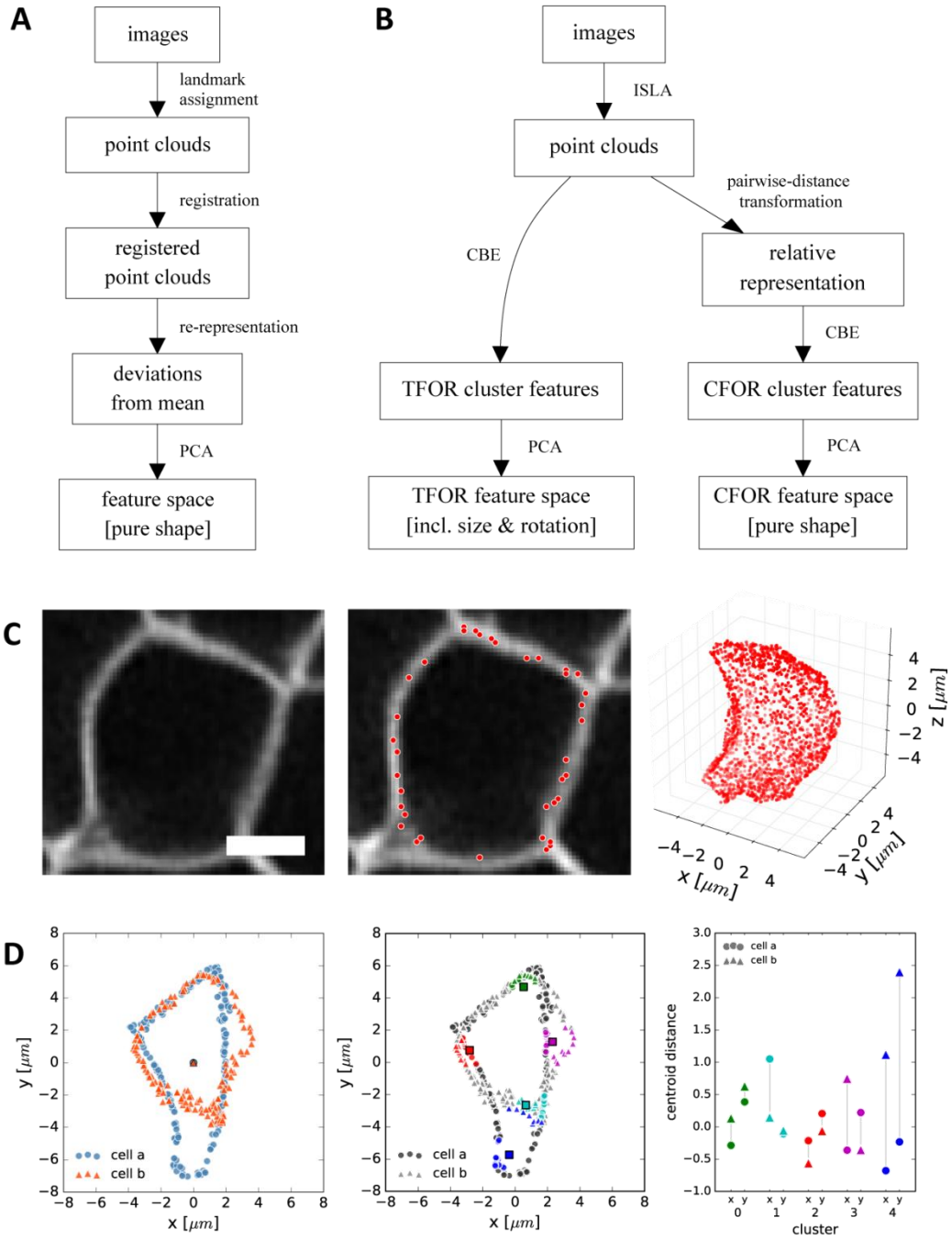


Figure 4

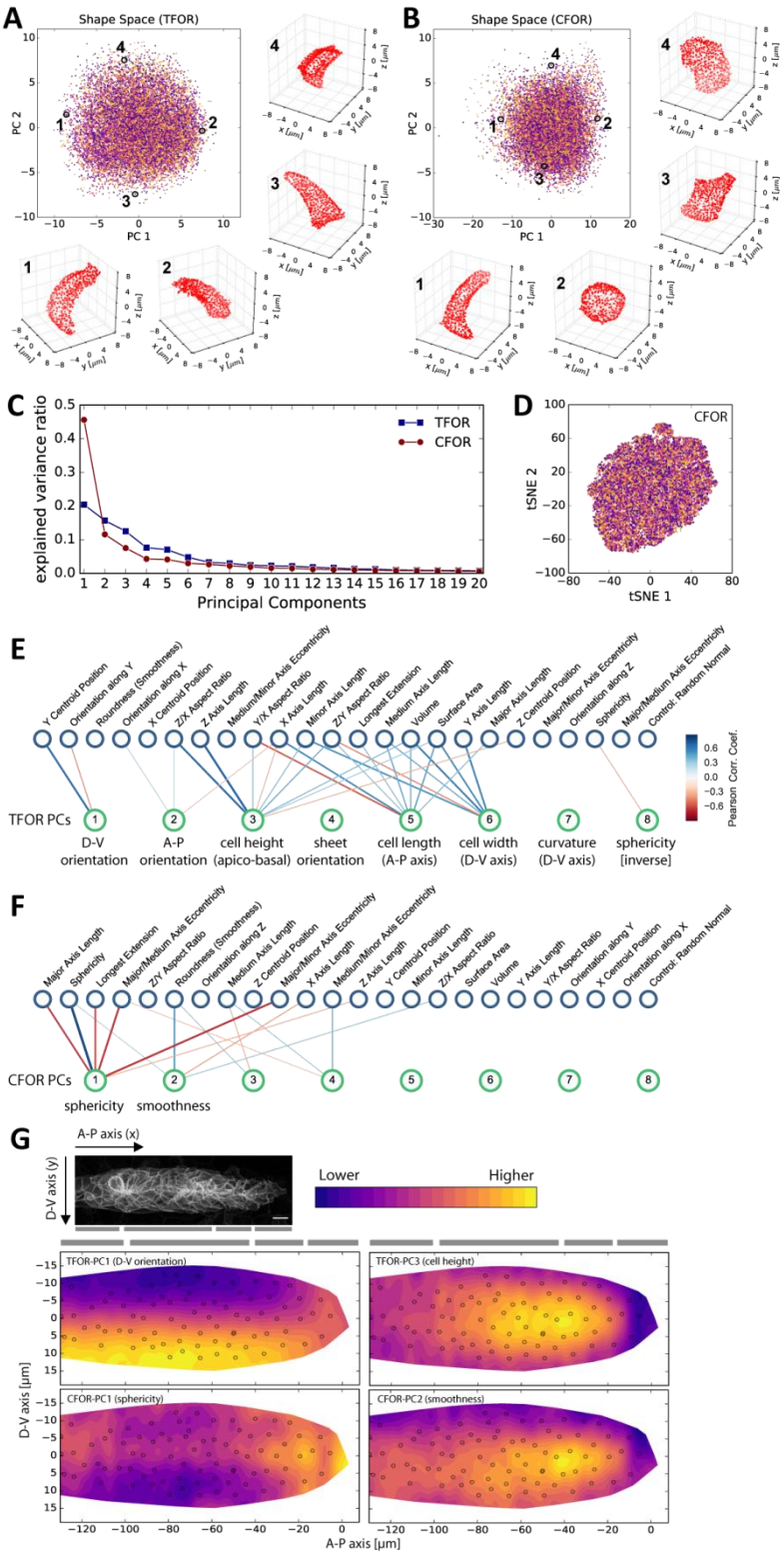


Figure 5

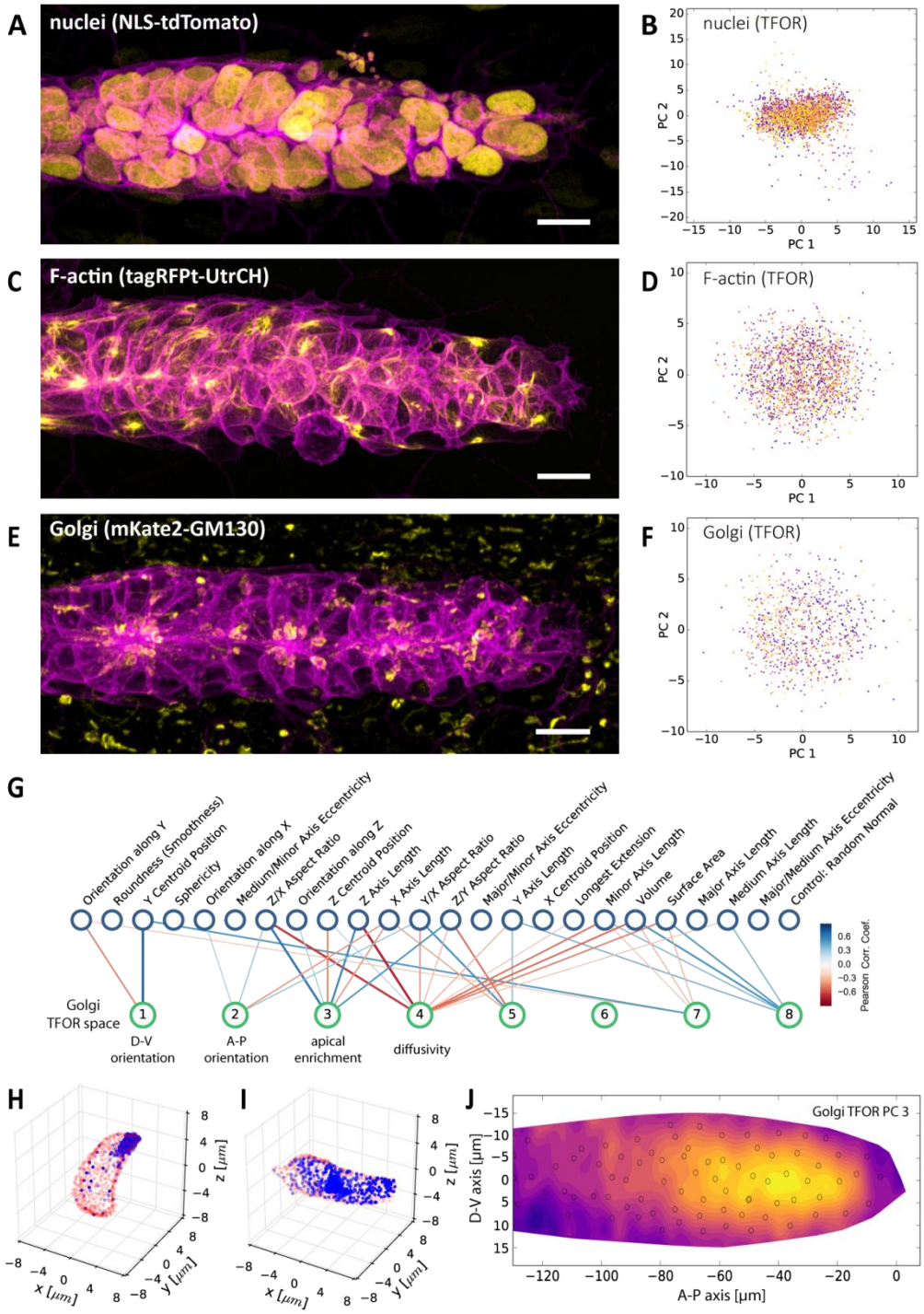


Figure 6

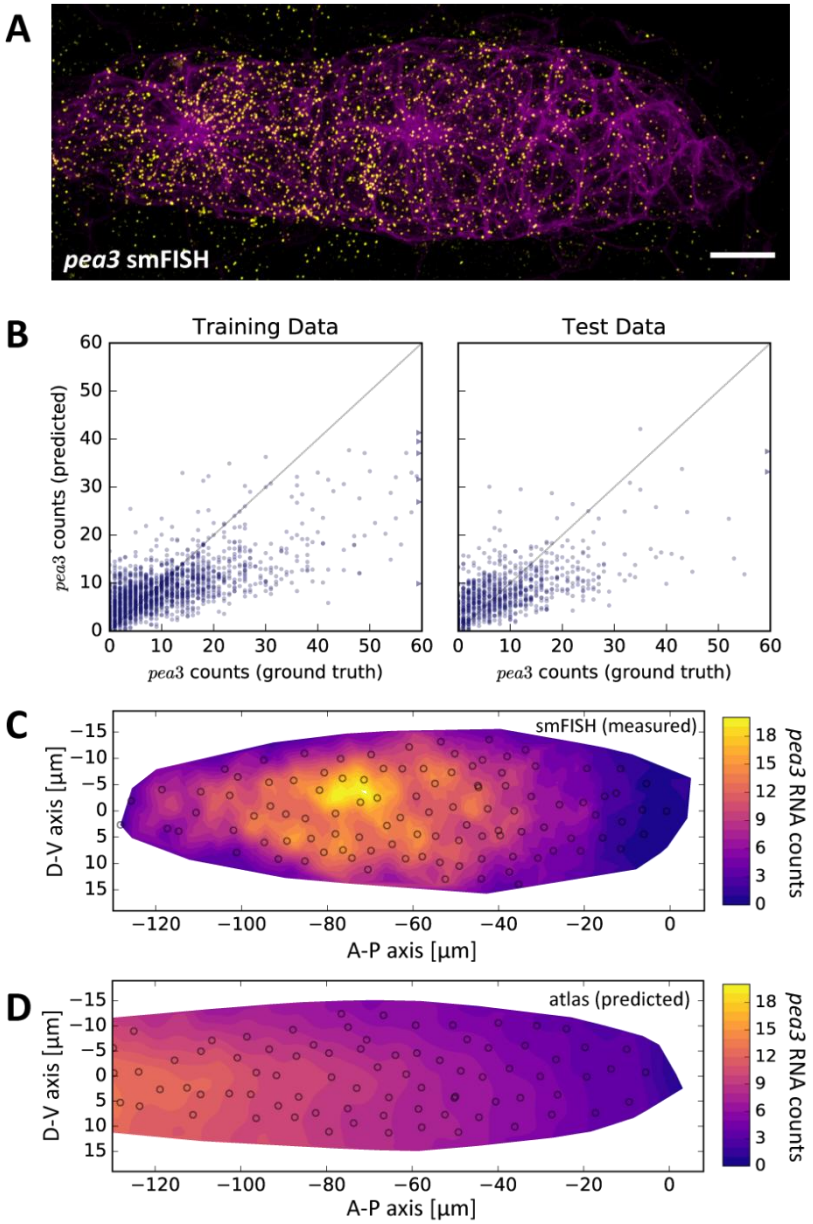
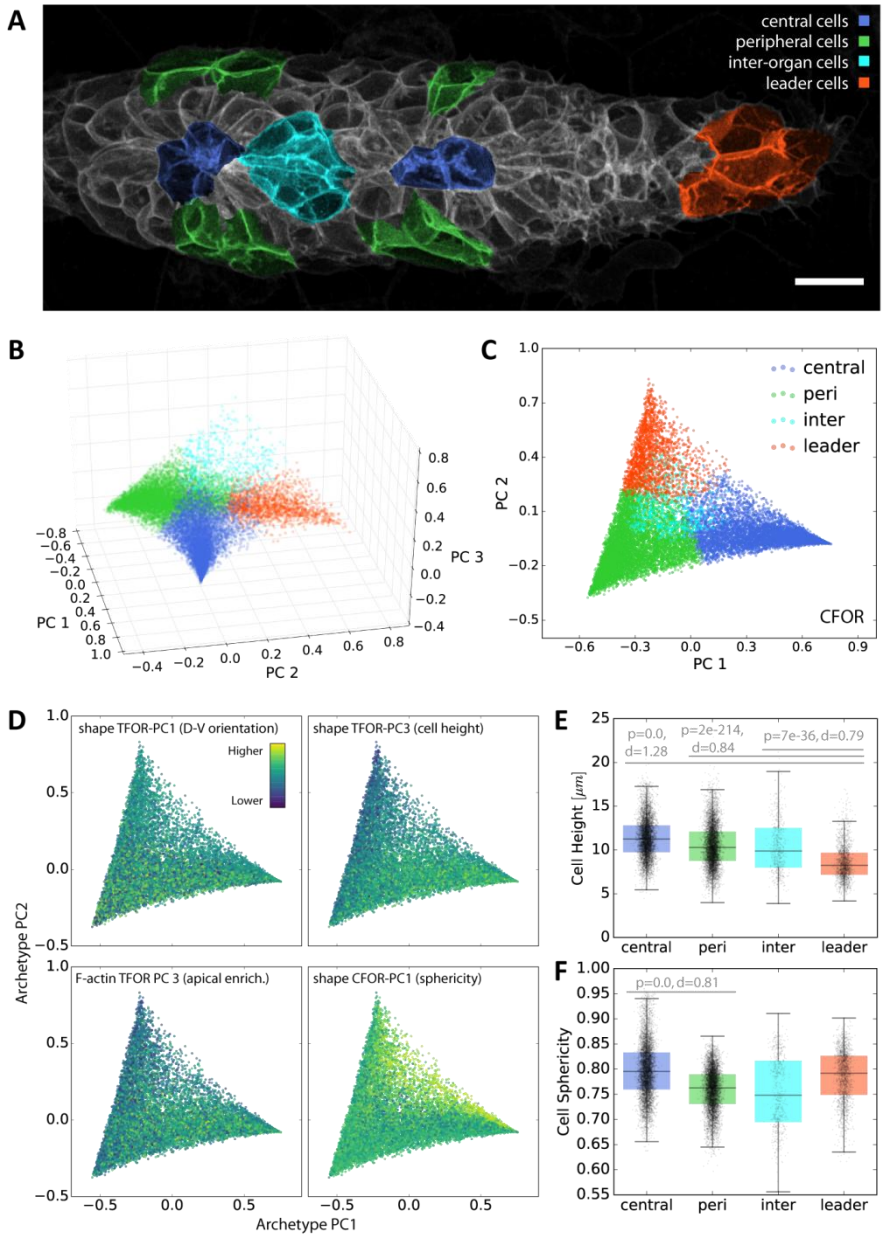


Figure 7



SUPPLEMENTARY FIGURE TITLES AND LEGENDS

Supplementary Figure 1: Flowcharts Illustrating the ISLA and CBE Algorithms.

(A) Flowchart of ISLA. To sample from intensity distributions, images are masked by setting voxels outside of the segmentation to zero and a simple background subtraction is performed. To sample from cell shapes, the 1vxl-wide outer shell of the segmentation is set to 1, all other voxels to zero. The resulting image is normalized and used to stochastically sample points for the point cloud. (B) Flowchart of CBE. Input point clouds of cells are either rotated according to a registration across tissues (Tissue Frame Of Reference, TFOR) or are volume-normalized and re-represented as a subset of the pairwise distances between points, removing size and rotational information (Cell Frame Of Reference, CFOR). A representative subset of the resulting clouds is overlaid and k-means clustering is performed on the overlay, yielding a set of common reference points. Finally, features are computed to describe each cell's point cloud relative to these common reference points, resulting in an embedded feature space. This feature space can be transformed with PCA to emphasize relevant variation across the sample population.

Supplementary Figure 2: Evaluation of the Expressiveness of CBE Embeddings With and Without Cell Frame of Reference (CFOR) Normalization

(A) Performance in predicting different generative parameters of point clouds in a synthetic dataset from either a raw or a size- and rotation-corrected (cell frame of reference, CFOR) embedding. As expected, CFOR normalization removes all information on cloud size ('scaling') and orientation ('rotation' 1-2). Interestingly, removing this information allows the regressor to perform far better when it comes to the shape parameters of the point cloud ('shape' 1-14). The 'random' parameter is a random Gaussian distribution and serves as a negative control. The regressor used is a Support Vector Regressor (SVR) with an RBF-kernel. (B) Evaluation of CBE compared to an alternative embedding strategy based on moments. Shown is how well the parameters used to synthetically generate point clouds can be predicted from embeddings of said clouds using different regression models (kNN: k-Nearest Neighbor regressor, l-SVR: linear Support Vector Regressor, r-SVR: RBF-kernel Support Vector Regressor). Black dots indicate results of 3-fold cross-validation, bars indicate the mean. Moments-based embedding is outperformed by CBE in all cases except with linear SVR, where the results are similar.

Supplementary Figure 3: Evaluation of Machine Learning Algorithms for Feature Space Atlas Mapping

(A) Performance of different algorithms at predicting the embedded feature spaces of different secondary marker channels from embeddings of cell shape. The left column contains predictions from the cell shape CFOR space to subcellular structure CFOR spaces, the right column from cell

shape TFOR space to subcellular structure TFOR spaces. Performance is quantified as variance-weighted average of r-squared values across target dimensions. Gray dots are the results from 3-fold cross-validation, blue bars are averages. The algorithms evaluated are a random assignment control (random), k Nearest Neighbors (kNN), the scikit-learn implementation of gradient boosting (boost), xgboost (xgb), random forest regression (forest), support vector regression (SVR), multi-layer perceptrons (MLP), multi-task Lasso regression (Lasso), and multi-task elastic nets (eNet). Note that TFOR predictions work far better than CFOR predictions, which may indicate that pure shape information is insufficient to predict key features of intracellular protein distributions, possibly because information on tissue context is lost. (B) Both for training and prediction, PCA-transformed feature spaces were used. Here, prediction quality is shown for each PC of an example channel, illustrating that high-variance PCs lend themselves to more accurate prediction than low-variance components, as expected given that the latter encode less meaningful variation and more noise. (C) Examples showing the correlation of resulting predictions with ground truths, again illustrating that high-variance PCs (left) can be fitted better than low-variance PCs (right).

Supplementary Figure 4: Spot Detection and Cell Shape Embedding for *pea3* smFISH Data

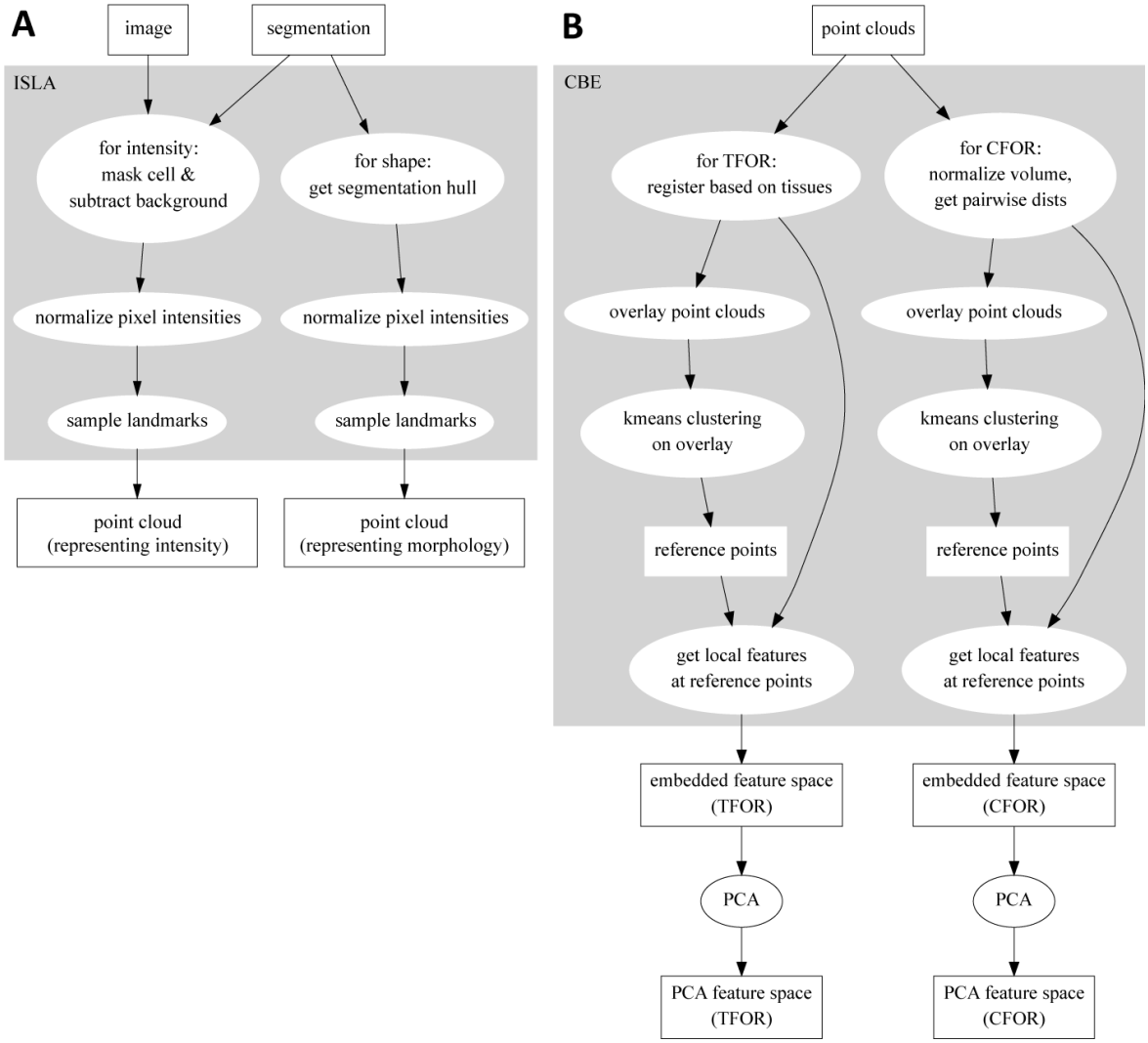
(A-C) Maximum z-projections of a two-color sample showing the *lyn-EGFP* membrane marker (A), the *pea3* smFISH probe (B), and the results of automated spot detection with red rings denoting detection events (C). Scale bars: 10µm. (D) Zoomed view of the region in the yellow box in (C). Scale bar: 2µm. (E) *pea3* smFISH spot counts for each cell, both from measured data (blue) and from predictions across the entire atlas dataset (purple). The left shows averages across primordia, which closely match those reported previously based on a different spot counting method [Durdu et al., 2014]. The individual cell counts on the right show that there is a long tail of cells with extremely high counts, which as one would expect is not captured in the SVR predictions. (F-H) Comparisons of three important cell shape variables between fixed smFISH samples and live samples (a subset of the main dataset), showing no significant difference.

Supplementary Figure 5: Evaluation of Morphological Archetype Prediction

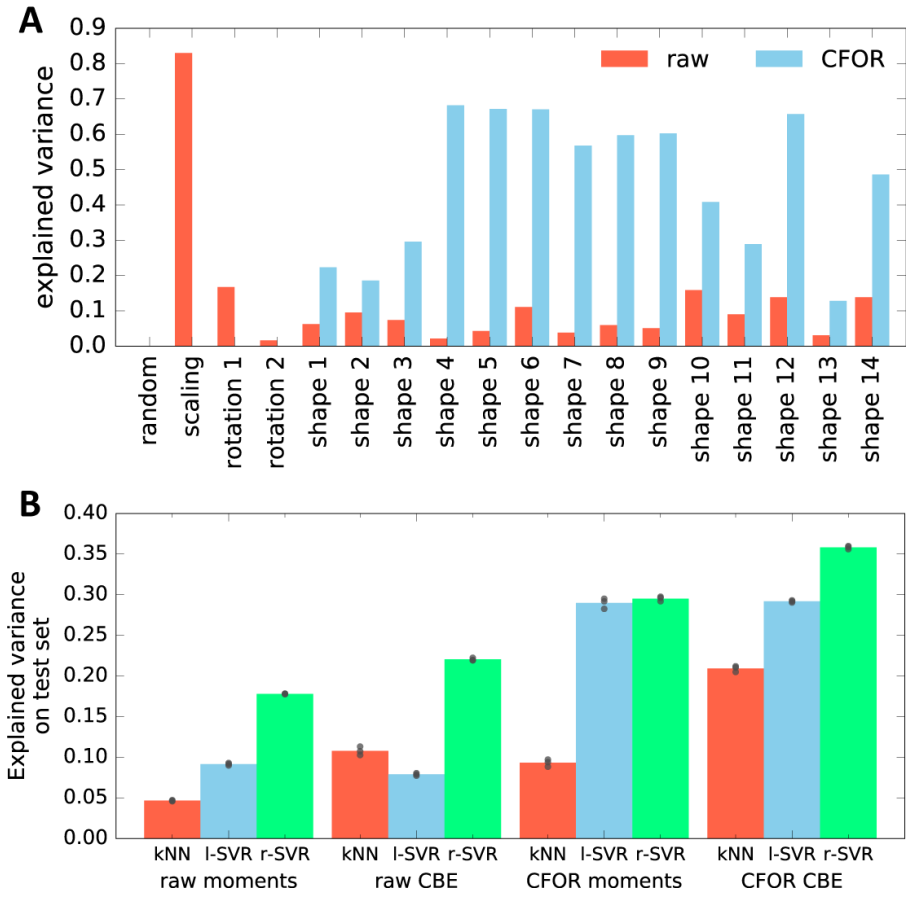
Confusion matrices for SVC archetype classification. The ground truth is based on manual annotation of high-confidence cases. Note that using TFOR features results in slightly better performance than using CFOR features, implying that rotational information and cell size are useful for prediction to some extent. Overall prediction accuracy is high but inter-organ cells are frequently mislabeled, in particular as peripheral cells. This indicates that most inter-organ cells participate in rosette formation in a similar fashion to peripheral cells at this stage and are therefore hard to distinguish based on morphology alone.

SUPPLEMENTARY FIGURES

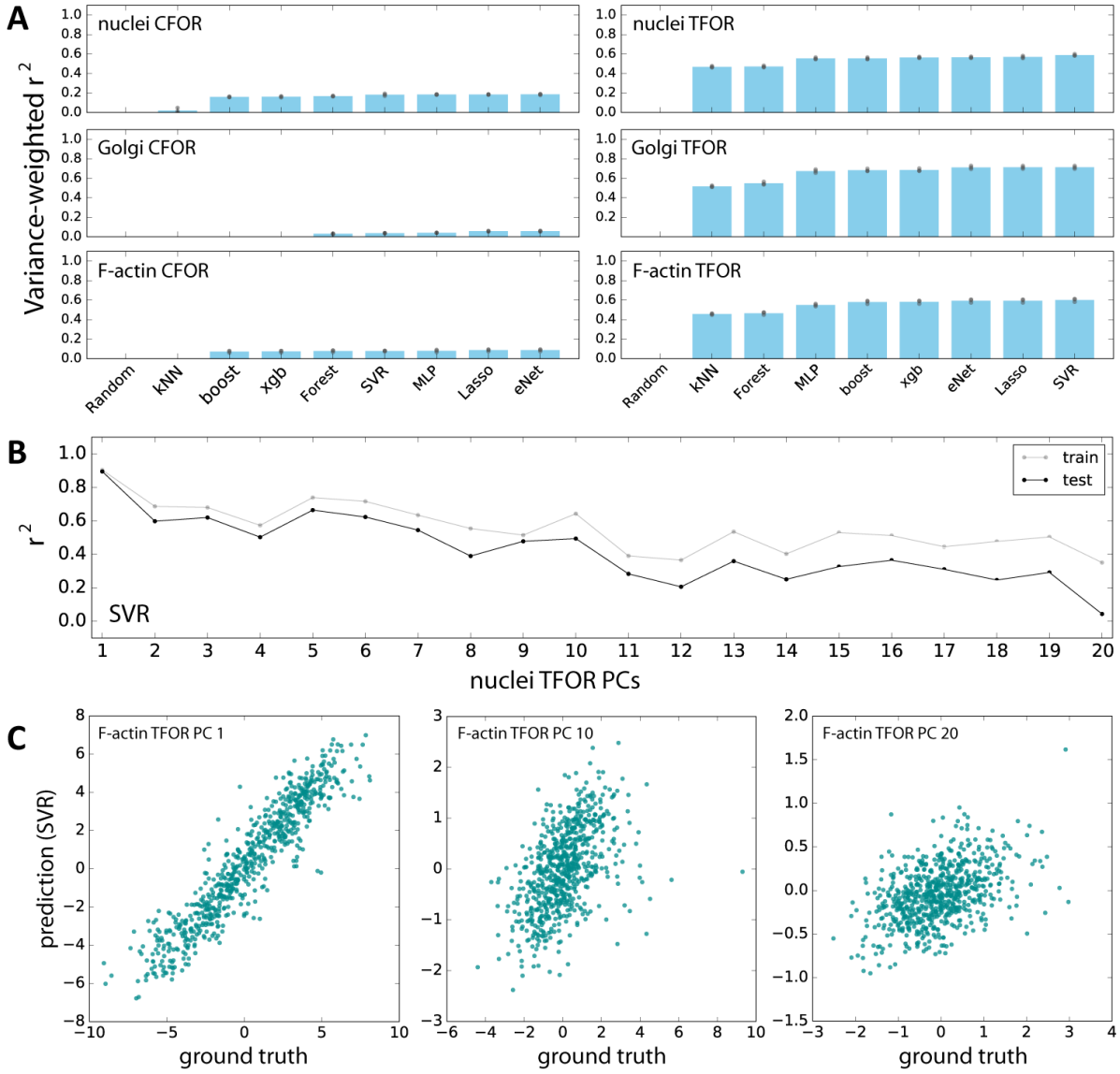
Supplementary Figure 1



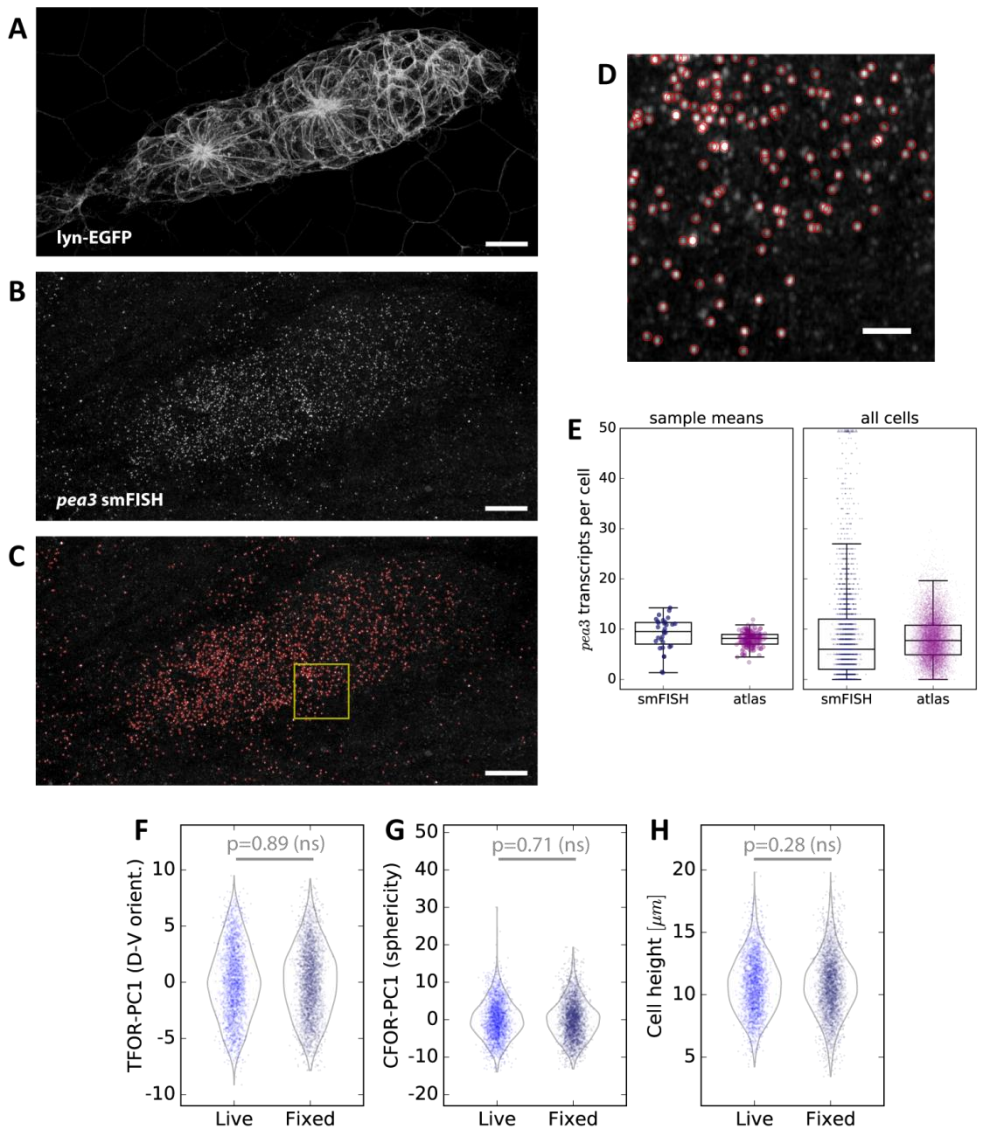
Supplementary Figure 2



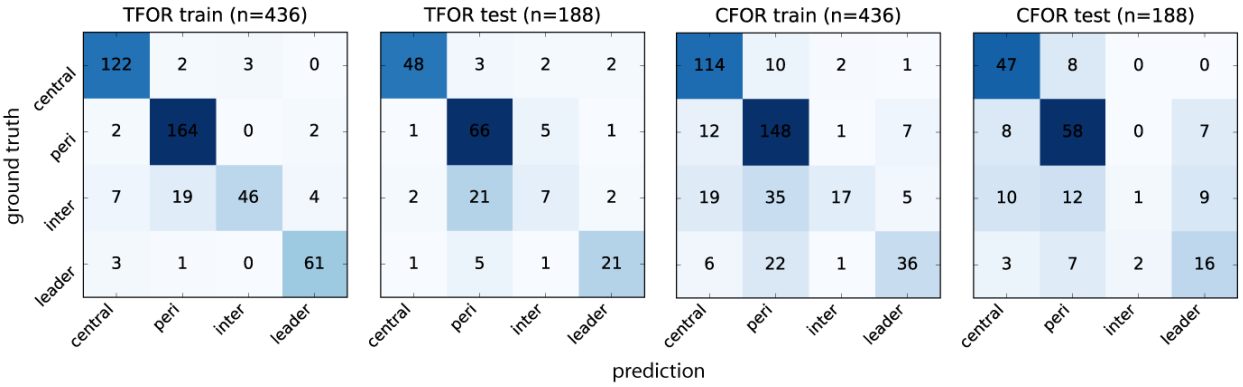
Supplementary Figure 3



Supplementary Figure 4



Supplementary Figure 5



SUPPLEMENTARY MOVIE TITLES AND LEGENDS

Supplementary Movie 1: Dynamic View 3D pLLP Stack

Movie of the stack shown in figure 2A. The slices of the stack are scanned through from bottom to top first, then a maximum z-projection is generated from top to bottom. Scale bar: 10µm.

Supplementary Movie 2: Dynamic View of 3D pLLP Segmentation

Movie of the single-cell segmentation shown in figure 2B. The slices of the stack are scanned through from bottom to top first, then a maximum z-projection is generated from top to bottom. Scale bar: 10µm.

Supplementary Movie 3: Dynamic View of 3D Expanded Segmentation

3D Movie of expanded segmentation shown in figure 2C. First, a rotation of the raw data is shown, followed by its expansion based on the single-cell segmentation. Scale bar: 10µm.

SUPPLEMENTARY TABLES

Supplementary table 1: Dataset Composition (After Quality Control)

Genotype	Structure	N Primordia	n Cells
cldnb:lyn-EGFP	membranes	24	2310
cldnb:lyn-EGFP *cxcr4b:NLS-tdTomato*	membranes nuclei	20	2528
cldnb:lyn-EGFP *Actb2:mKate2-Rab11a*	membranes recycling endosomes	19	1554
cldnb:lyn-EGFP RNA: *mKate2-Rab5a*	membranes early endosomes	14	1131
cldnb:lyn-EGFP RNA: *mKate2-GM130(rat)*	membranes *cis*-Golgi	11	866
cldnb:lyn-EGFP *LexOP:CDMPR-tagRFPt* *cxcr4b:LexPR (driver)*	membranes TGN & late endosomes	13	967
cldnb:lyn-EGFP *LexOP:B4GalT1(1-55Q)-tagRFPt* *cxcr4b:LexPR (driver)*	membranes *trans*-Golgi	10	789
cldnb:lyn-EGFP *atoh1a:dtomato*	membranes *atoh1a* expression	14	1524
cldnb:lyn-EGFP *6xUAS:tagRFPt-UtrCH* *ETL GA346 (driver)*	membranes F-actin	19	1876
cldnb:lyn-EGFP LysoTracker™ Deep Red staining	membranes lysosomes	21	1802
Total		**165**	**15347**
Additional smFISH dataset (fixed):			
cldnb:lyn-EGFP *pea3* smFISH staining	membranes *pea3* RNA molecules	31	3149

Supplementary table 2: Engineered Features

Name	Description
Y,X,Z Centroid Position	Position of cell centroids in TFOR.
Orientation along X,Y,Z	Normalized components (in TFOR) of the vector pointing from the cell centroid to the most distant point in the cellular point cloud.
A/B Axis Eccentricity	Eccentricity ratios of an ellipsoid fitted to the cellular point cloud. Major, medium and minor (A, B) refer to the length of the principal semi-axes, from largest to smallest.
A Axis Length	Length of the ellipsoid's principal semi-axes (A) in microns.
A/B Aspect Ratio	Ratios of cell extents along different axes (A, B) in TFOR.
X,Y,Z Axis Length	Cell extents along different axes in TFOR, in microns.
Longest Extension	Distances from the cell centroid to the most distant point in the cellular point cloud, in microns.
Volume	Cell volume (in cubic microns).
Surface Area	Cell surface area (in square microns)
Sphericity	Measure of how closely the distribution of the cellular point cloud approximates a sphere. It is the mean distance of points to a sphere centered on cell centroid with a radius equal to the mean distance of points from the centroid. The measure is linearly normalized to between 0 and 1, where 1 means perfectly spherical and <1 means less spherical.
Roundness (Smoothness)	Measure of how closely the distribution of the cellular point cloud approximates a circumscribed ellipsoid and hence how smooth the surface is. It is the mean distance of points to an ellipsoid fitted to the cloud. The measure is linearly normalized to between 0 and 1, where 1 means perfectly round/smooth and <1 means less round/smooth.
Control: Random Normal	A randomly sampled normal distribution with $\mu=0$ and $\sigma=1$. Used as a simple "negative control" for any code related to correlation measurements.