Zero-shot reconstruction of mutant spatial transcriptomes

Yasushi Okochi[1,2,*], Takaaki Matsui[3], Honda Naoki[1,4,5,6,*]

[1] Laboratory for Data-driven Biology, Graduate School of Integrated Sciences for Life, Hiroshima University, Higashihiroshima, Hiroshima, Japan

[2] Faculty of Medicine, Kyoto University, Sakyo, Kyoto, Kyoto, Japan

[3] Division of Biological Sciences, Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Nara, Japan

[4] Laboratory for Theoretical Biology, Graduate School of Biostudies, Kyoto University, Kyoto, Kyoto, Japan

[5] Theoretical Biology Research Group, Exploratory Research Center on Life and Living Systems (ExCELLS), National Institutes of Natural Sciences, Okazaki, Aichi, Japan

[6] Kansei-Brain Informatics Group, Center for Brain, Mind and Kansei Sciences Research (BMK Center), Hiroshima University, Kasumi, Minami-ku, Hiroshima 734-8551, Japan

*** Corresponding author**

Graduate School of Integrated Sciences for Life, Hiroshima University, Kagamiyama 1-3-1, Higashihiroshima, Hiroshima, 739-8526, Japan

Tel.: +81-75-753-4422

E-mail: nhonda@hiroshima-u.ac.jp

*** Co-corresponding author**

Faculty of Medicine, Kyoto University, Yoshidakonoe-cho, Sakyo-ku, Kyoto, Kyoto, 606-8315, Japan

Tel.: +81-75-647-9929

E-mail: ookouchi.yasushi.68m@st.kyoto-u.ac.jp

Abstract

Analysis of mutants is a core of biological/pathological research, and measuring their spatial gene expression can facilitate phenotypic understanding of their tissue disorganization. However, considering the existence of enormous numbers of mutants which are worth to be investigated, the cost- and technically-demanding nature of the experiments measuring spatial transcriptomes could be a bottleneck. Although spatial transcriptomes have been computationally predicted from single cell RNA sequencing (scRNA-seq) data by referring to teaching data of spatial gene expressions of some genes, this task is still challenging because the teaching data is not available in most of the mutants. Zero-shot learning shows potential for solving general prediction problems without teaching data in various machine learning tasks. Here, we provide the first zero-shot framework (ZENomix) to predict mutant spatial transcritptomes from mutant scRNA-seq data without teaching data (e.g. mutant spatial reference atlas). We validated ZENomix by demonstrating accurate prediction of spatial transcriptomes of Alzheimer's disease-model mice and mutant zebrafish embryos losing Nodal signaling. Furthermore, we proposed a spatially-informed screening approach based on ZENomix's prediction, which discovered new Nodal-downregulated genes in zebrafish. We expect ZENomix to reveal new phenotypic insights by utilizing massive mutant/disease scRNA-seq data that are and will be acquired.

Introduction

Identification of spatial gene expression profiles in mutants and diseases is crucial to understand how the tissue of interest is functioned and dysfunctioned. The recently developed spatially-resolved transcriptomics technologies (e.g. *in situ* RNA capturing for Next generation sequencing-based method and *in situ* RNA sequencing for *in situ* hybridization-based method)[1] enabled us to measure gene expression profiles with spatial context in high-throughput manner, and have revealed insights into the mechanism of tissue disorganization in diseases[2–5]. However, there exists enormous numbers of mutants, which are biologically/pathologically worth to be investigated, and comprehensive measurement of spatial transcriptomes of these mutants must be limited due to the cost- and technically-demanding nature of the technologies[6]. In addition, these technologies often suffer from the trade-off between gene detection sensitivity and the number of genes measured[7]. By contrast, there are methods to computationally reconstruct spatial transcriptomes from scRNA-seq data which have high gene detection sensitivity on the whole transcriptomes[8]. In this concept of reconstruction, dissociated scRNA-seq data is assembled by referring to teaching data of spatial expression patterns of some genes, e.g., *in situ* hybridization (ISH) atlas. However, in mutant tissues, most of which have no atlas of spatial gene expressions, teaching data is generally unavailable, making it impossible to apply this concept.

Prediction in case of no teaching data is a challenging problem in various tasks in image recognition and natural language processing, which require predicting previously unseen instances. In these fields, in order to compensate for the lacking teaching data, existing data not containing teaching signals of interest are trained

as side information to refine the prediction. This idea, called "zero-shot learning", has large potential for solving general prediction problems without teaching data, just as humans are able to predict a new instance without ever experiencing it[9].

In this study, we developed the first, to our knowledge, computational framework (ZENomix) for zero-shot reconstruction of mutant spatial transcriptomes, in which zero-shot learning is used to deal with the loss of mutant spatial reference atlas (i.e. zero teaching information). In our zero-shot framework, we used wildtype spatial reference atlas, which is much easier to access, as side information. We reasoned that, although wildtype has different gene expressions from mutant tissues, underlying spatial coordinates on tissues are the same if wildtype and mutant tissue morphology are similar. In ZENomix, wildtype reference atlas is used as landmark points for spatial coordinate in tissues, helping add spatial information to mutant scRNA-seq data. ZENomix learns the spatial information in tissue by embedding wildtype spatial reference atlas in the latent space, and subsequently maps mutant scRNA-seq data into the learned latent space. This spatial information in the latent space, in turn, is used to reconstruct mutant spatial transcriptomes.

In the following, we first demonstrated that ZENomix was correctly performed in a zero-shot manner using the simulated scRNA-seq data in Alzheimer's disease (AD)-model mice. Next, we applied ZENomix to mutant zebrafish early embryo scRNA-seq data. We confirmed ZENomix's predicted spatial transcriptomes by comparing with known *in situ* hybridization data in *MZoep* mutants. Through the prediction of spatial gene expression by ZENomix, we listed the previously unknown genes that could exhibit the spatially restricted gene expression changes, which are validated by *in situ* hybridization experiments. Thus, we concluded that ZENomix opened up a new concept by offering a zero-shot reconstruction of mutant spatial transcriptomes.

Results

Zero-shot reconstruction framework of ZENomix

ZENomix is a computational method for predicting mutant spatial transcriptomes through zero-shot learning using wildtype spatial reference data as a side information (**Fig. 1a**).

In general, it is a challenging task to predict mutant spatial transcriptomes without teaching data (i.e. mutant spatial reference atlas). To make the ZENomix's framework clear, we first focus on a simple, one-dimensional tissue with two spatial gene expression profiles (e.g. D-V axis patterning of the vertebrate neural tube by Shh and BMP/Wnt[10]) (**Fig. 1b**). In this tissue, the spatial information of each cell can be expressed along the trajectory in the gene expression space (**Fig. 1c**). By considering that wildtype spatial reference data has both gene expressions and spatial information data, wildtype spatial reference data points can be used as landmark points of the spatial information in gene expression spaces (**Fig. 1c**). It means that if mutant scRNA-seq data points can be put along this wildtype trajectory, the spatial information of mutant scRNA-seq data points can be recovered from the landmark points which have the spatial information. However, in a mutant tissue, the trajectory in the gene expression space should be distorted due to the difference of the gene expression profiles (**Fig. 1c**). Our idea is that we can calibrate the difference of these trajectories and recover the spatial information of mutant scRNA-seq data by matching the distributions of cells in the gene expression space (**Fig. 1d**).

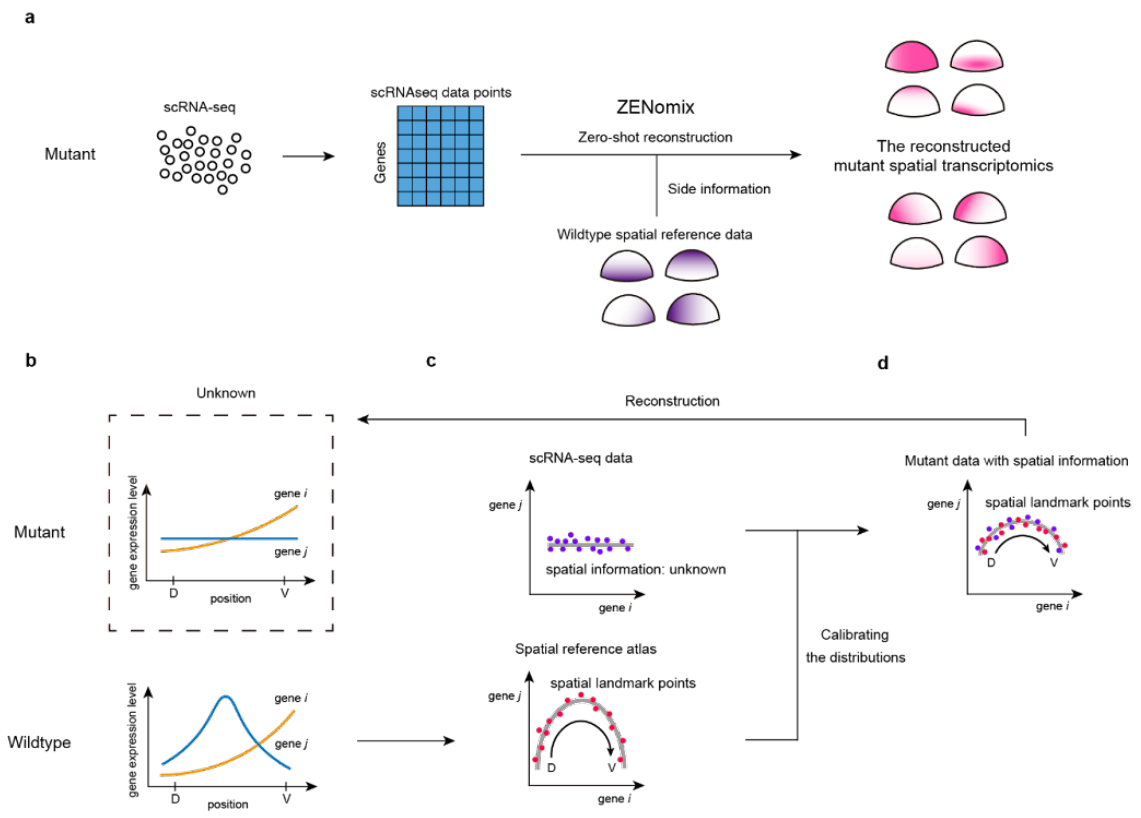


Figure 1: Concept of zero-shot reconstruction

(a) Reconstruction of mutant spatial transcriptome by ZENomix. (b) The simple, one-dimensional model of wildtype and mutant tissues. The model tissues have different gene expression along the dorso-ventral (D-V) axis. The blue and orange line indicate the expression profiles of gene i and j, respectively. (c) Different trajectories in gene expression space. In mutants, the gene expression levels are measured by scRNA-seq that loses spatial information. In wildtype, the gene expression levels are measured by *in situ* technology that has spatial information (black arrow), which are used as landmark points (postal codes) when adding spatial information to mutant scRNA-seq data. The red and blue points indicate the data points of *in situ* technology and scRNA-seq, respectively. (d) Mutant gene expression data with spatial information. The trajectories can be calibrated by matching the distributions of two data points. After calibrating the difference of the trajectories, the lost spatial information of the mutant scRNA-seq data points can be recovered by referencing the landmark points which have the spatial information.

The zero-shot framework in ZENomix involves two steps: the training and reconstruction. The training step is to extract the landmark points of the spatial information in tissue from wildtype spatial reference data on the above concept. In practice, the spatial reference data has tens to thousands of gene expression profiles on two- or three-dimensional tissue, e.g. 47 genes for zebrafish ISH data[11] and 12,337 genes for mouse olfactory bulb Spatial Transcriptomics (ST) data[12]. To deal with the high dimensional nature of these data, ZENomix embeds wildtype spatial reference data onto the latent space to get the landmark points of spatial information using a non-linear dimensionality reduction method, the Gaussian process-latent variable model (GPLVM[13]) (**Fig. 2a**). The difference in distributions between wildtypes and mutants is calibrated by minimizing the distance (maximum mean discrepancy (MMD)[14]) between two data distributions (**Fig. 2b**). The reconstruction step is to estimate mutant spatial gene expression profiles. ZENomix maps back the landmark points derived from the wildtype data to the mutant scRNA-seq space, and mutant spatial transcriptomes are reconstructed as the weighted-average of the mutant scRNA-seq data points by using the Gaussian process regression (arrows in **Fig. 2c**).

To estimate the parameters of ZENomix, we proposed a new inference scheme (vGPLVM-MMD) combining the variational Gaussian process latent variable model (vGPLVM[15,16]) and maximum mean discrepancy (MMD) (See *Methods*). vGPLVM-MMD is a Bayesian learning algorithm using the inducing points method and variational inference.

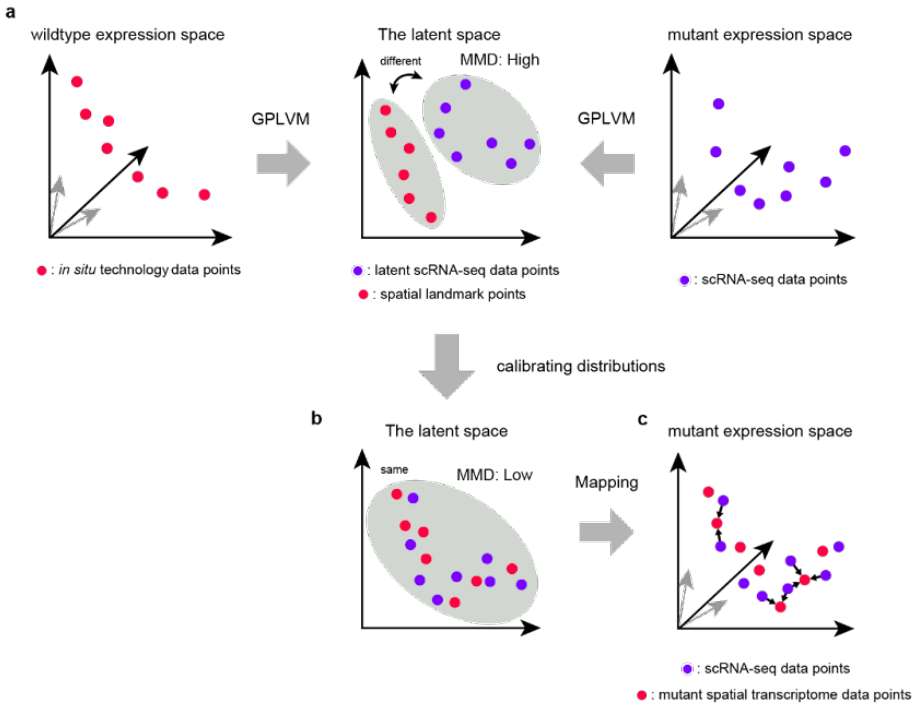


Figure 2: Data processing of ZENomix
(a-b) The first step of ZENomix. (a) The embedding to the latent space in ZENomix. In the wildtype expression space, there are high dimensional *in situ* technology data points (the red points; the left panel), and in the mutant expression space, there are high dimensional scRNA-seq data points (the blue points; the right panel). ZENomix uses GPLVM to embed the wildtype data points onto the latent space for getting the spatial landmark points. The mutant data points are also embedded onto the latent space, although the data distribution is different from that of the wildtypes. (b) Calibration of the difference in the distributions. After mapping onto the latent space, ZENomix calibrates the difference in the distributions of the two data by minimizing discrepancy between the distributions (MMD; see *Methods*). The gray-shaded region indicates the calibrated distribution. (c) The second step of ZENomix. After the first step, ZENomix maps the spatial landmark points derived from wildtype data onto the mutant expression space. The mutant spatial transcriptomes (the red points) are reconstructed by weighted-average of the mutant scRNA-seq data points (the blue points). The black arrows indicate the weights of scRNA-seq data points.

ZENomix correctly performed zero-shot learning

To check whether our **zero-shot** reconstruction framework works well, we first tested the performance of ZENomix using the simulated scRNA-seq data of AD-mutant (3xAD) mouse olfactory bulb (OB) (**Fig. 3a**). The data we addressed here was ST data in mouse OB, in which gene expression data was obtained at points spatially arranged in a lattice (**Fig. 3b**). The simulated scRNAseq data (1,409 data points) were generated by masking the true spatial coordinates of AD-mutant mouse OB ST data from Navarro *et al*., 2020.[3] whereas ST data of the wildtypes from Ståhl *et al*., 2016.[12] can be used for spatial reference data (**Fig. 3c**). It should be noted that we knew the ground-truth spatial coordinates of the simulated mutant scRNA-seq data as a benchmark.

Here, we first confirmed that ZENomix was able to calibrate the distributions of two data in latent space (**Fig. 3d**) and that the model parameters were converged (**Extended Data Fig. 1**). We then demonstrated

that ZENomix successfully predicted the spatial transcritptomes of the AD-mutant mouse OB (**Fig. 3e**). We repeated the same analysis using the simulated wildtype scRNA-seq data made from the wildtype ST data in Navarro *et al*., 2020 (**Fig. 3f, Extended Data Fig. 2**). To assess the accuracy of the ZENomix prediction for both AD-mutant and wildtype mice, we compared the predicted spatial transcriptome with the original ST data by computing Moran's I statistic, a measure of spatial autocorrelation, for all genes (n = 16,037 and 16,046 genes for AD-mutant and wildtype data, respectively) and found that the predicted and original spatial transcriptomes were well correlated (CC = 0.77 for the AD-mutant; CC = 0.79 for the wildtype; **Fig. 3g-h**). These results clearly supported that ZENomix has the ability to perform zero-shot learning for reconstruction of mutant spatial transcriptomes.

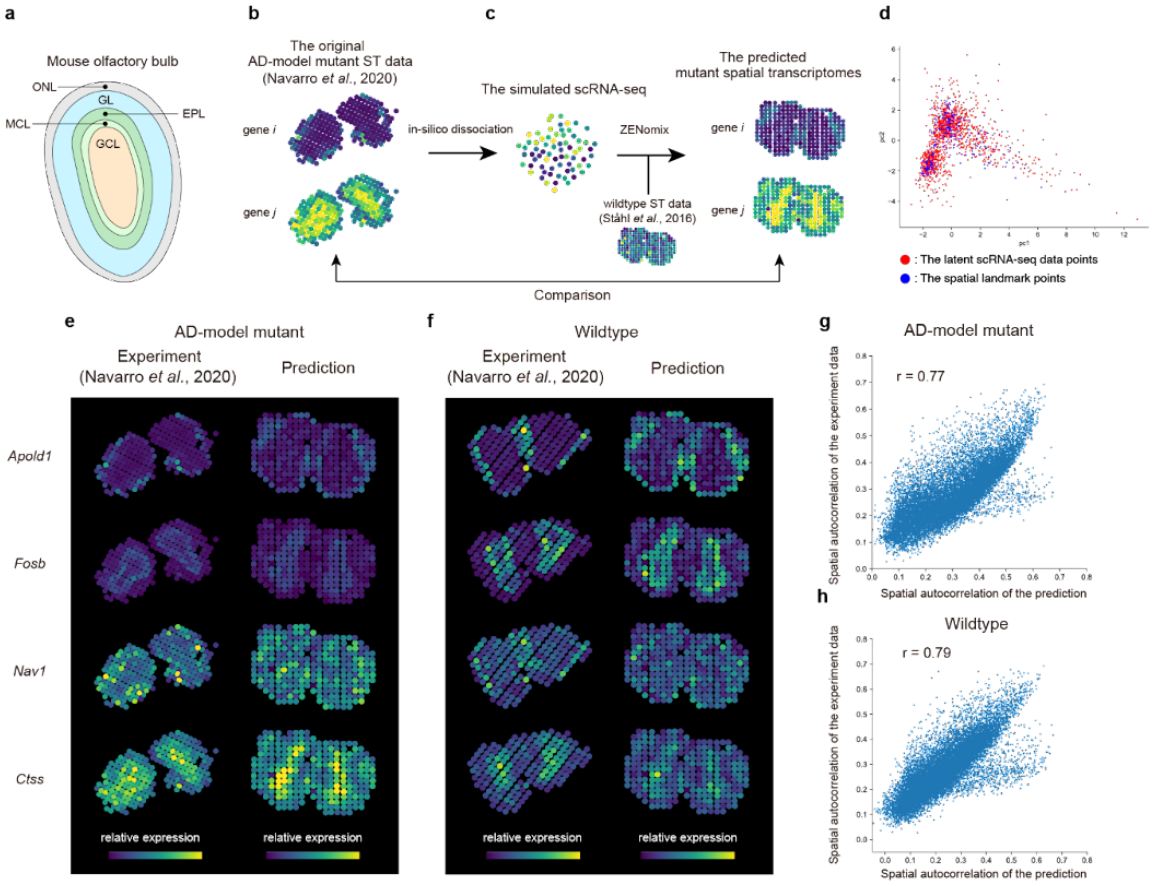


Figure 3: Zero-shot reconstruction of AD-mutant spatial transcriptomes
(a) Anatomy of a mouse olfactory bulb[17]. "GCL", "MCL", "EPL", "GL", and "ONL" indicate granular cell layer, mitral cell layer, external plexiform layer, glomerular layer, and olfactory nerve layer, respectively. (b-c) Flow of the experiment. (b) AD-mutant ST data from Navarro *et al*., iScience, 2020[3]. Simulated scRNA-seq data are generated by in-silico dissociation of the AD-mutant ST data. (c) The zero-shot reconstruction by ZENomix. ZENomix reconstructs the AD-mutant spatial transcriptomes from the simulated mutant scRNA-seq data from Navarro *et al*., iScience, 2020.[3]. The wildtype ISH data from Ståhl *et al*., Science, 2016.[12] is used as side information. The prediction accuracy can be evaluated by comparing the predicted and original AD-mutant spatial transcriptomes. (d) Scatter plot of the calibrated distributions of spatial landmark points and scRNA-seq data points (corresponding to Fig.2b). Principal component analysis was used to visualize the latent space. (e-f) The original and predicted spatial transcriptomes of (e) the AD-mutant and (f) the wildtype. (g-h)

Comparison between the original and predicted spatial transcriptomes for (g) the AD-mutant and (h) the wildtype. The original and predicted spatial profiles were quantitatively evaluated by the spatial autocorrelation, Moran's I value. Each dot indicates each gene (n = 16,037 and 16,046 genes for AD-mutant and wildtype data, respectively).

ZENomix predicted mutant spatial transcriptomes in MZoep zebrafish embryo

Next, we applied ZENomix to the scRNA-seq data of the mutant zebrafish embryo, to which the previous experimental method is difficult to apply due to the dome-shaped small tissue. We used the scRNA-seq data in an early embryo (at stage of 50% epiboly) of maternal-zygotic *one-eyed pinhead* (*MZoep*)[18] mutant obtained by Farrel, *et al*., 2018.[19]. *MZoep* mutant lacks the coreceptor of Nodal signaling, causing the defects in the dorsal organizer, which induces mesoendodermal formation[20,21] (**Fig. 4a**). For the spatial reference data, we used the binary ISH data including 47 genes obtained by Satija, *et al*, 2015.[11] (**Fig. 4b**).

First, we confirmed that ZENomix calibrated the difference in the distributions of two data in the latent space (**Fig. 4c, Extended Data Fig. 1-2**). Then, ZENomix tried to predict mutant spatial transcriptomes of several genes known to be perturbed in the *MZoep* mutant[18,22]. Notably, these all predictions were consistent with the previously published ISH images (**Fig. 4d**). Thus, ZENomix recapitulated spatial gene expression changes caused by mutation in Nodal signaling. These results clearly showed that ZENomix can successfully predict mutant spatial transcriptomes in a zero-shot manner.

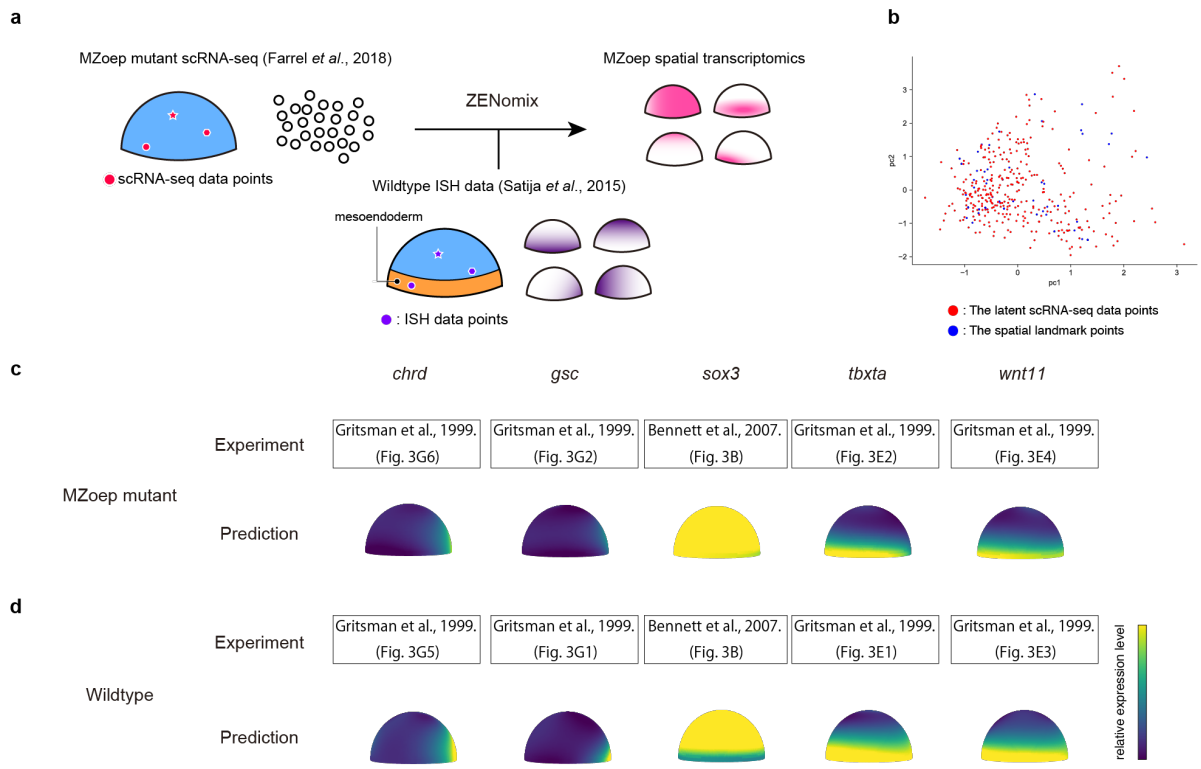


Figure 4: Prediction of spatial transcriptomes of zebrafish mutant embryos by ZENomix
(a) Flow of the experiment. ZENomix predicts the *MZoep*-mutant spatial transcriptomes from the mutant scRNA-seq data from Farrel *et al*., Science, 2018. The wildtype ISH data from Satija *et al*., Nat Biotech, 2015 is used as side information. (b) Scatter plot of the calibrated distributions of spatial landmark points and scRNA-seq data points (corresponding to Fig.2b). Principal component analysis was used to visualize the latent space. (c-d) The experiment and prediction of the spatial transcriptomes of (d) the *MZoep*-mutant and (e) the wildtype. For the experiment, please see the ISH images of

206 embryos in the similar developmental stages from Gritsman *et al*., Cell, 1999.[18] (*chrd*, *gsc*, *tbxta*, and *wnt11*; upper view)
207 and Bennett *et al*., Developmental biology, 2007.[22] (*sox3*; lateral view).
208

209 ***ZENomix's predictions identified the putative spatially differentially expressed genes***

210 To identify the spatially differentially expressed (DE) genes in *MZoep* embryos, we compared the mutant
211 reconstructed spatial transcriptome data with those of wildtype (**Fig. 5**). First, we computed the difference
212 between the reconstructed spatial transcriptomes in the wildtype and mutant (**Fig. 5a**). Then, we plotted the
213 maximum absolute value and standard deviation of expression changes in the *MZoep* mutant transcriptome for
214 each gene (n = 26,545 genes), and screened 142 Nodal-associated genes which could be associated with the
215 Nodal signaling (**Fig. 5b**). For further screening, we focused on the expression changes in the embryo margin
216 (red box in **Fig. 5c**), which is most affected by the defects in Nodal signaling. By plotting the average of the
217 expression changes only in the embryo margin, the Nodal-associated genes were able to be classified into two
218 groups: the 101 putative Nodal-upregulated (NU) genes and the 41 putative Nodal-downregulated (ND) genes
219 (**Fig. 5c, Extended Data Fig. 3-5**).

220 Next, to better understand the Nodal-associated gene expression changes, we performed the
221 hierarchical clusterings for each of the putative NU and ND genes (**Fig. 5d**). We identified the four and three
222 modules in the putative NU and ND genes, respectively (**Fig. 5e, Supplementary Table 1**). Finally, we
223 excluded the genes in module 1 of the putative NU genes and module 1 and 3 of the putative ND genes because
224 these genes showed the universal gene expression changes in the whole embryo, and obtained the putative 87
225 spatially DE genes (the 74 putative NU genes and the 13 putative ND genes). Note that 50.0 % (37/74) of these
226 putative NU genes and 15.3 % (2/13) of these putative ND genes were shared with the 72 genes and two genes,
227 respectively, identified by similar screening by bulk microarray in *MZoep* embryos[22], suggesting that our
228 screening was consistent with the previous study (**Fig. 5f**).

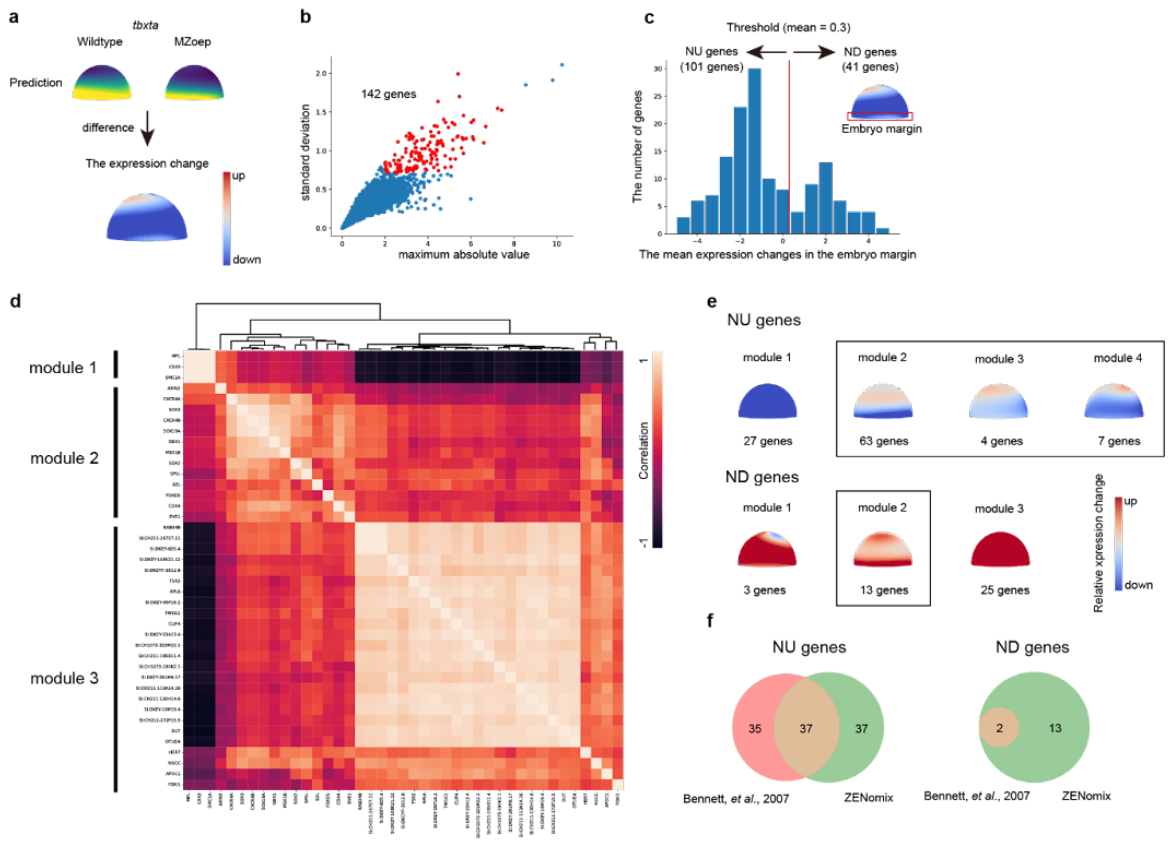


Figure 5: Screening of the putative spatially differentially expressed genes
(a) Example of expression changes calculated as subtracting the predicted spatial transcriptome of the wildtype from the predicted one of the *MZoep*-mutant embryos. (b) Scatter plot for the gene screening. The x and y axes indicate the maximum absolute value and standard deviation of the expression changes, respectively. Each dot indicates each gene (n = 26,545 genes). The red dots indicate the selected genes (Nodal-associated genes; n = 152 genes). (c) Histogram of the mean expression changes within the embryo margin. Red line (mean = 0.3) indicates the threshold of the classification between NU genes and ND genes. Redbox in the inset indicates the embryo margin, within which means are calculated. (d) The hierarchical clustering of the putative ND genes. The heatmap indicates the correlation of gene expression changes among the newly-screened ND genes. (e) Average of spatial gene expression changes for each module. The four modules for the NU genes and three modules for the ND genes were found in (e) and Extended Data figure *. The black circles indicate the modules that are not excluded from the further analysis. (f) Venn diagrams for sets of putative NU and ND genes screened by our and the previous study[22].

Discovery of the new genes repressed by Nodal signaling

As a function of Nodal signaling, it is important not only to induce and maintain the dorsal organizer, but also to repress ectodermal cell fates[22]. Although Bennett *et al*., 2007.[22] have suggested that Nodal represses some target genes, so far there are only two genes (*sox2* and *sox3*) identified as the genes downregulated by Nodal signaling. Notably, our screening by ZENomix discovered the 13 putative ND genes (11 unknown genes, and *sox2* and *sox3*) (**Fig. 5f**).

To validate our candidates of ND genes, we checked if the reconstructed spatial transcriptome of the 11 unknown candidate genes in the *MZoep*-mutant embryos follows the spatial gene expression profiles by the ISH experiment. Since *MZoep* is not maintained in our zebrafish facility, and since overexpression of *lefty1*, which encodes the Nodal inhibitor, can phenocopy *MZoep* mutants and *cyclops/squint* double mutants[23,24], we

used *lefty1* overexpressed embryos instead of *MZoep* mutants for the ISH evaluation of ZENomix predictions (the *lefty1* overexpressed embryos; see Methods). Surprisingly, we found that, among the 11 candidate genes, the predicted spatial gene expression patterns of the eight genes (*cdx4*, *cxcr4a*, *cxcr4b*, *eve1*, *foxd5*, *sox19a*, *sp5l*, and *szl*) were well-correlated with their ISH experiments, indicating that ZENomix discovered the eight new genes, which are repressed by Nodal signaling (**Fig. 6**). Therefore, we concluded that the mutant spatial transcritptomes predicted by ZENomix provided us with the new biological insight to the regulation of gene expressions.

For the other three candidate genes (*axin2*, *gbx1*, and *msx1b*), we found that their expression patterns were not consistent with their ISH experiments (**Extended Data Fig. 6**). Specifically, *axin2* showed the similar, ubiquitous expression patterns between wildtype and *lefty1*-injected embryos in the ISH experiments, although *axin2*-positive cells were increased in *MZoep*-mutant embryos compared with wildtype embryos in the scRNA-seq data. (**Extended Data Fig. 6**). *gbx1* and *msx1b* showed no expression for the both embryos in the ISH experiments, although these genes were expressed in the scRNA-seq data **(Extended Data Fig. 6)**. It is possibly because their expression properties were different between the *MZoep*-mutant embryos and the *lefty1* overexpressed embryos for some reason, e.g., measurement difference, batch effects, effects of mRNA injection, and phenotypic differences between *MZoep* and *lefty1* overexpression.

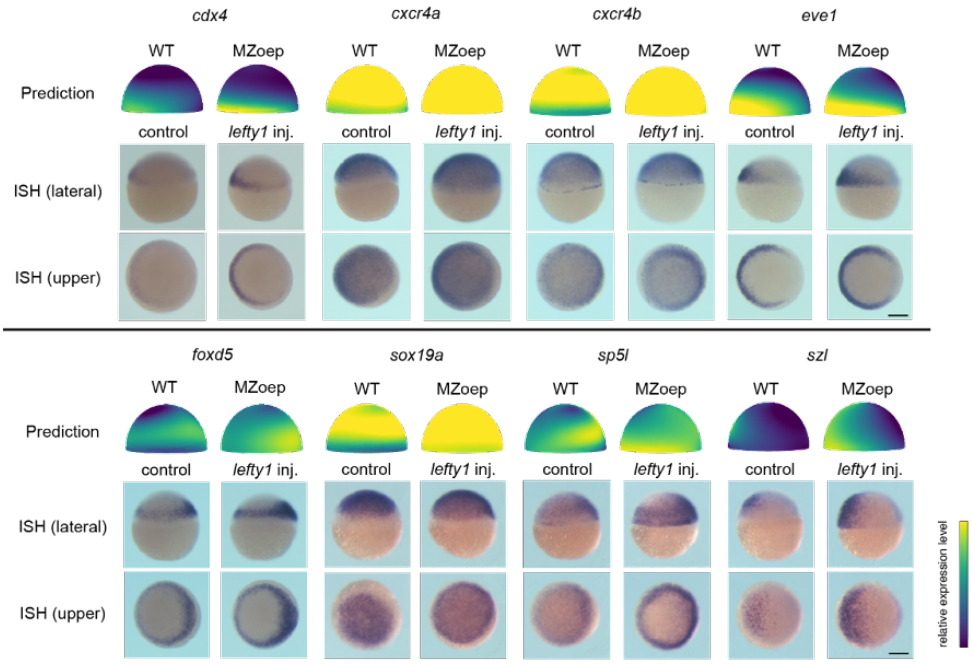


Figure 6: ISH validation of the Nodal-downregulated genes discovered from ZENomix-predicted mutant spatial transcriptomes

The whole-mount ISH experiments of the identified eight ND genes. Top to bottom: ZENomix's predicted expression pattern, a lateral view of the ISH, and an upper view of the ISH. For each gene, wildtype and mutant expression patterns are displayed (ZENomix's prediction: wildtype and *MZoep*-mutant embryos; ISH experiment: control and *lefty1*-injected embryos). Scale bar, 200 μm.

Discussion

We developed ZENomix, a computational framework to reconstruct mutant spatial transcriptomes without teaching data by introducing a new concept, zero-shot reconstruction. ZENomix used wildtype spatial gene expression atlas as side information, and extracted the landmark points of the spatial coordinates in tissues from wildtype spatial reference atlas, which enabled us to recover the spatial information of mutant scRNA-seq data. Using the simulated and real scRNA-seq data, we demonstrated that ZENomix can accurately predict spatial transcriptomes of AD-mutant model mice and *MZoep*-mutant zebrafish embryos, respectively. Furthermore, we offered a new spatially-informed screening approach based on ZENomix's prediction. Indeed, we discovered the new eight Nodal-downregulated genes in zebrafish early embryos. Therefore, we concluded that ZENomix must be useful tools to reveal new insights in mutant/diseased tissues by finding new genes which are perturbed.

The proposed zero-shot reconstruction contributes to the field of spatial transcriptomes in two ways. The first one is that our method can reuse published mutant/diseased scRNA-seq data. Because *in situ* data of mutant/diseased of interest is not needed in the cross-genotype integration, re-analysis of the public mutant/diseased scRNA-seq data can yield the mutant spatial transcriptomes. The importance of this point must be increased because a lot of patient-derived scRNA-seq data are collected these days[25]. The second contribution of ZENomix is its applicability to small, sphere- or dome-like tissues such as drosophila and zebrafish early embryos, which still have difficulty in applying spatial transcriptomics technology. To our best knowledge, there are only a few studies revealing spatial transcriptomes of small embryos of drosophila and zebrafish using spatial transcriptomics technology (Stereo-seq[26]), while these tissues are well studied using computational reconstruction methods[11,27]. With these two innovative points, the zero-shot reconstruction could be a breakthrough in spatial biology.

While many studies including our previous method, Perler, have reconstructed spatial transctiptomes from scRNA-seq data using various algorithms[11,28–35], it can be stated that no method can perform our zero-shot reconstruction of mutant spatial transcriptomes for two reasons. First, both scRNA-seq and *in situ* data are modeled by assuming to be in the same biological condition. For example, a neural network-based method, gimVI[32], assumes that *in situ* and scRNA-seq data are generated from the shared latent biological state by using the same function. ZENomix overcomes this problem by modeling mutant and wildtype data with different functions based on Gaussian Process (GP) prior. The second reason is that most of the previous methods optimized similarity measures (e.g. cosine similarity[34], mean squared error[35], inverse correlation[31], and mutual information[33]) between the *in situ* data and the reconstructed transcriptomes, making it impossible to apply to the integration between differently genotyped data. ZENomix relies on Bayesian estimation of spatial transcriptomes, enabling us to reconstruct them without any similarity measures.

In this study, we conducted a new screening approach based on ZENomix's prediction to reveal genes whose expressions are perturbed in mutant tissues. The important feature of our screening approach is that it can use the gene expression changes from wildtype to mutant at the specific region (e.g. embryonic

margin in *MZoep*), which is contrary to the bulk transcriptomics-based screening with microarray and RNA-seq. By using our spatially informed-screening, we discovered the new eight Nodal-downregulated genes, which has not been identified so far. Note that we also identified the 74 putative Nodal-upregulated genes, which are consistent with the previous bulk microarray-based screening[22] (Fig. 5f). These results highlight the importance of spatial information in knock down/out analysis, and suggest that ZENomix can find new biological mechanisms via zero-shot reconstruction.

Methods

Zebrafish

Wildptype zebrafish were used in this study. All zebrafish experiments were carried out with the approval of the Nara Institute of Science and Technology's Animal Studies Committee.

The whole-mount ISH experiment

Whole-mount ISH was performed as described previously[36,37]. Template DNAs for *axin2*, *cdx4*, *cxcr4a*, *eve1*, *foxd5*, *gbx1*, *msx1b*, *sox19a*, *sp5l*, and *szl* were obtained by PCR using primers that T3 and T7 sequences were added at the 5' ends of the forward and reverse primer, respectively. The sequence was confirmed using T3 and T7 primers. pCRII-*cxcr4b* was also used as a template. DIG antisense RNA probes were synthesized using T7 or SP6 RNA polymerase.

Overexpression of lefty1 in zebrafish embryos

lefty1 mRNAs were synthesized using a mMessage mMachine SP6 transcription kit (Thermo Fisher Scientific) from pCS2-*lefty1* (Kind gift from Dr. Masashiro Hibi) as a template[24]. The synthesized *lefty1* mRNAs (5 pg) were injected into the one-cell stage of zebrafish embryos and used for the ISH experiments.

ZENomix model

ZENomix takes two data, wildtype spatial reference data and mutant scRNA-seq data, as inputs. In the spatial reference data (e.g. ST data) generated by *in situ* methods, gene expression vectors are available for all regions/cells, whose locations in the tissue are known. The gene-expression vector of gene *j* is represented as $y^{(W)}{}_j = \left(y^{(W)}{}_{j,1}, \dots, y^{(W)}{}_{j,n_W}, \dots, y^{(W)}{}_{j,N_W}\right) \in R^{N_W}$, where cells are indexed by n_W ($n_W \in \{1,2,\dots, N_W\}$), and N_W is the total number of cells in the tissue of interest. By contrast, in an scRNA-seq dataset of genotype of interest, gene-expression vectors lack information regarding cell location in tissue. The gene-expression vector of gene *j* is represented as $y^{(M)}{}_j = \left(y^{(M)}{}_{j,1}, \dots, y^{(M)}{}_{j,n_M}, \dots, y^{(M)}{}_{j,N_M}\right) \in R^{N_M}$, where cells are indexed by n_M ($n_M \in \{1,2,\dots, N_M\}$, and N_M is the total number of cells used for scRNA-seq measurement. Note that the number of dimensions of these two data, p, are the same because we take the landmark genes.

Generative model

The measurements of scRNA-seq data and spatial reference data were modeled as a generative model:

$$P(X) = N(X|0, I), \tag{1}$$

$$P\left(Y^{(W)}, Y^{(M)} \middle| X, \sigma^2\right) = \Pi_j P\left({y^{(W)}}_j \middle| X, \sigma^2\right) P\left({y^{(M)}}_j \middle| X, \sigma^2\right), \quad (2)$$

where $N(0, I)$ is a standard Gaussian distribution with q dimensions, $Y^{(W)} \in R^{p \times N_W}$ and $Y^{(M)} \in R^{p \times N_M}$ indicate a *in situ* data matrix and scRNA-seq data matrix, respectively, X indicates the common latent variables matrix (spatial landmark points) whose dimensionality is q, and h indicates the genotype or the modality of data ($h \in$ {W (wildtype spatial reference data), M (mutant scRNA-seq data)}). Here, the gene expression is generated by functions of low-dimensional latent variables as

$$y^{(h)}{}_j = f^{(h)}{}_j(X) + \epsilon, \quad (3)$$

where ϵ indicates the Gaussian observation noise with zero mean and standard deviation σ. Note that σ is common among modalities of the data. The aim is to estimate the posterior distribution of latent variables given as $P\left(X \middle| Y^{(W)}, Y^{(M)}\right)$.

In practice, the Gaussian process prior were placed on the projection functions, $f^{(h)}{}_j$, to assume non-linear transformation from latent variables to gene expression levels, that is,

$$P\left(y_j^h \middle| X, \sigma\right) = N(0, K_h + \sigma^2 I), \quad (4)$$

where K_h is the gram matrix defined by the kernel function $k\left(x^{(h)}, x'^{(h)}; \theta_k\right)$ between two distinct latent variables, $x^{(h)}$ and $x'^{(h)}$ with kernel hyperparameters, θ_k, and $x^{(h)}$ is a latent variable corresponding to each observation, $y^{(h)}$.

Inference scheme

In the ZENomix model, the aim is to estimate the posterior probability of the common latent variables (spatial landmark points), $P\left(X \middle| Y^{(W)}, Y^{(M)}\right)$. However, a major problem exists for this estimation; it is difficult to estimate the latent variables shared by $Y^{(W)}$ and $Y^{(M)}$, because two data are unpaired, that is, we cannot use methods for paired data (e.g. canonical correlation analysis (CCA)[38] and Multimodal Mixture-of-Experts VAE[39]). To overcome this problem, a new inference scheme, vGPLVM-MMD, was proposed, which extracts spatial landmark points from wildtype data.

The vGPLVM-MMD scheme consists of two parts. The first part is to map $Y^{(W)}$ and $Y^{(M)}$ onto low-dimensional latent space, independently, and the second part to match these latent data distributions. In

$$P\left(X^{(h)}\right) = N\left(X^{(h)} \middle| 0, I\right), \quad (6)$$

$$P\left(Y^{(h)} \middle| X^{(h)}, \sigma^2\right) = \Pi_j P\left({y^{(h)}}_j \middle| X^{(h)}, \sigma^2\right), \quad (5)$$

the first part, the generative model above was separated into two independent models of $Y^{(W)}$ and $Y^{(M)}$ as where $X^{(h)}$ is a separated latent variable matrix corresponding to each data matrix, $Y^{(h)}$. Consequently, the gene expression is generated by function of low-dimensional latent variables, $X^{(h)}$, which is different for each data as,

$$y^{(h)}{}_j = f^{(h)}{}_j\left(X^{(h)}\right) + \epsilon, \tag{7}$$

where the Gaussian process prior were placed on the projection functions, $f^{(h)}{}_j$. From this model, we can estimate the posterior distributions $P\left(X^{(h)}\middle|Y^{(h)}\right)$ for each data. In practice, we estimated the approximated posterior distributions, $q\left(X^{(h)}\right)$, using variational inference instead of $P\left(X^{(h)}\middle|Y^{(h)}\right)$. Note that in this model, $q\left(X^{(W)}\right)$ and $q\left(X^{(M)}\right)$ are different distributions.

This model (Eq. (6-7)) is known as the Bayesian Gaussian Process latent variable model (Bayesian GPLVM)[15]. Because the posterior distribution is analytically intractable in this model, Titsias and Lawlence, 2009.[15] and Damianou *et al*., 2016.[16] developed the variational GPLVM using the inducing inputs. Following their methods, the generative model can be described by

$$P\left(X^{(h)}\right) = \Pi_{n_h} N\left(x^{(h)}{}_{n_h}\middle|0, I\right), \tag{11}$$

$$P(U|X_u) = \Pi_j N\left(u_j\middle|0, K_{uu}\right), \tag{8}$$

$$P\left(F^{(h)}\middle|U, X^{(h)}, X_u\right) = \Pi_j N\left(f_j{}^{(h)}\middle|K_{hu}K_{uu}{}^{-1}u_j, Tr(K_h) - K_{hu}K_{uu}{}^{-1}K_{uh}\right), \tag{9}$$

$$P\left(Y^{(h)}\middle|F^{(h)}\right) = \Pi_j N\left(y^{(h)}{}_j\middle|f_j{}^{(h)}, \sigma^2 I\right), \tag{10}$$

where $X_u\left(= \left(x^{(u)}{}_1 \dots x^{(u)}{}_m \dots x^{(u)}{}_{N_u}\right)\right) \in R^{q\times N_u}$ and $U\ (= (u_1\ \dots u_j\ \dots\ u_p)^T) \in R^{p\times N_u}$ indicate the inducing inputs in the latent space and extra samples in the observation space, respectively, and $F^{(h)}$ indicates the GP mapping values. The gram matrices, $K_{hu} \in R^{N_h\times N_u}$ and $K_{uu} \in R^{N_u\times N_u}$, in Eq. (9) are defined as

$$(K_{hu})_{nm} = k\left(x^{(h)}{}_n, x^{(u)}{}_m\right), \tag{12}$$

$$(K_{uu})_{mm'} = k\left(x^{(u)}{}_m, x^{(u)}{}_{m'}\right). \tag{13}$$

Here, variational inference is applied to approximate the true posterior, $P(X^{(h)}|Y^{(h)})$ by a Gaussian distribution:

$$q\left(X^{(h)}\right) = \prod_n^{N_h} N\left(x^{(h)}{}_n\middle|\mu^{(h)}{}_n, s^2 I\right), \tag{14}$$

where s indicates noise intensity common in all distributions. Then, the lower bound of $P(Y^{(h)})$ is given as follows:

$$F\left(q\left(X^{(h)}\right)\right) = \sum_j^p \hat{F}_j\left(q\left(X^{(h)}\right)\right) - KL\left(q\left(X^{(h)}\right)\middle|p\left(X^{(h)}\right)\right), \tag{15}$$

where

$$\hat{F}_j\left(q\left(X^{(h)}\right)\right) = log\left[\frac{(\beta)^{\frac{N_h}{2}}|K_{uu}|^{\frac{1}{2}}}{(2\pi)^{\frac{N_h}{2}}|\beta\Psi_2+K_{uu}|^{\frac{1}{2}}}e^{-\frac{1}{2}y_j{}^T W y_j}\right] - \frac{\beta\psi_0}{2} + \frac{\beta}{2}Tr\left(K_{uu}{}^{-1}\Psi_2\right), \tag{16}$$

$$W = \beta I - \beta^2\Psi_1(\beta\Psi_2 + K_{uu})^{-1}\Psi_1{}^T, \tag{17}$$

$$\psi_0 = Tr\left(< K_h >_{q\left(X^{(h)}\right)}\right), \tag{18}$$

$$\Psi_1 =< K_{hu} >_{q\left(X^{(h)}\right)}, \tag{19}$$

$$\Psi_2 =< K_{hu}K_{hu}{}^T >_{q\left(X^{(h)}\right)}. \tag{20}$$

Here, $<\cdot>_{p(x)}$ denotes expectation under the distribution $p(x)$. $KL(q(x)\,||\,p(x))$ indicates Kullback–Leibler (KL) divergence between distribution $q(x)$ and $p(x)$.

In our implementation, $\mathbf{\Psi}$ statistics (ψ_0, Ψ_1, and Ψ_2) is computed using Gauss-Hermite approximation as in Gpy[40]. By maximizing the variational bound for each genotype, *h*, the latent representation of each data can be obtained.

In the second part of the vGPLVM-MMD scheme, the separately derived posterior distributions, $q\left(X^{(W)}\right)$ and $q\left(X^{(M)}\right)$, are combined so that two distributions were nearly matched as $q\left(X^{(W)}\right) \approx q\left(X^{(M)}\right)$ to extract the common latent variables, X. To make the two distributions almost identical, a distance between the two distributions, $q\left(X^{(W)}\right)$ and $q\left(X^{(M)}\right)$ was minimized. The distance was calculated by Maximum Mean Discrepancy (MMD)[14] (see *Calculation of MMD*):

$$Dist\left(q\left(X^{(W)}\right), q\left(X^{(M)}\right)\right) \coloneqq MMD\left(M^{(W)}, M^{(M)}\right), \tag{21}$$

where $M^{(h)}(= (\mu^{(h)}{}_0 \ldots \mu^{(h)}{}_n \ldots \mu^{(h)}{}_{N_h}))$ is a mean of $q\left(X^{(h)}\right)$.

Calulation of MMD

MMD is a nonparametric distance between two sample distributions by embedding them in a Reproducing Kernel Hilbert Space (RKHS). Let assume general situation that there are two data samples X and Y with the same dimension. The empirical estimate of MMD between two data, X and Y is

$$MMD(X,Y) = ||\frac{1}{N_x}\sum_{n_x}^{N_x}\phi\left(x_{n_x}\right) - \frac{1}{N_y}\sum_{n_y}^{N_y}\phi\left(y_{n_y}\right)||^2, \tag{22}$$

where ϕ indicates the kernel-induced feature map. By expanding the Eq. (22) and replacing the inner products by their kernel values (the kernel trick), MMD is given as

$$MMD(X,Y) = \frac{1}{{N_x}^2}\sum_{n_x,n_{x'}} l(x_{n_x}, x_{n_{x'}}) - \frac{2}{N_x N_y}\sum_{n_x,n_y} l\left(x_{n_x}, y_{n_y}\right) + \frac{1}{{N_y}^2}\sum_{n_y,n_{y'}} l\left(y_{n_y}, y_{n_{y'}}\right), \tag{23}$$

where l(x, y) denotes the kernel function with MMD hyperparameters, θ_l. It is known that when RKHS is universal, MMD will asymptotically approach zero if and only if the two distributions are the same[14].

Variational inference

To estimate the parameters and hyperparameters of ZENomix, we combined the first and second **part** of vGPLVM-MMD. Thus, the cost function of vGPLVM-MMD is

$$L(X_u, M^{(W)}, M^{(M)}, S, \sigma, \theta_k) = F\left(q(X^{(W)})\right) + F\left(q(X^{(M)})\right) - Dist\left(q(X^{(W)}), q(X^{(M)})\right). \tag{24}$$

This cost function was maximized by a gradient-based optimization algorithm (see the *ZENomix implementation* section) to extract the common latent variables of the two data. The MMD hyperparameters, θ_l are fixed in ZENomix. Note that computational cost of evaluating $F\left(q(X^{(h)})\right)$ is $O(N_h * {N_u}^2)$ and that of $MMD(M^{(W)}, M^{(M)})$ is $O((N_I + N_R)^2)$. Then, if $N_u = O(\sqrt{N_I + N_R})$, computational complexity of evaluating L is $O((N_I + N_R)^2)$

Spatial reconstruction

In the second step of ZENomix, mutant spatial gene expression profiles of gene *j* are obtained by mapping the latent variables of wildtype *in situ* data, $X^{(W)}$, to mutant scRNA-seq space as ${F^{*(M)}}_j := {f^{(M)}}_j(X^{(W)})$. The posterior distribution of ${F^{*(M)}}_j$ can be inferred by using Eq. (11-10) as

$$\begin{aligned} P\left({F^{*(M)}}_j \middle| Y^{(W)}, Y^{(M)}\right) &= \int P\left(F^{(M)}, X^{(W)}, U \middle| Y^{(W)}, Y^{(M)}\right) dX^{(W)} dU \\ &= \int \left[\int P\left(F^{(M)} \middle| U, X^{(W)}\right) q(U) dU\right] q\left(X^{(W)}\right) dX^{(W)} \\ &= \int q\left(F^{(M)} \middle| X^{(W)}\right) q\left(X^{(W)}\right) dX^{(W)}. \end{aligned} \tag{25}$$

This integral is analytically intractable, but Titsias and Lawrence, 2009. and Damianou, *et al*., 2016. shows that mean and covariance of ${F^{*(M)}}_j$ can be calculated as follows:

$$E[{F^{*(M)}}_j] = (\Lambda^T {\Psi^{(W)}}_1)_j, \tag{26}$$

$$\begin{aligned} Cov[{F^{*(M)}}_j] &= \Lambda^T \left({\Psi^{(W)}}_2 - {\Psi^{(W)}}_1 {\Psi^{(W)}}_1^{\,T}\right)\Lambda + {\psi^{(W)}}_0 I \\ &- Tr\left(\left[{K_{uu}}^{-1} - \left(K_{uu} + \beta {\Psi^{(M)}}_2\right)^{-1}\right] {\Psi^{(W)}}_2\right) I, \end{aligned} \tag{27}$$

$$\text{where } \Lambda = \beta(K_{uu} + \beta\Psi^{(M)}{}_2)^{-1}\Psi^{(M)}{}_1{}^T Y^{(M)}, \tag{28}$$

$$\Psi^{(W)}{}_1 =< K_{Iu} >_{q(X^{(W)})}, \tag{29}$$

$$\Psi^{(W)}{}_2 =< K_{Iu}K_{Iu}{}^T >_{q(X^{(W)})}, \tag{30}$$

$$\Psi^{(M)}{}_1 =< K_{Ru} >_{q(X^{(M)})}, \tag{31}$$

$$\Psi^{(M)}{}_2 =< K_{Ru}K_{Ru}{}^T >_{q(X^{(M)})}. \tag{32}$$

ZENomix implementation

In our implementation, we choose the Matern 3/2 kernel as both kernel functions, $k(x, y)$ and $l(x, y)$ with length scale $1/\sigma^f{}_k(= \theta_k)$ and $1/\sigma^f{}_l(= \theta_l)$, respectively. To initialize the mean of the posterior distributions, M_I and M_R, the principal component analysis (PCA) was used for the input data. Before the first step of ZENomix, all input data are z-scored, and $\sum_h F\left(q\left(X^{(h)}\right)\right)$ and $MMD\left(M^{(W)}, M^{(M)}\right)$ are divided by their initial values for normalization. For the detailed initial parameter settings, see **Supplementary Table 2**. For the gradient-based optimization method, we used the L-BFGS-B algorithm implemented in Scipy.

Data collection and preprocessing

mouse olfactory bulb data

The mouse WT olfactory bulb ST (WT ST) data is originally generated by Ståhl *et al*., 2016.[12]. The normalized data and a list of highly variable genes were downloaded from SpatialDB[41]; http://www.spatialomics.org/SpatialDB/) and Rep11 were used for further analysis. The mouse AD olfactory bulb ST (AD ST) dataset was downloaded from Mendeley Data[3] (https://doi.org/10.17632/6s959w2zyr.1). The normalized mouse WT and AD data (WT and 3xAD , respectively) were used for further analysis (preprocessing and integration with WT ST data). For integration of WT ST data with AD ST data, spatial information of AD ST data was ignored and each spot was treated as a simulated scRNA-seq data point. As landmark genes, we used the highly variable genes of WT reference downloaded from SpatialDB. The genes, which are not included in the simulated scRNA-seq data or whose expressions are all zero in the simulated scRNA-seq data, were removed from the landmark genes measured in *in situ* hybridization data.

zebrafish early embryo data

The zebrafish early embryo *in situ* hybridization data was downloaded from the Satija lab homepage (https://satijalab.org/). For the zebrafish early embryo scRNA-seq data, we downloaded the raw data from the Gene Expression Omnibus (GEO) database with accession number GSE106587 and preprocessed as described in the original article[19] for both genotypes (WT and *MZoep*). The genes, which are not included in scRNA-seq

463 data or whose expressions are all zero in scRNA-seq data, were removed from the landmark genes measured
464 in *in situ* hybridization data. The geometry data are obtained from Cang *et al*., 2020.[42].
465
466 ***ZENomix and Downstream analysis***
467 ZENomix is built on Python 3.10.5 and distributed through GitHub ((URL). All other softwares used in the
468 ZENomix is publicly available: Numpy==1.22.4 (https://numpy.org/) for calculation; scipy==1.9.3
469 (https://scipy.org/) for calculation; jax==0.3.25 (https://jax.readthedocs.io/en/latest/index.html) for
470 calculation; scikit-learn==1.1.3 (https://scikit-learn.org/) for traditional machine learing (e.g. PCA);
471 pandas==1.5.1 (https://pandas.pydata.org/) for reading data frames. For downstream analysis, we used
472 Scanpy==1.9.1[43] (https://scanpy.readthedocs.io/en/stable/) for scRNA-seq data analysis, Squidpy==1.2.3[44]
473 (https://squidpy.readthedocs.io/en/stable/) for spatial transcriptome data analysis, which all are publicly
474 available.
475
476 ***Calculating Moran's I statistics***
477 The accuracy of the predicted spatial transcriptomes is evaluated by comparing the spatial autocorrelations
478 (Moran's I value) of the predicted and original spatial transcriptome data. Moran's I value is calculated as

$$\mathrm{I} = \frac{N}{W}\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{33}$$

479 where N is the number of data points indexed by i and j, x is the data point, $\bar{x}$ is the mean of x, w_{ij} is
480 a matrix of spatial weights with zeroes on the diagonal, and W is the sum of all w_{ij}. For implementation, we
481 used the *gr.spatial_autocorr* function with mode = "moran" in Squidpy.
482
483 ***Selecting the spatially differentially expressed genes***
484 We first calculated max absolute value and standard deviation of the expression changes between *MZoep*
485 mutant spatial transcriptomes and those of wildtype. Then, 142 genes were selected by manually thresholding
486 8 > *max_diff* > 2 and *std* > 0.72. These genes were classified into two groups, the putative NU and ND genes
487 by the mean expression changes of the embryo margin. The embryo margin is defined as tier = "1-2" or "3-4"
488 in the zebrafish geometry data. The spatial gene expression profiles of the putative NU and ND genes were
489 clustered by hierarchical clustering using the *cluster.hierarchy.fcluster* function in Scipy based on the
490 correlation matrices of the gene expression changes, and obtained the four and three modules in the putative
491 NU and ND genes, respectively. The module 1 of the putative NU genes and the module 1 and 3 of the putative
492 ND genes are excluded because these genes show the universal gene expression change, and we finally
493 obtained 87 spatially differentially expressed genes (74 genes for the putative NU genes and 13 genes for the
494 putative ND genes).
495
496 ***Data visualization***
497 Data visualization was performed by using Matplotlib==3.6.2 (https://matplotlib.org/) and Seaborn==0.12.1
498 (https://seaborn.pydata.org/), which all are publicly available. For mouse olfactory bulb data, we used spatial

location data of spatial transcriptomics data of the collected wildtype and AD-mutant model mouse OB data (see *Data collection and preprocessing* section). For visualizing zebrafish embryos data, we used plot_zf function created by Cang *et al*., 2020.[42], which interpolates the original 64 data points *in situ* data.

Acknowledgments

We are grateful to Prof. Masahiko Hibi and Dr. Ken Nakae for their valuable discussions, Ms. Maiko Yokouchi for technical assistance, and Prof. Masahiko Hibi for kindly gifting pCS2-*lefty1*. We also thank for the research opportunity provided by Prof. Michiyuki Matsuda and the Medical Scientist Training Program, Kyoto University [to Y.O.]. This study was supported in part by the Moonshot R&D–MILLENNIA Program [grant number JPMJMS2024-9 to H.N.] by Japan Science and Technology Agency (JST), Grant-in-Aid for Transformative Research Areas (B) [grant number 21H05170 to H.N.], Grant-in-Aid for Scientific Research (B) [grant number 21H03541 to H.N.] from the Japan Society for the Promotion of Science (JSPS), and Cooperative Study Program of Exploratory Research Center on Life and Living Systems (ExCELLS) [program number 19-102 to H.N.], and Grant-in-Aids for Scientific Research (B) and Challenging Exploratory Research [grant numbers 22H02821 and 21K19265 to T.M.] from the JSPS.

Data Availability

This study was a re-analysis of existing data. The websites where the data are collected are described in the *Data collection and preprocessing* section of *Methods*.

Code Availability

ZENomix is developed under python 3.10.5 and is distributed through GitHub (URL).

References

1. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
2. Maniatis, S. *et al.* Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science* **364**, 89–93 (2019).
3. Navarro, J. F. *et al.* Spatial Transcriptomics Reveals Genes Associated with Dysregulated Mitochondrial Functions and Stress Signaling in Alzheimer Disease. *iScience* **23**, 101556 (2020).
4. Chen, W.-T. *et al.* Spatial Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease. *Cell* **182**, 976-991.e19 (2020).
5. Boyd, D. F. *et al.* Exuberant fibroblast activity compromises lung function via ADAMTS4. *Nature* **587**, 466–471 (2020).
6. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).
7. Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R. & Haque, A. An introduction to spatial transcriptomics for biomedical research. *Genome Med.* **14**, 68 (2022).
8. Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* **22**, 627–644 (2021).
9. Chen, J. *et al.* Knowledge-aware Zero-Shot Learning: Survey and Perspective. (2021) doi:10.48550/ARXIV.2103.00070.
10. Briscoe, J. & Small, S. Morphogen rules: Design principles of gradient-mediated embryo patterning. *Dev. Camb.* **142**, 3996–4009 (2015).
11. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
12. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
13. Lawrence, N. Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data. in *Advances in Neural Information Processing Systems* vol. 16 (MIT Press, 2003).
14. Gretton, A., Borgwardt, KM., Rasch, M., Schölkopf, B. & Smola, AJ. A Kernel Approach to Comparing Distributions. in *Proceedings of the 22. AAAI Conference on Artificial Intelligence* 1637–1641 (Biologische Kybernetik, 2007).
15. Titsias, M. K. & Lawrence, N. D. Bayesian Gaussian Process Latent Variable Model. in *Proceedings of the Thirteenth International Workshop on Artificial Intelligence and Statistics* (eds. Teh, Y. W. & Titterington, D. M.) vol. 9 844–851 (PMLR, 2010).
16. Damianou, A. C., Titsias, M. K. & Lawrence, N. D. Variational Inference for Latent Variables and Uncertain Inputs in Gaussian Processes. *J. Mach. Learn. Res.* **17**, 1–62 (2016).
17. Imamura, F., Ito, A. & LaFever, B. J. Subpopulations of Projection Neurons in the Olfactory Bulb. *Front. Neural Circuits* **14**, (2020).
18. Gritsman, K. *et al.* The EGF-CFC Protein One-Eyed Pinhead Is Essential for Nodal Signaling. *Cell* **97**,

121–132 (1999).

19. Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).

20. Gritsman, K., Talbot, W. S. & Schier, A. F. *Nodal signaling patterns the organizer*. (2000).

21. Schier, A. F. Nodal Morphogens. *Cold Spring Harb. Perspect. Biol.* **1**, a003459 (2009).

22. Bennett, J. T. *et al.* Nodal signaling activates differentiation genes during zebrafish gastrulation. *Dev. Biol.* **304**, 525–540 (2007).

23. Chen, Y. & Schier, A. F. The zebrafish Nodal signal Squint functions as a morphogen. *Nature* **411**, 607–610 (2001).

24. Ryu, S.-L. *et al.* Regulation of dharma/bozozok by the Wnt Pathway. *Dev. Biol.* **231**, 397–409 (2001).

25. Potter, S. S. Single-cell RNA sequencing for the study of development, physiology and disease. *Nat. Rev. Nephrol.* **14**, 479–492 (2018).

26. Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* (2022) doi:10.1016/j.cell.2022.04.003.

27. Karaiskos, N. *et al.* The Drosophila embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).

28. Halpern, K. B. *et al.* Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).

29. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* (2019) doi:10.1016/j.cell.2019.05.031.

30. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).

31. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* 1–6 (2019) doi:10.1038/s41586-019-1773-3.

32. Lopez, Romain *et al.* A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *ICML Workshop Comput. Biol.* (2019).

33. Okochi, Y., Sakaguchi, S., Nakae, K., Kondo, T. & Naoki, H. Model-based prediction of spatial gene expression via generative linear mapping. *Nat. Commun.* **12**, 3731 (2021).

34. Biancalani, T. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat. Methods* **18**, 1352–1362 (2021).

35. Shengquan, C., Boheng, Z., Xiaoyang, C., Xuegong, Z. & Rui, J. stPlus: a reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics* **37**, i299–i307 (2021).

36. Matsui, T. *et al.* Noncanonical Wnt signaling regulates midline convergence of organ primordia during zebrafish development. *Genes Dev.* **19**, 164–175 (2005).

37. Matsui, T., Murakami, T. & Ohki, K. Transient neuronal coactivations embedded in globally propagating waves underlie resting-state functional connectivity. *Proc. Natl. Acad. Sci.* **113**, 6556–6561 (2016).

38. Hotelling, H. RELATIONS BETWEEN TWO SETS OF VARIATES. *Biometrika* **28**, 321–377 (1936).

39. Shi, Y., N, S., Paige, B. & Torr, P. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) vol. 32

(Curran Associates, Inc., 2019).

40. GPy. GPy: A Gaussian process framework in python. (2012).

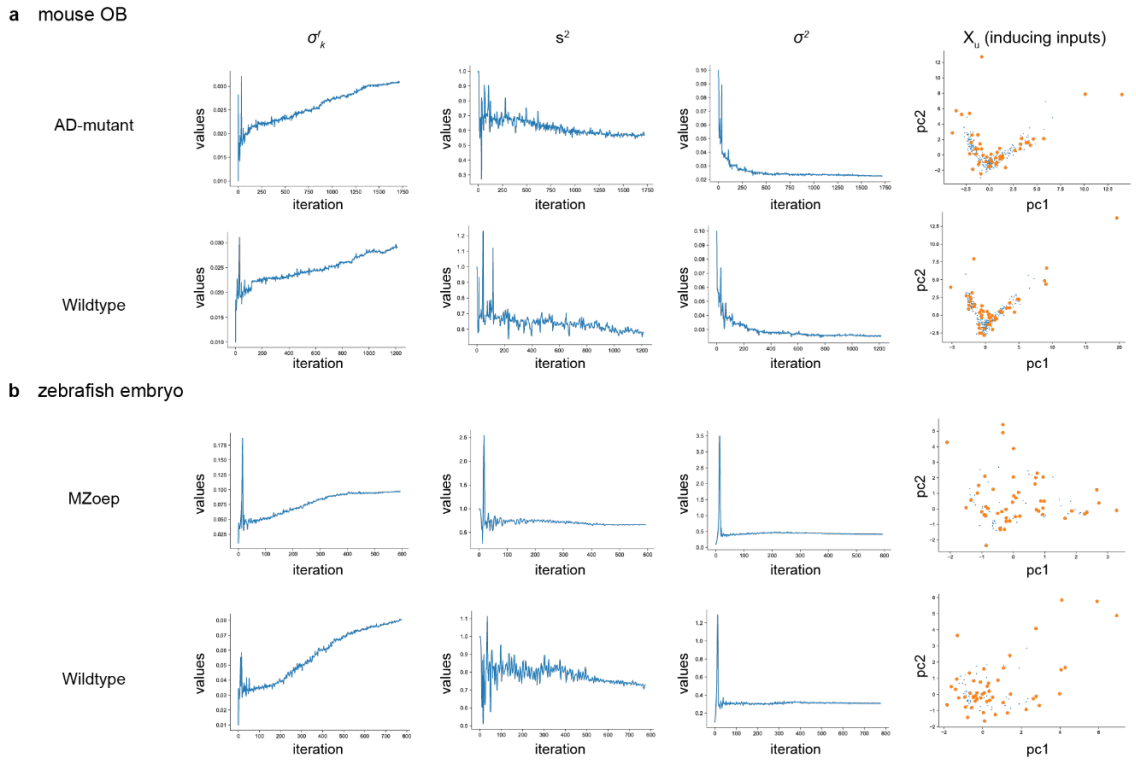
41. Fan, Z., Chen, R. & Chen, X. SpatialDB: a database for spatially resolved transcriptomes. *Nucleic Acids Res.* **48**, D233–D237 (2020).

42. Cang, Z. & Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun.* **11**, 2084 (2020).

43. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

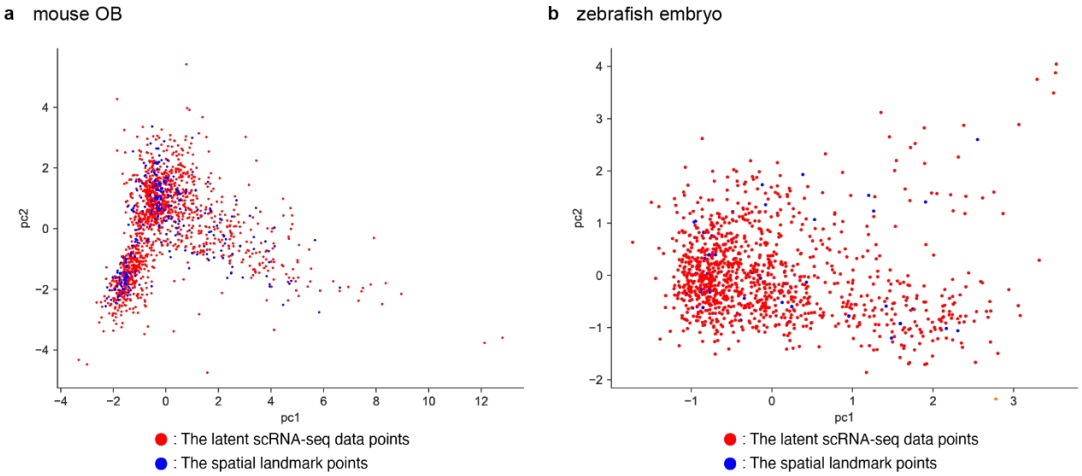
44. Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).

Extended Data Figures



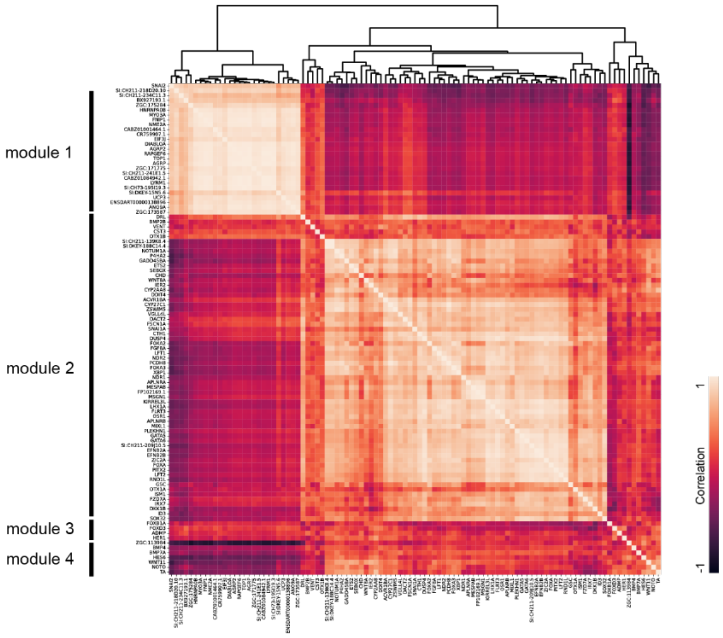
Extended Data Figure 1: Parameter convergence of ZENomix

Parameter values during the first step of ZENomix for the prediction of (a) mouse OB spatial transcriptomes and of (b) zebrafish embryo spatial transcriptomes. Each parameter is explained in the *Methods* section. For the inducing points, their final position is visualized by Principal component analysis. The big orange and small blue points indicate inducing points and spatial landmark points, respectively.



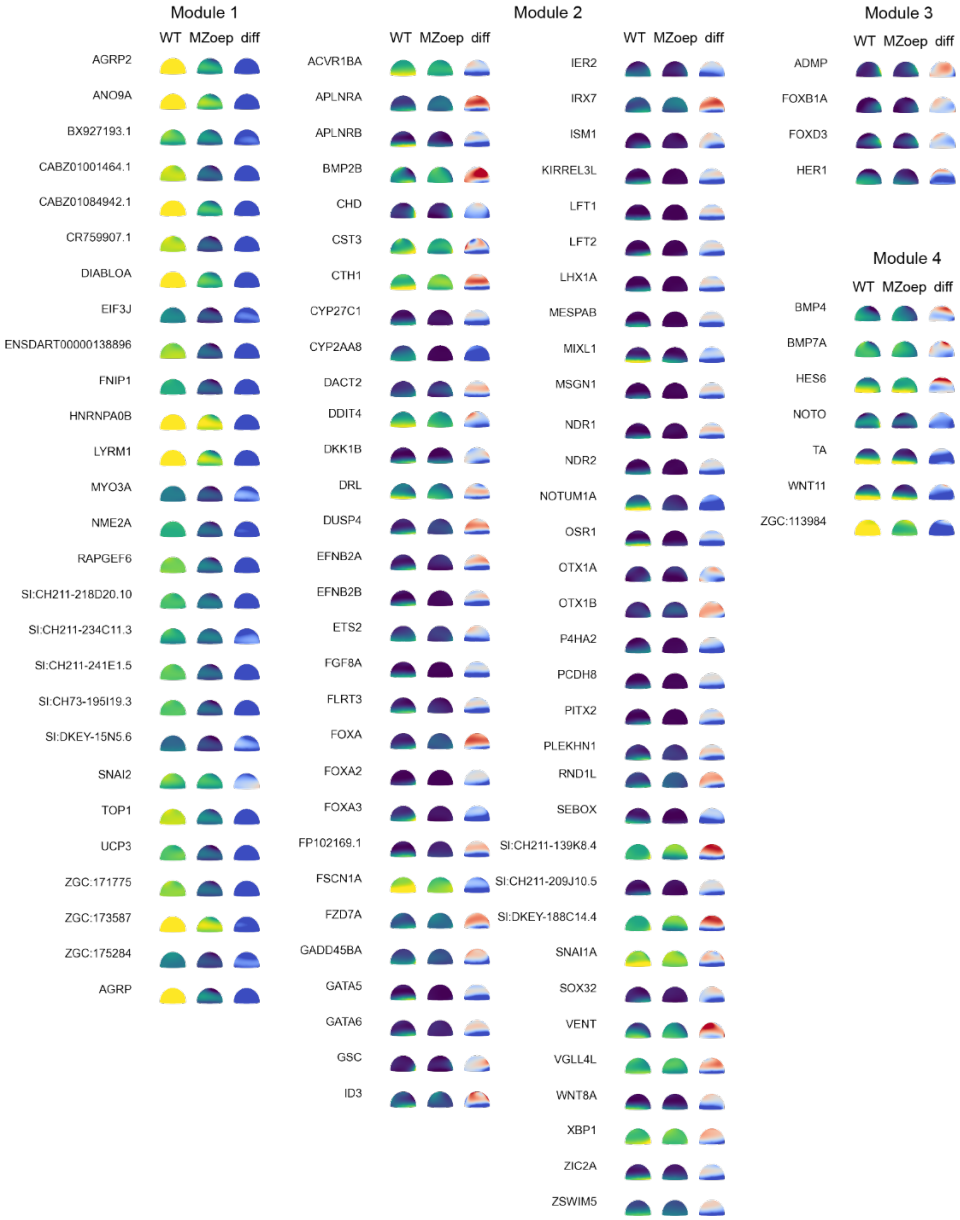
Extended Data Figure 2: latent data calibration in wildtype prediction

Scatter plot of the calibrated distributions of spatial landmark points and scRNA-seq data points (corresponding to Fig.2b) for the prediction of (a) wildtype mouse OB spatial transcriptomes and of (b) wildtype zebrafish embryo spatial transcriptomes. Principal component analysis was used to visualize the latent space.



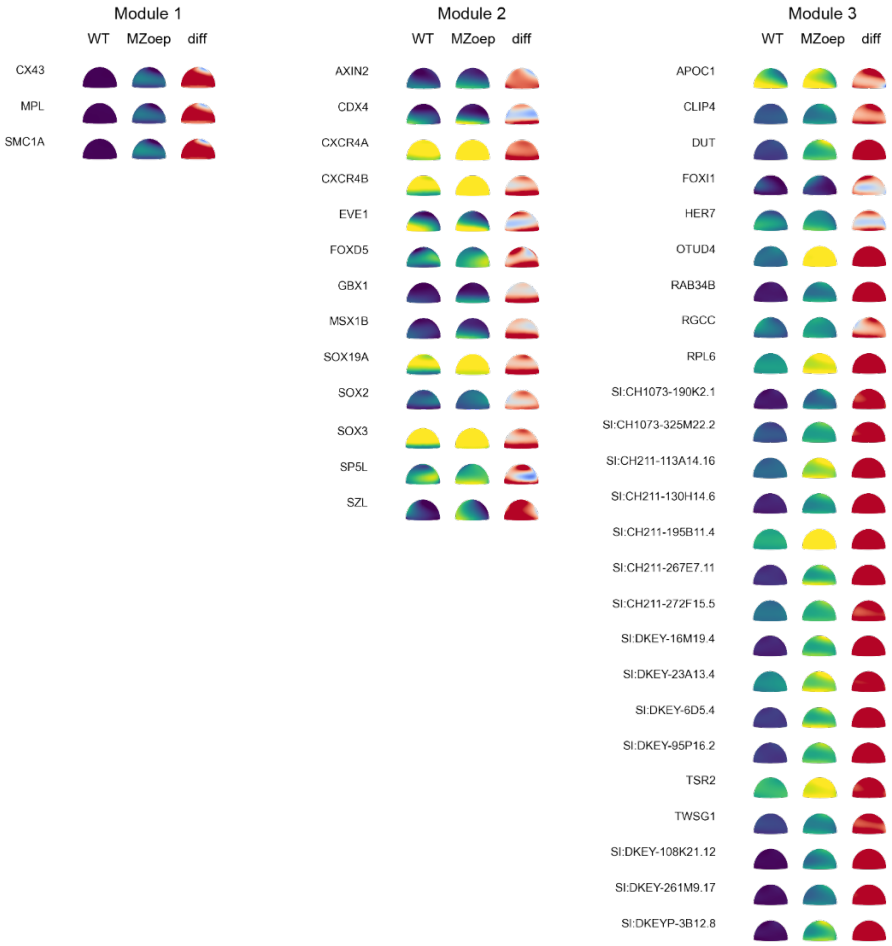
Extended Data Figure 3: Hierarchical clustering of the putative NU genes

(a) The hierarchical clustering of the putative NU genes (corresponding to Fig. 5d). The heatmap indicates the correlation of gene expression changes among the newly-screened NU genes.



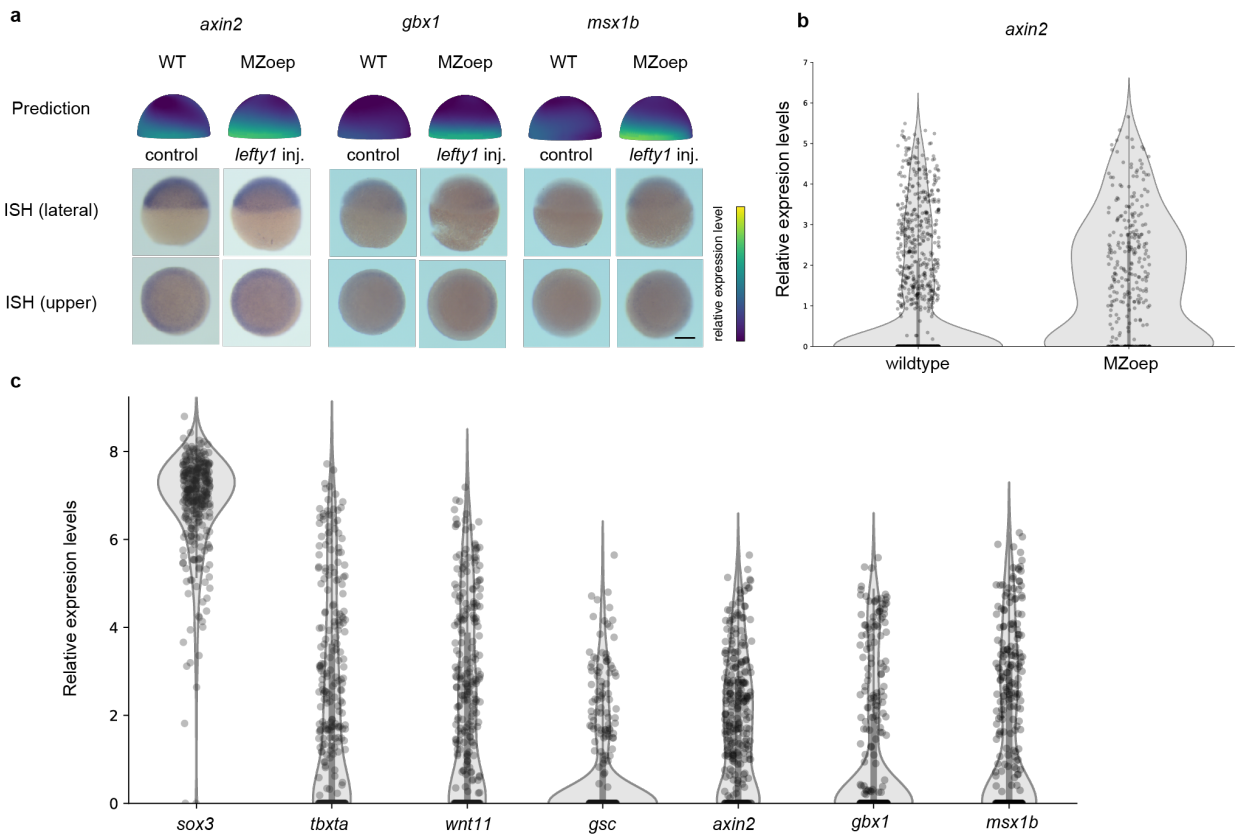
Extended Data Figure 4: All putative NU genes

The predicted spatial gene expressions of wildtype and MZoep zebrafish embryos for all putative NU genes. “diff” indicates the expression difference between MZoep and wildtype spatial transcriptomes.



Extended Data Figure 5: All putative ND genes

The predicted spatial gene expressions of wildtype and MZoep zebrafish embryos for all putative ND genes. “diff” indicates the expression difference between MZoep and wildtype spatial transcriptomes.



Extended Data Figure 6: The putative ND genes that are not consistent with ISH experiments
(a) The whole-mount ISH experiments of *axin2*, *gbx1*, and *msx1b*. Top to bottom: ZENomix's predicted expression pattern, a lateral view of the ISH, and an upper view of the ISH. For each gene, wildtype and mutant expression patterns are displayed (ZENomix's prediction: wildtype and *MZoep*-mutant embryos; ISH experiment: control and *lefty1*-injected embryos). Scale bar, 200 µm. (b) Violin Plot of the difference of *axin2*-expression levels between wildtype and MZoep scRNA-seq data. In scRNA-seq data, MZoep embryos had higher *axin2*-expression levels than wildtype embryos. (c) Violinplot of expression levels of *axin2*, *gbx1*, and *msx1b*. For reference, the expression levels of *sox3, tbxta, wnt11,* and *gsc* are also shown. *axin2*, *gbx1*, and *msx1b* are moderately expressed between the lowly expressed gene (*gsc*) and highly expressed genes (*sox3, tbxta*, and *wnt11*). Please see Figure 4 for the ISH images of *gsc*, *sox3, tbxta*, and *wnt11*.

650 # Supplementary Tables

NU genes							
Module 1	*AGRP2*	*ANO9A*	*BX927193.1*	*CABZ01001464.1*	*CABZ01084942.1*	*CR759907.1*	*DIABLOA*
	EIF3J	*ENSDART00000138896*	*FNIP1*	*HNRNPA0B*	*LYRM1*	*MYO3A*	*NME2A*
	RAPGEF6	*SI:CH211-218D20.10*	*SI:CH211-234C11.3*	*SI:CH211-241E1.5*	*SI:CH73-195I19.3*	*SI:DKEY-15N5.6*	*SNAI2*
	TOP1	*UCP3*	*ZGC:171775*	*ZGC:173587*	*ZGC:175284*	*AGRP*	
Module 2	*ACVR1BA*	*APLNRA*	*APLNRB*	*BMP2B*	*CHD*	*CST3*	*CTH1*
	CYP27C1	*CYP2AA8*	*DACT2*	*DDIT4*	*DKK1B*	*DRL*	*DUSP4*
	EFNB2A	*EFNB2B*	*ETS2*	*FGF8A*	*FLRT3*	*FOXA*	*FOXA2*
	FOXA3	*FP102169.1*	*FSCN1A*	*FZD7A*	*GADD45BA*	*GATA5*	*GATA6*
	GSC	*ID3*	*IER2*	*IRX7*	*ISM1*	*KIRREL3L*	*LFT1*
	LFT2	*LHX1A*	*MESPAB*	*MIXL1*	*MSGN1*	*NDR1*	*NDR2*
	NOTUM1A	*OSR1*	*OTX1A*	*OTX1B*	*P4HA2*	*PCDH8*	*PITX2*
	PLEKHN1	*RND1L*	*SEBOX*	*SI:CH211-139K8.4*	*SI:CH211-209J10.5*	*SI:DKEY-188C14.4*	*SNAI1A*
	SOX32	*VENT*	*VGLL4L*	*WNT8A*	*XBP1*	*ZIC2A*	*ZSWIM5*
Module 3	*ADMP*	*FOXB1A*	*FOXD3*	*HER1*			
Module 4	*BMP4*	*BMP7A*	*HES6*	*NOTO*	*TA*	*WNT11*	*ZGC:113984*
ND genes							
Module 1	*CX43*	*MPL*	*SMC1A*				
Module 2	*AXIN2*	*CDX4*	*CXCR4A*	*CXCR4B*	*EVE1*	*FOXD5*	*GBX1*
	MSX1B	*SOX19A*	*SOX2*	*SOX3*	*SP5L*	*SZL*	
Module 3	*APOC1*	*CLIP4*	*DUT*	*FOXI1*	*HER7*	*OTUD4*	*RAB34B*
	RGCC	*RPL6*	*SI:CH1073-190K2.1*	*SI:CH1073-325M22.2*	*SI:CH211-113A14.16*	*SI:CH211-130H14.6*	*SI:CH211-195B11.4*
	SI:CH211-267E7.11	*SI:CH211-272F15.5*	*SI:DKEY-16M19.4*	*SI:DKEY-23A13.4*	*SI:DKEY-6D5.4*	*SI:DKEY-95P16.2*	*TSR2*
	TWSG1	*SI:DKEY-108K21.12*	*SI:DKEY-261M9.17*	*SI:DKEYP-3B12.8*			

651 **Supplementary Table 1: The putative NU and ND genes**

652

	p	q	The numbers of Inducing points	Initial values of s^2	Initial values of σ^2	Initial values of σ^f_k	σ^f_l
AD-mutant Mouse OB (Figure 3)	62	30	40	1	0.1	0.01	0.01
Wildtype Mouse OB (Figure 3)	62	30	40	1	0.1	0.01	0.01
MZoep Zebrafish (Figure 4, 5, and 6)	47	20	40	1	0.1	0.01	0.01
Wildtype Zebrafish (Figure 4, 5, and 6)	47	20	40	1	0.1	0.01	0.01

Supplementary Table 2: The parameter values used in this study